Mengxi Yi
Klaus Nordhausen *Editors*

Robust and Multivariate Statistical Methods

Festschrift in Honor of David E. Tyler

Springer

Robust and Multivariate Statistical Methods

David E Tyler 1986 in Paris and 2015 in Myrtle Beach Published with the kind permission of
© Coleen Tyler, 2022. All Rights Reserved.

Mengxi Yi • Klaus Nordhausen
Editors

Robust and Multivariate Statistical Methods

Festschrift in Honor of David E. Tyler

Springer

Editors
Mengxi Yi
School of Statistics
Beijing Normal University
Beijing, China

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

ISBN 978-3-031-22686-1 ISBN 978-3-031-22687-8 (eBook)
https://doi.org/10.1007/978-3-031-22687-8

Mathematics Subject Classification: 62Hxx, 62F35, 62E20, 62J07

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.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

It is an honor and a pleasure to contribute this biographical note to the Festschrift dedicated to David (Dave) Tyler on the occasion of his pending retirement from Rutgers University.

I met Dave in the late 1980s through our mutual friend and colleague Javier Cabrera. Our frequent conversations regarding statistics and robustness led to the idea that workshops in robust statistics would be beneficial to the rapidly changing field of statistics. As a result, our first conference on *Robustness and Data Analysis* was held at Princeton University in 1994 with an outstanding list of invited speakers that included John Tukey and Frank Hampel. The success of this first conference encouraged us to continue with international meetings over more than a decade, and Dave's contributions were vital to our creation of the International Conference on Robust Statistics (ICORS) workshops. Today, with the effort and guidance of a younger generation of statisticians, these ICORS workshops continue to thrive.

But Dave's involvement in these conferences went well beyond the usual organizational stage. His deep understanding of statistical issues and his conviction that robust statistics is not merely a subfield of statistics, but rather a school of thought motivated by the realities of data analysis, provided a clear and firm foundation for these meetings. In every conference, Dave interacted extensively with the participants by exchanging ideas and engaging in discussions for the future direction of robust statistics as well as by stressing the importance to further integrate the concept of robustness into the everyday practice of data analysis.

Dave's insight and depth of understanding regarding statistical issues is attested by his many high-quality research publications. From his first paper, "Asymptotic theory of eigenvectors," published in the *Annals of Statistics* in 1981, to his latest research work on robust covariance matrices, his broad knowledge and independent approach are reflected in the long list of research papers published in the most prestigious statistical journals. Taken together, these papers illustrate the wide range of Dave's interests and his continuing influence in statistics, particularly in the areas of multivariate statistics and robustness. His international academic reputation in these areas is evidenced by his countless invitations to conferences, seminars, and special courses at major academic institutions worldwide. His distinguished academic credentials led to appointments of associate editor at some of the most prestigious statistical journals, including the *Annals of Statistics,* the *Journal of the Royal Statistical Society B,* and the *Journal of Multivariate Analysis.*

Dave's supportive and collaborative attitude toward his colleagues and his deep statistical knowledge earned him the respect and appreciation of major international researchers in statistics, as is documented by his joint publications with distinguished scholars around the world. In addition, his generosity with his ideas is reflected in the numerous doctoral students that he inspired and supervised with their PhD dissertations. The contributors of this Festschrift, many of whom had the privilege to have worked with Dave and to have benefitted from his knowledge as well as his company, enthusiastically offered their research papers for this volume, and this constitutes a testimony of their appreciation and admiration for his stature as a scholar and for him as a person.

Dave's childhood and adolescence merit comment. From the third oldest child of an impoverished family with ten children to a PhD in statistics from Princeton University to a Distinguished Professorship at Rutgers University, Dave's life and career have been quite unusual. He is the only member of his family to have achieved an advanced university degree.

Born and raised in Pittsburgh, Pennsylvania, Dave spent his early school years in Catholic schools, where he developed an early interest in mathematics. When he was 11 years old, due to family circumstances, he and his siblings were sent to a Catholic orphanage, and he remained there for two years. While poverty was a defining state, his mother provided a stabilizing influence in the family. After graduating from an urban public high school, Dave was admitted to Indiana University of Pennsylvania, where he earned a BA in mathematics in 1972 and where later, in 2015, he was awarded a Distinguished Alumni Award for his career accomplishments. In 1972, he continued with graduate studies at the University of Massachusetts, Amherst, and earned a Master's degree from the Mathematics Department in 1974. That same year, at age 24, he married Coleen McCullough, an aspiring young artist of similar religious and social background. Dave then pursued a doctoral program at Princeton University, where in 1979 he was awarded a PhD in statistics. After Princeton, he served as Assistant Professor at the University of Florida (1978,1979) and Old Dominion University (1979–1983). Finally, in 1983 he joined the Statistics Department of Rutgers University, where he was named Distinguished Professor of Statistics in 2004.

Among Dave's striking personal characteristics are modesty, humanity, and total honesty. This was evident not only during my work with him, but also in the social setting, where I met Coleen, an accomplished artist with whom I had a lasting friendship, and his son, Ed. The interaction between our families made me appreciate Dave's human dimension in addition to his outstanding scholarship. Moreover, firmly bound to his modest origins but dedicated to the field of statistics, Dave complemented his colorful personality with numerous interests and activities such as swimming, basketball, chess, hiking, biking, and boating, among others.

On the occasion of his pending retirement from Rutgers University, the institution where he spent most of his career, I wish Dave many more productive years and I look forward to enjoying the pleasure of his professional and personal company for many more years to come.

Philadelphia, PA, USA
June 2022

Luisa Fernholz

Preface

We are honoured and delighted to edit this Festschrift dedicated to David (Dave) E. Tyler, Distinguished Professor of Statistics at Rutgers University. The idea for this Festschrift was born around the occasion of Dave's 70th birthday and coming retirement from Rutgers to celebrate his outstanding career with many significant contributions to the field of statistics, especially in the areas of multivariate and robust methods.

Dave has a remarkable research career, which he started in 1978, after obtaining his PhD from Princeton University, as an assistant professor at the University of Florida. Via the Old Dominion University, he came in 1983 to Rutgers University, where he currently is a distinguished professor in statistics. In 1994, Dave was elected as an IMS fellow for his distinctive contributions in statistics regarding his independent work on M-estimation of scatter. In particular, most of his work was supported by various grants from, e.g. National Science Foundation (NSF). Dave has a reliable intuition and ability to identify interesting and challenging research questions which are of general importance and relevance. Then he develops his ideas in an insightful as well as rigorous manner addressing all possible details. His attention to detail, while keeping an eye on the big picture and the relevant questions, has been passed on to early career stage researchers, with whom he collaborated and mentored. It is therefore also no surprise that all seven PhD students of Dave embarked on their careers in academia, most of whom are now associate and full professors at universities around the world.

Contributed by Dave's students, friends, coauthors and colleagues, this book includes 22 peer-reviewed papers. The topics of the contributions are mainly motivated by the research interests of Dave. Accordingly, the book consists of four parts. Part I begins with an analysis of Dave's publication and coauthor networks, followed by a review article on Dave's famous *Tyler's shape estimator*. Parts II and III, as the main body of this book, cover some recent advances in multivariate and robust methods. The final part, Part IV, includes some various other topics such as supervised learning and normal extremes.

Speaking of these cutting-edge articles, we would like to express our gratitude to the efforts and patience of all contributors in the publishing process, especially because of the Covid-19 pandemic that disrupted most contributors' routine of work. Despite of those disruptions, upon joint work of authors and referees, we have reached a milestone with very interesting papers. We would like to thank therefore all contributors, who submitted their original and high-quality work to this Festschrift for Dave, and the referees, without whose generous help we would not have made it in time, given the tight schedule. We would like to thank also Veronika Rosteck and Daniel Ignatius from Springer who provided help and assistance whenever needed.

Finally, we want to salute Dave again for his intellectual contributions as well as his help as a mentor and as a friend. May Dave stay healthy and continue advancing the knowledge and boundaries of statistics!

Beijing, China
Jyväskylä, Finland
July 2022

Mengxi Yi
Klaus Nordhausen

Acknowledgements

The Editors would like to thank all the following referees for their excellent work:

Suchin Aeron
Andreas Alfons
Aurore Archimbaud
Andreas Artemiou
Marco Avella
James O. Berger
Germain van Bever
Ana Bianco
Eva Cantoni
Raymond Carroll
Raymond Chambers
Christophe Croux
Lutz Dümbgen
Kamila Facevicova
Hank Flury
Dominique Fourdeinier
Marc Hallin
Zifei Han
Jana Jureckova
Shogo Kato
John Kent
Arnab Kumar Laha
Gaorong Li
Chuan Liu
Markus Matilatinen
Peter McCullagh
Jaakko Nevalainen
Hannu Oja
Esa Ollila
Davy Paindaveine
Una Radojicic
Peter Rousseeuw
Gerta Rucker
Marcelo Ruiz
Anne Ruiz-Gazen
Richard L. Smith
Sara Taskinen
Beavers Traymon
Daniel Vogel
Changxi Wang
Christian Weiss
Ami Wiesel
Ines Wilms
Victor Yohai
Teng Zhang
Ting Zhang
Guang Zhu

Contents

Part I About David E. Tyler's Publications

Part II Multivariate Theory and Methods

Part III Robust Theory and Methods

Part IV Other Methods

List of Contributors

Ana M. Bianco Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina

Gracila Boente Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina

Howard Bondell School of Mathematics and Statistics, University of Melbourne, Parkville, Australia

Javier Cabrera Department of Statistics, Rutgers University, Piscataway, NJ, USA

Gonzalo Chebi Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina

Xinyan Chen Department of Statistics, Rutgers University, Piscataway, NJ, USA

Kenneth Edward Cherasia Department of Statistics, Rutgers University, Piscataway, NJ, USA

Lutz Dümbgen Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

Luisa T. Fernholz Department of Statistics, Temple University, Philadelphia, PA, USA

Robert Fernholz Intech Corp., Princeton, NJ, USA

Daniel Fischer Applied Statistical Methods, Natural Resources Institute Finland (Luke), Jokioinen, Finland

Gabriel Frahm Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany

Katrin Gysel SAKK Kompetenzzentrum, Bern, Switzerland

Marc Hallin ECARES and Département de Mathématique, Université libre de Bruxelles, Brussels, Belgium

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

Hyon-Jung Kim Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

John Kolassa Department of Statistics, Rutgers University, Piscataway, NJ, USA

Manuel Koller Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
Seminar für Statistik, ETH Zürich, Zürich, Switzerland

Ginette Lafit Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium

Thibault Laurent Toulouse School of Economics, French National Centre for Scientific Research, Toulouse, France

Peter Meer Department of Electrical Engineering, Rutgers University, Piscataway, NJ, USA

Camille Mondon Département de Mathématiques et Applications, Ecole Normale Supérieure, Paris, France

Gilles Mordant Institute for Mathematical Stochastics, Universität Göttingen, Göttingen, Germany

Javier Nogales Department of Statistics, Universidad Carlos III de Madrid, Getafe, Spain

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

Hannu Oja Department of Mathematics and Statistics, University of Turku, Turku, Finland

Esa Ollila School of Electrical Engineering, Aalto University, Espoo, Finland

Davy Paindaveine ECARES and Mathematics Department, Université libre de Bruxelles, Brussels, Belgium

Daniel Pena Department of Statistics, Universidad Carlos III de Madrid, Getafe, Spain

Fabrice Perler Bundesamt für Gesundheit BAG, Leibefeld, Switzerland

Setareh Ranjbar Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland

Elvezio Ronchetti Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland

Marcelo Ruiz Department of Mathematics, Universidad Nacional de Río Cuarto, Río Cuarto, Argentina

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

Yodit Seifu Bristol-Myers Squibb, Berkeley Heights, NJ, USA

Stefan Sperlich Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland

Werner A. Stahel Seminar für Statistik, ETH Zürich, Zürich, Switzerland

David S. Stoffer Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA

William E. Strawderman Department of Statistics, Rutgers University, Piscataway, NJ, USA

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

Christine Thomas-Agnan Toulouse School of Economics, University of Toulouse 1 Capitole, Toulouse, France

Thomas Verdebout Mathematics Department, Université libre de Bruxelles, Brussels, Belgium

Daniel Vogel MEDICE Arzneimittel Pütter GmbH & Co. KG, Iserlohn, Germany
Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, UK

Stuart J. Watt Mirador Analytics, Melrose, UK

Anna Wiedemann Department of Psychiatry, University of Cambridge, Cambridge, UK

Han Xiao Department of Statistics, Rutgers University, Piscataway, NJ, USA

Mengxi Yi School of Statistics, Beijing Normal University, Beijing, China

Victor J. Yohai Department of Mathematics and Instituto de Calculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

Weichang Yu School of Mathematics and Statistics, University of Melbourne, Parkville, Australia

Ruben Zamar Department of Statistics, University of British Columbia, Vancouver, BC, Canada

Dewei Zhong Anheuser-Busch Companies, New York, NY, USA

Yijun Zhu Department of Statistics, Rutgers University, Piscataway, NJ, USA

Stavros Zinonos Cardiovascular Institute of New Jersey, RWJMS, New Brunswick, NJ, USA

Part I
About David E. Tyler’s Publications

An Analysis of David E. Tyler's Publication and Coauthor Network

Daniel Fischer, Klaus Nordhausen, and Mengxi Yi

Abstract David E. Tyler can look back on an impressive career with many significant contributions to robust and multivariate statistical methods. In this paper we attempt to quantify his scientific impact by having a closer look at his publications and by analyzing his coauthor network.

Keywords Bibliography · Community detection

1 Introduction

David (Dave) E. Tyler is a driving force in the development of robust and multivariate statistical methods with an impressive publication record. In this article, we will give a brief overview of Dave's publications and analyze his coauthorship network as well. As Dave is still active in research, we anticipate and expect to see still many more significant contributions from him. Here, however, we consider only his publications until May 2022. By then, Dave has published 82 scientific papers[1] as listed in Appendix A.1, which could be roughly classified into statistical theory, methodology, application, and comments such as reviews and discussions.

[1] For the purpose of this paper we ignored Dave's applied papers resulting from consulting but included also methodological papers which are so far only available on Arxiv.

D. Fischer
Natural Resources Institute Finland (Luke), Applied Statistical Methods, Jokioinen, Finland
e-mail: Daniel.Fischer@luke.fi

K. Nordhausen
Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland
e-mail: klaus.k.nordhausen@jyu.fi

M. Yi (✉)
School of Statistics, Beijing Normal University, Beijing, China
e-mail: mxyi@bnu.edu.cn

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_1

Below, we refer to these works using the numbers provided in the Appendix and all the citation data are based on the Web of Science on 5.5.2022 (URL: http://apps.webofknowledge.com). Citation details were downloaded from Semantic Scholar (URL: https://www.semanticscholar.org).

Dave is an expert in robust statistics, especially in M-estimation, having also significant influence in other areas such as signal processing. His most cited work [10] has been cited 380 times, which is rather high considering the general low frequency in citations in the area of statistics. In this paper, Dave proposes a new M-estimator of scatter which is nowadays quite popular and often referred to as Tyler's M-estimator or Tyler's shape matrix and it is still being actively studied in statistics and signal processing and, for example, reviewed in Taskinen et al. (2023).

Meanwhile, Dave has contributed to the field of multivariate analysis, directional data analysis, spectral analysis of time series, and functional data analysis. For instance, his most cited methodological work [58] (with 76 citations) suggested the invariant co-ordinate selection (ICS) procedure to better explore multivariate data. An R package developed accordingly introduces this method to a broader audience; see [54] and the [R1] in Appendix A.2, which lists the R packages where Dave is involved in. Methodologically applied in the area of psychometrics, computer vision, and signal processing, Dave's work has become more and more appreciated beyond the community of statistics. Dave's contribution to the academia also embodies in some review and discussion papers that gain significant attention from peer researchers, among which the most cited work [63] has been referred 261 times so far.

Dave obtained his PhD from Princeton University 1979 for his dissertation entitled "Redundancy Analysis and Associated Asymptotic Distribution Theory" supervised by Lawrence S. Mayer. This makes Dave an academic descendant of various famous statisticians who worked among others on multivariate methods and robust nonparametric methods, topics which Dave developed further in his career. A pruned version of Dave's academic genealogical tree is given in Fig. 1 which lists also the seven students who Dave supervised so far.

As Dave's academic life has been devoted to the development of statistical theory and methodology, most of his works are published in highly ranked statistical journals including *The Annals of Statistics, Biometrika, Journal of the Royal Statistical Society Series B,* and *Journal of Multivariate Analysis*. Based on the titles and abstracts of the publications considered here we provide in Fig. 2 the corresponding word cloud, see, e.g., Seifert et al. (2008), which shows the most frequent 100 words after removing the standard stop words of the English language as defined by the package `tm` and the words *abstract, keywords, can, also,* and *given*. As the font size and color reflects the frequency of a word, Fig. 2 shows clearly that Dave's research interest centers around multivariate data and scatter matrices and emphasizes the robust aspect, such as the breakdown point. Theoretically, Dave considers especially asymptotic properties and the distribution of estimators. Notice also that Dave's publication record shows surprisingly consistency on the studies of M-estimation of scatter matrices, where Dave has published numerous single authored papers and left a strong influence on statistics and signal processing till now and beyond.

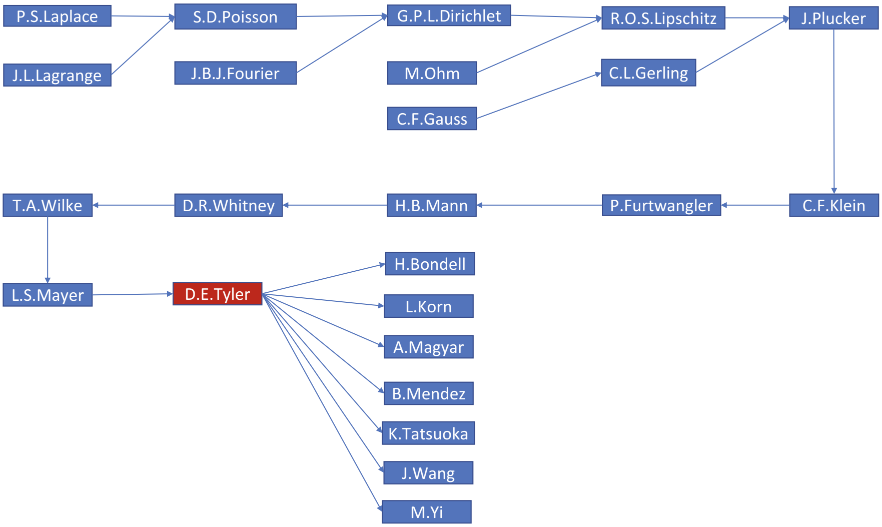


Fig. 1 Extract of Dave's genealogical tree including his seven academic children

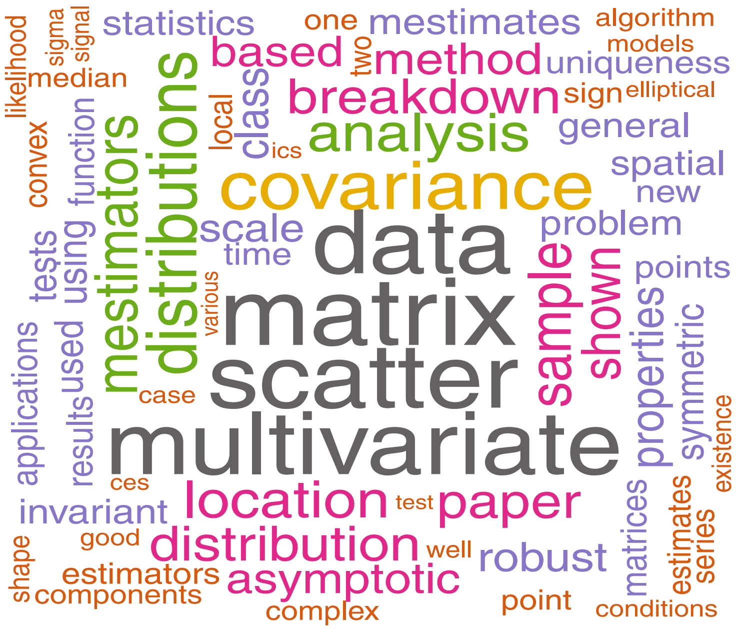


Fig. 2 Word cloud based on the abstracts and titles of Dave's publications

In the following sections, we apply community detection methods to investigate Dave's coauthorship networks and the impact of Dave's work in the community.The analysis is done in R (R Core Team 2022), using the R packages `bibtex` (Francois 2014), `igraph` (Csardi & Nepusz 2006; Kolaczyk & Csardi 2014), `circlize` (Gu 2014), `wordcloud` (Fellows 2018), `tm` (Feinerer & Hornik 2020), and `rworldmap` (South 2011).

2 David Tyler's Coauthor Network

We first review some basic facts of network theory. A network graph $G = (V, E)$ consists of a set V of vertices or nodes and a set E of edges or links. The number of vertices $n = |V|$ is the *order* of the network G and the number of edges $m = |E|$ is its *size*. Two vertices are *neighbors* or *adjacent* if they are connected by an edge. Networks are *undirected* if there is no ordering in the vertices defining an edge and are *weighted* if a real number is associated with each of the edges. If vertices are not allowed to be connected to themselves, the graph G is called a *simple* graph. A network can be partitioned into several subgraphs, where $G_r = (V_r, E_r)$ is called a *subgraph* of $G = (V, E)$ if $V_r \subset V$ and $E_r \subset E$.

One of the most important subgraphs is the *egocentric* network. This is a network created by selecting an ego-node and all of its connections. First, we are interested to build Dave's direct-coauthor network $G = (V, E)$ based on his publications [1]–[82]. Thus, Dave is the ego vertex, his direct coauthors are the other vertices, and two distinct authors are connected by an edge if they have written one joint paper with Dave. Table 1 presents Dave's collaboration frequencies, which shows that Dave has one collaborator with whom he has written 8 papers. Dave has 21 single authored papers that do not contribute to the network. In the remaining 61 publications, Dave has 51 coauthors.

The network of Dave's direct coauthors is visualized in Fig. 3. Here, no information about how his coauthors work together with him was used, it rather visualizes Dave's blossoming levels of collaboration, as each joint publication between coauthors is visualized with an own edge, the closer therefore an coauthor is in Fig. 3 to Dave, the more often he/she collaborated with him. This indicates that his inner circle of coauthors with more than 5 joint papers consists of L. Dümbgen, P. Meer, K.Nordhausen, H. Oja, and E. Ollila; the first four coauthors have published 6 papers with Dave and the last author has published 8 papers.

Table 1 Number of times Dave collaborated with coauthors for his publications. The value zero corresponds to single author papers

Number of joint papers	0	1	2	3	5	6	8
Frequency	21	29	11	5	1	4	1

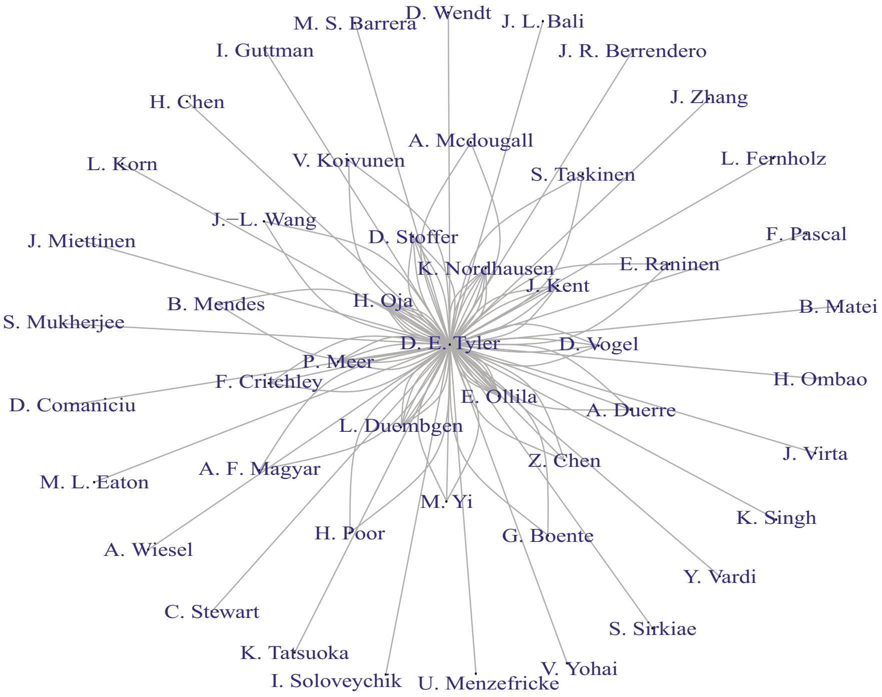


Fig. 3 Network of Dave's direct coauthors. Each connection corresponds to a collaboration for a joint paper

Next, we review some results on community detection methods. As the whole information about a network can be stored in matrix form, we could define the $n \times n$ *adjacency matrix* $\boldsymbol{A}$, for a network $G = (V, E)$, as follows

$$A_{ij} = \begin{cases} 1, & \text{if } \{i, j\} \in E, \\ 0, & \text{otherwise.} \end{cases}$$

For a simple network, the diagonal elements of the adjacency matrix are all zero. And the matrix $\boldsymbol{A}$ will be symmetric for undirected networks. If G is a weighted network, then A_{ij} represents the weight of the edge between i and j. Note also that the *degree* k_i of a vertex i, i.e., the number of its neighbors, can be given by $k_i = \sum_{j=1}^{n} A_{ij}$.

In order to detect significant community structure, or to identify good partitions of a network, it is useful to have a quality function to assess the goodness of a graph partition. In this way, the largest number given by the quality function means the partition is best. One of the most popular quality function is the *modularity* used in Newman (2006). It is based on the idea of finding divisions of the network in which the actual minus the expected number of edges over all pairs of vertices that belong to the same cluster is highest. Let c_i be the cluster or community to which vertex i belongs, the modularity is defined as

$$Q = \frac{1}{2m} \sum_{i}^{n} \sum_{j}^{n} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

where m is the total number of edges of the network and $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and 0 otherwise. The goal is then to decide the optimal number of partitions and to label the vertices by maximizing the modularity. Many methods have been suggested, in the literature, for this optimization problem; see, for example, in Fortunato (2010) for an overview. In the following we will describe the *multi-level modularity optimization algorithm* of Blondel et. al. (2008).

The algorithm consists of two phases. Consider initially to assign a different community to each node of the network. Then put each node to the community for which it gain maximum of the modularity. Repeat this process until no further improvement can be made. The second phase starts by regarding each community, found in the first phase, as a vertex and builds the network based on these nodes and links. The process stops until there are no more changes or a maximum of the modularity is attained.

By using the above-described community detection method, we build Dave's community graph, Fig. 4, based on Dave's direct coauthors. Here, in addition to the Dave's publication information, we also included relevant all coauthor publications that contribute an edge to the network to get the correct weights for the edges between the different coauthors. That means, the network visualizes the connections between the coauthors based on all joint papers and not only based on joint papers with Dave. In total we can identify eight different communities within this network; see Table 2. For instance, Community 7 corresponds to Dave's work on robust functional methods, Community 6 to Hannu Oja's group, Community 4 to computer vision groups, and Community 5 on his work in signal processing.

After considering the network of Dave's direct coauthors, we extend our search to include as well the coauthors from the coauthors. Also in this network, we take the direct connections between coauthors into consideration, so that we had to look at Dave's peers that are even three nodes away. For that extensive search Semantic Scholar granted us a API access and the search resulted in 495,864 peers with 253.5 million connections in total. For this network, we filter then to peers that are two levels away only; see Fig. 5. Here, we visualize $n = 2755$ nodes and $m = 364{,}469$ edges. And Table 3 lists the 5 most influential authors in each community. It becomes obvious that Dave is not part of a tiny community but has via his connections a huge reach into the scientific community. This is indicated that many of the communities detected are not directly linked. Crude interpretations for some of communities are possible. Community 3 could be dependent data like time series, Community 5 the British school of statistics and Community 15 multivariate nonparametric statistics while Community 12 could be summarized as signal processing.

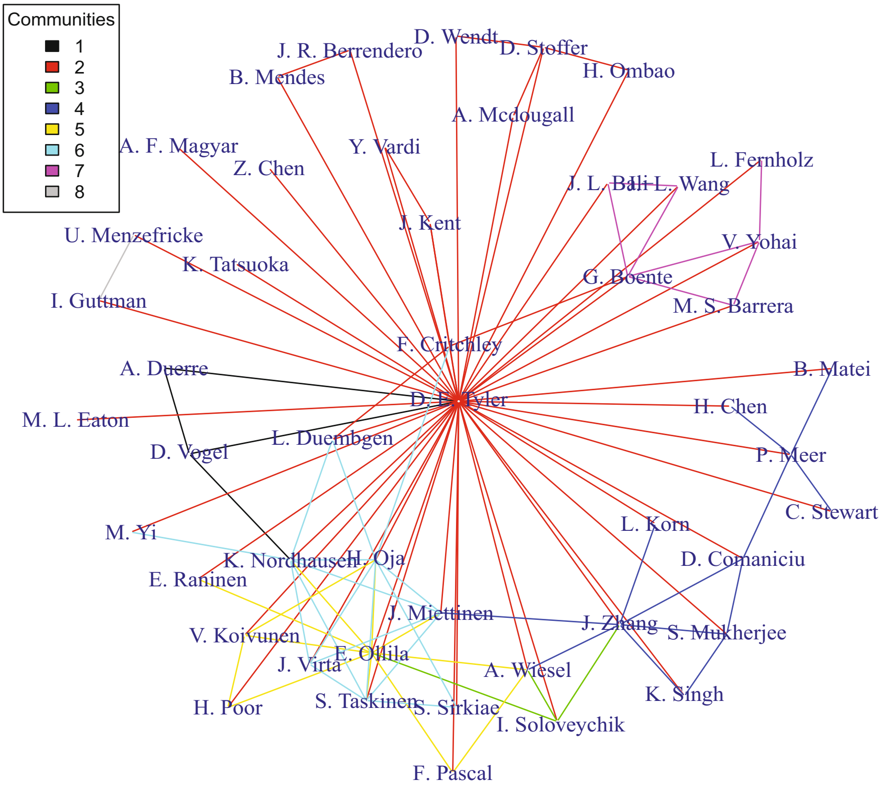


Fig. 4 Community Network of Dave’s direct coauthors, where members belonging to the same community are connected by edges in the same color

Table 2 Communities detected by using Dave’s direct coauthors. The table below lists the most prominent members of the communities

Community	1	2	3	4	5
1	D. Vogel	A. Duerre			
2	D. E. Tyler	D. Stoffer	J. Kent	F. Critchley	A. Mcdougall
3	I. Soloveychik	A. Wiesel			
4	J. Zhang	P. Meer	D. Comaniciu	S. Mukherjee	K. Singh
5	E. Ollila	V. Koivunen	H. Poor	F. Pascal	E. Raninen
6	H. Oja	S. Taskinen	K. Nordhausen	J. Miettinen	J. Virta
7	G. Boente	V. Yohai	M. S. Barrera	J. L. Bali	J.-L. Wang
8	I. Guttman	U. Menzefricke			

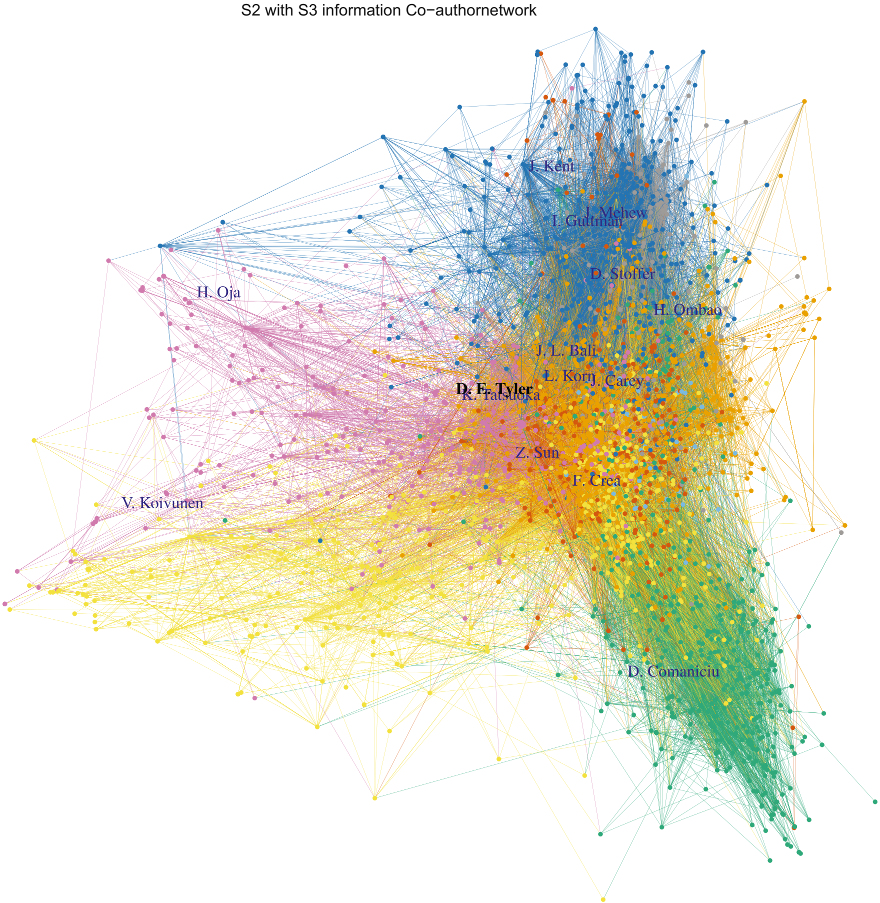


Fig. 5 Community Network of Dave's coauthors' coauthors. Top representatives from each community are highlighted with names

3 David Tyler's Influence

When considering the network spanned by Dave's coauthors' coauthors one can suspect that Dave's ideas have a wide reach. While it is quite a challenge to measure a researcher's impact we make an attempt by looking at the citations Dave's work got in Semantic Scholar. Based on this data, Dave's papers received in total 3739 citations from 2993 different authors. Hereby, his most citing peers are F. Pascal (184), K. Nordhausen (159), H. Oja (158), A. Breloy (99), D. Paindaveine (99), and E. Ollila (96). Further, his citations originate from 433 different journals, with *IEEE Transactions on Signal Processing* (179), *Journal of Multivariate Analysis* (146), *Signal Processing* (57), *Computational Statistics & Data Analysis* (55), and *IEEE Signal Processing Letters* (48) being the journals that contain the most citations

Table 3 Communities with more than 4 members detected by using Dave's coauthors–coauthors and giving their most prominent members

Community	Size	1	2	3	4	5
1	273	Z. Sun	Y. Chen	X. Luo	R. Sharma	Y. Pan
3	70	D. Stoffer	R. Fried	M. Thiel	R. Kliegl	M. Ding
4	223	F. Crea	K. Fox	H. Katus	S. Achenbach	S. Blankenberg
5	453	J. Kent	K. Mardia	F. Critchley	P. Diggle	C. Williams
6	86	L. Korn	S. Alimokhtari	S. Maberti	C. Weisel	A. Winer
7	70	J. L. Bali	L. Otero	E. Quel	P. Ristori	J. Salvador
8	47	I. Guttman	O. P. Aggarwal	D. Newman	H. Shapiro	M. Klamkin
9	393	H. Ombao	J. Williams	M. John	C. Brown	H. Zhao
10	35	J. Carey	P. Liedo	N. Papadopoulos	F. Molleman	B. Katsoyannos
11	343	D. Comaniciu	B. Georgescu	A. Kamen	P. Meer	J. Hornegger
12	241	V. Koivunen	F. Pascal	J. Ovarlez	A. Wiesel	H. Poor
14	163	K. Tatsuoka	C. Stewart	S. Adams	H. Martus	A. Olaharski
15	278	H. Oja	S. Taskinen	J. Miettinen	K. Nordhausen	L. Capranica
16	71	J. Mehew	S. Stehlíková	F. Gasperoni	F. J. G. Perez	V. Yohai

Table 4 Number of citations to Dave's work from different fields of sciences

Field	Citations	Field	Citation
Mathematics	2507	Psychology	9
Computer Science	1550	History	7
Medicine	145	Geology	6
Engineering	122	Art	4
Physics	53	Chemistry	3
Economics	45	Business	2
Biology	40	Political Science	2
Environmental Science	18	Sociology	2
Geography	16		
Materials Science	12		
Philosophy	12		

to Dave's work, which shows that his ideas are especially in signal processing of interest. Classifying these journals[2] into scientific fields, according to the system of Semantic Scholar, shows that these journals represent 19 main fields of study, ranging from Mathematics and Computer Science of Engineering, Biology and Physics towards Philosophy, History and Art; see Table 4 the frequencies. From Table 4, we find that Dave's most influential area is Mathematics, specifically in Statistics. Interesting to note is that his methods could also be applied in Art.

However, considering the huge amount of data that was collected to create these Figures, we relied heavily on an automatic data collection. Here it is possible that we missed for some authors a few publications, in case there is no consistent and traceable affiliation history available. Also, it might also happen that some wrong publications were assigned to authors based on a name mix-up.

A more detailed view of how Dave impacts the work of others could be revealed when we look at the keywords of the citing articles. For the most frequent ones ($\geq$20 occurrences), we create again another word cloud; see Fig. 6. Here, it is easily noticeable from the word cloud that statistical terms stand out.

While above we considered the total number of citations it is of course of interest to see how these papers distribute over Dave's 82 papers which are here under consideration. We visualize the corresponding information in a circos plot, a visualization type that is typically used in comparative genomics to show links between different chromosomes, see Krzywinski et al. (2009). In the circos plot given in Fig. 7 we therefore order the papers chronological and give a line from each paper to the year in which it was cited. As it is usual in mathematical sciences it usually takes some time before a paper gets cited. The blue lines in the figure correspond to citations of paper [10] where Dave introduced his famous shape matrix. The paper seems to have gotten increased interest starting from 2004 and interestingly then its popularity increased in three year cycles. Inquired which

[2] Note that one journal can be assigned by the system to several fields of science.

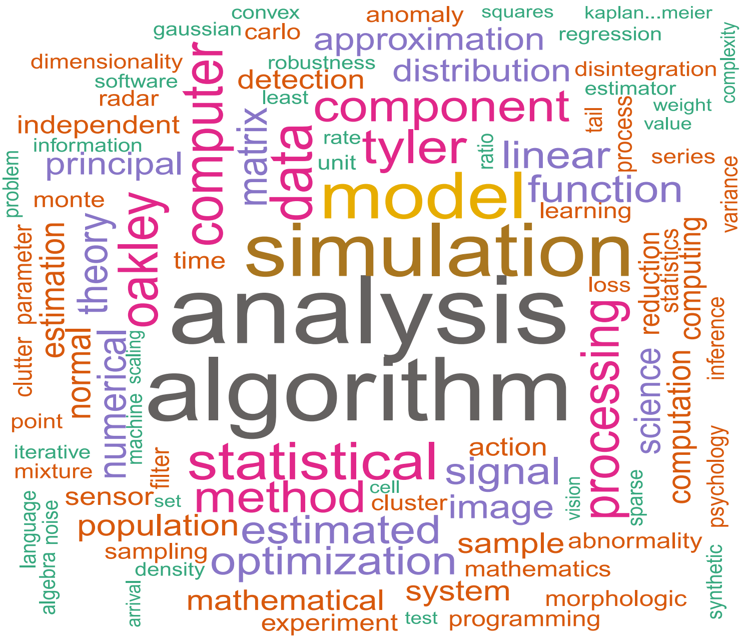


Fig. 6 Word cloud based on the keywords of articles citing Dave's papers

papers besides [10] have most important contributions, Dave listed his 10 papers [1, 5, 14, 18, 19, 24, 36, 58, 60, 63] which are marked in red in the figure. It is quite clear from the figure that all these papers have a steady impact over time but are less influential than the shape paper. Another feature of the figure is Dave's collaboration pattern, the comparison between collaborative papers and single authored ones. Interesting to see there is that while in the first 10 years of his career Dave mainly worked alone, after that almost all publications are made in collaborations. While this might be partly due to a shift in scientific conventions we believe it also reflects Dave's increased popularity and that Dave is an excellent collaborator willing to share with others his deep insight and many ideas as when looking of the list of collaborators one can see that many of them especially in the last 10 years are in much earlier career stages.

Also with the increase in collaborations Dave's output seems to have increased considerable since the mid-nineties. Notable in this context is also that when considering what we consider the current location of his coauthors that they are spread quite around the globe as visualized in Fig. 8.

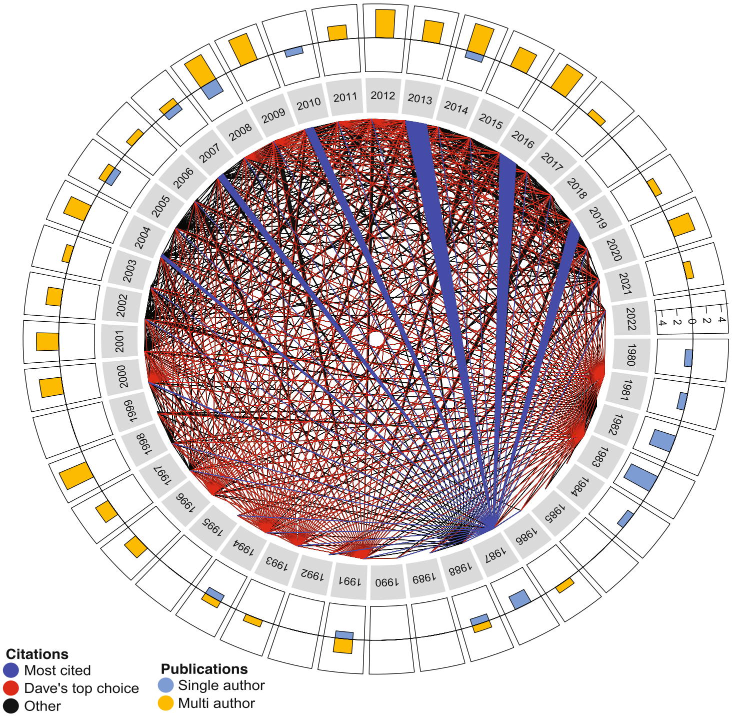


Fig. 7 Circos plot showing the influence of David's paper over time. The outer ring indicates the publications per year (blue are single author publications, yellow are multi-author papers). The red lines indicate citations of David's top 10 important papers and the blue lines are citations to Dave's shape matrix paper [10]

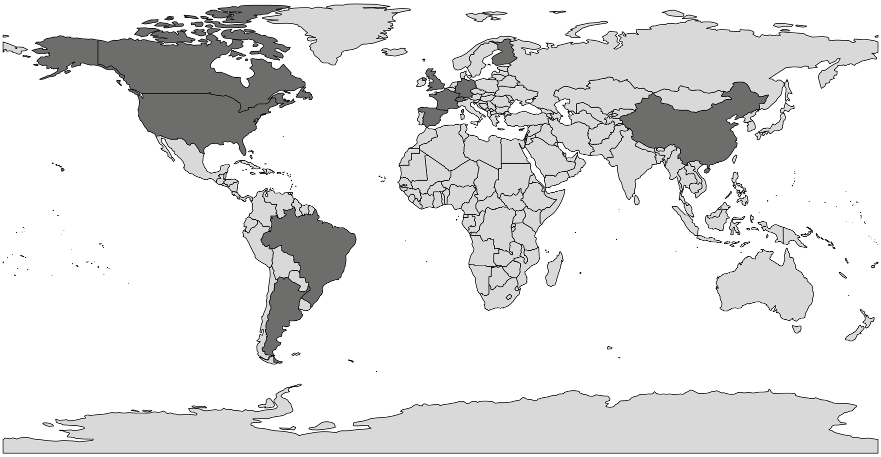


Fig. 8 Current location of Dave's coauthors, where the corresponding countries are in black

4 Concluding Remarks

Dave's academic pedigree raises high expectations and we think he has more than fulfilled them. Dave was and still is a driving force in multivariate and robust statistics with many important contributions to the field. While it is inherently difficult to measure scientific relevance and impact we attempted this by investigating Dave's network of coauthors and citation network. Our analysis shows Dave's wide interests having many different coauthors embedding him in large networks where his contributions are frequent and regularly cited. While it seems his research is most relevant in the fields of statistics and signal processing he nevertheless gets also many citations from other fields of science.

And we are happy to note that this analysis is just a snapshot and hopefully soon outdated as Dave is still active in research where his current interest is on M-estimation in high dimensions as, for example, his recent papers like [69,76,78,79,82] show and we are looking forward to read still many papers published by Dave!

Acknowledgments The work of Mengxi Yi is supported by the National Natural Science Foundation of China (No. 12101119). The authors are grateful for the comments given by the referees which helped to improve the paper.

Appendices

A.1 Publications of David E Tyler

1. David E Tyler. Asymptotic inference for eigenvectors. The Annals of Statistics, 9(4):725–736, 1981.
2. David E Tyler. Radial estimates and the test for sphericity. Biometrika, 69(2):429–436, 1982.
3. David E Tyler. A counterexample to Miller and Farr's algorithm for the index of redundancy. Multivariate Behavioral Research, 17(1):131–135, 1982.
4. David E Tyler. On the optimality of the simultaneous redundancy transformations. Psychometrika, 47(1):77–86, 1982.
5. David E Tyler. Robustness and efficiency properties of scatter matrices. Biometrika, 70(2):411–420, 1983.
6. David E Tyler. The asymptotic distribution of principal component roots under local alternatives to multiple roots. The Annals of Statistics, 11(4):1232–1242, 1983.
7. David E Tyler. A class of asymptotic tests for principal component vectors. The Annals of Statistics, 11(4):1243–1250, 1983.

8. Irwin Guttman, Ulrich Menzefricke, and David E Tyler. Magnitudinal effects in the normal multivariate model. The Annals of Statistics, 14(4):1555–1571, 1986.
9. David E Tyler. Breakdown properties of the M-estimators of multivariate scatter. ArXiv preprint arXiv:1406.4904, 2014, reprint of an unpublished 1986 Rutgers Technical Report.
10. David E Tyler. A distribution-free M-estimator of multivariate scatter. The Annals of Statistics, 15(1):234–251, 1987.
11. David E Tyler. Statistical analysis for the angular central Gaussian distribution on the sphere. Biometrika, 74(3):579–589, 1987.
12. David E Tyler. Some results on the existence, uniqueness, and computation of the M-estimates of multivariate location and scatter. SIAM Journal on Scientific and Statistical Computing, 9(2): 354–362, 1988.
13. John T Kent and David E Tyler. Maximum likelihood estimation for the wrapped Cauchy distribution. Journal of Applied Statistics, 15(2):247–254, 1988.
14. Morris L Eaton and David E Tyler. On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. The Annals of Statistics, 19(1): 260–271, 1991.
15. David E Tyler. Some issues in the robust estimation of multivariate location and scatter. In Directions in Robust Statistics and Diagnostics, pages 327–336. Springer, 1991.
16. John T Kent and David E Tyler. Redescending M-estimates of multivariate location and scatter. The Annals of Statistics, 19(4):2102–2119, 1991.
17. David S Stoffer, David E Tyler, Andrew J McDougall, and Gabriel Schachtel. Spectral analysis of DNA sequences. Bulletin of the International Statistical Institute, 49:345–361, 1993.
18. David S Stoffer, David E Tyler, and Andrew J McDougall. Spectral analysis for categorical time series: Scaling and the spectral envelope. Biometrika, 80(3):611–622, 1993.
19. David E Tyler. Finite sample breakdown points of projection based multivariate location and scatter statistics. The Annals of Statistics, 22(2):1024–1044, 1994.
20. Morris L Eaton and David E Tyler. The asymptotic distribution of singular-values with applications to canonical correlations and correspondence analysis. Journal of Multivariate Analysis, 50(2):238–264, 1994.
21. John T Kent, David E Tyler, and Yahuda Vardi. A curious likelihood identity for the multivariate t-distribution. Communications in Statistics – Simulation and Computation, 23(2):441–453, 1994.
22. David E Tyler. M-estimates, S-estimates and CM-estimates: A review. In IEEE Proceedings of NSF/AFPA Workshop: Performance versus Methodology in Computer Vision, pages 1–6. IEEE, 1994.
23. Beatriz VM Mendes and David E Tyler. Constrained M-estimation for regression. In Robust Statistics, Data Analysis, and Computer Intensive Methods, pages 299–320. Springer, 1996.

24. John T Kent and David E Tyler. Constrained M-estimation for multivariate location and scatter. The Annals of Statistics, 24(3):1346–1370, 1996.
25. David E Tyler. A more general framework for the EM algorithm? Discussion of the paper "The EM algorithm–an old folk-song sung to a fast new tune" by Xiaoli Meng and David van Dyk. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59(3):511–567, 1997.
26. Raymond Carroll, Luisa Fernholz, and David E Tyler. Robust statistics and data analysis. Journal of Statistical Planning and Inference, 1(57):3–4, 1997.
27. Andrew J McDougall, David S Stoffer, and David E Tyler. The spectral envelop for continuous-valued and multivariate time series. Journal of Statistical Planning and Inference, 57(2):195–214, 1996.
28. David S Stoffer and David E Tyler. Matching sequences: Cross-spectral analysis of categorical time series. Biometrika, 85(1):201–213, 1998.
29. Peter Meer and David E Tyler. Smoothing the gap between statistics and image understanding. Comments on the paper "Edge-preserving smoothers for image processing" by Chu CK, Glad IK, Godtliebsen F. and Marron JS. Journal of the American Statistical Association, 93:526–541, 1998.
30. Bogdan Matei, Peter Meer, and David E Tyler. Performance assessment by resampling: rigid motion estimators. In Empirical Evaluation Techniques in Computer Vision, pages 72–95. IEEE, 1998.
31. Beatriz VM Mendes and David E Tyler. Illustrating the behavior of CM-estimates of location and scale. Brazilian Journal of Probability and Statistics, 12:41–53, 1998.
32. David E Tyler. S-estimators. In Encyclopedia of Statistic Science, pages 659–662. Wiley, New York, 1999.
33. Dorin Comaniciu, Peter Meer, Kun Xu, and David E Tyler. Retrieval performance improvement through low rank corrections. In Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL'99), pages 50–54. IEEE, 1999.
34. Peter Meer, Charles V Stewart, and David E Tyler. Robust computer vision: An interdisciplinary challenge. Computer Vision and Image Understanding, 78(1):1–7, 2000.
35. David S Stoffer, David E Tyler, and David A Wendt. The spectral envelope and its applications. Statistical Science, 15(3):224–253, 2000.
36. Kay Tatsuoka and David E Tyler. The uniqueness of S and M-functionals under non-elliptical distributions. The Annals of Statistics, 28(4):1219–1243, 2000.
37. Leo R Korn and David E Tyler. Robust estimation for chemical concentration data subject to detection limits. In Statistics in Genetics and in the Environmental Sciences, pages 41–63. Springer, 2001.
38. John T Kent and David E Tyler. Regularity and uniqueness for constrained M-estimates and redescending M-estimates. The Annals of Statistics, 29(1):252–265, 2001.
39. Haifeng Chen, Peter Meer, and David E Tyler. Robust regression for data with multiple structures. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, pages I–I. IEEE, 2001.

40. David S Stoffer, Hernando C Ombao, and David E Tyler. Local spectral envelope: an approach using dyadic tree-based adaptive segmentation. Annals of the Institute of Statistical Mathematics, 54(1):201–223, 2002.
41. Zhiqiang Chen and David E Tyler. The influence function and maximum bias of Tukey's median. The Annals of Statistics, 30(6):1737–1759, 2002.
42. David E Tyler. High breakdown point multivariate M-estimation. Estadística, 54:213–247, 2002.
43. Dorin Comaniciu, Peter Meer, and David E Tyler. Dissimilarity computation through low rank corrections. Pattern Recognition Letters, 24(1–3):227–236, 2003.
44. Luisa Fernholz, David E Tyler, and Victor Yohai. Preface to "Contemporary data analysis: Theory and methods". Journal of Statistical Planning and Inference, 122:1–2, 2004.
45. Zhiqiang Chen and David E Tyler. On the behavior of Tukey's depth and median under symmetric stable distributions. Journal of Statistical Planning and Inference, 122(1–2):111–124, 2004.
46. Zhiqiang Chen and David E Tyler. On the finite sample breakdown points of redescending M-estimates of location. Statistics & Probability Letters, 69(3):233–242, 2004.
47. David E Tyler. Discussion of the paper "Breakdown and groups" by P Laurie Davies and Ursula Gathers. The Annals of Statistics, 33:1009–1015, 2005.
48. Lutz Dümbgen and David E Tyler. On the breakdown properties of some multivariate M-functionals. Scandinavian Journal of Statistics, 32(2):247–264, 2005.
49. Klaus Nordhausen, Hannu Oja, and David E Tyler. On the efficiency of invariant multivariate sign and rank tests. In Liski, E.P., Isotalo, J., Niemelä, J., Puntanen, S., and Styan, G.P.H. (editors) "Festschrift for Tarmo Pukkila on his 60th birthday", 217–231, University of Tampere, Tampere, 2006.
50. José R Berrendero, Beatriz VM Mendes, and David E Tyler. On the maximum bias functions of MM-estimates and constrained M-estimates of regression. The Annals of statistics, 35(1):13–40, 2007.
51. David E Tyler. Book review for "Robust statistical methods with R". Journal of the American Statistical Association, 102(478):759–760, 2007.
52. Anne Ruiz-Gazen and David E Tyler. Discussion on "Robustness and data analysis". Bulletin of the International Statistical Institute, 56, 2007.
53. Jue Wang and David E Tyler. A graphical method for detecting asymmetry. In Transactions of the 63rd Deming Conference, 2007.
54. Klaus Nordhausen, Hannu Oja, and David E Tyler. Tools for exploring multivariate data: The package ICS. Journal of Statistical Software, 28(6):1–31, 2008.
55. David E Tyler. Book review for "Robust statistics: Theory and methods". Journal of the American Statistical Association, 103(482):888–889, 2008.
56. Kesar Singh, David E Tyler, Jingshan Zhang, and Somnath Mukherjee. Quantile scale curves. Journal of Computational and Graphical Statistics, 18(1):92–105, 2009.

57. Seija Sirkiä, Sara Taskinen, Hannu Oja, and David E Tyler. Tests and estimates of shape based on spatial signs and ranks. Journal of Nonparametric Statistics, 21(2):155–176, 2009.
58. David E Tyler, Frank Critchley, Lutz Dümbgen, and Hannu Oja. Invariant coordinate selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(3):549–592, 2009.
59. David E Tyler. A note on multivariate location and scatter statistics for sparse data sets. Statistics & Probability Letters, 80(17–18):1409–1413, 2010.
60. Juan Lucas Bali, Graciela Boente, David E Tyler, and Jane-Ling Wang. Robust functional principal components: A projection-pursuit approach. The Annals of Statistics, 39(6):2852–2882, 2011.
61. Andrew Magyar and David E Tyler. The asymptotic efficiency of the spatial median for elliptically symmetric distributions. Sankhya B, 73(2):165–192, 2011.
62. Esa Ollila, David E Tyler, Visa Koivunen, and H Vincent Poor. Compound-Gaussian clutter modeling with an inverse Gaussian texture distribution. IEEE Signal Processing Letters, 19(12):876–879, 2012.
63. Esa Ollila, David E Tyler, Visa Koivunen, and H Vincent Poor. Complex elliptically symmetric distributions: Survey, new results and applications. IEEE Transactions on Signal Processing, 60(11):5597–5625, 2012.
64. Esa Ollila and David E Tyler. Distribution-free detection under complex elliptically symmetric clutter distribution. In 2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM), pages 413–416. IEEE, 2012.
65. Andrew Magyar and David E Tyler. The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. Biometrika, 101(3):673–688, 2014.
66. Alexander Dürre, Daniel Vogel, and David E Tyler. The spatial sign covariance matrix with unknown location. Journal of Multivariate Analysis, 130:107–117, 2014.
67. Daniel Vogel and David E Tyler. Robust estimators for nondecomposable elliptical graphical models. Biometrika, 101(4):865–882, 2014.
68. Graciela Boente, Matías Salibián Barrera, and David E Tyler. A characterization of elliptical distributions and some optimality properties of principal components for functional data. Journal of Multivariate Analysis, 131:254–264, 2014.
69. Esa Ollila and David E Tyler. Regularized M-estimators of scatter matrix. IEEE Transactions on Signal Processing, 62(22):6059–6070, 2014.
70. Esa Ollila, Frederic Pascal and David E Tyler. Complex elliptically symmetric distributions and their applications in signal processing. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-14, 2014.
71. Klaus Nordhausen and David E Tyler. A cautionary note on robust covariance plug-in methods. Biometrika, 102(3):573–588, 2015.
72. Jue Wang and David E Tyler. Generalized MM-tests for symmetry. In Nordhausen, K. and Taskinen, S. (editors) "Modern Nonparametric, Robust and Multivariate Methods. Festschrift in Honour of Hannu Oja", pages 133–148, Springer, 2015.

73. Jari Miettinen, Klaus Nordhausen, Sara Taskinen, and David E Tyler. On the computation of symmetrized M-estimators of scatter. In Agostinelli, C., Basu, A., Filzmoser, P. and Mukherje, D. (editors) "Recent Advances in Robust Statistics: Theory and Applications", pages 151–167, Springer India, New Delhi, 2016.
74. Alexander Dürre, David E Tyler, and Daniel Vogel. On the eigenvalues of the spatial sign covariance matrix in more than two dimensions. Statistics & Probability Letters, 111:80–85, 2016.
75. Esa Ollila, Ilya Soloveychik, David E Tyler, and Ami Wiesel. Simultaneous penalized M-estimation of covariance matrices using geodesically convex optimization. arXiv preprint arXiv:1608.08126, 2016.
76. Lutz Dümbgen and David E Tyler. Geodesic convexity and regularized scatter estimators. arXiv preprint arXiv:1607.05455, 2016.
77. Klaus Nordhausen, Hannu Oja, David E Tyler, and Joni Virta. Asymptotic and bootstrap tests for the dimension of the non-Gaussian subspace. IEEE Signal Processing Letters, 24(6):887–891, 2017.
78. David E Tyler and Mengxi Yi. Lassoing eigenvalues. Biometrika, 107(2):397–414, 2020.
79. Mengxi Yi and David E Tyler. Shrinking the covariance matrix using convex penalties on the matrix-log transformation. Journal of Computational and Graphical Statistics, 30(2):442–451, 2021.
80. Elias Raninen, Esa Ollila, and David E Tyler. On the variability of the sample covariance matrix under complex elliptical distributions. IEEE Signal Processing Letters, 28:2092–2096, 2021.
81. Klaus Nordhausen, Hannu Oja, and David E Tyler. Asymptotic and bootstrap tests for subspace dimension. Journal of Multivariate Analysis, 188:104830, 2022.
82. Elias Raninen, David E Tyler, and Esa Ollila. Linear pooling of sample covariance matrices. IEEE Transactions on Signal Processing, 70:659–672, 2022.

A.2 R Packages of David E Tyler

R1. Nordhausen, K., Oja, H., Tyler, D.E.: ICS: Tools for Exploring Multivariate Data via ICS/ICA. First release 2007. http://cran.r-project.org/package=ICS.

R2. Nordhausen, K., Sirkia, S., Oja, H., Tyler, D.E.: ICSNP: Tools for Multivariate Nonparametrics. First release 2007. http://cran.r-project.org/package=ICSNP.

R3. Nordhausen, K., Oja, H., Tyler, D. E., Virta, J.: ICtest: Estimating and Testing the Number of Interesting Components in Linear Dimension Reduction. First release 2016. http://cran.r-project.org/package=ICtest.

References

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10*, 10008.

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

Feinerer, I., & Hornik, K. (2020). tm: Text Mining Package. R package version 0.7–8. https://CRAN.R-project.org/package=tm

Fellows, I. (2018). wordcloud: Word Clouds. R package version 2.6. https://CRAN.R-project.org/package=wordcloud

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*, 75–174.

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

Gu, Z. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics, 30*(19), 2811–2812.

Kolaczyk, E. D., & Csardi, G. (2014). *Statistical analysis of network data with R*. New York: Springer Verlag.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., & Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research, 19*(9),1639–1645.

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

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Seifert, C., Kump, B., Kienreich, W., Granitzer, G., & Granitzer, M. (2008). On the beauty and usability of tag clouds. In *Proceedings of the 12th International Conference Information Visualisation* (pp. 17–25)

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

Taskinen, S., Frahm, G., Nordhausen, K., & Oja, H. (2023). A review of Tyler's shape matrix and its extensions. In M. Yi & K. Nordhausen (Eds.) *Robust and Multivariate Statistical Methods – Festschrift in Honor of David E. Tyler* (440p). Springer.

A Review of Tyler's Shape Matrix and Its Extensions

Sara Taskinen, Gabriel Frahm, Klaus Nordhausen, and Hannu Oja

Abstract In a seminal paper, Tyler (1987a) suggests an M-estimator for shape, which is now known as Tyler's shape matrix. Tyler's shape matrix is increasingly popular due to its nice statistical properties. It is distribution free within the class of generalized elliptical distributions. Further, under very mild regularity conditions, it is consistent and asymptotically normally distributed after the usual standardization. Tyler's shape matrix is still the subject of active research, e.g., in the signal processing literature, which discusses structured and regularized shape matrices. In this article, we review Tyler's original shape matrix and some recent developments.

Keywords M-estimator · Generalized elliptical distribution · High dimension · Robust estimator · Regularization

1 Introduction

Maronna (1976) and Huber (1981) propose robust M-estimators for location and scatter of multivariate elliptically distributed data. Since their seminal work, we can find many contributions finding new ways to estimate the location vector and scatter matrix. See Maronna et al. (2018) for a nice overview of robust multivariate methods.

S. Taskinen (✉) · K. Nordhausen
Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland
e-mail: sara.taskinen@jyu.fi; klaus.k.nordhausen@jyu.fi

G. Frahm
Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany
e-mail: frahm@hsu-hh.de

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

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_2

In this work, we focus on the robust M-estimator for shape, introduced by Tyler (1987a). We start by fixing some notation. Consider first a location-scatter model. This means that the p-variate observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent copies of

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{e},$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location vector and $\boldsymbol{\Omega} \in \mathbb{R}^{p\times q}$ is a transformation (or mixing) matrix with $\mathrm{rk}(\boldsymbol{\Omega}) = p$. Hence, we have that $p \leq q$ and the symmetric positive-definite matrix $\boldsymbol{\Sigma} := \boldsymbol{\Omega}\boldsymbol{\Omega}^\top \in \mathbb{R}^{p\times p}$ is referred to as the scatter matrix. Without loss of generality, we may choose the decomposition $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}$, where $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is the unique symmetric root of $\boldsymbol{\Sigma} > 0$.

Different multivariate models are obtained by making specific assumptions about the q-variate random vector $\mathbf{e}$ (Nordhausen and Oja 2018b). For example, it is typically assumed that $p = q$ and that $\mathbf{e}$ has a spherically symmetric absolutely continuous distribution on $\mathbb{R}^p$, i.e., the density function of $\mathbf{e}$ is of the form $f(\mathbf{e}) = \exp\{-\rho(||\mathbf{e}||)\}$ for some function $\rho : \mathbb{R}_0^+ \to \mathbb{R}$, where $\|\cdot\|$ denotes the Euclidean norm (Fang et al. 1990). Then, we can decompose $\mathbf{e}$ into a radial part and an angular part by $\mathbf{e} = r\mathbf{u}$, where the modulus, i.e., the radius, $r = \|\mathbf{e}\| > 0$ and the direction $\mathbf{u} = \|\mathbf{e}\|^{-1}\mathbf{e}$ are stochastically independent with $\mathbf{u}$ being uniformly distributed on the unit hypersphere in $\mathbb{R}^p$. The density of the modulus is proportional to $r^{p-1}\exp\{-\rho(r)\}$.

For all $\tau > 0$ we have that $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}r\mathbf{u} = \boldsymbol{\mu} + \boldsymbol{\Upsilon}s\mathbf{u}$ with $\boldsymbol{\Upsilon} := \boldsymbol{\Omega}/\tau$ and $s := \tau r$. Hence, the scatter matrix of $\mathbf{x}$ is defined only up to scale. To fix $\boldsymbol{\Sigma}$ we could assume that $\mathbf{E}(r^2) = p$ or $\mathbf{Med}(r^2) = \chi^2_{p,0.5}$, where $\chi^2_{p,0.5}$ is the median of the χ^2-distribution with p degrees of freedom. The first assumption requires that the second moment of r is finite, whereas the second assumption does not require any moment condition on r at all. If the first assumption is satisfied, we have that $\mathbf{COV}(\mathbf{e}) = \mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix, and $\mathbf{COV}(\mathbf{x}) = \boldsymbol{\Sigma}$. However, it is more common to impose the scaling condition

$$\mathbf{E}\big(\varphi(r^2)\big) = p \tag{1}$$

with $\varphi\big(r^2\big) := w\big(r^2\big)\,r^2$, where w is a real-valued partial function on $\mathbb{R}_0^+$.[1] In fact, this is typically done both in M-estimation and in ML-estimation of scatter (Frahm et al. 2020; Tyler 1982). The chosen weight function w is considered appropriate if and only if there exists no scaling constant $\tau \neq 1$ such that $\mathbf{E}\big(\varphi\big((\tau r)^2\big)\big) = p$.[2] In the special case of $w : r^2 \mapsto 1$, we obtain the simple scaling condition $\mathbf{E}(r^2) = p$ mentioned above.

[1] A partial function $f : D \to C$ is a function from a subset of D to C.

[2] See Frahm (2022) for a detailed explanation.

Under the above assumptions, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent copies of the p-variate random vector $\mathbf{x}$, which follows an elliptically symmetric distribution with density function

$$f(\mathbf{x}) = \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} g((\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})).$$

The function $g : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is referred to as the density generator of $\mathbf{x}$. Given that the first moment of r is finite, the location vector $\boldsymbol{\mu}$ is the mean vector of $\mathbf{x}$, and if the second moment of r is finite, $\mathbf{COV}(\mathbf{x}) = \mathbf{E}(r^2)/p \cdot \boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{x}$.

If we allow r to be negative and to depend on $\mathbf{u}$, then $\mathbf{x}$ is generalized elliptically distributed (Frahm and Jaekel 2010). It is worth noting that, in this case, we can no longer assume that $p = q$ without loss of generality. A particular generalized elliptical distribution, which will be of interest later on, is obtained by setting $\boldsymbol{\mu} = \mathbf{0}$ and $r = \|\boldsymbol{\Omega}\mathbf{u}\|^{-1}$ with $p = q$. The random vector $\mathbf{x} = \boldsymbol{\Omega}\mathbf{u}/\|\boldsymbol{\Omega}\mathbf{u}\|$ follows an angular central Gaussian distribution on the sphere (Tyler 1987b). In the bivariate case, i.e., $p = 2$, the angular central Gaussian distribution turns into the wrapped Cauchy distribution after angle doubling (Kent and Tyler 1988).

The last location-scatter model relevant later on is the so-called independent component model where it is assumed that the components of $\mathbf{e}$ are mutually independent. In independent component analysis the goal is to estimate $\mathbf{e}$ based on $\mathbf{x}$ alone (for an overview see for example Nordhausen and Oja 2018a). If not stated otherwise, in the following we will assume that $\mathbf{x}$ follows an elliptically symmetric distribution.

The scatter matrix $\boldsymbol{\Sigma}$ can be written as $\boldsymbol{\Sigma} = \sigma^2\mathbf{V}$, where $\sigma^2 = \sigma^2(\boldsymbol{\Sigma})$ represents the scale of $\boldsymbol{\Sigma}$. A scale function $\sigma^2(\cdot)$ is such that $\sigma^2(\mathbf{I}_p) = 1$ and $\sigma^2(\tau^2\boldsymbol{\Sigma}) = \tau^2\sigma^2(\boldsymbol{\Sigma})$ for all $\tau > 0$. Further, the matrix $\mathbf{V} = \boldsymbol{\Sigma}/\sigma^2(\boldsymbol{\Sigma})$ is the unique shape matrix associated with $\boldsymbol{\Sigma}$. Classical choices of $\sigma^2(\boldsymbol{\Sigma})$ are $\boldsymbol{\Sigma}_{11}$ (Hallin et al. 2006; Hallin & Paindaveine 2006; Hettmansperger & Randles 2002), $\mathrm{tr}(\boldsymbol{\Sigma})/p$ (Dümbgen 1998; Frahm & Jaekel 2015; Taskinen & Oja 2016; Tyler 1987a), and $\det(\boldsymbol{\Sigma})^{1/p}$ (Dümbgen & Tyler 2005; Paindaveine 2008; Salibián-Barrera et al. 2006; Taskinen et al. 2006; Tatsuoka & Tyler 2000).

Note that $\mathrm{tr}(\boldsymbol{\Sigma})/p$ and $\det(\boldsymbol{\Sigma})^{1/p}$ correspond to the arithmetic and geometric means of the eigenvalues of $\boldsymbol{\Sigma}$, respectively. The use of $\det(\boldsymbol{\Sigma})^{1/p}$ as a scale function yields a canonical definition of shape, meaning that the scale and shape estimators are asymptotically independent if the data are elliptically distributed (Paindaveine 2008). The scale describes the "size," whereas the shape describes the "orientation" of an elliptical distribution and it is well-known that several multivariate methods, such as principal component analysis, canonical correlation analysis, and multivariate regression, require the shape matrix only (Croux & Haesbroeck 2000; Salibián-Barrera et al. 2006; Taskinen et al. 2006).

In the robust-statistics literature, several functionals for multivariate distributions are proposed. Let $\mathbf{x}$ be a p-variate random vector with cumulative distribution function $F_{\mathbf{x}}$. Then a functional $\boldsymbol{\mu}(F_{\mathbf{x}}) \in \mathbb{R}^p$ is said to be a location vector if it is affine equivariant in the sense that $\boldsymbol{\mu}(F_{\mathbf{Ax+b}}) = \mathbf{A}\boldsymbol{\mu}(F_{\mathbf{x}}) + \mathbf{b}$ for any nonsingular matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and vector $\mathbf{b} \in \mathbb{R}^p$. A symmetric positive-definite functional $\mathbf{S}(F_{\mathbf{x}}) \in \mathbb{R}^{p\times p}$ is called a scatter matrix if $\mathbf{S}(F_{\mathbf{Ax+b}}) = \mathbf{A}\mathbf{S}(F_{\mathbf{x}})\mathbf{A}^\top$. Further, a symmetric positive-definite functional $\mathbf{V}(F_{\mathbf{x}}) \in \mathbb{R}^{p\times p}$ is referred to as a shape matrix if $\mathbf{V}(F_{\mathbf{x}}) = \mathbf{S}(F_{\mathbf{x}})/\sigma^2(\mathbf{S}(F_{\mathbf{x}}))$ and thus

$$\mathbf{V}(F_{\mathbf{Ax+b}}) = \frac{\mathbf{A}\mathbf{V}(F_{\mathbf{x}})\mathbf{A}^\top}{\sigma^2(\mathbf{A}\mathbf{V}(F_{\mathbf{x}})\mathbf{A}^\top)}.$$

Hence, in general, a shape matrix is not affine equivariant and $\mathbf{V}(F_{\mathbf{Ax+b}})$ even is not proportional to $\mathbf{A}\mathbf{V}(F_{\mathbf{x}})\mathbf{A}^\top$. However, if we use the canonical scale function $\det(\boldsymbol{\Sigma})^{1/p}$, we have that

$$\mathbf{V}(F_{\mathbf{Ax+b}}) = \frac{\mathbf{A}\mathbf{V}(F_{\mathbf{x}})\mathbf{A}^\top}{\sigma^2(\mathbf{A}\mathbf{A}^\top)}.$$

Thus, at least for $\sigma^2(\mathbf{A}\mathbf{A}^\top) = 1$, i.e., if not the scale, but only the shape of the distribution of $\mathbf{x}$ is affected by the transformation $\mathbf{A}$, the canonical shape matrix remains equivariant (Frahm 2009).

If the distribution of $\mathbf{x}$ is elliptically symmetric, then $\boldsymbol{\mu}(F_{\mathbf{x}}) = \boldsymbol{\mu}$. This means that all location vectors correspond to the same population quantity $\boldsymbol{\mu}$. By contrast, all scatter matrices are related to each other by $\mathbf{S}(F_{\mathbf{x}}) = \sigma^2(\mathbf{S}(F_{\mathbf{x}}))\mathbf{V}$, where $\mathbf{V}$ is the (unique) shape matrix of $\mathbf{x}$. Put another way, a scatter matrix is always a multiple of the shape matrix $\mathbf{V}$. Finally, if the functionals $\boldsymbol{\mu}(\cdot)$, $\mathbf{S}(\cdot)$, and $\mathbf{V}(\cdot)$ are applied to an empirical distribution function $\hat{F}_{\mathbf{x}}$, i.e., the empirical distribution of a random sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we obtain the corresponding estimators, which we denote by $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{F}_{\mathbf{x}})$, $\hat{\mathbf{S}} = \mathbf{S}(\hat{F}_{\mathbf{x}})$, and $\hat{\mathbf{V}} = \mathbf{V}(\hat{F}_{\mathbf{x}})$, respectively.

As mentioned above, several multivariate methods can be based on shape matrices only. Such matrices can be easily defined by normalizing any scatter matrix with a scale parameter. On the other hand, sometimes shape matrices arise naturally as a result of some estimation procedure. In this review we discuss Tyler's shape matrix, proposed in the seminal paper by Tyler (1987a), which was initially motivated via estimating equations utilizing spatial sign scores. Recall that spatial sign scores are defined as $\mathbf{U}(\mathbf{x}) = \mathbf{x}/||\mathbf{x}||$, for $\mathbf{x} \neq \mathbf{0}$, and $\mathbf{U}(\mathbf{0}) = \mathbf{0}$ (Möttönen and Oja 1995). We define Tyler's shape matrix and review its statistical properties in Sect. 2. Section 3 is devoted to some recent extensions of Tyler's shape matrix and the paper is concluded with some discussion on Sect. 4.

2 Definition and Statistical Properties

Assume that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ with $n > p$ is a random sample from a centered p-variate elliptical distribution, i.e., $\boldsymbol{\mu} = \mathbf{0}$. Further, suppose that r has no atom at $\mathbf{0}$, which means that $\mathrm{P}(r = 0) = 0$. In Tyler (1982, 1983) tests for sphericity and related shape estimators based on Huber's M-estimators were considered. It was noted that it is possible to use in the M-estimation procedure a weight function that yields a distribution-free test and estimate under the elliptical model. This served as a motivation in Tyler (1987a) to propose a shape matrix estimator $\hat{\mathbf{V}}$ as a solution of

$$\hat{\mathbf{V}} = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \hat{\mathbf{V}}^{-1} \mathbf{x}_i}. \tag{2}$$

Tyler (1987a) considers the solution of (2) an M-estimator for scatter, since it can be written as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} w\big(r_i^2\big) \mathbf{x}_i \mathbf{x}_i^\top$$

with $r_i^2 = \mathbf{x}_i^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x}_i$ for $i = 1, 2, \ldots, n$, where the weight function w is $r^2 \mapsto p/r^2$. However, we prefer to call it Tyler's shape matrix. This term is commonly used in the literature, too. The shape matrix can also be seen as the limit of two popular M-estimators for scatter, namely Huber's M-estimator and the ML-estimator under the assumption that the data have a multivariate t-distribution. More precisely, Huber's weight function is

$$w : r^2 \longmapsto \begin{cases} \gamma\,, & r^2 < \lambda \\ \gamma\lambda/r^2\,, & r^2 \geq \lambda\,, \end{cases}$$

where the parameters $\gamma, \lambda > 0$ are such that $\mathbf{E}\big(\varphi(\chi_p^2)\big) = p$. Now, Tyler's weight function occurs for $\lambda \searrow 0$, i.e., λ approaching zero from above. Alternatively, we obtain Tyler's weight function by setting $\nu = 0$ in the Student-type weight function $r^2 \mapsto (p + \nu)/(r^2 + \nu)$ or $\alpha = 1$ in the power M-weight function $r^2 \mapsto (r^2/p)^{-\alpha}$ proposed by Frahm et al. (2020).

Another way to write down the estimation equation in (2) is via spatial sign scores, which are defined in Möttönen and Oja (1995). Then Tyler's shape matrix $\hat{\mathbf{V}}$ solves

$$\frac{p}{n} \sum_{i=1}^{n} \frac{\hat{\mathbf{V}}^{-\frac{1}{2}} \mathbf{x}_i \mathbf{x}_i^\top \hat{\mathbf{V}}^{-\frac{1}{2}}}{||\hat{\mathbf{V}}^{-\frac{1}{2}} \mathbf{x}_i||^2} = \frac{p}{n} \sum_{i=1}^{n} \mathbf{U}(\mathbf{z}_i) \mathbf{U}(\mathbf{z}_i)^\top = \mathbf{I}_p$$

with $\mathbf{z}_i := \hat{\mathbf{V}}^{-\frac{1}{2}} \mathbf{x}_i$ for $i = 1, 2, ..., n$ and $\mathbf{U}(\mathbf{z}_i) := \|\mathbf{z}_i\|^{-1} \mathbf{z}_i$. This means that $\hat{\mathbf{V}}$ is chosen such that the spatial signs of the transformed observations $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ are, approximately, uniformly distributed on the unit hypersphere.

In any case, (2) can be re-written, equivalently, as

$$\hat{\mathbf{V}} = \frac{p}{n}\sum_{i=1}^{n} \frac{r_i^2 \mathbf{V}^{\frac{1}{2}}\mathbf{u}_i\mathbf{u}_i^\top(\mathbf{V}^{\frac{1}{2}})^\top}{r_i^2\mathbf{u}_i^\top(\mathbf{V}^{\frac{1}{2}})^\top\hat{\mathbf{V}}^{-1}\mathbf{V}^{\frac{1}{2}}\mathbf{u}_i} = \frac{p}{n}\sum_{i=1}^{n} \frac{\mathbf{V}^{\frac{1}{2}}\mathbf{u}_i\mathbf{u}_i^\top(\mathbf{V}^{\frac{1}{2}})^\top}{\mathbf{u}_i^\top(\mathbf{V}^{\frac{1}{2}})^\top\hat{\mathbf{V}}^{-1}\mathbf{V}^{\frac{1}{2}}\mathbf{u}_i},$$

which means that the sample observations $r_1, r_2, \ldots, r_n$ of the modulus r have no impact on Tyler's shape matrix at all. This holds true even if some r_i becomes negative, since r_i^2 does not depend on the sign of r_i, and also if r_i depends on $\mathbf{u}_i$. Hence, Tyler's shape matrix is distribution free if the data are generalized elliptically distributed—provided that r has no atom at $\mathbf{0}$ and we know the location vector $\boldsymbol{\mu}$ (Frahm & Jaekel 2015). Here, we have chosen $\mathbf{V}^{\frac{1}{2}}$ as a transformation matrix. Indeed, the decomposition of $\mathbf{V}$, i.e., the precise meaning of the $p \times p$ matrix $\mathbf{U}$ in $\mathbf{V} = \mathbf{U}\mathbf{U}^\top$, is *not* arbitrary if r depends on $\mathbf{u}$, but our arguments still remain valid if we choose any other decomposition of $\mathbf{V}$.

Originally, conditions for the existence of Tyler's shape matrix were listed in Tyler (1987a) and it was shown that the matrix is unique up to a positive scaling constant. In Tyler (1987a) the shape matrix was chosen so that $\mathrm{tr}(\hat{\mathbf{V}}) = p$ and in Tatsuoka and Tyler (2000) $\det(\hat{\mathbf{V}}) = 1$ was used. We use here the first option. Tyler's shape matrix can be computed simply by starting with an initial value, e.g., $\hat{\mathbf{V}}_0 = \mathbf{I}_p$, and then iterating

$$\mathbf{z}_i = \hat{\mathbf{V}}_{k-1}^{-\frac{1}{2}}\mathbf{x}_i,$$

$$\hat{\mathbf{V}}_k \leftarrow \hat{\mathbf{V}}_{k-1}^{\frac{1}{2}}\,\frac{p}{n}\sum_{i=1}^{n}\mathbf{U}(\mathbf{z}_i)\mathbf{U}(\mathbf{z}_i)^\top\,\hat{\mathbf{V}}_{k-1}^{\frac{1}{2}},$$

until convergence. The scale can be fixed either at each iteration step or in the end so that $\mathrm{tr}(\hat{\mathbf{V}}) = p$. In Tyler (1987a) weak conditions for the convergence are given. See also Kent & Tyler 1988 for the existence of the solution under general distributions.

Recently, in Wiesel (2012) a new viewpoint for the investigation of covariance matrices was developed. In this framework covariance matrices can be seen as elements of the Riemannian manifold of symmetric positive-definite matrices which can also be used to study Tyler's shape matrix. The use of the concept of geodesic convexity provides then a new set of tools to prove existence and uniqueness of Tyler's shape matrix. Dümbgen & Tyler 2016 give a very detailed treatment of the geodesic approach to M-estimation of scatter in general and to Tyler's shape matrix in particular. Another advantage of this framework is the development of fast Newton-Raphson type algorithms for Tyler's shape matrix (Dümbgen et al. 2016; Dümbgen and Tyler 2016) which are from a computational point of view more efficient than the fixed-point algorithm mentioned above. Franks and Moitra (2020) show the connection between Tyler's shape matrix and operator scaling. This connection is then used to derive non-asymptotic bounds and to show that the iterative algorithm from above converges in polynomially many steps. Other results concerning non-asymptotic performance are given in Soloveychik & Wiesel 2015.

Now, consider the limiting behavior of Tyler's shape matrix, more precisely, the consistency of $\hat{\mathbf{V}}$ and the asymptotic distribution of $\sqrt{n}\big(\hat{\mathbf{V}} - \mathbf{V}\big)$, where $\hat{\mathbf{V}}$ is based on a random sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \sim \mathbf{x}$. We need not require that $\mathbf{x}$ is elliptically distributed. The matrix $\mathbf{V}$ just represents a solution of

$$\mathbf{V} = p\mathbf{E}\left(\frac{\mathbf{x}\mathbf{x}^\top}{\mathbf{x}^\top\mathbf{V}^{-1}\mathbf{x}}\right).$$

This solution exists and is unique –up to scale– if the distribution of $\mathbf{x}$ is continuous (Tyler 1987a). Further, in this case, $\hat{\mathbf{V}}$ is strongly consistent, i.e., it converges almost surely to $\mathbf{V}$. In order to prove that $\sqrt{n}\big(\hat{\mathbf{V}} - \mathbf{V}\big) \to N_{p\times p}(\mathbf{0}, \mathbf{C})$ as $n \to \infty$, Tyler (1987a) applies the normalization $\hat{\mathbf{V}}_0 = p\hat{\mathbf{V}}/\operatorname{tr}(\mathbf{V}^{-1}\hat{\mathbf{V}})$. The asymptotic covariance matrix of $\sqrt{n}\big(\hat{\mathbf{V}}_0 - \mathbf{V}\big)$ is quite complicated (Tyler 1987a, Theorem 3.2).

If $\mathbf{x}$ is elliptically distributed, the asymptotic covariance matrix of $\sqrt{n}\big(\hat{\mathbf{V}}_0 - \mathbf{V}\big)$ simplifies, essentially. More precisely, it holds that

$$\mathbf{C} = \frac{p+2}{p}\big(\mathbf{I}_{p^2} + \mathbf{K}_{p^2}\big)\big(\mathbf{V}\otimes\mathbf{V}\big) + \frac{2(p+2)}{p^2}\operatorname{vec}(\mathbf{V})\operatorname{vec}(\mathbf{V})^\top,$$

where $\mathbf{I}_{p^2}$ is the $p^2 \times p^2$ identity matrix, $\mathbf{K}_{p^2}$ is the $p^2 \times p^2$ commutation matrix, and $\operatorname{vec}(\mathbf{V})$ is the p^2-variate vector obtained by stacking the columns of $\mathbf{V}$ on top of each other. Frahm (2009, Corollary 1) shows that we obtain the same asymptotic covariance matrix for $\sqrt{n}\big(\hat{\mathbf{V}} - \mathbf{V}\big)$ when choosing the canonical scale function, i.e., requiring that $\det(\hat{\mathbf{V}}) = 1$.

It can be seen that $\hat{\mathbf{V}}_0$ is affine equivariant. However, in general, this is no longer true if we choose another normalization of $\hat{\mathbf{V}}$. The chosen scale function has an essential impact on the asymptotic covariance matrix. More precisely, we have that

$$\mathbf{C} = \frac{p+2}{p}\psi\big(\mathbf{I}_{p^2} + \mathbf{K}_{p^2}\big)\big(\mathbf{V}\otimes\mathbf{V}\big)\psi^\top$$

with $\psi := \mathbf{I}_{p^2} - \operatorname{vec}(\mathbf{V})\mathcal{J}_{\sigma^2}$, where $\mathcal{J}_{\sigma^2}$ is the Jacobian of the scale function σ^2 (Frahm 2009; Frahm & Jaekel 2010; Frahm et al. 2020). See also Sirkiä et al. (2009) and Taskinen and Oja (2016), among others, for the limiting distributions of $\sqrt{n}\big(\hat{\mathbf{V}} - \mathbf{V}\big)$ under different choices of σ^2. In any case, since Tyler's shape matrix is distribution free within the class of generalized elliptical distributions, the asymptotic covariance matrix never depends on the distribution of the generating variate r. Further, the breakdown point of Tyler's shape matrix is between $1/(p+1)$ and $1/p$ (Dümbgen & Tyler 2005; Yohai & Maronna 1990). In Adrover (1998) Tyler's shape matrix is shown to be minimax bias-robust.

Tyler (1987a) points out that his shape matrix is the "most robust" estimator for the shape matrix of an absolutely continuous elliptical population. More precisely, let h be a real-valued, differentiable, and scale invariant function of $\mathbf{\Sigma} > 0$. That is, we have that $h(\alpha\mathbf{\Sigma}) = h(\mathbf{\Sigma})$ for all $\alpha > 0$ and $\mathbf{\Sigma} > 0$. Consider some parameter $\theta = h(\mathbf{\Sigma})$ and some estimator $\hat{\theta} = h(\hat{\mathbf{\Sigma}})$. It is clear that we can substitute $\mathbf{\Sigma}$ with $\mathbf{V}$ and $\hat{\mathbf{\Sigma}}$ with $\hat{\mathbf{V}}$. Now, Tyler's shape matrix minimizes the maximum asymptotic variance of $\hat{\theta} = h(\hat{\mathbf{V}})$ among all consistent estimators $\hat{\mathbf{V}}$ such that $\sqrt{n}\big(\hat{\mathbf{V}} - \mathbf{V}\big) \rightarrow N_{p\times p}(\mathbf{0}, \mathbf{C})$.

Tyler's shape matrix is usually introduced as a general M-estimator of shape; however, it can also be derived as the ML-estimator for $\mathbf{\Sigma}$ under the angular central Gaussian distribution, as shown in Tyler (1987b). Later Ollila and Tyler (2012) showed the similar result under the more general model of elliptical distributions of proportional covariance matrices. See also Gini and Greco (2002), Conte et al. (2002).

Above, we assumed that the location vector of the elliptical distribution is known. However, Tyler (1987a) considers also the case in which the location is unknown. One can, for example, center the observations using any $\sqrt{n}$-consistent location estimate before computing the shape matrix. The asymptotic properties of the resulting shape matrix estimate will remain the same. Tyler (1987a) also mentions a possibility of estimating the location vector and shape matrix simultaneously in a similar fashion as in Maronna (1976), Huber (1981) and recognizes the limitations of such an approach. We will discuss the simultaneous estimation in Sect. 3 along with other extensions of Tyler's shape matrix.

To conclude this section, note that Paindaveine and Van Bever (2019) introduce the concept of Tyler shape depth which can be used to order shape matrices. The deepest shape matrix is then related to the definition of Tyler's shape matrix.

3 Extensions

In the exposition above, Tyler's shape matrix was considered for real data observations with known location and for data without missing values. It was also assumed that the shape matrix does not follow any special structure and that the sample size n is larger than the dimension p. All the issues listed above have been recently addressed in the literature and in the following we will give an overview of the approaches that tackle these issues.

As custom in statistics we will continue focusing on real-valued data. Especially in the signal processing community the theory is, however, often developed considering complex-valued data and most of the methods described below can also be applied in such a context. The interested reader is referred, for example, to Kent (1997), Gini and Greco (2002), Conte et al. (2002), Pascal et al. (2008), Ollila and Tyler (2012), Ollila et al. (2012), and the references therein.

3.1 Joint Estimation of Location and Tyler's Shape Matrix: The Hettmansperger–Randles Estimators

Hettmansperger and Randles (2002) tackle the problem of simultaneous estimation of location vector and shape matrix utilizing spatial sign scores. Write now $\mathbf{z}_i = \hat{\mathbf{V}}^{-\frac{1}{2}}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$, $i = 1, \ldots, n$, for transformed observations. Then the Hettmansperger–Randles (HR) estimators of location vector and shape matrix, $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{V}}$, solve

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{U}(\mathbf{z}_i) = \mathbf{0} \quad \text{and} \quad \frac{p}{n}\sum_{i=1}^{n}\mathbf{U}(\mathbf{z}_i)\mathbf{U}(\mathbf{z}_i)^\top = \mathbf{I}_p, \tag{3}$$

and $\hat{\mathbf{V}}$ is standardized, for example, such that $\mathrm{tr}(\hat{\mathbf{V}}) = p$. The resulting location vector estimate is known as the transformation-retransformation (TR) spatial median (Chakraborty et al. 1998) and the shape matrix is the Tyler's shape matrix with respect to the TR spatial median. Notice that the classical spatial median, which solves $n^{-1}\sum_{i=1}^{n}\mathbf{U}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$, is only rotation equivariant, whereas the TR spatial median is affine equivariant. For the robustness properties and limiting distributions of HR estimates, see Hettmansperger and Randles (2002); Möttönen et al. (2010); Oja (2010).

HR estimates are easy to compute as estimating equations in (3) lead to the following iteration steps:

$$\mathbf{z}_i = \hat{\mathbf{V}}_{k-1}^{-\frac{1}{2}}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{k-1}),$$

$$\hat{\boldsymbol{\mu}}_k \leftarrow \hat{\boldsymbol{\mu}}_{k-1} + \frac{\hat{\mathbf{V}}_{k-1}^{\frac{1}{2}}\sum_{i=1}^{n}\mathbf{U}(\mathbf{z}_i)}{\sum_{i=1}^{n}||\mathbf{z}_i||^{-1}},$$

$$\hat{\mathbf{V}}_k \leftarrow \hat{\mathbf{V}}_{k-1}^{\frac{1}{2}}\,\frac{p}{n}\sum_{i=1}^{n}\mathbf{U}(\mathbf{z}_i)\mathbf{U}(\mathbf{z}_i)^\top\,\hat{\mathbf{V}}_{k-1}^{\frac{1}{2}}.$$

See also Hettmansperger and Randles (2002) for computation of HR estimates. Unfortunately, as far as we know, there is no proof for the convergence of the above algorithm. Also, as Tyler (1987a) already pointed out, the existence and uniqueness of the HR estimates remains an open question as the estimates do not satisfy the conditions that guarantee the existence and uniqueness of simultaneous M-estimates (Huber 1981; Maronna 1976). Motivated by this, Taskinen and Oja (2016) proposed k-step HR estimators for location and shape, that is, one starts with initial $\sqrt{n}$-consistent estimates $\hat{\boldsymbol{\mu}}_0$ and $\hat{\mathbf{V}}_0$ and repeats the above iteration steps k times. Resulting estimates are affine equivariant if the initial estimates are affine equivariant. The limiting distributions depend on the limiting distributions of the initial pair of estimates and those of HR estimates. The larger the k, the more similar are the distributions to those of the HR estimates (Taskinen & Oja 2016). For the robustness properties of k-step estimates, see Croux et al. (2010), Taskinen and Oja (2016).

3.2 The Symmetrized Variant of Tyler's Shape Matrix: Dümbgen's Estimator

Tyler (1987a) assumes that the location center is known or given. Dümbgen (1998) avoids this assumption and proposes a symmetrized version of the Tyler's shape matrix. Write now $\mathbf{z}_i = \hat{\mathbf{V}}^{-\frac{1}{2}}\mathbf{x}_i$, $i = 1, \ldots, n$. Dümbgen's shape matrix $\hat{\mathbf{V}}$ is then chosen to solve

$$\frac{1}{p}\mathbf{I}_p = \binom{n}{2}^{-1} \sum\sum_{i<j} \frac{\hat{\mathbf{V}}^{-\frac{1}{2}}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \hat{\mathbf{V}}^{-\frac{1}{2}}}{||\hat{\mathbf{V}}^{-\frac{1}{2}}(\mathbf{x}_i - \mathbf{x}_j)||^2}$$

$$= \binom{n}{2}^{-1} \sum\sum_{i<j} \mathbf{U}(\mathbf{z}_i - \mathbf{z}_j)\mathbf{U}(\mathbf{z}_i - \mathbf{z}_j)^\top$$

and standardized, for example, such that $\mathrm{tr}(\hat{\mathbf{V}}) = p$. This shape matrix is thus Tyler's shape matrix computed on pairwise differences.

Statistical properties of Dümbgen's shape matrix are studied in detail in Dümbgen (1998), Dümbgen and Tyler (2005), Rublik (2021), and later in Sirkiä et al. (2007), Dümbgen et al. (2015) under a framework of symmetrized M-estimators of scatter. The shape matrix obtained using pairwise differences is highly efficient (under the elliptical model). It also possesses the so-called joint (and block) independence properties which means that the matrix is a (block) diagonal matrix if the components of **x** are mutually (block) independent (Nordhausen & Tyler 2015). The joint independence property is rare among scatter and shape functionals and needed, for example, in the independent component model. The use of symmetrized scatter functionals for independent component analysis is discussed in Oja et al. (2006). Other multivariate methods that require the joint or block independence property are discussed in Nordhausen and Tyler (2015).

Dümbgen's shape matrix can be computed using the algorithm proposed in Tyler (1987a), that is, one can simply start with an initial value, e.g., $\hat{\mathbf{V}}_0 = \mathbf{I}_p$, and then iterate

$$\mathbf{z}_i = \hat{\mathbf{V}}_{k-1}^{-\frac{1}{2}}\mathbf{x}_i,$$

$$\hat{\mathbf{V}}_k \leftarrow \hat{\mathbf{V}}_{k-1}^{\frac{1}{2}} \binom{n}{2}^{-1} \sum\sum_{i<j} \mathbf{U}(\mathbf{z}_i - \mathbf{z}_j)\mathbf{U}(\mathbf{z}_i - \mathbf{z}_j)^\top \hat{\mathbf{V}}_{k-1}^{\frac{1}{2}}.$$

The standardization can be done after the algorithm has converged. Although the above algorithm is easy to apply in practice, it has a drawback of being highly intensive when the sample size is large. Due to this issue, several new computational approaches and variants of Dümbgen's shape matrix are introduced in the literature. For alternative algorithms, see Miettinen et al. (2016), Dümbgen et al. (2016). In

Taskinen et al. (2010) k-step estimator of Dümbgen's shape matrix was considered. Finally notice that iteration steps

$$\hat{\mathbf{V}}_k \leftarrow \hat{\mathbf{V}}_{k-1}^{\frac{1}{2}} \frac{1}{n(n-1)^2} \sum \sum \sum_{i \neq j, i \neq k} \mathbf{U}(\mathbf{z}_i - \mathbf{z}_j)\mathbf{U}(\mathbf{z}_i - \mathbf{z}_k)^\top \hat{\mathbf{V}}_{k-1}^{\frac{1}{2}}$$

yield a related shape matrix estimator based on spatial rank vectors (Möttönen & Oja 1995). We refer interested readers to Sirkiä et al. (2009) for more details.

3.3 Estimation Under Missing Data

In real-life applications, practitioners often face the problem that some data are missing. Nevertheless, it may be of interest to estimate the scatter matrix by using *all* available observations—not only the observations that are complete. Under the assumption that the data are missing at random, maximum-likelihood methods based on the so-called observed-data likelihood function are well-developed (Schafer 1997). In order to generalize Tyler's shape matrix to the case of incomplete data, Frahm and Jaekel (2010) use the fact that Tyler's shape matrix $\hat{\mathbf{V}}$ is a ML-estimator under the angular central Gaussian distribution. More precisely, they show that $\hat{\mathbf{V}}$ represents an observed-data ML-estimator under the assumption that the data stem from a generalized elliptical distribution. They also point out that the incomplete data must be missing *completely* at random to guarantee the consistency of $\hat{\mathbf{V}}$.

Frahm and Jaekel (2010) provide a fixed point algorithm for the computation of Tyler's shape matrix in the case of incomplete data. An extension to the case of the Hettmansperger–Randles estimator is also given. Since the notation convention in the missing-data framework is nonstandard, we omit details here and refer to Frahm and Jaekel (2010). Theoretical properties of M-estimators, in particular for Tyler's shape matrix, in the case of independent and dependent observations are derived by Frahm et al. (2020). The aforementioned authors show that, when applying M-weight functions to incomplete data, the critical scaling condition expressed by (1) must be satisfied, correspondingly, for each incomplete observation, in order to guarantee that the given M-estimator for scatter is consistent. They resolve the scaling problem by introducing the class of power M-estimators for location and scatter. Both the Gauss-type weight function, $r^2 \mapsto 1$, and Tyler's weight function, $r^2 \mapsto d/r^2$, represent two distinguished power M-weight functions, which implicitly satisfy the critical scaling condition for incomplete data.

3.4 Structured Tyler's Shape Estimation

In many applications there is some prior knowledge about the structure of the scatter/shape matrix available. Such structures include, for example, Toeplitz structure, spiked covariance structure, group symmetry, and Kronecker structure, among many others. Originally, structured estimation was considered in the context of the covariance matrix estimation for iid Gaussian data and it was shown that enforcing the structure improves the performance of the estimator. Recently, especially in the signal processing community, there has been an increasing interest in estimating robust structured scatter matrices in the context of elliptical distributions, and the research has focused especially on Tyler's shape matrix (see for example Mériaux et al. 2021; Soloveychik and Trushin 2016; Soloveychik et al. 2016; Soloveychik and Wiesel 2014; Sun et al. 2016, and the references therein). In general, algorithms to estimate the structured shape matrix are usually tailored for the specific structure. A lot depends on the convexity of the assumed structure. As the unstructured Tyler's shape matrix is geodesic convex, it can be concluded that the minimizer of the cost function under a constraint that is also geodesic convex leads to a global maximizer, which is, for example, the case under a group symmetric constraint (Soloveychik et al. 2016).

In the following assume a centered sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ with an unstructured estimate $\hat{\mathbf{V}}$ of Tyler's shape matrix and denote $\mathcal{S}$ as a known closed convex subset of all positive semi-definite $p \times p$ matrices $\mathcal{V}$ under an appropriate scale constraint, i.e., $\mathcal{S} \subset \mathcal{P}$. The subset $\mathcal{S}$ of $\mathcal{P}$ is closed but not necessarily convex. In the following we only outline some general approaches for structured estimation and provide some references for more details.

Convex projection projects an unconstrained estimate onto the closest element of the constrained set, that is, structured shape matrix $\mathbf{V}^s$ is found as a solution to

$$\min_{\mathbf{V}^s \in \mathcal{S}} ||\mathbf{V}^s - \hat{\mathbf{V}}||,$$

where $||\cdot||$ denotes some norm. This is a convex optimization problem, but it consists of a two-step procedure and therefore does not couple structural and distributional properties in the estimation process (Soloveychik and Wiesel 2014).

Convex constrained covariance estimation (COCA, Soloveychik and Wiesel 2014) is based on the general methods of moments (GMM) approach and it seeks an approximate solution to

$$\min_{\mathbf{V}^s \in \mathcal{S}} \left|\left| \mathbf{V}^s - \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \mathbf{V}^{s\,-1} \mathbf{x}_i} \right|\right|.$$

This problem is, however, not convex and for general practical computation a convex relaxation of the above equation is suggested which then allows the use of general optimizers for solving the problem. It is then shown that in the unconstrained case

COCA is equivalent to Tyler's shape matrix and in the constrained case the two matrices are asymptotically equivalent.

The most general form is the majorization-minimization (MM) approach of Sun et al. (2015, 2016), which starts from a log-likelihood point of view and aims at solving

$$\min_{\mathbf{V}^s \in \mathcal{S}'} \log\det(\mathbf{V}^s) + \frac{p}{n}\sum_{i=1}^{n}\log(\mathbf{x}_i^\top \mathbf{V}^{s-1}\mathbf{x}_i). \tag{4}$$

Due to the complexity of the problem, the MM approach searches for stationary points of (4) and therefore does not necessarily provide the global optimum. Sun et al. (2015, 2016) then provide many tailored MM algorithms for specific structures whose properties depend on the structure at hand. These include convex (e.g., Toeplitz structure, sum of rank one matrices structure) and non-convex structures (e.g., spiked covariance model structure, Kronecker structure).

3.5 Regularized Estimators

A topic of increasing interest in multivariate statistics is high-dimensionality, as the dimension p of modern data can be very large and increasingly often even larger than the available sample size n. Therefore, the behavior of multivariate methods is nowadays often investigated in settings where n and/or p grow.

A key result regarding scatter matrix estimation is given in Tyler (2010), which states that for finite data, if $n \leq p+1$ and the data is in general position, then any affine equivariant scatter matrix is proportional to the covariance matrix, where the proportionality factor does not depend on the data. The question is then, what is the behavior of Tyler's shape matrix if n and p grow, i.e., if $p/n \to c$ when $n \to \infty$ and $p \to \infty$. Dümbgen (1998), Frahm and Glombek (2012) consider the case $c = 0$ and show that the condition number of Tyler's shape matrix is $1 + 4\sqrt{p/n} + o(\sqrt{p/n})$ and that the spectral distribution of $\sqrt{n/p}(\hat{\mathbf{V}} - \mathbf{I}_p)$ converges weakly to a semicircle distribution. Further, Zhang et al. (2016) show that in the case $0 < c < 1$ the spectral distribution of Tyler's shape matrix converges weakly to the Marčenko–Pastur distribution. Notice that these results are derived in the context of iid samples from elliptical distributions while similar results for the covariance matrix require usually iid samples from the Gaussian distribution or are less useful in case of elliptical distributions (Karoui 2009; Zhang et al. 2016).

As the estimation in high-dimensional setting is quite challenging, often estimators are regularized in such a setup and include shrinkage. For Tyler's shape matrix basically three different options for shrinkage are considered: (i) shrinking the eigenvalues of an already computed shape matrix, (ii) adding an penalty term to the M-estimation objective function, or (iii) modifying the M-estimation equation.

In the following we outline some recent suggestions to regularize Tyler's shape matrix and refer for further details to the provided references. We first consider shrinking the eigenvalues which assumes a framework with $n > p$ and that we are able to compute Tyler's shape matrix $\hat{\mathbf{V}}$ with $\mathrm{tr}(\hat{\mathbf{V}}) = p$ for the centered sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The shrinkage regularized Tyler's shape matrix is then defined as

$$\hat{\mathbf{V}}^r_\alpha = \alpha\hat{\mathbf{V}} + (1-\alpha)\mathbf{I}_p,$$

where $\alpha \in [0, 1]$ is a regularization parameter. Thus $\hat{\mathbf{V}}^r_\alpha$ shares the same eigenvectors as $\hat{\mathbf{V}}$ but the eigenvalues of it are shrank towards the mean of the eigenvalues of $\hat{\mathbf{V}}$. This type of estimator is considered, for example, in Chen et al. (2011), Ollila et al. (2021). Ollila et al. (2021) suggest to choose α as

$$\alpha_o = \min_\alpha \mathrm{MSE}(\hat{\mathbf{V}}^r_\alpha),$$

where $\mathrm{MSE}(\hat{\mathbf{V}}^r_\alpha) = \mathbf{E}[||\hat{\mathbf{V}}^r_\alpha - \mathbf{V}||^2]$, for which a closed-form expression can be obtained in case of elliptical distributions or using cross validation (CV).

To allow $p > n$ Abramovich and Spencer (2007) suggest to load the diagonals in the fixed point algorithm of Tyler's shape matrix by modifying the updating steps as follows

$$\begin{aligned}
\tilde{\mathbf{V}}^r_{k,\beta} &\leftarrow \beta\,\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top(\hat{\mathbf{V}}^r_{k-1,\beta})^{-1},\,\beta\mathbf{x}_i} + (1-\beta)\mathbf{I}_p,\\
\hat{\mathbf{V}}^r_{k,\beta} &\leftarrow \frac{\tilde{\mathbf{V}}^r_{k,\beta}}{\mathrm{tr}(\tilde{\mathbf{V}}^r_{k,\beta})},
\end{aligned}$$

and iterate until convergence. Here $\beta \in [0, 1]$ is a shrinkage coefficient. Chen et al. (2011) establish uniqueness of the estimator and suggest a way to choose β. The above estimator has, however, been criticized for being heuristic as it is not related to any cost function which it would minimize. Therefore, Wiesel (2012) starts again from a log-likelihood point of view and suggests to minimize the following penalized log-likelihood function, that is, to solve

$$\min_{\mathbf{V}\in\mathcal{S}} \log\det(\mathbf{V}) + \frac{p}{n}\sum_{i=1}^{n}\log(\mathbf{x}_i^\top\mathbf{V}^{-1}\mathbf{x}_i) + \gamma P(\mathbf{V}), \tag{5}$$

where $P(\mathbf{V})$ is the penalty function and $\gamma \geq 0$ is a regularization parameter. Wiesel (2012) uses $P(\mathbf{V}) = p\log(\mathrm{tr}(\mathbf{V}^{-1}\mathbf{T})) + \log(|\mathbf{V}|)$, which has its minimum at $\mathbf{T}$ which is the desired target matrix towards which $\mathbf{V}$ should be shrunk to. The minimizer of (5) is denoted as $\hat{\mathbf{V}}^r_\gamma$. Wiesel (2012) and Dümbgen and Tyler (2016) also list several other penalty functions and discuss their appropriateness for different data settings. The regularization parameter can either be fixed or chosen data dependent using so-called oracle type estimator as discussed in Chen et al. (2011), Ollila and Tyler (2014). The use of cross validation was suggested in Dümbgen and Tyler (2016). It is shown that statistical properties, such as existence and uniqueness, depend on the used penalty function and special attention is given to penalty functions which are geodesic convex in $\mathbf{V}$. For example, Sun et al. (2014) show that if one uses as penalty the Kullback–Leibler distance between two zero mean Gaussian distributions with covariance matrices $\mathbf{V}$ and $\mathbf{T}$, an estimate very similar to the diagonal loading method mentioned above is obtained.

For further discussions on regularized Tyler's shape matrices we refer to Pascal et al. (2014), Couillet and McKay (2014), Sun et al. (2014), Ollila and Tyler (2014), Dümbgen and Tyler (2016), where maybe Dümbgen and Tyler (2016) provide the most general treatment of regularized Tyler's shape matrices and suggest also a cross validation procedure. Corresponding algorithms are discussed, for example, in Sun et al. (2014), Dümbgen and Tyler (2016). Robustness properties of previous regularized estimators are studied in Tyler and Yi (2020) showing that, under certain conditions on the tuning parameter, the breakdown point of regularized Tyler's shape matrix could be 1, if not estimating the center $\boldsymbol{\mu}$.

None of the above methods guarantees a sparse solution. To obtain a sparse estimate based on a (regularized) Tyler's shape matrix, Goes et al. (2020) discuss thresholding. Entry-wise thresholding of a matrix $\mathbf{A} = (a_{ij})$ and a threshold $t > 0$ is defined as

$$\boldsymbol{\tau}_t(\mathbf{A}) = (I(|a_{ij}| > t)a_{ij}).$$

Applying such an entry-wise thresholding to an estimate of Tyler's shape matrix, which can also be regularized, yields the thresholded estimate

$$\hat{\mathbf{V}}^t = \boldsymbol{\tau}_t(\hat{\mathbf{V}}),$$

where it is assumed that $\hat{\mathbf{V}}$ has unit trace. Under the assumption of elliptical data with approximately sparse scatter matrix, Goes et al. (2020) provide many properties of $\hat{\mathbf{V}}^t$, especially that these estimators are rate optimal, meaning that the rate coincides with the minimax rate for sparse covariance estimation for sub-Gaussian elliptical data but in addition holds also for heavy tailed elliptical data. There seems, however, to be no suggestion yet for choosing the threshold t in a data-driven fashion.

4 Discussion

The seminal paper introducing Tyler's shape matrix (Tyler 1987a) has been cited according to the Web of Science up to date 378 times.[3] Since its appearance, Tyler's shape matrix has been used in many application areas such as antenna array processing (Ollila and Koivunen 2003), radar detection (Ollila & Tyler 2012) or image analysis based subspace recovery (Zhang 2015). Applications in the field of finance are discussed, for example, in Frahm and Jaekel (2015) and Yang et al. (2015).

This paper is a short and restricted review which shows that due to its non-parametric nature with many excellent statistical properties and computational simplicity, Tyler's shape matrix is still, 35 years after its introduction, an active research area. Tyler's shape matrix continues to exhibit great promise and can be extended in different directions driven by the complex nature of modern data sets.

Acknowledgments The authors would like to thank the editor and referees for their insightful comments and suggestions.

References

Abramovich, Y. I., & Spencer, N. K. (2007). Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP '07* (vol. 3, pp. III–1105–III–1108).

Adrover, J. G. (1998). Minimax bias-robust estimation of the dispersion matrix of a multivariate distribution. *The Annals of Statistics*, *26*, 2301–2320.

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

Chen, Y., Wiesel, A., & Hero, A. O. (2011). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, *59*, 4097–4107.

Conte, E., De Maio, A., & Ricci, G. (2002). Recursive estimation of the covariance matrix of a compound-gaussian process and its application to adaptive CFAR detection. *IEEE Transactions on Signal Processing*, *50*(8), 1908–1915.

Couillet, R., & McKay, M. (2014). Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, *131*, 99–120.

Croux, C., Dehon, C., & Yadine, A. (2010). The k-step spatial sign covariance matrix. *Advances in Data Analysis and Classification*, *4*, 137–150.

Croux, C., & Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, *87*, 603–618.

Dümbgen, L. (1998). On Tyler's M-functional of scatter in high dimension. *Annals of the Institute of Statistical Mathematics*, *50*(3), 471–491.

Dümbgen, L., Nordhausen, K., & Schuhmacher, H. (2016). New algorithms for M-estimation of multivariate scatter and location. *Journal of Multivariate Analysis*, *144*, 200–217.

[3] Access date: 09.05.2022.

Dümbgen, L., Pauly, M., & Schweizer, T. (2015). M-functionals of multivariate scatter. *Statistics Surveys*, *9*, 32–105.

Dümbgen, L., & Tyler, D. E. (2005). On the breakdown properties of some multivariate M-functionals. *Scandinavian Journal of Statistics*, *32*, 247–264.

Dümbgen, L., & Tyler, D. E. (2016). Geodesic convexity and regularized scatter estimators. arXiv:1607.05455.

Fang, K., Kotz, S., & Ng, K. (1990). *Symmetric multivariate and related distributions*. London: Chapman & Hall.

Frahm, G. (2009). Asymptotic distributions of robust shape matrices and scales. *Journal of Multivariate Analysis*, *100*, 1329–1337.

Frahm, G. (2022). Power M-estimators for location and scatter. In M. Yi & K. Nordhausen (Eds.), *Robust and multivariate statistical methods: Festschrift in Honor of David E. Tyler*. Cham: Springer. https://doi.org/10.1007/978-3-031-22687-8_8

Frahm, G., & Glombek, K. (2012). Semicircle law of Tyler's M-estimator for scatter. *Statistics & Probability Letters*, *82*, 959–964.

Frahm, G., & Jaekel, U. (2010). A generalization of Tyler's M-estimators to the case of incomplete data. *Computational Statistics & Data Analysis*, *54*, 374–393.

Frahm, G., & Jaekel, U. (2015). Tyler's M-estimator in high-dimensional financial-data analysis. In K. Nordhausen, & S. Taskinen (Eds.), *Modern nonparametric, robust and multivariate methods: Festschrift in Honour of Hannu Oja* (pp. 289–305). Cham: Springer International Publishing.

Frahm, G., Nordhausen, K., & Oja, H. (2020). M-estimation with incomplete and dependent multivariate data. *Journal of Multivariate Analysis*, *176*, 104569.

Franks, W. C., & Moitra, A. (2020). Rigorous guarantees for Tyler's M-estimator via quantum expansion. In J. Abernethy & S. Agarwal (Eds.), *Proceedings of Thirty Third Conference on Learning Theory* (vol. 125). *Proceedings of Machine Learning Research* (pp. 1601–1632).

Gini, F., & Greco, M. (2002). Covariance matrix estimation for CFAR detection in correlated heavy tailed clutter. *Signal Processing*, *82*, 1847–1859.

Goes, J., Lerman, G., & Nadler, B. (2020). Robust sparse covariance estimation by thresholding Tyler's M-estimator. *The Annals of Statistics*, *48*, 86–110.

Hallin, M., Oja, H., & Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. II. Optimal R-estimation of shape. *The Annals of Statistics*, *34*, 2757–2789.

Hallin, M., & Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. I. optimal rank-based tests for sphericity. *The Annals of Statistics*, *34*, 2707–2756.

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

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Karoui, N. E. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, *19*, 2362–2405.

Kent, J. T. (1997). Data analysis for shapes and images. *Journal of Statistical Planning and Inference*, *57*, 181–193. Robust Statistics and Data Analysis, Part II.

Kent, J. T., & Tyler, D. E. (1988). Maximum likelihood estimation for the wrapped cauchy distribution. *Journal of Applied Statistics*, *15*, 247–254.

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

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibian-Barrera, M. (2018). *Robust statistics: theory and methods (with R)* (2nd ed.). Hoboken: Wiley.

Mériaux, B., Ren, C., Breloy, A., Korso, M. N. E., & Forster, P. (2021). Matched and mismatched estimation of Kronecker product of linearly structured scatter matrices under elliptical distributions. *IEEE Transactions on Signal Processing*, *69*, 603–616.

Miettinen, J., Nordhausen, K., Taskinen, S., & Tyler, D. E. (2016). On the computation of symmetrized M-estimators of scatter. In C. Agostinelli, A. Basu, P. Filzmoser, & D. Mukherjee

(Eds.), *Recent advances in robust statistics: Theory and applications* (pp. 151–167). New Delhi: Springer India.

Möttönen, J., Nordhausen, K., & Oja, H. (2010). Asymptotic theory of the spatial median. In J. Antoch, M. Huskova, & P. Sen (Eds.) *Nonparametrics and robustness in modern statistical inference and time series analysis: A Festschrift in honor of Professor Jana Jureckova* (pp. 182–193). Beachwood, Ohio, USA: Institute of Mathematical Statistics.

Möttönen, J., & Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, *5*, 201–213.

Nordhausen, K., & Oja, H. (2018a). Independent component analysis: A statistical perspective. *WIREs: Computational Statistics*, *10*, e1440.

Nordhausen, K., & Oja, H. (2018b). Robust nonparametric inference. *Annual Review of Statistics and Its Application*, *5*, 473–500.

Nordhausen, K., & Tyler, D. E. (2015). A cautionary note on robust covariance plug-in methods. *Biometrika*, *102*, 573–588.

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

Oja, H., Sirkiä, S., & Eriksson, J. (2006). Scatter matrices and independent component analysis. *Australian Journal of Statistics*, *35*, 175–189.

Ollila, E., & Koivunen, V. (2003). Robust antenna array processing using M-estimators of pseudo-covariance. In *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003* (vol. 3, pp. 2659–2663).

Ollila, E., Palomar, D. P., & Pascal, F. (2021). Shrinking the eigenvalues of M-estimators of covariance matrix. *IEEE Transactions on Signal Processing*, *69*, 256–269.

Ollila, E., & Tyler, D. E. (2012). Distribution-free detection under complex elliptically symmetric clutter distribution. In *2012 IEEE 7th sensor array and multichannel signal processing workshop (SAM)* (pp. 413–416).

Ollila, E., & Tyler, D. E. (2014). Regularized M-estimators of scatter matrix. *IEEE Transactions on Signal Processing*, *62*, 6059–6070.

Ollila, E., Tyler, D. E., Koivunen, V., & Poor, H. V. (2012). Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Transactions on Signal Processing*, *60*, 5597–5625.

Paindaveine, D. (2008). A canonical definition of shape. *Statistics & Probability Letters*, *78*, 2240–2247.

Paindaveine, D., & Van Bever, G. (2019). Tyler shape depth. *Biometrika*, *106*, 913–927.

Pascal, F., Chitour, Y., Ovarlez, J.-P., Forster, P., & Larzabal, P. (2008). Covariance structure maximum-likelihood estimates in compound gaussian noise: Existence and algorithm analysis. *IEEE Transactions on Signal Processing*, *56*, 34–48.

Pascal, F., Chitour, Y., & Quek, Y. (2014). Generalized robust shrinkage estimator and its application to STAP detection problem. *IEEE Transactions on Signal Processing*, *62*, 5640–5651.

Rublik, F. (2021). On jackknifing the symmetrized Tyler matrix. *Statistics*, *55*, 195–230.

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

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Sirkiä, S., Taskinen, S., & Oja, H. (2007). Symmetrised M-estimators of multivariate scatter. *Journal of Multivariate Analysis*, *98*, 1611–1629.

Sirkiä, S., Taskinen, S., Oja, H., & Tyler, D. E. (2009). Tests and estimates of shape based on spatial signs and ranks. *Journal of Nonparametric Statistics*, *21*, 155–176.

Soloveychik, I., & Trushin, D. (2016). Gaussian and robust Kronecker product covariance estimation: Existence and uniqueness. *Journal of Multivariate Analysis*, *149*, 92–113.

Soloveychik, I., Trushin, D., & Wiesel, A. (2016). Group symmetric robust covariance estimation. *IEEE Transactions on Signal Processing*, *64*, 244–257.

Soloveychik, I., & Wiesel, A. (2014). Tyler's covariance matrix estimator in elliptical models with convex structure. *IEEE Transactions on Signal Processing*, *62*, 5251–5259.

Soloveychik, I., & Wiesel, A. (2015). Performance analysis of Tyler's covariance estimator. *IEEE Transactions on Signal Processing*, *63*, 418–426.

Sun, Y., Babu, P., & Palomar, D. P. (2014). Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms. *IEEE Transactions on Signal Processing*, *62*, 5143–5156.

Sun, Y., Babu, P., & Palomar, D. P. (2015). Robust estimation of structured covariance matrix for heavy-tailed distributions. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5693–5697).

Sun, Y., Babu, P., & Palomar, D. P. (2016). Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Transactions on Signal Processing*, *64*, 3576–3590.

Taskinen, S., Croux, C., Kankainen, A., Ollila, E., & Oja, H. (2006). Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. *Journal of Multivariate Analysis*, *97*, 359–384.

Taskinen, S., & Oja, H. (2016). Influence functions and efficiencies of k-step Hettmansperger-Randles estimators for multivariate location and regression. In R. Y. Liu & J. W. McKean (Eds.), *Robust rank-based and nonparametric methods* (pp. 189–207). Cham: Springer International Publishing.

Taskinen, S., Sirkiä, S., & Oja, H. (2010). k-Step shape estimators based on spatial signs and ranks. *Journal of Statistical Planning and Inference*, *140*, 3376–3388.

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

Tyler, D. E. (1982). Radial estimates and the test for sphericity. *Biometrika*, *69*, 429–436.

Tyler, D. E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika*, *70*, 411–420.

Tyler, D. E. (1987a). A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics*, *15*, 234–251.

Tyler, D. E. (1987b). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika*, *74*, 579–589.

Tyler, D. E. (2010). A note on multivariate location and scatter statistics for sparse data sets. *Statistics & Probability Letters*, *80*, 1409–1413.

Tyler, D. E., & Yi, M. (2020). Breakdown points of penalized and hybrid M-estimators of covariance. arXiv (p. 2003.00078).

Wiesel, A. (2012). Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing*, *60*, 6182–6189.

Yang, L., Couillet, R., & McKay, M. (2015). A robust statistics approach to minimum variance portfolio optimization. *IEEE Transactions on Signal Processing*, *63*, 6684–6697.

Yohai, V. J., & Maronna, R. A. (1990). The maximum bias of robust covariances. *Communications in Statistics – Theory and Methods*, *19*, 3925–3933.

Zhang, T. (2015). Robust subspace recovery by Tyler's M-estimator. *Information and Inference*, 5, 1–21.

Zhang, T., Cheng, X., & Singer, A. (2016). Marčenko–Pastur law for Tyler's M-estimator. *Journal of Multivariate Analysis*, *149*, 114–123.

Part II
Multivariate Theory and Methods

On the Asymptotic Behavior of the Leading Eigenvector of Tyler's Shape Estimator Under Weak Identifiability

Davy Paindaveine and Thomas Verdebout

Abstract We consider point estimation in an elliptical Principal Component Analysis framework. More precisely, we focus on the problem of estimating the leading eigenvector $\boldsymbol{\theta}_1$ of the corresponding shape matrix. We consider this problem under asymptotic scenarios that allow the difference $r_n := \lambda_{n1} - \lambda_{n2}$ between both largest eigenvalues of the underlying shape matrix to converge to zero as the sample size n diverges to infinity. Such scenarios make the problem of estimating $\boldsymbol{\theta}_1$ challenging since this leading eigenvector is then not identifiable in the limit. In this framework, we study the asymptotic behavior of $\hat{\boldsymbol{\theta}}_1$, the leading eigenvector of Tyler's M-estimator of shape. We show that consistency and asymptotic normality survive scenarios where $\sqrt{n}r_n$ diverges to infinity as n does, although the faster the sequence (r_n) converges to zero, the poorer the corresponding consistency rate is. We also prove that consistency is lost if $r_n = O(1/\sqrt{n})$, but that $\hat{\boldsymbol{\theta}}_1$ still bears some information on $\boldsymbol{\theta}_1$ when $\sqrt{n}r_n$ converges to a positive constant. When $\sqrt{n}r_n$ diverges to infinity, we provide asymptotic confidence zones for $\boldsymbol{\theta}_1$ based on $\hat{\boldsymbol{\theta}}_1$. Our non-standard asymptotic results are supported by Monte Carlo exercises.

Keywords Principal component analysis · Point estimation · Confidence zone estimation · Robustness · Weak identifiability

1 Introduction

Many classical models in multivariate statistics include scatter parameters. The most common example is the elliptical model where observations are independent copies of a random p-vector $\mathbf{X}$ whose characteristic function is of the form

$$\mathbf{t} \mapsto e^{i\mathbf{t}'\boldsymbol{\mu}}\phi(\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}) \tag{1}$$

D. Paindaveine (✉) · T. Verdebout
ECARES and Mathematics Department, Brussels, Belgium
e-mail: Davy.Paindaveine@ulb.be; Thomas.Verdebout@ulb.be

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_3

for some *characteristic generator* $\phi : \mathbb{R}^+ \to \mathbb{R}$. Here, the p-vector $\boldsymbol{\mu}$ is a location parameter and the $p \times p$ symmetric and positive definite matrix $\boldsymbol{\Sigma}$ is a scatter parameter. A very popular instance is of course the p-variate normal model that is obtained with $\phi(s) = \exp(-s^2/2)$. Inference on the scatter parameter in the elliptical model has been the subject of many contributions: to cite only a few, Frahm (2009), Lopuhaa (1999), Tyler (1987) studied the asymptotic properties of robust estimators of $\boldsymbol{\Sigma}$, Cator and Lopuhaä (2010) provided several properties of the Minimum Covariance Determinant estimator of $\boldsymbol{\Sigma}$, sphericity tests have been studied in Hallin and Paindaveine (2006), Onatski et al. (2014), Taskinen et al. (2006) computed the influence functions of empirical canonical correlation coefficients, Hallin et al. (2010), Hallin et al. (2014), Salibián-Barrera et al. (2006), Tyler (1981), Tyler (1983) considered Principal Component Analysis based on estimators of $\boldsymbol{\Sigma}$, whereas Dürre et al. (2016), Tyler (1983) studied estimators of the eigenvalues of $\boldsymbol{\Sigma}$.

In the present paper, we consider estimation of the leading eigenvector of $\boldsymbol{\Sigma}$, that is, of the eigenvector, $\boldsymbol{\theta}_1$ say, associated with the largest eigenvalue of $\boldsymbol{\Sigma}$. This is of course the primary object of interest when conducting a Principal Component Analysis exercise. Since $\boldsymbol{\theta}_1$ does not change when $\boldsymbol{\Sigma}$ is replaced with $c\boldsymbol{\Sigma}$ for any $c > 0$, we actually want to estimate the leading eigenvector of the *shape matrix*

$$\mathbf{V} := \frac{\boldsymbol{\Sigma}}{(\det \boldsymbol{\Sigma})^{1/p}} \tag{2}$$

associated with $\boldsymbol{\Sigma}$, that is, of the version of $\boldsymbol{\Sigma}$ that is normalized to have determinant one (see Paindaveine 2008); note that this also takes care of the fact that, in (1), $\boldsymbol{\Sigma}$ was identified up to a positive scalar factor only. What makes our contribution original is that we will consider double asymptotic scenarios where, as the sample size n diverges to infinity, the underlying shape matrix $\mathbf{V} = \mathbf{V}_n$ has its two leading eigenvalues $\lambda_{n1} > \lambda_{n2}$ that satisfy $\lambda_{n1}/\lambda_{n2} \to 1$. This means that, while $\boldsymbol{\theta}_1$ is properly identifiable for any n (up to an unimportant sign change, as usual), it is no more identifiable in the limit as $n \to \infty$. Obviously, such *weak identifiability* scenarios make inference on $\boldsymbol{\theta}_1$ challenging for large n.

More precisely, we will consider throughout triangular arrays of observations $\mathbf{X}_{n1}, \ldots, \mathbf{X}_{nn}$, where, for each n, the $\mathbf{X}_{in}$'s form a random sample from the p-variate elliptical distribution with location $\boldsymbol{\mu}$, shape matrix

$$\mathbf{V}_n = \frac{\mathbf{I}_p + \delta_n \xi \boldsymbol{\theta}_1 \boldsymbol{\theta}_1'}{(\det(\mathbf{I}_p + \delta_n \xi \boldsymbol{\theta}_1 \boldsymbol{\theta}_1'))^{1/p}} = \frac{(1 + \delta_n \xi)}{(1 + \delta_n \xi)^{1/p}} \boldsymbol{\theta}_1 \boldsymbol{\theta}_1' + \frac{1}{(1 + \delta_n \xi)^{1/p}} (\mathbf{I}_p - \boldsymbol{\theta}_1 \boldsymbol{\theta}_1'), \tag{3}$$

and characteristic generator ϕ_n; in (3), $\boldsymbol{\theta}_1$ is a unit p-vector, ξ is a positive real number, δ_n is a bounded positive sequence, and $\mathbf{I}_\ell$ denotes the ℓ-dimensional identity matrix. We will denote the corresponding sequence of hypotheses as $\mathrm{P}_{\boldsymbol{\theta}_1, \delta_n, \xi, \phi_n}$. Throughout the paper, we tacitly assume that ϕ_n is such that $\mathbf{X}_{n1} \neq \mathbf{0}$ almost surely, which is needed to make Tyler's estimator of shape well-defined below. The second expression of $\mathbf{V}_n$ in (3) makes it clear that the leading eigenvalue of $\mathbf{V}_n$ is

$$\lambda_{n1} := (1+\delta_n\xi)^{(p-1)/p}, \tag{4}$$

with corresponding eigenvector $\boldsymbol{\theta}_1$, and that its remaining eigenvalues are

$$\lambda_{n2} = \ldots = \lambda_{np} := (1+\delta_n\xi)^{-1/p}, \tag{5}$$

with an eigenspace that is obviously the orthogonal complement to $\boldsymbol{\theta}_1$. If $\delta = 1$ for any n (which we will denote as $\delta \equiv 1$ in the sequel), then the classical setup in which λ_{n1} remains asymptotically well separated from the remaining eigenvalues is obtained. While we will cover this case as well, our main interest below will be on the weakly identifiable case where δ_n is $o(1)$, which provides $\lambda_{n1}/\lambda_{n2} \to 1$, hence makes $\boldsymbol{\theta}_1$ unidentifiable in the limit.

In the sequel, we will restrict to the case $\boldsymbol{\mu} = \mathbf{0}$, which is actually without any loss of generality in the distributional setup considered above. In elliptical models, the Fisher information matrix for location and shape parameters is indeed block-diagonal (see Hallin and Paindaveine 2006), which entails that asymptotic inference for the shape parameter can be conducted in the same way under specified and unspecified location (block-diagonality of the Fisher information matrix guarantees in particular that parametric efficiency bounds for shape under known and unknown $\boldsymbol{\mu}$ do coincide). In the specified location case, the results of this paper actually trivially extend to the *generalized elliptical distributions* introduced in Frahm (2004).

Quite naturally, $\boldsymbol{\theta}_1$ can be estimated by the leading eigenvector of a shape estimator $\hat{\mathbf{V}}_n$. For this purpose, we will focus in this paper on the shape estimator $\hat{\mathbf{V}}_n$ that was proposed by David Tyler in Tyler (1987). We will investigate the asymptotic behavior of the corresponding leading eigenvector $\hat{\boldsymbol{\theta}}_{n1}$ in the triangular distributional framework described above. In particular, we will show that $\hat{\boldsymbol{\theta}}_{n1}$ is consistent and asymptotically normal when $\sqrt{n}\delta_n \to \infty$, but that it is not consistent when $\delta_n = O(1/\sqrt{n})$. We will precisely derive the limiting distribution of $\hat{\boldsymbol{\theta}}_{n1}$ for any sequence (δ_n). Our results identify the same phase transitions as in the corresponding hypothesis testing framework, when testing $\mathcal{H}_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$ against $\mathcal{H}_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^0$ for some fixed unit p-vector $\boldsymbol{\theta}_1^0$; see Paindaveine et al. (2020a,b).

The rest of the paper is organized as follows: in Sect. 2, we recall the definition of Tyler's estimator of shape $\hat{\mathbf{V}}_n$ and provide its asymptotic distribution under weak identifiability. In Sect. 3, we derive the limiting behavior of $\hat{\boldsymbol{\theta}}_{n1}$ under weak identifiability and discuss the construction of confidence zones for $\boldsymbol{\theta}_1$ under sequences δ_n such that $\sqrt{n}\delta_n \to \infty$. In Sect. 4, we corroborate the results of Sect. 3 through Monte Carlo exercises. A technical appendix collects the proofs.

For the sake of convenience, we collect here the notation that will be used in the paper. Throughout, $\mathbf{e}_\ell$ will denote the ℓth vector of the canonical basis of $\mathbb{R}^p$, so that $\mathbf{K}_p := \sum_{i,j=1}^p (\mathbf{e}_i\mathbf{e}_j') \otimes (\mathbf{e}_j\mathbf{e}_i')$ is the usual *commutation matrix*. Denoting as $\operatorname{vec}\mathbf{A}$ the vector obtained by stacking the columns of the matrix $\mathbf{A}$ on top of each other, we let $\mathbf{J}_p := (\operatorname{vec}\mathbf{I}_p)(\operatorname{vec}\mathbf{I}_p)'$. We will write $\operatorname{diag}(a_1, \ldots, a_\ell)$ for the diagonal matrix collecting the real numbers $a_1, \ldots, a_\ell$ on its diagonal. For a symmetric and positive definite matrix $\mathbf{B}$, we will denote as $\mathbf{B}^{1/2}$ its symmetric and positive definite square root and as $\mathbf{B}^{-1/2}$ the inverse of this square root. Finally, $\to_{\mathcal{D}}$ will stand for convergence in distribution.

2 Tyler's Estimator of Shape Under Weak Identifiability

As explained above, we consider the problem of estimating the eigenvector $\boldsymbol{\theta}_1$ associated with the largest eigenvalue of the underlying shape matrix $\mathbf{V}_n$. To estimate $\boldsymbol{\theta}_1$, we will use the leading eigenvector $\hat{\boldsymbol{\theta}}_{n1}$ of Tyler's estimator of shape $\hat{\mathbf{V}}_n$ from Tyler (1987). Under specified location $\boldsymbol{\mu} = \mathbf{0}$, this shape estimator $\hat{\mathbf{V}}_n$ is defined as the solution of

$$\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{X}_{ni}\mathbf{X}'_{ni}}{\mathbf{X}'_{ni}\hat{\mathbf{V}}_n^{-1}\mathbf{X}_{ni}} = \hat{\mathbf{V}}_n, \tag{6}$$

normalized to have unit determinant. This can be seen as the estimator of shape for which the directions (or *spatial signs*) of the resulting sphericized observations

$$\frac{\hat{\mathbf{V}}_n^{-1/2}\mathbf{X}_{n1}}{\|\hat{\mathbf{V}}_n^{-1/2}\mathbf{X}_{n1}\|}, \ldots, \frac{\hat{\mathbf{V}}_n^{-1/2}\mathbf{X}_{nn}}{\|\hat{\mathbf{V}}_n^{-1/2}\mathbf{X}_{nn}\|}$$

have an empirical covariance matrix (with respect to specified location $\boldsymbol{\mu} = \mathbf{0}$) equal to $(1/p)\mathbf{I}_p$. Tyler's estimator of shape enjoys many nice properties. In particular, it is distribution-free in the (centered) elliptical model and it is consistent and asymptotically normal under a broad range of distributions without moment assumptions; see Tyler (1987). Distribution-freeness is an important property since it entails that the distribution of $\hat{\boldsymbol{\theta}}_{n1}$ does not depend on the underlying characteristic generator ϕ_n, that is, it does not depend on the type of elliptical distribution at hand (normal, t, etc.) nor on the scale of this elliptical distribution.

The following result provides the asymptotic distribution of Tyler's estimator of shape in a framework where $\boldsymbol{\theta}_1$ is possibly weakly identifiable.

Proposition 1 *Fix a unit vector $\boldsymbol{\theta}_1$, a positive real number ξ and a sequence (δ_n) that either is $\delta_n \equiv 1$ or is $o(1)$. Let $(\mathbf{V}_n)$ be the resulting sequence of shape matrices in (3). Let further (ϕ_n) be a sequence of characteristic generators. Then,*

$$\sqrt{n}\,\operatorname{vec}(\hat{\mathbf{V}}_n - \mathbf{V}_n) \to_{\mathcal{D}} \mathcal{N}\Big(\mathbf{0}, \Big(1+\frac{2}{p}\Big)\Big\{(\mathbf{I}_{p^2} + \mathbf{K}_p)(\mathbf{V}\otimes\mathbf{V}) - \frac{2}{p}\operatorname{vec}(\mathbf{V})\operatorname{vec}'(\mathbf{V})\Big\}\Big)$$

under $\mathrm{P}_{\boldsymbol{\theta}_1,\delta_n,\xi,\phi_n}$ as $n \to \infty$, where $\mathbf{V}$ is the limit of $(\mathbf{V}_n)$.

This result shows that root-n consistency of Tyler's estimator of shape $\hat{\mathbf{V}}_n$ is robust to arbitrarily weakly identifiable scenarios, that is, to scenarios where (δ_n) converges to zero arbitrarily fast. As we will show in the next section, this is not the case for the leading eigenvector of $\hat{\mathbf{V}}_n$.

3 Asymptotic Behavior of Tyler's Leading Eigenvector Under Weak Identifiability

The main goal of this section is to derive the asymptotic behavior of the leading eigenvector $\hat{\boldsymbol{\theta}}_{n1}$ of $\hat{\mathbf{V}}_n$ under weak identifiability. Denoting as $\hat{\lambda}_{nj}$, $j = 1, \ldots, p$, the eigenvalues of $\hat{\mathbf{V}}_n$ in decreasing order (these sample eigenvalues are pairwise different almost surely), we first provide the following result that shows that root-n consistency of these eigenvalues is robust to weak identifiability.

Proposition 2 *Fix a unit vector $\boldsymbol{\theta}_1$, a positive real number ξ and a sequence (δ_n) that either is $\delta_n \equiv 1$ or is $o(1)$. Let (ϕ_n) be a sequence of characteristic generators. Then, for any $j = 1, \ldots, p$, $\sqrt{n}(\hat{\lambda}_{nj} - \lambda_{nj})$ is $O_{\mathrm{P}}(1)$ as $n \to \infty$ under $\mathrm{P}_{\boldsymbol{\theta}_1,\delta_n,\xi,\phi_n}$.*

With $\boldsymbol{\theta}_1$ fixed, pick arbitrarily p-vectors $\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p$ such that $\boldsymbol{\Gamma} := (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)$ is orthogonal. Let further $\hat{\boldsymbol{\Gamma}}_n := (\hat{\boldsymbol{\theta}}_{n1}, \ldots, \hat{\boldsymbol{\theta}}_{np})$ stand for the orthogonal matrix whose jth column vector is an eigenvector of $\hat{\mathbf{V}}_n$ associated with eigenvalue $\hat{\lambda}_{nj}$. To unambiguously fix the "signs" of $\hat{\boldsymbol{\theta}}_{nj}$, $j = 1, \ldots, p$, we impose that, with probability one, all entries in the first column of

$$\mathbf{E}_n := \hat{\boldsymbol{\Gamma}}_n' \boldsymbol{\Gamma} = \begin{pmatrix} E_{n,11} & \mathbf{E}_{n,12} \\ \mathbf{E}_{n,21} & \mathbf{E}_{n,22} \end{pmatrix} \tag{7}$$

are positive (note that all entries of $\mathbf{E}_n$ are almost surely non-zero). The following result then provides the asymptotic behavior of $\mathbf{E}_n$ in the present context.

Proposition 3 *Fix a unit vector $\boldsymbol{\theta}_1$, a positive real number ξ and a sequence (δ_n) that either is $\delta_n \equiv 1$ or is $o(1)$. Let (ϕ_n) be a sequence of characteristic generators. Let $\mathbf{Z}$ be a $p \times p$ random matrix such that*

$$\mathrm{vec}(\mathbf{Z}) \sim \mathcal{N}\Big(\mathbf{0}, \Big(1 + \frac{2}{p}\Big)\Big\{(\mathbf{I}_{p^2} + \mathbf{K}_p) - \frac{2}{p}\mathbf{J}_p\Big\}\Big),$$

and let $\mathbf{E}(\xi) := (\mathbf{w}_1(\xi), \ldots, \mathbf{w}_p(\xi))'$, where $\mathbf{w}_j(\xi) = (w_{j1}(\xi), \ldots, w_{jp}(\xi))'$ is the unit eigenvector associated with the jth largest eigenvalue of $\mathbf{Z} + \mathrm{diag}(\xi, 0, \ldots, 0)$ and such that $w_{j1}(\xi) > 0$ almost surely. Then, we have the following as $n \to \infty$

under $\mathrm{P}_{\boldsymbol{\theta}_1,\delta_n,\xi,\phi_n}$*:*

(i) If $\delta_n \equiv 1$*, then* $n(E_{n,11}-1) = O_{\mathrm{P}}(1)$*,* $\mathbf{E}_{n,22}\mathbf{E}'_{n,22} = \mathbf{I}_{p-1} + o_{\mathrm{P}}(1)$*,* $\sqrt{n}\mathbf{E}_{n,21} = O_{\mathrm{P}}(1)$*, and both* $\sqrt{n}\mathbf{E}'_{n,22}\mathbf{E}_{n,21}$ *and* $\sqrt{n}\mathbf{E}'_{n,12}$ *are asymptotically normal with mean zero and covariance matrix* $\xi^{-2}(1+\xi)(1+\frac{2}{p})\mathbf{I}_{p-1}$*.*

(ii) If δ_n *is* $o(1)$ *with* $\sqrt{n}\delta_n \to \infty$*, then* $n\delta_n^2(E_{n,11}-1) = O_{\mathrm{P}}(1)$*,* $\mathbf{E}_{n,22}\mathbf{E}'_{n,22} = \mathbf{I}_{p-1} + o_{\mathrm{P}}(1)$*,* $\sqrt{n}\delta_n\mathbf{E}_{n,21} = O_{\mathrm{P}}(1)$*, and both* $\sqrt{n}\delta_n\mathbf{E}'_{n,22}\mathbf{E}_{n,21}$ *and* $\sqrt{n}\delta_n\mathbf{E}'_{n,12}$ *are asymptotically normal with mean zero and covariance matrix* $\xi^{-2}(1+\frac{2}{p})\mathbf{I}_{p-1}$*.*

(iii) If $\delta_n = 1/\sqrt{n}$*, then* $\mathbf{E}_n$ *converges weakly to* $\mathbf{E}(\xi)$*.*

(iv) If $\delta_n = o(1/\sqrt{n})$*, then* $\mathbf{E}_n$ *converges weakly to* $\mathbf{E} := \mathbf{E}(0)$*.*

The four regimes (i)–(iv) identified in this result will play a crucial role in the asymptotic behavior of $\hat{\boldsymbol{\theta}}_{n1}$ below. At this point, let us note that, in regimes (i)–(ii),

$$\|\sqrt{n}\delta_n(\hat{\boldsymbol{\theta}}_{n1}-\boldsymbol{\theta}_1)\|^2 = 2n\delta_n^2(1-\hat{\boldsymbol{\theta}}'_{n1}\boldsymbol{\theta}_1) = 2n\delta_n^2(1-E_{n,11}) = O_{\mathrm{P}}(1); \tag{8}$$

this is compatible with the well-known $\sqrt{n}$-consistency of $\hat{\boldsymbol{\theta}}_{n1}$ in the classical case obtained with $\delta_n \equiv 1$ and suggests that $\sqrt{n}$-consistency deteriorates into $(\sqrt{n}\delta_n)$-consistency in regime (ii), which in turn suggests that consistency is lost in regime (iii). The following result, which is the main result of the paper, shows that this is precisely what happens.

Theorem 1 *Fix a unit vector* $\boldsymbol{\theta}_1$*, a positive real number* ξ *and a sequence* (δ_n) *that either is* $\delta_n \equiv 1$ *or is* $o(1)$*. Let* (ϕ_n) *be a sequence of characteristic generators. Then, the leading eigenvector* $\hat{\boldsymbol{\theta}}_{n1}$ *of Tyler's estimator of shape satisfies the following as* $n \to \infty$ *under* $\mathrm{P}_{\boldsymbol{\theta}_1,\delta_n,\xi,\phi_n}$*:*

(i) If $\delta_n \equiv 1$*, then* $\sqrt{n}(\hat{\boldsymbol{\theta}}_{n1}-\boldsymbol{\theta}_1)$ *is asymptotically normal with mean zero and covariance matrix*

$$\frac{1}{\xi^2}(1+\xi)\left(1+\frac{2}{p}\right)(\mathbf{I}_p-\boldsymbol{\theta}_1\boldsymbol{\theta}'_1) = \left(1+\frac{2}{p}\right)\frac{\lambda_1\lambda_2}{(\lambda_1-\lambda_2)^2}(\mathbf{I}_p-\boldsymbol{\theta}_1\boldsymbol{\theta}'_1),$$

where λ_1 *and* λ_2 *are the eigenvalues in (4)–(5) with* $\delta_n \equiv 1$*.*

(ii) If δ_n *is* $o(1)$ *with* $\sqrt{n}\delta_n \to \infty$*, then* $\sqrt{n}\delta_n(\hat{\boldsymbol{\theta}}_{n1}-\boldsymbol{\theta}_1)$ *is asymptotically normal with mean zero and covariance matrix*

$$\frac{1}{\xi^2}\left(1+\frac{2}{p}\right)(\mathbf{I}_p-\boldsymbol{\theta}_1\boldsymbol{\theta}'_1). \tag{9}$$

(iii) If $\delta_n = 1/\sqrt{n}$*, then* $\hat{\boldsymbol{\theta}}_{n1}$ *converges weakly to the unit eigenvector associated with the largest eigenvalue of* $\mathbf{Z} + \xi\boldsymbol{\theta}_1\boldsymbol{\theta}'_1$*, where* $\mathbf{Z}$ *is as in the statement of Proposition 3.*

(iv) If $\delta_n = o(1/\sqrt{n})$, then $\hat{\boldsymbol{\theta}}_{n1}$ converges weakly to a random vector that is uniformly distributed over the unit sphere $\mathcal{S}^{p-1}$.

This result confirms that, while the consistency rate of $\hat{\boldsymbol{\theta}}_{n1}$ is of course $\sqrt{n}$ in the standard case $\delta_n \equiv 1$, this consistency rate goes down to $\sqrt{n}\delta_n$ when $\delta_n \to 0$ with $\sqrt{n}\delta_n \to \infty$. Asymptotic normality is obtained in both cases. In the threshold regime (iii) obtained with $\delta_n = 1/\sqrt{n}$, the estimator $\hat{\boldsymbol{\theta}}_{n1}$ is no more consistent for $\boldsymbol{\theta}_1$, yet it still bears some information on $\boldsymbol{\theta}_1$. Clearly, the larger ξ, the larger this information (in particular, the weak limit of $\hat{\boldsymbol{\theta}}_{n1}$ converges to the Dirac distribution at $\boldsymbol{\theta}_1$ as $\xi \to \infty$). Finally, if $\delta_n = o(1/\sqrt{n})$, then $\hat{\boldsymbol{\theta}}_{n1}$ behaves asymptotically as a random vector that is uniformly distributed on the unit sphere of $\mathbb{R}^p$, hence does not bear any information on $\boldsymbol{\theta}_1$. Incidentally, we stress that since $\boldsymbol{\theta}_{n1}$ (resp., $\hat{\boldsymbol{\theta}}_{n1}$) is a homogenous function of $\mathbf{V}_n$ (resp., $\hat{\mathbf{V}}_n$), Theorem 1 still holds true if, in (2), $\mathbf{V}_n$ is rather normalized so that it has trace p, or so that its upper-left entry is equal to one, etc.

The results in Theorem 1 allow one to build confidence zones for $\boldsymbol{\theta}_1$. Let us start with regime (i). Since the sample eigenvalues $\hat{\lambda}_{nj}$, $j = 1, 2$, are $\sqrt{n}$-consistent, confidence zones for $\boldsymbol{\theta}_1$ with asymptotic confidence level $1 - \alpha$ in this regime are given by

$$C_n^{1-\alpha} := \left\{ \boldsymbol{\theta}_1 \in \mathcal{S}^{p-1} : n\left(1+\frac{2}{p}\right)^{-1} \frac{(\hat{\lambda}_1 - \hat{\lambda}_2)^2}{\hat{\lambda}_1\hat{\lambda}_2} \hat{\boldsymbol{\theta}}'_{n1}(\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}'_1)\hat{\boldsymbol{\theta}}_{n1} \leq \chi^2_{p-1,1-\alpha} \right\},$$

where $\chi^2_{p-1,1-\alpha}$ denotes the upper-α quantile of the chi-square distribution with $p-1$ degrees of freedom. Now, in regime (ii),

$$\begin{aligned}
&\frac{(\hat{\lambda}_{n1} - \hat{\lambda}_{n2})}{\sqrt{\hat{\lambda}_{n1}\hat{\lambda}_{n2}}}\sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)\\
&\quad = \frac{\sqrt{n}(\hat{\lambda}_{n1} - \lambda_{n1}) - \sqrt{n}(\hat{\lambda}_{n2} - \lambda_{n2}) + \sqrt{n}(\lambda_{n1} - \lambda_{n2})}{\sqrt{\hat{\lambda}_{n1}\hat{\lambda}_{n2}}}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)\\
&\quad = \frac{\sqrt{n}(\lambda_{n1} - \lambda_{n2})}{\sqrt{\lambda_{n1}\lambda_{n2}}}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) + o_{\mathrm{P}}(1) = \delta_n\xi(1 + o(1))\sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) + o_{\mathrm{P}}(1)\\
&\quad \to_{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \left(1 + \frac{2}{p}\right)(\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}'_1)\right),
\end{aligned}$$

where we used the fact that $\sqrt{n}(\hat{\lambda}_{nj} - \lambda_{nj})$, $j = 1, 2$, are still $O_{\mathrm{P}}(1)$ in this regime (Proposition 2). A direct consequence is that the asymptotic confidence zones $C_n^{1-\alpha}$ above are still valid in regime (ii).

To conclude this section, we turn to robustness issues by considering the influence function of $\hat{\boldsymbol{\theta}}_{n1}$ in regimes (i)–(ii). Using (27) (resp., (28)) in regime (i) (resp., regime (ii)), jointly with (20), (23) and the fact that $E_{n,11} = 1 + o_{\mathrm{P}}(1)$ in regimes (i)–(ii), we obtain

$$\begin{aligned}\sqrt{n}\delta_n(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) &= (\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)\sqrt{n}\delta_n E_{n,11}\mathbf{E}'_{n,12} + o_{\mathrm{P}}(1)\\ &= -(\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)\sqrt{n}\delta_n \mathbf{E}'_{n,22}\mathbf{E}_{n,21} + o_{\mathrm{P}}(1)\\ &= (\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)\xi^{-1}(1+\delta_n\xi)^{1/p}(\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)'\sqrt{n}(\hat{\mathbf{V}}_n - \mathbf{V}_n)\boldsymbol{\theta}_1 + o_{\mathrm{P}}(1)\\ &= \xi^{-1}(1+\delta_n\xi)^{1/p}(\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}'_1)\sqrt{n}(\hat{\mathbf{V}}_n - \mathbf{V}_n)\boldsymbol{\theta}_1 + o_{\mathrm{P}}(1). \qquad (10)\end{aligned}$$

From (14) and (16) in the proof of Proposition 1, we have

$$\begin{aligned}&\sqrt{n}\,\mathrm{vec}\big(\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2} - \mathbf{I}_p\big)\\ &\qquad= \left(\mathbf{I}_{p^2} - \frac{1}{p}\mathbf{J}_p\right)\sqrt{n}\,\mathrm{vec}\left(\frac{p\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}}{\mathrm{tr}[\mathbf{V}_n^{-1}\hat{\mathbf{V}}_n]} - \mathbf{I}_p\right) + o_{\mathrm{P}}(1)\\ &\qquad= (p+2)\left(\mathbf{I}_{p^2} - \frac{1}{p}\mathbf{J}_p\right)\sqrt{n}\,\mathrm{vec}\big(\mathbf{S}_n(\mathbf{V}_n) - \tfrac{1}{p}\mathbf{I}_p\big) + o_{\mathrm{P}}(1),\end{aligned}$$

which yields

$$\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{V}}_n - \mathbf{V}_n) = (p+2)\big(\mathbf{V}_n^{1/2}\otimes\mathbf{V}_n^{1/2}\big)\left(\mathbf{I}_{p^2} - \frac{1}{p}\mathbf{J}_p\right)\sqrt{n}\,\mathrm{vec}\big(\mathbf{S}_n(\mathbf{V}_n) - \tfrac{1}{p}\mathbf{I}_p\big) + o_{\mathrm{P}}(1).$$

Since

$$(\boldsymbol{\theta}'_1 \otimes (\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}'_1))\big(\mathbf{V}_n^{1/2}\otimes\mathbf{V}_n^{1/2}\big) = \frac{(1+\delta_n\xi)^{1/2}}{(1+\delta_n\xi)^{1/p}}(\boldsymbol{\theta}'_1 \otimes (\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}'_1))$$

(which in particular entails that $(\boldsymbol{\theta}'_1\otimes(\mathbf{I}_p-\boldsymbol{\theta}_1\boldsymbol{\theta}'_1))\big(\mathbf{V}_n^{1/2}\otimes\mathbf{V}_n^{1/2}\big)(\mathrm{vec}\,\mathbf{I}_p) = \mathbf{0}$), plugging this in (10) then provides

$$\begin{aligned}\sqrt{n}\delta_n(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) &= \frac{(p+2)(1+\delta_n\xi)^{1/2}}{\xi}(\boldsymbol{\theta}'_1 \otimes (\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}'_1))\sqrt{n}\,\mathrm{vec}(\mathbf{S}_n(\mathbf{V}_n)) + o_{\mathrm{P}}(1)\\ &= \frac{(p+2)(1+\delta_n\xi)^{1/2}}{\xi\sqrt{n}}(\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}'_1)\sum_{i=1}^n \frac{\mathbf{V}_n^{-1/2}\mathbf{X}_{ni}\mathbf{X}'_{ni}\mathbf{V}_n^{-1/2}}{\|\mathbf{V}_n^{-1/2}\mathbf{X}_{ni}\|^2}\boldsymbol{\theta}_1 + o_{\mathrm{P}}(1).\end{aligned}$$

By applying the multivariate central limit theorem (and (15)), it is readily checked that this Bahadur representation result for $\sqrt{n}\delta_n(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)$ is compatible with the asymptotic normality statements in Theorem 1(i)–(ii). More importantly, this

Bahadur representation result shows that the boundedness of the influence function of $\hat{\boldsymbol{\theta}}_{n1}$ does not only hold in the standard regime (i) but also in the weakly identifiable regime (ii).

4 Numerical Illustration

In this section, we conduct Monte Carlo simulation exercises to validate the various asymptotic results in Theorem 1. For any $\ell \in \{0, 1, \ldots, 7\}$, we generated $M = 10,000$ independent random samples of size $n = 100,000$ from the bivariate ($p = 2$) normal distribution with mean vector zero and covariance matrix

$$\boldsymbol{\Sigma}_{n,\ell} = \mathbf{I}_2 + \delta_{n,\ell}\xi\boldsymbol{\theta}_1\boldsymbol{\theta}_1', \tag{11}$$

with $\delta_{n,\ell} = n^{-\ell/8}$, $\xi = 2$ and $\boldsymbol{\theta}_1 = \mathbf{e}_1 \in \mathbb{R}^2$. In each of these samples, we computed the leading eigenvector $\hat{\boldsymbol{\theta}}_{n1}$ of Tyler's estimator of scatter (still with respect to fixed location at the origin of $\mathbb{R}^p$); evaluation of Tyler's estimator of scatter was done by using the function `tyler.shape` from the R package *ICSNP* (Nordhausen et al. 2018). We first focus on Theorem 1(i)–(ii), hence on the cases $\ell \in \{0, 1, 2, 3\}$. For each such ℓ, we provide in Fig. 1 a histogram of the M corresponding values of

$$\sqrt{n}\delta_{n,\ell}\mathbf{e}_2'\hat{\boldsymbol{\theta}}_{n1} = \sqrt{n}\delta_{n,\ell}\mathbf{e}_2'(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1). \tag{12}$$

Clearly, the results nicely agree with the corresponding asymptotic distribution of (12) in Theorem 1, namely $\mathcal{N}(0, \frac{3}{2})$ for $\ell = 0$ (regime (i)) and $\mathcal{N}(0, \frac{1}{2})$ for $\ell = 1, 2, 3$ (regime (ii)). We then turn to Theorem 1(iii)–(iv), hence to the cases $\ell \in \{4, 5, 6, 7\}$. For these values of ℓ, Fig. 2 reports histograms of

$$\mathbf{e}_2'\hat{\boldsymbol{\theta}}_{n1}. \tag{13}$$

Here, the asymptotic distributions of (13) in Theorem 1(iii)–(iv) do not have a closed form density, and we are therefore plotting kernel density estimates obtained from a random sample of size 10^6 from the weak limit of (13) in Theorem 1(iii)–(iv). To avoid boundary effects (the support of this weak limit is of course $[-1, 1]$), we employed the function `kde.boundary` from the R package *ks* (Duong 2021) with default parameters, which returns the kernel density estimate using the second form of the Beta boundary kernel in Chen (1999). Irrespective of $\ell \in \{4, 5, 6, 7\}$, these empirical results fully support the corresponding asymptotic results in Theorem 1.

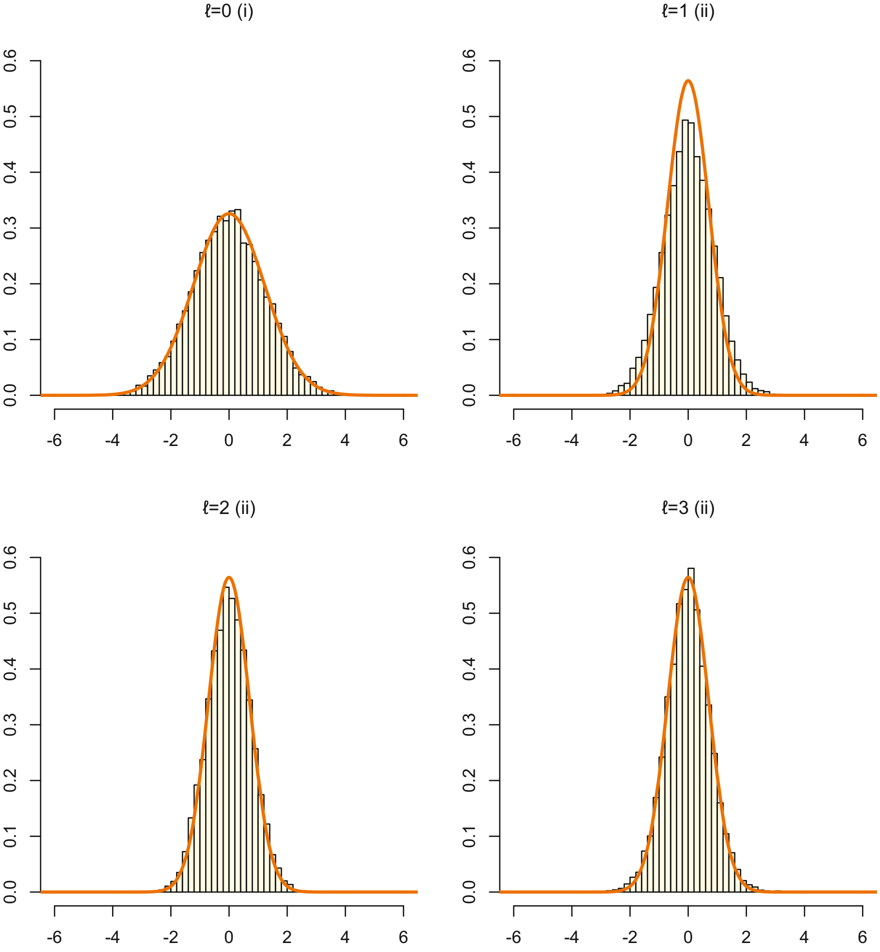


Fig. 1 For each $\ell \in \{0, 1, 2, 3\}$, histograms of the quantities $\sqrt{n}\delta_{n,\ell}\mathbf{e}_2'(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)$ computed from $M = 10{,}000$ independent random samples of size $n = 100{,}000$ from the bivariate normal distribution with mean vector zero and the covariance matrix $\boldsymbol{\Sigma}_{n,\ell}$ in (11), where $\hat{\boldsymbol{\theta}}_{n1}$ denotes the leading eigenvector of Tyler's estimator of scatter (with respect to fixed location at the origin of $\mathbb{R}^2$). In each panel, the solid curve is the density of the corresponding asymptotic distribution, namely $\mathcal{N}(0, \frac{3}{2})$ for $\ell = 0$ and $\mathcal{N}(0, \frac{1}{2})$ for $\ell = 1, 2, 3$

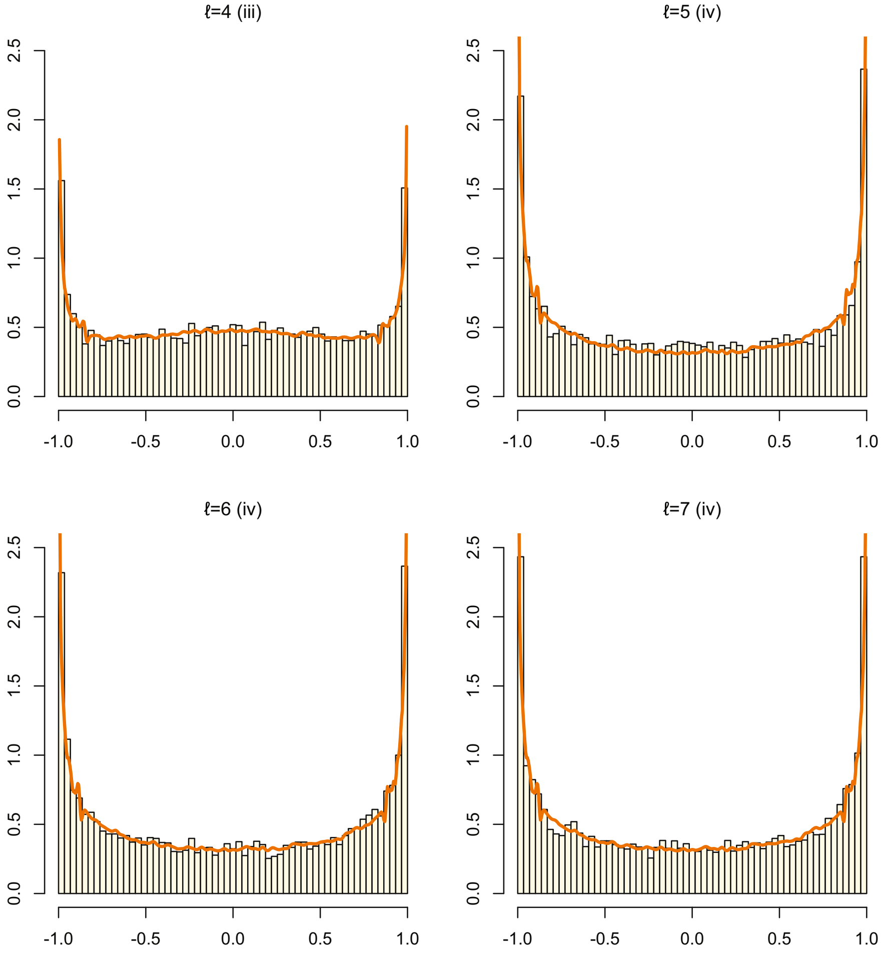


Fig. 2 For each $\ell \in \{4, 5, 6, 7\}$, histograms of the quantities $\mathbf{e}_2'\hat{\boldsymbol{\theta}}_{n1}$ computed from $M = 10{,}000$ independent random samples of size $n = 100{,}000$ from the bivariate normal distribution with mean vector zero and the covariance matrix $\boldsymbol{\Sigma}_{n,\ell}$ in (11), where $\hat{\boldsymbol{\theta}}_{n1}$ still denotes the leading eigenvector of Tyler's estimator of scatter (with respect to fixed location at the origin of $\mathbb{R}^2$). In each panel, the solid curve is a kernel estimate for the density of the corresponding weak limit obtained from Theorem 1(iii)–(iv); see Sect. 4 for details

Acknowledgments This research is supported by the Program of Concerted Research Actions (ARC) of the Université libre de Bruxelles.

Appendix

The proof of Proposition 1 requires the following preliminary result, which follows from (3.7)–(3.8) in Tyler (1987).

Lemma 1 *Fix a unit vector $\boldsymbol{\theta}_1$, a positive real number ξ and a sequence (δ_n) that either is $\delta_n \equiv 1$ or is $o(1)$. Let $(\mathbf{V}_n)$ be the resulting sequence of shape matrices in (3). Let further (ϕ_n) be a sequence of characteristic generators. Then, letting*

$$\mathbf{G}_p := \mathbf{I}_{p^2} - \tfrac{1}{p+2}(\mathbf{I}_{p^2} + \mathbf{K}_p - \mathbf{J}_p) \ \ and \ \ \mathbf{S}_n(\mathbf{V}) := \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{V}^{-1/2}\mathbf{X}_{ni}\mathbf{X}'_{ni}\mathbf{V}^{-1/2}}{\|\mathbf{V}^{-1/2}\mathbf{X}_{ni}\|^2},$$

we have that

$$\mathbf{G}_p\sqrt{n}\,\mathrm{vec}\bigg(\frac{p\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}}{\mathrm{tr}[\mathbf{V}_n^{-1}\hat{\mathbf{V}}_n]} - \mathbf{I}_p\bigg) = p\sqrt{n}\,\mathrm{vec}\big(\mathbf{S}_n(\mathbf{V}_n) - \tfrac{1}{p}\mathbf{I}_p\big) + o_{\mathrm{P}}(1),$$

under $\mathrm{P}_{\boldsymbol{\theta}_1,\delta_n,\xi,\phi_n}$ as $n \to \infty$.

In all proofs below, stochastic convergences are as $n \to \infty$ under $\mathrm{P}_{\boldsymbol{\theta}_1,\delta_n,\xi,\phi_n}$.

Proof of Proposition 1 Letting $\mathbf{H}_p := \mathbf{I}_{p^2} + \mathbf{K}_p - \frac{2}{p}\mathbf{J}_p$, we have $\mathbf{H}_p\mathbf{J}_p = \mathbf{0}$ and $\mathbf{H}_p\mathbf{K}_p(\mathrm{vec}\,\mathbf{B}) = \mathbf{H}_p(\mathrm{vec}\,\mathbf{B})$ for any symmetric matrix $\mathbf{B}$, so that

$$\mathbf{H}_p\mathbf{G}_p(\mathrm{vec}\,\mathbf{B}) = (\mathbf{H}_p - \tfrac{2}{p+2}\mathbf{H}_p)(\mathrm{vec}\,\mathbf{B}) = \tfrac{p}{p+2}\mathbf{H}_p(\mathrm{vec}\,\mathbf{B})$$

for any symmetric matrix $\mathbf{B}$. Lemma 1 thus yields that

$$\tfrac{p}{p+2}\mathbf{H}_p\sqrt{n}\,\mathrm{vec}\bigg(\frac{p\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}}{\mathrm{tr}[\mathbf{V}_n^{-1}\hat{\mathbf{V}}_n]} - \mathbf{I}_p\bigg) = p\sqrt{n}\,\mathbf{H}_p\mathrm{vec}\big(\mathbf{S}_n(\mathbf{V}_n) - \tfrac{1}{p}\mathbf{I}_p\big) + o_{\mathrm{P}}(1).$$

Using the fact that $\mathbf{J}_p(\mathrm{vec}\,\mathbf{B}) = (\mathrm{tr}[\mathbf{B}])(\mathrm{vec}\,\mathbf{I}_k)$ and $\mathbf{K}_p(\mathrm{vec}\,\mathbf{B}) = \mathrm{vec}\,\mathbf{B}$ for any symmetric matrix $\mathbf{B}$, this rewrites

$$\sqrt{n}\,\mathrm{vec}\bigg(\frac{p\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}}{\mathrm{tr}[\mathbf{V}_n^{-1}\hat{\mathbf{V}}_n]} - \mathbf{I}_p\bigg) = (p+2)\sqrt{n}\,\mathrm{vec}\big(\mathbf{S}_n(\mathbf{V}_n) - \tfrac{1}{p}\mathbf{I}_p\big) + o_{\mathrm{P}}(1). \qquad (14)$$

Now, Lemma A.3(ii) from Paindaveine et al. (2020a) states that

$$\sqrt{n}\,\mathrm{vec}\big(\mathbf{S}_n(\mathbf{V}_n) - \tfrac{1}{p}\mathbf{I}_p\big) \to_{\mathcal{D}} \mathcal{N}\bigg(\mathbf{0}, \frac{1}{p(p+2)}\mathbf{H}_p\bigg), \qquad (15)$$

so that

$$\sqrt{n}\,\mathrm{vec}\bigg(\frac{p\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}}{\mathrm{tr}[\mathbf{V}_n^{-1}\hat{\mathbf{V}}_n]} - \mathbf{I}_p\bigg) \to_{\mathcal{D}} \mathcal{N}\bigg(\mathbf{0}, \Big(1 + \frac{2}{p}\Big)\mathbf{H}_p\bigg).$$

Hence, the argument in the bottom of page 341 in Hallin and Paindaveine (2006) yields that

$$\sqrt{n}\,\mathrm{vec}\bigg(\frac{\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}}{(\det(\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}))^{1/p}}-\mathbf{I}_p\bigg)$$

$$=\Big(\mathbf{I}_{p^2}-\frac{1}{p}\mathbf{J}_p\Big)\sqrt{n}\,\mathrm{vec}\bigg(\frac{p\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}}{\mathrm{tr}[\mathbf{V}_n^{-1}\hat{\mathbf{V}}_n]}-\mathbf{I}_p\bigg)+o_{\mathrm{P}}(1) \tag{16}$$

$$\to_{\mathcal{D}} \mathcal{N}\Big(\mathbf{0},\,\frac{1}{p(p+2)}\mathbf{H}_p\Big),$$

that is,

$$\sqrt{n}\,\mathrm{vec}\big(\mathbf{V}_n^{-1/2}\hat{\mathbf{V}}_n\mathbf{V}_n^{-1/2}-\mathbf{I}_p\big)\to_{\mathcal{D}} \mathcal{N}\Big(\mathbf{0},\,\frac{1}{p(p+2)}\mathbf{H}_p\Big).$$

Since this rewrites

$$\big(\mathbf{V}_n^{-1/2}\otimes\mathbf{V}_n^{-1/2}\big)\sqrt{n}\,\mathrm{vec}\big(\hat{\mathbf{V}}_n-\mathbf{V}_n\big)\to_{\mathcal{D}} \mathcal{N}\Big(\mathbf{0},\,\frac{1}{p(p+2)}\mathbf{H}_p\Big),$$

we finally obtain that

$$\sqrt{n}\,\mathrm{vec}\big(\hat{\mathbf{V}}_n-\mathbf{V}_n\big)\to_{\mathcal{D}} \mathcal{N}\bigg(\mathbf{0},\,\Big(1+\frac{2}{p}\Big)\Big\{(\mathbf{I}_{p^2}+\mathbf{K}_p)(\mathbf{V}\otimes\mathbf{V})-\frac{2}{p}\mathrm{vec}(\mathbf{V})\mathrm{vec}'(\mathbf{V})\Big\}\bigg),$$

with $\mathbf{V}$ the limiting value of $(\mathbf{V}_n)$. □

We do not prove Proposition 2 here since the proof follows along the exact same lines as the proof of Lemma 2.2 in Paindaveine et al. (2020b). We thus turn to the proof of Proposition 3 that requires the following linear algebra result.

Lemma 2 *Let $\mathbf{A}$ be a $p\times p$ matrix. Assume that λ is an eigenvalue of $\mathbf{A}$ and that the corresponding eigenspace V_λ has dimension one. Denoting as $C=(C_{ij})$ the cofactor matrix of $\mathbf{A}-\lambda\mathbf{I}_p$, assume that $\mathbf{v}:=(C_{11},\ldots,C_{1p})'\neq\mathbf{0}$. Then $V_\lambda=\{t\mathbf{v}:t\in\mathbb{R}\}$.*

Proof of Lemma 2 For any $j=1,\ldots,p$, denote as $(\mathbf{A}-\lambda\mathbf{I}_p)_j$ the jth row of $\mathbf{A}-\lambda\mathbf{I}_p$. For $j=2,\ldots,p$,

$$(\mathbf{A}-\lambda\mathbf{I}_p)_j v=\det\begin{pmatrix}(\mathbf{A}-\lambda\mathbf{I}_p)_j\\(\mathbf{A}-\lambda\mathbf{I}_p)_2\\\vdots\\(\mathbf{A}-\lambda\mathbf{I}_p)_p\end{pmatrix}=0,$$

since this is the determinant of a matrix with (at least) twice the same row. Since λ is an eigenvalue of $\mathbf{A}$, this determinant is also zero for $j = 1$. Therefore, $(\mathbf{A} - \lambda \mathbf{I}_p)\mathbf{v} = 0$. The non-zero vector $\mathbf{v}$ thus belongs to V_λ. Since V_λ has dimension one by assumption, the result follows.

Proof of Proposition 3 In this proof, we put

$$\mathbf{Z}_n := \sqrt{n}\boldsymbol{\Gamma}'(\hat{\mathbf{V}}_n - \mathbf{V}_n)\boldsymbol{\Gamma}, \tag{17}$$

and $\boldsymbol{\Lambda}_n := \boldsymbol{\Gamma}'\mathbf{V}_n\boldsymbol{\Gamma} = \mathrm{diag}(\lambda_{n1}, \ldots, \lambda_{np})$. First note that since

$$\mathbf{E}_n = \hat{\boldsymbol{\Gamma}}_n'\boldsymbol{\Gamma} = \begin{pmatrix} E_{n,11} & \mathbf{E}_{n,12} \\ \mathbf{E}_{n,21} & \mathbf{E}_{n,22} \end{pmatrix}$$

is an orthogonal matrix, we easily obtain that

$$\mathbf{E}_{n,21} = -\frac{1}{E_{n,11}}\mathbf{E}_{n,22}\mathbf{E}_{n,12}', \tag{18}$$

$$\mathbf{E}_{n,22}\mathbf{E}_{n,22}' = \mathbf{I}_{p-1} - \mathbf{E}_{n,21}\mathbf{E}_{n,21}' \tag{19}$$

and

$$E_{n,11}\mathbf{E}_{n,12}' = -\mathbf{E}_{n,22}'\mathbf{E}_{n,21}. \tag{20}$$

We start with the proof of (i)–(ii). The random matrix $\mathbf{Y}_n := \sqrt{n}\boldsymbol{\Gamma}'\hat{\mathbf{V}}_n\boldsymbol{\Gamma} - \sqrt{n}\lambda_{n1}\mathbf{I}_p$ admits the eigenvectors $\mathbf{w}_{nj} := \boldsymbol{\Gamma}'\hat{\boldsymbol{\theta}}_{nj}$, $j = 1, \ldots, p$, with corresponding eigenvalues $\zeta_{nj} := \sqrt{n}(\hat{\lambda}_{nj} - \lambda_{n1})$, $j = 1, \ldots, p$. Thus, with probability one, we have $\zeta_{n1} > \zeta_{n2} > \ldots > \zeta_{np}$, and the eigenspace of

$$\mathbf{Y}_n = \mathbf{Z}_n + \sqrt{n}(\boldsymbol{\Lambda}_n - \lambda_{n1}\mathbf{I}_p) = \mathbf{Z}_n - \mathrm{diag}\Big(0, \frac{\sqrt{n}\delta_n\xi}{(1+\delta_n\xi)^{1/p}}, \ldots, \frac{\sqrt{n}\delta_n\xi}{(1+\delta_n\xi)^{1/p}}\Big) \tag{21}$$

associated with eigenvalue ζ_{n1} is spanned by

$$\mathbf{w}_{n1} = \boldsymbol{\Gamma}'\hat{\boldsymbol{\theta}}_{n1} = \begin{pmatrix} E_{n,11} \\ \mathbf{E}_{n,12}' \end{pmatrix}.$$

Partitioning $\mathbf{Z}_n$ into

$$\mathbf{Z}_n = \begin{pmatrix} Z_{n,11} & \mathbf{Z}_{n,21}' \\ \mathbf{Z}_{n,21} & \mathbf{Z}_{n,22} \end{pmatrix},$$

where $Z_{n,11}$ is a scalar and $\mathbf{Z}_{n,22}$ is a $(p-1)\times(p-1)$ matrix, Lemma 2 then yields that $\mathbf{w}_{n1}$ is proportional to the vector of cofactors associated with the first row of

$$\mathbf{M}_{n,1} := \begin{pmatrix} Z_{n,11} - \zeta_{n1} & \mathbf{Z}'_{n,21} \\ \mathbf{Z}_{n,21} & \mathbf{Z}_{n,22} - \frac{\sqrt{n}\delta_n\xi}{(1+\delta_n\xi)^{1/p}}\mathbf{I}_{p-1} - \zeta_{n1}\mathbf{I}_{p-1} \end{pmatrix}, \tag{22}$$

or equivalently, that $\mathbf{w}_{n1}$ is proportional to the vector of cofactors associated with the first row of

$$\begin{pmatrix} Z_{n,11} - \zeta_{n1} & \mathbf{Z}'_{n,21} \\ \frac{(1+\delta_n\xi)^{1/p}}{\sqrt{n}\delta_n\xi}\mathbf{Z}_{n,21} & \frac{(1+\delta_n\xi)^{1/p}}{\sqrt{n}\delta_n\xi}\mathbf{Z}_{n,22} - \mathbf{I}_{p-1} - \frac{(1+\delta_n\xi)^{1/p}}{\sqrt{n}\delta_n\xi}\zeta_{n1}\mathbf{I}_{p-1} \end{pmatrix}.$$

Since $\mathbf{Z}_{n,21}$ and $\mathbf{Z}_{n,22}$ are $O_{\mathrm{P}}(1)$ (Proposition 1) and so is ζ_{n1} (Proposition 2), we obtain that

$$\begin{pmatrix} E_{n,11} \\ \mathbf{E}'_{n,12} \end{pmatrix} = \mathbf{e}_1 + o_{\mathrm{P}}(1)$$

(recall that $E_{n,11} > 0$ almost surely and that $\mathbf{e}_1$ is the first vector of the canonical basis of $\mathbb{R}^p$) and that

$$\sqrt{n}\delta_n\mathbf{E}'_{n,12} = O_{\mathrm{P}}(1).$$

Using the fact that $\mathbf{E}_n$ is orthogonal, it follows that

$$n\delta_n^2(1 - E_{n,11}) = \frac{\|\sqrt{n}\delta_n\mathbf{E}'_{n,12}\|^2}{1 + E_{n,11}} = \frac{1}{2}\|\sqrt{n}\delta_n\mathbf{E}'_{n,12}\|^2 + o_{\mathrm{P}}(1) = O_{\mathrm{P}}(1).$$

Since $\mathbf{E}_{n,22}$ is bounded, it also directly follows from (18) that $\sqrt{n}\delta_n\mathbf{E}_{n,21} = O_{\mathrm{P}}(1)$. In view of (19), we then obtain that $\mathbf{E}_{n,22}\mathbf{E}'_{n,22} - \mathbf{I}_{p-1}$ is $o_{\mathrm{P}}(1)$. Now, letting $\hat{\mathbf{\Lambda}}_n := \hat{\mathbf{\Gamma}}'_n\hat{\mathbf{V}}_n\hat{\mathbf{\Gamma}}_n = \mathrm{diag}(\hat{\lambda}_{n1}, \ldots, \hat{\lambda}_{np})$, we have

$$\begin{aligned} \mathbf{Z}_{n,21} &= \sqrt{n}(\mathbf{\Gamma}'\hat{\mathbf{V}}_n\mathbf{\Gamma})_{21} = \sqrt{n}(\mathbf{\Gamma}'\hat{\mathbf{\Gamma}}_n\hat{\mathbf{\Lambda}}_n\hat{\mathbf{\Gamma}}'_n\mathbf{\Gamma})_{21} \\ &= \sqrt{n}(\mathbf{E}'_n\hat{\mathbf{\Lambda}}_n\mathbf{E}_n)_{21} = \sqrt{n}(\mathbf{E}'_{n,12}\,\mathbf{E}'_{n,22})\hat{\mathbf{\Lambda}}_n\begin{pmatrix} E_{n,11} \\ \mathbf{E}_{n,21} \end{pmatrix} \\ &= \sqrt{n}\hat{\lambda}_{n1}E_{n,11}\mathbf{E}'_{n,12} + \sqrt{n}\mathbf{E}'_{n,22}\mathrm{diag}(\hat{\lambda}_{n2}, \ldots, \hat{\lambda}_{np})\mathbf{E}_{n,21}. \end{aligned}$$

Writing $\ell_{nj} := \sqrt{n}(\hat{\lambda}_{nj} - \lambda_{nj})$ for $j = 1, \ldots, p$, using (4)–(5), then (20), thus provides

$$\begin{aligned}\mathbf{Z}_{n,21} &= \ell_{n1}\mathbf{E}'_{n,11}\mathbf{E}'_{n,12} + \mathbf{E}'_{n,22}\mathrm{diag}(\ell_{n2}, \ldots, \ell_{np})\mathbf{E}_{n,21} \\ &\qquad + \sqrt{n}(1+\delta_n\xi)^{(p-1)/p}E_{n,11}\mathbf{E}'_{n,12} + \sqrt{n}(1+\delta_n\xi)^{-1/p}\mathbf{E}'_{n,22}\mathbf{E}_{n,21} \\ &= \mathbf{E}'_{n,22}\mathrm{diag}(\ell_{n2} - \ell_{n1}, \ldots, \ell_{np} - \ell_{n1})\mathbf{E}_{n,21} - \sqrt{n}\delta_n\xi(1+\delta_n\xi)^{-1/p}\mathbf{E}'_{n,22}\mathbf{E}_{n,21},\end{aligned}$$

which, since the ℓ_{nj}'s are $O_{\mathrm{P}}(1)$ (Proposition 2), yields

$$\sqrt{n}\delta_n\mathbf{E}'_{n,22}\mathbf{E}_{n,21} = -\frac{(1+\delta_n\xi)^{1/p}}{\xi}\mathbf{Z}_{n,21} + o_{\mathrm{P}}(1). \tag{23}$$

Now, Proposition 1 directly entails that $\mathrm{vec}\,\mathbf{Z}_n = (\mathbf{\Gamma}' \otimes \mathbf{\Gamma}')\sqrt{n}\mathrm{vec}\,(\hat{\mathbf{V}}_n - \mathbf{V}_n)$ is asymptotically

$$\mathcal{N}\bigg(\mathbf{0}, \Big(1 + \frac{2}{p}\Big)\Big\{(\mathbf{I}_{p^2} + \mathbf{K}_p)(\mathbf{\Lambda} \otimes \mathbf{\Lambda}) - \frac{2}{p}(\mathrm{vec}\,\mathbf{\Lambda})(\mathrm{vec}\,\mathbf{\Lambda})'\Big\}\bigg)$$

in case (i), where $\mathbf{\Lambda} := \mathrm{diag}((1+\xi)^{(p-1)/p}, (1+\xi)^{-1/p} \ldots, (1+\xi)^{-1/p})$ and

$$\mathcal{N}\bigg(\mathbf{0}, \Big(1 + \frac{2}{p}\Big)\Big\{(\mathbf{I}_{p^2} + \mathbf{K}_p) - \frac{2}{p}\mathbf{J}_p\Big\}\bigg)$$

in case (ii). Therefore, straightforward computations yield

$$\mathbf{Z}_{n,21} = (\mathbf{e}_2, \ldots, \mathbf{e}_p)'\mathbf{Z}_n\mathbf{e}_1 = (\mathbf{e}'_1 \otimes (\mathbf{e}_2, \ldots, \mathbf{e}_p)')\mathrm{vec}\,\mathbf{Z}_n \to_{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{B}),$$

where

$$\mathbf{B} := \Big(1 + \frac{2}{p}\Big)(1+\xi)^{(p-2)/p}\mathbf{I}_{p-1} \quad \text{and} \quad \mathbf{B} := \Big(1 + \frac{2}{p}\Big)\mathbf{I}_{p-1}$$

in case (i) and in case (ii), respectively. In view of (23), the desired asymptotic normality result for $\sqrt{n}\delta_n\mathbf{E}'_{n,22}\mathbf{E}_{n,21}$ follows. The one for $\sqrt{n}\delta_n\mathbf{E}'_{n,12}$ then follows from (20) and the fact that $E_{n,11} = 1 + o_{\mathrm{P}}(1)$.

We turn to the proof of (iii)–(iv). As above, $\mathbf{w}_{n1} = \mathbf{\Gamma}'\hat{\boldsymbol{\theta}}_{n1} = \mathbf{E}'_n\mathbf{e}_1$ is the unit eigenvector associated with the eigenvalue $\zeta_{n1} = \ell_{n1} = \sqrt{n}(\hat{\lambda}_{n1} - \lambda_{n1})$ of $\mathbf{Y}_n$ in (21), or equivalently, with the eigenvalue

$$\tilde{\ell}_{n1} = \ell_{n1} + \frac{\sqrt{n}\delta_n\xi}{(1+\delta_n\xi)^{1/p}} = \sqrt{n}(\hat{\lambda}_{n1} - \lambda_{n2})$$

of

$$\mathbf{Y}_n + \frac{\sqrt{n}\delta_n\xi}{(1+\delta_n\xi)^{1/p}}\mathbf{I}_p = \mathbf{Z}_n + \mathrm{diag}\bigg(\frac{\sqrt{n}\delta_n\xi}{(1+\delta_n\xi)^{1/p}}, 0, \ldots, 0\bigg). \tag{24}$$

Similarly, $\mathbf{w}_{nj} := \boldsymbol{\Gamma}'\hat{\boldsymbol{\theta}}_{nj} = \mathbf{E}_n'\mathbf{e}_j$, $j = 2, \ldots, p$, are the unit eigenvectors associated with the $p-1$ smallest eigenvalues $\ell_{n2} = \sqrt{n}(\hat{\lambda}_{n2} - \lambda_{n2}), \ldots, \ell_{np} = \sqrt{n}(\hat{\lambda}_{np} - \lambda_{np})$ of (24). Consequently, the joint distribution of $\mathbf{w}_{nj}$, $j = 1, \ldots, p$—that is, the joint distribution of the columns of $\mathbf{E}_n'$—converges weakly to the joint distribution of the unit eigenvectors (associated with eigenvalues in decreasing order, and with the signs fixed as in the statement of the theorem) of

$$\mathbf{Z} + \lim_{n\to\infty} \operatorname{diag}\left(\frac{\sqrt{n}\delta_n \xi}{(1+\delta_n\xi)^{1/p}}, 0, \ldots, 0 \right)$$

(recall that, in cases (iii)–(iv), $\mathbf{Z}_n$ converges weakly to the random matrix $\mathbf{Z}$). This establishes the result.

Proof of Theorem 1 (i) In this regime, the eigenvalues λ_{nj}, $j = 1, \ldots, p$, are fixed and given by

$$\lambda_1 := (1+\xi)^{(p-1)/p} \quad \text{and} \quad \lambda_j := (1+\xi)^{-1/p}, \; j = 2, \ldots, p,$$

respectively; see (4)–(5). Since

$$\frac{1}{\xi^2}(1+\xi) = \frac{\lambda_1\lambda_2}{(\lambda_1-\lambda_2)^2},$$

Proposition 3(i) entails that

$$\sqrt{n}\mathbf{E}_{n,12}' \to_{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \left(1+\frac{2}{p}\right)\frac{\lambda_1\lambda_2}{(\lambda_1-\lambda_2)^2}\mathbf{I}_{p-1}\right). \tag{25}$$

Now, writing $\boldsymbol{\tau}_n := \sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)$, we have

$$\frac{\|\boldsymbol{\tau}_n\|^2}{2\sqrt{n}} = \frac{\boldsymbol{\tau}_n'\boldsymbol{\tau}_n}{2\sqrt{n}} = \frac{\sqrt{n}}{2}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)'(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) = \sqrt{n}(1 - \boldsymbol{\theta}_1'\hat{\boldsymbol{\theta}}_{n1}) = -\boldsymbol{\theta}_1'\boldsymbol{\tau}_n, \tag{26}$$

where we used the fact that $\hat{\boldsymbol{\theta}}_{n1}$ and $\boldsymbol{\theta}_1$ are unit vectors. Since $\boldsymbol{\tau}_n := \sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)$ is $O_{\mathrm{P}}(1)$, it follows that

$$(\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}_1')\boldsymbol{\tau}_n = \boldsymbol{\tau}_n - (\boldsymbol{\theta}_1'\boldsymbol{\tau}_n)\boldsymbol{\theta}_1 = \boldsymbol{\tau}_n + \frac{\|\boldsymbol{\tau}_n\|^2}{2\sqrt{n}} = \boldsymbol{\tau}_n + o_{\mathrm{P}}(1)$$

as $n \to \infty$. Therefore,

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) &= (\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}_1')\sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) + o_{\mathrm{P}}(1)\\ &= \Bigg(\sum_{j=2}^{p} \boldsymbol{\theta}_j\boldsymbol{\theta}_j'\Bigg)\sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) + o_{\mathrm{P}}(1)\\ &= \sqrt{n}\sum_{j=2}^{p} \boldsymbol{\theta}_j(\hat{\boldsymbol{\theta}}_{n1}'\boldsymbol{\theta}_j) + o_{\mathrm{P}}(1)\\ &= (\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)\sqrt{n}\mathbf{E}_{n,12}' + o_{\mathrm{P}}(1), \end{aligned}\tag{27}$$

so that the asymptotic normality result in (25) entails that $\sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)$ is asymptotically normal with mean zero and covariance matrix

$$\left(1 + \frac{2}{p}\right)\frac{\lambda_1\lambda_2}{(\lambda_1 - \lambda_2)^2}\sum_{j=2}^{p}\boldsymbol{\theta}_j\boldsymbol{\theta}_j' = \left(1 + \frac{2}{p}\right)\frac{\lambda_1\lambda_2}{(\lambda_1 - \lambda_2)^2}(\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}_1'),$$

as was to be shown. (ii) In this regime, $\boldsymbol{\tau}_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1)$ is $O_{\mathrm{P}}(1/\delta_n)$ (see (8)), so that (26) yields

$$(\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}_1')\delta_n\boldsymbol{\tau}_n = \delta_n\boldsymbol{\tau}_n + \frac{\delta_n\|\boldsymbol{\tau}_n\|^2}{2\sqrt{n}} = \delta_n\boldsymbol{\tau}_n + o_{\mathrm{P}}(1).$$

Therefore,

$$\begin{aligned}\sqrt{n}\delta_n(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) &= (\mathbf{I}_p - \boldsymbol{\theta}_1\boldsymbol{\theta}_1')\sqrt{n}\delta_n(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) + o_{\mathrm{P}}(1)\\ &= \Bigg(\sum_{j=2}^{p} \boldsymbol{\theta}_j\boldsymbol{\theta}_j'\Bigg)\sqrt{n}\delta_n(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_1) + o_{\mathrm{P}}(1)\\ &= (\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)\sqrt{n}\delta_n\mathbf{E}_{n,12}' + o_{\mathrm{P}}(1), \end{aligned}\tag{28}$$

so that the result follows from the fact that

$$\sqrt{n}\delta_n\mathbf{E}_{n,12}' \to_{\mathcal{D}} \mathcal{N}\bigg(\mathbf{0}, \frac{1}{\xi^2}\left(1 + \frac{2}{p}\right)\mathbf{I}_{p-1}\bigg).$$

in this regime; see Proposition 3(ii).

(iii) Let $\mathbf{Z}$ be as in the statement of Proposition 3 and write again $\boldsymbol{\Gamma} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$. In the regimes (iii)–(iv),

$$\boldsymbol{\Gamma}'\hat{\boldsymbol{\theta}}_{n1} = \begin{pmatrix} E_{n,11} \\ \mathbf{E}_{n,12}' \end{pmatrix}$$

converges weakly to the unit eigenvector associated with the largest eigenvalue of $\mathbf{Z} + \mathrm{diag}(\xi, 0, \ldots, 0)$ with $\xi > 0$ in regime (iii) and $\xi = 0$ in regime (iv). This directly entails that

$$\hat{\boldsymbol{\theta}}_{n1} = \boldsymbol{\Gamma} \begin{pmatrix} E_{n,11} \\ \mathbf{E}'_{n,12} \end{pmatrix}$$

converges weakly to the unit eigenvector associated with the largest eigenvalue of

$$\boldsymbol{\Gamma}(\mathbf{Z} + \mathrm{diag}(\xi, 0, \ldots, 0))\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}\mathbf{Z}\boldsymbol{\Gamma}' + \xi\boldsymbol{\theta}_1\boldsymbol{\theta}'_1. \tag{29}$$

Part (iii) of the result then follows from the fact that the distribution of $\mathbf{Z}$ is invariant with respect to orthogonal transformations, in the sense that $\mathbf{OZO}'$ has the same distribution as $\mathbf{Z}$ for any $p \times p$ orthogonal matrix $\mathbf{O}$. (iv) The proof for $\xi > 0$ in (iii) above applies for ξ and shows that, in regime (iv), $\hat{\boldsymbol{\theta}}_{n1}$ converges weakly to the unit eigenvector associated with the largest eigenvalue of $\mathbf{Z} = \mathbf{Z}(0)$. Now, the orthogonal invariance of the distribution of $\mathbf{Z} = \mathbf{Z}(0)$ entails that the joint distribution of its eigenvectors is the invariant Haar distribution on the group of $p \times p$ orthogonal matrices, which implies in particular that each of these eigenvectors is uniformly distributed over $\mathcal{S}^{p-1}$. This establishes Part (iv) of the result.

References

Cator, E. A., & Lopuhaä, H. P. (2010). Asymptotic expansion of the minimum covariance determinant estimators. *Journal of Multivariate Analysis, 101*, 2372–2388.

Chen, S. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis, 31*, 131–145.

Duong, T. (2021). ks: Kernel smoothing, https://CRAN.R-project.org/package=ks. R package version 1.13.2.

Dürre, A., Tyler, D. E., & Vogel, D. (2016). On the eigenvalues of the spatial sign covariance matrix in more than two dimensions. *Statistics & Probability Letters, 111*, 80–85.

Frahm, G. (2004). Generalized elliptical distributions: Theory and applications. PhD thesis, Universität zu Köln.

Frahm, G. (2009). Asymptotic distributions of robust shape matrices and scales. *Journal of Multivariate Analysis, 100*, 1329–1337.

Hallin, M., & Paindaveine, D. (2006). Parametric and semiparametric inference for shape: The role of the scale functional. *Statistics & Decisions, 24*, 327–350.

Hallin, M., & Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *The Annals of Statistics, 34*, 2707–2756.

Hallin, M., Paindaveine, D., & Verdebout, T. (2010). Optimal rank-based testing for principal components. *The Annals of Statistics, 38*, 3245–3299.

Hallin, M., Paindaveine, D., & Verdebout, T. (2014). Efficient R-estimation of principal and common principal components. *Journal of the American Statistical Association, 109*, 1071–1083.

Lopuhaa, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *The Annals of Statistics*, 1638–1665.

Nordhausen, K., Sirkia, S., Oja, H., & Tyler, D. E. (2018). ICSNP: Tools for multivariate nonparametrics, https://CRAN.R-project.org/package=ICSNP. R package version 1.1-1.

Onatski, A., Moreira, M., & Hallin, M. (2014). Signal detection in high dimension: The multispiked case. *The Annals of Statistics, 42*, 225–254.

Paindaveine, D. (2008). A canonical definition of shape. *Statistics & Probability Letters, 78*, 2240–2247.

Paindaveine, D., Remy, J., & Verdebout, T. (2020a). Sign tests for weak principal directions. *Bernoulli, 26*, 2987–3016.

Paindaveine, D., Remy, J., & Verdebout, T. (2020b). Testing for principal component directions under weak identifiability. *The Annals of Statistics, 48*, 324–345.

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

Taskinen, S., Croux, C., Kankainen, A., Ollila, E., & Oja, H. (2006). Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. *Journal of Multivariate Analysis, 97*, 359–384.

Tyler, D. (1983). The asymptotic distribution of principal component roots under local alternatives to multiple roots. *The Annals of Statistics, 11*, 1232–1242.

Tyler, D. (1987). A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics, 15*, 234–251.

Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics, 9*, 725–736.

Tyler, D. E. (1983). A class of asymptotic tests for principal component vectors. *The Annals of Statistics, 11*, 1243–1250.

On Minimax Shrinkage Estimation with Variable Selection

Stavros Zinonos and William E. Strawderman

Abstract We study minimax estimators of the mean vector of a spherically symmetric distribution that also perform variable selection by estimating certain components as 0. The basic class of estimators developed is closely related to, and generalizes, classes considered by Zhou and Hwang (2005) and Maruyama (2014) in the Gaussian setting. The class of distributions studied includes scale mixtures of normals (e.g., Student-t) as well as the general class of spherically symmetric distributions with a residual vector. Certain subclasses of these estimators based on truncated order statistics are shown to be particularly effective when some information on the sparsity is known.

Keywords Shrinkage estimation · Variable selection · Minimaxity · Quadratic loss

1 Introduction

We study minimax estimators of the mean vector of a spherically symmetric distribution which dominate the standard minimax estimator $\delta_0(\mathbf{X}) = \mathbf{X}$ under squared error loss. We are particularly interested in minimax estimators whose positive part adaptively estimates a certain subset of the mean vector as 0, while shrinking the remaining coordinates. The results may be viewed as a modification based on order statistics of extensions of Zhou and Hwang (2005) and of Maruyama (2014) from the Gaussian case to the broader class of spherically symmetric distributions, and also to a somewhat wider class of estimators. The modification of this wider class of estimators seems effective in risk reduction and variable selection when

S. Zinonos (✉)
Cardiovascular Institute of New Jersey, RWJMS, New Brunswick, NJ, USA
e-mail: szinonos@rwjms.rutgers.edu

W. E. Strawderman
Department of Statistics and Biostatistics, Rutgers University, New Brunswick, NJ, USA
e-mail: straw@stats.rutgers.edu

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_4

information on sparsity is available. Specifically, let $\mathbf{X}$ be a spherically symmetric distribution with density given by $f(\|\mathbf{x}-\boldsymbol{\theta}\|^2)$ where $dim(\mathbf{X}) = dim(\boldsymbol{\theta}) = p \geq 3$, $cov(\mathbf{X}) = \boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I_p}$ where σ^2 is known, and let the loss function for estimation of $\boldsymbol{\theta}$ be given by

$$L(\boldsymbol{\theta}, \mathbf{d}) = \|\mathbf{d}-\boldsymbol{\theta}\|^2. \tag{1}$$

We study minimaxity of estimators of the form:

$$\delta(\mathbf{X}) = (\delta_1(\mathbf{X}), \delta_2(\mathbf{X}), \ldots, \delta_p(\mathbf{X}))',$$

where

$$\delta_i(\mathbf{X}) = (1-\phi_i(\mathbf{X}))X_i, \tag{2}$$

and where

$$\phi_i(\mathbf{X}) = \psi_i(X_1^2, X_2^2, \ldots, X_p^2). \tag{3}$$

If $f(\cdot)$ is unimodal and $\delta(\cdot)$ is minimax, the positive-part estimator $\delta^+(\mathbf{X})$ with

$$\delta_i^+(\mathbf{X}) = (1-\phi_i(\mathbf{X}))_+ X_i$$

(where $a_+ = max(0, a)$) is also minimax (and in fact dominates $\delta(\mathbf{X})$) and additionally may allow adaptively selected subsets of the coordinates to be estimated by 0. Hence minimaxity and variable selection are simultaneously achieved. In particular Theorem 2 establishes minimaxity of certain shrinkage estimators with coordinates of the form

$$\delta_i^+(\mathbf{x}) = \begin{cases} (1 - c\sigma^2 v(\sum_{i=1}^p h(x_i^2 \wedge z_k^2))w(x_i^2))_+ x_i & \text{if } |x_i| \leq z_k \\ (1 - \frac{c\sigma^2 v(\sum_{i=1}^p h(x_i^2 \wedge z_k^2))w(z_k^2)z_k}{|x_i|})_+ x_i & \text{if } |x_i| > z_k, \end{cases} \tag{4}$$

where $x_i \wedge x_j = min(x_i, x_j)$ and $\mathbf{Z} = (Z_{(1)}, Z_{(2)}, \cdots, Z_{(p)})'$ with $Z_i = |X_i|$. Hence $\delta_i^+(\mathbf{x})$ is set to zero whenever

$$\begin{cases} w(x_i^2) > \frac{1}{c\sigma^2 v(\sum_{i=1}^p h(x_i^2 \wedge z_k^2))} & \text{if } |x_i| \leq z_k \\ \frac{w(z_k^2)z_k}{|x_i|} > \frac{1}{c\sigma^2 v(\sum_{i=1}^p h(x_i^2 \wedge z_k^2))} & \text{if } |x_i| > z_k. \end{cases}$$

Zhou and Hwang (2005) and Maruyama (2014) have studied classes of such procedures given by (4) for the case $k = dim(\mathbf{X}) = p$ in the Gaussian case, basing shrinkage on the ℓ_d-norm of $\mathbf{X}$: $1 < d < 2$ for Zhou and Hwang, and general d for

Maruyama. In particular the form of (4) when $k = dim(\mathbf{X}) = p$ is given by

$$\delta_i^+(\mathbf{X}) = (1 - c\sigma^2 v(\sum_{i=1}^{p} h(x_i^2))w(x_i^2))_+ x_i. \tag{5}$$

Hence $\delta_i^+(\mathbf{X})$ is set to 0 whenever

$$w(x_i^2) > \frac{1}{c\sigma^2 v(\sum_{i=1}^{p} h(x_i^2))}.$$

Applications of classes of these procedures given in (4) when $k = p$ were used by Zhou and Hwang (2005) in estimation of functions via wavelets. For other estimators in the context of wavelet denoising see, e.g., Donoho and Johnstone (1994), Donoho and Johnstone (1995), and Cai (1999).

This type of modification of shrinkage estimators is due to Stein (1981). While Stein introduced the idea to limit the amount of shrinkage, we use it to limit the size of the denominator of the shrinkage factor, so that it is easier for coordinates with small $|X_i|$ to be deselected (estimated as 0). A simulation study indicated that this class can be particularly successful in reducing risk and in variable selection in certain subspaces reflecting knowledge of the sparsity of the model. A notable feature of this modified estimator as shown in the simulation studies is that, in certain subspaces, the asymptotic risk is substantially less than the minimax risk. Furthermore the asymptotic probability of selection of inactive (i.e., mean 0) variables is strictly less than 1. Both of these asymptotic behaviors do not occur for the Zhou-Hwang or Maruyama estimators.

In addition to extending the classes of minimax estimators we extend the results to scale mixtures of normal distributions. We also extend the result to the general class of spherically symmetric distributions with a residual vector that allows estimation of an unknown scale. In particular let

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{U}\end{pmatrix} \sim f(\|\mathbf{x} - \boldsymbol{\theta}\|^2 + \|\mathbf{u}\|^2), \tag{6}$$

where $dim(\mathbf{X}) = dim(\boldsymbol{\theta}) = p \geq 3$, $dim(\mathbf{U}) = m \geq 1$, and covariance matrix $\boldsymbol{\sigma^2}\mathbf{I_{p+m}}$ with σ^2 unknown. The vector $\mathbf{U}$ is referred to as the residual vector and is used to estimate the scale parameter σ^2. The model (6) is a canonical form of the general linear model with error vector $\epsilon \in \mathbb{R}^{p+m}$ where $\epsilon \sim f(\|\epsilon\|^2)$. Let

$$L(\boldsymbol{\theta}, \mathbf{d}) = \frac{\|\mathbf{d} - \boldsymbol{\theta}\|^2}{E\|\mathbf{U}\|^2}.$$

In this case the dominating minimax estimators are of the form

$$\delta_i(\mathbf{X}, \mathbf{U}) = (1 - \frac{\mathbf{U}'\mathbf{U}\phi_i(\mathbf{X})}{m+2})X_i$$

and

$$\delta_i^+(\mathbf{X}, \mathbf{U}) = (1 - \frac{\mathbf{U}'\mathbf{U}\phi_i(\mathbf{X})}{m+2})_+ X_i.$$

As a special case of this extension the estimator $\delta_i^+(\mathbf{X}, \mathbf{U})$ will be minimax in the Gaussian case with an unknown scale. The basic tool in the Gaussian setting is the Stein unbiased estimator of risk technique as in Zhou and Hwang (2005) and Maruyama (2014). Generalizations of this technique to the spherically symmetric setting form the basis of the results for more general spherically symmetric distributions. See Fourdrinier et al. (2018) for a general discussion of shrinkage estimations for such models.

The paper is organized as follows: Sect. 2 considers classes of minimax estimators in the Gaussian case. The generalization will include classes of pseudo-bayes estimators. Section 3 studies extensions to the class of scale mixtures of normal distributions. Section 4 extends the results of Sect. 2 to spherically symmetric distributions (including the normal) with a residual vector. Section 5 simulates the risk and probability that $|\hat{\theta}_i| \neq 0$ for certain classes of estimators developed in Sect. 2. Section 6 gives some concluding remarks.

2 Results for the Normal Case, Known Scale

In this section, $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2\mathbf{I_p})$ with σ^2 known and loss given by (1). We focus on modifications of estimators of the form

$$\delta_M(\mathbf{X})_i = (1 - \sigma^2 c v(D) w(x_i^2))x_i = (1 - \psi_i(x_1^2, \ldots, x_p^2))x_i, \tag{7}$$

where

$$D = \sum_{i=1}^{p} h(x_i^2) \tag{8}$$

and on their positive parts. The estimators of Zhou and Hwang (7) and Marauyama (5) are of this form. The modifications are based on the ordered values of the $|X_i|$'s. Specifically let $Y_i = |X_i|$ for $i = 1, \ldots, p$ and let $Z_k = Y_{(k)}$, where $Y_{(i)}$ is the i^{th} order statistic and k is fixed, and such that $p \geq k \geq 3$. The form of the truncated version is

$$\delta_i(\mathbf{X}) = \begin{cases} (1 - \sigma^2 c v(D) w(x_i^2))x_i & \text{if } |x_i| \leq z_k \\ (1 - \sigma^2 c v(D) w(z_k^2)\frac{z_k}{|x_i|})x_i & \text{if } |x_i| > z_k, \end{cases} \tag{9}$$

where $D = \sum_{i=1}^{p} h(x_i^2 \wedge z_k^2)$.

The class of estimators considered in (7) contains classes of estimators which are sometimes referred to as pseudo-Bayes estimators, by which we mean (in this paper) estimators of the form

$$\delta(\mathbf{X}) = \mathbf{X} + \sigma^2 g(\mathbf{X}),$$

where

$$g(\mathbf{X}) = \frac{\nabla m(\mathbf{X})}{m(\mathbf{X})}.$$

The function $m(\mathbf{X})$ is referred to as a pseudo-marginal. If $m(\mathbf{X})$ were a true marginal distribution corresponding to a generalized or proper prior $\pi(\boldsymbol{\theta})$, then δ would be a generalized or proper Bayes estimator. This can be seen by setting

$$m(t) = j^2(t)$$

so that

$$\frac{(\nabla m(D))_i}{m(D)} = \frac{4j'(D)h'(x_i^2)x_i}{j(D)} \tag{10}$$

with D as in (8). Hence,

$$cv(D) = \frac{4j'(D)}{j(D)}, \text{ and} \tag{11}$$

$$w(x_i^2) = h'(x_i^2) \tag{12}$$

so that the estimator (7) may be interpreted as a pseudo-Bayes estimator with pseudo-marginal $m(\mathbf{x}) = j^2(D)$.

Theorem 1 gives general conditions under which thresholding each of the coordinates of an estimator $\delta(\mathbf{X})$ of $\boldsymbol{\theta}$ results in an estimator, $\delta^+(\mathbf{X})$ with smaller risk under square error loss than $\delta(\mathbf{X})$. Various versions, including Theorem 4 in Zhou and Hwang (2005), have appeared in the literature. The current version is broad enough to apply to other (non-Gaussian) distribution studied in this paper.

Theorem 1 *Suppose* $\mathbf{X}$ *is a random variable in* $\mathbb{R}^p$ *with density* $f(\|\mathbf{x}-\boldsymbol{\theta}\|^2)$ *where* f *is symmetric, and unimodal in each of the coordinates separately for each fixed value of the other coordinates. Let* $\delta(\mathbf{X}) = (\delta_1(\mathbf{X}), \dots, \delta_p(\mathbf{X}))'$ *be an estimator of* $\boldsymbol{\theta}$ *satisfying (2) and (3). Let*

$$\delta^+(\mathbf{X}) = ((1-\phi_1(\mathbf{X}))_+ X_1, \dots, (1-\phi_p(\mathbf{X}))_+ X_p)' =$$

$$(\delta_1^+(\mathbf{X}), \dots, \delta_p^+(\mathbf{X}))'$$

then

$$E_{\boldsymbol{\theta}}[\theta_i - \delta_i^{+}(\mathbf{X}))^2] \leq E_{\boldsymbol{\theta}}[(\theta_i - \delta_i(\mathbf{X}))^2] \textit{for } i = 1, \ldots, p.$$

Furthermore if there exists an i such that $P_{\boldsymbol{\theta}}(\phi_i(\mathbf{X}) > 1) > 0$, *then*

$$E_{\boldsymbol{\theta}}[(\theta_i - \delta_i^{+}(\mathbf{X}))^2] < E_{\boldsymbol{\theta}}[(\theta_i - \delta_i(\mathbf{X}))^2].$$

Proof The proof is similar to that in Zhou and Hwang (2005). The details are omitted. ■

The following standard Lemma follows easily from the covariance inequality (see, e.g., Casella and Berger 2002, Theorem 4.7.9).

Lemma 1 *Let f and h be continuous monotonic functions defined over* $[a, b] \subseteq \mathbb{R}$ *into* $\mathbb{R}$. *For any finite collection* $\{x_i\}_{i=1,\ldots,n} \subseteq [a, b]$

(i) If f and h are both monotonic increasing functions, then

$$(\sum_{i=1}^{n} f(x_i))(\sum_{i=1}^{n} h(x_i)) \leq n \sum_{i=1}^{n} f(x_i)h(x_i).$$

(ii) If f is a monotonic increasing function and h is a monotonic decreasing function, then

$$n \sum_{i=1}^{n} f(x_i)h(x_i) \leq (\sum_{i=1}^{n} f(x_i))(\sum_{i=1}^{n} h(x_i)).$$

■

Theorem 2 is the main result of this section. Sufficient conditions are given so that estimators of the form (7) and their positive-part versions are minimax under squared error loss, $\|\mathbf{d} - \boldsymbol{\theta}\|^2$.

Theorem 2 *Let* $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2\mathbf{I_p})$ *with* σ^2 *known and* $\boldsymbol{\theta}$ *unknown. Let* $\delta(\mathbf{X}) = \mathbf{X} + \sigma^2 g(\mathbf{X})$ *be of the form (9) with* $g(\mathbf{X})$ *weakly differentiable and* $E_{\boldsymbol{\theta}}[\|g(\mathbf{X})\|^2] < \infty$. *Also, assume the following:*

k is fixed and such that $p \geq k \geq 3$.

Assumptions on $h(\cdot)$*:*

(H1) h(t) is differentiable in t for all $t \geq 0$.

(H2) h(t) and $h'(t)$ *are non-negative for all* $t \geq 0$.

(H3) There exists a constant $B > 0$ *such that* $h'(t)t \leq Bh(t)$ *for all* $t \geq 0$.

Assumptions on $w(\cdot)$*:*

(W1) w(t) is differentiable with respect to t for all $t \geq 0$.

(W2) $w(t) \geq 0$ *and* $w'(t) \leq 0$ *for all* $t \geq 0$.

(W3) *There exists a constant* $A < 0$ *such that* $Aw(t) \leq w'(t)t$ *for all* $t \geq 0$.

(W4) *For* $t \geq 0$ *w(t)t is monotonic increasing in t.*

Assumptions on $v(\cdot)$*:*

(V1) *The first derivative with respect to t,* $v'(t)$*, exists for all* $t > 0$.

(V2) *For* $t > 0$*,* $v(t) > 0$ *and* $v'(t) < 0$.

(V3) *There exists a constant* $F > 0$ *such that* $-4v'(t)t \leq Fv(t)$ *for* $t > 0$.

(V4) *There exists a constant* $H > 0$ *such that* $v(D)\sum_{i=1}^{p}\{w(x_i^2 \wedge z_k^2)(x_i^2 \wedge z_k^2)\} \leq H$ *for all* $x_i \neq 0$*,*

where $D = \sum_{i=1}^{p} h(x_i^2 \wedge z_k^2)$*. Then the estimator* $\delta(\mathbf{X})$ *is minimax under squared error loss provided:*

$$0 < 4k(A + \frac{1}{2}) - FB$$

and

$$0 < c \leq \frac{4k(A + \frac{1}{2}) - FB}{H}$$

is satisfied. Furthermore the estimator $\delta^{+}(\mathbf{X})$ *will also be minimax.*

Comment Before giving the proof, we comment on the possible value of the truncating the estimators based on the ordered $|X_i|$'s. Without truncation, if both $v(t)$ and $w(t)$ are strictly monotonic decreasing functions, and if $h(t)$ is a strictly monotonic increasing function (as in the case of Maruyama, i.e., Example 1), for any $|x_i|$ sufficiently large,

$$w(x_i^2) < \frac{1}{\sigma^2 cv(D)}$$

so that for sufficiently large $\|\mathbf{X}\|^2$ no coordinates of $\delta_i(\mathbf{X})$ will be set to 0. If an investigator believes at most s of $|\theta_i| > 0$, setting $k <= p - s$ allows some control on the size of $v(D)$ so that coordinates in non-active sets have a higher probability of being set to 0 for large values of $\|\mathbf{X}\|^2$, while coordinates in active sets with $|x_i|$ sufficient large satisfy

$$\frac{w(z_k^2)}{|x_i|} < \frac{1}{\sigma^2 cv(D)}$$

and are not set to 0. The idea of truncation was used by Stein (1981) to modify the shrinkage pattern of James–Stein estimators so that the amount of shrinkage of large

$|X_i|$'s was controlled. In this context we use it as a means to control the thresholding in sparse sets to produce estimators with more favorable model selection properties.

Proof The proof is based on Stein's unbiased estimator of risk (Stein 1981), i.e., $\triangle = E[\|\mathbf{X}+\sigma^2 g(\mathbf{X})-\boldsymbol{\theta}\|^2]-E[\|\mathbf{X}-\boldsymbol{\theta}\|^2] = \sigma^4 E[\|g(\mathbf{X})\|^2+2div(g(\mathbf{X}))]$. Hence to prove minimaxity of $\mathbf{X}+\sigma^2 g(\mathbf{X})$ it suffices to show $\|g(\mathbf{X})\|^2+2div(g(\mathbf{X})) \leq 0$, since $E_{\boldsymbol{\theta}}\|g(\mathbf{X})\|^2 < \infty$ is sufficient for the risk of $\delta(\mathbf{X})$ to be finite. Without loss of generality we assume that $\sigma^2 = 1$. Since

$$\frac{\partial g_i}{\partial x_i} = \begin{cases} -cv(D)[2w'(x_i^2)x_i^2 + w(x_i^2)] - 2cv'(D)w(x_i^2)h'(x_i^2)x_i^2 & \text{if } |x_i| < Z_k \\ -cv(D)[2w'(z_k^2)z_k^2 + w(z_k^2)] - 2c(p-k+1)v'(D)w(z_k^2)h'(z_k^2)z_k^2 & \text{if } |x_i| = Z_k \\ 0 & \text{if } |x_i| > Z_k \end{cases}$$

$\|g(\mathbf{x})\|^2 + 2div_{\mathbf{x}}(g(\mathbf{x}))$ is expressible as

$$c^2v^2(D)[\sum_{j=1}^{k-1} w^2(x_{(j)}^2)x_{(j)}^2 + (p-k+1)w^2(z_k^2)z_k^2]-$$

$$4cv'(D)\{\sum_{j=1}^{k-1} w(x_{(j)}^2)h'(x_{(j)^2})x_{(j)}^2 + (p-k+1)w(z_k^2)h'(z_k^2)z_k^2\}-$$

$$\{4cv(D)[\sum_{j=1}^{k-1}(w'(x_{(j)}^2)x_{(j)}^2) + \frac{w(x_{(j)}^2)}{2})] + (w'(z_k^2)z_k^2 + \frac{w(z_k^2)}{2})\} =$$

$$cV(D)[\sum_{j=1}^{k-1} w(x_{(j)}^2) + (p-k+1)w(z_k^2)]\{L_1 + L_2 - L_3,\} \tag{13}$$

where

$$L_1 = \frac{cv(D)[\sum_{j=1}^{k-1} w^2(x_{(j)}^2)x_{(j)}^2 + (p-k+1)w^2(z_k^2)z_k^2]}{\sum_{j=1}^{k-1} w(x_{(j)}^2) + (p-k+1)w(z_k^2)} \tag{14}$$

$$L_2 = \frac{-4v'(D)\{\sum_{j=1}^{k-1} w(x_{(j)}^2)h'(x_{(j)}^2)x_{(j)}^2 + (p-k+1)w(z_k^2)h'(z_k^2)z_k^2\}}{V(D)[\sum_{j=1}^{k-1} w(x_{(j)}^2) + (p-k+1)w(z_k^2)]} \tag{15}$$

$$L_3 = \frac{4\{\sum_{j=1}^{k-1}(w'(x_{(j)}^2)x_{(j)}^2) + \frac{w(x_{(j)}^2)}{2})] + (w'(z_k^2)z_k^2 + \frac{w(z_k^2)}{2})\}}{\sum_{j=1}^{k-1} w(x_{(j)}^2) + (p-k+1)w(z_k^2)} \tag{16}$$

in (13). Bounds on L_1, L_2, and L_3 are given so that $L_1+L_2 \leq L_3$ which implies (13)$\leq$ 0.

From assumption W3 and arranging the sum to include first k terms

$$4(A+\frac{1}{2})(\frac{\sum_{j=1}^{k} w(x_{(j)}^2)}{\sum_{j=1}^{k} w(x_{(j)}^2)+(p-k)w(z_k^2)}) \le L_3.$$

Let $\bar{w}_n = \frac{\sum_{j=1}^{n} w(X_{(j)}^2)}{n}$. Since $w(t)$ is decreasing in t (assumption W2), and $|X_{(j)}| \le Z_k$ for $j \le k$, $(p-k)\bar{w}_k \ge (p-k)w(Z_k)$, so that

$$4(A+\frac{1}{2})\frac{k}{p} = 4(A+\frac{1}{2})\frac{k(\bar{w}_k)}{k(\bar{w}_k)+(p-k)\bar{w}_k} \le \tag{17}$$

$$4(A+\frac{1}{2})(\frac{\sum_{j=1}^{k} w(x_{(j)}^2)}{\sum_{k=1}^{k} w(x_{(j)}^2)+(p-k)w(z_k^2)}) \le L_3. \tag{18}$$

Since $w(t)t$ is increasing in t (assumption W4) and w(t) is decreasing in t (assumption W2) Lemma 1 implies the p terms in the sum of L_1 satisfy

$$L_1 \le cv(D)\{\frac{(\sum_{j=1}^{k-1}(w(x_{(j)}^2)x_{(j)}^2)+(p-k+1)w(z_k^2)z_k^2)(\sum_{j=1}^{k-1} w(x_{(j)}^2)+(p-k+1)w(z_k^2))}{p(\sum_{j=1}^{k-1} w(x_{(j)}^2)+(p-k+1)w(z_k^2))}\} \le$$

$$\text{(assumption V4)}\ \frac{cH}{p}. \tag{19}$$

By assumptions V3 and H3

$$L2 \le \frac{FB(\sum_{j=1}^{k-1} w(x_{(j)}^2)h(x_{(j)}^2)+(p-k+1)w(z_k^2)h(z_k^2))}{D(\sum_{j=1}^{k-1} w(x_{(j)}^2)+(p-k+1)w(z_k^2))} \le \tag{20}$$

$$\frac{FB\{\sum_{i=1}^{k-1} w(x_{(j)}^2))+(p-k+1)w(z_k^2)\}D}{D\{\sum_{j=1}^{k-1} w(x_{(j)}^2)+(p-k+1)w(z_k^2)\}p} = \frac{FB}{p}, \tag{21}$$

where (21) follows from (20) form Lemma 1 using the p terms in the sum for L_2 since w is decreasing (assumption W2) and h is increasing (assumption H2). Therefore once

$$0 \le L_1 + L2 \le \frac{cH+FB}{p} \le 4(A+\frac{1}{2})\frac{k}{p} \le L_3$$

is satisfied $\delta(\mathbf{X})$ is minimax. ■

Theorem 1 implies the positive-part estimators, or equivalently the adaptively thresholded estimators, are also minimax. Hence in general under the stated condition on h(), w(), and v(), the shrinkage estimator given by (9) remains minimax when each coordinate is adaptively thresholded (and set to 0).

Example 1 (Truncated Maruyama Estimator) In this example we retrieve the results of Maruyama (2014) and Zhou and Hwang (2005) as a special cases of Theorem 2. Let $\mathbf{X} \sim N_p(\boldsymbol{\theta}, \mathbf{I_p})$. The estimators considered by Zhou and Hwang have coordinates of the form

$$\delta_{ZH,i}(\mathbf{X}, a, c) = (1 - \frac{c}{\|\mathbf{X}\|_{2-a}^{2-a}|X_i|^a})X_i, \tag{22}$$

where

$$0 < c \leq 2(p - 2 - a(p-1)),$$

and are minimax provided $p \geq 3$ and $0 < a < \frac{p-2}{p-1}$. Here $\|\mathbf{x}\|_a = [\sum |x_i|^a]^{\frac{1}{a}}$. As Maruyama (2014) notes, this is a subclass of his estimators, $\delta_M(\mathbf{X}, a, d, c)$, with i^{th} coordinates of the form

$$\delta_{M,i}(\mathbf{X}, a, d, c) = (1 - \frac{c}{\|\mathbf{X}\|_d^{2-a}|X_i|^a})X_i \tag{23}$$

for $d > 0$. Maruyama (2014) gives the condition

$$0 < c * max\{1, p^{1-(\frac{2-a}{d})}\} \leq 2(p - 2 - a(p-1))$$

for $p \geq 3$, $0 < a < \frac{p-2}{p-1}$, and $d > 0$ for the minimaxity of $\delta_M(\mathbf{X}, a, d, c)$. This implies the Zhou and Hwang result since

$$max\{1, p^{1-\frac{2-a}{d}}\} = 1$$

for $d = 2 - a$ with $0 < a < \frac{p-2}{p-1}$.

Interestingly, the Zhou and Hwang estimator can be derived as a pseudo-Bayes estimator of the form (10)- (12) where $j(t) = t^{-\frac{a_1}{4}}$ and $h(t) = t^{\frac{d_1}{2}}$ so that

$$cv(D) = 4\frac{j'(D)}{j(D)} = \frac{-a_1}{4D} = \frac{-a_1}{\sum_{i=1}^p (x_i^2)^{\frac{d_1}{2}}},$$

and

$$w(x_i^2) = h'(x_i^2) = -\frac{d_1}{2}(x_i^2)^{\frac{d_1-2}{2}}.$$

Hence

$$\mathbf{X} + \frac{\nabla m(\mathbf{X})}{m(\mathbf{X})} = (1 - \frac{a_1(\frac{d_1}{2})}{[\sum_{i=1}^p |x_i|^{d_1}]|x_i|^{2-d_1}})x_i.$$

Let $Y_i = |X_i|$, and $Z_k = Y_{(k)}$ denote the k^{th} largest value of Y_i.

Example 2 shows the estimator

$$\delta_i(\mathbf{x}, a, c, d, k) = \begin{cases} x_i - \frac{cx_i}{[\sum_{i=1}^{p}(|x_i|^d \wedge z_k^d)]^{\frac{2-a}{d}}|x_i|^a} & \text{if } |x_i| \le z_k \\ x_i - \frac{cz_k * sign(x_i)}{[\sum_{i=1}^{p}(|x_i|^d \wedge z_k^d)]^{\frac{2-a}{d}} z_k^a} & \text{if } |x_i| > z_k \end{cases} \tag{24}$$

is a minimax estimator for $0 < c * max\{1, p^{1-(\frac{2-a}{d})}\} \le 2(k-2-a(k-1))$, $0 < a < \frac{k-2}{k-1}$, $d > 0$, and $3 \le k \le p = dim(\mathbf{X})$. Setting $k = p$ in (24) yields back Maruyama's estimator (23) with the same bounds for minimaxity under squared error loss.

Example 2 (Truncated Maruyama Continued)

As in Example 1, let $\sigma^2 = 1$. Since $sign(x_i) = \frac{x_i}{|x_i|}$, (24) can be re-expressed as

$$\delta_i(\mathbf{x}, a, c, d, k) = \begin{cases} (1 - \frac{c}{[\sum_{i=1}^{p}(|x_i|^d \wedge z_k^d)]^{\frac{2-a}{d}}|x_i|^a})x_i & \text{if } |x_i| \le z_k \\ (1 - \frac{cz_k}{[\sum_{i=1}^{p}(|x_i|^d \wedge z_k^d)]^{\frac{2-a}{d}} z_k^a |x_i|})x_i & \text{if } |x_i| > z_k. \end{cases} \tag{25}$$

The minimaxity of δ^+ follows from Theorem 2 as follows, with the identification $v(t) = t^{-(\frac{2-a}{d})}$, $h(t) = t^{\frac{d}{2}}$, $w(t) = t^{-\frac{a}{2}}$.

Condition on $h(\cdot)$:

With $h(t) = t^{\frac{d}{2}}$ conditions H1, and H2 are satisfied once $d > 0$, and condition H3 is satisfied with $B = \frac{d}{2}$.

Conditions on $w(\cdot)$:

With $w(t) = t^{-(\frac{a}{2})}$ and $0 \le a \le 2$, conditions W1-W4 are satisfied with $A = -\frac{a}{2}$.

Conditions on $v(\cdot)$:

With $v(t) = t^{-(\frac{2-a}{d})}$ V1-V3 are satisfied with $0 < a < 2, d > 0$, and $F = 4(\frac{2-a}{d})$. Selection of the constant H which satisfies V4 can be separated into 2 cases:

Case 1: $\frac{2-a}{d} > 1$. Since

$$[\sum_{i=1}^{p}(x_i^2)^{\frac{d}{2}}]^{\frac{2-a}{d}} \ge \sum_{i=1}^{p}(x_i^2)^{\frac{2-a}{2}} \tag{26}$$

we can choose H to be 1.

Case 2: $\frac{2-a}{d} < 1$.

From Jensen's inequality for a concave functions (i.e., $c(x)$ that $c(EX) \ge Ec(X)$)

$$[\frac{\sum_{i=1}^{p}(x_i^2)^{\frac{d}{2}}}{p}]^{(\frac{2-a}{d})} \ge \frac{\sum_{i=1}^{p} x_i^{2(\frac{2-a}{2})}}{p}$$

which implies

$$\frac{\sum_{i=1}^{p}(x_i^2)^{(\frac{2-a}{2})}}{[\sum_{i=1}^{p}(x_i^2)^{(\frac{d}{2})}]^{(\frac{2-a}{d})}} \leq (\frac{1}{p})^{(\frac{2-a}{d})-1}$$

so that $H = p^{1-(\frac{2-a}{d})}$.
From Case 1 and 2, $H = max\{1, p^{1-(\frac{2-a}{d})}\}$.
From Theorem 2, (25) will be minimax once:

$$0 < c * \max\{1, p^{1-(\frac{2-a}{d})}\} \leq 4k(\frac{1-a}{2}) - 4(\frac{2-a}{d})\frac{d}{2},$$

or equivalently,

$$0 < c * \max\{1, p^{1-(\frac{2-a}{d})}\} \leq 2(k - 2 - a(k-1))$$

for $0 < a < \frac{k-2}{k-1}$, $3 \leq k \leq p$, and $d > 0$.

A simulation study in Sect. 5 indicates that, when knowledge of the sparsity, $p - k$, is available, the positive-part version of estimator (25) markedly improves both the risk and probability of non-inclusion of inactive variables in certain subspaces.

3 Scale Mixtures of Normal Distributions

In this section we extend Theorem 2 to the case of scale mixtures of normals. In particular the distributions studied have the hierarchical structure

$$\mathbf{X}|\sigma^2 \sim N_p(\boldsymbol{\theta}, \boldsymbol{\sigma}^2\mathbf{I_p}) \tag{27}$$

$$\sigma^2 \sim G(\sigma^2), \tag{28}$$

where $E[\sigma^2] < \infty$ and $E[\frac{1}{\sigma^2}] < \infty$. The estimators studied have coordinates of the form (analogous to (9))

$$\delta(\mathbf{x})_i = x_i - \begin{cases} \frac{c}{E[\frac{1}{\sigma^2}]} v(\sum_{i=1}^{p} h(x_i^2))w(x_i^2)x_i & \text{if } |x_i| \leq z_k \\ \frac{c}{E[\frac{1}{\sigma^2}]} v(\sum_{i=1}^{p} h(x_i^2 \wedge z_k^2))w(z_k^2)z_k * sign(x_i) & \text{if } |x_i| > z_k. \end{cases} \tag{29}$$

Here is the main result of this section. Its proof uses similar techniques found in Strawderman (1974) which proved extensions to Baranchik-type shrinkage estimators for scale mixture of normal distributions.

Theorem 3 *Let* $\mathbf{X}$ *have the distribution given in (27)–(28) where* $E[\sigma^2]$ *and* $E[\frac{1}{\sigma^2}]$ *are finite. Let* $\delta(\mathbf{X})$ *be an estimator for* $\boldsymbol{\theta}$ *with* i^{th} *coordinate given by (29) such that v, h, and w satisfy the assumptions of Theorem 2 with* $k \le p$, $4k(A+\frac{1}{2}) - FB > 0$. *Also assume that* $E[\sigma^2 \sum_{i=1}^{p} v(\sum_{i=1}^{p} h(x_i^2 \wedge z_k^2))w(x_i^2 \wedge z_k^2)|\sigma^2]$ *is a monotonic increasing function of* σ^2. *Then the estimator* $\delta(\mathbf{X})$ *is minimax for*

$$0 < c \le \frac{4k(A+\frac{1}{2}) - FB}{H}.$$

Proof Let

$$r = \frac{c}{E[\frac{1}{\sigma^2}]}$$

so that the ith coordinate of the estimator δ is expressible as

$$\delta_i(\mathbf{x}) = x_i - \begin{cases} rv(\sum_{i=1}^{p} h(x_i^2))w(x_i^2)x_i & \text{if } |x_i| \le z_k \\ rv(\sum_{i=1}^{p} h(x_i^2 \wedge z_k^2))w(z_k^2)z_k * sign(x_i) & \text{if } |x_i| > z_k. \end{cases}$$

The conditional (on σ^2) difference in risk, $\triangle$, between this estimator and $\mathbf{X}$ is expressible as

$$\triangle = E[\|g(\mathbf{X}) + 2(\mathbf{X}-\boldsymbol{\theta})'g(\mathbf{X})\|^2] = E[E[\|g(\mathbf{X})\|^2 + 2\sigma^2 div_{\mathbf{x}}(g(\mathbf{X}))|\sigma^2]]$$

$$= E[E[r\sigma^2 V(D)[\sum_{j=1}^{k-1} w(x_{(j)}^2) + (p-k+1)w(z_k^2)]\{\frac{L_1}{\sigma^2} + L_2 - L_3\}|\sigma^2]]$$

with L_1,L_2, and L_3 as in (14) - (16) of the proof of Theorem 2. As in the proof of Theorem 2,

$$L_1 \le \frac{rH}{\sigma^2 p},\ L_2 \le \frac{FB}{p},\ \text{ and, } L_3 \ge 4(A+\frac{1}{2})\frac{k}{p}, \tag{30}$$

where the bounds in (30) are established in (17)–(21) in the proof of Theorem 2. Let

$$M_k = \sigma^2 rV(D)[\sum_{j=1}^{k-1} w(x_{(j)}^2) + (p-k+1)w(z_k^2),$$

$$T_1(\sigma^2) = E[M_k|\sigma^2], \text{ and}$$

$$T_2(\sigma^2) = E[\frac{rH}{\sigma^2 p} + \frac{FB}{p} - 4(A+\frac{1}{2})\frac{k}{p}].$$

Therefore

$$\triangle = E[E[M_k\{\frac{L_1}{\sigma^2} + L_2 + L_3\}|\sigma^2]]$$

$$\leq E[E[M_k\{\frac{rH}{\sigma^2 p} + \frac{FB}{p} - 4(A+\frac{1}{2})\frac{k}{p}\}|\sigma^2]] \tag{31}$$

$$\leq E[T_1(\sigma^2)]E[T_2(\sigma^2)], \tag{32}$$

where (32) follows from (31) by the correlation inequality since $T_1(\sigma^2)$ is a monotonic increasing function of σ^2, and $T_2(\sigma^2)$ is a monotonic decreasing function of σ^2. Expression (32) is non-positive once

$$0 \leq r \leq \frac{4k(A+\frac{1}{2}) - FB}{HE[\frac{1}{\sigma^2}]}$$

or equivalently

$$0 \leq c \leq \frac{4k(A+\frac{1}{2} - FB)}{H}$$

establishing the result. ■

Example 3 (Extension of Truncated Maruyama Estimator) Let $\mathbf{X}|\sigma^2 \sim N_p(\boldsymbol{\theta}, \boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I_p})$ and assume that the distribution of σ^2 satisfies $E[\sigma^2]$ and $E[\frac{1}{\sigma^2}]$ are finite. Then the estimator $\delta(\mathbf{X})$ of $\boldsymbol{\theta}$ with i^{th} coordinate of the form

$$\delta_i^+(\mathbf{x}, a, c, d, k) = \begin{cases} (1 - \frac{c}{[E[\frac{1}{\sigma^2}]\sum_{i=1}^p(|x_i|^d \wedge z_k^d)]^{\frac{2-a}{d}}|x_i|^a})_+ x_i & \text{if } |x_i| \leq z_k \\ (1 - \frac{cz_k}{E[\frac{1}{\sigma^2}][\sum_{i=1}^p(|x_i|^d \wedge z_k^d)]^{\frac{2-a}{d}} z_k^a|x_i|})_+ x_i & \text{if } |x_i| > z_k \end{cases}$$

where $z_k = y_{(k)}$, $3 \leq k \leq p$, and $y_i = |x_i|$, is minimax for

$$0 < c * max\{1, p^{1-\frac{2-a}{d}}\} \leq 2(k - 2 + a(k-1)),$$

$0 < a < \frac{k-2}{k-1}$, and $d > 0$. This follows from Theorem 3 and Example 1 with the identification $h(t) = t^{\frac{d}{2}}$, $v(t) = t^{-(\frac{2-a}{d})}$, and $w(t) = t^{-\frac{a}{2}}$. Example 1 established the conditions on c, so that $cH + FB \leq 4k(A + \frac{1}{2})$. When $k =$

$p\ E[\sigma^2 \frac{c}{E[\frac{1}{\sigma^2}]} V(D) \sum_{i=1}^{p} w(x_i^2)|\sigma^2]$ is a monotonic increasing function of σ^2. To show this monotonicity, note that

$$E[\sigma^2 \frac{c}{E[\frac{1}{\sigma^2}]} v(D) \sum_{i=1}^{p} w(x_i^2)|\sigma^2]$$

$$= \frac{c}{E[\frac{1}{\sigma^2}]} E[\sigma^2 \frac{\sum_{i=1}^{p}(x_i^2)^{-\frac{a}{2}}}{(\sum_{i=1}^{p}(x_i^2)^{\frac{d}{2}})^{\frac{2-a}{d}}}|\sigma^2](\text{letting } y_i^2 = \frac{x_i^2}{\sigma^2})$$

$$= \frac{c}{E[\frac{1}{\sigma^2}]} E[\frac{\sum_{i=1}^{p}(y_i^2)^{-\frac{a}{2}}}{(\sum_{i=1}^{p}(y_i^2)^{\frac{d}{2}})^{\frac{2-a}{d}}}|\sigma^2]$$

$$= \frac{c}{E[\frac{1}{\sigma^2}]} E[\frac{\sum_{k\neq i}(y_k^2)^{-\frac{a}{2}}}{((y_i^2)^{\frac{d}{2}} + \sum_{k\neq i}(y_k^2)^{\frac{d}{2}})^{\frac{2-a}{d}}} + \frac{(y_i^2)^{-\frac{a}{2}}}{((y_i^2)^{\frac{d}{2}} + \sum_{k\neq i}(y_k^2)^{\frac{d}{2}})^{\frac{2-a}{d}}}|\sigma^2]$$

$$= \frac{c}{E[\frac{1}{\sigma^2}]} E[U_1(y_i^2, \ldots, y_p^2) + U_2(y_i^2, \ldots, y_p^2)].$$

Here, both U_1 and U_2 are monotonic decreasing in each y_i^2 for every fixed value of the other coordinates. Since $\{Y_i^2; i = 1, \ldots, p\}$ is a collection of independent random variables such that

$$Y_i^2 \sim \chi_1^2(\frac{\theta_i^2}{\sigma^2} = \nu_i)$$

is stochastically increasing in ν_i and hence stochastically decreasing in σ^2 (see, e.g., Lehmann and Romano 2005, Lemma 3.4.2), and each $U_i(y_i^2, \cdots, y_p^2)$ is a decreasing function of each of its coordinates, the random variables U_i are stochastically increasing in σ^2. It follows $\frac{c}{E[\frac{1}{\sigma^2}]} E[U|\sigma^2]$ is a monotonic increasing function of σ^2. When $3 \leq k < p$, the monotonicity of

$$E[\frac{\sum_{j=1}^{k-1}[(Y_{(j)}^2)^{\frac{-a}{2}}) + (p-k+1)(Y_{(k)}^2)^{-\frac{a}{2}}]}{[\sum_{j=1}^{k-1}(Y_{(j)}^2)^{\frac{d}{2}} + (p-k+1)(Y_{(k)}^2)^{\frac{d}{2}}]^{\frac{2-a}{d}}}|\sigma^2] = E[U(Y_{(1)}^2, Y_{(2)}^2, \ldots, Y_{(k)}^2)]$$

with respect to σ^2 where $Y_i^2 \sim \chi_1^2(\frac{\theta_i^2}{\sigma^2} = \nu_i)$, follows from U being an increasing function of each coordinate for fixed values of the other coordinates and the order statistics originate from a collection of independent random variable $\{Y_i\}$, where each Y_i^2 is stochastically increasing in ν_i (stochastically decreasing in σ^2) so that $Y_{(i)}^2$ is stochastically increasing in ν_i (stochastically decreasing in σ^2).

4 Spherically Symmetric Distributions with Residual

This section extends the results of Sect. 2 to the general spherically symmetric case with a residual vector. In this section we do not assume a known covariance matrix for the distribution and use the residual vector to estimate the unknown scale. In particular suppose

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{U}\end{pmatrix} \sim SS_{p+m}\left(\begin{pmatrix}\boldsymbol{\theta}\\ \mathbf{0}\end{pmatrix}\right)$$

with $dim(\mathbf{X}) = dim(\boldsymbol{\theta}) = p \geq 3$ and $dim(\mathbf{U}) = dim(\mathbf{0}) = m \geq 1$, and let the loss be

$$L(\mathbf{d}, \boldsymbol{\theta}) = \frac{\|\mathbf{d} - \boldsymbol{\theta}\|^2}{E[\|\mathbf{U}\|^2]}. \tag{33}$$

The covariance matrix will be of the form $\boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I}_{\mathbf{p}+\mathbf{m}}$ with σ^2 unknown. The estimate $E[\frac{\|\mathbf{U}\|^2}{m+2}]$ serves as an estimate for the unknown scale parameter σ^2. Loss (33) is chosen as an invariant loss function where the estimator $\mathbf{X}$ of the location parameter θ is a minimax. As a special case, the results of this section apply when the underlying distribution is a normal distribution with residual vector $\mathbf{U}$, and covariance $\boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I}_{\mathbf{p}+\mathbf{m}}$. The following Lemma from Fourdrinier et al. (2006) applies to general estimators of the form

$$\delta(\mathbf{X}, S) = \mathbf{X} + \frac{S}{m+2} g(\mathbf{X}), \tag{34}$$

where $S = \|\mathbf{U}\|^2$.

Lemma 2 (Fourdrinier et al. 2006) *Let* $\begin{pmatrix}\mathbf{X}\\ \mathbf{U}\end{pmatrix} \sim SS_{p+m}\left(\begin{pmatrix}\boldsymbol{\theta}\\ \mathbf{0}\end{pmatrix}\right)$ *where* $dim(\mathbf{X}) = dim(\boldsymbol{\theta}) = p \geq 3$ *and* $dim(\mathbf{U}) = dim(\mathbf{0}) = m \geq 1$. *Assume* $E[\|\mathbf{U}\|^2] < \infty$ *and that* $g(\mathbf{X})$ *is such that*

(i) $g(\mathbf{X})$ *is a weakly differentiable function,*
(ii) $E[\|\mathbf{U}\|^4 \|g\|^2] < \infty$
(iii) $\|g(\mathbf{X})\|^2 + 2div_{\mathbf{x}}(g(\mathbf{X})) \leq 0$ *a.e.*

Then the estimator (34) is minimax under the loss (33).

Theorem 4 is a straightforward consequence of Lemma 2 and Theorem 2. It is the main result of this section.

Theorem 4 *Let* $\begin{pmatrix}\mathbf{X}\\ \mathbf{U}\end{pmatrix} \sim SS_{p+m}(\begin{pmatrix}\boldsymbol{\theta}\\ \mathbf{0}\end{pmatrix})$ *where* $dim(\mathbf{X}) = dim(\boldsymbol{\theta}) = p \geq 3$ *and* $dim(\mathbf{U}) = dim(\mathbf{0}) = m \geq 1$ *such that all second moments exist. Let* $Z_k = Y_{(k)}$

where $Y_i = |X_i|$ *for* $i = 1, \ldots, p$ *and* $3 \leq k \leq p$. *Then the estimator*

$$\delta(\mathbf{X}, \|\mathbf{U}\|^2) = \begin{cases} (1 - \frac{\|\mathbf{U}\|^2 cv(D)w(X_i^2)}{(m+2)})X_i & \textit{if } |X_i| \leq Z_k \\ \\ (1 - \frac{\|\mathbf{U}\|^2 cv(D)w(Z_k^2)Z_k}{(m+2)|X_i|})X_i & \textit{if } |X_i| > Z_k, \end{cases} \tag{35}$$

where $D = \sum_{i=1}^{p} h(X_i^2 \wedge Z_k^2)$, *and v, w, and h satisfy the assumptions of Theorem 2 with,*

$$4k(A + \frac{1}{2}) - FB \geq 0$$

and

$$0 < c \leq \frac{4k(A + \frac{1}{2}) - FB}{H}$$

is minimax under loss (33) provided assumption ii) of Lemma 2 is also satisfied. If, in addition, $\begin{pmatrix} \mathbf{X} \\ \mathbf{U} \end{pmatrix}$ *has a density, f that is unimodal, the positive-part version of this estimator dominates the original estimator,* **X**.

Proof With the identification

$$g_i(\mathbf{x}) = \begin{cases} cv(D)x(x_i^2)x_i & \text{if } |x_i| < z_k \\ \frac{cV(D)w(z_k^2)z_k x_i}{|x_i|} & \text{if } |x_i| \geq z_k \end{cases}$$

Equation (35) is expressible as $\delta(\mathbf{X}, S) = \mathbf{X} + \frac{Sg(\mathbf{X})}{m+2}$, where $S = \|\mathbf{U}\|^2$. Therefore the difference in risk between the minimax estimator, **X**, and $\boldsymbol{\delta}$ is proportional to

$$E\left[\frac{S^2}{(m+2)^2}\|g(\mathbf{X})\|^2 + 2\frac{S}{m+2}(\mathbf{X} - \boldsymbol{\theta})' g(\mathbf{X})\right]. \tag{36}$$

Lemma 2 establishes sufficient conditions in which estimator (35) is minimax by expressing (36) as

$$E\left[\frac{S^2}{(m+2)^2}\|g(\mathbf{X})\|^2 + 2\frac{S^2}{(m+2)^2} div_{\mathbf{X}}(g(\mathbf{X}))\right]$$

using the equality $E[S(\mathbf{X} - \boldsymbol{\theta})' g(\mathbf{X})] = \frac{1}{(m+2)} E[S^2 div_{\mathbf{x}}(g(\mathbf{X}))]$. Therefore once the assumptions of Theorem 2 are satisfied, assumption iii) of Lemma 2 is satisfied, proving the result. Assumption ii) of Lemma 2 is sufficient for the risk of the estimator to exist.

5 A Simulation Study

In this section we study the risk function and the probability that $|\hat{\theta}_i| \neq 0$ for $\delta(\mathbf{X}, a, d, c, k)$ with coordinates of the form

$$\delta_i(\mathbf{x}, a, c, d, k) = \begin{cases} (1 - \frac{c}{[\sum_{i=1}^{p}(|x_i|^d \wedge z_k^d)]^{\frac{2-a}{d}} |x_i|^a})_+ x_i & \text{if } |x_i| < z_k \\ (1 - \frac{cz_k}{[\sum_{i=1}^{p}(|x_i|^d \wedge z_k^d)]^{\frac{2-a}{d}} z_k^a |x_i|})_+ x_i & \text{if } |x_i| \geq z_k \end{cases} \tag{37}$$

in the Gaussian case. We assume without loss of generality that $\sigma = 1$, since for all estimators under consideration $R(\boldsymbol{\theta}, \sigma, \delta) = R(\frac{\boldsymbol{\theta}}{\sigma}, 1, \delta)$. The dimension, p, in each of the simulations is 12.

For $\delta(\mathbf{X}, a, d, c, k)$ in (37) we consider $a = 0.2$ and $d = 1.8$ with $k = 6, 9, 11$, and 12, and the following one dimensional subspaces:

$$D_1 = \langle \mathbf{e_1} \rangle, \tag{38}$$

and

$$D_3 = \langle \mathbf{e_1} + \mathbf{e_2} + \mathbf{e_3} \rangle, \tag{39}$$

where $\{\mathbf{e_i}\}_{i=1,2,\ldots,12}$ denotes the standard Euclidean basis in $\mathbb{R}^{12}$. The shrinkage constant, c, is chosen to be $c = 2(k - 2 - a(k - 1))$, i.e., the largest value leading to minimaxity. The case $k = 12$ corresponds to the Zhou and Hwang estimate. To illustrate the effect of basing shrinkage on the ℓ_{2-a} norm we also include the positive-part James–Stein estimators (corresponding to $a = 0$ and $k = 12, d = 2$) with the shrinkage constant $2(p-2)$. The mvrnorm (Genz et al. 2020) function in the MASS (Venables and Ripley 2002) package for R version 4.02 was used to generate 70,000 simulations from a multivariate normal distribution with mean $0.1j\mathbf{e_1}$, and $0.25j(\mathbf{e_1}+\mathbf{e_2}+\mathbf{e_3})$, and covariance equal to $\mathbf{I_{12}}$ for every $j \in \{0, 1, 2, \ldots, 99\}$. After each simulation the estimator was computed and the loss along with the output of an indicator function specifying if $|\delta_i| > 0$ was recorded.

Figures 1 and 2 plot the risks of $\delta(\mathbf{X})$ when $a = 0.2$ for the varying parameter k, and compares it to the positive-part James–Stein estimator over the spaces D_1 and D_3, respectively, as functions of $\|\theta\|^2$.

Note that the risks of the positive-part James–Stein estimators depend only on $\|\boldsymbol{\theta}\|^2$ and not on the particular one dimensional subspace, however, the risk of the two Zhou-Hwang estimators, along with their truncated versions depend both on $\|\boldsymbol{\theta}\|^2$ and on the particular one dimensional subspace D_i. The risk of the Zhou and Hwang estimator as well as the positive-part James–Stein estimator approaches the risk of the minimax estimator $\mathbf{X}$ from below for sufficiently large $\|\boldsymbol{\theta}\|^2$. For each D_i, at the origin, the positive-part James–Stein estimator with $c = 20$ has the smallest risk, followed by the Zhou-Hwang estimator, then the three truncated versions of

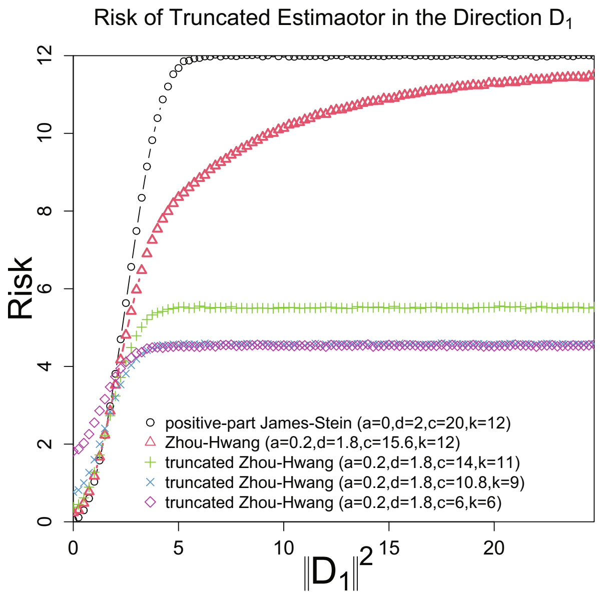


Fig. 1 Risk of the estimator δ (37), for $a = 0.2$, $d = 1.8$, $p = 12$, and varying k, vs positive-part James–Stein estimator, in the direction D_1 (38)

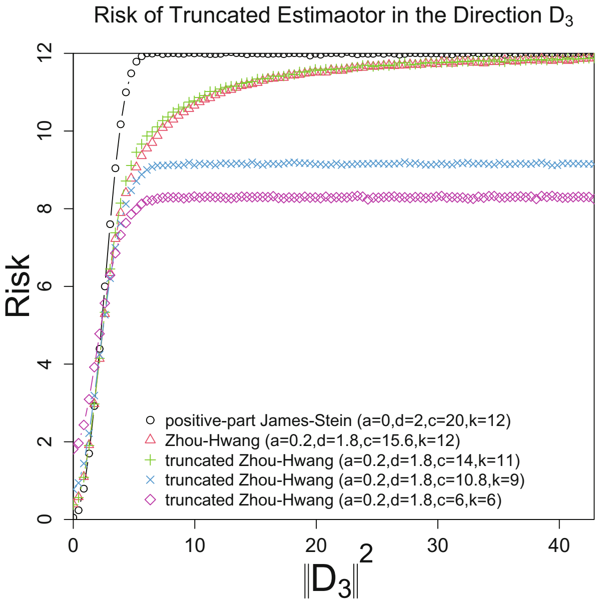


Fig. 2 Risk of the estimator δ (37), for $a = 0.2$, $d = 1.8$, $p = 12$, and varying k, vs positive-part James–Stein estimator, in the direction D_3 (39)

the Zhou-Hwang Estimator with $k = 11$ having smaller risk than when $k = 9$, and $k = 6$ having the highest risk.

When $\boldsymbol{\theta} \in D_1$ and $\|\theta\|^2$ is sufficiently large, the risks of the truncated versions of the Zhou and Hwang estimators have asymptotes that are less than the risk of the

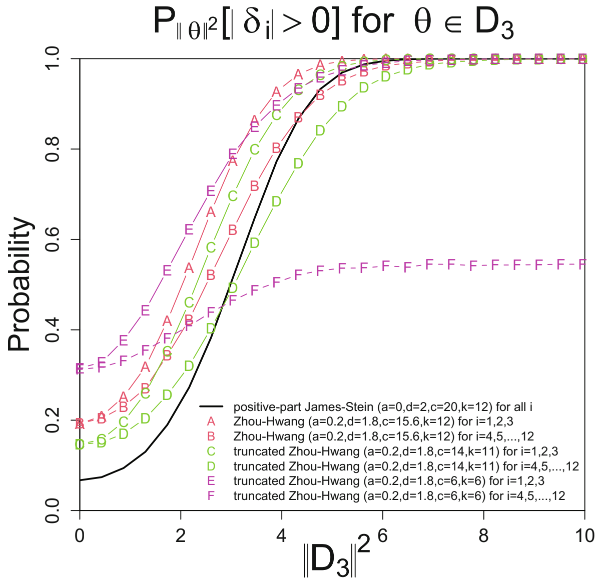


Fig. 3 $P_{\|\boldsymbol{\theta}\|^2}(|\delta_i| > 0)$ for δ given by (37) vs positive-part James–Stein estimator, for $\theta \in D_3$ (39) when $p = 12$

minimax estimator, **X**. For sufficiently large $\|\boldsymbol{\theta}\|^2$ the risk of the truncated estimator when $k = 9$ and 6 is similar to and less than the risk of the truncated estimator when $k = 11$. For $\boldsymbol{\theta} \in D_3$ the risks of the truncated versions of the Zhou-Hwang estimators when $k = 6$ and $k = 9$, approach values less than the risk of the minimax estimator **X** with $k = 6$ having smaller asymptote than when $k = 9$), however, when $k = 11$ the truncated version of the Zhou-Hwang estimator approaches the risk of the estimator, **X**, from below. This implies when k is misspecified and $p - k < s$, where s denotes the number of θ_i in the active set, the estimators will have risk comparable to the untruncated versions of the Zhou-Hwang estimators for sufficiently large $\|\boldsymbol{\theta}\|^2$.

To study the probability of correct selection, Fig. 3 plots the $P_{\|\boldsymbol{\theta}\|^2}(|\delta_i| > 0)$ for the positive-part James–Stein estimator, the Zhou-Hwang estimator, and the truncated versions of the Zhou-Hwang estimator for $k = 11$ and 6 from Figs. 1 and 2 in the subspace D_3. Note that for the positive-part James–Stein estimator this probability only depends on $\|\boldsymbol{\theta}\|^2$ and not on the coordinate θ_i. For the Zhou-Hwang estimators (both truncated and untruncated), however, this probability differs from coordinate to coordinate depending on $\|\boldsymbol{\theta}\|^2$. For example, the curves for $\theta_4, \theta_5, \ldots, \theta_{12}$ coincide and differ from that for θ_1, θ_2, and θ_3.

Note that for $\boldsymbol{\theta} \in D_3$, curve A (corresponding to $i = 1, 2, 3$) is larger than curve B (corresponding to $i = 4, 5,\ldots,12$), curve C (corresponding to $i = 1, 2, 3$) is larger than curve D (corresponding to $i = 4, 5,\ldots,12$), and curve E (corresponding to $i = 1, 2, 3$) is larger than curve F (corresponding to $i = 4, 5,\ldots,12$). This indicates that the Zhou-Hwang procedure has noticeably higher probability of including non-zero θ_i's in the model than those θ_i's such that $\theta_i = 0$. However, only the truncated versions, when $p - k \geq s$, where s denoted the number of non-zero θ_i's, is able to do so for large values of $\|\boldsymbol{\theta}\|^2$.

Although analytic bounds on the risk of estimator (37) over the parameter subspace generated by the non-zero θ_i's, when k is correctly specified and $3 \leq k \leq p - s$, seems intractable, we conjecture that the risk of truncated Zhou-Hwang will have an asymptote that is strictly less than the minimax estimator, $\mathbf{X}$ due to fact that the truncated Zhou-Hwang procedure will estimate non-active coordinates as 0 with a non-zero probability over the parameter space, thereby decreasing the risk to a value less than the minimax estimator, $\mathbf{X}$, as there will be no contribution to the loss when $|\hat{\theta}_i|$ is estimate as 0 when $\theta_i = 0$.

6 Summary and Conclusion

In this paper we extend the results of Zhou and Hwang (2005) and Maruyama (2014) for the Gaussian case with covariance matrix $\boldsymbol{\sigma^2 \mathbf{I}}$, to broad classes of spherically symmetric distributions including the case of scale mixtures of normal distributions, and to the general class of spherically symmetric distributions with residual vector and unknown scale (including the normal distribution with a residual vector and covariance equal to an unknown scale times identity). Extension to the class of estimators discussed by Zhou and Hwang (2005) and Maruyama (2014) is also given. A numerical study suggests that certain classes of these estimators preform favorably in terms of risk when compared to the positive-part James–Stein estimator and are also successful in differentially estimating null coordinate parameters as 0. An interesting feature is the development of such estimators based on the order statistics of the $|\mathbf{X}|$'s. In certain subspaces, these estimators have asymptotic risks which are strictly less than the minimax risk and probability of selection of inactive (mean 0) coordinates strictly less than 1.

Acknowledgments Dave Tyler has made significant contributions to the theory and applications of multivariate analysis and robustness. To the first author he has been a valued professor and to the second author, a valued colleague and friend for more than 40 years. We are pleased to contribute to this volume in his honor. This work was partially supported by grants from the Simons Foundation (# 209035 and # 418098 to William Strawderman).

References

Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *The Annals of Statistics, 27*(3), 255–264.

Casella, G., & Berger, R. (2002). *Statistical inference* (2nd ed.) Duxbury, California.

Donoho, D., & Johnstone, I. (1994). Ideal spatial adaption via wavelet shrinkage. *Biometrika, 81*, 425–455.

Donoho, D., & Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association, 90*(432), 1200–1224.

Fourdrinier, D., Strawderman, W., & Wells, M. (2006). Estimation of a location parameter with restrictions or "vague information" for spherically symmetric distributions. *AISM, 58*, 73–92.

Fourdrinier, D., Strawderman, W., & Wells, M. (2018). *Shrinkage estimation* (1st ed.) Switzerland: Springer Nature.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2020). mvtnorm: Multivariate normal and t distributions. https://CRAN.R-project.org/package=mvtnorm.

Lehmann, E., & Romano, J. (2005). *Testing statistical hypothesis* (3rd ed.). New York: Springer-Verlag.

Maruyama, Y. (2014). ℓ_p – based James-Stein estimation with minimaxity and sparsity. Arxiv (pp. 735–750)

Stein, C. (1981). Estimation of the mean of multivariate normal distribution. *The Annals of Statistics, 9*, 1135–1151.

Strawderman, W. (1974). Minimax estimation of location parameters for certain spherically symmetric distributions. *Journal of Multivariate Analysis, 4*, 255–264.

Venables, W., & Ripley, B. (2002). *Modern applied statistics with S* (4th ed.) New York: Springer.

Zhou, H., & Hwang, J. (2005). Minimax estimation with thresholding and its application to wavelet analysis. *The Annals of Statistics, 33*, 101–125.

On the Finite-Sample Performance of Measure-Transportation-Based Multivariate Rank Tests

Marc Hallin and Gilles Mordant

Abstract Extending to dimension two and higher the dual univariate concepts of ranks and quantiles has remained an open problem for more than half a century. Based on measure transportation results, a solution has been proposed recently under the name *center-outward ranks and quantiles*. Contrary to previous proposals, center-outward ranks enjoy all the properties that make univariate ranks a successful tool for statistical inference. Just as their univariate counterparts (to which they reduce in dimension one), they allow for the construction of distribution-free and asymptotically efficient tests for a variety of problems where the density of some underlying noise or innovation remains unspecified. The actual implementation of these tests involves the somewhat arbitrary choice of a grid. While the asymptotic impact of that choice is nil, its finite-sample consequences are not. In this note, we investigate this finite-sample impact in the typical context of the multivariate two-sample location problem.

Keywords Measure transportation · Multivariate ranks · Two-sample rank tests

1 Introduction

1.1 David Tyler, Beyond Affine Equivariance and Elliptical Symmetry

The closely related concepts of affine equivariance and elliptical symmetry played a central role in the development of robust multivariate statistics over the

M. Hallin (✉)
ECARES and Département de Mathématique, Université libre de Bruxelles, Brussels, Belgium
e-mail: mhallin@ulb.ac.be

G. Mordant
IMS, Universität Göttingen, Göttingen, Germany

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_5

past 60 years.[1] A critical attitude toward this dominant role of elliptical densities constitutes a red thread running through all of David's contributions to multivariate analysis[2]—an attitude that actually takes place in a broader debate on the ordering of the real space in dimension $d \geq 2$. Such ordering is an essential issue if the univariate concepts of distribution and quantile functions, ranks, and signs, all heavily depending on the canonical ordering of the real line, are to be extended to a multivariate context.

1.2 Ordering the Real Space in Dimension d ≥ 2

The problem of ordering $\mathbb{R}^d$ for $d \geq 2$, hence ranking multivariate observations, has a long history in statistics. Many attempts have been made to define adequate multivariate concepts of ranks.

The notion of rank, however, cannot be an isolated one and is inseparable from that of empirical quantiles, quantile regions and quantile contours (collections of points with ranks less than or equal to, or equal to, some given value), and (collections of points having the same rank). A sound definition thus should include all these concepts, along with their population versions—the population distribution and quantile functions F and $Q := F^{-1}$—and their mutual relations (a quantile function is the inverse of a distribution function; a population distribution function and its empirical version are asymptotically related via a Glivenko–Cantelli result, etc.). Among the key properties of any successful concept are the distribution-freeness (within the class of absolutely continuous distributions P, say) of the ranks and the *push-forward*[3] of a distribution P by its distribution function F. Without that property, the level of a quantile $Q(\tau) = F^{-1}(\tau)$ depends on the distribution P characterized by F and can be anything larger or smaller than τ: as a quantile, thus, it is totally meaningless.

Appealing as they are, none of the attempts that had been made until recently—marginal ranks, spatial ranks, elliptical (or Mahalanobis) ranks, etc. ...—is satisfying the desired properties; actually, none of them is even enjoying distribution-freeness. Nor do the various depth concepts: the probability content of a depth contour of given depth strongly depends on the underlying P, which hinders its interpretation as a quantile contour.

[1] Tukey (1960), Huber (1964), and Hampel (1968) generally are considered as laying the foundations of modern robust statistics; see Ronchetti (2006) for a historical perspective and Stigler (1973) for an account of the pre-Tukey era.

[2] Significantly, *"Robust Multivariate Statistics: Beyond Ellipticity and Affine Equivariance"* is the title of one of David's NSF grants.

[3] We adopt here the convenient terminology and notation of measure transportation: the *push-forward* $F\#\mathrm{P}$ of P by F is the distribution of $F(Z)$ where $Z \sim \mathrm{P}$, i.e., $F(Z) \sim F\#\mathrm{P}$ if $Z \sim \mathrm{P}$.

Based on measure transportation results (mainly, a theorem by McCann 1995), Chernozhukov et al. (2017), Hallin (2017), and Hallin et al. (2021) recently introduced the concepts of *center-outward ranks* and *signs*, *distribution* and *quantile functions* which, for the first time, satisfy all the desired properties (see Hallin et al. 2021 and the review by Hallin 2022 for details) and further triggered the development of several appealing multivariate, distribution-free statistical procedures, among which (Deb et al. 2021, 2020; Deb & Sen 2022; Faugeras & Rüschendorf 2017; Hallin et al. 2022a,b,c; Shi et al. 2022a, 2021, 2022b). These are the concepts we are considering, under various versions, in this note and describe in Sect. 2. Section 3 provides details pertaining to the simulations of Sects. 4 and 5. These simulations, by shedding some light on the power of center-outward rank-based tests, constitute the main contribution of this paper.

2 Center-Outward Ranks and Signs

For the simplicity of exposition, we throughout consider distributions P on $\mathbb{R}^d$ in the family $\mathcal{P}^d$ of Lebesgue-absolutely continuous distributions with *nonvanishing densities*, that is, with a density f such that for all $B > 0$, there exist $m_B^- \leq m_B^+$ such that $0 < m_B^- \leq f(\mathbf{z}) \leq m_B^+ \leq \infty$ for all $\mathbf{z}$ such that $\|\mathbf{z}\| \leq B$. That assumption can be relaxed, though, see del Barrio et al. (2020).

2.1 Measure Transportation-Based Concepts of Distribution and Quantile Functions

The basic idea behind the definitions of the center-outward distribution and quantile functions of a probability measure $\mathrm{P} \in \mathcal{P}^d$ is quite simple. For $d = 1$, the distribution function F of P is the unique monotone increasing function pushing P forward to the uniform $\mathrm{U}_{[0,1]}$ over $[0, 1]$—namely, $F\#\mathrm{P} = \mathrm{U}_{[0,1]}$. Rather than F, however, which is based on a left-to-right ordering of $\mathbb{R}$ that does not extend to $\mathbb{R}^d$ for $d \geq 2$, we consider the *center-outward distribution function* $F_\pm := 2F - 1$, which contains the same information as F and is the unique monotone increasing function pushing P forward to the uniform $\mathrm{U}_{[-1,1]}$ over $[-1, 1]$. A monotone increasing function is the gradient (the derivative) of a convex function: the center-outward distribution function $F_\pm$ actually, is the unique gradient of a convex function such that $F_\pm\#\mathrm{P} = \mathrm{U}_{[-1,1]}$. The interval $[-1, 1]$ is, for $d = 1$, the closed unit ball $\bar{\mathbb{S}}_d$, where $\mathbb{S}_d := \left\{\mathbf{u} | \|\mathbf{u}\| < 1\right\}$ and, denoting by U_d the spherical uniform[4] over $\bar{\mathbb{S}}_d$, the spherical uniform U_1 over $\bar{\mathbb{S}}_1$ coincides with the Lebesgue uniform $\mathrm{U}_{[-1,1]}$ over $[-1, 1]$.

[4] The *spherical uniform* U_d over $\bar{\mathbb{S}}_d$ is the spherical distribution with center $\mathbf{0}$ and radial density the uniform over $[0, 1]$: it is thus the product of a uniform over $[0, 1]$ for the distances to the origin and a uniform over the unit (hyper)sphere for the directions.

A celebrated theorem by McCann (1995) tells us that, for arbitrary dimension $d \in \mathbb{N}$ and arbitrary $\mathrm{P} \in \mathcal{P}^{\mathrm{d}}$, there exists a (P-a.s., here Lebesgue-a.e.) unique gradient of a convex function $\mathbf{F}_{\pm}$ such that $\mathbf{F}_{\pm}\#\mathrm{P} = \mathrm{U}_d$. Obviously, for $d = 1$, $\mathbf{F}_{\pm}$ coincides with the univariate $F_{\pm}$, whence the notation. Call $\mathbf{F}_{\pm}$ the *center-outward distribution function* of P. It follows from Figalli (2018) that—except perhaps at $\mathbf{F}_{\pm}^{-1}(\mathbf{0})$ (a set of points with Lebesgue measure zero)—$\mathbf{F}_{\pm}\#\mathrm{P}$ is a homeomorphism and hence admits a continuous (except perhaps at $\mathbf{0}$) inverse $\mathbf{Q}_{\pm} := \mathbf{F}_{\pm}^{-1}$: let $\mathbf{Q}_{\pm}(\mathbf{0}) := \mathbf{F}_{\pm}^{-1}(\{\mathbf{0}\})$ and call $\mathbf{Q}_{\pm}$ the *center-outward quantile function* of P. Clearly, $\mathbf{Q}_{\pm}\#\mathrm{U}_d = \mathrm{P}$.

This, with the spherical uniform U_1 U_d (extending $\mathrm{U}_{[-1,1]}$) as a reference distribution, is the concept proposed in Hallin (2017) and Hallin et al. (2021), where we refer to for further properties of $\mathbf{F}_{\pm}$ and $\mathbf{Q}_{\pm}$ justifying their qualification as distribution and quantile functions.

Other choices are possible for the reference U, though. Replacing U_d with an arbitrary compactly supported absolutely continuous reference distribution U, Chernozhukov et al. (2017), in a very general approach, propose, under the name of *Monge-Kantorovich vector rank* and *Monge-Kantorovich quantile* functions, measure-transportation-based definitions of a broad class of analogues, $\mathbf{F}_{\mathrm{MK}}$ and $\mathbf{Q}_{\mathrm{MK}}$, say, of $\mathbf{F}_{\pm}$ and $\mathbf{Q}_{\pm}$. For nonspherical U's, however, the Monge–Kantorovich quantile functions do not enjoy all the features expected from a quantile function;[5] Chernozhukov et al. (2017) therefore also introduce a concept of *Monge-Kantorovich depth* D_{MK}—a transformation–retransformation version (based on the Monge–Kantorovich vector rank function) of the classical Tukey depth D_{Tukey}. For spherical U's, the Monge–Kantorovich depth and quantile contours coincide. More precisely, defining $\delta(\tau) := D_{\mathrm{Tukey}}(\mathbf{u}_{\tau})$, where $\mathrm{U}\big(\big\{\mathbf{u}\,\big|\,\|\mathbf{u}\| \leq \|\mathbf{u}_{\tau}\|\big\}\big) = \tau$, one has $\Big\{\mathbf{z}\,\Big|\,\|\mathbf{F}_{\mathrm{MK}}\| = \tau\Big\} = D_{\mathrm{MK}}^{-1}(\delta(\tau))$. Recurring to depth in order to construct quantile regions and contours, thus, is not necessary in the case of a spherical reference U which, in that respect, offers a better conceptual coherence between the resulting notions of vector ranks and quantiles. As far as rank tests are concerned, however, this can be considered a minor concern.

The choice for U of the nonspherical Lebesgue uniform $\mathrm{U}_{[0,1]^d}$ over the unit (in the canonical basis) hypercube $[0, 1]^d$—call it the *cubic uniform*—yields a vector rank function $\mathbf{F}_{\mathrm{MK}}$ that reduces, for $d = 1$, to the classical distribution function F just as $\mathbf{F}_{\pm}$ reduces to $F_{\pm}$. Despite poor equivariance properties,[6] its use has been advocated by several authors: see, e.g., Faugeras and Rüschendorf (2017), Carlier et al. (2016), Deb et al. (2021, 2020), and Deb and Sen (2022).

[5] On this point, see Section 3.4 in Hallin (2022).

[6] Contrary to $\mathbf{F}_{\pm}$, which is nicely equivariant, the rank vector function $\mathbf{F}_{\mathrm{MK}}$ associated with the cubic uniform $\mathrm{U}_{[0,1]^d}$ is highly non-equivariant under orthogonal transformations.

2.2 Multivariate Ranks and Signs

Denote by $\mathbf{Z}^{(n)} := (\mathbf{Z}_1^{(n)}, \ldots, \mathbf{Z}_n^{(n)})$ an i.i.d. sample with distribution $\mathrm{P} \in \mathcal{P}^{\mathrm{d}}$. The empirical counterpart $\mathbf{F}_{\pm}^{(n)}$ of $\mathbf{F}_{\pm}$ is obtained as the solution of an optimal pairing problem between the sample values $\mathbf{Z}_1^{(n)}, \ldots, \mathbf{Z}_n^{(n)}$ and a "regular" grid $\mathfrak{G}^{(n)}$ with gridpoints $\mathfrak{s}_1^{(n)}, \ldots, \mathfrak{s}_n^{(n)}$. Precisely, $\left(\mathbf{F}_{\pm}^{(n)}(\mathbf{Z}_1^{(n)}), \ldots, \mathbf{F}_{\pm}^{(n)}(\mathbf{Z}_n^{(n)})\right)$ is defined as the minimizer $\left(\mathfrak{s}_{\pi^*(1)}^{(n)}, \ldots, \mathfrak{s}_{\pi^*(n)}^{(n)}\right)$, over the $n!$ possible permutations $\pi \in \Pi_n$ of the integers $\{1, \ldots, n\}$, of $\sum_{i=1}^{n} \left\| \mathbf{Z}_i^{(n)} - \mathfrak{s}_{\pi(i)}^{(n)}) \right\|^2$.

The choice of the grid $\mathfrak{G}^{(n)}$, of course, depends on the reference distribution U adopted in the definitions of Sect. 2.1: in particular, the uniform discrete distribution over the n gridpoints $\mathfrak{s}_1^{(n)}, \ldots, \mathfrak{s}_n^{(n)}$ should converge weakly to U as $n \to \infty$. Our objective is to investigate the finite-sample performance of the two-sample location tests based on

- (Ti) the empirical center-outward distribution function $\mathbf{F}_{\pm}^{(n)}$ associated with the spherical uniform reference distribution $\mathrm{U} = \mathrm{U}_{\mathrm{d}}$;
- (Tii) the empirical Monge–Kantorovich vector ranks $\mathbf{F}_{\Box}^{(n)}$ associated with the cubic uniform reference distribution $\mathrm{U} = \mathrm{U}_{[0,1]^{\mathrm{d}}}$;
- (Tiii) the empirical Monge–Kantorovich vector ranks $\mathbf{F}_{\pm\mathcal{N}}^{(n)}$ associated with the Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ reference considered as a spherical distribution;
- (Tiv) the empirical Monge–Kantorovich vector ranks $\mathbf{F}_{\Box\mathcal{N}}^{(n)}$ associated with the Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ reference considered as a product of univariate standard normal distributions.

The grids we are using for these four cases are constructed as follows (see Sect. 3.1 for details on Halton sequences and the choice of n_R and n_S):

- ($\mathfrak{G}$i) $\mathrm{U} = \mathrm{U}_{\mathrm{d}}$: (a) factorize n into $n = n_R n_S + n_0$ with $n_0 < \min(n_R, n_S)$, (b) generate a Halton sequence $\mathfrak{S}_{(n_S)} := (\mathbf{u}_1, \ldots \mathbf{u}_{n_S})$ over the unit (hyper)-sphere $\mathcal{S}_{d-1}$, and (c) the grid $\mathfrak{G}^{(n)}$ consists of the intersections of these n_S unit vectors with the n_R hyperspheres centered at $\mathbf{0}$, with radii $j/(n_R + 1)$, $j = 1, \ldots, n_R$—along with n_0 copies of the origin;
- ($\mathfrak{G}$ii) $\mathrm{U} = \mathrm{U}_{[0,1]^{\mathrm{d}}}$: the grid $\mathfrak{G}^{(n)}$ is a Halton sequence over $[0, 1]^d$;
- ($\mathfrak{G}$iii) $\mathrm{U} = \mathcal{N}(\mathbf{0}, \mathbf{I}_{\mathrm{d}})$, spherical grid: the grid $\mathfrak{G}^{(n)}$ is the image, by the radial transformation $\mathbf{z} \mapsto \sqrt{F_{\chi_d^2}^{-1}(\|\mathbf{z}\|)}\mathbf{z}$, of the spherical grid constructed in (i), where $F_{\chi_d^2}$ denotes the chi-square distribution function with d degrees of freedom;

($\mathfrak{G}$iv) U $= \mathcal{N}(\mathbf{0}, \mathbf{I}_{\rm d})$, cubic grid: the grid $\mathfrak{G}^{(n)}$ is the image, by componentwise application of the standard normal quantile transformation $z_i \mapsto \Phi^{-1}(z_i)$,[7] of a Halton sequence over $[0, 1]^d$.

Remark 1 The grid $\mathfrak{G}^{(n)}$ in ($\mathfrak{G}$i) reduces, for $d = 1$, to

$$\{\pm 1/(\lceil n/2\rceil + 1), \ldots, \pm\lceil n/2\rceil/(\lceil n/2\rceil + 1)\}$$

along with the origin in case n is odd; that grid is of the form

$$\{2(1/(n + 1)) - 1, \ldots, 2(n/(n + 1)) - 1\},$$

where $\{(1/(n + 1)), \ldots, (n/(n + 1))\}$ is the grid producing traditional univariate ranks to which the grid $\mathfrak{G}^{(n)}$ in ($\mathfrak{G}$ii) also reduces for $d = 1$.

Remark 2 In ($\mathfrak{G}$i) and ($\mathfrak{G}$iii), the grid $\mathfrak{G}^{(n)}$ is spherical; as a consequence, $\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)$ and $\mathbf{F}^{(n)}_{\pm\mathcal{N}}(\mathbf{Z}^{(n)}_i)$ in (Ti) and (Tiii) naturally factorize as

$$\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i) = \|\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)\| \frac{\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)}{\|\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)\|} =: \frac{R^{(n)}_{i\pm}}{n_R + 1}\mathbf{S}^{(n)}_{i\pm},$$

where $R^{(n)}_{i\pm} = (n_R + 1)\|\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)\|$, ranging from 0 or 1 (according as $n_0 \neq 0$ or $n_0 = 0$) to n_R, is the *center-outward rank* of $\mathbf{Z}^{(n)}_i$ and $\mathbf{S}^{(n)}_{i\pm}$ (a unit vector) has the interpretation of a (multivariate) *center-outward sign* and

$$\mathbf{F}^{(n)}_{\pm\mathcal{N}}(\mathbf{Z}^{(n)}_i) = \|\mathbf{F}^{(n)}_{\pm\mathcal{N}}(\mathbf{Z}^{(n)}_i)\| \frac{\mathbf{F}^{(n)}_{\pm\mathcal{N}}(\mathbf{Z}^{(n)}_i)}{\|\mathbf{F}^{(n)}_{\pm\mathcal{N}}(\mathbf{Z}^{(n)}_i)\|} =: J_{\rm vdW}\left(\frac{R^{(n)}_{i\pm\mathcal{N}}}{n_R + 1}\right)\mathbf{S}^{(n)}_{i\pm\mathcal{N}}, \qquad (1)$$

where $J_{\rm vdW} = \sqrt{F^{-1}_{\chi^2_d}}$ is the univariate *normal* or *van der Waerden score function*, $R^{(n)}_{i\pm\mathcal{N}}$ the rank of $\|\mathbf{F}^{(n)}_{\pm\mathcal{N}}(\mathbf{Z}^{(n)}_i)\|$ among the n_R distinct values of $\|\mathbf{F}^{(n)}_{\pm\mathcal{N}}(\mathbf{Z}^{(n)}_i)\|$ for $i = 1, \ldots, n$, and $\mathbf{S}^{(n)}_{i\pm\mathcal{N}}$ similarly has the interpretation of a multivariate sign. Being based on different transport maps, however, neither $R^{(n)}_{i\pm}$ and $R^{(n)}_{i\pm\mathcal{N}}$ nor $\mathbf{S}^{(n)}_{i\pm}$ and $\mathbf{S}^{(n)}_{i\pm\mathcal{N}}$ need to coincide.

Remark 3 No similar factorization into ranks and signs occurs with the vector ranks $\mathbf{F}^{(n)}_{\square}$ and $\mathbf{F}^{(n)}_{\square\mathcal{N}}$ in (Tii) and (Tiv).

[7] As usual, we denote by Φ the standard normal distribution function and by Φ^{-1} the standard normal quantile function.

2.3 Distribution-Free Tests Based on Center-Outward Ranks and Signs

Hallin et al. (2022a) propose, for multiple-output regression models with unspecified noise distribution $\mathrm{P} \in \mathcal{P}^{\mathrm{d}}$, fully distribution-free yet, for adequate choice of scores, parametrically efficient center-outward rank tests of the null hypothesis of no-treatment effect based on the empirical center-outward distribution functions $\mathbf{F}_{\pm}^{(n)}$ (hence, the center-outward ranks and signs).

The particular case of two-sample location is treated by Deb et al. (2021) who also consider tests based on the empirical Monge–Kantorovich vector ranks $\mathbf{F}_{\mathrm{MK}}^{(n)}$ associated with various reference distributions.

2.3.1 Score Functions

In line with the classical theory developed, e.g., by Hájek and Šidák (1967), rank-based statistics, irrespective of the reference distribution, involve *score functions* or *scores*. Depending on the context, a score function is a mapping $\mathbf{J}$ from the unit ball $\mathbb{S}_d$ or the unit cube $[0, 1]^d$ to $\mathbb{R}^d$ satisfying some mild regularity assumptions (continuity, square integrability, etc., see, e.g., Hallin et al. (2022a), Assumption 3.1). The only score functions we are considering here are the *Wilcoxon*, the *spherical van der Waerden*, and the *marginal van der Waerden score functions*

$$\mathbf{J}_{\mathrm{W}}(\mathbf{u}) := \mathbf{u}, \quad \mathbf{J}^{\pm}_{\mathrm{vdW}}(\mathbf{u}) := \sqrt{F^{-1}_{\chi^2_d}(\|\mathbf{u}\|)}\,\frac{\mathbf{u}}{\|\mathbf{u}\|}, \text{ and } \mathbf{J}^{\square}_{\mathrm{vdW}}(\mathbf{u}) := \left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

respectively, where $F_{\chi^2_d}$ and Φ stand for the (univariate) chi-square (d degrees of freedom) and standard normal distribution functions.

2.3.2 Test Statistics

For simplicity, our investigation here is limited to the particular case of two-sample location models, where the n observations are i.i.d. under the null and consist of two samples, $\mathbf{Z}_1^{(n)}, \ldots, \mathbf{Z}_{n_1}^{(n)}$ and $\mathbf{Z}_{n_1+1}^{(n)}, \ldots, \mathbf{Z}_{n_1+n_2}^{(n)}$, with $n_1 + n_2 = n$. The classical procedure for this problem is Hotelling's test, based on a quadratic statistic of the form

$$\left(T^{(n)}_{\mathrm{Hot}}\right)^2 := \boldsymbol{\Delta}^{(n)\prime}_{\mathrm{Hot}}\left(\boldsymbol{\Sigma}^{(n)}_{\mathrm{Hot}}\right)^{-1}\boldsymbol{\Delta}^{(n)}_{\mathrm{Hot}},$$

where $\boldsymbol{\Sigma}^{(n)}_{\mathrm{Hot}}$ is the estimated (under the null) covariance matrix of

$$\boldsymbol{\Delta}^{(n)}_{\mathrm{Hot}} := \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{Z}^{(n)}_i - \frac{1}{n_2}\sum_{i=n_1+1}^{n}\mathbf{Z}^{(n)}_i .$$

The Hotelling test is parametrically efficient under Gaussian assumptions; it remains asymptotically valid,[8] however, under mild moment assumptions and therefore qualifies as a pseudo-Gaussian procedure.

For score functions $\mathbf{J}$, the center-outward rank-based test statistics in Section 5.3.1 of Hallin et al. (2022a) are of the form

$$\left(\underset{\sim}{T}^{(n)}_{\mathbf{J}\pm}\right)^2 = \underset{\sim}{\boldsymbol{\Delta}}^{(n)\prime}_{\mathbf{J}\pm}\left(\boldsymbol{\Sigma}_{\underset{\sim}{\boldsymbol{\Delta}}_{\mathbf{J}\pm}}\right)^{-1}\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\mathbf{J}\pm} \qquad (2)$$

where $\boldsymbol{\Sigma}_{\underset{\sim}{\boldsymbol{\Delta}}_{\mathbf{J}}}$ is the exact or asymptotic covariance of

$$\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\mathbf{J}\pm} := \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{J}(\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)) - \frac{1}{n_2}\sum_{i=n_1+1}^{n}\mathbf{J}(\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)). \qquad (3)$$

Since the quadratic form (2) is invariant under affine transformations of $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\mathbf{J}\pm}$ and since the sum $\sum_{i=1}^{n}\mathbf{J}(\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i))$ is a deterministic constant that only depends on $\mathbf{J}$ and the grid used in the definition of $\mathbf{F}^{(n)}_{\pm}$, the same test statistic can be based on

$$\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\mathbf{J}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{J}(\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{J}(\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i)),$$

yielding the test statistic described in Section 5.3.1 of Hallin et al. (2022a), which, in the particular case of Wilcoxon and van der Waerden scores $\mathbf{J}_{\mathrm{W}}$ and $\mathbf{J}^{\pm}_{\mathrm{vdW}}$, we denote as $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, respectively.

For the same testing problem, Deb et al. (2021) consider statistics of the form (2) but also

(a) Based on the empirical Monge–Kantorovich vector ranks $\mathbf{F}^{(n)}_{\square}$ associated with the cubic uniform reference $\mathrm{U} = \mathrm{U}_{[0,1]^{\mathrm{d}}}$, statistics $\underset{\sim}{T}^{(n)}_{\mathbf{J}\square}$ and $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\mathbf{J}\square}$ of the same form as $\underset{\sim}{T}^{(n)}_{\mathbf{J}\pm}$ and $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\mathbf{J}\pm}$ in (2) and (3) but with $\mathbf{J}(\mathbf{F}^{(n)}_{\square}(\mathbf{Z}^{(n)}_i))$ instead of $\mathbf{J}(\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i))$; denote by $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$ the particular cases of the

[8] Asymptotically valid, here, means pointwise (with respect to the actual density of the observations) asymptotically correct nominal probability levels, not *uniformly* asymptotically correct nominal probability levels.

Wilcoxon and cubic van der Waerden statistics, obtained for the scores $\mathbf{J}_{\mathrm{W}}$ and $\mathbf{J}^{\Box}_{\mathrm{vdW}}$, respectively

(b) Based on the empirical Monge–Kantorovich vector ranks $\mathbf{F}^{(n)}_{\pm\mathcal{N}}$ and $\mathbf{F}^{(n)}_{\Box\mathcal{N}}$ associated with the spherical Gaussian reference $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ considered as spherical or as a product of independent uniforms, statistics $\underset{\sim}{T}^{(n)}_{\mathbf{J}\pm\mathcal{N}}$ and $\underset{\sim}{T}^{(n)}_{\mathbf{J}\Box\mathcal{N}}$ of the same form as $\underset{\sim}{T}^{(n)}_{\mathbf{J}\pm}$ and $\underset{\sim}{\mathbf{\Delta}}^{(n)}_{\mathbf{J}\pm}$ in (2) and (3) but with $\mathbf{J}(\mathbf{F}^{(n)}_{\pm\mathcal{N}}(\mathbf{Z}^{(n)}_i))$ and $\mathbf{J}(\mathbf{F}^{(n)}_{\Box\mathcal{N}}(\mathbf{Z}^{(n)}_i))$, respectively, substituting $\mathbf{J}(\mathbf{F}^{(n)}_{\pm}(\mathbf{Z}^{(n)}_i))$; this, for Wilcoxon scores $\mathbf{J}(\mathbf{u}) = \mathbf{u}$, yields the van der Waerden statistics $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\Box\mathcal{N}}\right)^2$.

Remark 4 Although they are based on Wilcoxon (identity) scores, the terminology "van der Waerden statistic" for $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}$ and $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\Box\mathcal{N}}$ seems more appropriate than the terminology "Wilcoxon statistic" used by Deb et al. (2021) and is in line with the traditional terminology of rank-based inference. Both $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}$ and $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}$ indeed result from a transport from the sample values to a grid of Gaussian quantiles of the form ($\mathfrak{G}$iii). For $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}$, the transport is $\mathbf{J} \circ \mathbf{F}^{(n)}_{\pm}$, which, as a rule, is not an optimal one (not the gradient of a convex function), while, for $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}$, the transport is the optimal one $\mathbf{F}^{(n)}_{\Box\mathcal{N}}$; the difference between $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}$ and $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}$ thus essentially consists in the way the transport to the spherical Gaussian grid is performed. A similar remark can be made for $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\Box}$ and $\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\Box\mathcal{N}}$.

3 Finite-Sample Performance: Two-Sample Location Simulations

It clearly appears that choices are to be made before performing a rank test based on the concepts of multivariate ranks developed in the previous sections: center-outward ranks? vector ranks? which ones? with which scores? The analysis of asymptotic performance does not help much, as the same local powers are achieved irrespective of such choices. The objective of this chapter is to determine whether finite-sample performance can help us with these choices. We restrict ourselves to the two-sample location problem, Wilcoxon and van der Waerden scores, but the conclusions are quite likely to hold for other score functions and in the general case of the multiple-output linear models considered in Hallin et al. (2022a).

Before explaining how simulations were conducted, let us provide some details on the way the grids described in Sect. 2.2 were constructed. Recall that the aim of these grids is to provide a discrete approximation of the chosen continuous reference distribution.

3.1 Halton Sequences on the Cube and the Sphere (($\mathfrak{G}$ii) and ($\mathfrak{G}$iv) Grids)

The grid constructions ($\mathfrak{G}$ii) and ($\mathfrak{G}$iv) involve Halton sequences on the hypercube $[0, 1]^d$. Halton sequences are pseudo-random numbers with low discrepancy, which are routinely used in methods such as Monte Carlo simulations. We used the implementation available in the package SDraw by McDonald and McDonald (2020). The grid construction in ($\mathfrak{G}$i), hence also in ($\mathfrak{G}$iii), requires an n_S-point "Halton sequence" over the hypersphere $\mathcal{S}^{d-1}$. To obtain such a grid, we first generate an n_S-point Halton sequence over $[0, 1]^{d-1}$ and then componentwise perform the standard normal quantile transformation $u_j \mapsto z_j := \Phi^{-1}(u_j)$. This yields an n_S-tuple of points $\mathbf{z}_1, \ldots, \mathbf{z}_{n_S}$, with

$$\mathbf{z}_j := (\Phi^{-1}(u_{j1}), \ldots, \Phi^{-1}(u_{jd})).$$

The resulting unit vectors $\mathbf{z}_j/\|\mathbf{z}_j\|$, $j = 1, \ldots, n_S$, constitute the desired sequence over $\mathcal{S}^{d-1}$.[9]

3.2 Factorization of *n* (($\mathfrak{G}$i) and ($\mathfrak{G}$iii) Grids)

As for the grid constructions ($\mathfrak{G}$i) and ($\mathfrak{G}$iii), they require a factorization of n into $n_R n_S + n_0$ with $n_0 < \min(n_R, n_S)$. Intuition suggests choosing n_R and n_S of order $n^{1/d}$ and $n^{(d-1)/d}$, respectively. This, however, is of little help for finite n. Since the grid is supposed to provide an approximation of the spherical uniform, we rather proceed by minimizing the Wasserstein distance to the spherical uniform as proposed in Mordant (2021). More precisely, considering the grid with n_R radial points described in ($\mathfrak{G}$i), denote by $\mathrm{G}^{(n)}_{n_R}$ the discrete measure placing a probability mass $1/n$ on each of the n gridpoints except for the origin which receives probability mass n_0/n. As suggested in Mordant (2021), we select the grid with n_R^* radial points, where

$$n_R^* := \underset{1 \le n_R \le n}{\arg\min}\; W_2(\mathrm{G}^{(n)}_{n_R}, \mathrm{U}_d) \tag{4}$$

(W_2, as usual, stands for the Wasserstein distance of order two). For $d \geq 3$, that distance $W_2(\mathrm{G}^{(n)}_{n_R}, \mathrm{U}_d)$ does not only depend on n_R (hence on n_S) but also on the n_S points chosen (as explained in Sect. 3.1) on the hypersphere $\mathcal{S}^{d-1}$. The minimization

[9] The justification is the fact that if the distribution of $\mathbf{Z}$ is a product of independent univariate standard normal marginals, then $\mathbf{Z}$ is spherical Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and hence $\mathbf{Z}/\|\mathbf{Z}\|$ is uniform over $\mathcal{S}^{d-1}$.

Table 1 Optimal (in the sense of (4)) values of n_R, n_S, and n_0 as functions of the sample size n, the dimension d, and the reference distributions (($\mathfrak{G}$i) or ($\mathfrak{G}$iii) grids)

Reference distribution	d	$n = 50$	$n = 100$	$n = 200$	$n = 300$	$n = 400$
U_d, ($\mathfrak{G}$i) grid	2	$n_R = 4$	$n_R = 6$	$n_R = 9$	$n_R = 11$	$n_R = 12$
		$n_S = 12$	$n_S = 16$	$n_S = 22$	$n_S = 27$	$n_S = 33$
		$n_0 = 2$	$n_0 = 4$	$n_0 = 2$	$n_0 = 3$	$n_0 = 4$
$\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, ($\mathfrak{G}$iii) grid	2	$n_R = 4$	$n_R = 7$	$n_R = 11$	$n_R = 14$	$n_R = 18$
		$n_S = 12$	$n_S = 14$	$n_S = 18$	$n_S = 21$	$n_S = 22$
		$n_0 = 2$	$n_0 = 2$	$n_0 = 2$	$n_0 = 6$	$n_0 = 4$
$n^{1/d}$		$n^{1/2} = 7.071$	$n^{1/2} = 10$	$n^{1/2} = 14.142$	$n^{1/2} = 17.321$	$n^{1/2} = 20$
U_d, ($\mathfrak{G}$i) grid	5	$n_R = 2$	$n_R = 2$	$n_R = 2$	$n_R = 3$	$n_R = 3$
		$n_S = 25$	$n_S = 50$	$n_S = 100$	$n_S = 100$	$n_S = 133$
		$n_0 = 0$	$n_0 = 0$	$n_0 = 0$	$n_0 = 0$	$n_0 = 1$
$\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, ($\mathfrak{G}$iii) grid	5	$n_R = 1$	$n_R = 1$	$n_R = 1$	$n_R = 2$	$n_R = 2$
		$n_S = 50$	$n_S = 100$	$n_S = 200$	$n_S = 150$	$n_S = 200$
		$n_0 = 0$	$n_0 = 0$	$n_0 = 0$	$n_0 = 0$	$n_0 = 0$
$n^{1/d}$		$n^{1/5} = 2.187$	$n^{1/5} = 2.512$	$n^{1/5} = 2.885$	$n^{1/5} = 3.129$	$n^{1/5} = 3.314$

in (4) is feasible, as n, n_R, n_S, and n_0 all are integers. A similar strategy is adopted for the construction of the spherical Gaussian grids ($\mathfrak{G}$iii).

Table 1 provides, for dimensions $d = 2$ and $d = 5$, various sample sizes, and various reference distributions, the spherical uniform (($\mathfrak{G}$i) grids) and the spherical Gaussian (($\mathfrak{G}$iii) grids), the "optimal values" obtained via (4) for n_R, n_S, and n_0. These values are in line with the intuition that the "optimal" n_R behaves like $n^{1/d}$, while the role of distances to the center rapidly decreases as the dimension d increases.

3.3 *Simulations*

Based on the grids obtained along the lines described in Sects. 3.1 and 3.2, the distribution-free critical values of the various rank tests under study were computed from 40,000 replications. Throughout, we chose $n_1 = n_2 = n/2$. The optimal maps between the sample and the grids were obtained via an exact solver relying on the so-called Hungarian method as implemented in the R-package clue by Hornik (2022). We now turn to the empirical evaluation of the performance of the various rank-based Wilcoxon and van der Waerden tests for the two-sample location problem.

The objective of our simulations is, essentially, obtaining empirical answers to the following two questions:

(a) should we use spherical grids (($\mathfrak{G}$i) or ($\mathfrak{G}$iii)) or cubic (($\mathfrak{G}$ii) or ($\mathfrak{G}$iv)) ones?
(b) should we, in line with the Hájek tradition, privilege transports to the uniform combined with scores (as in $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$), or, as recommended by Deb et al. (2021), should we rather consider transports to the "scored distribution," that is, choose as reference distribution the push-forward of the uniform by the score (as in $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$)?

4 Wilcoxon-Type Tests

The Wilcoxon tests are based on the identity score function $\mathbf{J}(\mathbf{u}) = \mathbf{u}$ and uniform (either spherical or cubic) reference distributions, yielding (see Sect. 2.3.2) the test statistics $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$.

4.1 The Bivariate Case

In this section, we evaluate the performance of the bivariate Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$ for samples of size $n_1 = n_2 = n/2$ with $n = 100$, 200, and 400. The first sample is drawn from a centered distribution, the second one from a shifted version with shift $(\eta, \eta)'$, $\eta > 0$ of the same. The number of replications is $N = 500$.

4.1.1 Spherical Gaussian Samples

The first sample is drawn from $\mathcal{N}((0, 0)', \mathbf{I}_2)$ and the second one from $\mathcal{N}((\eta, \eta)^\top, \mathbf{I}_2)$ with $\eta > 0$. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 1. All three tests display, essentially, the same performance: although Wilcoxon, in principle, is strictly less powerful than Hotelling (which in this case is finite-sample optimal), no significant loss of efficiency is detected.

4.1.2 Nonspherical Gaussian Samples

The first sample is drawn from an $\mathcal{N}((0, 0)', \mathbf{\Sigma})$ distribution, and the second sample is drawn from an $\mathcal{N}((\eta, \eta)', \mathbf{\Sigma})$ one; $\mathrm{vech}(\mathbf{\Sigma}) = (1, 0.8, 1)'$. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 2. The results are

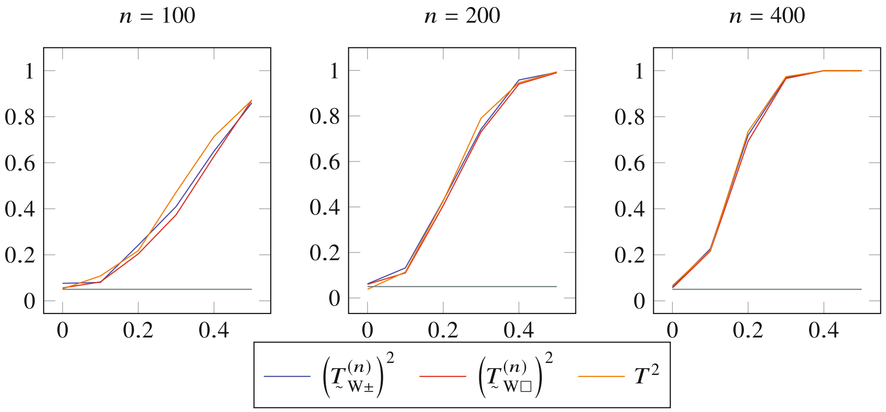


Fig. 1 Rejection frequencies, for bivariate spherical Gaussian samples and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}{}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}{}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

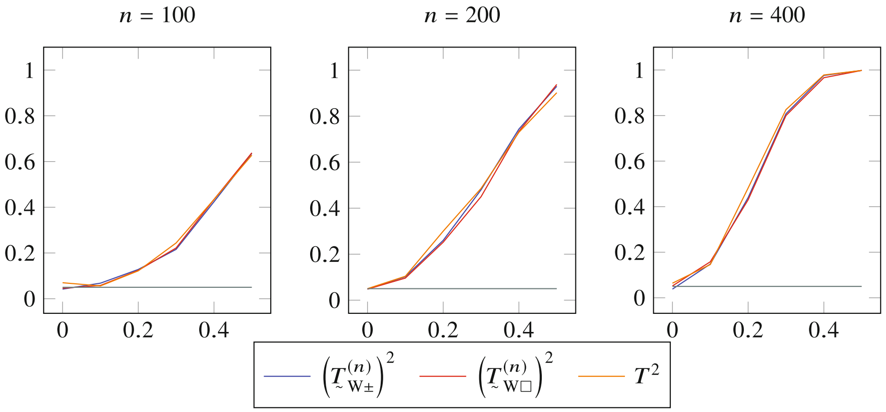


Fig. 2 Rejection frequencies, for bivariate samples of nonspherical Gaussian distributions and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}{}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}{}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

essentially the same as in the spherical case (Sect. 4.1.1). Note the loss of power in the three tests under study, due to the non-specification of the population covariance matrix; that loss, however, is uniform over the three tests.

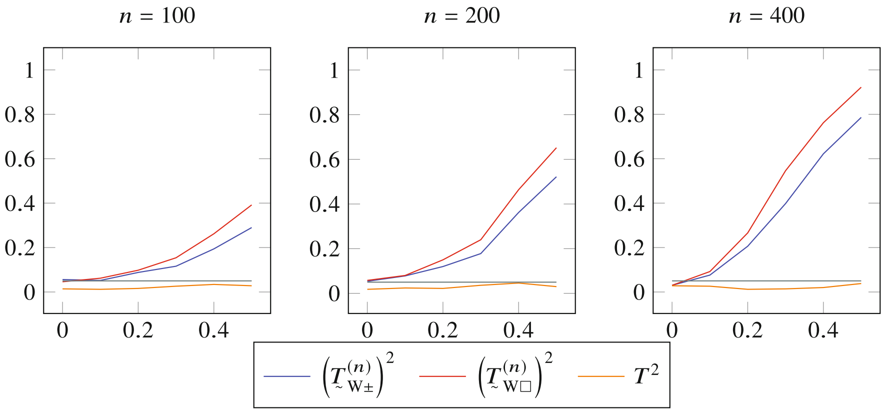


Fig. 3 Rejection frequencies, for bivariate samples with independent Cauchy marginals and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

4.1.3 Samples with Independent Cauchy Marginals

The first sample is drawn from a product of two independent Cauchy and the second one from the shifted version of the same distribution. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 3. With a rejection probability uniformly less than the nominal 5% level, Hotelling, as expected, performs miserably. In this independent component situation, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$ does outperform $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$.

4.1.4 Spherical Cauchy Samples

The first sample is drawn from a centered spherical student with one degree of freedom $t_1((0, 0)', \mathbf{I}_2)$ (spherical Cauchy) and the second one from the shifted version $t_1((\eta, \eta)', \mathbf{I}_2)$ of the same distribution. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 4. The performance of Hotelling, again, is a disaster; although the actual distribution is spherical, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$ still outperforms $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$.

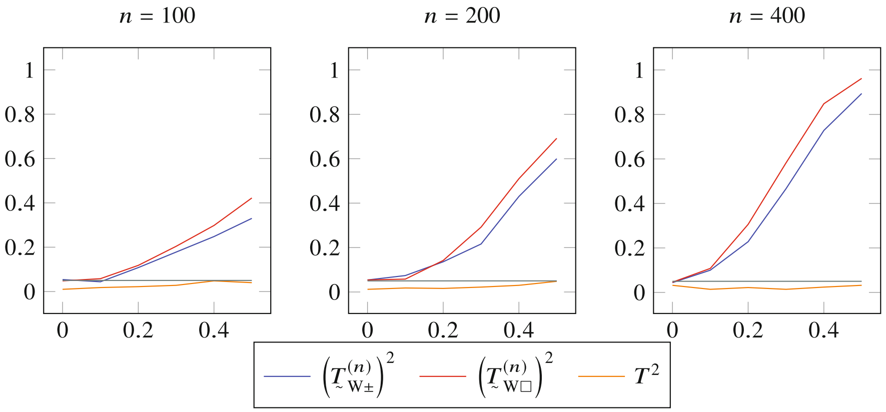


Fig. 4 Rejection frequencies, for bivariate samples of spherical Cauchy distributions (Sect. 4.1.4) and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

4.1.5 "Banana-Shaped" Samples

The first sample is drawn from a centered "banana-shaped" mixture

$$0.3\,\mathcal{N}_2\left(\begin{pmatrix}0\\-0.7\end{pmatrix},\begin{pmatrix}0.35^2 & 0\\ 0 & 0.35^2\end{pmatrix}\right)+0.35\,\mathcal{N}_2\left(\begin{pmatrix}-0.9\\0.3\end{pmatrix},\begin{pmatrix}0.358 & -0.55\\-0.55 & 1.02\end{pmatrix}\right)$$
$$+\,0.35\,\mathcal{N}_2\left(\begin{pmatrix}0.9\\0.3\end{pmatrix},\begin{pmatrix}0.358 & 0.55\\0.55 & 1.02\end{pmatrix}\right)$$

of three Gaussian components. The second sample is drawn from a shifted version (shift $(\eta,\eta)'$, $\eta > 0$) of the same mixture. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 5. The conclusions are the same as in the previous case, except that the (slight) advantage now belongs to $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$, despite the fact that the actual distribution is highly nonspherical.

4.2 *Wilcoxon-Type Statistics in Dimension d = 5*

We essentially adopted the same simulation settings as before, with $n_1 = n_2 = n/2$. A sample size of $n = 100$ in dimension $d = 5$ is very small, though, and we considered sample sizes $n = 200$, 400, and 800.

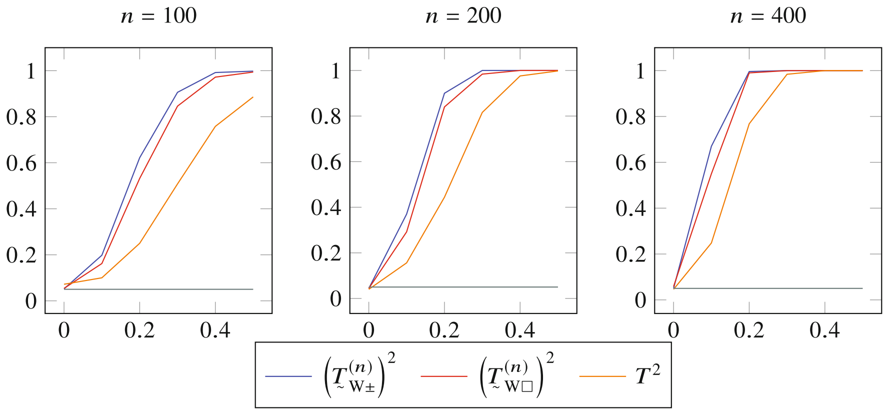


Fig. 5 Rejection frequencies, for bivariate "banana-shaped" samples and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

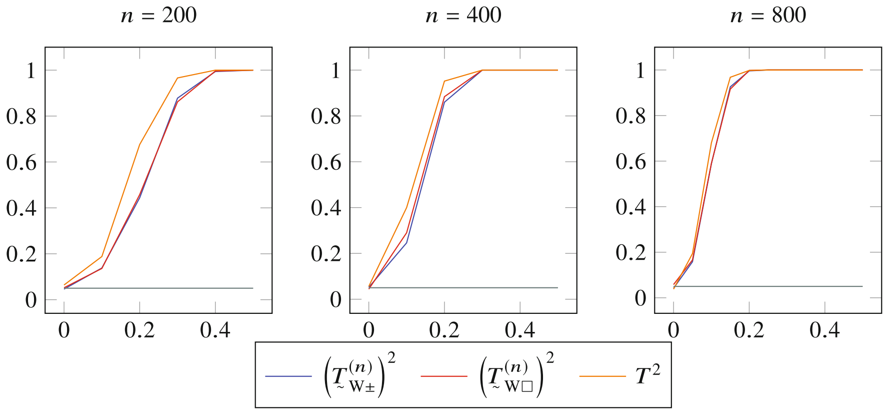


Fig. 6 Rejection frequencies, for samples with five-dimensional spherical Gaussian distributions (see Sect. 4.1.1) and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

4.2.1 Spherical Gaussian Samples

Here, the first sample is drawn from the $\mathcal{N}(\mathbf{0}, \mathbf{I}_5)$ distribution and the second one from the $\mathcal{N}(\eta\mathbf{1}, \mathbf{I}_5)$ distribution, where $\mathbf{1}$ denotes a 5-variate vector of ones. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 6. In the "small sample" case ($n = 200$), the optimality of Hotelling over Wilcoxon is perceptible (more so than in dimension $d = 2$); this superiority, however, fades away with growing n: again, under Gaussian assumptions, abandoning the parametrically

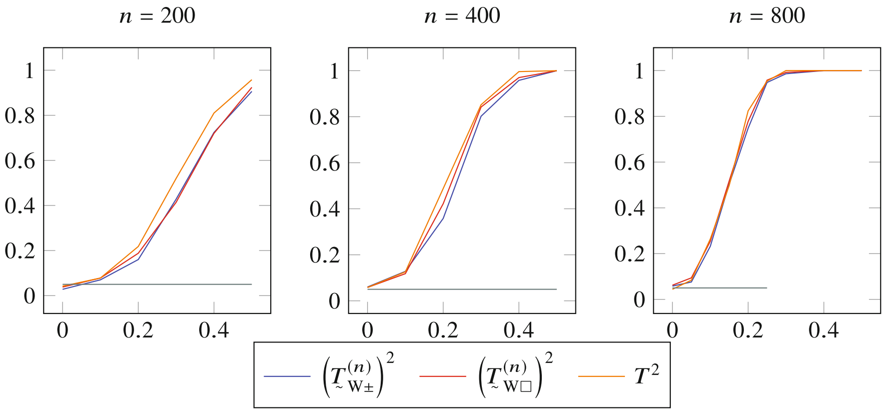


Fig. 7 Rejection frequencies, for nonspherical five-dimensional Gaussian distributions and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

optimal Hotelling test in favor of the rank-based Wilcoxon one has no visible cost in terms of power.

4.2.2 Nonspherical Gaussian Samples

The first sample is drawn from the $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ distribution and the second one from the $\mathcal{N}(\eta\mathbf{1}, \boldsymbol{\Sigma})$ distribution, where $\mathbf{1}$ denotes a 5-variate vector of ones and $\boldsymbol{\Sigma}$ is a correlation matrix with all off-diagonal entries equal to 0.5. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 7. Here again, the slight advantage of Hotelling over Wilcoxon very rapidly fades away with growing n, and the three tests yield very similar performances; in particular, no significant difference can be detected between $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$.

4.2.3 Samples with Independent Cauchy Marginals

The first sample is drawn from a product of five independent Cauchy distributions and the second one from the shifted version of the same. Rejection frequencies over 500 replications are shown (as functions of η) in Fig. 8. The performance of Hotelling, as in dimension $d = 2$, is terrible. The advantage (which is in line with the independent component nature of the distribution) of $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$ over $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ is even more significant than in dimension $d = 2$.

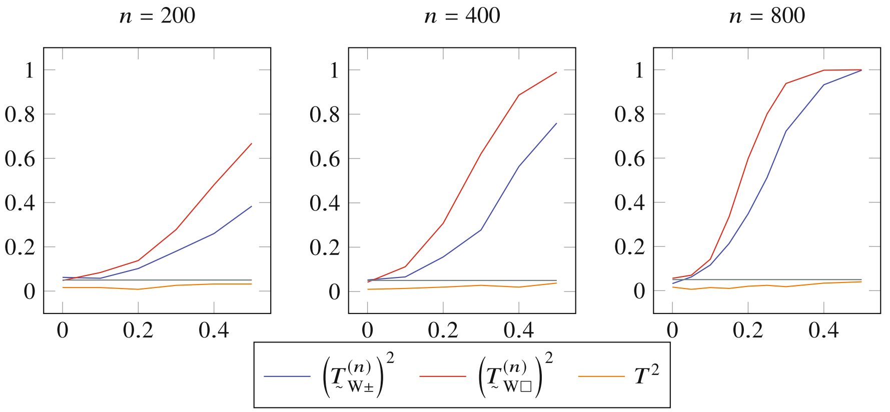


Fig. 8 Rejection frequencies, for five-dimensional distributions with independent Cauchy marginals and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

4.3 Wilcoxon-Type Statistics in Dimension d = 30

We essentially adopted the same simulation settings as before, with $n_1 = n_2 = n/2$ and sample sizes $n = 200$, 400, and 800.

4.3.1 Spherical Gaussian Samples

Here, the first sample is drawn from the $\mathcal{N}(\mathbf{0}, \mathbf{I}_{30})$ distribution, the second one from the $\mathcal{N}(\eta\mathbf{1}, \mathbf{I}_{30})$ distribution. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 9. The strict optimality of Hotelling, which was not perceptible for $d = 2$ and hardly so for $d = 5$, is now quite visible; as for the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, they achieve essentially the same performance.

4.3.2 Nonspherical Gaussian Samples

The first sample is drawn from the $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ distribution, the second one from the $\mathcal{N}(\eta\mathbf{1}, \mathbf{\Sigma})$ distribution, where $\mathbf{1}$ denotes a 30-variate vector of ones and $\mathbf{\Sigma}$ is a correlation matrix with all off-diagonal entries equal to 0.5. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 10. Powers are significantly less than in the spherical case: estimating a 30 × 30 covariance matrix is costly—more costly, under sample sizes 200 and 400, for Hotelling than for Wilcoxon. The two versions of Wilcoxon yield similar results.

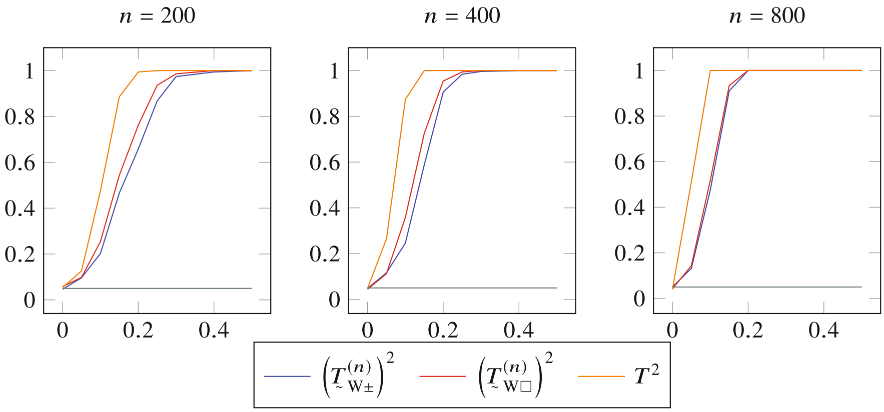


Fig. 9 Rejection frequencies, for samples with 30-dimensional spherical Gaussian distributions and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

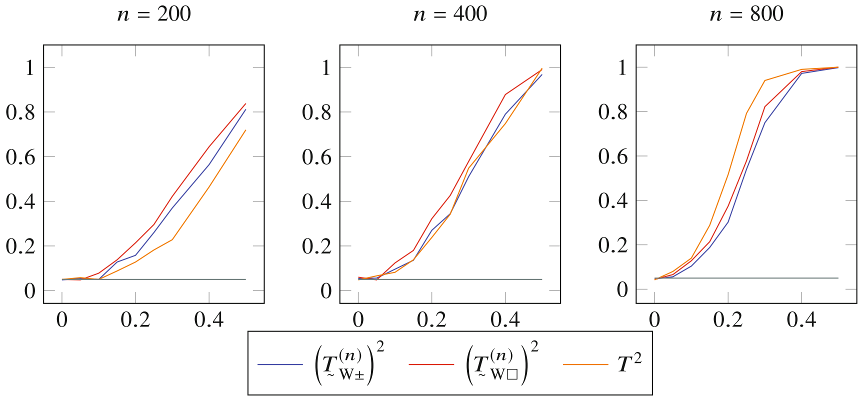


Fig. 10 Rejection frequencies, for nonspherical 30-dimensional Gaussian distributions and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

4.3.3 Samples with Independent Cauchy Marginals

The first sample is drawn from a product of thirty independent Cauchy distributions and the second one from the shifted version of the same. Rejection frequencies over 500 replications are shown (as functions of η) in Fig. 11. As expected, Hotelling fails miserably; here again, in line with the independent component nature of the distribution, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$ outperforms $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$.

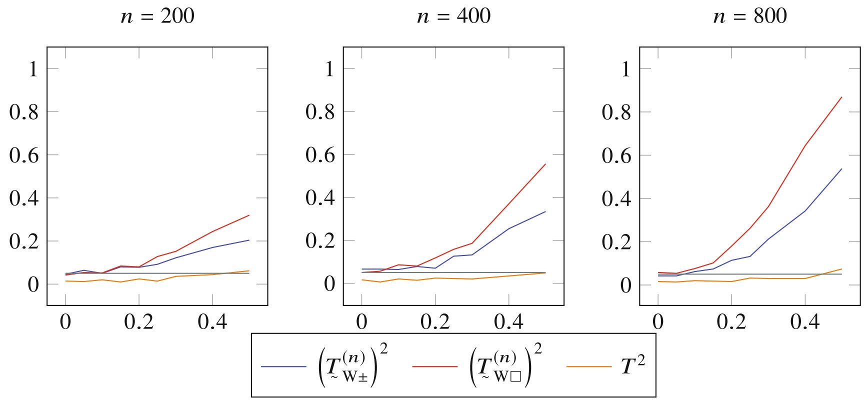


Fig. 11 Rejection frequencies, for 30-dimensional distributions with independent Cauchy marginals and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{W\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{W\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

4.4 Wilcoxon-Type Statistics in Dimension d = **100**

We essentially adopted the same simulation settings as before, with $n_1 = n_2 = n/2$ and $n = 200$, 400, and 800, which, in dimension 100, are quite small sample sizes.

4.4.1 Spherical Gaussian Samples

The first sample is drawn from the $\mathcal{N}(\mathbf{0}, \mathbf{I}_{100})$ distribution and the second one from the $\mathcal{N}(\eta\mathbf{1}, \mathbf{I}_{100})$ distribution, where $\mathbf{1}$ denotes a 100-dimensional vector of ones. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 12. Hotelling is losing its advantage over Wilcoxon for $n = 200$; as n grows, the three tests are essentially equivalent. Again, under Gaussian assumptions, abandoning the parametrically optimal Hotelling test in favor of the rank-based Wilcoxon one has no visible cost in terms of power.

4.4.2 Nonspherical Gaussian Samples

The first sample is drawn from the $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ distribution and the second one from the $\mathcal{N}(\eta\mathbf{1}, \mathbf{\Sigma})$ distribution, where $\mathbf{1}$ denotes a 5-variate vector of ones and $\mathbf{\Sigma}$ is a correlation matrix with all off-diagonal entries equal to 0.5. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 13. Hotelling, here, is dramatically lagging behind the two versions of Wilcoxon; the latter, in terms of power, are quite equivalent and obviously less sensitive to the problematic estimation of a 100 × 100 covariance matrix than their parametric counterpart.

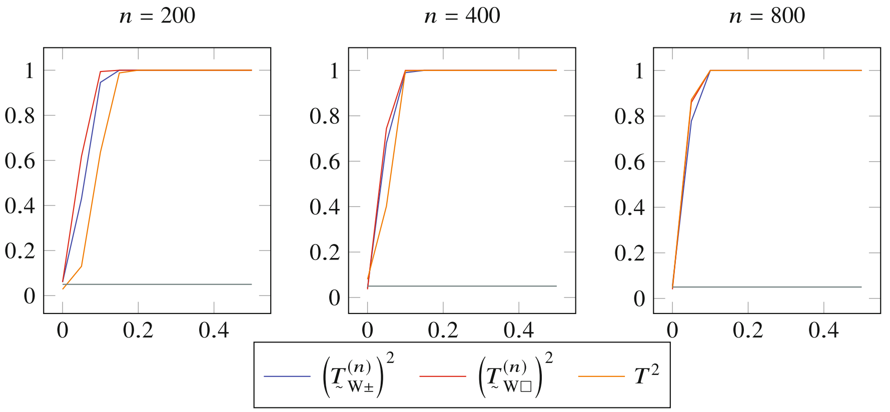


Fig. 12 Rejection frequencies, for samples with 100-dimensional spherical Gaussian distributions (see 4.2.1) and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

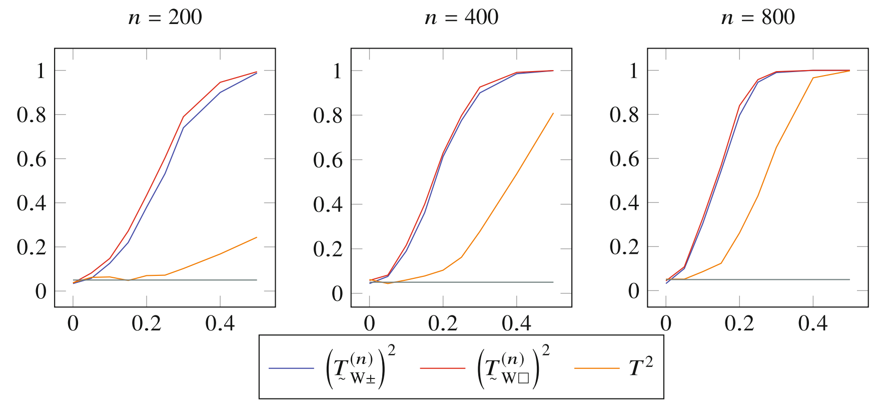


Fig. 13 Rejection frequencies, for nonspherical 100-dimensional Gaussian distributions and various sample sizes, of the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

4.4.3 Samples with Independent Cauchy Marginals

The first sample is drawn from a product of $d = 100$ independent Cauchy distributions and the second one from the shifted version of the same. Rejection frequencies over 500 replications are shown (as functions of η) in Fig. 14. The results are surprisingly comparable with those obtained in Fig. 11 for dimension $d = 30$, and the conclusions are the same.

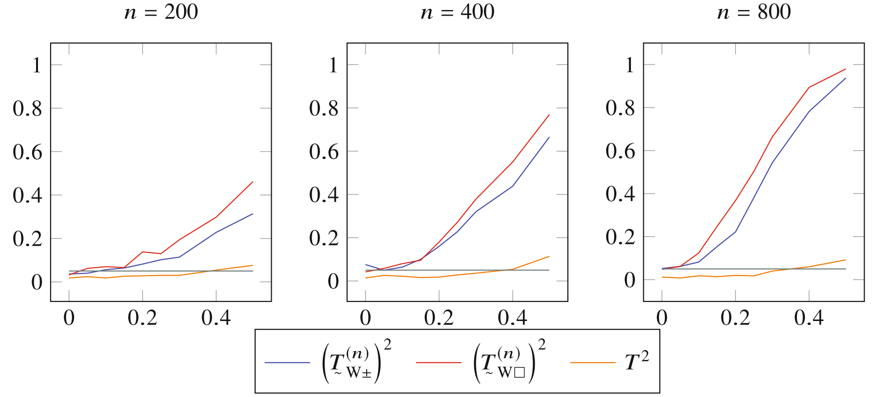


Fig. 14 Rejection frequencies, for 100-dimensional distributions with independent Cauchy marginals and various sample sizes, of Hotelling's test based on T^2 and the Wilcoxon tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{W}\square}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

5 van der Waerden-Type Tests

In this section, we are considering four distinct tests of the van der Waerden type, based (see Sect. 2.3.2) on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$ (spherical uniform reference density, ($\mathfrak{G}$i) grid), $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$ (cubic uniform reference density, ($\mathfrak{G}$ii) grid), $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$ (spherical Gaussian reference density, spherical grid ($\mathfrak{G}$iii)), and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$ (spherical Gaussian reference density, cubic grid ($\mathfrak{G}$iv)).

5.1 *Bivariate Case*

5.1.1 Spherical Gaussian Samples

The same Gaussian samples as in Sect. 4.1.1 are used. Figure 15 shows the rejection frequencies over $N = 500$ replications of Hotelling and the various rank-based tests: as expected, performing rank-based van der Waerden tests instead of Hotelling ones does not imply any loss of efficiency in the Gaussian case.

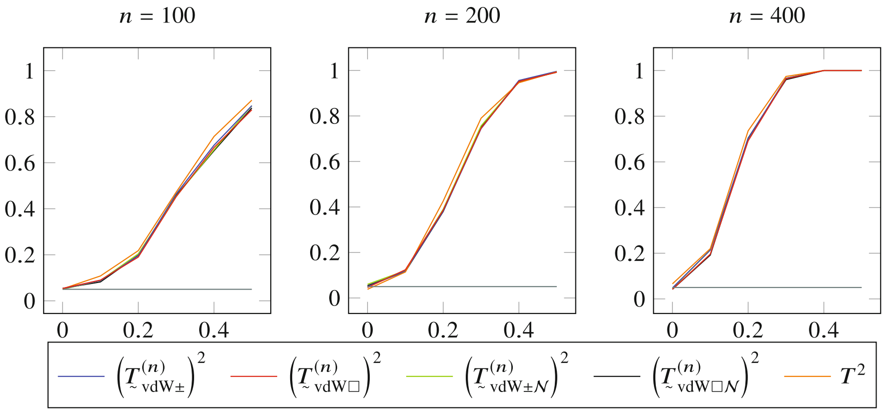


Fig. 15 Rejection frequencies, for bivariate spherical Gaussian samples (see 5.1.1) and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

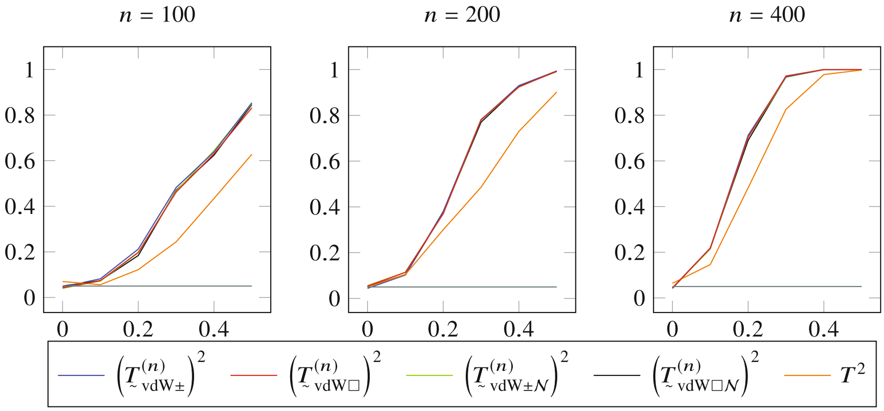


Fig. 16 Rejection frequencies, for bivariate nonspherical Gaussian samples and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

5.1.2 Nonspherical Gaussian Samples

The same correlated Gaussian samples as in Sect. 4.1.2 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 16. The non-specification of the covariance matrix apparently has no impact on the comparative performance of Hotelling and its rank-based van der Waerden competitors, which all coincide.

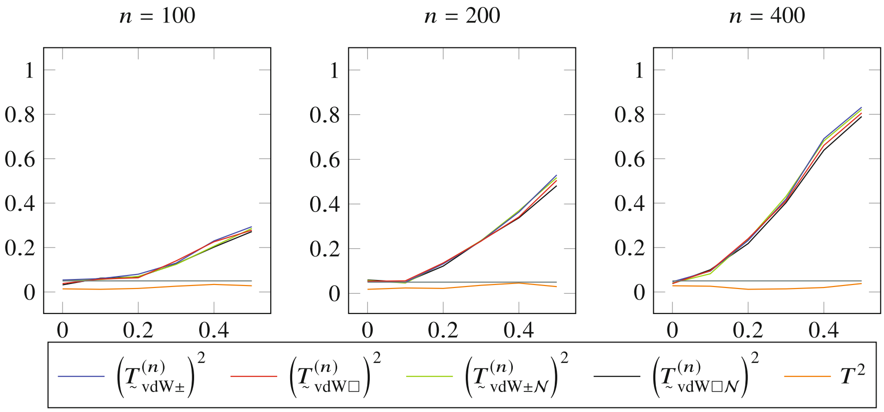


Fig. 17 Rejection frequencies, for bivariate samples of independent Cauchy marginals and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

5.1.3 Samples with Independent Cauchy Marginals

The same samples with independent Cauchy marginals as in Sect. 4.1.3 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 17. Again, all powers are much less than in the Gaussian case, but Hotelling is totally inefficient.

5.1.4 Spherical Cauchy Samples

The same spherical Cauchy samples as in Sect. 4.1.4 are used. Rejection frequencies over 500 replications are shown (as functions of η) in Fig. 18. All tests perform similarly except, of course, for Hotelling, which fails completely.

5.1.5 "Banana-Shaped" Samples

The same "banana-shaped" mixtures as in Sect. 4.1.5 are used. Rejection frequencies over 500 replications are shown (as functions of η) in Fig. 19. The empirical power curves of the four van der Waerden tests are essentially indistinguishable, while significantly outperforming the Hotelling ones.

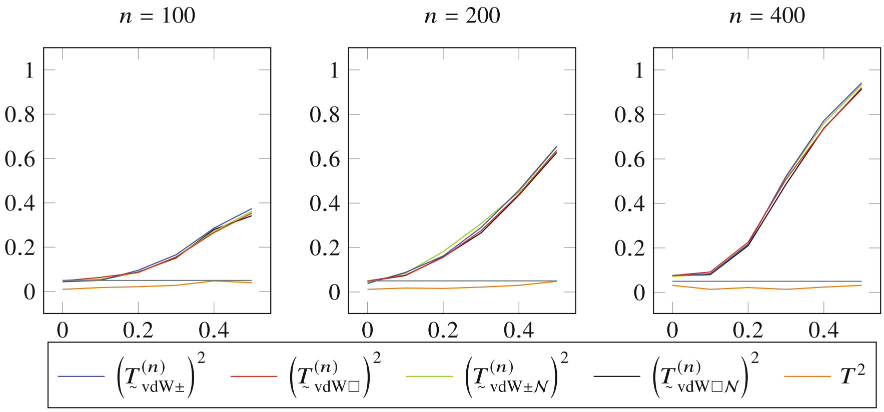


Fig. 18 Rejection frequencies, for bivariate spherical Cauchy samples and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications.

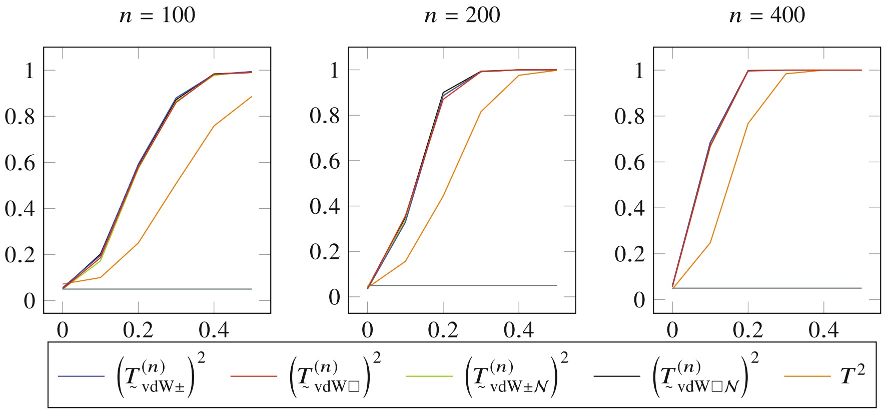


Fig. 19 Rejection frequencies, for bivariate "banana-shaped" samples and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

5.2 *van der Waerden-Type Statistics in Dimension d = 5*

5.2.1 Spherical Gaussian Samples

The same spherical Gaussian samples as in Sect. 4.4.1 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 20. A "small sample" superiority of Hotelling for $n = 200$ rapidly disappears as n increases; all van der Waerden tests yield the same performance.

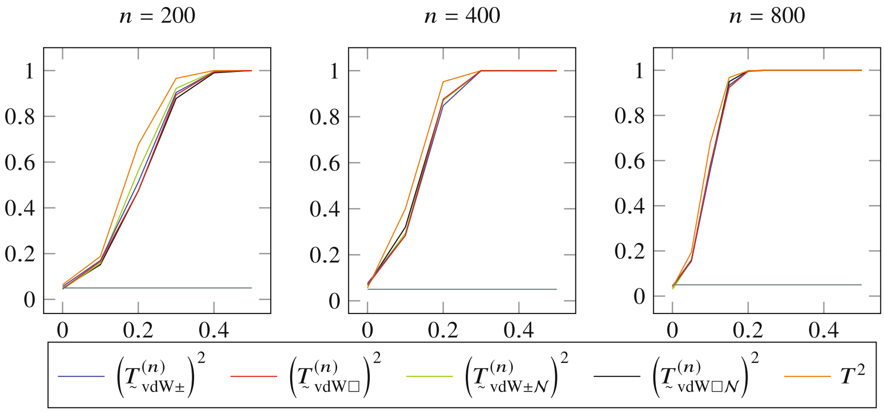


Fig. 20 Rejection frequencies, for samples with five-dimensional spherical Gaussian distributions (see Sect. 5.4.1) and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

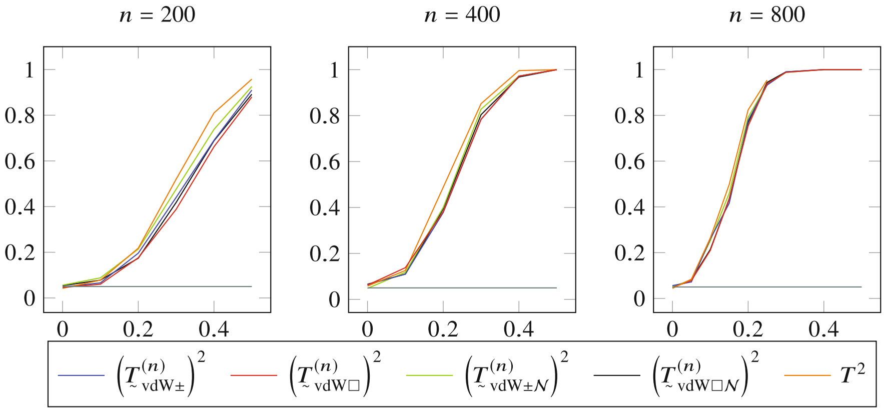


Fig. 21 Rejection frequencies, for nonspherical five-dimensional Gaussian samples (see Sect. 5.4.2) and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

5.2.2 Nonspherical Gaussian Samples

The same spherical Gaussian samples as in Sect. 4.4.2 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 21. The slight advantage of Hotelling for $n = 200$ under spherical Gaussian is fading away. All van der Waerden tests yield similar performance.

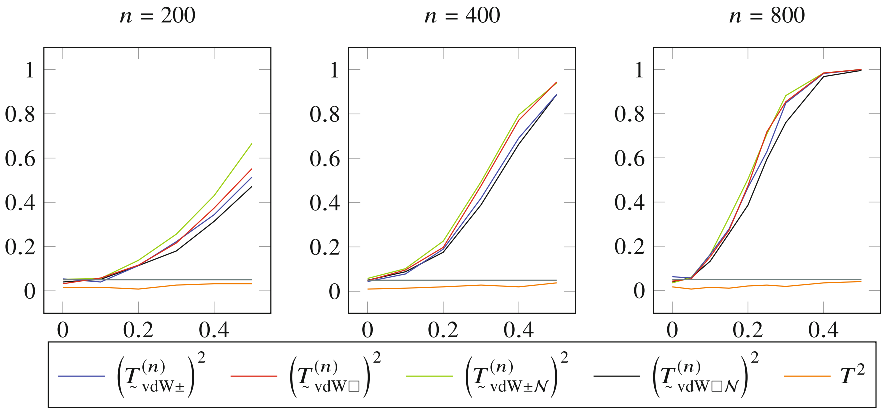


Fig. 22 Rejection frequencies, for five-dimensional samples with independent Cauchy marginals, (see Sect. 5.4.3) and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

5.2.3 Samples with Independent Cauchy Marginals

The same Cauchy samples as in Sect. 4.4.3 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 22. The conclusions drawn for $d = 2$ still hold, with a very slight superiority of the "direct transportation" test $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$ over the Gaussian score ones $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$ and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$.

5.3 van der Waerden-Type Statistics in Dimension $d = 30$

5.3.1 Spherical Gaussian Samples

The same spherical Gaussian samples as in Sect. 4.4.1 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 23. Hotelling, which is optimal, outperforms the four van der Waerden tests, which all yield the same performance. The dimension increase thus slows down the local convergence of van der Waerden to Hotelling.

5.3.2 Nonspherical Gaussian Samples

The same spherical Gaussian samples as in Sect. 4.4.2 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 24.

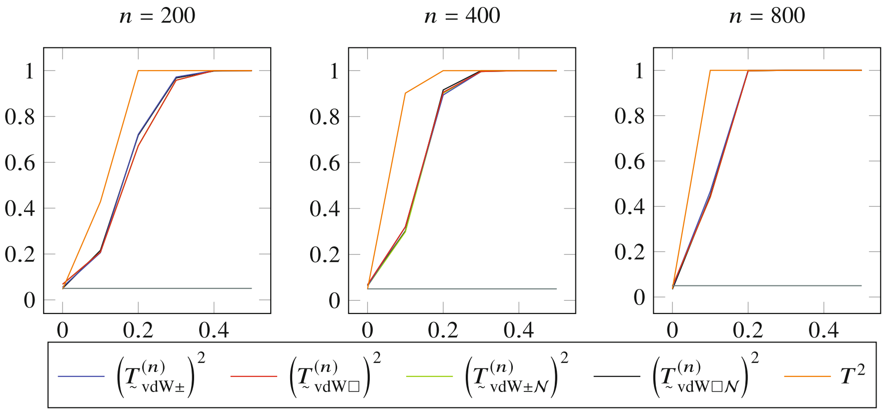


Fig. 23 Rejection frequencies, for samples with 30-dimensional spherical Gaussian distributions (see Sect. 5.4.1) and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

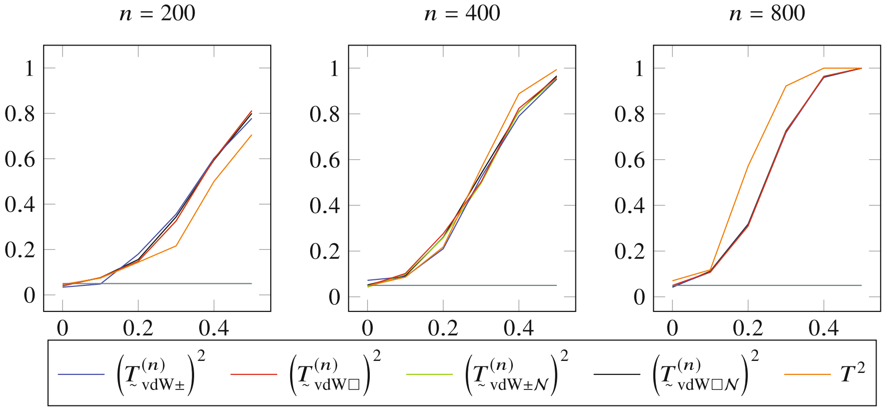


Fig. 24 Rejection frequencies, for nonspherical 30-dimensional Gaussian samples and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

The advantage of Hotelling under spherical Gaussian is hampered by the need to estimate a 30 × 30 covariance matrix and only reappears as n increases to 800; still no differences among the four van der Waerden tests.

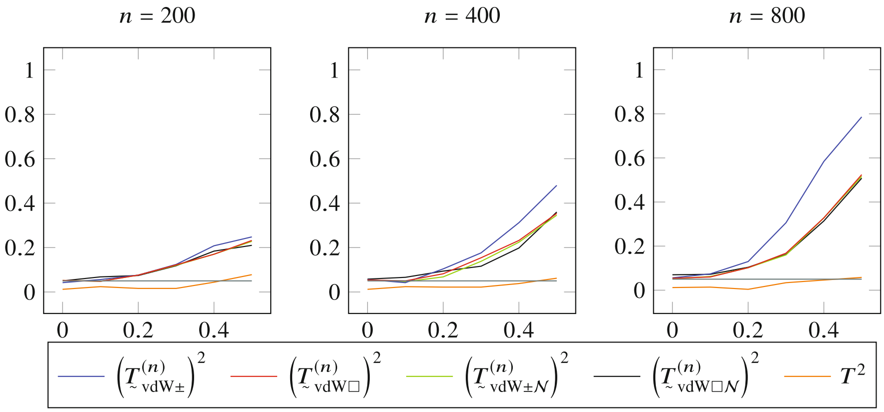


Fig. 25 Rejection frequencies, for 30-dimensional samples with independent Cauchy marginals, and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

5.3.3 Samples with Independent Cauchy Marginals

The same Cauchy samples as in Sect. 4.4.3 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 25. Conclusions are the same as in dimension $d = 5$.

5.4 *van der Waerden-Type Statistics in Dimension d* = 100

5.4.1 Spherical Gaussian Samples

The same spherical Gaussian samples as in Sect. 4.4.1 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 26. Hotelling is outperformed for "small" $n = 200$; for $n = 400$ and 800, all tests yield the same performance.

5.4.2 Nonspherical Gaussian Samples

The same nonspherical Gaussian samples as in Sect. 4.4.2 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 27. The cost of estimating a 100×100 covariance matrix has (even for $n = 800$) a quite significant impact on Hotelling but a much milder one on the four van der Waerden test which, furthermore, yield very similar performances.

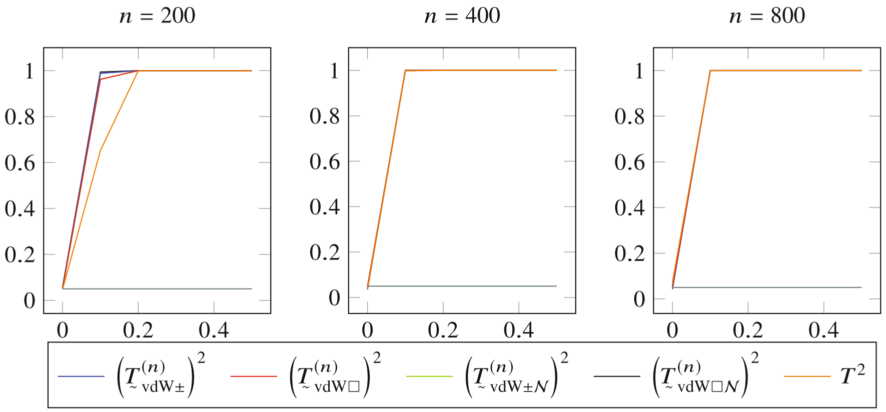


Fig. 26 Rejection frequencies, for samples with 100-dimensional spherical Gaussian distributions and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

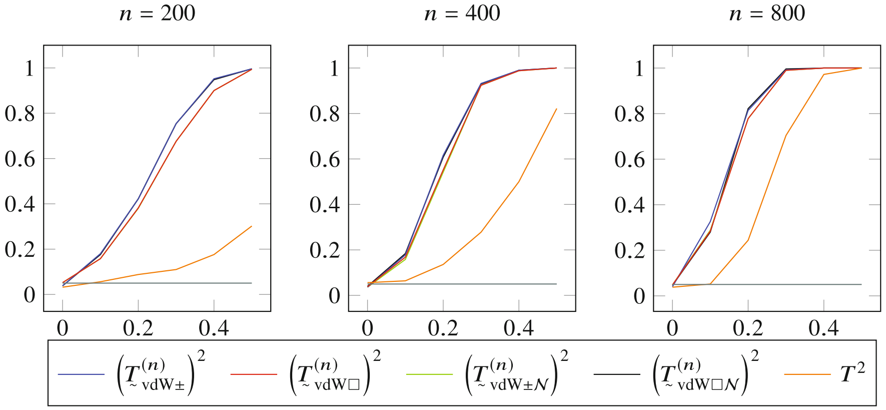


Fig. 27 Rejection frequencies, for nonspherical 100-dimensional Gaussian samples and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

5.4.3 Samples with Independent Cauchy Marginals

The same Cauchy samples as in Sect. 4.4.3 are used. Rejection frequencies over $N = 500$ replications are shown (as functions of η) in Fig. 28. Conclusions are the same as for $d = 30$; a slight advantage for the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\mathrm{vdW}\pm}\right)^2$.

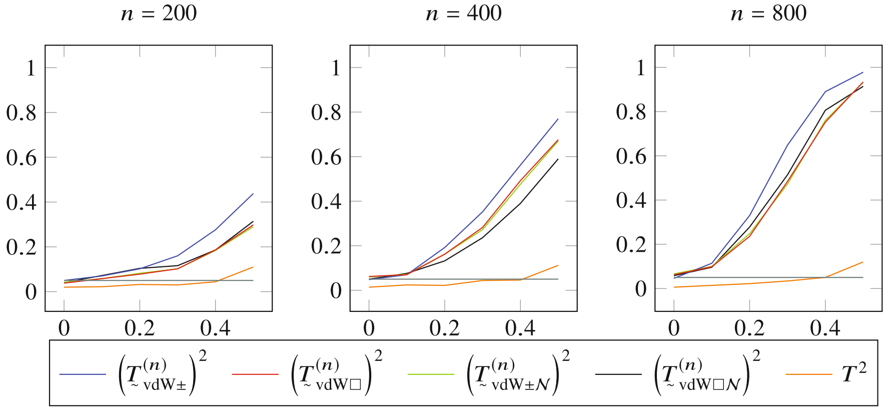


Fig. 28 Rejection frequencies, for 100-dimensional samples with independent Cauchy marginals and various sample sizes, of Hotelling's test based on T^2 and the van der Waerden tests based on $\left(\underset{\sim}{T}^{(n)}_{\text{vdW}\pm}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\text{vdW}\square}\right)^2$, $\left(\underset{\sim}{T}^{(n)}_{\text{vdW}\pm\mathcal{N}}\right)^2$, and $\left(\underset{\sim}{T}^{(n)}_{\text{vdW}\square\mathcal{N}}\right)^2$, respectively, as functions of the shift η; $N = 500$ replications

6 Conclusions

While confirming the advantages and excellent performance of rank tests over their daily practice pseudo-Gaussian counterparts, the simulations of the previous sections provide empirical answers to several questions of practical importance.

The choice of the grid (whether spherical ($\mathfrak{G}$i), cubic ($\mathfrak{G}$ii), or Gaussian (($\mathfrak{G}$iii) and ($\mathfrak{G}$iv)) seems to have relatively little impact on the performance of the corresponding Wilcoxon tests and no impact at all on the performance of van der Waerden tests. In particular, there is no evidence that Wilcoxon tests based on spherical grids ($\mathfrak{G}$i) are preferable under spherical distributions, while Wilcoxon tests based on cubic grids ($\mathfrak{G}$ii) are preferable under distributions with independent components: see, e.g., the Cauchy case (Sects. 4.1.4 and 4.1.3).

Similarly, little difference is detected among the four versions of van der Waerden test statistics, whether based on score functions or "direct transportation."

Supplemental Material Additional material presenting visualizations of different grids and the contours they induce as well as the codes used for the simulations are available on https://sites.google.com/view/gillesmordant.

References

Carlier, G., Chernozhukov, V., & Galichon, A. (2016). Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, *44*, 1165–92. https://doi.org/10.1214/15-AOS1401.

Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, *45*, 223–256. https://doi.org/10.1214/16-AOS1450.

Deb, N., Bhattacharya, B. B., & Sen, B. (2021). Efficiency lower bounds for distribution-free Hotelling-type two-sample tests based on optimal transport. ArXiv:2104.01986.

Deb, N., Ghosal, P., & Sen, B. (2020). Measuring association on topological spaces using kernels and geometric graphs. ArXiv:2010.01768.

Deb, N., & Sen, B. (2022). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*. In press. https://doi.org/10.1080/01621459.2021.1923508.

del Barrio, E., González-Sanz, A., & Hallin, M. (2020). A note on the regularity of optimal-transport-based center-outward distribution and quantile functions. *Journal of Multivariate Analysis*, *180*, 104671, 13. https://doi.org/10.1016/j.jmva.2020.104671.

Faugeras, O., & Rüschendorf, L. (2017). Markov morphisms: a combined copula and mass transportation approach to multivariate quantiles. *Mathematica Applicanda*, *45*, 21–63.

Figalli, A. (2018). On the continuity of center-outward distribution and quantile functions. *Nonlinear Analysis*, *177*, 413–21. https://doi.org/10.1016/j.na.2018.05.008.

Hájek, J., & Šidák, Z. (1967). *Theory of Rank Tests*. New York: Academic Press.

Hallin, M. (2017). On distribution and quantile functions, ranks, and signs in $\mathbb{R}^d$: A measure transportation approach. ideas.repec.org/p/eca/wpaper/2013-258262.html.

Hallin, M. (2022). Measure transportation and statistical decision theory. *Annual Review of Statistics and Its Applications*, *9*, 401–424. https://doi.org/10.1146/annurev-statistics-040220-105948.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., & Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, *49*, 1139–1165.

Hallin, M., Hlubinka, D., & Hudecová, Š. (2022a). Fully distribution-free center-outward rank tests for multiple-output regression and MANOVA. *Journal of the American Statistical Association*. In press.

Hallin, M., La Vecchia, D., & Liu, H. (2022b). Center-outward R-estimation for semiparametric VARMA models. *Journal of the American Statistical Association*, *117*, 925–938. https://doi.org/10.1080/01621459.2020.1832501.

Hallin, M., La Vecchia, D., & Liu, H. (2022c). Rank-based testing for semiparametric VAR models: A measure transportation approach. *Bernoulli*. In press, *29*, 229–273.

Hampel, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. thesis, University of California, Berkeley.

Hornik, K. (2022). Package clue: Cluster ensembles, R package version 0.3-63. https://CRAN.R-project.org/package=clue.

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*, 73–101.

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, *80*, 309–324.

McDonald, T., & McDonald, A. (2020). Package SDraw: Spatially balanced samples of spatial objects. R package version 2.1.13. https://CRAN.R-project.org/package=SDraw.

Mordant, G. (2021). Transporting Probability Measures: some contributions to statistical inference. Ph.D. thesis, Université catholique de Louvain.

Ronchetti, E. (2006). The historical development of robust statistics. In A. Rossman & B. Chance (Eds.), *ICOTS-7 Proceedings*. IASE.

Shi, H., Drton, M., & Han, F. (2022a). Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, *117*, 395–410.

Shi, H., Hallin, M., Drton, M., & Han, F. (2021). Center-outward sign- and rank-based quadrant, Spearman, and Kendall tests for multivariate independence. arXiv:2111.15567.

Shi, H., Hallin, M., Drton, M., & Han, F. (2022b). On universally consistent and fully distribution-free rank tests of vector independence. *Annals of Statistics*, *50*, 1933–1959.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, *68*, 872–879.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin (Ed.), *Contributions to probability and statistics*, (pp. 448–485). Palo Alto: Stanford University Press.

Refining Invariant Coordinate Selection via Local Projection Pursuit

Lutz Dümbgen, Katrin Gysel, and Fabrice Perler

Abstract Invariant coordinate selection (ICS), introduced by Tyler et al. (J. Roy. Stat. Soc. B 71(3):549–592, 2009), is a powerful tool to find potentially interesting projections of multivariate data. In some cases, some of the projections proposed by ICS come close to really interesting ones, but little deviations can result in a blurred view which does not reveal the feature (e.g., a clustering), which would otherwise be clearly visible. To remedy this problem, we propose an automated and localized version of projection pursuit (PP), cf. Huber (Ann. Stat. 13(2):435–525, 1985). Precisely, our local search is based on gradient descent applied to estimated differential entropy as a function of the projection matrix.

Keywords Differential entropy · Gradient descent · Local perturbation · Local search · Orthogonal matrix · Orthogonal projection

1 Projection Pursuit

Suppose our data consist of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$, and we view these as independent copies from a random vector $\boldsymbol{X}$ with unknown distribution P. If $p \in \{1, 2, 3\}$, there are various ways to display the data graphically and find interesting structures, e.g., two or more separated clusters or manifolds close to which the data are concentrated. If we use a particular method to visualize data sets in dimension $d \in \{1, 2, 3\}$ but $p > d$, we want to find a d-dimensional linear projection of the data which exhibits interesting structure. This task has been coined “projection pursuit”

L. Dümbgen (✉)
University of Bern, Bern, Switzerland
e-mail: duembgen@stat.unibe.ch

K. Gysel
SAKK, Bern, Switzerland

F. Perler
Bundesamt für Gesundheit, Bern, Switzerland

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_6

(PP) by Friedman and Tukey (1974). We also refer to the excellent discussion papers of Huber (1985) and Jones and Sibson (1987) for different aspects and variants of this paradigm.

More formally, if the distribution P has been standardized already in some way, our goal is to determine a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times d}$ with orthonormal columns such that the distribution $P_{\boldsymbol{A}}$ of $\boldsymbol{A}^\top \boldsymbol{X}$ is "interesting." In what follows, such a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times d}$, that is, $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_d$, is called a "($d$-dimensional) projection matrix," and the distribution $P_{\boldsymbol{A}}$ is called a "projection of P (via $\boldsymbol{A}$)."

An obvious question is how to measure whether a distribution Q on $\mathbb{R}^d$ is "interesting." To answer this, let us summarize some of the considerations of Huber (1985). Suppose that Q has density g with respect to d-dimensional Lebesgue measure. Shannon's (differential) entropy of Q is defined as

$$H(Q) := -\int \log(g(\boldsymbol{y}))g(\boldsymbol{y})\,d\boldsymbol{y}.$$

It is well known that among all distributions Q with given mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and nonsingular covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$, the Gaussian distribution $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the unique maximizer of $H(Q)$. Coming back to the distribution P, that its projection $Q = P_{\boldsymbol{A}}$ is non-interesting if it is Gaussian is also supported by the so-called Diaconis–Freedman effect, cf. Diaconis and Freedman (1984). Under mild assumptions on P, most projections $P_{\boldsymbol{A}}$ look Gaussian. Precisely, if $\boldsymbol{A}$ is uniformly distributed on the manifold of all d-dimensional projection matrices, then for fixed d,

$$P_{\boldsymbol{A}} \to_{w,\mathbb{P}} \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{I}_d) \quad \text{as } p \to \infty,\ p^{-1}\|\boldsymbol{X}\|^2 \to_{\mathbb{P}} 1,\ p^{-1}\boldsymbol{X}^\top\tilde{\boldsymbol{X}} \to_{\mathbb{P}} 0,$$

where $\tilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$; see also Dümbgen and Del Conte-Zerial (2013).

In view of these considerations, a possible strategy is to find a projection matrix $\boldsymbol{A}$ such that $\hat{H}(\boldsymbol{A}^\top\boldsymbol{x}_1, \ldots, \boldsymbol{A}^\top\boldsymbol{x}_n)$ is minimal, where $\hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ is an estimator of $H(Q)$, based on observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^d$.

In this chapter, we focus on the entropy $H(Q)$ and the estimated entropy $\hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. As discussed by Huber (1985), Jones and Sibson (1987), and many other authors, there are numerous proposals of "PP indices" measuring how interesting a distribution Q or the empirical distribution of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ is. As explained at the end of Sect. 4, our local search method can be easily adapted to different PP indices.

2 Invariant Coordinate Selection as a Starting Point

Invariant coordinate selection (ICS), introduced as a generalization of independent component analysis by Tyler et al. (2009), may be described as a two-step procedure. In a first step, the centered raw observations $\boldsymbol{x}_1^{\mathrm{raw}}, \ldots, \boldsymbol{x}_n^{\mathrm{raw}}$ are standardized by means of some scatter estimator $\hat{\boldsymbol{\Sigma}}_0 = \hat{\boldsymbol{\Sigma}}_0(\boldsymbol{x}_1^{\mathrm{raw}}, \ldots, \boldsymbol{x}_n^{\mathrm{raw}}) \in \mathbb{R}^{p\times p}_{\mathrm{sym},+}$, where

$\mathbb{R}^{p\times p}_{\mathrm{sym},+}$ stands for the set of symmetric, positive definite matrices in $\mathbb{R}^{p\times p}$. Having determined $\hat{\boldsymbol{\Sigma}}_0 = \boldsymbol{B}_0\boldsymbol{B}_0^\top$, we replace the raw observations $\boldsymbol{x}_i^{\mathrm{raw}}$ with the standardized observations $\boldsymbol{x}_i := \boldsymbol{B}_0^{-1}\boldsymbol{x}_i^{\mathrm{raw}}$. Strictly speaking, these standardized observations $\boldsymbol{x}_i$ are no longer stochastically independent, but this is not essential for the subsequent considerations.

To the preprocessed observations $\boldsymbol{x}_i$, we apply a different estimator of scatter to obtain $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{p\times p}_{\mathrm{sym},+}$. Now, we compute a spectral decomposition $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i^\top$ with eigenvalues $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p > 0$ and an orthonormal basis $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_p$ of $\mathbb{R}^p$. Then, the resulting invariant coordinates correspond to the mappings

$$\mathbb{R}^q \ni \boldsymbol{x}^{\mathrm{raw}} \mapsto \hat{\boldsymbol{u}}_k^\top \boldsymbol{B}_0^{-1}\boldsymbol{x}^{\mathrm{raw}} \in \mathbb{R}, \qquad 1 \le k \le p.$$

The results of Tyler et al. (2009) suggest to look at the $d+1$ particular projection matrices

$$\begin{aligned} \boldsymbol{A}_0 &:= \big[\hat{\boldsymbol{u}}_1 \ldots \hat{\boldsymbol{u}}_d\big], \\ \boldsymbol{A}_1 &:= \big[\hat{\boldsymbol{u}}_1 \ldots \hat{\boldsymbol{u}}_{d-1}\, \hat{\boldsymbol{u}}_p\big], \\ &\vdots \\ \boldsymbol{A}_d &:= \big[\hat{\boldsymbol{u}}_{p-d+1} \ldots \hat{\boldsymbol{u}}_p\big], \end{aligned}$$

that is, the columns of $\boldsymbol{A}_k$ are the vectors $\hat{\boldsymbol{u}}_i$ with $1 \le i \le d-k$ or $p-k < i \le p$. One could also consider all $\binom{p}{d}$ matrices

$$\boldsymbol{A} = [\hat{\boldsymbol{u}}_{i(1)} \cdots \hat{\boldsymbol{u}}_{i(d)}] \quad \text{with} \quad 1 \le i(1) < \cdots < i(d) \le p.$$

3 Estimation of Entropy

For given observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^d$ with unknown distribution Q and density g, a standard estimator of g would be a kernel density estimator with standard Gaussian kernel,

$$\hat{g}_h(\boldsymbol{y}) := n^{-1}\sum_{j=1}^{n} \phi_h(\boldsymbol{y} - \boldsymbol{y}_j),$$

where $\phi_h(\boldsymbol{y}) := h^{-d}\phi(h^{-1}\boldsymbol{y})$ with $\phi(\boldsymbol{y}) := (2\pi)^{-d/2}\exp(-\|\boldsymbol{y}\|^2/2)$, and $h = h(n) > 0$ is some bandwidth to be specified later. Then, a possible estimator of $H(Q)$ is given by

$$\hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) := -n^{-1}\sum_{i=1}^{n} \log \hat{g}_h(\boldsymbol{y}_i).$$

Note that $\hat{H}(\cdot)$ is continuously differentiable with

$$\begin{aligned}
&\hat{H}(\boldsymbol{y}_1 + \boldsymbol{\delta}_1, \ldots, \boldsymbol{y}_n + \boldsymbol{\delta}_n) \\
&\quad = \hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \\
&\qquad + \ n^{-1}h^{-2} \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} (\boldsymbol{\delta}_i - \boldsymbol{\delta}_j)^\top (\boldsymbol{y}_i - \boldsymbol{y}_j) \phi_h(\boldsymbol{y}_i - \boldsymbol{y}_j)}{\sum_{j=1}^{n} \phi_h(\boldsymbol{y}_i - \boldsymbol{y}_j)} \\
&\qquad + \ O\big(\|\boldsymbol{\delta}_1\|^2 + \cdots + \|\boldsymbol{\delta}_n\|^2\big)
\end{aligned}$$

as $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n \to \boldsymbol{0}$ because $\phi_h(\boldsymbol{y} + \boldsymbol{\delta}) = -h^{-2}\boldsymbol{\delta}^\top \boldsymbol{y}\phi_h(\boldsymbol{y}) + O(\|\boldsymbol{\delta}\|^2)$ as $\boldsymbol{\delta} \to \boldsymbol{0}$. This expansion will be useful for local optimization.

The smoothing parameter h has an impact, of course. Suppose that the underlying distribution Q is the standard Gaussian $\mathcal{N}_d(\boldsymbol{0}, \boldsymbol{I}_d)$. Then, the expected value of $\hat{g}(\boldsymbol{y})$ equals $\phi_{(1+h^2)^{1/2}}(y)$, the density of the convolution of Q and $\mathcal{N}_d(\boldsymbol{0}, h^2\boldsymbol{I}_d)$. Hence, $\hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d)$ may be viewed as an estimator of

$$-\int \log \phi_{(1+h^2)^{1/2}}(\boldsymbol{y})\phi(\boldsymbol{y})\, d\boldsymbol{y} \ = \ (d/2)\big((1 + h^2)^{-1} + \log(1 + h^2) + \log(2\pi)\big). \tag{1}$$

4 Local Optimization

For our purposes, it is convenient to over-parametrize the search problem by writing

$$\boldsymbol{A} \ = \ \boldsymbol{U}\boldsymbol{\Pi} \tag{2}$$

with an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ and the standard projection matrix

$$\boldsymbol{\Pi} \ := \ \begin{bmatrix} \boldsymbol{I}_d \\ \boldsymbol{0} \end{bmatrix} \ \in \ \mathbb{R}^{p\times d},$$

which reduces a vector $\boldsymbol{x} \in \mathbb{R}^p$ to its subvector $\boldsymbol{\Pi}^\top \boldsymbol{x} = (x_i)_{i=1}^d$. Instead of looking for a suitable projection matrix $\boldsymbol{A}$ directly, we are looking for a suitable orthogonal matrix $\boldsymbol{U}$ such that $\hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{U}^\top \boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{U}^\top \boldsymbol{x}_n)$ is particularly small.

For a rigorous description of the local search strategy, we need to introduce some notation and geometry. Recall first that any matrix space $\mathbb{R}^{p\times q}$ becomes a Euclidean space by means of the inner product $\langle \boldsymbol{B}, \boldsymbol{C}\rangle := \mathrm{trace}(\boldsymbol{B}^\top \boldsymbol{C}) = \sum_{i,j} B_{ij}C_{ij}$, and the resulting norm is the Frobenius norm $\|\boldsymbol{B}\|_F = \big(\sum_{i,j} B_{ij}^2\big)^{1/2}$. In the special case of $p = q$, it is well known that $\mathbb{R}^{p\times p}$ is the sum of the orthogonal linear spaces $\mathbb{R}_{\mathrm{sym}}^{p\times p}$ and $\mathbb{R}_{\mathrm{anti}}^{p\times p}$ of symmetric and antisymmetric matrices, respectively. Indeed, any matrix $\boldsymbol{\Delta} \in \mathbb{R}^{p\times p}$ can be written as $\boldsymbol{\Delta} = \boldsymbol{\Delta}_{\mathrm{sym}} + \boldsymbol{\Delta}_{\mathrm{anti}}$ with the symmetric matrix $\boldsymbol{\Delta}_{\mathrm{sym}} := 2^{-1}(\boldsymbol{\Delta} + \boldsymbol{\Delta}^\top)$ and the antisymmetric matrix $\boldsymbol{\Delta}_{\mathrm{anti}} := 2^{-1}(\boldsymbol{\Delta} - \boldsymbol{\Delta}^\top)$.

Searching locally means that a given candidate $\boldsymbol{U}$ in (2) is multiplied from the right with another orthogonal matrix $\boldsymbol{V}$ which is close to the identity matrix $\boldsymbol{I}_p$. Specifically, let $\boldsymbol{V}$ be equal to

$$\exp(\boldsymbol{\Delta}) \;=\; \sum_{k=0}^{\infty}(k!)^{-1}\boldsymbol{\Delta}^k$$

for $\boldsymbol{\Delta} \in \mathbb{R}^{p\times p}_{\mathrm{anti}}$. It is well known that $\exp(\cdot)$ defines a surjective mapping from $\mathbb{R}^{p\times p}_{\mathrm{anti}}$ to the set of orthogonal matrices in $\mathbb{R}^{p\times p}$ with determinant 1, where $\exp(\boldsymbol{0}) = \boldsymbol{I}_p$. Moreover, $\exp(\boldsymbol{\Delta})^\top = \exp(-\boldsymbol{\Delta})$, and

$$\exp(\boldsymbol{\Delta}) \;=\; \boldsymbol{I}_p + \boldsymbol{\Delta} + O(\|\boldsymbol{\Delta}\|_F^2).$$

To find a promising new orthogonal matrix $\boldsymbol{U}\exp(\boldsymbol{\Delta})$, we may assume without loss of generality that $\boldsymbol{U} = \boldsymbol{I}_p$ because

$$\boldsymbol{\Pi}^\top(\boldsymbol{U}\exp(\boldsymbol{\Delta}))^\top\boldsymbol{x}_i \;=\; \boldsymbol{\Pi}^\top\exp(\boldsymbol{\Delta})^\top(\boldsymbol{U}^\top\boldsymbol{x}_i).$$

Hence, we may replace $\boldsymbol{x}_i$ with $\boldsymbol{U}^\top\boldsymbol{x}_i$ for $1 \le i \le p$ and then look for a promising perturbation $\exp(\boldsymbol{\Delta})$ of $\boldsymbol{I}_p$. To this end, let

$$\boldsymbol{x}_i \;=\; \begin{bmatrix}\boldsymbol{y}_i\\ \boldsymbol{z}_i\end{bmatrix} \quad \text{with } \boldsymbol{y}_i \in \mathbb{R}^d,\ \boldsymbol{z}_i \in \mathbb{R}^{p-d}, \tag{3}$$

that is, $\boldsymbol{y}_i = \boldsymbol{\Pi}^\top\boldsymbol{x}_i$. If we write

$$\boldsymbol{\Delta} \;=\; \begin{bmatrix}\boldsymbol{\Delta}_1 & -\boldsymbol{C}^\top\\ \boldsymbol{C} & \boldsymbol{\Delta}_2\end{bmatrix} \tag{4}$$

with arbitrary matrices $\boldsymbol{\Delta}_1 \in \mathbb{R}^{d\times d}_{\mathrm{anti}}$, $\boldsymbol{\Delta}_2 \in \mathbb{R}^{(p-d)\times(p-d)}_{\mathrm{anti}}$, and $\boldsymbol{C} \in \mathbb{R}^{(p-d)\times d}$, then

$$\boldsymbol{\Pi}^\top\exp(\boldsymbol{\Delta})^\top\boldsymbol{x}_i \;=\; \boldsymbol{y}_i + \boldsymbol{\Delta}_1^\top\boldsymbol{y}_i + \boldsymbol{C}^\top\boldsymbol{z}_i + O(\|\boldsymbol{\Delta}\|_F^2).$$

Consequently, it follows from the general expansion of $\hat{H}(\cdot)$ in the previous section that

$$\begin{aligned}
&\hat{H}\big(\boldsymbol{\Pi}^\top\exp(\boldsymbol{\Delta})^\top\boldsymbol{x}_1,\ldots,\boldsymbol{\Pi}^\top\exp(\boldsymbol{\Delta})^\top\boldsymbol{x}_n\big) - \hat{H}(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)\\
&= n^{-1}h^{-2}\sum_{i=1}^n \frac{\sum_{j=1}^n\big(\boldsymbol{\Delta}_1^\top(\boldsymbol{y}_i-\boldsymbol{y}_j) + \boldsymbol{C}^\top(\boldsymbol{z}_i-\boldsymbol{z}_j)\big)^\top(\boldsymbol{y}_i-\boldsymbol{y}_j)\phi_j(\boldsymbol{y}_i-\boldsymbol{y}_j)}{\sum_{j=1}^n\phi_h(\boldsymbol{y}_i-\boldsymbol{y}_j)}\\
&\qquad + \;O(\|\boldsymbol{\Delta}\|_F^2)
\end{aligned}$$

$$= \langle \boldsymbol{C}, \hat{\boldsymbol{C}} \rangle + O(\|\boldsymbol{\Delta}\|_F^2) \tag{5}$$

as $\boldsymbol{\Delta} \to \boldsymbol{0}$, where

$$\hat{\boldsymbol{C}} := n^{-1} h^{-2} \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} \phi_h(\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{z}_i - \boldsymbol{z}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top}{\sum_{j=1}^{n} \phi_h(\boldsymbol{y}_i - \boldsymbol{y}_j)} \in \mathbb{R}^{(p-q)\times d}. \tag{6}$$

In the last step, we used the representations

$$\begin{aligned} (\boldsymbol{\Delta}_1^\top(\boldsymbol{y}_i - \boldsymbol{y}_j))^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) &= \big\langle \boldsymbol{\Delta}_1, (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top \big\rangle, \\ (\boldsymbol{C}^\top(\boldsymbol{z}_i - \boldsymbol{z}_j))^\top(\boldsymbol{y}_i - \boldsymbol{y}_j) &= \big\langle \boldsymbol{C}, (\boldsymbol{z}_i - \boldsymbol{z}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top \big\rangle \end{aligned}$$

and the fact that $\boldsymbol{\Delta}_1 \in \mathbb{R}^{d\times d}_{\mathrm{anti}}$ is perpendicular to $(\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^\top \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$. Since

$$\langle \boldsymbol{C}, \hat{\boldsymbol{C}} \rangle = 2^{-1} \langle \boldsymbol{\Delta}, \hat{\boldsymbol{\Delta}} \rangle \quad \text{with} \quad \hat{\boldsymbol{\Delta}} := \begin{bmatrix} \boldsymbol{0} & -\hat{\boldsymbol{C}}^\top \\ \hat{\boldsymbol{C}} & \boldsymbol{0} \end{bmatrix},$$

expansion (5) shows that the gradient of the mapping

$$\mathbb{R}^{p\times p}_{\mathrm{anti}} \ni \boldsymbol{\Delta} \mapsto \hat{H}\big(\boldsymbol{\Pi}^\top \exp(\boldsymbol{\Delta})^\top \boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \exp(\boldsymbol{\Delta})^\top \boldsymbol{x}_n\big)$$

at $\boldsymbol{\Delta} = \boldsymbol{0}$ is given by $2^{-1}\hat{\boldsymbol{\Delta}}$. Consequently, promising candidates for the factor $\exp(\boldsymbol{\Delta})$ are given by

$$\exp(-t\hat{\boldsymbol{\Delta}}) = \exp(t\hat{\boldsymbol{\Delta}})^\top, \quad t \geq 0.$$

The explicit computation of $\exp(t\hat{\boldsymbol{\Delta}})$ is rather convenient when working with a singular value decomposition of $\hat{\boldsymbol{C}}$. With $m := \min(d, p-d)$, suppose that

$$\hat{\boldsymbol{C}} = \boldsymbol{W} \operatorname{diag}(\boldsymbol{\sigma}) \boldsymbol{V}^\top$$

with matrices $\boldsymbol{W} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_m] \in \mathbb{R}^{(p-d)\times m}$ and $\boldsymbol{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_m] \in \mathbb{R}^{d\times m}$ of singular vectors such that $\boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_m$ and a vector $\boldsymbol{\sigma} \in [0,\infty)^m$ of singular values. If we define

$$\hat{\boldsymbol{v}}_i := \begin{bmatrix} \boldsymbol{v}_i \\ \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^p \quad \text{and} \quad \hat{\boldsymbol{w}}_i := \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{w}_i \end{bmatrix} \in \mathbb{R}^p,$$

then $\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{v}}_i = \sigma_i \hat{\boldsymbol{w}}_i$ and $\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{w}}_i = -\sigma_i \hat{\boldsymbol{v}}_i$ for $1 \leq i \leq m$, while $\hat{\boldsymbol{\Delta}}\boldsymbol{x} = \boldsymbol{0}$ for $\boldsymbol{x} \perp \{\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_m, \hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_m\}$. From this, one can deduce that for $1 \leq i \leq m$,

$$\exp(t\hat{\boldsymbol{\Delta}})\hat{\boldsymbol{v}}_i = \cos(t\sigma_i)\hat{\boldsymbol{v}}_i + \sin(t\sigma_i)\hat{\boldsymbol{w}}_i,$$

$$\exp(t\hat{\boldsymbol{\Delta}})\hat{\boldsymbol{w}}_i \;=\; -\sin(t\sigma_i)\hat{\boldsymbol{v}}_i + \cos(t\sigma_i)\hat{\boldsymbol{w}}_i,$$

while $\exp(t\hat{\boldsymbol{\Delta}})\boldsymbol{x} = \boldsymbol{x}$ for $\boldsymbol{x} \perp \{\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_m, \hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_m\}$. In other words,

$$\begin{aligned}\exp(t\hat{\boldsymbol{\Delta}}) \;&=\; \begin{bmatrix} \boldsymbol{I}_d + \boldsymbol{V}\operatorname{diag}(\cos(t\boldsymbol{\sigma}) - 1)\boldsymbol{V}^\top & -\boldsymbol{V}\operatorname{diag}(\sin(t\boldsymbol{\sigma}))\boldsymbol{W}^\top \\ \boldsymbol{W}\operatorname{diag}(\sin(t\boldsymbol{\sigma}))\boldsymbol{V}^\top & \boldsymbol{I}_{p-d} + \boldsymbol{W}\operatorname{diag}(\cos(t\boldsymbol{\sigma}) - 1)\boldsymbol{W}^\top \end{bmatrix} \\ &=\; \begin{bmatrix} \boldsymbol{I}_d - \boldsymbol{V}\operatorname{diag}(2\sin(t\boldsymbol{\sigma}/2)^2)\boldsymbol{V}^\top & -\boldsymbol{V}\operatorname{diag}(\sin(t\boldsymbol{\sigma}))\boldsymbol{W}^\top \\ \boldsymbol{W}\operatorname{diag}(\sin(t\boldsymbol{\sigma}))\boldsymbol{V}^\top & \boldsymbol{I}_{p-d} - \boldsymbol{W}\operatorname{diag}(2\sin(t\boldsymbol{\sigma}/2)^2)\boldsymbol{W}^\top \end{bmatrix},\end{aligned}$$

where the functions of $t\boldsymbol{\sigma}$ are computed component-wise. Note that the upper left block $\boldsymbol{I}_d + \boldsymbol{V}\operatorname{diag}(\cos(t\boldsymbol{\sigma}) - 1)\boldsymbol{V}^\top$ equals $\boldsymbol{V}\operatorname{diag}(\cos(t\boldsymbol{\sigma}))\boldsymbol{V}^\top$ in case of $d = m$, and the lower right block $\boldsymbol{I}_{p-d} + \boldsymbol{W}\operatorname{diag}(\cos(t\boldsymbol{\sigma}) - 1)\boldsymbol{W}^\top$ equals $\boldsymbol{W}\operatorname{diag}(\cos(t\boldsymbol{\sigma}))\boldsymbol{W}^\top$ in case of $p - d = m$.

For the explicit choice of $t \geq 0$, we propose an Armijo–Goldstein procedure, see Nocedal and Wright (2006). Specifically, recall that the auxiliary function $\hat{h}(t) := \hat{H}\big(\boldsymbol{\Pi}^\top \exp(t\hat{\boldsymbol{\Delta}})\boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \exp(t\hat{\boldsymbol{\Delta}})\boldsymbol{x}_n\big)$ satisfies $\hat{h}(0) = \hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and $\hat{h}'(0) = -\|\hat{\boldsymbol{C}}\|_F^2$. Now, we choose $t = 2^{-k}$ with the smallest integer $k \geq 0$ such that the improvement $\hat{h}(0) - \hat{h}(2^{-k})$ is at least $-2^{-k}\hat{h}'(0)/3 = 2^{-k}\|\hat{\boldsymbol{C}}\|_F^2/3$.

Using Arbitrary PP Indices Suppose we replace estimated entropy with an arbitrary continuously differentiable function $\hat{H} : (\mathbb{R}^d)^n \to \mathbb{R}$. Then, there exist vectors $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_i(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \in \mathbb{R}^d$ such that for arbitrary perturbations $\boldsymbol{\delta}_i \in \mathbb{R}^d$,

$$\hat{H}(\boldsymbol{y}_1 + \boldsymbol{\delta}_1, \ldots, \boldsymbol{y}_n + \boldsymbol{\delta}_n) \;=\; \hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) + \sum_{i=1}^n \boldsymbol{\gamma}_i^\top \boldsymbol{\delta}_i + o\Big(\sum_{i=1}^n \|\boldsymbol{\delta}_i\|\Big)$$

as $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n \to \boldsymbol{0}$. With $\boldsymbol{x}_i$ and $\boldsymbol{\Delta}$ as in (3) and (4),

$$\begin{aligned}&\hat{H}\big(\boldsymbol{\Pi}^\top \exp(\boldsymbol{\Delta})^\top \boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \exp(\boldsymbol{\Delta})^\top \boldsymbol{x}_n\big) \\ &\quad=\; \hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) + \sum_{i=1}^n \boldsymbol{\gamma}_i^\top (\boldsymbol{\Delta}_1^\top \boldsymbol{y}_i + \boldsymbol{C}^\top \boldsymbol{z}_i) + o(\|\boldsymbol{\Delta}\|_F)\end{aligned}$$

as $\boldsymbol{\Delta} \to \boldsymbol{0}$. If $\hat{H}$ is orthogonally invariant in the sense that $\hat{H}(\boldsymbol{V}\boldsymbol{y}_1, \ldots, \boldsymbol{V}\boldsymbol{y}_n) = \hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ for arbitrary orthogonal matrices $\boldsymbol{V} \in \mathbb{R}^{d\times d}$, then $\sum_{i=1}^n \boldsymbol{\gamma}_i^\top \boldsymbol{\Delta}_1^\top \boldsymbol{y}_i = 0$. This can be seen by considering the special case $\boldsymbol{C} = \boldsymbol{0}$ and $\boldsymbol{\Delta}_2 = \boldsymbol{0}$, because then $\boldsymbol{\Pi}^\top \exp(\boldsymbol{\Delta})^\top \boldsymbol{x}_i = \exp(\boldsymbol{\Delta}_1)^\top \boldsymbol{y}_i$, and $\exp(\boldsymbol{\Delta}_1)$ is orthogonal. Consequently, the previous expansion of $\hat{H}$ simplifies to

$$\hat{H}\big(\boldsymbol{\Pi}^\top \exp(\boldsymbol{\Delta})^\top \boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \exp(\boldsymbol{\Delta})^\top \boldsymbol{x}_n\big) = \hat{H}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) + \langle \boldsymbol{C}, \hat{\boldsymbol{C}} \rangle + o(\|\boldsymbol{\Delta}\|_F)$$

as $\boldsymbol{\Delta} \to \mathbf{0}$, where

$$\hat{\boldsymbol{C}} := \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{y}_i^\top \in \mathbb{R}^{(p-d)\times d}.$$

Hence, our version of local optimization may be applied to any smooth PP index $\hat{H}$ which is orthogonally invariant.

5 The Complete Procedure(s)

The complete procedure consists of three different basic procedures.

Basic Procedure 1 (Pre-whitening) Given the centered raw data $\boldsymbol{x}_1^{\mathrm{raw}}, \ldots, \boldsymbol{x}_n^{\mathrm{raw}}$, we compute the preliminary scatter estimator $\hat{\Sigma}_0(\boldsymbol{x}_1^{\mathrm{raw}}, \ldots, \boldsymbol{x}_n^{\mathrm{raw}}) = \boldsymbol{B}_0 \boldsymbol{B}_0^\top$. Then, we set

$$\boldsymbol{x}_i^{\mathrm{pre}} := \boldsymbol{B}_0^{-1} \boldsymbol{x}_i^{\mathrm{raw}}.$$

Basic Procedure 2 (ICS) Now, we compute the second scatter estimator and its spectral decomposition: $\hat{\Sigma}(\boldsymbol{x}_1^{\mathrm{pre}}, \ldots, \boldsymbol{x}_n^{\mathrm{pre}}) = \hat{\boldsymbol{U}} \operatorname{diag}(\hat{\boldsymbol{\lambda}}) \hat{\boldsymbol{U}}^\top$ with an orthogonal matrix $\hat{\boldsymbol{U}} \in \mathbb{R}^{p\times p}$ and a vector $\hat{\boldsymbol{\lambda}} \in (0, \infty)^p$ of eigenvectors. Then, we set

$$\boldsymbol{x}_i^{\mathrm{ics}} := \hat{\boldsymbol{U}}^\top \boldsymbol{x}_i^{\mathrm{pre}}.$$

Basic Procedure 3 (Local PP) For given indices $j(1) < \cdots < j(d)$ in $\{1, 2, \ldots, p\}$, let $\ell(1) < \cdots < \ell(p-d)$ be the elements of $\{1, 2, \ldots, p\} \setminus \{j(1), \ldots, j(d)\}$. With the standard basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ of $\mathbb{R}^p$, we define the permutation matrix $\boldsymbol{U} := [\boldsymbol{e}_{j(1)} \cdots \boldsymbol{e}_{j(d)}\, \boldsymbol{e}_{\ell(1)} \cdots \boldsymbol{e}_{\ell(p-d)}]$ and set

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \leftarrow (\boldsymbol{U}^\top \boldsymbol{x}_1^{\mathrm{ics}}, \ldots, \boldsymbol{U}^\top \boldsymbol{x}_n^{\mathrm{ics}}).$$

Now, we start the following iterative algorithm with some small threshold $\delta_o > 0$:

```
$\hat{H} \leftarrow \hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{x}_n)$
$\hat{\boldsymbol{C}} \leftarrow \hat{\boldsymbol{C}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
$\delta \leftarrow \|\hat{\boldsymbol{C}}\|_F^2$
while $\delta \geq \delta_o$ do
    $(\boldsymbol{V}, \boldsymbol{\sigma}, \boldsymbol{W}) \leftarrow \mathrm{SVD}(\hat{\boldsymbol{C}})$
    $\boldsymbol{U} \leftarrow \mathrm{Exp}(\boldsymbol{V}, \boldsymbol{\sigma}, \boldsymbol{W})$
    $(\boldsymbol{x}_1^{\mathrm{tmp}}, \ldots, \boldsymbol{x}_n^{\mathrm{tmp}}) \leftarrow (\boldsymbol{U}\boldsymbol{x}_1, \ldots, \boldsymbol{U}\boldsymbol{x}_n)$
    $\hat{H}^{\mathrm{tmp}} \leftarrow \hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{x}_1^{\mathrm{tmp}}, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{x}_n^{\mathrm{tmp}})$
    while $\hat{H} - \hat{H}^{\mathrm{tmp}} < \delta/3$ do
        $\delta \leftarrow \delta/2$
        $\boldsymbol{\sigma} \leftarrow \boldsymbol{\sigma}/2$
        $\boldsymbol{U} \leftarrow \mathrm{Exp}(\boldsymbol{V}, \boldsymbol{\sigma}, \boldsymbol{W})$
        $(\boldsymbol{x}_1^{\mathrm{tmp}}, \ldots, \boldsymbol{x}_n^{\mathrm{tmp}}) \leftarrow (\boldsymbol{U}\boldsymbol{x}_1, \ldots, \boldsymbol{U}\boldsymbol{x}_n)$
        $\hat{H}^{\mathrm{tmp}} \leftarrow \hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{x}_1^{\mathrm{tmp}}, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{x}_n^{\mathrm{tmp}})$
    end while
    $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \leftarrow (\boldsymbol{x}_1^{\mathrm{tmp}}, \ldots, \boldsymbol{x}_n^{\mathrm{tmp}})$
    $\hat{H} \leftarrow \hat{H}^{\mathrm{tmp}}$
    $\hat{\boldsymbol{C}} \leftarrow \hat{\boldsymbol{C}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
    $\delta \leftarrow \|\hat{\boldsymbol{C}}\|_F^2$
end while
```

Here, $\hat{\boldsymbol{C}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{(p-d)\times d}$ is given by (6), $\mathrm{SVD}(\hat{\boldsymbol{C}})$ yields the ingredients for the singular value decomposition $\hat{\boldsymbol{C}} = \boldsymbol{W}\,\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{V}^\top$, and $\mathrm{Exp}(\boldsymbol{V}, \boldsymbol{\sigma}, \boldsymbol{W})$ computes

$$\exp\left(\begin{bmatrix} \boldsymbol{0} & -\boldsymbol{V}\,\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{W}^\top \\ \boldsymbol{W}\,\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{V}^\top & \boldsymbol{0} \end{bmatrix}\right).$$

The inner while-loop is the Armijo–Goldstein step size correction mentioned before.

The iterative algorithm will always converge. In each instance of the outer while-loop, the data $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are replaced with $(\boldsymbol{U}\boldsymbol{x}_1, \ldots, \boldsymbol{U}\boldsymbol{x}_n)$ with some orthogonal matrix $\boldsymbol{U}$ such that $\hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{x}_n)$ decreases strictly. Since the set of all orthogonal matrices is a compact differentiable manifold and since $\hat{\boldsymbol{C}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a continuous function of its input data which is closely related to the gradient of $\hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{U}\boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{U}\boldsymbol{x}_n)$ as a function of $\boldsymbol{U}$, the condition $\|\hat{\boldsymbol{C}}\|_F^2 < \delta_o$ has to be satisfied after finitely many steps.

Basic procedure 3 is executed with $d + 1$ or $\binom{p}{d}$ different choices of $i(1) < \cdots < i(d)$. It is also possible to compute first $\hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{x}_1, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{x}_n)$ for all these choices and then start local PP only for the choice with minimal initial value of $\hat{H}$. Alternatively, one can inspect a scatter plot matrix of the data $\boldsymbol{x}_i^{\mathrm{ics}}$ visually and then run a local search, either with $d = 1$ and the most promising index j or with $d = 2$ and the most promising indices $1 \le j(1) < j(2) \le p$.

The Result Running basic procedures 1, 2, and 3 leads to transformed observations $\boldsymbol{x}_i = \boldsymbol{B}^\top \boldsymbol{x}_i^{\mathrm{raw}}$ such that the d-dimensional observations $\boldsymbol{\Pi}^\top \boldsymbol{x}_i$ reveal (hopefully) some interesting feature of the raw data.

The nonsingular matrix $\boldsymbol{B} \in \mathbb{R}^{p\times p}$ may be recovered quickly via multivariate least squares: with the data matrices $\underline{\boldsymbol{X}}_{\mathrm{raw}} = [\boldsymbol{x}_1^{\mathrm{raw}} \ldots \boldsymbol{x}_n^{\mathrm{raw}}]^\top$ and $\underline{\boldsymbol{X}} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_n]^\top$ in $\mathbb{R}^{n\times p}$, the matrix $\boldsymbol{B}$ satisfies $\underline{\boldsymbol{X}}_{\mathrm{raw}}\boldsymbol{B} = \underline{\boldsymbol{X}}$, whence $\boldsymbol{B} = (\underline{\boldsymbol{X}}_{\mathrm{raw}}^\top \underline{\boldsymbol{X}}_{\mathrm{raw}})^{-1} \underline{\boldsymbol{X}}_{\mathrm{raw}}^\top \underline{\boldsymbol{X}}$.

We did not specify how the raw data have been centered. In our numerical experiments, we used the sample mean, but any estimator of location would be possible, provided that the scatter estimators $\hat{\Sigma}_0(\boldsymbol{x}_1^{\mathrm{raw}}, \ldots, \boldsymbol{x}_n^{\mathrm{raw}})$ and $\hat{\Sigma}(\boldsymbol{x}_1^{\mathrm{pre}}, \ldots, \boldsymbol{x}_n^{\mathrm{pre}})$ are invariant under translations of the input data. Note that $\hat{H}$ has this invariance property as well.

Global PP In our numerical experiments, we also tried a global version of PP. This consists of basic procedure 1 (pre-whitening) and basic procedure 3 (local PP) applied to the observations $\boldsymbol{x}_i^{\mathrm{pre}}$ instead of the observations $\boldsymbol{x}_i^{\mathrm{ics}}$. Of course, one could extend this by applying basic procedure 3 several times to the observations $\boldsymbol{V}_s^\top \boldsymbol{x}_i^{\mathrm{pre}}$, where $\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3, \ldots$ are independent random orthogonal matrices in $\mathbb{R}^{p\times p}$, independent from the data.

6 Numerical Examples

The subsequent numerical examples are similar to examples presented by Tyler et al. (2009), but with higher dimensions. We always started with the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_0$, and $\hat{\boldsymbol{\Sigma}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ was a one-step symmetrized M-estimator of scatter,

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) := C \sum_{1\le i<j\le n} \frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top}{(\nu + \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)^\gamma} \tag{7}$$

with some (irrelevant for us) scaling factor $C = C_{n,p,\nu,\gamma} > 0$ and parameters $\nu, \gamma > 0$. If $\nu > 0$, $\gamma = 1$ and $C = \nu + p$, then $\hat{\boldsymbol{\Sigma}}$ is a one-step approximation of the symmetrized maximum-likelihood estimator of a centered multivariate t distribution with ν degrees of freedom, see Kent and Tyler (1991). If $\nu = 0$ and $\gamma = 1$, then $\hat{\boldsymbol{\Sigma}}$ corresponds to the symmetrized distribution-free scatter estimator of Tyler (1987). Our numerical experiments indicate that it is worthwhile to try ν close to 0 and various parameters $\gamma > 1$, although $\hat{\boldsymbol{\Sigma}}$ does not correspond to a robust M-estimator of scatter then.

Another question is the choice of the bandwidth $h > 0$. The larger the bandwidth h, the smoother is the target function $\hat{H}$, whereas small bandwidths h result in many irrelevant local minima of $\hat{H}$. Our numerical experiments indicate that to detect clusters, values h between 0.3 and 0.5 work quite well. However, if the structure to be detected is on a rather small scale, e.g., data lying on several parallel hyperplanes, then one needs smaller bandwidths h (and probably exponents $\gamma > 1$) to detect such features.

For all subsequent examples, we simulated data sets of size $n = 500$ in different dimensions p, and we searched for interesting projections in dimension $d = 2$. The underlying distribution was chosen such that a scatter plot of the raw or pre-whitened data would not reveal the interesting structure. As to basic procedure 3, we considered all $p(p-1)/2$ standard projections $\boldsymbol{x} \mapsto \boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}$, where $1 \leq j < k \leq p$ and $\boldsymbol{U}_{jk} = [\boldsymbol{e}_j, \boldsymbol{e}_k, \boldsymbol{e}_{\ell(1)}, \ldots, \boldsymbol{e}_{\ell(p-2)}]$ with the elements $\ell(1) < \cdots < \ell(p-2)$ of $\{1, \ldots, p\} \setminus \{j, k\}$. Local PP was run with threshold $\delta_o = 10^{-11}$.

Example 1 We simulated data in dimension $p = 8$. After pre-whitening the data, we first tried a global PP with $h = 0.5$, which yields a value of 2.8610 for (1). The initial values $\hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}_1^{\mathrm{pre}}, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}_n^{\mathrm{pre}})$ ranged from 2.8251 to 2.8423. The minimal value was obtained for $(j, k) = (1, 5)$. Running a local PP with this starting point revealed three clusters, and the final value of $\hat{H}$ was 2.4735. This is shown in the upper panels of Fig. 1; the scatter plots show the projections before (left) and after (right) local PP.

Now, we applied the procedure we advocate in this manuscript. We performed ICS with $\nu = 0$ and $\gamma = 1$. Then, the values $\hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}_1^{\mathrm{ics}}, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}_n^{\mathrm{ics}})$ ranged from 2.7112 to 2.8395. The minimal value was obtained for $(j, k) = (1, 2)$, and the corresponding scatter plot indicates already some clustering, see the lower left panel of Fig. 1. Running local PP revealed essentially the same structure as the global PP, see the lower right panel.

In this example, global PP without ICS seems to be just as good as our three-stage procedure. But note that local PP starting from components 1 and 5 of the data $\boldsymbol{x}_i^{\mathrm{pre}}$ resulted in 24 iterations, whereas local PP starting from components 1 and 2 of the data $\boldsymbol{x}_i^{\mathrm{ics}}$ resulted in 10 iterations only.

Example 2 We simulated data in dimension $p = 16$. Again, we tried first a global PP without ICS, which means we started local PP from some standard projections of the data $\boldsymbol{x}_i^{\mathrm{pre}}$. With the same bandwidth $h = 0.5$ as in Example 1, the initial values $\hat{H}_{jk} := \hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}_1^{\mathrm{pre}}, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}_n^{\mathrm{pre}})$ ranged from 2.8133 to 2.8472. The minimal value was obtained for $(j, k) = (3, 15)$. The upper left panel of Fig. 2 shows that projection of the data $\boldsymbol{x}_i^{\mathrm{pre}}$, and starting local PP from this projection led to the scatter plot in the upper right panel with a value 2.6599 of $\hat{H}$. The number of iterations was 112.

Next, we tried other index pairs (j, k), ordered by the initial values $\hat{H}_{jk}$. The detected structures were similar for the next 12 pairs, but starting local PP from $(j, k) = (3, 6)$ revealed the true underlying structure, a uniform distribution on a two-dimensional circle with a value 2.3871 of $\hat{H}$, see the lower panels of Fig. 2. The number of iterations was 29.

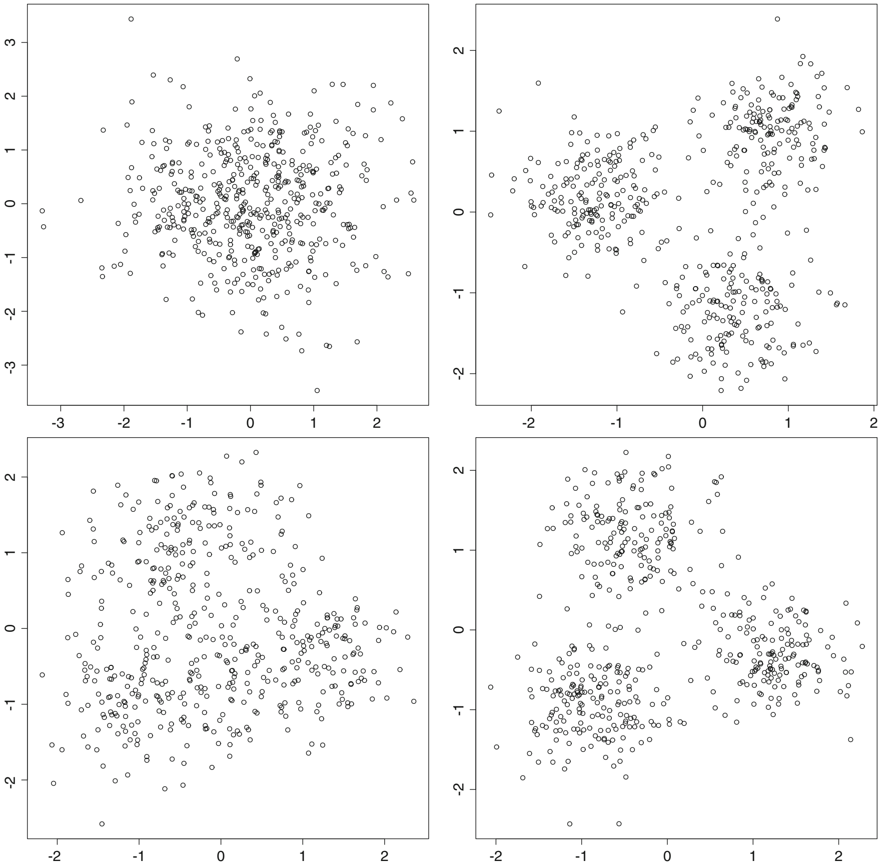


Fig. 1 PP for Example 1: two-dimensional projections of the pre-whitened data $\boldsymbol{x}_i$ before and after local PP (upper left and right panels) and of the preprocessed data $\boldsymbol{x}_i^{\mathrm{ics}}$ before and after local PP (lower left and right panels)

Finally, we applied the procedure advocated in this manuscript. After running ICS with $\nu = 0$ and $\gamma = 1$, the initial values $\hat{H}(\boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}_1^{\mathrm{ics}}, \ldots, \boldsymbol{\Pi}^\top \boldsymbol{U}_{jk}^\top \boldsymbol{x}_n^{\mathrm{ics}})$ ranged from 2.7651 to 2.8476. The minimal value was obtained with $(j, k) = (1, 2)$, and the corresponding scatter plot is shown in the upper left panel of Fig. 3. Running local PP with this starting point revealed quickly the underlying structure. The total number of iterations was 13, but already 4 iterations gave away the uniform distribution on the circle, see the lower right panel of Fig. 3.

This example illustrates nicely the benefit of using ICS as a means to find promising starting points for local PP.

Example 3 Our final example concerns a structure which is surprisingly difficult to detect even in moderate dimension. Here, the dimension is $p = 6$. Starting local PP starting from all $p(p-1)/2 = 15$ pairs of two components of the pre-whitened data

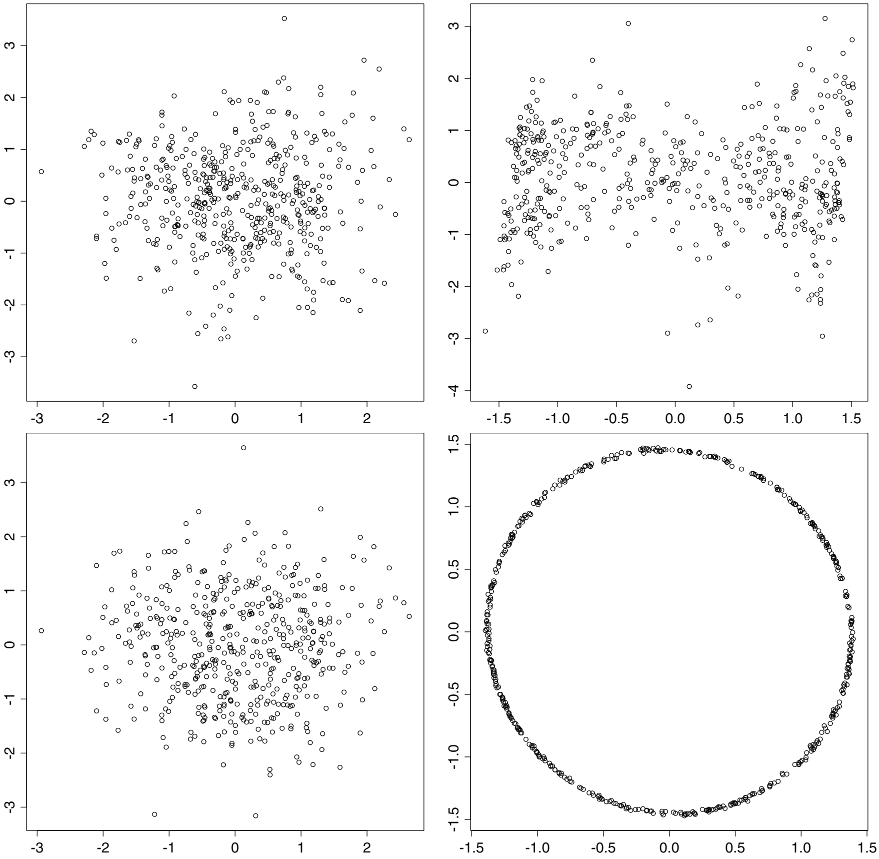


Fig. 2 PP for Example 2: two-dimensional projections of the data $\boldsymbol{x}_i^{\text{pre}}$ before (left panels) and after (right panels) local PP. The upper row corresponds to components $(j, k) = (3, 15)$ and the lower row to components $(j, k) = (3, 6)$

$\boldsymbol{x}_i^{\text{pre}}$ did not reveal anything, neither for $h = 0.5$ nor for $h = 0.2$. Also, our three-stage procedure with $\nu = 0$ and $\gamma = 1$ led nowhere. But with $\gamma = 4$, an interesting structure became visible for components $(j, k) = (1, 5)$ of the observations $\boldsymbol{x}_i^{\text{ics}}$, see the upper left panel of Fig. 4. The initial value of $\hat{H}$ was 2.6202. After 85 iterations of local PP, we ended up with the projection shown in the lower right panel, and the estimated entropy was 2.5797.

Acknowledgments Part of this research was supported by Swiss National Science Foundation. The authors are grateful to two referees for constructive comments.

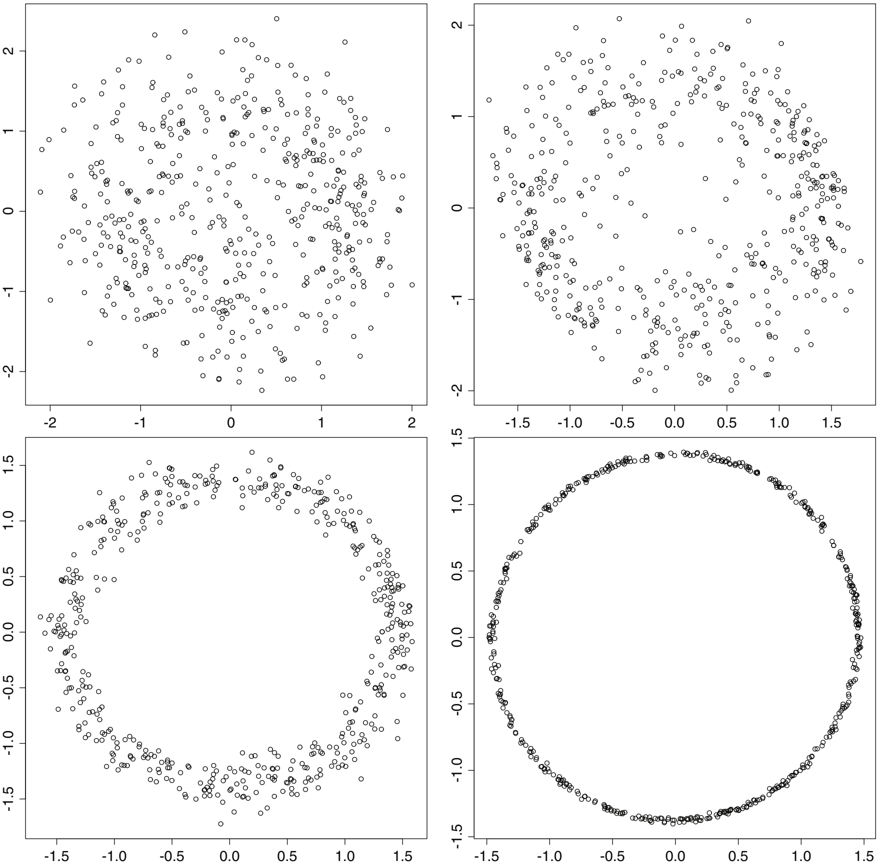


Fig. 3 PP for Example 2: Two-dimensional projections of the data $\boldsymbol{x}_i^{\mathrm{ics}}$, starting from components $(j, k) = (1, 2)$ (upper left panel) and after 1 (upper right panel), 2 (lower left panel) and 4 (lower right panel) iterations of local PP

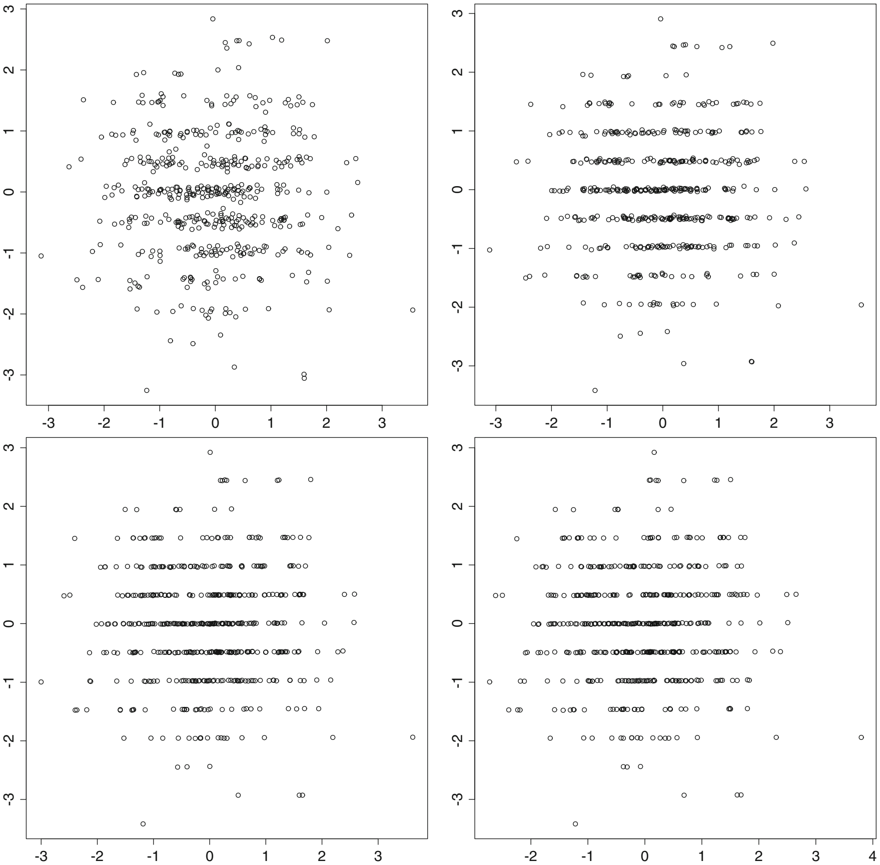


Fig. 4 PP for Example 3: Two-dimensional projections of the data $\boldsymbol{x}_i^{\text{ics}}$, starting from components $(j, k) = (1, 5)$ (upper left panel) and after 1 (upper right panel), 4 (lower left panel) and 85 (lower right panel) iterations of local PP

References

Diaconis, P., & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, *12*(3), 793–815.

Dümbgen, L., & Del Conte-Zerial, P. (2013). On low-dimensional projections of high-dimensional distributions. In M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii, & M. H. Maathuis (Eds.) *From probability to statistics and back: High-dimensional models and processes. A Festschrift in Honor of Jon Wellner*, vol. 9 of *IMS collections* (pp. 91–104). Hayward, California: Institute of Mathematical Statistics.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, *C-23*, 881–890.

Huber, P. J. (1985). Projection pursuit. *Annals of Statistics*, *13*(2), 435–525. With discussion.

Jones, M. C., & Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society: Series A*, *150*(1), 1–36. With discussion.

Kent, J. T., & Tyler, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Annals of Statistics*, *19*(4), 2102–2119.

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed.). Springer Series in Operations Research and Financial Engineering. Springer, New York.

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, *15*(1), 234–251.

Tyler, D. E., Critchley, F., Dümbgen, L., & Oja, H. (2009). Invariant coordinate selection (with discussion). *Journal of the Royal Statistical Society: Series B*, *71*(3), 549–592.

Directional Distributions and the Half-Angle Principle

John T. Kent

Abstract Angle halving, or alternatively the reverse operation of angle doubling, is a useful tool when studying directional distributions. It is especially useful on the circle where, in particular, it yields an identification between the wrapped Cauchy distribution and the angular central Gaussian distribution, as well as a matching of their parameterizations. The operation of angle halving can be extended to higher dimensions, but its effect on distributions is more complicated than on the circle. In all dimensions, angle halving provides a simple way to interpret stereographic projection from the sphere to Euclidean space.

Keywords Angular central Gaussian distribution · Gnomonic projection · Möbius transformation · Multivariate t distribution · Stereographic projection · Wrapped Cauchy distribution

1 Introduction

The wrapped Cauchy (WC) distribution on the circle is a remarkable distribution that appears in a wide variety of seemingly unrelated settings in probability and statistics. The angular central Gaussian (ACG) distribution is another important distribution in directional statistics. It was used by Tyler (1987a,b) to construct and study a robust estimator of a covariance matrix, or more generally a scatter matrix, for q-dimensional multivariate data. Hence, it is a pleasure to include this chapter in a volume dedicated to Dave Tyler's many contributions to statistical methodology.

As noted in Kent and Tyler (1988), the ACG distribution in $q = 2$ dimensions (i.e., on the circle) can be identified with the WC distribution after angle doubling. Equivalently, WC distribution can be identified with the ACG distribution after angle halving. Hence, algorithms to estimate the parameters of one distribution

J. T. Kent (✉)
University of Leeds, Leeds, UK
e-mail: j.t.kent@leeds.ac.uk

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_7

can be used with little change to estimate the parameters of the other distribution. Several algorithms to compute the maximum likelihood estimates based on the EM algorithm have been explored in Kent and Tyler (1988) and Kent et al. (1994). See also Arslan et al. (1995) for further discussion.

The current chapter extends the analysis as follows:

- To use angle halving on the circle to recast the Möbius transformation in terms of a rescaled linear transformation of the plane, a result that additionally allows us to match the parameterizations of the WC and ACG distributions
- To extend angle halving to higher dimensions and to show the connection between gnomonic projection and stereographic projection
- To note that the ACG distribution under gnomonic projection maps to a multivariate Cauchy distribution and to contrast it with the spherical Cauchy distribution of Kato and McCullagh (2020), which under stereographic projection maps to a multivariate t-distribution
- To summarize some further properties of the WC distribution

To set the scene for the main investigation of the work, recall some basic properties of the WC and ACG distributions on the circle S_1, with points on the circle represented by either an angle $0 \leq \theta < 2\pi$ or a unit vector $(\cos\theta, \sin\theta)^T$. The WC distribution, written WC(λ), has the probability density function (p.d.f.)

$$f_{\mathrm{WC}}(\theta; \lambda) = (2\pi)^{-1} \frac{1-\lambda^2}{1+\lambda^2 - 2\lambda\cos\theta}, \quad \theta \in S_1. \tag{1}$$

Here, $0 \leq |\lambda| < 1$ is a concentration parameter. The distribution has been centered to have its mode at $\theta = 0$ if $\lambda > 0$ and $\theta = \pi$ if $\lambda < 0$; it reduces to the uniform distribution if $\lambda = 0$.

The ACG distribution on S_1, written ACG(b), has probability density function (p.d.f.)

$$\begin{aligned} f_{\mathrm{ACG}}(\varphi; b) &= (2\pi)^{-1} b / \{b^2 \cos^2\varphi + \sin^2\varphi\} \\ &= \pi^{-1} b / \{b^2(1+\cos 2\varphi) + (1-\cos 2\varphi)\} \\ &= \pi^{-1} b / \{(1+b^2) - (1-b^2)\cos 2\varphi\}, \ \varphi \in S_1. \end{aligned} \tag{2}$$

Here, $0 < b < \infty$ is a concentration parameter. The density is antipodally symmetric, $f(\varphi) = f(\varphi + \pi)$. The distribution has been centered to have its modes at $\varphi = 0, \pi$ if $b < 1$ and $\varphi = \pm\pi/2$ if $b > 1$; it reduces to the uniform distribution if $b = 1$.

If

$$b = (1-\lambda)/(1+\lambda), \tag{3}$$

it can be checked that (2) is the same as (1) under the angle doubling relation $\theta = 2\varphi$. That is, if Φ is a random angle following the ACG(b) distribution and (3) holds, then $\Theta = 2\Phi$ is a random angle following the WC(λ) distribution. The relation (3) between b and λ will be assumed throughout the chapter.

The chapter is organized as follows. The basic transformations of the circle are defined and examined in Sect. 2. These transformations are used in Sect. 3 to obtain the ACG and WC distributions on the circle as transformations of the uniform distribution. Angle doubling is extended to the sphere in Sect. 4 and interpreted through two projections in Sect. 5. The spherical version of the ACG distribution is studied in Sect. 6 and a spherical analog of the WC distribution is constructed in Sect. 7. Section 8 gives a discussion of transformation groups on the sphere and shows how the ACG and spherical Cauchy distributions can be obtained as transformations of the uniform distribution. Finally, Sect. 9 summarizes some further derivations and motivations for the WC distribution on the circle.

For some standard background on directional distributions, see, e.g., Mardia and Jupp (2000) and Chikuse (2003). For basic results from multivariate analysis, see, e.g., Mardia et al. (1979). A fundamental reference is McCullagh (1996), which goes further than the current chapter in exploring how the family of WC distributions is closed under the group of Möbius transformations on the unit circle. See also Downs (2009) for a broader discussion of Möbius transformations. The use of the Möbius transformation in directional regression models was proposed in Downs and Mardia (2002) and Downs (2003).

2 Basic Operations on the Circle

A point on the circle can be written as an angle φ, where without loss of generality, $\varphi \in (-\pi, \pi]$. The point can also be expressed as a unit vector

$$\boldsymbol{x} = (x_1, x_2)^T = (\cos\varphi, \sin\varphi)^T = \pm(1, r)^T/\sqrt{1+r^2}, \quad r = \tan\varphi, \tag{4}$$

or as a complex number $x_1 + ix_2 = C(\boldsymbol{x})$. It is convenient to denote the mappings between vector and angular representations by

$$\varphi = \operatorname{Arg}(\boldsymbol{x}), \quad \boldsymbol{x} = \operatorname{vec}(\varphi). \tag{5}$$

For later use, note that the derivatives of the mappings between φ and $r = \tan\varphi$ are given by

$$dr/d\varphi = \sec^2\varphi = 1/\cos^2\varphi = 1 + r^2, \quad d\varphi/dr = 1/(1 + r^2). \tag{6}$$

Another important representation of an angle, where this time the angle is denoted θ, is in terms of the tangent of the half-angle, $s = \tan(\theta/2)$. Square both sides and use the double angle formulas to get

$$s^2 = \tan^2(\theta/2) = \frac{\sin^2(\theta/2)}{\cos^2(\theta/2)} = \frac{1-\cos\theta}{1+\cos\theta}, \tag{7}$$

which can be inverted to give

$$\cos\theta = \frac{1-s^2}{1+s^2},$$

so that $1 + \cos\theta = 2/(1 + s^2)$.

Throughout the chapter, we assume that θ and φ are related by the double angle condition, $\theta = 2\varphi$, so that $r = s$. However, it is helpful to use both notations r and s to emphasize that r is obtained from φ and s is obtained from θ.

Three important mappings from S_1 to itself are as follows.

(a) *Squaring*, denoted $D(\boldsymbol{x})$, where D stands for the doubling of the angle. In vector form the transformation is defined by

$$D(\boldsymbol{x}) = (x_1^2 - x_2^2, 2x_1x_2)^T, \quad \boldsymbol{x} \in S_1. \tag{8}$$

If $\boldsymbol{y} = D(\boldsymbol{x})$, then in complex arithmetic $y_1 + iy_2 = (x_1 + ix_2)^2$. Furthermore, if $\varphi = \mathrm{Arg}(\boldsymbol{x})$ and $\theta = \mathrm{Arg}(\boldsymbol{y})$ are the two points in angular coordinates, then $\theta = 2\varphi$. Hence, squaring is a two-to-one mapping of S_1 to itself.

(b) The *rescaled diagonal linear transformation*, denoted $L(\boldsymbol{x}; b)$, where $b > 0$ is a scaling constant. In vector form the transformation is defined by

$$L(\boldsymbol{x}; b) = (x_1, bx_2)^T / \sqrt{x_1^2 + b^2x_2^2}. \tag{9}$$

That is, the second component of $\boldsymbol{x}$ is scaled by a factor b, and the resulting vector is rescaled to be a unit vector. The rescaled diagonal linear transformation can also be described as follows. If $\boldsymbol{z} = L(\boldsymbol{x}; b)$, then

$$\tan \mathrm{Arg}(\boldsymbol{z}) = b \tan \mathrm{Arg}(\boldsymbol{x}). \tag{10}$$

(c) The *diagonal Möbius transformation*, denoted $M(\boldsymbol{y}; \lambda)$. Here, in vector form, the transformation is defined for $\lambda > 0$ by

$$M(\boldsymbol{y}; \lambda) = (2\lambda + (1+\lambda^2)y_1, (1-\lambda^2)y_2)^T/(1+\lambda^2+2\lambda y_1), \quad \boldsymbol{y} \in S_1. \tag{11}$$

If $\boldsymbol{w} = M(\boldsymbol{y}; \lambda)$ where $\mathrm{Arg}(\boldsymbol{y}) = \theta$ and $\mathrm{Arg}(\boldsymbol{w}) = \eta$, then θ and η are related by

$$\tan \eta/2 = b \tan \theta/2, \tag{12}$$

where b and λ are related by (3). That is, the Möbius transformation is the same as the rescaled diagonal linear transformation after the angles θ and η are divided by 2. The Möbius transformation is most commonly defined using complex arithmetic,

$$C(M(\boldsymbol{y}; \lambda)) = \frac{y_1 + iy_2 + \lambda}{\lambda(y_1 + iy_2) + 1}, \quad \boldsymbol{y} \in S_1, \tag{13}$$

where for our purposes here $0 < \lambda < 1$ is restricted to being real.

These transformations can be combined to give the following result, which is helpful to call the *fundamental diagonal Möbius identity*:

$$M(D(\boldsymbol{x}); \lambda) = D(L(\boldsymbol{x}; b)), \quad \boldsymbol{x} \in S_1, \tag{14}$$

where b and λ are related by (3). That is, a rescaled diagonal linear transformation followed by squaring is the same as squaring followed by a diagonal Möbius transformation.

The identity in (14) has been stated for diagonal case. However, it is possible to construct a more general version by allowing rotations before and after the relevant transformation. Let

$$\boldsymbol{R}_\alpha = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \tag{15}$$

denote a 2×2 rotation matrix by an angle α. Also, recall that any 2×2 matrix $\boldsymbol{B}$ with positive determinant can be written using the singular value decomposition as

$$\boldsymbol{B} = c\boldsymbol{R}_\alpha \mathrm{diag}(1, b)\boldsymbol{R}_\beta^T, \tag{16}$$

where $c > 0$ and $b > 0$. Note that if $\boldsymbol{x} = \mathrm{vec}(\varphi)$, then $\boldsymbol{R}_\beta^T \boldsymbol{x} = \mathrm{vec}(\varphi - \beta)$ and $D(\boldsymbol{R}_\beta^T \boldsymbol{x}) = \boldsymbol{R}_\beta^{2T} D(\boldsymbol{x}) = \mathrm{vec}(2(\varphi - \beta))$.

Define more general versions of the rescaled diagonal linear and Möbius transformations by

$$\begin{aligned} L(\boldsymbol{x}; \boldsymbol{B}) &= \boldsymbol{Bx}/||\boldsymbol{Bx}|| = \boldsymbol{R}_\alpha L(\boldsymbol{R}_\beta^T \boldsymbol{x}; b), \\ M(\boldsymbol{x}; \lambda, \exp(2i\alpha), \exp(2i\beta)) &= \boldsymbol{R}_\alpha^2 M(\boldsymbol{R}_\beta^{2T} \boldsymbol{x}; \lambda), \end{aligned} \tag{17}$$

where $||\boldsymbol{x}||^2 = \boldsymbol{x}^T\boldsymbol{x}$. In complex notation, the Möbius transformation becomes

$$M(\boldsymbol{y}; \lambda, \exp(2i\alpha), \exp(2i\beta)) = \exp(2i(\alpha - \beta))\frac{y_1 + iy_2 + \lambda\exp(2i\beta)}{\lambda\exp(-2i\beta)(y_1 + iy_2) + 1}.$$

Note that L now depends on the matrix $\boldsymbol{B}$ and M now depends on a real number and two complex numbers. The more general version of the fundamental Möbius identity becomes

$$M(D(\boldsymbol{x}); \lambda, \exp(2i\alpha), \exp(2i\beta)) = D(L(\boldsymbol{x}; \boldsymbol{B})). \tag{18}$$

3 Transformations of Distributions on the Circle

Let Φ^* follow a uniform distribution on the circle, with density $f(\varphi^*) = 1/(2\pi)$, $-\pi < \varphi^* < \pi$. Let $R^* = \tan\Phi^*$ and $\boldsymbol{X}^* = \text{vec}(\Phi^*)$ denote the corresponding tangent of the angle and the Euclidean coordinates. Consider the rescaled diagonal linear transformation $\boldsymbol{X} = L(\boldsymbol{X}^*; b)$, where $b > 0$, and let $\Phi = \text{Arg}(\boldsymbol{X})$ and $R = \tan(\Phi)$ denote the corresponding angular and tangent values.

The inverse transformation between $\boldsymbol{X}$ and $\boldsymbol{X}^*$ is $\boldsymbol{X}^* = L(\boldsymbol{X}; 1/b)$. Then, the p.d.f. of Φ is given by

$$\begin{aligned}\frac{1}{2\pi}\frac{d\varphi^*}{d\varphi} &= \frac{1}{2\pi}\frac{d\varphi^*}{dr^*}\frac{dr^*}{dr}\frac{dr}{d\varphi} \\ &= \frac{1}{2\pi}\frac{1}{1+r^{*2}}b^{-1}(1+r^2) \\ &= \frac{1}{2\pi b}\frac{\cos^2\varphi}{\cos^2\varphi + b^{-2}\sin^2\varphi}\frac{1}{\cos^2\varphi} \\ &= \frac{b}{2\pi}\frac{1}{b^2\cos^2\varphi + \sin^2\varphi} = f_{\text{ACG}}(\varphi; b),\end{aligned} \tag{19}$$

where we have used the fact that $r^{*2} = b^{-2}r^2 = b^{-2}\sin^2\varphi/\cos^2\varphi$, and $1/(1+r^2) = \cos^2\varphi$. In other words, Φ follows the ACG(b) distribution.

If Φ^* follows a uniform distribution, then so does $\Theta^* = 2\Phi^*$. Hence,

$$\Theta = \text{Arg}(M(\text{vec}(\Theta^*), \lambda)) = 2\Phi = 2\text{Arg}(L(\text{vec}(\Phi^*), b))$$

has p.d.f. (19) as a function of φ (the factor 1/2 from the Jacobian $d\varphi^*/d\theta^*$ cancels the factor 2 which arises since the mapping from φ^* to θ^* is two to one). After writing the p.d.f. in terms of θ, the wrapped Cauchy density $f_{\text{WC}}(\theta; \lambda)$ in (1) is obtained, where λ is related to b by (3).

In particular, if $0 < \lambda < 1$, i.e., $0 < b < 1$, the diagonal Möbius mapping $\boldsymbol{Y} = M(\boldsymbol{Y}^*, \lambda)$ pulls probability mass toward the direction $\theta = 0$; similarly, the rescaled diagonal linear mapping $\boldsymbol{X} = L(\boldsymbol{X}^*; b)$ pulls probability mass toward the directions $\varphi = 0$ and π. Hence, the WC distribution for $\boldsymbol{Y}$ has a mode in the zero direction, and the ACG distribution for $\boldsymbol{X}$ has its modes in the directions 0 and π.

In summary, both the ACG and WC distributions can be obtained from suitable transformations of the uniform distribution. For simplicity, attention has been focused on the centered distributions in this section, but rotations of the modal direction can be easily included.

4 Basic Operations on the Sphere

To deal with higher dimensional spheres, more notation is needed. Let $S_{q-1} = \{\boldsymbol{x} \in \mathbb{R}^q : \boldsymbol{x}^T\boldsymbol{x} = 1\}$ denote the unit sphere in $\mathbb{R}^q$, $q \geq 2$, in a unit vector notation. The surface area of S_{q-1} is given by

$$\pi_q = 2\pi^{q/2}/\,\Gamma(q/2). \tag{20}$$

A point $\boldsymbol{x} \in S_{q-1}$ can be written in polar form about the north pole $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$ as

$$\boldsymbol{x} = \pm\begin{bmatrix} \cos\varphi \\ \sin\varphi\, \boldsymbol{u} \end{bmatrix}, \quad 0 \leq \varphi \leq \pi, \tag{21}$$

where $\boldsymbol{u}$ is a unit $(q-1)$-dimensional vector. If $q = 2$, then $u = \pm 1$ is just a scalar.

Using the polar representation (21), the surface measure on S_{q-1}, written $[d\boldsymbol{x}]$, say, can be written recursively as

$$[d\boldsymbol{x}] = \sin^{q-2}\varphi\, d\varphi\, [d\boldsymbol{u}]. \tag{22}$$

When $q = 2$, the formula simplifies to $[d\boldsymbol{x}] = d\varphi$. However, note (2) used a slightly different convention for φ; the scalar $u = \pm 1$ was not present and the angle φ was allowed to range through the whole circle, $-\pi < \varphi \leq \pi$.

For all dimensions $q \geq 2$, changing φ to $\pi - \varphi$ and $\boldsymbol{u}$ to $-\boldsymbol{u}$ changes $\boldsymbol{x}$ to $-\boldsymbol{x}$. Hence, when studying antipodally symmetric p.d.f.s, it is sufficient to restrict φ to the range $0 \leq \varphi < \pi/2$.

Let $\boldsymbol{y}$ be another point in S_{q-1} with polar representation

$$\boldsymbol{y} = \begin{bmatrix} \cos\theta \\ \sin\theta\, \boldsymbol{u} \end{bmatrix}. \tag{23}$$

If $\boldsymbol{u}$ is the same as in (21) and $\theta = 2\varphi$, then $\boldsymbol{y}$ can be said to be obtained from $\boldsymbol{x}$ by *doubling the angle*, where “angle” here means the colatitude φ. Write

$$\boldsymbol{y} = D_q(\boldsymbol{x}) \tag{24}$$

by analogy with the corresponding operation (8) on the circle.

In dimensions $q > 2$, the concept of doubling the angle is less general than the squaring operation on the circle ($q = 2$) given in (8). In particular, when $q > 2$, the operation of doubling the angle has a Jacobian which depends on the choice of north pole; see (32).

For use below, consider the following linear function of a q-dimensional unit vector $\boldsymbol{y}$:

$$P(\boldsymbol{y}) = P(\boldsymbol{y}; \lambda, \boldsymbol{\mu}_0) = 1 + \lambda^2 - 2\lambda\, \boldsymbol{y}^T \boldsymbol{\mu}_0, \tag{25}$$

and partition the unit vector $\boldsymbol{\mu}_0 = (\mu_1, \boldsymbol{\mu}_2^T)^T$ in terms of a scalar and a $(q-1)$-vector. Using (21) and (23), $P(\boldsymbol{y})$ can be rewritten as a quadratic function of $\boldsymbol{x}$ as follows:

$$\begin{aligned}
P(\boldsymbol{y}) &= 1 + \lambda^2 - 2\lambda\, \boldsymbol{y}^T \boldsymbol{\mu}_0 \\
&= (1 + \lambda^2) - 2\lambda\mu_1 \cos\theta - 2\lambda(\boldsymbol{\mu}_2^T \boldsymbol{u}) \sin\theta \\
&= (1 + \lambda^2)(\cos^2\varphi + \sin^2\varphi) - 2\lambda\mu_1(\cos^2\varphi - \sin^2\varphi) - 4\lambda(\boldsymbol{\mu}_2^T \boldsymbol{u}) \sin\varphi\cos\varphi \\
&= (1 + \lambda^2)(x_1^2 + \boldsymbol{x}_2^T \boldsymbol{x}_2) - 2\lambda\mu_1(x_1^2 - \boldsymbol{x}_2^T \boldsymbol{x}_2) - 4\lambda(\boldsymbol{\mu}_2^T \boldsymbol{x}_2)x_1 \\
&= (1 + \lambda^2 - 2\lambda\mu_1)x_1^2 + (1 + \lambda^2 + 2\lambda\mu_1)\boldsymbol{x}_2^T \boldsymbol{x}_2 - 4\lambda(\boldsymbol{\mu}_2^T \boldsymbol{x}_2)x_1 \\
&= Q(\boldsymbol{x}), \text{ say,}
\end{aligned} \tag{26}$$

a homogeneous quadratic form $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$ with matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 + \lambda^2 - 2\lambda\mu_1 & -2\lambda\boldsymbol{\mu}_2^T \\ -2\lambda\boldsymbol{\mu}_2 & (1 + \lambda^2 + 2\lambda\mu_1)I_{q-1} \end{bmatrix}. \tag{27}$$

Since $\boldsymbol{\mu}_0^T \boldsymbol{\mu}_0 = 1$ and $|\lambda| < 1$, it can be checked that $\boldsymbol{A}$ is positive definite.

5 Projections from the Sphere to Euclidean Space

In this section, we look at two standard tangent projections from the sphere to Euclidean space. It is convenient to set up the definitions and notation for all dimensions $q \geq 2$. We can then specialize to the case $q = 2$ and describe how the projections are connected to the transformations of Sect. 3.

The first is *gnomonic projection*, taking the open hemisphere $H_{q-1} = \{\boldsymbol{x} \in S_{q-1} : x_1 > 0\}$ to $\mathbb{R}^{q-1}$. If $\boldsymbol{x}$ is a unit q-vector in the open hemisphere, it can be written in the form (21) where $0 \le \varphi < \pi/2$ and $\boldsymbol{u}$ is a unit $(q-1)$-vector. As in (4), let $r = \tan\varphi$. Then, the gnomonic projection of $\boldsymbol{x}$ is defined by

$$\boldsymbol{v} = r\,\boldsymbol{u} = \frac{\sin\varphi}{\cos\varphi}\boldsymbol{u} = \frac{\sin\varphi}{x_1}\boldsymbol{u}. \tag{28}$$

The second is stereographic projection, taking the sphere S_{q-1}, minus the point at $-\boldsymbol{e}_1$, to $\mathbb{R}^{q-1}$. If $\boldsymbol{y} \in S_{q-1}$ is a unit vector other than $-\boldsymbol{e}_1$, write it in the form (23), where $-\pi < \theta < \pi$. As in (7), let $s = \tan(\theta/2)$. Then, the *stereographic projection* of $\boldsymbol{y}$ is defined by

$$\boldsymbol{w} = s\boldsymbol{u} = \frac{\sin(\theta/2)}{\cos(\theta/2)}\boldsymbol{u} = \frac{\sin\theta}{1+y_1}\,\boldsymbol{u} \tag{29}$$

since $\sin\theta = 2\sin(\theta/2)\cos(\theta/2)$ and $1 + y_1 = 1 + \cos\theta = 2\cos^2(\theta/2)$.

If $\boldsymbol{y}$ is obtained from $\boldsymbol{x}$ by angle doubling, then the two projections are identical. That is, if $\theta = 2\varphi$, then $r = s$ and $\boldsymbol{v} = \boldsymbol{w}$. However, the mapping of the uniform measure on the sphere to Euclidean space is different for the two projections. For gnomonic projection, the polar coordinate representation $\boldsymbol{v} = r\,\boldsymbol{u}$ states that r is the radial part of $\boldsymbol{v}$ so that the Lebesgue measure in the tangent space $\mathbb{R}^{q-1}$ is related to the uniform measure on the sphere by

$$\begin{aligned} d\boldsymbol{v} &= r^{q-2}dr\,[d\boldsymbol{u}] \\ &= (\sin\varphi/\cos\varphi)^{q-2}(dr/d\varphi)\,d\varphi\,[d\boldsymbol{u}] \\ &= \cos^{-q}\varphi\{\sin^{q-2}\varphi\,d\varphi\,[d\boldsymbol{u}]\} \\ &= \cos^{-q}\varphi\,[d\boldsymbol{x}], \end{aligned} \tag{30}$$

using (22) and $dr/d\varphi = \sec^2\varphi$. On the other hand, for stereographic projection, the polar coordinate representation $\boldsymbol{w} = s\,\boldsymbol{u}$ implies

$$\begin{aligned} d\boldsymbol{w} &= s^{q-2}ds\,[d\boldsymbol{u}] \\ &= \{\sin(\theta/2)/\cos(\theta/2)\}^{q-2}\,(ds/d\theta)\,d\theta\,[d\boldsymbol{u}] \\ &= \frac{1}{2}\{\sin(\theta/2)/\cos(\theta/2)\}^{q-2}\{\cos(\theta/2)\}^{-2}\sin^{-(q-2)}\theta\{\sin^{q-2}\theta\,d\theta\,[d\boldsymbol{u}]\} \\ &= \left(\frac{1}{2}\right)^{q-1}\cos^{-2(q-1)}(\theta/2)[d\boldsymbol{y}] \end{aligned} \tag{31}$$

since $ds/d\theta = (1/2)\sec^2(\theta/2)$ and $\sin\theta = 2\sin(\theta/2)\cos(\theta/2)$. Except on the circle $q = 2$, the two differentials involve different powers of $\cos(\theta/2) = \cos\varphi$.

Since $d\boldsymbol{v} = d\boldsymbol{w}$ both represent the Lebesgue measure in $\mathbb{R}^{q-1}$, (30) and (31) can be combined to describe the effect of angle doubling on the sphere,

$$[d\boldsymbol{y}] = 2^{q-1} \cos^{q-2} \varphi \, [d\boldsymbol{x}]. \tag{32}$$

The reason for the cosine factor is straightforward to understand intuitively. For example, consider the case $q = 3$ corresponding to the usual sphere. For a constant value of a colatitude φ, the longitude can range between 0 and 2π, and the corresponding points on the sphere lie on a small circle of circumference $2\pi \sin\varphi$. If φ is near $\pi/2$, the corresponding small circle for $\boldsymbol{x}$ is near the equator, a circle with circumference 2π. However, the corresponding value of $\theta = 2\varphi$ is near π, and the corresponding small circle for $\boldsymbol{y}$ lies near the south pole with circumference close to 0. The cosine factor in (32) accounts for this change in circumference.

Figure 1 illustrates the two projections on the circle, where $\theta = 2\varphi$. The gnomonic projection of φ is obtained by following the ray from the origin O through $(\cos\varphi, \sin\varphi)^T$ to the vertical line tangent to the circle at B. The stereographic projection of θ is obtained by following the ray from A through $(\cos\theta, \sin\theta)^T$ to the same vertical line and dividing the result by 2. Note that the stereographic projection of θ is the same as the gnomonic projection of φ.

The diagonal transformations on the circle in Sect. 2 can be as given simple interpretations in terms of these projections. First, the rescaled diagonal linear transformation of a unit vector vec(φ) can be obtained by applying the following three transformations:

(a) Gnomonic projection, $\varphi \to \tan\varphi$
(b) Scale change, $\tan\varphi \to b\tan\varphi$
(c) Inverse gnomonic projection, $b\tan\varphi \to \text{atan}(b\tan\varphi)$

Similarly, the diagonal Möbius transformation of a unit vector vec(θ) can be obtained by applying the following three transformations:

(a) Stereographic projection, $\theta \to \tan(\theta/2)$
(b) Scale change, $\tan(\theta/2) \to b\tan(\theta/2)$
(c) Inverse stereographic projection, $b\tan(\theta/2) \to 2\text{atan}\{b\tan(\theta/2)\}$

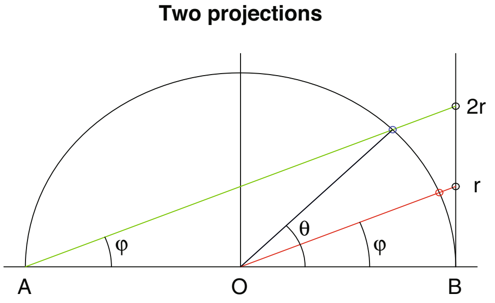


Fig. 1 Two projections, gnomonic and stereographic, from the circle to the vertical line tangent to the circle at point B. If $\varphi = \theta/2$, then $r = \tan\varphi = \tan\theta/2$ is both the gnomonic projection of φ and the stereographic projection of θ

If $\theta = 2\varphi$, these two mappings are essentially the same as each other, thus confirming the fundamental Möbius identity (14).

6 The ACG Distribution on the Sphere

This section takes a closer look at the ACG distribution on the sphere $S_{q-1},\ q \geq 2$, and in particular derives its behavior under gnomonic projection. First, it is useful to recall some results about quadratic forms.

6.1 Review of Quadratic Forms in the Multivariate Normal Distribution

Let $\boldsymbol{x} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$ be a q-dimensional vector partitioned into two parts of dimensions q_1 and q_2. Similarly, partition a $q \times q$ positive definite matrix as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

If $\boldsymbol{x}$ follows a multivariate normal distribution, $\boldsymbol{x} \sim N_q(\boldsymbol{0}, \boldsymbol{\Sigma})$, then $\boldsymbol{x}_1 \sim N_{q_1}(\boldsymbol{0}, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{x}_2|\boldsymbol{x}_1 \sim N_{q_2}(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}_1, \boldsymbol{\Sigma}_{22.1})$ (e.g., Mardia et al. 1979, p. 63), where $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. Writing the joint density of $\boldsymbol{x}$ as a product of a marginal and a conditional density, $f(\boldsymbol{x}) = f_1(\boldsymbol{x}_1)f(\boldsymbol{x}_2|\boldsymbol{x}_1)$ yields an identity for quadratic forms,

$$Q = Q_1 + Q_{2.1}, \tag{33}$$

where

$$\begin{aligned} Q &= \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \\ Q_1 &= \boldsymbol{x}_1^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}_1, \\ Q_{2.1} &= (\boldsymbol{x}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}_1)^T\boldsymbol{\Sigma}_{22.1}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{x}_1). \end{aligned} \tag{34}$$

If $q_1 = 1,\ q_2 = q - 1$, then $\boldsymbol{x}_1 = x_1$ is a scalar, $\boldsymbol{\Sigma}_{11} = \boldsymbol{\sigma}_{11}$ is a scalar, and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\sigma}_{21}$ is a vector. This case will be useful in the next section when studying gnomonic projection.

6.2 Basic Properties of the ACG Distribution

This section reviews some basic facts about the ACG distribution. Let $\boldsymbol{\Sigma}$ be a symmetric $q \times q$ positive definite matrix with inverse $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. The ACG distribution on S_{q-1} is defined by the density (with respect to the uniform measure on S_{q-1}) by

$$f_{\mathrm{ACG}}(\boldsymbol{x}) = f_{\mathrm{ACG}}(\boldsymbol{x}; \boldsymbol{\Omega}) = \pi_q^{-1} |\boldsymbol{\Omega}|^{1/2} / (\boldsymbol{x}^T \boldsymbol{\Omega} \boldsymbol{x})^{q/2}, \tag{35}$$

where π_q is given in (20). The parameter $\boldsymbol{\Omega}$ is defined up to a multiplicative scalar. If $\boldsymbol{\Omega}$ has spectral decomposition $\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Gamma}$ is an orthogonal matrix containing the eigenvectors and $\boldsymbol{\Delta}$ is a diagonal matrix containing the eigenvalues, then it is possible to separate out the orientation and the concentration parts of the model. The ACG distribution is antipodally symmetric, $f_{\mathrm{ACG}}(\boldsymbol{x}) = f_{\mathrm{ACG}}(-\boldsymbol{x})$.

If $q = 2$ and $\boldsymbol{\Omega} = \mathrm{diag}(b^2, 1)$ is a diagonal matrix with $0 < b < 1$, then the density in polar coordinates reduces to (2). A similar expansion can be carried out in higher dimensions $q > 2$. Suppose $\boldsymbol{\Omega}$ is partitioned as

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \boldsymbol{\omega}_{21}^T \\ \boldsymbol{\omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix},$$

and partition a unit vector $\boldsymbol{x} \in S_{q-1}$ as in (21). The quadratic form becomes

$$\boldsymbol{x}^T \boldsymbol{\Omega} \boldsymbol{x} = \omega_{11} \cos^2 \varphi + 2 \sin \varphi \cos \varphi \, (\boldsymbol{\omega}_{21}^T \boldsymbol{u}) + \sin^2 \varphi \, \boldsymbol{u}^T \boldsymbol{\Omega}_{22} \boldsymbol{u}. \tag{36}$$

If, in addition, $\boldsymbol{\omega}_{21} = \boldsymbol{0}$, then ω_{11} is an eigenvalue. If ω_{11} is the smallest eigenvalue, then the density has its modes at $\varphi = 0, \pi$.

6.3 ACG Distribution Under Gnomonic Projection

Next, consider gnomonic projection of the ACG distribution. Equations (33) and (34) can be used to show that the ACG distribution on the sphere is transformed to a multivariate Cauchy distribution in $\mathbb{R}^{q-1}$. To verify this result, recall the identities in (4). Then, the quadratic form $Q = Q(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}$ in (36), after dividing by $\cos^2 \varphi = 1/(1 + r^2)$, becomes

$$\begin{aligned}(1 + r^2) Q &= \omega_{11} + 2 \boldsymbol{v}^T \boldsymbol{\omega}_{21} + \boldsymbol{v}^T \boldsymbol{\Omega}_{22} \boldsymbol{v} \\ &= \omega_{11} - \boldsymbol{\omega}_{21}^T \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21} + (\boldsymbol{v} + \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21})^T \boldsymbol{\Omega}_{22} (\boldsymbol{v} + \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21}) \\ &= \sigma_{11}^{-1} + (\boldsymbol{v} - \boldsymbol{\sigma}_{21}/\sigma_{11})^T \boldsymbol{\Sigma}_{22.1}^{-1} (\boldsymbol{v} - \boldsymbol{\sigma}_{21}/\sigma_{11}),\end{aligned} \tag{37}$$

using the identities $\boldsymbol{\sigma}_{21}/\sigma_{11} = -\boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21}$, $\sigma_{11}^{-1} = \omega_{11} - \boldsymbol{\omega}_{21}^T \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\omega}_{21}$ and $\boldsymbol{\Sigma}_{22.1}^{-1} = \boldsymbol{\Omega}_{22}$ for the inverse of a partitioned matrix (e.g., Mardia et al. 1979, p. 459). Without loss of generality, we can rescale $\boldsymbol{\Sigma}$ so that $\sigma_{11} = 1$.

The $(q-1)$-dimensional multivariate t-distribution, with location parameter $\boldsymbol{\mu}$, scatter matrix $\boldsymbol{B}$, and degrees of freedom $\kappa > 0$, written $t_{q-1}(\boldsymbol{\mu}, \boldsymbol{B}, \kappa)$, has density proportional to

$$f(\boldsymbol{v}) \propto \{1 + \kappa^{-1}(\boldsymbol{v} - \boldsymbol{\mu})^T \boldsymbol{B}^{-1}(\boldsymbol{v} - \boldsymbol{\mu})\}^{-(q-1+\kappa)/2} \tag{38}$$

(e.g., Mardia et al. 1979, p. 57). If $\kappa = 1$, the distribution is known as the multivariate Cauchy distribution.

Using (30), (35), and (37) to give the p.d.f. of the ACG($\boldsymbol{\Sigma}$) distribution after gnomonic projection yields

$$f_{\text{ACG,gnomonic}}(\boldsymbol{v}) \propto Q^{-q/2} \cos^q \varphi = Q^{-q/2}(1 + r^2)^{-q/2},$$

with respect to the Lebesgue measure $d\boldsymbol{v}$ in the tangent plane, which is the same as (38) with $\kappa = 1$. That is, the gnomonic projection follows a multivariate Cauchy distribution $t_{q-1}(\boldsymbol{\sigma}_{21}, \boldsymbol{\Sigma}_{22.1}^{-1}, 1)$.

7 The Spherical Cauchy Distribution

Kato and McCullagh (2020) have defined the *spherical Cauchy (SC) distribution* on S_{q-1} to have the p.d.f.

$$f_{\text{SC}}(\boldsymbol{y}; \lambda, \boldsymbol{\mu}_0) = \pi_q^{-1} \left\{ \frac{1 - \lambda^2}{P(\boldsymbol{y}; \lambda, \boldsymbol{\mu}_0)} \right\}^{q-1}, \quad \boldsymbol{y} \in S_{q-1}, \tag{39}$$

where π_q is given in (20) and $P(\boldsymbol{y}; \lambda, \boldsymbol{\mu}_0)$ is given in (25). Here, $0 \leq \lambda < 1$ is a measure of concentration and $\boldsymbol{\mu}_0$ is a unit q-vector representing the modal direction. When $q = 2$, the SC distribution reduces to the WC distribution (1).

Write $\boldsymbol{\mu}_0 = (\mu_1, \boldsymbol{\mu}_2^T)^T$, where μ_1 is a scalar and $\boldsymbol{\mu}_2$ is a $(q-1)$-vector and $\mu_1^2 + \boldsymbol{\mu}_2^T \boldsymbol{\mu}_2 = 1$. Then, similarly to the expansion in (26), the quantity $P(\boldsymbol{y})$ in (25) can be written in stereographic coordinates $\boldsymbol{v}$ as

$$\begin{aligned} P = P(\boldsymbol{y}) &= 1 + \lambda^2 - 2\lambda \boldsymbol{y}^T \boldsymbol{\mu}_0 \\ &= (1 + \lambda^2) - 2\lambda(\mu_1 \cos\theta + \boldsymbol{u}^T \boldsymbol{\mu}_2 \sin\theta) \\ &= \frac{1}{1 + r^2}\{(1 + \lambda^2)(1 + r^2) - 2\lambda[(1 - r^2)\mu_1 + 2\boldsymbol{v}^T \boldsymbol{\mu}_2)]\} \\ &= \frac{1}{1 + r^2}\{\gamma + \delta r^2 - 4\lambda \boldsymbol{v}^T \boldsymbol{\mu}_2\} \\ &= \frac{1}{1 + r^2}\{\gamma - (4\lambda^2/\delta)\boldsymbol{\mu}_2^T \boldsymbol{\mu}_2 + \delta(\boldsymbol{v} - (2\lambda/\delta)\boldsymbol{\mu}_2)^T(\boldsymbol{v} - (2\lambda/\delta)\boldsymbol{\mu}_2)\} \\ &= \frac{\gamma^*}{1 + r^2}\{1 + (\boldsymbol{v} - \boldsymbol{m})^T(\boldsymbol{v} - \boldsymbol{m})/\sigma^2\}, \end{aligned} \tag{40}$$

where in the fourth line

$$\gamma = 1 + \lambda^2 - 2\lambda\mu_1, \quad \delta = 1 + \lambda^2 + 2\lambda\mu_1$$

and in the final line

$$\gamma^* = \gamma - (4\lambda^2/\delta)\boldsymbol{\mu}_2^T\boldsymbol{\mu}_2 = (1-\lambda^2)/\delta, \quad \boldsymbol{m} = (2\lambda/\delta)\,\boldsymbol{\mu}_2, \quad \sigma = (1-\lambda^2)/\delta.$$

In addition, the identities $\boldsymbol{v}^T\boldsymbol{v} = r^2\boldsymbol{u}^T\boldsymbol{u} = r^2$, $\cos^2\varphi = 1/(1+r^2)$, $\cos\theta = (1-r^2)/(1+r^2)$, and $\sin\theta = 2\sin\varphi\cos\varphi = (2\tan\varphi)/(1+r^2)$ have been used.

Using the change of variables formula (31), the distribution of the stereographic projection of $\boldsymbol{y}$ has density

$$f_{\gamma,\text{stereo}}(\boldsymbol{v}) \propto P^{-(q-1)}\cos^{2(q-1)}(\theta/2) = \{(1+r^2)P\}^{-(q-1)},$$

which as a function of $\boldsymbol{v}$ can be identified with the density of the multivariate t-distribution $t_{q-1}(\boldsymbol{m}, (q-1)^{-1}\sigma^2\boldsymbol{I}_{q-1}, q-1)$ with $\kappa = q-1$ degrees of freedom. Note that the identification is valid even if $\boldsymbol{\mu}_0 \neq \boldsymbol{e}_1$, i.e., even if the mode of the SC distribution does not lie in the direction of the first coordinate axis. This result was proved in Kato and McCullagh (2020); see also McCullagh (1996) for a deeper study of the circular case.

Note that the factor $(1+r^2)^{-(q-1)}$ in the density has the right power to combine with $P^{-(q-1)}$ in the density. This property explains why the SC distribution was defined by raising P to the power $-(q-1)$, and not some other power, in (39).

When $q \neq 2$, the SC distribution can never be identified with the ACG distribution under angle doubling. In particular, the gnomonic projections of an ACG distribution follows a multivariate Cauchy distribution (i.e., a multivariate t-distribution with 1 degree of freedom). In contrast, the stereographic projection of an SC distribution follows a multivariate t-distribution with $q-1$ degrees of freedom.

Finally, an anonymous referee has noted that there is another definition of a spherical Cauchy distribution as an exit distribution in diffusion theory. Consider a q-dimensional Brownian motion starting at the point $\lambda\boldsymbol{\mu}_0$, where $0 \le \lambda < 1$ and $\boldsymbol{\mu}_0$ is a unit vector. The position of the Brownian motion when it first hits the sphere S_{q-1} has the density

$$f_{\text{BMSC}}(\boldsymbol{y}; \lambda, \boldsymbol{\mu}_0) = \pi_q^{-1}\frac{1-\lambda^2}{P(\boldsymbol{y}; \lambda, \boldsymbol{\mu}_0)^{q/2}}, \quad \boldsymbol{y} \in S_{q-1}, \tag{41}$$

(e.g., Durrett 1984, Section 1.10) where $P(\boldsymbol{y}; \lambda, \boldsymbol{\mu}_0)$ is given in (25) and where the subscript BMSC stands for *Brownian motion spherical Cauchy*. Except on the circle, $q = 2$, (41) is different from (39) and its stereographic projection cannot be identified with any t-distribution.

8 Transformation Groups on the Sphere

This section extends some results involving the rescaled linear and Möbius transformations from the circle to higher dimensional spheres $S_{q-1},\ q > 2$.

Start with the *general rescaled linear transformations* of the form

$$\boldsymbol{x} \to \boldsymbol{Bx}/||\boldsymbol{Bx}||, \quad \boldsymbol{x} \in S_{q-1}, \tag{42}$$

where $\boldsymbol{B}(q \times q)$ is nonsingular with positive determinant. It is easy to see that these transformations form a group under composition where the group operation corresponds to matrix multiplication. This group can be used to facilitate simulation. For example, if $\boldsymbol{x}$ is uniformly distributed on S_{q-1} and $\boldsymbol{B} = \boldsymbol{\Omega}^{-1/2} = \boldsymbol{\Sigma}^{1/2}$, then $\boldsymbol{Bx}/||\boldsymbol{Bx}||$ follows the ACG distribution in (35).

Of special interest are the *rescaled diagonal linear transformations* for which $\boldsymbol{B}$ is assumed to have the form

$$\boldsymbol{B} = \mathrm{diag}(1, b\boldsymbol{I}_{q-1}), \quad b > 0. \tag{43}$$

That is, the scaling factor for the first coordinate direction is different from the common scaling factor for the other coordinate directions. If $\boldsymbol{x} = (\cos\varphi, \sin\varphi\, \boldsymbol{u}^T)^T$ as in (21), the rescaled diagonal linear transformation of $\boldsymbol{x}$ can be written as

$$L_q(\boldsymbol{x}; b) = (\cos\varphi^*, \sin\varphi^*\, \boldsymbol{u}^T)^T, \text{ where } \mathrm{vec}(\varphi^*) = L(\mathrm{vec}(\varphi); b)$$

in terms of the corresponding transformation L on the circle in (9).

It is also possible to extend Möbius transformations to higher dimensions. In this case, it is simplest to start with the *diagonal Möbius transformations*. If $\boldsymbol{y} = (\cos\theta, \sin\theta\, \boldsymbol{u}^T)^T$ as in (23), the diagonal Möbius transformation of $\boldsymbol{y}$ can be written as

$$M_q(\boldsymbol{y}; \lambda) = (\cos\theta^*, \sin\theta^*\, \boldsymbol{u}^T)^T, \text{ where } \mathrm{vec}(\theta^*) = M(\mathrm{vec}(\theta); \lambda)$$

in terms of the corresponding transformation M on the circle in (11).

This transformation can be used to facilitate simulation. If $\boldsymbol{y}$ is uniformly distributed on S_{q-1}, then $M_q(\boldsymbol{y}; \lambda)$ follows the spherical Cauchy distribution (39). See also Downs (2009), who used this property to motivate the definition of the spherical Cauchy distribution.

It is also possible to define a *general Möbius transformation* consisting of three operations: (a) a rotation, followed by (b) a diagonal Möbius transformation, followed by (c) another rotation. Although it is not immediately obvious, the set of general Möbius transformations forms a group under composition.

The fundamental diagonal Möbius identity (14) on the circle between the rescaled diagonal linear transformations and the diagonal Möbius transformations carries over with a little change. It becomes

$$M_q(D_q(\boldsymbol{x}); \lambda) = D_q(L_q(\boldsymbol{x}); b), \quad \boldsymbol{x} \in S_{q-1}, \tag{44}$$

where D_q is defined in (24). Furthermore, the interpretation of this identity in terms of gnomonic and stereographic projections given at the end of Sect. 5 carries over immediately to higher dimensions.

However, two notes of caution are needed . First, it is not possible to usefully extend (18) to give a version of the general Möbius identity in dimensions $q > 2$. In particular, even if $\boldsymbol{B} = \boldsymbol{I}$ is the identity matrix, the singular value decomposition $\boldsymbol{B} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Gamma}$ is any rotation matrix, is not unique, leading to ambiguities in the construction of the general Möbius transformation.

Second, it should be emphasized that the fundamental Möbius identity does not lead to a natural pairing of distributions when $q > 2$. If $\boldsymbol{x}$ follows the ACG distribution with $\boldsymbol{B}$ given by (43) and if $\boldsymbol{y} = D_q(\boldsymbol{x})$, then $\boldsymbol{y}$ does not follow a spherical Cauchy distribution. The underlying reason is because the Jacobian term in (32) is not constant.

9 Parameterizations and Motivations for the Wrapped Cauchy Distribution on S_1

The WC(λ) distribution on the circle arises in a variety of settings in statistics. Here, we give a brief review. The standard one-dimensional Cauchy distribution with scale parameter b^2 and written $t_1(0, b^2, 1)$ in (38) plays a key role in two of the settings.

(a) *Angle doubling*. This topic has been the main theme of the chapter. In particular, the WC(λ) distribution can be obtained from the ACG(b) distribution by angle doubling, where b and λ are related by (3).

(b) *Stereographic projection*. As noted in Sects. 6 and 7, the WC(λ) distribution can be obtained from the Cauchy distribution by inverse stereographic projection when b is related to λ by (3).

(c) *Wrapping*. If $Z \sim t_1(0, b^2, 1)$, set $\Theta = Z \bmod 2\pi$. Recall that the Cauchy distribution has Fourier transform $\hat{f}(t) = \exp(-b|t|),\ t \in \mathbb{R}$, and its wrapped version has Fourier coefficients $\hat{f}(m),\ m \in \mathbb{Z}$. Since the WC($\lambda$) distribution has Fourier coefficients, $\lambda^{|m|},\ m \in \mathbb{Z}$, it follows that $\Theta \sim$ WC(λ) distribution with $\lambda = \exp(-b)$. Note that this value of λ is different from the value in (b).

(d) *AR(1) process*. Consider the first-order autoregression AR(1) model in time series,

$$X_{t+1} = \lambda X_t + \epsilon_t, \quad t \in \mathbb{Z},$$

where the innovation sequence $\{\epsilon_t\}$ consists of independent identically distributed $N(0, \sigma_\epsilon^2)$ random variables with ϵ_t independent of $X_s,\ s < t$. For $|\lambda| < 1$, the model describes a stationary Gaussian process with spectral density (after standardizing it to be a probability density) given by the WC(λ) density.

Table 1 Various parameterizations of the wrapped Cauchy distribution

Number	Parameter	A	B	C	Setting
1	$0 \le \lambda < 1$	$1 - \lambda^2$	$1 + \lambda^2$	2λ	Wrapped Cauchy, AR(1)
2	$0 < b \le 1$	$2b$	$1 + b^2$	$1 - b^2$	Doubled ACG, stereographic projection
3	$0 < \mu \le \pi/2$	$\sin\mu$	1	$\cos\mu$	Angular rep
4	$0 \le \alpha < 1/2$	$\sqrt{1 - 4\alpha^2}$	1	2α	CAR(1)

(e) *CAR(1) process*. Consider the first-order conditional autoregression CAR(1) model, defined by the conditional distributions

$$X_t | \{X_s,\ s \neq t\} \sim N(\alpha(X_{t-1} + X_{t+1}), \sigma_\eta^2),$$

indexed by $t \in \mathbb{Z}$. For $|\alpha| < 1/2$, this model defines a stationary process which is the same as the stationary AR(1) process. The parameters are related by $\alpha = \lambda/(1 + \lambda^2)$.

(f) *Exit distribution for Brownian motion*. For a standard Brownian motion in the plane starting from a point inside S_1, the exit distribution on S_1 has a wrapped Cauchy distribution; see (41).

Several of these settings involve different ways to parameterize the WC distribution. Note that the WC(λ) density for $0 \le \lambda < 1$ can be written in the form

$$f_{\mathrm{WC}}(\theta; \lambda) = \frac{1}{2\pi} \frac{A}{B - C\cos\theta}, \quad \theta \in S_1, \tag{45}$$

where $A, B > 0$ and $C \ge 0$. Provided $B^2 = A^2 + C^2$, the density integrates to 1. Furthermore, the density is unchanged if the parameters are multiplied by the same scalar constant. Hence, there is only one free parameter. Table 1 lists some common choices for A, B, C. Furthermore, by interchanging A and C, as has already been done for Parameterizations 1 and 2, the number of parameterizations can be doubled.

Parameterization 1 is the standard representation. As noted in (a), Parameterization 2 is motivated by doubling the angle in the ACG distribution with its standard parameterization. As noted in (b), it is also motivated by the standard parameterization of the Cauchy distribution after inverse stereographic projection. Parameterization 3 is the simplest algebraically. Parameterization 4 is motivated by the CAR(1) model in (e).

Acknowledgments I am grateful to Peter Jupp and two anonymous referees for helpful comments on the presentation.

References

Arslan, O., Constable, P. D. L., & Kent, J. T. (1995). Convergence behaviour of the EM algorithm for the multivariate t-distribution. *Communications in Statistics: Theory and Methods*, *24*, 2981–3000.

Chikuse, Y. (2003). *Statistics on special manifolds*, vol. 174 of *Lecture notes in statistics*. New York: Springer.

Downs, T. D. (2003). Spherical regression. *Biometrika*, *90*, 655–668.

Downs, T. D. (2009). Cauchy families of directional distributions closed under location and scale transformations. *The Open Statistics & Probability Journal*, *1*, 76–92.

Downs, T. D., & Mardia, K. V. (2002). Circular regression. *Biometrika*, *89*, 683–697.

Durrett, R. (1984). *Brownian motion and martingales in analysis*. Belmont, CA: Wadsworth.

Kato, S., & McCullagh, P. (2020). Some properties of a Cauchy family on the sphere derived from Möbius transformation. *Bernoulli*, *26*, 3224–3248.

Kent, J. T., & Tyler, D. E. (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *Journal of Applied Statistics*, *15*, 247–254.

Kent, J. T., Tyler, D. E., & Vardi, Y. (1994). A curious likelihood identity for the multivariate t-distribution. *Communications in Statistics: Simulation and Computation*, *23*, 441–453.

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics*. Chichester: Wiley.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

McCullagh, P. (1996). Möbius transformation and Cauchy parameter estimation. *Annals of Statistics*, *24*, 787–808.

Tyler, D. E. (1987a). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, *15*, 234–251.

Tyler, D. E. (1987b). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika*, *74*, 579–589.

Part III
Robust Theory and Methods

Power M-Estimators for Location and Scatter

Gabriel Frahm

Abstract Power M-estimators for location and scatter are studied by Frahm et al. (J. Multivariate Anal. 176:104569, 2020) in the context of missing data. It is shown that they are identical to the corresponding ML-estimators under the assumption that the data possess a re-scaled multivariate power-exponential distribution. Further, the asymptotic distributions for the power M-estimators are simplified. As a by-product, the asymptotic distributions for power M-estimators for scale-invariant functions of scatter are provided, too.

Keywords Location · M-estimation · Multivariate power-exponential distribution · Scatter; Tyler's M-estimator

1 Motivation

Power M-estimators are used by Frahm et al. (2020) to estimate the location and scatter of incomplete elliptically distributed data. Here, it is assumed that the data are complete in order to obtain closed-form expressions. The contribution is threefold: (i) It is shown that the power M-estimators for location and scatter are identical to the corresponding ML-estimators under the assumption that the data possess a re-scaled multivariate power-exponential distribution. (ii) The asymptotic distributions for the power M-estimators given by Frahm et al. (2020) are simplified. (iii) Further, the asymptotic distributions for power M-estimators for scale-invariant functions of scatter are derived.

In fact, many applications of multivariate analysis are based on the estimation of scale-invariant functions of scatter, e.g., principal component analysis, canonical correlation analysis, linear discriminant analysis, and linear regression (see, e.g., Croux & Haesbroeck 1999; Hallin & Paindaveine 2006; Oja 2003; Paindaveine

G. Frahm (🖂)
Chair of Applied Stochastics and Risk Management, Department of Mathematics and Statistics,
Helmut Schmidt University, Hamburg, Germany
e-mail: frahm@hsu-hh.de

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_8

2008; Taskinen et al. 2006). Thus, it seems worth emphasizing that the associated asymptotic distributions are quite simple and important from a practical viewpoint.

2 Prerequisites

Let X be a d-dimensional random vector possessing an elliptical distribution, i.e., $X = \mu + \Lambda\mathcal{R}U$, where $\mu \in \mathbb{R}^d$, $\Lambda \in \mathbb{R}^{d\times k}$, U is a k-dimensional random vector that is uniformly distributed on the unit hypersphere in $\mathbb{R}^k$, and $\mathcal{R}$ is a nonnegative random variable being stochastically independent of U (Cambanis et al. 1981; Fang et al. 1990, p. 42). Throughout this chapter, it is implicitly assumed that $\mathrm{P}(\mathcal{R} > 0) > 0$, which means that the distribution of X is not degenerate. Further, we can only observe the realizations of X, whereas the realizations of $\mathcal{R}$ and U are unobservable.

The distribution of X depends on Λ only through the $d \times d$ matrix $\Sigma := \Lambda\Lambda' \geq 0$. This is referred to as the scatter matrix of X, whereas μ is said to be its location vector. The random variable $\mathcal{R}$ is called the generating variate of X. If $\mathbf{E}(\mathcal{R}^2) < \infty$, the covariance matrix of X is given by $\mathbf{Cov}(X) = \mathbf{E}(\mathcal{R}^2)\Sigma/k$. In any case, the linear dependence structure of X can be described by the scatter matrix Σ. I assume that Σ is positive definite, i.e., $\mathrm{rk}(\Lambda) = d$, and that $k = d > 1$, without loss of generality. Thus, I focus on the multivariate case, although many of the results presented here are valid also for the univariate case, i.e., $d = 1$.

It is typically supposed that the distribution of $\mathcal{R}$ is absolutely continuous. In this case, the density of $\mathcal{R}$ is given by

$$f(r) = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)}\, r^{d-1} g\left(r^2\right), \qquad r \geq 0,$$

where g is a nonnegative function on $\mathbb{R}_0^+$ referred to as the density generator of X.[1] It is implicitly assumed that μ and Σ do not have any influence on g, i.e., the density generator is fixed. Now, the density of the random vector X can be written as

$$p(x) = \sqrt{\det \Sigma^{-1}}\, g\left((x-\mu)'\Sigma^{-1}(x-\mu)\right).$$

The class of elliptical distributions is a broad and fundamental generalization of the multivariate normal distribution (Cambanis et al. 1981; Fang et al. 1990; Kelker 1970). For all $\tau > 0$, we have that $X = \mu + \Lambda\mathcal{R}U = \mu + V\mathcal{S}U$ with $V := \Lambda/\tau$ and $\mathcal{S} := \tau\mathcal{R}$. Hence, if X has the scatter matrix Σ, there always exists an equivalent representation of X with the scatter matrix Σ/τ^2. Thus, Σ can be identified only if either the scale of Σ or the scale of $\mathcal{R}$ is fixed, in some appropriate way.

More precisely, consider some function $\psi : D \to \mathbb{R}^k$, where D is an open subset of the Euclidean space. The function ψ is said to be (positively) homogeneous of degree $\gamma \in \mathbb{R}$ if and only if $\psi(\kappa x) = \kappa^\gamma \psi(x)$ for all $\kappa > 0$ and $x \in D$. In particular,

[1] The quantity $\Gamma\left(\frac{d}{2}\right)/\left(2\pi^{\frac{d}{2}}\right)$ corresponds to the surface area of the unit hypersphere in $\mathbb{R}^d$.

it is said to be:

- Scale invariant if and only if it is homogeneous of degree 0.
- Linearly homogeneous if and only if it is homogeneous of degree 1.

Let $\mathcal{P}^d$ be the set of all symmetric positive-definite $d \times d$ matrices and $\sigma^2 : \mathcal{P}^d \to \mathbb{R}^+$ be a linearly homogeneous function, i.e., $\sigma^2(\kappa\Gamma) = \kappa\sigma^2(\Gamma)$ for all $\kappa > 0$ and $\Gamma \in \mathcal{P}^d$. The so-called scale function σ^2 is assumed to be differentiable, and it is also required that $\sigma^2(\mathbf{I}_d) = 1$.[2] Now, $\sigma^2(\Sigma)$ quantifies the scale of Σ, and thus "fixing the scale of Σ" means to require that $\sigma^2(\Sigma) = 1$. Typical scale functions are $\sigma^2(\Gamma) = \Gamma_{11}$, $\sigma^2(\Gamma) = \mathrm{tr}(\Gamma)/d$, and $\sigma^2(\Gamma) = \det(\Gamma)^{1/d}$. For more details, see Frahm (2009) and Paindaveine (2008).

Alternatively, given some appropriate real-valued partial function w on $\mathbb{R}_0^+$,[3] "fixing the scale of $\mathcal{R}$" means to require that the generating variate satisfies the scaling condition

$$\mathbf{E}\big(\varphi(\mathcal{R}^2)\big) = d \tag{1}$$

with $\varphi\big(\mathcal{R}^2\big) := w\big(\mathcal{R}^2\big)\,\mathcal{R}^2$. The function w is referred to as a weight function.[4] It is considered "appropriate" if and only if there exists no scaling constant $\tau \neq 1$ such that $\mathbf{E}\big(\varphi\big((\tau\mathcal{R})^2\big)\big) = d$. A sufficient condition is that φ is strictly increasing.

A prominent exception is Tyler's weight function $r^2 \mapsto d/r^2$, which can be used whenever $\mathcal{R}$ has no atom at 0, i.e., $\mathrm{P}(\mathcal{R} = 0) = 0$ (Tyler 1987a,b). In this case, obviously, it holds that $\mathbf{E}\big(\varphi\big((\tau\mathcal{R})^2\big)\big) = d$ for all $\tau > 0$, and thus we must, instead, fix the scale of Σ, i.e., require that $\sigma^2(\Sigma) = 1$ for any scale function σ^2. By contrast, the Gauss-type weight function $r^2 \mapsto 1$ is clearly appropriate. In particular, this weight function implies that

$$\Sigma = \frac{\mathbf{E}\big(\varphi(\mathcal{R}^2)\big)}{d}\,\Sigma = \frac{\mathbf{E}\big(\mathcal{R}^2\big)}{d}\,\Sigma = \mathbf{Cov}(X),$$

i.e., the scatter matrix Σ corresponds to the covariance matrix of X.

3 Power M-Estimators for Location and Scatter

Throughout this section, let the random vectors $X_1, X_2, \ldots, X_n \sim X$ be independent, where X has an elliptical distribution on $\mathbb{R}^d$ with $d > 1$, location vector μ, scatter matrix $\Sigma > 0$, and generating variate $\mathcal{R}$ without atom at 0.

[2] Here, $\mathbf{I}_d$ represents the $d \times d$ identity matrix.

[3] A partial function from A to B is a function from a subset of A to B.

[4] The reason why will become clear in the next section.

3.1 ML-Estimation

In the context of ML-estimation, it is assumed that the distribution of $\mathcal{R}$, i.e., the generating distribution of X, is known. Further, it is supposed to be absolutely continuous and so X has a density function. For example:

- $\sqrt{\chi^2_d}$ is the generating variate of the multivariate normal distribution.
- $\sqrt{dF_{d,\nu}}$ with $F_{d,\nu} \sim F(d,\nu)$ represents the generating variate of the multivariate t-distribution with $\nu > 0$ degrees of freedom.
- $G^{1/(2\beta)}_{d/(2\beta),2}$ with $G_{d/(2\beta),2} \sim \text{Gamma}\big(\frac{d}{2\beta}, 2\big)$ generates the multivariate power-exponential distribution with shape parameter $\beta > 0$.[5]

For $\nu \to \infty$, we obtain $dF_{d,\nu} \overset{\text{d}}{\to} \chi^2_d$, and for $\beta = 1$, it holds that $G^{1/\beta}_{d/(2\beta),2} \sim \chi^2_d$, too, which leads us to the multivariate normal distribution.

The density generator of the multivariate normal distribution is

$$r^2 \longmapsto (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}r^2\right).$$

Further, the density generator of the multivariate t-distribution is given by

$$r^2 \longmapsto \frac{1}{(\nu\pi)^{\frac{d}{2}}} \frac{\Gamma(\frac{d+\nu}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{r^2}{\nu}\right)^{-\frac{d+\nu}{2}},$$

whereas for the multivariate power-exponential distribution (Gómez et al. 1998), we have that

$$r^2 \longmapsto \frac{1}{2^{\frac{d}{2\beta}}\pi^{\frac{d}{2}}} \frac{\Gamma(\frac{d}{2}+1)}{\Gamma\big(\frac{d}{2\beta}+1\big)} \exp\left(-\frac{1}{2}r^{2\beta}\right).$$

Hence, we obtain the multivariate normal distribution for $\beta = 1$, whereas the given distribution is light tailed if $\beta > 1$. The latter case is ignored throughout this chapter.

In order to compute the corresponding ML-estimates for μ and Σ, one usually tries to solve the system

$$0 = \frac{1}{n}\sum_{i=1}^{n} w\big(r_i^2\big)(X_i - \hat{\mu})$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} w\big(r_i^2\big)(X_i - \hat{\mu})(X_i - \hat{\mu})'$$

[5] Here, $d/(2\beta)$ is the shape parameter and 2 is the scale parameter of the Gamma distribution.

of ML-estimating equations,[6] where $r_i = \sqrt{(X_i - \hat{\mu})'\widehat{\Sigma}^{-1}(X_i - \hat{\mu})}$ for $i = 1, 2, \ldots, n$. In some cases, it can happen that the simultaneous ML-estimating equations have no solution at all. I will come back to this pathological case below.

The corresponding ML-weight function w is given by $r^2 \mapsto -2g'(r^2)/g(r^2)$ for all $r^2 \geq 0$ with $g(r^2) > 0$ (Tyler 1982), provided that the density generator g is differentiable wherever it is positive. ML-estimation presumes that the distribution of $\mathcal{R}$ is known, and so the scatter matrix Σ is identifiable by construction. In fact, the scaling condition $\mathbf{E}\big(\varphi(\mathcal{R}^2)\big) = d$, which is given by Eq. 1, is implicitly satisfied if the ML-estimator for Σ is Fisher consistent, i.e., if we have that

$$\mathbf{E}\left(\frac{\partial \log p(X; \mu, \Sigma)}{\partial \Sigma}\right) = 0,$$

where $p(\cdot; \mu, \Sigma)$ is the density function of X, given its location vector μ and scatter matrix Σ (Frahm et al. 2020). Fisher consistency is an essential requirement of ML-estimation. If an ML-estimator, understood as the solution of an ML-estimating equation, is Fisher inconsistent, it cannot be consistent (in the usual sense) at all.

The ML-weight function w associated with the multivariate normal distribution is $r^2 \mapsto 1$. Further, for the multivariate t-distribution, we obtain the weight function $r^2 \mapsto (d + \nu)/(r^2 + \nu)$, whereas the multivariate power-exponential distribution leads us to the weight function $r^2 \mapsto \beta r^{2(\beta-1)}$. Hence, in the latter case, the corresponding ML-estimating equations for μ and Σ are

$$0 = \frac{\beta}{n} \sum_{i=1}^{n} \frac{X_i - \hat{\mu}}{\big[(X_i - \hat{\mu})'\widehat{\Sigma}^{-1}(X_i - \hat{\mu})\big]^{1-\beta}} \tag{2}$$

and

$$\widehat{\Sigma} = \frac{\beta}{n} \sum_{i=1}^{n} \frac{(X_i - \hat{\mu})(X_i - \hat{\mu})'}{\big[(X_i - \hat{\mu})'\widehat{\Sigma}^{-1}(X_i - \hat{\mu})\big]^{1-\beta}}. \tag{3}$$

For $\beta = 1$, i.e., if X is multivariate normally distributed, we obtain the constant weight function $r^2 \mapsto 1$. Then the ML-estimators for μ and Σ simply turn into the sample mean vector and sample covariance matrix.

I already mentioned above that the simultaneous ML-estimating equations could have no solution at all. The multivariate power-exponential distribution is continuous on $\mathbb{R}^d$. This means that there is no lower-dimensional hyperplane $\mathcal{H} \subset \mathbb{R}^d$ such that $\mathrm{P}(X \in \mathcal{H}) > 0$. Now, let $x_1, x_2, \ldots, x_n$ (with $n > d$) be some data points in $\mathbb{R}^d$ being generated by a multivariate power-exponential distribution with known shape parameter $0 < \beta \leq 1$. Further, let $\hat{\mu}$ and $\widehat{\Sigma}$ be the ML-estimates of

[6] In this chapter, the symbol "0" represents a zero scalar, a zero vector, or a zero matrix. Its particular meaning should always be clear from the context.

μ and Σ, respectively, based on the assumption that X has a multivariate power-exponential distribution with shape parameter β. To be more precise, $\hat{\mu}$ and $\widehat{\Sigma}$ maximize the *log-likelihood function*

$$(\mu, \Sigma) \longmapsto c + n \log \det \Sigma^{-\frac{1}{2}} - \frac{1}{2} \sum_{i=1}^{n} \left[(x_i - \mu)'\Sigma^{-1}(x_i - \mu)\right]^{\beta}$$

with some constant c, but they do not necessarily solve the ML-estimating equations 2 and 3. In the case of $\beta \leq \frac{1}{2}$, the log-likelihood function of the multivariate power-exponential family is peaked and thus not differentiable at each $\mu \in \{x_1, x_2, \ldots, x_n\}$. Thus, solving the ML-estimating equations, in order to compute the ML-estimates, is inappropriate if the shape parameter β is not greater than $\frac{1}{2}$. Then it could even happen that $\hat{\mu} = x_i$ with positive probability, in which case the ML-estimators $\hat{\mu}$ and $\widehat{\Sigma}$ cannot satisfy the simultaneous ML-estimating equations 2 and 3.[7]

3.2 M-Estimation

Now, we drop the assumption that the generating distribution of X is known and absolutely continuous. In this case, we can estimate μ and Σ by solving the system

$$0 = \frac{1}{n} \sum_{i=1}^{n} v(r_i)(X_i - \hat{\mu})$$

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} w\left(r_i^2\right)(X_i - \hat{\mu})(X_i - \hat{\mu})'$$

of M-estimating equations, where v and w are some real-valued partial weight functions on $\mathbb{R}_0^+$ (see, e.g., Huber & Ronchetti 2009; Maronna 1976, Chapter 8).

The population version of the given system of M-estimating equations is

$$0 = \mathbf{E}\big(v(\mathcal{R})(X - \mu)\big)$$

$$\Sigma = \mathbf{E}\big(w\big(\mathcal{R}^2\big)(X - \mu)(X - \mu)'\big),$$

where $\mathcal{R} = \sqrt{(X - \mu)'\Sigma^{-1}(X - \mu)}$. From:

- $\Sigma = \Lambda\Lambda'$
- $\mathbf{E}(U) = 0$

[7] I would like to thank an anonymous referee for this important hint.

- $\mathbf{E}(UU') = \mathbf{I}_d/d$
- $X - \mu = \Lambda \mathcal{R} U$
- $(X - \mu)(X - \mu)' = \mathcal{R}^2 \Lambda U U' \Lambda'$

we conclude that the first equation is always satisfied, provided that $\mathbf{E}\big(v(\mathcal{R})\mathcal{R}\big) < \infty$, whereas the second equation requires the critical scaling condition given by Eq. 1.

Most of the results contained in this chapter do not depend on the choice of v. Nonetheless, it is important to choose v carefully—as we will see at the end of this section. Whenever I refer to some weight function without any further remark, I mean the weight function w. The following M-weight functions can frequently be found in the literature (see, e.g., Kent & Tyler 1991; Tyler 1987a):

- The Gauss-type weight function $r^2 \mapsto 1$.
- The Student-type weight function $r^2 \mapsto (d + \nu)/(r^2 + \nu)$ with $\nu > 0$.
- Tyler's weight function $r^2 \mapsto d/r^2$.
- Huber's weight function

$$r^2 \longmapsto \begin{cases} \gamma\,, & r^2 < \lambda \\ \gamma\lambda/r^2\,, & r^2 \geq \lambda\,, \end{cases}$$

  where the parameters $\gamma, \lambda > 0$ are such that $\mathbf{E}\big(\varphi(\chi^2_d)\big) = d$.

As already mentioned in the last section, if we choose $r^2 \mapsto -2g'(r^2)/g(r^2)$ for all $r^2 \geq 0$ with $g(r^2) > 0$, the M-weight function reduces to the ML-weight function associated with the density generator g. I call the weight function $r^2 \mapsto 1$ "Gauss-type" because it is the ML-weight function under the assumption that the data have a multivariate normal distribution. Similarly, the weight function $r^2 \mapsto (d + \nu)/(r^2 + \nu)$ is called "Student-type" because it is the ML-weight function given that the data have a multivariate t-distribution. See Dümbgen et al. (2015) for a comprehensive survey on M-estimation of scatter.

The power M-weight functions for μ and Σ proposed by Frahm et al. (2020) are given by:

- $v : r \mapsto r^{-\alpha}$
- $w : r^2 \mapsto \left(\frac{r^2}{d}\right)^{-\alpha}$

respectively, where $0 \leq \alpha \leq 1$ represents a so-called tail index. If α is lower than 1, the function $\varphi : r^2 \mapsto d^{\alpha} r^{2(1-\alpha)}$ is strictly increasing. Hence, the power M-weight function for scatter, w, with tail index $\alpha < 1$ is appropriate in the sense that we are able to fix the scale of $\mathcal{R}$ by applying the scaling condition expressed by Eq. 1.

The power M-estimators $\hat{\mu}$ and $\widehat{\Sigma}$ are the solutions of

$$0 = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \hat{\mu}}{\big[(X_i - \hat{\mu})'\widehat{\Sigma}^{-1}(X_i - \hat{\mu})\big]^{\frac{\alpha}{2}}} \tag{4}$$

and

$$\widehat{\Sigma} = \frac{d^\alpha}{n}\sum_{i=1}^{n}\frac{(X_i-\hat{\mu})(X_i-\hat{\mu})'}{\left[(X_i-\hat{\mu})'\widehat{\Sigma}^{-1}(X_i-\hat{\mu})\right]^\alpha}. \tag{5}$$

Once again, it could happen that the power M-estimators do not exist, almost surely, which depends on the chosen tail index α and the distribution of X.

The weight function v with tail index $\alpha = 0$ leads us to the sample mean vector $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimator for μ, whereas for $\alpha = 1$ we get the M-estimator for location proposed by Hettmansperger and Randles (2002). This means that $\hat{\mu}$ is the solution of

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\hat{\mu}}{\sqrt{(X_i-\hat{\mu})'\widehat{\Sigma}^{-1}(X_i-\hat{\mu})}},$$

where $\widehat{\Sigma}$ is the power M-estimator for Σ with tail index $\alpha = 1$, i.e., Tyler's M-estimator

$$\widehat{\Sigma} = \frac{d}{n}\sum_{i=1}^{n}\frac{(X_i-\hat{\mu})(X_i-\hat{\mu})'}{(X_i-\hat{\mu})'\widehat{\Sigma}^{-1}(X_i-\hat{\mu})}$$

for scatter. The solutions of both equations do not depend on the chosen scale of $\widehat{\Sigma}$.

Correspondingly, the Gauss-type weight function $w\colon r^2 \mapsto 1$ appears for $\alpha = 0$, whereas Tyler's weight function $w\colon r^2 \mapsto d/r^2$ can be found on the boundary $\alpha = 1$. Any choice of α between 0 and 1 is a compromise between the sample covariance matrix $\frac{1}{n}\sum_{i=1}^{n}\big(X_i-\hat{\mu}\big)\big(X_i-\hat{\mu}\big)'$, i.e., the typical nonrobust estimator, and Tyler's M-estimator, i.e., the most robust estimator for Σ (Tyler 1987a, Remark 3.1). Simply put, the tail index α determines the robustness of the power M-estimators.

In the introduction, I mentioned that power M-estimators are used by Frahm et al. (2020) to estimate the location and scatter of incomplete elliptically distributed data. Now, I would like to explain, shortly, the reason why the class of power M-estimators is particularly suitable for missing-data analysis. Suppose that we observe only $k \in \{1, 2, \ldots, d\}$ components of X. In fact, the observed subvector of X is elliptically distributed, too, but with generating variate $\mathcal{R}\sqrt{\mathcal{B}}$, where $\mathcal{B} \sim \mathrm{Beta}\big(\frac{k}{2}, \frac{d-k}{2}\big)$ is independent of $\mathcal{R}$ (Cambanis et al. 1981). Hence, when applying an M-weight function for scatter to incomplete data, the critical scaling condition given by Eq. 1 turns into $\mathbf{E}\big(\varphi_k(\mathcal{R}^2\mathcal{B})\big) = k$ with $\varphi_k(\mathcal{R}^2\mathcal{B}) := w_k(\mathcal{R}^2\mathcal{B})(\mathcal{R}^2\mathcal{B})$, where w_k denotes the M-weight for scatter given that one observes only k components of X. To guarantee that the resulting M-estimator is consistent, the scaling condition must be satisfied for $k = 1, 2, \ldots, d$. Let $\mathcal{R}$ be such that

$$\mathbf{E}\left(\left(\frac{\mathcal{R}^2}{d}\right)^{-\alpha}\mathcal{R}^2\right) = d$$

for some tail index $0 \leq \alpha \leq 1$. Hence, given the power M-weight function for scatter, $\mathcal{R}$ satisfies the scaling condition expressed by Eq. 1, which is a basic requirement in the complete-data case. Now, Theorem 3 of Frahm et al. (2020) states that

$$\mathbf{E}\left(\frac{\mathrm{B}\left(\frac{k}{2}+1,\frac{d-k}{2}\right)}{\mathrm{B}\left(\frac{k}{2}+1-\alpha,\frac{d-k}{2}\right)}\left(\frac{\mathcal{R}^2\mathcal{B}}{d}\right)^{-\alpha}\mathcal{R}^2\mathcal{B}\right)=k,$$

where $\mathrm{B}(a,b)$ denotes Euler's beta function with parameters $a,b>0$. Thus, a natural choice of the power M-weight function in the case of incomplete data is

$$w_k\colon\ r^2\longmapsto\frac{\mathrm{B}\left(\frac{k}{2}+1,\frac{d-k}{2}\right)}{\mathrm{B}\left(\frac{k}{2}+1-\alpha,\frac{d-k}{2}\right)}\left(\frac{r^2}{d}\right)^{-\alpha},\qquad 0\leq\alpha\leq 1.$$

It holds that $\mathrm{B}(a,x)/\,\mathrm{B}(b,x)=1$ as $x\searrow 0$, and so we may define $\mathrm{B}(a,0)/\,\mathrm{B}(b,0)=1$ for all $a,b>0$. Hence, in the complete-data case, i.e., $k=d$, w_k turns into the usual power M-weight function for scatter, i.e., $w\colon r^2\mapsto\left(\frac{r^2}{d}\right)^{-\alpha}$. Further, for $\alpha=0$ and any $k\in\{1,2,\ldots,d\}$, we obtain the Gauss-type weight function $w_k : r^2\mapsto 1$, whereas for $\alpha=1$ we have that $w_k : r^2\mapsto k/r^2$ (Frahm et al. 2020). To sum up, scaling the power M-weight function w in an appropriate way guarantees that the resulting M-estimator for scatter remains consistent in the incomplete-data case.

A seeming weakness of the power M-estimating equations 4 and 5 is that both

$$\frac{x_i-\mu}{\left[(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]^{\frac{\alpha}{2}}}\tag{6}$$

and

$$\frac{(x_i-\mu)(x_i-\mu)'}{\left[(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]^{\alpha}}\tag{7}$$

are not (yet) defined for $\mu=x_i$, where $x_i\in\mathbb{R}^d$ is any data point. The Euclidean norm of the vector in (6) is $\mathcal{O}\left(\|\mu-x_i\|^{1-\alpha}\right)$. Hence, in the case of $0\leq\alpha<1$, that vector vanishes as $\mu\to x_i$, and thus we can set it to 0 whenever $\mu\in\{x_1,x_2,\ldots,x_n\}$. By contrast, in the case of $\alpha=1$, the min–max theorem tells us that the Euclidean norm of the vector in (6) is bounded below by the square root of the minimum eigenvalue of Σ, which is positive. Hence, in this case, the vector does *not* vanish as $\mu\to x_i$.

The same argument holds true for the matrix in (7), since this can be written as

$$\left[\frac{x_i-\mu}{\left[(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]^{\frac{\alpha}{2}}}\right]\left[\frac{x_i-\mu}{\left[(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]^{\frac{\alpha}{2}}}\right]'.$$

To sum up, the power M-estimating equations are always well defined, provided that the tail index α is lower than 1. More precisely,

$$\frac{X_i - \hat{\mu}}{\big[(X_i - \hat{\mu})'\widehat{\Sigma}^{-1}(X_i - \hat{\mu})\big]^{\frac{\alpha}{2}}}$$

can be considered zero if $\hat{\mu} = X_i$ for some $i \in \big\{1, 2, \ldots, n\big\}$. However, for the tail index $\alpha = 1$, there seems to be no obvious solution in the pathological case $\hat{\mu} = X_i$.[8]

The above arguments can be summarized as follows. Let

$$S_\alpha(y) = \begin{cases} y/\|y\|^\alpha, & y \neq 0 \\ 0, & y = 0 \end{cases}$$

with $0 \leq \alpha < 1$ be the spatial power sign of the data point $y \in \mathbb{R}^d$, whereas the spatial power sign for $\alpha = 1$, i.e., the spatial sign $S_1(y) = y/\|y\|$, remains undefined at $y = 0$. Hence, in the case of $0 \leq \alpha < 1$, the power M-estimating equations 4 and 5 can be re-written, equivalently, as

$$\frac{1}{n}\sum_{i=1}^{n} S_\alpha(Y_i) = 0 \qquad \text{and} \qquad \frac{d^\alpha}{n}\sum_{i=1}^{n} S_\alpha(Y_i)S_\alpha(Y_i)' = \mathbf{I}_d$$

with $Y_i := \hat{\Lambda}^{-1}(X_i - \hat{\mu})$ for $i = 1, 2, \ldots, n$ and $\hat{\Lambda}\hat{\Lambda}' = \widehat{\Sigma}$. This works also if $Y_i = 0$ for some $i \in \big\{1, 2, \ldots, n\big\}$. By contrast, re-writing the power M-estimating equations makes not much sense at all for $\alpha = 1$.

Being able to choose the weight functions for location and scatter differently is a major advantage of M-estimation. The solution proposed above is not applicable to the ML-estimators for μ and Σ under the assumption that the data have a multivariate power-exponential distribution with shape parameter $\beta \leq \frac{1}{2}$. The problem is the ML-estimating equation for μ, i.e., Eq. 2. The Euclidean norm of the vector

$$\frac{x_i - \mu}{\big[(x_i - \mu)'\Sigma^{-1}(x_i - \mu)\big]^{1-\beta}}$$

is $O\big(\|\mu - x_i\|^{2\beta - 1}\big)$. Hence, in the case of $\beta \leq \frac{1}{2}$, which I already mentioned at the end of the last section, the vector cannot be considered zero for $\mu = x_i$. Once again, this underpins the observation that applying the ML-estimating equations 2 and 3, instead of the power M-estimating equations 4 and 5, is inappropriate if X has a multivariate power-exponential distribution with shape parameter $\beta \leq \frac{1}{2}$.

[8] Tyler (1987a) simply suggests to disregard all data points that equal $\hat{\mu}$, since they do not contain any directional information at all.

3.3 Main Result

Tyler (1983) considers his M-weight function a limit of Huber's M-weight function, whereas Tyler (1987b) derives it as an ML-weight function by observing that the distribution of the random vector $S = (X - \mu)/\|X - \mu\|$ does not depend on $\mathcal{R} > 0$. In fact, we have that

$$S = \frac{X - \mu}{\|X - \mu\|} = \frac{\mathcal{R}\Lambda U}{\|\mathcal{R}\Lambda U\|} = \frac{\Lambda U}{\|\Lambda U\|},$$

provided that $\mathcal{R}$ has no atom at 0. Tyler (1987b) calls the distribution of S angular central Gaussian on the sphere. Its density function is

$$\phi : s \longmapsto \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{\frac{d}{2}}} \sqrt{\det \Sigma^{-1}} \sqrt{s'\Sigma^{-1}s}^{\,-d}$$

for all $s \in \mathbb{R}^d$ with $\|s\| = 1$. Frahm (2004, Ch. 4.2.1) calls ϕ a spectral density function and notes that S possesses a generalized elliptical distribution. Frahm and Jaekel (2010) call ϕ a characteristic density function, since the eigenvectors and eigenvalues of Σ are characterized by the stationary points of ϕ (Frahm & Jaekel 2015, p. 299). Moreover, in Mardia and Jupp (2000, p. 178), the distribution of S is referred to as the projected (or offset) normal. In the case of $d = 2$, it turns into the wrapped Cauchy distribution after angle doubling (Kent and Tyler 1988).

Another way to obtain Tyler's weight function is to set $\nu = 0$ in the Student-type weight function $r^2 \mapsto (d + \nu)/(r^2 + \nu)$ or to set $\alpha = 1$ in the power M-weight function $r^2 \mapsto \left(\frac{r^2}{d}\right)^{-\alpha}$ proposed by Frahm et al. (2020). A main contribution of this chapter is the observation that the power M-weight function for scatter with tail index $0 \leq \alpha < 1$ represents the ML-weight function based on the density generator g with

$$g(r^2) \propto \exp\left(-\frac{d^{\alpha} r^{2(1-\alpha)}}{2(1-\alpha)}\right), \tag{8}$$

which represents the density generator of a *re-scaled* multivariate power-exponential distribution with shape parameter $\beta = 1 - \alpha > 0$.

By contrast, as already mentioned in Sect. 3.1, Gómez et al. (1998) originally choose the density generator g with

$$g(r^2) \propto \exp\left(-\frac{1}{2} r^{2\beta}\right)$$

with shape parameter $\beta > 0$ to define the multivariate power-exponential family. More precisely, the original choice of the generating variate $\mathcal{R}$ is such that

$$\mathcal{R}^{2\beta} \sim \text{Gamma}\left(\frac{d}{2\beta}, 2\right).$$

After substituting β with $1 - \alpha$ and multiplying $\mathcal{R}$ by $\left(\frac{1-\alpha}{d^\alpha}\right)^{1/[2(1-\alpha)]}$, we obtain the density generator given by Eq. 8. Hence, the stochastic representation of the corresponding random vector X is

$$X = \mu + \Lambda\left(\frac{1-\alpha}{d^\alpha}\right)^{\frac{1}{2(1-\alpha)}} \mathcal{R}U,$$

which means that X has a multivariate power-exponential distribution with location vector μ, scatter matrix $\Upsilon = \left(\frac{1-\alpha}{d^\alpha}\right)^{1/(1-\alpha)}\Sigma$, and shape parameter $1-\alpha$. However, the scatter matrix of X is still Σ if we consider the random variable $\left(\frac{1-\alpha}{d^\alpha}\right)^{1/[2(1-\alpha)]}\mathcal{R}$ its generating variate. Thus, by using the power M-weight function, we estimate Σ, i.e., the scatter matrix of the *re-scaled* multivariate power-exponential distribution, not the scatter matrix Υ in the sense of Gómez et al. (1998).

Theorem 1 *Let the density generator g of an elliptically distributed random vector X with location vector $\mu \in \mathbb{R}^d$, positive-definite scatter matrix $\Sigma \in \mathbb{R}^{d\times d}$, and generating variate $\mathcal{R}$ be such that*

$$g(r^2) \propto \exp\left(-\frac{d^\alpha r^{2(1-\alpha)}}{2(1-\alpha)}\right), \qquad 0 \le \alpha < 1.$$

Then the density function of X is given by

$$x \longmapsto c(d,\alpha)\sqrt{\det\Sigma^{-1}}\exp\left(-\frac{d^\alpha}{2(1-\alpha)}\left[(x-\mu)'\Sigma^{-1}(x-\mu)\right]^{1-\alpha}\right)$$

with

$$c(d,\alpha) = \frac{1}{\pi^{\frac{d}{2}}}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma\left(\frac{d}{2(1-\alpha)}+1\right)}\left(\frac{d^\alpha}{2(1-\alpha)}\right)^{\frac{d}{2(1-\alpha)}}.$$

This is the density function of a multivariate power-exponential distribution with location vector μ, scatter matrix $\Upsilon = \left(\frac{1-\alpha}{d^\alpha}\right)^{1/(1-\alpha)}\Sigma$, and shape parameter $1-\alpha$. Further, the generating variate $\mathcal{R}$ is such that

$$\mathcal{R}^{2(1-\alpha)} \sim \textit{Gamma}\left(\frac{d}{2(1-\alpha)}, \frac{2(1-\alpha)}{d^\alpha}\right).$$

The associated ML-weight function is the power M-weight function for scatter with tail index α, i.e., $r^2 \mapsto \left(\frac{r^2}{d}\right)^{-\alpha}$, and it holds that $\mathbf{E}\big(\varphi(\mathcal{R}^2)\big) = d$ for all $\alpha \in [0, 1]$.

We conclude that the ML-estimator for Σ under the assumption that X has a rescaled multivariate power-exponential distribution with shape parameter $1 - \alpha > 0$, according to Theorem 1, corresponds to the power M-estimator $\widehat{\Sigma}$ with tail index $0 \leq \alpha < 1$ given by Eq. 5. Nonetheless, the power M-estimator $\hat{\mu}$ in Eq. 4 is *not* the ML-estimator for μ under the same distributional assumption about X. Actually, the corresponding ML-estimator is the solution of

$$0 = \frac{d^\alpha}{n} \sum_{i=1}^{n} \frac{X_i - \hat{\mu}}{\big[(X_i - \hat{\mu})'\widehat{\Sigma}^{-1}(X_i - \hat{\mu})\big]^\alpha}.$$

Put another way, it equals the ML-estimator for μ given that X has a multivariate power-exponential distribution with shape parameter $\beta = 1 - \alpha$ (see Eq. 2).

4 Asymptotic Distributions

4.1 Theoretical Results

The theoretical results presented in this section about the asymptotic distributions for the power M-estimators $\hat{\mu}$ and $\widehat{\Sigma}$, which solve the simultaneous M-estimating equations 4 and 5, require some basic notation of multivariate analysis:

- As already mentioned, $\mathbf{I}_d$ denotes the $d \times d$ identity matrix.
- The $d^2 \times d^2$ matrix $\mathbf{J}_{d^2}$ is defined as $\sum_{i=1}^{d} \mathbf{e}_{ii} \otimes \mathbf{e}_{ii}$, where $\otimes$ is the Kronecker product and $\mathbf{e}_{ij}$ is the $d \times d$ matrix with 1 in the ijth position and 0 elsewhere.
- The commutation matrix $\mathbf{K}_{d^2}$ is the $d^2 \times d^2$ matrix given by $\sum_{i,j=1}^{d} \mathbf{e}_{ij} \otimes \mathbf{e}_{ji}$.
- For any $d \times d$ matrix M, the d^2-dimensional vector vec$M$ is obtained by stacking the columns of M on top of each other.

In order to obtain closed-form expressions, I assume that the data are complete, independent, and identically distributed. For more details on the case of incomplete and (serially or spatially) dependent data, see Frahm et al. (2020). Moreover, I presume that the sample size, n, is sufficiently large. Otherwise, the probability that the simultaneous power M-estimating equations 4 and 5 have no solution at all would be positive. I assume also that, under the given elliptical distribution, $\hat{\mu}$ and $\widehat{\Sigma}$ exist and that they are unique, almost surely, for each sufficiently large sample size n.

If $\Sigma \in \mathbb{R}^{d \times d}$ is a positive-definite matrix and $\widehat{\Sigma}$ is some estimator for Σ, then

$$\sqrt{n}\big(\widehat{\Sigma} - \Sigma\big) \longrightarrow N_{d \times d}(0, C), \qquad n \longrightarrow \infty,$$

means that $\sqrt{n}\big(\mathrm{vec}\widehat{\Sigma} - \mathrm{vec}\Sigma\big)$ converges weakly to a multivariate normally distributed random vector $\xi \sim N_{d^2}(0, C)$ with $0 \in \mathbb{R}^{d^2}$ and $C \in \mathbb{R}^{d^2 \times d^2}$.

Let $\psi : \mathcal{P}^d \to \mathbb{R}^k$ be any differentiable function of the scatter matrix Σ, and suppose that the parameter $\theta = \psi(\Sigma)$ is scale invariant, i.e., $\psi\big(\Sigma/\tau^2\big) = \psi(\Sigma)$ for all $\tau > 0$. The Jacobian of ψ at Σ is given by the $k \times d^2$ matrix $\partial\psi(\Sigma)/\partial\mathrm{vec}(\Sigma)'$. Each off-diagonal element of the lower triangular part of Σ represents an implicit function of the corresponding off-diagonal element of its upper triangular part and vice versa. Hence, the total differential of ψ is $\mathrm{d}\psi(\Sigma) = \mathcal{J}_\psi \mathrm{dvec}(\Sigma)$ with

$$\mathcal{J}_\psi := \frac{\partial\psi(\Sigma)}{\partial\mathrm{vec}(\Sigma)'}\frac{1}{2}\big(\mathbf{I}_{d^2} + \mathbf{J}_{d^2}\big),$$

where $\frac{1}{2}(\mathbf{I}_{d^2} + \mathbf{J}_{d^2})$ adjusts for the redundancy caused by the symmetry of Σ.

Throughout this section, I assume that the scale of the generating variate $\mathcal{R}$ is fixed such that

$$\mathbf{E}\big(\varphi(\mathcal{R}^2)\big) = \mathbf{E}\Big(\big(\mathcal{R}^2/d\big)^{-\alpha}\mathcal{R}^2\Big) = d,$$

which means that $\mathcal{R}$ satisfies the general scaling condition expressed by Eq. 1. Hence, we can substitute $\mathbf{E}\big(\mathcal{R}^{2(1-\alpha)}\big)$ with $d^{1-\alpha}$ and thus simplify the asymptotic moments of $\sqrt{n}\big(\hat{\mu} - \mu\big)$ and $\sqrt{n}\big(\widehat{\Sigma} - \Sigma\big)$ provided by Frahm et al. (2020).

Theorem 2 *Suppose that the random vectors $X_1, X_2, \ldots, X_n \sim X$ are independent, where X has an elliptical distribution on $\mathbb{R}^d$ with $d > 1$, location vector μ, scatter matrix $\Sigma > 0$, and generating variate $\mathcal{R}$ without atom at 0 such that:*

1. $\mathbf{E}\big(\mathcal{R}^{-\alpha}\big) < \infty$
2. $\mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big) < \infty$
3. $\mathbf{E}\big(\mathcal{R}^{2(1-\alpha)}\big) = d^{1-\alpha}$

Let $\hat{\mu}$ and $\widehat{\Sigma}$ be the power M-estimators for location and scatter with tail index $0 \leq \alpha \leq 1$. Further, assume that $\hat{\mu}$ and $\widehat{\Sigma}$ exist and that they are unique, almost surely, for each sufficiently large sample size n. Then we have that

$$\sqrt{n}\big(\hat{\mu} - \mu\big) \longrightarrow N_d\left(0, \frac{d^{2-\alpha}}{(d-\alpha)^2}\frac{1}{\mathbf{E}^2\big(\mathcal{R}^{-\alpha}\big)}\Sigma\right), \qquad n \longrightarrow \infty\,.$$

Further, it holds that $\sqrt{n}\big(\widehat{\Sigma} - \Sigma\big) \to N_{d\times d}(0, A)$ as $n \to \infty$ with

$$A = \gamma_1\big(\mathbf{I}_{d^2} + \mathbf{K}_{d^2}\big)\big(\Sigma \otimes \Sigma\big) + \gamma_2\mathrm{vec}(\Sigma)\mathrm{vec}(\Sigma)',$$

provided that $0 \leq \alpha < 1$. *In this case, the numbers* γ_1 *and* γ_2 *are given by*

$$\gamma_1 = \frac{d+2}{d}\,\eta(d,\alpha) \quad \textit{and} \quad \gamma_2 = \frac{1}{(1-\alpha)^2}\left[\frac{d+2(1-\alpha)(1+\alpha)}{d}\,\eta(d,\alpha) - 1\right]$$

with

$$\eta(d,\alpha) = \left(\frac{d^\alpha}{d+2(1-\alpha)}\right)^2 \mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big).$$

Moreover, $\sqrt{n}\big(\hat{\mu}-\mu\big)$ *and* $\sqrt{n}\big(\widehat{\Sigma}-\Sigma\big)$ *are asymptotically independent.*

Let $\psi : \mathcal{P}^d \to \mathbb{R}^k$ *be a scale-invariant differentiable function of* Σ *and* $\hat{\theta} = \psi\big(\widehat{\Sigma}\big)$ *be the estimator for* $\theta = \psi(\Sigma)$. *Then we have that* $\sqrt{n}\big(\hat{\theta}-\theta\big) \to N_{k\times k}\big(0, B\big)$ *as* $n \to \infty$ *with*

$$B = 2\,\frac{d+2}{d}\left(\frac{d^\alpha}{d+2(1-\alpha)}\right)^2 \mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big)\,\mathcal{J}_\psi\,(\Sigma\otimes\Sigma)\mathcal{J}_\psi'\,.$$

Finally, $\sqrt{n}\big(\hat{\mu}-\mu\big)$ *and* $\sqrt{n}\big(\hat{\theta}-\theta\big)$ *are asymptotically independent, too.*

The given result concerning $\sqrt{n}\big(\widehat{\Sigma}-\Sigma\big)$ requires $0 \leq \alpha < 1$, whereas regarding $\sqrt{n}\big(\widehat{\theta}-\theta\big)$, it is valid also for $\alpha = 1$, in which case the power M-estimator for Σ is Tyler's M-estimator. As we can see, there are three moment conditions:

1. The inlier condition $\mathbf{E}\big(\mathcal{R}^{-\alpha}\big) < \infty$
2. The outlier condition $\mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big) < \infty$
3. The scaling condition $\mathbf{E}\big(\mathcal{R}^{2(1-\alpha)}\big) = d^{1-\alpha}$

The inlier condition states that the generating distribution of X must not be too heavily concentrated around 0, whereas the outlier condition requires that its right tail must not be too heavy. If we use the sample moments for μ and Σ, i.e., $\alpha = 0$, the inlier condition disappears, whereas the outlier condition requires $\mathcal{R}$ to have a finite fourth moment. By contrast, if we apply the Hettmansperger–Randles M-estimator for μ and Tyler's M-estimator for Σ, i.e., $\alpha = 1$, the outlier condition disappears, whereas the inlier condition states that $\mathbf{E}\big(\mathcal{R}^{-1}\big) < \infty$. Thus, choosing an appropriate tail index $0 \leq \alpha \leq 1$ means to make a trade-off between the inlier and outlier conditions. To be more precise, if the data are heavy tailed, α should be close to 1, but then the data points must not be too much concentrated around the center. For $\alpha = 1$, this is already observed by Tyler (1987a).

For example, suppose that $\mathcal{R} \sim \text{Gamma}(\varsigma, \zeta)$ with shape parameter $\varsigma > 0$ and scale parameter $\zeta > 0$. Then, we have that

$$\mathbf{E}\big(\mathcal{R}^{-\alpha}\big) = \zeta^{-\alpha}\,\frac{\Gamma(\varsigma-\alpha)}{\Gamma(\varsigma)}.$$

This means that the shape parameter of the gamma distribution, ς, must exceed α, since otherwise the inlier condition is violated. However, the outlier condition is always satisfied if $\mathcal{R}$ is Gamma-distributed.

By contrast, if $\mathcal{R} = \sqrt{dF_{d,\nu}}$ with $F_{d,\nu} \sim F(d,\nu)$ and $\nu > 0$, i.e., if we choose the generating variate of the multivariate t-distribution with ν degrees of freedom, we obtain

$$\mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big) = d^{2(1-\alpha)}\mathbf{E}\big(F_{d,\nu}^{2(1-\alpha)}\big).$$

It is well known that $\mathbf{E}\big(F_{d,\nu}^{2(1-\alpha)}\big)$ is finite if and only if $\nu > 4(1-\alpha)$. Hence, the outlier condition is violated for each $\nu \leq 4(1-\alpha)$. Further, the inlier condition requires $\mathbf{E}\big(F_{d,\nu}^{-\alpha/2}\big)$ to be finite, which is true if and only if $d > \alpha$. This means that the inlier condition is always satisfied for the multivariate t-distribution.

Finally, suppose that $\mathcal{R} = G_{d/(2\beta),2}^{1/(2\beta)}$ with $G_{d/(2\beta),2} \sim \text{Gamma}\big(\frac{d}{2\beta}, 2\big)$, which is the generating variate of the multivariate power-exponential distribution with shape parameter $\beta > 0$. In this case, we have that

$$\mathbf{E}\big(\mathcal{R}^{-\alpha}\big) = \mathbf{E}\Big(\big(\mathcal{R}^{2\beta}\big)^{-\frac{\alpha}{2\beta}}\Big) = 2^{-\frac{\alpha}{2\beta}}\,\frac{\Gamma\big(\frac{d-\alpha}{2\beta}\big)}{\Gamma\big(\frac{d}{2\beta}\big)}.$$

Hence, once again, d must exceed α, which is true for all $0 \leq \alpha \leq 1$. Further, all positive moments of the gamma distribution are finite. We conclude that both the inlier condition and the outlier condition are satisfied if the data are multivariate power-exponentially distributed.

Theorem 2 shows that the asymptotic covariance matrix of $\sqrt{n}\big(\hat{\theta} - \theta\big)$ is much simpler than the asymptotic covariance matrix of $\sqrt{n}\big(\widehat{\Sigma} - \Sigma\big)$. To the best of my knowledge, this crucial and highly relevant observation is made first by Tyler (1983). It follows that the asymptotic relative efficiency of $\hat{\theta}$ that is based on Tyler's M-estimator, compared to some power M-estimator for Σ with tail index $0 \leq \alpha < 1$, amounts to

$$\text{ARE} = \left(\frac{d^{\alpha}}{d + 2(1-\alpha)}\right)^2 \mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big).$$

We conclude that, in order to estimate θ, Tyler's M-estimator is preferable whenever

$$\mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big) > \left(\frac{d + 2(1-\alpha)}{d^{\alpha}}\right)^2.$$

In particular, Tyler's M-estimator is preferable compared to the sample covariance matrix, i.e., the power M-estimator with $\alpha = 0$, if $\mathbf{E}\big(\mathcal{R}^4\big) > (d+2)^2$. In some practical applications, it can happen that no choice of $\alpha < 1$ is favorable, which depends on the heaviness of the right tail of the distribution of $\mathcal{R}$ and the number d of dimensions.

An evident question is why we do not always use the ML-estimator for scatter, i.e., the power M-estimator given by Eq. 5, since from a theoretical point of view this is preferable to any (other) M-estimator. Thus, suppose that $\mathcal{R}$ is the generating variate of the re-scaled multivariate power-exponential distribution with shape parameter $1 - \alpha > 0$. Then, according to Theorem 1, we have that

$$\mathcal{R}^{2(1-\alpha)} \sim \text{Gamma}\left(\frac{d}{2(1-\alpha)}, \frac{2(1-\alpha)}{d^{\alpha}}\right).$$

By using the recurrence property of the gamma function, i.e., $\Gamma(x+1) = x\Gamma(x)$, we conclude that $\mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big) = d^{2(1-\alpha)}$. Hence, the asymptotic relative efficiency of $\hat{\theta}$ that is based on Tyler's M-estimator, compared to the ML-estimator, amounts to

$$\text{ARE} = \left(\frac{d}{d+2(1-\alpha)}\right)^{2},$$

provided that the random vector X, in fact, possesses a re-scaled multivariate power-exponential distribution with known shape parameter $1 - \alpha > 0$. The asymptotic relative efficiency is always lower than 1, which means that Tyler's M-estimator for scatter cannot be preferable to the ML-estimator for Σ. However, the problem is that, in most real-life applications, the generating distribution of X is unknown. In this case, the chosen "ML-estimator" actually represents an M-estimator, and if the data are heavy tailed or the number of dimensions is high, Tyler's M-estimator usually turns out to be the better alternative (Frahm & Jaekel 2010; Frahm et al. 2020).

4.2 A Simple Application

It seems worth illustrating a simple application of Theorem 2. Let Y be some random variable and X be a k-dimensional random vector such that the $(k+1)$-dimensional random vector $Z = (Y, X)$ has a multivariate normal distribution with covariance matrix

$$\Gamma = \begin{bmatrix} \varrho & \rho' \\ \rho & \Sigma \end{bmatrix} \in \mathcal{P}^{k+1}.$$

Here, ϱ symbolizes the variance of Y, $\Sigma \in \mathcal{P}^k$ is the covariance matrix of X, and $\rho \in \mathbb{R}^k$ is the vector of covariances between X and Y. Hence, the linear-regression equation of X and Y is $Y = \beta_0 + \beta' X + \varepsilon$ with $\beta_0 \in \mathbb{R}$ being the intercept and $\beta = (\beta_1, \beta_2, \ldots, \beta_k) \in \mathbb{R}^k$ the vector of regression coefficients.[9] It holds that $\beta =$

[9] Here, I use the traditional symbol "β" for the vector of regression coefficients. This is not to be confounded with the shape parameter β of the multivariate power-exponential distribution.

$\Sigma^{-1}\rho$, and thus β represents a scale-invariant function of Γ. The OLS-estimator for β is given by $\hat{\beta} = \widehat{\Sigma}^{-1}\hat{\rho}$, where $\widehat{\Sigma}$ and $\hat{\rho}$ are the sample moments, i.e., the power M-estimators with tail index $\alpha = 0$, respectively.[10] It is well known that

$$\sqrt{n}\big(\hat{\beta} - \beta\big) \longrightarrow N_k\big(0, \sigma_\varepsilon^2\Sigma^{-1}\big), \qquad n \longrightarrow \infty,$$

with $\sigma_\varepsilon^2 := \mathbf{Var}(\varepsilon) = \varrho - \beta'\rho$.

To calculate the asymptotic covariance matrix of $\sqrt{n}\big(\hat{\beta} - \beta\big)$ by Theorem 2, we have to set $\alpha = 0$. Since Z has a $(k+1)$-dimensional normal distribution, we have that $\mathcal{R}^2 = \chi^2_{k+1}$, which satisfies the general scaling condition required by Eq. 1, viz.

$$\mathbf{E}\big(\varphi(\mathcal{R}^2)\big) = \mathbf{E}\big(\mathcal{R}^2\big) = \mathbf{E}\big(\chi^2_{k+1}\big) = k+1.$$

Moreover, it holds that $\mathbf{E}\big(\chi^4_{k+1}\big) = (k+1)(k+3)$, and thus Theorem 2 leads us to the asymptotic covariance matrix $B = 2\mathcal{J}_\psi\big(\Sigma \otimes \Sigma\big)\mathcal{J}_\psi' = \sigma_\varepsilon^2\Sigma^{-1}$.

This result can readily be used to determine the asymptotic covariance matrix of $\sqrt{n}\big(\hat{\beta} - \beta\big)$ if Z has any other elliptical distribution or if $0 < \alpha \leq 1$, provided the moment conditions of Theorem 2 are satisfied. More precisely, by applying the power M-estimators with tail index $0 < \alpha \leq 1$, we leave the area of OLS-estimation and step into M-estimation of linear-regression coefficients. The asymptotic covariance matrix of $\sqrt{n}\big(\hat{\beta} - \beta\big)$ corresponds to

$$B = \frac{k+3}{k+1}\left(\frac{(k+1)^\alpha}{k+1+2(1-\alpha)}\right)^2 \mathbf{E}\big(\mathcal{R}^{4(1-\alpha)}\big)\,\sigma_\varepsilon^2\Sigma^{-1}.$$

In particular, if we use the Hettmansperger–Randles M-estimator for μ and Tyler's M-estimator for Γ, i.e., set $\alpha = 1$, then we obtain the simple expression

$$B = \frac{k+3}{k+1}\,\sigma_\varepsilon^2\Sigma^{-1}.$$

Obviously, this does not dependent on the generating distribution of Z. The power M-estimators with $\alpha = 1$ are preferable, compared to the sample moments, which are obtained by setting $\alpha = 0$, whenever $\mathbf{E}\big(\mathcal{R}^4\big) > (k+3)^2$.

Finally, note that also $\sigma_\varepsilon^2\Sigma^{-1}$ is a scale-invariant function of the covariance matrix or, to put it more generally, of the scatter matrix Γ of Z. Further, the coefficient of determination, i.e., $R^2 = \beta'\rho/\varrho$, is a scale-invariant function of Γ, too. This means that the above arguments hold true regarding the M-estimation of $\sigma_\varepsilon^2\Sigma^{-1}$ and R^2.

[10] It is implicitly assumed that $n > k$, where n is the sample size. Thus, $\widehat{\Sigma}$ is regular, almost surely.

5 Proofs

Proof of Theorem 1 According to Gómez et al. (1998), the density generator of the multivariate power-exponential distribution is proportional to $\exp\big(-\frac{1}{2}r^{2\beta}\big)$ with $\beta > 0$, which leads us to the density function

$$x \longmapsto \frac{1}{2^{\frac{d}{2\beta}}\pi^{\frac{d}{2}}}\frac{\Gamma\big(\frac{d}{2}+1\big)}{\Gamma\big(\frac{d}{2\beta}+1\big)}\sqrt{\det\Sigma^{-1}}\exp\bigg(-\frac{1}{2}\big[(x-\mu)'\Sigma^{-1}(x-\mu)\big]^{\beta}\bigg).$$

The density generator given by the theorem creates a density that is proportional to

$$\bigg(\frac{1-\alpha}{d^{\alpha}}\bigg)^{\frac{d}{2(1-\alpha)}}\sqrt{\det\Upsilon^{-1}}\exp\bigg(-\frac{1}{2}\big[(x-\mu)'\Upsilon^{-1}(x-\mu)\big]^{1-\alpha}\bigg)$$

with $\Upsilon = \big(\frac{1-\alpha}{d^{\alpha}}\big)^{1/(1-\alpha)}\Sigma$. Thus, we have that

$$\frac{1}{2^{\frac{d}{2(1-\alpha)}}\pi^{\frac{d}{2}}}\frac{\Gamma\big(\frac{d}{2}+1\big)}{\Gamma\big(\frac{d}{2(1-\alpha)}+1\big)} = c(d,\alpha)\bigg(\frac{1-\alpha}{d^{\alpha}}\bigg)^{\frac{d}{2(1-\alpha)}},$$

i.e.,

$$c(d,\alpha) = \frac{1}{\pi^{\frac{d}{2}}}\frac{\Gamma\big(\frac{d}{2}+1\big)}{\Gamma\big(\frac{d}{2(1-\alpha)}+1\big)}\bigg(\frac{d^{\alpha}}{2(1-\alpha)}\bigg)^{\frac{d}{2(1-\alpha)}}.$$

The resulting density function

$$x \longmapsto c(d,\alpha)\sqrt{\det\Sigma^{-1}}\exp\bigg(-\frac{d^{\alpha}}{2(1-\alpha)}\big[(x-\mu)'\Sigma^{-1}(x-\mu)\big]^{1-\alpha}\bigg)$$

corresponds to

$$x \longmapsto \frac{1}{2^{\frac{d}{2(1-\alpha)}}\pi^{\frac{d}{2}}}\frac{\Gamma\big(\frac{d}{2}+1\big)}{\Gamma\big(\frac{d}{2(1-\alpha)}+1\big)}\sqrt{\det\Upsilon^{-1}}\exp\bigg(-\frac{1}{2}\big[(x-\mu)'\Upsilon^{-1}(x-\mu)\big]^{1-\alpha}\bigg),$$

which is the density function of a multivariate power-exponential distribution with location vector μ, scatter matrix

$$\Upsilon = \bigg(\frac{1-\alpha}{d^{\alpha}}\bigg)^{\frac{1}{1-\alpha}}\Sigma,$$

and shape parameter $1-\alpha$. Hence, the generating variate of X is $\left(\frac{1-\alpha}{d^\alpha}\right)^{1/[2(1-\alpha)]}$ times the generating variate of a multivariate power-exponential distribution with shape parameter $1-\alpha$, i.e., $G_{d/[2(1-\alpha)],2}^{1/[2(1-\alpha)]}$. Put another way, $\mathcal{R}$ is such that

$$\mathcal{R}^{2(1-\alpha)} = \frac{1-\alpha}{d^\alpha} G_{d/[2(1-\alpha)],2} \sim \text{Gamma}\left(\frac{d}{2(1-\alpha)}, \frac{2(1-\alpha)}{d^\alpha}\right).$$

The ML-weight function w is given by

$$r^2 \longmapsto -2\,\frac{g'(r^2)}{g(r^2)} = \frac{\mathrm{d}\,d^\alpha r^{2(1-\alpha)}/(1-\alpha)}{\mathrm{d}\,r^2} = \left(\frac{r^2}{d}\right)^{-\alpha},$$

which corresponds to the power M-weight function with tail index $0 \le \alpha < 1$. Finally, it holds that

$$\mathbf{E}\big(\varphi(\mathcal{R}^2)\big) = \mathbf{E}\Big(\big(\mathcal{R}^2/d\big)^{-\alpha}\mathcal{R}^2\Big) = d^\alpha \mathbf{E}\big(\mathcal{R}^{2(1-\alpha)}\big) = d^\alpha \frac{d}{2(1-\alpha)} \frac{2(1-\alpha)}{d^\alpha} = d$$

for all $0 \le \alpha < 1$ and also $\mathbf{E}\big(\varphi(\mathcal{R}^2)\big) = \mathbf{E}\big(\big(\mathcal{R}^2/d\big)^{-1}\mathcal{R}^2\big) = d$ for $\alpha = 1$. Q.E.D.

Proof of Theorem 2 In the case of $0 \le \alpha < 1$, the joint asymptotic distribution of $\sqrt{n}\big(\hat{\mu} - \mu\big)$ and $\sqrt{n}\big(\widehat{\Sigma} - \Sigma\big)$ follows by the Central Limit Theorem together with Theorem 4 in Frahm et al. (2020), after substituting "$\mathbf{E}\big(V^{2(1-\alpha)}\big)$" with "$m^{1-\alpha}$" and thus observing that $\tau_2 = 1-\alpha$. Moreover, for $\alpha = 1$, the asymptotic distribution of $\sqrt{n}\big(\hat{\mu} - \mu\big)$ is provided by Hettmansperger and Randles (2002). Since ψ is scale invariant, Euler's theorem leads us to $\mathcal{J}_\psi \operatorname{vec}\Sigma = 0$, which makes the second part of A superfluous. This holds true also for the limiting case $\alpha = 1$, in which the number γ_2 is undefined. Hence, for $0 \le \alpha \le 1$, we have that

$$B = \gamma_1 \mathcal{J}_\psi \big(\mathbf{I}_{d^2} + \mathbf{K}_{d^2}\big)\big(\Sigma \otimes \Sigma\big)\mathcal{J}_\psi'.$$

Since Σ is symmetric, it holds that $\mathbf{K}_{d^2}\mathcal{J}_\psi' = \mathcal{J}_\psi'$ and thus $\mathcal{J}_\psi\big(\mathbf{I}_{d^2} + \mathbf{K}_{d^2}\big) = 2\mathcal{J}_\psi$, i.e., $B = 2\gamma_1 \mathcal{J}_\psi\big(\Sigma \otimes \Sigma\big)\mathcal{J}_\psi'$. According to Theorem 4 in Frahm et al. (2020), $\sqrt{n}\big(\hat{\mu} - \mu\big)$ and $\sqrt{n}\big(\widehat{\Sigma} - \Sigma\big)$ are asymptotically independent if $0 \le \alpha < 1$, whereas their asymptotic independence for $\alpha = 1$ is proved by Hettmansperger and Randles (2002). Since $\hat{\theta}$ is a function of $\widehat{\Sigma}$, and thus $\sqrt{n}\big(\hat{\theta} - \theta\big)$ is a function of $\sqrt{n}\big(\widehat{\Sigma} - \Sigma\big)$, $\sqrt{n}\big(\hat{\mu} - \mu\big)$ and $\sqrt{n}\big(\hat{\theta} - \theta\big)$ are asymptotically independent, too. Q.E.D.

Acknowledgments I would like to thank the anonymous referees very much for their valuable comments.

References

Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, *11*, 368–385.

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

Dümbgen, L., Pauly, M., & Schweizer, T. (2015). M-functionals of multivariate scatter. *Statistics Surveys*, *9*, 32–105.

Fang, K., Kotz, S., & Ng, K. (1990). *Symmetric multivariate and related distributions*. Chapman & Hall.

Frahm, G. (2004). *Generalized Elliptical Distributions: Theory and Applications*. Ph.D. thesis, University of Cologne.

Frahm, G. (2009). Asymptotic distributions of robust shape matrices and scales. *Journal of Multivariate Analysis*, *100*, 1329–1337.

Frahm, G., & Jaekel, U. (2010). A generalization of Tyler's M-estimators to the case of incomplete data. *Computational Statistics and Data Analysis*, *54*, 374–393.

Frahm, G., & Jaekel, U. (2015). Tyler's M-estimator in high-dimensional financial-data analysis. In K. Nordhausen, & S. Taskinen (Eds.) *Modern nonparametric, robust and multivariate methods*, Chap. 17, (pp. 289–305). Springer.

Frahm, G., Nordhausen, K., & Oja, H. (2020). M-estimation with incomplete and dependent multivariate data. *Journal of Multivariate Analysis*, *176*. https://doi.org/10.1016/j.jmva.2019.104569.

Gómez, E., Gómez-Villegas, M., & Marín, J. (1998). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics: Theory and Methods, 27*, 589–600.

Hallin, M., & Paindaveine, D. (2006). Parametric and semiparametric inference for shape: the role of the scale functional. *Statistics and Decisions*, *24*, 327–350.

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

Huber, P., & Ronchetti, E. (2009). *Robust statistics*. John Wiley.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya A*, *32*, 419–430.

Kent, J., & Tyler, D. (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *Journal of Applied Statistics*, *15*, 247–254.

Kent, J., & Tyler, D. (1991). Redescending M-estimates of multivariate location and scatter. *Annals of Statistics*, *19*, 2102–2119.

Mardia, K., & Jupp, P. (2000). *Directional statistics*. John Wiley.

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

Oja, H. (2003). Multivariate M-estimates of location and shape. In R. Höglund, M. Jäntti, & G. Rosenqvist (Eds.) *Statistics, econometrics and society. Essays in Honor of Leif Nordberg.*. Statistics Finland.

Paindaveine, D. (2008). A canonical definition of shape. *Statistics and Probability Letters*, *78*, 2240–2247.

Taskinen, S., Croux, C., Kankainen, A., Ollila, E., & Oja, H. (2006). Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices. *Journal of Multivariate Analysis*, *97*, 1219–1243.

Tyler, D. (1982). Radial estimates and the test for sphericity. *Biometrika*, *69*, 429–436.

Tyler, D. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika*, *70*, 411–420.

Tyler, D. (1987a). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, *15*, 234–251.

Tyler, D. (1987b). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika*, *74*, 579–589.

On Robust Estimators of a Sphericity Measure in High Dimension

Esa Ollila and Hyon-Jung Kim

Abstract The need to test (or estimate) sphericity arises in various applications in statistics, and thus the problem has been investigated in numerous papers. Recently, estimates of a sphericity measure are needed in high-dimensional shrinkage covariance matrix estimation problems, wherein the (oracle) shrinkage parameter minimizing the mean squared error (MSE) depends on the unknown sphericity parameter. The purpose of this chapter is to investigate the performance of robust sphericity measure estimators recently proposed within the framework of elliptically symmetric distributions when the data dimensionality, p, is of similar magnitude as the sample size, n. The population measure of sphericity that we consider here is defined as the ratio of the mean of the squared eigenvalues of the scatter matrix parameter relative to the mean of its eigenvalues squared. We illustrate that robust sphericity estimators based on the spatial sign covariance matrix (SSCM) or M-estimators of scatter matrix provide superior performance for diverse covariance matrix models compared to sphericity estimators based on the sample covariance matrix (SCM) when distributions are heavy-tailed and $n = O(p)$. At the same time, they provide equivalent performance when the data are Gaussian. Our examples also illustrate the important role that the sphericity plays in determining the attainable accuracy of the SCM.

Keywords Elliptical distributions · High-dimensional statistics · M-estimators of scatter matrix · Robust statistics · Sign covariance matrix · Sphericity parameter

E. Ollila (✉)
School of Electrical Engineering, Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
e-mail: esa.ollila@aalto.fi

H.-J. Kim
Faculty of Information Technology and Communication Sciences, Unit of Computing Sciences, Tampere University, Tampere, Finland
e-mail: hyon-jung.kim@tuni.fi

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_9

1 Introduction

Suppose we observe independent and identically distributed (i.i.d.) p-variate real-valued random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Then a commonly occurring problem in multivariate analysis is to analyze the degree of sphericity of the underlying sampling distribution. A random vector $\mathbf{z}$ is said to have a spherically symmetric distribution iff $\mathbf{z} =_d \mathbf{Pz}$ for all orthogonal $p \times p$ matrices $\mathbf{P}$ (i.e., $\mathbf{P}^\top \mathbf{P} = \mathbf{I}$), where $=_d$ should be read as "has the same distribution as." The probability density function (pdf) of a spherical random vector $\mathbf{z}$, given it exists, is of the form $g(\mathbf{z}^\top \mathbf{z})$, where $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ is a function, called the density generator (Fang et al. 1990), verifying $\int_0^\infty t^{p/2-1} g(t) \mathrm{d}t < \infty$.

In this chapter, we assume that the sample is generated from an (absolutely continuous) elliptically symmetric (ES) distribution. A random vector $\mathbf{x}$ with an ES distribution has same distribution as an affine transformation of a spherical random vector $\mathbf{z}$ (Fang et al. 1990):

$$\mathbf{x} =_d \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \mathbf{z}, \tag{1}$$

where a positive-definite symmetric $p \times p$ matrix $\boldsymbol{\Sigma}$ and a vector $\boldsymbol{\mu} \in \mathbb{R}^p$ are parameters of the ES distribution, called the scatter matrix and the symmetry center. Above $\boldsymbol{\Sigma}^{1/2}$ denotes the unique positive-definite symmetric matrix square root of $\boldsymbol{\Sigma}$. The pdf of $\mathbf{x}$, given it exists, is then of the form

$$f(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} g((\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})),$$

and we denote this case by $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$. For example, the multivariate normal (MVN) distribution, $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is a particular instance of the elliptical distribution obtained when $g(t) = (2\pi)^{-p/2} \exp(-t/2)$. The sphericity hypothesis is then true if and only if $\boldsymbol{\Sigma} \propto \mathbf{I}$.

Assuming that $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ has finite 2nd-order moments, then its mean vector is $\boldsymbol{\mu} = \mathsf{E}[\mathbf{x}]$ and its covariance matrix, $\mathrm{cov}(\mathbf{x}) = \mathsf{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]$, is

$$\mathrm{cov}(\mathbf{x}) = \sigma_{\mathrm{cov}} \cdot \boldsymbol{\Sigma} \quad \text{for} \quad \sigma_{\mathrm{cov}} = \frac{\mathsf{E}[\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|^2]}{p}, \tag{2}$$

which follows by applying the stochastic decomposition (1). Thus the scatter matrix $\boldsymbol{\Sigma}$ is proportional to the covariance matrix $\mathrm{cov}(\mathbf{x})$, given it exists. Note that, for some ES distributions (multivariate Cauchy distribution, for example), the covariance matrix does not exist, yet the scatter matrix is a well-defined parameter that determines the shape and orientation of the elliptical equidensity contours. Also note that in the MVN case, $\sigma_{\mathrm{cov}} = 1$.

Most measures of sphericity are based on discrepancy of eigenvalues λ_i of $\boldsymbol{\Sigma}$ from a scale parameter of $\boldsymbol{\Sigma}$, such as the mean of the eigenvalues:

$$\eta = \frac{\mathrm{tr}(\boldsymbol{\Sigma})}{p} = \frac{1}{p}\sum_{i=1}^{p} \lambda_i. \tag{3}$$

See Paindaveine (2008) for a detailed account on scale statistics. Formally, $\eta \equiv \eta(\boldsymbol{\Sigma})$ is a scale parameter if it verifies $\eta(\mathbf{I}) = 1$ and $\eta(a\boldsymbol{\Sigma}) = a\eta(\boldsymbol{\Sigma})$ for all $a > 0$. Furthermore, the shape matrix (or normalized covariance matrix) is defined by

$$\boldsymbol{\Lambda} = \frac{\boldsymbol{\Sigma}}{\eta} = \frac{p\boldsymbol{\Sigma}}{\mathrm{tr}(\boldsymbol{\Sigma})}, \tag{4}$$

and note that $\mathrm{tr}(\boldsymbol{\Lambda}) = p$. One commonly used measure of sphericity is

$$\gamma = \frac{p\,\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2} = \frac{\|\boldsymbol{\Lambda}\|_{\mathrm{F}}^2}{p} = \frac{\frac{1}{p}\sum_{i=1}^{p}\lambda_i^2}{\left(\frac{1}{p}\sum_{i=1}^{p}\lambda_i\right)^2}, \tag{5}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius matrix norm ($\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})}$ for any matrix $\mathbf{A}$) and $\mathrm{tr}(\cdot)$ denotes the matrix trace, $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{p} a_{ii}$. Thus the sphericity measure (5) is the ratio of the mean of the squared eigenvalues of $\boldsymbol{\Sigma}$ relative to the mean of its eigenvalues squared. Letting $s^2 = \frac{1}{p}\sum_{i=1}^{p}(\lambda_i - \eta)^2$ denote the sample variance of the eigenvalues, we may express γ as

$$\gamma = 1 + \frac{s^2}{\eta^2} = 1 + \frac{1}{p}\|\boldsymbol{\Lambda} - \mathbf{I}\|_{\mathrm{F}}^2.$$

The first identity illustrates that γ measures the degree of variability of the eigenvalues around their mean, while the second identity illustrates that γ measures distance of $\boldsymbol{\Sigma}$ from $c\mathbf{I}$ for any $c > 0$ (i.e., distance of $\boldsymbol{\Lambda}$ from $\mathbf{I}$). It is important to notice that γ is invariant to scaling of $\boldsymbol{\Sigma}$, and thus one may replace $\boldsymbol{\Sigma}$ in (5) by $c \cdot \boldsymbol{\Sigma}$ for any $c > 0$, e.g., the covariance matrix, without changing its value.

The sphericity parameter gets values in the range $[1, p]$ and attains its minimum if and only if $\boldsymbol{\Sigma}$ is a scaled identity matrix (so $\lambda_i = \lambda_j$) and its maximum for a rank one matrix. Indeed, if all the eigenvalues are identical, then their sample variance is $s^2 = 0$, and consequently, $\gamma = 1 + s^2/\eta^2 = 1$. On the other hand, if $\boldsymbol{\Sigma}$ is of rank 1, so it has only one non-zero eigenvalue, then the sample variance is $s^2 = \eta^2(p-1)$, and consequently, $\gamma = 1 + s^2/\eta^2 = p$. The fact that γ is lower-bounded by $\gamma \leq p$ is easiest seen by recalling the submultiplicativity of the matrix trace; namely, for any positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$, it holds that $\mathrm{tr}(\mathbf{AB}) \leq \mathrm{tr}(\mathbf{A})\,\mathrm{tr}(\mathbf{B})$. Thus $\mathrm{tr}(\boldsymbol{\Lambda}^2) \leq \mathrm{tr}(\boldsymbol{\Lambda})^2 = p^2$, and consequently, $\|\boldsymbol{\Lambda}\|^2/p = \mathrm{tr}(\boldsymbol{\Lambda}^2)/p \leq p$.

Let $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$ denote the sample covariance matrix (SCM). We note that $\mathbf{S}$ is an unbiased estimate of $\mathrm{cov}(\mathbf{x})$, so $\mathsf{E}[\mathbf{S}] = \mathrm{cov}(\mathbf{x})$ for any p-

variate distribution with finite 2nd-order moments. In cases when $\boldsymbol{\mu}$ is known, and assuming $\boldsymbol{\mu} = \mathbf{0}$ without any loss of generality (w.l.o.g.), the SCM is defined as $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top$. An obvious plug-in estimate of sphericity γ is then

$$\hat{\gamma}_{\text{John}} = \frac{p\,\text{tr}(\mathbf{S}^2)}{\text{tr}(\mathbf{S})^2}, \tag{6}$$

which is the sphericity statistic originally proposed and studied by John (1971, 1972). However, John's sphericity statistics fails to be consistent estimator of γ when $n = O(p^\delta)$, $0 < \delta \le 1$ (Srivastava 2005). At the same time, the estimator is highly non-robust. This has led to many authors consider sphericity tests that are robust (Hallin & Paindaveine 2006; Sirkiä et al. 2009; Tyler 1982) and/or consistent in the large sample regime, where both n and p are large, but often comparable in size (Chen et al. 2010a; Jung & Marron 2009; Ledoit & Wolf 2002; Paindaveine & Verdebout 2016; Srivastava 2005; Virta 2021; Zou et al. 2014). More recently, instead of constructing robust sphericity tests, the focus has shifted toward finding accurate estimates of the sphericity parameter γ under various conditions, e.g., in the works proposing high-dimensional shrinkage covariance matrix estimators as in Chen et al. (2010b), Zhang and Wiesel (2016), Ollila (2017), Ollila and Raninen (2019), Ollila et al. (2021).

This chapter is organized as follows. Section 3 reveals the role that the sphericity parameter γ plays in describing the attainable accuracy of the SCM in high dimensions, while Sect. 3 reviews estimators of sphericity parameter γ that are based on the SCM. Then Sects. 4 and 5 review robust sphericity estimators based on the spatial sign covariance matrix and M-estimators of scatter matrix, respectively. In Sect. 6, we investigate the performance of the considered sphericity estimators in diverse setups using simulations. Finally, Sect. 7 concludes.

2 On the Role of Sphericity on the Accuracy of SCM in High Dimension

Sphericity γ plays an important role in determining the accuracy of the SCM $\mathbf{S}$ in high dimensions. Namely, Ollila and Raninen (2019, Theorem 4) showed that the normalized MSE of $\mathbf{S}$ when sampling from an elliptical distribution $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ with finite 4th-order moments and unknown $\boldsymbol{\mu}$ is

$$\text{NMSE}(\mathbf{S}) \equiv \frac{\mathsf{E}\big[\|\mathbf{S} - \boldsymbol{\Sigma}\|_\text{F}^2\big]}{\|\boldsymbol{\Sigma}\|_\text{F}^2} = \left(1 + \frac{p}{\gamma}\right)\left(\frac{1}{n-1} + \frac{\kappa}{n}\right) + \frac{\kappa}{n}, \tag{7}$$

where κ is elliptical kurtosis parameter, defined formally in (12). The expression of NMSE($\mathbf{S}$) when $\boldsymbol{\mu}$ is known ($\boldsymbol{\mu} = \mathbf{0}$) is given in Ollila (2017, Lemma 1), while the NMSE of $\mathbf{S}$ when sampling from complex ES distributions can be found in Raninen et al. (2021a).

In order to obtain more insight on the effect of sphericity γ on MSE of $\mathbf{S}$, we consider the following instructive example, where the scatter matrix parameter has an autoregressive model (**AR(1)**) structure

$$(\boldsymbol{\Sigma})_{ij} = \eta \varrho^{|i-j|}, \tag{8}$$

where η is the scale (3) and ϱ is the correlation parameter, $\varrho \in (-1, 1)$. The sphericity is then (Raninen et al. 2021b, Prop. 3)

$$\gamma = \frac{p - p\varrho^4 - 2\varrho^2 + 2(\varrho^2)^{p+1}}{p(\varrho^2 - 1)^2}. \tag{9}$$

Figure 1a displays the NMSE when sampling from p-variate normal distribution, $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for $p = 40$ fixed. In this case, $\kappa = 0$. As can be noted, the accuracy of the SCM $\mathbf{S}$ depends heavily on value of γ. When $\gamma \approx 1$ (i.e., the distributions are close to being spherical, so $\varrho \approx 0$), the NMSE is largest and rises steeply when $n < p$.

Next we consider the large sample asymptotic limit,

$$c = \frac{p}{n} \to c_0, \quad 0 < c_0 < \infty, \quad \text{as } p, n \to \infty. \tag{10}$$

Noting that $\gamma \to \gamma_0 = (1 + \varrho^2)/(1 - \varrho^2)$ as $p \to \infty$, and using (7), it immediately follows that the limiting NMSE under asymptotic regime (10) is

$$\text{NMSE}(\mathbf{S}) \to \frac{1 - \varrho^2}{1 + \varrho^2}(1 + \kappa)c_0 = \frac{1 + \kappa}{\gamma_0} c_0.$$

Since the limit is a positive constant, it follows that $\mathbf{S}$ *is not a consistent estimator* of $\boldsymbol{\Sigma}$ possessing an AR(1) structure, unless $c = p/n \to 0$. This is illustrated in Fig. 1b that displays the limiting NMSE as a function of γ_0 for different cases of c_0 ranging from $1/10$ to 10. Again the limiting NMSE is largest when $\boldsymbol{\Sigma}$ is close to being spherical ($\varrho \approx 0$). Moreover, if $c_0 > 1$, the limiting NMSE can be very large.

3 Sphericity Estimator Based on the Sample Covariance Matrix

As was already mentioned, $\hat{\gamma}_{\text{John}}$ defined in (6) is not a consistent estimator of γ when $n = O(p^\delta)$, $0 < \delta \leq 1$. Srivastava (2005) showed that a consistent estimator of γ can be obtained using

$$\hat{\gamma} = \frac{(n-1)^2}{(n-2)(n+1)} \left(\frac{p \,\text{tr}(\mathbf{S}^2)}{\text{tr}(\mathbf{S})^2} - \frac{n}{n-1} \frac{p}{n} \right) = b_n(\hat{\gamma}_{\text{John}} - a_n c), \tag{11}$$

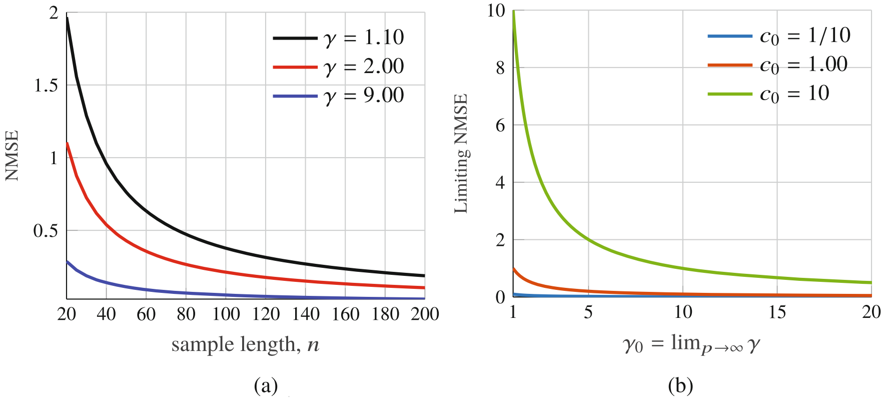


Fig. 1 The effect of sphericity γ on NMSE of SCM $\mathbf{S}$ when sampling from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ having an AR(1) structure. (**a**) Displays the NMSE for $p = 40$, while (**b**) displays the limiting NMSE as a function of limiting sphericity value, $\gamma \to \gamma_0$, as $p/n \to c_0$ as $p, n \to \infty$. (**a**) $p = 40$, n varies. (**b**) $p/n \to c_0$ as $p, n \to \infty$

where $c = p/n$, under the assumption that the samples are generated from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This result was extended in Ollila and Raninen (2019) for general elliptical distributions $\mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ with finite 4th-order moments.

Before discussing the approach of Ollila and Raninen (2019), we need to introduce some notation. We recall that the elliptical kurtosis (Muirhead 1982) is defined by

$$\kappa = \frac{\mathsf{E}\big[\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\|^4\big]}{\big(\mathsf{E}\big[\|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\|^2\big]\big)^2} \cdot \frac{p}{p+2} - 1, \tag{12}$$

where we assume that the elliptical random vector $\mathbf{x}$ has finite 4th-order moments. The elliptical kurtosis shares properties similar to the kurtosis of a real random variable. Namely if $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\kappa = 0$. This result becomes more obvious when one notices the following relationship of κ with the marginal (excess) kurtosis, $\mathrm{kurt}(x_i)$, of any component of x_i of $\mathbf{x} \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ (Ollila et al. 2021, Lemma 3):

$$\kappa = \frac{1}{3}\cdot \mathrm{kurt}(x_i) = \frac{1}{3}\left(\frac{\mathsf{E}[(x_i-\mu_i)^4]}{\mathsf{E}[(x_i-\mu_i)^2]^2} - 3\right). \tag{13}$$

The lower bound for the kurtosis parameter is $\kappa^{\mathrm{LB}} = -2/(p+2)$ (Bentler & Berkane 1986).

The generalized sphericity estimator proposed in Ollila and Raninen (2019) is defined by

$$\hat{\gamma} = \hat{b}_n(\hat{\gamma}_{\mathrm{John}} - \hat{a}_n c), \tag{14}$$

where $\hat{b}_n = a_n(\hat{\kappa})$ and $\hat{a}_n = a_n(\hat{\kappa})$ are constants defined as

$$\hat{a}_n = \frac{n}{n+\hat{\kappa}}\left(\frac{n}{n-1} + \hat{\kappa}\right) \quad \text{and} \quad \hat{b}_n = \frac{(\hat{\kappa}+n)(n-1)^2}{(n-2)(3\hat{\kappa}(n-1)+n(n+1))},$$

and $\hat{\kappa}$ is an estimate of elliptical kurtosis κ described in Ollila and Raninen (2019, Sect. 4). The estimator (14) reduces to (11) when using $\hat{\kappa} = 0$, i.e., by assuming the data have MVN distribution. The benefit of (14) is that it does not assume that data follow any specific ES distribution. This is illustrated in Sect. 6.

4 Sphericity Estimator Based on the Spatial Sign Covariance Matrix

The spatial sign covariance matrix (SSCM) has been used for constructing robust estimates or tests of sphericity in many works, see Hallin and Paindaveine (2006), Zou et al. (2014), Paindaveine and Verdebout (2016), Zhang and Wiesel (2016), Ollila and Raninen (2019), Raninen et al. (2021b), Ollila and Breloy (2022).

The SSCM is an estimate of the shape matrix (or normalized covariance matrix) (4). The scaled[1] SSCM is defined as (Visuri et al. 2000)

$$\hat{\mathbf{\Lambda}} = \frac{p}{n}\sum_{i=1}^{n} \frac{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top}{\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|^2}, \tag{15}$$

where $\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}\|$ is the sample spatial median (Brown 1983). When $\boldsymbol{\mu}$ is known ($\boldsymbol{\mu} = \mathbf{0}$), the SSCM is defined as $\hat{\mathbf{\Lambda}} = \frac{p}{n}\sum_{i=1}^{n} \frac{\mathbf{x}_i\mathbf{x}_i^\top}{\|\mathbf{x}_i\|^2}$.

One of the major selling points of SSCM is its impeccable robustness properties: it possesses the highest possible breakdown point of 1 with fixed location (Magyar & Tyler 2014) and breakdown point of 1/2 when using the spatial median to estimate the location (Croux et al. 2010). This can be contrasted to M-estimators of scatter for which the best possible breakdown point is $1/p$ and obtained by Tyler's M-estimator (Dümbgen & Tyler 2005).

Raninen et al. (2021b) studied the following estimate of sphericity based on the SSCM (when $\boldsymbol{\mu}$ is known),

$$\hat{\gamma} = \frac{n}{n-1}\left(\frac{\|\hat{\mathbf{\Lambda}}\|_{\mathrm{F}}^2}{p} - \frac{p}{n}\right), \tag{16}$$

[1] The common definition is without the multiplier p.

and showed that (16) is asymptotically (as $p \to \infty$) unbiased when sampling from elliptical distribution under the assumption $\gamma/p \to 0$ as $p \to \infty$. This assumption is sufficiently general and holds for many scatter matrix models (Raninen et al. 2021b, Prop. 3). For example, for AR(1) covariance matrix (8), the sphericity given in (9) verifies $\gamma = O(1) = o(p)$.

When location is not known, centering via the spatial median $\hat{\boldsymbol{\mu}}$ results in nonnegligible error in the sphericity estimate in the high-dimensional setting, and a bias correction is needed (Zou et al. 2014). In this case, the sphericity estimate is

$$\hat{\gamma}^* = \hat{\gamma} - p\delta, \tag{17}$$

where

$$\delta = \frac{1}{n^2} \cdot \left(2 - 2\frac{q_2}{q_1^2} + \left(\frac{q_2}{q_1^2}\right)^2\right) + \frac{1}{n^3} \cdot \left(8\frac{q_2}{q_1^2} - 6\left(\frac{q_2}{q_1^2}\right)^2 + 2\frac{q_2 q_3}{q_1^5} - 2\frac{q_3}{q_1^3}\right)$$

and $q_m = (1/n)\sum_{i=1}^n \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|^{-m}$. When computational simplicity is desired, it is also possible to use $\delta \approx n^{-2} + 2n^{-3}$, which is often a good approximation (Zou et al. 2014).

5 Sphericity Estimators Based on M-Estimators of Scatter

Assume now that $n > p$ and $\boldsymbol{\mu} = \mathbf{0}$, so we consider that the location is known. An M-estimator of scatter matrix (Maronna 1976) is defined as positive-definite symmetric $p \times p$ matrix $\hat{\boldsymbol{\Sigma}}$ that solves an estimating equation

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n u(\mathbf{x}_i^\top \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top, \tag{18}$$

where $u : [0, \infty) \to [0, \infty)$ is a non-increasing weight function. An M-estimator is an adaptively weighted SCM with weights determined by function $u(\cdot)$. Using $u(t) = 1\ \forall t$, the solution is the SCM $\mathbf{S}$, and for $u(t) = -2 \cdot g'(t)/g(t)$, one obtains the maximum likelihood estimate (MLE) of $\boldsymbol{\Sigma}$ of centered elliptical distribution $\mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$. To guarantee existence of the solution, it is required that the data verify the condition stated in Kent and Tyler (1991).

Define a function $\psi(t) = u(t)t$ and

$$\psi_1 = \frac{1}{p(p+2)}\mathsf{E}\left[\psi\left(\frac{r^2}{\sigma}\right)^2\right], \tag{19}$$

where the statistical expectation is w.r.t. the distribution of $r = \|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\|$, and $\sigma > 0$ is a solution to an equation

$$\mathsf{E}\left[\psi\left(\frac{r^2}{\sigma}\right)\right] = p. \tag{20}$$

Note that population parameter corresponding to $\hat{\boldsymbol{\Sigma}}$ is $\sigma\boldsymbol{\Sigma}$ when $x \sim \mathcal{E}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, i.e., an M-estimator is Fisher consistent to scatter matrix up to a scalar σ.

Then Ollila et al. (2021) proposed the following estimator:

$$\hat{\gamma} = \hat{b}_n\left(\frac{p\,\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^2)}{\mathrm{tr}(\hat{\boldsymbol{\Sigma}})^2} - \hat{\psi}_1\hat{a}_n\frac{p}{n}\right), \tag{21}$$

where

$$\hat{a}_n = \frac{n}{n + \hat{\psi}_1 - 1} \quad \text{and} \quad \hat{b}_n = \frac{n}{n-1}\left(\frac{n - 1 + \hat{\psi}_1}{n - 1 + 3\hat{\psi}_1}\right), \tag{22}$$

and $\hat{\psi}_1$ is an estimate of ψ_1. Below we will discuss possible choices of weight functions $u(\cdot)$ that may be used to construct robust sphericity estimators.

Huber's weight function is defined as

$$u_{\mathrm{H}}(t; c) = \begin{cases} 1/b, & \text{for } t \leqslant c^2 \\ c^2/(tb), & \text{for } t > c^2 \end{cases}, \tag{23}$$

where $c > 0$ is a user-defined tuning constant that determines the robustness and efficiency of the estimator and b is a scaling factor defined by

$$b = F_{\chi^2_{p+2}}(c^2) + c^2(1 - F_{\chi^2_p}(c^2))/p,$$

where $F_{\chi^2_p}(\cdot)$ denotes the cumulative distribution function (cdf) of chi-squared distribution with p d.o.f. This choice of b guarantees that $\hat{\boldsymbol{\Sigma}}$ is Fisher consistent to the covariance matrix when sampling from MVN distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, i.e., $\sigma = 1$ when $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Since $r^2 = \|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\|^2$ has a χ^2_p-distribution when $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, the tuning constant c^2 is chosen as qth upper quantile of χ^2_p-distribution:

$$q = \Pr(r^2 \leq c^2) \Leftrightarrow F^{-1}_{\chi^2_p}(q) = c^2 \tag{24}$$

for some $q \in (0, 1]$. We use $q = 0.7$ in the simulations. Computation of the estimate $\hat{\psi}_1$ in this case is discussed in detail in Ollila et al. (2021, Section IVB).

Another popular choice is *MVT-weight function* (Kent & Tyler 1991):

$$u_{\mathrm{T}}(t;\nu) = \frac{p+\nu}{\nu+t} \tag{25}$$

in which case the corresponding M-estimator $\hat{\mathbf{\Sigma}}$ is also the MLE of scatter matrix of multivariate t (MVT) distribution with $\nu > 0$ degrees of freedom (d.o.f.). The d.o.f. parameter ν is estimated adaptively from the data using the method described by Ollila et al. (2021, Algorithm 1) and $\hat{\psi}_1 = (p+\hat{\nu})/(2+p+\hat{\nu})$.

Finally, another classic choice with nice robustness properties is Tyler's (Tyler 1987) M-estimator, in which case the weight function is

$$u_{\mathrm{Tyl}}(t) = \frac{p}{t}. \tag{26}$$

Both Huber's and MVT-weight functions yield Tyler's weight function as special cases; namely, for $\nu \to 0$, one notices that $u_{\mathrm{T}}(t;\nu \to 0) = u_{\mathrm{Tyl}}(t)$, and in the limit case, as $c \to 0$, Huber's weight function tends to Tyler's weight function. In this case, $\hat{\psi}_1 = p/(p+2)$.

6 Simulation Studies

The purpose of this simulation study is to compare the performance of different sphericity estimators when the covariance matrix has different structures and the data is from different types (light-tailed or heavy-tailed) elliptical distributions. The estimators included in the study are:

- γ-MVN: estimator (11) using SCM and assuming MVN distribution
- γ-Ell: estimator (14) using SCM and assuming elliptical distribution
- γ-SSCM: estimator (16) using SSCM
- γ-MVT, -Tyl, -Hub: estimator (21) using an M-estimator of scatter based on MVT, Tyler's or Huber's weight function, respectively

We assume that $\boldsymbol{\mu}$ is known and used the noncentered estimator in each case. For Huber's estimator, we use a fixed value for c^2 using $q = 0.7$ in (24). Recall that for MVT-weight function, the d.o.f. parameter ν is adaptively estimated from the data.

In the first experiment, the scatter matrix parameter $\mathbf{\Sigma}$ has an AR(1) structure (8). Recall that $\varrho = 0$ implies that the distribution is spherical ($\mathbf{\Sigma} = \eta\mathbf{I}$), so $\gamma = 1$, while as ϱ gets larger than 0, the distribution becomes distinctively non-spherical, with $\gamma \to p$ as $\varrho \to 1$. Performance of estimators is tested using different values of ϱ.

Figure 2 shows the mean values of different sphericity estimators as a function of n in the case that $p = 50$, and the data are from a heavy-tailed MVT distribution with $\nu = 3$ degrees of freedom. We can observe from Fig. 2a that in the spherical

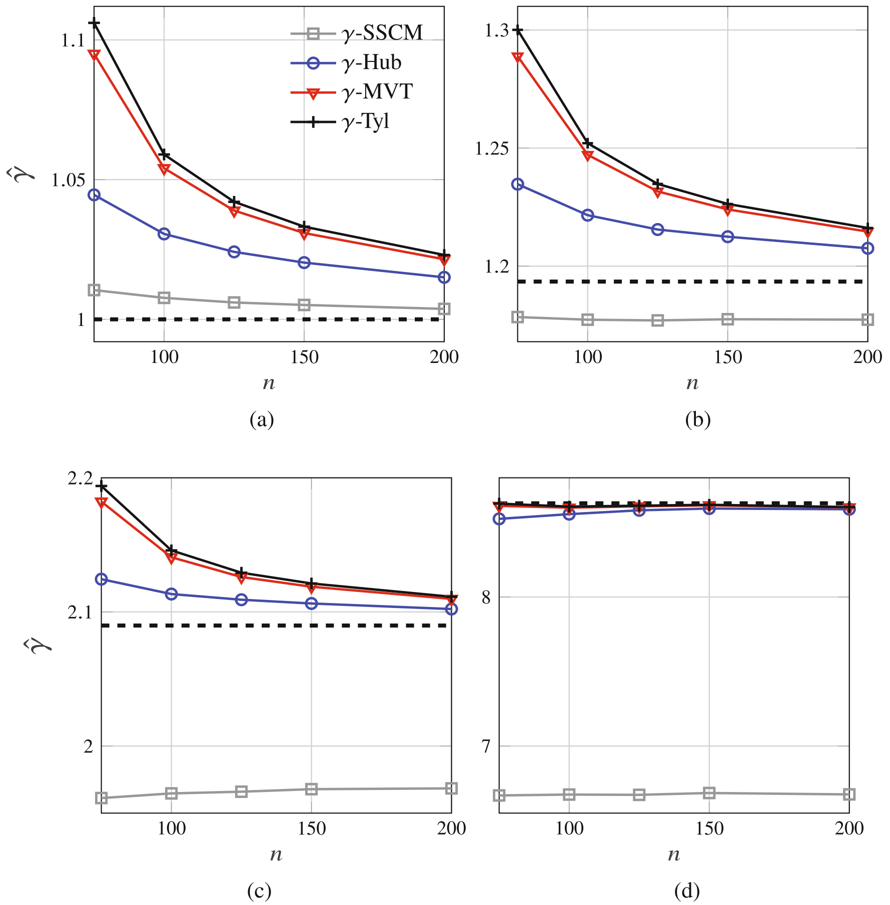


Fig. 2 Mean of sphericity estimates when sampling from a MVT distribution with $\nu = 3$ d.o.f. when $\boldsymbol{\Sigma}$ has an AR(1) structure; $p = 50$ and 5000 MC trials. Dashed line corresponds to true sphericity . (**a**) $\varrho = 0$. (**b**) $\varrho = 0.3$. (**c**) $\varrho = 0.6$. (**d**) $\varrho = 0.9$

case ($\varrho = 0$), the SSCM-based sphericity estimator is performing clearly the best. However, as ϱ increases, its performance gradually deteriorates, and at $\varrho = 0.9$, it provides severely downward biased estimate of sphericity. This does not come as a surprise, since in Raninen et al. (2021b, Theorem 2) it was shown that the bias of the SSCM is of the order of the sphericity, $\gamma = \mathrm{tr}(\boldsymbol{\Lambda}^2)/p$, and becomes negligible as the dimension increases assuming $\gamma = o(p)$. Since γ measures how close the shape matrix is to an identity matrix, the implication is that the bias of the SSCM is smaller for approximately spherical covariance matrices (γ small) and larger for spiked covariance matrices, where there are only a few large eigenvalues (γ large).

When we compare the performance of sphericity estimators based on M-estimators, we notice that an M-estimator based on Huber's function is generally

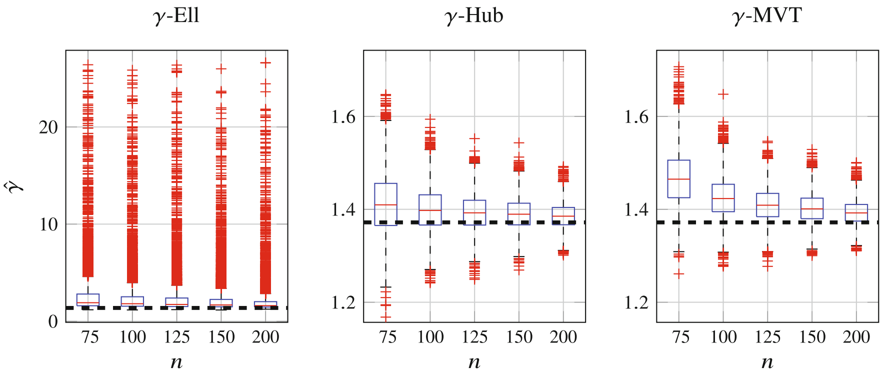


Fig. 3 Boxplots when sampling from a MVT distribution with $\nu = 3$ d.o.f. when $\mathbf{\Sigma}$ has an AR(1) structure with $\varrho = 0.4$; $p = 50$ and 5000 MC trials. Dashed line corresponds to true sphericity γ

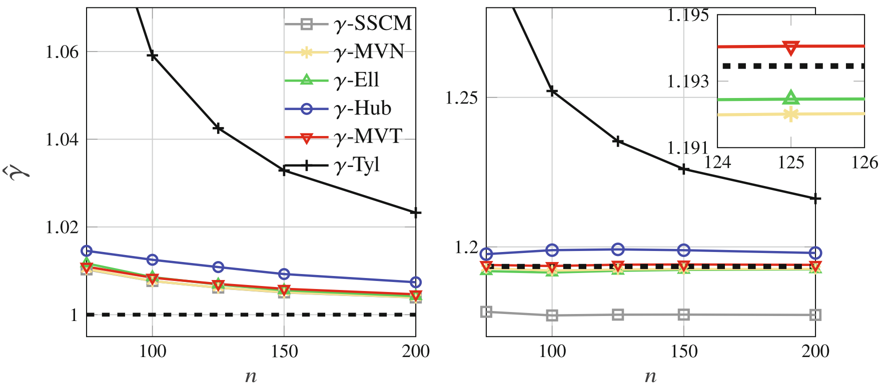
performing very well for a large range of ϱ values. The performance of sphericity estimator based on MVT weight is not on par which may be related with the property that the method estimates the degrees of freedom parameter ν adaptively from the data, and obtaining an accurate estimate of ν is a difficult task in heavy-tailed setting ($\nu = 3$).

Sphericity estimators γ-SCM and γ-Ell are left out from plots in Fig. 2 due to their poor performance. This is not surprising as they require finite 4th-order moments, which does not hold for MVT distribution with $\nu = 3$ d.o.f. To obtain a better view about the variability of sphericity estimators, Fig. 3 displays boxplots for the non-robust γ-Ell and the robust γ-Hub, and γ-MVT estimators when $\varrho = 0.4$. As can be noted, γ-Ell that uses the SCM-based estimator has large variability (due to heavy-tailed data), and it is clearly not consistent. The robust methods, γ-Hub and γ-MVT, on the other hand are providing accurate and consistent estimation of γ.

Next we consider the same setting, but the data are generated from MVN distribution. Figure 4 displays the performance of all estimators for $\varrho = 0$ and $\varrho = 0.3$. In the spherical case, four of the estimators, namely γ-MVN, -Ell, -SSCM, -MVT, are all providing similar top performance. When $\mathbf{\Sigma}$ becomes non-spherical (case $\varrho = 0.3$), the performance of γ-SSCM deteriorates, and it provides clearly downward biased estimate of sphericity. Surprisingly, γ-MVT has slightly better accuracy than γ-MVN that assumes Gaussianity, which is also illustrated in the zoomed-in subplot of Fig. 4b. This is probably due to the used adaptive estimation of the d.o.f. parameter ν used by the M-estimator of scatter with MVT-weight function.

Next we assume that scatter matrix $\mathbf{\Sigma}$ has the compound symmetry (**CS**) structure:

$$(\mathbf{\Sigma})_{ij} = \begin{cases} \eta\varrho, & \text{for } i \neq j \\ \eta, & \text{for } i = j \end{cases}, \tag{27}$$



(a)

(b)

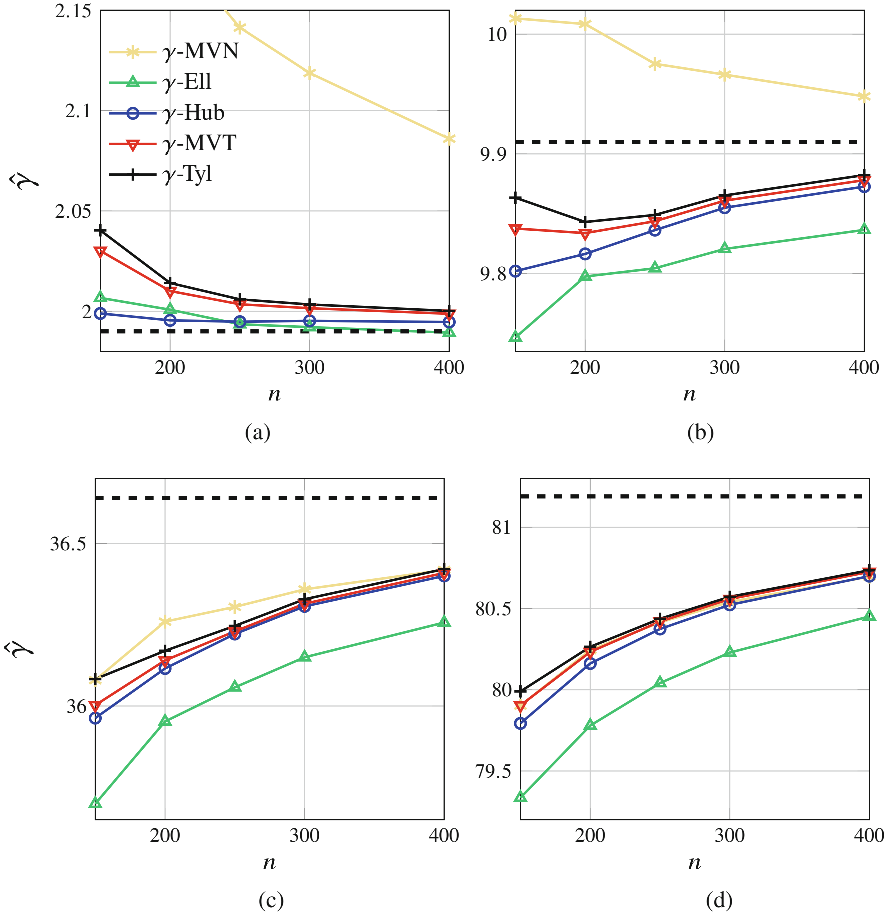
Fig. 4 Mean of sphericity estimates when sampling from a MVN distribution, when $\mathbf{\Sigma}$ has an AR(1) structure; $p = 50$ and 5000 MC trials. Dashed line corresponds to true sphericity γ. (**a**) Spherical case: $\varrho = 0$. (**b**) $\varrho = 0.3$

where $\varrho \in (-(p-1)^{-1}, 1)$ is the correlation parameter and η is the scale parameter, which are both fixed. Note that $\varrho = 1$ implies that $\mathbf{\Sigma}$ is of rank 1 (so $\gamma = p$), and when $\varrho = 0$, the distribution is spherical (so $\gamma = 1$). The structure of $\mathbf{\Sigma}$ implies that $\mathbf{x}$ consists of (since $\mathbf{\Sigma} \propto \text{cov}(\mathbf{x})$) equally correlated variables with equal variances. The eigenvalues of $\mathbf{\Sigma}$ are $\lambda_1 = \eta(\varrho p + 1 - \varrho)$ with multiplicity 1 and $\lambda_2 = \eta(1 - \varrho)$ with multiplicity $p - 1$. The restriction on $\varrho > -1/(p-1)$ is thus needed for $\mathbf{\Sigma}$ to be positive definite (so $\lambda_i > 0$). The CS model is an example of a spiky spectral distribution, where there exists a large concentration of small eigenvalues and a single large eigenvalue. The spectral gap, so separation of the largest eigenvalue from second largest eigenvalue grows with dimension. Since sphericity is invariant to scaling of $\mathbf{\Sigma}$, we can assume w.l.o.g. $\eta = 1$ and compute γ as

$$\gamma = \frac{1}{p}\sum_{i=1}^{p}\lambda_i^2 = (\varrho p + 1 - \varrho)^2 + (p-1)(1-\varrho)^2 = 1 + (p-1)\varrho^2,$$

and thus $\gamma = O(p)$. Since $\gamma \neq o(p)$, the SSCM-based sphericity estimator is not useful for this model.

Performance of sphericity estimators is now tested using different values of ϱ. We generated data from a MVT distribution with $\nu = 9$ d.o.f., which is only mildly heavier tailed than the MVN distribution, having kurtosis $\text{kurt}(x_i) = 1.2$. Figure 5 displays the performance for several cases of correlation parameter ϱ when the dimension is $p = 100$. As can be noted, in the nearly spherical case ($\varrho = 0.1$), all estimators are performing relatively well. As expected, γ-MVN has the worst performance (as Gaussianity assumption is not valid), while γ-Ell showcases the highest accuracy, especially for large sample lengths. When ϱ increases (and hence the spectral gap increases), the estimators start to behave very similarly and attain



(a)

(b)

(c)

(d)

Fig. 5 Mean of sphericity estimates when sampling from a MVT distribution with $\nu = 9$ d.o.f., when $\boldsymbol{\Sigma}$ has a CS structure; $p = 100$ and average is over 5000 MC trials. Dotted line corresponds to true sphericity. (**a**) $\varrho = 0.1$. (**b**) $\varrho = 0.3$. (**c**) $\varrho = 0.6$. (**d**) $\varrho = 0.9$

very similar performance when $\varrho = 0.9$. However, γ-Ell underestimates the true sphericity more than others. This is probably due to the fact that for spiked spectrum model, the kurtosis parameter is more difficult to estimate.

We excluded γ-SSCM from plots in Fig. 5 since it is not a consistent estimator of γ when $\gamma = O(p)$. To get an idea of its bias, we display boxplots of γ-Ell, γ-SSCM, and γ-Hub in Fig. 6 in the case that $\varrho = 0.6$. As can be noted, the sphericity estimator based on the SSCM is severely downward biased and fails to provide a sensible estimator of sphericity. Indeed, the condition $\gamma = o(p)$ required by SSCM implies that $\gamma \ll p$ in large dimensions, and essentially means that the eigenvalues of $\boldsymbol{\Sigma}$ tend to be similar as p grows. This is not the case for CS model whose first eigenvalue is much larger than the others in high dimensions (Fig. 6).

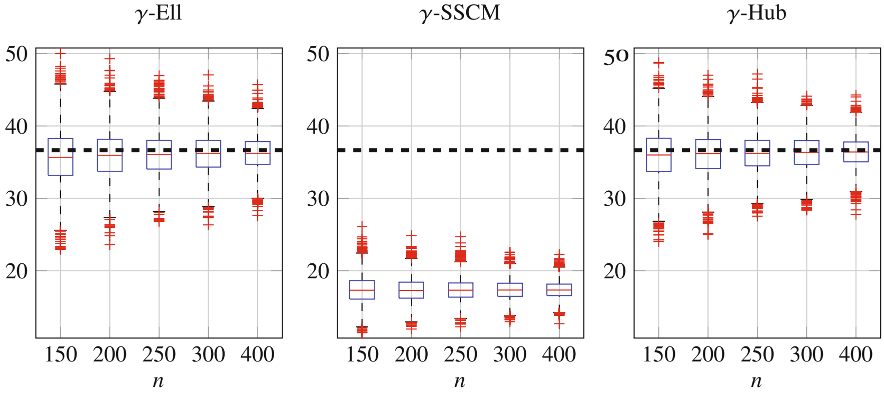


Fig. 6 Boxplots when sampling from a MVT distribution with $\nu = 9$ d.o.f. when $\mathbf{\Sigma}$ has an CS structure; $\varrho = 0.6$, $p = 100$, and 5000 MC trials. Dashed line corresponds to the true sphericity

7 Conclusions

In this chapter, the performance of robust sphericity estimators was tested when $n = O(p)$. It was observed that sphericity estimators based on M-estimators provide the best performance when the distribution is heavy-tailed and are on par with the SCM-based sphericity estimators when sampling from a MVN distribution. Thus when $n = O(p)$, the sphericity estimators based on M-estimators can always be favored.

M-estimators are however not applicable when $n < p$. If they would be, then that would contradict with the finding of Tyler (2010) that affine equivariant scatter matrices (assuming here known location) must be proportional to the sample covariance matrix when $n \leq p$. In this case, the SSCM-based sphericity estimator is useful as it can be applied in high-dimensional low-sample-size ($n \ll p$) regime. It is not consistent however in spiked models with a large spectral gap (which often implies that $\gamma = O(p)$). It was also observed that it provided severely downward biased estimate when the covariance matrix was distinctively non-spherical ($\gamma \gg 1$, e.g., as in AR(1) model for large ϱ (cf. Fig. 2c,d). In order to reduce the bias of the SSCM, we therefore recommend using a bias correction to its eigenvalues, e.g., using the method proposed recently in Raninen and Ollila (2021) or Dürre et al. (2017), in order to improve its accuracy.

References

Bentler, P. M., & Berkane, M. (1986). Greatest lower bound to the elliptical theory kurtosis parameter. *Biometrika*, *73*(1), 240–241.

Brown, B. (1983). Statistical Uses of the Spatial Median. *Journal of the Royal Statistical Society: Series B*, *45*(1), 25–30.

Chen, S. X., Zhang, L.-X., & Zhong, P.-S. (2010a). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, *105*(490), 810–819.

Chen, Y., Wiesel, A., Eldar, Y. C., & Hero, A. O. (2010b). Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, *58*(10), 5016–5029.

Croux, C., Dehon, C., & Yadine, A. (2010). The k-step spatial sign covariance matrix. *Advances in Data Analysis and Classification*, *4*(2), 137–150.

Dümbgen, L., & Tyler, D. E. (2005). On the breakdown properties of some multivariate m-functionals. *Scandinavian Journal of Statistics*, *32*(2), 247–264.

Dürre, A., Fried, R., & Vogel, D. (2017). The spatial sign covariance matrix and its application for robust correlation estimation. *Austrian Journal of Statistics*, *46*(3–4), 13–22.

Fang, K.-T., Kotz, S., & Ng, K.-W. (1990). *Symmetric multivariate and related distributions*. London: Chapman and Hall.

Hallin, M., & Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape I. Optimal rank-based tests for sphericity. *Annals of Statistics*, *34*(6), 2707–2756.

John, S. (1971). Some optimal multivariate tests. *Biometrika*, *58*(1), 123–127.

John, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika*, *59*(1), 169–173.

Jung, S., & Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Annals of Statistics*, *37*(6B), 4104–4130.

Kent, J. T., & Tyler, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Annals of Statistics*, *19*(4), 2102–2119.

Ledoit, O., & Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, *30*(4), 1081–1102.

Magyar, A. F., & Tyler, D. E. (2014). The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. *Biometrika*, *101*(3), 673–688.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Annals of Statistics*, *5*(1), 51–67.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory* (704 p.). New York: Wiley.

Ollila, E. (2017). Optimal high-dimensional shrinkage covariance estimation for elliptical distributions. In *Proc. European Signal Processing Conference (EUSIPCO 2017)*, (pp. 1689–1693). Kos, Greece.

Ollila, E., & Breloy, A. (2022). Regularized tapered sample covariance matrix. *IEEE Transactions on Signal Processing*, *70*, 2306–2320.

Ollila, E., Palomar, D. P., & Pascal, F. (2021). Shrinking the eigenvalues of M-estimators of covariance matrix. *IEEE Transactions on Signal Processing*, *69*, 256–269.

Ollila, E., & Raninen, E. (2019). Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions. *IEEE Transactions on Signal Processing*, *67*(10), 2707–2719.

Paindaveine, D. (2008). A canonical definition of shape. *Statistics & Probability Letters*, *78*(14), 2240–2247.

Paindaveine, D., & Verdebout, T. (2016). On high-dimensional sign tests. *Bernoulli*, *22*(3), 1745–1769.

Raninen, E., & Ollila, E. (2021). Bias adjusted sign covariance matrix. *IEEE Signal Processing Letters*, *29*, 339–343.

Raninen, E., Ollila, E., & Tyler, D. E. (2021a). On the variability of the sample covariance matrix under complex elliptical distributions. *IEEE Signal Processing Letters*, *28*, 2092–2096.

Raninen, E., Tyler, D. E., & Ollila, E. (2021b). Linear pooling of sample covariance matrices. *IEEE Transactions on Signal Processing*, *70*, 659–672.

Sirkiä, S., Taskinen, S., Oja, H., & Tyler, D. E. (2009). Tests and estimates of shape based on spatial signs and ranks. *Journal of Nonparametric Statistics*, *21*(2), 155–176.

Srivastava, M. S. (2005). Some tests concerning the covariance matrix in high dimensional data. *Journal of the Japan Statistical Society*, *35*(2), 251–272.

Tyler, D. E. (1982). Radial estimates and the test for sphericity. *Biometrika*, *69*(2), 429–436.

Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics*, *15*(1), 234–251.

Tyler, D. E. (2010). A note on multivariate location and scatter statistics for sparse data sets. *Statistics & Probability Letters*, *80*(17–18), 1409–1413.

Virta, J. (2021). Testing for subsphericity when n and p are of different asymptotic order. *Statistics & Probability Letters*, *179*, 109209.

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

Zhang, T., & Wiesel, A. (2016). Automatic diagonal loading for Tyler's robust covariance estimator. In *IEEE Statistical Signal Processing Workshop (SSP'16)*, (pp. 1–5).

Zou, C., Peng, L., Feng, L., & Wang, Z. (2014). Multivariate sign-based high-dimensional tests for sphericity. *Biometrika*, *101*(1).

Detecting Outliers in Compositional Data Using Invariant Coordinate Selection

Anne Ruiz-Gazen, Christine Thomas-Agnan, Thibault Laurent, and Camille Mondon

Abstract Invariant coordinate (or component) selection (ICS) is a multivariate statistical method introduced by Tyler et al. (J R Stat Soc Ser B (Stat Methodol) 71(3):549–592, 2009) and based on the simultaneous diagonalization of two scatter matrices. A model-based approach of ICS, called invariant coordinate analysis, has already been adapted for compositional data in Muehlmann et al. (Independent component analysis for compositional data. In Daouia, A, Ruiz-Gazen A (eds) Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan. Springer, New York, pp. 525–545, 2021). In a model-free context, ICS is also helpful at identifying outliers Nordhausen and Ruiz-Gazen (J Multivar Anal 188:104844, 2022). We propose to develop a version of ICS for outlier detection in compositional data. This version is first introduced in coordinate space for a specific choice of isometric log-ratio coordinate system associated to a contrast matrix and follows the outlier detection procedure proposed by Archimbaud et al. (Comput Stat Data Anal 128:184–199, 2018a). We then show that the procedure is independent of the choice of contrast matrix and can be defined directly in the simplex. To do so, we establish some properties of the set of matrices satisfying the zero-sum property and introduce a simplex definition of the Mahalanobis distance and the one-step M-estimators class of scatter matrices. We also need to define the family of elliptical distributions in the simplex. We then show how to interpret the results directly in the simplex using two artificial datasets and a real dataset of market shares in the automobile industry.

A. Ruiz-Gazen (✉) · C. Thomas-Agnan
Toulouse School of Economics, University of Toulouse 1 Capitole, Toulouse, France
e-mail: anne.ruiz-gazen@tse-fr.eu; christine.thomas@tse-fr.eu

T. Laurent
Toulouse School of Economics, CNRS, Toulouse, France
e-mail: thibault.laurent@tse-fr.eu

C. Mondon
Ecole Normale Supérieure (ENS), Paris, France
e-mail: camille.mondon@ens.fr

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_10

Keywords Anomaly detection · Elliptical distribution · Log-ratio transformation · Market share data · Scatter matrix · Ternary diagram

1 Introduction

Compositional data are by nature multivariate. Indeed, vectors with positive components are considered as compositional data when the interest lies in the relative information between their components: this last fact implies that they can be represented by a unique element in a simplex by dividing the components by their sum. Classical statistical techniques need to be adapted to deal with these constraints (positivity, sum equal to one). A common approach consists in transforming the data using the centered log-ratio (clr) or the isometric log-ratio (ilr) transformations (see Egozcue et al. 2011), and in applying standard techniques in this coordinate space. Filzmoser et al. (2012) propose to use the ilr transformation and detect outliers with the usual or the robust version of the Mahalanobis distance. Because of the affine invariance property of the Mahalanobis distance, the authors notice that the identified outliers do not depend on the choice of the ilr transformation. Moreover, they propose some graphical tools in coordinate space based on robust principal component analysis (PCA) and biplots representation in order to interpret the outliers. Their interpretation is only done in coordinate space. This is also the case for Filzmoser et al. (2014) who propose tools based on pairwise Mahalanobis distances for detecting local outliers in data that are compositional and spatial at the same time. In the present work, we consider adapting the invariant coordinate selection (ICS) technique for outlier detection to compositional data. ICS is a multivariate statistical method based on the joint diagonalization of two scatter matrices and aimed at detecting interesting features in multivariate datasets such as outliers or clusters (see, e.g., Tyler et al. 2009 and Archimbaud et al. 2018a). Compared to the Mahalanobis distance criterion, ICS includes a dimension reduction step. Compared to PCA, the components of ICS are invariant under affine transformations. We first propose to introduce ICS in coordinate space using an ilr transformation. Following Archimbaud et al. (2018a), we focus on the case of a small proportion of outliers and use the invariant components associated with the largest eigenvalues of the joint diagonalization of two particular scatter matrices. As with the Mahalanobis distance, the identification of outliers with ICS does not depend on the choice of the ilr transformation (see also Muehlmann et al. 2021). In order to go beyond coordinate space and interpret the outliers in the simplex, we introduce new algebra tools and define eigen-elements of endomorphisms of the simplex. We also introduce a class of one-step M-scatter estimators and elliptical distributions in the simplex. Thanks to these tools, we are able to write a reconstruction formula of the data in the simplex decomposing the data in a proper way for outlier identification and interpretation using ternary diagrams. In Sect. 2, we recall some facts about the ICS method and its application to outlier detection. Section 3 is a reminder about compositional data analysis. In Sect. 4, we develop some tools necessary for Sect. 5. First come some properties of the algebra of $D \times D$ matrices with the zero-sum property: in particular, their rank, their inverses, and their eigen-elements. Then Sect. 4.2 defines one-step M-scatter functionals for simplex-valued random variables together with an adapted version of Mahalanobis distance. Finally, Sect. 4.3 introduces the family of elliptical distributions in the simplex. Section 5 first introduces ICS in coordinate space and then reformulates ICS directly in the simplex. In Sect. 5.3, we present a formula for reconstructing the data from ICS in coordinate space and in the simplex. Section 6 is dedicated to three applications, with two toy datasets (with small and large dimensions) and a real marketing application from the automobile industry.

2 Reminder About ICS and Outlier Detection

Invariant coordinate (or component) selection is a multivariate statistical method based on the simultaneous diagonalization of two scatter matrices. As detailed in Nordhausen and Ruiz-Gazen (2022), the method belongs to a large family of multivariate statistical methods and is useful in particular for outlier detection as described below.

2.1 Scatter Matrices

The family of scatter matrices generalizes the notion of covariance matrix (see Nordhausen and Tyler 2015; Tyler et al. 2009, among others), and it has the following functional definition. For a p-dimensional vector $\mathbf{X}$ with distribution function $F_{\mathbf{X}}$, a functional $\mathbf{S}(F_{\mathbf{X}})$, also denoted by $\mathbf{S}(\mathbf{X})$, is called a scatter functional if it is a $p \times p$ symmetric positive-definite and affine equivariant matrix. We recall that an affine equivariant matrix $\mathbf{S}(\mathbf{X})$ is such that

$$\mathbf{S}(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}\mathbf{S}(\mathbf{X})\mathbf{A}^T,$$

where ^{T} denotes the transpose operator, $\mathbf{A}$ is any full rank $p \times p$ matrix, and $\mathbf{b}$ any p-vector.

For a p-variate dataset $\mathbf{X}_n = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, the empirical version $\mathbf{S}(\mathbf{X}_n)$ of a scatter functional is the scatter functional $\mathbf{S}(F_n)$, where F_n is the empirical distribution function. Thus, a scatter matrix estimator is a $p \times p$ symmetric positive-definite and affine equivariant matrix such that

$$\mathbf{S}(\mathbf{X}_n\mathbf{A} + \mathbf{1}_n\mathbf{b}^T) = \mathbf{A}^T\mathbf{S}(\mathbf{X}_n)\mathbf{A},$$

where $\mathbf{A}$ is any full rank $p \times p$ matrix, $\mathbf{b}$ any p-vector, and $\mathbf{1}_n$ an n-dimensional vector of ones.

There exist many scatter matrices as detailed for example in Tyler et al. (2009). The most well-known scatter matrix is the covariance matrix. As many other scatter matrices, the covariance involves the mean that is an affine equivariant location estimator. We recall that an affine equivariant location estimator **T** is such that:

$$\mathbf{T}(\mathbf{AX}+\mathbf{b}) = \mathbf{AT}(\mathbf{X})+\mathbf{b},$$

for the functional version, and

$$\mathbf{T}(\mathbf{X}_n\mathbf{A}+\mathbf{1}_n\mathbf{b}^T) = \mathbf{A}^T\mathbf{T}(\mathbf{X}_n)+\mathbf{b},$$

for the empirical version where **A** is any full rank $p \times p$ matrix and **b** any p-vector.

A general class of scatter matrices is the class of one-step M-estimators with a functional defined by

$$\mathbf{COV}_w(\mathbf{X}) = \mathbf{E}\Big[w(M^2(\mathbf{X}))(\mathbf{X}-\mathbf{E}(\mathbf{X}))(\mathbf{X}-\mathbf{E}(\mathbf{X}))^T\Big],$$

where w is a non-negative and continuous weight function and

$$M^2(\mathbf{X}) = (\mathbf{X}-\mathbf{E}(\mathbf{X}))^T\mathbf{COV}(\mathbf{X})^{-1}(\mathbf{X}-\mathbf{E}(\mathbf{X})) \tag{1}$$

is the square Mahalanobis distance with $\mathbf{E}(\mathbf{X})$ the expectation of **X** and $\mathbf{COV}(\mathbf{X})$ its covariance matrix. The sample version of one-step M-estimators is

$$\mathbf{COV}_w(\mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^{n} w(M^2(\mathbf{x}_i))(\mathbf{x}_i-\bar{\mathbf{x}}_n)(\mathbf{x}_i-\bar{\mathbf{x}}_n)^T,$$

where $\bar{\mathbf{x}}_n = 1/n\sum_{i=1}^n \mathbf{x}_i$ is the empirical mean and

$$M^2(\mathbf{x}_i) = (\mathbf{x}_i-\bar{\mathbf{x}}_n)^T\mathbf{COV}(\mathbf{X}_n)^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}_n)$$

is the empirical version of the square Mahalanobis distance.

Note that the covariance matrix **COV** is obtained with $w(d) = 1$, while the fourth-moment-based estimator $\mathbf{COV}_4$ is obtained with $w(d) = d/(p+2)$. $\mathbf{COV}_4$ is widely used in the blind source separation literature (see, e.g., Nordhausen and Virta 2019; Theis and Inouye 2006) but also in the context of outlier detection (see Archimbaud et al. 2018a).

For elliptical distributions with second moments, scatter functionals are all proportional to the covariance matrix (see, e.g., Bilodeau and Brenner 2008). We recall that an elliptical distribution is obtained as an affine transformation of a spherical distribution that is a distribution invariant by orthogonal transformation. Multivariate normal and Student distributions belong to this family of distributions.

2.2 ICS Principle

Let $\mathbf{S}_1(\mathbf{X})$ and $\mathbf{S}_2(\mathbf{X})$ be two scatter functionals. ICS consists of the simultaneous diagonalization of $\mathbf{S}_1(\mathbf{X})$ and $\mathbf{S}_2(\mathbf{X})$. If the random vector $\mathbf{X}$ follows an elliptical distribution, the two scatter matrices will be proportional, and the result will be useless. However, as mentioned in Tyler et al. (2009), comparing two different scatter functionals may help revealing interesting departures from an elliptical distribution. This is the case in particular for anomaly detection. The method searches for a $p \times p$ matrix $\mathbf{H}(\mathbf{X})$ and a diagonal matrix $\mathbf{\Lambda}(\mathbf{X})$ so that

$$\mathbf{H}(\mathbf{X})^T\mathbf{S}_1(\mathbf{X})\mathbf{H}(\mathbf{X}) = \mathbf{I}_p \quad \text{and} \quad \mathbf{H}(\mathbf{X})^T\mathbf{S}_2(\mathbf{X})\mathbf{H}(\mathbf{X}) = \mathbf{\Lambda}(\mathbf{X}), \tag{2}$$

where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. The matrix $\mathbf{\Lambda}(\mathbf{X})$ contains the eigenvalues of $\mathbf{S}_1(\mathbf{X})^{-1}\mathbf{S}_2(\mathbf{X})$ in decreasing order, while the columns of the matrix $\mathbf{H}(\mathbf{X}) = (\mathbf{h}_1, \ldots, \mathbf{h}_p)$ contain the corresponding eigenvectors so that

$$\mathbf{S}_2(\mathbf{X})\mathbf{H}(\mathbf{X}) = \mathbf{S}_1(\mathbf{X})\mathbf{H}(\mathbf{X})\mathbf{\Lambda}(\mathbf{X})$$

$$\text{or equivalently} \quad \mathbf{S}_1(\mathbf{X})^{-1}\mathbf{S}_2(\mathbf{X})\mathbf{H}(\mathbf{X}) = \mathbf{H}(\mathbf{X})\mathbf{\Lambda}(\mathbf{X}).$$

These eigenvalues and eigenvectors can also be derived through the spectral decomposition of the following symmetric matrix:

$$\mathbf{S}_1(\mathbf{X})^{-1/2}\mathbf{S}_2(\mathbf{X})\mathbf{S}_1(\mathbf{X})^{-1/2} = \mathbf{U}(\mathbf{X})\mathbf{\Lambda}(\mathbf{X})\mathbf{U}(\mathbf{X})^T, \tag{3}$$

with $\mathbf{U}(\mathbf{X})$ a $p \times p$ orthogonal matrix and the same eigenvalues in the diagonal matrix $\mathbf{\Lambda}(\mathbf{X})$. We have

$$\mathbf{H}(\mathbf{X}) = \mathbf{S}_1(\mathbf{X})^{-1/2}\mathbf{U}(\mathbf{X})$$

and

$$\mathbf{H}(\mathbf{X})\mathbf{H}(\mathbf{X})^T = \mathbf{S}_1(\mathbf{X})^{-1} \text{ and } \mathbf{H}(\mathbf{X})^{-1} = \mathbf{U}(\mathbf{X})^T\mathbf{S}_1(\mathbf{X})^{1/2}.$$

Tyler et al. (2009) give an interesting interpretation of the eigenvalues $\lambda_1, \ldots, \lambda_p$ in terms of kurtosis. Using the optimality property of eigen-elements, it is easy to see that $\mathbf{h}_1$ maximizes the ratio:

$$\frac{\mathbf{h}^T\mathbf{S}_2(\mathbf{X})\mathbf{h}}{\mathbf{h}^T\mathbf{S}_1(\mathbf{X})\mathbf{h}}$$

over all possible vectors $\mathbf{h}$ in $\mathbb{R}^p$ and that λ_1 is equal to the maximum ratio. This ratio of two scale measures can be viewed as a generalized measure of kurtosis, and λ_1 can thus be interpreted as a maximum kurtosis. The other eigenvalues and eigenvectors can be defined in a similar way by maximizing the same ratio over vectors **h** that verify additional orthogonality conditions (see Tyler et al. (2009) for details).

Using any affine equivariant location estimator $\mathbf{T}(\mathbf{X})$, the ICS scores $\mathbf{Z} = (z_1, \dots, z_p)^T$ are defined by

$$\mathbf{Z} = \mathbf{H}(\mathbf{X})^T(\mathbf{X} - \mathbf{T}(\mathbf{X})),$$

or equivalently by $z_k =< \mathbf{h}_k, \mathbf{X} - \mathbf{T}(\mathbf{X}) >$ where $< ., . >$ denotes the standard scalar product. The scores define the affine invariant coordinates or components. The square Euclidian norm of these coordinates is given by

$$\begin{aligned}\mathbf{Z}^T\mathbf{Z} &= (\mathbf{X} - \mathbf{T}(\mathbf{X}))^T\mathbf{H}(\mathbf{X})\mathbf{H}(\mathbf{X})^T(\mathbf{X} - \mathbf{T}(\mathbf{X}))\\ &= (\mathbf{X} - \mathbf{T}(\mathbf{X}))^T\mathbf{S}_1(\mathbf{X})^{-1}(\mathbf{X} - \mathbf{T}(\mathbf{X})).\end{aligned}$$

The last expression is a generalization of the Mahalanobis distance (1) of **X** with the location parameter $\mathbf{T}(\mathbf{X})$ (instead of $\mathbf{E}(\mathbf{X})$) and with respect to the scatter matrix $\mathbf{S}_1(\mathbf{X})$ (instead of $\mathbf{COV}(\mathbf{X})$). In the special case where $\mathbf{T}(\mathbf{X}) = \mathbf{E}(\mathbf{X})$ and $\mathbf{S}_1(\mathbf{X}) = \mathbf{COV}(\mathbf{X})$, we have

$$\mathbf{Z}^T\mathbf{Z} = \sum_{k=1}^{p} z_k^2 = M^2(\mathbf{X}). \tag{4}$$

The empirical version of ICS consists of the joint diagonalization of a scatter pair of estimators $\mathbf{S}_1(\mathbf{X}_n)$ and $\mathbf{S}_2(\mathbf{X}_n)$. For a $p \times p$ matrix $\mathbf{H}(\mathbf{X}_n)$ and a diagonal matrix $\mathbf{\Lambda}(\mathbf{X}_n)$, we have

$$\mathbf{H}(\mathbf{X}_n)^T\mathbf{S}_1(\mathbf{X}_n)\mathbf{H}(\mathbf{X}_n) = \mathbf{I}_p \quad \text{and} \quad \mathbf{H}(\mathbf{X}_n)^T\mathbf{S}_2(\mathbf{X}_n)\mathbf{H}(\mathbf{X}_n) = \mathbf{\Lambda}(\mathbf{X}_n).$$

Using any affine equivariant location estimator $\mathbf{T}(\mathbf{X}_n)$, the ICS scores are given by

$$\mathbf{Z}_n = (\mathbf{z}_1, \dots, \mathbf{z}_n)^T = (\mathbf{X}_n - \mathbf{1}_n\mathbf{T}(\mathbf{X}_n)^T)\mathbf{H}(\mathbf{X}_n)$$

and are affine invariant. As in (4), if $T(\mathbf{X}_n) = \bar{\mathbf{x}}_n$ and $\mathbf{S}_1(\mathbf{X}_n) = \mathbf{COV}(\mathbf{X}_n)$, we have

$$M^2(\mathbf{x}_i) = \mathbf{z}_i^T\mathbf{z}_i.$$

2.3 ICS for Outlier Detection

As already stated in Tyler et al. (2009), one possible application of ICS is outlier detection. The Mahalanobis distance is a well-known tool to detect outliers (see Rousseeuw and Van Zomeren 1990), but it does not offer the possibility of dimension reduction. ICS gives the possibility of selecting components that are helpful in detecting anomalies (see Archimbaud et al. (2018a) for details). In the case of a small proportion of outliers, the theoretical properties of ICS (see Archimbaud et al. (2018a) for details) lead us to only focus on the invariant components associated with the largest kurtosis and thus the largest eigenvalues. In this context, Archimbaud et al. (2018a) show that the scatter pair $\mathbf{S}_1(\mathbf{X}) = \mathbf{COV}(\mathbf{X})$ and $\mathbf{S}_2(\mathbf{X}) = \mathbf{COV}_4(\mathbf{X})$ is not only simple and fast to compute but also effective in detecting outliers when compared to other pairs that involve robust scatter estimators. Archimbaud et al. (2018a) propose different automatic procedures for invariant components selection based on hypothesis testing. In short, the idea is to test sequentially the normality of each of the invariant components using some classical tests such as the D'Agostino test and to select the first k components that reject the Gaussian assumption. After selecting k invariant components among p, the last step of the procedure is the outlier identification. Let us consider the empirical version of ICS. For each observation $i = 1, \ldots, n$, the square "ICS distance" is the square Euclidian norm in the invariant coordinate system accounting for the k first coordinates:

$$(\text{ICS distance})^2_{i,k} = \sum_{j=1}^{k} \left(z_i^j\right)^2, \tag{5}$$

where z_i^j denotes the jth value of the score $\mathbf{z}_i$. In Archimbaud et al. (2018a), an observation is flagged as an outlier when its ICS distance using k components is larger than a cutoff based on Monte Carlo simulations from the standard Gaussian distribution. Given a data dimension, a scatter pair, and a number k of selected components, many Gaussian samples are generated, and the ICS distances are computed. A cutoff is derived for a fixed level γ as the mean of the $(1-\gamma)$-quantiles of these distances over the replications. The whole ICS procedure for outlier detection is available in the `R` package `ICSOutlier` described in Archimbaud et al. (2018b) and used in Sect. 6 below.

3 Reminder About Compositional Data Analysis

A D composition $\mathbf{u}$ is a vector of D parts (or shares) of some whole that carries relative information. There exists a unique representation of this vector in the unit simplex space

$$\mathbf{S}^D = \left\{ \mathbf{u} = (u_1, \ldots, u_D)^T \; : u_m > 0, m = 1, ..., D; \sum_{m=1}^{D} u_m = 1 \right\}.$$

For any vector $\mathbf{w} \in \mathbb{R}^{+D}$, its representer in the simplex is obtained by the closure operation

$$C(\mathbf{w}) = \left(\frac{w_1}{\sum_{m=1}^{D} w_m}, \cdots, \frac{w_D}{\sum_{m=1}^{D} w_m} \right).$$

The following operations endow the unit simplex with a vector space structure:

1. $\oplus$ is the perturbation operation, corresponding to the addition in $\mathbb{R}^D$:

$$\text{For} \quad \mathbf{u}, \mathbf{v} \in \mathbf{S}^D, \mathbf{u} \oplus \mathbf{v} = C(u_1 v_1, \ldots, u_D v_D).$$

2. $\odot$ is the power operation, corresponding to the scalar multiplication in $\mathbb{R}^D$:

$$\text{For} \quad \lambda \in \mathbb{R}, \mathbf{u} \in \mathbf{S}^D \quad \lambda \odot \mathbf{u} = C(u_1^\lambda, \ldots, u_D^\lambda).$$

The subtraction operation can be naturally defined by $\mathbf{u} \ominus \mathbf{v} = C(u_1/v_1, \ldots, u_D/v_D)$. Compositional data analysis uses log-ratio transformations such as the centered log-ratio (clr) and the isometric log-ratio (ilr) transformations. The clr vector components specify the relative dominance of each compositional part over the whole composition, see for example Filzmoser et al. (2018). Formally, the clr transformation of a vector $\mathbf{u} \in \mathbf{S}^D$ is defined by

$$\text{clr}(\mathbf{u}) = \mathbf{G}_D \ln \mathbf{u},$$

where $\mathbf{G}_D = \mathbf{I}_D - \frac{1}{D}\mathbf{1}_D {\mathbf{1}_D}^T$, $\mathbf{I}_D$ is a $D \times D$ identity matrix, and $\mathbf{1}_D$ is the D-vector of ones and where the logarithm of $\mathbf{u} \in \mathbf{S}^D$ is understood componentwise.

For a vector $\mathbf{u}$ in the orthogonal space $\mathbf{1}_D^\perp$ (orthogonality with respect to the standard scalar product of $\mathbb{R}^D$), the inverse clr transformation is defined by

$$\text{clr}^{-1}(\mathbf{u}) = C(\exp(\mathbf{u})).$$

The simplex $\mathcal{S}^D$ of dimension $D - 1$ can be equipped with the Aitchison scalar product

$$< \mathbf{u}, \mathbf{v} >_A = < \text{clr}(\mathbf{u}), \text{clr}(\mathbf{v}) >,$$

where the right-hand side scalar product is the standard scalar product in $\mathbb{R}^D$.

The clr coordinates sum up to zero inducing a degeneracy. For this reason, the class of isometric log-ratio coordinates has been introduced providing orthonormal and non-singular coordinates. For any given orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$, orthonormality being understood with respect to the Aitchison scalar product here, one can define a so-called contrast matrix $\mathbf{V}$ of dimension $D \times (D - 1)$ (e.g. Pawlowsky-Glahn et al. 2015) given by $\mathbf{V} = \text{clr}(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$, where clr is understood columnwise. To each such matrix, $\mathbf{V}$ is associated an isometric log-ratio transformation by

$$\mathrm{ilr}_V(\mathbf{u}) = \mathbf{V}^T \ln(\mathbf{u}).$$

The inverse transformation, for any vector $\mathbf{u}^*$ of $\mathbb{R}^{D-1}$, is given by

$$\mathbf{u} = \mathrm{ilr}_V^{-1}(\mathbf{u}^*) = \mathcal{C}(\exp(\mathbf{V}\mathbf{u}^*)).$$

The link between the ilr and clr transformations is $\mathrm{clr}(\mathbf{u}) = \mathbf{V}\mathrm{ilr}_V(\mathbf{u})$.

4 Multivariate Tools for Compositional Data

For working with scatter matrices for compositional data, we are going to need some algebra tools concerning matrices of endomorphisms in the simplex.

4.1 Algebra of Endomorphisms of the Simplex and Eigendecomposition

Let $\mathcal{A}$ be the set of $D \times D$ matrices such that $\mathbf{A}\mathbf{1}_D = \mathbf{0}_D$ and $\mathbf{A}^T\mathbf{1}_D = \mathbf{0}_D$, where $\mathbf{0}_D$ denotes the D-dimensional column vector of zeros: this condition is called the zero-sum property. Pawlowsky-Glahn et al. (2015) define endomorphisms of the simplex using the ilr transformation and prove that they can be associated to a matrix belonging to $\mathcal{A}$, see Property 4.16 and pages 55–58. The linearity here refers to the vector space structure of the simplex based on the perturbation and powering operations. Let us introduce an equivalent formulation based on the clr transformation: for $\mathbf{u} \in \mathbf{S}^D$ and $\mathbf{A} \in \mathcal{A}$, endomorphisms of the simplex are defined by maps $\mathbf{u} \mapsto \mathbf{A} \boxdot \mathbf{u} := \mathrm{clr}^{-1}(\mathbf{A}\mathrm{clr}(u))$.

The composition of endomorphisms corresponds to the ordinary matrix product since it is clear that $\mathbf{A}\boxdot(\mathbf{B}\boxdot\mathbf{u}) = \mathbf{A}\mathbf{B}\boxdot\mathbf{u}$, and therefore, $\mathcal{A}$ is an algebra with neutral element $\mathbf{G}_D$. We are now going to extend the definition of the ilr transformation to matrices of $\mathcal{A}$.

Theorem 1 *Let $\mathbf{V}$ be a $D \times (D-1)$ contrast matrix, and let $\mathbf{P}_V$ be the $D \times D$ block matrix $[\mathbf{V}\ \frac{1}{\sqrt{D}}\mathbf{1}_D]$. For a $D \times D$ matrix $\mathbf{A} \in \mathcal{A}$, the $(D-1) \times (D-1)$ matrix $\mathbf{A}^* :=$ $ilr_V(\mathbf{A}) = \mathbf{V}^T\mathbf{A}\mathbf{V}$ is such that $\mathbf{A} = ilr_V^{-1}(\mathbf{A}^*) = \mathbf{V}\mathbf{A}^*\mathbf{V}^T = \mathbf{P}_V \begin{pmatrix} \mathbf{A}^* & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix} \mathbf{P}_V^T$ and satisfies the following properties:*

1. *The rank of $\mathbf{A}$ is equal to the rank of $ilr_V(\mathbf{A})$.*

2. *If* $ilr_V(\mathbf{A})$ *is invertible, then* $\mathbf{A}$ *is invertible in* $\mathcal{A}$*, and we have the following expressions for its* $\mathcal{A}$*-inverse*

$$\mathbf{A}^{-1} = (\mathbf{A}+\frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T)^{-1}-\frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T = \mathbf{V}(\mathbf{V}^T\mathbf{A}\mathbf{V})^{-1}\mathbf{V}^T = \mathbf{P}_V\begin{pmatrix}\mathbf{A}^{*-1} & \mathbf{0}_{D-1}\\ \mathbf{0}_{D-1}^T & 0\end{pmatrix}\mathbf{P}_V^T.$$

3. $ilr_V(\mathbf{AB}) = ilr_V(\mathbf{A})ilr_V(\mathbf{B})$*. If* $\mathbf{A}$ *is invertible, then* $ilr_V(\mathbf{A}^{-1}) = (ilr_V(\mathbf{A}))^{-1}$*. If* $(ilr_V(\mathbf{A}))^{1/2}$ *exists, then* $ilr_V(\mathbf{A}^{1/2}) = (ilr_V(\mathbf{A}))^{1/2}$*.*

Note that a matrix $\mathbf{A}$ of the algebra $\mathcal{A}$ is never invertible in the space of matrices in the classical sense. But it may be invertible in the sense of the algebra, and its $\mathcal{A}$-inverse then coincides with the Moore–Penrose pseudo-inverse of $\mathbf{A}$ in the usual sense. The matrix $\text{ilr}_V(\mathbf{A})$ is simply the matrix corresponding to $\mathbf{A}$ in coordinate space when the coordinates are defined by ilr_V. We also extend the definition of the clr transformations to matrices.

Theorem 2 *For a* $D \times D$ *matrix* $\mathbf{B}$*, let us define its clr transformation by*

$$clr(\mathbf{B}) = \mathbf{G}_D\mathbf{B}\mathbf{G}_D.$$

We then have the following properties:

1. *If* $\mathbf{A} \in \mathcal{A}$*, then* $clr(\mathbf{A}) = \mathbf{A}$*.*
2. *If* $\mathbf{B} \notin \mathcal{A}$*, then* $clr(\mathbf{B}) \in \mathcal{A}$ *and for any* $\mathbf{x} \in \mathcal{S}^D$

$$\mathbf{B} \boxdot \mathbf{x} := clr^{-1}(clr(\mathbf{B})clr(\mathbf{x})) = clr(\mathbf{B}) \boxdot \mathbf{x}. \qquad (6)$$

3. *If* $\mathbf{B} \notin \mathcal{A}$*, then the unique element* $\mathbf{A} \in \mathcal{A}$ *such that* $ilr_V(\mathbf{A}) = ilr_V(\mathbf{B})$ *is* $\mathbf{A} = clr(\mathbf{B})$*.*
4. *For any contrast matrix* $\mathbf{V}$ *and any* $\mathbf{A} \in \mathcal{A}$*, we have* $clr(\mathbf{A}) = \mathbf{V}ilr_V(\mathbf{A})\mathbf{V}^T$*.*

Note that the matrix product $\boxdot$ can be defined even when the matrix $\mathbf{B}$ does not belong to $\mathcal{A}$, but in that case it is not linear. Note also that the ilr and clr transformations preserve symmetry.

The next proposition links the eigen-elements of $\mathbf{A}$ to those of $\text{ilr}(\mathbf{A})$. Let us first define the notion of $\mathcal{A}$-diagonalizable for a matrix of $\mathcal{A}$.

Definition 1 A matrix $\mathbf{A} \in \mathcal{A}$ is said $\mathcal{A}$-diagonalizable if there exists a basis $\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}$ of $\mathcal{S}^D$ and $D-1$ reals λ_j $(j = 1, \ldots, D-1)$ such that

$$\mathbf{A} \boxdot \mathbf{e}_j = \lambda_j \odot \mathbf{e}_j \quad \forall j = 1, \ldots, D-1. \qquad (7)$$

We will say that $\mathbf{e}_j$ is an $\mathcal{A}$-eigenvector of $\mathbf{A}$. It is clear that $\text{clr}(\mathbf{e}_j)$ is then an eigenvector of $\text{clr}(\mathbf{A}) = \mathbf{A}$ and that for any contrast matrix $\mathbf{V}$, $\text{ilr}_V(\mathbf{e}_j)$ is an eigenvector of $\text{ilr}_V(\mathbf{A})$. Note that $\mathbf{1}_D$ is an eigenvector of $\mathcal{A}$ associated to the eigenvalue 0. It is natural to say that a matrix $\mathbf{A} \in \mathcal{A}$ is diagonal in a given basis $\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}$ of $\mathcal{S}^D$ if Eq. (7) is satisfied for these vectors.

Theorem 3 *Let* $\mathbf{V}$ *be a* $D \times (D-1)$ *contrast matrix. For a* $D \times D$ *matrix* $\mathbf{A} \in \mathcal{A}$, *we have the following properties:*

1. *If* $\mathbf{e}_j^* \in \mathbb{R}^{D-1}$ *is an eigenvector of* $ilr_V(\mathbf{A})$, *then* $\mathbf{e}_j = ilr^{-1}(\mathbf{e}_j^*) \in \mathcal{S}^D$ *is an* $\mathcal{A}$*-eigenvector of* $\mathbf{A}$ *and* $\mathbf{w}_j = clr(\mathbf{e}_j) \in \mathbb{R}^D$ *an eigenvector of* $\mathbf{A}$.
2. *The set of eigenvalues of* $\mathbf{A}$ *contains the eigenvalue* 0. *The other* $D-1$ *eigenvalues of* $\mathbf{A}$ *coincide with the eigenvalues of* $ilr_V(\mathbf{A})$ *for any contrast matrix* $\mathbf{V}$.
3. $ilr_V(\mathbf{A})$ *is diagonalizable if and only if* $\mathbf{A}$ *is diagonalizable, and if and only if* $\mathbf{A}$ *is* $\mathcal{A}$*-diagonalizable.*

All symmetric matrices in $\mathcal{A}$ are $\mathcal{A}$-diagonalizable. Note that the vectors $\mathbf{e}_j = \text{clr}^{-1}(\mathbf{e}_j^*)$ are independent of the contrast matrix $\mathbf{V}$. Let $\mathbf{A}$ be a symmetric matrix of $\mathcal{A}$. Since the vector $\mathbf{1}_D$ is an eigenvector of $\mathbf{A}$, $\mathbf{A}$ cannot be diagonal in the canonical basis of $\mathbb{R}^D$, but it can be diagonal in a basis obtained by completing $\mathbf{w}_D = \frac{1}{D}\mathbf{1}_D$ with $D-1$ orthogonal eigenvectors in $\mathbf{1}_D^{\perp}$, say $\mathbf{w}_1, \ldots, \mathbf{w}_{D-1}$. Then $\mathbf{e}_j = \text{clr}^{-1}(\mathbf{w}_j) \in \mathcal{S}^D$ $(j = 1, \ldots, D-1)$ is an orthonormal basis of $\mathcal{S}^D$ for the Aitchison metric since $< \mathbf{e}_i, \mathbf{e}_j >_A = < \mathbf{w}_i, \mathbf{w}_j >_E = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise, and these vectors are $\mathcal{A}$-eigenvectors of $\mathbf{A}$. If $\mathbf{W} = [\mathbf{w}_1 \ldots, \mathbf{w}_{D-1}]$ is the corresponding contrast matrix, then $\text{ilr}_W(\mathbf{A})_{ij} = \mathbf{w}_i^T \mathbf{A} \mathbf{w}_j = \lambda_j \mathbf{w}_i^T \mathbf{w}_j = \lambda_i \delta_{ij}$, which shows that $\text{ilr}_{\mathbf{W}}(\mathbf{A})$ is the $(D-1) \times (D-1)$ diagonal matrix $\mathbf{\Lambda}$ with the λ_i as diagonal elements. Then using Theorem 1, we can write that $\mathbf{A} = \mathbf{P}_W \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix} \mathbf{P}_W^T$ showing that $\mathbf{A}$ is similar to the diagonal matrix $\begin{pmatrix} \mathbf{\Lambda} & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix}$. This last result gives us the general form of diagonal matrices of $\mathcal{A}$ with the corresponding spectral representation $\mathbf{A} = \sum_{i=1}^{D-1} \lambda_i \mathbf{w}_i \mathbf{w}_i^T$.

4.2 One-Step M-Scatter Functionals of a Compositional Random Vector

For a simplex-valued random vector $\mathbf{X}$ (see Pawlowsky-Glahn et al. 2015), let us recall the following definition of expectation:

$$\mathbf{E}^{\oplus}\mathbf{X} := \text{clr}^{-1}(\mathbf{E}\text{clr}(\mathbf{X}))$$

and the following definition of the (clr-)covariance matrix $\mathbf{COV}^{\oplus}\mathbf{X}$ (see Aitchison 1982) given by the $D \times D$ matrix

$$\mathbf{COV}^{\oplus}\mathbf{X} := \mathbf{COV}(\text{clr}(\mathbf{X})).$$

Note that, by Theorem 2, we can see that $\mathbf{COV}^{\oplus}\mathbf{X}$ is also equal to $\mathrm{clr}^{-1}(\mathbf{COV}(\mathrm{clr}(\mathbf{X})))$. Using the same principles, let us now introduce a simplex adapted definition of the square Mahalanobis distance as being the square Mahalanobis distance in the usual sense of the clr coordinates of $\mathbf{X}$

$$M^2(\mathbf{X}) = (\mathrm{clr}(\mathbf{X}) - \mathbf{E}\mathrm{clr}(\mathbf{X}))^T(\mathbf{COV}^{\oplus}\mathbf{X})^{-1}(\mathrm{clr}(\mathbf{X}) - \mathbf{E}\mathrm{clr}(\mathbf{X})).$$

In the same line, let us define the following one-step M-scatter matrix of a simplex-valued random vector as the corresponding scatter of its clr coordinates

$$\begin{aligned}\mathbf{COV}_w^{\oplus}\mathbf{X} &:= \mathbf{COV}_w(\mathrm{clr}(\mathbf{X}))\\ &= \mathbf{E}[w(M^2(\mathbf{X}))(\mathrm{clr}(\mathbf{X}) - \mathbf{E}\mathrm{clr}(\mathbf{X}))(\mathrm{clr}(\mathbf{X}) - \mathbf{E}\mathrm{clr}(\mathbf{X}))^T].\end{aligned}$$

For $w(d) = d/(D+2)$, we get the fourth-moment-based scatter matrix $\mathbf{COV}_4^{\oplus}\mathbf{X}$:

$$\begin{aligned}\mathbf{COV}_4^{\oplus}\mathbf{X} &:= \mathbf{COV}_4(\mathrm{clr}(\mathbf{X}))\\ &= \frac{1}{D+2}\mathbf{E}[M^2(\mathbf{X})(\mathrm{clr}(\mathbf{X}) - \mathbf{E}\mathrm{clr}(\mathbf{X}))(\mathrm{clr}(\mathbf{X}) - \mathbf{E}\mathrm{clr}(\mathbf{X}))^T].\end{aligned}$$

All these characteristics can also be expressed using the ilr coordinates associated to any contrast matrix $\mathbf{V}$ by the following formulas:

$$\mathbf{E}^{\oplus}\mathbf{X} = \mathrm{ilr}_V^{-1}(\mathbf{E}\mathrm{ilr}_V(\mathbf{X})),$$

$$\mathbf{COV}^{\oplus}\mathbf{X} = \mathbf{COV}(\mathrm{clr}(\mathbf{X})) = \mathbf{COV}(\mathbf{V}\mathrm{ilr}_V(\mathbf{X})),$$

and thus

$$\mathbf{COV}^{\oplus}\mathbf{X} = \mathbf{V}\mathbf{COV}(\mathrm{ilr}_V(\mathbf{X}))\mathbf{V}^T = \mathrm{ilr}_V^{-1}(\mathbf{COV}(\mathrm{ilr}_V(\mathbf{X}))),$$

$$M^2(\mathbf{X}) = M^2(\mathrm{ilr}_V(\mathbf{X})),$$

and similarly

$$\mathbf{COV}_w^{\oplus}(\mathbf{X}) = \mathbf{V}\mathbf{COV}_w(\mathrm{ilr}_V(\mathbf{X}))\mathbf{V}^T = \mathrm{ilr}_V^{-1}(\mathbf{COV}_w(\mathrm{ilr}_V(\mathbf{X}))).$$

Note that the scatter functionals $\mathbf{COV}_w^{\oplus}\mathbf{X}$ belong to the algebra $\mathcal{A}$, and thus we also have

$$\mathbf{COV}_w^{\oplus}(\mathbf{X}) = \mathrm{clr}^{-1}(\mathbf{COV}_w(\mathrm{clr}(\mathbf{X}))).$$

Given a sample of size n, the empirical versions of the previous scatter matrices can be derived easily.

4.3 Elliptical Distribution in the Simplex

Nguyen (2019) introduces the Student distribution in the simplex with an application to political economy. Mateu-Figueras et al. (2021) review some distributions in the simplex including the multivariate Student distribution. We define a new family of elliptical distributions with second moment in the simplex. A random vector $\mathbf{X}$ with values in $\mathcal{S}^D$ is said to follow an elliptical distribution if any of its ilr coordinates follows an elliptical distribution with second moment in $\mathbb{R}^{D-1}$. This definition makes sense due to the following theorem.

Theorem 4 *Given two contrast matrices* $\mathbf{V}$ *and* $\mathbf{W}$, *if* $\mathbf{X}^*_V = ilr_V(\mathbf{X})$ *follows an elliptical distribution with parameters* $\boldsymbol{\mu}^*_V = \mathbf{E}(\mathbf{X}^*_V)$ *and* $\boldsymbol{\Sigma}^*_V = \mathbf{COV}(\mathbf{X}^*_V)$*, then* $\mathbf{X}^*_W = ilr_W(\mathbf{X})$ *follows an elliptical distribution with parameters* $(\boldsymbol{\mu}^*_W, \boldsymbol{\Sigma}^*_W)$ *with*

$$\begin{aligned} \mathbf{W}\boldsymbol{\mu}^*_W &= \mathbf{V}\boldsymbol{\mu}^*_V, \\ \mathbf{W}\boldsymbol{\Sigma}^*_W\mathbf{W}^T &= \mathbf{V}\boldsymbol{\Sigma}^*_W\mathbf{V}^T, \\ \mathbf{W}{\boldsymbol{\Sigma}^*_W}^{-1}\mathbf{W}^T &= \mathbf{V}{\boldsymbol{\Sigma}^*_V}^{-1}\mathbf{V}^T. \end{aligned}$$

From this theorem, we can say that $\boldsymbol{\mu}_{\text{clr}} = \mathbf{V}\boldsymbol{\mu}^* = \mathbf{E}(\text{clr}(X))$ is an invariant that characterizes the location parameter of the elliptical distribution in clr coordinate space, and $\boldsymbol{\mu} = \text{clr}^{-1}(\boldsymbol{\mu}_{\text{clr}}) = \mathbf{E}^{\oplus}(X)$ is an invariant that characterizes the location parameter in the simplex. Moreover, $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Sigma}^*_V\mathbf{V}^T = \text{ilr}_V^{-1}(\boldsymbol{\Sigma}^*_V) = \mathbf{COV}^{\oplus}\mathbf{X}$ is an invariant that characterizes the scatter matrix in the simplex. Similarly, $\mathbf{Q} = \mathbf{V}{\boldsymbol{\Sigma}^*_V}^{-1}\mathbf{V}^T = \text{ilr}_V^{-1}({\boldsymbol{\Sigma}^*_V}^{-1})$ is an invariant that characterizes the precision matrix of this distribution in the simplex. As in Pawlowsky-Glahn et al. (2015), it is easy to write the density of this distribution with respect to Lebesgue measure in coordinate space as well as with respect to the Aitchison measure in the simplex. As in Comas-Cufí et al. (2016), we can extend this definition to a mixture of elliptical distributions.

5 ICS for Compositional Data

5.1 ICS in Coordinate Space

With the definitions introduced in Sect. 4.2, we can now define ICS for a compositional random vector $\mathbf{X}$. For a given choice of contrast matrix $\mathbf{V}$, let $\mathbf{X}^* = \text{ilr}_V(\mathbf{X})$. In the ilr coordinate space, ICS consists of the joint diagonalization of two scatter matrices $\mathbf{S}_1(\mathbf{X}^*)$ and $\mathbf{S}_2(\mathbf{X}^*)$. Following Archimbaud et al. (2018a), let us focus on $\mathbf{S}_1(\mathbf{X}^*) = \mathbf{COV}(\mathbf{X}^*)$ and $\mathbf{S}_2(\mathbf{X}^*) = \mathbf{COV}_4(\mathbf{X}^*)$. From Eq. (3) in Sect. 2.2, we can derive the affine invariant coordinates by diagonalizing the $(D-1) \times (D-1)$ symmetric matrix

$$\mathbf{L}^* = \mathbf{COV}(\mathbf{X}^*)^{-1/2}\mathbf{COV}_4(\mathbf{X}^*)\mathbf{COV}(\mathbf{X}^*)^{-1/2}.$$

Let $\lambda_1 \geq \ldots \geq \lambda_{D-1}$ be the eigenvalues of $\mathbf{L}^*$ in descending order, and $\boldsymbol{\Lambda}$ be the $(D-1) \times (D-1)$ diagonal matrix with the vector of eigenvalues on its diagonal. Let $\mathbf{u}_k^*$, k ranging from 1 to $D-1$, be the $D-1$ corresponding eigenvectors of $\mathbf{L}^*$ and $\mathbf{U}^* = [\mathbf{u}_1^* \ldots \mathbf{u}_{D-1}^*]$ be the matrix whose columns are these eigenvectors. By construction, the matrix $\mathbf{U}^*$ is orthogonal (with respect to the standard scalar product in $\mathbb{R}^{D-1}$). We have for all $k = 1, \ldots, D-1$:

$$\mathbf{L}^*\mathbf{u}_k^* = \lambda_k \mathbf{u}_k^*.$$

If we denote by $\mathbf{h}_k^*$, $k = 1, \ldots, D-1$ the column vectors of $\mathbf{H}^* = \mathbf{COV}(\mathbf{X}^*)^{-1/2}\mathbf{U}^*$, we have

$$\mathbf{H}^{*T}\mathbf{COV}(\mathbf{X}^*)\mathbf{H}^* = \mathbf{I}_{D-1}, \tag{8}$$

$$\mathbf{H}^{*T}\mathbf{COV}_4(\mathbf{X}^*)\mathbf{H}^* = \boldsymbol{\Lambda}. \tag{9}$$

Equations (8) and (9) correspond to the joint diagonalization of $\mathbf{COV}(\mathbf{X}^*)$ and $\mathbf{COV}_4(\mathbf{X}^*)$. As for Eq. (2), we also have

$$\mathbf{COV}_4(\mathbf{X}^*)\mathbf{H}^* = \mathbf{COV}(\mathbf{X}^*)\mathbf{H}^*\boldsymbol{\Lambda}(\mathbf{X}).$$

The scores or invariant coordinates of $\mathbf{X}^*$ are given by

$$\mathbf{Z}^* = \mathbf{H}^{*T}(\mathbf{X}^* - \mathbf{E}\mathbf{X}^*) \tag{10}$$

or equivalently by $z_k^* =< \mathbf{h}_k^*, \mathbf{X}^* - \mathbf{E}\mathbf{X}^* >, k = 1, \ldots, D-1$.

5.2 ICS in the Simplex

Let us now use Sect. 4 to obtain a formulation of the previous results back in the simplex. This presentation of ICS involves elements (scatter matrices, eigenvalues, and eigenvectors) that are independent of the particular choice of contrast matrix, thus justifying this approach. Let us denote by $\mathbf{L}$ the following matrix:

$$\mathbf{L} = (\mathbf{COV}^{\oplus}\mathbf{X})^{-1/2}\mathbf{COV}_4^{\oplus}\mathbf{X}(\mathbf{COV}^{\oplus}\mathbf{X})^{-1/2}. \tag{11}$$

By Theorem 1, we have that

$$\mathrm{ilr}_V(\mathbf{L}) = \mathbf{L}^*, \tag{12}$$

and by Theorem 3, we have that, for $k = 1, \ldots, D$,

$$\mathbf{L} \boxdot \mathbf{u}_k = \lambda_k \odot \mathbf{u}_k,$$

where $\mathbf{u}_k = \mathrm{ilr}_V^{-1}(\mathbf{u}_k^*)$ for $k = 1, \ldots, D-1$, and $\mathbf{u}_D = \mathbf{1}_D/\sqrt{D}$ corresponding to $\lambda_D = 0$. We have $< \mathbf{u}_k, \mathbf{u}_l >_A = \delta_{kl}$, for $k, l = 1, \ldots, D$. The vectors $\mathbf{u}_k$ are the $\mathcal{A}$-eigenvectors of $\mathbf{L}$. We can write the following spectral representation of $\mathbf{L}$:

$$\mathbf{L} = \sum_{k=1}^{D-1} \lambda_k \mathrm{clr}(\mathbf{u}_k)\mathrm{clr}(\mathbf{u}_k)^T.$$

If we denote by $\mathbf{h}_k = \mathrm{ilr}_V^{-1}(\mathbf{h}_k^*) = (\mathbf{COV}^{\oplus}\mathbf{X})^{-1/2} \boxdot \mathbf{u}_k$, $k = 1, \ldots, D$, we get

$$\mathbf{COV}_4^{\oplus}\mathbf{X} \boxdot \mathbf{h}_k = \lambda_k \odot \mathbf{COV}^{\oplus}\mathbf{X} \boxdot \mathbf{h}_k$$

and

$$(\mathbf{COV}^{\oplus}\mathbf{X})^{-1}\mathbf{COV}_4^{\oplus}\mathbf{X} \boxdot \mathbf{h}_k = \lambda_k \odot \mathbf{h}_k.$$

The scores $\mathbf{Z}^* = (z_1^*, \ldots, z_{D-1})$ defined by (10) do not depend on the contrast matrix as already mentioned in Muehlmann et al. (2021) and are given by

$$z_k^* =< \mathbf{h}_k^*, \mathbf{X}^* - \mathbf{E}\mathbf{X}^* > = < \mathbf{h}_k, \mathbf{X} \ominus \mathbf{E}^{\oplus}\mathbf{X} >_A . \tag{13}$$

This equation shows that the scores can be used for outlier detection independently of the contrast matrix.

5.3 Reconstruction Formula

From (10), it is easy to derive the reconstruction formula in coordinate space:

$$\mathbf{X}^* = \mathbf{E}\mathbf{X}^* + (\mathbf{H}^{*T})^{-1}\mathbf{Z}^*. \tag{14}$$

Let $\mathbf{a}_k^*$ denote the column vectors of the matrix $(\mathbf{H}^{*T})^{-1} = \mathbf{COV}(\mathbf{X}^*)^{1/2}\mathbf{U}^*$ for $k = 1, \ldots, D-1$. Let us define the scalar product with respect to the metric $\mathbf{COV}(\mathbf{X}^*)^{-1}$ by

$$< \mathbf{u}^*, \mathbf{v}^* >_{\mathbf{COV}(\mathbf{X}^*)^{-1}} = \mathbf{u}^{*T}\mathbf{COV}(\mathbf{X}^*)^{-1}\mathbf{v}^*.$$

Equation (8) shows that the vectors $\mathbf{a}_k^*$, $k = 1, \ldots, D-1$, are orthonormal in the sense of this scalar product since the equation can be rewritten as

$$(\mathbf{H}^*)^{-1}\mathbf{COV}(\mathbf{X}^*)^{-1}(\mathbf{H}^{*T})^{-1} = \mathbf{I}_{D-1}. \tag{15}$$

This orthogonality implies that the reconstruction formula can also be obtained by

$$\mathbf{X}^* - \mathbf{EX}^* = \sum_{k=1}^{D-1} < \mathbf{a}_k^*, \mathbf{X}^* - \mathbf{EX}^* >_{\mathbf{COV}(\mathbf{X}^*)^{-1}} \mathbf{a}_k^*. \tag{16}$$

The scalar products $< \mathbf{a}_k^*, \mathbf{X}^* - \mathbf{EX}^* >_{\mathbf{COV}(\mathbf{X}^*)^{-1}}$, $k = 1, \ldots, D-1$, are the coordinates of the $(D-1)$ vector:

$$(\mathbf{H}^*)^{-1}\mathbf{COV}(\mathbf{X}^*)^{-1}(\mathbf{X}^* - \mathbf{EX}^*).$$

Using (14), this vector can be written:

$$(\mathbf{H}^*)^{-1}\mathbf{COV}(\mathbf{X}^*)^{-1}(\mathbf{X}^* - \mathbf{EX}^*) = (\mathbf{H}^*)^{-1}\mathbf{COV}(\mathbf{X}^*)^{-1}(\mathbf{H}^{*T})^{-1}\mathbf{Z}^*. \tag{17}$$

Using (17) and (15), we get

$$(\mathbf{H}^*)^{-1}\mathbf{COV}(\mathbf{X}^*)^{-1}(\mathbf{X}^* - \mathbf{EX}^*) = \mathbf{Z}^*,$$

and thus

$$< \mathbf{a}_k^*, \mathbf{X}^* - \mathbf{EX}^* >_{\mathbf{COV}(\mathbf{X}^*)^{-1}} = z_k^*,$$

where $(z_1^*, \ldots, z_{D-1}^*)$ denote the coordinates of $\mathbf{Z}^*$.

Combining with (16), we get the final reconstruction formula in coordinate space

$$\mathbf{X}^* = \mathbf{EX}^* + \sum_{k=1}^{D-1} z_k^* \mathbf{a}_k^*. \tag{18}$$

Applying ilr_V^{-1} to Eq. (18), we get the following simplex version of the reconstruction formula

$$\mathbf{X} = \mathbf{E}^{\oplus}\mathbf{X} \bigoplus_{k=1}^{D-1} z_k^* \odot \mathbf{a}_k, \tag{19}$$

where

$$\mathbf{a}_k = \mathrm{ilr}_V^{-1}(\mathbf{a}_k^*) = (\mathbf{COV}^{\oplus}\mathbf{X})^{1/2} \boxdot \mathbf{u}_k. \tag{20}$$

The vectors $\mathbf{a}_k$ are related to the $\mathscr{A}$-eigenvectors $\mathbf{u}_k$ of $\mathbf{L}$ by (20). They generate simplex lines called ICS axes that are the sets of vectors $\alpha \odot \mathbf{a}_k$, for $\alpha \in \mathbb{R}$. In the next section, we use the empirical version of the reconstruction formula (19) in order to plot the projection of the data on the vector $\mathbf{a}_1$ in some ternary diagrams in situations where the number of selected invariant coordinates is one.

We can also write (19) in terms of the vectors $\mathbf{h}_k$:

$$\mathbf{X} = \mathbf{E}^{\oplus}\mathbf{X} \bigoplus_{k=1}^{D-1} < \mathbf{h}_k, \mathbf{X} \ominus \mathbf{E}^{\oplus}\mathbf{X} >_A \odot \big(\mathbf{COV}^{\oplus}\mathbf{X} \boxdot \mathbf{h}_k\big).$$

6 Examples of Application

We first consider two artificial datasets following a mixture of two normal distributions with 10% of observations that differ from the 90% constituting the main bulk of the data. The dimension is $D = 3$ for the first example and $D = 20$ for the second one. The contrast matrices we use for the ilr transformations in this section are triangular Helmert matrices corresponding to the original ilr transformation defined by Egozcue et al. (2003).

6.1 Toy Examples

For the first example, the contrast matrix is given by $\mathbf{V}^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{2}\sqrt{\frac{2}{3}} & -\frac{1}{2}\sqrt{\frac{2}{3}} & \sqrt{\frac{2}{3}} \end{pmatrix}$.

In this toy example, $n = 100$ observations are generated in the ilr space with $D - 1 = 2$ dimensions from a mixture of two Gaussian distributions. The mean and the covariance matrix of the first 90% of the observations (sample 1) are, respectively,

$$\boldsymbol{\mu}_1^* = (0, 0)^T \text{ and } \boldsymbol{\Sigma}_1^* = 0.02\mathbf{I}_2 + 0.02\mathbf{1}_2\mathbf{1}_2^T,$$

while the mean vector and the covariance matrix of the remaining 10% (sample 2) are

$$\boldsymbol{\mu}_2^* = \left(\frac{2}{\sqrt{2}}\log 2, \frac{-1}{\sqrt{6}}\log 2\right)^T \text{ and } \boldsymbol{\Sigma}_2^* = 0.05\mathbf{I}_2.$$

Figure 1 on the left (resp., in the middle) shows the dataset in the simplex $\mathcal{S}^3$ (resp., in the ilr space). The points in cyan (resp., magenta) belong to sample 1 (resp., sample 2), and we can see that component x_2 has higher values in sample 2 than in sample 1, to the detriment of x_1 and x_3. We perform the ICS method

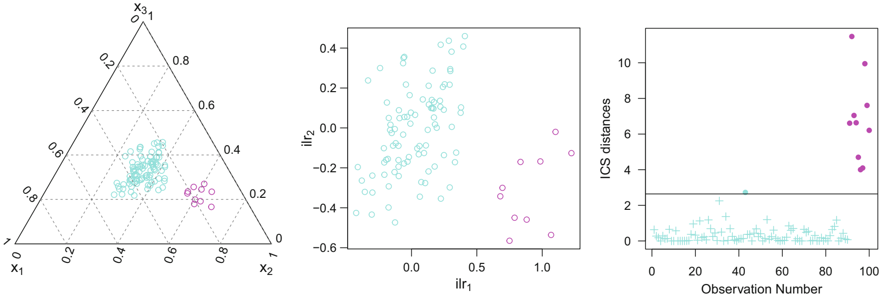


Fig. 1 First toy example: data in the simplex (left), data in the ilr space (middle), identification of the outlying observations using ICS (right)

in the ilr space using the `ICSOutlier` package (Archimbaud et al. 2018b). The eigenvalues are 1.57 and 0.81, and the D'Agostino test for normality leads to the selection of a single invariant component. Note that this test is based on the ICS scores and thus does not depend on the ilr transformation (see Archimbaud et al. 2018a, for more details). Figure 1 on the right reports the ICS distances as in Eq. (5) for each observation. The horizontal line represents a cutoff value based on Monte Carlo simulations and a 90% quantile. The choice of the quantile order can be done with respect to the expected percentage of outliers in the data. The ICS distances and the cutoff are also independent of the ilr transformation since they depend on the ICS scores only. Figure 1 on the right allows us to identify outliers represented by filled circles. On this example, all 10 observations from sample 2 are identified as outliers, whereas only 1 out of the 90 observations from sample 1 is incorrectly identified (at the limit of the cutoff value).

The two vectors generating the ICS axes (dashed lines in Fig. 2) are equal to $\mathbf{a}_1^* = (0.31, -0.1)$ and $\mathbf{a}_2^* = (0.12, 0.22)$ in the ilr space and $\mathbf{a}_1 = (0.27, 0.43, 0.30)$ and $\mathbf{a}_2 = (0.28, 0.33, 0.39)$ in the simplex space. To better understand the role of the ICS components and how they discriminate the observations, we represent in Fig. 2 the projections of the observations on the first ICS axis (left plots) and the second ICS axis (right plots) in the ilr space (top plots) and in the simplex space (bottom plots). The first ICS axis allows to discriminate the observations with a high value of x_2 relatively to the other shares and results in a good discrimination of the two groups. On the contrary, the second axis that seems to separate observations with high values of x_1 against observations with high values of x_3 does not allow to discriminate the two groups.

Finally, using the cutoff value, we represent in gray in Fig. 3 the zones or areas of the ilr space (left plot) and of the simplex (right plot) where the observations are considered as outliers. This confirms that the observations with a large or a small value of x_2 relatively to the other shares are in the outlying zone.

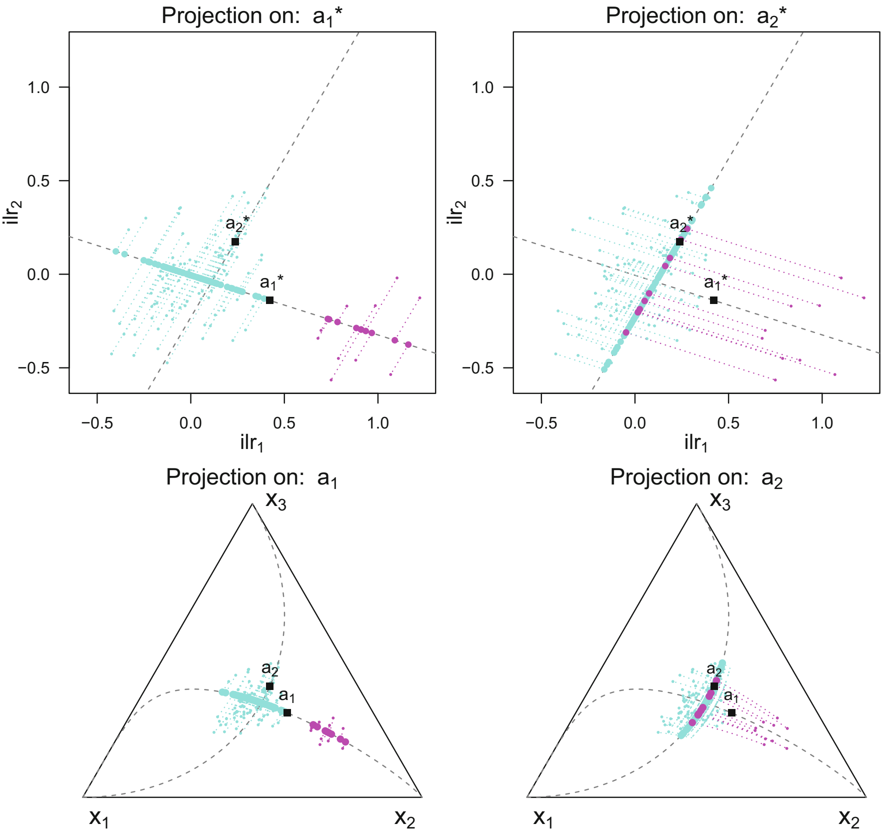


Fig. 2 First toy example: plot of the ICS axes and projections of the data on the ICS axes (ICS 1 on the left and ICS 2 on the right) in the ilr space (top plots) and in the simplex (bottom plots)

For the second toy example, we generate a higher dimensional dataset with $D = 20$, using two multivariate Gaussian distributions. The first sample is of size $n_1 = 90$ with

$$\boldsymbol{\mu}_1^* = (0, 0, \ldots, 0)^T \quad \text{and} \quad \boldsymbol{\Sigma}_1^* = 0.02\mathbf{I}_{D-1} + 0.02\mathbf{1}_{D-1}\mathbf{1}_{D-1}^T,$$

and the second sample is of size $n_2 = 10$ with

$$\boldsymbol{\mu}_2^* = \left(\frac{2}{\sqrt{2}}\log 2, \frac{-1}{\sqrt{6}}\log 2, 0, \ldots, 0\right)^T \text{ and}$$

$$\boldsymbol{\Sigma}_2^* = \begin{pmatrix} 0.05\mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & 0.02\mathbf{I}_{D-1} + 0.02\mathbf{1}_{D-3}\mathbf{1}_{D-3}^T \end{pmatrix}.$$

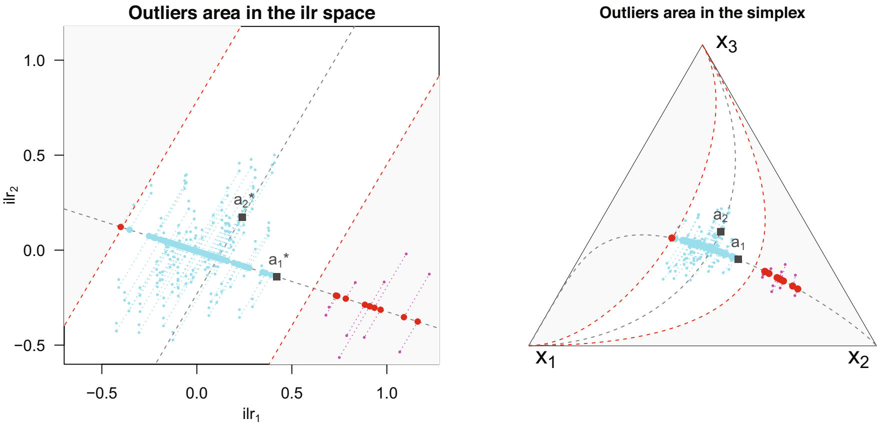


Fig. 3 First toy example: outliers zones in gray in the ilr space (left) and in the simplex space (right)

When $D > 3$, several options can be used for representing compositional data. One possibility is to plot ternary diagrams using sub-compositions as described in van den Boogaart and Tolosana-Delgado (2008). An alternative is to plot a ternary diagram with x_1, x_2 and the sum of the remaining parts $x_3 + \ldots + x_D$. Another possibility is to replace the sum of the remaining parts by their geometric mean. If $D > 3$ is not too large, these sub-ternary diagrams can be gathered in a square matrix of dimension $D(D-1)/2$.

In order to identify the outliers, we implement the ICS method using `ICSOutlier` in coordinate space. The procedure selects only the first invariant component. The left plot of Fig. 4 displays the ICS distances and the cutoff value as an horizontal line to identify outliers. This plot is the same for all ilr transformations. 9 observations out of 10 are detected as outliers in sample 2, while none of the observations from sample 1 are identified as outliers. The symbols for the points are as in Fig. 1.

The right plot represents several sub-ternary diagrams, but not all of them because of the large dimension $D = 20$. The selected ternary diagrams plot two parts among x_1 to x_5 against the geometric mean of the rest denoted by $*$. However, the diagrams that are not shown are very similar to the ones that focus on x_3, x_4, and x_5 (see the rows and columns 3, 4, and 5 on the matrix plot). Observations with the cross (resp., circle) symbol belong to sample 1 (resp., sample 2). The sub-ternary diagrams confirm that x_1 and x_2 are the composition parts playing a role in explaining the outlyingness of the red points. In fact, the observations of sample 1 are clearly visible and separated from the other group when considering the diagram with the x_1 and x_2 components and the geometric mean of the other parts. On the contrary, when looking at the ternary diagrams that do not take x_1 and x_2 separately from the other parts, the outliers are not distinct from the other observations.

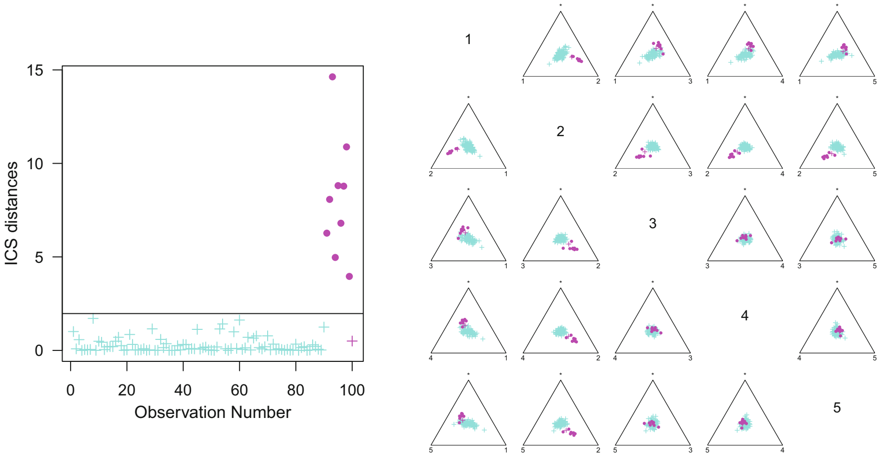


Fig. 4 Second toy example: ICS distances (left), sub-ternary diagrams of the first five composition parts (right), with cyan (resp., magenta) for sample 1 (resp., 2) and filled circles for detected outliers

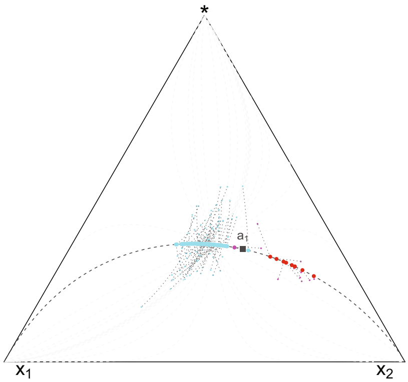


Fig. 5 Second toy example: plot of the ICS axis $\mathbf{a}_1$ and projections of the data on this axis in the ternary diagram $(x_1, x_2, *)$

We represent in Fig. 5 the sub-ternary diagram $(x_1, x_2, *)$ (where $*$ represents the geometric mean of the rest), with small circles in cyan (resp., magenta) for sample 1 (resp., sample 2). The vector $\mathbf{a}_1$ is plotted together with the ICS axis represented by a dashed line. We see that the data projected on the first ICS axis are clearly discriminated by high values of x_2 relatively to x_1.

6.2 Market Shares Example

This market share dataset has been simulated from a model fitted on the real European cars market in 2015 and is available in Barreiro et al. (2022). The plot on the top of Fig. 6 represents the shares in the French automobile market of 5 segments ($D = 5$), from January 2003 to August 2015, denoted by A, B, C, D, and E (European cars market segments, from the cheapest cars to the most powerful and luxury ones). We perform the ICS method in the ilr space and represent in the bottom of Fig. 6 the ICS distances for each observation. The normality test of the ICS procedure reveals that only the first component is important for outlier identification. The cutoff value is based on the quantile of order 97.5%. All the identified outliers are concentrated in a time interval between September 2008 and May 2009. It turns out that during this period, the global automobile market was undergoing a crisis with worldwide sales significantly down and political solutions have been provided such as the scrapping bonus at the end of 2008.

As before, in Fig. 7, we represent the matrix of sub-ternary diagrams with detected outliers in red. The ternary diagram vertices consist of two selected parts, and the third part indicated by $*$ corresponds to the geometric mean of the remaining

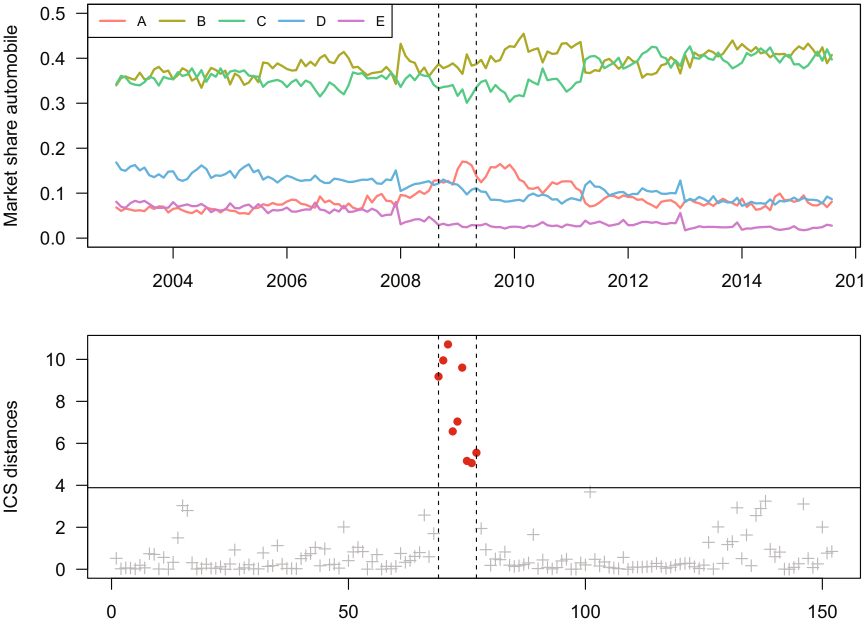


Fig. 6 French Market automobile shares example: from January 2003 to August 2015 in 5 segments (top) and identification of the outlying observations using ICS distances (bottom). The dotted vertical lines represent the period during which outliers were identified (September 2008 to May 2009)

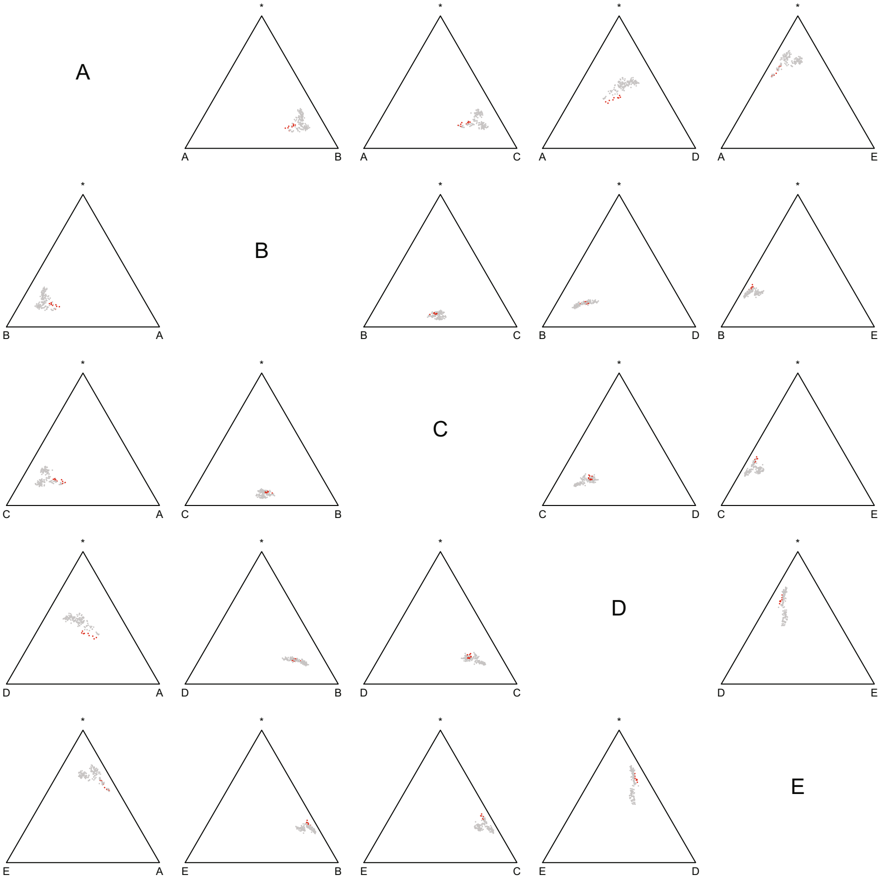


Fig. 7 French Market automobile shares example: outlier identification on the matrix sub-ternary diagram

parts. It seems that among all ternary diagrams, the ones including segment A are the best possible in order to identify the outliers. More precisely, the sub-ternary diagram that includes segments A, D, and the others separates the most the two groups. Thus, we plot in Fig. 8 the data in the sub-ternary diagram $(A, D, *)$. We also represent the vector $\mathbf{a}_1$, the ICS axis, and the projections of the data on this axis.

The time points that are detected as outlying correspond to observations with high values of segment A, compared to more normal values of D and low values of the geometric mean of B, C and E. This interpretation is confirmed when looking at the top plots of Fig. 6.

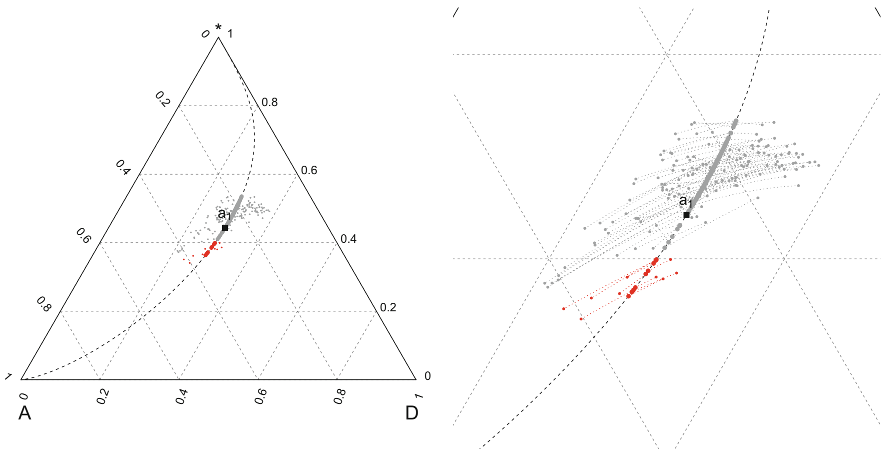


Fig. 8 French Market automobile shares example: projection of the data on the first ICS axis $\mathbf{a}_1$ in the sub-ternary diagram defined by A, D, and the amalgamation of other components (left). Zoom on the interesting part of the ternary diagram (right)

7 Conclusion

The present contribution extends ICS for outlier detection to the context of compositional data. As for standard data, ICS with the scatter pair **COV** and $\mathbf{COV}_4$ is a powerful tool to detect a small proportion of outliers. The definition of ICS in coordinate space is straightforward, and the identification of outliers does not depend on the choice of the isometric log-ratio transformation. The definition of ICS in the simplex is more challenging, and some algebra tools have been introduced to tackle the problem. Using a reconstruction formula, ICS axes can be plotted on ternary diagrams that help interpreting the outliers. Further interpretation tools are work in progress. Among the perspectives, we can mention the extension of ICS to compositional functional data (see Rieser and Filzmoser (2022) and Archimbaud et al. (2022)). Some supplementary material is available on https://github.com/tibo31/ics_coda in order to permit the reproducibility of the empirical analyses contained in the present paper.

Acknowledgments The authors would like to thank Pr. Dr. David Tyler for all his work in multivariate data analysis, which is an immense source of inspiration. We would like to thank also the editors and two reviewers for their suggestions and comments that helped us to improve the paper. We acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

Appendix

Proof of Theorem 1

1. Let $\mathbf{P}_V$ be the $D \times D$ block matrix $[\mathbf{V}\ \frac{1}{\sqrt{D}}\mathbf{1}_D]$. Then $\mathbf{P}_V^T\mathbf{P}_V = \mathbf{I}_D$ and $\mathbf{P}_V\mathbf{P}_V^T = \mathbf{V}\mathbf{V}^T + \frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T = \mathbf{I}_D$; therefore, $\mathbf{P}_V$ is invertible, and its inverse is equal to $\mathbf{P}_V^T$. If $\mathbf{A} = \mathbf{V}\mathbf{A}^*\mathbf{V}^T$ for a $(D-1)\times(D-1)$ matrix $\mathbf{A}^*$, then $\mathbf{A} = \mathbf{P}_V\begin{pmatrix} \mathbf{A}^* & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix}\mathbf{P}_V^T = \mathbf{P}_V\begin{pmatrix} \mathbf{A}^* & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix}\mathbf{P}_V^{-1}$; therefore, $\mathbf{A}$ is similar to $\mathbf{A}^*$ and their rank is equal.
2. If $\mathbf{A}^*$ is invertible, by the previous property, $\mathbf{A} = \mathbf{V}\mathbf{A}^*\mathbf{V}^T$ is also invertible. Then, let us first prove that $(\mathbf{A} + \frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T)$ is invertible. We can write

$$\mathbf{P}_V\begin{pmatrix} \mathbf{A}^* & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 1 \end{pmatrix}\mathbf{P}_V^T = \mathbf{A} + \frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T.$$

The rank of the central matrix is D; therefore, $\mathbf{A} + \frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T$ is invertible, and its inverse is given by

$$\begin{aligned}\left(\mathbf{P}_V\begin{pmatrix} \mathbf{A}^* & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 1 \end{pmatrix}\mathbf{P}_V^T\right)^{-1} &= \left(\mathbf{P}_V\begin{pmatrix} \mathbf{A}^* & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 1 \end{pmatrix}\mathbf{P}_V^{-1}\right)^{-1} \\ &= \mathbf{P}_V\begin{pmatrix} \mathbf{A}^{*-1} & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 1 \end{pmatrix}\mathbf{P}_V^T.\end{aligned}$$

Then let us check that the inverse of $\mathbf{A}$ in $\mathcal{A}$ is given by $\mathbf{P}_V\begin{pmatrix} \mathbf{A}^{*-1} & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix}\mathbf{P}_V^T$. Indeed $\mathbf{P}_V\begin{pmatrix} \mathbf{A}^{*-1} & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix}\mathbf{P}_V^T\mathbf{A} = \mathbf{P}_V\begin{pmatrix} \mathbf{A}^{*-1} & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix}\mathbf{P}_V^T\mathbf{P}_V\begin{pmatrix} \mathbf{A}^* & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix}\mathbf{P}_V^T = \mathbf{V}\mathbf{V}^T = \mathbf{G}_D$. Same for the other direction. Since $\mathbf{P}_V\begin{pmatrix} \mathbf{0}_{D-1}\mathbf{0}_{D-1}^T & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 1 \end{pmatrix}\mathbf{P}_V^T = \frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T$, we have

$$\begin{aligned}\mathbf{A}^{-1} &= \mathbf{P}_V\begin{pmatrix} \mathbf{A}^{*-1} & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 0 \end{pmatrix}\mathbf{P}_V^T \\ &= \mathbf{P}_V\begin{pmatrix} \mathbf{A}^{*-1} & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 1 \end{pmatrix}\mathbf{P}_V^T - \mathbf{P}_V\begin{pmatrix} \mathbf{0}_{D-1}\mathbf{0}_{D-1}^T & \mathbf{0}_{D-1} \\ \mathbf{0}_{D-1}^T & 1 \end{pmatrix}\mathbf{P}_V^T\end{aligned}$$

and thus $\mathbf{A}^{-1} = (\mathbf{A} + \frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T)^{-1} - \frac{1}{D}\mathbf{1}_D\mathbf{1}_D^T$. An alternative formula is

$$\mathbf{A}^{-1} = \mathbf{V}(\mathbf{V}^T\mathbf{A}\mathbf{V})^{-1}\mathbf{V}^T.$$

3. $\mathrm{ilr}_V(\mathbf{A})\mathrm{ilr}_V(\mathbf{B}) = \mathbf{V}^T\mathbf{A}\mathbf{V}\mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{V}^T\mathbf{A}\mathbf{B}\mathbf{V} = \mathrm{ilr}_V(\mathbf{A}\mathbf{B})$. If $\mathbf{A}$ is invertible, then $\mathrm{ilr}_V(\mathbf{A}^{-1}) = \mathbf{V}^T\mathbf{V}(\mathbf{V}^T\mathbf{A}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{V} = (\mathbf{V}^T\mathbf{A}\mathbf{V})^{-1} = (\mathrm{ilr}_V(\mathbf{A}))^{-1}$. If $(\mathrm{ilr}_V(\mathbf{A}))^{1/2}$ exists, let us define $\mathbf{A}^{1/2} = \mathrm{ilr}^{-1}\big((\mathrm{ilr}_V(\mathbf{A}))^{1/2}\big) = \mathbf{V}(\mathrm{ilr}_V(\mathbf{A}))^{1/2}\mathbf{V}^T$. We have $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{V}(\mathrm{ilr}_V(\mathbf{A}))^{1/2}\mathbf{V}^T\mathbf{V}(\mathrm{ilr}_V(\mathbf{A}))^{1/2}\mathbf{V}^T = \mathbf{V}\mathrm{ilr}_V(\mathbf{A})\mathbf{V}^T = \mathbf{A}$.

Proof of Theorem 2

1. 1 is a clear consequence of the fact that $\mathbf{G}_D$ is the neutral element of $\mathcal{A}$.
2. It is clear that $\mathrm{clr}(\mathbf{B})\mathrm{clr}(\mathbf{x}) \in \mathbf{1}_D^{\perp}$; hence, by definition, $\mathrm{clr}^{-1}(\mathrm{clr}(\mathbf{B})\mathrm{clr}(\mathbf{x})) = \mathrm{clr}(\mathbf{B}) \boxdot \mathbf{x}$.
3. If $\mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{V}^T\mathbf{A}\mathbf{V}$, then multiplying on the left by $\mathbf{V}$ and on the right by $\mathbf{V}^T$ and using the fact that $\mathbf{V}\mathbf{V}^T = \mathbf{G}_D$, we get $\mathbf{G}_D\mathbf{B}\mathbf{G}_D = \mathbf{G}_D\mathbf{A}\mathbf{G}_D$, and hence, $\mathrm{clr}(\mathbf{A}) = \mathrm{clr}(\mathbf{B})$. Then if $\mathbf{A} \in \mathcal{A}$, then $\mathrm{clr}(\mathbf{A}) = \mathbf{A} = \mathrm{clr}(\mathbf{B})$.
4. By Theorem 1, we have $\mathbf{A} = \mathbf{V}\mathrm{ilr}_V(\mathbf{A})\mathbf{V}^T$ and $\mathbf{A} = \mathrm{clr}(\mathbf{A})$ by 1.

Proof of Theorem 3 $\mathbf{A}^*$ is diagonalizable if there exists a basis $\mathbf{v}_1^*, \ldots, \mathbf{v}_{D-1}^*$ of $\mathbb{R}^{D-1}$ and $D-1$ real values λ_j such that $\mathbf{A}^*\mathbf{v}_j^* = \lambda_j\mathbf{v}_j^*$. Then let $\mathbf{e}_j = \mathrm{ilr}_V^{-1}(\mathbf{v}_j^*)$; we get by applying ilr^{-1}: $\mathbf{A} \boxdot \mathbf{e}_j = \mathrm{ilr}_V^{-1}(\lambda_j\mathbf{v}_j^*) = \lambda_j \odot \mathrm{ilr}_V^{-1}(\mathbf{v}_j^*) = \lambda_j \odot \mathbf{e}_j$ so that $\mathbf{e}_j$ is an $\mathcal{A}$-eigenvector of $\mathbf{A}$. Now applying the clr transformation, we also get that if $\mathbf{w}_j := \mathrm{clr}(\mathbf{e}_j)$, then $\mathbf{A}\mathrm{clr}(\mathbf{e}_j) = \lambda_j\mathrm{clr}(\mathbf{e}_j)$ so that $\mathbf{A}\mathbf{w}_j = \lambda_j\mathbf{w}_j$ showing that $\mathbf{w}_j$ is an eigenvector of $\mathbf{A}$. $\mathbf{1}_D/\sqrt{D}$ is an eigenvector of $\mathbf{A}$ associated to the eigenvalue 0 when $\mathbf{A} \in \mathcal{A}$, and this completes the basis in $\mathbb{R}^D$ since the vectors $\mathbf{w}_j$ belong to $\mathbf{1}_D^{\perp}$, $j = 1, \ldots, D-1$.

Proof of Theorem 4 The density of the elliptical distribution of $\mathbf{X}_V^* = \mathrm{ilr}_V(\mathbf{X})$ is a function of $R = (\mathrm{ilr}_V(\mathbf{X}) - \boldsymbol{\mu}_V^*)^T{\boldsymbol{\Sigma}_V^*}^{-1}(\mathrm{ilr}_V(\mathbf{X}) - \boldsymbol{\mu}_V^*)$. Since $\mathrm{ilr}_V(\mathbf{X}) = \mathbf{V}^T\mathrm{clr}(\mathbf{X})$, an alternative formulation for R is

$$R = (\mathrm{clr}(\mathbf{X}) - \mathrm{clr}(\boldsymbol{\mu}))^T\mathbf{V}^T{\boldsymbol{\Sigma}_V^*}^{-1}\mathbf{V}(\mathrm{clr}(\mathbf{X}) - \mathrm{clr}(\boldsymbol{\mu})).$$

Now if we let $\boldsymbol{\mu}_W^* = \mathbf{W}^T\mathbf{V}\boldsymbol{\mu}_V^*$, we have $\mathbf{W}\boldsymbol{\mu}_W^* = \mathbf{V}\boldsymbol{\mu}_V^*$. Similarly, let $\boldsymbol{\Sigma}_W^* = \mathbf{W}^T\mathbf{V}\boldsymbol{\Sigma}_V^*\mathbf{V}^T\mathbf{W}$, and we have $\mathbf{W}\boldsymbol{\Sigma}_W^*\mathbf{W}^T = \mathbf{V}\boldsymbol{\Sigma}_V^*\mathbf{V}^T$. Therefore, substituting this expression in R, we see that R is invariant to the specification of the contrast matrix, and going backward, we can rewrite $R = (\mathrm{ilr}_W(\mathbf{X}) - \boldsymbol{\mu}_W^*)^T{\boldsymbol{\Sigma}_W^*}^{-1}(\mathrm{ilr}_W(\mathbf{X}) - \boldsymbol{\mu}_W^*)$, which shows that $\mathrm{ilr}_V(\mathbf{X})$ follows an elliptical distribution with parameters $\boldsymbol{\mu}_W^*$ and $\boldsymbol{\Sigma}_W^*$. Now using the properties of contrast matrices $\mathbf{V}\mathbf{V}^T = \mathbf{G}_D$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{D-1}$, we have

$$(\mathbf{W}^T\mathbf{V}\boldsymbol{\Sigma}_V^*\mathbf{V}^T\mathbf{W})(\mathbf{W}^T\mathbf{V}{\boldsymbol{\Sigma}_V^*}^{-1}\mathbf{V}^T\mathbf{W}) = \mathbf{I}_{D-1},$$

which proves the last part of the theorem.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological), 44*(2), 139–160.

Archimbaud, A., Boulfani, F., Gendre, X., Nordhausen, K., Ruiz-Gazen, A., & Virta, J. (2022). ICS for multivariate functional anomaly detection with applications to predictive maintenance and quality control. *Econometrics and Statistics*, In press.

Archimbaud, A., Nordhausen, K., & Ruiz-Gazen, A. (2018a). ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis, 128*, 184–199.

Archimbaud, A., Nordhausen, K., & Ruiz-Gazen, A. (2018b). ICSOutlier: Unsupervised outlier detection for low-dimensional contamination structure. *The R Journal, 10*(1), 234–250.

Barreiro, I. R., Laurent, T., & Thomas-Agnan, C. (2022). *Regression models involving compositional variables*. R package, https://github.com/tibo31/codareg.

Bilodeau, M., & Brenner, D. (2008). *Theory of Multivariate Statistics*. New York: Springer.

Comas-Cufí, M., Martín-Fernández, J. A., & Mateu-Figueras, G. (2016). Log-ratio methods in mixture models for compositional data sets. *Sort, 1*, 349–374.

Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology, 35*(3), 279–300.

Egozcue, J. J., Barceló-Vidal, C., Martín-Fernández, J. A., Jarauta-Bragulat, E., Díaz-Barrero, J. L., & Mateu-Figueras, G. (2011). Elements of simplicial linear algebra and geometry. In V. Pawlowsky-Glahn, & A. Buccianti (Eds.), *Compositional data analysis*, chapter 11 (pp. 139–157). New York: Wiley.

Filzmoser, P., Hron, K., & Reimann, C. (2012). Interpretation of multivariate outliers for compositional data. *Computers & Geosciences, 39*, 77–85.

Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis: With worked examples in R*. Berlin: Springer.

Filzmoser, P., Ruiz-Gazen, A., & Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers, 55*(1), 29–47.

Mateu-Figueras, G., Monti, G. S., & Egozcue, J. (2021). Distributions on the simplex revisited. In *Advances in Compositional Data Analysis* (pp. 61–82). Berlin: Springer.

Muehlmann, C., Fačevicová, K., Gardlo, A., Janečková, H., & Nordhausen, K. (2021). Independent component analysis for compositional data. In A. Daouia, & A. Ruiz-Gazen (Eds.), *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan* (pp. 525–545). New York: Springer.

Nguyen, T. H. A. (2019). *Contribution to the statistical analysis of compositional data with an application to political economy*. PhD thesis, TSE, University Toulouse 1 Capitole.

Nordhausen, K. & Ruiz-Gazen, A. (2022). On the usage of joint diagonalization in multivariate statistics. *Journal of Multivariate Analysis, 188*, 104844.

Nordhausen, K. & Tyler, D. E. (2015). A cautionary note on robust covariance plug-in methods. *Biometrika, 102*(3), 573–588.

Nordhausen, K. & Virta, J. (2019). An overview of properties and extensions of FOBI. *Knowledge-Based Systems, 173*, 113–116.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modelling and Analysis of Compositional Data*. New York: Wiley.

Rieser, C. & Filzmoser, P. (2022). Outlier detection for pandemic-related data using compositional functional data analysis. In *Pandemics: Insurance and Social Protection* (pp. 251–266). Cham: Springer.

Rousseeuw, P. J. & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*(411), 633–639.

Theis, F. J. & Inouye, Y. (2006). On the use of joint diagonalization in blind signal processing. In *IEEE International Symposium on Circuits and Systems* (pp. 3589–3593). New York: IEEE.

Tyler, D. E., Critchley, F., Dümbgen, L., & Oja, H. (2009). Invariant co-ordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71*(3), 549–592.

van den Boogaart, K. G. & Tolosana-Delgado, R. (2008). "Compositions": A unified R package to analyze compositional data. *Computers & Geosciences, 34*(4), 320–338.

Robust Forecasting of Multiple Time Series with One-Sided Dynamic Principal Components

Daniel Peña and Víctor J. Yohai

Abstract Given a high-dimensional vector of time series, we define a class of robust forecasting procedures based on robust one-sided dynamic principal components. Peña et al. (J Am Stat Assoc 114(528):1683–1694, 2019) defined one-sided dynamic principal components as linear combinations of the present and past values of the series with optimal reconstruction properties. In order to make the estimation of these components robust to outliers, we propose here to compute the principal components by minimizing the sum of squares of the M-scales of the reconstruction errors of all the variables. The resulting robust components are called scale one-sided dynamic principal components (S-ODPC), and an alternating weighted least squares algorithm to compute them is presented. We prove that when both the number of series and the sample size tend to infinity, if the data follow a dynamic factor model, the mean of the squares of the M-scales of the reconstruction errors of the S-ODPC converges to the mean of the squares of the M-scales of the idiosyncratic terms, with rate $m^{1/2}$, where m is the number of dimensions. A Monte Carlo study shows that the S-ODPC introduced in this chapter can be successfully used for forecasting high-dimensional multiple time series, even in the presence of outlier observations.

Keywords Outliers · Dimensionality reduction · S-estimators · Dynamic factor models

D. Peña
Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain
e-mail: danielpena@uc3m.es

V. J. Yohai (✉)
Department of Mathematics and Instituto de Calculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina
e-mail: vyohai@dm.uba.ar

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_11

1 Introduction

High-dimensional sets of correlated time series are nowadays automatically generated in many fields, from engineering to environmental science and economics. These large data sets are often collected by using wireless sensor networks that may fail to record the data correctly due to depletion of batteries or environmental influence. These failures will produce outliers in the time series recorded that can modify strongly the forecasting generated from these contaminated data. Thus, using robust forecasting procedures based on robust estimation methods, which can deal with large outliers, is very important in this high-dimensional data sets.

Robust estimation of multivariate data sets generated by vector autoregressive (VAR) models was studied by Muler and Yohai (2013). They generalized to the multivariate case the robust estimation procedure proposed by Muler et al. (2009) for ARMA univariate models. However, the multivariate approach requires the estimation of a VAR/VARMA model for the vector of time series, and for these models, the number of parameters grows at least with the square of the number of series, turning their estimation unfeasible for high-dimensional sets of time series. Therefore, other alternatives for robust estimation in these situations have been explored. Dynamic factor models have been shown useful to model high-dimensional sets of time series, and some procedures have been proposed for the robust estimation of these models, see Fan et al. (2019), Alonso et al. (2020), Fan et al. (2021), and Trucíos et al. (2021).

Peña and Yohai (2016), following Brillinger's idea of dynamic principal components, see Brillinger (1964, 1981), proposed a new class of dynamic principal components that provide an optimal reconstruction of the observed set of series. These dynamic components are more general than those of Brillingers's, as they are computed without the assumption of being linear combinations of the data. In Peña and Yohai (2016) was also developed a robust estimation procedure for these principal components. However, although they are useful for reconstruction of the set of time series, these components are not expected to work well in forecasting problems, as their last values will be computed with a smaller number of observations than the central ones.

In order to have dynamic components useful for forecasting, (Peña et al. 2019) proposed the one-sided dynamic principal components (ODPC) that are defined as linear combinations of the observations based on a one-sided filter of past and present observations, instead of the double filter of past and future values, as proposed by Brillinger (1964). See also Forni et al. (2015, 2017) for a related approach to build one-sided filters for dynamic factor models.

Since the estimation procedure applied in Peña et al. (2019) minimizes the mean square error of the series reconstruction, it can be very sensitive to outliers. To overcome this drawback, in this chapter, we introduce a robust ODPC procedure that is based on the minimization of the sum of squares of M-scales of the reconstruction error of all the variables. Thus, the resulting forecasting procedure can be applied to automatic forecasting of large high-dimensional data sets of time series. The M-scales introduced by Huber (1964) are robust estimators that measure how large in absolute value are the elements of a sample. These scales may have a 50 % breakdown point against outliers and inliers. Therefore, they protect for both types of anomalies. We call this procedure S-ODPC, and by means of a Monte Carlo procedure, we show that it produces accurate forecasts even with outlier contaminated data.

In Sect. 2 of this chapter, we review the one-sided dynamic principal components (ODPC) proposed by Peña et al. (2019). In Sect. 3, we introduce its robust version, the S-ODPC, and in Sect. 4, we describe an alternating weighted least squares to compute them. In Sect. 5, we discuss how to use the S-ODPC to forecast future values of a multiple time series. In Sect. 6, two possible robust strategies to determine the number of components and the number of lags to define each component and to reconstruct the time series are described. In Sect. 7, we show that asymptotically, when both the number of series and the sample size go to infinity, if the data follow a dynamic factor model, the reconstruction obtained with S-ODPC converges, in mean squared error, to the common part of a dynamic factor model. In Sects. 8 and 9, we illustrate with Monte Carlo simulations and with a real data example that in the presence of outliers the forecasting procedure based on S-ODPC performs much better than the one based on ODPC. Finally, Sect. 10 contains conclusions. An Appendix includes the mathematical proofs.

2 One-Sided Dynamic Principal Components

Consider a zero mean vector of stationary time series $\mathbf{z}_1, \ldots, \mathbf{z}_T$, where $\mathbf{z}_t = (z_{t,1}, \ldots, z_{t,m})'$. Let $\mathbf{Z}$ be the data matrix of dimension $T \times m$ where each row is $\mathbf{z}_t'$. We will use $\mathbf{E}$ for the expectation operator, $\|\cdot\|$ for the Euclidean norm of vectors and the spectral norm of matrices and $\|\cdot\|_F$ for the Frobenius norm of a matrix. Consider an integer number $k_1 \geq 0$, and let $\mathbf{a} = (\mathbf{a}_0', \ldots, \mathbf{a}_{k_1}')'$, where $\mathbf{a}_h' = (a_{h,1}, \ldots, a_{h,m})$ is a vector of dimension m. Following Peña et al. (2019), the scores of the first one-sided dynamic principal component are of the form

$$f_t(\mathbf{a}) = \sum_{h=0}^{k_1} \mathbf{a}_h' \mathbf{z}_{t-h}, \quad t = k_1 + 1, \ldots, T. \tag{1}$$

This component, built with k_1 lags, is used to reconstruct each observation $z_{t,j}$ using $k_2 \geq 0$ of its lags and the respective loading coefficients by

$$\widehat{z}_{t,j}(\mathbf{a}, \mathbf{B}) = \varphi_j + \sum_{h=0}^{k_2} b_{j,h} f_{t-h}(\mathbf{a}). \tag{2}$$

Let $\mathbf{B}$ be the $m \times (k_2+2)$ loading matrix with row j equal to $(\varphi_j, b_{j,0}, \ldots, b_{j,k_2})$. Then if

$$\widehat{z}_t(\mathbf{a},\mathbf{B}) = \widehat{z}_{t,1}(\mathbf{a},\mathbf{B}), \ldots, \widehat{z}_{t,m}(\mathbf{a},\mathbf{B}))'$$

and $\mathbf{F}_t(\mathbf{a}) = (1, f_t(\mathbf{a}), \ldots, f_{t-k_2}(\mathbf{a}))$, (2) can be written as

$$\widehat{\mathbf{z}}_t(\mathbf{a},\mathbf{B}) = \mathbf{B}\mathbf{F}_t(\mathbf{a}).$$

Note that we can consider the reconstruction (2) as the predictions of the common component in a dynamic factor model (DFM). This equation can be interpreted as the forecast of the common component of a DFM with one dynamic factor and k_2 lags. The loadings are given by the matrix $\mathbf{B}$. The dynamic factor is assumed to be a linear combination of the observations and their k_1 lags, defined by the $\mathbf{a}$ weights in (1).

We suppose here that k_1 and k_2 are given, and in Sect. 6, we will propose a method to choose them. We will call $T^* = T - (k_1 + k_2)$ to the number of observations that can be reconstructed. The population optimal values of $\mathbf{a}$ and $\mathbf{B}$ were defined by Peña et al. (2019) as those that minimize the mean squared error in the reconstruction of the data

$$(\mathbf{a}^*, \mathbf{B}^*) = \arg\min_{\mathbf{a},\mathbf{B}} \mathbf{E}(\|\mathbf{z}_t - \widehat{\mathbf{z}}_t(\mathbf{a},\mathbf{B})\|^2).$$

Since if $(\mathbf{a},\mathbf{B})$ is a solution of (2), then $(c\mathbf{a}, \mathbf{B}/c)$, for $c \neq 0$, will be one as well, we can normalize the vector $\mathbf{a}$, so that $\|\mathbf{a}\| = 1$, although, as in standard principal components $-\mathbf{a}$, $-\mathbf{B}$ works as well as $\mathbf{a}, \mathbf{B}$. Given a sample, $\mathbf{z}_1, \ldots, \mathbf{z}_T$, the estimators $\widehat{\mathbf{a}}$, and $\widehat{\mathbf{B}}$ of the optimal values $(\mathbf{a}^*, \mathbf{B}^*)$ are defined as

$$(\widehat{\mathbf{a}}, \widehat{\mathbf{B}}) = \arg \min_{\|\mathbf{a}\|=1,\mathbf{B}} \frac{1}{T^*} \sum_{t=(k_1+k_2)+1}^{T} \|\mathbf{z}_t - \widehat{\mathbf{z}}_t(\mathbf{a},\mathbf{B})\|^2, \tag{3}$$

and the estimated first dynamic principal component is given by

$$\widehat{f}_t = f_t(\widehat{\mathbf{a}}) = \sum_{h=0}^{k_1} \widehat{\mathbf{a}}_h' \mathbf{z}_{t-h}, \; k_1 + 1 \leq t \leq T, \tag{4}$$

and $\widehat{\mathbf{z}}_t = \widehat{\mathbf{z}}_t(\widehat{\mathbf{a}}, \widehat{\mathbf{B}})$ will provide an estimated optimal reconstruction of $\mathbf{z}_t$ using k_2 of its lags at periods t, $(k_1 + k_2) + 1 \leq t \leq T$.

The second and higher orders of one-sided dynamic principal components are defined similarly. Let $k_1^{(h)}$ and $k_2^{(h)}$ be the number of lags to define the h-th component, and suppose that we have already computed the first l principal components. Denote by $\mathbf{r}_t^{(l)}$, $\max_{1\leq h\leq l}(k_1^{(h)} + k_2^{(h)}) + 1 \leq t \leq T$, the residual vector at time t using the first l components. Then, the $(l+1)$ one-sided dynamic estimated component is a vector with components of the form

$$f_t(\mathbf{a}) = \mathbf{a}_0'\mathbf{z}_t + \mathbf{a}_1'\mathbf{z}_{t-1} + \ldots + \mathbf{a}'_{k_1^{(l+1)}}\mathbf{z}_{t-k_1^{(l+1)}}, \max_{1\le h\le l+1}(k_1^{(h)} + k_2^{(h)}) + 1 \le t \le T, \tag{5}$$

where the vector $\mathbf{a} = (\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{k_1})$ is chosen so that it optimizes the reconstruction of $\mathbf{r}_t^{(l)}$, $\max_{1\le h\le l+1}(k_1^{(h)} + k_2^{(h)}) + 1 \le t \le T$. More precisely, consider reconstructions of $\mathbf{r}_t^{(l)}$ of the form $\widehat{\mathbf{r}}_t^{(l)}(\mathbf{a}, \mathbf{B}) = \mathbf{BF}_t(\mathbf{a})$, where $\mathbf{B}$ is a $m \times (k_2^{(l)} + 2)$ matrix, and then the $(l+1)$-th principal component has values $f_t(\widehat{\mathbf{a}}^{(l+1)})$, where $\widehat{\mathbf{a}}^{(l+1)}$ is defined so that there exists a $m\times$ $(k_2^{(l+1)} + 2)$ matrix $\widehat{\mathbf{B}}^{(l+1)}$ such that

$$(\widehat{\mathbf{a}}^{(l+1)}, \widehat{\mathbf{B}}^{(l+1)}) = \arg \min_{||\mathbf{a}||=1,\mathbf{B}} \frac{1}{T^*} \sum_{t=k_1^{(l+1)}+k_2^{(l+1)}+1}^{T} \left\| \mathbf{r}_t^{(l)} - \widehat{\mathbf{r}}_t^{(l)}(\mathbf{a}, \mathbf{B}) \right\|^2.$$

More technical details can be found in Peña et al. (2019).

We will consider here only the estimating equations of the first component, and therefore, we will drop the superscript (l). The estimating equations of higher order principal components can be found in Peña et al. (2019). Let $\mathbf{Z}_h$ be the $T^* \times m$ data matrix of T^* consecutive observations

$$\mathbf{Z}_h = \begin{pmatrix} \mathbf{z}'_{h+1} \\ \mathbf{z}'_{h+2} \\ \vdots \\ \mathbf{z}'_{h+T^*}, \end{pmatrix} \tag{6}$$

and we will consider this matrix for $h = 0, \ldots, (k_1 + k_2)$. Merging these matrices, we can write in a compact way the data used in the estimation. Note that the matrix $\mathbf{Z}_{k_1+k_2}$ includes all the values of the series to be reconstructed.

Second, we will consider the larger matrix $\mathbf{Z}_{l,k_1} = \left[\mathbf{Z}_h, \mathbf{Z}_{h-1}, \ldots, \mathbf{Z}_{h-k_1}\right]$ of dimension $T^* \times m(k_1 + 1)$ that includes the observations and also their k_1 lags required for the computation of the first component. The matrix of values of the components used for the reconstruction is the $T^* \times (k_2 + 2)$ matrix $\mathbf{F}_{k_1,k_2}(\mathbf{a})$, which has as rows $\mathbf{F}_t(\widehat{\mathbf{a}})$, $k_1 + k_2 + 1 \le t \le T$, and the reconstruction of the values $\mathbf{Z}_{k_1+k_2}$ is made with the $T^* \times m$ matrix $\widehat{\mathbf{Z}}_{k_1+k_2}$ computed as

$$\widehat{\mathbf{Z}}_{k_1+k_2} = \mathbf{F}_{k_1,k_2}(\widehat{\mathbf{a}})\widehat{\mathbf{B}}'.$$

Third, let $\mathbf{Z}^B$ be the matrix with dimensions $T^*(k_1 + 1) \times m(k_1 + 1)$ given by

$$\mathbf{Z}^B = \begin{pmatrix} \mathbf{Z}_{k_1+k_2,k_1} \\ \vdots \\ \mathbf{Z}_{k_1,k_1} \end{pmatrix}, \tag{7}$$

$\mathbf{B}_1$ the matrix $\mathbf{B}$ with its first column deleted and $\mathbf{I}_d$ the $d \times d$ identity matrix. Define the matrix of products of loadings and data values by

$$\mathbf{X}(\mathbf{B}) = (\mathbf{B}_1 \otimes \mathbf{I}_{T^*})\mathbf{Z}^B \tag{8}$$

as a $mT^* \times m(k_1+1)$ matrix with rank $m(k_1+1)$. Then, in Peña et al. (2019), it is shown that $(\widehat{\mathbf{a}}, \widehat{\mathbf{B}})$ are values $(\mathbf{a}, \mathbf{B})$ satisfying

$$\mathbf{a} = (\mathbf{X}(\mathbf{B})'\mathbf{X}(\mathbf{B}))^{-1}\mathbf{X}(\mathbf{B})' vec(\mathbf{Z}_{k_1+k_2}), \tag{9}$$

and this vector is standardized to unit norm. On the other hand, $\widehat{\mathbf{B}}$ can also be computed by least squares by

$$\mathbf{B}' = (\mathbf{F}'_{k_1,k_2}(\mathbf{a})\mathbf{F}_{k_1,k_2}(\mathbf{a}))^{-1}\mathbf{F}'_{k_1,k_2}(\mathbf{a})\mathbf{Z}_{k_1+k_2}. \tag{10}$$

An alternating least squares algorithm to compute $(\widehat{\mathbf{a}}, \widehat{\mathbf{B}})$ can be carried out as follows. Given an initial value of $\widehat{\mathbf{a}}$, the matrix of values of the component $\mathbf{F}_{k_1,k_2}(\mathbf{a})$ is computed by (9), and the matrix $\widehat{\mathbf{B}}$ is obtained by (10). Then, this matrix allows to compute a new value of $\widehat{\mathbf{a}}$, by first applying (8) and then using (9). This alternating process is continued until convergence. A similar algorithm can be applied to obtain the i-th component for $i > 1$.

3 Robust One-Sided Dynamic Principal Components

Since the ODPC estimator described in Sect. 2 is based on the minimization of the reconstruction mean square error, this estimator is very sensitive to the presence of outliers in the sample. To address this problem, we are going to propose a class of robust one-sided principal components that will be called S-ODPC and that are based on a M-scale.

Given a random variable x, the M-scale $S_M(x)$ is defined by

$$\mathbf{E}\left(\rho\left(\frac{x}{S_M(x)}\right)\right) = b,$$

where $\rho\colon \mathbb{R} \to \mathbb{R}^+$ and ρ and b satisfy: (a) $\rho(0) = 0$, (b) $\rho(-x) = \rho(x)$, (c) $\rho(x)$ is non-decreasing for $x \geq 0$, (d) $\lim_{x\to\infty} \rho(x) = 1$, and (e) $0 < b < 1$. The scale $S_M(\mathbf{x})$ is a measure of how large are the values that x takes. Note that if $\rho(x) = x^2$ and $b = 1$, then $S_M(\mathbf{x})$ is the L_2-scale given by $S_M^2(\mathbf{x}) = \mathbf{E}(x^2)$.

Given a sample $\mathbf{x} = (x_1, x_2, \ldots x_n)$ of x, $S_M(x)$ may be estimated by $s_M(\mathbf{x})$ satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{x_i}{s_M(\mathbf{x})}\right) = b. \tag{11}$$

Generally, it is required that $E_{\Phi}(\rho(x)) = b$, where Φ is the standard normal distribution. This condition implies that if $\mathbf{x_n}$ is a sample of size n of a N(0,σ^2) distribution, then $\lim_{n\to\infty} s_M(\mathbf{x}_n) = \sigma^2$.

A family of ρ functions satisfying these properties is the Tukey biweight family defined by

$$\rho_c^{BS}(x) = \left[1 - \left(1 - \left(\frac{x}{c}\right)^2\right)^3\right] I(|x| \le c).$$

The breakdown point of a M-scale is $\varepsilon^* = \min(b.1 - b)$ and is maximized when $\delta = 0.5$, and in this case, $\varepsilon^* = 0.5$. The consistency condition $E_{\Phi}(\rho_c^{BS}(x)) = 0.5$ is satisfied when $c = 1.547$. This is the function used in the simulations in Sect. 8 and in the example in Sect. 9.

The M-scales have been used to define robust estimators for many statistical problems. We will mention here two classes of estimators based on a M-scale: S-estimators for regression (see Rousseeuw and Yohai 1984) and S-estimators of the scatter matrix and multivariate location (see Davies 1987).

When the first component is used to reconstruct the series, the reconstruction error of the variable j at the period t is given by

$$r_{t,j}(\mathbf{a}, \mathbf{B}) = z_{t,j} - \widehat{z}_{t,j}(\mathbf{a}, \mathbf{B}), \quad k_1 + k_2 + 1 \le t \le T. \tag{12}$$

The first population S-ODPC is defined by the values $\mathbf{a}_S^*, \mathbf{B}_S^*$ given by

$$\left(\mathbf{a}_S^*, \mathbf{B}_S^*\right) = \arg \min_{||\mathbf{a}||=1, B} \mathcal{K}_M(\mathbf{a}, \mathbf{B}),$$

where

$$\mathcal{K}_M(\mathbf{a}, \mathbf{B}) = \sum_{j=1}^{m} S_M^2(r_{t,j}(\mathbf{a}, \mathbf{B})).$$

Given a sample $\mathbf{z}_t$, $1 \le t \le T$, $\left(\mathbf{a}_S^*, \mathbf{B}_S^*\right)$ can be estimated by

$$\left(\widehat{\mathbf{a}}_S, \widehat{\mathbf{B}}_S\right) = \arg \min_{||\mathbf{a}||=1, B} k_M(\mathbf{a}, \mathbf{B}),$$

where

$$k_M(\mathbf{a},\mathbf{B}) = \sum_{j=1}^{m} s_M^2(\mathbf{r}_{.j}(\mathbf{a},\mathbf{B}))$$

and $\mathbf{r}_{.j}(\mathbf{a},\mathbf{B}) = (r_{t,j}(\mathbf{a},B))_{k_1+k_2+1\le t\le T}$.

Remark 1 Observe that if the performance of the principal components is measured by the sum of the squares of the M-scales of the empirical reconstruction residuals, then the S-ODPC defined above is still optimal when dealing with nonstationary series. Therefore, it may be used for forecasting even in this case.

In the Appendix, we prove that differentiating $S(\mathbf{a},\mathbf{B})$, $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{B}}$ satisfy expressions similar to (9) and (10), respectively. In fact, it can be shown that $\widehat{\mathbf{a}}_S$ and $\widehat{\mathbf{B}}_S$ are values $\mathbf{a}$ and $\mathbf{B}$ satisfying the following weighted least squares relationships:

$$\mathbf{a}^* = (\mathbf{X}(\mathbf{B})'\mathbf{W}(\mathbf{a},\mathbf{B})\mathbf{X}(\mathbf{B}))^{-1}\mathbf{X}(\mathbf{B})'\mathbf{W}(\mathbf{a},\mathbf{B})vec(\mathbf{Z}_{k_1+k_2}),$$

$$\widehat{\mathbf{a}} = \mathbf{a}^*/\big|\big|\mathbf{a}^*\big|\big|, \tag{13}$$

where $\mathbf{W}$ is a $mT^* \times mT^*$ diagonal matrix of weights. On the other hand, fixing $\mathbf{F}_{k_1,k_2}(\mathbf{a})$, the optimal $\mathbf{B}$ also satisfies a weighted least squares expression. Then, $\mathbf{B}' = (\mathbf{b}_1,\ldots,\mathbf{b}_m)$, where

$$\mathbf{b}_j = (\mathbf{F}_{k_1,k_2}(\mathbf{a})'\mathbf{W}_j(\mathbf{a},\mathbf{B})\mathbf{F}_{k_1,k_2}(\mathbf{a}))^{-1}\mathbf{F}_{k_1,k_2}(\mathbf{a})'\mathbf{W}_j(\mathbf{a},\mathbf{B})\mathbf{z}_j^*\ , \tag{14}$$

where $\mathbf{z}_j^*$ is the j-th column of $\mathbf{Z}_{k_1+k_2}$ and $\mathbf{W}_j$ is a $T^* \times T^*$ matrix of weights. The matrices $\mathbf{W}_j$, $1 \le j \le m$, are defined as follows: let $\psi(u) = \rho'(u)$ and $w(u) = \psi(u)/u$ if $u \ne 0$ and $w(0) = \lim_{u\to 0} w(u)$. Given $\mathbf{a}$ and $\mathbf{B}$, let $\sigma_j(\mathbf{a},\mathbf{B}) = s_M(\mathbf{r}_{.j}(\mathbf{a},\mathbf{B}))$ and $\mathbf{w}_j(\mathbf{a},\mathbf{B}) = (w_{t,j}(\mathbf{a},\mathbf{B}))_{k_1+k_2+1\le t\le T}$, $1 \le j \le m$, where

$$w_{t,j}(\mathbf{a},\mathbf{B}) = \frac{w(r_{t,j}(\mathbf{a},\mathbf{B})/\sigma_j(\mathbf{a},\mathbf{B}))}{\frac{1}{T^*}\sum_{k=k_1+k_{2+1}}^{T} w(r_{k,j}(\mathbf{a},\mathbf{B})/\sigma_j(\mathbf{a},\mathbf{B}))}. \tag{15}$$

Then, $\mathbf{W}_j(\mathbf{a},\mathbf{B})$ and $\mathbf{W}(\mathbf{a},\mathbf{b})$ have as diagonals $\mathbf{w}_j(\mathbf{a},\mathbf{B})'$ and $\mathbf{w}(\mathbf{a},\mathbf{B}) = (\mathbf{w}_1(\mathbf{a},\mathbf{B})',\ldots,\mathbf{w}_m(\mathbf{a},\mathbf{B})')'$, respectively. These weights penalize outliers reducing or removing their influence on the estimators.

Observe that $\mathbf{w}_j$, $\mathbf{W}$, $\mathbf{F}_{k_1,k_2}$, and $\mathbf{X}$ depend on $\mathbf{a}$ and $\mathbf{B}$, and therefore, Eqs. (13) and (14) cannot be used to directly compute them. In the next section, we propose an alternating weighted least squares algorithm that overcomes this problem.

The second and higher order S-ODPC can be defined in a similar way as that in the non-robust ODPC.

4 Computing Algorithm for the S-ODPC

We propose here an alternating weighted least squares algorithm for computing the first S-ODPC. Let $\mathbf{a}^{(i)}$, $\mathbf{B}^{(i)}$, and $\mathbf{f}^{(i)}$ be the values of $\mathbf{a}$, $\mathbf{B}$, and $\mathbf{f} = (f_{k_1+1,} \ldots, f_T)^{'}$ corresponding to the i-th iteration and $\delta \in (0, 1)$, a tolerance parameter to stop the iterations. In order to define the algorithm, it is enough to give: (1) initial values $\mathbf{a}^{(0)}$ and $\mathbf{B}^{(0)}$, (2) a rule that given the values of the i-th iteration $\mathbf{a}^{(i)}$ and $\mathbf{B}^{(i)}$, $1 \leq j \leq m$, establishes how to compute $\mathbf{a}^{(i+1)}$ and $\mathbf{B}^{(i+1)}$, $1 \leq j \leq m$, and (3) a stopping rule. Then, the iterated algorithm is as follows:

1. To obtain the initial values, we first compute a standard (non- dynamic) robust principal component $\mathbf{f}^{(0)} = (f_t^{(0)})_{1\leq t\leq T}$, for example using the proposal of Maronna (2005). Then, we take $\mathbf{B}^{(0)} = (\mathbf{b}_1^{(0)}, \ldots, \mathbf{b}_m^{(0)})$, where $\mathbf{b}_j^{(0)}, 1 \leq j \leq m$, is an S-regression estimator (see Rousseeuw and Yohai 1984) using as dependent variable $\mathbf{z}_{j,k_1+k_2} = (z_{k_1+k_2+1,j}, \ldots, z_{T,j})$ and as matrix of independent variables $\mathbf{F}_{k_1,k_2}^{(0)}$ with rows $\mathbf{F}_t^{(0)} = (1, f_t^{(0)}, f_{t-1}^{(0)}, \ldots, f_{t-k_2}^{(0)})$, $k_1 + k_2 + 1 \leq t \leq T$. Once obtained $\mathbf{B}^{(0)}$, we compute the matrix of residuals $\mathbf{R} = \mathbf{Z}_{k_1+k_2} - \mathbf{F}_{k_1,k_2}^{(0)}\mathbf{B}^{(0)\prime}$. This matrix is used to define the weights $w_{t,j}$ as in (15) and the diagonal matrix $\mathbf{W}^{(0)}$. We compute $\mathbf{a}^{(0)} = \mathbf{a}^*/||\mathbf{a}^*||$, where
$$\mathbf{a}^* = (\mathbf{X}(\mathbf{B}^{(0)})'\mathbf{W}^{(0)}\mathbf{X}(\mathbf{B}^{(0)})^{-1}\mathbf{X}(\mathbf{B}^{(0)})'\mathbf{W}^{(0)}vec(\mathbf{Z}_{k_1+k_2})$$
and $\mathbf{X}(\mathbf{B})$ is defined as in (8).
2. Given $\mathbf{a}^{(i)}$ and $\mathbf{B}^{(i)}$, we compute the reconstruction residuals matrix $\mathbf{R} = \mathbf{Z}_{k_1+k_2} - \mathbf{F}_{k_1,k_2}^{(i)}\mathbf{B}^{(i)\prime}$ and the corresponding new M- scales $\sigma_j^{(i)}$, $1 \leq j \leq m$. Using these residuals and scales, we obtain new weights $w_{t,j}(\mathbf{a}^{(i)}, \mathbf{B}^{(i)})$, $k_1 + k_2 + 1 \leq t \leq T$, and the corresponding diagonal matrices $\mathbf{W}_j(\mathbf{a}^{(i)}, \mathbf{B}^{(i)})$, $1 \leq j \leq m$ and $\mathbf{W}(\mathbf{a}^{(i)}, \mathbf{B}^{(i)})$. Then $\mathbf{B}^{(i+1)} = (\mathbf{b}_1^{(i+1)}, \ldots, \mathbf{b}_m^{(i+1)})$ is defined by (14) with $\mathbf{a} = \mathbf{a}^{(i)}$. To compute $\mathbf{a}^{(i+1)}$, we use (13) with $\mathbf{B} = \mathbf{B}^{(i+1)}$.
3. The stopping rule is as follows: stop when
$$\frac{S(\mathbf{a}^{(i)}, \mathbf{B}^{(i)}) - S(\mathbf{a}^{(i+1)}, \mathbf{B}^{(i+1)})}{S(\mathbf{a}^{(i)}, \mathbf{B}^{(i)})} \leq \delta.$$

As in Salibian-Barrera and Yohai (2006), it can be shown at each step the MSE decreases, and therefore, it converges to a local minimum. To obtain a global minimum, the initial value $\mathbf{f}^{(0)}$ should be close enough to the optimal one.

Note that, since the matrix $\mathbf{X}(\mathbf{B}) = (\mathbf{B}_1 \otimes \mathbf{I}_{T^*})\mathbf{Z}^B$ has dimensions $mT^* \times m(k_1 + 1)$, solving the associated least squares problem can be time- consuming for high-dimensional (large m) problems. The iterative nature of the algorithm we propose implies that this least squares problem will have to be solved several times for different $\mathbf{B}$ matrices. However, as the matrix $\mathbf{B}' \otimes \mathbf{I}_{T*}$ is sparse, it can be stored efficiently, and multiplying it with a vector is relatively fast. We found that for problems with a moderately large m, the following modification of our algorithm works generally faster: instead of finding the optimal $\mathbf{a}^{(i+1)}$ corresponding to $\mathbf{B}^{(i+1)}$, just do one iteration of coordinate descent for $\mathbf{a}^{(i+1)}$.

5 Forecasting Using the S-ODPC

Suppose that we have computed estimators of Q robust dynamic principal components and that the lags used for the q component were $(k_1^{(q)}, k_2^{(q)})$. Let $\widehat{\mathbf{f}}^{(q)} = (\widehat{f}_t^{(q)})_{k_1^{(q)}+1\le t\le T}$, $1 \le q \le Q$, be the estimated S-ODPC's and $\widehat{\mathbf{B}}^q$ the estimated reconstruction matrices. We will show now how we can predict the values of $\mathbf{z}_{T+1}, \dots, \mathbf{z}_{T+h}$ for some $h \ge 1$. For that purpose, fit a time series model for each component $\widehat{\mathbf{f}}^{(q)}$, $1 \le q \le Q$ (e.g., an ARMA model), using a robust procedure, and with this model, obtain predictions $\widetilde{f}_{T+l}^{(q)}$ of $f_{T+l}^{(q)}$, $1 \le q \le Q$, $1 \le l \le h$. We fit in the simulations AR models for each component, and these models were estimated using the filtered τ-estimation procedure described in Chapter 8 of Maronna et al. (2019). This procedure selects automatically the order of the AR model, gives robust estimators of its coefficients, and provides a filtered series $\widetilde{\mathbf{f}}^{(q)} = (\widetilde{f}_t^{(i)})_{k_1^{(q)}+1\le t\le T}$, $1 \le q \le Q$ cleaned of the detected outliers. With the help of these filtered series, we can obtain robust predictions as follows. Let $\widetilde{\mathbf{F}}_{T+l}^{(q)} = (1, \widetilde{f}_{T+l}^{(q)}, \widetilde{f}_{T+l-1}^{(q)}, \dots, \widetilde{f}_{T+l-k_2^q}^{(q)})'$; then a robust prediction $\mathbf{z}_{T+l}$ given the first T observations is

$$\widehat{\mathbf{z}}_{T+l|T} = \sum_{q=1}^{Q} \widehat{\mathbf{B}}^{(q)}\widetilde{\mathbf{F}}_{t+l}^{(q)}, \; 1 \le l \le h.$$

The filtered τ-estimation procedure is implemented in the function arima.rob of the R package robarima.

6 Selecting the Number of Lags and the Number of Components

An important problem is the selection of the number of dynamics components Q to use for prediction, and the number of lags $k_1^{(q)}$ and $k_2^{(q)}$, $1 \le q \le Q$, required to define each component. In order to simplify the presentation, we will assume that $k_1^{(q)} = k_2^{(q)} = k^{(q)}$. We can use two possible methods to select these values: (a) an information criterion and (b) cross-validation. Simulations performed for the ODPC in Peña et al. (2019) show that both procedures have similar efficiencies; However,

(a) is much faster than (b). For that reason, we propose to use (a). In what follows, we describe implementations of both procedures.

6.1 Selection Using an Information Criterion

In Peña et al. (2019), an adaptation of the Bai and Ng (2002) criterion for factor model was used to choose $k^{(q)}$ and Q for ODPC. Here, we modify this procedure for its use in S-ODPC. We should start given a maximum value K for $k^{(q)}$. We compute the first S-ODPC with $k^{(q)} = k$ for all values of k such that $0 \leq k \leq K$. For each of these values of k, we compute the residuals $(r_{t,j}^{(1,k)})_{2k+1 \leq t \leq T,} \, 1 \leq j \leq m$ and compute the M-scale $\mathrm{S}_M(\mathbf{r}_{.j}^{(1,k)})$ of each vector $\mathbf{r}_{.j}^{(1,k)} = (r_{2k+1,j}^{(1,k)}, \ldots, r_{T,j}^{(1,k)})$, $1 \leq j \leq m$. Let $\widehat{\sigma}_{1,k} = ((1/m)\sum_{j=1}^{m} \mathrm{S}_M(\mathbf{r}_{.j}^{(1,k)}))^{1/2}$. Let $T_{1,k}^{*} = T - 2k$; then we choose as $k^{(1)}$ the value of k among $0, \ldots, K$ that minimizes

$$\mathrm{BNG}_{1,k} = \log(\widehat{\sigma}_{1,k)}^{2}) + (k+1)\frac{\log(\min(T_{1,k}^{*}, m))}{\min(T_{1,k}^{*}, m)}.$$

Suppose we have already computed $q-1$ dynamic principal components, where the component i uses $k^{(i)}$ lags. Then we compute the q component with i lags for each $0 \leq i \leq K$ and the corresponding residuals matrix $(r_{t,j}^{(q,k)})_{h_{q,k}+1 \leq t \leq T, 1 \leq j \leq m}$, where $h_{q,k} = 2\Big(k + \sum_{i=1}^{q-1} k^{(i)}\Big)$. Let $T_{q,k}^{*} = T - h_{q,k}$, $\mathbf{r}_{.j}^{(q,k)} = (r_{h_{q,k}+1,j}^{(q,k)}, \ldots, r_{T,j}^{(q,k)})$ and $\widehat{\sigma}_{q,k} = ((1/m)\sum_{j=1}^{m} \mathrm{S}_M(\mathbf{r}_{.j}^{(q,k)}))^{1/2}$; then the value of $k^{(q)}$ is the value of $k, 0 \leq k \leq K,$ which minimize the following robustification of the Bai and Ng criterion

$$\mathrm{BNG}_{q,k} = \log(\widehat{\sigma}_{q,k)}^{2}) + (\sum_{i=1}^{q-1}((k^{(i)}+1) + k + 1)\frac{\log(\min(T_{q,k}^{*}, m))}{\min(T_{q,k}^{*}, m)}.$$

The selected number of components is $Q = q - 1$, where q is the minimum value q such that $\mathrm{BNG}_{q,k^{(q)}} \geq \mathrm{BNG}_{q,k^{(q-1)}}$.

6.2 Selection Using Robust Cross-validation

Suppose that we are interested in using the S-ODPC to predict $\mathbf{z}_{T+1}, \ldots, \mathbf{z}_{T+h}$. We can apply the following robust cross-validation procedure for selecting the number of components, Q, and $k^{(q)}$, $1 \leq q \leq Q$, the number of lags used for each component. Suppose that the first $T_1 < T$ observations are chosen as the training set, and the last $T - T_1$ observations as testing set. Then the training set will be used to compute all the loading vectors $\mathbf{a}^{(q,k)}$ for the q component with k lags and the testing set to evaluate the prediction power of any choice of the number of lags $k^{(q)}$ for each component q and the number of components Q.

The cross-validation procedure starts choosing k_1 as follows. For $0 \leq d \leq T - T_1 - i$, $1 \leq j \leq m$, $1 \leq i \leq h$ and $k \geq 0$, let $\widehat{z}^{(1,k)}_{T_1+d+i,j|T_1+d}$ be the prediction of $z_{T_1+d+i,j}$ using the first component with k lags and loading vector $\mathbf{a}^{(1,k)}$ corresponding to the periods $t \leq T_1 + d$, and let $\widehat{r}^{(1,k)}_{T_1+d+i,j|T_1+d} = \widehat{z}^{(1,k)}_{T_1+d+i,j|T_1+d} - z_{T_1+d+i,j}$ the corresponding prediction error. We evaluate the quality of the predictions up to h periods ahead using the first component with k lags by

$$SS^{(1)}_k = \sum_{j=1}^{m} \sum_{i=1}^{h} S_{\mathrm{M}}(\widehat{\mathbf{r}}^{(1,k,i)}_{.j}), \tag{16}$$

where $\widehat{\mathbf{r}}^{(1,k,i)}_{.j} = (\widehat{r}^{(1,k)}_{T_1+i,j|T_1}, \ldots, \widehat{r}^{(1,k)}_{T,j|T-i})$ is the vector of all the i periods ahead predictions. We select as $k^{(1)}$ the first k, such that $SS\,^{(1)}_k \leq SS^{(1)}_{k+1}$. Suppose now that we have already computed $q - 1$ components with lags $k^{(1)}, \ldots, k^{(q-1)}$. To obtain $k^{(q)}$, we proceed as when computing $k^{(1)}$, that is, for each k, we compute

$$SS^{(q)}_k = \sum_{j=1}^{m} \sum_{i=1}^{h} S_{\mathrm{M}}(\widehat{\mathbf{r}}^{(q,k,i)}_{.j}), \tag{17}$$

where $\widehat{\mathbf{r}}^{(q,k,i)}_{.j} = (\widehat{r}^{(q,k)}_{T_1+i,j|T_1}, \ldots, \widehat{r}^{(q,k)}_{T-h+i,j|T-h})$, $\widehat{r}^{(q,k)}_{T_1+d+i,j|T_1+d} = \widehat{z}^{(q,k)}_{T_1+d+i,j|T_1+d} - z_{T_1+d+i,j}$, and $\widehat{z}^{(q,k)}_{T_1+d+i,j|T_1+d}$ is the prediction of $z_{T_1+d+i,j}$ assuming that $\mathbf{z}_1, \ldots, \mathbf{z}_{T_1+d}$ are known, using the first $q-1$ components with lags $k^{(1)}, \ldots, k^{(q-1)}$ and the q component with k lags. The number of lags $k^{(q)}$ for the q component is defined as the first value of k, such that $SS_k \leq SS_{k+1}$. Similarly, the number of components Q is chosen as the first q such that $SS^{q}_{k(q)} \leq SS^{q+1}_{k(q+1)}$. The robustness of this procedures follows from the fact that all the options selected by the procedure are evaluated using a robust scale, see (16) and (17). The procedure to make the forecasting is described in Sect. 5. The case where k_1 may be different of k_2 can be treated similarly but with more computational effort.

The forecasting of a particular time series can be improved if we add a specific or idiosyncratic component that explains the residuals of the series. For that purpose, we may fit for each variable an ARMA model for the respective S-ODPC reconstruction residuals.

7 Asymptotic Behavior of the S-ODPC in Factor Models

Let $z_{t,j}^{(m)}$, $1 \leq j \leq m, m > 1$, be observations generated by a dynamic one-factor model with k lags, that is, they satisfy

$$z_{t,j}^{(m)} = \varphi_j^{(m)} + c_{j,0}^{(m)} f_t + \ldots + c_{j,k}^{(m)} f_{t-k} + u_{t,j}, 1 \leq j \leq m, \tag{18}$$

where f_t and $u_{t,j}$, $1 \leq j \leq m$, are independent stationary process. We also have $\mathbf{E}(u_{t,j}) = \mathbf{E}(f_t) = 0$, $\text{var}(f_t) = \tau^2$, and var $(u_{t,j}) = \sigma_j^2$.

In this section, we study the behavior of the first population S-ODPC when m tends to infinite. This is stated more precisely in Theorem 1, which is the analogous of Theorem 3 of Peña et al. (2019), but using S-ODPC instead of ODPC. Consider the following Assumptions:

A1. There exist $\varepsilon > 0$ and A_1 such that $0 < \varepsilon < \sigma_j^2 < A_1 < \infty$ for all j.

A2. Let $s_j = S_M(u_{tj})$; then

$$0 < s_j \leq A_2. \tag{19}$$

A3. The function ρ has a derivative ψ that is continuous and bounded. Then $A_3 = \sup \psi < \infty$.

A4. $A_4 = \inf_j \mathbf{E}(u_{t,j} \psi(u_{t,j})) > 0$.

A5. There exists C such that $\sup_m \sup_{1 \leq j \leq m, 0 \leq i \leq k} |c_{j,i}^{(m)}| \leq C$ and $\sup_m \sup_{1 \leq j \leq m} |\varphi_j^{(m)}| \leq C$.

A6. Let $\mathbf{c}_i^{(m)} = (c_{1,i}^{(m)}, ..., c_{m,i}^{(m)}), 0 \leq i \leq k$, and $E^{(m)}$ the subspace of $\mathbb{R}^m$ generated by $\mathbf{c}_i^{(m)}$, $1 \leq i \leq m$. Then, we can write $\mathbf{c}_0^{(m)} = \mathbf{d}^{(m)} + \mathbf{e}^{(m)}$,where $\mathbf{d}^{(m)}$ is orthogonal to $E^{(m)}$ and $\mathbf{e}^{(m)} \in E^{(m)}$. Then, there exists δ such that $||\mathbf{d}^{(m)}||^2 \geq m\delta$ for all m. This condition implies that the common part of $\mathbf{z}_t^{(m)} = (z_{t,1}^{(m)}, \ldots, z_{t,m}^{(m)})$ does not get close to the $k-1$- dimensional subspace $E^{(m)}$ when m increases.

Theorem 1 *Assume A1–A6. Let* $\mathbf{z}_t^{(m)} = (z_{t,1}^{(m)}, \ldots, z_{t,m}^{(m)})$ *generated as the dynamic one-factor model given in (18) and* $\mathbf{z}_t^{(m)*} = (z_{t,1}^{(m)*}, \ldots, z_{t,m}^{(m)*})$ *its population optimal reconstruction using the first S-ODPC with* k_1 *equal to any nonnegative integer and* k_2 *equal to* k*, the number of lags of the factor model. Then, there exists a constant* K *independent of* m *such that*

$$\frac{1}{m^{1/2}} \left(\sum_{j=1}^{m} S_M^2 (z_{t,j}^{(m)} - z_{t,j}^{*(m)})^2 - \sum_{j=1}^{m} S_M^2 (u_{t,j}) \right) \leq K. \tag{20}$$

Remark 2 This theorem can be generalized to a model with k factors, if the first k S-ODPC are defined simultaneously instead of sequentially so that they minimize the sum of the squares of the population M-scales. The proof is similar to the case of one factor.

8 Simulation Results

We generate matrices $(z_{t,j})$ of T =102 observations and $m = 50$ time series using the following dynamic factor model:

$$z_{t,j} = d_{j,1} f_t + d_{j,2}\, f_{t-1} + d_{j,3} f_{t-2} + 0.2u_{t,j},\ \ 1 \le t \le 102, 1 \le j \le 50,$$

where $u_{t,j}$ are i.i.d. $N(0, 1)$ and f_t follows the autoregressive model

$$f_t = 0.85 f_{t-1} + v_t$$

and v_t are i.i.d. $N(0, 1)$. The coefficients $d_{j,i}$ are generated in each replication as i.i.d. with uniform distribution in [0, 1].

The first 100 observations are used to obtain the one-sided dynamic components (ODPC and S-ODPC), and the last two to evaluate the prediction performance of these methods.

These values $z_{t,j}$ are contaminated as follows: for $t = kT/H, k = 1, 2, \ldots, H-1$, the values of the observed series are

$$z^*_{t,j} = z_{t,j} + K.$$

Three values of H are considered, 0, 5, and 10, that is, 0%, 4%, and 9% of outliers, respectively, and the values of K are 3, 5, 10, and 15.

The number of replications is 500. For each case, we compute the first ODPC and the first S-ODPC with $k_1 = 1$ and $k_2 = 2$. The performance of both procedures for predicting the values of the series one and two periods ahead is evaluated by the sum of squares of the M-scales of the respective prediction error. Let $s_{M,j}$, $1 \le j \le 50$, be the M-scale of the prediction errors of the j-th variable. Then the performance of each procedure is evaluated by

$$\sum_{j=1}^{50} s^2_{M,j}.$$

The results are shown in Table 1.

We observe that under outlier contamination the forecasting error using the S-ODPC is much smaller than that using ODPC.

Table 1 Sum of the squares of the M-scales of the prediction errors. Pr1 and Pr2 stand for one and two steps ahead, respectively

%out	K	Pr1 ODPC	Pr1 S-ODPC	Pr2 ODPC	Pr2 S-ODPC
0		21.23	22.69	74.19	77.20
4	3	36.63	23.90	108.73	83.27
	5	53.63	23.13	128.54	81.92
	10	85.54	23.52	157.62	79.94
	15	133.15	23.33	202.20	78.43
9	3	51.14	24.79	127.61	98.63
	5	66.69	24.33	138.69	82.47
	10	147.56	23.34	208.53	79.51
	15	254.02	24.17	299.28	82.75

9 Example with a Real Data Set

We consider a multiple time series $z_{t,j}$, $1 \leq t \leq 678$, $1 \leq j \leq 24$, the electricity price in the Connecticut region, New England, during the Thursdays of 676 weeks for each of the 24 hours of the day. Then we have 24 series of 676 observations. The data can be obtained at www.iso-ne.com and were previously considered for clustering time series by Alonso and Peña (2019).

Figure 1 shows plots of 12 of these 24 series. The series appear to be highly correlated, and therefore, dimension reduction is expected to be useful. We observe that at every hour of the day there are outliers especially between the weeks 500 and 600, and therefore, a robust procedure seems to be appropriate for these data.

For each d, $1 \leq d \leq 177$, we apply the ODPC and S-ODPC procedures to the set $z_{t,j}$, $1 \leq t \leq 500 + d$, and predict the values of $z_{500+d+1,j}$, $1 \leq j \leq 24$.

Figure 2 shows that in general the M-scales of the prediction errors of the S-ODPC are smaller than those of the ODPC specially between 10 am and 8 pm.

As indicated by one of the referees, the large values around weeks 500 and 600 may be due to some interesting facts that, if considered, can provide important insights in the data analysis. We have not investigated this important issue for a rigorous analysis of these data, as our objective here is to illustrate the performance of our robust procedure. Also, the plot of the series suggests that they are not stationary. However, as is mentioned in Remark 1, the S-ODPC may be a useful tool to predict future values of the series even in this case.

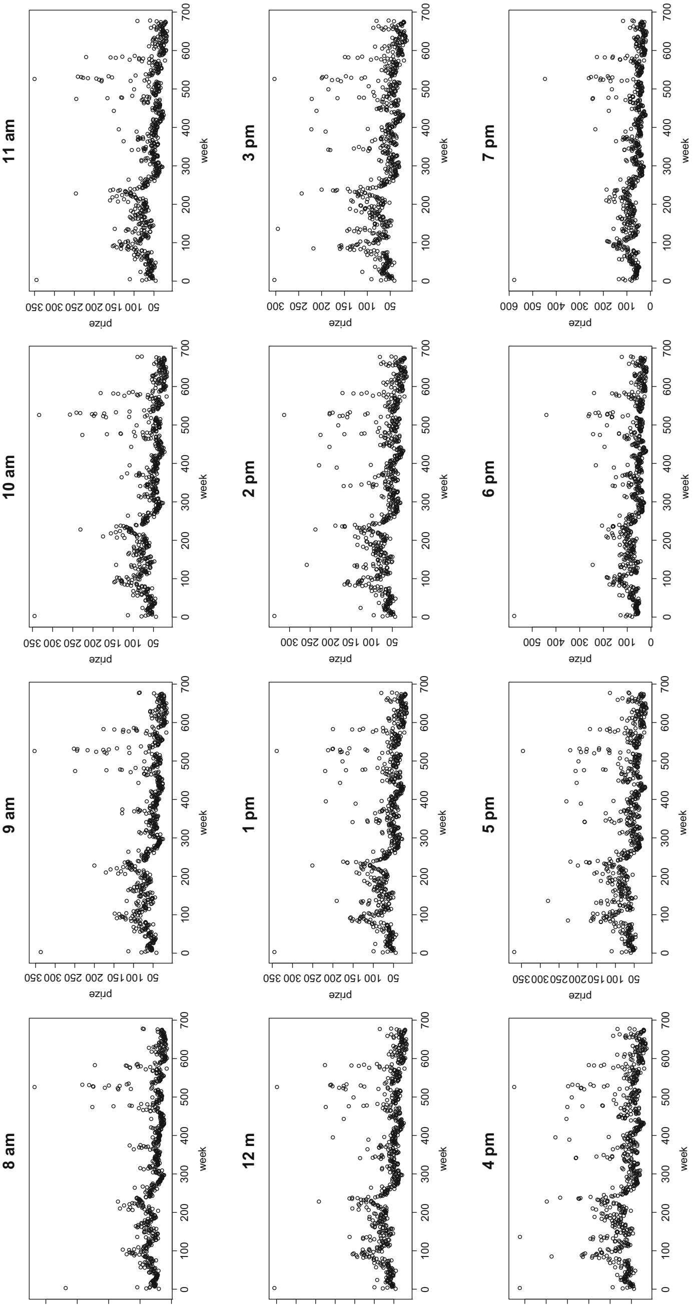


Fig. 1 Electricity data

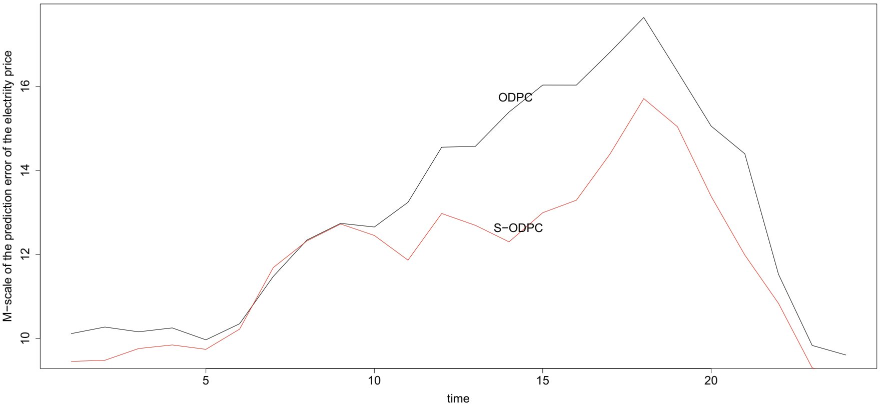


Fig. 2 M-scales of the one step ahead prediction errors

10 Conclusions

Given a vector series $\mathbf{z}_t$, $\leq t \leq T$, we have introduced the S one-sided dynamic principal components (S-ODPC) f_t, $k_1+1 \leq t \leq T$, defined as a linear combination of $\mathbf{z}_t$, $\mathbf{z}_{t-1} \dots \mathbf{z}_{t-k_1}$ that have the following properties:

- It allows the reconstruction $\widehat{\mathbf{z}}_t$ of the series $\mathbf{z}_t$ for $k_1 + k_2 - 1 \leq t \leq T$ as a linear combination of $f_t, \dots f_{t-k_2}$.
- It allows the forecasting of $\mathbf{z}_{T+h}$.
- The reconstruction of the series and the forecasting of future values of $\mathbf{z}_t$ can be improved taking higher order components.
- The values of k_1, k_2 and the number of components q can be chosen by: (a) a robust version of the Bai and Ng (2002) criterion for factor models or (b) cross-validation.
- The procedure S-ODPC is robust, that is, it can be applied successfully even in the presence of outliers.

Acknowledgments Daniel Peña was partially supported by the Spanish Agencia Nacional de Evaluación under grant PID2019-109196GB-I00, and Víctor J. Yohai was partially supported by Grants 20020170100330BA from the Universidad de Buenos Aires and PICT-201-0377 from ANPYCT.

Appendix

Derivation of the Estimating Equations

Call $\boldsymbol{\theta} = (\mathbf{a}, \mathbf{B})$; then the value of $\boldsymbol{\theta}$ for the S-ODPC estimator is obtained minimizing

$$k_M(\boldsymbol{\theta}) = \sum_{j=1}^{m} s_M^2(\mathbf{r}_{.j}(\boldsymbol{\theta})), \tag{21}$$

where $\mathbf{r}_{.j}(\boldsymbol{\theta}) = (r_{k_1+k_2+1,j}(\boldsymbol{\theta}), \ldots, r_{T,j}(\boldsymbol{\theta}))$, $r_{t,j}(\boldsymbol{\theta}) = z_{t,j} - \widehat{z}_{t,j}(\boldsymbol{\theta})$ satisfies

$$\frac{1}{T^*} \sum_{t=k_1+k_2+1}^{T} \rho\left(\frac{r_{t,j}(\boldsymbol{\theta})}{s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))}\right) = b, \tag{22}$$

with $T^* = T - k_1 - k_2$. Differentiating (21), we get that the optimal $\boldsymbol{\theta}$ for the S-ODPC satisfies

$$\sum_{j=1}^{m} s_M(\mathbf{r}_{.j}(\boldsymbol{\theta})) \frac{\partial s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = 0. \tag{23}$$

Let us differentiate (22). We have

$$\sum_{t=k_1+k_2+1}^{T} \psi\left(\frac{r_{t,j}(\boldsymbol{\theta})}{s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))}\right) \frac{s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))\partial r_{t,j}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} - r_{tj}(\boldsymbol{\theta})\partial s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))/\partial\boldsymbol{\theta}}{s_M^2(\mathbf{r}_{.j}(\boldsymbol{\theta}))} = 0$$

and

$$\frac{\partial s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \sum_{t=k_1+k_2+1}^{T} \psi\left(r_{t,j}(\boldsymbol{\theta})/s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))\right) \frac{s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))\partial r_{t,j}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{\sum_{t=k_1+k_2+1}^{T} \psi\left(r_{t,j}(\boldsymbol{\theta})/s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))\right) r_{t,j}(\boldsymbol{\theta})}.$$

This can also be written as

$$\frac{\partial s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))}{\sum_{t=k_1+k_2+1}^{T} w\left(r_{t,j}(\boldsymbol{\theta})/s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))\right) r_{t,j}^2(\boldsymbol{\theta})}$$
$$\times \sum_{t=k_1+k_2+1}^{T} w\left(\frac{r_{t,j}(\boldsymbol{\theta})}{s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))}\right) r_{t,j}(\boldsymbol{\theta}) \frac{\partial r_{t,j}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \; ,$$

where $w(u) = \psi(u)/u$. Then the estimating Eq. (23) satisfies

$$\sum_{t=k_1+k_2+1}^{T} w_{tj}^*(\boldsymbol{\theta}) r_{t,j}(\boldsymbol{\theta}) \frac{\partial r_{t,j}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \ , \tag{24}$$

where

$$w_{t,j}^*(\,\boldsymbol{\theta}) = \frac{s_M^2(\mathbf{r}_{.j}(\boldsymbol{\theta}))}{\frac{1}{T^*}\sum_{t=k_1+k_2+1}^{T} w\big(r_{t,j}(\boldsymbol{\theta})/s_M(\mathbf{r}_{.j}(\boldsymbol{\theta}))\big) r_{t,j}^2(\boldsymbol{\theta})} w(r_{t,j}(\boldsymbol{\theta})).$$

Note that when $w_{t,j}^*(\,\boldsymbol{\theta}) = 1$ for all (t, j), (24) is the estimating equation of the ODPC. Therefore, the only difference between the estimating equation of the RODPC and the one of the ODPC is that the least squares solutions to obtain the optimal values of **a** and **B** for the RODPC give weight $w_{t,j}^*(\,\boldsymbol{\theta})$ to the observation (t, j). Then we obtain (13) and (14).

Proof of Theorem 1

Lemma 1 *Suppose that $z_{t,j}$ satisfies A1, A4, A5, and A6; then there exist $\mathbf{a}^{(m)} \in \mathbb{R}^m$ and a $m \times (k+2)$ matrix $\mathbf{B}^{(m)}$ such that if $g_t^{(m)} = \mathbf{a}^{(\mathbf{m})\prime}\mathbf{z}_t$ and $\mathbf{F}_t^{(m)} = (1, g_t^{(m)}, \ldots, g_{t-k}^{(m)})$, the reconstruction $\widetilde{\mathbf{z}}_t^{(m)} = \mathbf{B}^{(m)}\mathbf{F}_t^{(m)}$ satisfies*

$$\mathbf{E}(z_{tj}^{(m)} - \widetilde{z}_{t,j}^{(m)} - u_{t,j})^2 \leq \frac{K_1}{m} \tag{25}$$

for some constant K_1.

Proof Let $\mathbf{u}_t^{(m)} = (u_{t,1}, \ldots, u_{t.m})$ and $\boldsymbol{\varphi}^{(m)} = (\varphi_1^{(m)}, \ldots, \varphi_m^{(m)})$; then

$$\mathbf{z}_t^{(m)} = \boldsymbol{\varphi}^{(m)} + f_t(\mathbf{d}^{(m)} + \mathbf{e}^{(m)}) + f_{t-1}\mathbf{c}_1^{(m)} + \ldots + f_{t-k}\mathbf{c}_k^{(m)} + \mathbf{u}_t^{(m)},$$

where $\mathbf{d}^{(m)}$ and $\mathbf{e}^{(m)}$ are as in A6. Let $\mathbf{a}^{(m)} = \mathbf{d}^{(m)}/||\mathbf{d}^{(m)}||^2$; then by A6,

$$||\mathbf{a}^{(m)}|| \leq 1/(m^{1/2}\delta^{1/2}). \tag{26}$$

Define $g_t^{(m)} = \mathbf{a}^{(m)\prime}\mathbf{z}_t^{(m)}$, and observe that

$$g_t^{(m)} = f_t + p^{(m)} + \boldsymbol{\eta}^{(m)},$$

where $\eta_t^{(m)} = \mathbf{a}^{(m)\prime}\mathbf{u}_t^{(m)}$ and $p^{(m)} = \mathbf{a}^{(m)\prime}\boldsymbol{\varphi}^{(m)}$. Then, by (26) and A1, we have

$$\mathbf{E}(\eta_t^{(m)2}) \leq D/m \tag{27}$$

with $D = A_1/\delta$. Let us reconstruct $\mathbf{z}_t^{(m)}$ using $g_t^{(m)}$ as follows:

$$\widetilde{\mathbf{z}}_t^{(m)} = (\boldsymbol{\varphi}^{(m)} - p^{(m)}(\mathbf{c}_0 + \ldots + \mathbf{c}_k)) + g_t^{(m)}\mathbf{c}_0^{(m)} + g_{t-1}^{(m)}\mathbf{c}_1^{(m)} + \ldots + g_{t-k}^{(m)}\mathbf{c}_k^{(m)}.$$

That is, if $\mathbf{B}^{(m)}$ is the $m \times (k+2)$ with columns $\boldsymbol{\varphi}^{(m)} - p^{(m)}(\mathbf{c}_0 + \ldots + \mathbf{c}_k)$ and $\mathbf{c}_i$, $0 \leq i \leq k$, we have

$$\widetilde{\mathbf{z}}_t^{(m)} = \mathbf{B}^{(m)}\mathbf{F}_t^{(m)}$$

and

$$\begin{aligned} z_{t,j}^{(m)} - \widetilde{z}_{t,j}^{(m)} &= -c_{j,0}\eta_t^{(m)} - c_1\eta_{t-1}^{(m)} - \ldots - c_{j,k}\eta_{t-k.}^{(m)} + u_{t,j} \\ &= v_{t,j}^{(m)} + u_{t,j}, \end{aligned}$$

where

$$v_{t,j}^{(m)} = -c_{j,0}\eta_t^{(m)} - c_1\eta_{t-1}^{(m)} - \ldots - c_{j,k}\eta_{t-k.}^{(m)}. \tag{28}$$

Then by (27) and A5, there exists K_1 such that for all $1 \leq j \leq m$

$$\mathbf{E}(v_{t,j}^{(m)2}) \leq \frac{K_1}{m},$$

and this proves the Lemma.

Lemma 2 *Suppose that $z_{t,j}$ satisfies A1–A6, and let $\mathbf{a}^{(m)}$, $\mathbf{B}^{(m)}$, and $\widetilde{\mathbf{z}}_t^{(m)}$ as in Lemma 1, and then there exists K_2 independent of m such that*

$$S_M(z_{tj}^{(m)} - \widetilde{z}_{t,j}^{(m)}) \leq S_M(u_{t,j}) + \frac{K_2}{m} \tag{29}$$

for some constant K_2.

It is enough to show that there exists m_0 such that for $m \geq m_0$ (29) holds. Let for $k > 0$ and v

$$L_{tj}(v, k) = \rho\left(\frac{v + u_{tj}}{s_j + k}\right),$$

where $s_j = S_M(u_{t,j})$.

Using the mean value theorem at (0, 0), we get

$$L_{tj}(v, k) = L_{tj}(0, 0) + v\frac{\psi\left(\frac{v^* + u_{tj}}{s_j + k^*}\right)}{s_j + k^*} - k\frac{\psi\left(\frac{v^* + u_{tj}}{s_j + k^*}\right)\frac{v^* + u_{tj}}{s_j + k^*}}{s_j + k^*}, \tag{30}$$

where $|v^*| \leq |v|$ and $0 \leq k^* \leq k$. Put $k = K_2/m^{1/2}$. Since $z_{t,j}^{(m)} - \widetilde{z}_{t,j}^{(m)} = v_{t,j}^{(m)} + u_{t,j}$, where $v_{t,j}^{(m)}$ is defined in Eq. (28) of Lemma 1, then using (30), we can write

$$\begin{aligned}\mathbf{E}\left(\rho\left(\frac{z_{t,j}^{(m)} - \widetilde{z}_{t,j}^{(m)}}{s_j + K_2/m^{1/2}}\right)\right) &= \mathbf{E}_{tj}(L(v_{t,j}^{(m)}, K_2/m^{1/2})) \\ &= b + \frac{\mathbf{E}\left(v_{t,j}^{(m)}\psi\left(\frac{v^*+u_{t.j}}{s_j+K^*/m^{1/2}}\right)\right)}{s_j + K^*/m^{1/2}} \\ &\quad - \frac{K_2}{m^{1/2}}\frac{\mathbf{E}\left(\psi\left(\frac{v^*+u_{t,j}}{s_j+K^*/m^{1/2}}\right)\frac{v^*+u_{tj}}{s_j+K^*/m^{1/2}}\right)}{s_j + K^*/m^{1/2}},\end{aligned} \tag{31}$$

where $|v^*| \leq |v_{t,j}|$ and $K^* \leq K_2$.

Put

$$K_2 = 2A_3 K_1^{!/2}/A_4. \tag{32}$$

Since by Lemma 1 $\mathbf{E}(v_{t,j}^{(m)2}) \leq K_1/m$ and by A3 $A_3 = \max \psi(u)$, we have

$$\begin{aligned}\frac{\left|\mathbf{E}\left(v_{t,j}^{(m)}\psi\left(\frac{v^*+u_{t,j}}{s_j+K^*/m^{1/2}}\right)\right)\right|}{s_j + K^*/m^{1/2}} &\leq \frac{\mathbf{E}(v_{t,j}^{(m)2})^{1/2}\max_u \psi(u)\ )}{s_j + K^*/m^{1/2}} \\ &\leq \frac{A_3 K_1^{1/2}/m^{1/2}}{s_j + K^*/m^{1/2}}.\end{aligned} \tag{33}$$

Take

$$K_2 > A_3 K_1^{1/2}/A_4;\ , \tag{34}$$

then since $K^* \leq K_2$ and $|v^*| \leq |v_{t,j}|$ by Lemma 1 and A4, there exists m_0 such that for $m \geq m_0$

$$\begin{aligned}&\frac{K_2}{m^{1/2}s_j + K^*/m^{1/2}}\mathbf{E}\left(\psi\left(\frac{v^* + u_{t,j}}{s_j + K^*/m^{1/2}}\right)\frac{v^* + u_{tj}}{s_j + K^*/m^{1/2}}\right) \\ &\geq \frac{K_2 A_4/2}{m^{1/2}s_j + K^*/m^{1/2}}.\end{aligned} \tag{35}$$

Then by (31), (33), and (35), we have

$$\left(\rho\left(\frac{z_{t,j}^{(m)} - \widetilde{z}_{t,j}^{(m)}}{s_j + K_2/m^{1/2}}\right)\right) < b \qquad ,$$

and therefore, $S_M(z_{t,j}^{(m)} - \widetilde{z}_{t,j}^{(m)}) \leq s_j + K_2/m^{1/2}$. This proves the Lemma.

Proof of Theorem 1

From Lemma 2, it can be derived that if $z_{t,j}^{*(m)}$ is the reconstruction with the optimal first S-ODPC with any k_1 and $k_2 = k$, we should have

$$\begin{aligned}
\frac{1}{m^{1/2}} \sum_{j=1}^{m} S_M^2(z_{t,j}^{(m)} - z_{t,j}^{*(m)}) &\leq \frac{1}{m^{1/2}} \sum_{j=1}^{m} S_M^2(z_{t,j}^{(m)} - \widetilde{z}_{t,j}^{(m)}) \\
&\leq \sum_{j=1}^{m} \left(s_j^2 + \frac{K_2^2}{m} + 2\frac{K_2}{m^{1/2}} s_j \right) \\
&\leq \frac{1}{m^{1/2}} \sum_{j=1}^{m} \left(s_j^2 + \frac{K}{m^{1/2}} \right) \\
&= \frac{1}{m^{1/2}} \sum_{j=1}^{m} s_j^2 + K,
\end{aligned}$$

where $K = \max(K_2^2, K_3)$ and $K_3 = 2K_2 \max_j s_{j.}$. This proves Theorem 1.

References

Alonso, A. M., Galeano, P., & Peña, D. (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics, 216*(1), 35–52.

Alonso, A. M. & Peña, D. (2019). Clustering time series by linear dependency. *Statistics and Computing, 29*(4), 655–676.

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica, 70*(1), 191–221.

Brillinger, D. R. (1964). The generalization of the techniques of factor analysis, canonical correlation and principal components to stationary time series. In *Invited Paper at the Royal Statistical Society Conference in Cardiff, Wales*.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

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

Fan, J., Wang, K., Zhong, Y., & Zhu, Z. (2021). Robust high dimensional factor models with applications to statistical machine learning. *Statistical Science, 36*(2), 303.

Fan, J., Wang, W., & Zhong, Y. (2019). Robust covariance estimation for approximate factor models. *Journal of Econometrics, 208*(1), 5–22.

Forni, M., Hallin, M., Lippi, M., & Zaffaroni, P. (2015). Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. *Journal of Econometrics, 185*(2), 359–371.

Forni, M., Hallin, M., Lippi, M., & Zaffaroni, P. (2017). Dynamic factor models with infinite-dimensional factor space: Asymptotic analysis. *Journal of Econometrics, 199*(1), 74–92.

Huber, P. J. (1964). *Robust statistics*. New York: Wiley.

Maronna, R. (2005). Principal Components and Orthogonal Regression Based on Robust Scales. *Technometrics, 47*(3), 264–273.

Maronna, R., Martin, D., Yohai, V., & Salibian-Barrera, M. (2019). *Robust Statistics*. New York: Wiley.

Muler, N., Peña, D., & Yohai, V. J. (2009). Robust estimation for ARMA models. *The Annals of Statistics, 37*(2), 816–840.

Muler, N. & Yohai, V. J. (2013). Robust estimation for vector autoregressive models. *Computational Statistics & Data Analysis, 65*, 68–79.

Peña, D., Smucler, E., & Yohai, V. J. (2019). Forecasting Multiple Time Series with One-Sided Dynamic Principal Components. *Journal of the American Statistical Association, 114*(528), 1683–1694.

Peña, D. & Yohai, V. J. (2016). Generalized dynamic principal components. *Journal of the American Statistical Association, 111*(515), 1121–1131.

Rousseeuw, P. J. & Yohai, V. J. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle, & D. Martin (Eds.), *Robust and Nonlinear Time Series*. Lecture Notes in Statistics (vol. 26, pp. 256–272). New York: Springer.

Salibian-Barrera, M. & Yohai, V. J. (2006). A Fast Algorithm for S-Regression Estimates. *Journal of Computational and Graphical Statistics, 15*(2), 414–427.

Trucíos, C., Mazzeu, J. H., Hotta, L. K., Pereira, P. L. V., & Hallin, M. (2021). Robustness and the general dynamic factor model with infinite-dimensional space: identification, estimation, and forecasting. *International Journal of Forecasting, 37*(4), 1520–1534.

Robust and Sparse Estimation of Graphical Models Based on Multivariate Winsorization

Ginette Lafit, Javier Nogales, Marcelo Ruiz, and Ruben Zamar

Abstract We propose the use of a robust covariance estimator based on multivariate Winsorization in the context of the Tarr–Müller–Weber framework for sparse estimation of the precision matrix of a Gaussian graphical model. Likewise Croux–Öllerer's precision matrix estimator, our proposed estimator attains the maximum finite-sample breakdown point of 0.5 under cellwise contamination. We conduct an extensive Monte Carlo simulation study to assess the performance of ours and the currently existing proposals. We find that ours has a competitive behavior, regarding the estimation of the precision matrix and the recovery of the graph. We demonstrate the usefulness of the proposed methodology in a real application to breast cancer data.

Keywords Gaussian graphical model · Precision matrix · Sparse robust estimation · Cellwise contamination · Winsorization

G. Lafit
University of Leuven, Leuven, Belgium
e-mail: ginette.lafit@kuleuven.be

J. Nogales
Universidad Carlos III de Madrid, Getafe, Spain
e-mail: fcojavier.nogales@uc3m.es

M. Ruiz (✉)
Universidad Nacional de Río Cuarto, Río Cuarto, Argentina
e-mail: mruiz@exa.unrc.edu.ar

R. Zamar
University of British Columbia, Vancouver, BC, Canada
e-mail: ruben@stat.ubc.ca

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_12

1 Introduction

Let $\boldsymbol{X} = (X_1, \ldots, X_p)'$ be a p-variate random vector with Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We assume that $\boldsymbol{\Sigma}$ is positive definite ($\boldsymbol{\Sigma} \succ 0$), and its inverse, the precision matrix, will be denoted by $\boldsymbol{\Omega} = (\omega_{ij})_{i,j=1\ldots,p}$; furthermore, we assume that $\boldsymbol{\mu} = \boldsymbol{0}$. Abbreviated, the model is

$$\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{1}$$

Given $V = \{1, \ldots, p\}$, let $V^2 = V \times V$ and $V^2_{-d} = \{(i, l) \in V^2 : i \neq l\}$. For a given pair $(i, l) \in V^2_{-d}$, let $V \backslash \{i, l\} = \{j \in V : i \neq j \neq l\}$ and $X_{V\backslash\{i,l\}} = \{X_j : j \in V\backslash\{i, l\}\}$.

For a random vector $\boldsymbol{X}$ satisfying (1), a *Gaussian graphical model* (GGM) is the undirected graph (V, E), where V is the set of nodes and E is the set of edges, which is defined by

$$(i, l) \in E \text{ if and only if } \operatorname{corr}\big(X_i, X_l | X_{V\backslash\{i,l\}}\big) \neq 0, \tag{2}$$

where $\operatorname{corr}\big(X_i, X_l | X_{V\backslash\{i,l\}}\big)$ is the conditional correlation coefficient of X_i and X_l given $X_{V\backslash\{i,l\}}$.

So, the set of edges E can be expressed as

$$E = \Big\{(i, l) \in V^2_{-d} : \ \operatorname{corr}\big(X_i, X_l | X_{V\backslash\{i,l\}}\big) \neq 0\Big\}. \tag{3}$$

It is well known that there exists a characterization of the conditional correlation in terms of the elements of the precision matrix. More specifically,

$$\forall (i, l) \in V^2_{-d} : \ \operatorname{corr}\big(X_i, X_l | X_{V\backslash\{i,l\}}\big) = -\frac{\omega_{il}}{\sqrt{\omega_{ii}\omega_{ll}}}. \tag{4}$$

Hence, we have the following parametrization for E:

$$E = \{(i, l) \in V^2_{-d} : \omega_{i,l} \neq 0\}. \tag{5}$$

Given a sample of $\boldsymbol{X}$, the goal of covariance selection is to estimate the conditional dependence structure by determining the set of nonzero entries of the precision matrix $\boldsymbol{\Omega}$ (see Dempster 1972; Edwards 2000; Lauritzen 1996). Generally, in high-dimensional statistics, it is assumed that there are just a few entries of $\boldsymbol{\Omega}$ that are different from zero, that is, that $\boldsymbol{\Omega}$ is sparse.

Until a few decades ago, statistical procedures assumed that datasets included many observations of a few and carefully chosen variables. Nowadays, datasets may contain a large number of variables relative to the sample size, bringing along blessings but also curses of dimensionality (Donoho 2000, 2017). Therefore, in a high-dimensional setting, the estimation of precision matrices faces significant challenges.

Let $\mathbb{X} = \left(\boldsymbol{x}_1', \ldots \boldsymbol{x}_n'\right)'$ be a $n \times p$ data matrix where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is a sample. If $n > p$, then the sample covariance matrix $\mathbf{S} = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$ is well conditioned and a well-known optimal estimate of $\boldsymbol{\Sigma}$. Moreover, $\mathbf{S}^{-1}$ can be used to define an unbiased estimator of $\boldsymbol{\Omega}$, provided $n > p + 2$. See Anderson (2003, pp. 272–274) or Muirhead (2005, p 137). On the other hand, when $p > n$, the sample covariance matrix is not invertible.

To deal with this problem, several covariance selection procedures based on regularization have been developed under the assumption that $\boldsymbol{\Omega}$ is sparse. See Friedman et al. (2008); Yuan and Lin (2007) and Banerjee et al. (2008).

For instance, if $\widehat{\boldsymbol{\Sigma}}$ is an estimator of $\boldsymbol{\Sigma}$, the graphical lasso (Glasso) proposed by Friedman et al. (2008) is defined by

$$\widehat{\boldsymbol{\Omega}} = \operatorname{argmin}_{\{\mathbf{U}:\mathbf{U}'=\mathbf{U},\mathbf{U}\succ 0\}}\left\{\operatorname{tr}(\mathbf{U}\widehat{\boldsymbol{\Sigma}}) - \operatorname{logdet}(\mathbf{U}) + \lambda \parallel \mathbf{U} \parallel_1\right\}, \tag{6}$$

where the optimization is over the set of symmetric positive-definite matrices,

$$\parallel \mathbf{U} \parallel_1 =: \sum_{i,j} |u_{ij}| \quad \text{for } i, j = 1, \ldots, p \tag{7}$$

is the ℓ_1 norm of the matrix $\mathbf{U} = (u_{ij})_{i,j=1\ldots,p}$, and $\lambda \geq 0$ is a regularization or penalty parameter usually determined by cross-validation. Following Friedman et al. (2008), we penalize all the elements of $\boldsymbol{\Omega}$ including the diagonal terms. Note that the larger the value of λ is, the more sparse the precision matrix estimate becomes.

For $\lambda = 0$, if $\widehat{\boldsymbol{\Sigma}} = \mathbf{S} \succ 0$, then the solution of (6) is the classical maximum likelihood estimate of $\boldsymbol{\Omega}$. On the other hand, (Banerjee et al. 2008) proved that, for any symmetric and positive semidefinite matrix $\widehat{\boldsymbol{\Sigma}}$ and $\lambda > 0$, Eq. (6) has a strictly positive-definite solution $\widehat{\boldsymbol{\Omega}}$ even if $p > n$.

In contrast to univariate datasets, in multivariate settings, outliers can appear in complex ways. In this regard, two types of contamination mechanisms have been introduced in the robustness literature: the Tukey–Huber contamination model (THCM) and the independent contamination model (ICM). In the THCM, it is assumed that a relative small proportion ϵ ($\epsilon < 1/2$) of the rows in the data table are contaminated. In the ICM, introduced by Alqallaf et al. (2009), each cell of the data matrix has a probability to be independently contaminated. This second mechanism is a better fit for the high-dimensional setting where the variables are likely to be obtained from different sources and measured separately (Agostinelli et al. 2015).

The vast majority of the work in the area of robust statistics has concentrated on the estimation of the covariance matrix under these two types of contamination models. Robust conditional correlation coefficient estimation has been studied when p is small. Rao and Sievers (1995) introduced a measure that uses residuals based on rank estimates of regression parameters when $p = 3$. Only recently a few papers have focused on estimation of the precision matrix in the context of ICM.

Tarr et al. (2016) and Öllerer and Croux (2015) showed that Glasso is not robust in the presence of cellwise outliers. Therefore, in order to obtain a robust estimate of the precision matrix, they proposed a plug-in approach, using a robust covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}$ in Eq. (6). There are several robust estimators of $\boldsymbol{\Sigma}$, but, unfortunately, their computation is very time-consuming and may not be possibly well defined when the dimension p is high (Khan et al. 2007). To overcome this problem, resistant pairwise procedures can be used to avoid sensitivity to two-dimensional outliers, like in Tarr et al. (2016) and Öllerer and Croux (2015) proposals. Tarr et al. (2016) proposed to use pairwise robust covariances estimates, whereas (Öllerer and Croux 2015) used pairwise robust correlation estimates.

Huber (2011) proposed a robust estimator of the correlation coefficient by using one-dimensional Winsorization. Alqallaf et al. (2002) proposed the use of Huberized pairwise correlation coefficients based on one-dimensional Winsorization. A limitation of this approach is that the pairwise Huberized estimates and covariance estimates do not take into account the orientation of the (pairwise) bivariate data. To overcome this limitation, Khan et al. (2007) developed an adjusted bivariate Winsorization estimation, obtaining a robust estimator of the correlation matrix under cellwise contamination. Here, we use this estimator to introduce a new robust graphical lasso procedure, RGlassoWinsor. We compare the performance of our method with other existing approaches under cellwise and casewise contamination.

Section 2 discusses the main differences between the THCM and ICM. Section 3 introduces our proposal. Section 4 presents the results of an extensive simulation experiment comparing the currently existing estimators of the robust precision matrix with our new robust graphical lasso procedure. Section 5 contains an application to breast cancer data. Section 6 concludes with some remarks.

2 Outliers in High-Dimensional Data

In this section, we briefly outline the main differences between THCM and ICM.

Consider a set of n independent observations of the multivariate Gaussian vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ satisfying (1), let $\epsilon \in (0, 1)$ be the fraction of contamination, and define the random vector

$$\boldsymbol{B} = (B_1, \ldots, B_p)' \text{ with } B_j \sim \text{Bernoulli}(\epsilon),\ j = 1, \ldots, p. \tag{8}$$

Suppose that instead of $\boldsymbol{X}$, we observe

$$\boldsymbol{Y} = (\mathbf{I} - \mathbf{D})\boldsymbol{X} + \mathbf{D}\boldsymbol{Z}, \tag{9}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix, $\boldsymbol{Z}$ is a p-variate random vector with an arbitrary and unspecified outlier generating distribution, and $\mathbf{D}$ is a diagonal matrix with diagonal elements $B_1, \ldots, B_p$. Moreover, we assume that $\boldsymbol{X}$, $\boldsymbol{B}$ and $\boldsymbol{Z}$ are independent.

The classical THCM assumes that the random vector $\boldsymbol{B} = (B_1, \ldots, B_p)'$ satisfies $P(B_1 = B_2 = \ldots = B_p) = 1$. So, we see either a perfect realization of the random vector $\boldsymbol{X}$, with probability $1-\epsilon$, or a realization of the random vector $\boldsymbol{Z}$, with probability ϵ.

Motivated by the THCM, robust procedures identify and downweight possibly contaminated cases. However, in a high-dimensional setting, this strategy is inconvenient for two reasons. The most obvious is that in high dimension, when n is relatively small compared with p, discarding a single observation may result in a substantial loss of information. A perhaps less obvious reason was highlighted by Alqallaf et al. (2009), where they argued that there are situations where the contaminating mechanism may be independent for different variables. Consequently, they proposed the ICM that assumes that $B_1, \ldots, B_p$ are independent random variables and satisfy

$$P(B_1 = 1) = \ldots = P(B_p = 1) = \epsilon. \tag{10}$$

Hence, a case is uncontaminated, $\boldsymbol{Y} = \boldsymbol{X}$, with probability $P(\boldsymbol{B} = \boldsymbol{0}) = (1-\epsilon)^p$, which quickly decreases below $1/2$ as p increases. Equivalently, the probability that at least one component of $\boldsymbol{Y}$ is contaminated is $1-(1-\epsilon)^p$. For example, if $p = 60$ and $\epsilon = 0.05$, this probability equals to 0.95. If $p \geq 200$ (not an uncommon case these days), this probability becomes nearly 1.

The indicator matrix $\mathbf{D}$, whose diagonal is a sequence of Bernoulli random variables, determines the structure of the contamination model. Figure 1 shows a representation of a sample of size $n = 100$ of $\boldsymbol{B}$ with dimension $p = 60$, contamination fraction $\epsilon = 0.10$, under both contamination models: THCM in

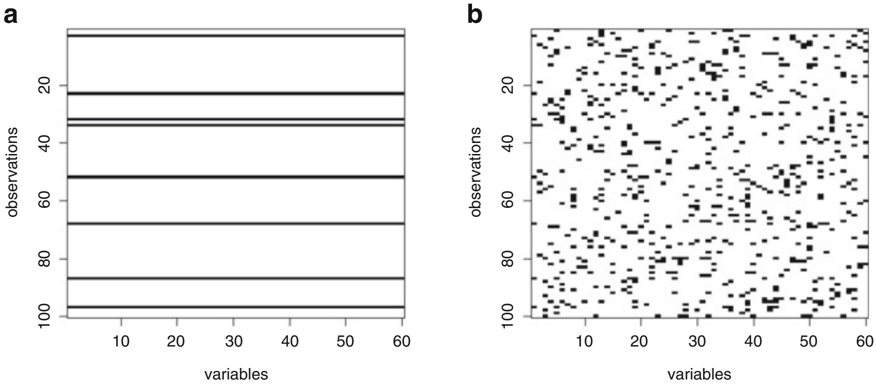


Fig. 1 Panels (**a**) and (**b**) represent the data matrix of dimension 100 × 60 corresponding to the random vector $\boldsymbol{B}$ of dimension 60 given in (8) generated under THCM and ICM, respectively. Uncontaminated cells are in color white, and contaminated cells are in color black

panel (a) and ICM in panel (b). On each panel, uncontaminated cells are in color white and contaminated cells are in color black. For THCM, the actual proportion of contaminated cells is 0.08, coinciding with the percentage of contaminated observations (rows). But, for ICM, the proportion of contaminated cells is 0.10, but all the observations have at least one contaminated cell ($\approx 1 - (0.9)^{60}$); hence, the totality of the cases or rows is contaminated. This phenomenon is called "propagation of outliers" in Alqallaf et al. (2009).

THCM is also called casewise contamination model, where a minority of observations or cases (rows) of the data matrix contains outliers and the size of this minority does not depend on the number p of variables. ICM is also denominated cellwise contamination model because the contamination is produced randomly affecting the cells of the data table.

The classical robustness theory based on the affine equivariant Tukey–Huber contamination model relays and enforces the concept of equivariance. Alqallaf et al. (2009) showed that under the cellwise contamination model, standard high-breakdown affine equivariant estimators propagate outliers, and this causes their very poor performance when p is large. The reason is that affine equivariant robust estimators depend on linear combinations of the observations that have a very high probability of being contaminated under ICM for moderate and large p. Notice that under ICM, the majority of cases will have at least some contaminated component. Agostinelli et al. (2015) addressed the problem of robust estimation of location and scatter under the two contamination models.

3 Robust Lasso for Precision Matrices

In this section, we introduce the robust covariance estimator based on bivariate Winsorization, and we define the robust precision matrix estimator of a Gaussian graphical model based on (Tarr et al. 2016) framework.

3.1 Plug-in Strategy

Hereafter, $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^{'}, i = 1, \ldots, n$, denotes a sample of observations of a p-multivariate random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ satisfying (9), and let $\mathbb{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n) \in \mathbb{R}^{n \times p}$ be the corresponding data table. Let $\mathbf{R}$ denote the correlation matrix; i.e., if $\boldsymbol{\Sigma} = (\Sigma_{ij})$, then $\mathbf{R} = (R_{ij})$ with $R_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$.

Following Tarr et al. (2016) and using (6), we will construct a robust estimation procedure of the precision matrix as follows:

$$\widehat{\boldsymbol{\Omega}} = \operatorname{argmin}_{\{\mathbf{U}:\mathbf{U}'=\mathbf{U},\mathbf{U}\succ 0\}} \operatorname{tr}(\mathbf{U}\widehat{\boldsymbol{\Sigma}}) - \operatorname{logdet}(\mathbf{U}) + \lambda \parallel \mathbf{U} \parallel_1, \tag{11}$$

where $\widehat{\boldsymbol{\Sigma}}$ is a robust estimator of the covariance matrix.

3.2 Adjusted Multivariate Winsorization

To control the effect of bivariate outliers on the pairwise estimation of $\mathbf{\Sigma}$, we apply the procedure proposed by Khan (2006). In this procedure, the robust estimator $\widehat{\mathbf{R}}^W$ of the correlation matrix $\mathbf{R}$ is defined in two steps by first computing the pairwise correlation matrix, $\widehat{\mathbf{R}}^0$, using an adjusted Winsorization scheme, which takes into consideration the orientation of bidimensional data. Later, based on $\widehat{\mathbf{R}}^0$, a robust estimator of the covariance matrix $\mathbf{\Sigma}$ is defined.

The two steps to compute $\widehat{\mathbf{R}}^W$ are given below:

(1) Initial estimate $\widehat{\mathbf{R}}^0$.

Given $j, k \in \{1, \ldots, p\}$, with $j \neq k$, let consider the bivariate sample $\{(y_{ij}, y_{ik})^{'}, i = 1, \ldots n\}$ and compute for every $l = j, k$

$$m_l = \text{median}(y_{1l}, \ldots, y_{nl}), s_l = \text{mad}(y_{1l}, \ldots, y_{nl}), \tag{12}$$

where "mad" denotes the median absolute deviation. Define now the standardized samples

$$\tilde{y}_{il} = \frac{y_{il} - m_l}{s_l}, \; i = 1, \ldots, n, \tag{13}$$

for every $l = j, k$.

As Khan et al. (2007) noted, one-dimensional Winsorization does not account for the orientation of the bidimensional data and does not address the effect of bivariate outliers. Therefore, they propose a bivariate adjusted Winsorization that uses two tuning constants denoted c_1 and c_2. The constant c_1 is used on the two quadrants that contain the majority of the standardized data, and the constant c_2, smaller than c_1, is used on the other two quadrants. Typically, $c_1 = 2$ or 2.5 and $c_2 = \sqrt{h}c_1$ with $h = n_2/n_1$, where n_1 is the number of observations in the two major quadrants and $n_2 = n - n_1$.

The bivariate Winsorized data $(v_{ij}, v_{ik})^{'}$, $i = 1, \ldots, n$, are computed as follows. If $(\tilde{y}_{ij}, \tilde{y}_{ik})$ lies in one of the major (more populated) quadrants, let

$$v_{il} = \psi_{c_1}(\tilde{y}_{il}), \; i = 1, \ldots, n; \; l = j, k, \tag{14}$$

where ψ_{c_1} is the Huber function $\psi_c(x) = \min\{\max\{-c, x\}, c\}$ with tuning constant $c = c_1$. On the other hand, if $(\tilde{y}_{ij}, \tilde{y}_{ik})$ lies in one of the minor (less populated) quadrants, then

$$v_{il} = \psi_{c_2}(\tilde{y}_{il}), \; i = 1, \ldots, n; \; l = j, k. \tag{15}$$

The elements $\widehat{R}^0_{jk}$ of the matrix $\widehat{\mathbf{R}}^0$ are now defined as follows. For $j = k$, we set $\widehat{R}^0_{jj} = 1$, and for $j \neq k$, we set

$$\widehat{R}^0_{jk} = \text{corr}(\mathbf{v}_j, \mathbf{v}_k),$$

where $\mathbf{v}_j = (v_{1j}, \dots, v_{nj})'$ and $\mathbf{v}_k = (v_{1k}, \dots, v_{nj})'$.

(2) Final estimate $\widehat{\mathbf{R}}^W$.
As before, consider $\{(y_{ij}, y_{ik})', \ i = 1, \dots n\}$ a bivariate sample of the two variables Y_j and Y_k, with $j \neq k$ (columns j and k of the data table). Let

$$A_{jk} = \begin{pmatrix} 1 & \widehat{R}^0_{jk} \\ \widehat{R}^0_{kj} & 1 \end{pmatrix}$$

be the 2×2 submatrix of $\widehat{\mathbf{R}}^0$. Perform now, for every $l = j, k$, the following bivariate transformation:

$$u_{il} = y_{il} \min\Big(\sqrt{c/D_{jk}(y_{ij}, y_{ik})}, 1\Big), i = 1, \dots, n; l = j, k, \tag{16}$$

where D_{jk} is the Mahalanobis distance based on the correlation matrix A_{jk} and evaluated in (y_{ij}, y_{ik}). The tuning constant $c = 5.99$ corresponds to the 95% quantile of a χ^2_2 distribution. By this transformation, the outliers are shrunken to the border of an ellipse, including the majority of the data.
We now define the Winsorized correlation estimate $\widehat{\mathbf{R}}^W = (\widehat{R}^W_{jk})$ as follows. For $j \neq k$, we set

$$\widehat{R}^W_{jk} = \text{corr}(\mathbf{u}_j, \mathbf{u}_k),$$

where $\mathbf{u}_j = (u_{1j}, \dots, u_{nj})'$ and $\mathbf{u}_k = (u_{1k}, \dots, u_{nj})'$, and for $j = k$, we set $\widehat{R}^W_{jj} = 1$.

Finally, based on $\widehat{\mathbf{R}}^W$, a robust estimator of $\mathbf{\Sigma}$ is defined as

$$\widehat{\mathbf{\Sigma}}^W = \text{diag}(s_1, \dots, s_p)\widehat{\mathbf{R}}^W \text{diag}(s_1, \dots, s_p), \tag{17}$$

where s_j is the robust estimator of the dispersion introduced in (12). In order to guarantee positive definiteness of $\widehat{\mathbf{\Sigma}}^W$, we compute the nearest positive-definite matrix (Higham 2002). Finally, the robust Glasso estimator of the precision matrix based on bivariate adjusted Winsorization, called and denoted by $\widehat{\mathbf{\Omega}}^W$, is defined by (11) with $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^W$.

Remark 1 By Theorem 19.1 and Proposition 19.1 in Öllerer and Croux (2015), the finite-sample breakdown point under ICM of $\widehat{\mathbf{\Omega}}^W$ satisfies

$$\epsilon_n\left(\widehat{\mathbf{\Omega}}^W\right) \geq \epsilon_n^+(\widehat{\mathbf{\Sigma}}^W) \geq \max_{j=1,\ldots,p} \epsilon_n^+(s_j) = 1/2,$$

where $\epsilon_n^+(\widehat{\mathbf{\Sigma}}^W)$ is the explosion finite-sample breakdown point (EBP) under ICM contamination of $\widehat{\mathbf{\Sigma}}^W$ and $\epsilon_n^+(s_j)$ is the EBP of the univariate scale estimator scale s_j, $j = 1, \ldots, p$.

4 Simulation Experiment and Numerical Results

We conducted a Monte Carlo simulation experiment to investigate the performance of RGlassoWinsor compared with other procedures.

4.1 Simulation Settings

In the following, we describe the precision matrix models, the contamination scenarios, and the precision matrix estimation procedures considered in our simulation study.

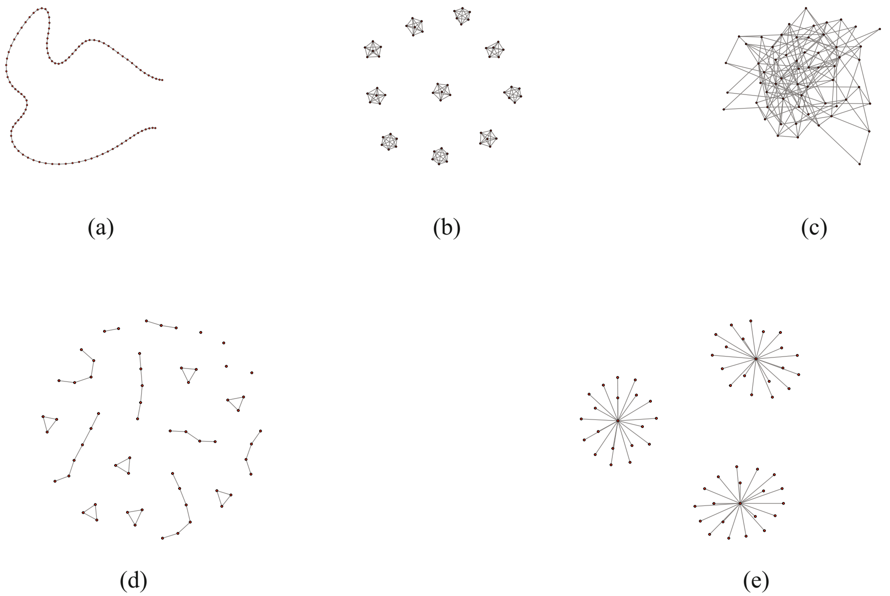
Precision Matrix Models

We consider two-dimension values ($p = 60, 200$) and five $\mathbf{\Omega}$ models.

Model 1. Autoregressive model of order 1, denoted AR(1). In this case, we set $\Sigma_{ij} = 0.4^{|i-j|}$ for $i, j = 1, \ldots p$ and $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$.

Model 2. Block-diagonal matrix model, denoted BG. In this case, the precision matrix $\mathbf{\Omega}$ has q blocks of size p/q. Each block has diagonal elements equal to 1 and off-diagonal elements equal to 0.5. For $p = 60, 200$, we use $q = 10$ and 40 blocks, respectively.

Model 3. Random model, denoted Rand. Jiang et al. (2021), in the R package huge, compute the $\mathbf{\Omega}$ matrix of this model as follows. First they consider $\mathbf{\Theta} = (\theta_{ij})$ an adjacency matrix of dimension p such that every diagonal entry $\theta_{ii} = 0$, and each pair of off-diagonal elements is randomly set $\theta_{ij} = \theta_{ji} = 1$ with probability prob $= 3/p$ (the default value) and defined as 0 otherwise. Then they define the set of edges of the graph, establishing that two different nodes, i and j, are connected if and only if $\theta_{ij} = 1$. Finally, given Θ, is possibly to choose real constants v and s such that $\mathbf{\Omega} = v\mathbf{\Theta} + (|e| + 0.1 + s)I_p$ is positive definite, with I_p the identity matrix, e the smallest eigenvalue of $v\mathbf{\Theta}$, and v and s set to the default values 0.3 and 0.1, respectively.



(a) (b) (c)

(d) (e)

Fig. 2 Graphs of AR(1), BG, Rand, NN(2), and Hub, graphical models for $p = 60$ nodes. (**a**) Model AR(1). (**b**) Model BG. (**c**) Model Rand. (**d**) Model NN(2). (**e**) Model Hub

Model 4. Nearest neighbors model of order 2, denoted NN(2). For each node, we randomly select two neighbors and choose a pair of symmetric entries of $\mathbf{\Omega}$ using the "NeighborOmega" function of the R package `Tlasso` (Sun et al. 2016).

Model 5. Hub model, denoted Hub. As in Model 3, consider $\mathbf{\Theta} = (\theta_{ij})$ an adjacency matrix defined as follows. The row/columns are evenly partitioned into 3 (10) disjoint groups if $p = 60$ (if $p = 200$). Each group is associated with a "center" row i in that group. Each pair of off-diagonal elements, $i \neq j$, is set $\theta_{ij} = \theta_{ij} = 1$ if j also belongs to the same group as i and 0 otherwise. It results in 57 (190) edges in E if $p = 60$ (if $p = 200$). The precision matrix $\mathbf{\Omega}$ is defined as in Rand model and computed using the same R package `huge`.

Figure 2 displays graphs from Models 1–5 with $p = 60$.

Contamination Scenarios

As in (9), let $\boldsymbol{Y} = (\mathbf{I} - \mathbf{B})\boldsymbol{X} + \mathbf{B}\boldsymbol{Z}$, and consider the following scenarios:

(i) *Clean data.* $\boldsymbol{Y} = \boldsymbol{X} \sim \text{N}(\mathbf{0}, \mathbf{\Sigma})$ corresponding to $\epsilon = 0$.

(ii) *Cellwise or ICM.* Here $\boldsymbol{Z} \sim \text{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{\Sigma})$, where $\boldsymbol{\mu}_1 = (10, \dots, 10)'$, $\sigma = 0.2$, and contamination fractions ϵ : 0.01, 0.05, 0.10.

(iii) *Casewise or THCM*. Let $\boldsymbol{Z} = \boldsymbol{z}$ with $\boldsymbol{z} = k\boldsymbol{v}$, $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}$ satisfying $\boldsymbol{v}'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{v} = 1$ and $k = 100$. We consider ϵ : 0.05 and 0.10.

For every p, ϵ, and $\boldsymbol{\Omega}$ model, we generate $N = 100$ random samples $\mathbb{Y}_1, \ldots, \mathbb{Y}_N$, of size $n = 100$ of $\boldsymbol{Y}$.

Precision Matrix Estimators

We will compare the performance of the following estimators of $\boldsymbol{\Omega}$:

1. The classical Glasso estimator defined by (6).
2. TWM estimators. Tarr et al. (2016) estimate a robust initial covariance matrix based on the approach proposed by Gnanadesikan and Kettenring (1972). Noting that the covariance of two random variables X and Y can be written as

$$\text{cov}(X, Y) = \frac{1}{4\alpha\beta}[\text{Var}(\alpha X + \beta Y) - \text{Var}(\alpha X - Y)], \tag{18}$$

where $\alpha = 1/\sqrt{\text{Var}(X)}$ and $\beta = 1/\sqrt{\text{Var}(Y))}$, a robust estimate of the bivariate covariance $\widehat{\Sigma}_{lj}$ can be obtained by replacing Var in (18) with a robust variance estimator like Q_n or τ-scale estimators defined by Maronna and Zamar (2002) and Rousseeuw and Croux (1993). Based on these robust estimators of the covariance matrix, using (11), Tarr et al. (2016) derived a robust estimator of $\boldsymbol{\Omega}$, denoted by RGlassoQn and RGlassotau. We use the R package `robustbase` to compute the robust variance estimators Q_n and τ-scale (Maechler et al. 2022).
3. OC estimators. Öllerer and Croux (2015) proposed a robust estimator $\widehat{\Sigma}_{lj}^R$ of the bivariate correlations

$$\widehat{\Sigma}_{lj}^R = \text{scale}(\mathbf{y}_l)\text{scale}(\mathbf{y}_j)r(\mathbf{y}_l, \mathbf{y}_k), \tag{19}$$

where $r(\cdot)$ and scale$(\cdot)$ are robust correlation and scale estimators, respectively. For instance, scale$(\cdot)$ is Q_n (or the mad), and for $r(\cdot)$, there are different possibilities, such as Gaussian rank correlation, Spearman correlation, and Quadrant correlation. This proposal leads, using (11), to three robust estimators called RGlassoGauss, RGlassoSpearman, and RGlassoQuadrant. As in Öllerer and Croux (2015), and based on Croux and Dehon (2010), to obtain Fisher consistency at the bivariate normal distribution, Quadrant and Spearman correlations need to be transformed.
4. Our proposal, RGlassoWinsor estimator. To compute the robust bivariate adjusted correlation estimator defined in steps 1 and 2 of Sect. 3, we use the function "corhuber" of the R package `robustHD` (Alfons 2021).

In proposals (1) and (3), to make the pairwise correlation matrices positive definite, we compute the nearest positive-definite matrix using the function "nearPD" of the R package `Matrix` (Bates and Maechler 2019). To solve the regularized equation (11), we use the R package `huge`. There are different alternatives to select the optimal regularization parameter, and we use 5-fold cross-validation as it is indicated by Jiang et al. (2021) and Öllerer and Croux (2015).

Estimation Performance Evaluation

We wish to evaluate two different features of the procedures: (i) their performance as estimates of $\boldsymbol{\Omega}$; and (ii) how well they recover the true graphical model graph.

The *numerical performance* of $\widehat{\boldsymbol{\Omega}}$ is measured by the mean squared error (MSE) defined by the Frobenius norm of the difference between $\boldsymbol{\Omega}$ and the predicted precision matrix $\widehat{\boldsymbol{\Omega}}$

$$m_F = ||\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}||_F = \sqrt{\sum_{ij} |\omega_{ij} - \hat{\omega}_{ij}|^2}$$

and also quantified by the Kullback–Leibler divergence

$$D_{KL} = \frac{1}{2}\Big(\mathrm{tr}\Big\{\widehat{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1}\Big\} - \log\Big\{\det\Big[\widehat{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1}\Big]\Big\} - p\Big).$$

To evaluate the *graph recovery or classification performance*, we compute the true positive and true negative rates—also called sensitivity and specificity, respectively—defined by

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\#E} \text{ and } \mathrm{TNR} = \frac{\mathrm{TN}}{\#NE},$$

where $E = \big\{(i, j) \in V_{-d}^2 : \omega_{ij} \neq 0\big\}$ is the set of edges, $NE = \big\{(i, j) \in V_{-d}^2 : \omega_{ij} = 0\big\}$ is the set of non-connected nodes, and

$$\mathrm{TP} = \#\Big\{(i, j) \in V_{-d}^2 : \hat{\omega}_{ij} \neq 0 \wedge \omega_{ij} \neq 0\Big\},$$

$$\mathrm{TN} = \#\Big\{(i, j) \in V_{-d}^2 : \hat{\omega}_{ij} = 0 \wedge \omega_{ij} = 0\Big\}$$

denote the sizes of the sets of true positives and true negatives, respectively.

A related measure is the Matthews correlation coefficient (MCC) given by

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}},$$

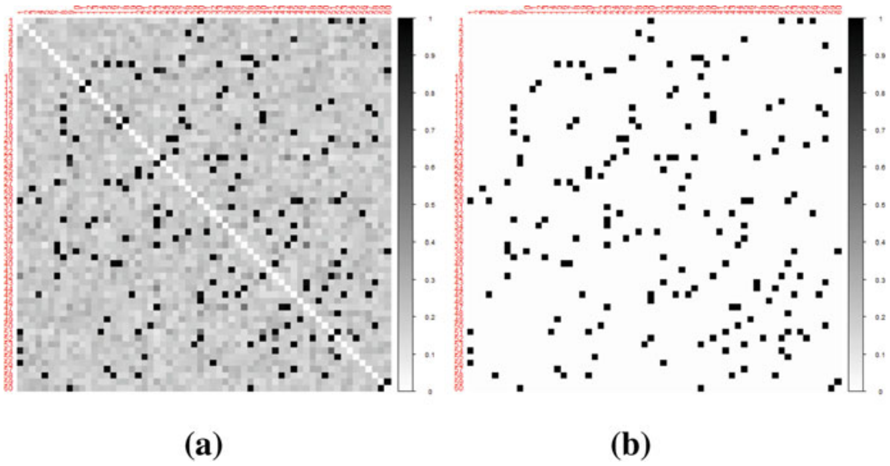


Fig. 3 Heatmap for the frequency of adjacency for each pair of nodes with $p = 60$ and true graph of model Rand. The axes display the graph p-nodes in a given order. (**a**) Estimated model Rand. (**b**) True graph of model Rand

where

$$\text{FP} = \#\Big\{(i, j) \in V_{-d}^2 : \hat{\omega}_{ij} \neq 0 \wedge \omega_{ij} = 0\Big\},$$

$$\text{FN} = \#\Big\{(i, j) \in V_{-d}^2 : \hat{\omega}_{ij} = 0 \wedge \omega_{ij} \neq 0\Big\}$$

denote the number of false positives and false negatives sets, respectively.

Note that larger values of TPR, TNR, and MCC indicate better performances (Baldi et al. 2000; Fan et al. 2009).

Heatmaps are useful to visualize the graph recovery performance of a given procedure. As an example, for $p = 60$, the axes in the panels of Fig. 3 display the graph nodes in a given order. Panel (a) shows $N = 100$ estimated Rand models by Glasso from simulation replicates where each cell displays a gray level proportional to how frequently the corresponding pair of nodes appears in the estimated graph in the N simulation replicates. So, a white color in a given cell (i, j) means that nodes i and j are never adjacent in the simulated graphs, and a pair of nodes that are always adjacent in the simulated graphs is represented by a black colored cell. The heatmap of Panel (a) is compared with the figure of Panel (b) that represents the graph of true model Rand where a black or white cell corresponds to a pair of connected or non-connected nodes, respectively.

Finally, Fig. 4 represents the five true model graphs.

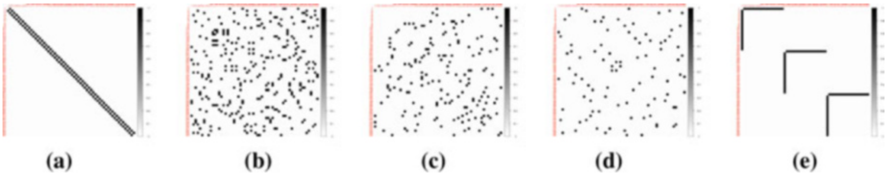


Fig. 4 True model graphs with $p = 60$. The axes display the graph p-nodes in a given order. (**a**) AR(1). (**b**) BG. (**c**) Rand. (**d**) NN(2). (**e**) Hub

4.2 *Estimation and Graph Recovery Performances*

In this section, we analyze the numerical and graph recovery performances of the estimation of the different GGMs, represented by its precision matrix $\boldsymbol{\Omega}$, for clean data and under both contamination scenarios.

To fix some ideas, we first focus on the estimation results for the AR(1) model. Tables 1, 2, 3, 4, 5, and 6 show the estimation performance under ICM. Lafit et al. (2022) report the results under THCM (see Tables 40 to 45, Appendix B).

In terms of numerical performance, Glasso is slightly better than other methods for clean data, but it is clearly non-robust under both contamination models for all positive contamination fractions. In both contamination models, our proposal, RGlassoWinsor, has the best numerical performance. Note that the mean squared error, m_F, and the Kullback–Leibler divergence, D_{KL}, grow when the dimension

Table 1 Model AR(1) under ICM. Comparison of means and standard deviations (in brackets) of m_F and D_{KL} over $N = 100$ replicates. $p = 60$, $n = 100$

	$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.05$		$\epsilon = 0.10$	
	D_{KL}	m_F	D_{KL}	m_F	D_{KL}	m_F	D_{KL}	m_F
RGlassoWinsor	5.223	4.365	5.739	4.689	8.336	5.798	13.089	7.053
	(0.039)	(0.024)	(0.046)	(0.024)	(0.069)	(0.024)	(0.102)	(0.022)
Glasso	4.232	4.063	30.007	8.960	76.465	10.828	103.705	11.241
	(0.028)	(0.021)	(0.339)	(0.030)	(0.240)	(0.005)	(0.187)	(0.002)
RGlassoQ_n	8.118	5.830	10.314	6.477	29.604	9.100	57.220	10.406
	(0.080)	(0.027)	(0.131)	(0.034)	(0.450)	(0.040)	(0.428)	(0.013)
RGlassoTau	5.687	4.737	7.071	5.373	24.044	8.548	71.010	10.742
	(0.044)	(0.023)	(0.070)	(0.030)	(0.501)	(0.054)	(0.593)	(0.013)
RGlassoGauss	4.595	4.278	5.732	4.854	10.540	6.516	16.375	7.697
	(0.033)	(0.021)	(0.048)	(0.025)	(0.080)	(0.022)	(0.095)	(0.016)
RGlassoSpearman	4.968	4.478	5.889	4.936	10.303	6.455	16.274	7.670
	(0.042)	(0.025)	(0.049)	(0.025)	(0.076)	(0.021)	(0.096)	(0.016)
RGlassoQuad	10.545	6.560	11.682	6.843	16.151	7.693	22.521	8.515
	(0.073)	(0.020)	(0.093)	(0.023)	(0.109)	(0.019)	(0.130)	(0.015)

Table 2 Model AR(1) under ICM. Comparison of means and standard deviations (in brackets) of m_F and D_{KL} over $N = 100$ replicates. $p = 200, n = 100$

	$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.05$		$\epsilon = 0.10$	
	D_{KL}	m_F	D_{KL}	m_F	D_{KL}	m_F	D_{KL}	m_F
RGlassoWinsor	23.481	9.541	25.191	9.986	33.635	11.564	49.845	13.611
	(0.125)	(0.038)	(0.126)	(0.035)	(0.193)	(0.034)	(0.233)	(0.025)
Glasso	19.469	8.784	94.502	16.112	257.189	19.867	350.501	20.628
	(0.085)	(0.044)	(0.467)	(0.025)	(0.576)	(0.007)	(0.409)	(0.003)
RGlassoQ_n	63.930	14.998	78.345	15.856	149.160	18.274	255.526	19.914
	(0.335)	(0.024)	(0.465)	(0.025)	(0.746)	(0.017)	(1.105)	(0.011)
RGlassoTau	29.859	11.082	38.890	12.440	135.343	17.916	306.799	20.354
	(0.169)	(0.031)	(0.250)	(0.034)	(0.834)	(0.022)	(1.245)	(0.009)
RGlassoGauss	21.102	9.295	25.163	10.216	41.158	12.711	60.306	14.603
	(0.107)	(0.039)	(0.106)	(0.028)	(0.180)	(0.026)	(0.217)	(0.019)
RGlassoSpearman	23.131	9.794	26.254	10.438	41.564	12.756	61.643	14.686
	(0.116)	(0.035)	(0.110)	(0.029)	(0.189)	(0.027)	(0.222)	(0.019)
RGlassoQuad	48.458	13.612	53.087	14.044	70.442	15.322	91.330	16.416
	(0.181)	(0.021)	(0.214)	(0.022)	(0.254)	(0.019)	(0.336)	(0.017)

Table 3 Model AR(1) under ICM. Comparison of means and standard deviations (in brackets) of TPR and TNR over $N = 100$ replicates. $p = 60, n = 100$

	$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.05$		$\epsilon = 0.10$	
	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
RGlassoWinsor	0.991	0.842	0.989	0.853	0.962	0.887	0.799	0.934
	(0.001)	(0.003)	(0.001)	(0.003)	(0.003)	(0.003)	(0.009)	(0.003)
Glasso	0.997	0.816	0.140	0.986	0.033	0.985	0.045	0.968
	(0.001)	(0.003)	(0.015)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)
RGlassoQ_n	0.865	0.952	0.786	0.968	0.131	0.998	0.003	1.000
	(0.006)	(0.002)	(0.011)	(0.002)	(0.018)	(0.000)	(0.001)	(0.000)
RGlassoTau	0.960	0.875	0.928	0.902	0.397	0.988	0.012	1.000
	(0.003)	(0.002)	(0.004)	(0.002)	(0.025)	(0.001)	(0.001)	(0.000)
RGlassoGauss	0.996	0.834	0.987	0.835	0.888	0.882	0.655	0.924
	(0.001)	(0.003)	(0.002)	(0.003)	(0.005)	(0.003)	(0.008)	(0.003)
RGlassoSpearman	0.990	0.850	0.983	0.851	0.914	0.890	0.718	0.927
	(0.001)	(0.003)	(0.002)	(0.003)	(0.004)	(0.002)	(0.008)	(0.002)
RGlassoQuad	0.729	0.934	0.688	0.941	0.553	0.957	0.339	0.978
	(0.008)	(0.002)	(0.010)	(0.002)	(0.012)	(0.002)	(0.012)	(0.001)

p increases, for both clean and contaminated data. D_{KL} and m_F are higher for cellwise contamination model than the casewise contamination model.

Even when there is no contamination and considering the MCC as graph recovery measure, the performance of Glasso is poor. Under cellwise contamination model, MCC means produced by RGlassoWinsor and those produced by OC estimators

Table 4 Model AR(1) under ICM. Comparison of means and standard deviations (in brackets) of MCC over $N = 100$ replicates. $p = 60$, $n = 100$

	$\epsilon = 0$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.10$
RGlassoWinsor	0.389	0.402	0.444	0.471
	(0.004)	(0.004)	(0.005)	(0.005)
Glasso	0.360	0.147	0.026	0.013
	(0.003)	(0.013)	(0.003)	(0.003)
RGlassoQ_n	0.567	0.594	0.261	0.023
	(0.005)	(0.005)	(0.017)	(0.005)
RGlassoTau	0.420	0.455	0.441	0.056
	(0.004)	(0.004)	(0.014)	(0.006
RGlassoGauss	0.380	0.378	0.399	0.360
	(0.003)	(0.003)	(0.004)	(0.004)
RGlassoSpearman	0.398	0.397	0.424	0.401
	(0.004)	(0.004)	(0.003)	(0.004)
RGlassoQuad	0.428	0.425	0.391	0.331
	(0.005)	(0.004)	(0.005)	(0.006)

Table 5 Model AR(1) under ICM. Comparison of means and standard deviations (in brackets) of TPR and TNR over $N = 100$ replicates. $p = 200$, $n = 100$

	$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.05$		$\epsilon = 0.10$	
	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
RGlassoWinsor	0.971	0.932	0.961	0.941	0.904	0.958	0.620	0.982
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.001)	(0.009)	(0.001)
Glasso	0.986	0.914	0.207	0.984	0.024	0.989	0.028	0.983
	(0.001)	(0.002)	(0.010)	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
RGlassoQ_n	0.041	1.000	0.021	1.000	0.001	1.000	0.000	1.000
	(0.003)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
RGlassoTau	0.867	0.963	0.796	0.976	0.017	1.000	0.002	1.000
	(0.004)	(0.001)	(0.006)	(0.001)	(0.002)	(0.000)	(0.000)	(0.000)
RGlassoGauss	0.978	0.931	0.959	0.937	0.794	0.952	0.487	0.973
	(0.001)	(0.002)	(0.002)	(0.001)	(0.005)	(0.001)	(0.008)	(0.001)
RGlassoSpearman	0.967	0.935	0.957	0.936	0.842	0.952	0.565	0.973
	(0.002)	(0.002)	(0.002)	(0.001)	(0.004)	(0.001)	(0.008)	(0.001)
RGlassoQuad	0.648	0.965	0.607	0.970	0.455	0.979	0.304	0.985
	(0.005)	(0.001)	(0.006)	(0.001)	(0.009)	(0.001)	(0.008)	(0.001)

remain almost constant and even slightly increase when the contamination fractions increase as it is shown in Tables 4 and 6. Conversely, MCC means of the TWM estimators dramatically decrease when the contamination fraction ϵ increases. A better explanation can be found by looking at Tables 3 and 5, while the mean of TPR remains relatively high for RGlassoWinsor and for the estimators of OC, and the mean of TPR goes to zero for the estimators of TWM. Note that, although not

Table 6 Model AR(1) under ICM. Comparison of means and standard deviations (in brackets) of MCC over $N = 100$ replicates. $p = 200$, $n = 100$

	$\epsilon = 0$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.10$
RGlassoWinsor	0.342	0.365	0.405	0.405
	(0.004)	(0.004)	(0.005)	(0.003)
Glasso	0.312	0.137	0.013	0.008
	(0.004)	(0.005)	(0.001)	(0.001)
RGlassoQ_n	0.185	0.121	0.016	0.001
	(0.008)	(0.008)	(0.003)	(0.001)
RGlassoTau	0.404	0.451	0.108	0.017
	(0.004)	(0.004)	(0.006)	(0.003)
RGlassoGauss	0.345	0.350	0.332	0.269
	(0.004)	(0.003)	(0.003)	(0.002)
RGlassoSpearman	0.348	0.347	0.352	0.307
	(0.003)	(0.003)	(0.004)	(0.002)
RGlassoQuad	0.311	0.317	0.284	0.227
	(0.002)	(0.003)	(0.003)	(0.003)

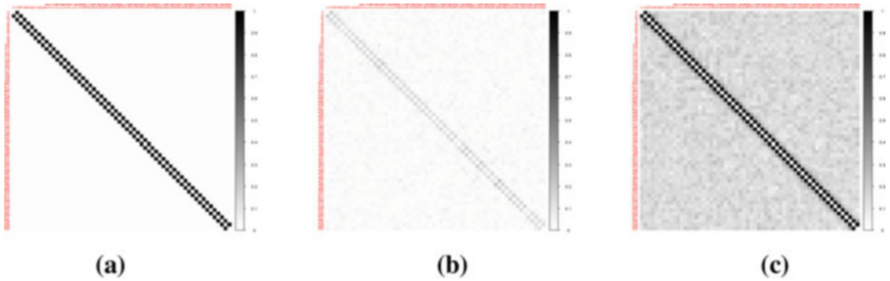


Fig. 5 Graph of true model AR(1) and heatmaps for the frequency of adjacency for each pair of nodes over $N = 100$ replicates. $p = 60$ and $n = 100$. ICM with $\epsilon = 0.01$. The axes display the graph p-nodes in a given order. (**a**) True AR(1). (**b**) Glasso. (**c**) RGlassoWinsor

so extreme, a similar phenomenon occurs under THCM, as it is shown in Lafit et al. (2022) (Tables 44 and 45).

Figures 5, 6, and 7 show the performance of Glasso and RGlassoWinsor for contaminated data under ICM. Notice that for contaminated data, Glasso cannot recover the true set of edges, introducing a large number of false negatives. Although RGlassoWinsor introduces false positives, it better recovers the true set of edges.

In the following paragraphs, we set general conclusions about the behavior of the estimators for all analyzed $\mathbf{\Omega}$ models, based on Tables 1, 2, 3, 4, 5, and 6 of this section, Tables 16 to 19 of Appendix A (ICM), and Tables 40 to 69 of Appendix B (THCM) reported in Lafit et al. (2022). Tables 7, 8, 9, 10, 11, 12, 13, and 14 below report the average ranks for all the compared estimation methods, evaluated across all the considered precision matrix models. Rank 1 and 7 correspond to the

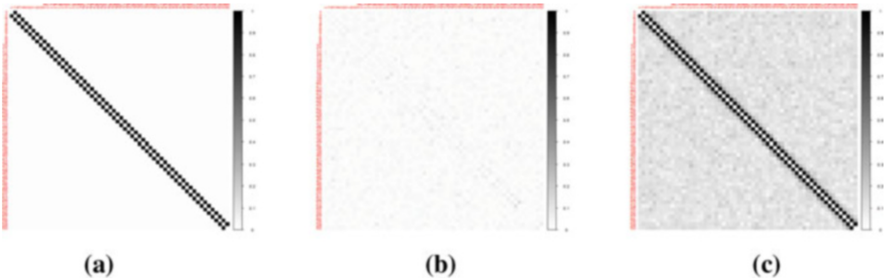


Fig. 6 Graph of true model AR(1) and heatmaps for the frequency of adjacency for each pair of nodes over $N = 100$ replicates. $p = 60$ and $n = 100$. ICM with $\epsilon = 0.05$. The axes display the graph p-nodes in a given order. (**a**) True AR(1). (**b**) Glasso. (**c**) RGlassoWinsor

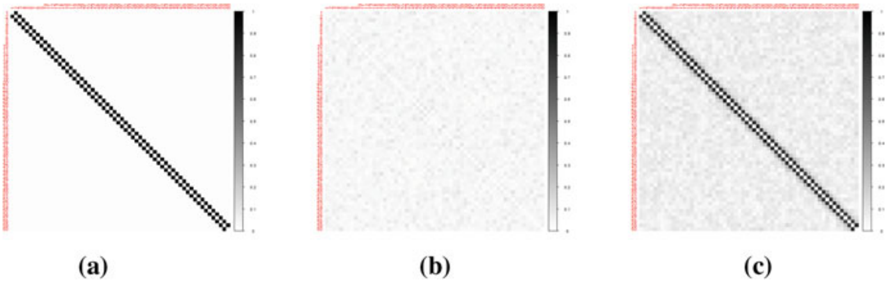


Fig. 7 Graph of true model AR(1) and heatmaps for the frequency of adjacency for each pair of nodes over $N = 100$ replicates. $p = 60$ and $n = 100$. ICM with $\epsilon = 0.10$. The axes display the graph p-nodes in a given order. (**a**) True AR(1). (**b**) Glasso. (**c**) RGlassoWinsor

best and worst performing methods, respectively. The average ranks of the best two performing methods are shown in boldface.

For $\epsilon = 0$, Glasso performs slightly better than the other estimators and shows a non-robust performance, being the worst ranked for contaminated data. For almost all contamination scenarios, RGlassoWinsor is the best ranked, and the estimators of TWM, specially RGlassoSpearman and RGlassoGaus, have the closest rankings. Note that for dimension $p = 200$, RGlassoSpearman has a slightly better average rank than RGlassoWinsor, under THCM.

4.3 Timing Comparisons

In this section, we report timing comparisons of the estimation methods. Tables 15 and 16 give the average running time and standard error (in bracket) in seconds for every method based on $R = 100$ replications. In this simulation, we used the AR(1) model with $p = 60, 200$. The times were obtained using the R function `Sys.time`.

Note that RGlassoWinsor and RGlassoTau are computationally intensive.

Table 7 Average rank of the estimation methods based on m_F and D_{KL} under ICM. $p = 60$, $n = 100$

	$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.05$		$\epsilon = 0.10$	
	m_F	D_{KL}	m_F	D_{KL}	m_F	D_{KL}	m_F	D_{KL}
RGlassoWinsor	3.4	4	**1.6**	**2**	**1**	**1**	**1**	**1**
Glasso	**1**	**1**	6.4	6.4	7	7	6.6	6.6
RGlassoQ_n	6.4	6.4	5.8	6	6	6	5.2	5.2
RGlassoTau	5.4	5.4	4.2	4.8	5	5	6.2	6.2
RGlassoGauss	**2.2**	**2**	**2.2**	**1.2**	**2.4**	**2.6**	2.6	**2.4**
RGlassoSpearman	3.4	3	2.6	2.8	2.6	2.4	**2.4**	2.6
RGlassoQuad	6.2	6.2	5.2	5	4	4	4	4

Table 8 Average rank of the estimation methods based on MCC under ICM. $p = 60$, $n = 100$

	$\epsilon = 0$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.10$
RGlassoWinsor	**3.2**	**2.2**	**1.2**	**1.4**
Glasso	4.4	6.4	7	6.8
RGlassoQ_n	**3.4**	3.4	5.8	5.4
RGlassoTau	3.8	**3.2**	4.6	6.2
RGlassoGauss	3.8	4.2	3.5	3
RGlassoSpearman	4.2	3.6	**2**	**1.6**
RGlassoQuad	5.6	5	4.2	4

Table 9 Average rank of the estimation methods based on m_F and D_{KL} under ICM. $p = 200$, $n = 100$

	$\epsilon = 0$		$\epsilon = 0.01$		$\epsilon = 0.05$		$\epsilon = 0.10$	
	m_F	D_{KL}	m_F	D_{KL}	m_F	D_{KL}	m_F	D_{KL}
RGlassoWinsor	3.6	3.4	**1.6**	**1.8**	**1**	**1**	**1.2**	**1**
Glasso	**1**	**1.4**	6.2	6.2	6.6	6.6	6.6	6.8
RGlassoQ_n	6.8	6.8	6	6.2	5.6	5.8	5.4	5
RGlassoTau	5.6	5.2	5	5	5.4	5.2	6	6.2
RGlassoGauss	**2**	**2**	**1.6**	**1.2**	**2**	**2**	**2**	**2.2**
RGlassoSpearman	3.4	3.4	2.8	3	3	3	2.8	2.8
RGlassoQuad	5.6	5.8	4.8	4.6	4.4	4.4	4	4

Table 10 Average rank of the estimation methods based on MCC under ICM. $p = 200$, $n = 100$

	$\epsilon = 0$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.10$
RGlassoWinsor	3.2	**1.4**	**1**	**1.6**
Glasso	3.6	6.2	6.6	6.4
RGlassoQ_n	4.2	5.4	4.4	4.6
RGlassoTau	**3**	2.6	4.2	6
RGlassoGauss	**2.4**	**3**	3.2	3
RGlassoSpearman	3.4	3.4	**2.2**	**1.8**
RGlassoQuad	6.2	5.2	4.6	4.4

Table 11 Average rank of the estimation methods based on m_F and D_{KL} under THCM. $p = 60, n = 100$

	$\epsilon = 0.05$		$\epsilon = 0.10$	
	m_F	D_{KL}	m_F	D_{KL}
RGlassoWinsor	**1.2**	**1.8**	**1.4**	**2**
Glasso	5.2	7	7	7
RGlassoQ_n	6	5.8	5.8	5.8
RGlassoTau	4.6	4.4	4.6	4.4
RGlassoGauss	**2.2**	2.4	**2.2**	2.4
RGlassoSpearman	2.6	**1.8**	2.4	**1.8**
RGlassoQuad	4.8	4.8	4.6	4.6

Table 12 Average rank of the estimation methods based on MCC under THCM. $p = 60, n = 100$

	$\epsilon = 0.05$	$\epsilon = 0.10$
RGlassoWinsor	**2.6**	**2**
Glasso	6.8	7
RGlassoQ_n	3.6	4.6
RGlassoTau	3	3.4
RGlassoGauss	4.6	4.8
RGlassoSpearman	**2.6**	**2.6**
RGlassoQuad	3.8	3.6

Table 13 Average rank of the estimation methods based on m_F and D_{KL} under THCM. $p = 200, n = 100$

	$\epsilon = 0.05$		$\epsilon = 0.10$	
	m_F	D_{KL}	m_F	D_{KL}
RGlassoWinsor	**2**	**1**	**1.4**	**2**
Glasso	4.8	7	5.2	6.2
RGlassoQ_n	7	6	6.6	6.4
RGlassoTau	4.8	5	4.8	5.2
RGlassoGauss	**1**	**2**	**1.4**	**1.4**
RGlassoSpearman	3	3	2.8	2.4
RGlassoQuad	5.2	4	4.6	4.2

Table 14 Average rank of the estimation methods based on MCC under THCM. $p = 200, n = 100$

	$\epsilon = 0.05$	$\epsilon = 0.10$
RGlassoWinsor	**2.4**	**2.2**
Glasso	5.4	5.6
RGlassoQ_n	6.4	5.6
RGlassoTau	3	3.8
RGlassoGauss	3.2	2.6
RGlassoSpearman	**2**	**2.2**
RGlassoQuad	5	5.2

Table 15 Average running time and estimated standard error (in bracket) in seconds for each method based on $R = 100$ replications. AR(1) model with $p = 60$ under ICM

	$\epsilon = 0$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.10$
RGlassoWinsor	17.111	16.373	23.178	19.875
	(0.051)	(0.022)	(0.194)	(0.046)
Glasso	4.931	3.940	4.239	4.464
	(0.048)	(0.004)	(0.017)	(0.005)
RGlassoQ_n	14.430	12.312	13.653	14.708
	(0.058)	(0.011)	(0.065)	(0.019)
RGlassoTau	37.920	43.234	43.045	43.256
	(0.115)	(0.516)	(0.100)	(0.027)
RGlassoGauss	4.202	4.513	4.549	4.791
	(0.015)	(0.004)	(0.006)	(0.017)
RGlassoSpearman	4.279	4.573	4.623	5.091
	(0.015)	(0.006)	(0.004)	(0.015)
RGlassoQuad	4.315	4.672	4.709	4.862
	(0.018)	(0.007)	(0.005)	(0.005)

Table 16 Average running time and estimated standard error (in bracket) in seconds for each method based on $R = 100$ replications. AR(1) model with $p = 200$ under ICM

	$\epsilon = 0$	$\epsilon = 0.01$	$\epsilon = 0.05$	$\epsilon = 0.10$
RGlassoWinsor	180.510	161.564	161.452	162.035
	(1.585)	(0.062)	(0.075)	(0.159)
Glasso	5.631	5.315	5.378	5.300
	(0.010)	(0.023)	(0.018)	(0.035)
RGlassoQ_n	107.061	111.841	110.646	109.240
	(0.757)	(0.280)	(0.273)	(0.269)
RGlassoTau	403.468	419.055	416.680	417.736
	(1.254)	(5.350)	(5.331)	(5.241)
RGlassoGauss	6.015	6.407	6.268	6.231
	(0.011)	(0.017)	(0.021)	(0.013)
RGlassoSpearman	7.355	7.701	7.711	7.634
	(0.012)	(0.021)	(0.021)	(0.019)
RGlassoQuad	7.660	7.522	7.735	7.663
	(0.010)	(0.054)	(0.022)	(0.015)

5 Real Data Example

In preoperative chemotherapy, when all invasive cancer cells are eradicated, the patient is said to have reached the state of *pathological complete response*, abbreviated as pCR. This pCR is associated with the long-term cancer-free survival of a person. On the contrary, residual disease (RD) indicates that the disease has not

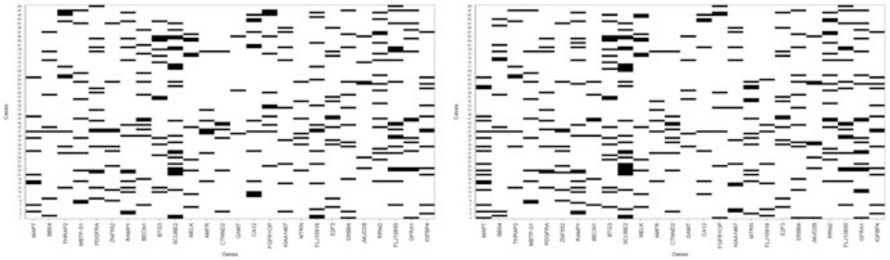


Fig. 8 Cellwise outliers detected by "cellHandler" for the RD class based on (**a**) $\widehat{\boldsymbol{\Sigma}}^W$ and (**b**) $\widehat{\boldsymbol{\Sigma}}^G$. A black colored cell indicates an outlier

been eradicated. Measurements of the expression level (activity) of genes may be able to predict if a patient can reach a pCR.

Hess et al. (2006) use normalized gene expression data of patients in stages I–III of breast cancer, to identify patients that may achieve pCR under preoperative chemotherapy. Their database has 22283 gene expression levels for 133 patients, with 34 pCR and 99 RD. Hess et al. (2006) and Natowicz et al. (2008) identify 26 important genes for predicting survival and response to adjuvant chemotherapy. Following Ambroise et al. (2009) and Tang et al. (2021), we estimate the precision matrix for the 26 key genes on the two classes pCR and RD.

Raymaekers and Rousseeuw (2022) proposed a method that detects cellwise outliers, implemented in the R package `cellWise` (Raymaekers and Rousseeuw 2021). The function "cellHandler" of `cellWise` flags cellwise outliers in the data matrix, based on robust estimates of the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ with 0.95% as cutoff used in the detection of cellwise outliers. We compare the performance of RGlassoWinsor and RGlassoGauss because both have shown similar rankings. Using the sample median and a robust estimates of $\boldsymbol{\Sigma}$ provided by Winsorization, $\widehat{\boldsymbol{\Sigma}}^W$, and Gaussian rank correlations, $\widehat{\boldsymbol{\Sigma}}^G$ (see (17) and (19)), we first detect outliers in the dataset.

Figure 8 illustrates cellwise outliers flagged by "cellHandler" based on both robust covariance estimates for the RD class. The rows represent the patients or cases, and the columns represent the variables or gene expressions. A black colored cell indicates that its value is an outlier.

Of the total of 2574 (99 × 26) cells of the data matrix of the RD class, 307 (12%) are contaminated according to "cellHandler" based on $\widehat{\boldsymbol{\Sigma}}^W$. The first five most contaminated variables correspond to genes SCUBE2, FLJ12650, RRM2, FLJ10916, and PDGFRA. Using $\widehat{\boldsymbol{\Sigma}}^G$, 325 (almost 13%) are flagged as contaminated, and the first six most contaminated variables correspond to genes SCUBE2, GFRA1, RRM2, BTG3, MAPT, and FLJ12650 (i.e., the last two genes have the same numbers of cellwise outliers)

A similar procedure shows that for the pCR group, using $\widehat{\boldsymbol{\Sigma}}^W$, of the total of 884 (34 × 26) cells of the data matrix, "cellHandler" flags 166 (19%) cells as contaminated, and the first five most contaminated variables correspond to genes

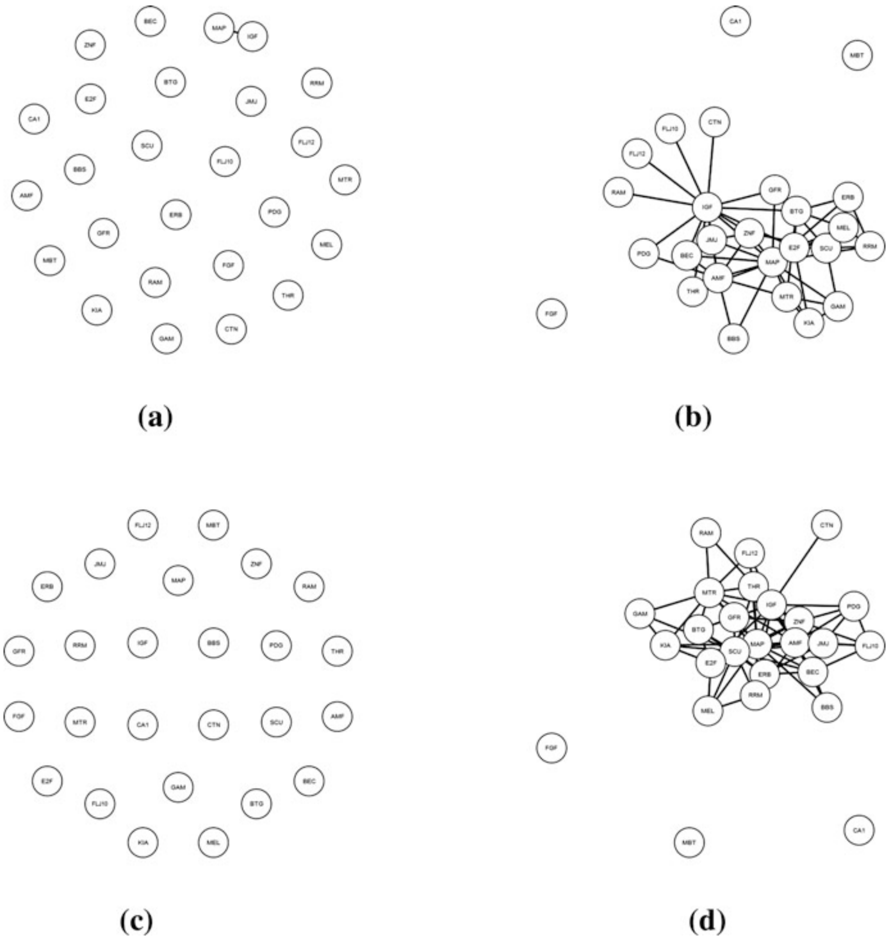


Fig. 9 Estimated graph of the GGM for the 26 genes corresponding to RD class. (**a**) Glasso. (**b**) RGlassoWinsor. (**c**) RGlassoTau. (**d**) RGlassoGauss

PDGFRA, CA12, SCUBE2, BBS4, and IGFBP4. Using $\widehat{\boldsymbol{\Sigma}}^G$, 108 (almost 12 %) cells are flagged as contaminated, and the first five most contaminated variables correspond to genes CA12, SCUBE2, IGFBP4, KIAA1467, and MTRN.

Figures 9 and 10 display the resulting network obtained using Glasso, RGlassoWinsor, RGlassoTau, and RGlassoGauss, the latter two representing TWM and OC of procedures. Table 17 exhibits the estimated network density for the 26 genes for each class, for all procedures, using a regularization parameter chosen by 5-fold cross-validation.

Excluding the estimated networks by RGlassoQ_n and RGlassoTau, the undirected graphs differ according to the class membership that may suggest that genes

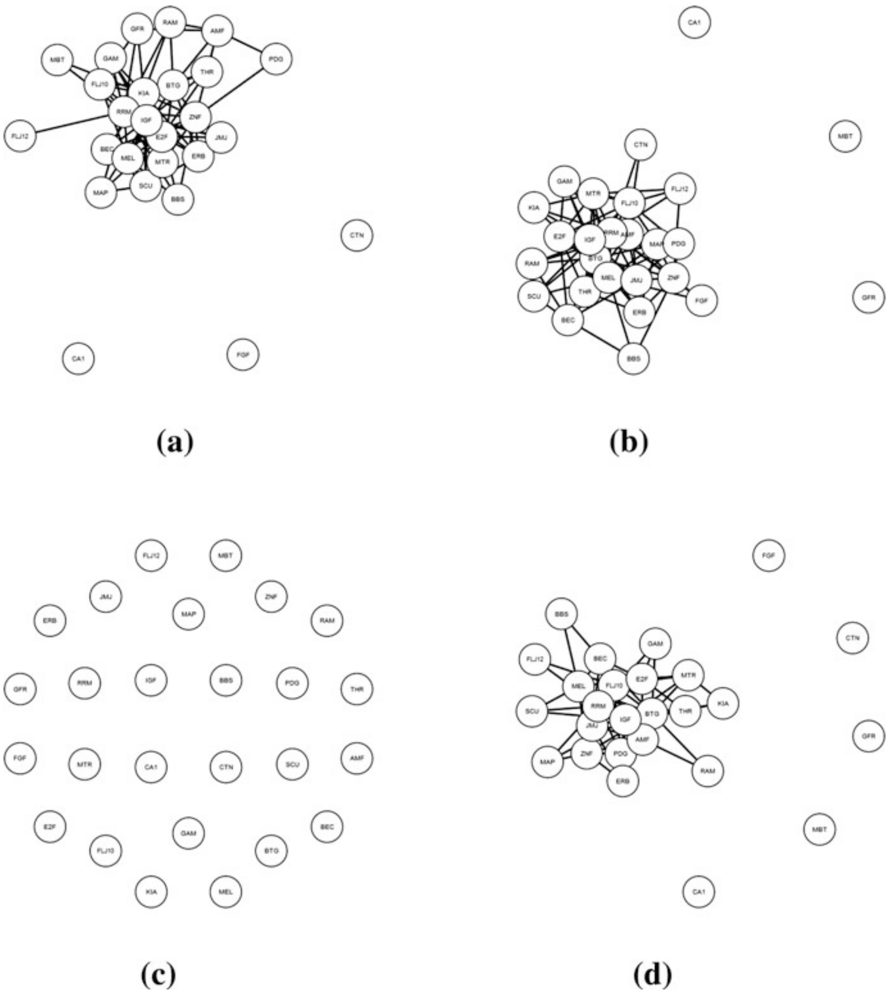


Fig. 10 Estimated graph of the GGM for the 26 genes corresponding to PCR class. (**a**) Glasso. (**b**) RGlassoWinsor. (**c**) RGlassoTau. (**d**) RGlassoGauss

regulation differs according the participants response to the treatment (Ambroise et al. 2009).

In the pCR class, RGlassoWinsor produces a less sparse network than Glasso and RGlassoSpearman, but a similar structure. But, in the RD class, while Glasso and RGlassoTau do not detect any conditional relationship between nodes (genes), RGlassoWinsor and the procedures of OC detect several edges between genes.

Table 17 Estimated network density for the 26 genes from breast cancer gene expressions data

	pCR class	RD class
RGlassoWinsor	0.280	0.169
Glasso	0.243	0.003
RGlassoQ_n	0.000	0.000
RGlassoTau	0.000	0.000
RGlassoGauss	0.203	0.249
RGlassoSpearman	0.197	0.237
RGlassoQuad	0.117	0.234

6 Concluding Remarks

This chapter introduces a new robust graphical lasso procedure called RGlassoWinsor based on adjusted bivariate Winsorization estimation of the covariance matrix for high-dimension covariance selection or precision matrix estimation.

RGlassoWinsor is compared with the currently existing robust estimators of the precision matrix, introduced by Tarr et al. (2016) and Öllerer and Croux (2015), by using different performance measures regarding graph recovery and sparse estimation of the precision matrix.

Our proposal shows a good performance for all the precision models, dimensions, and contamination scenarios considered in this research. For clean data, Glasso is slightly better than other methods, but it is clearly non-robust. Under contamination and for almost all performance measures, our proposal, RGlassoWinsor, has the best overall performance.

Moreover, our procedure attains the maximum finite-sample breakdown point of $1/2$ under cellwise contamination.

Finally, we demonstrate the usefulness of RGlassoWinsor in an application to the analysis of breast cancer data.

References

Agostinelli, C., Andy, L., Yohai, V., & Zamar, R. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test, 24*, 441–461.

Alfons, A. (2021). robustHD: An R package for robust regression with high-dimensional data. *Journal of Open Source Software, 6*(67), 3786.

Alqallaf, F., Aelst, S. V., Yohai, V. J., & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics, 37*(1), 311–331.

Alqallaf, F., Konis, K., Martin, D., & Zamar, R. H. (2002). Scalable robust covariance and correlation estimates for data mining. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta* (pp. 14–23).

Ambroise, C., Chiquet, J., and Matias, C. (2009). Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics, 3*, 205–238.

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics, 16*(5), 412–424.

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research, 9*, 485–516.

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-18.

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods and Applications, 19*, 497–515.

Dempster, A. P. (1972). Covariance selection. *Biometrics, 28*(1), 157–175. https://doi.org/10.2307/2528966

Donoho, D. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Lectures, 1*(2000), 32.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*, 745–763.

Edwards, D. (2000). *Introduction to Graphical Modelling*. New York: Springer.

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *The Annals of Applied Statistics, 3*(2), 521–541.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432–441.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics, 28*(1), 81–124. https://doi.org/10.2307/2528963

Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., et al. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology, 24*(26), 4236–4244.

Higham, N. (2002). Computing the nearest correlation matrix—a problem from finance. *Journal of Numerical Analysis, 22*, 329—343.

Huber, P. J. (2011). *Robust Statistics*. New York: Springer.

Jiang, H., Fei, X., Liu, H., Roeder, K., Lafferty, J., Wasserman, L., Li, X., & Zhao, T. (2021). *huge: High-Dimensional Undirected Graph Estimation*. R package version 1.3.5.

Khan, J. A., Van, S. Aelst, & Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association, 102*(480), 1289–1299.

Khan, M. J. A. (2006). *Robust Linear Model Selection for High-dimensional Datasets*. Ph. D. thesis, Canada: University of British Columbia.

Lafit, G., Nogales, F., Ruiz, M., & Zamar, R. (2022). Robust graphical lasso based on multivariate winsorization, pp. 1–33. *ArXiv* http://arxiv.org/abs/2201.03659.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Oxford University Press.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., & Anna di Palma, M. (2022). *robustbase: Basic Robust Statistics*. R package version 0.95-0.

Maronna, R., & Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Journal of Machine Learning Research, 44*, 307–317.

Muirhead, R. (2005). *Aspects of Multivariate Statistical*. New York: Wiley.

Natowicz, R., Incitti, R., Horta, E. G., Charles, B., Guinot, P., Yan, K., Coutant, C., Andre, F., Pusztai, L., & Rouzier, R. (2008). Prediction of the outcome of preoperative chemotherapy in breast cancer using DNA probes that provide information on both complete and incomplete responses. *BMC Bioinformatics, 9*(1), 1–17.

Öllerer, V., & Croux, C. (2015). Robust high-dimensional precision matrix estimation. In K. Nordhausen, & S. Taskinen (Eds.), *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja* (pp. 325–350). Cham: Springer.

Rao, S., & Sievers, G. (1995). A robust partial correlation measure. *Nonparametric Statistics, 5*, 1–20.

Raymaekers, J., & Rousseeuw, P. (2021). *cellWise: Analyzing Data with Cellwise Outliers*. R package version 2.2.5.

Raymaekers, J., & Rousseeuw, P. J. (2022). Handling cellwise outliers by sparse regression and robust covariance. *Journal of Data Science, Statistics, and Visualisation, 1*(3), 1–30.

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

Sun, W. W., Wang, Z., Lyu, X., Liu, H., & Cheng, G. (2016). *Tlasso: Non-convex Optimization and Statistical Inference for Sparse Tensor Graphical Models*. R package version 1.0.1.

Tang, P., Jiang, H., Kim, H., & Deng, X. (2021). Robust estimation of sparse precision matrix using adaptive weighted graphical lasso approach. *Journal of Nonparametric Statistics, 33*(2), 249–272.

Tarr, G., Müller, S., & Weber, N. C. (2016). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis, 93*, 404–420.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika, 94*(1), 19–35.

Robustly Fitting Gaussian Graphical Models—the R Package robFitConGraph

Daniel Vogel, Stuart J. Watt, and Anna Wiedemann

Abstract This chapter gives a tutorial-style introduction to the R package robFitConGraph, which provides a robust goodness-of-fit test for Gaussian graphical models. Its use is demonstrated at a data example on music performance anxiety, which also illustrates *why* one would want to fit a Gaussian graphical model—and why one should do so robustly. The underlying theory is briefly explained, much of which has been developed by David Tyler.

Keywords Covariance selection model · Deviance test · M-estimator · Music performance anxiety · Partial correlation

1 Introduction

The first two Sects. 1 *Introduction* and 2 *A Case Study* are intended for a general audience, assuming neither deeper familiarity with graphical models nor robustness. Section 3 *Background and Theory* discusses some aspects in detail.

D. Vogel (✉)
MEDICE Arzneimittel Pütter GmbH & Co. KG, Iserlohn, Germany

Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, UK
e-mail: daniel.vogel@tu-dortmund.de

S. J. Watt
Mirador Analytics, Melrose, UK

A. Wiedemann
Department of Psychiatry, University of Cambridge, Cambridge, UK
e-mail: aw778@medschl.cam.ac.uk

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_13

1.1 Gaussian Graphical Modeling

Graphical models are an important tool for analyzing the dependence structure of several variables. Gaussian graphical models are employed for continuously distributed data, where a multivariate normal, or Gaussian, distribution is adequate.

A graph encodes the dependence structure of a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ as follows: The nodes represent the individual variables, and an edge between two nodes represents partial correlation, or the absence of an edge zero partial correlation. Partial correlation is a measure for conditional dependence between two variables, conditional on all remaining variables, which, in a very loose sense, can be understood as a measure for dependence *not* explained by joint dependence on other variables. As soon as more than two variables are considered, conditional dependence is arguably more important than marginal dependence: It suggests and helps to verify causal relationships. While ice cream sales and criminal assault rates per day—at any given place in the temperate climate zone—are certainly positively correlated, their conditional independence given a suitable mediator variable, such as outside mean temperature, provides evidence for why this may be so.

Graphical modeling refers to the statistical task of finding an appropriate graph that describes the dependence structure of a given data set, i.e., identifying all zero partial correlations. The aim is to find a parsimonious graph, i.e., one with few edges, which does not contradict the data. A full graph, with all edges present, means no restriction and contains no structural information. A completely empty graph, with no edges at all, means all variables are independent.

Let Σ denote the covariance matrix of $\boldsymbol{X}$. Throughout, we assume that all second moments of $\boldsymbol{X}$ are finite and that furthermore Σ is positive definite and thus may be inverted to yield the *concentration matrix* or *precision matrix* $K = \Sigma^{-1}$. The assumption of strict positive definiteness is a mild one and is equivalent to the probability mass not being concentrated on a lower-dimensional affine linear subspace of $\mathbb{R}^p$. We make the same assumption on the data points $\mathbb{X}_n = (\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)^\top$, and hence, the sample covariance matrix $\hat{\Sigma}_n$ is positive definite, and the sample concentration matrix $\hat{K}_n = \hat{\Sigma}_n^{-1}$ exists.

The basis for Gaussian graphical modeling is the following characterization: The variables X_i and X_j are partially uncorrelated given all remaining $p-2$ variables if and only if $K_{i,j} = 0$. The *partial correlation* $p_{i,j}$ of X_i and X_j given the remaining components of $\boldsymbol{X}$ is defined as the correlation between the residuals of X_i and X_j when regressing both on the remaining components. Some matrix calculus yields that

$$p_{i,j} = -\frac{K_{i,j}}{\sqrt{K_{i,i}K_{j,j}}}, \tag{1}$$

see, e.g., Whittaker (1990, Chapter 5). Thus the pairwise partial correlations are obtained from the inverse covariance matrix K in a very similar fashion as the pairwise correlations are obtained from the covariance matrix Σ itself. The only difference is the minus sign. Consequently, an absent edge in the graph means a zero entry in K, and finding the graph for given data comes down to finding the zero pattern in the inverse of the true covariance matrix Σ. Three basic sub-tasks of graphical modeling can be identified:

(T1) Finding an appropriate graph
(T2) Determining if a given graph fits the data
(T3) Estimating the (remaining) partial correlations under a given graph structure

Whether (T1) or (T2) is considered more important may be debatable and mainly depends on whether one pursues an explorative or an inferential analysis. Task (T3) may appear of lesser importance, but it is intrinsically linked to (T2). We briefly outline how these tasks are approached, in reverse order, starting in with (T3).

Task (T3) We require some mathematical notation. Define a graph $G = (V, E)$ as a set of vertices $V = \{1, \ldots, p\}$ and a set of undirected edges $E \subseteq \{\{i, j\} : i, j = 1, \ldots, p,\ i < j\}$. Let $\mathscr{S}_p$ denote the set of all symmetric $p \times p$ matrices and $\mathscr{S}_p^+$ the set of all positive-definite, symmetric $p \times p$ matrices. For any graph $G = (V, E)$, let $\mathscr{S}_p^+(G)$ be the set of matrices $A \in \mathscr{S}_p^+$ with zero entries at off-diagonal positions specified by G, i.e., $A_{i,j} = 0$ for all $i, j = 1, \ldots, p$, $i \neq j$, with $\{i, j\} \notin E$. We call any set of p-dimensional probability measures with the common property that they possess a concentration matrix $K \in \mathscr{S}_p^+(G)$ a *covariance selection model* induced by G. We call a covariance selection model consisting of all regular p-variate Gaussian distributions a *Gaussian graphical model* and denote it by $N_p(G)$, i.e., $N_p(G) = \{N_p(\boldsymbol{\mu}, \Sigma) : \boldsymbol{\mu} \in \mathbb{R}^p, \Sigma^{-1} \in \mathscr{S}_p^+(G)\}$. The maximum-likelihood estimator $\hat{\Sigma}_G$ of Σ within the parametric family $N_p(G)$ is given as the solution to

$$\hat{\Sigma}_G = \underset{\Sigma^{-1} \in \mathscr{S}_p^+(G)}{\arg\min} \left\{ \log\det\Sigma + \frac{1}{n}\sum_{i=1}^{n} \operatorname{trace}\Big(\hat{\Sigma}_n\, \Sigma^{-1}\Big) \right\}. \tag{2}$$

This optimization problem leads to the estimation equations

$$\begin{cases} \Big[\hat{\Sigma}_G\Big]_{i,j} = \Big[\hat{\Sigma}_n\Big]_{i,j} & \text{for } \{i, j\} \in E \text{ or } i = j, \\ \Big[\hat{\Sigma}_G^{-1}\Big]_{i,j} = 0 & \text{for } \{i, j\} \notin E \text{ and } i \neq j. \end{cases}$$

The solution $\hat{\Sigma}_G$ depends on the data only through the sample covariance matrix $\hat{\Sigma}_n$. This approach is due to Dempster (1972), and the optimization problem (2) has since been thoroughly studied (e.g., Speed and Kiiveri 1986). Algorithms to compute $\hat{\Sigma}_G$ for arbitrary graphs G can be found, e.g., in Lauritzen (1996, Chapter 5) or Hastie et al. (2009, Chapter 17). For decomposable graphs G, there is also an explicit solution, i.e., $\hat{\Sigma}_G$ can be computed in a finite number of steps. For details, see also Lauritzen (1996, Chapter 5). In R (R Core Team 2022), these algorithms are implemented in the function `fitConGraph` in the package ggm (Marchetti et al. 2020). With a graph-constrained[1] estimate $\hat{\Sigma}_G$ for the covariance matrix available, the remaining non-zero partial correlations are computed from $\hat{\Sigma}_G$ as unconstrained partial correlations are computed from $\hat{\Sigma}_n$ by virtue of (1).

Task (T2) Within the parametric framework described above, the likelihood-ratio test for testing G against the full model is given by the test statistic

$$D_n^{\Sigma}(G) = n\Big(\log\det\hat{\Sigma}_G - \log\det\hat{\Sigma}_n\Big), \tag{3}$$

which, under the null hypothesis that G is the true graph, converges to a χ^2_q distribution as $n \to \infty$, where q is the number of missing edges in G. The quantity D_n^{Σ} is also called *deviance* and this likelihood-ratio test hence *deviance test*. It simultaneously tests the absence of all edges not in G avoiding any multiple-testing problems. The deviance is also returned by `fitConGraph`.

Task (T1) While statistical theory provides rather precise and unambiguous solutions to tasks (T3) and (T2), this is not the case for (T1), which is already reflected by the phrasing of *finding an appropriate graph* rather than *finding the best-fitting graph*. Deciding on an appropriate graph may also be influenced by interpretability aspects and relevant domain knowledge. A multitude of approaches exist. A basic idea, also initiated by Dempster (1972), is the iterative application of the deviance test. For instance, one starts with the full graph, then removes one or several edges (with small absolute partial correlations), and keeps the new candidate graph if the deviance test accepts it. This may be iterated until no further edge removal leads to an accepted graph. The opposite search direction is also possible: One starts with the empty graph and successively adds edges until a graph is obtained which is accepted by the deviance test. For further reading, see the textbooks by Whittaker (1990) or Edwards (2000). Elaborate model search strategies have been proposed (e.g., Drton and Perlman 2008; Edwards and Havránek 1985).

Many other model selection approaches use L_1-regularization and are aimed at finding sparse graphs in high-dimensional settings. For instance, Meinshausen and Bühlmann (2006) propose a node-wise LASSO-regression. Yuan and Lin (2007) and Friedman et al. (2008) add an L_1-penalty for $K = \Sigma^{-1}$ to the optimization problem (2). Various algorithms have since been proposed for an efficient computation of such high-dimensional optimization problems (e.g., Cai et al. 2011; Sun and Zhang 2013; Yuan 2010). Within this framework, the regularization parameter must be chosen, usually by means of cross-validation. Liu and Wang (2017) particularly address the latter issue.

[1] That is, it obeys the zero pattern in the inverse induced by G.

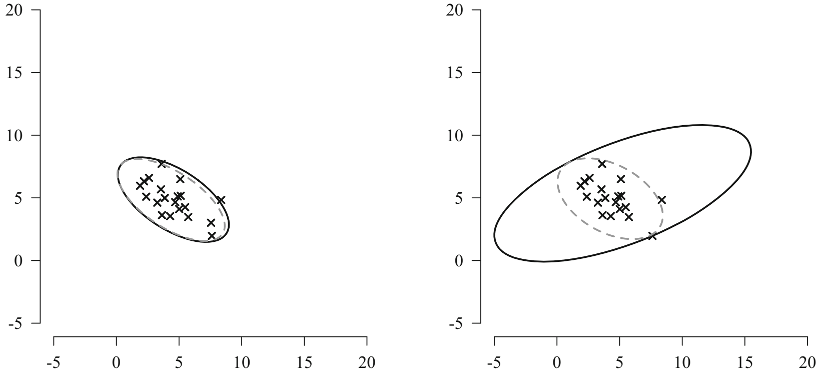


Fig. 1 The non-robust sample covariance matrix (solid line) and the robust t_3 M-estimate (dashed line). The data on the left-hand and right-hand panels differ only by one point

1.2 Robustness

Robustness in general terms is the property of a statistical method to yield sensible results if its assumptions are violated. In a more specific sense, it means insensitivity to outliers. Starting with the pioneering work of Huber (1964) and Hampel (1971, 1974), robust statistics has evolved into a large research area, see, e.g., the textbooks by Huber and Ronchetti (2009) or Maronna et al. (2019). For our purposes, the important fact to note is that the sample covariance matrix $\hat{\Sigma}_n$ is not robust. In Fig. 1, the left-hand panel shows a small data set of 20 two-dimensional observations. The black ellipse visualizes the sample covariance matrix, i.e., the 95% probability ellipse of the thus fitted normal model. In the right-hand panel, one single observation has been moved from the center to the upper right corner. The covariance estimate has tremendously changed, suggesting even a positive rather than a negative correlation. The dashed curve, in contrast, represents an alternative, robust estimator of multivariate scatter (a t_3 M-estimator, see below) and is little altered by the outlier.

With the sample covariance matrix $\hat{\Sigma}_n$ being the main ingredient of essentially all graphical modeling tasks, they all inherit its lack of robustness. The good news is: robust alternatives exist. The work on robust multivariate location and scatter estimation has been originated by Maronna (1976), who developed Huber's M-estimation approach for the multivariate setting. Since then, many proposals have been made.[2] Here we consider only one rather simple and easy-to-compute robust scatter estimator, the t_ν M-estimator, which is also already mentioned in Maronna's paper. This is an M-estimator, whose loss function stems from the maximum-likelihood estimator within the elliptical t_ν model. The parameter ν is referred to

[2] With numerous contributions by David Tyler (e.g., Kent and Tyler 1996; Tyler 1987).

as the *degrees of freedom* and may be any positive real number. The smaller the ν, the heavier-tailed the t_ν distribution and, consequently, the more outlier-resistant the corresponding M-estimator. The parameter ν is usually not inferred from the data but selected by the data analyst. A common choice is $\nu = 3$. This is not extremely heavy-tailed, second moments are finite (and hence the covariance matrix is properly defined), but it is sufficiently heavy-tailed to yield a strongly outlier-resistant estimator.

The t_ν M-estimator of scatter $\hat{S}_n$, along with the corresponding estimate of location $\hat{\boldsymbol{\mu}}_n$, is defined as the solution to the optimization problem

$$(\hat{\boldsymbol{\mu}}_n, \hat{S}_n) = \underset{\boldsymbol{\mu}\in\mathbb{R}^p, S\in\mathscr{S}_p^+}{\arg\min} \left[\sum\nolimits_{i=1}^{n} \rho_{\nu,p}\left\{(\boldsymbol{x}_i - \boldsymbol{\mu})^\top S^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right\} + n \log\det S\right] \tag{4}$$

with $\rho_{\nu,p}(x) = (\nu + p)\log(1 + x/\nu)$. This yields the estimation equations

$$\begin{cases} 0 = \sum\nolimits_{i=1}^{n} \psi_{\nu,p}(\hat{r}_i)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n), \\ \hat{S}_n = n^{-1}\sum\nolimits_{i=1}^{n} \psi_{\nu,p}(\hat{r}_i)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)^\top, \end{cases}$$

where $\psi_{\nu,p}(x) = \rho'_{\nu,p}(x) = (\nu+p)/(\nu+x)$ and $\hat{r}_i = (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)^\top \hat{S}_n^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)$. The t_ν M-estimator can be computed, e.g., by a fixed-point algorithm. Implementations in R can be found, e.g., in the functions `cov.trob` from the package MASS (Venables and Ripley 2002), `tM` from the package ICS (Nordhausen et al. 2008), and `MVTMLE` from the package fastM (Dümbgen et al. 2016, 2018). The latter uses a partial Newton–Raphson algorithm.

With robust alternatives $\hat{S}_n$ being available, an intuitive path to a robust analysis is the plug-in approach: First solving (4) and then plugging the thus obtained estimate $\hat{S}_n$ instead of $\hat{\Sigma}_n$ into (2) to obtain a graph-constrained robust estimate $\hat{S}_G$ and the corresponding robust (pseudo-)deviance

$$D_n^S(G) = n\Big(\log\det\hat{S}_G - \log\det\hat{S}_n\Big). \tag{5}$$

This is accomplished in the package robFitConGraph by the function of the same name. Plug-in robustifications for the L_1-regularization methods have equally been proposed (Öllerer and Croux 2015; Tarr et al. 2016). Alternatively, Finegold and Drton (2011) regularize the elliptical t_ν log density.

Before providing further details on the function `robFitConGraph` and the underlying theory in Sect. 3, its use shall be demonstrated at a data example on music performance anxiety.

2 A Case Study: Music Performance Anxiety

The fear about one's ability to perform a specific task, such as giving a presentation or sitting an exam, affects almost everyone. Pressure can be particularly high in certain professions where performing in front of others is an integral part of day-to-day life. While some levels of stress and anxiety are normal and actually help us to achieve optimal performance, severe levels of stress and anxiety are debilitating and can develop into a disorder. Professional musicians are often exposed to extreme pressure where maintaining top-quality performances is not just essential to keeping their job, but to progress in their careers. Music performance anxiety (MPA) can be understood as a continuum ranging from low to high anxiety levels. The latter poses a serious problem to the profession and is the subject of ongoing clinical research, see, e.g., Fernholz et al. (2019) for a recent review.

MPA is often considered to be a form of social anxiety that, loosely speaking, is the overwhelming fear of social situations (e.g., Cox and Kenardy 1993; Dobos et al. 2019; Kenny 2011; Nicholson et al. 2015). This is underlined by the fact that the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association 2013) now acknowledges evidence of individuals suffering exclusively from performance anxiety as a distinct sub-type of social anxiety disorder. However, some researchers and clinicians have questioned this description, as MPA is a complex phenomenon caused by the interaction of many different factors, and the fear of social judgment is not necessarily always the main problem.

Wiedemann et al. (2022) analyzed a data set consisting of eight numerical variables measured at $n = 82$ students at German music colleges: music performance anxiety (MPA), agoraphobia (AG), generalized anxiety disorder (GAD), panic disorder (PD), separation anxiety disorder (SEP), specific phobia (SP), social anxiety disorder (SAD) as well as illness anxiety disorder (ILL). Each variable is a summary score from a self-assessment inventory with Likert-scale items. A higher value signifies a higher severity of the condition. MPA was assessed using the German version of the Kenny Music Performance Anxiety Inventory (K-MPAI; Kenny 2009) translated by Spahn et al. (2016). All other anxieties were assessed using the German translation of the disorder-specific anxiety measures (Beesdo-Baum et al. 2012; Lebeau et al. 2012) for the dimensional anxiety scales of the DSM-5. The data set is also included in the package robFitConGraph as `anxieties`.

Figure 2 shows the pairwise scatter plots of the data set. While most participants score low on most anxiety scales—as we would expect and hope—there are a few very high values in all variables. The normality assumption can be seen to be violated in a similar manner as in Fig. 1 with the same implications for any sample-covariance-based analysis. For instance, removing the outlier in the variable AG reduces the Pearson correlation between AG and SEP from 0.612 to 0.407. Computing the correlation coefficient from a t_3 M-estimator, we obtain the value 0.410, which reduces to 0.350 when removing said outlier. A robust analysis is highly recommended for these data. All results reported in the following are based

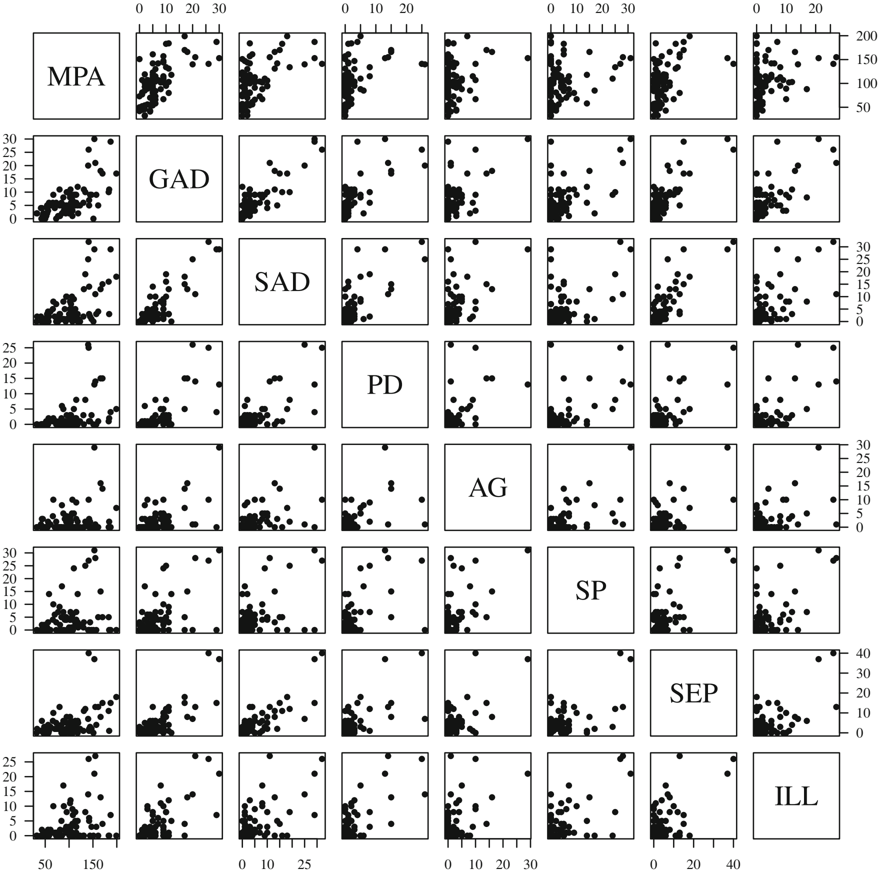


Fig. 2 Pairwise scatter plots of the anxieties data

on a t_3 M-estimator of scatter. It yields the correlation coefficients given in Table 1. They can also be obtained by the function `robFitConGraph` by supplying the full model as adjacency matrix

```
> library(robFitConGraph)
> data(anxieties)
> p <- ncol(anxieties)
> Shat <- robFitConGraph(X = anxieties,
+     amat = matrix(1, ncol = p, nrow = p),
+     df = 3)$Shat
> round(cov2cor(Shat), d = 2)
```

All variables are positively correlated. This is neither surprising nor uncommon for multivariate data. The corresponding partial correlations, as given in Table 2, shed further light on the dependence structure of the variables.

Table 1 Pairwise correlations computed from t_3-M-estimate

	MPA	GAD	SAD	PD	AG	SP	SEP	ILL
MPA	·	0.62	0.37	0.43	0.17	0.26	0.38	0.32
GAD	0.62	·	0.66	0.65	0.40	0.35	0.60	0.47
SAD	0.37	0.66	·	0.50	0.51	0.36	0.64	0.42
PD	0.43	0.65	0.50	·	0.51	0.44	0.48	0.46
AG	0.17	0.40	0.51	0.51	·	0.49	0.41	0.36
SP	0.26	0.35	0.36	0.44	0.49	·	0.37	0.34
SEP	0.38	0.60	0.64	0.48	0.41	0.37	·	0.29
ILL	0.32	0.47	0.42	0.46	0.36	0.34	0.29	·

Table 2 Pairwise *partial* correlations computed from t_3-M-estimate

	MPA	GAD	SAD	PD	AG	SP	SEP	ILL
MPA	·	0.44	−0.05	0.05	−0.14	0.08	0.05	0.05
GAD	0.44	·	0.33	0.34	−0.02	−0.06	0.19	0.15
SAD	−0.05	0.33	·	−0.05	0.25	0.00	0.37	0.13
PD	0.05	0.34	−0.05	·	0.24	0.14	0.09	0.16
AG	−0.14	−0.02	0.25	0.24	·	0.30	0.04	0.07
SP	0.08	−0.06	0.00	0.14	0.30	·	0.11	0.12
SEP	0.05	0.19	0.37	0.09	0.04	0.11	·	−0.12
ILL	0.05	0.15	0.13	0.16	0.07	0.12	−0.12	·

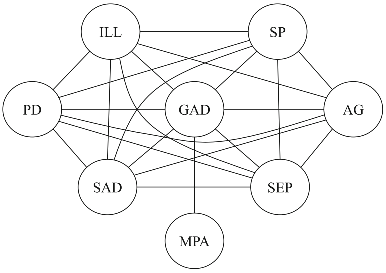
```
> Phat <- -cov2cor(solve(Shat))
> diag(Phat) <- 1
> round(Phat,d = 2)
```

Many of the partial correlations are near zero, suggesting conditional independences, i.e., their association may be fully mediated by other variables in the data set.

2.1 *Inferential Analysis: MPA and Social Anxiety*

One hypothesis examined by Wiedemann et al. (2022) is whether MPA is primarily related to a social anxiety disorder (SAD). Based on the positive correlation of 0.37, a short-sighted analysis may deduce a strong connection between MPA and SAD, which in light of other, even larger correlations, is right away questionable. As we will see, the data in fact carry no evidence for a particularly strong connection between MPA and SAD.

For that purpose, one tests the hypothesis that MPA and SAD are conditionally independent given all remaining six variables, i.e., testing a graphical model with only one missing edge between MPA and SAD by means of a robust pseudo-deviance test. If this test rejects, there is evidence for a strong link between MPA and



	MPA	GAD	SAD	PD	AG	SP	SEP	ILL
MPA	1	1	0	0	0	0	0	0
GAD	1	1	1	1	1	1	1	1
SAD	0	1	1	1	1	1	1	1
PD	0	1	1	1	1	1	1	1
AG	0	1	1	1	1	1	1	1
SP	0	1	1	1	1	1	1	1
SEP	0	1	1	1	1	1	1	1
ILL	0	1	1	1	1	1	1	1

Fig. 3 The graph encoding the hypothesis MPA is conditionally independent of all other variables given GAD (left) and the corresponding adjacency matrix (right)

SAD. Their dependence can then not be fully explained by their associations with other anxiety types. However, this hypothesis is *not* rejected (a p-value of 0.66), and we find no such evidence in the data. The p-value is returned by the function `robFitConGraph` via the named list element `pval`, see also the next code chunk below.

One can even further test the stronger hypothesis that MPA is, given GAD only, conditionally independent of all other six specific anxiety scales. This is done by testing the graph in Fig. 3, which can equivalently be expressed by its adjacency matrix on the right-hand side of Fig. 3.

Almost unnoticed, we make here use of a non-trivial result, which is the real merit and the real beauty of graphical models: Each absent edge in the graph denotes a conditional independence of two individual variables given the respective remaining six variables. This is indeed equivalent to MPA and (SAD, PD, AG, SP, SEP, ILL) being conditionally independent given GAD only, because in the graph GAD separates MPA from (SAD, PD, AG, SP, SEP, ILL). This is known as the equivalence between the local Markov property and the global Markov property. For details, see Lauritzen (1996, Chapter 3).

The pseudo-deviance test based on the t_3 M-estimator for the graph in Fig. 3 is carried out as follows:

```
> amat <- matrix(1, ncol = p, nrow = p)
> rownames(amat) <- colnames(anxieties)
> colnames(amat) <- colnames(anxieties)
> amat["MPA", c("SAD", "PD", "AG", "SP", "SEP", "ILL")] <- 0
> amat[c("SAD", "PD", "AG", "SP", "SEP", "ILL"), "MPA"] <- 0
> robFitConGraph(X = anxieties, amat = amat, df = 3)$pval
[1] 0.881159
```

With a p-value of 0.88, this hypothesis is also *not* rejected. There is no evidence against the null hypothesis of MPA being conditionally independent of the specific anxiety types given generalized anxiety (GAD). Accepting the null hypothesis, GAD is fully sufficient for predicting MPA. When already knowing a person's GAD score, i.e., how anxious the person generally is, additionally knowing any of their specific anxieties scores provides no further information about their MPA score. So

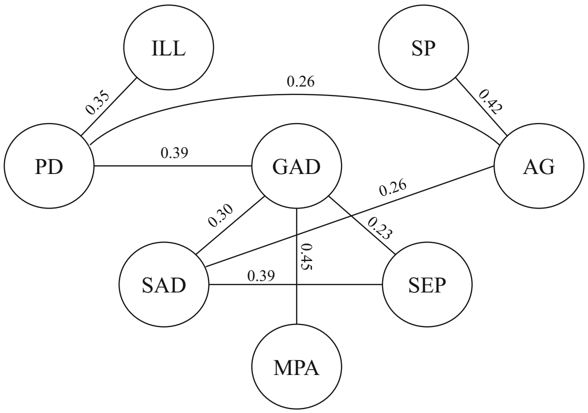


Fig. 4 A fitting graph with 9 edges based on a t_3 scatter estimator with p-value 0.45

this analysis does not support the hypothesis that MPA is foremost related to social anxiety.

The analysis is complemented by testing if MPA and GAD are conditionally independent given the remaining six variables. This hypothesis *is* rejected with a p-value below 0.01, which corresponds to the partial correlation of 0.44 in Table 2.[3] So there is a strong connection between MPA and GAD, this connection is not mediated by any of the other variables, and GAD explains the connection between MPA and the remaining variables. Altogether, GAD is the link between MPA and the other anxieties.

2.2 Explorative Analysis

Naturally, the question arises which other edges may be removed. For that purpose, we remove all edges that have absolute partial correlation below 0.15 (Table 2), corresponding to an individual p-value above 0.2. The resulting graph is depicted in Fig. 4, which is not rejected by the pseudo-deviance test with a p-value of 0.45. The partial correlations along the edges in Fig. 4 are fitted under the graph, i.e., they are estimated taking the graph structure into account. They are different from the partial correlations given in Table 2.

```
> amat <- abs(Phat) > 0.15
> Shat_G <- robFitConGraph(X = anxieties,
+     amat = amat, df = 3)$Shat
> Phat_G <- -cov2cor(solve(Shat_G))
> diag(Phat_G) <- 1
> round(Phat_G, d = 2)
```

[3] For a single missing edge, the deviance test can indeed be expressed in terms of corresponding partial correlation only.

The information learned from such a fitted graph is, e.g., that GAD is central within this set of variables: It has many edges to other vertices, and it has much explanatory power about the other variables. If one were to retain only a single variable (as a very simple dimension reduction approach, say), GAD would be a natural candidate based on this "vertex degree criterion."

The fitted graph in Fig. 4 with 9 edges (out of 28 possible) is one parsimonious graph that fits the data. Generally, there is no well-defined *most parsimonious*, *best-fitting* graph, as obviously fit and parsimony are contradicting goals. However, one may elaborate upon the initial search by checking if other (equally or more) parsimonious graphs may fit as well. For instance, increasing the initial partial correlation threshold to 0.17, say, which results in a further removal of the edge ILL–PD, leads to a p-value of 0.01 and hence should be rejected. Alternatively, removing the "next smallest edge" from the candidate graph in Fig. 4, which is GAD–SEP, results in a p-value of 0.17, which is still acceptable.

Despite using deviance-test p-values as decision criterion whether to accept a graph or not, this analysis is purely explorative. We have not *tested* the graph of Fig. 4. Hypotheses to be tested must be formed a priori. Testing for a graph that is the result of model selection procedure is as prohibitive as testing if the observed sample mean is the population mean.

So far, we have used `robFitConGraph` with `df = 3`, which is also the default setting in case `df` is not specified. Generally, the results are not very sensitive to variations in ν. A smaller value of ν downweights outliers more strongly. In the present example, taking $\nu = 1$, we find the hypothesis of Fig. 3 equally accepted with a p-value of 0.58. The explorative graph in Fig. 4 is accepted with a p-value of 0.26.

2.3 The Classical Analysis

After having cautioned its use at the beginning, we close this section by remarking that a classical sample-covariance-based analysis leads qualitatively to the same findings. The classical deviance test gives a p-value of 0.63 for the hypothesized graph of Fig. 3. An explorative analysis analogous to the one above leads to the graph in Fig. 5 with a p-value of 0.51, which is also communicated by Wiedemann et al. (2022). The graph is different, but it equally shows the centrality of GAD as well as the marginality of MPA within the set of variables. It is generally advisable to try varying parameter settings and different methods, robust and non-robust, in any statistical analysis. Conclusions unanimously obtained by several methods may be considered even more trustworthy.

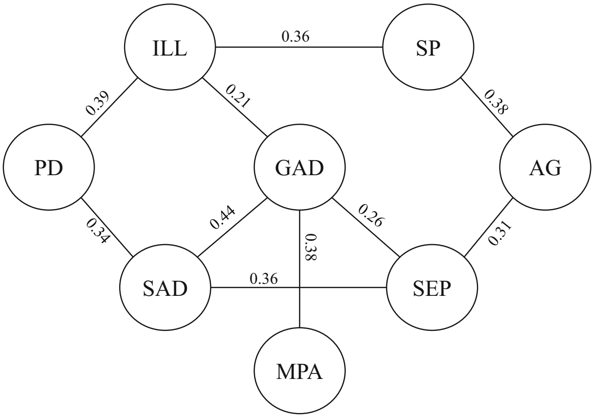


Fig. 5 A fitting graph with 10 edges based on the sample covariance matrix with p-value 0.51

3 Background and Theory

The centerpiece of the package robFitConGraph is the function by the same name. For a given data set $\mathbb{X}_n$ and a given graph G, it simultaneously provides a robust graph-constrained matrix $\hat{S}_G$ and the p-value of the corresponding pseudo-deviance test. But so far, the actual worth of the function `robFitConGraph` has not become apparent: It appears, its plug-in functionality is equally achieved by a combination of, say, `cov.trob` and `fitConGraph`, as the latter takes the sample covariance matrix as input. However, there are two good reasons for `robFitConGraph`, which both require a slightly deeper foray into statistical theory:

(1) The limit distribution of the pseudo-deviance $D_n^S(G)$ does not follow strictly a χ_q^2 distribution under the graph G, but a χ_q^2-variate multiplied by a constant σ_1. This constant is required to obtain p-values and depends on the data dimension p and the degrees of freedom ν of the estimator. It is computed by the function `find_sigma1`.
(2) The plug-in estimator $\hat{S}_G$ is just one approach to a graph-constrained robust scatter estimator. The function `robFitConGraph` also provides an alternative approach, referred to as the *direct estimator* and denoted by $\tilde{S}_G$ below.

Another good reason is speed. Being implemented in C++, `robFitConGraph` is considerably faster than the combination of `cov.trob` and `fitConGraph`. A runtime comparison is given by Watt (2019).

3.1 The Constant σ_1

Recall the class of all p-dimensional, continuous, elliptical distributions, i.e., distributions possessing a p-dimensional Lebesgue density f of the form

$$f(\boldsymbol{x}) = \det(S)^{-\frac{1}{2}} g\big\{(\boldsymbol{x}-\boldsymbol{\mu})^\top S^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big\} \tag{6}$$

for some $\boldsymbol{\mu} \in \mathbb{R}$, $S \in \mathscr{S}_p^+$, and $g : [0,\infty) \to [0,\infty)$, such that f integrates to 1. Let $E_p(\boldsymbol{\mu}, S, g)$ denote the distribution described by (6). The univariate function g is called the *elliptical generator* and S the *scatter* or *shape matrix* of $E_p(\boldsymbol{\mu}, S, g)$. It is proportional to the covariance matrix if $E_p(\boldsymbol{\mu}, S, g)$ has second moments. In the present paper, we only consider two examples of elliptical distributions: the normal distribution with

$$g_N(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{x}{2}\Big), \qquad 0 \le x,$$

and the elliptical t_ν distribution with

$$g_{\nu,p}(x) = c_{\nu,p}\Big(1+\frac{x}{\nu}\Big)^{-(\nu+p)/2}, \qquad 0 \le x,$$

where the normalizing constant $c_{\nu,p}$ is given in the Technical Appendix at the end of the chapter. Assume the independent observations $\mathbb{X}_n = (\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)^\top$ to stem from an elliptical distribution $E_p(\boldsymbol{\mu}, S, g)$, and let $\hat{V}_n$ be an *arbitrary* $p \times p$ scatter matrix estimator fulfilling two fairly natural conditions:

(1) $\hat{V}_n$ is affine equivalent, i.e., $\hat{V}_n(\mathbb{X}_n A^\top + \mathbf{1}_n \boldsymbol{b}^\top) = A\hat{S}_n(\mathbb{X}_n)A^\top$ for any $\boldsymbol{b} \in \mathbb{R}^p$ and full rank $A \in \mathbb{R}^{p\times p}$, where $\mathbf{1}_n$ is the n-vector consisting of ones.

(2) $\hat{V}_n$ is $\sqrt{n}$-consistent and asymptotically normal at $E_p(\boldsymbol{\mu}, S, g)$, i.e., there exist matrices $V \in \mathscr{S}_p$ and $W \in \mathscr{S}_{p^2}$ such that

$$\sqrt{n}\mathrm{vec}\{\hat{V}_n(\mathbb{X}_n) - V\} \ \to \ N_{p^2}(\mathbf{0},\ W)$$

in distribution as $n \to \infty$.

Tyler (1982) showed that under these two conditions, V and W take on specific forms: $V = \eta S$ for some $\eta > 0$, and

$$W = 2\eta^2\sigma_1 \mathcal{M}_p(S \otimes S) \ + \ \eta^2\sigma_2 \mathrm{vec(S)vec(S)}^\top, \tag{7}$$

where $\otimes$ is the Kronecker product, $\sigma_1 \ge 0$ and $\sigma_2 \ge -2\sigma_1/p$ are scalar constants independent of $\boldsymbol{\mu}$ and S, and $\mathcal{M}_p$ is a fixed $p^2 \times p^2$ matrix defined in the Technical Appendix. The latter formula greatly simplifies the asymptotic efficiency comparison of any two affine equivariant scatter estimators at elliptical distributions.

In case of $\hat{V}_n$ being an elliptical M-estimator, with a general loss function ρ instead of $\rho_{\nu,p}$ in (4), Tyler (1983) gives the following expressions for the scalars η, σ_1, and σ_2: Letting $\psi(x) = \rho'(x)$, $\phi(x) = x\psi(x)$, and $R = (\boldsymbol{X} - \boldsymbol{\mu})^\top S^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ for $\boldsymbol{X} \sim E_p(\boldsymbol{\mu}, S, g)$, the scalar η is the solution to $\mathbf{E}\{\phi(R/\eta)\} = p$. Letting further

$$\gamma_1 = \frac{\mathbf{E}\{\phi^2(R/\eta)\}}{p(p+2)}, \qquad \gamma_2 = \frac{1}{p}\mathbf{E}\left\{\frac{R}{\eta}\phi'\left(\frac{R}{\eta}\right)\right\},$$

the scalars σ_1 and σ_2 are

$$\sigma_1 = \frac{(p+2)^2\gamma_1}{(2\gamma_2+p)^2}, \qquad \sigma_2 = \gamma_2^{-2}\left[\gamma_1 - 1 - \frac{2\gamma_1(\gamma_2-1)\{p+(p+4)\gamma_2\}}{(2\gamma_2+p)^2}\right].$$

Tyler (1983) further showed that the asymptotic variance of any scale-invariant, continuously differentiable function h of $\hat{V}_n$ only depends on σ_1 and not on σ_2. A *scale-invariant* function $h : \mathscr{S}_p \to \mathbb{R}$ satisfies $h(\alpha V) = h(V)$ for any $\alpha > 0$. The pseudo-deviance-test statistic (5) is such a scale-invariant function of the scatter estimator $\hat{S}_n$. Dependence is an inherently scale-free concept. So this equally applies to any aspect of multivariate scatter that quantifies dependence in one way or another, may it be correlations, partial correlations, canonical correlations, principal components, etc.

In the package robFitConGraph, the functions `find_eta` and `find_sigma1` compute the scalars η and σ_1, respectively, for t_{ν_1} M-estimators in case the data stem from a normal or an elliptical t_{ν_2} distribution. Note that the degrees of freedom ν_1 of the estimator loss function $\rho_{\nu_1,p}$ and the degrees of freedom of the population distribution ν_2 can be generally different. If they coincide, the t_{ν_1} M-estimator is a maximum-likelihood estimator.

The value of ν_1 is specified by `df_est` and ν_2 by `df_data`. For both, `Inf` is allowed, which corresponds to the sample covariance matrix and the normal distribution, respectively. For `find_sigma1`, the input `df_est = 0` is also allowed. This corresponds to Tyler's distribution-free M-estimator of scatter (Tyler 1987). In this case, $\sigma_1 = 1 + 2/p$ regardless of the elliptical population distribution. There is no t_0 distribution; hence, `df_data = 0` is not allowed.

The value σ_1 can be given directly to `robFitConGraph` via the optional argument `sigma1`. If none is provided, `robFitConGraph` calls `find_sigma1` with `df_data = Inf` (i.e., assuming Gaussian data) and its argument `df` being passed on as `df_est` to the function `find_sigma1`. The argument `df` of `robFitConGraph` is optional with the default `df = 3`.

3.2 The Direct vs. the Plug-in Estimate

Instead of solving (4) and (2) sequentially, an alternative approach is to directly solve the optimization problem

$$(\tilde{\boldsymbol{\mu}}_G, \tilde{S}_G) = \underset{\boldsymbol{\mu} \in \mathbb{R}^p, S^{-1} \in \mathscr{S}_p^+(G)}{\arg\min} \left[\sum\nolimits_{i=1}^n \rho_{\nu,p} \left\{ (\boldsymbol{x}_i - \boldsymbol{\mu})^\top S^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right\} + n \log\det S \right], \tag{8}$$

with $\rho_{\nu,p}$ as in (4), leading to the estimation equations

$$\begin{cases} 0 = \sum\nolimits_{i=1}^n \psi_{\nu,p}(\tilde{r}_i)(\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_G), & \\ \left[\tilde{S}_G\right]_{j,k} = \left[n^{-1} \sum\nolimits_{i=1}^n \psi_{\nu,p}(\tilde{r}_i)(\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_G)(\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_G)^\top \right]_{j,k}, & \text{for } \{j,k\} \in E \ \text{ or } \ j = k, \\ \left[\tilde{S}_G^{-1}\right]_{j,k} = 0, & \text{for } \{j,k\} \notin E \ \text{ and } \ j \neq k, \end{cases} \tag{9}$$

where $\tilde{r}_i = (\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_G)^\top \tilde{S}_n^{-1} (\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_G)$ and, as before, $\psi_{\nu,p}(x) = \rho'_{\nu,p}(x)$.

The estimator $\tilde{S}_G$ shall be called the *direct estimator*, which is short for *direct graph-constrained t_ν M-estimator*, and is an alternative to the plug-in graph-constrained t_ν M-estimator. Using the function `robFitConGraph`, the direct estimator is invoked by setting the option `direct = TRUE` or `plug_in = FALSE`. In case of conflicting specifications, `plug_in` has priority, and a message will be displayed. Contrary to the plug-in estimator $\hat{S}_G$, the direct estimator $\tilde{S}_G$ is not a function of the corresponding unconstrained estimate $\hat{S}_n$ alone.

One main theoretical result of Vogel and Tyler (2014) is the asymptotic equivalence of $\hat{S}_G$ and $\tilde{S}_G$ under elliptical population distributions. Hence, the limiting distribution of the pseudo-deviance and the constant σ_1 are the same in both cases. It may be argued that this asymptotic equivalence result favors the plug-in estimator: Considering the elliptical distribution (6) with the fixed generator $g_{\nu,p}$, one defines the elliptical-t_ν graphical model $E_p(g_{\nu,p}, G)$ analogously to the Gaussian graphical model $N_p(G)$. The direct graphical t_ν M-estimator $(\tilde{\boldsymbol{\mu}}_G, \tilde{S}_G)$ with the corresponding loss function $\rho_{\nu,p}$ is then the maximum-likelihood estimator within this parametric family.[4] So direct graphical M-estimators generalize maximum-likelihood estimators, which are known to be first-order efficient. Plug-in estimators are popular with practitioners as they are easily applied, fast to compute, and now it turns out that they also possess desirable asymptotic properties. Direct graphical M-estimators are generally harder to compute. They may be solved by a double-loop, iteratively

[4] The parameters of interest are $\boldsymbol{\mu}$ and S. The degrees of freedom ν are held fixed.

reweighted least-squares algorithm, where the Gaussian model fit is nested into the M-estimation loop.

The other half of the story is that direct graphical M-estimators can be substantially more efficient in small samples as is demonstrated by simulations in Vogel and Tyler (2014).

3.3 Ellipticity vs. Normality

Multivariate data containing outliers may be modeled conceptually in two different ways: Either by a corrupted Gaussian distribution, i.e., a few observations are erroneous and stem from a different, outlier-generating distribution, or by a heavy-tailed elliptical distribution, which generates outliers itself. Using the t_ν M-estimator implicitly suggests the latter viewpoint. However, two issues arise:

(1) Many data sets exhibit features such as the anxiety data set: It is clearly not normal as it contains outliers, but it is clearly not elliptical either as it is skewed. Hence, we do adopt the viewpoint of corrupted normal data. In the analysis, however, we apply outlier-resistant methods that have been derived from considerations in the elliptical t_ν model. We work on the plausible, but not formalized assumption that they provide outlier-resistance regardless of the outliers being scattered symmetrically around the center or not.

(2) An absent edge in a Gaussian graphical model, i.e., a zero partial correlation, has the interpretation of conditional independence, a notion we have repeatedly used in Sect. 2. Under ellipticity, an absent edge has the slightly weaker interpretation of conditional uncorrelatedness. This is occasionally mentioned as a limitation of elliptical graphical models. However, this limitation is largely void. The conclusions of any statistical analysis are always a combination of the information contained in the data and the modeling assumptions. When performing a purely linear analysis based on partial correlations, concluding conditional independences due the normality assumption certainly fall into the latter category.

Technical Appendix

The normalizing constant in the elliptical t_ν density is

$$c_{\nu,p} = \frac{\Gamma\{(\nu + p)/2\}}{(\nu\pi)^{p/2}\Gamma(\nu/2)},$$

where Γ is the gamma function, see, e.g., Bilodeau and Brenner (1999, Chapter 13).

The Kronecker product $A \otimes B$ of two matrices $A, B \in \mathbb{R}^{p \times p}$ is defined as the $p^2 \times p^2$ matrix with entry $a_{i,j} b_{k,l}$ at position $((i-1)p+k, (j-1)p+l)$. For a matrix $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p) \in \mathbb{R}^{p \times p}$, the notation vec(A) means the p^2-vector obtained by stacking the columns of A, i.e., $\mathrm{vec}(A) = (\boldsymbol{a}_1^\top, \ldots, \boldsymbol{a}_p^\top)^\top$. The matrix $\mathcal{M}_p$ in (7) is defined as

$$\mathcal{M}_p = \frac{1}{2}\big(\mathcal{I}_{p^2} + \mathcal{K}_p\big),$$

where $\mathcal{I}_{p^2}$ denotes the $p^2 \times p^2$ identity matrix and

$$\mathcal{K}_p = \sum\nolimits_{i=1}^{p} \sum\nolimits_{j=1}^{p} \boldsymbol{e}_i \boldsymbol{e}_j^\top \otimes \boldsymbol{e}_j \boldsymbol{e}_i^\top,$$

where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ denote the Euclidean basis vectors in $\mathbb{R}^p$. The matrix $\mathcal{K}_p$, commonly referred to as the *commutation matrix*, is orthogonal and corresponds to the transpose operator $\mathcal{K}_p \mathrm{vec}(A) = \mathrm{vec}(A^\top)$. The idempotent matrix $\mathcal{M}_p$ is called the *symmetrization matrix* since it maps vec(A) to $\mathrm{vec}(A + A^\top)/2$.

Acknowledgments We thank the anonymous referees for many valuable suggestions and spotting mistakes in the originally submitted manuscript.

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Beesdo-Baum, K., Klotsche, J., Knappe, S., Craske, M. G., LeBeau, R. T., Hoyer, J., Strobel, A., Pieper, L., & Wittchen, H.-U. (2012). Psychometric properties of the dimensional anxiety scales for DSM-V in an unselected sample of German treatment seeking patients. *Depression and Anxiety, 29*(12), 1014–1024.

Bilodeau, M., & Brenner, D. (1999). *Theory of Multivariate Statistics.* Springer Texts in Statistics. New York: Springer.

Cai, T., Liu, W., & Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association, 106*(494), 594–607.

Cox, W. J., & Kenardy, J. (1993). Performance anxiety, social phobia, and setting effects in instrumental music students. *Journal of Anxiety Disorders, 7*(1), 49–60.

Dempster, A. P. (1972). Covariance Selection. *Biometrics, 28*, 157–175.

Dobos, B., Piko, B. F., & Kenny, D. T. (2019). Music performance anxiety and its relationship with social phobia and dimensions of perfectionism. *Research Studies in Music Education, 41*(3), 310–326.

Drton, M., & Perlman, M. D. (2008). A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference, 138*(4), 1179–1200.

Dümbgen, L., Nordhausen, K., & Schuhmacher, H. (2016). New algorithms for M-estimation of multivariate scatter and location. *Journal of Multivariate Analysis, 144*, 200–217.

Dümbgen, L., Nordhausen, K., & Schuhmacher, H. (2018). *fastM: Fast Computation of Multivariate M-Estimators*. R package version 0.0-4.

Edwards, D. (2000). *Introduction to graphical modelling.* Springer Texts in Statistics. New York, NY: Springer.

Edwards, D., & Havránek, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika, 72*, 339–351.

Fernholz, I., Mumm, J. L., Plag, J., Noeres, K., Rotter, G., Willich, S. N., Ströhle, A., Berghöfer, A., & Schmidt, A. (2019). Performance anxiety in professional musicians: a systematic review on prevalence, risk factors and clinical treatment effects. *Psychological Medicine, 49*(14), 2287–2306.

Finegold, M., & Drton, M. (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Annals of Applied Statistics, 5*(2A), 1057–1080.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432–441.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics, 42*, 1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*, 383–393.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Series in Statistics. Springer, New York.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*(1), 73–101.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics.* Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley.

Kenny, D. T. (2009). The factor structure of the revised Kenny Music Performance Anxiety Inventory. In *International Symposium on Performance Science* (pp. 37–41). The Netherlands: Association Européenne des Conservatoires Utrecht.

Kenny, D. T. (2011). *The Psychology of Music Performance Anxiety*. Oxford: Oxford University Press.

Kent, J. T., & Tyler, D. E. (1996). Constrained M-estimation for multivariate location and scatter. *Annals of Statistics, 24*(3), 1346–1370.

Lauritzen, S. L. (1996). *Graphical models.* Oxford Statistical Science Series. Oxford: Oxford University Press.

Lebeau, R. T., Glenn, D. E., Hanover, L. N., Beesdo-Baum, K., Wittchen, H.-U., & Craske, M. G. (2012). A dimensional approach to measuring anxiety for DSM-5. *International Journal of Methods in Psychiatric Research, 21*(4), 258–272.

Liu, H., & Wang, L. (2017). TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models. *Electronic Journal of Statistics, 11*(1), 241–294.

Marchetti, G. M., Drton, M., & Sadeghi, K. (2020). *GGM: Graphical Markov Models with Mixed Graphs*. R package version 2.5.

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

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with R)* (2nd ed.). New York: Wiley.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics, 34*(3), 1436–1462.

Nicholson, D. R., Cody, M. W., & Beck, J. G. (2015). Anxiety in musicians: On and off stage. *Psychology of Music, 43*(3), 438–449.

Nordhausen, K., Oja, H., & Tyler, D. E. (2008). Tools for exploring multivariate data: The package ICS. *Journal of Statistical Software, 28*(6), 1–31.

Öllerer, V., & Croux, C. (2015). Robust high-dimensional precision matrix estimation. In K. Nordhausen, & S. Taskinen (eds.), *Modern Nonparametric, Robust and Multivariate Methods* (pp. 325–350). Berlin: Springer.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Spahn, C., Walther, J.-C., & Nusseck, M. (2016). The effectiveness of a multimodal concept of audition training for music students in coping with music performance anxiety. *Psychology of Music, 44*(4), 893–909.

Speed, T. P., & Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Annals of Statistics, 14*, 138–150.

Sun, T., & Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research, 14*(1), 3385–3418.

Tarr, G., Müller, S., & Weber, N. C. (2016). Robust estimation of precision matrices under cellwise contamination. *Computational Statistics & Data Analysis, 93*, 404–420.

Tyler, D. E. (1982). Radial estimates and the test for sphericity. *Biometrika, 69*, 429–436.

Tyler, D. E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika, 70*, 411–420.

Tyler, D. E. (1987). A Distribution-Free M-Estimator of Multivariate Scatter. *Annals of Statistics, 15*(1), 234–251.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer. ISBN 0-387-95457-0.

Vogel, D., & Tyler, D. E. (2014). Robust estimators for non-decomposable elliptical graphical models. *Biometrika, 101*(4), 865–882.

Watt, S. J. (2019). *Algorithms for data analysis*, B.Sc. thesis. Scotland: University of Aberdeen.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester, etc.: Wiley.

Wiedemann, A., Vogel, D., Voss, C., & Hoyer, J. (2022). How does music performance anxiety relate to other anxiety disorders? *Psychology of Music, 50*, 204–2017.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research, 11*, 2261–2286.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika, 94*(1), 19–35.

Robust Estimation of General Linear Mixed Effects Models

Manuel Koller and Werner A. Stahel

Abstract The classical REML estimator for fitting a general linear mixed effects model is modified by bounding the terms appearing in the scoring equations. This leads to a generally applicable robust M-type estimator that we call robust scoring equations estimator. It requires only minor assumptions on the covariance matrices (block diagonal for the random effects and diagonal, known up to scale for the residual errors) additional to those of the classical methods. The structure of the data is arbitrary as long as the model is estimable in the classical sense. The estimator can detect and contain the effect of outliers in moderately contaminated datasets. Contamination is detected and treated at all levels of variability of the model, e.g., at both the subject and the observation level for a one-way ANOVA model. The estimator's properties are studied by simulation and two examples. One example implies crossed random effects, for which the known robust methods are not applicable.

Keywords Mixed model · Variance components · Hierarchical models · Crossed random effects · Robustness · Huberizing

1 Introduction

Classical estimators of linear mixed models are sensitive to contamination in the data. Simple screening of residuals from a classical fit has shown to be of limited use even in simpler settings (Maronna et al. 2019). Mixed effects models allow

M. Koller (✉)
Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

Seminar für Statistik, ETH Zürich, Zürich, Switzerland
e-mail: koller@stat.math.ethz.ch

W. A. Stahel
Seminar für Statistik, ETH Zürich, Zürich, Switzerland
e-mail: stahel@stat.math.ethz.ch

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_14

for variability on multiple levels of the data, which makes it even easier for contamination to hide in the data. Robust estimation methods provide the means to automatically flag and contain the effect of outliers or other contamination. Armed with the additional information, users of robust methods can investigate flagged datapoints and gain insight that would otherwise have been overlooked. Therefore, robust estimation methods are needed.

The simplest linear mixed effects model is the one-way analysis of variance model

$$Y_{hi} = \beta_0 + B_h + \varepsilon_{hi}, \tag{1}$$

where $\varepsilon_{hi} \sim \mathcal{N}\big(0, \sigma^2\big)$ and $B_h \sim \mathcal{N}\big(0, \sigma_b^2\big)$, $h = 1, \ldots, H$, $i = 1, \ldots, N_h$. It can be written in vector form,

$$\mathbf{Y}_h = \boldsymbol{X}_h \boldsymbol{\beta} + \boldsymbol{\delta}_h, \tag{2}$$

with $\boldsymbol{X}_h = \mathbf{1}$, $\boldsymbol{\beta} = \beta_0$ and correlated error vector $\boldsymbol{\delta}_h \sim \mathcal{N}_{N_h}\big(\mathbf{0}, \boldsymbol{\Sigma}_h\big)$, $\boldsymbol{\Sigma}_h = \sigma^2 \boldsymbol{I} + \sigma_b^2 \mathbf{1}\mathbf{1}^\mathsf{T}$. For $h \neq h'$ the vectors $\boldsymbol{\delta}_h$ and $\boldsymbol{\delta}_{h'}$, and, therefore, $\mathbf{Y}_h$ and $\mathbf{Y}_{h'}$, are independent.

The general form of the mixed effects model is written as $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{B} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ collects all the n observations of the target variable, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrices for the fixed and random effects. The vector $\boldsymbol{\varepsilon}$ of random errors fulfills $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathbf{VAR}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$, or, more generally, $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathbf{VAR}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V}_e$ with known $\boldsymbol{V}_e$. We will assume that $\boldsymbol{V}_e$ is diagonal. Finally, $\mathbf{B}$ is the vector of random effects with $\mathbf{E}(\mathbf{B}) = \mathbf{0}$, $\mathbf{VAR}(\mathbf{B}) = \boldsymbol{\Sigma}_b$. If all random variables are normally distributed, then the entire vector $\mathbf{Y}$ has a multivariate normal distribution, $\mathbf{Y} \sim \mathcal{N}\big(\boldsymbol{X}\boldsymbol{\beta}\,,\ \boldsymbol{Z}\boldsymbol{\Sigma}_b \boldsymbol{Z}^\mathsf{T} + \sigma^2 \boldsymbol{V}_e\big)$.

This general form encompasses models with several hierarchical factors of random effects (also called variance components) as well as crossed factors, pedigree models, and more.

Remark 1 This form embraces even the general model of geostatistics, where $\boldsymbol{Z} = \boldsymbol{I}$, $\boldsymbol{\Sigma}_b$ has a spatial structure and $\boldsymbol{\varepsilon}$ is called the nugget effect. This forms the basis for a robust procedure for this context, see Papritz et al. (2013). It may also help to obtain robust estimators in time series with respect to the so-called isolated gross errors.

In the simple case (2), $\boldsymbol{\Sigma}_b = \sigma_b^2 \boldsymbol{I}$. More generally, $\boldsymbol{\Sigma}_b$ is often diagonal, and the random effects split into independent groups g of iid components: Let $\mathcal{J}(g)$ be the index set of the random effects composing group g. Then,

$$\mathbf{VAR}\Big(\mathbf{B}_{\mathcal{J}(g)}\Big) = \sigma_g^2 \boldsymbol{I}\ . \tag{3}$$

If the design $\boldsymbol{Z}$ is hierarchical, the variances σ_g^2 are known as the variance components, and if there is only one group, the model again reduces to (2).

Considering deviations from the assumptions of normal distributions, a natural model is to assume long-tailed or "gross error" distributions for the random effects $\mathbf{B}_j$ and the random errors ε_i. (For the general case of correlated $\mathbf{B}_j$, see Sect. 2.) We call this model the "Random Effects Contamination Model." In this model, contamination comes from different sources. Residual error contamination only influences a single observation. Contamination of a random effect has an influence on all the observations that contain it, but the joint distribution of these observations, given the random effect, is not altered.

If the observations $\mathbf{Y}$ split into independent subvectors $\mathbf{Y}_h$, as is the case in (2), then a multivariate contamination of the $\mathbf{Y}_h$ can be assumed. This idea will be called the "Multivariate Contamination Model."

There are several papers on robust estimation of linear mixed models in the literature. An earlier survey can be found in Heritier et al. (2009). Most of the work is based on the assumption of independent subvectors $\mathbf{Y}_h$, i.e., on the Multivariate Contamination Model. A first line of research follows Richardson (1997). Pinheiro et al. (2001) developed a robust mixed model estimator based on multivariate t-distributions of the random effects and the residual errors. They also published an R package called `heavy`. A second line applies high breakdown estimators of multivariate location and scatter to the context of linear mixed models, see Copt and Victoria-Feser (2006), Chervoneva and Vishnyakov (2011), and Chervoneva and Vishnyakov (2014). The notion of high breakdown applies not to single observations but to the independent groups $\mathbf{Y}_h$. Any multivariate observation $\mathbf{Y}_h$ can be contaminated by just a single observation $\mathbf{Y}_{hi}$. As known from the multivariate location and scatter situation, this entails that a small fraction of contaminated observations $\mathbf{Y}_{hi}$ can lead to a breakdown of these methods, see Maronna et al. (2019, Section 6.4.2). A third line of research applies robustness on the single observation level, called the independent contamination model in Alqallaf et al. (2009). The Composite-τ estimators introduced in Agostinelli and Yohai (2016), see also Maronna et al. (2019, Section 6.15.4), are based on the independent contamination model. The method is robust against contamination of both the random errors and effects, but only applicable to balanced datasets.

The method proposed in this paper has been compared in simulation studies in Agostinelli and Yohai (2016) to their method and the one proposed by Copt and Victoria-Feser (2006). Mason et al. (2021) do the same, after evaluating two variants of bootstrap to compute confidence intervals for the estimated parameters.

To the best knowledge of the authors, there is just one proposal based on the random effects contamination model, introduced by Fellner (1986) and followed up by Stahel and Welsh (1997). It has been introduced and studied only in the multi group situation (1) and is limited to a diagonal covariance matrix of the random effects. The present paper generalizes this approach.

This paper is an excerpt of Manuel Koller's dissertation (Koller 2013). The simulation studies presented in Sects. 4.2 and 4.4 are new. A more accessible introduction to the method presented here together with practical instructions on tuning the method has been published in Koller (2016).

The paper is organized as follows. The classical estimating equations for the general mixed effects model are reviewed in the next section. A robust version is introduced in Sect. 3, and its properties are evaluated in Sect. 4. We analyze two examples in Sect. 5 and conclude with Sect. 6.

2 The Model and Classical Estimation

The *linear mixed effects model* is, as introduced above, $\mathbf{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}\mathbf{B} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}\big(\mathbf{0}, \sigma^2 \boldsymbol{V}_e\big)$, where $\boldsymbol{V}_e$ is assumed to be diagonal and known a priori. For later convenience, we write the covariance matrix of $\mathbf{B}$ as $\sigma^2 \boldsymbol{V}_b(\boldsymbol{\theta})$, $\mathbf{B} \sim \mathcal{N}\big(\mathbf{0}, \sigma^2 \boldsymbol{V}_b(\boldsymbol{\theta})\big)$. $\mathbf{B}$ shall be independent of $\boldsymbol{\varepsilon}$. The notation $\boldsymbol{V}_b(\boldsymbol{\theta})$ reflects that $\boldsymbol{V}_b$ depends on parameters $\boldsymbol{\theta}$. In the case (3) of grouped random effects, $\sigma^2\theta_g^2 = \sigma_g^2$ is the variance of the group $\mathbf{B}_{\mathcal{J}(g)}$ of the random effects. For most of our development, we will assume that $\boldsymbol{V}_b(\boldsymbol{\theta})$ is diagonal or block diagonal.

As convention for indices, we use

$i = 1, \ldots, n$ for observations Y_i.
$j = 1, \ldots, J$ for random effects $\mathbf{B}_j$.
$k = 1, \ldots, K$ for diagonal blocks in $\boldsymbol{V}_b(\boldsymbol{\theta})$ of size $m(k)$; if $\boldsymbol{V}_b(\boldsymbol{\theta})$ is diagonal, then $k = j$.
$\ell = 1, \ldots, L$ for the covariance parameters θ_ℓ, and
$g = 1, \ldots, G$ for any independent groups of random effects, with index sets $\mathcal{J}(g)$ and $\mathbf{VAR}\Big(\mathbf{B}_{\mathcal{J}(g)}\Big) = \sigma_g^2 \boldsymbol{I}$ (see (3)); if $\mathbf{B}$ only consists of such groups and $\boldsymbol{V}_b$ is diagonal, then $g = \ell$.

Following Bates (2010), we use an alternative formulation of the model based on spherical random effects $\mathbf{B}^*$. They are defined by the relation $\mathbf{B} = \boldsymbol{U}_b(\boldsymbol{\theta})\mathbf{B}^*$, where $\boldsymbol{U}_b(\boldsymbol{\theta})$ is the lower triangular Cholesky factor of $\boldsymbol{V}_b(\boldsymbol{\theta})$. This avoids numerical problems with vanishing variance components. The model then is $\mathbf{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}\boldsymbol{U}_b(\boldsymbol{\theta})\mathbf{B}^* + \boldsymbol{U}_e\boldsymbol{\varepsilon}^*$, $\mathbf{B}^* \sim \mathcal{N}\big(\mathbf{0}, \sigma^2\boldsymbol{I}_q\big)$, $\boldsymbol{\varepsilon}^* \sim \mathcal{N}\big(\mathbf{0}, \sigma^2\boldsymbol{I}_n\big)$, $\mathbf{B}^* \perp \boldsymbol{\varepsilon}^*$. Here, we have also replaced $\boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon}^* = \boldsymbol{U}_e^{-1}\boldsymbol{\varepsilon}$, where $\boldsymbol{V}_e = \boldsymbol{U}_e\boldsymbol{U}_e^\intercal$.

Next, we need a suitable form of the likelihood. To be able to separate the random components as discussed in the Introduction, we consider the likelihood treating the random effects as observed and insert the best linear unbiased predictor (BLUP) of the random effects (for a derivation, see Searle et al. 1992, Chapter 7). This gives a pseudo-likelihood

$$\widehat{d}(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma, \mathbf{b}^*|\mathbf{y}) = n\log 2\pi + \log\Big|\sigma^2\big[\boldsymbol{Z}\boldsymbol{U}_b(\boldsymbol{\theta})(\boldsymbol{Z}\boldsymbol{U}_b(\boldsymbol{\theta}))^\intercal + \boldsymbol{V}_e\big]\Big| + \big(\boldsymbol{\varepsilon}^*(\boldsymbol{\beta}, \mathbf{b}^*)^\intercal\boldsymbol{\varepsilon}^*(\boldsymbol{\beta}, \mathbf{b}^*) + \mathbf{b}^{*\intercal}\mathbf{b}^*\big)/\sigma^2\,, \tag{4}$$

where $\boldsymbol{\varepsilon}^*(\boldsymbol{\beta}, \mathbf{b}^*) = \boldsymbol{U}_e^{-1}(\mathbf{y} - \boldsymbol{X\beta} - \boldsymbol{Z}\boldsymbol{U}_b(\boldsymbol{\theta})\mathbf{b}^*)$. We drop the dependency of $\boldsymbol{U}_b(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ in our notation from now on and only write $\boldsymbol{U}_b$ instead.

We will robustify on the level of estimating equations. To get them, we need the partial derivatives of (4). The MLE results from equating them to zero. It is well known that this leads to biased estimates for σ^2 and $\sigma^2 \boldsymbol{V}_b(\boldsymbol{\theta})$. The restricted maximum likelihood (REML) estimates are obtained by equating partial derivatives to their expected values instead of 0 (Stahel & Welsh 1997). After some more rewriting this yields

$$\begin{aligned}
\boldsymbol{X}^{\mathsf{T}} \boldsymbol{U}_e^{-\mathsf{T}} \boldsymbol{\varepsilon}^*(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}^*)/\widehat{\sigma} &= \mathbf{0} \,, \\
\Big(\boldsymbol{U}_b^{\mathsf{T}} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{U}_e^{-\mathsf{T}} \boldsymbol{\varepsilon}^*(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}^*) - \widehat{\mathbf{b}}^*\Big)/\widehat{\sigma} &= \mathbf{0} \,, \\
\widehat{\boldsymbol{\varepsilon}}^{*\mathsf{T}} \widehat{\boldsymbol{\varepsilon}}^* &= \mathbf{E}_0\Big[\widehat{\boldsymbol{\varepsilon}}^{*\mathsf{T}} \widehat{\boldsymbol{\varepsilon}}^*\Big] \,, && (5) \\
\widehat{\mathbf{b}}^{*\mathsf{T}} \boldsymbol{Q}_\ell(\widehat{\boldsymbol{\theta}}) \widehat{\mathbf{b}}^* &= \operatorname{tr}\!\Big(\mathbf{E}_0\Big[\widehat{\mathbf{b}}^* \widehat{\mathbf{b}}^{*\mathsf{T}}\Big] \boldsymbol{Q}_\ell(\widehat{\boldsymbol{\theta}})\Big) \qquad \text{for } \ell = 1, \dots, L \,, && (6)
\end{aligned}$$

where

$$\boldsymbol{Q}_\ell(\boldsymbol{\theta}) = \boldsymbol{U}_b(\boldsymbol{\theta})^{-1} \frac{\partial \boldsymbol{U}_b(\boldsymbol{\theta})}{\partial \theta_\ell} \,,$$

and $\widehat{\boldsymbol{\varepsilon}}^* = \boldsymbol{\varepsilon}^*(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}^*)$. $\mathbf{E}_0$ denotes expectations with respect to the standard normal distribution of ε^* or b^*. These expectations are computed using the linear approximations developed in Appendix.

Note that in the diagonal case with the iid group structure (3), $\boldsymbol{Q}_\ell(\boldsymbol{\theta})$ reduces to ones and zeroes and, therefore, each instance ℓ of the fourth equation reduces to one that looks like the third,

$$\widehat{\mathbf{b}}^{*\mathsf{T}}_{\mathcal{J}(g)} \widehat{\mathbf{b}}^*_{\mathcal{J}(g)} = \mathbf{E}\Big[\widehat{\mathbf{b}}^{*\mathsf{T}}_{\mathcal{J}(g)} \widehat{\mathbf{b}}^*_{\mathcal{J}(g)}\Big] \,.$$

3 The Robust Scoring Equations Estimator

We will first robustify the above estimating equations assuming a diagonal covariance matrix $\boldsymbol{V}_b(\boldsymbol{\theta})$ of the random effects. In the second step, we will relax this assumption.

3.1 Estimation for Diagonal V_b

To robustify the first two estimating equations, we replace the spherical random effects and the residuals by a bounded function ψ of themselves. According to our assumptions, all the terms are independent and we apply the ψ-function componentwise, using the notation $\boldsymbol{\psi}(\mathbf{x}) = [\psi(x_1), \dots, \psi(x_i), \dots]^{\mathsf{T}}$. The first two estimating equations then read

$$\boldsymbol{X}^{\mathsf{T}} \boldsymbol{U}_e^{-\mathsf{T}} \boldsymbol{\psi}_e\big(\widehat{\boldsymbol{\varepsilon}}^*/\widehat{\sigma}\big) = \mathbf{0} \,, \tag{7}$$

$$\boldsymbol{U}_b^{\mathsf{T}} \boldsymbol{Z}^{\mathsf{T}} \boldsymbol{U}_e^{-\mathsf{T}} \boldsymbol{\psi}_e\big(\widehat{\boldsymbol{\varepsilon}}^*/\widehat{\sigma}\big)/\lambda_e - \boldsymbol{\psi}_b\big(\widehat{\mathbf{b}}^*/\widehat{\sigma}\big)/\lambda_b = \mathbf{0} \,, \tag{8}$$

where $\lambda_e = \mathbf{E}_0[\psi_e']$ and $\lambda_b = \mathbf{E}_0[\psi_b']$, $\mathbf{E}_o$ denoting the expectation for the central model with scale 1—the standard normal distribution. These factors are introduced to balance the $\widehat{\boldsymbol{\varepsilon}}^*$ and $\widehat{\mathbf{b}}^*$ terms in case different ψ-functions are used, by making the influences of the two terms approximately independent of the choice ψ-functions and thus their ratio equal to its value in the classical case (for more detail, see Koller 2013, p.32). If there are groups g of random effects, different ψ_b-functions can be used for different groups, and this leads to a straightforward generalization of (8) covered by the formulas of the next subsection.

For the estimating equations for σ and $\boldsymbol{\theta}$, we apply the basic idea of the Design Adaptive Scale (DAS) approach described in Koller and Stahel (2011). Its starting point is the well-known fact that the variances of the residuals are smaller than the error variance and they vary individually according to the design. Let the covariance matrix of the residuals $\widehat{\boldsymbol{\varepsilon}}^*$ (for given σ and $\boldsymbol{\theta}$) be $\sigma^2 \boldsymbol{V}_e^*$, and analogously, $\mathbf{VAR}(\widehat{\mathbf{b}}^*) = \sigma^2 \boldsymbol{V}_b^*$. Approximations to $\boldsymbol{V}_e^*$ and $\boldsymbol{V}_b^*$ are given in Appendix. The idea of the DAS approach is to focus on standardized residuals $\widehat{\varepsilon}_i^*/(\tau_{e,i}\sigma)$ with $\tau_{e,i}^2 = \boldsymbol{V}_{e,ii}^*$ and to write (5) as $\sum_i \tau_{e,i}^2[(\widehat{\varepsilon}_i^*/(\tau_{e,i}\sigma))^2 - 1] = 0$. Robust estimation is achieved by introducing robustness weights of standardized residuals in the form

$$\sum_i \tau_{e,i}^2 w^{(\sigma)}(\widehat{\varepsilon}_i^* / (\tau_{e,i}\widehat{\sigma}))\Big[(\widehat{\varepsilon}_i^* / (\tau_{e,i}\widehat{\sigma}))^2 - \kappa_\sigma\Big] = 0 \tag{9}$$

$$\kappa_\sigma = \mathbf{E}_0\Big[w^{(\sigma)}(z) z^2\Big] \Big/ \mathbf{E}_0\Big[w^{(\sigma)}(z)\Big] ,$$

where $w^{(\sigma)}$ is a weighting function. We discuss choices of weighting functions in Sect. 3.4. Analogously, we write for the estimation of $\boldsymbol{\theta}$ in the case of diagonal $\boldsymbol{V}_b$

$$\sum_j \tau_{b,j}^2 w^{(\ell)}\Big(\widehat{b}_j^* / (\tau_{b,j}\widehat{\sigma})\Big) Q_{\ell,jj}(\widehat{\theta}) \left[\Big(\widehat{b}_j^* / (\tau_{b,j}\widehat{\sigma})\Big)^2 - \kappa_\ell\right] = 0 , \qquad \ell = 1, \ldots, L , \tag{10}$$

where $\tau_{b,j}^2 = \boldsymbol{V}_{b,jj}^*$ and κ_ℓ is determined as κ_σ is.

The more sophisticated version of the DAS approach determines the scaling factor $\tau_{e,i}$ by zeroing the expectation of each term of the sum in (9) in a more precise fashion than is achieved by standardizing the $\widehat{\varepsilon}_i^*$. To this end, $\widehat{\varepsilon}_i^*$ is split according to (15) into a term containing ε_i^* and a remainder R_i,

$$\begin{aligned}\widehat{\varepsilon}_i^*/\sigma &\approx \varepsilon_i^*/\sigma - (\boldsymbol{A}_{ee})_{ii}\psi_e\big(\varepsilon_i^*/\sigma\big) - R_i \\ R_i &= \sum_{h \neq i} (\boldsymbol{A}_{ee})_{ih}\psi_e\big(\varepsilon_h^*/\sigma\big) + \sum_j (\boldsymbol{A}_{eb})_{ij}\psi_b\Big(b_j^*/\sigma\Big) .\end{aligned}$$

The remainder R_i is independent of ε_i^* and has approximately a normal distribution, the variance of which is obtained in a straightforward manner. Then, τ_i is determined by calculating expectations based on integration over the distribution of ε_i^*/σ and R_i. The idea also applies to the estimation of $\boldsymbol{\theta}$. For details, see Appendix.

The method proposed in this paper is defined as the simultaneous solution of Eqs. (7) to (10). Since these are robustified versions of the classical scoring equations, we call the method the Robust Scoring Equations (RSE) estimator.

3.2 Estimation for Block Diagonal $\boldsymbol{V}_b$

If the covariance matrix $\boldsymbol{V}_b(\boldsymbol{\theta})$ of the random effects is block diagonal instead of completely diagonal, bounding and weighting is based on a multivariate view of each block. If $\mathbf{b}_k$ denotes the sub-vector corresponding to the kth diagonal block $\boldsymbol{V}_k$ of $\boldsymbol{V}_b$, the bounding and weighting functions shall depend on the (squared) Mahalanobis norm of $\mathbf{b}_k$ or, equivalently, on the (scaled) norm of $\mathbf{b}_k^*$. Sensible weighting will depend on the dimension $m(k)$ of $\mathbf{b}_k^*$. We, therefore, choose a weighting function $w_{b,m}$ and let

$$\begin{aligned}\boldsymbol{\psi}_k(\mathbf{z}_k) &= w_k\big(\mathbf{z}_k^\top \mathbf{z}_k\big)\mathbf{z}_k\ , \quad \mathbf{z}_k \in R^{m(k)}\\ \boldsymbol{\psi}_b(\mathbf{z}) &= (\psi_k(\mathbf{z}_k))_{k=1,\ldots,K}\,,\end{aligned}$$

where w_k usually depends on k only through $m(k)$, $w_k = w_{b,m(k)}$. The second estimating equation (8) is

$$\boldsymbol{U}_b^\top \mathbf{Z}^\top \boldsymbol{U}_e^{-\top} \boldsymbol{\psi}_e(\widehat{\boldsymbol{\varepsilon}}^*/\sigma)/\lambda_e - \boldsymbol{\Lambda}_b^{-1}\boldsymbol{\psi}_b\big(\mathbf{b}^*/\sigma\big) = \mathbf{0}\ . \tag{11}$$

Here, $\boldsymbol{\Lambda}_b$ is a diagonal matrix with elements depending on the functions $w_{k(j)}$ through the block size $m_{k(j)}$,

$$\begin{aligned}\boldsymbol{\Lambda}_b &= \mathbf{Diag}\big(\lambda_b\big(k(j), m_{k(j)}\big)\big)_{j=1,\ldots,J}\\ \lambda_b(k,m) &= \mathbf{E}_{0,m}\big[2w_k'(u)u\big]\big/m + \mathbf{E}_{0,m}[w_k(u)]\ ,\end{aligned}$$

where $\mathbf{E}_{0,m}$ denotes the expectation over the χ_m^2 distribution.

The fourth equation determines the estimated covariance matrix of the blocks of random effects $\mathbf{b}_k$. Equivariant M-estimation of a covariance matrix is, following Stahel (1987), based on choosing two weighting functions $w^{(\eta)}$ and $w^{(\tau)}$, which determine the influences on the shape and the size of the covariance matrix, respectively. It is convenient to introduce a third, derived, weight function $w^{(\delta)}$ to simplify notation in the estimating equation below,

$$w^{(\delta)}(u) = \Big(uw^{(\eta)}(u) - \Big(u - m\kappa_b^{(\tau)}\Big)w^{(\tau)}\Big(u - m\kappa_b^{(\tau)}\Big)\Big)\Big/m\ ,$$

where $\kappa_b^{(\tau)}$ fulfills

$$\mathbf{E}\Big[\Big(u - m\kappa_b^{(\tau)}\Big)w^{(\tau)}\Big(u - m\kappa_b^{(\tau)}\Big)\Big] = 0 \quad \text{for } u \sim \chi_m^2\ .$$

Then, the estimating equation for a covariance matrix $\boldsymbol{\Sigma}$ is

$$\sum_i w^{(\eta)}\Big(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}\mathbf{x}_i\Big)\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}} = \boldsymbol{\Sigma} \sum_i w^{(\delta)}\Big(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}\mathbf{x}_i\Big) \; .$$

Alternatively, one can search for the (lower triangular) standardization matrix $\boldsymbol{U}$ such that

$$\sum_i w^{(\eta)}(\mathbf{z}_i^{\mathsf{T}}\mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}} = \boldsymbol{I}\sum_i w^{(\delta)}(\mathbf{z}_i^{\mathsf{T}}\mathbf{z}_i) \; ,$$

with $z_i = \boldsymbol{U}x_i$, and then estimate $\boldsymbol{\Sigma}$ by $\boldsymbol{U}^{-1}\boldsymbol{U}^{-\mathsf{T}}$.

Using this idea, we turn to the estimation of the covariance matrix $\sigma^2\boldsymbol{V}_k$ of $\mathbf{B}_k$, which can be formulated as determining $\boldsymbol{U}_k$ such that $\mathbf{VAR}\big(\mathbf{B}_k^*\big) = \boldsymbol{I}$. We again propose to use the covariance structure $\boldsymbol{V}_b^*$ of the estimated random effects $\widehat{\mathbf{b}}^*$ to standardize them. This leads to the robustified fourth equation

$$\sum_k \Big[w_{m(k)}^{(\eta)}\Big(\|\boldsymbol{T}_k^{-1}\widehat{\mathbf{b}}_k^*/\widehat{\sigma}\|^2\Big)\widehat{\mathbf{b}}_k^{*\mathsf{T}}\, \boldsymbol{Q}_{\ell,k}(\widehat{\boldsymbol{\theta}})\widehat{\mathbf{b}}_k^*/\widehat{\sigma}^2 - w_{m(k)}^{(\delta)}\Big(\|\boldsymbol{T}_k^{-1}\widehat{\mathbf{b}}_k^*/\widehat{\sigma}\|^2\Big)\,\mathrm{tr}\big(\boldsymbol{V}_k^*\,\boldsymbol{Q}_{\ell,k}(\widehat{\boldsymbol{\theta}})\big)\Big] = 0 \; , \tag{12}$$

where $\boldsymbol{Q}_{\ell,k}(\widehat{\boldsymbol{\theta}})$ is the $m(k) \times m(k)$ submatrix of $\boldsymbol{Q}_\ell(\widehat{\boldsymbol{\theta}})$ which acts on block k and $\boldsymbol{T}_k$ is a square root of $\boldsymbol{V}_k^* = \mathbf{E}_{0,m}\big[\widehat{\mathbf{b}}_k^*\widehat{\mathbf{b}}_k^{*\mathsf{T}}\big]$.

The refined DAS approach can also be extended to this case, see Appendix.

Remark 2 In the classical case, the linear approximations for $\widehat{\mathbf{b}}^*$ and $\widehat{\boldsymbol{\varepsilon}}^*$ are exact, and the estimating equations (9, 12) reduce to the REML estimating equations (5, 6).

Remark 3 In the case of general $\boldsymbol{\Sigma}_b$ in the sense of Remark 1, robustification of the random effects and the respective parameters cannot be achieved along these lines, since there are no independent replications of parts of $\mathbf{B}$. Robust methods would need to rely on formalizing a notion of vicinity in space or time.

3.3 Computation

Solutions to these equations are best found by using a nested iterative reweighting algorithm. The outer loop is updating $\widehat{\boldsymbol{\theta}}$ until it converges. For each new value of $\widehat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{b}}^*$ and then $\widehat{\sigma}$ are updated. A complete description of the algorithm can be found in Koller (2013, Section 3.1.3 and Section 3.2.3).

Initial estimates are required to start the above procedure. For bounded, monotone ψ-functions, the solution can be expected to be unique aside from pathological, easily discarded solutions, whence starting from the classical estimates is fine. To get a high breakdown point, one would have to use redescending ψ-functions and a suitable high breakdown initial estimate. The authors do not know of any such estimator for the general model. It would be interesting to explore possibilities of generalizing the idea of S-estimation to the setup of estimating more than one scale parameter (variance component) in the framework presented here.

The methods are implemented in the R package `robustlmm` (Koller 2016).

3.4 Choices of ψ and w

A simple and reliable choice of a ψ-function is what we call the smoothed Huber function. It is defined as

$$\psi(x; c, s) = \begin{cases} x & |x| \le c_0 \\ \operatorname{sign}(x)\left(c - \frac{1}{(|x| - c_1)^s}\right) & \text{otherwise} \end{cases},$$

where $c_0 = c - s^{-s/(s+1)}$ and $c_1 = c_0 - s^{1/(s+1)}$. Compared to the regular Huber ψ-function, it has a smooth transition from the ascending linear to the flat part. (We use $s = 10$ throughout this paper.) This modification of the classical Huber function is meant to avoid numerical problems.

We will now discuss how to choose the tuning constant c for the ψ—and w-functions for each estimating equation in turn.

For ψ_e used in (7), the tuning parameter c can be chosen as a suitable quantile of the standard normal distribution that determines the expected proportion of trimmed residual errors, or, as we prefer, according to efficiency. To simplify the computation of asymptotic efficiency, one may consider the estimating equation (7) on its own as a simple M-estimator of regression with known scale. Formulas to compute asymptotic efficiencies and tables for a range of choices of c can be found in Koller (2013, Appendix B.2.1 and B.3) and Koller (2016, Appendix A). For the Huber and the smoothed Huber ψ-function, 95% asymptotic efficiency is reached for $c = 1.345$.

For $w^{(\sigma)}$ used in (9), maximum likelihood for long-tailed distributions of errors and random effects would lead to $w^{(\sigma)}(x) = \psi_e(x)/x$. For monotone ψ, this leads to unbounded $w^{(\sigma)}(x)x^2$ and consequently to unbounded influences on $\widehat{\sigma}$. We recommend to use the squared robustness weights instead, $w^{(\sigma)}(x) = (\psi_e(x)/x)^2$, just as in Huber's Proposal 2 for estimating location and scale. The squaring comes with an efficiency loss, so one has to use a larger tuning parameter in order to reach the same efficiency as with regular robustness weights. Similar to the case for ψ_e considered above, one may treat the estimating equation as simple M-estimator of scale in order to approximate the efficiency. For the Huber and the smoothed Huber ψ-function, 95% asymptotic efficiency is reached for $c = 2.28$. For increased robustness, one may also use the same tuning parameter as for ψ_e, but this would lead to an asymptotic efficiency of 71%.

For ψ_b used in (8) or (11), the choice of c depends on whether the covariance matrix $\boldsymbol{V}_b(\boldsymbol{\theta})$ is diagonal or block diagonal. In the diagonal case, one may use the same tuning parameter considerations as for ψ_e. We generally use the same tuning parameter as for ψ_e. In the block-diagonal case, the (squared) Mahalanobis norm is used as an argument, which requires considerably larger tuning parameters that also depend on the size of each block. Using squared robustness weights is not necessary in the block-diagonal case. We prefer to choose the tuning parameter according to the efficiency of the weight function $w^{(\tau)}$ that controls the size of the covariance matrix. With $c = 5.14$, this efficiency is about 95%.

For $w^{(\ell)}$ used in (10) in the diagonal case, the same considerations as for $w^{(\sigma)}$ apply. We prefer to use the same tuning parameter as for $w^{(\sigma)}$.

For $w^{(\eta)}$ and $w^{(\tau)}$ used in (12), one would technically have to use different tuning parameters to achieve about the same efficiency. But we prefer to keep things simple and use the same tuning parameter as for ψ_b for both of them as well.

3.5 Robust Tests

For testing fixed effects in mixed effects models, exact tests are only known for some special data structures even in the classical setup. For the general case, only approximating methods are available. For the robust RSE estimator proposed here, an approximate covariance matrix $\boldsymbol{V}_\beta = \mathbf{VAR}\big(\widehat{\boldsymbol{\beta}}\big)$ of the estimated fixed effects is given by (17) in Appendix. This leads in the usual way to a Wald type test with test statistic

$$T_W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\mathsf{T} \widehat{\boldsymbol{V}_\beta}^{-1} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approx\sim \chi_p^2 \,,$$

where p is the dimension of $\boldsymbol{\beta}$, and to an approximate confidence region of $\Big\{\boldsymbol{\beta} \mid (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\mathsf{T} \widehat{\boldsymbol{V}_\beta}^{-1} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \le q\Big\}$, where q is the suitable quantile of χ_p^2.

Testing for variance components against 0 is difficult in the given setting, where only estimating equations are used and no likelihood is available. Note that testing for zero variance components is inherently difficult already in the classical case as the null hypothesis is on the boundary of the parameter space.

The wild bootstrap provides a general tool for constructing tests and confidence intervals. It has been applied to the linear mixed model by Mason et al. (2021). Since the bootstrap samples are by construction contamination-free, one can use the classical method to fit the model for the bootstrap samples. Mason et al. (2021) have found that the confidence intervals computed using wild bootstrap for the RSE method proposed here have an equal coverage to those computed using parametric or wild bootstrap for the classical method when there is no contamination. In case of contamination, the confidence interval coverage is affected more for the classical method than for the RSE estimator.

4 Properties of the Robust Scoring Equations Estimator

Robust estimators should show limited influence of single observations or small subgroups and be reasonably unbiased and efficient at the central model as well as less biased and more efficient than the classical method for contaminated distributions. We examine these issues in turn. All the R code needed to replicate the simulation studies is available in Koller (2022).

The following subsections feature two differently tuned versions of the proposed RSE method as well as up to three other estimators for comparison. We introduce them here, including the tuning parameters used, so that cross-comparisons between subsections are easily possible. The estimators are

lme the classical Restricted Maximum Likelihood estimator as implemented in the `lme4` R package by Bates et al. (2015).

RSEa the robust scoring equations (RSE) estimator with smoothed Huber ψ and weighting functions. We adjust $w^{(\sigma)}$ and $w^{(1)}$ to obtain the same efficiency for $\widehat{\sigma}$ and $\widehat{\theta}$ as for the effects. The tuning parameters are $c = 1.345$ for ψ_e and $c = 2.28$ for $w^{(\sigma)}$. In the diagonal case (Sects. 4.1 and 4.2), we use $c = 1.345$ for ψ_b and $c = 2.28$ for $w^{(1)}$. In the block-diagonal case (Sect. 4.4), we use $c = 5.14$ for ψ_b, $w^{(\eta)}$ and $w^{(\tau)}$. We use squared robustness weights for $w^{(\sigma)}$ and $w^{(1)}$, the latter in the diagonal case only.

RSEn the same method, without efficiency adjustment, that is, using $c = 1.345$ also for the weighting functions $w^{(\sigma)}$ and $w^{(1)}$ in the diagonal case and analogue in the block-diagonal case.

cTau the Composite-τ estimator of Agostinelli and Yohai (2016) as implemented in the `robustvarComp` R package by Agostinelli and Yohai (2019), using the *optimal* ψ-function with tuning parameters $c_1 = 1$ and $c_2 = 1.643168$. As initial estimator we use `covOGK`.

S the S-estimator of Copt and Victoria-Feser (2006), using the implementation in the `robustvarComp` R package using the *Rocke* ψ-function with asymptotic rejection probability set to 0.01. As initial estimator we use `covOGK`.

4.1 Sensitivity Curves

Sensitivity curves give intuitive, valuable insight into the way a method achieves robustness.

For one-way ANOVA-type datasets using model (1), there are three ways of obtaining sensitivity curves:

(a) Changing the response of a single observation
(b) Moving the responses of a whole group, i.e., changing the random effect corresponding to this group

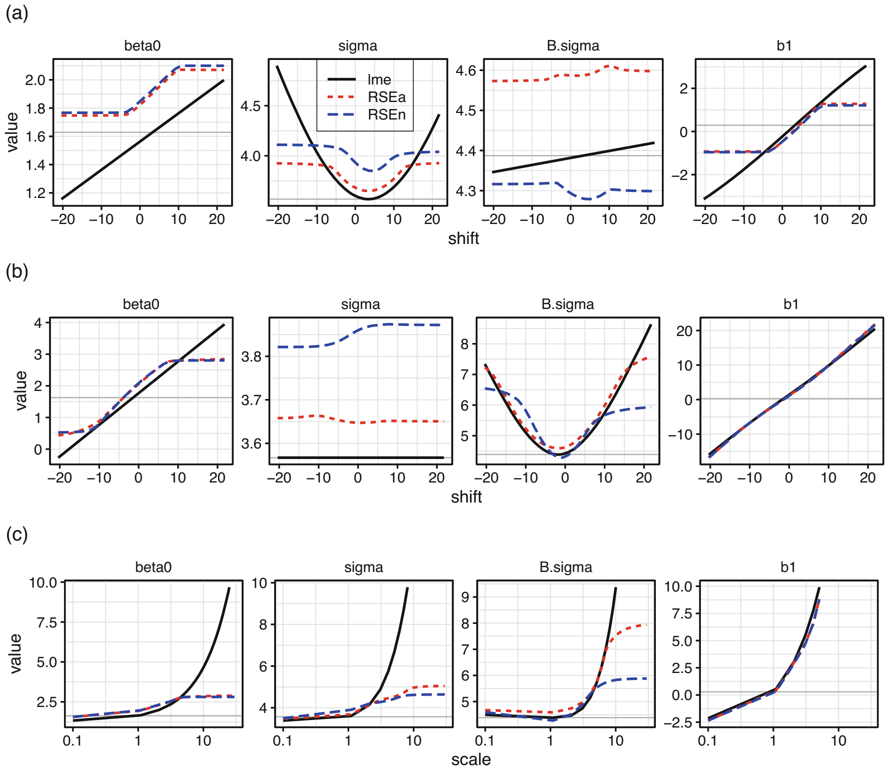


Fig. 1 Sensitivity curves for a balanced one-way dataset with 10 groups of 5 observations. The gray lines indicate the true values. Explanation of (**a**) Shift of first observation. (**b**) Shift of first group. (**c**) Scaling of first group, see text

(c) Changing the spread of the observations of a group around their expectation given the (simulated) random effect

For a randomly generated dataset of 10 groups with 5 observations each, the three sensitivity curves are given in Fig. 1. The values of the three estimators for the unchanged sample –zero shift or unit scale– are naturally different, and only the shapes of the curves should be interpreted.

The robust estimators show the flattening of the curves towards extreme changes, except for the estimated random effect b_1 when the whole first group is shifted or scaled. As is to be expected, the efficiency adjusted version, which uses a larger tuning constant for $w^{(1)}$, is more sensitive to the changes than the unadjusted version.

4.2 Efficiency and Robustness, Diagonal Case

We examine properties for a completely crossed design with a fixed effects factor of 3 levels, a random effects factor of 5 levels, and 10 observations for each combination, leading to 150 observations in total. The true values were 1 for the intercept (β_0) and for the group contrasts (β_1, β_2), 4 for σ and 1 for θ, where $\sigma_b = \sigma\theta$ is the standard deviation of the random effects. For $1,000$ replicates, we compute the estimators listed in Sect. 4. The methods proposed by Pinheiro et al. (2001) using heavy-tailed distributions (`heavyLme`) and by Geraci and Bottai (2014) using a linear quantile mixed model (`lqmm`) were also computed but showed poor performance here and for the block-diagonal case in the next subsection.

For the error and random effects distributions, we used the appropriately scaled versions of

N the standard normal distribution
CN a "fixed mixture" or "contamponent" of 90% standard normal samples with 10% from a shifted normal, $\mathcal{N}(4, 1)$—the "contaminated normal" distribution
t3 the t distribution with 3 degrees of freedom
skt3 the skewed t distribution with 3 degrees of freedom and skewing parameter $\gamma = 2$ as introduced by Fernández and Steel (1998). We use the implementation in the R package `skewt` by King and Anderson (2021)

The CN and skt3 distributions were centered by shifting them such that the Huber Proposal 2 location functional, with tuning constant $c = 1.345$, becomes zero for the shifted distribution. These and the t3 were scaled in order to bring the Proposal 2 scale to one.

An analogous simulation has been carried out for the one-way model considered in the last Sect. 4.1. Since the results have been very similar, they are not reported here.

Figure 2 shows the results for selected combinations of distributions of the random errors and effects. The results show that

1. All estimators perform very similarly to the classical estimator for the normal (N/N) model and also if the random effects alone are contaminated (N/CN). The two scenarios show the largest differences for σ_b. The efficiency adjusted RSE estimator `RSEa` is closest to the classical estimator `lme`, whereas the unadjusted RSE estimator `RSEn` is even slightly closer to the true value than `lme`. The Composite-τ estimator `cTau` and the S-estimator are biased downwards for σ_b in both the normal model and also if the random effects alone are contaminated. `cTau` consistently produces the smallest average estimate for σ_b over all scenarios.
2. The overall performance for the two versions of the RSE estimator and `cTau` is quite comparable, but the S-estimator is less robust. For σ, the S-estimator's bias is very similar to `lme`'s. The S-estimator loses a little less efficiency than `lme` for the regular and skewed t_3 scenarios, but clearly more than the other estimators.

3. The RSE estimator is more efficient than the others for β_1 and β_2. Both cTau and S have difficulty with contaminated normal errors (CN/N and CN/CN), their efficiency loss is almost the same as for lme.
4. The efficiency adjustment for RSEa leads to an efficiency gain and loss of degree of robustness for σ and σ_b that is to be expected.

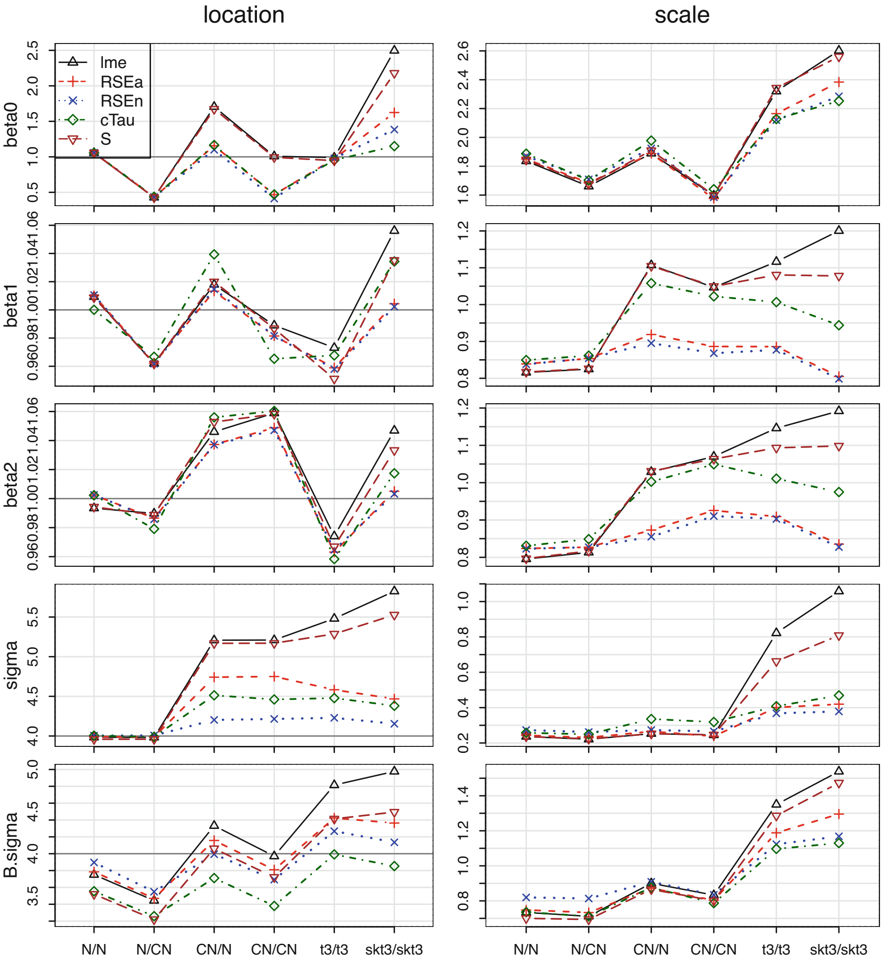


Fig. 2 Simulation results for the diagonal case. The left column shows a robust location estimate of the simulated estimates for the diverse methods and the five parameters β_0, β_1, β_2, σ, σ_b. The true values are indicated by gray horizontal lines. Deviations from them thus represent biases. The right column shows robust scale parameters –measures of simulated standard errors– in an analogous way

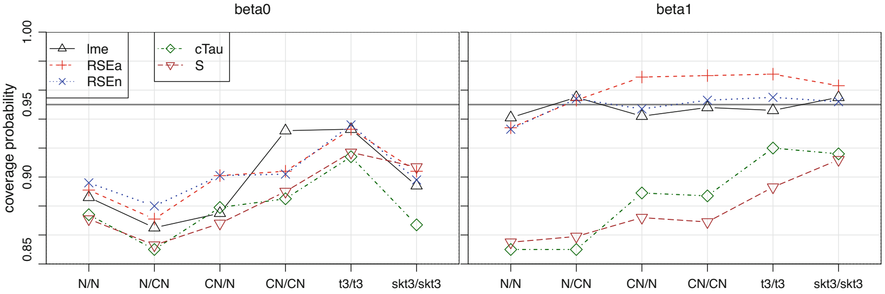


Fig. 3 Simulation results for the diagonal case showing empirical coverage probabilities for the intercept β_0 (left) and β_1 (right). The expected level of 0.95 is shown by a gray line

4.3 Coverage Probabilities, Diagonal Case

In order to examine the validity of tests and confidence intervals for the fixed effects, coverage probabilities for the latter were simulated for the preceding model. The results for β_0 and β_1 are shown in Fig. 3. (Results for β_2 are very similar to those for β_1.)

For the normal model (N/N) as well as contaminated effects (N/CN), the RSE method performs very similarly to the classical `lme`. For other distributions of the random error term, the method with efficiency adjustment of the scale parameter shows conservative coverage, whereas the version using the unadjusted tuning constant performs well. The results for `cTau` and `S` are clearly worse. The block-diagonal case discussed next produced similar results, except for `cTau`, which performed about the same as `lme` and RSE.

4.4 Efficiency and Robustness, Block-Diagonal Case

For the block-diagonal case, the model of the Sleep Study example (see Sect. 5.2) forms the basis of simulations. In this example, 18 subjects (blocks k) produce Y values for 10 time points x. The model for the blocks is a simple linear dependency of the target variable on time, with random intercept and slope,

$$\mathbf{Y}_k = \beta_0 + \beta_1 \mathbf{x} + B_{0,k} + B_{1,k}\mathbf{x} + \boldsymbol{\varepsilon}_h \ , \qquad \mathbf{x} = [1, 2, \ldots, 10]^\mathsf{T} \ ,$$

$$\boldsymbol{U}_k = \begin{bmatrix} \theta_1 & 0 \\ \theta_2 & \theta_3 \end{bmatrix}, \qquad \boldsymbol{V}_k = \mathbf{VAR}\big([B_{0,k}, B_{1,k}]^\mathsf{T}\big)/\sigma^2 = \begin{bmatrix} \theta_1^2 & \theta_1\theta_2 \\ \theta_1\theta_2 & \theta_2^2 + \theta_3^2 \end{bmatrix}. \tag{13}$$

Simulations consisted on 1, 000 replicates, setting the classically estimated values as true parameters. The same distributions and estimators as in the diagonal case are examined, using the ψ—and w-functions described in Sect. 4.

The results, shown in Fig. 4, show that the first two conclusions of the diagonal case are confirmed. The slight advantage of the RSE estimators over the Composite-τ disappears as well as the efficiency gain of the non-adjusted version.

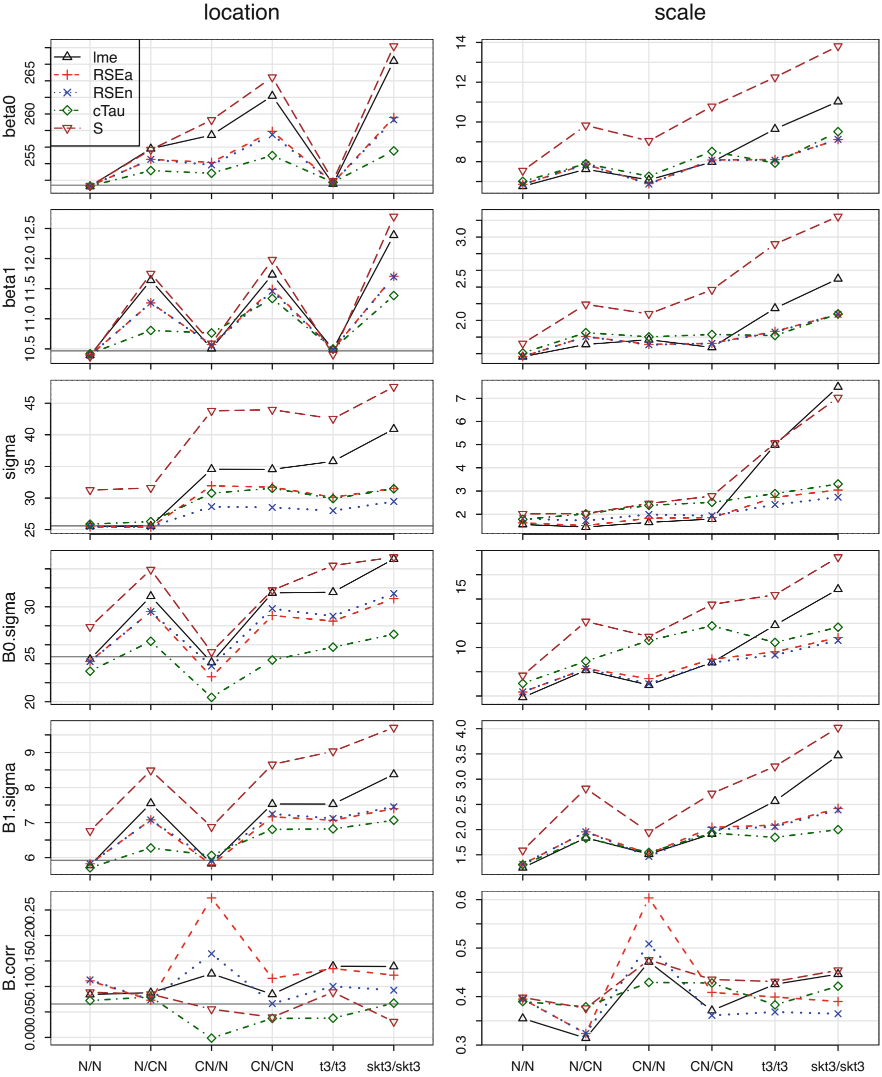


Fig. 4 Simulation results for the block-diagonal case, shown as in Fig. 2

The results for the correlation `B.corr` between the random intercept and slope with the combination of contaminated random errors and normal random effects (CN/N) are surprising. A closer look at the distributions of the estimated correlation in Fig. 5 (last row, 3rd column) reveals that the value is often 1, and in fact, the algorithm then often converges to a phony solution, for which the estimating equations are not fulfilled. This also happens for other combinations of distributions and notably also to the classical `lme` estimator. Table 1 shows the frequencies of this event. Further research is needed to analyze this problem and potentially find a remedy.

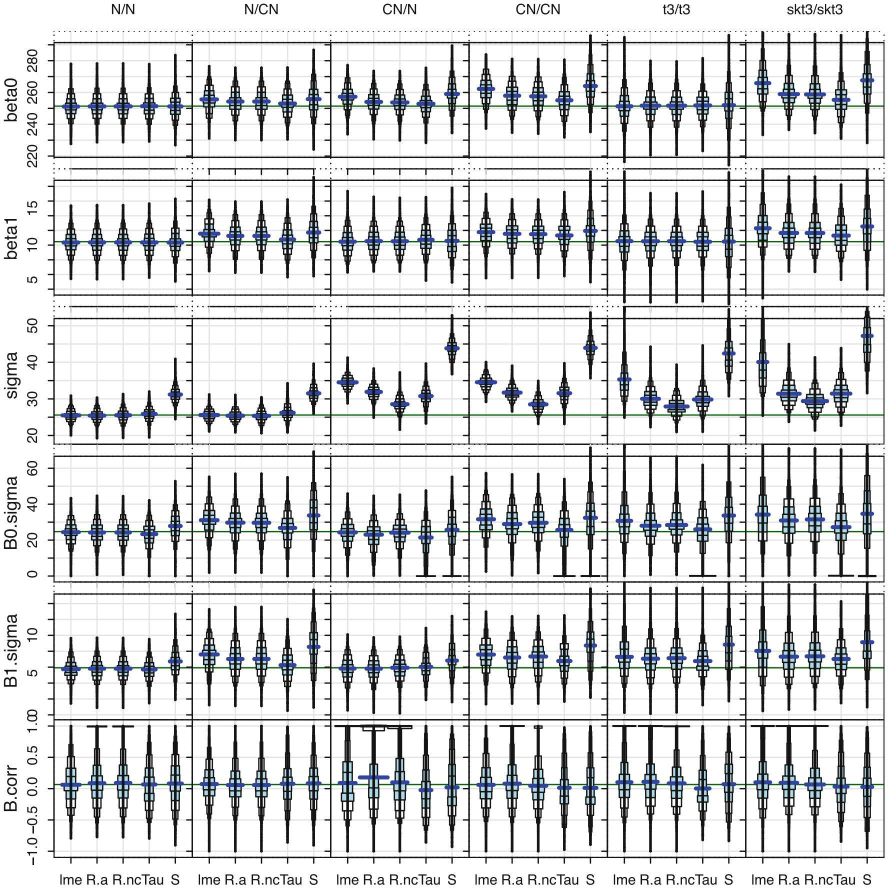


Fig. 5 Simulated distribution of estimates in the block-diagonal case. The green horizontal line marks the true value. The plotting function applies a robust "inner" plotting range in order to avoid the dominance of rare, extreme values. Where plotting overshoots the range defined by the solid plotting box, the respective elements of the plot are transformed nonlinearly to appear in the respective margin anyway

Table 1 Percentages of estimated correlation of (almost) 1 or -1 of the random effects B_0 and B_1 for the examined combinations of distributions

Method	N/N	N/CN	CN/N	CN/CN	t3/t3	skt3/skt3
lme	1.6	0.6	8.6	3.0	7.2	10.1
RSEa	8.0	3.6	25.9	11.7	12.0	13.0
RSEn	7.9	3.7	17.7	7.7	9.2	11.1
cTau	0.1	0.0	0.1	0.0	0.0	0.0
S	0.1	0.1	0.0	0.0	0.0	0.0

5 Examples

We apply the methods to two examples, a dataset with crossed random effects and a longitudinal dataset with a random intercept and slope.

5.1 *Penicillin Data*

The study of Davies and Goldsmith (1972), used as an example in Bates (2010), was run to …

> …assess the variability between samples of penicillin by the *B. subtilis* method. In this test method a bulk-inoculated nutrient agar medium is poured into a Petri dish of approximately 90 mm. diameter, known as a plate. When the medium has set, six small hollow cylinders or pots (about 4 mm. in diameter) are cemented onto the surface at equally spaced intervals. A few drops of the penicillin solutions to be compared are placed in the respective cylinders, and the whole plate is placed in an incubator for a given time. Penicillin diffuses from the pots into the agar, and this produces a clear circular zone of inhibition of growth of the organisms, which can be readily measured. The diameter of the zone is related in a known way to the concentration of penicillin in the solution.

The dataset thus implies a balanced two-way ANOVA model with two random effects: *sample* with six levels and *plate* with 24 levels. These random effects are completely crossed. Other current robust estimating methods, except for the one of Fellner (1986), cannot be applied here since the observations cannot be split into independent groups.

To make things more interesting, we slightly modified the original dataset. We scaled the first plate's observation values down and we moved one further observation down to the lowest original observation. The modified dataset is shown in Fig. 6.

We fit the model in R (R Core Team 2014) using function `rlmer` of the R package `robustlmm` published on CRAN (Koller 2016). The results of the classical and robust fits are shown in Table 2. For the robust fit, we use the RSE method tuned for 95% asymptotic efficiency. The robust method detects both alterations and contains their effects. The *plate* variance component is only slightly elevated, much less so than with the classical estimator. Increasing the severity of the contamination, i.e., shifting down the first group even more, only slightly influences the robust estimates, since the influences of the contamination on the estimates have already

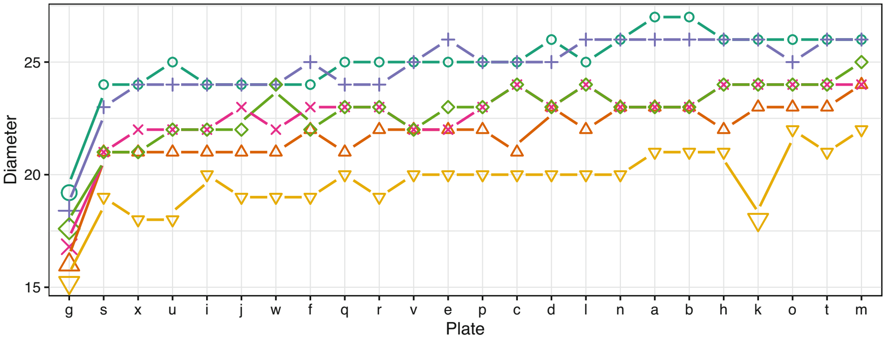


Fig. 6 Diameters of growth inhibition zones of 6 samples applied to each of 24 agar plates to assess penicillin concentration. The lines join the observations of the same sample. The plates have been reordered according to their mean response values. The observations displayed with larger symbols have been modified to introduce some contamination

Table 2 Fitted models for the Penicillin example. The classical fits were computed with `lmer`. The robust fits were computed using `rlmer` and the smoothed Huber function with tuning constant $c = 1.345$ for both ψ_e and ψ_b. For the variance components, weights were used with $c = 2.28$ (adjusted efficiency). The results are shown for the original and the modified data

	Original		Modified	
Data estimation	Classical	Robust	Classical	Robust
Intercept (std. error)	23.0 (0.809)	23.0 (0.843)	22.8 (0.85)	23.0 (0.848)
Random effects				
B0.sd (plate)	0.847	0.869	1.409	0.939
B1.sd (sample)	1.932	1.964	1.955	1.967
σ	0.550	0.545	0.609	0.566

reached the plateau. The classical estimates, on the other hand, will show a dramatic increase in the estimated variance component—rendering the estimate useless just because of one contaminated group.

5.2 *Sleep Study*

Belenky et al. (2003) studied the effects of sleep deprivation. 18 subjects chosen from a population of long distance drivers were allowed to sleep for only three hours each night for 10 days in a row. Each subject's reaction time was measured several times and averaged on each day of the trial.

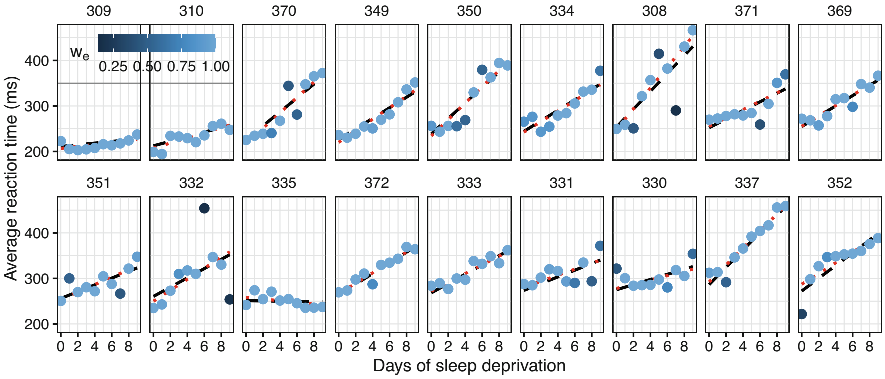


Fig. 7 Reaction times of subjects versus days of sleep deprivation. Each subject is shown in a separate facet. The dashed black and dotted red lines show the fixed plus random effects obtained by the classical and the robust methods, respectively. The darkness of the bullets reflects the observation level robustness weights w_e. The subjects are ordered by increasing intercept

A model to study the average increase of reaction time per day allows for a random intercept and a random linear effect of the day for each subject. The expected values of these two effects form the fixed effects. This results in a block-diagonal covariance matrix of the random effects, described by Eq. (13).

The data, the fitted values for classical and the robust fits as well as a robust per-subject fit are shown in Fig. 7. For the robust fit, we use the RSE method tuned for 90% asymptotic efficiency. While most of the subjects follow the general population fit quite closely, others, such as subject 335, show even a negative trend. Nevertheless, the robustness weights for the random effects do not show any clear outliers. Subject 309 is given a robustness weight of 0.69, the lowest of all the subjects. Subject 335 is given a weight of 0.81. There are only a small fraction of observations with a low robustness weights on the random error level. The observation on day 6 for Subject 332 is given the lowest robustness weight (value 0.15).

The estimated classical and robust coefficients are summarized in Table 3. The estimates for the fixed effects are nearly identical. The robust estimator almost doubles the random effects' standard deviations and reduces the random errors' standard deviation, compared to the classical fit. The correlation of the two random effects turns from slightly positive to slightly negative. This negative trend is clearly visible in the scatterplot of the predicted random effects shown in Fig. 8.

Table 3 Fitted models for the Sleep Study example. Approximate standard errors are shown in parentheses. Estimators are the same as for Table 2, except for the tuning constants. They are $c = 1$ for ψ_e, $c = 2.09$ for $w^{(\sigma)}$, $c = 3.011$ for ψ_b (w_k), and $c = 3.8$ for $w^{(\eta)}$ and $w^{(\tau)}$

Estimation	Classical	Robust
Fixed effects		
Intercept β_0	251.4	251.8
(std. error)	(6.82)	(7.87)
Days β_1	10.5	10.8
(std. error)	(1.55)	(1.72)
Random effects		
B0.sd	24.74	29.59
B1.sd	5.92	6.59
Correlation b.corr	0.0656	−0.0751
σ	25.6	19.6

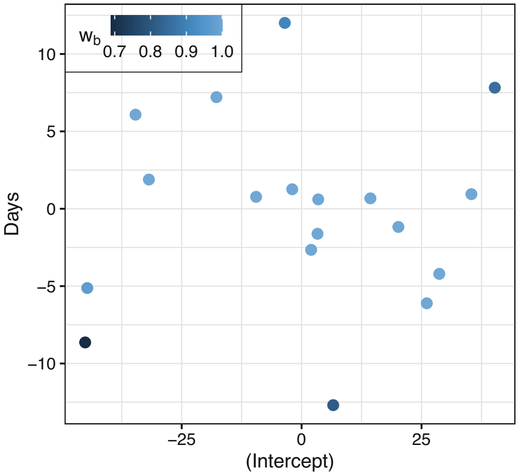


Fig. 8 Scatterplot of the predicted random effects of the robust fit for the Sleep Study example

6 Conclusions

We have developed a robust and flexible method of estimating linear mixed effects models robustly by robustifying the classical estimating equations. Aside from minor assumptions on the covariance matrices of the random effects (requiring a block-diagonal matrix), and the residual errors (diagonal matrix, know up to a scale), we do not make any additional assumptions to those for the classical methods. The structure in the data, given by the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, is arbitrary as long as the model is estimable in the classical sense.

The main advantage of the proposed RSE estimator lies in its generality. The competitors all cover only special mixed models. This is even true for the Composite-τ, which turns out to perform equally well as the robust estimating equations method in the simulations.

The limited assumptions make it difficult to derive some sort of asymptotic results. As in the classical case, such results will have to be established on a case by case basis (Miller 1977). However, the simulation study discussed in Sect. 4 suggests that the asymptotic properties are as one expects from such a method.

The method has been implemented in the R package `robustlmm` and is freely available on the official repository, CRAN. It comes with a vignette containing more detailed examples and information on how to choose the ψ-functions. The main function has the same arguments as `lmer` of the R package `lme4` by Bates et al. (2015). This enables a quick and easy way of checking the classical estimates for biases caused by contamination.

Acknowledgments The authors would like to thank Kali Tal for providing editorial help with an earlier version of the manuscript.

Appendix

Linear Approximation of Estimated Quantities

In this section, we develop linear approximations to the residuals $\widehat{\boldsymbol{\varepsilon}}^*$ and the estimated random effects $\widehat{\mathbf{b}}^*$. We use these linear approximations to compute the expected values in the estimating equations as well as the scaling factors τ used in the DAS approach.

Let $\Delta\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, $\Delta\mathbf{b}^* = \widehat{\mathbf{b}}^* - \mathbf{b}^*$, $\boldsymbol{\psi}_e^* = \boldsymbol{\psi}_e(\boldsymbol{\varepsilon}^*/\sigma)/\lambda_e$, $\boldsymbol{\psi}_b^* = \boldsymbol{\Lambda}_b^{-1}\boldsymbol{\psi}_b(\mathbf{b}^*/\sigma)$, $\boldsymbol{D}_e = \mathbf{Diag}\big(\boldsymbol{\psi}_e'(\boldsymbol{\varepsilon}^*/\sigma)\big)/\lambda_e$, $\boldsymbol{D}_b = \boldsymbol{\Lambda}_b^{-1}\mathbf{Diag}(w_b'(u_k)\mathbf{b}_k^*\mathbf{b}_k^{*\mathsf{T}}/\sigma^3 + w_b(u_k)\boldsymbol{I}_{J_k}/\sigma)_{k=1,\dots,K}$ with $u_k = \mathbf{b}_k^{*\mathsf{T}}\mathbf{b}_k^*/\sigma^2$, $\boldsymbol{X}^* = \boldsymbol{U}_e^{-1}\boldsymbol{X}$, and $\boldsymbol{Z}^* = \boldsymbol{U}_e^{-1}\boldsymbol{Z}\boldsymbol{U}_b$.

We linearize around $\boldsymbol{\beta}$ and $\mathbf{b}^*$, which will be the "true" $\boldsymbol{\beta}$ and $\mathbf{B}^*$ later on,

$$\boldsymbol{\psi}_e(\widehat{\boldsymbol{\varepsilon}}^*/\sigma)/\lambda_e \approx \boldsymbol{\psi}_e^* - \boldsymbol{D}_e\big(\boldsymbol{X}^*\Delta\boldsymbol{\beta} + \boldsymbol{Z}^*\Delta\mathbf{b}^*\big)/\sigma \ , \quad \text{and} \quad \boldsymbol{\Lambda}_b^{-1}\boldsymbol{\psi}_b\big(\widehat{\mathbf{b}}^*/\sigma\big) \approx \boldsymbol{\psi}_b^* + \boldsymbol{D}_b\Delta\mathbf{b}^*/\sigma \ .$$

Plugging these expressions into the estimating equations (7), divided by λ_e, and (11) and combining both equations into one yields

$$\begin{bmatrix} \boldsymbol{M}_{XX} & \boldsymbol{M}_{XZ} \\ \boldsymbol{M}_{ZX} & \widehat{\boldsymbol{M}}_{ZZ} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\beta}/\sigma \\ \Delta\mathbf{b}^*/\sigma \end{bmatrix} \approx \begin{bmatrix} \boldsymbol{X}^{*\mathsf{T}}\boldsymbol{\psi}_e^* \\ \boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{\psi}_e^* - \boldsymbol{\psi}_b^* \end{bmatrix},$$

where

$$\boldsymbol{M}_{XX} = \boldsymbol{X}^{*\mathsf{T}}\boldsymbol{D}_e\boldsymbol{X}^* \ , \qquad \widehat{\boldsymbol{M}}_{ZZ} = \boldsymbol{M}_{ZZ} + \boldsymbol{D}_b \ , \qquad \boldsymbol{M}_{ZZ} = \boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{D}_e\boldsymbol{Z}^* \ ,$$

$$\boldsymbol{M}_{XZ} = \boldsymbol{X}^{*\mathsf{T}}\boldsymbol{D}_e\boldsymbol{U}_e^{-1}\boldsymbol{Z}^* \ , \quad \text{and} \quad \boldsymbol{M}_{ZX} = \boldsymbol{M}_{XZ}^{\mathsf{T}} \ .$$

Using the formula for the inversion of a partitioned matrix, we have

$$\begin{bmatrix} \Delta\boldsymbol{\beta}/\sigma \\ \Delta\mathbf{b}^*/\sigma \end{bmatrix} \approx \begin{bmatrix} \boldsymbol{M}_{\beta\beta} & \boldsymbol{M}_{\beta b} \\ \boldsymbol{M}_{b\beta} & \boldsymbol{M}_{bb} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^{*\mathsf{T}}\boldsymbol{\psi}_e^* \\ \boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{\psi}_e^* - \boldsymbol{\psi}_b^* \end{bmatrix}, \tag{14}$$

where

$$\boldsymbol{M}_{bb} = \left(\widehat{\boldsymbol{M}}_{ZZ} - \boldsymbol{M}_{ZX}\boldsymbol{M}_{XX}^{-1}\boldsymbol{M}_{XZ}\right)^{-1}, \quad \boldsymbol{M}_{\beta\beta} = \boldsymbol{M}_{XX}^{-1} + \boldsymbol{M}_{XX}^{-1}\boldsymbol{M}_{XZ}\boldsymbol{M}_{bb}\boldsymbol{M}_{ZX}\boldsymbol{M}_{XX}^{-1},$$
$$\boldsymbol{M}_{\beta b} = -\boldsymbol{M}_{XX}^{-1}\boldsymbol{M}_{XZ}\boldsymbol{M}_{bb}, \quad \text{and} \qquad \boldsymbol{M}_{b\beta} = \boldsymbol{M}_{\beta b}^{\mathsf{T}}$$

or, equivalently,

$$\boldsymbol{M}_{\beta\beta} = \left(\boldsymbol{M}_{XX} - \boldsymbol{M}_{XZ}\widehat{\boldsymbol{M}}_{ZZ}^{-1}\boldsymbol{M}_{ZX}\right)^{-1}, \quad \boldsymbol{M}_{bb} = \widehat{\boldsymbol{M}}_{ZZ}^{-1} + \widehat{\boldsymbol{M}}_{ZZ}^{-1}\boldsymbol{M}_{ZX}\boldsymbol{M}_{\beta\beta}\boldsymbol{M}_{XZ}\widehat{\boldsymbol{M}}_{ZZ}^{-1},$$
$$\boldsymbol{M}_{b\beta} = -\widehat{\boldsymbol{M}}_{ZZ}^{-1}\boldsymbol{M}_{ZX}\boldsymbol{M}_{\beta\beta}, \quad \text{and} \qquad \boldsymbol{M}_{\beta b} = \boldsymbol{M}_{b\beta}^{\mathsf{T}}.$$

Plugging this into (14), we get an approximation for the residuals and for the estimated random effects,

$$\widehat{\boldsymbol{\varepsilon}}^* = \boldsymbol{U}_e^{-1}\left(\mathbf{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\boldsymbol{U}_b\widehat{\mathbf{b}}^*\right) = \boldsymbol{\varepsilon}^* - \boldsymbol{X}^*\Delta\boldsymbol{\beta} + \boldsymbol{Z}^*\Delta\mathbf{b}^* \approx \boldsymbol{\varepsilon}^* - \sigma\boldsymbol{A}_{ee}\boldsymbol{\psi}_e^* + \sigma\boldsymbol{A}_{eb}\boldsymbol{\psi}_b^* \tag{15}$$
$$\widehat{\mathbf{b}}^* \approx \mathbf{B}^* + \sigma\boldsymbol{A}_{be}\boldsymbol{\psi}_e^* - \sigma\boldsymbol{A}_{bb}\boldsymbol{\psi}_b^* \tag{16}$$

with

$$\begin{aligned}\boldsymbol{A}_{ee} &= \boldsymbol{X}^*\boldsymbol{M}_{\beta\beta}\boldsymbol{X}^{*\mathsf{T}} + \boldsymbol{X}^*\boldsymbol{M}_{\beta b}\boldsymbol{Z}^{*\mathsf{T}} + \boldsymbol{Z}^*\boldsymbol{M}_{\beta b}^{\mathsf{T}}\boldsymbol{X}^{*\mathsf{T}} + \boldsymbol{Z}^*\boldsymbol{M}_{bb}\boldsymbol{Z}^{*\mathsf{T}} \\ &= \boldsymbol{X}^*\boldsymbol{M}_{XX}^{-1}\boldsymbol{X}^{*\mathsf{T}} + \left(\boldsymbol{X}^*\boldsymbol{M}_{XX}^{-1}\boldsymbol{M}_{ZX} - \boldsymbol{Z}^*\right)\boldsymbol{M}_{bb}\left(\boldsymbol{X}^*\boldsymbol{M}_{XX}^{-1}\boldsymbol{M}_{ZX} - \boldsymbol{Z}^*\right)^{\mathsf{T}},\end{aligned}$$
$$\boldsymbol{A}_{bb} = \boldsymbol{M}_{bb}, \quad \boldsymbol{A}_{eb} = \boldsymbol{X}^*\boldsymbol{M}_{\beta b} + \boldsymbol{Z}^*\boldsymbol{M}_{bb} = \left(\boldsymbol{Z}^* - \boldsymbol{X}^*\boldsymbol{M}_{XX}^{-1}\boldsymbol{M}_{XZ}\right)\boldsymbol{M}_{bb}, \quad \text{and} \quad \boldsymbol{A}_{be} = \boldsymbol{A}_{eb}^{\mathsf{T}}.$$

Covariance Matrices

The approximations (15) and (16) are used in the computation of covariance matrices. In simpler setups, covariance matrices are calculated on the basis of influence functions **IF** by integrating **IF IF**$^{\mathsf{T}}$. **IF** is obtained, in the same way as for any M-estimator, from a linear approximation and results proportional to the ψ-function, the factor being the integral of its derivative, $\lambda = \mathbf{E}_0[\psi_e']$. Even though we have no rigorous proof for a generalization to our case, we apply this idea here.

The expected values of $\boldsymbol{D}_e$ and $\boldsymbol{D}_b$ are the identity matrices. When these expected values are used as approximations, the matrices $\boldsymbol{M}_{..}$ and $\boldsymbol{A}_{..}$ depend only on $\boldsymbol{\theta}$. The calculation of covariance matrices is then straightforward. They will contain the following expectations under the standard normal distribution:

$$\gamma_{.}^{(1)} = \mathbf{E}_o[z\psi_e(z)]/\lambda_. \qquad \gamma_{.}^{(2)} = \mathbf{E}_o\Big[\psi_e(z)^2\Big]/\lambda_.^2$$

where the dot (.) stands for e or b. The corresponding expressions for the block-diagonal case are $\boldsymbol{\Gamma}_b^{(1)}{=}\boldsymbol{\Lambda}_b^{-1}\mathbf{E}_o\big[\boldsymbol{\psi}(\mathbf{b}^*)\mathbf{b}^{*\mathsf{T}}\big]$ and $\boldsymbol{\Gamma}_b^{(2)}{=}\boldsymbol{\Lambda}_b^{-1}\mathbf{E}_o\big[\boldsymbol{\psi}(\mathbf{b}^*)\boldsymbol{\psi}(\mathbf{b}^*)^{\mathsf{T}}\big]\boldsymbol{\Lambda}_b^{-1}$. These are diagonal matrices with entries $\gamma_b^{(1)}\big(k(j), m_{k(j)}\big)$ and $\gamma_b^{(2)}\big(k(j), m_{k(j)}\big)$, respectively, which depend on the dimensions of the blocks k. They are given by

$$\gamma_b^{(p)}(k, m) = m^{-1}\mathbf{E}_{0,m}\big[w_k(u)^p u\big]/\lambda_b(k, m) \qquad p = 1, 2\,.$$

For fully diagonal $\boldsymbol{V}_b$, $m = 1$ and these formulas reduce to $\gamma_b^{(1)}$ and $\gamma_b^{(2)}$.

The covariance matrix of the estimated fixed effects is

$$\begin{aligned}
\mathbf{VAR}\big(\widehat{\boldsymbol{\beta}}\big) &= \sigma^2\mathbf{VAR}(\Delta\boldsymbol{\beta}/\sigma) = \sigma^2\mathbf{VAR}\big(\boldsymbol{M}_{\beta\beta}\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{\psi}_e^* + \boldsymbol{M}_{\beta b}\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{\psi}_e^* - \boldsymbol{M}_{\beta b}\boldsymbol{\psi}_b^*\big)\\
&= \sigma^2\big(\boldsymbol{M}_{\beta\beta}\boldsymbol{X}^{*\mathsf{T}} + \boldsymbol{M}_{\beta b}\boldsymbol{Z}^{*\mathsf{T}}\big)\mathbf{VAR}\big(\boldsymbol{\psi}_e^*\big)\big(\boldsymbol{X}^*\boldsymbol{M}_{\beta\beta} + \boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{M}_{b\beta}\big) + \sigma^2\boldsymbol{M}_{\beta b}\mathbf{VAR}\big(\boldsymbol{\psi}_b^*\big)\boldsymbol{M}_{b\beta}\\
&= \sigma^2\gamma_e^{(2)}\,\big(\boldsymbol{M}_{\beta\beta}\boldsymbol{M}_{XX}\boldsymbol{M}_{\beta\beta} + \boldsymbol{M}_{\beta\beta}\boldsymbol{M}_{XZ}\boldsymbol{M}_{b\beta} + \boldsymbol{M}_{\beta b}\boldsymbol{M}_{ZX}\boldsymbol{M}_{\beta\beta} + \boldsymbol{M}_{\beta b}\boldsymbol{M}_{ZZ}\boldsymbol{M}_{b\beta}\big)\\
&\quad + \sigma^2\boldsymbol{M}_{\beta b}\mathbf{E}_0\big[\boldsymbol{\psi}_b^*\boldsymbol{\psi}_b^{*\mathsf{T}}\big]\boldsymbol{M}_{b\beta}\\
&= \sigma^2\Big(\gamma_e^{(2)}\boldsymbol{M}_{\beta\beta} + \boldsymbol{M}_{\beta b}\Big(\boldsymbol{\Gamma}_b^{(2)} - \gamma_e^{(2)}\boldsymbol{I}\Big)\boldsymbol{M}_{b\beta}\Big).
\end{aligned} \tag{17}$$

For the derivation of the last equality, we have used the following two identities:

$$\begin{aligned}
&\boldsymbol{M}_{\beta\beta}\boldsymbol{M}_{XX}\boldsymbol{M}_{\beta\beta} + \boldsymbol{M}_{\beta\beta}\boldsymbol{M}_{XZ}\boldsymbol{M}_{b\beta} = \big(\boldsymbol{M}_{\beta\beta}\boldsymbol{M}_{XX} + \boldsymbol{M}_{\beta b}\boldsymbol{M}_{ZX}\big)\boldsymbol{M}_{\beta\beta} =\\
&\Big(\boldsymbol{I} + \boldsymbol{M}_{XX}^{-1}\boldsymbol{M}_{XZ}\boldsymbol{M}_{bb}\boldsymbol{M}_{ZX} - \boldsymbol{M}_{XX}^{-1}\boldsymbol{M}_{XZ}\boldsymbol{M}_{bb}\boldsymbol{M}_{ZX}\Big)\boldsymbol{M}_{\beta\beta} = \boldsymbol{M}_{\beta\beta}\;, \quad \text{and}
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{M}_{\beta b}\boldsymbol{M}_{ZX}\boldsymbol{M}_{\beta\beta} + \boldsymbol{M}_{\beta b}\boldsymbol{M}_{ZZ}\boldsymbol{M}_{b\beta} =\\
&\boldsymbol{M}_{\beta b}\Big(\boldsymbol{M}_{ZX}\boldsymbol{M}_{\beta\beta} - \big(\widehat{\boldsymbol{M}}_{ZZ} - \boldsymbol{D}_b\big)\widehat{\boldsymbol{M}}_{ZZ}^{-1}\boldsymbol{M}_{ZX}\boldsymbol{M}_{\beta\beta}\Big) = -\boldsymbol{M}_{\beta b}\boldsymbol{D}_b\boldsymbol{M}_{b\beta}.
\end{aligned}$$

For the DAS standardization, we need the covariance matrix of the residuals $\widehat{\varepsilon}_i^*$ and the $\widehat{b}_j^*$,

$$\begin{aligned}
\mathbf{VAR}(\widehat{\boldsymbol{\varepsilon}}^*) &\approx \boldsymbol{V}_e^* = \sigma^2\Big(\boldsymbol{I} - 2\gamma_e^{(1)}\boldsymbol{A}_{ee} + \gamma_e^{(2)}\boldsymbol{A}_{ee}\boldsymbol{A}_{ee} + \boldsymbol{A}_{eb}\boldsymbol{\Gamma}_b^{(2)}\boldsymbol{A}_{be}\Big)\,, \quad \text{and}\\
\mathbf{VAR}\big(\widehat{\mathbf{b}}^*\big) &\approx \boldsymbol{V}_b^* = \sigma^2\Big(\boldsymbol{I} - 2\boldsymbol{\Gamma}_b^{(1)}\boldsymbol{A}_{bb} + \boldsymbol{A}_{bb}\boldsymbol{\Gamma}_b^{(2)}\boldsymbol{A}_{bb} + \gamma_e^{(2)}\boldsymbol{A}_{be}\boldsymbol{A}_{eb}\Big)\,.
\end{aligned}$$

Refined Design Adaptive Scale

We first write down the equation determining $\tau_{e,i}$ for the determination of $\widehat{\sigma}$ through (9). The requirement that the ith term in the sum should be zero in expectation translates to the implicit equation

$$\begin{aligned}&\int \psi^{(\sigma)}\Big((e-\psi_e(e)-r)/\tau_{e,i}\Big)\varphi\Big(r/\sigma_i^{(R)}\Big)/\sigma_i^{(R)}\mathrm{d}r\varphi(e)\mathrm{d}e\\ &=\kappa_\sigma\int w^{(\sigma)}\Big((e-\psi_e(e)-r)/\tau_{e,i}\Big)\varphi\Big(r/\sigma_i^{(R)}\Big)/\sigma_i^{(R)}\mathrm{d}r\varphi(e)\mathrm{d}e\end{aligned}$$

for τ_i, where $\psi^{(\sigma)}(e)=e^2w^{(\sigma)}(e)$, φ is the standard normal density, $\sigma_i^{(R)}$ is the standard deviation of R_i, and κ_σ is defined below (9). The modification for the variance components θ_ℓ in the case of diagonal $\boldsymbol{V}_b$ is straightforward.

For random effects with block-diagonal covariance structure, we have

$$\begin{aligned}\widehat{\mathbf{b}}_k^*/\sigma &\approx \mathbf{B}_k^*/\sigma-(\boldsymbol{A}_{bb})_{kk}\lambda_b(k,m_k)\psi_k\big(\mathbf{B}_k^*/\sigma\big)-\mathbf{R}_k\\ \mathbf{R}_k &= \sum_{h\neq k}(\boldsymbol{A}_{bb})_{kh}\lambda_b(k,m_k)\psi_k\big(\mathbf{B}_h^*/\sigma\big)+\sum_i(\boldsymbol{A}_{be})_{ki}\psi_e\big(\varepsilon_i^*/\sigma\big)/\lambda_e\end{aligned}$$

and $\boldsymbol{T}_k$ is determined by

$$\begin{aligned}&\int \psi_{m(k)}^{(\eta)}\Big(\boldsymbol{T}_k^{-1}(\mathbf{b}-\psi_k(\mathbf{b})-\mathbf{r})\Big)\psi_{m(k)}^{(\eta)}\Big(\boldsymbol{T}_k^{-1}(\mathbf{b}-\psi_k(\mathbf{b})-\mathbf{r})\Big)^\intercal\exp\Big(-\mathbf{r}^\intercal(\boldsymbol{V}_k^{(R)})^{-1}\mathbf{r}/2\Big)\mathrm{d}\mathbf{r}\exp\Big(-\|\mathbf{b}\|^2/2\Big)\mathrm{d}\mathbf{b}\\ &=\kappa_\ell\int w_{m(k)}^{(\delta)}\Big(\|\boldsymbol{T}_k^{-1}(\mathbf{b}-\psi_k(\mathbf{b})-\mathbf{r})\|^2\Big)\exp\Big(-\mathbf{r}^\intercal(\boldsymbol{V}_k^{(R)})^{-1}\mathbf{r}/2\Big)\mathrm{d}\mathbf{r}\exp\Big(-\|\mathbf{b}\|^2/2\Big)\mathrm{d}\mathbf{b}\cdot\boldsymbol{V}_k^*\,,\end{aligned}$$

where $\psi^{(\eta)}(\mathbf{b})=\mathbf{b}\;w^{(\eta)}(\|\mathbf{b}\|)^{1/2}$ and $\boldsymbol{V}_k^{(R)}$ is the covariance matrix of $\mathbf{R}_k$. (Note that the normalizing constants of the densities cancel.) Integration thus extends over $2m(k)$ dimensions. With this choice of $\boldsymbol{T}_k$, each term in the sum (12) has approximate expectation zero. To see this, note that $\mathbf{b}_k^{*\intercal}\boldsymbol{Q}_{\ell,k}(\widehat{\boldsymbol{\theta}})\mathbf{b}_k^*=\mathrm{tr}\big(\mathbf{b}_k^*\mathbf{b}_k^{*\intercal}\boldsymbol{Q}_{\ell,k}(\widehat{\boldsymbol{\theta}})\big)$. Therefore, multiplying the last equation by $\boldsymbol{Q}_{\ell,k}(\widehat{\boldsymbol{\theta}})$ from the right and forming the trace proves the result.

The last equation resembles the problem of estimating a robust covariance matrix and can be computed along the same lines.

References

Agostinelli, C., & Yohai, V. J. (2016). Composite robust estimators for linear mixed models. *Journal of the American Statistical Association, 111*(516), 1764–1774.

Agostinelli, C., & Yohai, V. J. (2019). robustvarComp: Robust estimation for variance component models. R package version 0.1-5.

Alqallaf, F., Van Aelst, S., Yohai, V. J., & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 311–331.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bates, D. M. (2010). lme4: Mixed-Effects Modeling with R. http://lme4.r-forge.r-project.org/book/.

Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., & Balkin, T. J. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. *Journal of Sleep Research, 12*, 1–12.

Chervoneva, I., & Vishnyakov, M. (2011). Constrained S-estimators for linear mixed effects models with covariance components. *Statistics in Medicine, 30*(14), 1735–1750.

Chervoneva, I., & Vishnyakov, M. (2014). Generalized S-estimators for linear mixed effects models. *Statistica Sinica*, 1257–1276.

Copt, S., & Victoria-Feser, M. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association, 101*(473), 292–300.

Davies, O. L., & Goldsmith, P. L., (Eds.) (1972). *Statistical methods in research and production* (4th edn.) Hafner.

Fellner, W. (1986). Robust estimation of variance components. *Technometrics, 28*(1), 51–60.

Fernández, C., & Steel, M. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association, 93*(441), 359–371.

Geraci, M., & Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing, 24*(3), 461–479.

Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M. (2009). *Robust methods in biostatistics*. John Wiley & Sons.

King, R., & Anderson, E. (2021). skewt: The skewed Student-t distribution. R package version 1.0.

Koller, M. (2013). Robust estimation of linear mixed models. Diss., ETH Zürich, Nr. 20997, 2013.

Koller, M. (2016). robustlmm: an R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software, 75*, 1–24.

Koller, M. (2022). Replication code for simulation studies. https://CRAN.R-Project.org/package=robustlmm. Vignette included in R package robustlmm, version 3.0.

Koller, M., & Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis, 55*(8), 2504–2515.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with R)*. John Wiley & Sons.

Mason, F., Cantoni, E., & Ghisletta, P. (2021). Parametric and semi-parametric bootstrap-based confidence intervals for robust linear mixed models. *Methodology, 17*(4), 271–295.

Miller, J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics, 5*(4), 746–762.

Papritz, A., Künsch, H. R., Schwierz, C., & Stahel, W. A. (2013). Robust geostatistical analysis of spatial data. In *EGU General Assembly Conference Abstracts* (p. 14145).

Pinheiro, J., Liu, C., & Wu, Y. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics, 10*(2), 249–276.

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

Richardson, A. (1997). Bounded influence estimation in the mixed linear model. *Journal of the American Statistical Association, 92*(437).

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. John Wiley & Sons.

Stahel, W. (1987). Estimation of a covariance matrix with location: Asymptotic formulas and optimal B-robust estimators. *Journal of Multivariate Analysis, 22*(2), 296–312.

Stahel, W., & Welsh, A. (1997). Approaches to robust estimation in the simplest variance components model. *Journal of Statistical Planning and Inference, 57*(2), 295–319.

Asymptotic Behaviour of Penalized Robust Estimators in Logistic Regression When Dimension Increases

Ana M. Bianco, Graciela Boente, and Gonzalo Chebi

Abstract In the framework of logistic regression in order to obtain sparse models and automatic variable selection, penalized M-estimators that bound the deviance have been previously studied for fixed dimension. In this chapter, we consider a wide class of M-estimators that involves some well-known robust proposals and study their asymptotic behaviour when the covariates dimension grows to infinity with the sample size. Among other results, we obtain consistency, rates of convergence, and we explore the oracle properties of the regularized M-estimators, for penalty functions of different nature. Specifically, under suitable conditions, we prove that, with probability tending to 1, these estimators only select variables corresponding to non-null true coefficients, and we derive their asymptotic distribution.

Keywords Logistic regression · High-dimensional covariates · Penalty functions · Robust estimation · Sparsity

1 Introduction

A common practice to reduce the complexity of a regression model is to bet on sparsity. In this situation, it is assumed that the number of actually relevant predictors, k, is lower than the number p of measured covariates. Sparse models have been extensively studied in linear regression, but they are not limited to them. In particular, in high-dimensional logistic regression, practitioners usually have to face the challenge of robustly estimating sparse models, which is the topic of this chapter.

Logistic regression is a widely studied problem in statistics and has been useful to classify data. In the non-sparse scenario, the maximum likelihood estimator (MLE) of the regression coefficients is very sensitive to outliers, meaning that we cannot

A. M. Bianco · G. Boente (✉) · G. Chebi
Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina
e-mail: abianco@dm.uba.ar; gboente@dm.uba.ar; gonzalo.chebi@gmail.com

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_15

accurately classify a new observation based on these estimators, neither identify those covariates with important information for assignation. Robust methods for logistic regression have been introduced and discussed in Bianco and Yohai (1996), Cantoni and Ronchetti (2001), Croux and Haesbroeck (2003) and Bondell (2005, 2008), among others. The minimum divergence proposal due to Basu et al. (2017) may be seen as a particular case of the Bianco and Yohai (1996) estimator with a properly defined loss function. However, all these methods are not reliable under collinearity, and they do not allow for automatic variable selection when only a few number of covariates are relevant. These topics become more challenging when the number of covariates is close to the sample size or even larger.

Some robust estimators for logistic regression in the sparse regressors framework have already been proposed in the literature, in the last decade. Among others, we can mention Chi and Scott (2014) who considered a least squares estimator with a Ridge and Elastic Net penalty and Kurnaz et al. (2018) who proposed estimators based on a trimmed sum of the deviances with an Elastic Net penalty. It is worth noticing that the least squares estimator in Chi and Scott (2014) corresponds to a particular choice of the loss function bounding the deviance considered in Bianco and Yohai (1996). Finally, Tibshirani and Manning (2013) introduced a real-valued shift factor to protect against the possibility of mislabelling, while Park and Konishi (2016) considered a weighted deviance approach with weights based on the Mahalanobis distance computed over a lower-dimensional principal component space and that includes an Elastic Net penalty. In the present framework, the statistical challenge is not only to provide new statistical procedures, but also to show that they effectively provide variable selection and lead to the same asymptotic distribution as its oracle counterpart. Some results in that direction were obtained recently in Guo et al. (2017) and Avella-Medina and Ronchetti (2018) who treated the situation of penalized M-estimators in generalized linear models by bounding the quasi-likelihood. In this setting, Avella-Medina and Ronchetti (2018) considered penalties that are a deterministic sum of univariate functions, while Guo et al. (2017) proposed a penalty related to the ADALASSO one. Both of them studied the asymptotic behaviour of penalized robust quasi-likelihood type estimators, when the dimension p increases with the sample size n. Basu et al. (2021) considered robust estimators based on the density power divergence using an adaptively weighted LASSO penalty. Finally, Bianco et al. (2021) proposed a general family of penalized estimators based on bounding the deviance with a general penalty term, possible random, to produce sparse estimators and studied their asymptotic behaviour for fixed p. In this sense, our aim is to fill the gap by studying the asymptotic behaviour of the penalized robust estimators defined in Bianco et al. (2021) when the dimension increases with the sample size. Unlike Guo et al. (2017), according to a natural point of view in robustness, we do not assume that the parameter space is a compact subset of $\mathbb{R}^p$ and weaker assumptions on the penalty are required. Besides, our results are not restricted to the LASSO or ADALASSO penalties as in Avella-Medina and Ronchetti (2018) or Guo et al. (2017). Indeed, they are stated in a general penalty framework that allows to include not only the two penalties already mentioned but also SCAD and MCP penalties. The rest of this chapter is organized as follows. In Sect. 2, we briefly review the robust penalized logistic regression estimators defined in Bianco et al. (2021). Sections 3 and 4 summarize the asymptotic properties of the proposal. Proofs are relegated to the Appendix or to the technical report available at Bianco et al. (2022).

2 Preliminaries: Robust Penalized Estimators

Throughout this chapter, we consider a sequence of logistic regression models, where the number of covariates $p = p_n$ diverges to infinity. To be more precise, we consider a triangular array of independent Bernoulli random variables $\{y_{n,i} : 1 \le i \le n,\ n \ge 1\}$ and the corresponding triangular array of explanatory variables $\{\mathbf{x}_{n,i} : 1 \le i \le n,\ n \ge 1\}$, where $\mathbf{x}_{n,i} \in \mathbb{R}^p$ and $y_{n,i}|\mathbf{x}_{n,i} \sim Bi(1, \pi_{0,n,i})$ with $\pi_{0,n,i} = \mathbb{P}(y_{n,i} = 1|\mathbf{x}_{n,i}) = F(\mathbf{x}_{n,i}^{\mathrm{T}}\boldsymbol{\beta}_{0,n})$ and $F(t) = \exp(t)(1 + \exp(t))^{-1}$. The sequence $\{\boldsymbol{\beta}_{0,n} : n \ge 1\}$ corresponds to the true regression coefficient parameters. We will assume that for each n, $(y_{n,i}, \mathbf{x}_{n,i})$, $1 \le i \le n$, are independent and identically distributed.

Denote $\textsc{dev}(y, t) = -\log(F(t))y - \log(1 - F(t))(1 - y)$ the deviance and $\rho : \mathbb{R}_{\ge 0} \to \mathbb{R}$ a loss function that is bounded, differentiable, and nondecreasing with derivative $\psi = \rho'$. For sparse models when the dimension is fixed, Bianco et al. (2021) defined a family of regularized robust estimators as

$$\widehat{\boldsymbol{\beta}}_n = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \phi(y_{n,i}, \mathbf{x}_{n,i}^{\mathrm{T}}\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta})\,, \tag{1}$$

with the goal of penalizing candidates with many non-zero components. In (1), $I_{\lambda_n}(\boldsymbol{\beta})$ is a penalty function, chosen by the user, depending on a tuning parameter λ_n that measures the estimated logistic regression model complexity, while

$$\begin{aligned}\phi(y, t) &= \rho(\textsc{dev}(y, t)) + G(F(t)) + G(1 - F(t))\\ &= y\rho(-\log[F(t)]) + (1 - y)\rho(-\log[1 - F(t)]) + G(F(t)) + G(1 - F(t))\,,\end{aligned} \tag{2}$$

with $G(t) = \int_0^t \psi(-\log u)\, du$. Note that $G(F(t)) + G(1 - F(t))$ is the correction term needed to guarantee Fisher consistency in the non-regularized case, as introduced in Bianco and Yohai (1996). When the model contains an intercept, this component is usually not penalized. For that reason and for the sake of simplicity, when deriving the asymptotic properties of the estimators, we will assume that the model has no intercept. When the penalty function is properly chosen, the penalized M-estimator defined in (1) is well defined even when $p > n$ and leads to sparse models as we will show below.

As mentioned in Bianco et al. (2021), the estimators given through (1) define a wide family that includes, beyond the penalized maximum likelihood estimator, the penalized least squares estimator proposed in Chi and Scott (2014), since it corresponds to the bounded function $\rho(t) = 1-\exp(-t)$ and the Elastic Net penalty $I_\lambda(\boldsymbol{\beta}) = \lambda\big(a\|\boldsymbol{\beta}\|_1 + [(1-a)/2]\|\boldsymbol{\beta}\|_2^2\big)$. When $I_\lambda(\boldsymbol{\beta}) \equiv 0$, taking as loss function in (2) $\rho(t) = \rho_{\text{DIV}}(t) = (1+1/c_0)\{1-\exp(-c_0\,t)\}$, we obtain the minimum divergence estimators defined in Basu et al. (2017).

As it is well known, LASSO penalty tends to over-penalize large coefficients, resulting in a larger and biased model. In contrast, the choice of an appropriate non-convex penalty function can overcome this drawback. Among other non-convex penalties, we can mention the Bridge penalty introduced in Frank and Friedman (1993) and defined as $I_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_q^q$ that is non-convex for $0 < q < 1$, the smoothly clipped absolute deviation (SCAD) penalty defined in Fan and Li (2001), and the minimax concave penalty (MCP) proposed by Zhang (2010). Both SCAD and MCP penalties can be written as $I_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^p J_\lambda(|\beta_j|)$, where $J_\lambda(\cdot)$ is a non-negative, twice differentiable function in $(0,\infty)$. More precisely, for any non-negative real number b, the function $J_\lambda(b)$ equals $\text{SCAD}_{\lambda,a}(b)$ in the first case and $\text{MCP}_{\lambda,a}(b)$ in the latter, where

$$\text{SCAD}_{\lambda,a}(b) = \lambda b\mathbf{1}_{\{b\le\lambda\}} + \frac{1}{a-1}\left(a\,\lambda\,b - \frac{b^2+\lambda^2}{2}\right)\mathbf{1}_{\{\lambda<b\le a\lambda\}} + \frac{\lambda^2(a^2-1)}{2(a-1)}\mathbf{1}_{\{b>a\lambda\}}\,,$$

$$\text{MCP}_{\lambda,a}(b) = \left(\lambda b - \frac{b^2}{2a}\right)\mathbf{1}_{\{b\le a\lambda\}} + \frac{a\lambda^2}{2}\mathbf{1}_{\{b>a\lambda\}}\,,$$

with $\mathbf{1}_A$ the indicator of the set A. For both penalties, the positive constant a, which must be larger than 2 for SCAD, is selected by the user. Since the loss functions and penalties in this chapter are non-convex, we will also consider the following restricted estimator:

$$\widehat{\boldsymbol{\beta}}_{n,\,R} = \underset{\|\boldsymbol{\beta}\|_1\le R}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^n \phi(y_{n,i}, \mathbf{x}_{n,i}^{\mathrm{T}}\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta}), \tag{3}$$

where $R > 0$ is a fixed constant and ϕ is the function given in (2). This type of restrictions has been considered in Loh (2017) and Elsener and van de Geer (2018)), when the minimization problem involves a non-convex function. As it will be shown, consistency properties are easier to obtain for these restricted estimators. However, in this chapter, we also give consistency results for the unrestricted estimator defined in (1).

2.1 Assumptions

In order to derive the asymptotic results, the following assumptions on the loss function ρ used in (2) will be needed.

R1 $\rho : \mathbb{R}_{\ge 0} \to \mathbb{R}$ is bounded and continuously differentiable with bounded derivative ψ and $\rho(0) = 0$.

R2 $\psi(t) \geq 0$, and there exists some $c \geq \log 2$ such that $\psi(t) > 0$ for all $0 < t < c$.

R3 ρ is bounded, twice continuously differentiable with bounded derivatives, i.e., ψ and $\psi' = \rho''$ are bounded. Moreover, $\rho(0) = 0$.

R4 $\psi(t) \geq 0$, and there exist values $c \geq \log 2$ and $\tau > 0$ such that $\psi(t) > \tau$ for every $0 < t < c$.

R5 ρ is bounded, three times continuously differentiable, with bounded derivatives ψ, ψ', and ψ'' and $\rho(0) = 0$.

Remark 1 Assumption **R5** entails that the function $\phi(y, t)$ defined in (2) is three times differentiable with respect to t and that the related derivatives are bounded for $y \in \{0, 1\}$. On the other hand, if $\psi(0) \neq 0$ and assumptions **R1** and **R2** hold for some constant $c > \log(2)$, then condition **R4** holds.

When considering the penalized minimum divergence estimators with tuning parameter c_0, the loss function $\rho(t) = \rho_{\text{DIV}}(t) = (1 + 1/c_0)\{1 - \exp(-c_0\, t)\}$ automatically satisfies conditions **R1** to **R5** since $\psi_{\text{DIV}}(t) = \rho'_{\text{DIV}}(t) = c_0\,(1 + 1/c_0)\exp(-c_0\, t) > 0$ for all t. Moreover, **R1** to **R5** also hold when considering the function

$$\rho_{c_0}(t) = \left(t - \frac{15}{16\,c_0}t^2 + \frac{5}{16\,c_0^3}t^4 - \frac{1}{16\,c_0^5}t^6\right)\mathbf{1}_{0\leq t\leq c_0} + \frac{5}{16}c_0\,\mathbf{1}_{t>c_0}\,,$$

which corresponds to truncating the identity function and adding a polynomial term to ensure smoothness. It is worth mentioning that this loss function is a modification of the one considered in Bianco and Yohai (1996) to ensure that **R5** holds.

For the sake of simplicity and to avoid burden notation, henceforth, we will omit the subscript n unless necessary. For instance, we will write $\boldsymbol{\beta}_0$ instead of $\boldsymbol{\beta}_{0,n}$.

We assume, without loss of generality, that only the first k covariates are relevant for prediction purposes, i.e., $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,\text{A}}^{\text{T}}, \mathbf{0}_{p-k}^{\text{T}})^{\text{T}}$, where $\boldsymbol{\beta}_{0,\text{A}} \in \mathbb{R}^k$ corresponds to the active components, that is, it has all its coordinates different from zero. It is worth mentioning that the number $k = k_n$ of non-zero components may depend on n, eventually growing with the sample size. For that reason, in order to obtain results regarding the asymptotic distribution of our estimators, conditions on the quantity $m_{0,n}$ that involves only the coefficients in $\boldsymbol{\beta}_{0,\text{A}}$ and is defined as

$$m_{0,n} = \min\{|\beta_{0,j}| : \beta_{0,j} \neq 0\} \tag{4}$$

will be required. As mentioned in Bühlmann and van de Geer (2011), variable selection properties depend on the fact that the minimum signal $m_{0,n}$ does not tend to zero too fast.

To be consistent with the notation used for $\boldsymbol{\beta}_0$, we will partition a vector of covariates $\mathbf{x}$ as $\mathbf{x} = (\mathbf{x}_{\text{A}}^{\text{T}}, \mathbf{x}_{\text{NA}}^{\text{T}})^{\text{T}}$, where $\mathbf{x}_{\text{A}} \in \mathbb{R}^k$ and $\mathbf{x}_{\text{NA}} \in \mathbb{R}^{p-k}$. Besides, as done for the covariates, we will also write the estimator of $\boldsymbol{\beta}_0$ as $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,\text{A}}^{\text{T}}, \widehat{\boldsymbol{\beta}}_{n,\text{NA}}^{\text{T}})^{\text{T}}$, where $\widehat{\boldsymbol{\beta}}_{n,\text{A}} \in \mathbb{R}^k$ corresponds to the active components of $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_{n,\text{NA}} \in \mathbb{R}^{p-k}$ to the null ones.

Given a symmetric and positive semi-definite matrix $\mathbf{C} \in \mathbb{R}^{p\times p}$, the smallest and largest eigenvalues of $\mathbf{C}$ will be denoted as $\iota_1(\mathbf{C})$ and $\iota_p(\mathbf{C})$, respectively. From now on, we denote as $\Psi(y,t) = \partial\phi(y,t)/\partial t$ and $\chi(y,t) = \partial\Psi(y,t)/\partial t$. Note that $\Psi(y,t) = -[y - F(t)]\nu(t)$, while $\chi(y,t) = F(t)(1-F(t))\nu(t)-(y-F(t))\nu'(t)$, with

$$\nu(t) = \psi(-\log F(t))[1 - F(t)] + \psi(-\log[1 - F(t)])F(t)\,. \tag{5}$$

The function $\chi(y,t)$ always exists for the minimum divergence estimators, while for other choices of the loss function ρ, it is well defined when ρ is twice continuously differentiable as required in **R3**. We also have that $\chi(0,s) = \chi(1,-s)$. To lighten the notation in the next assumptions, let $(y_n, \mathbf{x}_n)$ be such that $(y_n, \mathbf{x}_n) \sim (y_{n,1}, \mathbf{x}_{n,1})$ and denote

$$\mathbf{H} = \mathbf{H}_n = \mathbb{E}\big(\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\big)\,. \tag{6}$$

We will also consider the following hypotheses regarding the distribution of the covariates.

Z1 $\mathbb{E}\left(\max\limits_{1\le j\le p}\sum\limits_{i=1}^{n} x_{n,ij}^2/n\right) = O(1)$, where $x_{n,ij}$ is the jth coordinate of the random vector $\mathbf{x}_{n,i}$.
Z2 There exists a constant $K_1 > 0$ not depending on n such that $\iota_p(\mathbf{H}) \le K_1$.
Z3 There exists a constant $\tau_1 > 0$ not depending on n such that $\iota_1(\mathbf{H}) \ge \tau_1$.
Z4 There exists a constant $K_2 > 0$ not depending on n such that $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{H}\boldsymbol{\beta}_0 \le K_2^2$.
Z5 $\mathbf{x}_n$ has a centred elliptical distribution with characteristic function $\phi_{\mathbf{x}_n}(\mathbf{t}) = \xi(\mathbf{t}^{\mathrm{T}}\boldsymbol{\Gamma}_p\,\mathbf{t})$ for some semi-definite symmetric matrix $\boldsymbol{\Gamma}_p \in \mathbb{R}^{p\times p}$ and some function $\xi : \mathbb{R} \to \mathbb{R}$ that does not depend on n, where to avoid burden notation, $\boldsymbol{\Gamma}_p = \boldsymbol{\Gamma}_{p_n}$.
Z6 There exists a constant $K_3 > 0$ not depending on n such that $\mathbb{E}\,\|\mathbf{x}_{n,\mathrm{A}}\|_2^6 \le K_3$.
Z7 There exists a constant $\tau_2 > 0$ not depending on n such that $\iota_1(\mathbf{B}_{\mathrm{A}}) \ge \tau_2$, where $\mathbf{B}_{\mathrm{A}} = \mathbf{B}_{n,\mathrm{A}} = \mathbb{E}\big[\Psi^2(y_n, \mathbf{x}_{n,\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{0,\mathrm{A}})\mathbf{x}_{n,\mathrm{A}}\mathbf{x}_{n,\mathrm{A}}^{\mathrm{T}}\big]$.

Remark 2 Assumption **Z1** is needed to obtain rates of convergence with order $(p\log p/n)^{1/2}$ without requiring additional bounding conditions on the eigenvalues of $\mathbf{H}$. This assumption holds, for example, if $\mathbf{x}_n \sim N(\mathbf{0}_p, \mathbf{I}_p)$ and $a_n = \log p\,/n \to 0$. Indeed, let $V_j = \sum_{i=1}^{n} x_{n,ij}^2$, $V_1,\dots,V_p$ are independent $V_j \sim \chi_n^2$. Then, inequality (7) in Dasarathy (2011) allows to obtain the bound

$$\mathbb{E}\left(\max_{1\le j\le p}\frac{1}{n}\sum_{i=1}^{n} x_{n,ij}^2\right) = \frac{1}{n}\,\mathbb{E}\left(\max_{1\le j\le p} V_n\right) \le \frac{4\,a_n}{1-\exp(-2\,a_n)}\,.$$

Using the fact that $1 - x \ge \exp(-2x)$ for $0 < x \le 1/2$, we get that **Z1** holds if $a_n = \log p\,/n \to 0$.

Assumptions **Z3**, **Z4**, and **Z5** will be used to derive consistency results for the unrestricted estimator defined in (1). It is worth mentioning that, under **Z3**, the matrix $\boldsymbol{\Gamma}_p$ in **Z5** is nonsingular.

Note that **Z3** and **Z4** imply that $\tau_1 \|\boldsymbol{\beta}_0\|^2 \leq \text{VAR}(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_0) \leq K_2^2$, which together with the fact that $\|\boldsymbol{\beta}_0\|^2 = \|\boldsymbol{\beta}_{0,\mathrm{A}}\|^2$, leads to $\sum_{j=1}^{k} \beta_{0,j}^2 \leq K_2^2/\tau_1$ for all n (even if k grows with the sample size). In particular, $\max\{|\beta_{0,j}| : \beta_{0,j} \neq 0\}$ is bounded and $m_{0,n} = O(1/\sqrt{k})$, with $m_{0,n}$ defined in (4). Then, if $k \to \infty$, as the sample size increases, and assumptions **Z3** and **Z4** hold, we have that $m_{0,n} \to 0$.

Assumption **Z2** is required to obtain rates of convergence with order $\sqrt{n/p}$ (see Theorem 2b)). Finally, **Z6** and **Z7** will be used to derive the asymptotic normality of the estimators when using the SCAD or MCP penalties.

Remark 3 It is worth mentioning that assumption **Z5** holds if $\mathbf{x}_n$ is a scale mixture of normal distributions of the form $\mathbf{x}_n \sim S_n\, \mathbf{z}_n$, where S_n and $\mathbf{z}_n$ are independent, $\mathbb{P}(S_n > 0) = 1$, $\mathbf{z}_n \sim N(\mathbf{0}_p, \boldsymbol{\Gamma}_p)$, and, in addition, S_n has a distribution that does not depend on n, i.e., for all n, $S_n \sim S$, for some positive random variable S. Among others, assumption **Z5** includes the contaminated normal and the multivariate Student's T_m with degrees of freedom m not depending on n.

Remark 4 Analogous arguments to those considered in Remark 2.1 in Boente et al. (2014) allow to show that if assumption **Z5** holds and $\boldsymbol{\Gamma}_p > 0$ (which arises if **Z3** also holds), then $\mathbf{x}_n$ is a scale mixture of normals with the structure described in Remark 3. The proof of this statement may be found in Appendix 1.

Let $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^{\mathrm{T}}$. To obtain the asymptotic distribution of the estimators of $\boldsymbol{\beta}_{0,\mathrm{A}}$ and the oracle property of the penalized estimators, we consider the following assumptions regarding the growth of n, k, λ_n and the coefficients in $\boldsymbol{\beta}_{0,\mathrm{A}}$.

N1 $m_{0,n}\, \sqrt{n/k} \to \infty$.
N2 $m_{0,n}/\lambda_n \to \infty$.
N3 $k/n = O(\lambda_n^2)$.

Remark 5 It is worth mentioning that, if k and $\boldsymbol{\beta}_{0,\mathrm{A}}$ are fixed, **N1** holds, whereas **N2** is equivalent to $\lambda_n \to 0$. On the other hand, if there exists $m_0 > 0$ (independent of n) such that $m_{0,n} > m_0$, then $k/n \to 0$ and $\lambda_n \to 0$ imply **N1** and **N2**, respectively. If additionally $m_{0,n}$ has a finite upper bound, these conditions are equivalent. Finally, if $m_{0,n} = O(1/\sqrt{k})$, as it is the case when **Z3** and **Z4** hold, then **N1** and **N2** imply $k^2/n \to 0$ and $k\, \lambda_n^2 \to 0$. Note that the two latter conditions are the same when $\lambda_n = O(1/\sqrt{n})$. For other convergence rates of the penalty parameter, **N1** and **N2** give a relationship between the penalty parameter and the speed at which the number of non-zero coordinates increases.

3 Consistency and Rates of Convergence

Recall that $(y_n, \mathbf{x}_n) \sim (y_{n,1}, \mathbf{x}_{n,1})$, and then define $L_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi(y_{n,i}, \mathbf{x}_{n,i}^{\mathrm{T}}\boldsymbol{\beta})/n$ and $\mathbb{L}_n(\boldsymbol{\beta}) = \mathbb{E}\phi(y_n, \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})$. Note that $L_n(\boldsymbol{\beta})$ is the empirical counterpart of $\mathbb{L}_n(\boldsymbol{\beta})$. It is worth mentioning that since the distribution of $\mathbf{x}_n$ and the dimension of $\mathbf{x}_n$ and $\boldsymbol{\beta}_0$ depend on n through $p = p(n)$, the function $\mathbb{L}_n(\cdot)$ also depends on n. However, to avoid burden notation, from now on, we will omit the dependence on n and write $\mathbb{L}(\boldsymbol{\beta})$ instead of $\mathbb{L}_n(\boldsymbol{\beta})$, unless necessary. In order to give a measure of closeness between two predicted probabilities, given $\boldsymbol{\beta}_j \in \mathbb{R}^p$, $j = 1, 2$, we define $d_n^2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathbb{E}[F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_1) - F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_2)]^2$, where the index n is used to make explicit the dependence on the sample size. Note that $d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0)$ can be written as $d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) = \mathbb{E}\left\{[F(\mathbf{x}_n^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n) - F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_0)]^2 \middle| (y_{n,1}, \mathbf{x}_{n,1}), \dots, (y_{n,n}, \mathbf{x}_{n,n})\right\}$.

The following result shows that the estimators defined in (1) and (3) lead to consistent predictions. The weak consistency of the unrestricted estimator defined through (1), in the sense that $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \xrightarrow{p} 0$, is also derived under additional assumptions.

Theorem 1 *Suppose that* **R1** *and* **R4** *hold. Let* $\widehat{\boldsymbol{\beta}}_n$ *and* $\widehat{\boldsymbol{\beta}}_{n,\,R}$ *be the estimators defined through* (1) *and* (3)*, respectively. Then, if* $r_n = \sqrt{p/n} + I_{\lambda_n}(\boldsymbol{\beta}_0)$*, we have that:*

(a) $d_n^2(\widehat{\boldsymbol{\beta}}_{n,\,R}, \boldsymbol{\beta}_0) = O_{\mathbb{P}}(r_n)$ *and* $d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) = O_{\mathbb{P}}(r_n)$.

(b) If, in addition, $\iota_1(\mathbf{H}) > 0$*, where* $\mathbf{H}$ *is defined in* (6)*,* $\|\boldsymbol{\beta}_0\|_1 \leq R$*, and there exists a constant* $M > 0$ *such that* $P(\|\mathbf{x}_n\|_\infty \leq M) = 1$*, for all* $n \geq 1$*, then* $\|\widehat{\boldsymbol{\beta}}_{n,\,R} - \boldsymbol{\beta}_0\|_2^2 = O_{\mathbb{P}}(r_n/\iota_1(\mathbf{H}))$*. Therefore, if* **Z3** *holds,* $\|\widehat{\boldsymbol{\beta}}_{n,\,R} - \boldsymbol{\beta}_0\|_2^2 = O_{\mathbb{P}}(r_n)$.

(c) When **Z3***,* **Z4***, and* **Z5** *also hold,* $p/n \to 0$ *and* $I_{\lambda_n}(\boldsymbol{\beta}_0) \to 0$*, then* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \xrightarrow{p} 0$.

Remark 6 Note that Theorem 1(a) implies that $F(\mathbf{x}_n^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n)$ is consistent in the L_2-norm if $p/n \to 0$ and $I_{\lambda_n}(\boldsymbol{\beta}_0) \to 0$. However, in contrast to the linear regression setting where from **Z3** this convergence implies the consistency of $\widehat{\boldsymbol{\beta}}_n$, some additional assumptions will be needed for the unrestricted estimator. The main reason for this difference is the nature of the logistic regression model, where the link function F is such that $F'(t)$ converges to 0 when $|t| \to \infty$.

Note that according to Theorem 1 for the restricted estimator defined through (3) the situation is different than that for the unrestricted one. Effectively, as for regression models, a bound of the type

$$\mathbb{E}\left[(F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}) - F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_0))^2\right] \geq C\ \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \qquad (7)$$

may be obtained for any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\|_1 \leq R$, only requiring **Z3**. Hence, for the restricted estimator, all the results in the paper hold without **Z5**. It should be highlighted that in Loh and Wainwright (2015), results for restricted penalized maximum likelihood estimators are obtained requiring a sub-Gaussian condition to the covariates, in addition to **Z3**.

In contrast, when considering the unrestricted estimator defined through (1) assumption **Z5** on the covariates distribution is also needed to obtain (7). In particular, Theorem 1(c) and Remark 3 imply that the consistency of $F(\mathbf{x}_n^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n)$ leads to the consistency of the unrestricted estimator defined in (1), when $\mathbf{x}_n$ is a scale mixture of normals.

It is also worth mentioning that the result given in Theorem 1(b) provides a preliminary rate of convergence for the restricted estimator that will be improved in Theorem 2, under suitable conditions.

From now on, $\mathcal{B}_s(\boldsymbol{\theta}, \delta)$ stands for the closed s-dimensional ball, with respect to the usual $\|\cdot\|_2$, centred at $\boldsymbol{\theta}$ with radius δ, that is, $\mathcal{B}_s(\boldsymbol{\theta}, \delta) = \{\mathbf{z} \in \mathbb{R}^s : \|\mathbf{z} - \boldsymbol{\theta}\|_2 \leq \delta\}$. Moreover, when $\delta = 1$, we will write $\mathcal{B}_s(\boldsymbol{\theta})$ instead of $\mathcal{B}_s(\boldsymbol{\theta}, 1)$. In order to obtain rates of convergence for Lipschitz penalties, such as LASSO, or for bounded differentiable ones, such as SCAD or MCP, but under weaker conditions for λ_n in the latter case, we will need the following additional assumptions:

P1 There exist $\epsilon > 0$ and a constant K that does not depend on λ_n nor on n, such that $\left|I_{\lambda_n}(\boldsymbol{\beta}_1) - I_{\lambda_n}(\boldsymbol{\beta}_2)\right| \leq \lambda_n\, K\, \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1$, for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}_p(\boldsymbol{\beta}_0, \epsilon)$.

P2 There exist a positive constant value $\widetilde{\delta}$ and non-negative sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, such that, for any $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \widetilde{\delta}$, the penalty I_{λ_n} satisfies

$$I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) \geq -a_n\sqrt{k}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 - b_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \,. \tag{8}$$

Remark 7 As mentioned in Remark 4 in Bianco et al. (2021), Ridge, Elastic Net, LASSO, SCAD, and MCP penalties satisfy **P1**, while LASSO, SCAD, or MCP penalties satisfy **P2**. Indeed, for the LASSO penalty, this assumption holds taking $a_n = \lambda_n$ and $b_n = 0$. SCAD or MCP penalties can be written as $I_{\lambda_n}(\boldsymbol{\beta}) = \sum_{j=1}^{p} J_{\lambda_n}(|\beta_j|)$, where $J_{\lambda_n}(\cdot)$ is a non-negative, twice differentiable function in $(0, \infty)$, $J'_{\lambda_n}(|\beta_{0,\ell}|) \geq 0$, and $J_{\lambda_n}(0) = 0$. Given $\delta_0 > 0$, define

$$\begin{aligned} a_n &= \max\left\{J'_{\lambda_n}(|\beta_{0,\ell}|) : 1 \leq \ell \leq p \ \text{ and } \ \beta_{0,\ell} \neq 0\right\} = \max\left\{J'_{\lambda_n}(|\beta_{0,\ell}|) : 1 \leq \ell \leq k\right\} \\ b_n &= b_n(\delta_0) = \sup\{|J''_{\lambda_n}(|\beta_{0,\ell}| + \tau\delta_0)| : \tau \in [-1, 1]\ ,\ 1 \leq \ell \leq p \ \text{ and } \ \beta_{0,\ell} \neq 0\} \\ &= \sup\{|J''_{\lambda_n}(|\beta_{0,\ell}| + \tau\delta_0)| : \tau \in [-1, 1]\ ,\ 1 \leq \ell \leq k\}\,. \end{aligned}$$

Using same arguments as those considered in the proof of Theorem 2(b) in Bianco et al. (2021), it may be shown that (8) holds.

When considering the SCAD or MCP penalties, $J'_{\lambda_n}(t)$ and $J''_{\lambda_n}(t)$ are equal to zero if $t > a\,\lambda_n$ where a is the second tuning parameter of this penalty function (which is assumed to be fixed by the user). Hence, if $m_{0,n} > a\lambda_n$ for $n \geq n_0$, where $m_{0,n}$ is defined in (4), we have that $a_n = 0$ and $b_n = 0$ for a sufficiently large n. In particular, this holds if there exists $m_0 > 0$ such that it does not depend on n, $m_{0,n} > m_0$, and $\lambda_n \to 0$ or if **N2** holds. Moreover, observe that, since $m_{0,n} = O(1/\sqrt{k})$, if **Z3** and **Z4** hold, there exists a value M such that if $\sqrt{k}\; m_{0,n} \leq M$ for all n, so the condition $m_{0,n} > a\lambda_n$ for $n \geq n_0$ implies that $\lambda_n = O(1/\sqrt{k})$.

Theorem 2 gives convergence rates for our estimators. Its proof is based on bounds for the increments of empirical processes given in Bühlmann and van de Geer (2011) and Theorem 3.2.5 from van der Vaart and Wellner (1996), which uses the so-called peeling device.

Theorem 2 *Assume that* ***R1*** *holds and that there exist constants* $\eta > 0$ *and* $\tau > 0$ *such that if* $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \eta$*, then* $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$ *for all* $n \geq 1$*. Let* $\widehat{\boldsymbol{\beta}}_n$ *be the estimator defined in* (1) *or* (3)*, and assume that* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \xrightarrow{p} 0$*.*

(a) If ***P1*** *and* ***Z1*** *hold and* $\lambda_n = O(\sqrt{\log p/n})$*, then*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log p}{n}}\right). \tag{9}$$

(b) Under ***P1*** *and* ***Z2****, if* $\lambda_n = O(\sqrt{1/n})$*, then*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p}{n}}\right). \tag{10}$$

(c) Assume that ***P2*** *is satisfied and* $b_n \to 0$*, then:*

(i) If ***Z1*** *holds and* $a_n\sqrt{k} = O(\sqrt{p\log p/n})$*,* (9) *is verified.*
(ii) If ***Z2*** *holds and* $a_n\sqrt{k} = O(\sqrt{p/n})$*,* (10) *is verified.*

An important requirement in Theorem 2 is that there exist positive real numbers η and τ such that $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$ whenever $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \eta$. This inequality states, in some sense, a local strong convexity condition to the population risk $\mathbb{L}$. In this way, unlike van de Geer and Müller (2012), we avoid requiring convexity to the function $t \mapsto \phi(y, t)$, for each fixed y. Lemma 1 gives conditions ensuring that this assumption holds and its proof can be found in Bianco et al. (2022).

Lemma 1 *Assume that* ***Z2*** *to* ***Z5*** *hold and that the loss function* ρ *satisfies* ***R1*** *and* ***R4****. Then, there exist positive constants* η *and* τ *such that* $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$ *when* $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \eta$*.*

Remark 8 From Remark 7, we get that Theorem 2(a) and (b) may be applied to the Ridge, LASSO, Elastic Net, SCAD, and MCP penalties. However, taking into account that LASSO, SCAD, and MCP also verify **P2**, Theorem 2(c) allows to obtain the rates of convergence given in (a) and (b), but with milder assumptions for λ_n. In particular, for the LASSO penalty, to obtain the considered convergence rates, the parameter λ_n must satisfy $\lambda_n\sqrt{n\,k/p} = O(1)$ instead of $\lambda_n\sqrt{n} = O(1)$, while for the SCAD and MCP the required rate for λ_n is easily derived from Remark 7. The differences between uniformly Lipschitz penalties, that is, penalties satisfying **P1** and those verifying **P2** play an important role in the variable selection properties of the estimator. The distinction between the theoretical properties of penalized likelihood estimators using LASSO and folded-concave penalties, such as SCAD and MCP, was already discussed thoroughly by Fan and Peng (2004) and Fan and Lv (2011), for generalized lineal models.

4 Variable Selection and Asymptotic Distribution

In this section, we derive the asymptotic distribution of the considered estimators. In particular, we show that for the SCAD and MCP penalties, the robust penalized estimator has the oracle property, that is, that the penalized M-estimator of the non-null components of $\boldsymbol{\beta}_0$, $\widehat{\boldsymbol{\beta}}_{n,\mathrm{A}}$ has the same asymptotic distribution as that of the non-penalized estimator obtained assuming that the last components of $\boldsymbol{\beta}_0$ are equal to 0 and using this restriction in the logistic regression model. As in other settings, a key step is to derive variable selection properties, that is, to show that the procedure correctly identifies variables related to non-null coefficients. When the penalty parameter has a suitable rate of convergence, the variable selection property is obtained for penalties satisfying inequality (12) below. Even though this inequality trivially holds for the LASSO penalty, the rate conditions on λ_n are not fulfilled for this penalty, as it will be mentioned in Remark 9. With respect to SCAD and MCP, Corollary 1 shows that inequality (12) also holds for these two penalties, while in Remark 10 below we discuss the rates for λ_n required to obtain a procedure with automatic variable selection.

For notation simplicity, given a vector $\mathbf{b} = (\mathbf{b}_1^{\mathrm{T}}, \mathbf{b}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{b}_1 \in \mathbb{R}^k$ and $\mathbf{b}_2 \in \mathbb{R}^{p-k}$, we will denote $I_\lambda(\mathbf{b}_1, \mathbf{b}_2) = I_\lambda(\mathbf{b})$, for any $\lambda > 0$. The proofs of Theorem 3 and Corollaries 1 to 3 are omitted and may be found in Bianco et al. (2022). In particular, that of Theorem 3 follows similar arguments to those considered in the proof of Theorem 3 in Bianco et al. (2021), but adapted to the fact that the dimension increases with the sample size.

Theorem 3 *Let $\widehat{\boldsymbol{\beta}}_n$ be the estimator defined in (1) or (3), where $\phi(y,t)$ is given in (2) and the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies **R3**. Let $\{\ell_n\}_{n\in\mathbb{N}}$ be a sequence such that $\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$, and define $\mathbf{B} = \mathbb{E}\{\Psi^2(y_n, \mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\} = \mathbb{E}\{F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_0)\left[1 - F(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_0)\right]\nu^2(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\}$ and*

$$c_n = \frac{\sqrt{\iota_p(\mathbf{B})}}{\sqrt{n}} + \frac{\iota_p(\mathbf{H})}{\ell_n}\,, \tag{11}$$

where the function ν is defined in (5). Assume that for each $C > 0$, there exist constants $K_C > 0$ and $N_C \in \mathbb{N}$ such that for any $n \geq N_C$ and all vectors $\mathbf{u}_1 \in \mathbb{R}^k$ and $\mathbf{u}_2 \in \mathbb{R}^{p-k}$ satisfying $\|\mathbf{u}_1\|_2^2 + \|\mathbf{u}_2\|_2^2 \leq C^2$, the following inequality holds

$$I_{\lambda_n}\left(\boldsymbol{\beta}_{0,\mathrm{A}} + \frac{\mathbf{u}_1}{\ell_n}, \frac{\mathbf{u}_2}{\ell_n}\right) - I_{\lambda_n}\left(\boldsymbol{\beta}_{0,\mathrm{A}} + \frac{\mathbf{u}_1}{\ell_n}, \mathbf{0}_{p-k}\right) \geq K_C\,\frac{\lambda_n}{\ell_n}\,\|\mathbf{u}_2\|_2\,. \tag{12}$$

Then, if $\lambda_n c_n^{-1} \to \infty$, we have that $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{NA}} = \mathbf{0}_{p-k}) \to 1$.

Remark 9 It is worth mentioning that if $\ell_n = \sqrt{n/p}$ and there exists a constant $K > 0$ not depending on n such that $\max\{\iota_p(\mathbf{H}), \iota_p(\mathbf{B})\} \le K$), then $c_n = O(\sqrt{p/n})$, so $\lambda_n c_n^{-1} \to \infty$ if $\lambda_n\sqrt{n/p} \to \infty$. Moreover, if there exists a constant $K^\star > 0$ such that $\min\{\iota_p(\mathbf{H}), \iota_p(\mathbf{B})\} \ge K^\star$, both conditions $\lambda_n c_n^{-1} \to \infty$ and $\lambda_n\sqrt{n/p} \to \infty$ are equivalent, which implies that, in this case, the order required to the penalty parameter in Theorem 3 is analogous to the one obtained in Bianco et al. (2021), for fixed p.

Recall that, to obtain estimators with rate of convergence $\sqrt{n/p}$, Theorem 2(b) requires that the penalty parameter has order $\lambda_n = O(\sqrt{1/n})$ entailing that $\lambda_n\sqrt{n/p} \to 0$ when the dimension increases with the sample size that collides with the rate condition $\lambda_n\sqrt{n/p} \to \infty$. In particular, since Elastic Net satisfies **P1**, variable selection cannot be derived from Theorems 2(b) and 3.

Finally, when considering LASSO, Theorem 2(c) entails that to attain a rate of convergence $\ell_n = \sqrt{n/p}$, the penalty parameter should satisfy $\lambda_n\sqrt{n/p} = O(1/k)$, which again contradicts the requirement $\lambda_n\sqrt{n/p} \to \infty$ needed to obtain the selection variable property. Several authors have treated the variable selection properties when using the LASSO penalty. Among others, we can mention the discussions given in Fan and Li (2001), Fan and Peng (2004), Leng et al. (2006), Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Zou (2006), Yuan and Lin (2007), and Fan and Lv (2011).

In particular, when considering penalized least squares estimators under a linear regression model, Leng et al. (2006) showed that the LASSO penalty does not give consistent model selection for fixed dimension p, normal errors, and orthogonal designs. In fact, the lack of the variable selection property was already conjectured in Fan and Li (2001) and proved later in Zou (2006). Yuan and Lin (2007) also suggested that LASSO penalty must be carefully used as a variable selection method. Zhao and Yu (2006) derived selection and strong sign consistency under a strong irrepresentable condition for both small and large p settings. As discussed in Zhao and Yu (2006), when the strong irrepresentable condition fails, the non-active covariates are sufficiently correlated with the active ones so as to be picked up by LASSO, leading to the lack of the variable selection property of these penalized estimators. It is worth mentioning that, recently, Lahiri (2021) in a key paper provides necessary and sufficient conditions for variable selection consistency of the LASSO least squares method in high-dimensional linear regression models.

As noted in Fan and Lv (2011) who studied penalized likelihood estimators in the framework of generalized linear models, LASSO penalty does not generally lead to the rate needed to achieve the oracle property. According to the results obtained in Fan and Lv (2011) and in Theorem 3 above, this problem with LASSO penalty arises for both penalized maximum likelihood and robust estimators.

The following corollary states that the variable selection property holds for the SCAD and MCP penalties.

Corollary 1 *Let $\widehat{\boldsymbol{\beta}}_n$ be the estimator defined in (1) or (3), where $\phi(y,t)$ is given in (2) and the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies **R3**. Let $\{\ell_n\}_{n\in\mathbb{N}}$ be a sequence such that $\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$, and define c_n as in (11). Assume that $\lambda_n c_n^{-1} \to \infty$, $\lambda_n \ell_n \to \infty$ and that $I_{\lambda_n}(\boldsymbol{\beta})$ is the SCAD or MCP penalty. Then, $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{NA}} = \mathbf{0}_{p-k}) \to 1$.*

Remark 10 As mentioned in Remark 9, when we only assume that **P1** holds, the order of convergence required to λ_n in Theorem 2(a) and (b) in order to derive convergence rates for the robust penalized estimators is incompatible with the condition $\lambda_n c_n^{-1} \to \infty$. However, according to Theorem 2, when using the SCAD or MCP penalties, convergence rates for $\widehat{\boldsymbol{\beta}}_n$ are obtained by just requiring $\lambda_n \to 0$ whenever $m_{0,n} = \min\{|\beta_{0,j}| : \beta_{0,j} \neq 0\} > m_0$ for every n. According to Remark 7, under **Z3** and **Z4**, $m_{0,n} = O(1/\sqrt{k})$, so $\lambda_n = O(1/\sqrt{k})$, which is not contradictory with the order for λ_n required in Corollary 1. In particular, if $\ell_n = \sqrt{n/p}$ and there exist positive constants K and $K^\star$ such that $K^\star \leq \min\big\{\iota_p(\mathbf{H}), \iota_p(\mathbf{B})\big\} \leq \max\big\{\iota_p(\mathbf{H}), \iota_p(\mathbf{B})\big\} \leq K$, the condition $\lambda_n c_n^{-1} \to \infty$ is equivalent to $n/(k\,p) \to \infty$ when assumptions **Z3** and **Z4** hold, while, if **N2** holds, the condition $\lambda_n c_n^{-1} \to \infty$ implies $m_{0,n}\sqrt{n/p} \to \infty$.

It is worth mentioning that if **Z3** holds, then the condition $\lambda_n \ell_n \to \infty$ required in Corollary 1 is a consequence of $\lambda_n c_n^{-1} \to \infty$.

From Theorems 2 and 3, we can obtain the following corollary that allows to improve the convergence rate of the estimators defined in (1) or (3). First, observe that $I_\lambda(\cdot) : \mathbb{R}^p \to \mathbb{R}$, so in all the previous results the penalties constitute a sequence of functions, not only by their dependence on λ_n, but also because their domains depend on the sample size. However, to avoid the use of heavy notation, we will not make this distinction explicit, so $I_\lambda(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathbb{R}^p$ or $I_\lambda(\mathbf{b})$ with $\mathbf{b} \in \mathbb{R}^k$ will refer to penalties with different domains. For the sake of clarity, we will use the subindex k to indicate vectors in $\mathbb{R}^k$. To state Corollary 2, define

$$\widehat{\mathbf{b}}_k = \frac{1}{n} \operatorname*{argmin}_{\mathbf{b}_k \in \mathbb{R}^k} \sum_{i=1}^{n} \phi(y_{n,i}, \mathbf{x}_{n,i,\mathrm{A}}^{\mathrm{T}} \mathbf{b}_k) + I_{\lambda_n}(\mathbf{b}_k)\,, \tag{13}$$

and consider the following assumption on the penalty function.

P3 If $\mathbf{b}_k \in \mathbb{R}^k$ is such that $\mathbf{b}_k \neq \mathbf{0}_k$, then $I_\lambda(\mathbf{b}_k) = I_\lambda\big(\mathbf{b}_k, \mathbf{0}_{p-k}\big)$.

Corollary 2 *Let $\widehat{\boldsymbol{\beta}}_n$ be the estimator defined in (1) or in (3). Assume that $\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{NA}} = \mathbf{0}_{p-k}) \to 1$ when $n \to \infty$. Assume that **P3** holds and that $\|\widehat{\mathbf{b}}_k - \boldsymbol{\beta}_{0,\mathrm{A}}\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$, where $\widehat{\mathbf{b}}_k$ is defined through (13). Then, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$.*

Remark 11 First, observe that assumption **P3** holds for the LASSO, SCAD, and MCP penalties. More generally, it holds for every penalty function that can be written as $I_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} J_\lambda(|\beta_j|)$, where $J_\lambda(0) = 0$.

On the other hand, Theorem 2 gives conditions that guarantee $\|\widehat{\mathbf{b}}_k - \boldsymbol{\beta}_{0,\mathrm{A}}\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$. In fact, to obtain $\|\widehat{\mathbf{b}}_k - \boldsymbol{\beta}_{0,\mathrm{A}}\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$, assumptions **Z1** to **Z5** can be replaced by analogous versions in which only the first k coordinates of $\mathbf{x}_n$ and $\boldsymbol{\beta}_0$ are considered. Moreover, denoting $\mathbf{H}_\mathrm{A} = \mathbf{H}_{n,\mathrm{A}} = \mathbb{E}\big(\mathbf{x}_{n,\mathrm{A}}\mathbf{x}_{n,\mathrm{A}}^\mathrm{T}\big)$, we get that **Z4** is equivalent to $\boldsymbol{\beta}_{0,\mathrm{A}}^\mathrm{T}\mathbf{H}_\mathrm{A}\boldsymbol{\beta}_{0,\mathrm{A}} \le K_2^2$, which already gives a condition for the first k components of $\mathbf{x}_n$ and $\boldsymbol{\beta}_0$.

We now proceed to study the asymptotic distribution of the estimator $\widehat{\boldsymbol{\beta}}_n$ defined through (1). Theorem 4 states that, for certain penalties that include SCAD and MCP, the robust penalized M-estimator has the oracle property.

Given a vector $\mathbf{b}_k \in \mathbb{R}^k$, denote $\mathbf{A}_\mathrm{A}^{(k)}(\mathbf{b}_k) \in \mathbb{R}^{k\times k}$ and $\mathbf{B}_\mathrm{A}^{(k)}(\mathbf{b}_k) \in \mathbb{R}^{k\times k}$ the matrices

$$\mathbf{A}_\mathrm{A}^{(k)}(\mathbf{b}_k) = \mathbb{E}\Big[\chi(y_n, \mathbf{x}_{n,\mathrm{A}}^\mathrm{T}\mathbf{b}_k)\mathbf{x}_{n,\mathrm{A}}\mathbf{x}_{n,\mathrm{A}}^\mathrm{T}\Big] \qquad \text{and} \qquad \mathbf{B}_\mathrm{A}^{(k)}(\mathbf{b}_k) = \mathbb{E}\Big[\Psi^2(y_n, \mathbf{x}_{n,\mathrm{A}}^\mathrm{T}\mathbf{b}_k)\mathbf{x}_{n,\mathrm{A}}\mathbf{x}_{n,\mathrm{A}}^\mathrm{T}\Big].$$

In addition, we define $\mathbf{A}_\mathrm{A}^{(k)} = \mathbf{A}_\mathrm{A}^{(k)}(\boldsymbol{\beta}_{0,\mathrm{A}})$ and $\mathbf{B}_\mathrm{A}^{(k)} = \mathbf{B}_\mathrm{A}^{(k)}(\boldsymbol{\beta}_{0,\mathrm{A}})$, where the latter equals the matrix $\mathbf{B}_\mathrm{A}$ defined in **Z7**. Note that, in this case, given $\mathbf{v}_k \in \mathbb{R}^k$ with $\|\mathbf{v}_k\|_2 = 1$, the value $t^2 = \mathbf{v}_k^\mathrm{T}\mathbf{B}_\mathrm{A}^{(k)}\mathbf{v}_k$ also depends on n. However, to simplify the notation in Theorem 4 and Corollary 3, we will write t instead of t_n. Besides, for $\mathbf{b}_k = (b_1, \dots, b_k)^\mathrm{T} \in \mathbb{R}^k$ with $b_j \ne 0$, $1 \le j \le k$, we define $\nabla I_\lambda(\mathbf{b}_k) = \partial I_\lambda\big(\mathbf{b}_k, \mathbf{0}_{p-k}\big)/\partial\mathbf{b}_k$.

Theorem 4 *Let* $\mathbf{v}_k \in \mathbb{R}^k$ *be a vector such that* $\|\mathbf{v}_k\|_2 = 1$, *and denote* $t^2 = \mathbf{v}_k^\mathrm{T}\mathbf{B}_\mathrm{A}^{(k)}\mathbf{v}_k$. *Assume that* $\lim_{n\to\infty}\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\mathrm{NA}} = \mathbf{0}_{p-k}) = 1$ *and that* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$. *Moreover, assume that* ***N1****,* ***R5****,* ***Z6****, and* ***Z7*** *hold. Then, if* $k^2/n \to 0$ *and*

$$\sqrt{n}\|\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\mathrm{A}})\|_2 \xrightarrow{p} 0, \tag{14}$$

we have that $\sqrt{n}\, t^{-1}\mathbf{v}_k^\mathrm{T}\mathbf{A}_\mathrm{A}^{(k)}(\widehat{\boldsymbol{\beta}}_{n,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}) \xrightarrow{D} N(0,1)$.

Finally, Corollary 3 shows that the conclusion of Theorem 4 holds when considering SCAD or MCP. Its proof is obtained by showing that MCP and SCAD satisfy (14) under assumptions **N2** and **N3**.

Corollary 3 *Let* $\mathbf{v}_k \in \mathbb{R}^k$ *be a vector such that* $\|\mathbf{v}_k\|_2 = 1$ *and* $t^2 = \mathbf{v}_k^\mathrm{T}\mathbf{B}_\mathrm{A}^{(k)}\mathbf{v}_k$. *Assume that* $\lim_{n\to\infty}\mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\mathrm{NA}} = \mathbf{0}_{p-k}) = 1$ *and that* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\sqrt{k/n})$. *Furthermore, assume that* ***N1****,* ***N2****,* ***N3****,* ***R5****,* ***Z6****, and* ***Z7*** *hold. If* $k^2/n \to 0$ *and* I_{λ_n} *is the SCAD or MCP penalties, then* $\sqrt{n}\, t^{-1}\mathbf{v}_k^\mathrm{T}\mathbf{A}_\mathrm{A}^{(k)}(\widehat{\boldsymbol{\beta}}_{n,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}) \xrightarrow{D} N(0,1)$.

Remark 12 It is worth mentioning that the asymptotic normality stated in Theorem 4 and Corollary 3 still holds if we require as convergence rate the rate derived in Theorem 2, that is, $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(\sqrt{p/n})$, and we replace the condition $k^2/n \to 0$ by $p^2/n \to 0$ and assumptions **N1** and **N3** by the requirements $m_{0,n}\sqrt{n/p} \to \infty$ and $p/n = O(\lambda_n^2)$, respectively.

5 General Comments

An important issue when implementing penalized estimators is the choice of the penalty parameter λ_n, since it tunes the model complexity. Efron et al. (2004), Meinshausen (2007), and Chi and Scott (2014) discussed this topic. Bianco et al. (2021) recommended a robust K-fold criterion to select the penalty parameter and showed the importance of considering a robust cross-validation criterion in order to achieve reliable prediction and preserve the robustness of the whole procedure. For the sake of completeness, we briefly describe their robust proposal. We begin by randomly splitting the data into K disjoint subsets of approximately equal sizes, with indices $\mathcal{I}_j$, $1 \le j \le K$, the jth subset having size $n_j \ge 2$, so that $\bigcup_{j=1}^K \mathcal{I}_j = \{1, \dots, n\}$ and $\sum_{j=1}^K n_j = n$. Let $\widetilde{\Lambda} \subset \mathbb{R}$ be the set of possible values for λ to be considered. Denote as $\widehat{\boldsymbol{\beta}}_\lambda^{(j)}$ the robust penalized estimator of $\boldsymbol{\beta}_0$, computed with penalty parameter $\lambda \in \widetilde{\Lambda}$ without using the observations with indices in $\mathcal{I}_j$. In order to provide a robust alternative to the classical K-fold procedure, it seems natural to apply the same loss function ρ as in (1) to the predicted deviance residuals $d(y_{n,i}, \mathbf{x}_{n,i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\lambda^{(j)})$. Hence, a robust cross-validation criterion may be defined as

$$RCV(\lambda) = \frac{1}{n} \sum_{1 \le j \le K} \sum_{i \in \mathcal{I}_j} \phi(y_{n,i}, \mathbf{x}_{n,i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_\lambda^{(j)}) \, .$$

The penalty parameter λ_n is obtained through the minimization of $RCV(\lambda)$ over $\widetilde{\Lambda}$. When $K = n$, this method leads to the robust version of the leave-one-out cross-validation procedure, which is a popular choice, but with a more expensive computational cost. Bianco et al. (2021) numerically showed that, even when $\boldsymbol{\beta}_0$ is robustly estimated, the classical cross-validation criterion obtained using $\rho(t) = t$ may lead to a poor variable selection result.

The penalized M-estimators may be implemented using a cyclical descent algorithm. A detailed description of the algorithm as well as a suggestion on how to choose the initial value to start it is given in Section S.6 in the supplementary material of Bianco et al. (2021). The code allowing to compute the estimators is publicly available online at https://github.com/gonzalochebi/penMlogistic.git.

Usually, when considering robust procedures for linear models, the estimators are calibrated to attain a given efficiency preserving their robustness properties. Some generalized linear models have particular features in this aspect, and the logistic regression one is not an exception. Note that already in Avella-Medina and Ronchetti (2018) the calibration problems arising under a Poisson model are mentioned in their numerical study.

To illustrate the logistic case, let us consider the situation where neither the number of active variables k nor the distribution of $\mathbf{x}_{n,\mathrm{A}}$ depends on n. The latter happens, for instance, when $\boldsymbol{\Gamma}_p = \mathbf{I}_p$ in **Z5** (see Remark 3). Denote as $\mathbf{x}^{(k)}$ a random vector with the same distribution than $\mathbf{x}_{n,\mathrm{A}}$. The oracle property stated in Corollary 3 implies that, when considering SCAD or MCP penalties, $\sqrt{n}\ (\widehat{\boldsymbol{\beta}}_{n,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})$ has the same asymptotic distribution as the M-estimator we would obtain with observations following the same distribution as $(y, \mathbf{x}^{(k)})$, where $y|\mathbf{x}^{(k)} \sim F(\mathbf{x}^{(k)\,\mathrm{T}}\boldsymbol{\beta}_{0,\mathrm{A}})$. More precisely, its asymptotic covariance matrix equals $(\mathbf{A}_{\mathrm{A}}^{(k)})^{-1}\mathbf{B}_{\mathrm{A}}^{(k)}(\mathbf{A}_{\mathrm{A}}^{(k)})^{-1}$, where the matrices $\mathbf{A}_{\mathrm{A}}^{(k)} = \mathbf{A}_{\mathrm{A}}^{(k)}(\boldsymbol{\beta}_{0,\mathrm{A}})$ and $\mathbf{B}_{\mathrm{A}}^{(k)} = \mathbf{B}_{\mathrm{A}}^{(k)}(\boldsymbol{\beta}_{0,\mathrm{A}})$ defined in Sect. 4 may be written as $\mathbf{A}_{\mathrm{A}}^{(k)} = \mathbb{E}\big(F\big(\boldsymbol{\beta}_{0,\mathrm{A}}^{\mathrm{T}}\mathbf{x}^{(k)}\big)\big[1 - F\big(\boldsymbol{\beta}_{0,\mathrm{A}}^{\mathrm{T}}\mathbf{x}^{(k)}\big)\big]\nu\big(\boldsymbol{\beta}_{0,\mathrm{A}}^{\mathrm{T}}\mathbf{x}^{(k)}\big)\,\mathbf{x}^{(k)}\mathbf{x}^{(k)\,\mathrm{T}}\big)$, with the function ν is given in (5), and $\mathbf{B}_{\mathrm{A}}^{(k)} = \mathbb{E}\big(F(\boldsymbol{\beta}_{0,\mathrm{A}}^{\mathrm{T}}\mathbf{x}^{(k)})\big[1 - F(\boldsymbol{\beta}_{0,\mathrm{A}}^{\mathrm{T}}\mathbf{x}^{(k)})\big]\nu^2(\boldsymbol{\beta}_{0,\mathrm{A}}^{\mathrm{T}}\mathbf{x}^{(k)})\,\mathbf{x}^{(k)}\mathbf{x}^{(k)\,\mathrm{T}}\big)$. These expressions show that, as in the fixed dimension case, in the logistic regression model, the asymptotic efficiency of the robust estimator of the active components depends on the true value of the parameter. Hence, in practical applications, unlike under the usual linear model, these estimators cannot be universally calibrated.

A numerical study designed to compare the robustness properties of the penalized M-estimators defined through (1), for finite samples, can also be found in Bianco et al. (2021) and its supplementary material. With respect to the summary measures considered therein, their simulation results reveal that, either for clean or contaminated samples, the estimators using the penalties SCAD or MCP perform quite similarly and outperform the procedure based on LASSO. The situation of independent and correlated variables, as well as two values for $\boldsymbol{\beta}_{0,\mathrm{A}}$, is also reported in their supplementary file.

Appendix 1: Proofs of Remark 4 and of the Results in Sect. 3

Proof of Remark 4 Along this proof, for clarity, we strength the dependence of the dimension p on n. Let $\mathbf{w}_{p_n} = \boldsymbol{\Gamma}_{p_n}^{-1/2}\mathbf{x}_n$. It is enough to show that $\mathbf{w}_{p_n} \sim S\ \mathbf{z}_{p_n}$ for some positive random variable S whose distribution does not depend on n and some p_n-dimensional random vector $\mathbf{z}_{p_n}$ independent of S and such that $\mathbf{z}_{p_n} \sim N(\mathbf{0}_{p_n}, \mathbf{I}_{p_n})$.

Using that the random vector $\mathbf{w}_{p_n} = (w_{n,1}, \ldots, w_{n,p_n})^{\mathrm{T}} = \boldsymbol{\Gamma}_{p_n}^{-1/2}\mathbf{x}_n$, we get that $\mathbf{w}_{p_n}$ has a spherical distribution in $\mathbb{R}^{p_n}$ with characteristic function given by $\phi_{\mathbf{w}_{p_n}}(\mathbf{t}) = \xi(\|\mathbf{t}\|^2)$, $\mathbf{t} \in \mathbb{R}^{p_n}$. As it is well known, the function ξ is the characteristic function of $w_{n,1}$. The fact that $\mathbf{w}_{p_n}$ has a spherical distribution entails that $\mathbf{w}_{p_n} = T_{p_n}\mathbf{u}_{p_n}$, where $\mathbf{u}_{p_n} = \mathbf{w}_{p_n}/\|\mathbf{w}_{p_n}\|$ has a uniform distribution on the p_n-dimensional unit sphere, and $T_{p_n} = \|\mathbf{w}_{p_n}\|$ is a non-negative random variable independent from $\mathbf{u}_{p_n}$. The distribution of $\mathbf{u}_{p_n}$ may be represented as $\mathbf{u}_{p_n} \sim \mathbf{z}_{p_n}/\|\mathbf{z}_{p_n}\|$, where $\mathbf{z}_{p_n} \sim N(\mathbf{0}_{p_n}, \mathbf{I}_{p_n})$ and is independent of T_n. Hence, we have that $\mathbf{w}_{p_n} \sim D_n S_n \mathbf{z}_{p_n}$, where $D_n = \sqrt{p_n}/\|\mathbf{z}_{p_n}\|$, $S_n = T_n/\sqrt{p_n}$, and $\mathbf{z}_{p_n} = (z_1, \ldots, z_{p_n})^{\mathrm{T}} \sim N(\mathbf{0}_{p_n}, \mathbf{I}_{p_n})$ independent of S_n. Thus, $w_{n,1} \sim D_n S_n z_1$, with $z_1 \sim N(0, 1)$ independent of S_n. The weak law of large numbers and the fact that $p_n \to \infty$ as $n \to \infty$ entail that $D_n \xrightarrow{p} 1$. Since the distributions of $w_{n,1}$ and z_1 do not depend on n, S_n must converge in distribution to a random variable S, with S being independent of z_1. Thus, $w_{n,1} \sim S\, z_1$, and the fact that $\mathbf{w}_{p_n}$ is spherically distributed allows to conclude that $\mathbf{w}_{p_n} \sim S\, \mathbf{z}_{p_n}$, which is a scale mixture of normals, where the distribution of S does not depend on n, concluding the proof. Note that, as mentioned in Kingman (1972), the spherical symmetry of $\mathbf{w}_{p_n}$ entails that ξ is a radial characteristic function of dimension p_n, which is only possible if $\xi(t) = \int_0^\infty \exp\left(-u\, t^2\right) dG(u)$ for some distribution function G, and this property also leads to the desired result. ■

From now on, in order to lighten the notation and when there is no confusion, we will omit the subindex n in $(y_{n,i}, \mathbf{x}_{n,i})$, $1 \le i \le n$, as well as in $(y_n, \mathbf{x}_n)$ that has the same distribution as $(y_{n,1}, \mathbf{x}_{n,1})$. Lemmas 2 to 4 in this section are useful to prove our theoretical results; their proofs can be found in Bianco et al. (2022).

To prove the consistency of the proposed estimators when $p \to \infty$, we will make use of Theorem 2.14.1 from van der Vaart and Wellner (1996). It is worth to remind that since the dimension p diverges to infinity, the usual limit theorems such as the law of large numbers or the central limit theorem are no longer useful. Instead, in this context, we will need explicit bounds for the empirical process for a fixed n, as the one obtained in Lemma 2 for the family of functions $\mathcal{F} = \{f(y, \mathbf{x}) = \phi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^p\}$. Given a function $f : \mathbb{R}^{p+1} \to \mathbb{R}$, we use the usual empirical process notation, that is, $P_n f = (1/n)\sum_{i=1}^n f(y_i, \mathbf{x}_i)$ and $Pf = \mathbb{E}[f(y, \mathbf{x})]$.

Lemma 2 *Let ϕ be defined as in (2) and $\mathcal{F} = \{f(y, \mathbf{x}) = \phi(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^p\}$. Under* ***R1*** *and* ***R2****, we have that $\mathbb{E}\Big[\sup_{f\in\mathcal{F}} |(P_n - P)(f)|\Big] \le C_1\sqrt{p/n}$ for some constant C_1 independent of n and p.*

Lemma 3 *For $(\pi, \pi_0) \in (0, 1) \times [0, 1]$, define $M(\pi, \pi_0)$*

$$M(\pi, \pi_0) = \pi_0 \rho(-\log \pi) + (1-\pi_0)\rho(-\log(1-\pi)) + G(\pi) + G(1-\pi)\,. \tag{15}$$

(a) If assumptions ***R1*** *and* ***R2*** *hold, then the function $M(\pi, \pi_0)$ can be extended to a continuous function on $[0, 1] \times [0, 1]$.*

(b) If assumptions ***R1*** *and* ***R4*** *hold, then there exists a constant $\tau > 0$ such that, for each $0 < \pi < 1$, $M(\pi, \pi_0) - M(\pi_0, \pi_0) \ge \tau(\pi - \pi_0)^2$.*

Lemma 4 *Let $\mathbf{z} = (Z_1, Z_2)^{\mathrm{T}} \in \mathbb{R}^2$ be a random vector with a centred elliptical distribution and characteristic function $\phi_{\mathbf{z}}(\mathbf{u}) = \xi(\mathbf{u}^{\mathrm{T}}\boldsymbol{\Upsilon}\mathbf{u})$. Assume that $\mathbb{E}(Z_j^2) < \infty$, for $j = 1, 2$, and denote $\boldsymbol{\Sigma} = \mathbf{COV}(\mathbf{z})$, that is, $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = -2\xi'(0)\boldsymbol{\Upsilon}$, with $|\rho| \le 1$. Additionally, assume that $\sigma_2 > 0$ and that there exists a constant $K_2 > 0$ such that $\sigma_2 \le K_2$ and the distribution of $\mathbf{Z}$ verifies*

$\mathbb{E}\big\{[F(Z_1) - F(Z_2)]^2\big\} < [F(4\,K_2) - F(2\,K_2)]^2/4$, *where* $F(t) = \exp(t)/(1 + \exp(t))$. *Then:*

(a) There exists a constant C_0 *that only depends on* ξ *such that* $\sigma_1 \le C_0\,K_2$.
(b) There exists a constant C_2 *that only depends on* K_2 *and* ξ *such that* $\mathbb{E}[(Z_1 - Z_2)^2] \le C_2\,\mathbb{E}[(F(Z_1) - F(Z_2))^2]$.

Proof of Theorem 1 We will prove (a) only for the unrestricted estimator, and the proof for the restricted one is similar. Using the definition of $\widehat{\boldsymbol{\beta}}_n$, we have that

$$L_n(\widehat{\boldsymbol{\beta}}_n) \le L_n(\widehat{\boldsymbol{\beta}}_n) + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) \le L_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}_0)\,,$$

which implies $\mathbb{L}(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\boldsymbol{\beta}_0) \le \big[L_n(\boldsymbol{\beta}_0) - \mathbb{L}(\boldsymbol{\beta}_0)\big] - \big[L_n(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\widehat{\boldsymbol{\beta}}_n)\big] + I_{\lambda_n}(\boldsymbol{\beta}_0)$. Let C_1 be the constant from Lemma 2, which we will assume, without loss of generality, to be greater than one. Consider the event $\mathcal{A}_{n,T} = \big\{\sup_{\boldsymbol{\beta}} |L_n(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta})| \le C_1\,T\sqrt{p/n}\big\}$. From Lemma 2 and Markov's inequality, we get that $\mathbb{P}(\mathcal{A}_{n,T}) \ge 1 - 1/T$ for $T > 1$. Thus, restricting to the event $\mathcal{A}_{n,T}$, we obtain

$$\mathbb{L}(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\boldsymbol{\beta}_0) \le 2\,C_1\,T\sqrt{\frac{p}{n}} + I_{\lambda_n}(\boldsymbol{\beta}_0) \le 2\,C_1\,T\left\{\sqrt{\frac{p}{n}} + I_{\lambda_n}(\boldsymbol{\beta}_0)\right\}. \tag{16}$$

Straightforward calculations show that

$$\mathbb{L}(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\boldsymbol{\beta}_0) = \mathbb{E}\Big\{M(F(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n), F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)) - M(F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0), F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0))\Big|(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\Big\}\,,$$

with M defined in (15). Then, using Lemma 3, we obtain that there exists a constant $\tau > 0$ independent from n such that

$$\mathbb{L}(\widehat{\boldsymbol{\beta}}_n) - \mathbb{L}(\boldsymbol{\beta}_0) \ge \tau\;\mathbb{E}\Big\{\big[F(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n) - F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\big]^2\Big|(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\Big\} = \tau\,d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0)\,,$$

which together with (16) concludes the proof for (a).

To prove (b), observe that if $\max\{\|\widehat{\boldsymbol{\beta}}_n\|_1, \|\boldsymbol{\beta}_0\|_1\} \le R$ and $\|\mathbf{x}\|_\infty \le A$, then the Hölder inequality implies $\max\{|\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n|, |\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0|\} \le A\,R$. Using the fact that $F'(t)$ is an even function, increasing in $(-\infty, 0]$ and decreasing in $[0, \infty)$, we obtain that $\mathbb{E}\big[(F(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n) - F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0))^2\big] \ge (F'(A\,R))^2\;\iota_1(\mathbf{H})\;\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2$. In particular, when (**Z3**) holds, we have $\mathbb{E}\big[(F(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n) - F(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0))^2\big] \ge \tau_1\;(F'(A\,R))^2\;\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2$, and the desired result follows from (a).

Finally, we prove (c). It suffices to show that given $\varepsilon > 0$ and $\delta > 0$, there exists n_0 such that if $n \ge n_0$, then

$$\mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 \le \varepsilon) > 1 - \delta\,. \tag{17}$$

Assumption **Z5** implies that, for every $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{z}_{\boldsymbol{\beta}} = (\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)^{\mathrm{T}}$ has a centred elliptical distribution with finite second moments and generating function ξ. From assumption **Z3**, we obtain that $\text{VAR}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) \neq 0$. On the other hand, since **Z4** holds, $\text{VAR}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0) \leq K_2$, where K_2 does not depend on n.

Using that $p/n \to 0$ and $I_{\lambda_n}(\boldsymbol{\beta}_0) \to 0$, from (a), we conclude that $d_n(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \xrightarrow{p} 0$. Define the event $\mathcal{B}_n = \left\{d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq [F(4K_2) - F(2K_2)]^2/8 \right\}$. If $\omega \in \mathcal{B}_n$, then $\mathbf{z}_{\widehat{\boldsymbol{\beta}}(\omega)} = (\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_n(\omega), \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)^{\mathrm{T}}$ satisfies the conditions of Lemma 4. Hence, from item (b) of that Lemma, there exists C_2 that only depend on ξ and K_2 (and is independent from ω and n) such that $D_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq C_2 d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0)$, where $D_n^2(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = \mathbb{E}[(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)^2] = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. Moreover, define the event $\mathcal{A}_{\varepsilon,n} = \left\{d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq \varepsilon\, \tau_1/C_2\right\}$, where τ_1 is given in **Z3**. The fact that $d_n(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \xrightarrow{p} 0$ implies that $\lim_n \mathbb{P}(\mathcal{A}_{\varepsilon,n}) = \lim_n \mathbb{P}(\mathcal{B}_n) = 1$, so $\lim_n \mathbb{P}(\mathcal{A}_{\varepsilon,n} \cap \mathcal{B}_n) = 1$, and there exists n_0 such that if $n \geq n_0$, $\mathbb{P}(\mathcal{A}_{\varepsilon,n} \cap \mathcal{B}_n) > 1 - \delta$. Note that for any $\omega \in \mathcal{A}_{\varepsilon,n} \cap \mathcal{B}_n$, we have that $D_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq C_2\ d_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \leq \varepsilon \tau_1$. Besides, $D_n^2(\widehat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}_0) \geq \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2\ \iota_1(\mathbf{H}) \geq \tau_1 \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2$, which implies $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 < \varepsilon$. Thus, (17) holds, concluding the proof. ■

Proof of Theorem 2 We will only prove the result for the estimator $\widehat{\boldsymbol{\beta}}_n$ defined in (1), since the proof for the restricted estimator given in (3) is analogous. It is worth mentioning that **R1** implies $\mathbb{L}(\boldsymbol{\beta}) < \infty$ for all $\boldsymbol{\beta}$.

Let us show (a). Define $v_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta})$ and $\ell_n = \sqrt{n/(p \log p)}$. We will begin by bounding the increments of the empirical process v_n and for that aim define $\gamma(y, s) = \phi(y, s)$ and $y \in \{0, 1\}$. Observe that $\gamma(y, s)$ is differentiable with respect to its second argument with derivative $\gamma'(y, s) = \Psi(y, s)$, where $\Psi(y, t) = \partial\phi(y, t)/\partial t = -[y - F(t)]\nu(t)$, so $\|\gamma'\|_\infty \leq 4\|\psi\|_\infty < \infty$. The mean value theorem implies that $|\gamma(y, s) - \gamma(y, \widetilde{s})| \leq C_\gamma |s - \widetilde{s}|$, for any $s, \widetilde{s} \in \mathbb{R}$ with $C_\gamma = 4\|\psi\|_\infty$. Thus, Lemma 14.20 from Bühlmann and van de Geer (2011) allows to conclude that for every $M > 0$,

$$\mathbb{E}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1 \leq M} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \right) \leq 4MC_\gamma \sqrt{\frac{2\log(2p)}{n}}\, \mathbb{E}\left(\max_{1 \leq j \leq p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \right) \leq MC_1 \sqrt{\frac{\log p}{n}}\,,$$

where the last inequality follows from **Z1** and the constant C_1 does not depend on neither n nor p.

Using that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \delta$ implies $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \leq \sqrt{p}\,\delta$, we obtain

$$\mathbb{E}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2 \leq \delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \right) \leq \mathbb{E}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1 \leq \sqrt{p}\,\delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \right) \leq \frac{C_1\,\delta}{\ell_n}\,.$$

Thus, from Markov's inequality, we conclude that for each $C > 0$,

$$\mathbb{P}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2 \leq \delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| > C \right) \leq \frac{C_1\,\delta}{\ell_n\, C}\,. \tag{18}$$

The proof follows using the same arguments as those considered in the proof of Theorem 3.2.5 in van der Vaart and Wellner (1996) and is based on the so-called peeling device. More precisely, let $c_n = \mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq \eta)$, where $\eta > 0$ is such that $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$ for each $n \geq 1$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \eta$. Since $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \xrightarrow{p} 0$, we have that $c_n \to 0$. For $j \in \mathbb{N}$, define the sets $A_{n,j} = \{\boldsymbol{\beta} \in \mathbb{R}^p : 2^{j-1} \leq \ell_n \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq 2^j\}$. Let $M \in \mathbb{N}$. Using that $\widehat{\boldsymbol{\beta}}_n$ minimizes $L_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta})$, we obtain $L_n(\widehat{\boldsymbol{\beta}}_n) + I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n) \leq L_n(\boldsymbol{\beta}_0) + I_{\lambda_n}(\boldsymbol{\beta}_0)$. Thus, after some straightforward calculations and denoting $w_n(\boldsymbol{\beta}) = v_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta}) + \mathbb{L}(\boldsymbol{\beta})$, we get

$$\mathbb{P}(\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq 2^M) \leq c_n + \sum_{\substack{j \geq M+1 \\ 2^j \leq \ell_n \eta}} \mathbb{P}\Big(\widehat{\boldsymbol{\beta}}_n \in A_{n,j}\Big) \leq c_n + \sum_{\substack{j \geq M+1 \\ 2^j \leq \ell_n \eta}} \mathbb{P}\Big(\inf_{\boldsymbol{\beta} \in A_{n,j}} w_n(\boldsymbol{\beta}) - w_n(\boldsymbol{\beta}_0) \leq 0\Big) .$$

Note that **P1** implies $I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0) \geq -|I_{\lambda_n}(\boldsymbol{\beta}) - I_{\lambda_n}(\boldsymbol{\beta}_0)| \geq -\lambda_n K \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1$. Besides, given $\boldsymbol{\beta} \in A_{n,j}$, $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq \eta$ if $2^j \leq \ell_n \eta$, so $\mathbb{L}(\boldsymbol{\beta}) - \mathbb{L}(\boldsymbol{\beta}_0) \geq \tau \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$. Then, if $\boldsymbol{\beta} \in A_{n,j}$,

$$w_n(\boldsymbol{\beta}) - w_n(\boldsymbol{\beta}_0) \geq -|v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| - \lambda_n K \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \tau \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 , \tag{19}$$

allowing to conclude that $\mathbb{P}(\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq 2^M) \leq c_n + d_n$, where $d_n = \sum_{j \geq M+1 \, , \, 2^j \leq \ell_n \eta} d_{n,j}$ with $d_{n,j} = \mathbb{P}\Big(-\sup_{\boldsymbol{\beta} \in A_{n,j}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)|$ $-K \lambda_n \sup_{\boldsymbol{\beta} \in A_{n,j}} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \tau \inf_{\boldsymbol{\beta} \in A_{n,j}} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \leq 0\Big)$. Observe that if $\boldsymbol{\beta} \in A_{n,j}$, $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \geq 2^{2j-2}/\ell_n^2$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \leq \sqrt{p}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \sqrt{p}\, 2^j/\ell_n$, then $-K \lambda_n \sup_{\boldsymbol{\beta} \in A_{n,j}} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 + \tau \inf_{\boldsymbol{\beta} \in A_{n,j}} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 \geq \alpha_n$, where $\alpha_n = -K \sqrt{p}\, \lambda_n 2^j/\ell_n + \tau\, 2^{2j-2}/\ell_n^2$, which entails that $d_{n,j} \leq \mathbb{P}\big(\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq 2^j/\ell_n} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \geq \alpha_n\big)$. Since $\lambda_n = O(\sqrt{\log p/n})$, we get that there exists a constant $D > 0$ such that $\lambda_n \leq D\sqrt{\log p/n}$ for all n, so choosing $M \geq 1 + \log\big(8\, K\, D \tau^{-1}\big)/\log 2 = M_0$, we have that $\alpha_n > 0$. Using (18), we obtain that for all $j \geq M+1$, $d_{n,j} \leq C_1 2^j/(\ell_n^2 \alpha_n)$. From $\lambda_n \leq D\sqrt{\log p/n}$ for all n, we conclude $\lambda_n \sqrt{p} \leq D/\ell_n$, which implies that $\ell_n^2 \alpha_n \geq 2^j\big(\tau 2^{j-2} - K D\big) > \tau\, 2^{2j}/8$ if $j \geq M+1$, so $d_{n,j} \leq 2^{-j}\ (8\, C_1)/\tau$. Given $\varepsilon > 0$, let $N_\varepsilon \in \mathbb{N}$ be such that if $n \geq N_\varepsilon$, $c_n \leq \varepsilon/2$. Besides, let $M_\varepsilon \in \mathbb{N}$, be such that $M_\varepsilon \geq M_0$ and $\sum_{j \geq M_\varepsilon} 2^{-j} < \tau\varepsilon/(16 C_1)$. Hence, for any $n \geq N_\varepsilon$, we have $\mathbb{P}\big(\ell_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \geq 2^{M_\varepsilon}\big) \leq \varepsilon$, which concludes the proof of (a).

To derive (b), define $\ell_n = \sqrt{n/p}$ and denote ι_p the maximum eigenvalue of $\mathbf{H}$. Note that for any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \delta$, we have that $\mathbb{E}[(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)^2] \leq \delta^2 \iota_p$. Lemma 14.19 in Bühlmann and van de Geer (2011) implies that

$$e_n = \mathbb{E}\Bigg(\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq \delta} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)|\Bigg) \leq \mathbb{E}\Bigg(\sup_{\mathbb{E}[(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)^2] \leq \delta^2 \iota_p} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)|\Bigg) \leq 4 C_\gamma \delta \sqrt{\iota_p} \sqrt{\frac{p}{n}} .$$

Assumption **Z2** ensures that $\iota_p(\mathbf{H}) \le K_1$ for all n; hence, $e_n \le 4C_\gamma\sqrt{K_1}\delta\,\sqrt{p/n} = 4C\sqrt{K_1}\delta/\ell_n$. The proof follows now using the same arguments considered above in the proof of (a).

To prove (c)(i), take $\ell_n = \sqrt{n/(p\log p)}$. From the inequality chain used in (a), when bounding (19) but using (8) instead of assumption **P1**, we obtain that $V_n(\boldsymbol{\beta}) = w_n(\boldsymbol{\beta}) - w_n(\boldsymbol{\beta}_0)$ can be bounded as $V_n(\boldsymbol{\beta}) \ge -|v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| + \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 - a_n\sqrt{k}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 - b_n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2$. Hence, if $\alpha_n = \tau\, 2^{2j-2}/\ell_n^2 - a_n\sqrt{k}\,2^j/\ell_n - b_n\,2^{2j}/\ell_n^2$, we have that $V_n(\boldsymbol{\beta}) \ge -\sup_{\boldsymbol{\beta}\in A_{n,j}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| + \alpha_n$, for any $\boldsymbol{\beta} \in A_{n,j}$. Thus, $\mathbb{P}(\ell_n\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \ge 2^M) \le c_n + d_n$, where $c_n = \mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \ge \widetilde{\eta})$, $d_n = \sum_{j\ge M+1,\, 2^j\le \ell_n\widetilde{\eta}} d_{n,j}$, with $d_{n,j} = \mathbb{P}(\inf_{\boldsymbol{\beta}\in A_{n,j}} V_n(\boldsymbol{\beta}) \le 0) \le \mathbb{P}(\sup_{\boldsymbol{\beta}\in A_{n,j}} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \ge \alpha_n)$ and $\widetilde{\eta} = \min(\eta, \widetilde{\delta})$ with $\widetilde{\delta}$ given in **P2**. The fact that $a_n\sqrt{k} = O(1/\ell_n)$ entails that there exists $D > 0$ such that $a_n\sqrt{k} \le D/\ell_n$ for all n. Let $n_0 \in \mathbb{N}$ be such that for any $n \ge n_0$, $b_n \le \tau/8$, and let $M \ge 1 + \log\bigl(16\,D\tau^{-1}\bigr)/\log 2 = M_0$. Then, if $n \ge n_0$ and $M \ge M_0$, we get that $\alpha_n \ge \tau\,2^{2j}/(16\,\ell_n^2)$. Thus, $d_{n,j} \le \mathbb{P}(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_2\le 2^j/\ell_n} |v_n(\boldsymbol{\beta}) - v_n(\boldsymbol{\beta}_0)| \ge \tau\,2^{2j}/(16\ell_n^2))$, which together with (18) entails that $d_{n,j} \le (16\,C_1)/(\tau\,2^j)$, and the result follows as in (a).

Finally, the proof of (c)(ii) is completely analogous, taking $\ell_n = \sqrt{n/p}$. ■

Appendix 2: Proof of Theorem 4

To prove Theorem 4, we will need the following two lemmas whose proof may be found in Bianco et al. (2022). The first one is a direct extension of Hölder's inequality to the case of the product of three random variables that we include for the sake of completeness. The second result is analogous to Lemma 2, but now the family of functions is indexed over a compact set in $\mathbb{R}^{3k}$.

Lemma 5 *Let p, q, and r be real positive values such that* $(1/p)+(1/q)+(1/r) = 1$. *Let U, V, and W be random variables that satisfy* $\mathbb{E}|U|^p < \infty$, $\mathbb{E}|V|^q < \infty$, *and* $\mathbb{E}|W|^r < \infty$. *Then,* $\mathbb{E}|U\,V\,W| \le (\mathbb{E}|U|^p)^{1/p}\,(\mathbb{E}|V|^q)^{1/q}\,(\mathbb{E}|W|^r)^{1/r}$.

Recall that $\mathcal{B}_k(\boldsymbol{\beta}_{0,\mathrm{A}})$ stands for the closed unit ball in $\mathbb{R}^k$ centred at $\boldsymbol{\beta}_{0,\mathrm{A}}$, and denote as $\mathcal{S}^{k-1} = \{\mathbf{b} \in \mathbb{R}^k : \|\mathbf{b}\|_2 = 1\}$ the unit sphere centred in $\mathbf{0}_k$. To avoid the use of heavy notation, the vectors $\mathbf{b}_k, \mathbf{v}_k, \mathbf{w}_k \in \mathbb{R}^k$ will be denoted as $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$, respectively.

Lemma 6 *Let* $h_{\mathbf{w},\mathbf{b},\mathbf{u}} : \{0,1\} \times \mathbb{R}^k \to \mathbb{R}$ *be defined as* $h_{\mathbf{w},\mathbf{b},\mathbf{u}}(y,\mathbf{z}) = \chi(y, \mathbf{z}^{\mathrm{T}}\mathbf{b})\mathbf{w}^{\mathrm{T}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{u}$, *and consider the family of functions*

$$\mathcal{H} = \{h_{\mathbf{w},\mathbf{b},\mathbf{u}}\,, \quad \mathbf{b} \in \mathcal{B}_k(\boldsymbol{\beta}_{0,\mathrm{A}}), \mathbf{w}, \mathbf{u} \in \mathcal{S}^{k-1}\}\,. \tag{20}$$

Then, under ***R5*** *and* ***Z6***, $\mathbb{E}\bigl[\sup_{h\in\mathcal{H}} |(P_n - P)(h)|\bigr] \le C\sqrt{k/n}$ *for some constant C independent of n and p.*

Proof of Theorem 4 Let $\mathcal{A}_n = \{\widehat{\beta}_{n,j} \neq 0 \text{ for all } 1 \leq j \leq k\}$ and $\mathcal{B}_n = \{\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{NA}} = \mathbf{0}_{p-k}\}$. Note that $\mathbb{P}(\mathcal{A}_n^c) = \mathbb{P}(\widehat{\beta}_{n,j} = 0 \text{ for some } 1 \leq j \leq k) \leq \mathbb{P}(\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 > m_{0,n})$. Using that $\sqrt{n/k}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}(1)$ and assumption **N1**, we obtain that $\mathbb{P}(\mathcal{A}_n^c) \to 0$. On the other hand, as $\mathbb{P}(\mathcal{B}_n) = \mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{NA}} = \mathbf{0}_{p-k}) \to 1$, we get that $\mathbb{P}(\mathcal{B}_n \cap \mathcal{A}_n) \to 1$, and for any $\omega \in \mathcal{B}_n \cap \mathcal{A}_n$, every component of $\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}}$ is different from zero. Then, using that $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}}^{\mathrm{T}}, \mathbf{0}_{p-k}^{\mathrm{T}})^{\mathrm{T}}$ is the minimizer of $L_n(\boldsymbol{\beta}) + I_{\lambda_n}(\boldsymbol{\beta})$, we conclude that $\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}}$ minimizes $L_n\big(\mathbf{b}, \mathbf{0}_{p-k}\big) + I_{\lambda_n}\big(\mathbf{b}, \mathbf{0}_{p-k}\big)$ over $\mathbf{b} \in \mathbb{R}^k$, where we used for L_n the same notation introduced in Sect. 4 for I_λ. Therefore, we obtain that $\mathbf{0}_k = \nabla\Big(\sum_{i=1}^n \phi(y_i, \mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}})\Big)/n + \nabla\big(I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}})\big) + \mathbf{r}_n$, where $\mathbb{P}(\mathbf{r}_n = 0) \to 1$. Hence, for any $\mathbf{v} \in \mathbb{R}^k$ with $\|\mathbf{v}\|_2 = 1$, we have $0 = \sum_{i=1}^n \Psi(y_i, \mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}})\ \mathbf{v}^{\mathrm{T}}\mathbf{x}_{i,\mathrm{A}}/n + \mathbf{v}^{\mathrm{T}}\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}}) + \mathbf{v}^{\mathrm{T}}\mathbf{r}_n$. Given $\mathbf{b} \in \mathbb{R}^k$, denote $M_n(\theta) = \sum_{i=1}^n \Psi\Big(y_i, \mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\big[\theta\,\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} + (1-\theta)\,\boldsymbol{\beta}_{0,\mathrm{A}}\big]\Big)\mathbf{v}^{\mathrm{T}}\mathbf{x}_{i,\mathrm{A}}/n$ and $\mathbf{A}_{n,\mathrm{A}}(\mathbf{b}) = \sum_{i=1}^n \chi(y_i, \mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\mathbf{b})\mathbf{x}_{i,\mathrm{A}}\mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}/n$. Using the mean value theorem, we get that $M_n(1) = M_n(0) + M_n'(\alpha)$ for some $\alpha \in [0, 1]$. Thus,

$$0 = \frac{1}{n}\sum_{i=1}^n \Psi(y_i, \mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{0,\mathrm{A}})\mathbf{v}^{\mathrm{T}}\mathbf{x}_{i,\mathrm{A}} + \mathbf{v}^{\mathrm{T}}\mathbf{A}_{n,\mathrm{A}}(\boldsymbol{\beta}_{\mathrm{A}}^*)(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}) + \mathbf{v}^{\mathrm{T}}\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}}) + \mathbf{v}^{\mathrm{T}}\mathbf{r}_n\,, \tag{21}$$

where $\boldsymbol{\beta}_{\mathrm{A}}^* = \alpha\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} + (1-\alpha)\boldsymbol{\beta}_{0,\mathrm{A}}$ for some $\alpha \in [0, 1]$. Observe that $\sqrt{n}\ t_n^{-1}\mathbf{v}^{\mathrm{T}}\mathbf{A}_{\mathrm{A}}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}) = S_{1,n} + S_{2,n} + S_{3,n}$, where, to make the dependence on n explicit, we wrote t_n instead of $t = (\mathbf{v}^{\mathrm{T}}\mathbf{B}_{\mathrm{A}}^{(k)}\mathbf{v})^{1/2}$, and $S_{1,n} = \sqrt{n}\ t_n^{-1}\mathbf{v}^{\mathrm{T}}(\mathbf{A}_{\mathrm{A}} - \mathbf{A}_{\mathrm{A}}(\boldsymbol{\beta}_{\mathrm{A}}^*))(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})$, $S_{2,n} = \sqrt{n}\ t_n^{-1}\mathbf{v}^{\mathrm{T}}(\mathbf{A}_{\mathrm{A}}(\boldsymbol{\beta}_{\mathrm{A}}^*) - \mathbf{A}_{n,\mathrm{A}}(\boldsymbol{\beta}_{\mathrm{A}}^*))(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})$ and $S_{3,n} = \sqrt{n}\ t_n^{-1}\mathbf{v}^{\mathrm{T}}\mathbf{A}_{n,\mathrm{A}}(\boldsymbol{\beta}_{\mathrm{A}}^*)(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})$. We will show that $S_{3,n}\xrightarrow{D}N(0, 1)$ and that $S_{j,n}\xrightarrow{p}0$, for $j = 1, 2$.

We start by proving that $S_{1,n}\xrightarrow{p}0$. Given positive real numbers $\varepsilon, \delta > 0$, we need to show that $\mathbb{P}(|S_{1,n}| < \varepsilon) > 1 - \delta$ for n large enough. The fact that $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 = O_{\mathbb{P}}\big(\sqrt{k/n}\big)$ implies that, for any $\delta > 0$, there exists $C_1 > 0$ such that $\mathbb{P}(\mathcal{D}_n) > 1 - \delta/4$ for all n, where

$$\mathcal{D}_n = \{\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \leq C_1\sqrt{k/n}\}\,. \tag{22}$$

Note that from **R5**, $\chi_1(y, s) = (\partial/\partial s)\chi(y, s)$ is bounded. Then,

$$|S_{1,n}| \leq \sqrt{n}\ t_n^{-1}\mathbb{E}\big|\chi_1(y, \mathbf{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{A}}^{**})\mathbf{x}_{\mathrm{A}}^{\mathrm{T}}(\boldsymbol{\beta}_{0,\mathrm{A}} - \boldsymbol{\beta}_{\mathrm{A}}^*)\mathbf{v}^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}\mathbf{x}_{\mathrm{A}}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})\big|\,,$$

where $\boldsymbol{\beta}_{\mathrm{A}}^{**} = \alpha_1\boldsymbol{\beta}_{\mathrm{A}}^* + (1-\alpha_1)\boldsymbol{\beta}_{0,\mathrm{A}}$ for some $\alpha_1 \in [0, 1]$, and the expected value in the last equality is taken only with respect to y and $\mathbf{x}_{\mathrm{A}}$. Hence, using the fact that χ_1 is bounded and applying Lemma 5 to the random variables $U = (\boldsymbol{\beta}_{0,\mathrm{A}} - \boldsymbol{\beta}_{\mathrm{A}}^*)^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}$,

$V = \mathbf{v}^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}$ and $W = (\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}$ (taking $p = q = r = 3$), we obtain

$$\begin{aligned} |S_{1,n}| &\leq \|\chi_1\|_\infty \sqrt{n}\, t_n^{-1}\mathbb{E}\Big|(\boldsymbol{\beta}_{0,\mathrm{A}} - \boldsymbol{\beta}_{\mathrm{A}}^*)^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}\, \mathbf{v}^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}\, (\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}\Big| \\ &\leq \|\chi_1\|_\infty \sqrt{n}\, t_n^{-1}\mathbb{E}\Big[|(\boldsymbol{\beta}_{0,\mathrm{A}} - \boldsymbol{\beta}_{\mathrm{A}}^*)^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}|^3\Big]^{1/3}\mathbb{E}\Big[|\mathbf{v}^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}|^3\Big]^{1/3}\mathbb{E}\Big[|(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}|^3\Big]^{1/3} \\ &\leq \|\chi_1\|_\infty \sqrt{n}\, t_n^{-1}\mathbb{E}\Big[\|\mathbf{x}_{\mathrm{A}}\|_2^3\Big]\|\boldsymbol{\beta}_{0,\mathrm{A}} - \boldsymbol{\beta}_{\mathrm{A}}^*\|_2\|\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}\|_2\,, \end{aligned}$$

where, in the last inequality, we used Cauchy–Schwarz's inequality and the fact that $\|\mathbf{v}\|_2 = 1$. Therefore, for any $\omega \in \mathcal{D}_n$, from (22) we have that $|S_{1,n}| \leq \|\chi_1\|_\infty C_1^2 K_3^{1/2}\, t_n^{-1} k/\sqrt{n}$, where K_3 is the constant given in assumption **Z6**. Noticing that **Z7** entails that $t_n = \mathbf{v}^{\mathrm{T}}\mathbf{B}_{\mathrm{A}}^{(k)}\mathbf{v} \geq \iota_1\left(\mathbf{B}_{\mathrm{A}}^{(k)}\right) \geq \tau_2$, we conclude that $|S_{1,n}| \leq \|\chi_1\|_\infty C_1^2 K_3^{1/2}\, \tau_2^{-1}\, k/\sqrt{n}$. Finally, using that $k^2/n \to 0$, we get that there exists $n_0 \in \mathbb{N}$ such that for any $n \geq n_0$, $\mathcal{D}_n \subset \{\|S_{1,n}\| \leq \varepsilon\}$, which concludes the proof.

Let us show that $S_{2,n} \xrightarrow{p} 0$. Define $\mathbf{u}_n = (\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}})/\|\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}\|_2$, and note that

$$\begin{aligned} S_{2,n} &= \sqrt{n}\, t_n^{-1}\|\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}\|_2\left\{\mathbb{E}\Big[\chi(y, \mathbf{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{A}}^*)\mathbf{v}^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}\mathbf{x}_{\mathrm{A}}^{\mathrm{T}}\mathbf{u}_n\Big] - \frac{1}{n}\sum_{i=1}^{n}\chi(y_i, \mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{A}}^*)\mathbf{v}^{\mathrm{T}}\mathbf{x}_{i,\mathrm{A}}\mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\mathbf{u}_n\right\} \\ &= \sqrt{n}\, t_n^{-1}\|\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}\|_2(P - P_n)(h_{\mathbf{v},\boldsymbol{\beta}_{\mathrm{A}}^*,\mathbf{u}_n})\,, \end{aligned}$$

where the function $h_{\mathbf{v}_n,\boldsymbol{\beta}_{\mathrm{A}}^*,\mathbf{u}_n}$ is defined in (20). Let ε and δ be positive real numbers. Using that $\mathbb{P}(\mathcal{B}_n) = \mathbb{P}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{NA}} = \mathbf{0}_{p-k}) \to 1$, we get that there exists $n_0 \in \mathbb{N}$ such that, for any $n \geq n_0$, we have $\mathbb{P}(\mathcal{B}_n) > 1 - \delta/4$. On the other hand, recall that $\mathbb{P}(\mathcal{D}_n) > 1 - \delta/4$, where $\mathcal{D}_n$ is defined in (22). Hence, if $\widetilde{\mathcal{D}_n} = \{\|\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}} - \boldsymbol{\beta}_{0,\mathrm{A}}\|_2 \leq C_1\sqrt{k/n}\}$, we have that $\mathcal{B}_n \cap \mathcal{D}_n \subset \widetilde{\mathcal{D}_n}$ leading to $\mathbb{P}(\widetilde{\mathcal{D}_n}) > 1 - \delta/2$. Moreover, define the event $\mathcal{C}_n = \left\{\sup_{h\in\mathcal{H}}|(P_n - P)(h)| < (2C/\delta)\sqrt{k/n}\right\}$, where $\mathcal{H}$ is defined in (20) and C is the constant from Lemma 6. Applying Markov's inequality, we get that $\mathbb{P}(\mathcal{C}_n) > 1 - \delta/2$, which implies that $\mathbb{P}(\widetilde{\mathcal{D}_n} \cap \mathcal{C}_n) > 1 - \delta$. Let $n_1 \geq n_0$ be such that, for any $n \geq n_1$, $C_1\sqrt{k/n} < 1$. Then, restricting to the event $\widetilde{\mathcal{D}_n} \cap \mathcal{C}_n$, we obtain $|S_{2,n}| \leq C_3\, k/(\delta\sqrt{n})$, where $C_3 = 2CC_1/\tau_2$, and we again used that $t_n \geq \tau_2$. Finally, the fact that $k^2/n \to 0$ entails that there exists $n_2 \geq n_1$ such that $\widetilde{\mathcal{D}_n} \cap \mathcal{C}_n \subset \{|S_{2,n}| \leq \varepsilon\}$ for all $n \geq n_2$, thus $S_{2,n} \xrightarrow{p} 0$.

To conclude the proof, it remains to see that $S_{3,n} \xrightarrow{D} N(0,1)$. Using (21), we have that $S_{3,n} = \sum_{j=1}^{3} S_{3j,n}$, where $S_{31,n} = n^{-1/2}\, t_n^{-1}\sum_{i=1}^{n}\mathbf{v}^{\mathrm{T}}\Psi(y_i, \mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{0,\mathrm{A}})\mathbf{x}_{i,\mathrm{A}}$, $S_{32,n} = -\sqrt{n}\, t_n^{-1}\mathbf{v}^{\mathrm{T}}\nabla I_{\lambda_n}(\widehat{\boldsymbol{\beta}}_{n,\,\mathrm{A}})$, and $S_{33,n} = -\sqrt{n}\, t_n^{-1}\mathbf{v}_n^{\mathrm{T}}\mathbf{r}_n$. Using that $\mathbb{P}(\mathbf{r}_n = 0) \to 1$, the fact that (14) and **Z7** hold, it is easy to see that $S_{32,n} \xrightarrow{p} 0$ and $S_{33,n} \xrightarrow{p} 0$. It remains to show that $S_{31,n} \xrightarrow{D} N(0,1)$. Write $S_{31,n} = \sum_{i=1}^{n} W_{n,i}$, where $W_{n,i} = -t_n^{-1}\Psi(y_i, \mathbf{x}_{i,\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{0,\mathrm{A}})\mathbf{v}^{\mathrm{T}}\mathbf{x}_{i,\mathrm{A}}/\sqrt{n}$. Note that $\mathbb{E}W_{n,i} = 0$ for all

$n \in \mathbb{N}$ and $1 \le i \le n$, whereas $n\,\mathbb{E}W_{n,i}^2 = t_n^{-2}\ \mathbf{v}^{\mathrm{T}}\mathbb{E}\big[\Psi^2(y, \mathbf{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{0,\mathrm{A}})\mathbf{x}_{\mathrm{A}}\mathbf{x}_{\mathrm{A}}^{\mathrm{T}}\big]\mathbf{v} = t_n^{-2}\ \mathbf{v}^{\mathrm{T}}\mathbf{B}_{\mathrm{A}}^{(k)}\mathbf{v} = 1$, which implies $\sum_{i=1}^n \mathbb{E}W_{n,i}^2 = 1$. To apply the central limit theorem for triangular arrays, we will show that the Lyapunov's condition holds, that is, that there exists a value $\delta > 0$ such that $\lim_{n\to\infty}\sum_{i=1}^n \mathbb{E}\big[|W_{n,i}|^{2+\delta}\big] = 0$. Note that $t_n^{2+\delta}\ n^{1+\delta/2}\mathbb{E}|W_{n,i}|^{2+\delta} = \mathbb{E}\big[|\Psi(y, \mathbf{x}_{\mathrm{A}}^{\mathrm{T}}\boldsymbol{\beta}_{0,\mathrm{A}})|^{2+\delta}|\mathbf{v}^{\mathrm{T}}\mathbf{x}_{\mathrm{A}}|^{2+\delta}\big]$. Hence, using the fact that $\|\mathbf{v}\|_2 = 1$, Ψ is bounded and Cauchy–Schwarz's inequality, we obtain

$$\sum_{i=1}^{n}\mathbb{E}[|W_{n,i}|^{2+\delta}] \le \frac{1}{t_n^{2+\delta}}\frac{1}{n^{\frac{\delta}{2}}}\|\Psi\|_\infty^{2+\delta}\mathbb{E}\|\mathbf{x}_{\mathrm{A}}\|^{2+\delta} \le \frac{1}{n^{\frac{\delta}{2}}}\frac{1}{\tau_2^{2+\delta}}\|\Psi\|_\infty^{2+\delta}K_3^{\frac{2+\delta}{6}}\ ,$$

where the last inequality is a consequence of the fact that t_n^{-1} is bounded and assumption **Z6** holds. Hence, Lyapunov's condition holds, and using the Lindeberg–Feller's central limit theorem for triangular arrays, we conclude that $S_{31,n}\xrightarrow{D}N(0,1)$ and the desired result follows. ■

Acknowledgments The authors wish to thank the anonymous referees for their valuable comments that led to an improved version of the original paper. This research was partially supported by Grants 20020170100022BA from the Universidad de Buenos Aires and PICT 2018-00740 from ANPCYT at Buenos Aires, Argentina and also by the Spanish Project PID2020-116587GB-I00, CoDyNP from the Ministry of Science and Innovation (MCIN/ AEI /10.13039/501100011033) (Ana M. Bianco and Graciela Boente).

References

Avella-Medina, M., & Ronchetti, E. (2018). Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika, 105*, 31–44.

Basu, A., Gosh, A., Jaenada, M., & Pardo, L. (2021). Robust adaptive Lasso in high-dimensional logistic regression with an application to genomic classification of cancer patients. Available at https://arxiv.org/abs/2109.03028.

Basu, A., Gosh, A., Mandal, A., Martin, N., & Pardo, L. (2017). A Wald–type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. *Electronic Journal of Statistics, 11*, 2741–2772.

Bianco, A., Boente, G., & Chebi, G. (2021). Penalized robust estimators in logistic regression with applications to sparse models. *Test*. https://doi.org/10.1007/s11749-021-00792-w.

Bianco, A., Boente, G., & Chebi, G. (2022). Asymptotic behaviour of penalized robust estimators in logistic regression when dimension increases. Available at http://arxiv.org/abs/2201.12449.

Bianco, A., & Yohai, V. (1996). Robust estimation in the logistic regression model. *Lecture Notes in Statistics, 109*, 17–34.

Boente, G., Salibián-Barrera, M., & Tyler, D. (2014). A characterization of elliptical distributions and some optimality properties of principal components for functional data. *Journal of Multivariate Analysis, 131*, 254–264.

Bondell, H. D. (2005). Minimum distance estimation for the logistic regression model. *Biometrika, 92*, 724–731.

Bondell, H. D. (2008). A characteristic function approach to the biased sampling model, with application to robust logistic regression. *Journal of Statistical Planning and Inference, 138*, 742–755.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Berlin: Springer Science and Business Media.

Cantoni, E., & Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association, 96*, 1022–1030.

Chi, E. C., & Scott, D. W. (2014). Robust parametric classification and variable selection by a minimum distance criterion. *Journal of Computational and Graphical Statistics, 23*, 111–128.

Croux, C., & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis, 44*, 273–295.

Dasarathy, G. (2011). A simple probability trick for bounding the expected maximum of *n* random variables. Technical report of the Arizona State University. Available at http://www.public.asu.edu/~gdasarat/files/maxGaussians.pdf.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics, 32*, 407–499.

Elsener, A., & van de Geer, S. (2018). Sharp oracle inequalities for stationary points of nonconvex penalized M-estimators. *IEEE Transactions on Information Theory, 65*, 1452–1472.

Fan, J., & Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Fan, J., & Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory, 57*, 5467–5484.

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics, 32*, 928–961.

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35*, 109–135.

Guo, C., Yang, H., & and Lv, J. (2017). Robust variable selection for generalized linear models with a diverging number of parameters. *Communications in Statistics – Theory and Methods, 46*, 2967–2981.

Kingman, J. F. C. (1972). On random sequences with spherical symmetry on random sequences with spherical symmetry. *Biometrika, 59*, 492–494.

Kurnaz, F. S., Hoffmann, I., & Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems, 172*, 211–222.

Lahiri, S. (2021). Necessary and sufficient conditions for variable selection consistency of the Lasso in high dimensions. *Annals of Statistics, 49*, 820–844.

Leng, C., Lin, Y., & Wahba, G. (2006). A note on the Lasso and related procedures in model selection. *Statistica Sinica, 16*, 1273–1284.

Loh, P. L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics, 45*, 866–896.

Loh, P.-L., & Wainwright, M. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research, 16*, 559–616.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis, 52*, 374–393.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics, 34*, 1436–1462.

Park, H., & Konishi, S. (2016). Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *Journal of Statistical Computation and Simulation, 86*, 1450–1461.

Tibshirani, J., & Manning, C. D. (2013). Robust logistic regression using shift parameters. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 124–129).

van de Geer, S., & Müller, P. (2012). Quasi-likelihood and/or robust estimation in high dimensions. *Statistical Science, 27*, 469–480.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

Yuan, M., & Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*, 143–161.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics, 38*, 894–942.

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.

Conditional Distribution-Based Downweighting for Robust Estimation of Logistic Regression Models

Weichang Yu and Howard D. Bondell

Abstract We propose a new approach to robust estimation for logistic regression based on the conditional covariate distribution given the binary response. Instead of downweighting observations based on the marginal distribution often adopted in common approaches (e.g., the Mallows class), we propose using the conditional distributions that align with a case-control perspective of binary regression. We justify our proposed weighting scheme by showing that our method based on this new perspective leads to sensible weights in scenarios where the existing methods perform poorly. Through simulated and real datasets, we show that our proposed estimator achieves superior performance over the existing estimators in terms of both robustness and efficiency.

Keywords Case-control · Efficiency · M-estimation · Robustness

1 Introduction

Binary regression is an appealing model choice among the existing classification models in recent years. Much of its appeal is attributed to its explainability, particularly the direct correspondence between a research hypothesis and a statement about the regression coefficients. Moreover, when the canonical link function is specified, the resultant model is known as *logistic regression*, and its coefficients can be directly interpreted as an odds ratio.

The most common method for estimating logistic regression coefficients is the maximum likelihood estimation method whereby the estimator minimizes the sum of deviances. However, the maximum likelihood estimate (MLE) is very sensitive to influential outliers (Pregibon 1981, 1982). A small number of influential points can

W. Yu (✉) · H. Bondell
School of Mathematics and Statistics, University of Melbourne, Parkville, VIC, Australia
e-mail: weichang.yu@unimelb.edu.au; howard.bondell@unimelb.edu.au

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_16

have huge effects on the coefficient estimates as evidenced in several real datasets (Pregibon 1982).

A diverse range of methods has been proposed to circumvent this susceptibility to outliers. One approach is to assess the outlyingness of each observation through a statistic and delete highly influential outliers. Some examples of statistics for assessing influence include the coefficient gradient (Pregibon 1981), symmetric Kullback–Leibler divergence (Johnson 1985), and Cook's distance (Cook 1977; Martin & Pardo 2009). While these methods allow us to simply use the MLE after removing outliers, there is no guarantee that they identify all outliers and avoid erroneous deletion of non-outliers. Moreover, these methods may not be able to identify outliers that exert minor influence individually but are collectively influential as is the case of some datasets in Pregibon (1982).

Other methods include minimizing a robust function of the deviance (Bianco & Yohai 1996; Pregibon 1982), deleting extreme outliers, and then maximizing a trimmed correlation (Feng et al. 2014), or incorporating the distribution of the contamination in estimating the regression coefficients (Copas 1988). In particular, Bianco and Yohai (1996) proposed an estimator that minimizes a functional of the deviance to ensure Fisher consistency—a desirable property for robust estimators whereby the estimator yields the true parameter value if data from the entire population are used. Moreover, their proposed method does not require any computation of robust covariance matrix estimates and has comparable efficiency to the MLE in the absence of outliers.

A popular approach toward robust estimation for logistic regression is to find the root of a weighted score function. Here, the idea is to downweight observations with high leverage or is highly influential in some sense. Stefanski et al. (1986) and Künsch et al. (1989) propose the *bounded-influence estimators* in which the weights depend on a residual component and an elliptical contour component. Their proposed estimators minimize the asymptotic variance–covariance of an M-estimator subject to an upper bound constraint on the influence function (defined as the effect of an infinitesimal point mass contamination of the joint distribution of the covariate and the response). The conditionally unbiased bounded-influence estimator (CUBIF; Künsch et al. 1989) introduces a observation-specific perturbation to ensure conditional Fisher consistency—an even stronger version of the Fisher consistency in the regression context whereby the estimator yields the true parameter value if data from the entire population are used (for any fixed covariate value). Carroll and Pederson (1993) propose the use of a weighting scheme in which observations are downweighted according to their leverage or predicted probabilities. In fact, they show that their estimator's bias is close to the MLE for small samples without outliers.

Several of the downweighting-based methods utilize elliptical contours based on the marginal covariate moments or predicted probabilities to identify outliers. While this strategy would lead to good performance when the outliers are very extreme, they may fail in two cases—(i) the outlier is extreme with respect to the conditional distribution $\mathbf{X}|Y$ but is non-extreme with respect to the marginal distribution of $\mathbf{X}$, (ii) the sample is unbalanced. The reason for their potential failure can be better understood from a generative model perspective. In case (i), such outliers are allocated high weight despite having a strong effect on the estimate. An example is when a group 1 observation is located near the group 0 centroid. In case (ii), all observations from the smaller group would be allocated low weights as the marginal likelihood of the smaller group observations is very small.

To address this issue, Bondell (2005) proposed an estimator that minimizes the Cramer–von-Mises distance between the empirical distribution under the hypothesized logistic regression model and the nonparametric empirical distribution. The proposed estimator yields highly efficient estimates when the model is true and demonstrates robustness under small deviations.

We propose an alternative solution to address the limitations with the existing downweighting-based methods. Our proposed method is a modification of the Mallows class estimator discussed in Carroll and Pederson (1993). The resulting estimator is no longer a member of the Mallows class as its weight depends on a group-specific Mahalanobis distance. We show that the proposed estimator leads to substantial improvement in both efficiency and robustness.

In Sect. 2, we review various classes of M-estimators for logistic regression and provide details on two major limitations. In Sect. 3, we describe our proposed estimator and how they address the two major limitations with the existing downweighting-based methods. In Sect. 4, we present a numerical study to compare the performance of our proposed estimator with the existing methods. Section 5 concludes.

2 M-Estimators for Logistic Regression

Consider the set of independently and identically distributed data $\mathcal{D} = \{(\mathbf{x}_i, Y_i)\}_{i=1}^n$, where the response $Y \in \{0, 1\}$, the covariates $\mathbf{x} \in \mathbb{R}^p$, and our interest is to estimate the unknown parameters of the conditional distribution of Y given $\mathbf{x}$:

$$p(Y_i = 1 \mid \mathbf{x}_i) = Q(\mathbf{x}_i^\top \boldsymbol{\theta}),$$

where $Q(z) = 1/(1 + e^{-z})$. The class of M-estimators $\boldsymbol{\theta}$ for logistic regression is the solution to the equation:

$$\sum_{i=1}^n \Psi(\mathbf{x}_i, y_i, \boldsymbol{\theta}, \xi) = \mathbf{0}, \tag{1}$$

where $\Psi(\mathbf{x}_i, y_i, \boldsymbol{\theta}, \boldsymbol{\xi}) = w(\mathbf{x}_i, y_i, \boldsymbol{\theta}, \boldsymbol{\xi})\{y_i - Q(\mathbf{x}_i^\top \boldsymbol{\theta}) - c(\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\xi})\}\mathbf{x}_i$ and $\boldsymbol{\xi}$ is a vector of tuning parameters for the weight function. Observe that the LHS is a weighted sum of observation-specific score contributions. Note that the correction function c ensures that the estimating equation is conditionally unbiased, i.e.,

$\mathbf{E}[\Psi(\mathbf{X}, Y, \boldsymbol{\theta}, \xi) \,|\, \mathbf{X}] = \mathbf{0}$, and hence, we have

$$c(\mathbf{x}, \boldsymbol{\theta}, \xi) = \frac{E\{w(\mathbf{x}, Y, \boldsymbol{\theta}, \xi)(Y - Q(\mathbf{x}^\top \boldsymbol{\theta})) \,|\, \mathbf{x}\}}{E\{w(\mathbf{x}, Y, \boldsymbol{\theta}, \xi) \,|\, \mathbf{x}\}}.$$

In the existing literature, a well-known measure of robustness for M-estimators is the influence function:

$$\mathrm{IF}(\mathbf{x}, y; \Psi_{\boldsymbol{\xi}}, F_{\boldsymbol{\theta}}) = \mathbf{A}(\boldsymbol{\theta})^{-1} \Psi(\mathbf{x}, y, \boldsymbol{\theta}, \boldsymbol{\xi}),$$

where $\mathbf{A}(\boldsymbol{\theta}) = -\mathbf{E}_{\mathbf{X},Y}\{\frac{\partial}{\partial \boldsymbol{\theta}} \Psi(\mathbf{X}, Y, \boldsymbol{\theta}, \boldsymbol{\xi})\}$ and $(\mathbf{X}, Y) \sim F_{\boldsymbol{\theta}}$. The influence function measures the effect of contaminating the original data distribution $F_{\boldsymbol{\theta}}$ by an infinitesimal point mass contamination at $(\mathbf{x}, y)$, i.e., the effect on the resultant estimator if the true distribution is $\epsilon \delta_{(\mathbf{x},y)} + (1 - \epsilon) F_{\boldsymbol{\theta}}$ instead of $F_{\boldsymbol{\theta}}$, where $\epsilon \to 0$.

Here, the different types of M-estimators differ by the quantities which the weights depend on. For the *Schweppe class*, $w = w(\mathbf{x}_i, y_i, \mathbf{x}_i^\top \boldsymbol{\theta})$ and $c(\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\xi}) \neq 0$ for most specifications in this class, where $\boldsymbol{\xi}$ is a vector of tuning parameter. This class of estimators includes the CUBIF, where

$$w_{CUBIF} = w_b\left(|y_i - Q(\mathbf{x}_i^\top \boldsymbol{\theta}) - c(\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\xi})| \sqrt{\mathbf{x}_i^\top \mathbf{B}^{-1} \mathbf{x}_i} \right),$$

$w_b(a) = \max\{-b/a, \min(1, b/a)\}$ for $a > 0$, b is a tuning parameter, and $\mathbf{B}$ may be obtained via equations (2.8) and (2.9) in Künsch et al. (1989). The CUBIF estimator is an example of a bounded-influence estimator that minimizes the variance of Ψ among all estimators with bounded influence in some sense, where the tuning parameter b corresponds to a bound for IF.

For the *Mallows class*, the weights depend on the study design and/or predicted probabilities, i.e., they do not depend on Y. For example, Carroll and Pederson (1993) proposed two alternative weights

$$w(\mathbf{x}_i, \boldsymbol{\xi}) = \{1 - d(\mathbf{x}_i, \boldsymbol{\xi})/(b^2(p-1))\} \mathbb{I}\{d(\mathbf{x}_i, \boldsymbol{\xi}) \leq b^2(p-1)\},$$

where d_i is a robust distance metric and

$$w(\mathbf{x}^\top \boldsymbol{\theta}, c, \lambda) = [Q(\mathbf{x}^\top \boldsymbol{\theta})\{1 - Q(\mathbf{x}^\top \boldsymbol{\theta})\}]^c [Q(\mathbf{x}^\top \boldsymbol{\theta})^\lambda + \{1 - Q(\mathbf{x}^\top \boldsymbol{\theta})\}^\lambda].$$

It can be shown that the above Mallows class estimators have bounded IF $\|\mathbf{x}\| w$ is bounded. Here, the Mallows class estimator has several advantages. First, the weight functions are simpler to work with as they do not depend on Y unlike the Schweppe class estimators. Moreover, the resultant estimating equation is less complex as the correction factor equals zero. However, a disadvantage is that without using the response, points are downweighted due to their potential outlyingness, rather than actually being an outlier. This results in a loss of efficiency at the true model. Our formulation partially addresses this drawback.

2.1 A New Perspective to Outlier Downweighting

Most robust logistic regression estimators are proposed to attenuate the effects of outliers that are traditionally defined as observations with large residuals or high leverage (with respect to the marginal covariate moments). This definition is directly adapted from the linear regression context (Copas 1988). However, this implicit definition does not take into account the difference between the logistic regression and the linear regression in terms of their covariate distributions. Consequently, observations that are unusual with respect to the conditional distributions may still be considered plausible based on the marginal contours. This calls for a re-thinking of the definition and generating process of an outlier. Copas (1988) provided a discriminant-based definition to an outlier—any observation with $y = 1$ and predicted probability close to 0 or $y = 0$ and predicted probability close to 1. However, this definition does not directly address the problem due to the mixture profile of the covariate distribution in the logistic regression context. Here, we provide a definition of outliers based on a generative model perspective: an observation $(\mathbf{x}, y)$ is an outlier if $p(\mathbf{x} \mid Y = y)$ is small. In Fig. 1, we illustrate the difference between two weighting schemes for the covariate distributions in (3) and $p(Y = 1) = 0.5$ by plots of 99% probability contours. Observe that group 1 observations located in the neighborhood around $(-1.5, -1.5)$ are allocated

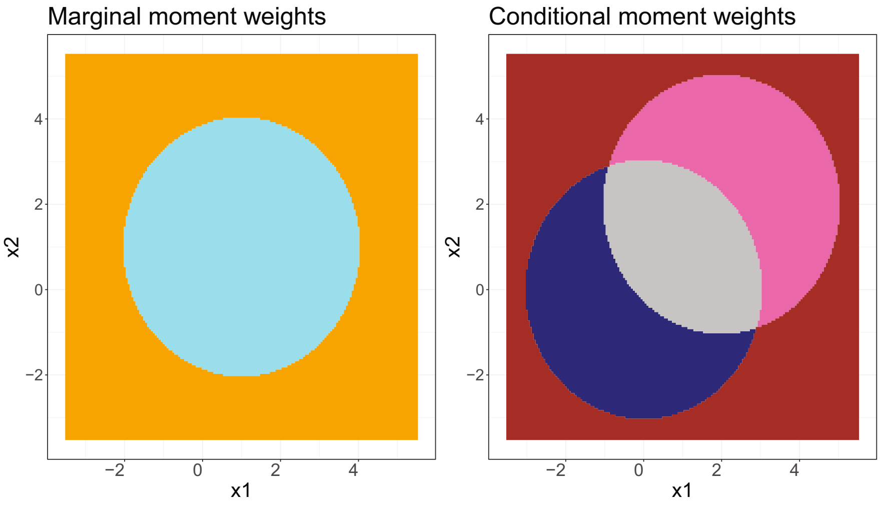


Fig. 1 (Left) Marginal weighting scheme: plot of 99% probability contour based on marginal covariate moments. Observations in the cyan region are allocated substantial weights, whereas observations in the orange region are allocated low weights. (Right) Conditional weighting scheme: plot of 99% probability contour based on marginal covariate moments. Observations in the gray (brown) region are allocated substantial (low) weights regardless of its class label. Observations in the pink region are allocated substantial (low) weight if it belongs to group 1 (0). Observations in the blue region are allocated substantial (low) weight if it belongs to group 0 (1)

substantial weights by the marginal weighting scheme but are allocated low weights by the conditional weighting scheme.

We provide details on how our new perspective to outlier identification can be incorporated into the existing Mallows class weighting scheme in the next section.

3 Modified Mallows Class Approach

We propose a modified Mallows estimator that circumvents the issues described in the previous sections. In particular, the proposed weighting scheme depends on a group-specific Mahalanobis distance, i.e.,

$$w(\mathbf{x}, Y, \boldsymbol{\theta}, \xi) = w(\mathbf{x}, Y, \xi) = \nu\{1 - (\nu/h)^2\}^3 \mathbb{I}\{|\nu| \leq h\},$$

where $\nu = \sqrt{d/(p-1)}$,

$$d = \begin{cases} (\mathbf{x} - \widehat{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \widehat{\mu}_1), \text{ if } Y = 1; \\ (\mathbf{x} - \widehat{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\mathbf{x} - \widehat{\mu}_0), \text{ if } Y = 0, \end{cases}$$

and $\widehat{\mu}_k$ and $\widehat{\Sigma}_k$ are robust estimators of the group-specific mean and variance for $Y = k$, and $k = 0, 1$. Note that the weight function w is adapted from Carroll and Pederson (1993). Since the weights depend on the response, our proposed estimator does not belong to the Mallows class. The weights taper toward 0 as the distance of the data point increases with respect to its group-specific mean. The corresponding correction function is

$$\begin{aligned} c(\mathbf{x}, \boldsymbol{\theta}, \xi) &= \frac{E\{w(\mathbf{x}, Y, \boldsymbol{\theta}, \xi)(Y - Q(\mathbf{x}^\top \boldsymbol{\theta})) \,|\, \mathbf{x}\}}{E\{w(\mathbf{x}, Y, \boldsymbol{\theta}, \xi) \,|\, \mathbf{x}\}} \\ &= \frac{w_1(1 - Q(\mathbf{x}^\top \boldsymbol{\theta}))Q(\mathbf{x}^\top \boldsymbol{\theta}) + w_0\big\{-Q(\mathbf{x}^\top \boldsymbol{\theta})\big\}(1 - Q(\mathbf{x}^\top \boldsymbol{\theta}))}{w_1 Q(\mathbf{x}^\top \boldsymbol{\theta}) + w_0(1 - Q(\mathbf{x}^\top \boldsymbol{\theta}))} \\ &= C(w_0, w_1, Q(\mathbf{x}^\top \boldsymbol{\theta})), \end{aligned}$$

where $w_k = w(\mathbf{x}, k, \boldsymbol{\theta}, \xi)$ and $C(t, u, v) = \{(1-v)t + vu\}^{-1} v(1-v)(u-t)$. Note that the second equality follows by taking expectation with respect to $p(Y|\mathbf{x}) = Q(\mathbf{x}^\top \boldsymbol{\theta})$. Hence, the modified Mallows estimator $\widetilde{\boldsymbol{\theta}}_M$ satisfies the equation:

$$\sum_{i=1}^{n} w_{y_i} \Big\{ y_i - Q(\mathbf{x}_i^\top \widetilde{\boldsymbol{\theta}}_M) - C(w_{i0}, w_{i1}, Q(\mathbf{x}_i^\top \widetilde{\boldsymbol{\theta}}_M)) \Big\} \mathbf{x}_i = \mathbf{0}.$$

Since $\widetilde{\boldsymbol{\theta}}_M$ has no closed-form expression, we may utilize a Newton's algorithm to compute its numerical value. Denote

$$F(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_{y_i} \Big\{ y_i - Q(\mathbf{x}_i^\top \boldsymbol{\theta}) - C(w_{0i}, w_{1i}, Q(\mathbf{x}_i^\top \boldsymbol{\theta})) \Big\} \mathbf{x}_i .$$

The derivative of F is

$$\tfrac{\partial}{\partial \boldsymbol{\theta}} F(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_{y_i} \mathbf{x}_i \Big\{ -\tfrac{\partial}{\partial \boldsymbol{\theta}} Q(\mathbf{x}_i^\top \boldsymbol{\theta}) - \tfrac{\partial}{\partial \boldsymbol{\theta}} C(w_{0i}, w_{1i}, Q(\mathbf{x}_i^\top \boldsymbol{\theta})) \Big\},$$

where $\frac{\partial}{\partial \boldsymbol{\theta}} Q(\mathbf{x}_i^\top \boldsymbol{\theta}) = Q(\mathbf{x}_i^\top \boldsymbol{\theta})\{1 - Q(\mathbf{x}_i^\top \boldsymbol{\theta})\}\mathbf{x}_i^\top$ and $\frac{\partial}{\partial \boldsymbol{\theta}} C(w_{i0}, w_{i1}, Q(\mathbf{x}_i^\top \boldsymbol{\theta})) = (w_{i1} - w_{i0})\{(1 - Q(\mathbf{x}_i^\top \boldsymbol{\theta}))^2 w_{i0} - Q(\mathbf{x}_i^\top \boldsymbol{\theta})^2 w_{i1}\}\{(1 - Q(\mathbf{x}_i^\top \boldsymbol{\theta}))w_{i0} + Q(\mathbf{x}_i^\top \boldsymbol{\theta}) w_{i1}\}^{-2}$ $\frac{\partial}{\partial \boldsymbol{\theta}} Q(\mathbf{x}_i^\top \boldsymbol{\theta})$. Thus, the Newton's algorithm is

$$\boldsymbol{\theta}^{(\tau+1)} = \boldsymbol{\theta}^{(\tau)} - \left[\tfrac{\partial}{\partial \boldsymbol{\theta}} F(\boldsymbol{\theta}^{(\tau)}) \right]^{-1} F(\boldsymbol{\theta}^{(\tau)}).$$

The influence function of our proposed modified Mallows estimator is given by

$$IF(\mathbf{x}, y; \boldsymbol{\theta}) = \{A(\boldsymbol{\theta})\}^{-1} \Psi(\mathbf{x}, y, \boldsymbol{\theta}, \xi),$$

where

$$A(\boldsymbol{\theta}) = \int Q(\mathbf{x}^\top \boldsymbol{\theta})\{1 - Q(\mathbf{x}^\top \boldsymbol{\theta})\} \eta(w_0, w_1, Q(\mathbf{x}^\top \boldsymbol{\theta})) \mathbf{x}\mathbf{x}^\top \; F(\mathrm{d}\mathbf{x}),$$

where $\eta(t, u, v) = (1 - v)u + vt - (u - t)C(t, u, v)$. Under standard regularity conditions, we have

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_M - \boldsymbol{\theta}^\star) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}^\star)), \tag{2}$$

where $\boldsymbol{\theta}^\star$ denotes the data true coefficient value,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}^\star) = \{A(\boldsymbol{\theta}^\star)\}^{-1} B(\boldsymbol{\theta}^\star) \{A(\boldsymbol{\theta}^\star)^{-1}\}^\top,$$

and

$$B(\boldsymbol{\theta}) = \int \Big[Q(\mathbf{x}^\top \boldsymbol{\theta}) w_1^2 \{1 - Q(\mathbf{x}^\top \boldsymbol{\theta}) - C(w_0, w_1, Q(\mathbf{x}^\top \boldsymbol{\theta}))\}^2 \\ + (1 - Q(\mathbf{x}^\top \boldsymbol{\theta})) w_0^2 \{-Q(\mathbf{x}^\top \boldsymbol{\theta}) - C(w_0, w_1, Q(\mathbf{x}^\top \boldsymbol{\theta}))\}^2 \Big] \mathbf{x}\mathbf{x}^\top \; F(\mathrm{d}\mathbf{x}).$$

4 Numerical Study

We compare the performance of our proposed estimator with the maximum likelihood estimator (MLE), the Mallows estimator (Carroll & Pederson 1993), and the CUBIF estimator (Künsch et al. 1989) in three simulation scenarios and one real dataset. We use the implementations of the Mallows and CUBIF estimators in the R package `robust`. The tuning parameters of each robust estimator are specified such that its estimated relative efficiency with respect to the MLE is approximately 90.5%, based on 500 simulated balanced datasets ($n_1 = 100$; $n_0 = 100$) without outliers and the covariate distributions:

$$\mathbf{x}_i \,|y_i \sim \begin{cases} \mathrm{N}\left(\begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right), \text{ if } y_i = 1; \\ \\ \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right), \text{ if } y_i = 0. \end{cases} \tag{3}$$

This procedure leads to the following specifications: `bpar`$=$ 0.824 and `cpar`$=$ 2 for CUBIF; `wt.tuning`$=$ 3.005 for Mallows; $h = 5.3$ for modified Mallows. For our proposed method, we used the Mallows estimate as the starting value as they are known to be fast to compute. In practice, any robust estimator will make a good starting value. In addition, under standard conditions of uniqueness of the MLE, our estimating equation will also have a unique solution, provided that one specifies positive-definite estimates for the group-specific covariance matrices.

4.1 Simulation Settings

By using the aforementioned calibration procedure to achieve similar efficiency, we assess the robustness of the methods with the following scenarios. In scenario 1, we compare the performance of the estimators on 500 simulated balanced datasets ($n_1 = 100$; $n_0 = 100$). Each dataset contains one outlier that belongs to group 1 and is positioned at various points along the line $x_1 = x_2$. In scenario 2, we compare the performance of the estimators on 500 simulated unbalanced datasets ($n_1 = 40$; $n_0 = 160$). There are no outliers in these datasets. In scenario 3, we compare the performance of the estimators on 500 simulated unbalanced datasets ($n_1 = 40$; $n_0 = 160$). Each dataset contains one outlier that belongs to group 1 and is positioned at various points along the line $x_1 = x_2$.

The covariate distributions for all scenarios are in Eq. (3). To assess the performance of the estimators, we compute the mean squared error (MSE) and estimated coverage probability of each method for each scenario.

4.2 Simulation Results

The MSEs of all methods for scenario 1 are presented in Fig. 2. When the outlier is positioned close to its own centroid and hence fits the model well, all robust methods demonstrate lower efficiency in comparison to the MLE. However, the key comparison is in the region of moderate outliers (outliers positioned between $(-2.1, -2.1)$ and $(-0.4, -0.4)$) that are challenging to identify and downweight. When the outlier is in this region, our proposed modified Mallows estimator outperforms the other robust estimators. This is because outliers located along this segment are mostly allocated low weights by the modified Mallows method, whereas they are allocated large weights by CUBIF and Mallows estimators. For example, the Mallows estimator would allocate the same weight to a group 1 outlier at $(-1, -1)$ and a group 1 non-outlier at $(3, 3)$. All robust methods perform comparably well when the outlier is very extreme. Nonetheless, we observe a

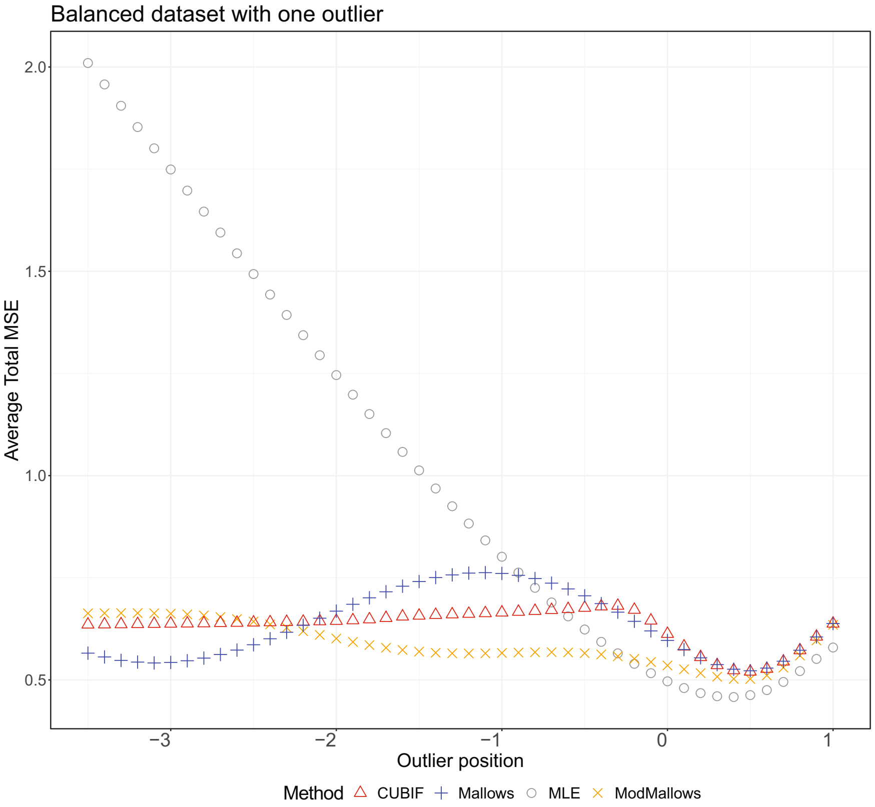


Fig. 2 Average total MSE against outlier position along the line $x_1 = x_2$. The average total MSE is computed over 500 balanced datasets, each consisting of one outlier

slight increase followed by a plateau for the modified Mallows estimator as the outlier position becomes more extreme. This is attributed to the susceptibility of the modified Mallows estimator to near-complete linear separability. More specifically, the modified Mallows estimator tends to allocate lower weights to data points near the true decision boundary than those data points near its respective centroid. When the dataset is near-complete separable, then a higher weight on the outlier ironically improves the accuracy of the estimate.

In scenario 2, the estimated relative efficiencies of the robust estimators with respect to the MLE are: 92.8% (mod-Mallows), 84.2% (CUBIF), and 83.0% (Mallows). Based on these relative efficiencies, it is evident that the modified Mallows estimator does not suffer from a loss of efficiency in the unbalanced case. The poorer performance by both CUBIF and Mallows estimators is due to low weights inadvertently allocated to the data points from the minor group. This is consequential to the unbalanced design that skews the overall centroid toward the centroid of the larger group. The results of scenario 3 are presented in Fig. 3. Observe that the results of the comparison between the robust methods are similar to

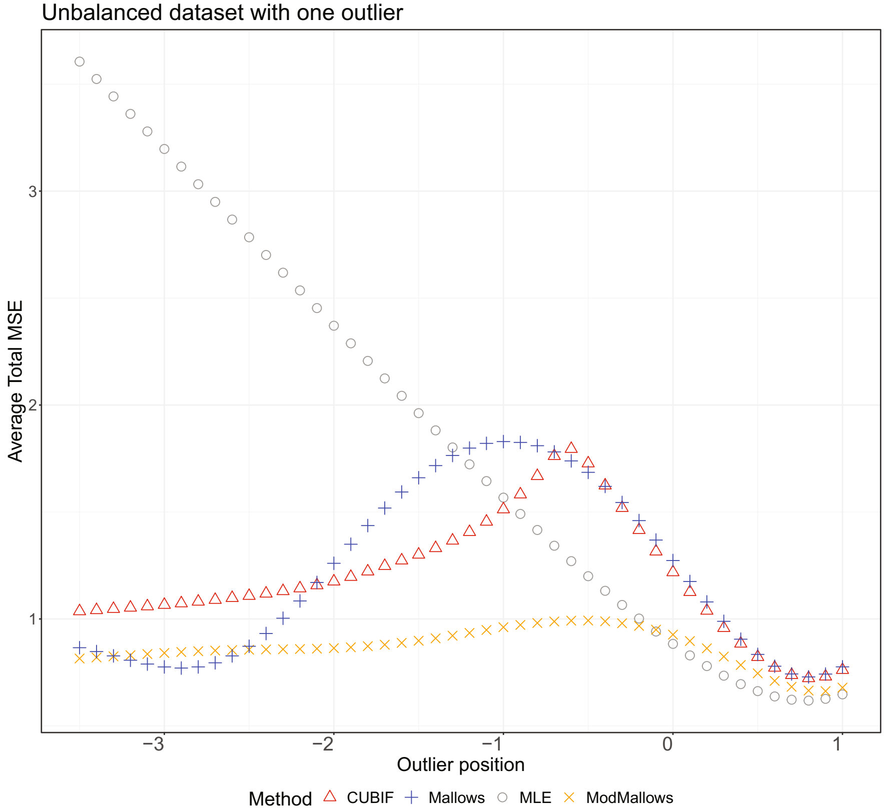


Fig. 3 Average total MSE against outlier position along the line $x_1 = x_2$. The average total MSE is computed over 500 unbalanced datasets, each consisting of one outlier

that of scenario 1, albeit a large MSE for CUBIF and Mallows due to the unbalanced sample and the presence of an outlier.

The estimated coverage probabilities for scenario 1 are provided in Fig. 4. Our proposed method yields confidence intervals that are significantly more accurate than both the Mallows class and CUBIF confidence intervals when the outlier is not extreme (located along the segment between $(-1.5, -1.5)$ and $(0, 0)$). We note that the CUBIF confidence interval is the most inaccurate among all the robust confidence intervals when the outlier is extreme. The estimated coverage probabilities for scenario 3 are provided in Fig. 5. Here, the differences in coverage probabilities between the various robust methods are more pronounced. When the outlier is not extreme, our proposed method yields confidence intervals that are about 10% to 25% closer to the 95% line than the CUBIF and Mallows confidence intervals.

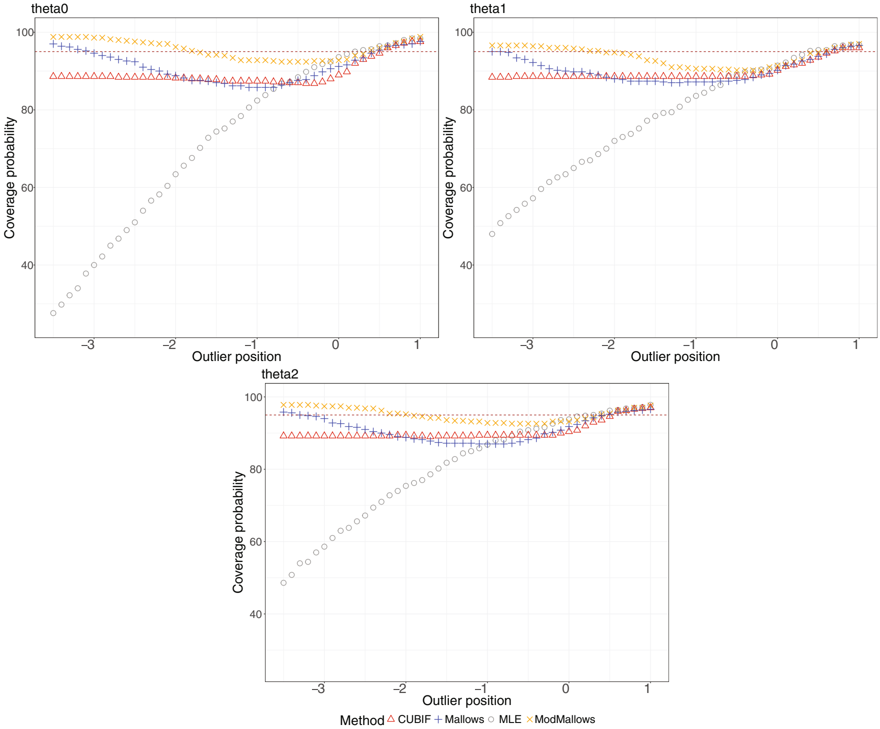


Fig. 4 Empirical coverage probabilities of 95% confidence intervals for each method (based on 500 balanced datasets with one outlier). The red dashed line marks a probability of 95%. A coverage probability that is close to 95% indicates accurate coverage

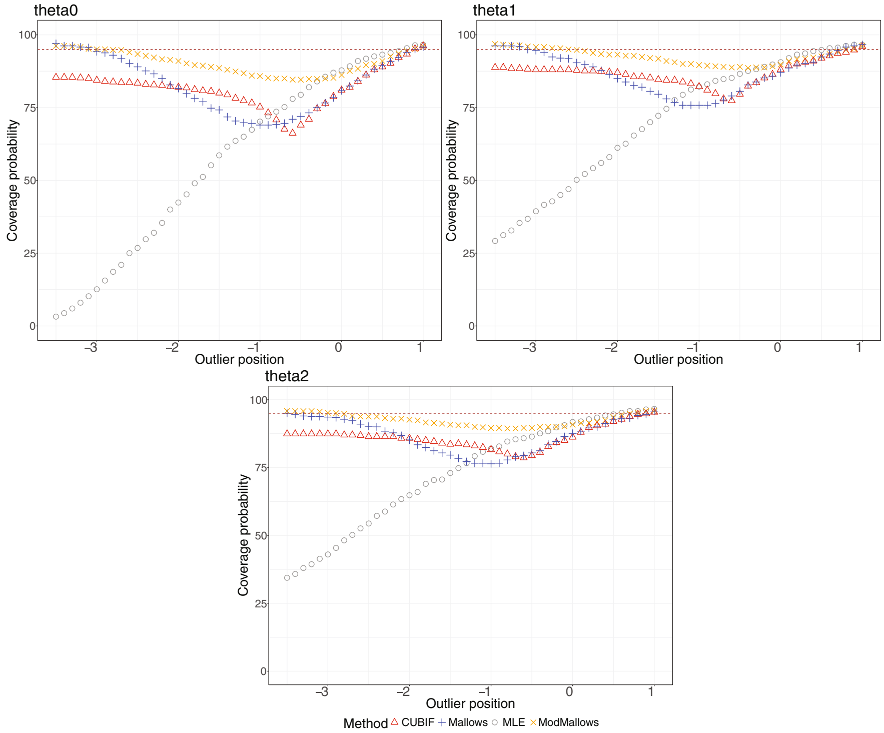


Fig. 5 Empirical coverage probabilities of 95% confidence intervals for each method (based on 500 unbalanced datasets with one outlier). The red dashed line marks a probability of 95%. A coverage probability that is close to 95% indicates accurate coverage

4.3 Leukemia Dataset

We assess the performance of the robust logistic regression estimators in the leukemia dataset that has been previously analyzed by Cook and Weisberg (1982), Johnson (1985), and Bianco and Yohai (1996). The dataset consists of white blood cell count (WBC, X_1), the albumin–globulin status of the white blood cells (AG, X_2), and a binary survival outcome variable Y. From Fig. 6, it is evident that there is an outlier from group 1. In fact, the dataset is often used in the literature for comparing the performance of robust logistic regression estimators. We consider the following logistic regression model for the data:

$$p(Y_i = 1 \,|\, \mathbf{x}_i) = Q(\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2}).$$

We compare the performance of the robust estimators against the MLE computed using all observations and the MLE computed without the outlier. The coefficient estimates and their corresponding standard errors are provided in Table 1. Results

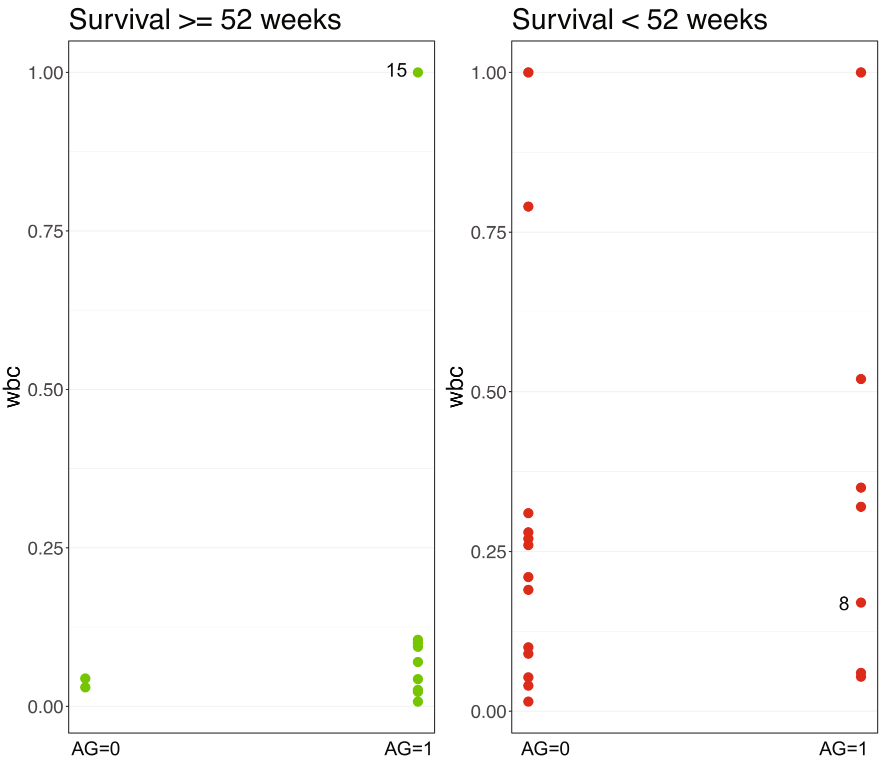


Fig. 6 Scatter plot of covariates for patients who survived ≥52 weeks (left) and who survived <52 weeks (right). Note the outlier at $(wbc, ag, y) = (100000, 1, 1)$ that corresponds to $i = 15$. Observation $i = 8$ is misclassified under the MLE computed using the full dataset and is correctly classified under the MLE computed without the outlier

Table 1 Coefficient estimates and standard errors of various logistic regression estimation methods

Coefficient	MLE	MLE w/o outlier	CUBIF	Mallows	ModMallows
θ_0	−1.31	0.21	−0.63	0.13	0.09
	(0.81)	(1.08)	(0.91)	(1.09)	(1.10)
θ_1	−3.18	−23.54	−9.61	−21.61	−19.94
	(1.86)	(13.54)	(5.20)	(13.49)	(14.26)
θ_2	2.26	2.56	2.26	2.50	2.43
	(0.95)	(1.23)	(1.04)	(1.21)	(1.22)

indicate that a more accurate classifier may be obtained by downweighting the outlier ($i = 15$). In particular, observation number 8 is misclassified under the MLE computed using all observations and is correctly classified under the MLE computed without the outlier.

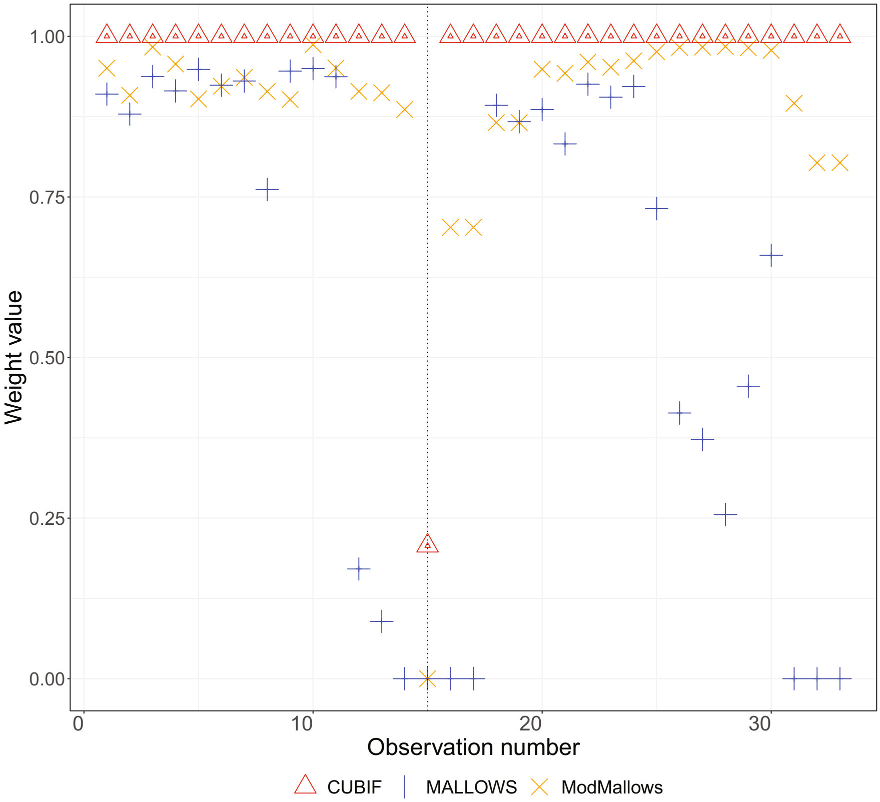


Fig. 7 Observation-specific weights allocated by the robust estimation methods. The vertical dotted line marks the weights for the outlier

The coefficient estimates for the Mallows and modified Mallows estimators are similar to the MLE without the outlier. The CUBIF estimator demonstrated some robustness, but its estimates are not as close to the MLE (without outlier) as its weight on the outlier is not as small (refer to Fig. 7). Also, note that as expected, the Mallows estimator tends to downweight more observations to achieve its robustness, while the proposed modification achieves the robustness while still keeping the weights higher. This reflects its ability to maintain higher efficiency alongside robustness.

5 Discussion

In this chapter, we propose a robust logistic regression estimator that involves modifying the weighting scheme of the Mallows class estimators. Our proposed approach is motivated by a new perspective to robust logistic regression modeling in which outliers are downweighted using the conditional covariate distribution. We justify our proposed method by showing that it leads to a sensible weighting scheme, whereas the existing robust estimators inadvertently downweight observations in the smaller group in an unbalanced dataset and fail to downweight outliers positioned near the mode of the complement group. We show that our proposed estimator is competitive with several existing methods. In particular, our proposed estimator achieves lower MSE when the outliers are positioned near the mode of the complementary group, whereas it achieves similar MSE when the outlier is either very extreme or in the absence of outlier.

We acknowledge that our proposed estimator may not perform well when the dataset without the outlier is nearly separable. To circumvent this issue, one may introduce a carefully positioned pseudo-observation to mitigate the effects of the separability. We recommend computing the group-specific leverages to assess the presence of a potential outlier before choosing a suitable estimator.

Taken together, the advantages of our proposed estimator suggest that it is viable option for robust logistic regression estimation.

Supplementary Codes
Supplementary codes for this article are available on the following website: https://github.com/weichangyu10/ConditionalWeightedLogistic.

References

Bianco, A., & Yohai, V. (1996). Robust estimation in logistic regression model. In H. Reider (Ed.), *Robust statistics, data analysis, and computer intensive methods*, Chap. Lecture Notes in Statistics 109, (pp. 17–34). New York: Springer-Verlag.

Bondell, H. D. (2005). Minimum distance estimation for the logistic regression model. *Biometrika*, *92*(3), 724–731.

Carroll, R. J., & Pederson, S. (1993). On robust estimation in the logistic regression model. *Journal of the Royal Statistical Society, Series B*, *55*(3), 693–706.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, *19*, 15–18.

Cook, R. F., & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman & Hall.

Copas, J. B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society, Series B*, *50*(2), 225–265.

Feng, J., Xu, H., Mannor, S., & Yan, S. (2014). Robust logistic regression and classification. In *Advances in Neural Information Processing Systems* (pp. 482–494).

Johnson, W. (1985). Influence measures for logistic regression: another point of view. *Biometrika*, *72*(1), 59–65.

Künsch, H. R., Stefanski, L. A., & Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, *84*(406), 460–466.

Martin, N., & Pardo, L. (2009). On the asymptotic distribution of cook's distance in logistic regression models. *Journal of Applied Statistics*, *36*(10), 1119–1146.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, *9*(4), 705–724.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, *38*(2), 485–498.

Stefanski, L. A., Carroll, R. J., & Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, *73*, 413–425.

Bias Calibration for Robust Estimation in Small Areas

Setareh Ranjbar, Elvezio Ronchetti, and Stefan Sperlich

Abstract It is well known that the existence of outliers in a sample can significantly affect the estimation of population parameters. Intuition suggests that this is even more the case in the context of small area estimation. If influential, outliers may heavily affect parameter estimates for areas in which they occur, especially when the domain-sample size is tiny. An obvious remedy is to use robust estimators but with the drawback of a potential bias. We compare different approaches, including some new ones, for bias calibration in this context. Among other findings, the simulations indicate that the new proposals can lead to more efficient estimators compared to existing methods. We conclude the study applying our estimators to obtain *Gini* coefficients in labor market areas of the Tuscany region of Italy. The new methods reveal a different picture than existing methods. We extend our ideas to predictions for non-sampled areas.

Keywords Asymmetric Huber function · Non-linear population parameters · Robust estimation · Robust prediction · Small area estimation

S. Ranjbar
HEC, University of Lausanne, Lausanne, Switzerland
e-mail: setareh.ranjbar@chuv.ch

E. Ronchetti (✉)
Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland
e-mail: elvezio.ronchetti@unige.ch

S. Sperlich
Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland
e-mail: stefan.sperlich@unige.ch

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_17

1 Introduction

Robust Statistics has been an important research area of Dave Tyler, whose contributions have been sometimes outside the mainstream literature. In our contribution for this Festschrift in his honor, we discuss robustness issues related to prediction, an area somewhat neglected in the robustness literature.

Small Area Estimation (SAE) has developed rapidly in recent years such that nowadays it is used in all kinds of official statistics, ranging from business decisions to attribution of health services or allocation of government funds. This is partly due to the high demand of statistics by policy makers on the one side, but also to the increasing data availability together with recent computational advances on the other side. Using sample surveys is a cost effective tool to provide estimates for characteristics of interest at population and sub-populations' (area/domain) level. This information, coming along with auxiliary data through administrative channels, is used for a better estimation of domain level parameters. Consequently, when sample sizes in the individual domains are too small ("small areas") to obtain reasonable mean square errors by means of direct estimates, then SAE "borrows strength from other existing sources of information." For a comprehensive review on this subject we refer to Tzavidis et al. (2018), Rao and Molina (2015), Chambers and Clark (2012), Longford (2005). Indirect estimators based on an explicit linking model are referred to as model-based estimators. Among these we concentrate on mixed effects models (MEM) with area-specific random effects that capture the between area variation beyond what is accounted for by auxiliary covariates, see Rao (2008), Datta (2009), Pratesi (2016), Jiang and Lahiri (2006), Pfeffermann (2013).

SAE techniques are intrinsically sensitive to outliers due to the small samples considered. Therefore, robust estimators have been proposed and developed in this field. Main streams of research on this topic are the robust version of the EBLUP (REBLUP) proposed by Sinha and Rao (2009), and the M-quantile (MQ) approach of Chambers and Tzavidis (2006). The former is based on bounded estimating equations for MEM. The latter captures the between area variation through the estimation of area-specific quantiles as coefficients; besides being robust against outliers it avoids problems associated with random effect prediction. For more recent developments on this method see Salvati et al. (2012), Pratesi et al. (2009), or Marchetti et al. (2017).

In robust estimation of finite population parameters, Chambers (1986) in his seminal paper, distinguished between projective and predictive estimation. The former refers to classical robust estimation where outliers are down-weighted or discarded in the estimation. In contrast, predictive estimation accounts for the so-called representative outliers, i.e., in-sample extreme observations which are likely to occur also among the non-sampled units. Calibration is necessary for the bias caused by down-weighting or disregarding these observations. A general bias calibration approach for the estimation of the finite population Cumulative Distribution Function (CDF) is proposed by Chambers and Dunstan (1986), and its robust version is presented in Welsh and Ronchetti (1998). For the SAE context, Tzavidis et al. (2010) introduced a general approach for the bias correction of existing robust estimators, and Chambers et al. (2014) discussed different methods to estimate the mean squared error (MSE) for those bias calibrated estimators. Although we concentrate here on MEM based SAE methods, it should be mentioned that many national statistical institutes have implemented other model-based estimators like GREG, and combined them with the so-called winsorization approach for robustness; see a recent paper by Favre-Martinoz et al. (2021) and the comparison of one-sided winsorization with M-estimation via asymmetric Huberization by Clark et al. (2017). We are not aware of a winsorization approach for bias calibration, and we will consider instead asymmetric Huberization.

There is an increasing literature on data transformation in SAE with MEM which is related to the issue mentioned above. Its main motivation is the fact that most of the inferential methods rely, some quite heavily, on the normality assumption for both, random area and individual effects. While taking the log of income and expenditures is quite common in empirical studies also for reasons of interpretation and theory about the underlying model, Box-Cox and Zellner transformations are typically not justified on theoretical grounds, but rather a device to achieve the desired distribution. If the variable of interest, however, is the untransformed one, a bias problem pups up here too (due to Jensen's inequality). Quite recently, Tzavidis et al. (2018) and Rojas-Perilla et al. (2020) introduce these transformations with data-driven parameters and automatic bias calibration. Kreutzmann et al. (2019) implemented these methods in the R-package *emdi*. These methods seem to work very efficiently for poverty and inequality mapping. Though their objective and nature are quite different from that of robust estimation in the presence of outliers, this seems to be a promising alternative approach. For comparison and completeness we therefore included them in our study, even though one has to be careful with the resulting conclusions.

In this article we first review the existing bias calibration methods for robust SAE. Then we propose two novel ideas for calibration of non-linear parameter estimators. First we argue that in situations where the errors come from highly skewed, heavy tailed distributions, one should use an asymmetric calibration to reflect the data generating process. We show how this can be implemented when deriving non-linear area parameters such as the *Gini* index. Then we suggest to linearize the area parameter by means of a von Mises linear approximation obtained by means of the Influence Function (IF) of the statistic, and then apply a calibration on this linear parameter. In both cases, bias calibration leads to more efficient estimators if the correction is done using the asymmetric Huber function with data-driven tuning parameter. Simulations indicate that the latter offers the lowest absolute bias, while the former achieves the lowest MSE compared to the existing bias calibration techniques. This proposal can be generally used for any other linear or non-linear parameter in small areas.

Finally note that in SAE in the absence of closed form formulae or good approximations, the MSE is often estimated by bootstrap; see, for example, Hall and Maiti (2006b), Hall and Maiti (2006a), and Chatterjee et al. (2008). These bootstrap methods or other Monte Carlo methods introduced for MEM in SAE are typically used to estimate the MSE of robust, bias calibrated estimators. Therefore we do not discuss the explicit MSE calculation for the different bias calibrated robust estimators. This again would be interesting but beyond the scope of this comparative study.

Section 2 introduces the general framework and the notation. In Sect. 3 we propose approaches to deal with the non-linear population parameters, using an asymmetric calibration of the estimates. Section 4 compares the performance of these and other existing methods by means of simulations. Section 5 discusses practical issues like the optimal tuning for the asymmetric calibration. In Sect. 6 we apply the methods to the EU-SILC survey and the census in Italy, to estimate *Gini* coefficients in the Tuscany region. Section 7 concludes.

2 General Framework and Notation

Consider the entire population (so-called super-population) $\mathcal{U}$ of size N, that is partitioned in d mutually disjoint sub-populations $\mathcal{U}_j$ of size N_j, corresponding to our small areas $j = 1, \cdots d$. For each area j we observe the outcome of interest Y_{ij} for a sub-sample, s_j, of individuals $i = 1, \ldots, n_j$ but not for the so-called unsampled subset r_j of size $N_j - n_j$. However, we assume that the auxiliary information $\mathbf{X}$, is available for all units, providing predictive power for the unobserved part of the population. This assumption could be avoided if we focus only on the linear parameters such as the mean of the area. The $\mathbf{x_{ij}}$ for individual i in area j is a column vector of dimension p that has 1 as its first component. We are interested in doing inference on Y_{ij} for the area-level. When n_j is too small for direct estimation, which would lead to large variances, or is not appropriate for other reasons, then model-based small area estimators are used. They apply a model on the super-population, typically to predict the unobserved Y_{ij} for the subsets r_j. Being interested in the distribution and the corresponding non-linear parameters (see Tzavidis et al. 2010), we do not consider area-level models like those in Fay and Herriot (1979), Dick (1995) or Pratesi and Salvati (2008). Instead, we consider unit level models that link the unit outcomes y_{ij} to the unit-specific covariates $\mathbf{x_{ij}}$, see, e.g., Battese et al. (1988).

As a basic setting, assume that the following mixed effect model (MEM) is in place for the sampled as well as for the unsampled units (i.e., without sampling selection bias):

$$y_{ij} = \mathbf{x_{ij}}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T u_j + \epsilon_{ij}, \qquad \forall j = 1, \ldots, d, \qquad \& \qquad \forall i = 1, \ldots, n_j, \tag{1}$$

where $\boldsymbol{\beta}$ is the p-dimensional vector of fixed effects, and u_j the random effects of the same dimension as $\mathbf{z}_{ij} \subset \mathbf{x}_{ij}$. In our application we concentrate on the commonly used nested error models, where z_{ij} contains only the 1. Standard assumptions are $u_j \overset{i.i.d}{\sim} (0, \sigma_u)$, and $\epsilon_{ij} \overset{i.i.d}{\sim} (0, \sigma_e)$ being individual error terms independent from the random effects. In our setting, however, to be more realistic we deviate from this assumption and allow for error terms that may belong to a heavily skewed distribution with potentially heavy tails for which the mean is not necessarily equal to zero. In addition, heteroscedasticity might be present.

Fitting the model to the sample at hand, one obtains estimates of the model parameters which are used to predict the unobserved Y_{ij}. By the substitution principle, once the Cumulative Distribution Function (CDF) for each area is estimated, further distribution related quantities (functional statistics) can be derived. Tzavidis et al. (2010) pointed out that the CDF estimate is particularly useful in cases where there are extreme values in the small area sample data, or if the small area distribution is highly skewed. The area-specific true CDF for a finite population in area j can be expressed as:

$$\begin{aligned} F_j(t) &= N_j^{-1}\Big[\textstyle\sum_{i\in s_j} \mathbb{1}\{y_{ij} \le t\} + \sum_{k\in r_j} \mathbb{1}\{y_{kj} \le t\}\Big] \\ &= N_j^{-1}\Big[\textstyle\sum_{i\in s_j} \mathbb{1}\{y_{ij} \le t\} + (N_j - n_j)F_j^{(2)}(t)\Big]. \end{aligned} \tag{2}$$

Population parameter that can be expressed as a functional of $F_j(t)$ can consequently be estimated as a functional of $\widehat{F}_j(t)$. In a naive setting we may use a plug-in estimator to obtain

$$\widehat{F}_j(t) = N_j^{-1}\left[\sum_{i\in s_j} \mathbb{1}\{y_{ij} \le t\} + (N_j - n_j)\widehat{F}_j^{(2)}(t)\right], \tag{3}$$

$$\widehat{F}_j^{(2)}(t) = \frac{1}{N_j - n_j}\sum_{k\in r_j} \mathbb{1}\{\widehat{y}_{kj} \le t\}.$$

In this case, the estimation of the distribution is obtained by predicting the unobserved units as $\hat{y}_{kj}$. This may be done by using different prediction methods suggested in the literature such as EBLUP, EB, HB, etc. In the presence of outliers or heavy tailed distribution, one would rather replace unobserved Y_{ij} by robust predictors. For instance, we could use robust mixed linear models to get an estimate of the model parameters and predict the robust version of EBLUP (called REBLUP) introduced by Sinha and Rao (2009). Alternatively one may use the M-quantile approach of Chambers and Tzavidis (2006) for estimation, and proceed accordingly. As mentioned above, other robust predictors can be used as the above distribution estimators are not predictor specific; these include estimators based on the one- or two-sided winsorization approach or on asymmetric Huberization.

3 Bias Calibration for Non-linear Parameter Estimates

Estimators for linear population parameters such as the mean or the total are well studied in the SAE literature. Estimating non-linear functionals is much more involved, and the calculations are not straightforward anymore. In such a case, a rather general approach is to use the estimate of the area-specific CDF F_j and then compute the statistics of interest by $\hat{T}_j = T(\hat{F}_j)$. However, in Eq. (3) using the expected value of any robust estimators to predict the outcome for non-sampled units i in area j results in a cumulative bias in the estimator $\widehat{F}_j$. Specifically in the presence of heteroskedastic and/or asymmetric errors, the bias will not cancel out by summation, see Tzavidis et al. (2010). The problem is even more prominent when there exist representative outliers (Chambers 1986), because these are extreme observations in the sample which are likely to occur also among the non-sampled units. To account for such bias, a calibration step is needed which has also the side-effect of causing some efficiency gains; see Chambers and Dunstan (1986), Welsh and Ronchetti (1998), Rao et al. (1990), Jiongo et al. (2013) for the SAE context.

A basic bias calibration for the CDF was proposed by Chambers and Dunstan (1986), namely

$$\widehat{F}_j^{CD}(t) = N_j^{-1}\left[\sum_{i\in s_j}\mathbb{1}\{y_{ij} < t\} + n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j}\mathbb{1}\{\widehat{y}_{kj} + (y_{ij} - \widehat{y}_{ij}) < t\}\right]. \tag{4}$$

In this case the effect of residuals, $y_{ij} - \widehat{y}_{ij}$, is not bounded. Welsh and Ronchetti (1998) extended this idea to obtain a bounded version of the prediction of a finite population CDF

$$\widehat{F}_j^{BC}(t) = N_j^{-1}\left[\sum_{i\in s_j}\mathbb{1}\{y_{ij} < t\}\right. \tag{5}$$

$$\left. + \, n_j^{-1}\sum_{i\in s_j}\sum_{k\in r_j}\mathbb{1}\{\widehat{y}_{kj}^{Rob} + w_j\phi_j\{(y_{ij} - \widehat{y}_{ij}^{Rob})/w_j\} < t\}\right],$$

where $\widehat{y}_{ij}^{Rob}$ and $\widehat{y}_{kj}^{Rob}$ are robust predictions of the observed and unobserved outcome, respectively, and w_j are robust estimates of the scale of the residuals in their area, like the median absolute deviation (MAD). Here, ϕ_j is a bounded influence function that can change over areas. Welsh and Ronchetti (1998) focus on one finite population; we extend this to several areas. They illustrate that in order to get a more efficient estimate for a finite population CDF, the truncation constant must change at different quantiles of the CDF, with larger constants for more extreme quantiles. Other calibration approaches are given for instance in Rao et al. (1990) and Jiongo et al. (2013).

Practitioners can use any robust method that has been developed in the field of SAE for estimation of the parameters in (1) and the prediction of $\widehat{y}_{ij}^{Rob}$ and $\widehat{y}_{kj}^{Rob}$ in (5). In this article we focus on the two most commonly used methods, namely REBLUP (Sinha and Rao 2009) and MQ (Chambers and Tzavidis 2006) in (5), together with a Symmetric Bias Calibration. We refer to them as REBLUP-BC and MQ-BC, respectively. Further building upon (5), we propose a skewed calibration that accounts for asymmetry of the error terms, and we extend this idea to correct for the bias in a linearized version of non-linear parameter estimates. This could also be interpreted as an extension of the calibration method of Welsh and Ronchetti (1998) to Eq. (5).

When there is extra knowledge available to the researcher, she should exploit this information to better calibrate the estimated CDF, and thereby its statistical functionals, say T_j for area j. For instance, it is common knowledge that income, wealth, or expenditure distributions are strongly skewed with a heavy tail to the right. One can use this information when predicting the distribution of each domain by applying an asymmetric calibration procedure. This requires two truncation constants for the skewed version of the Huber function:

$$\psi_{c,\gamma}(r) = \begin{cases} -c(\frac{2}{\gamma^2+1}) & if\ r \leq -c, \\ \frac{2}{\gamma^2+1}r & if\ -c < r < 0, \\ \frac{2\gamma^2}{\gamma^2+1}r & if\ 0 \leq r < c, \\ c(\frac{2\gamma^2}{\gamma^2+1}) & if\ r > c. \end{cases} \tag{6}$$

Here, c defines the width of the truncation window, and γ the degree of skewness. Like in the classical case of symmetric calibration, one chooses the optimal c and γ by minimizing $MSE(T_j)$. In presence of heteroskedasticity, we recommend to consider area-specific sets (c_j, γ_j).

The idea behind $\psi_{c,\gamma}(.)$ is the general presentation of skewed distributions along Fernandez and Steel (1998). The tuning parameter γ is always positive; while $\gamma = 1$ represent the original Huber function, values greater and smaller than 1 provide left and right skewed windows, respectively. From the definition of $\widehat{F}_{j|\widehat{u}_j}^{BC}$, r is a standardized residual divided by a robust estimate of its scale. Several choices of the latter are available. We use the one proposed by Rousseeuw and Croux (1993) which is based on the absolute pairwise differences of the residuals. It is an alternative to more traditional robust estimates and performs better for skewed distributions. Looking closer at $\psi_{c,\gamma}(.)$ one can see that this is very similar to the skewed Huber function of Chambers and Tzavidis (2006) defining the M-quantile method; namely

$$\psi_{c,q}(r) = 2\phi_c(r)[q\mathbb{1}\{r > 0\} + (1-q)\mathbb{1}\{r \leq 0\}],$$

where $\phi_c(.)$ is the classical Huber influence function, and $q = \frac{\gamma^2}{\gamma^2+1}$ the quantile index of the conditional outcome distribution, cf. Fig. 1. Notice, however, that here the skewed Huber function is used for calibration, not for estimation. We keep the

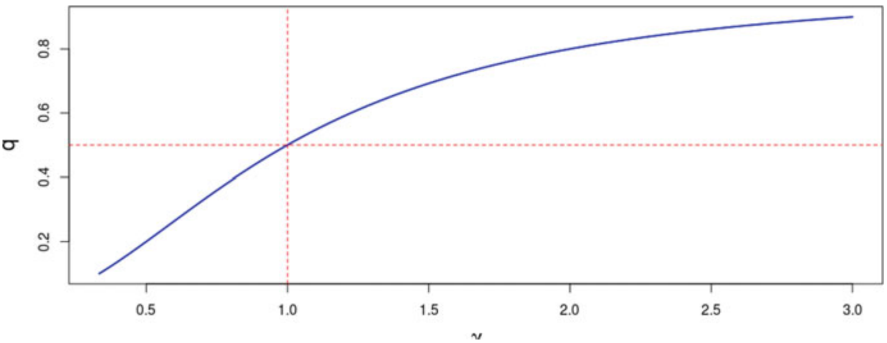


Fig. 1 The relation between q in $\psi_{c,q}(.)$ and γ in $\psi_{c,\gamma}(.)$

residuals effect bounded when searching for the shape of the true distribution. In practice, the optimal tuning constants are chosen by considering a mesh of a (c, γ) plane, and estimate the MSE (via bootstrapping) for each combination. As long as one allows for $\gamma = 1$, our method includes symmetric calibration.

Provided with the tuning parameter(s), the new area-specific CDF estimates are

$$\widehat{F}^{ABC}_{j|\widehat{u}_j}(t) = \frac{1}{N_j}\left[\sum_{i\in s_j} \mathbb{1}\{y_{ij} < t\}\right. \tag{7}$$

$$\left. + \, (n_j)^{-1} \sum_{i\in s_j} \sum_{k\in r_j} \mathbb{1}\left\{\widehat{y}^{Rob}_{kj} + w_j \psi_{c_j,\gamma_j}\left(\frac{y_{ij} - \widehat{y}^{Rob}_{ij}}{w_j}\right) < t\right\}\right]$$

with ψ_{c_j,γ_j} as in (6) but area-specific. Functionals and parameters like the *Gini* index can be calculated subsequently for each area. When REBLUP or MQ is used to predict $\widehat{y}^{Rob}_{kj}$ in Eq. (7), we refer to this method as REBLUP-ABC or MQ-ABC, respectively. As already discussed above, these are not the only options; other predictors or robustness approaches like, e.g., winsorization can be used for $\hat{y}^{Rob}_{ij}$. One may also argue that using an asymmetric Huberization could be useful in some applications throughout, not only for bias calibration. Our choices here, admittedly, were also driven by illustration and presentation aspects.

3.1 Linearization by the Influence Function

As mentioned above, we propose to first linearize the parameter of interest by means of the IF before applying our calibration. As the idea applies to all kind of predictors, we can suppress the superscript *Rob* here. We illustrate the idea with the popular example of the *Gini* coefficient. Consider the first order expansion introduced by von Mises (1947):

$$T(G) - T(F) = \int_{-\infty}^{+\infty} IF(y; T, F)d(G(y) - F(y)) + O\Big(\|G - F\|_2^2\Big), \tag{8}$$

where F is the (model) distribution, G a distribution in its neighborhood, and $IF(.; T, F)$ the influence function as defined by Hampel (1974). For $G := \widehat{F}_j$, this gives for $z_{ij} := IF(y_{ij}; T, F_j)$

$$\widehat{T}_j := T(\widehat{F}_j) \cong T(F_j) + \frac{1}{N_j}\sum_{i=1}^{N_j} IF(y_{ij}; T, F_j) = T(F_j) + \frac{1}{N_j}\sum_{i=1}^{N_j} z_{ij}. \tag{9}$$

Replace the unknown population parameter in (9) with a robust version $\widetilde{T}_j$, i.e.,

$$\widehat{T}_j = \widetilde{T}_j + \frac{1}{N_j}\left[\sum_{i\in s_j} z_{ij} + \sum_{k\in r_j} \widehat{z}_{kj}\right], \tag{10}$$

where $\widehat{z}_{kj} := IF(\widehat{y}_{kj}; T, F_j)$. Substituting robust predictors for all unobserved units, the calibration (hereafter IF-BC) is obtained by

$$\widehat{T}_j = \widetilde{T}_j + \frac{1}{N_j}\left[\sum_{i\in s_j} z_{ij} + \sum_{k\in r_j} \widehat{z}_{kj} + \frac{N_j - n_j}{n_j}\sum_{i\in s_j} w_j\phi(\frac{z_{ij} - \widehat{z}_{ij}}{w_j})\right], \tag{11}$$

where w_j is a robust estimate of the scale of the pseudo-residuals $\zeta_{ij} = z_{ij} - \widehat{z}_{ij}$ in area j, and $\phi(.)$ the Huber function. Extending this idea to asymmetric calibrations (see Sect. 3) gives

$$\widehat{T}_j = \widetilde{T}_j + \frac{1}{N_j}\left[\sum_{i\in s_j} z_{ij} + \sum_{k\in r_j} \widehat{z}_{kj} + \frac{N_j - n_j}{n_j}\sum_{i\in s_j} w_j\psi_{c_j,\gamma_j}(\frac{z_{ij} - \widehat{z}_{ij}}{w_j})\right], \tag{12}$$

with ψ_{c_j,γ_j} as in (6) with area-specific c_j, γ_j. This bias calibration is referred to as IF-ABC.

Calibration of the Gini Coefficient

In Sect. 6 we apply this to the *Gini* index defined as being twice the area between the 45° line and the Lorenz curve:

$$T(F) = 2\cdot\frac{I(F)}{\mu(F)} - 1,$$

where $I = I(F) = \int_0^{+\infty} tF(t)dF(t)$, and $\mu = \mu(F) = \int_0^{+\infty} tdF(t)$.

Suppressing the area sub-index j, the influence function of this functional is

$$IF(y; T, F) = 2 \cdot \Big(\frac{1}{\mu}\Big[\int_{y}^{+\infty} t dF(t) - I\Big]\Big) + 2 \cdot \frac{y}{\mu}\Big[F(y) - \frac{I}{\mu}\Big]; \tag{13}$$

see Appendix (7) for the derivation of (13). Then, using (9) one obtains

$$\begin{aligned}\widehat{T} &\cong T(F) + \frac{1}{N}\sum_{i=1}^{N} 2 \cdot \Big(\frac{1}{\mu}\Big[\int_{y_i}^{+\infty} t dF(t) - I\Big]\Big) + 2 \cdot \frac{y_i}{\mu}\Big[F(y_i) - \frac{I}{\mu}\Big] \\ &= T(F) + \frac{-4I}{\mu} + \frac{2}{\mu} \cdot \frac{1}{N}\sum_{i=1}^{N}\Big[\int_{y_i}^{+\infty} t dF(t) + y_i F(y_i)\Big],\end{aligned}$$

where in the last equality we approximate $\frac{1}{N}\sum_{i=1}^{N} y_i$ by $\mu(F)$. Replacing $T(F) = 2 \cdot \frac{I}{\mu} - 1$ gives

$$\widehat{T} \cong -T(F) - 2 + \frac{2}{\mu} \cdot \frac{1}{N}\sum_{i=1}^{N}\Big[\int_{y_i}^{+\infty} t dF(t) + y_i F(y_i)\Big]. \tag{14}$$

Setting $z_i = \int_{y_i}^{+\infty} t dF(t) + y_i F(y_i)$, the *Gini* (our non-linear parameter) in (14) is approximated by a linear function in z_i, which suggests an alternative estimator for the *Gini* coefficient. By replacing the unknown population parameters in (14) with robust estimates, substituting robust predictors for the unsamples units, and denoting by $\widetilde{T}$ and $\widetilde{\mu}$ the resulting estimates of the *Gini* coefficient and the population mean, respectively, we obtain a robust but biased estimate

$$\widehat{T} = -\widetilde{T} - 2 + \frac{2}{\widetilde{\mu}} \cdot \frac{1}{N}\Big[\sum_{i \in s} z_i + \sum_{k \in r} \widehat{z}_k\Big],$$

where $\widehat{z}_k := IF(\widehat{y}_k; T, F)$ as in Eq. (10). Adding calibration (12), in area-specific notation, i.e., reintroducing the sub-index j, one finally obtains

$$\begin{aligned}\widehat{T}_j^{ABC} &= -\widetilde{T}_j - 2 \\ &+ \frac{2}{\widetilde{\mu}_j} \cdot \frac{1}{N_j}\Bigg[\sum_{i \in s_j} z_{ij} + \sum_{k \in r_j} \widehat{z}_{kj} + \frac{N_j - n_j}{n_j}\sum_{i \in s_j} w_j \psi_{c_j,\gamma_j}\Big(\frac{z_{ij} - \widehat{z}_{ij}}{w_j}\Big)\Bigg].\end{aligned} \tag{15}$$

Summarizing, the implementation steps for the Gini estimate are:

Step 1. Use a robust estimator of the MEM to get robust predictions for unobserved y.

Step 2. In each area j, use observed y_{ij} for sampled, and robustly predicted for unsampled outcomes. Denote this vector by $\widetilde{Y}_j$ and calculate $\widetilde{T}_j, \widetilde{\mu}_j$.

Step 3. Put $\widetilde{Y}_j$ in ascending order, say $\widetilde{Y}_{(i)j}$, and compute $z_{ij} = \frac{1}{N}\sum_{h\geq i}\tilde{y}_{(h)j} + \frac{i}{N}\tilde{y}_{(i)j}$ for the sampled units, and $\widehat{z}_{kj} = \frac{1}{N}\sum_{h\geq k}\tilde{y}_{(h)j} + \frac{i}{N}\tilde{y}_{(k)j}$ for the unsampled ones. Now take only predicted outcomes for all units, sort them, and define $\widehat{z}_{ij} = \frac{1}{N}\sum_{h\geq i}\hat{y}_{(h)j} + \frac{i}{N}\hat{y}_{(i)j}$.

Step 4. Use (15) to get the bias calibrated estimates of the *Gini* coefficient for each area.

Slight modifications of Step 3 could be used for defining $\hat{z}_{ij}$ for the sampled units, but this is the computationally simplest version that we implemented in the next sections.

4 Model-based Simulation Study

To assess and compare the performance of existing and new proposals we conducted a series of simulations. We focus on bias and MSE of *Gini* estimates of various small area estimators under different scenarios. These methods can be grouped into "New-BC": REBLUP-ABC, MQ-ABC, and IF-ABC, denoting the new proposals introduced above, the most popular "Uncalibrated" estimators: EBLUP (Battese et al. 1988), REBLUP (Sinha & Rao 2009), MQ (Chambers & Tzavidis 2006), their bias corrected version "Classical-BC": REBLUP-BC and MQ-BC (Chambers et al. 2014), and "TF" which contains only the transformation based EBP (Rojas-Perilla et al. 2020) provided by the *emdi* R-package (see Sect. 1). It uses a EBP with data-driven Cox-Box transformation followed by a bias correction through parametric bootstrap. As said, this is not explicitly made for robust estimation in the presence of outliers, but particularly appropriate for asymmetric heavy tailed distributions.

There is no need to add simulations to the existing studies that look at standard scenarios for robust estimation like normal errors contaminated by extreme values. It is clear by construction that robust methods outperform non-robust ones and that bias calibrated versions will have smaller bias then. It has been studied less, however, what happens in the so-called representative outliers case with asymmetric extreme errors, in particular if the basic error distribution is not normal. As we look at the CDF in each small area, i.e., in the case of very small samples, there is hardly a difference in simulating directly a heavily skewed distribution with fat tails or a less skewed error distribution with slim tails but contaminated by extremes. This also explains why we included in our study the data-driven transformation based EBP (referred to as TF) and why it is sufficient to present only simulation results corresponding to the scenarios described below. Simulations with other, partly more complex scenarios, did not provide deeper insight. We generate a population of size $N = 15'000$, composed of $D = 50$ areas with $N_j = 300$ units in each area $j = 1, ..., D$. Then we draw a sample of size $n_j = 15$ from each area using SRSWOR (simple random sampling without replacement). The auxiliary variables

X_{ij} are i.i.d. $logNormal(mean = 1, sd = 0.5)$, and the outcome Y is generated by $y_{ij} = 100 + 5x_{ij} + u_j + \epsilon_{ij}$. We repeat each procedure $T = 1000$ times to calculate the relative prediction error (RPE) resulting from the difference between *Gini* predictions and their true values for each area j:

$$\mathrm{RPE}(Gini_j^{(t)}) = \frac{\text{predicted value}(Gini_j^{(t)}) - \text{true value}(Gini_j)}{\text{true value}(Gini_j)}.$$

The expected value of these relative errors over repeated sampling provides an estimate of the relative bias and MSE in each area, namely

$$\mathrm{RB}_j = \frac{1}{T}\sum_{t=1}^{T}\mathrm{RPE}(Gini_j^{(t)}) \text{ , } \mathrm{RRMSE}_j = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\Big(\mathrm{RPE}(Gini_j^{(t)})\Big)^2}.$$

As discussed in Sect. 3, the proposed new methods are expected to be especially beneficial when we are dealing with asymmetric distributions as in the case of incomes or expenditures. To create realistic scenarios, the error terms are generated from skewed t distributions with different degrees of freedom (df) and/or measures of skewness (λ) to generate the inequality we observe in different countries. Notice that the Gini coefficient typically ranges between 0.2 and 0.5. To avoid confusion, we use a different notation for the skewness parameter here than we used for tuning in the bias calibration. In all scenarios we keep the distribution of the random area effects as $u \sim N(0, 1)$. Note that McCulloch and Neuhaus (2011) showed that for linear MEM which only contain random intercepts uncorrelated with error terms or covariates, a misspecification of the random effect distribution introduces no or just a relatively small bias to the estimators of model parameters; see their paper for more details. This implies that the choice of the distribution of u_j would not impact our general conclusions and this is in accordance with our original simulations (not shown here). Table 1 summarizes the scenarios we considered in our simulations.

Table 1 Summary of scenarios: each simulation has $D = 50$ areas with population size of $N_j = 300$ units, and a sample size of $n_j = 15$ units from each area. Random effects are $u \sim N(0, 1)$, error terms are $\epsilon \sim St(\mathrm{df}, \lambda)$, where $St(\cdot, \cdot)$ denotes a right-skewed t distribution with *df* degrees of freedom, and $\lambda > 0$ its skewness parameter as introduced in Fernandez and Steel (1998). *Population Gini coefficient* refers to the average of calculated Gini coefficients using population data over all areas

Scenarios	df	λ	Population Gini coefficient
1-a	3	45	0.23
1-b	3	75	0.36
1-c	3	105	0.50
2-a	5	45	0.18
2-b	5	75	0.28
2-c	5	105	0.40

Relative Bias and RRMSE for the 50 areas are shown in the form of box plots for each method and scenario in Figs. 2 and 3. We summarize the results in Table 2 by giving the median of Relative Bias and RRMSE over the 50 areas under each scenario for all considered estimators. Looking at the distances between the boxplots and the red line, the first finding is that the calibrating methods outperform by far the three uncalibrated methods. The second finding is that the new calibrating methods clearly outperform the existing ones with symmetric calibration. Perhaps this is not surprising as we constructed the ABC methods such that they data-adaptively nest the BC methods. In spite of being purely data-driven, the box plots do not generally show an increase in the variability. An exception is the IF-ABC method, which in turn seems to be the best method in many scenarios for minimizing the bias. The third finding is that the data-adaptive-transformation based EBP, TF, achieves the minimum RRMSE in all the cases. It seems that our simulation scenarios (asymmetries and heavy tails) can be handled better by such a data-adaptive transformation than by correcting for outliers coming from a contamination. However, as said in our discussion above, in small samples (referring to the n_j) a contamination can hardly be distinguished from a heavy tail and/or asymmetry. As in practice we never know the truth, we would consider these two methods (TF and IF-ABC) as interesting complements. If the samples are sufficiently large, then working with mixtures (Chakraborty et al. 2019) would be another alternative.

5 Some Practical Issues

Before we apply these methods to our data for estimating the inequality in the different Labor Market Areas (LMAs) in Tuscany, we need to briefly address two practical issues. The first one arises when we need to provide a robust prediction for out-of-sample areas. In our data set (EU-SILC 2008), out of the 57 LMAs in Tuscany, only 29 are sampled. The second issue is of technical nature, as the proposed methods require two tuning parameters for calibration. Here we provide practical solutions for both problems.

5.1 Full Calibration vs. Partial Calibration

Recall that calibration is based on fitted model residuals: once a model is accepted to reflect well the Data Generating Process of the super-population and in the absence of sample selection problems, it is fitted to the sample to predict the unobserved outcomes using the auxiliary data. Clearly, one can also calculate robust predictions for the units with observed outcomes. The difference between the observed outcomes and their predictions is used for calibration. There are several ways to use these residuals to account for (representative) outliers. When

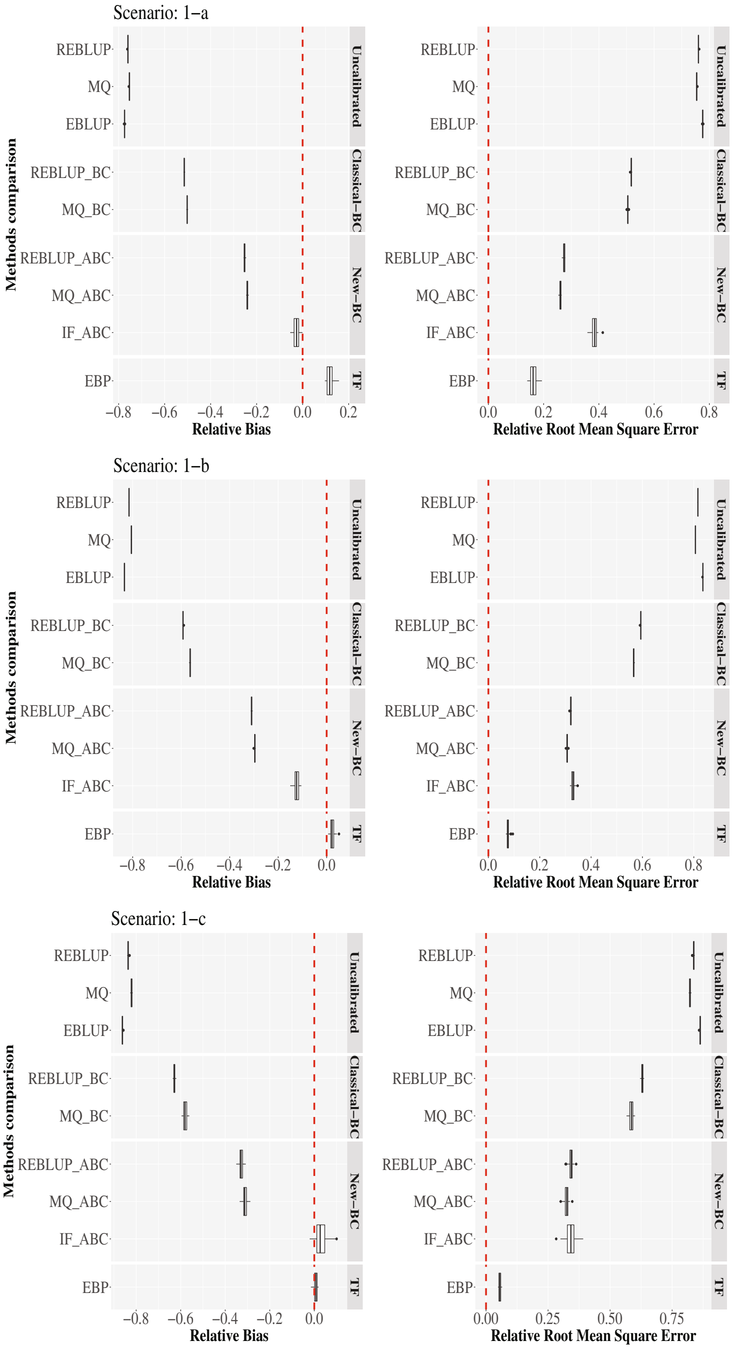


Fig. 2 The relative bias and square root relative MSE of the *Gini* coefficients, under scenarios 1-a:c from top to bottom

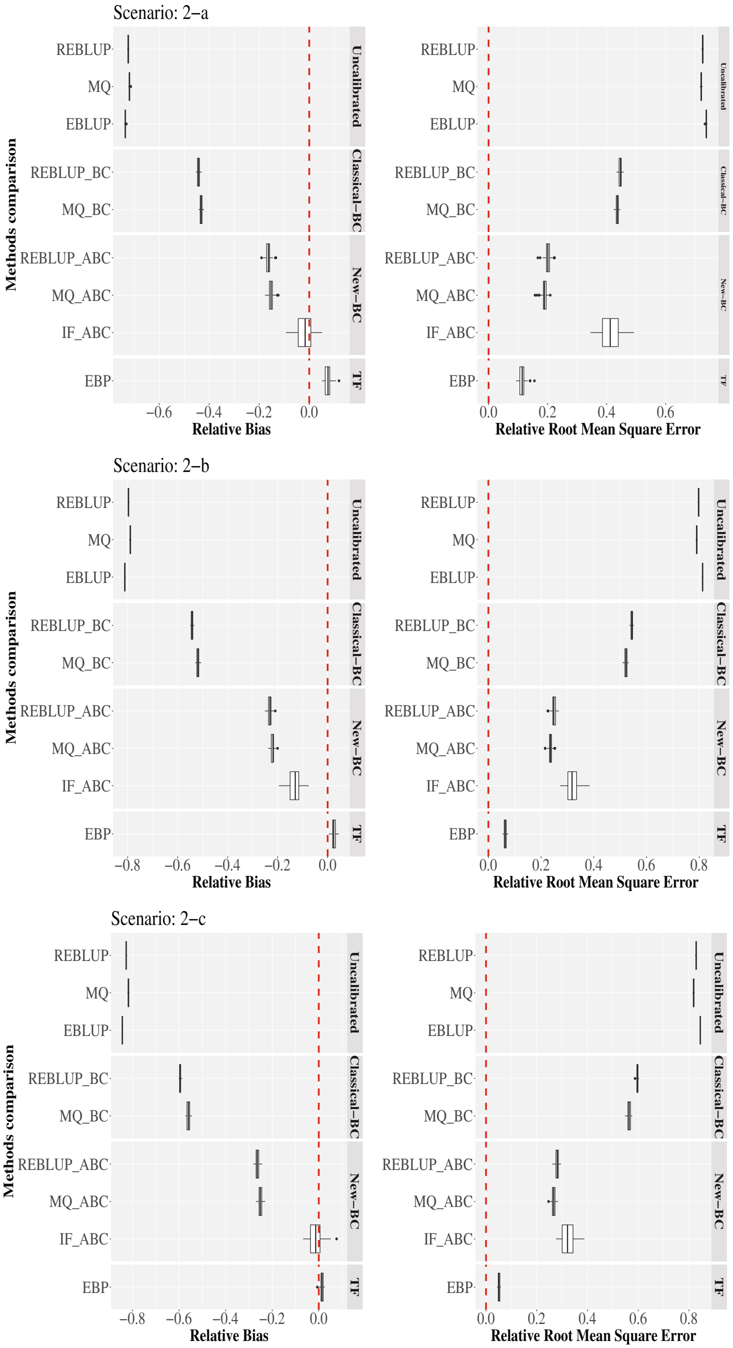


Fig. 3 The relative bias and square root relative MSE of the *Gini* coefficients, under scenarios 2-a:c from top to bottom

Table 2 Median of the areas' relative Bias and RRMSE, for the methods compared. For the exact data generating process of each scenario (1-a) to (2-c) see Table 1

Scenario	(1-a)		(1-b)		(1-c)	
Method	Rel. bias	RRMSE	Rel. bias	RRMSE	Rel. bias	RRMSE
EBLUP	−0.77	0.78	−0.83	0.84	−0.86	0.86
REBLUP	−0.76	0.76	−0.81	0.82	−0.84	0.84
MQ	−0.75	0.75	−0.80	0.81	−0.82	0.82
REBLUP-BC	−0.51	0.52	−0.59	0.59	−0.63	0.63
MQ-BC	−0.50	0.51	−0.56	0.57	−0.58	0.58
REBLUP-ABC	−0.25	0.27	−0.31	0.32	−0.33	0.34
MQ-ABC	−0.24	0.26	−0.30	0.31	−0.31	0.32
IF-ABC	−0.03	0.38	−0.12	0.33	0.03	0.3
TF	0.11	0.16	0.02	0.07	0.002	0.05
Scenario	(2-a)		(2-b)		(2-c)	
Method	Rel. Bias	RRMSE	Rel. Bias	RRMSE	Rel. Bias	RRMSE
EBLUP	−0.73	0.74	−0.81	0.81	−0.84	0.84
REBLUP	−0.72	0.73	−0.80	0.80	−0.83	0.83
MQ	−0.72	0.72	−0.79	0.79	−0.82	0.82
REBLUP-BC	−0.44	0.45	−0.54	0.54	−0.59	0.60
MQ-BC	−0.43	0.44	−0.52	0.52	−0.56	0.56
REBLUP-ABC	−0.17	0.20	−0.23	0.25	−0.26	0.28
MQ-ABC	−0.16	0.19	−0.22	0.24	−0.25	0.27
IF-ABC	−0.02	0.41	−0.14	0.32	−0.014	0.32
TF	0.06	0.10	0.02	0.06	0.01	0.05

we introduced our method, these residuals were used area-wise to correct for the bias in each specific area separately. Let us call this procedure "Partial Calibration." An alternative is to use each time the entire set of residuals (see Jiongo et al. 2013) and we call this procedure "Full Calibration." There are some differences between our concept and the one of Jiongo et al. (2013), where they also correct for the bias in the prediction of random effects. In our case, predicted area effects are considered as fixed, because we focus on the conditional CDF of each area. This is related, but again different, to what they call "Conditional Calibration." We introduce here a full calibration as partial calibration does not (or hardly) work for areas with no (or hardly any) observed outcome. Our proposal for these is to combine the full calibration idea with ours by choosing area-specific tuning constants for calibration even if using all residuals. This leads to a compromise that seems to work well (simulations not shown). Both, partial and this flexible full calibration are used and compared in the application, see next section. The full calibration analogues for (7) and (15) are

$$\begin{aligned}
\widehat{F}_{j|\widehat{u}_j}(t) &= \frac{1}{N_j}\left[\sum_{i\in s_j} I(y_{ij} < t)\right. \\
&\quad \left. + n^{-1} \sum_{i\in\bigcup_h s_h} \sum_{k\in r_j} I\left\{\widehat{y}_{kj}^{Rob} + w\psi_{c,\gamma}\left(\frac{y_{ih} - \widehat{y}_{ih}^{Rob}}{w}\right) < t\right\}\right] \\
\widehat{T}_j &= -\widetilde{T}_j - 2 \qquad\qquad (16) \\
&\quad + \frac{2}{\widetilde{\mu}_j}\cdot\frac{1}{N_j}\left[\sum_{i\in s_j} z_{ij} + \sum_{k\in r_j}\widehat{z}_{kj} + \frac{N_j - n_j}{n}\sum_{i\in\bigcup_h s_h} w\psi_{c,\gamma}(\frac{z_{ih} - \widehat{z}_{ih}}{w})\right],
\end{aligned}$$

where $z_{ij} = \int_{y_{ij}}^{+\infty} t dF_j(t) + y_{ij}F_j(y_{ij})$, and w a robust estimate of the scale of the entire vector of pseudo-residuals.

5.2 *Choice of the Tuning Parameters*

The choice of tuning parameters can play an important role in robust estimation and calibration. Therefore we provide a partly data-driven guideline to find optimal parameters for tuning, namely c and γ in (6). In symmetric calibration the convention is to use a rule of thumb for the width of truncating windows. But there also exist some guidelines for the best choice of tuning constants for calibrating certain population parameters, see, e.g., Welsh and Ronchetti (1998).

Generally, optimum tuning should minimize the MSE of the final estimator. For estimating the MSE of *linear* population parameter estimators, some analytic approximations have been proposed like first order Taylor expansion (Prasad & Rao 1990), defining the estimator as pseudo linear parameter (Chandra et al. 2011), or other approximations (Chambers et al. 2014). However, these do not account for the calibration, as they focus on the variance, not the bias. More importantly, there exist no general closed form expressions for the MSE of *non-linear* parameter estimators. Therefore, it is common to use the bootstrap; see Sect. 1. A non-parametric bootstrap (Hall & Maiti 2006a) can be used to obtain the MSE for our bias calibrated estimators.

One can estimate the MSE for different values of (c_j, γ_j) and select the tuning parameters that minimize the MSE. Note that rough MSE approximations will do, as long as they lead to the correct ranking. Such a procedure is explained in the Appendix. The drawback of this technique is that it can be computationally expensive. For cases where the computational burden is too heavy, we suggest the following alternative. Fix c_j as in the case of symmetric calibration along existing rules-of-thumb, see Chambers et al. (2014). For γ_j, needed for the asymmetric calibration, we propose

$$\widehat{\gamma}_j = \sqrt{n_j^- / n_j^+}, \tag{17}$$

where n_j^- and n_j^+ are the numbers of negative and positive centered residuals in area j. When using IF-ABC from Eq. (15), these refer to the residuals of the z_{ij}s. Appendix 7 provides some details on the derivation of this formula, which follows ideas used by Fernandez and Steel (1998) to estimate a transformation parameter to achieve a given skewness for a distribution.

In the simulations of Sect. 4, all c_j were fixed to 4 and 12 for symmetric and asymmetric calibration, respectively. In the case of IF-ABC the optimal choice had been between 3 to 4, depending on the scenario. For all asymmetric scenarios the area-specific estimate of γ_j was calculated according to (17). When γ_j is not stable due to very small samples or no observations, we may take $\widehat{\gamma} = \sqrt{n^- / n^+}$,where n^-, n^+ are the numbers of negative and positive centered residuals over the entire sample.

6 Estimating the Gini Coefficient for Labor Market Areas in Tuscany

In the following income study, our main interest focuses on the income inequality in LMAs regions of Tuscany, Italy. We apply the newly developed methods to estimate the *Gini* coefficient for all LMAs being provided with the EU-SILC 2008 sample survey of Italy and the 2001 census as an auxiliary source of information. From the survey we model the household equivalised disposable income on other explanatory variables at household and individual level. Since both, EU-SILC sample and census, have comparable covariates for individual characteristics, we can exploit the unit level model for this SAE. Specifically, the set of explanatory variables included in this study are gender, marital status, employment status and the years of education of the head of the household (household representative in the survey), as well as household size and household ownership status of the residence.

LMAs do not match with administrative boundaries, such that, though graphically and economically of great interest, they are not necessarily considered in the survey planning such as the EU-SILC database. Consequently, most of these regions must be regarded as small areas; see Table 3 for the number of observations in each area, and their ratio to the population size. In all LMAs less than 1% of the population is sampled. Moreover, for the 57 LMAs regions of Tuscany in the census, only for 29 of them we find observations in the sample. For the remaining 28 LMAs, direct estimation is not even possible. For these out-of-the-sample areas

Table 3 Description of EU-SILC data: population and sample size for the 29 sampled LMAs areas

Area	Population size	Sample size	Percentage sampled
1	13,265	75	0.57%
2	26,237	27	0.10%
3	29,875	17	0.06%
4	57,848	80	0.14%
5	45,010	33	0.07%
6	43,300	59	0.14%
7	47,363	73	0.15%
8	18,772	25	0.13%
9	15,081	48	0.32%
10	35,350	35	0.10%
11	281,036	261	0.09%
12	28,929	27	0.09%
13	70,240	59	0.08%
14	23,590	25	0.11%
15	71,461	95	0.13%
16	4619	25	0.54%
17	38,736	24	0.06%
18	33,258	29	0.09%
19	49,371	57	0.12%
20	11,577	27	0.23%
21	12,511	23	0.18%
22	44,078	118	0.27%
23	10,243	22	0.21%
24	42,662	75	0.18%
25	13,087	26	0.20%
26	38,111	35	0.09%
27	14,204	15	0.11%
28	2829	13	0.46%
29	92,408	132	0.14%

we use our fully calibrated indirect model predictions. For the other 29 regions we can alternatively also use partial calibration. We compare the results obtained with direct estimators, robust indirect estimators without calibration, REBLUP with symmetric and asymmetric calibration, as well as IF-ABC. For the sake of brevity we only show some selected results. Our rules-of-thumb (Sect. 5.2) suggest $c = 3$ and $c = 2$ for REBLUP-ABC and IF-ABC, respectively. We also study the effect of choosing the γ_j and γ according to the proposed method (17) compared to a range of alternative values.

6.1 Results for LMAs in Sample Areas with Partial Calibration

We first estimate the parameters for the 29 sampled LMAs (see Table 3) based on the presumingly more precise partial calibration. Apart from the robustness study regarding the choice of γ_j, Figs. 4 and 5 show the differences in the estimation of the *Gini* coefficient due to different calibration methods. Since we do not know the true values, we compare the results with the direct estimates which is supposed to be unbiased but with a large variance. We further compare them to indirect robust estimates. While Fig. 4 illustrates how Gini estimates of the areas vary over the different methods, Fig. 5 shows how asymmetrically bias calibrated estimates change with the choice of tuning parameters. It shows nicely that our bias calibrated estimators are not just alternatives to direct or robust indirect estimators, but actually offer an extremely useful compromise: while applying the smoothing-out of outlier effects, the bias calibration recovers the variation of the *Gini* coefficient over areas. The γ_j parameters allow us to move smoothly from one extreme to the other. The estimator (17) has a clear trend towards keeping the bias small, which is typically in the spirit of what practitioners would demand.

6.2 Results for All Areas with Full Calibration

If we want to predict the *Gini* coefficient for the 28 unsampled LMAs, we have to switch to full calibration. This can result in a heavy smoothing, making all areas looking quite similar unless the distributions of the covariates change dramatically over areas. For comparison reasons we give the estimates of the *Gini* coefficients with full calibration for all LMA areas, i.e., sampled and non-sampled—even though in practice one would probably take partial calibration for the sampled ones. In the 28 unsampled areas we predict income for all households by setting the area random effect $\hat{u}_{.} = \hat{u}_{(0.5)}$, i.e., to the median of predicted random effects. Then we use the entire vector of residuals to correct for the bias. Regarding tuning parameters, c is fixed as before, but γ is now estimated once for all areas (i.e., for simplicity not even varying for sampled areas) using again the entire vector of residuals in the algorithm introduced in Sect. 5.

In Figs. 6 and 7 we see the stronger smoothing effect of full calibration. Note that the scales in the maps are automatically set by R, and therefore different to those in Fig. 4. Perhaps surprisingly, here REBLUP-ABC is closer to IF-ABC than to REBLUP-BC. Figure 7 is confirming both, what we already found in the context of partial calibration (i.e., a strong bias correction effect) as well as what we have seen in Fig. 7 (i.e., the variation over areas has been strongly smoothed). Not surprisingly, the effect of the γ choice seems to be attenuated.

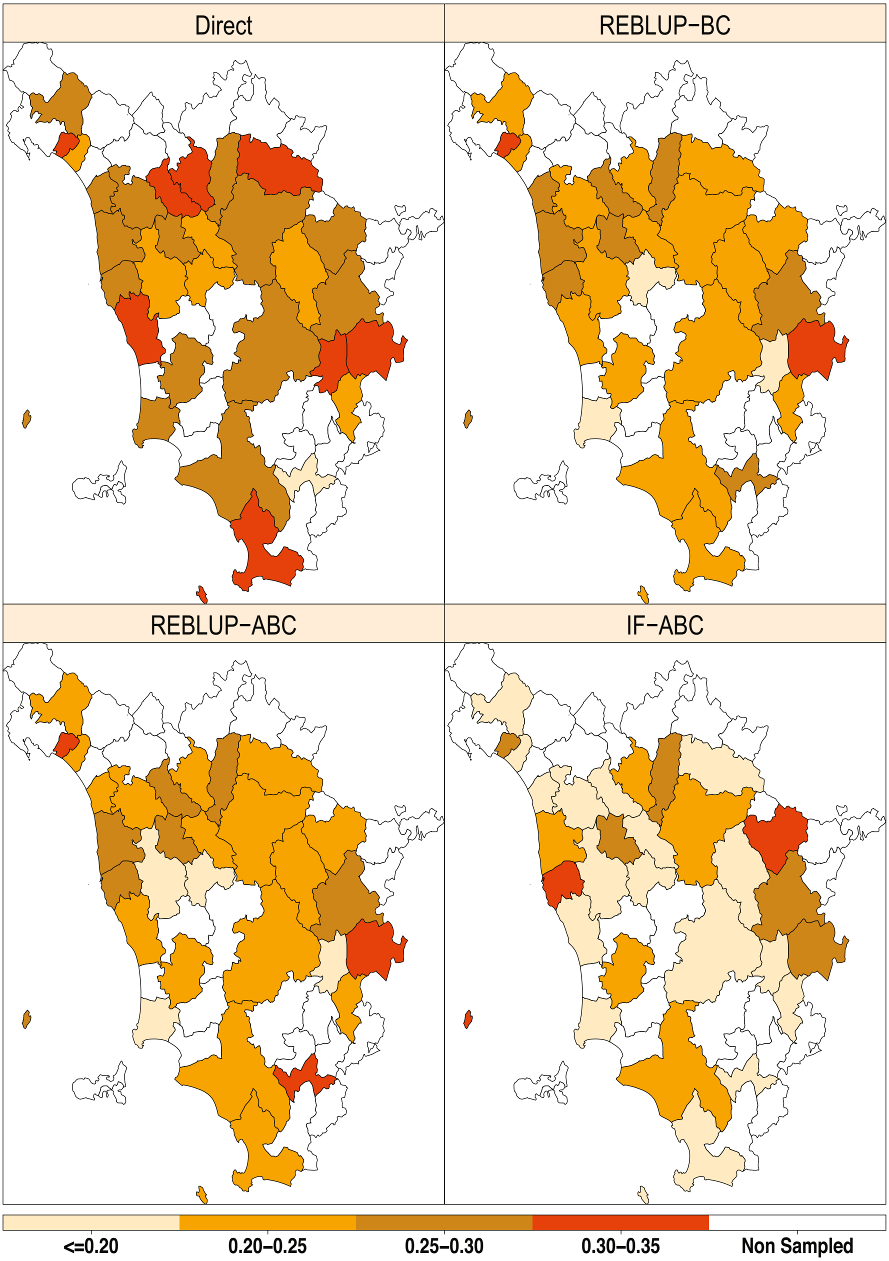


Fig. 4 Gini estimates for the 29 sampled LMAs of Tuscany using different estimation method with partial calibration

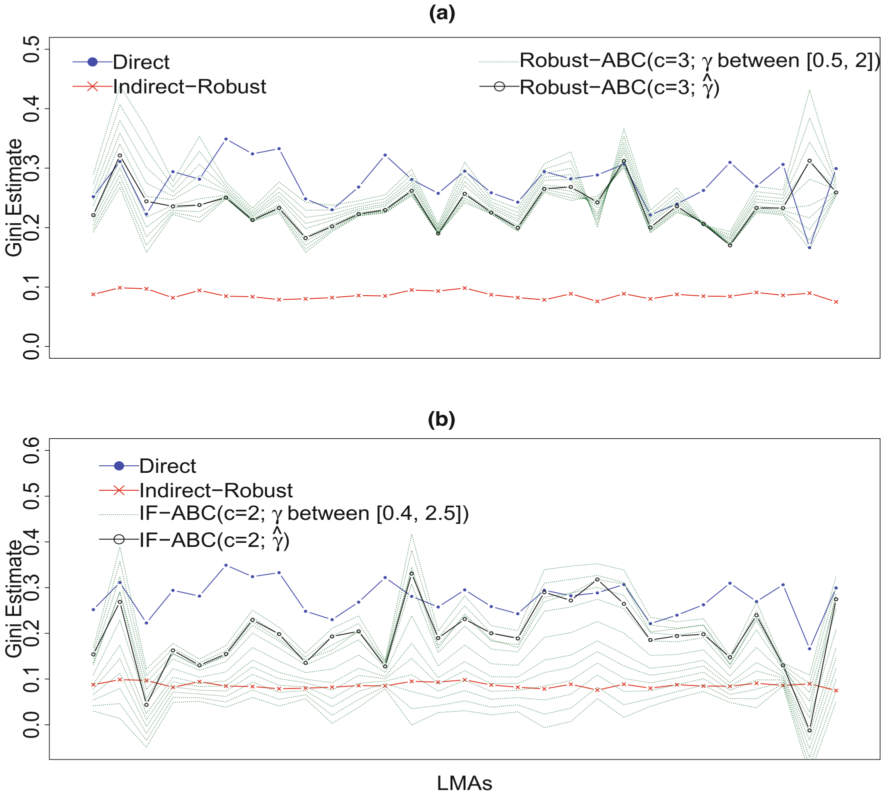


Fig. 5 Gini estimates, comparing direct and robust indirect estimators with REBLUP-ABC (**a**) and IF-ABC (**b**) calibration; the 29 sampled areas using partial calibration

Taking all together, we can clearly recommend the use of asymmetric bias calibration for the indirect robust estimators in SAE. One may prefer the calibration via CDF when the aim is to minimize the MSE, but the asymmetric calibration through IF if the aim is to minimize the bias. In both cases, full calibration is only recommended for out-of-sample areas to provide some bias correction also for these areas.

There might be some situations where the combination of the two full and partial calibration is more beneficial but this is a potential for future research and is out of the scope of this paper.

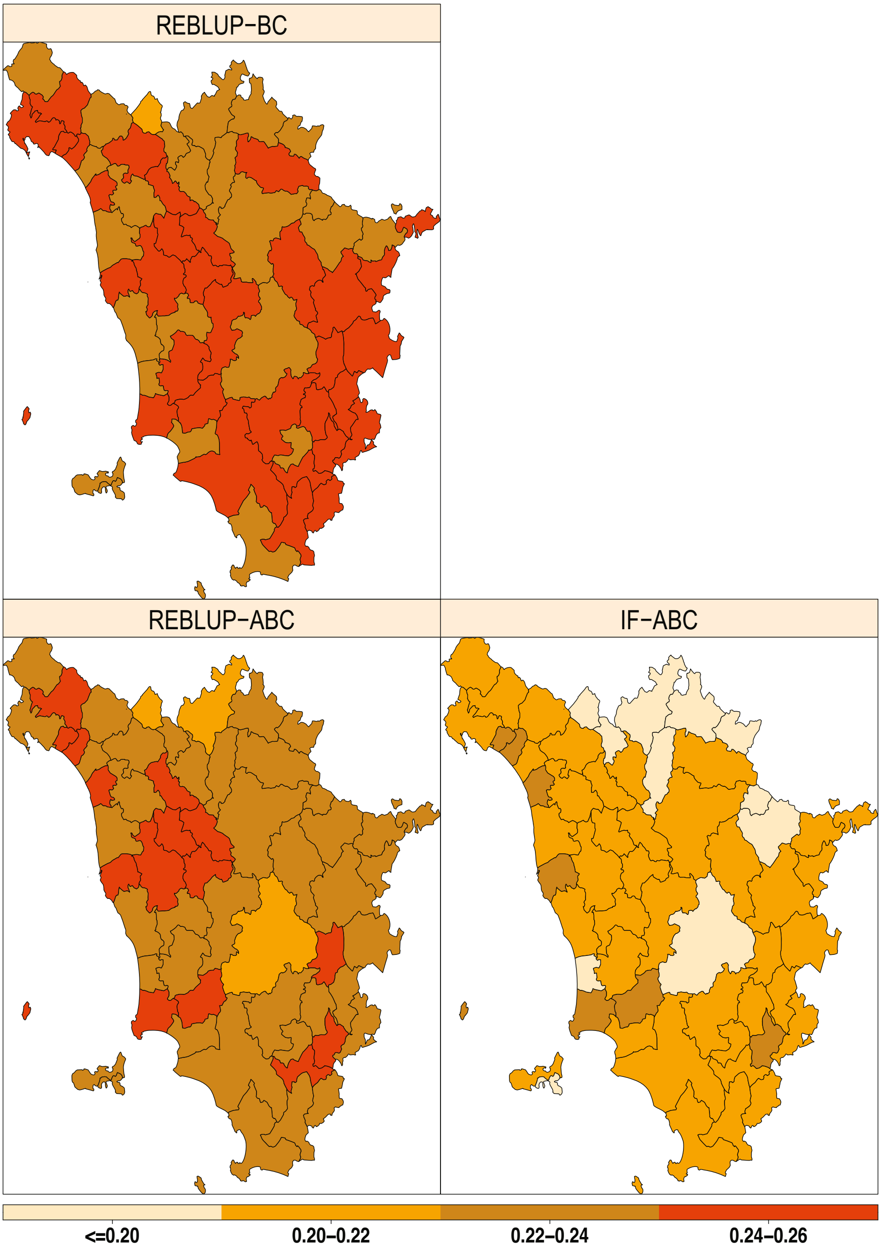


Fig. 6 *Gini* estimates for all 57 LMAs of Tuscany using different estimation methods with full calibration

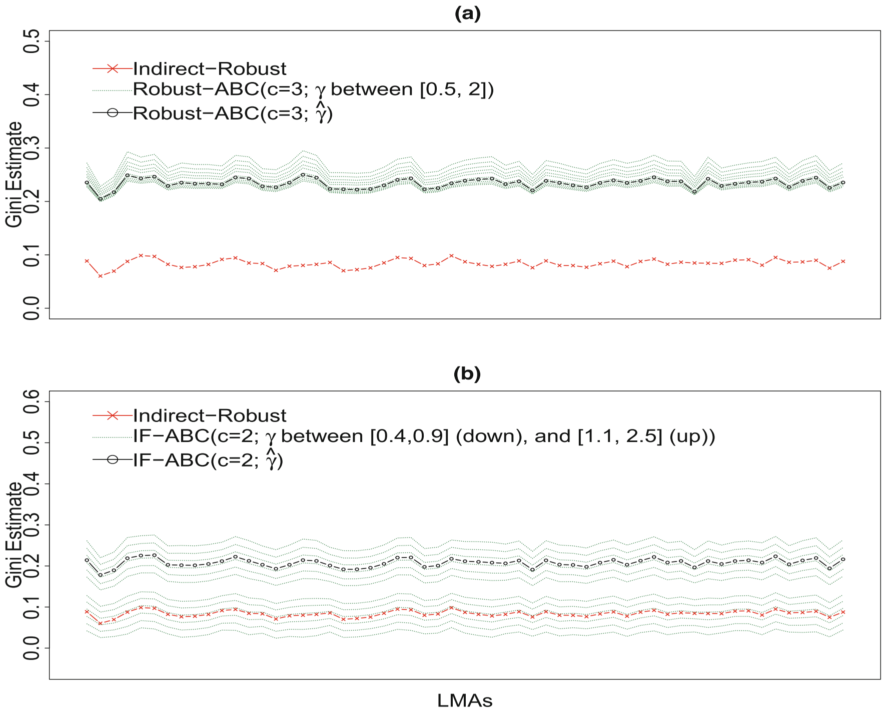


Fig. 7 *Gini* estimates, comparing robust indirect estimators with REBLUP-ABC (**a**) and IF-ABC (**b**) calibration; all areas using full calibration

7 Conclusions and Further Discussion

We review bias calibration methods that exist in the literature of SAE and extend these by the use of an asymmetric Huber function. This nests also the existing methods but can be quite beneficial when the practitioner is dealing with skewed distributions. Furthermore, for the problem of estimating non-linear parameters we propose to use linear approximations through the influence function to robustify and calibrate these linearized versions. Necessary tuning parameters can be chosen data-adaptively, and modifications of the calibration allow its application to non-sampled areas. Our simulations confirm the efficiency gain using these approaches compared to the existing bias calibration methods. While this was mainly shown along the objective of estimating the Gini coefficient, it is clear that these methods can be applied to other settings. However, we also find that the quite recently proposed transformation based EBP with Monte Carlo based calibration (Kreutzmann et al. 2019; Rojas-Perilla et al. 2020) is a very attractive alternative or complement, at least in the context of right skewed distributions. Further note that if quantile estimation is the main focus, the paper of Chen and Liu (2019) provides various alternatives to address non-normality.

We use these methods to estimate the income inequality for all LMAs in Tuscany, Italy. In this application we can illustrate the usefulness of calibration which exhibits quite serious shifts indicating important bias corrections. Also, it shows that full calibration, though useful for doing bias calibration in the non-sampled areas, has a strong smoothing effect. Thus, partial calibration should be the preferred choice, where applicable.

Acknowledgments We would like to thank the editors and two anonymous referees for their useful remarks, which improved the original version of the manuscript.

Appendix

Influence Function of the Gini Coefficient

Consider the *Gini* coefficient $T(F) = 2 \cdot \frac{I(F)}{\mu(F)} - 1$, where $I = I(F) = \int_0^{+\infty} tF(t)dF(t)$, $\mu = \mu(F) = \int_0^{+\infty} tdF(t)$, and define $F_{\epsilon,y}(t) = (1-\epsilon)F(t) + \epsilon\delta_y(t)$, where $\delta_y(t) = \mathbb{1}\{t \geq y\}$. Then

$$\begin{aligned}
T(F_{\epsilon,y}) &= \frac{2\int_0^{+\infty} t\big((1-\epsilon)F(t) + \epsilon\delta_y(t)\big)d\big[(1-\epsilon)F(t)+\epsilon\delta_y(t)\big]}{\int_0^{+\infty} td\big[(1-\epsilon)F(t)+\epsilon\delta_y(t)\big]} - 1 \\
&= \frac{2}{(1-\epsilon)\int_0^{+\infty} tdF(t) + \epsilon\int_0^{+\infty} td\delta_y(t)}\Bigg\{(1-\epsilon)^2\int_0^{+\infty} tF(t)dF(t) \\
&+ \epsilon(1-\epsilon)\int_0^{+\infty} t\delta_y(t)dF(t) + \epsilon(1-\epsilon)\int_0^{+\infty} tF(t)d\delta_y(t) \\
&+ \epsilon^2\int_0^{+\infty} t\delta_y(t)d\delta_y(t)\Bigg\} - 1.
\end{aligned}$$

Using the definition of Dirac delta function we obtain

$$\int_0^{+\infty} tF(t)d\delta_y(t) = y \cdot F(y), \qquad \int_0^{+\infty} t\delta_y(t)d\delta_y(t) = y.$$

Therefore, it follows

$$T(F_{\epsilon,y}) = 2 \cdot \frac{(1-\epsilon)^2 I + \epsilon(1-\epsilon)\int_y^{+\infty} tdF(t) + \epsilon(1-\epsilon)yF(y) + \epsilon^2 y}{(1-\epsilon)\mu + \epsilon y} - 1, \tag{18}$$

and by definition given in Hampel (1974), the influence function of the functional T is

$$IF(y;T,F) = \frac{d}{d\epsilon}T(F_{\epsilon,y})\mid_{\epsilon=0} = 2\cdot\frac{\left(-2I + yF(y) + \int_y^{+\infty} tdF(t)\right)\cdot\mu - I\cdot(y-\mu)}{\mu^2}$$
$$= 2\cdot\left(\frac{1}{\mu}\left[\int_y^{+\infty} tdF(t) - I\right]\right) + 2\cdot\frac{y}{\mu}\left[F(y) - \frac{I}{\mu}\right].$$

Bootstrap for RMSE and Tuning Parameter Selection

Bootstrap procedures are quite popular in SAE as they account for the dependence structure of the data, see Hall and Maiti (2006a) and Sperlich and José Lombardia (2010). Notice that the so-called naive or pair bootstrap is not adequate in this context. Given the large literature on bootstrap-based MSE estimation, we concentrate here on the conditional RMSE estimation. We propose a bootstrap method that is computationally inexpensive but helps us in approximating the MSE or RMSE for choosing the tuning parameters. Since we focus on the conditional distribution for each area (see (7)), we only sample from the error terms but not from the random effects. Once random effects are predicted, we consider them as fixed. This will, indeed, disregard the between area variation and will lead to an underestimation of the RMSE. However, recall that the aim is not to provide a precise estimate of the RMSE, but to find the tuning parameters minimizing RMSE.

It is well known that the presence of outliers in the residuals can harm the residual bootstrapping procedure. To avoid this problem, we do the bootstrap sampling from the pool of huberized residuals, see Singh (1998) for the breakdown theory of bootstrap quantiles. To evaluate the RMSE for a given pair (c_b, γ_b), we need, say, "a more relaxed" pair $(c = c_2 > c_b, \gamma = 1)$, to huberize the residuals a priori for the bootstrap sampling, see step 4 below. Specifically, the bootstrap algorithm consists of the following steps:

1. Fit the model with a robust estimator without calibration to get estimates and predictions for fixed and random parameters.
2. Pick a combination of c_{boot} and γ_{boot} from a mesh over a predefined domain.
3. Estimate the bias calibrated parameter of interest for each area, say $\widehat{Gini}_j^{BC}$, called "the original estimate" hereafter. Now, $\{y_{ij}\} \cup \{\widehat{y}_{kj}\}$ for $j \in s_j$ and $k \in r_j$ are considered to be the original population values for area j.
4. Get the residuals for each area from the original fit, and huberize them as

$$\widehat{res}_{ij} = \psi_{c_2,1}\big((y_{ij} - \widehat{y}_{ij})/\widehat{w}_j\big)\cdot\widehat{w}_j, \qquad \text{with } c_2 > c_{boot},$$

$\widehat{w}_j$ being robust estimates of the scale for the residuals of area j, e.g.,

$$\widehat{w}_j = \big(1.4826 \times median(\mid \widehat{\epsilon_{ij}} \mid)\big).$$

5. Sample randomly with replacement, separately from each area set of the huberized residuals, and stack them to build a vector of bootstrap residuals res^*_{ij}.
6. Construct the bootstrap sample by setting $y^*_{ij} = \widehat{y}_{ij} + res^*_{ij}$.
7. Using this bootstrap sample together with the original design of x, fit the bootstrap sample to predict the unobserved units, y^*_{kj}, for $k \in r_j$.
8. The bootstrap population outcome set is $\mathcal{U}^{*(b)}_j = \{y_{ij}\} \cup \{y^*_{kj}\}$. For this population, estimate the parameters of interest, e.g., $Gini^{*(b)}_j$ where the calibration is done by means of the bootstrap model residuals.
9. Repeat steps 4.-7., B times, each time calculating the error $\left(\frac{Gini^{*(b)}_j - \widehat{Gini}^{BC}_j}{\widehat{Gini}^{BC}_j}\right)$.
10. The estimated RRMSE and Bias are approximated by

$$RRMSE(c_{boot}, \gamma_{boot}) = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{Gini^{*(b)}_j - \widehat{Gini}^{BC}_j}{\widehat{Gini}^{BC}_j}\right)^2,$$

$$Bias(c_{boot}, \gamma_{boot}) = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{Gini^{*(b)}_j - \widehat{Gini}^{BC}_j}{\widehat{Gini}^{BC}_j}\right).$$

11. Repeat steps 2.-10. for all sensible combinations of $\{c_{boot}$ and $\gamma_{boot}\}$, and choose the pair $\{c_{boot}, \gamma_{boot}\}$ that gives the smallest RRMSE.

Details on the Estimator for Tuning Parameters

Let $f(.)$ be a unimodal symmetric distribution around 0. Fernandez and Steel (1998) define a class of asymmetric distributions by

$$p(\epsilon \mid \gamma) = \frac{2}{\gamma + \frac{1}{\gamma}}\left\{f(\frac{\epsilon}{\gamma})\mathbb{1}_{(-\infty,0)}(\epsilon) + f(\gamma\epsilon)\mathbb{1}_{[0,\infty)}(\epsilon)\right\}. \tag{19}$$

It follows that

$$\frac{Pr(\epsilon < 0 \mid \gamma)}{Pr(\epsilon \geq 0 \mid \gamma)} = \frac{\int_{-\infty}^{0}\frac{2}{\gamma+\frac{1}{\gamma}}f(\frac{\epsilon}{\gamma})d\epsilon}{\int_{0}^{\infty}\frac{2}{\gamma+\frac{1}{\gamma}}f(\gamma\epsilon)d\epsilon},$$

and by change of variables

$$\frac{Pr(\epsilon < 0 \mid \gamma)}{Pr(\epsilon \geq 0 \mid \gamma)} = \frac{\int_{-\infty}^{0}f(z)\gamma dz}{\int_{0}^{\infty}f(z)\frac{1}{\gamma}dz} = \gamma^2, \tag{20}$$

where the last equality holds since $f(.)$ is symmetric around 0. Assume that our model residuals, used for calibration, follow distribution (19), and try to estimate the two probabilities involved in (20) by

$$\widehat{Pr}(\epsilon < 0 \mid \gamma) = \frac{n^{-}}{N}, \qquad \widehat{Pr}(\epsilon \geq 0 \mid \gamma) = \frac{n^{+}}{N},$$

where n^{-}, n^{+}, and N are the numbers of positive, negative, and total residuals. Therefore, a heuristic estimation of γ, i.e., the skewness factor for the residuals is $\widehat{\gamma} = \sqrt{n^{-}/n^{+}}$. A feasible algorithm to obtain data-driven tuning parameters in (6) is the following

1. Center the block of residuals in each area.
2. Fix the constant c at a given value. Values between 2 and 4 seem to provide a good performance in practice, see Chambers et al. (2014).
3. Count the number of positive and negative centered residuals in each area: n_j^{+} and n_j^{-} for area j and set $\widehat{\gamma}_j = \sqrt{n_j^{-}/n_j^{+}}$.

References

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28–36.

Chakraborty, A., Datta, G., & Mandal, A. (2019). Robust hierarchical Bayes small area estimation for the nested error linear regression model. *International Statistical Review*, *87*(S1), S158–S176.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, *81*(396), 1063–1069.

Chambers, R. L., Chandra, H., Salvati, N., & Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 47–69.

Chambers, R. L., & Clark, R. (2012). *An introduction to model-based survey sampling with applications*. Oxford Statistical Science Series.

Chambers, R. L., & Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, *73*(3), 597–604.

Chambers, R. L., & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, *93*(2), 255.

Chandra, H., Tzavidis, N., & Chambers, R. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, *37*(2), 153–170.

Chatterjee, S., Lahiri, P., & Li, H. (2008). Parametric bootstrap approximation to the distribution of eblup and related prediction intervals in linear mixed models. *Annals of Statistics*, *36*(3), 1221–1245.

Chen, J., & Liu, Y. (2019). Small area quantile estimation. *International Statistical Review*, *87*(S1), S219–S238.

Clark, R., Kokic, P., & Smith, P. (2017). A comparison of two robust estimation methods for business surveys. *International Statistical Review*, *85*(2), 270–289.

Datta, G. S. (2009). Model-based approach to small area estimation. *Handbook of Statistics*, *29*, 251–288. Handbook of Statistics.

Dick, P. (1995). Modelling net undercoverage in the 1991 canadian census. *Survey Methodology*, *21*(1), 45–54.

Favre-Martinoz, C., Haziza, D., & Beaumont, J.-F. (2021). Efficient nonparametric estimation for skewed distributions. *The Canadian Journal of Statistics*, *49*(2), 471–496.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277.

Fernandez, C., & Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, *93*(441), 359–371.

Hall, P., & Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics*, *34*(4), 1733–1750.

Hall, P., & Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 221–238.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383–393.

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST*, *15*(1), 1.

Jiongo, V. D., Haziza, D., & Duchesne, P. (2013). Controlling the bias of robust small-area estimators. *Biometrika*, *100*(4), 843–858.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, *91*(7).

Longford, N. T. (2005). *Missing data and small-area estimation*. London: Springer-Verlag.

Marchetti, S., Giusti, C., Salvati, N., & Pratesi, M. (2017). Small area estimation based on m-quantile models in presence of outliers in auxiliary variables. *Statistical Methods and Applications*, *26*, 1–25.

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, *26*(3), 388–402.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, *28*(1), 40–68.

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*(409), 163–171.

Pratesi, M. (2016). *Analysis of poverty data by small area estimation*. John Wiley and Sons.

Pratesi, M., Ranalli, M. G., & Salvati, N. (2009). Nonparametric m-quantile regression using penalised splines. *Journal of Nonparametric Statistics*, *21*(3), 287–304.

Pratesi, M., & Salvati, N. (2008). Small area estimation: The EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, *17*(1), 113–141.

Rao, J. N. K. (2008). Some methods for small area estimation. *Rivista Internazionale di Scienze Sociali*, *116*(4), 387–406.

Rao, J. N. K., Kovar, J. G., & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, *77*(2), 365–375.

Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation*. John Wiley and Sons Ltd.

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(1), 121–148.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283.

Salvati, N., Tzavidis, N., Pratesi, M., & Chambers, R. L. (2012). Small area estimation via M-quantile geographically weighted regression. *TEST*, *21*(1), 1–28.

Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The Annals of Statistics*, *26*(5), 1719–1732.

Sinha, S. K., & Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, *37*(3), 381–399.

Sperlich, S., & José Lombardia, M. (2010). Local polynomial inference for small area statistics: Estimation, validation and prediction. *Journal of Nonparametric Statistics*, *22*(5), 633–648.

Tzavidis, N., Marchetti, S., & Chambers, R. L. (2010). Robust estimation of small-area means and quantiles. *Australian and New Zealand Journal of Statistics*, *52*(2), 167–186.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(4), 927–979.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, *18*(3), 309–348.

Welsh, A., & Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(2), 413–428.

The Diverging Definition of Robustness in Statistics and Computer Vision

Peter Meer

Abstract Statistics and computer vision have a different role for robustness. Statisticians are primarily concerned with the theoretical properties of estimators when models are only approximately true. In computer vision, performance takes precedence over theoretical considerations. This divergence is exemplified in statistics by the robust M-estimator in contrast to the RANdom SAmple Consensus (RANSAC) and the Multiple Input Structures with Robust Estimator (MISRE) in computer vision. All three have defined algorithms, but the M-estimator is based on theoretical results, while RANSAC and MISRE only emphasize recovering significant inlier structures. I offer suggestions for how the theory of the M-estimator can be further applied to MISRE, or MISRE can be applied to M-estimator.

Keywords Robustness · M-estimator · RANSAC · MISRE

1 Collaborations

My friendship with Dave Tyler goes back almost 30 years. In 1991, I took a position in the Department of Electrical and Computer Engineering at Rutgers University. My research focused on robust computer vision and, after settling in, I began to attend seminars at the Department of Statistics. I collaborated with Javier Cabrera on the application of bootstrapping in computer vision problems and got to know others in the department as well.

I do not remember exactly when I first met Dave Tyler, but we had fascinating discussions and quickly became friends. We were interested in how each other's fields treated robustness. As an expert on M-estimators, Dave taught me a lot in these conversations. Most importantly, perhaps, I learned that you do not have to be good at statistics, you just have to be friends with someone who is good at statistics.

P. Meer (✉)
Department of Electrical Engineering, Rutgers University, Piscataway, NJ, USA
e-mail: meer@soe.rutgers.edu

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_18

And Dave dove into the applications of robustness in computer vision, leading to some fruitful collaborations.

Dave and I worked together on two National Science Foundation grants ("Statistical Problems in 3D structure recovery," with Javier Cabrera, and "Modern statistical techniques for computer vision," between 1996 and 2003). He served on four Ph.D. committees for my students, three of whom also co-authored with him. That work brought in approaches from statistics that were little-known in computer vision, showing the value of cross-field interaction. In a chapter of "*Empirical Evaluation Techniques in Computer Vision*," Matei et al. (1998) analyzed the performance assessment when resampling 3D rigid motion of points. Starting from the Cook diagnostics in the volume of Rousseeuw and Leroy (1987, p. 227), the empirical influence functions $\widehat{EIF}$-s were computed by bootstrapping each point separately. This gives a stricter measure and the covariance matrices for rotation and translation, as well as more precise confidence regions of the input. In a conference about *Content-based Access of Image and Video Libraries* (CBAIVL-99), Fort Collins, CO, 1999 and in a journal paper in *Pattern Recognition Letters*, 2003, Comaniciu et al. (2003) used the Bhattacharyya distance between two arbitrary distributions, the query and the database, in a completely different way than it was already used in computer vision. The average recognition rates were better when applying it for VisTex textures and the Brodatz database. In the conference paper in *2001 IEEE Conference on Computer Vision and Pattern Recognition*, Chen et al. (2001) used the M-estimator with S-estimator for the auxiliary scale in very simple experiments, only three synthetic 2D lines. The limitations in the images were not the mirrors from the theory of statistical robust estimators.

Dave's natural curiosity, flexibility, and interest in computer vision led him to be an editor for the special issues on *"Robust computer vision"* in the journal *"Computer Vision and Image Understanding"* in 2000. He also participated in computer vision conferences in Seattle, WA, in 1994 and Wadern, Germany, in 1998. His invited lecture in Seattle, *M-estimates, S-estimates and CM-estimates: A Review*, showed how robustness is interpreted in statistics. At that point, M-estimation was still considered as a potentially successful approach in computer vision.

Meanwhile, I was also participating in statistics conferences in Raleigh, NC, 1995; Halifax, Canada, 1996; Anaheim, CA, 1997; and Leeds, England, 2000. My approach was always to emphasize the computer vision point of view. We collaborated to write *Smoothing the gap between statistics and image understanding*, a comment on *Edge preserving smoothers for image processing* in *J. Amer. Statist. Assoc.* in 1998, Vol. 93, pp. 526–541. We wrote that "in any interdisciplinary endeavor, successful cross-fertilization requires much more than just applying statistical tools to another problem domain." And we approached the question of image smoothing from two viewpoints: that of a user of the technique –the computer vision approach– and the methodological contribution—the statistics approach. That notion is at the heart of this essay: how is robustness viewed through the lenses of these different fields?

2 Statistics vs. Computer Vision

Dave noted the idea that computer vision had an “independent development of robust statistics” Tyler (2013, p. 84). Robustness in statistics means that a given distribution, generally Gaussian, is corrupted with a few outliers which have to be deleted before the final estimation. The outliers may have the same distribution with other parameters, another distribution, or no distribution at all. The theoretical properties for robustness are analyzed by statisticians; for example, the *breakdown point* is the largest part of the data beyond which a robust estimator becomes useless; *efficiency* measures of quality of the robust estimator with $O(n^2)$ being better than $O(n^3)$; the *influence function* is an infinitesimal perturbation which in order to remain robust and result in a small, smooth, and rescinding output. All these theoretical considerations are taken into account in robust statistical estimation. For example, the largest breakdown point in statistics is 50%, for least median of squares, Rousseeuw (1984), which has low efficiency $O(n^3)$. But in computer vision, the emphasis is on recovering the inlier structure without being particularly concerned about theory.

Dave discussed two computer vision robust estimators in Tyler (2013, pp. 86–92). The Hough transform appeared in Hough (1959) and the RANdom SAmple Consensus (RANSAC) in Fischler and Bolles (1981). The Hough transform is not really robust, since it mainly works for lines in 2D and 3D, the space has to be quantized before applying the transform and the background noise in the data is very important, among other problems. RANSAC is discussed in greater detail below.

Stigler (1973) had a wonderful paper about the origin of robustness in statistics. It was only in 1953 that George E.P. Box gave a formal statistical definition of robustness, but statisticians had used the concept for at least 250 years. In 1763, James Short, an English astronomer, had estimated the sun’s parallax based on observations of the transit of Venus. For the correct results he averaged three means: the sample mean, the mean of all observations with residuals less than one second, and the mean of those with residuals less than half a second. In 1818, Pierre-Simon Laplace proved that errors which are too large relative to the others should be rejected before estimation. In 1886, Simon Newcomb introduced a more robust estimator that gave “less weight to the more discordant observations.” The Huber (1964) M-estimator is similar to this pioneering work. And in 1888, Francis Edgeworth showed that “the median may possess an advantage over the sample mean.” These historical examples would be considered “robust” today.

John W. Tukey wrote in 1960, as Stigler (2010, p. 278) recalled, that the estimation of the scale of the normal distribution is less efficient if a few values at a distance of three standard deviations are contaminated. Huber (1964) translated this observation into the first robust statistical paper where contributions of the more distant points were systematically reduced. As Morgenthaler (2007, p. 277) explained, statisticians know the pitfalls of reliance on model assumptions over experimental facts. “All statisticians today are aware of the dangers of a data analysis that owes more to model assumptions than experimental facts." They know that the experimental procedure is at least as important as the chosen model.

The M-estimator has an objective function $\rho(u)$ which is nonnegative $\rho(u) \geq 0$ with $\rho(0) = 0$, symmetric $\rho(-u) = \rho(u)$ and nondecreasing with increasing of $|u|$. M-estimator starts with the whole dataset and uses iterative reweighted least squares at each step to eliminate outliers above a given threshold. Convergence is achieved when the deviation becomes under a given very small threshold between the steps.

In the early 1990s, researchers in computer vision still borrowed robust estimators from statistics: the M-estimator and the least median of squares (LMedS) of Rousseeuw (1984). The use of LMedS was diminishing because the procedure was not powerful enough to always detect more than one structure in a real image. Since the correct recovery of the inlier structure is the main focus in computer vision, the underlying theory behind the statistical estimators was never taken into account.

Much of Dave's work is about multivariate location, scatter, regression, and symmetric clutter distribution, mostly in relation to the different types of M-estimation. His work exemplifies the robustness literature in statistics. As mentioned above, Dave pointed out that computer vision has a different view of robustness, with which I am in complete agreement. This difference is driven by the different goals of the two fields. Statisticians are concerned with the theoretical properties of the estimators. As Dave put it in Tyler (2013, p.4), "are statistical methods which are good under the model reasonably good if the model is only approximately true?" But computer vision is concerned with performance and flexibility, and less so with the underlying theory.

To emphasize this point, I will describe both the commonly used RANdom SAmple Consensus (RANSAC) estimator which was conceived more than 40 years ago, and the new Multiple Input Structures with Robust Estimator (MISRE). Both of them, like the M-estimation, return parametric structures, but they approach the problem in different ways. I conclude by proposing a direction towards an M-estimator/MISRE combination which may lead to fruitful research on robustness in computer vision or statistics.

3 RANSAC

The RANdom SAmple Consensus (RANSAC) estimator, Fischler and Bolles (1981), is similar to very old methods that even predate least squares estimation. Around 1749, the German cartographer and astronomer Tobias Mayer (1723–1762) derived 27 equations to study the orbit of the moon from observations of a crater on the moon. To find the three unknowns, he used the method of averages, summing up groups of nine equations into a new equation. This approach increased the accuracy of his observations, Hald (2007, p. 44) and Stigler (1986, pp. 16–25).

The Croatian polymath Roger Joseph Boscovich (1711–1787) published a method to measure the ellipticity of the earth's oblate shape in 1755. With measurements from five locations, he obtained solutions for two unknowns in all ten pairs. The average of the solutions was incorrect and he removed two “so different from the others” pairs of points. The average of the eight remaining solutions was satisfactory, Hald (2007, pp. 45–46) and Stigler (1986, pp. 39–50).

Both Mayer and Boscovich reduced the scope of the estimation to the number of unknown. But Mayer used sums, while Boscovich computed the minimum number of points needed to solve the problem. Boscovich also eliminated two pairs. These were the forerunners of elemental subsets and the removal of data above a threshold.

The computation of RANSAC is very different from the computation of the statistical M-estimator. The objective function at the input, which has to be recovered robustly, can be a polynomial line, a nonlinear function like an ellipse, or a 3 × 3 matrix for 2D homography, etc. A three-dimensional example will illustrate this concept in the second part of the paper. If the objective function is nonlinear, RANSAC first transforms it into the linearized function, associating each term with a separate variable.

The 2D ellipse, the computer vision vector $\mathbf{y} = [x \;\; y]^\top$ becomes

$$f(\mathbf{y}) = (\mathbf{y} - \mathbf{y}_c)^\top \mathbf{Q}(\mathbf{y} - \mathbf{y}_c) - 1 \quad \longrightarrow \quad \sum_{i=1}^{5} x_i\theta_i - \alpha \tag{1}$$

$$\mathbf{x} = [x \;\; y \;\; x^2 \;\; xy \;\; y^2]^\top \quad \text{and} \quad 4\theta_3\theta_5 - \theta_4^2 > 0.$$

The five-dimensional $\boldsymbol{\theta}$ multiplies the elements of $\mathbf{x}$, and $\alpha = (\mathbf{y}_c^\top \mathbf{Q}\mathbf{y}_c - 1)$ is the scalar term of the linearized function.

The 2D homography, connecting the projective coordinates of two planes in the two images, has the input $\mathbf{y} = [x \;\; y \;\; x' \;\; y']^\top$ and the 3 × 3 objective function $\mathbf{H} = [\mathbf{h}_1 \;\; \mathbf{h}_2 \;\; \mathbf{h}_3]^\top$ becomes

$$\mathbf{f}(\mathbf{y}) = \begin{bmatrix} x_h' \\ y_h' \\ w_h' \end{bmatrix} - \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \mathbf{h}_3^T \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad x' = \frac{[x \;\; y \;\; 1]\mathbf{h}_1}{[x \;\; y \;\; 1]\mathbf{h}_3} \quad y' = \frac{[x \;\; y \;\; 1]\mathbf{h}_2}{[x \;\; y \;\; 1]\mathbf{h}_3} \tag{2}$$

$$\mathbf{x}^{[1]} = [-x \quad -y \quad -1 \quad 0 \quad 0 \quad 0 \quad x'x \quad x'y \quad x']^\top$$
$$\mathbf{x}^{[2]} = [0 \quad 0 \quad 0 \quad -x \quad -y \quad -1 \quad y'x \quad y'y \quad y']^\top,$$

where $[x_h' \; y_h' \; w_h']^\top$ are the projective coordinates and $[x' = \frac{x_h'}{w_h'} \;\; y' = \frac{y_h'}{w_h'}]^\top$ are the measured coordinates. If $\mathbf{f}(\mathbf{y})$ is equal to zero and x' and y' are divisions, the two nine-dimensional vectors can be seen.

Robust estimation in computer vision has the *elemental subset* as the building block. An elemental subset is a randomly chosen minimum number of input points required to estimate the linearized objective function. An 2D ellipse needs five points. A 2D homography needs only four points because each point has two equations. The returned parameters are correct only for an inlier structure. The total number of elemental subsets in real-world problems is very large. Therefore, only a reduced number M elemental subsets are taken, with M given before estimation. M can range from a few hundred to a few thousand, depending on the specific problem; using a larger M will not increase the accuracy of the estimation.

Before estimation, the *scale of the inliers*, σ, is also given. Points inside $\pm\sigma$ from the scalar term are retained for an elemental subset. The σ is unique, therefore, RANSAC will not work if multiple inlier structures with very different noise processes are present.

There is no deep theory behind RANSAC. The estimator is considered to be correct if the significant inlier structures are outputted correctly with very few outliers classified as inliers or vice versa. Because the amount of noise corrupting the inliers is unknown, there is no theory for predicting a minimum number for M. Each objective function class, like ellipse or homography, can have vastly different M even if they have the same number of inliers and outliers.

Assume n data points. RANSAC is computed by repeating the following procedure M times:

- Choose an elemental subset by random sampling without replacement.
- Define a linear model candidate by the minimum number of points.
- Assume the candidate is valid for all n points. Compute the distances between each point and the model.
- Distances less than $\pm\sigma$ from the α give the inlier consensus set.

The largest consensus set after M trials, i.e., the smallest set of outliers, is a returned elemental subset. Apply total least squares (TLS) to the selected inlier points and obtain the RANSAC estimate. If the input was nonlinear, project the estimate back to the input.

Figure 1 shows a noisy 2D line estimation with RANSAC in the presence of outliers (Fig. 1a). The least squares fit is completely erroneous (Fig. 1b). Before the estimation, the user selects M, the number of elemental subsets, and σ, the scale of the inliers. Pick a minimum of two points for a model; measure the distances for each point from the model; retain only the points inside $\pm\sigma$ from the model (Fig. 1c–e). Only a few points are retained. After M repetitions, a few models will be around the sought inlier structure. The one with the minimum number of outliers is the RANSAC estimate (Fig. 1f).

RANSAC can fail for several reasons: if the σ is not in the range of permissible values; if in a sequence of images the scale is significantly changed; if there are asymmetric outliers in the image; if an elemental subset cannot be defined uniquely from the minimum number of points; or if the data contains too many outliers, with the limiting value set by the input.

RANSAC uses randomly chosen minimum elemental subsets to build the largest consensus set, given a number of trials. In the last 25 years, RANSAC became the primary approach for robust computer vision. To eliminate some problems with RANSAC, many types of similar estimators were proposed which are more complicated. The theories behind those newer approaches were generally taken from statistics. But they emphasized the issues that were convenient for the authors'

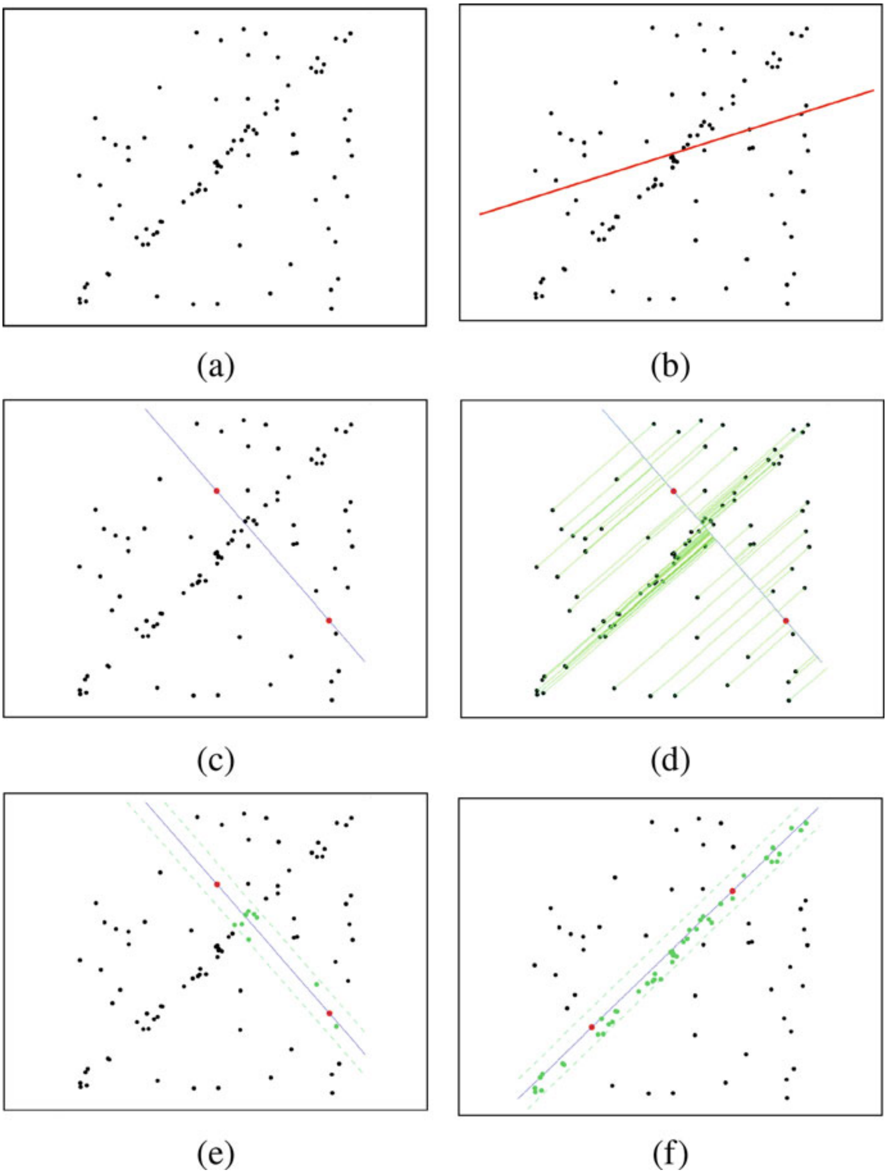


Fig. 1 RANSAC estimate for a noisy 2D line. (**a**) Many outliers are present in the input data. (**b**) Least squares fit is completely erroneous. (**c**) Define a model by taking two points (colored red). (**d**) Distances for all input points are measured from the model. (**e**) Retain only those points which have maximum $\pm\sigma$ distance from the model. (**f**) Repeat the procedure M times. A few models will be close to the sought inlier structure. The one with the minimum number of outliers is the solution

purposes and rarely focused on the limitations; if a method recovered significant inlier structures, it was considered suitable. A new RANSAC-type estimator was compared to similar approaches based on a comparison of a few images and objective functions, and unsurprisingly, authors almost invariably found that their proposed approach was superior.

RANSAC-type estimators used at least one of the following: guided sampling, distributions for the outliers and/or for the inliers, optimized model verification, detection of the degenerate configurations, and global regularization functionals. Jin et al. (2021) write "...it remains unclear how well they perform in real-world settings, compared to a well-tuned RANSAC." While these new RANSAC-type estimators were cited frequently, they were seldom used in real-world situations.

4 MISRE

In RANSAC, the input parameters have to be tuned correctly for an estimator to work. But in the M-estimator, reweighted least squares corrects the estimate in a few steps. Any RANSAC-type estimator aims to achieve consensus maximization for an inlier structure, which is only as good as the presumed separation between the inliers and outliers. Yang et al. (2021) recently published a new algorithm, the Multiple Input Structures with Robust Estimator (MISRE), which is more universal than RANSAC-type estimators. The inlier structures or outlier parts are treated similarly and therefore each of them is an independent iteration. MISRE robustly recovers a mathematical function for the significant inlier structures.

The universality of MISRE is divergent from RANSAC, which often needs different input parameters. MISRE uses the same two constants for all 2D and 3D estimation. But universality of MISRE cannot take into account pre-processing or post-processing of the images, which need specific thresholds. MISRE begins like RANSAC. The original objective function is linearized; elemental subsets are taken; and the user specifies the number of elemental subset trials. However, beside M no input parameter is given.

A vector objective function is recovered, like the 2D homography. There are a total of n points at the beginning and the input noise for the inliers can have arbitrary scales.

There are a total of ζ different Jacobian matrices, $\mathbf{C}_i^{[c]}$ $c = 1, \ldots \zeta$, which project the nonlinear input to the linear function and have to be processed with the same elemental subset $\boldsymbol{\theta}, \alpha$. The Mahalanobis distances in the null-space start from the scalar term α and are computed without the unknown scale.

$$d_i^{[c]} = \frac{|\mathbf{x}_i^{[c]\top}\boldsymbol{\theta} - \alpha|}{\sqrt{\boldsymbol{\theta}^\top \mathbf{C}_i^{[c]}\boldsymbol{\theta}}} \geq 0 \qquad c = 1, \ldots \zeta \quad i = 1, \ldots n. \tag{3}$$

The ζ different Jacobian matrices give different $d_i^{[c]}$ distances for the different $\mathbf{x}_i^{[c]}$. The most conservative choice, the largest Mahalanobis distance, is chosen for each point.

These Mahalanobis distances are ordered ascendingly for each of the M trials. Take the elemental subset which have the *minimum sum of the Mahalanobis distances* at five percent of the total data. This is the first constant of MISRE and is the same for all iterations in this problem. Therefore is no intrinsic bias in the algorithm since each iteration starts from the same number of points.

The elemental subset is $\hat{\boldsymbol{\theta}}_w$ and $\hat{\alpha}_w$. Divide the ordered sequence into equal Mahalanobis distances, denoted Δd_5, where Δd_5 corresponds to the first five percentage of the Mahalanobis points starting from α. The k-th Δd_5 segment has n_k points.

Expand the sequence, increasing with another Δd_5 each time. The second MISRE constant is *equal to number two*: the expansion finishes when the average number of points in the already-processed segments is larger than twice the number of points in the next segment.

$$\frac{1}{k}\sum_{i=1}^{k} n_i > 2\, n_{k+1} \qquad k = 1, 2, 3 \ldots \tag{4}$$

Increasing with 1% at the start the total number of data points, $\eta = 6\%, 7\% \ldots$, the same process is repeated again and again. Each expansion is independent because the input data is independent from other starting data.

When the second constant is satisfied, the expansions stop. The region of interest is defined from five percent to the total points till where the second constant is already satisfied at the starting point, giving η_f with a distance extended k_{t_η} times. The *largest Mahalanobis distance* in the region of interest is the estimated standard deviation.

$$\hat{\sigma} = \max_{\eta=5\%,\ldots,\eta_f} k_{t_\eta} \Delta d_\eta . \tag{5}$$

For inlier structures, the $\hat{\sigma}$ is relative small, while for an outlier part the $\hat{\sigma}$ is large.

The $\hat{\sigma}$ was based on a single elemental subset. Take another $0.1M$ elemental subsets from the points between $\hat{\alpha}_w \pm \hat{\sigma}$. For each subset find the closest mode to α by mean shifts, Comaniciu and Meer (2002). All n points participate.

The total least squares (TLS) for $\hat{\alpha}^{tls}$ has n_{st} points and standard deviation $\hat{\sigma}^{tls}$, giving the density of the structure as the ratio

$$\frac{n_{st}}{\hat{\sigma}^{tls}}, \tag{6}$$

and n_{st} points are removed from the input. The processing of a next iteration with $n - n_{st}$ begins if there are more from 5% of the total input data.

If the unprocessed points are already less than five percent of the total input data, the structures are sorted in descending order based on the densities (6). Until this point, inlier structures and outlier parts are not distinguished. The user specifies the cutoff between the significant inlier structures and the first outlier part, based on where the $\hat{\sigma}^{tls}$ increase is substantial.

The stronger inlier structures are always recovered, even when a weaker inlier structure turns into outliers. Classifying based on the standard deviations and not the densities can introduce small units of inliers or outliers between other significant inlier structures. Fusing two structures needs specific thresholds.

In the paper Yang et al. (2021), the pseudocode of the algorithm is also given, along with numerous examples for 2D images and 3D sequences. Circular cylinder estimation in 3D starts from 22 images in 2D with a 2D image is at the top-left of Fig. 2. The 6500 point are identifier in 3D (top-right, Fig. 2). Using the paper of Beder and Förstner (2006) start with $\mathbf{P}$, the most general nine-dimensional solution

$$\mathbf{P} = \begin{bmatrix} \mathbf{D} & \mathbf{d} \\ \mathbf{d}^\top & d \end{bmatrix}, \tag{7}$$

where $\mathbf{P}$ the 4×4 symmetric matrix and linearize $\mathbf{P}$ with the 9×3 Jacobian matrix. A circular cylinder has five degrees of freedom, four for the axes and one for the radius. If $\mathbf{P}$ is a circular cylinder, the 3×3 matrix $\mathbf{D}$ has to have two identical singular

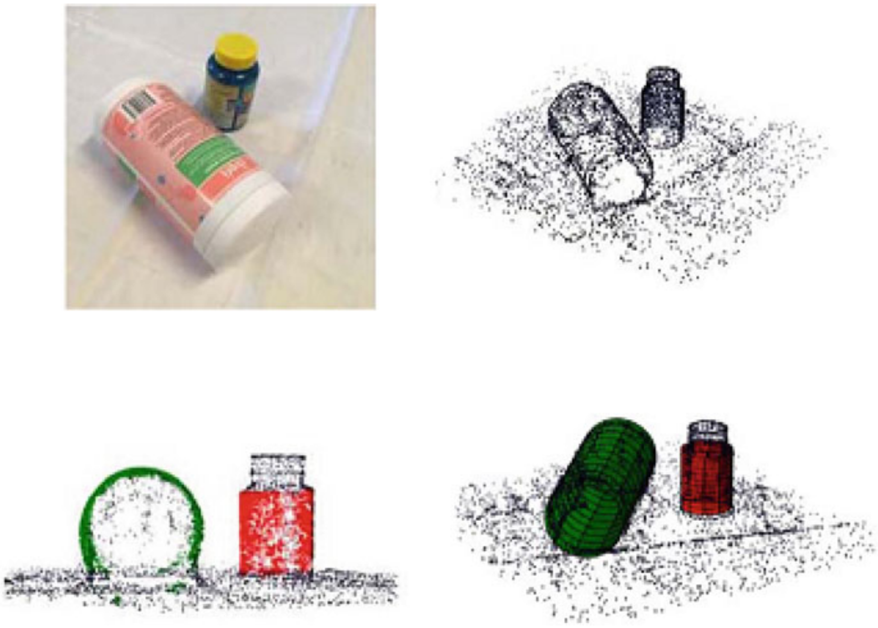


Fig. 2 3D recovery of circular cylinders. Top-left: one of the 22 images in 2D. Top-right: the 6500 points recovered in 3D. Bottom left and right: The 2262 inlier points viewed in two different angles. Images partially © 2021 IEEE. Reprinted, with permission, from Yang et al. (2021)

values, and the third one is equal to zero. The three-dimensional vector **d** has to be an eigenvector of the 3×3 matrix **D**. These four constraints have to be verified, up to a small threshold, for every nine-dimensional elemental subset chosen.

Take $M = 2000$ and process the points till below 325 points, which is equal 5%. The first two structures are inliers with a total of 2262 points (bottom, Fig. 2). The number of outliers is almost three times as many as the number of inliers. This can occur when a lot of outliers belong to another objective function, like a plane; this can be addressed in the algorithm.

5 Possible Cooperation

All three estimators, M-estimator, RANSAC, and MISRE, return mathematical objective functions based on well-defined but vastly different algorithms. The M-estimator starts from the total number of the input points and after a few iterations converge to a robust solution with fewer points which are mostly inliers. Before the estimation, the user has to give a threshold (scale value) for the inlier structure. MISRE picks up a sufficiently large number of elemental subsets M randomly, with each subset giving a minimal solution. The scale value for the inlier structure is obtained from the largest Mahalanobis distance in the region of interest. This comes from a single elemental subset.

In each iteration, the M-estimator computes an iterative reweighted least squares where only the weight of the input points are potentially changed. To achieve to the final estimate, MISRE has to do total least squares over all the already participating input datapoints.

Assume that both estimates have the same input data, the same inlier structure and the same noise. Then, the two estimates would have very similar M-estimator and MISRE input sequences and the returned objective functions at the output will be very close one to the other.

Can this parallel approach yield more insight into these different approaches to robustness? How do the breakdown point, the efficiency and the influence function change when applied to MISRE, especially for multiple inlier structures where M-estimator can fail? Are the new values sensible? If MISRE's scale value is used in the M-estimator, will the estimation change? If the M-estimator's scale is applied to MISRE, how large the change will be? Many questions can be examined, but the answers should come only through the experimentation.

Dave Tyler's influence goes far beyond his work on the M-estimator, even if this essay is limited to that contribution. As ideas of robustness in statistics and computer vision diverged, collaboration between the two fields diminished though –more importantly– Dave and I remain good friends. Perhaps a new generation can reignite the fruitful sharing of ideas between our fields.

Acknowledgments I thank Jonathan Meer for all the suggestions which made this essay possible.

References

Beder, C., & Förstner, W. (2006). Direct solutions for computing cylinders from minimal sets of 3D points. In *2006 European Conference on Computer Vision* (vol. 3952, pp. 135–146). Springer.

Chen, H., Meer, P., & Tyler, D. E. (2001). Robust regression for data with multiple structures. In *2001 IEEE Computer Vision and Pattern Recognition* (pp. 1069–1075).

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 603–619.

Comaniciu, D., Meer, P., & Tyler, D. E. (2003). Dissimilarity computation through low rank corrections. *Pattern Recognition Letters*, *24*, 227–236.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*, 381–395.

Hald, A. (2007). *A history of parametric statistical inference from Bernoulli to Fisher, 1713 to 1935*. Springer.

Hough, P. (1959). Machine analysis of bubble chamber pictures. In *International Conference on High Energy Accelerators and Instrumentation*. CERN.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*, 73–101.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K. M., & Trulls, E. (2021). Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, *129*, 517–547.

Matei, B., Meer, P., & Tyler, D. E. (1998). Performance assessment by resampling: Rigid motion estimators. In K. W. Bowyer & P. J. Phillips (Eds.), *Empirical Evaluation Techniques in Computer Vision* (pp. 72–95). IEEE CS Press.

Morgenthaler, S. (2007). A survey of robust statistics. Discussion. *Statistical Methods and Applications*, *15*, 271–293.

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

Rousseeuw, P. J., & Leroy, A. (1987). *Robust Regression and Outlier Detection*. John Viley & Sons.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, *68*, 872–879.

Stigler, S. M. (1986). *The history of statistics. The measurement of uncertainty before 1900.* Harvard University Press.

Stigler, S. M. (2010). The changing history of robustness. *The American Statistician*, *64*, 277–281.

Tyler, D. E. (2013). A short course on robust statistics. http://www.stat.rutgers.edu/home/dtyler/ShortCourse.pdf. On-line, Rutgers University.

Yang, X., Meer, P., & Meer, J. (2021). A new approach to robust estimation of parametric structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*, 3754–3769.

Part IV
Other Methods

Power Calculations and Critical Values for Two-Stage Nonparametric Testing Regimes

John Kolassa, Xinyan Chen, Yodit Seifu, and Dewei Zhong

Abstract Interim analysis techniques for clinical trials provide improved power with smaller average sample sizes. These techniques crucially require multivariate probability calculations for determining critical values. Most existing techniques rely on multivariate normal approximations to the joint null distribution of test statistics evaluated on potential interim and full data sets. More accurate critical values for nonparametric testing with an interim analysis are given, using a new multivariate Cornish–Fisher expansion. While earlier authors demonstrated that such an expansion is possible, it has never been implemented before this manuscript. Generally, the superior accuracy of power calculations via an Edgeworth series is demonstrated. Example calculations giving sample sizes from desired power are provided. Calculations are implemented in an R package.

Keywords Cornish–Fisher expansion · Multi-stage testing

1 Introduction

With more and more drugs being developed for rare diseases (Pharmaceutical Research and Manufacturers of America 2019), the rarity of the disease and the need for new therapy can potentially result in small sample sizes for pivotal trials (Chow and Huang 2020; Rom and McTague 2020). For trials with small sample sizes, if the efficacy response is not approximately normally distributed, designing

J. Kolassa (✉) · X. Chen
Rutgers, The State University of New Jersey, Piscataway, NJ, USA
e-mail: kolassa@stat.rutgers.edu; orchid12@scarletmail.rutgers.edu

D. Zhong
Anheuser-Busch Companies, LLC, New York, NY, USA

Y. Seifu
Bristol-Myers Squibb, Berkeley Heights, NJ, USA
e-mail: Yodit.Seifu@bms.com

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_19

a study that is based on mean differences under the normality assumption might result in inflating the type-I error. In such cases, it may be more accurate to design a trial based on testing for differences in location using a nonparametric procedure such as the Mann–Whitney test. One of the advantages of such rank-based testing is that these procedures are robust to aspects of the distribution of the underlying distributions, like asymmetry, because the statistics are distribution-free. One of the disadvantages of rank-based testing is that desirable test properties of parametric tests may not apply in this case; Amrhein (1995) demonstrates that a related test is biased. If the hypothesized efficacy is present, there is also a need to make these therapies available to patients as soon as possible, especially if the indication does not otherwise have a standard therapy. In this chapter, we propose a group-sequential trial design when comparing two treatments, using a Mann–Whitney test at each stage. One interim analysis is done to potentially stop early for efficacy. The proposed analysis methodology is highly accurate and computationally efficient when compared to available exact methods.

Our proposed methods are appropriate to interim analyses in which each subject has one assessment, and study enlargement considered is via the recruiting of new subjects that will presumably act independently and not via more assessment on existing subjects.

While trials for rare diseases tend to be quite small, small trials exist in other contexts, including phase I and II trials, and particularly those in which early stopping is considered important.

This manuscript addresses the design of a trial incorporating such an interim analysis. The present work might be extended to allow for early stopping for safety or futility and for the consideration of multiple interim analyses. It may also be extended to multi-stage designs to select among multiple treatments (Whitehead and Jaki 2009).

Wilding et al. (2011) discuss exact inference for two-stage trials, using the Wilcoxon statistic (Mann and Whitney 1947), and present a recursive algorithm for probability atoms for the resulting bivariate test statistic. Exact inference is slow, even for moderate sample sizes, making it difficult to routinely use for derivation of sample sizes and power. The present work follows up on Kolassa (1995), who provided power and sample size calculations for an application of the Wilcoxon statistic with no interim analysis. Rahardja et al. (2009) and Shieh et al. (2006) summarize work on power calculations for one-stage tests.

Section 2 of this chapter sets out assumptions and notation for later calculations. Section 3 reviews existing multi-dimensional exact calculations. Section 4 reviews Edgeworth probability function approximate calculations. Section 5 reviews univariate Cornish–Fisher quantile approximations. Section 6 presents a new approximation to multivariate quantiles. Section 7 applies this new approximation to quantile estimation for multivariate rank tests. Section 8 discusses continuity correction. Section 9 provides guidance on sample size. Section 10 provides an example calculation for the critical values. Section 11 presents accuracies for the probability calculations for some experimental configurations. Section 12 discusses errors in critical values. Section 13 discusses errors in calculation of powers. Section 14 presents conclusions.

2 Assumptions

This section lays out notation for the two-stage rank-based testing problem and specifies the assumptions under which the analysis is performed.

Consider a study with two treatments (active treatment versus control). Under the two-sample testing plan with an interim analysis, the investigator initially collects observations from each of the treatment and control groups. Denote the initial number of control observations as m_1, and denote the control observations as $X_1, \dots, X_{m_1}$. Similarly, denote the initial number of treatment observations as n_1, and denote the treatment observations as $Y_1, \dots, Y_{n_1}$.

Suppose that the control observations have distribution function F and that treatment observations have distribution function G. The null hypothesis is that there is no difference in distribution between treatment and control observations:

$$H : F(x) = G(x)\ \forall x,$$

and the alternative hypothesis is that the treatment observations are systematically greater than the control observations:

$$K : F(x) \geq G(x) \forall x,\ F(y) > G(y) \text{ for some } y.$$

A more specific alternative hypothesis is that the distribution function for treatment subjects is offset that of the control subjects, in the direction leading to larger values under the alternative:

$$K' : F(x) = G(x - \Delta)\ \forall x,$$

for some $\Delta > 0$. Note that both the general alternative K and the specific alternative K' are inherently one-sided.

The first step in assessing the presence or absence of shift in distribution uses the Wilcoxon statistic

$$U_1 = \sum_{i=1}^{m_1} \sum_{j=1}^{n_1} I(X_i < Y_j), \tag{1}$$

where I takes the value 1 if its argument is true and 0 otherwise. The investigator compares U_1 to the critical value $c_1(m_1, n_1)$ such that

$$P_0[U_1 \geq c_1] \leq \alpha_1. \tag{2}$$

If $U_1 \geq c_1$, the investigator declares treatment superior to controls and terminates the trial. Otherwise, the investigator collects and additional m_2 observations $X_{m_1+1}, \ldots, X_{m_1+m_2}$ from controls, and n_2 observations $Y_{n_1+1}, \ldots, Y_{n_1+n_2}$ from treated subjects, and calculates

$$U_2 = \sum_{i=1}^{m_1+m_2} \sum_{j=1}^{n_1+n_2} I(X_i < Y_j), \tag{3}$$

the Wilcoxon statistic for the combined data set. Assume that

$$X_1, \ldots, X_{m_1+m_2}, Y_1, \ldots, Y_{n_1+n_2} \tag{4}$$

are jointly independent with continuous distributions, that $X_1, \ldots, X_{m_1+m_2}$ identically distributed, and that $Y_1, \ldots, Y_{n_1+n_2}$ identically distributed. Then, U_2 is compared to the critical value c_2, for $c_2(m_1, n_1, m_2, n_2)$ such that

$$P_0[U_1 \geq c_1 \text{ or } U_2 \geq c_2] \leq \alpha_2. \tag{5}$$

If $U_2 \geq c_2$, the investigator declares treatment superior to controls.

Subscript 0 on probabilities in (2) and (5) indicates that probabilities are calculated under the null hypothesis. Hence, the critical values c_1 and c_2 are calculated under the null hypothesis that $X_1, \ldots, X_{m_1+m_2}, Y_1, \ldots, Y_{n_1+n_1}$ are all identically and independently distributed, with a continuous distribution.

The primary contribution of this manuscript is in describing the procedure for calculating c_2 of (5).

The next section discusses asymptotic approximations to the critical values that avoid exact calculations discussed in §1.

3 Existing Probability Calculations

Fix and Hodges (1955) give recursion relations for probabilities for the univariate statistic, and an expression for the first four null univariate moments, for use in an Edgeworth series. As with the univariate Mann–Whitney statistic, the bivariate statistic has null probabilities that can be calculated using a recursion relation; that recursion is similar to that presented by Wilding et al. (2011). This recursion is too slow for routine use in moderate samples, but available for assessing the accuracy of the probability approximations presented in this manuscript, if one is willing to devote enough computing time. This recursion, described in the appendix and specified by (22) and (23), was used to calculate true probability atoms with no Monte Carlo error. Our fastest implementation of this recursion (in Fortran), using integer arithmetic and significant caching of intermediate values, took 10 h to generate probabilities for the largest configuration studied, with $m_1 = n_1 =$

$m_2 = n_2 = 9$. Hence, while this recursion is useful for studying the behavior of the technique, it is too slow for routine clinical trial design use; that is, it is too slow for power calculation and far too slow to invert for sample size calculation.

Approximate calculations based on Edgeworth series techniques will be substituted for calculating the probabilities involved in (2) and (5).

4 Approximating Corner Probabilities

Solution to (2) and (5) for c_1 and c_2 will begin with an approximation to the two probabilities. This will be done using an Edgeworth series, using the moments to order 4. This approximation is well known McCullagh (1987), although seldom implemented in practice for dimensions higher than one. The approximation is most easily described in terms of variables rescaled to give expectation 0 and marginal variances 1:

$$\mathbf{V} = (V_1, V_2) \text{ for } V_1 = (U_1 - \mu_1)/\sigma_1,\ V_2 = (U_2 - \mu_2)/\sigma_2 \tag{6}$$

for

$$\left.\begin{aligned} &\mu_1 = m_1 n_1/2,\ \sigma_1 = \sqrt{m_1 n_1 (m_1 + n_1 + 1)/12} \\ &\mu_2 = (m_1 + m_1)(n_2 + m_2)/2, \\ &\sigma_2 = \sqrt{(m_1 + m_2)(n_1 + n_2)(m_1 + m_2 + n_1 + n_2 + 1)/12} \end{aligned}\right\}. \tag{7}$$

Let $\kappa^{i,\cdots,k}$ be the cumulant of components $i, \ldots, k$ of (V_1, V_2). Sundrum (1954) gives expressions for the first four univariate moments, including under alternative hypotheses, using combinatoric arguments. Zhong and Kolassa (2017) give bivariate moments, under both the null and alternative distributions. Cumulants are calculated from moments via the standard multivariate relations. Assume that a cumulant of order r is of size $O(N^{2-r})$, as is the case for the Wilcoxon statistics. Here,

$$N = m_1 + n_1 + m_2 + n_2, \tag{8}$$

with m_1, n_1, m_2, and n_2 roughly proportional to N.

McCullagh (1987) approximates the density for $\mathbf{V} = (V_1, V_2)$ by

$$\begin{aligned} e_4(\mathbf{v}) = \phi(\mathbf{v}, \rho)\Bigg\{1 &+ \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{\kappa^{i,j,k} h_{ijk}(\mathbf{v})}{3!} + \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} \frac{\kappa^{i,j,k,l} h_{ijkl}(\mathbf{v})}{4!} \\ &+ \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{m=1}^{2}\sum_{n=1}^{2} \frac{\kappa^{i,j,k}\kappa^{l,m,n} h_{ijklmn}(\mathbf{v})[10]}{6!}\Bigg\} + o\left(\frac{1}{N}\right) \end{aligned} \tag{9}$$

with $\rho = \kappa^{12}$, $\phi(\cdot, \rho)$ the bivariate normal density with expectations 0, marginal variances 1, and covariance ρ. The functions h_{ijk} are defined as

$$h_{ij}(\mathbf{v}) = \frac{d}{dv_i}\frac{d}{dv_j}\phi(\mathbf{v}, \rho)/\phi(\mathbf{v}, \rho)$$
$$h_{ijk}(\mathbf{v}) = -\frac{d}{dv_i}\frac{d}{dv_j}\frac{d}{dv_k}\phi(\mathbf{v}, \rho)/\phi(\mathbf{v}, \rho)$$
$$h_{ijkl}(\mathbf{v}) = \frac{d}{dv_i}\frac{d}{dv_j}\frac{d}{dv_k}\frac{d}{dv_l}\phi(\mathbf{v}, \rho)/\phi(\mathbf{v}, \rho);$$

these polynomials depend on ρ, but this dependence is suppressed.

This density approximation may be integrated term-wise to get an approximation to tail probabilities. Define

$$\bar{E}_4(\mathbf{v}) = \int_{v_1}^{\infty}\int_{v_2}^{\infty} e_4(\mathbf{w})\, dw_1 dw_2. \tag{10}$$

The leading term of (9) integrates to $\Phi(\mathbf{v}, \rho)$, the bivariate normal cumulative distribution function with expectations 0, variances 1, and correlation ρ. Terms involving h functions with indices including both 1 and 2 integrate to terms of the same form, with one index equal to 1 and one index equal to 2 both dropped.

Terms involving the functions h with all indices 1, or with all indices 2, are more complicated. Take, for instance, a term like $\kappa^{1,1,1}h_{111}(\mathbf{v})\phi(\mathbf{v}, \rho)$. Note that

$$\phi(\mathbf{v}, \rho) = \frac{\exp(-v_1^2/2)}{(2\pi)^{1/2}} \times \frac{\exp(-(v_2 - \rho v_1)^2/(2(1-\rho^2)))}{(2\pi)^{1/2}\sqrt{1-\rho^2}}.$$

Integration with respect to the first argument gives $\kappa^{1,1,1}h_{11}(\mathbf{v})\phi(\mathbf{v}, \rho)$. Subsequent integration with respect to the second argument gives

$$\kappa^{1,1,1}h_{11}(\mathbf{v})\phi(v_1)\bar{\Phi}((v_2 - \rho v_1)/\sqrt{1-\rho^2}),$$

where ϕ and Φ with a scalar first argument and without a correlation second argument refer to the one-dimensional standard normal density and distribution function, respectively. Hence, these terms without indices for each dimension as superscripts to h are polynomial multiples of a normal density and a normal conditional distribution function, rather than of the bivariate normal density. Careful accounting for such terms can give a distribution function counterpart to (9). Kolassa (2003) gives more details about these calculations. Furthermore, Wold (1934) provides adjustments, called Sheppard's corrections, to multivariate cumulants, and Kolassa and McCullagh (1990) argued that these were the appropriate cumulant corrections for an improved Edgeworth approximation in the univariate case. In this case, Shepard's corrections are of smaller order than $O(1/N)$.

Kolassa and McCullagh (1990) argue that Edgeworth approximations to lattice variables (that is, variables that take values on an affine transformation of the integers) require continuity correction. Kolassa (1989) extends this argument to random vectors. While their work was specific to sums of independent and identically distributed variables, the argument holds in the Mann–Whitney–Wilcoxon case as well.

Because of this complication, the bivariate Edgeworth approximation to tail probabilities is cumbersome and was constructed using the symbolic computation software Mathematica Wolfram Research, Inc. (2018). The series was constructed by expanding the exponentiated cumulant generating function and applying Fourier inversion term-wise. The symbolic computation software writes mathematical expressions into code for a computer language (in this case, Fortran) accessible from our R package *TwoStage*. This package and the package on which it depends, bivcornish, are hosted on Github; if one runs

```
library(devtools)
install_github("kolassa-dev/bivcornish")
install_github("kolassa-dev/TwoStage")
```

they will be installed.

Obtaining a one-dimensional distribution function approximation by term-wise integration avoids these complications and is given by

$$\bar{\Phi}(v) + \phi(v)\left\{\frac{\kappa^{1,1,1}h_{11}(v)}{3!} + \frac{\kappa^{1,1,1,1}h_{111}(v)}{4!} + \frac{\kappa^{1,1,1}\kappa^{1,1,1}h_{11111}(v)}{72}\right\} + o(1/N). \tag{11}$$

5 Existing Approximate Critical Values

Equations (2) and (5) define the critical values. In cases with the distribution of $\mathbf{V}$ continuous, (2) and (5) may be interpreted as holding with equality, and Takeuchi (1978) and Takemura and Takeuchi (1988) prove that expansions of the form

$$c_1 = \mu_1 + \sigma_1(x_{10} + x_{11}/\sqrt{N} + x_{12}/N + o(1/N)), \tag{12}$$

$$c_2 = \mu_2 + \sigma_2(x_{20} + x_{21}/\sqrt{N} + x_{22}/N + o(1/N)) \tag{13}$$

hold, although we know of no published work that determines x_{21} or x_{22}. The univariate expansion of Cornish and Fisher (1938), obtained by approximately inverting (11), provides values for x_{10}, x_{11}, and x_{12}:

$$\left.\begin{aligned} x_{10} &= z_{\alpha_1}, \qquad x_{11} = \frac{\kappa^{1,1,1}(z_{\alpha_1}^2 - 1)}{6\sqrt{N}}, \\ x_{12} &= \frac{3\kappa^{1,1,1,1}(z_{\alpha_1}^3 - 3z_{\alpha_1}) - 2(\kappa^{1,1,1})^2(2z_{\alpha_1}^3 - 5z_{\alpha_1})}{72N} \end{aligned}\right\}. \tag{14}$$

6 A New Bivariate Quantile Approximation

The new bivariate Cornish–Fisher expansion is determined by equating $\bar{E}_4(x_{10} + x_{11}N^{-1/2} + x_{12}N^{-1}, x_{20} + x_{21}N^{-1/2} + x_{22}N^{-1}) = 1 - \alpha_2$, for $\bar{E}_4$ defined in (10), and expanding the result in powers of $N^{-1/2}$ and equating terms. The leading terms determine x_{20} to be z_{α_1,α_2}, for z_{α_1,α_2} solving

$$\Phi((z_{\alpha_1}, z_{\alpha_1,\alpha_2}), \rho) = 1 - \alpha_2. \tag{15}$$

One equates factors multiplying $N^{-1/2}$ and N^{-1} to zero and solves for x_{21} and x_{22}. Then,

$$\begin{aligned} x_{21} = \frac{1}{6}(1-\rho^2)^{-2}\Big[&\kappa_{222}\Big(z^2_{\alpha_1,\alpha_2} - 1\Big)\Big(\rho^2 - 1\Big)^2 + \sqrt{1-\rho^2}m^{-1} \\ &\Big((\kappa_{111}z_{\alpha_1} + \kappa_{222}z_{\alpha_1,\alpha_2})\rho^3 + (\kappa_{222}z_{\alpha_1} + \kappa_{111}z_{\alpha_1,\alpha_2})\rho^2 - (2\kappa_{111}z_{\alpha_1} \\ &+3\kappa_{122}z_{\alpha_1} + 3\kappa_{112}z_{\alpha_1,\alpha_2} + 2\kappa_{222}z_{\alpha_1,\alpha_2})\rho \\ &+3(\kappa_{112}z_{\alpha_1} + \kappa_{122}z_{\alpha_1,\alpha_2})\Big)\Big], \end{aligned} \tag{16}$$

where $m = \bar{\Phi}((z_{\alpha_1} - \rho z_{\alpha_1,\alpha_2})/\sqrt{1-\rho^2})/\phi((z_{\alpha_1} - \rho z_{\alpha_1,\alpha_2})/\sqrt{1-\rho^2})$. Symbolic computation software was used to invert the Edgeworth series to give the expression for x_{22}. The expression for x_{22} is extensive enough that we do not include it. Mathematica Wolfram Research, Inc. (2018) was used to write x_{22} into Fortran code called by the R package `TwoStage`. The appendix presents a continuous, asymmetric example of this new Cornish–Fisher expansion.

7 Application to Rank Tests

When planning a multi-stage two-sample test, one first chooses an overall test level, α_2, and a smaller level for the first stage of the test, α_1. One then approximates the critical values c_1 and c_2, defined in (2) and (5), via the Cornish–Fisher expansion (12) and (13). In this case, moments (and hence the correlation ρ) are given by Zhong and Kolassa (2017). The null value for ρ is

$$\frac{\sqrt{m_1 n_1(m_1 + n_1 + m_2 + n_2 + 1)}}{\sqrt{(m_1 + m_2)(n_1 + n_2)(m_1 + n_1 + 1)}}$$

(Spurrier and Hewett 1976). This manuscript gives only the $O(1/\sqrt{N})$ term for (13) in (16), while computer code to calculate the $O(1/N)$ term is incorporated in the R package `TwoStage`. Again, N is the total sample size in both stages combined, as in (8).

Power is calculated using the Edgeworth approximation (10); our calculations below use the formulae coded into `TwoStage`.

8 Continuity Correction for the Two-Stage Wilcoxon Statistic

The continuity correction to c_1 involves moving critical values to the nearest integer plus half. However, moving this critical value changes approximation to the level of the test in the first look at the results, reflecting the fact that in this case with a discrete test statistic, only a discrete set of test levels is possible. We propose adjusting this first test level α_1 to reflect the actual tail probability; we do this approximately via the univariate Edgeworth approximation (11).

9 Sample Size Calculation

Sample levels are generally determined by first specifying a proportion of the observations to be treatments and controls in the first and second stages of the experiment. Choose the proportions

$$\lambda_{11} = m_1/N,\ \lambda_{21} = n_1/N,\ \lambda_{12} = m_2/N,\ \lambda_{22} = n_2/N. \tag{17}$$

This framework allows for differing ratios of treated individuals to control individuals in the two stages. For example, if one wants twice as many treated as control individuals in the first stage, and the same number of treated and control in the second stage, with both stages involving the same total number of subjects, then $\lambda_{11} = 1/6$, $\lambda_{12} = 1/3$, $\lambda_{21} = 1/4$, and $\lambda_{22} = 1/4$. Under the null hypothesis, $P\big[Y_i \geq X_j\big] = 1/2$. The Mann–Whitney–Wilcoxon test has power under alternatives for which this probability is different from 1/2; specify the alternative as

$$P\big[Y_i \geq X_j\big] = \omega + 1/2 \tag{18}$$

with $\omega > 0$ measuring the difference from the null hypothesis. This quantification of the degree by which the alternative hypothesis differs from the null is entirely general, in that every pair of continuous distributions yields a value of ω in [0, 1], but a more intuitive parameterization is given by Kolassa and Seifu (2013).

By analogy with the single-stage approach, one might take the joint statistic vector (U_1, U_2), subtract the null expectations, and divide by the standard deviations, to obtain components that are marginally standard normal, and note that, under the alternative hypothesis, components of this vector have an expectation that is a multiple of ω and $\sqrt{N}$. One then might equate the bivariate probability of rejecting the null hypothesis to the desired power and solve for N.

Specifically, consider $\mathbf{V}$ of (6), with the approximate null moments

$$\left.\begin{aligned} &\mu_1 = \lambda_{11}\lambda_{21}N^2/2, \sigma_1 = N^{3/2}\sqrt{\lambda_{11}\lambda_{21}\lambda_{\cdot 1}/12} \\ &\mu_2 = \frac{1}{2}\lambda_{1\cdot}\lambda_{2\cdot}N^2, \sigma_2 = N^{3/2}\sqrt{\lambda_{1\cdot}\lambda_{2\cdot}/12} \end{aligned}\right\}. \tag{19}$$

Then, the null distribution of $\mathbf{V}$ is approximately bivariate normal, with unit variances, and correlation

$$\frac{\sqrt{\lambda_{11}\lambda_{21}(1+1/N)}}{\sqrt{\lambda_{1\cdot}\lambda_{2\cdot}(\lambda_{\cdot 1}+1/N)}} = \tilde{\rho} + O(1/N),$$

for $\tilde{\rho} = \frac{\sqrt{\lambda_{11}\lambda_{21}}}{\sqrt{\lambda_{1\cdot}\lambda_{2\cdot}\lambda_{\cdot 1}}}$. As above, reject the null hypothesis if $V_1 \geq z_{\alpha_1}$ or if $V_2 \geq z_{\alpha_1,\alpha_2}$.

The mean of $\mathbf{V}$ under the alternative hypothesis given by (18) is $\mu_1 = \sqrt{12N\lambda_{11}\lambda_{21}/\lambda_{\cdot 1}}$ and $\mu_2 = \sqrt{12N\lambda_{1\cdot}\lambda_{2\cdot 1}}$. As is the case in classical one-stage power calculation, variation in the mean of the test statistic as one moves from the null to the alternative hypothesis has more bearing on sample size than does movement in the covariance matrix and so initial power calculation is made leaving standard deviations and the correlation at their null values.

Power is given by $1 - \Phi((z_{\alpha_1}, z_{\alpha_1,\alpha_2}) - \omega\sqrt{12N}(\sqrt{\lambda_{11}\lambda_{21}}/\sqrt{\lambda_{\cdot 1}}, \sqrt{\lambda_{1\cdot}\lambda_{2\cdot}}))$. An approximation to the sample size yielding power $1-\beta$ is the solution N to

$$\Phi((z_{\alpha_1}, z_{\alpha_1,\alpha_2}) - \omega\sqrt{12N}(\sqrt{\lambda_{11}\lambda_{21}}/\sqrt{\lambda_{\cdot 1}}, \sqrt{\lambda_{1\cdot}\lambda_{2\cdot}}) = \beta. \tag{20}$$

In the simple one-stage case, one would apply the normal quantile function to both sides of the analog to (20), to obtain a power formula involving the normal quantile associated with the desired power; in this multi-stage case, this approach is not possible, since no univariate quantile function exists in this case.

However, one might bound the desired N below by the value associated with the Wilcoxon test with no intermediate assessment; Rahardja et al. (2009) give this as

$$N = (z_{\alpha_2} + z_\beta)^2/(12\lambda_{1\cdot}\lambda_{2\cdot}\omega^2). \tag{21}$$

This might be taken as the starting value for a search using (12) and (13) to calculate critical values and (10) to determine a refined approximation to power.

One routinely chooses the overall significance level α_2 to be the same as one would use in a study without an interim analysis. The quantity α_1 ought to be chosen in light of the costs associated with extension of the study to the second stage and in light of the expected effect size. If one expects a large effect, then α_1 might be chosen relatively large, in order to avoid the necessity of the second stage; if a small effect is expected, α_1 might be taken smaller.

10 An Example Calculation

The tools developed in this manuscript are designed to help with planning a multi-stage clinical trial. The experience of subjects in this study is to be summarized by a single continuous measurement, in order to test the null hypothesis that the distribution of this measurement in the control arm is the same as the distribution in the treatment arm vs. the alternative hypothesis that the distribution of treatment measurements is stochastically larger than the distribution for control subjects. In order to avoid assumptions of normality of the responses, the observations are compared using the Mann–Whitney–Wilcoxon test. Testing is performed in two stages, according to the scheme outlined in the introduction. Below we demonstrate the application of this sequential trial design through a recent pivotal clinical trial where the sample size was small. Small trials are conducted in order to speed clinical discovery, to avoid unnecessary suboptimal treatment regimes for patients, and to control trial costs. The group-sequential trial design is of importance for the same reasons.

Henricson et al. (2012) present data including results for a six-minute walk test for patients with Duchenne muscular dystrophy (DMD) and controls. Participants are male, between 4 and 12 years. Patients with DMD included some treated with a variety of oral corticosteroids. The 22 control subjects showed a six-minute walk result, in meters, with a mean of 623 and a standard deviation of 66, and the 17 DMD subjects showed a mean of 352 and a standard deviation of 87. These results were measured after one year of follow-up.

Cahalin et al. (2012), reporting on this same test in a different population, report negligible skewness, but significant excess kurtosis, for six-minute walk values, and so a tool that does not require normality of these observations is important.

In this section, we demonstrate how to design a hypothetical study with a two-stage two-arm analysis, using the multivariate Cornish–Fisher and Edgeworth techniques. We consider a population like the DMD population and a conceptual treatment that might improve walk test performance by cutting the decline by 40%. Patients like those with DMD in historical DMD studies will be randomized to treatment or control. The study will be powered to detect an increase in six-minute walk results, that is, 40% of the decline found by Henricson et al. (2012). We approximate the walk results for both control and treated individuals as standard normal, with approximate standard deviation $\sigma = 70$ and expectations $\mu_1 = 352$ and $\mu_2 = 352 + .4 \times (623 - 352) = 460$, respectively. Hence, the null hypothesis is equal distributions for treated and controls, approximately normal with expectations 352, and common standard deviation 75, and the alternative hypothesis is that treated individuals will have six-minute walk tests with expectations 352 and 460 and common standard deviation 72. Note that $(\mu_2 - \mu_1)/\sigma = 1.5$.

The randomization will be one to one. We plan this study with the same number of observations in each arm and at each stage (that is, $\lambda_{ij} = 1/4$ for $i, j = 1, 2$ in (17)) with first-stage one-sided size $\alpha_1 = 0.02$, as defined in (2), and ultimate one-sided size $\alpha_2 = 0.05$, as defined in (5), and 85% power.

This implies that $\omega = \Phi(1.5/\sqrt{2}) - .5 = 0.3556$, and approximation (21) gives $N = 20.69$. Dividing by four, and rounding to the nearest integer, gives 5 subjects per arm in each of the two stages. The uncorrected Gaussian approximation to the first critical value c_1 is 22.32, which, when rounded to the nearest midpoint of the integer lattice, gives 22.5. (Recall that the lattice midpoints are of the form an integer plus one half; the Edgeworth approximations are designed to have the proper error behavior if evaluated at such continuity-corrected points.) The conventional Cornish–Fisher approximation is 22.147, which is also rounded to 22.5.

Both first-stage critical values are rounded to an integer lattice plus half and so the approximate first-stage test level is changed. The Gaussian approximation to the nominal 0.02 level for the rounded Gaussian critical value is 0.0184. The Edgeworth approximation to the tail beyond the rounded Cornish–Fisher critical value is 0.0108.

The uncorrected bivariate Gaussian approximation to the second critical value, with first-stage test level 0.0184, is 73.5. The new Cornish–Fisher approximation to the second critical value, using the first-stage test level of 0.0108, is 72.5.

True levels for the resulting tests are calculated via Monte Carlo, taking 1,000,000 sets, each consisting of two groups of 10 standard normals, are given in Table 1, and show a slight improvement for the new method. The improvement is larger for the second example in Table 1, with test levels 0.01 and 0.025.

Powers are calculated in the same manner as test levels, but with the second group shifted by 1.5. The recommended total sample size, 20, is quite small and is smaller than that of Henricson et al. (2012). Never the less, it is larger than other recent clinical trials on DMD Peripheral and Central Nervous System Drugs Advisory Committee (2016).

Various authors suggest different strategies for setting critical values for the various analysis stages. Pocock (1977) suggests choosing critical values to keep nominal test levels at each stage equal. O'Brien and Fleming (1979), in the case with equal sample sizes in each of two stages, suggest an approach fixing $c_1 = c_2\sqrt{2}$, making stopping early somewhat rarer. The proposed calculations are also exhibited for approximate nominal levels chosen to be equal in Table 1. The stopping rule of O'Brien and Fleming (1979) is generally more powerful than that of Pocock (1977), as is expected.

11 Results

In this section, we evaluate approximations to critical values for one-sided tests of size 0.05 and evaluate approximations to test power. Tables with stages 1 and 2 group totals all 2 (that is, $m_1 = n_1 = m_2 = n_2 = 2$), through group totals all 9 (that is, $m_1 = n_1 = m_2 = n_2 = 9$), are examined. This upper limit on group sizes was determined by the largest experimental configurations for which counts given by (22) and (23) could be performed using long integer arithmetic. Hence, we

Table 1 Exact and approximate test levels. First-stage test levels as suggested by Pocock (1977) are marked with [a], and as suggested by O'Brien and Fleming (1979) are marked with [b]. The first six columns are target test level for stage 1 of the test, target test level after stage 2, critical values for the two stages using the bivariate normal distribution, and critical values using the Cornish–Fisher expansion. The following five columns are actual test level after stage 2 using the uncorrected bivariate normal critical values (labeled Unc.), the actual test level using corrected critical values from the Cornish–Fisher expansion (labeled Corr.), true power for the Cornish–Fisher critical values, approximate power calculated by applying a bivariate normal approximation to the Cornish–Fisher critical values, and approximate power calculated by applying the bivariate Edgeworth approximation to Cornish–Fisher critical values

		Critical values				Levels		Power		
		Uncorrected bivariate normal		Corrected Cornish–Fisher						
α_1	α_2					Unc.	Corr.	True	Normal	Edge.
(%)	(%)	c_1	c_2	c_1	c_2	(%)	(%)	(%)	(%)	(%)
3 subjects at each time in each arm										
5.00	10.00	0.5	9.5	0.5	9.5	11.5	11.5	86.33	85.66	84.60
5 subjects at each time in each arm										
2.00	5.0	22.5	73.5	22.5	72.5	4.68	5.31	91.23	93.74	92.4
1.00	2.5	23.5	77.5	23.5	76.5	2.31	2.67	83.66	85.69	85.36
3.01[a]	5.0	21.5	74.5	21.5	74.5	4.99	4.99	90.29	90.73	89.25
1.46[a]	2.5	22.5	80.5	22.5	79.5	2.23	2.42	77.33	77.47	78.46
0.88[b]	5.0	23.5	72.5	23.5	72.5	4.84	4.84	92.33	93.68	92.59
0.26[b]	2.5	25.5	76.5	24.5	76.5	2.16	2.40	85.30	85.57	85.66
8 control and 7 treatment subjects at each time										
2.00	5.0	53.5	186.5	53.5	186.5	5	5	98.34	99.2	98.25
1.00	2.5	56.5	193.5	55.5	193.5	2.34	2.58	96.23	97.48	96.22

examined $8^4 = 4096$ configurations in all, all with one-sided critical values with stage 1 size .02 and ultimate size .05.

12 Errors in Levels for Approximate Critical Values

We compare bivariate normal and Cornish–Fisher approximations to critical values. Generally, because of the discrete nature of the values of the Wilcoxon test statistic, the corrected and uncorrected tests tend to be conservative.

Approximately, one quarter of these tables exhibited a different integer value for one or both critical values. Absolute errors in levels of normal and Cornish–Fisher critical values for these tables in which the critical values differed are given in Fig. 1.

The Cornish–Fisher approximation provides better size control for most configurations in which approximations differ and for all configurations in which either approximation gives substantial errors.

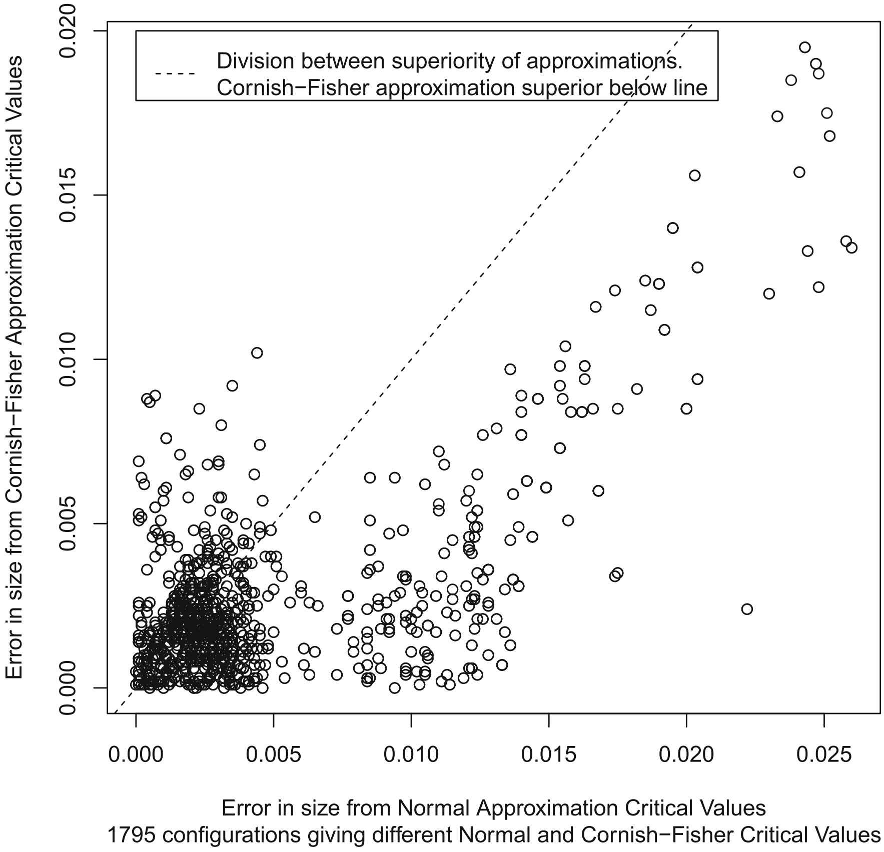


Fig. 1 Error in sizes of approximations to test

13 Errors in Approximations to Power

We compare normal and Edgeworth power approximations. The Cornish–Fisher critical values are used for both. The alternative model features independent univariate normals with mean offset by δ, with δ chosen to give approximate .8 power for the final stage Mann–Whitney test using all $m_1 + n_1 + m_2 + n_2$ observations, ignoring the first stage. All 4096 configurations were examined. For almost all configurations, the Edgeworth approximation proved more accurate than the bivariate normal. These absolute errors are displayed in Fig. 2.

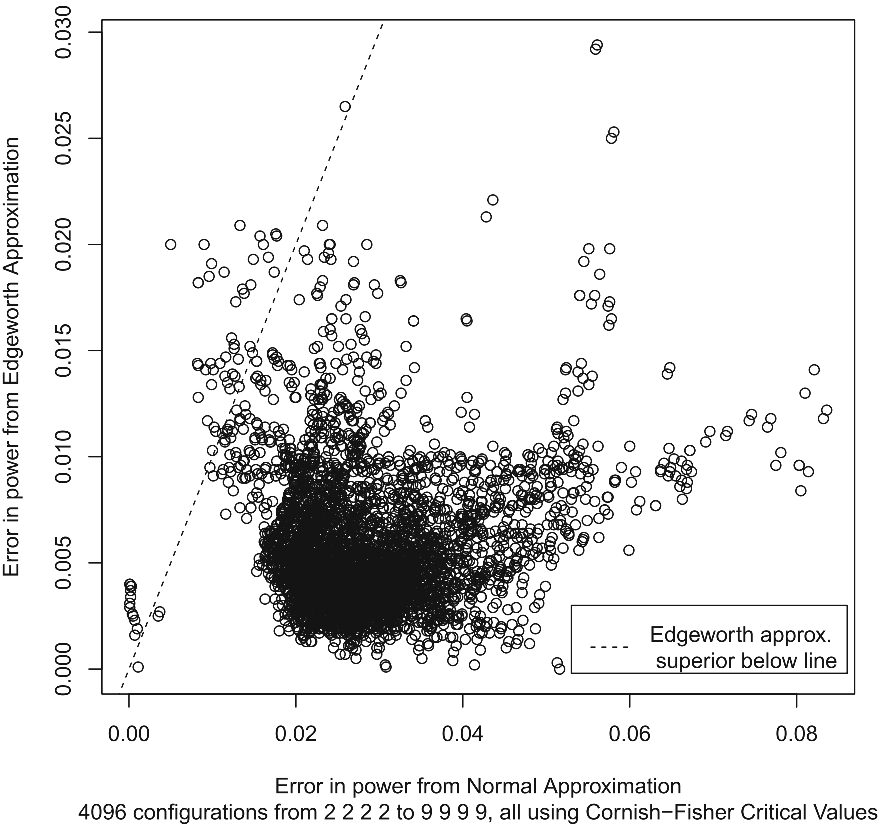


Fig. 2 Errors in approximations of powers of tests

14 Conclusions

With more and more targeted therapies being developed to treat rare diseases (e.g., Eteplirsen for DMD and Symdeko for cystic fibrosis), small sample sizes for confirmatory trials are becoming more common. Furthermore, there is a need for making these therapies available to patients as soon as possible, if the efficacy is very promising. Hence, sequential analyses that are more accurate and easier to implement such as the one developed here is a step in the direction of bringing promising therapies to patients sooner.

For constructing the two-stage Mann–Whitney–Wilcoxon procedure, the bivariate Cornish–Fisher expansion provides improved control of test size over its bivariate normal competitor. The bivariate Edgeworth series provides a more accurate approximation to powers of this test. Both of these approaches should be used with continuity correction.

Acknowledgments The authors thank Yuzo Maruyama for an English summary of Takeuchi (1978) and Todd Kuffner for pointers to references Takemura and Takeuchi (1988) and Takemura and Takeuchi (1988). John Kolassa thanks NSF DMS 1712839 for partial support for this project. The authors thank editors and referees for helpful comments.

Appendix 1: A Bivariate Recursion for Exact Probabilities

The univariate recursion is constructed by counting the number of data set orderings leading to the statistic value and decomposing them into orderings based on one fewer value from the first group and one fewer value from the second group. The bivariate recursion is similar to the above univariate recursion and is similar to that of Wilding et al. (2011).

Let $b(u_1, u_2, m_1, n_1, m_2, n_2)$ represent the number of orderings of (4) for which $U_1 = u_1$ and $U_2 = u_2$. This number is zero if any of the sample sizes m_1, n_1, m_2, n_2 is negative, if either statistic value is negative, or if either statistic value is larger than its maximum value. It is also zero if both additional sample sizes for stage 2 are zero but the second statistic value exceeds the first. If all sample sizes are zero, then the sums in (1) and (3) are empty, and both statistic values are zero; hence $b(0, 0, 0, 0, 0, 0) = 1$. These end conditions are given by

$$b(u_1, u_2, m_1, n_1, m_2, n_2) := \begin{cases} 0, & \text{if any of } m_1, m_2, n_1, n_2, u_1, u_2 \text{ is negative or} \\ & \ \text{if } u_1 > m_1 n_1 \text{ or } u_2 > (m_1 + m_2)(n_1 + n_2) \\ 0, & \text{if } n_1 = 0 \text{ or } m_1 = 0, \text{ and } u_1 > 0 \\ 1, & \text{if } m_1 = 0 \text{ and } n_1 = 0 \text{ and } u_1 = 0 \\ & \ \text{and } m_2 = 0 \text{ and } n_2 = 0 \text{ and } u_2 = 0. \end{cases} \tag{22}$$

Otherwise, the number of rearrangements of the data (4) giving rise to statistic values u_1 and u_2 are the sum of four contributions. First, add those with sample sizes $m_1 - 1, n_1, m_2, n_2$, with an additional value from the first group in the first sample that exceeds all values in the sample, and hence leaves the statistic value unchanged. Second, add those with sample sizes $m_1, n_1 - 1, m_2, n_2$, with an additional value from the second group in the first sample that exceeds all values in the sample, and hence increases the first statistic by m_1 and increases the second statistic by $m_1 + m_2$. Third, add those with sample sizes $m_1, n_1, m_2 - 1, n_2$, with an additional value from the first group in the second sample added that exceeds all values in the sample, and hence leaves the statistic value unchanged. Fourth, add those with sample sizes $m_1, n_1, m_2, n_2 - 1$, with an additional value from the second group in the second sample added that exceeds all values in the sample, and hence leaves U_1 unchanged, and increases the second statistic by $m_1 + m_2$. This leads to the following recursion:

$$\begin{aligned} b(u_1, u_2, m_1, n_1, m_2, n_2) := {} & m_1 b(u_1, u_2, m_1 - 1, n_1, m_2, n_2) \\ & + n_1 b(u_1 - m_1, u_2 - m_1 - m_2, m_1, n_1 - 1, m_2, n_2) \\ & + m_2 b(u_1, u_2, m_1, n_1, m_2 - 1, n_2) \\ & + n_2 b(u_1, u_2 - m_1 - m_2, m_1, n_1, m_2, n_2 - 1). \end{aligned} \tag{23}$$

Numerical examples in Figs. 1 and 2 exhibit comparisons of probability approximations to bivariate probabilities and quantiles for $\mathbf{U} = (U_1, U_2)$, to the exact values.

Appendix 2: A Continuous Example with Nonzero Skewness

Our aim in determining the expansion for c_2 is to apply the techniques to Wilcoxon testing, but the same quantile approximation may be used more generally. Before application to the Wilcoxon statistic, which is somewhat atypical, because the third cumulants are zero, leading to a less dramatic effect, and because the Wilcoxon statistic is discrete, and hence lacks the continuity that the technique was developed for. Instead, we present a more general example consisting of a continuous distribution with nonzero third order cumulants.

Consider Y_1, Y_2, Y_3 independent exponentials. Let $U_1 = Y_1 + Y_3$, and $U_2 = Y_2 + Y_3$. Figure 3 compares Edgeworth (E) or Cornish–Fisher (CF), normal (N), and Monte Carlo (MC, taken with 500,000 samples and treated as the truth). Panels are:

(a) compares difference of E and N upper tail univariate probability approximation from the MC approximation, as a function of the MC approximation.
(b) compares error of E and N upper tail bivariate probability approximation, as a function of the ordinate.
(c) Gives contours of MC upper tail probabilities.
(d) Represents CF and N approximation to upper tail, vs. MC value. CF and N values exhibit some dependence on target for first univariate tail, and so are represented as a range.

Note from panel b that the Edgeworth approximation fails to dominate the normal approximation only for a narrow band between the contours marked 0 in the middle of the plot.

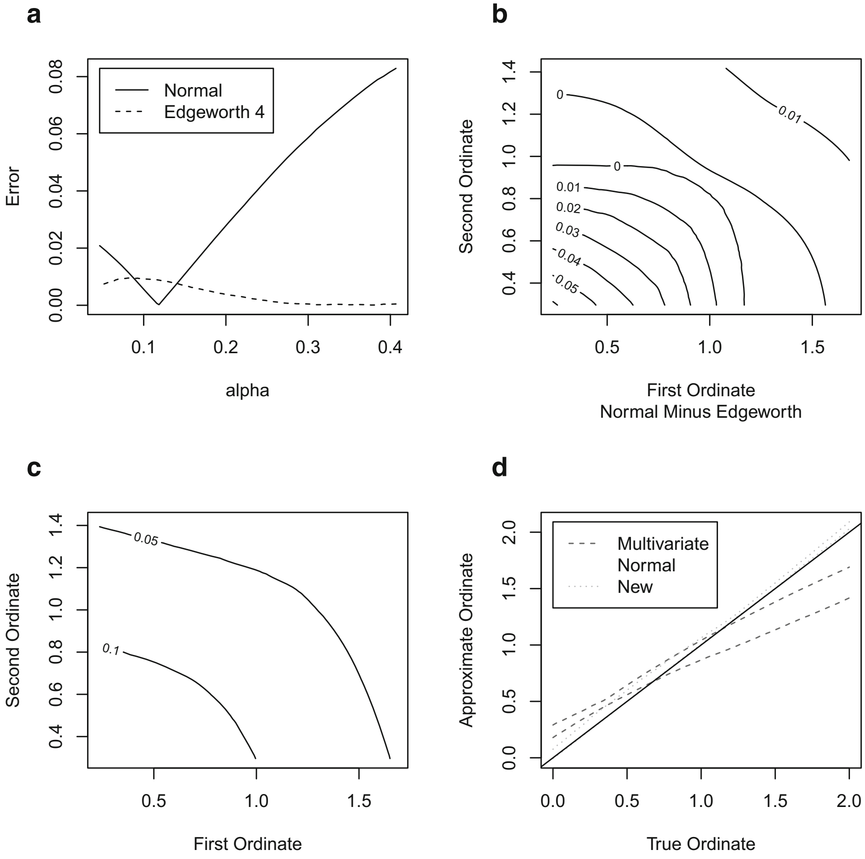


Fig. 3 Exponential example, sample size 1. (**a**) Absolute error for approximations. (**b**) Difference in absolute error. (**c**) True bivariate probability. (**d**) Second ordinate approximation range

References

Amrhein, P. (1995). An example of a two-sided wilcoxon signed rank test which is not unbiased. *Annals of the Instatute of Statistical Mathematics*, *47*(1), 167–170.

Cahalin, L. P., Arena, R., & Guazzi, M. (2012). Comparison of heart rate recovery after the six-minute walk test to cardiopulmonary exercise testing in patients with heart failure and reduced and preserved ejection fraction. *The American Journal of Cardiology*, *110*(3), 467–468. http://www.sciencedirect.com/science/article/pii/S0002914912013318.

Chow, S.-C., & Huang, Z. (2020). Innovative design and analysis for rare disease drug development. *Journal of Biopharmaceutical Statistics*, *30*(3), 537–549. PMID: 32065047. https://doi.org/10.1080/10543406.2020.1726371.

Cornish, E. A., & Fisher, R. A. (1938). Moments and cumulants in the specification of distributions. *Revue de l'Institut International de Statistique/Review of the International Statistical Institute*, *5*(4), 307–320. http://www.jstor.org/stable/1400905.

Fix, E., & Hodges, J. L. (1955). Significance probabilities of the wilcoxon test. *Annals of Mathematical Statistics*, *26*(2), 301–312. https://doi.org/10.1214/aoms/1177728547.

Henricson, E., Abresch, R., Han, J. J., Nicorici, A., Goude Keller, E., Elfring, G., Reha, A., Barth, J., & McDonald, C. M. (2012). Percent-predicted 6-minute walk distance in duchenne muscular dystrophy to account for maturational influences. *PLOS Currents*, *4*, RRN1297–RRN1297. https://www.ncbi.nlm.nih.gov/pubmed/22306689.

Kolassa, J. (1989). Topics in Series Approximations to Distribution Functions. Ph.D. thesis, University of Chicago.

Kolassa, J. E. (1995). A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data. *Statistics in Medicine*, *14*(14), 1577–1581.

Kolassa, J. E. (2003). Multivariate saddlepoint tail probability approximations. *The Annals of Statistics*, *31*(1), 274–286. http://www.jstor.org/stable/3448375.

Kolassa, J. E., & McCullagh, P. (1990). Edgeworth series for lattice distributions. *The Annals of Statistics*, *18*(2), 981–985. https://doi.org/10.1214/aos/1176347637.

Kolassa, J. E., & Seifu, Y. (2013). Nonparametric multivariate inference on shift parameters. *Academic Radiology*, *20*(7), 883–888. http://www.sciencedirect.com/science/article/pii/S1076633213001645.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491.

McCullagh, P. (1987). *Tensor methods in statistics*. Monographs on statistics and applied probability. Chapman and Hall. https://books.google.com/books?id=JEjvAAAAMAAJ.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549–556. http://www.jstor.org/stable/2530245.

Peripheral and Central Nervous System Drugs Advisory Committee (2016). Nda 206488: Eteplirsen. Technical report. U.S. Food and Drug Administration.

Pharmaceutical Research and Manufacturers of America (2019). Spurring innovation in rare diseases. Downloaded 22 March 2020. https://www.phrma.org/-/media/Project/PhRMA/PhRMA-Org/PhRMA-Org/PDF/P-R/RareDisease_Backgrounder1.pdf.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199. https://doi.org/10.1093/biomet/64.2.191.

Rahardja, D., Zhao, Y. D., & Qu, Y. (2009). Sample size determinations for the Wilcoxon-Mann-Whitney test: A comprehensive review. *Statistics in Biopharmaceutical Research*, *1*(3), 317–322. https://doi.org/10.1198/sbr.2009.0016.

Rom, D. M., & McTague, J. A. (2020). Exact critical values for group sequential designs with small sample sizes. *Journal of Biopharmaceutical Statistics*, *30*(4), 1–13. PMID: 32151177. https://doi.org/10.1080/10543406.2020.1730878.

Shieh, G., Jan, S., & Randles, R. H. (2006). On power and sample size determinations for the Wilcoxon-Mann-Whitney test. *Journal of Nonparametric Statistics*, *18*(1), 33–43. https://doi.org/10.1080/10485250500473099.

Spurrier, J. D., & Hewett, J. E. (1976). Two-stage Wilcoxon tests of hypotheses. *Journal of the American Statistical Association*, *71*(356), 982–987. https://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480981.

Sundrum, R. M. (1954). A further approximation to the distribution of Wilcoxon's statistic in the general case. *Journal of the Royal Statistical Society. Series B (Methodological)*, *16*(2), 255–260. http://www.jstor.org/stable/2984051.

Takemura, A., & Takeuchi, K. (1988). Some results on univariate and multivariate Cornish-Fisher expansion: Algebraic properties and validity. *Sankhyā: The Indian Journal of Statistics, Series A (1961–2002)*, *50*(1), 111–136. http://www.jstor.org/stable/25050684.

Takeuchi, K. (1978). A multivariate generalization of Cornish Fisher expansion and its applications (in Japanese). *Keizaigaku Ronshu*, *44*(2), 1–12.

Whitehead, J., & Jaki, T. (2009). One- and two-stage design proposals for a phase II trial comparing three active treatments with control using an ordered categorical endpoint. *Statistics in Medicine*, *28*(5), 828–847. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3508.

Wilding, G. E., Shan, G., & Hutson, A. D. (2011). Exact two-stage designs for phase ii activity trials with rank-based endpoints. *Contemporary Clinical Trials*, *33*(2), 332–341. https://doi.org/10.1016/j.cct.2011.10.008.

Wold, H. (1934). Sheppard's correction formulae in several variables. *Scandinavian Actuarial Journal*, *1934*(1), 248–255. https://doi.org/10.1080/03461238.1934.10419243.

Wolfram Research, Inc. (2018). Mathematica, Version 11.3. Champaign, IL.

Zhong, D., & Kolassa, J. (2017). Moments and Cumulants of The Two-Stage Mann-Whitney Statistic. Technical report.

Data Nuggets in Supervised Learning

Kenneth Edward Cherasia, Javier Cabrera, Luisa T. Fernholz, and Robert Fernholz

Abstract Big data presents many challenges in modern statistics and data analysis. While a large number of observations can lead to increased precision in statistical parameter estimation and prediction, computational and storage costs may present a problem. Since there is often significant redundancy in large data in a lower dimensional setting, it seems reasonable that big datasets can be compressed to a smaller number of observations with comparable statistical performance, where the amount of compression scales with the dimensionality. We propose an extension of the "data nuggets" methodology of Beavers et al. (2020) for a compression-based approach to statistical modeling in big data. We utilize the linear regression model to showcase the idea, establish a theoretical foundation, and explore finite-sample performance via simulation analysis. Data nuggets are shown to provide a significant improvement over random sampling in model parameter estimation and out-of-sample prediction performance in the linear regression setting, and the concept is promising for other models as well.

Keywords Big data · Data nuggets · Supervised learning · Asymptotic efficiency

K. E. Cherasia · J. Cabrera (✉)
Department of Statistics, Rutgers University, Piscataway, NJ, USA
e-mail: kecherasia@outlook.com; cabrera@stat.rutgers.edu

L. T. Fernholz
Department of Statistics, Temple University, Philadelphia, PA, USA
e-mail: luisa.fernholz@temple.edu

R. Fernholz
Intech Corp., Princeton, NJ, USA
e-mail: bob@bobfernholz.com

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_20

1 Introduction

A significant amount of modern statistics and data analysis is focused on big data, which has become common in every industry from finance and economics to science and medicine. This big data can be in the form of a large dataset consisting of millions of observations, such as a hospital database containing demographic and medical information on many patients, or even constant-flowing streaming data, as will be collected on the order of Exabytes, or millions of Terabytes, using the Square Kilometer Array (Zhang and Zhao 2015). Limitations in analyzing such data include but are not limited to computational time, memory constraints, and long-term storage.

A reasonable approach to avoid such limitations is reducing the size of the data. For example, a simple random sample of the data could be used; however, this will likely result in decreased precision of estimation or prediction, increased uncertainty, and variability among samples, depending on the desired amount of reduction. For these reasons, a completely random sample is often not desirable unless the size is very large, which defeats the purpose.

1.1 Literature Overview

A better solution to the problem of representing big data is using a representative sample, instead of a completely random sample. By a representative sample, we refer to a set of points that accurately reflect the full data in terms of distribution. For example, B. A. Flury introduced the concept of "principal points," a set of points that minimize the expected Euclidean distance of a random vector to the nearest point within the set (Flury 1990, 1997). Tibshirani (1992) introduced the concept of "principal curves," essentially a generalization of linear principal components (Tibshirani 1992). Mak proposed the concept of "support points," a similar idea to Flury's "principal points" except with a different distance measure (Mak and Joseph 2018).

Another closely related idea is data compression. DuMouchel et al. (1999) introduced the concept of "data squashing," a type of lossy data compression that results in a set of fewer weighted observations (DuMouchel et al. 1999). There are several published papers related to this concept of "data squashing" (DuMouchel 2002). These include Owen (1999) and Madigan et al. (2001). All of these approaches involve the same underlying framework of likelihood approximation with compressed data.

1.2 Data Nuggets

The aforementioned methods for representative sampling and data compression inspired the concept of “data nuggets” by Beavers et al. (2020), which falls in the intersection of both frameworks (Beavers et al. 2020). Data nuggets are a set of points in a higher dimensional space that represent a reduction in size, i.e., the number of observations, of the original data while retaining the general structure of the data, including on the periphery. Each data nugget corresponds to a set of observations in the original data and is described by three parameters: center, scale, and weight. Essentially, each data nugget is a set of summary statistics corresponding to the observations in one subset in a partition of the data. The nuggets are created in such a way that within-nugget distance between points is minimized while keeping computational time feasible.

The original creators of the data nuggets methodology were motivated by a big data example in flow cytometry, in this case, researching the level of expression of specific proteins on the surface of B-cells using clustering and principal component analysis (Beavers et al. 2020). The primary issue was estimation of the covariance matrix due to the size of the data, since this particular dataset contained over one million observations; however, the authors showed that the covariance matrix can be reasonably recovered using a much smaller number of data nuggets, since the within-nugget variation is minimized during the generation of these nuggets.

Beavers et al. (2020) applied the data nuggets methodology in an unsupervised learning setting for purposes of clustering or principal component analysis. However, compression-based methodology in general also has applications in supervised learning, e.g., statistical modeling and prediction. We extend the idea of data nuggets to supervised learning for purposes of modeling and prediction. In this setting, the data contains a response variable, which must be factored in during nugget generation. Otherwise, the central idea is similar—create a set of points that represents the original data, where each point is a summary of a set of observations in the original dataset according to some partition, and the partitioning process is designed to minimize internal variability. We will focus on the linear regression model to illustrate the concept, but the idea is applicable in the general setting with some adjustments.

2 Setup

Consider data $\{(x_k^*, y_k^*)\}_{k=1}^N$ consisting of N observations, where x_k^* is a p-dimensional covariate vector with numeric or binary entries and y_k^* is a continuous response. Suppose the data is partitioned into $M << N$ data nuggets $\{D_i\}_{i=1}^M$, where D_i is a set of summary statistics for nugget i and $N = \sum_{i=1}^M n_i$. Relabel the data as (x_{ij}, y_{ij}) for $i = 1, \ldots, M$ and $j = 1, \ldots, n_i$ so that the data subset $\{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$

for fixed i corresponds to nugget D_i. We want to fit a linear regression model using only the information contained in the data nuggets.

There are two key components to this process that require specification. First, it must be decided how the data nuggets are formed, relating to both the partitioning of the observations and summarization of observations within each nugget. Second, an appropriate estimator for model parameters using only the nugget information must be determined. We discuss both of these issues here.

2.1 Formation of Nuggets

First, the N observations must be partitioned into M disjoint sets, where M is the desired number of nuggets. There are two readily available options for partitioning using Beavers' algorithm: use only the feature variables in distance computation, or use both the features and response. There are several problems with the latter, particularly how to generalize to different classes of response and appropriately weigh the response in distance calculations as the number of features increases. Therefore, we will use Beavers' algorithm with only the features for distance computation.

Second, assuming that a partition has been completed, the observations within each nugget must be represented by a set of summary statistics. One of these must be the weight or the number of observations within the nugget. For a center or location parameter, several options were explored, including:

A. X-center defined as within-nugget covariate centroid
Y-center defined as within-nugget mean response
B. X-center defined as within-nugget covariate centroid
Y-center defined by within-nugget model prediction at X-center
C. X-center defined as random point within nugget
Y-center defined by within-nugget model prediction at X-center

Option C had variable performance and is not recommended. Options A and B performed similarly in most cases, since the prediction at the X-center is likely close to the mean due to the minimization of internal feature distance in nugget creation. Variable parameters can include different specifications for X, Y, and XY variation. The exact parameters required depend on the estimator that will be used.

2.2 Estimation with Nuggets

Assume that the data have been reduced to M nuggets. Let (x_i, y_i) be center or centroid; s_i^x the x-scale (vector) or internal x-variability, i.e., corresponding entry on the within-nugget covariance matrix for covariates only; s_i^y the y-scale or internal y-

variability; s_i^{xy} the xy-scale (vector) or internal xy-covariability; and n_i the weight of the i^{th} data nugget. The following estimators were considered:

A. A weighted least-squares (WLS) estimator, with weights $w_i = n_i$.
The minimum nugget parameters are (x_i, y_i, n_i).

B. A WLS estimator, with weights $w_i \propto 1/s_i^x$ such that $\sum\limits_{i=1}^{M} w_i = N$.
The minimum nugget parameters are (x_i, y_i, n_i, s_i^x).

C. An ordinary least-squares (OLS) "mimic" estimator, which tries to approximate the OLS estimator of β in the full data case, $\hat{\beta}_{LS}$.
The minimum nugget parameters are $(x_i, y_i, n_i, s_i^x, s_i^y, s_i^{xy})$.

Option A seemed to perform the best. For Option C, the "mimic" estimator is simply the ordinary least-squares estimator in the full data case where the within-nugget covariation matrix for X is approximated by a diagonal matrix. The general approach can be inferred from Theorem 1 in the next section. This option did not have good performance in general and required too many parameters.

3 Asymptotics

We examine the theoretical asymptotic properties of the previously defined data nugget regression parameter estimators in the homoscedastic linear regression setting. This includes consistency of the coefficient and variance estimators and asymptotic normality.

The following results will be consistent with the notation in the previous section. Given N points $\{(x_{ij}, y_{ij})\}$, $M << N$ nuggets of the form $D_i = (x_i, y_i, n_i)$ are formed, where $x_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$ and $y_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ for $i \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, n_i\}$.

3.1 Intuition

Our goal is to perform linear regression and associated statistical inference using only the information contained in the nuggets $\{D_i\}_{i=1}^{M}$. The following result provides some intuition for performing that statistical inference.

Theorem 1 *Consider the linear model* $y_{ij} = x_{ij}'\beta + \epsilon_{ij}$ *for* $i = 1, \ldots, M$ *and* $j = 1, \ldots, n_i$. *Define* $x_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$, $y_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$, *and* $s_i^y = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (y_{ij} - y_i)^2$.

1. *The residual sum of squares (RSS) for this model can be written in the form:*

$$RSS = \underbrace{\sum_{i=1}^{M} n_i (y_i - x_i'\beta)^2}_{(A)} + \underbrace{\sum_{i=1}^{M}(n_i - 1)s_i^y}_{(B)} + \underbrace{\beta'\left(\sum_{i=1}^{M}(n_i - 1)Cov(X_i)\right)\beta}_{(C)}$$
$$+ \underbrace{2\beta'\left(\sum_{i=1}^{M}\left(n_i x_i y_i - \sum_{j=1}^{n_i} x_{ij} y_{ij}\right)\right)}_{(D)}$$

where $Cov(X_i)$ is the covariance matrix of $(x_{i1}, \ldots, x_{in_i})$.

2. *The ordinary least-squares (OLS) estimator of β, denoted by $\hat{\beta}_{LS}$, is given explicitly by the following equation, provided the model matrix is of full rank:*

$$\hat{\beta}_{LS} = \left[\left(\frac{1}{N}\sum_{i=1}^{M} n_i x_i x_i'\right) + \left(\frac{1}{N}\sum_{i=1}^{M}(n_i - 1)Cov(X_i)\right)\right]^{-1} *$$
$$\left[\left(\frac{1}{N}\sum_{i=1}^{M} n_i x_i y_i\right) - \left(\frac{1}{N}\sum_{i=1}^{M}\left(n_i x_i y_i - \sum_{j=1}^{n_i} x_{ij} y_{ij}\right)\right)\right]$$

Proof The derivation of the mathematical form for the residual sum of squares in (1) is straightforward. Let $RSS(i)$ denote the contribution of $\{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$ to the RSS such that $RSS = \sum_{i=1}^{M} RSS(i)$. Note that the RSS can be represented as

$$RSS(i) = \sum_{j=1}^{n_i}(y_{ij} - x_{ij}'\beta)^2 = \sum_{j=1}^{n_i}\left[(y_i - x_i'\beta) + (y_{ij} - y_i) + (x_i - x_{ij})'\beta\right]^2$$

Upon expansion, we obtain the following expression:

$$RSS(i) = \underbrace{\sum_{j=1}^{n_i}(y_i - x_i'\beta)^2}_{\text{(A)}} + \underbrace{\sum_{j=1}^{n_i}(y_{ij} - y_i)^2}_{\text{(B)}} + \underbrace{\sum_{j=1}^{n_i}((x_{ij} - x_i)'\beta)^2}_{\text{(C)}}$$
$$+ 2\sum_{j=1}^{n_i}\left[\underbrace{(y_{ij} - y_i)(x_i - x_{ij})'\beta}_{\text{(D)}} + \underbrace{(y_i - x_i'\beta)(y_{ij} - y_i)}_{=0} + \underbrace{(y_i - x_i'\beta)(x_i - x_{ij})'\beta}_{=0}\right]$$

where the last two terms are identically zero, since $\sum_{j=1}^{n_i}(x_{ij} - x_i) = \sum_{j=1}^{n_i}(y_{ij} - y_i) = 0$ by definition. Re-expressing terms (B), (C), and (D) and noting that $RSS =$

$\sum_{i=1}^{M} RSS(i)$ conclude the proof of claim (1). Regarding claim (2), we know that $\hat{\beta}_{LS} = \arg\min_{\beta}\{RSS\}$ and is a solution to the following equation:

$$0 = \frac{1}{2}\left(\frac{\partial RSS}{\partial\beta}\right) = -\sum_{i=1}^{M} n_i x_i y_i + \left(\sum_{i=1}^{M} n_i x_i x_i'\right)\beta + \left(\sum_{i=1}^{M} (n_i - 1)\mathrm{Cov}(X_i)\right)\beta$$
$$+\sum_{i=1}^{M}\left(n_i x_i y_i - \sum_{j=1}^{n_i} x_{ij} y_{ij}\right)$$

In the full-rank case, rearrangement and re-scaling yield our claim (2). It can be verified that $\hat{\beta}_{LS}$ is indeed a minimizer by checking the second derivative. □

This theorem provides a decomposition of the least-squares estimator in the full data case into terms involving nugget parameters and terms involving internal nugget variability (which may be unknown after nugget formation). A reasonable estimator of β is the case-weighted least-squares estimator $\hat{\beta}_N$ defined by

$$\hat{\beta}_N = \left(\sum_{i=1}^{M} n_i x_i x_i'\right)^{-1}\left(\sum_{i=1}^{M} n_i x_i y_i\right)$$

Note that $\hat{\beta}_N$ is unbiased for estimating β, which is a desirable property. The corresponding response variance estimator $\hat{\sigma}_N^2$ for σ^2 is

$$\hat{\sigma}_N^2 = \frac{1}{M-(p+1)}\sum_{i=1}^{M} n_i (y_i - x_i'\hat{\beta}_N)^2$$

For relatively large M, it is reasonable to replace $M-(p+1)$ by just M in the denominator. We will use the latter for simplicity of notation in the development of asymptotic results, but the derived asymptotic results are the same for both estimators. In the following section, we will establish the asymptotic properties of these estimators as both the number of observations N and number of nuggets M tend toward infinity.

3.2 Consistency of Coefficient Estimator

The following theorem provides conditions under which $\hat{\beta}_N$ is consistent for the estimation of β.

Theorem 2 *Consider the linear model* $y_{ij} = x'_{ij}\beta + \epsilon_{ij}$ *for* $i = 1, \dots, M$ *and* $j = 1, \dots, n_i$, *where* $(\epsilon_{ij})_{(i,j)}$ *are i.i.d. with* $E[\epsilon_{ij}|x_{ij}] = 0$ *and* $Var[\epsilon_{ij}|x_{ij}] = \sigma^2 < \infty$. *Assume that the following conditions are satisfied:*

1. $\lim\limits_{M\to\infty}\left(\max\limits_{1\le i\le M} s_i^x\right) = 0$
2. $\max\limits_{1\le i\le M}\{n_i\} < K$ *for some constant* $1 \le K < \infty$

Then, $\hat{\beta}_N$ *is consistent for estimating* β, *i.e.,* $\hat{\beta}_N - \beta = o_p(1)$.

To simplify the argument, we will first introduce a lemma.

Lemma 1 *Define* $\delta_{ij} = x_{ij} - x_i$. *If* $\lim\limits_{M\to\infty} s_i^x = 0$ *and* $n_i < K$ *for some positive constant* $K < \infty$, *then* $\lim\limits_{M\to\infty}\left\|\delta_{ij}\right\|_{L_2} = 0$ *for all* $j = 1, \dots, n_i$.

Proof If $n_i = 1$, then trivially $s_i^x = 0$ and $\left\|\delta_{ij}\right\|_{L_2} = 0$ for $j = 1, \dots, n_i$. If $n_i \ge 2$, then s_i^x can be expressed as

$$s_i^x = \text{tr}(Cov(X_i)) = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\text{tr}(\delta_{ij}\delta'_{ij}) = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left\|\delta_{ij}\right\|_{L_2}^2$$

and, in the limiting case, we have

$$\lim_{M\to\infty}\left[\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left\|\delta_{ij}\right\|_{L_2}^2\right] = \lim_{M\to\infty} s_i^x = 0$$

Since $2 \le n_i < K$ and $\left\|\delta_{ij}\right\|_{L_2}^2 \ge 0$ for all (i, j), the result follows. □

Now, we use the lemma to prove the consistency result using our decomposition.

Proof It is known that $\hat{\beta}_{LS} = \beta + o_p(1)$, so by Theorem 1 and the continuous mapping theorem, it suffices to show that:

1. $\lim\limits_{M\to\infty}\dfrac{1}{N}\sum\limits_{i=1}^{M}(n_i - 1)\text{Cov}(X_i) = 0$
2. $\dfrac{1}{N}\sum\limits_{i=1}^{M}\left(n_i x_i y_i - \sum\limits_{j=1}^{n_i} x_{ij}y_{ij}\right) = o_p(1)$

We start with claim (1). Since $\lim\limits_{M\to\infty} s_i^x = 0$ for all i and thus $\lim\limits_{M\to\infty}\left\|\delta_{ij}\right\|_{L_2} = 0$ for all (i, j) by our lemma, we know that

$$\lim_{M\to\infty}\left\|\delta_{ij}\delta'_{ij}\right\|_F = 0$$

The within-nugget covariance matrix can be represented in terms of deviations δ_{ij}:

$$\frac{1}{N}\sum_{i=1}^{M}(n_i-1)\mathrm{Cov}(X_i)=\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\delta_{ij}\delta_{ij}'$$

By the triangle inequality, in the limiting case, we have

$$\lim_{M\to\infty}\left\|\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\delta_{ij}\delta_{ij}'\right\|_F \leq \lim_{M\to\infty}\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\|\delta_{ij}\delta_{ij}'\right\|_F=0$$

This implies claim (1). Next, we prove claim (2). Note that we can represent $x_{ij}=x_i+\delta_{ij}$, $y_{ij}=x_i'\beta+\delta_{ij}'\beta+\epsilon_{ij}$, and $y_i=x_i'\beta+\epsilon_i$ for $\epsilon_i=n_i^{-1}\sum_{j=1}^{n_i}\epsilon_{ij}$ and so

$$\sum_{j=1}^{n_i}x_{ij}y_{ij}=\sum_{j=1}^{n_i}(x_i+\delta_{ij})(x_i'\beta+\delta_{ij}'\beta+\epsilon_{ij})$$

After some algebraic manipulation, our quantity of interest can be represented as

$$\frac{1}{N}\sum_{i=1}^{M}\left(n_ix_iy_i-\sum_{j=1}^{n_i}x_{ij}y_{ij}\right)=\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\Big(x_i\delta_{ij}'\beta+\delta_{ij}x_i'\beta+\delta_{ij}\delta_{ij}'\beta+\delta_{ij}\epsilon_{ij}\Big)$$

Observe that several terms have limit zero by the Cauchy–Schwarz inequality:

$$\lim_{M\to\infty}\left\|\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}x_i\delta_{ij}'\right\|_F=\lim_{M\to\infty}\left\|\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\delta_{ij}x_i'\right\|_F=\lim_{M\to\infty}\left\|\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\delta_{ij}\delta_{ij}'\right\|_F=0$$

By the WLLN under our finite variance assumption,

$$\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\delta_{ij}\epsilon_{ij}=o_p(1)$$

Claim (2) follows from the decomposition and these limiting results. □

3.3 Consistency of Variance Estimator

The following theorem provides conditions under which $\hat{\sigma}_N^2$ is consistent for the estimation of σ^2. For simplicity of notation, we will use M instead of $M-(p+1)$ in the denominator of $\hat{\sigma}_N^2$, but the developed asymptotic results are the same.

Theorem 3 *Consider the linear model* $y_{ij} = x_{ij}'\beta + \epsilon_{ij}$ *for* $i = 1, \ldots, M$ *and* $j = 1, \ldots, n_i$, *where* $(\epsilon_{ij})_{(i,j)}$ *are i.i.d. with* $E[\epsilon_{ij}|x_{ij}] = 0$ *and* $Var[\epsilon_{ij}|x_{ij}] = \sigma^2 < \infty$. *Assume that the following conditions are satisfied:*

1. $\lim\limits_{M\to\infty} s_i^x = 0$ *for all* $i = 1, \ldots, M$
2. $\max\limits_{1\le i\le M} \{n_i\} < K$ *for some positive constant* $K < \infty$
3. $\lim\limits_{M\to\infty}\left[\frac{1}{M}\sum\limits_{i=1}^{M} n_i x_i x_i'\right] = G$, *where* $\|G\|_F < \infty$

Then, $\hat{\sigma}_N^2$ *is consistent for estimating* σ^2, *i.e.,* $\hat{\sigma}_N^2 - \sigma^2 = o_p(1)$.

Proof First, note that our assumptions imply consistency of $\hat{\beta}_N$. Define $\epsilon_i = n_i^{-1}\sum_{j=1}^{n_i}\epsilon_{ij}$, where $\text{Var}[\epsilon_i|x_{i1}, \ldots, x_{in_i}] = n_i^{-1}\sigma^2$. The variance is

$$\hat{\sigma}_N^2 = \frac{1}{M}\sum_{i=1}^{M} n_i(y_i - x_i'\hat{\beta}_N)^2 = \frac{1}{M}\sum_{i=1}^{M} n_i((y_i - x_i'\beta) + x_i'(\beta - \hat{\beta}_N))^2$$

Upon expansion of terms, we have

$$\hat{\sigma}_N^2 = \underbrace{\frac{1}{M}\sum_{i=1}^{M} n_i(y_i - x_i'\beta)^2}_{(1)} + \underbrace{(\hat{\beta}_N - \beta)'\left[\frac{1}{M}\sum_{i=1}^{M} n_i x_i x_i'\right](\hat{\beta}_N - \beta)}_{(2)}$$
$$\underbrace{-\frac{2}{M}\sum_{i=1}^{M} n_i(y_i - x_i'\beta)x_i'(\hat{\beta}_N - \beta)}_{(3)}$$

Regarding (1), by the WLLN we have convergence in probability to σ^2:

$$\frac{1}{M}\sum_{i=1}^{M} n_i(y_i - x_i'\beta)^2 = \frac{1}{M}\sum_{i=1}^{M}(\sqrt{n_i}\epsilon_i)^2 = \sigma^2 + o_p(1)$$

where we have used the fact that $E[(\sqrt{n_i}\epsilon_i)^2|x_i] = \text{Var}(\sqrt{n_i}\epsilon_i|x_i) = \sigma^2$ for all $i = 1, \ldots, M$. Term (2) is $o_p(1)$ by Assumption 3 and consistency of $\hat{\beta}_N$. Term (3) is $o_p(1)$ by similar reasoning, since by the WLLN,

$$\frac{1}{M}\sum_{i=1}^{M} n_i \epsilon_i x_i = o_p(1)$$

Combining these results, the consistency of $\hat{\sigma}_N^2$ follows. □

3.4 Asymptotic Normality of Coefficient Estimator

The following theorem establishes the asymptotic normality of $\hat{\beta}_N$.

Theorem 4 *Consider the linear model $y_{ij} = x_{ij}'\beta + \epsilon_{ij}$ for $i = 1, \ldots, M$ and $j = 1, \ldots, n_i$, where $(\epsilon_{ij})_{(i,j)}$ are i.i.d. with $E[\epsilon_{ij}|x_{ij}] = 0$ and Var$[\epsilon_{ij}|x_{ij}] = \sigma^2$. Assume that the following conditions are satisfied:*

1. $\lim_{M\to\infty} s_i^x = 0$ *for all* $i = 1, \ldots, M$
2. $\max_{1\le i\le M} \{n_i\} < K$ *for some constant* $1 \le K < \infty$
3. $\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} n_i x_i x_i' = G$*, where* $0 < \|G\| < \infty$ *and* $|G| \neq 0$

Then, $\hat{\beta}_N$ has an asymptotic normal distribution; in particular,

$$\sqrt{M}(\hat{\beta}_N - \beta) \xrightarrow{\mathcal{L}} N\Big(0, G^{-1}\Big)$$

Furthermore, the asymptotic variance is the same whether $\hat{\beta}_N$ or $\hat{\beta}_{LS}$ is used.

Proof First, consider the quantity $\hat{\beta}_N - \beta$, for which

$$\sqrt{M}(\hat{\beta}_N - \beta) = \left(\frac{1}{M}\sum_{i=1}^{M} n_i x_i x_i'\right)^{-1}\left(\frac{1}{\sqrt{M}}\sum_{i=1}^{M} n_i x_i \epsilon_i\right)$$

Regarding the first term, by assumption we know that

$$\lim_{M\to\infty}\left(\frac{1}{M}\sum_{i=1}^{M} n_i x_i x_i'\right)^{-1} = \left(\lim_{M\to\infty}\frac{1}{M}\sum_{i=1}^{M} n_i x_i x_i'\right)^{-1} = G^{-1}$$

Regarding the second term, note that by the central limit theorem (CLT):

$$\frac{1}{\sqrt{M}}\sum_{i=1}^{M} n_i x_i \epsilon_i \xrightarrow{\mathcal{L}} N\Big(0, \sigma^2 G\Big)$$

Therefore, by Slutsky's theorem, we have

$$\sqrt{M}(\hat{\beta}_N - \beta) \xrightarrow{\mathcal{L}} N\Big(0, \sigma^2 G^{-1}\Big)$$

It remains to show that $\mathrm{Var}(\hat{\beta}_N) - \mathrm{Var}(\hat{\beta}_{LS}) = o_p(1)$. By the continuous mapping theorem, it suffices to show the following, where (1) was established by Theorem 3:

1. $\lim\limits_{M\to\infty} \hat{\sigma}_N^2 = \lim\limits_{M\to\infty} \hat{\sigma}_{LS}^2 = \sigma^2$
2. $\lim\limits_{M\to\infty} \left[\left(\sum\limits_{i=1}^{M} n_i x_i x_i' \right)^{-1} - \left(\sum\limits_{i=1}^{M} \sum\limits_{j=1}^{n_i} x_{ij} x_{ij}' \right)^{-1} \right] = 0$

Regarding statement (2), if we again define $\delta_{ij} = x_{ij} - x_i$, so $x_{ij} = x_i + \delta_{ij}$, then

$$\sum_{i=1}^{M}\sum_{j=1}^{n_i} x_{ij}x_{ij}' = \sum_{i=1}^{M}\left(n_i x_i x_i' + \sum_{j=1}^{n_i}\Big(x_i \delta_{ij}' + \delta_{ij} x_i' + \delta_{ij}\delta_{ij}' \Big) \right)$$

Since $\lim\limits_{M\to\infty} s_i^x = 0$, by our lemma, we have

$$\lim_{M\to\infty}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\Big(x_i \delta_{ij}' + \delta_{ij} x_i' + \delta_{ij}\delta_{ij}' \Big) = 0$$

In the limiting case, this implies that

$$\lim_{M\to\infty}\sum_{i=1}^{M}\sum_{j=1}^{n_i} x_{ij}x_{ij}' = \lim_{M\to\infty}\sum_{i=1}^{M} n_i x_i x_i'$$

and statement (2) follows. Thus, the asymptotic variances are the same. □

These results show that the nugget estimators have similar asymptotic properties to the full-data estimators. Finite-sample performance will be explored next.

4 Example

We present a sample analysis using data nuggets in linear regression. Our data consist of a few hundred thousand hospital patients with liver cancer from the national inpatient sample (NIS) database. We want to predict the length of a patient's hospital stay, in days, based on demographics, comorbidities, insurance, and some other relevant information.

A sample of 5000 data nuggets was formed from the original data consisting of over 200,000 observations, which is roughly a 98% reduction in the number of observations. Coefficient estimates and standard errors using the full data and 5000 nuggets in a model with ten parameters are compared. Table 1 displays the results using the untransformed response Y. Table 2 displays the results using the transformed response $\log(Y+1)$, which was decided after an exploratory analysis.

The results show that 5000 nuggets perform very well in estimating the parameters. The point estimates are very close to the least-squares estimate in the full data. The standard errors are inflated by roughly 30%; however, the number of data nuggets is only roughly 2% of the number of original observations, and a random sample would provide estimates with standard errors several times larger. For sufficiently large samples, since standard errors are inversely proportional to the square root of the sample size, the standard error using a random sample of size M from a dataset of size N is expected to be inflated by a multiple of $\sqrt{N/M}$; in this case, a random sample of $M = 5000$ from a dataset of size $N \gtrapprox 200{,}000$ is expected to provide standard errors approximately greater than or equal to $\sqrt{40} > 6$ times as large as when using the full data, whereas the standard errors using 5000 data nuggets did not even double.

Table 1 Regression for length of hospital stay, untransformed response

Term	Est-full	Est-nugget	SE-full	SE-nugget
(Intercept)	15.32693	15.32693	0.02409	0.03958
mxraceWTRUE	0.94200	0.93929	0.02468	0.04057
mxAGE	4.25668	4.29443	0.02940	0.05007
mxV3	1.47874	1.47388	0.02578	0.04290
mxpay1TRUE	3.94812	3.92996	0.02822	0.04675
mxZIPINC.QRTL2	−0.14651	−0.14620	0.02850	0.04682
mxZIPINC.QRTL3	−0.11484	−0.11469	0.02832	0.04652
mxZIPINC.QRTL4	0.09829	0.09857	0.02826	0.04642
mxZIPINC.QRTLA	0.02658	0.02655	0.02410	0.03960
mxhospdTRUE	1.29200	1.29103	0.02426	0.03986
mxV4	−0.51078	−0.51163	0.02420	0.03980

Table 2 Regression for length of hospital stay, transformed response

Term	Est-full	Est-nugget	SE-full	SE-nugget
(Intercept)	2.492691	2.492691	0.001125	0.001478
mxraceWTRUE	0.046703	0.046588	0.001152	0.001514
mxAGE	0.304493	0.305995	0.001372	0.001869
mxV3	0.072390	0.071835	0.001204	0.001601
mxpay1TRUE	0.148963	0.148325	0.001318	0.001745
mxZIPINC.QRTL2	−0.012482	−0.012477	0.001330	0.001748
mxZIPINC.QRTL3	−0.015472	−0.015478	0.001322	0.001737
mxZIPINC.QRTL4	−0.006428	−0.006432	0.001319	0.001733
mxZIPINC.QRTLA	0.002652	0.002650	0.001125	0.001478
mxhospdTRUE	0.084561	0.084516	0.001132	0.001488
mxV4	−0.029690	−0.029648	0.001130	0.001486

5 Simulations

We performed a simulation analysis to evaluate the performance of data nuggets in linear regression parameter estimation and prediction. The number of nuggets used and the number of features were varied to compare performance. Results using simple random samples are provided for comparison.

Consider the following homoscedastic linear regression model in which the response Y depends linearly on two covariates X_1 and X_2:

$$Y = 10 + 3X_1 + X_2 + \epsilon, \qquad \epsilon \sim N(0, 2^2)$$

Let $X_1, \ldots, X_P$ for $P \geq 2$ be the feature variables in our data, where P is specified for each simulation. Note that if $P > 2$, then the model coefficients are zero for $i \in \{3, \ldots, P\}$. All covariates are generated independently from standard normal distributions. The mean and variance of these distributions are irrelevant, since centering and scaling could always be performed.

We simulate $N = 10^5$, i.e., one hundred thousand, observations for $P \in \{2, 5, 20, 50\}$ feature variables $X_1, \ldots, X_P$ and continuous response Y according to the model. We present the results of simulations for both estimation and prediction.

5.1 First Simulation Set: Prediction

The first set of simulations evaluates the out-of-sample predictive performance of linear regression models fit using various numbers of data nuggets from the same original dataset. Model performance was evaluated using root-mean-square prediction error (RMSE) and mean absolute prediction error (MAE) measures under fivefold cross-validation. For each data split, a model is trained using data nuggets

Table 3 Data nugget simulation: prediction, two covariates

Random sample				
	RMSE		MAE	
N	AVG	SD	AVG	SD
50	1.9084	0.1905	1.6110	0.1605
100	1.9601	0.1328	1.6043	0.1319
500	2.0095	0.0773	1.6109	0.0668
1000	2.0011	0.0497	1.6020	0.0396
5000	1.9988	0.0183	1.5957	0.0163
10,000	1.9992	0.0151	1.5952	0.0149
1e+05	1.9996	0.0000	1.5950	0.0000

Data nuggets				
	RMSE		MAE	
M	AVG	SD	AVG	SD
50	1.9996	0.0098	1.5950	0.0060
100	1.9996	0.0098	1.5950	0.0060
500	1.9996	0.0098	1.5950	0.0060
1000	1.9996	0.0098	1.5950	0.0060
5000	1.9996	0.0098	1.5950	0.0060
10,000	1.9996	0.0098	1.5950	0.0060

Table 4 Data nugget simulation: prediction, five covariates

Random sample				
	RMSE		MAE	
N	AVG	SD	AVG	SD
50	2.0592	0.2554	1.7213	0.2139
100	2.0438	0.1603	1.6661	0.1406
500	1.9762	0.0678	1.5866	0.0583
1000	2.0067	0.0383	1.6037	0.0360
5000	1.9900	0.0213	1.5901	0.0172
10,000	1.9969	0.0109	1.5933	0.0095
1e+05	1.9938	0.0000	1.5908	0.0000

Data nuggets				
	RMSE		MAE	
M	AVG	SD	AVG	SD
50	1.9938	0.0178	1.5908	0.0174
100	1.9938	0.0178	1.5908	0.0173
500	1.9938	0.0178	1.5909	0.0173
1000	1.9938	0.0178	1.5909	0.0173
5000	1.9938	0.0178	1.5908	0.0174
10,000	1.9938	0.0178	1.5908	0.0173

generated from the training data, while the predictive measure is evaluated based on that training model and the testing data. The mean and standard deviation of trial results are reported to account for variation between runs. Results generated using data nuggets are a summary of five runs, while results generated using random samples are a summary of twenty-five runs.

Tables 3, 4, 5, and 6 display the prediction results for $P \in \{2, 5, 20, 50\}$. For each value of P, two tables are provided: the left table displaying results for random samples, including the full data, and the right table displaying results for data nuggets. For $P \in \{2, 5\}$, the average prediction error when using even a small number of nuggets is practically identical compared to using the entire dataset. For $P \in \{20, 50\}$, the average prediction error when using a very small number of nuggets is at least comparable to a random sample of $5\% - 10\%$ of the original data. Data nuggets also have the advantage of providing more consistent prediction performance, while random sampling results are generally much more variable, especially when the random sample is relatively small.

Table 5 Data nugget simulation: prediction, twenty covariates

Random sample				
	RMSE		MAE	
N	AVG	SD	AVG	SD
50	2.3915	0.4608	1.9913	0.3915
100	2.1791	0.1387	1.7937	0.1055
500	2.0305	0.0669	1.6301	0.0556
1000	2.0147	0.0435	1.6134	0.0356
5000	1.9997	0.0228	1.5966	0.0183
10,000	1.9937	0.0105	1.5896	0.0089
1e+05	1.9945	0.0000	1.5908	0.0000

Data nuggets				
	RMSE		MAE	
M	AVG	SD	AVG	SD
50	1.9957	0.0078	1.5919	0.0094
100	1.9953	0.0084	1.5915	0.0101
500	1.9950	0.0080	1.5912	0.0101
1000	1.9951	0.0079	1.5911	0.0100
5000	1.9948	0.0082	1.5911	0.0102
10,000	1.9949	0.0080	1.5912	0.0101

Table 6 Data nugget simulation: prediction, fifty covariates

Random sample				
	RMSE		MAE	
N	AVG	SD	AVG	SD
100	2.8849	0.4049	2.3625	0.3340
500	2.1146	0.0698	1.6927	0.0537
1000	2.0494	0.0402	1.6376	0.0296
5000	2.0108	0.0186	1.6022	0.0159
10,000	2.0057	0.0160	1.5987	0.0126
1e+05	1.9977	0.0000	1.5933	0.0000

Data nuggets				
	RMSE		MAE	
M	AVG	SD	AVG	SD
100	2.0061	0.0052	1.6001	0.0065
500	2.0002	0.0047	1.5950	0.0054
1000	2.0002	0.0037	1.5953	0.0051
5000	1.9995	0.0041	1.5947	0.0053
10,000	1.9986	0.0043	1.5939	0.0055

5.2 *Second Simulation Set: Estimation*

The second set of simulations evaluates the estimation performance in linear regression models fit using data nuggets. Estimation results for the two non-zero coefficients $\beta_1 = 3$ and $\beta_2 = 1$ were obtained. Only the results for β_1 are shown here, but similar trends hold for β_2. The results include the absolute estimation error for the parameter and standard error magnitude. The mean and standard deviation of trial results are reported to account for variation between runs. Results generated using data nuggets are a summary of five runs, while results generated using random samples are a summary of twenty-five runs.

Tables 7, 8, 9, and 10 display the estimation results for $P \in \{2, 5, 20, 50\}$. For each value of P, four tables are provided that compare the estimation of β_1 using data nuggets and random samples of varying size. For $P \in \{2, 5\}$, a very small number of nuggets performs at least as well as a random sample of size 10,000. For $P \in \{20, 50\}$, a small number of nuggets performs around as well as 5000–10,000 nuggets. Again, data nugget performance is generally much more consistent between runs compared to random sample performance.

Table 7 Data nugget simulation: estimation, two covariates

Random sample				
	B1 EST ERR		B1 STD ERR	
N	AVG	SD	AVG	SD
50	0.2461	0.1503	0.2919	0.0292
100	0.1326	0.1121	0.2112	0.0224
500	0.0517	0.0431	0.0890	0.0033
1000	0.0442	0.0309	0.0633	0.0021
5000	0.0213	0.0144	0.0284	0.0004
10,000	0.0112	0.0089	0.0200	0.0002
1e+05	0.0041	0.0000	0.0063	0.0000

Data nuggets				
	B1 EST ERR		B1 STD ERR	
M	AVG	SD	AVG	SD
50	0.0052	0.0022	0.0064	0.0006
100	0.0051	0.0003	0.0063	0.0004
500	0.0036	0.0006	0.0065	0.0002
1000	0.0039	0.0003	0.0064	0.0000
5000	0.0041	0.0001	0.0064	0.0000
10,000	0.0042	0.0001	0.0063	0.0000

Table 8 Data nugget simulation: estimation, five covariates

Random sample				
	B1 EST ERR		B1 STD ERR	
N	AVG	SD	AVG	SD
50	0.2445	0.1110	0.2985	0.0566
100	0.2083	0.1371	0.2031	0.0147
500	0.0489	0.0363	0.0893	0.0037
1000	0.0467	0.0367	0.0635	0.0019
5000	0.0197	0.0152	0.0283	0.0004
10,000	0.0160	0.0113	0.0200	0.0002
1e+05	0.0069	0.0000	0.0063	0.0000

Data nuggets				
	B1 EST ERR		B1 STD ERR	
M	AVG	SD	AVG	SD
50	0.0134	0.0047	0.0090	0.0005
100	0.0092	0.0039	0.0074	0.0008
500	0.0078	0.0017	0.0073	0.0002
1000	0.0070	0.0037	0.0069	0.0001
5000	0.0071	0.0031	0.0066	0.0001
10,000	0.0076	0.0007	0.0065	0.0000

Table 9 Data nugget simulation: estimation, twenty covariates

Random sample				
	B1 EST ERR		B1 STD ERR	
N	AVG	SD	AVG	SD
50	0.2892	0.2158	0.3734	0.0779
100	0.1574	0.1170	0.2194	0.0278
500	0.0695	0.0490	0.0923	0.0065
1000	0.0508	0.0354	0.0636	0.0022
5000	0.0172	0.0142	0.0284	0.0004
10,000	0.0169	0.0116	0.0200	0.0002
1e+05	0.0020	0.0000	0.0063	0.0000

Data nuggets				
	B1 EST ERR		B1 STD ERR	
M	AVG	SD	AVG	SD
50	0.0098	0.0103	0.0139	0.0027
100	0.0040	0.0038	0.0135	0.0024
500	0.0069	0.0033	0.0105	0.0004
1000	0.0077	0.0068	0.0101	0.0001
5000	0.0029	0.0016	0.0087	0.0001
10,000	0.0024	0.0014	0.0082	0.0001

Table 10 Data nugget simulation: estimation, fifty covariates

Random sample				
	B1 EST ERR		B1 STD ERR	
N	AVG	SD	AVG	SD
100	0.1784	0.1516	0.2782	0.0324
500	0.0630	0.0511	0.0934	0.0044
1000	0.0467	0.0371	0.0646	0.0018
5000	0.0287	0.0166	0.0284	0.0004
10,000	0.0112	0.0126	0.0201	0.0002
1e+05	0.0036	0.0000	0.0063	0.0000

Data nuggets				
	B1 EST ERR		B1 STD ERR	
M	AVG	SD	AVG	SD
100	0.0194	0.0156	0.0265	0.0042
500	0.0113	0.0035	0.0157	0.0008
1000	0.0107	0.0055	0.0144	0.0003
5000	0.0086	0.0059	0.0117	0.0002
10,000	0.0051	0.0040	0.0105	0.0000

6 Extensions

The focus of this chapter has been on the utilization of data nuggets methodology to predict a continuous response using a linear regression model in big data. There are many additional scenarios that can be considered. These include different classes of response; for example, a binary or categorical response, for which a logistic or multinomial regression model could be used. There are also so-called machine learning models, such as decision trees and ensembles, which are often used for prediction in big data.

Consider data consisting of N observations of P covariates $X_1, \ldots, X_P$ and a response Y, where the class or nature of Y is currently unspecified. We want to form M data nuggets, $M << N$, in a similar fashion to the methodology showcased in the case of linear regression with a continuous response. The partitioning process will remain the same, but the summarization of the within-nugget response information will depend on the specific scenario. The estimation procedure will depend on the model. We explore a few additional scenarios.

Suppose the response Y is binary with two possible outcomes, where we will assume $Y \in \{0, 1\}$ with $Y = 1$ termed a "success" and $Y = 0$ a "failure." The within-nugget response can be summarized by either the number of successes or proportion of success within the nugget, both of which are equivalent since the nugget weight is already stored. If response variability is an issue, the nuggets can be further split such that there is no response variability within each nugget. Regarding estimation, a binomial regression model using the compressed data would be appropriate, since the sum of independent Bernoulli random variables has a binomial distribution.

As a generalization of the previous case, suppose the response Y is categorical with at least two possible outcomes, where we will assume $Y \in \{0, 1, \ldots, K - 1\}$ with $Y = 0$ as a reference category. The within-nugget response can be summarized by a vector of length $K - 1$ indicating the number of occurrences or proportion of occurrence of each outcome aside from the reference category, which can be recovered with the nugget weight. A multinomial regression model would be appropriate.

An extension to survival data is important to consider but difficult to visualize. At worst, the entirety of the within-nugget response information can be stored. For small nuggets, this is highly recommended. For large nuggets, it is reasonable to consider summarizing the within-nugget response by some model, such as a simple Weibull model if the fit is appropriate.

For non-classical predictive modeling methods such as decision trees and ensembles, similar procedures can be used. A continuous or binary response can be summarized by a mean or proportion. Quantities such as the likelihood can be approximated by using the nugget weights and possibly within-nugget variability estimates. Since these methods often rely more heavily on local behavior, a reasonably large number of nuggets (appropriately scaled with the data dimension) are recommended.

7 Discussion

We have discussed the extension of Beavers' mathematically informed, compression-based "data nugget" methodology to supervised learning, using linear regression as a motivating example but discussing other models as well. The examples and simulations presented show that this methodology provides a significant improvement over random sampling in model parameter point and interval estimation, as well as out-of-sample prediction performance, in the linear regression setting. The methodology is promising for simplifying computation in large-scale predictive data analysis.

We are currently developing an R package which implements this methodology. This includes functions to form data nuggets in the presence of a response and perform statistical inference with nuggets using statistical modeling. The package calls upon the *datanugget* package by Beavers et al. to partition the data into sets, which will become data nuggets in the supervised setting (Beavers et al. 2020). We are also working on a related Shiny application to demonstrate the methodology.

In the future, we would like to investigate the performance of data nuggets in other settings that were discussed, including logistic regression, modeling of time-to-event data, and non-classical predictive modeling. This includes a rigorous simulation analysis for point estimation, confidence interval estimation, and out-of-sample prediction, as well as the establishment of theoretical asymptotic guarantees. We believe that non-classical supervised methods (Amaratunga et al. 2014) such as SVM, random forest, boosting, and deep learning would benefit most from data nuggets due to the significantly greater computational complexity compared to least-squares regression, which is the most economical computationally among supervised methods. Another potential application is functional data analysis (Boente et al. 2014). There will always be finite computing power, and direct analysis of big data using these supervised methods may not be possible, which our methodology attempts to address.

Regarding the development of asymptotics, we have assumed that the sample size tended to infinity. Tukey (Fernholz and Morgenthaler 2003) used to say that if the population distribution is known, then asymptotics are not necessary. While we do not often know the exact distribution of the population, a big data sample should provide a close approximation. Even if the exact population distribution is known, statistical computations on the full data may not be possible. Thus, in the future, we are interested in exploring asymptotics as the number of data nuggets converges to a known large finite population size.

Another vital area of future research is a comprehensive study of factors related to the compression or estimation that may affect quality of interest. For example, more nuggets are required to maintain similar quality of inference as the data dimensionality increases, likely due to the difficulty of distance computation and clustering in higher dimensions. Other factors such as correlation structure may play a role in the quality of compression and inference. This would require further simulation work.

8 Conclusion

We have presented an extension of Beavers' "data nugget" methodology for data reduction and compression to the field of supervised learning with a focus on linear regression with a continuous response. We established theoretical asymptotic guarantees in terms of consistency and asymptotic normality such that our coefficient estimator has the same asymptotic distribution as the ordinary least-square estimator in the big data setting as the number of nuggets grows toward the sample size, which demonstrates that these estimators perform as expected asymptotically. For finite-sample performance, several simulation studies were conducted to show that a small number of nuggets provide excellent prediction performance and standard errors many times smaller than random samples of comparable size, closely approximating the standard errors in the full data case in many scenarios. Thus, data nuggets provide a significant decrease in the size of the data without significantly increasing the uncertainty in estimates. There are many opportunities for future research and investigation in this area, including further development of asymptotics and extensions to other statistical models, including classic regression models for different response classes and even predictive machine learning models. The development here provides a solid basis for handling big data in statistical modeling, which we believe in general is one of the most pressing issues in applied statistical data analysis today.

References

Amaratunga, D., Cabrera, J., & Shkedy, Z. (2014). *Exploration and analysis of DNA microarray and other high-dimensional data*. Wiley.

Beavers, T., Cabrera, J., & Lubomirski, M. (2020). datanugget: Create, Refine, and Cluster Data Nuggets. R package version 1.2.0.

Beavers, T., et al. (2020). Data Nuggets: A Method for Reducing Large Datasets While Maintaining Data Structure. Under revision.

Boente, G., Barrera, M. S., & Tyler, D. E. (2014). A characterization of elliptical distributions and some optimality properties of principal components for functional data. *Journal of Multivariate Analysis, 131*, 254–264.

Cherasia, K. E. (2021). Computational Methods for Data Analysis in Clinical Studies. PhD thesis. Rutgers University.

DuMouchel, W. (2002). Data squashing: Constructing summary data sets. *Massive Computing*, 579–591.

DuMouchel, W., et al. (1999). Squashing flat files flatter. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Fernholz, L. T., & Morgenthaler, S. (2003). A conversation with John Tukey. *Statistical Science, 18*(3).

Flury, B. A. (1990). Principal points. *Biometrika, 77*(1), 33–41.

Flury, B. (1997). A first course in multivariate statistics. In *Springer texts in statistics*.

Madigan, D., et al. (2001). Instance construction via likelihood-based data squashing. In *Instance Selection and Construction for Data Mining* (pp. 209–226).

Mak, S., & Joseph, R. V. (2018). Support points. *The Annals of Statistics, 46*(6A), 2562–2592.

Owen, A. (1999). Data squashing by empirical likelihood. *Data Mining and Knowledge Discovery, 7*, 101–113.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing, 2*(4), 183–190.

Zhang, Y., & Zhao, Y. (2015). Astronomy in the big data era. *Data Science Journal, 14*, 11.

Improved Convergence Rates of Normal Extremes

Yijun Zhu and Han Xiao

Abstract It is well known that the convergence of the normal extremes to the limiting Gumbel distribution is extremely slow, at the rate of $(\log n)^{-1}$. We show that after a monotone transform, the convergence rate of the squared normal extremes can be improved to $(\log n)^{-3}$. Simulations confirm that the convergence is much faster than existing results uniformly, especially when the sample is of moderate sizes around hundreds or thousands. More importantly, it is observed that the convergence rate at the upper tail is substantially improved, which has direct implications for hypothesis tests based on the maximum type test statistics.

Keywords Extreme value theory · Gumbel distribution · Normal distribution · Rate of convergence

1 Introduction

Let $X_1, X_2, \ldots$ be a sequence of independent standard normal random variables, and let $M_n := \max\{X_1, X_2, \ldots, X_n\}$ be the maximum of the first n of them. According to the extreme value theory (see Leadbetter et al. 1983, for an overview), after proper centering and rescaling, the limiting distribution of M_n is the extreme value distribution of type I, or the so-called Gumbel distribution, with the distribution function $G_1(x) = \exp(-e^{-x})$. In fact, if we define

$$\alpha_n = (2\log n)^{-1/2}$$

$$\beta_n = \sqrt{2\log n} - \frac{\log(\log n) + \log(4\pi)}{2\sqrt{2\log n}},$$

Y. Zhu · H. Xiao (✉)
Rutgers University, Piscataway, NJ, USA
e-mail: yijunzhu@stat.rutgers.edu; hxiao@stat.rutgers.edu

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_21

then $\alpha_n^{-1}(M_n - \beta_n)$ converges to G_1 in distribution, i.e.,

$$\lim_{n\to\infty} P\left[\alpha_n^{-1}(M_n - \beta_n) \leq x\right] = \lim_{n\to\infty} \Phi^n(\alpha_n x + \beta_n) = \exp\left(-e^{-x}\right), \quad x \in \mathbb{R}, \tag{1}$$

where $\Phi(\cdot)$ is the distribution function of $N(0, 1)$.

The rate of convergence in (1) is extremely slow. The fact was noted by Fisher and Tippett (1928) and studied more precisely by Hall (1979), who proved that the convergence rate in (1) is no better than $(\log\log n)^2/\log n$. Hall (1979) also found that if β_{n1} is the solution of the equation

$$2\pi\beta_{n1}^2 \exp(\beta_{n1}^2) = n^2 \tag{2}$$

and $\alpha_{n1} = \beta_{n1}^{-1}$, then

$$\frac{C_1}{\log n} < \sup_{-\infty<x<\infty}\left|P[\alpha_{n1}^{-1}(M_n - \beta_{n1}) \leq x] - G_1(x)\right| < \frac{C_2}{\log n}, \tag{3}$$

where C_1 and C_2 are absolute constants. In other words, the convergence rate can be improved to $(\log n)^{-1}$ by choosing a better centering constant β_{n1}. In the same paper, it was further proved that the rate cannot be better than $(\log n)^{-1}$ by choosing a different sequence of normalizing constants.

It is equivalent and sometimes more convenient to study the limiting behavior of M_n through its squared version M_n^2. There are counterparts of (1) and (3) for M_n^2. More importantly, Hall (1980) found that with suitably chosen constants a_n and b_n, the normalized sequence $a_n^{-1}(M_n^2 - b_n)$ converges to $G_1(x)$ with the rate $(\log n)^{-2}$. A detailed overview of the progression regarding the convergence rates of normal extremes will be provided in Sect. 2.3 via the squared version M_n^2.

While the aforementioned results are all on the uniform convergence rates, the convergence to G_1 in the upper tail is of particular interests when performing hypothesis tests using maximum type statistics. For example, the stepdown procedure of Romano and Wolf (2005) for multiple testing requires the knowledge about the upper quantiles of the maximum test statistic. Cai et al. (2014) used the maximum coordinate-wise difference of two transformed sample mean vectors to test the equality of two high-dimensional means.

In Fig. 1, we plot the empirical distributions of M_n^2 with different choices of normalizing sequences. The black line is the theoretical cumulative distribution function (CDF) G_1, the dashed, red, and green lines (labeled by b_{n1}, b_{n2}, and b_{n3}, *respectively*) are empirical CDF corresponding to convergence rates in (1), (3), and $(\log n)^{-2}$, respectively. Figure 2 zooms in on the upper tails. Despite the fact that the red line is associated with a faster convergence rate than that of the dashed one, Fig. 2 shows that it is consistently farther from the theoretical CDF in the upper tail, even when the sample size is as large as 10^5. This need not contradict the theories on the uniform convergence rates because we see in Fig. 1 that the dashed line deviates apparently from the black one in the lower tail. However, tests based on the statistic in (3) will be quite off and have no advantage over the statistic in (1). On the other hand, the green line, corresponding to the rate $(\log n)^{-2}$, shows the potential to outperform the dashed one, when the sample size is sufficiently large, as shown in the bottom right panel of Fig. 2. The issue is that the green line is below the theoretical CDF, indicating that the corresponding asymptotic test is not conservative.

Our major finding is that the convergence rate can be further improved to $(\log n)^{-3}$ by applying a monotone transform to M_n^2. Let $b_n := \frac{1}{2}[\Phi^{-1}(1 - 1/n)]^2$. Define Y_n through the following transform of M_n^2:

$$Y_n := \left[1 - \left(1 + \frac{M_n^2 - 2b_n}{8b_n^2}\right)^{-1}\right]\left(4b_n^2 + 2b_n - 2\right).$$

The results in Sect. 2 imply the following rate of convergence:

$$\sup_{-\infty<x<\infty} |P(Y_n \leq x) - G_1(x)| < \frac{C_3}{(\log n)^3}.$$

The blue lines in Fig. 1 give empirical CDF of Y_n, which are almost identical with G_1 even when the sample size is as small as 200. When zoomed into the upper tail in Fig. 2, the faster convergence of Y_n is more clearly seen. Furthermore, if Y_n is used as the test statistic for the asymptotic test, it is not only more accurate but also always conservative, since the blue curve sits above the black one (for G_1) in the upper tail.

The rest of this chapter is organized as follows. We present and prove the pointwise and uniform convergence rates of Y_n in Sects. 2.1 and 2.2, respectively. In Sect. 2.3, we demonstrate how the faster convergence rate is achieved by comparing with existing results. Similar convergence rates regarding the k-th maxima are presented in Sect. 2.4. Numerical analysis and an application on testing the covariance structure are given in Sect. 3. Additional figures, tables, and some technical results are relegated in the Appendix.

We conclude this section by a brief review of the literature on the convergence rates of normal extremes. Cohen (1982b) showed that the penultimate approximation can achieve the $(\log n)^{-2}$ rate and considered the extension to other types of extreme value distributions in Cohen (1982a). Daniels (1982) proposed another nonlinear transformation which leads to faster convergence. Rootzén (1983) investigated the convergence rates of the extremes from a stationary Gaussian process. Hall (1991) found that the extreme of a continuous time Gaussian process also has a logarithmic convergence rate. For convergence rates of extremes from a non-Gaussian sequence, we refer to Hall and Wellner (1979), Smith (1982), Leadbetter et al. (1983), de Haan and Resnick (1996), Peng et al. (2010) and the references therein.

2 Main Results

We will first consider the pointwise convergence rates in Sect. 2.1 and then illustrate how the faster rates are achieved by modifying the normalizing constants and applying a transform of M_n^2 in Sect. 2.3. The uniform convergence rates are given in Sect. 2.2. In Sect. 2.4, we present the corresponding results for the k-th maxima. We make the convention that $C, C_1, C_2, \ldots$ are generic absolute constants, whose values may vary from place to place.

2.1 Pointwise Convergence Rates

Let b_n be the solution of the equation $1 - \Phi(\sqrt{2b_n}) = 1/n$. Recall that Y_n is defined as

$$Y_n := \left[1 - \left(1 + \frac{M_n^2 - 2b_n}{8b_n^2}\right)^{-1}\right]\left(4b_n^2 + 2b_n - 2\right). \tag{4}$$

According to the definition, $\sqrt{2b_n}$ is the $(1-1/n)$-th quantile of the standard normal distribution. Since $M_n^2 \geq 0$ and $b_5 \approx .35$, the transform given in (4) is strictly monotone when $n \geq 5$, which we shall assume in the sequel.

Using the Newton–Raphson approximation (see Appendix 4 for detailed derivations), it can be shown that

$$b_n = \log n - \tfrac{1}{2}\log\log n - \tfrac{1}{2}\log 4\pi + O(\log\log n/\log n).$$

We first prove the pointwise convergence rate of Y_n to G_1. It is convenient to express the result through b_n, which is of the order $\log n$.

Theorem 1 *For each fixed* $-\infty < x < \infty$*,*

$$P(Y_n \leq x) - G_1(x) = G_1(x)e^{-x} \cdot \frac{4x^3 + 15x^2 + 30x}{24b_n^3} + O(b_n^{-4}).$$

Proof Define the function $g_n(x)$ as the inverse transform of (4)

$$g_n(x) = \left[\left(1 - \frac{x}{4b_n^2 + 2b_n - 2}\right)^{-1} - 1\right] \cdot 8b_n^2 + 2b_n. \tag{5}$$

Since (4) is a monotone transform, the event $[Y_n \leq x]$ is equivalent to $[M_n^2 \leq g_n(x)]$. It can be shown that

$$g_n(x) = 2b_n + 2x - \frac{x}{b_n} + \frac{x^2 + 3x}{2b_n^2} - \frac{2x^2 + 5x}{4b_n^3} + O(b_n^{-4}). \tag{6}$$

When n is large enough, $g_n(x) > 0$, and we let $x_n = [g_n(x)]^{1/2}$. Note that

$$P(M_n \leq x_n) > P(M_n^2 \leq x_n^2) = P(M_n \leq x_n) - P(M_n < -x_n) > P(M_n \leq x_n) - 2^{-n}. \tag{7}$$

According to Lemma 2.4.1 in Leadbetter et al. (1983), for any $0 \leq z \leq n$,

$$0 \leq e^{-z} - \left(1 - \frac{z}{n}\right)^n \leq \frac{z^2 e^{-z}}{2} \cdot \frac{1}{n-1}. \tag{8}$$

Let $\tau_n(x) = n[1 - \Phi(x_n)]$, and it follows that

$$P(M_n \leq x_n) = [1 - (1 - \Phi(x_n))]^n = \exp[-\tau_n(x)] + O(n^{-1}). \tag{9}$$

To evaluate $\tau_n(x)$, we make use of the following series expansion of the normal tail probability (Abramowitz and Stegun 1964): for any $z > 0$ and any positive integer m,

$$1 - \Phi(z) = \frac{\phi(z)}{z}\left\{1 - \frac{1}{z^2} + \frac{1 \cdot 3}{z^4} + \cdots + \frac{(-1)^m 1 \cdot 3 \ldots (2n-1)}{z^{2m}} + R_m\right\},$$

where

$$R_m = (-1)^{m+1}(2m+1)!! \int_z^\infty \frac{\phi(t)}{t^{2m+2}} dt,$$

which is less in absolute value than the first neglected term. In particular, when $m = 3$, it holds that for any $z > 0$,

$$\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5} - \frac{15}{z^7}\right)\phi(z) < 1 - \Phi(z) < \left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5} - \frac{15}{z^7} + \frac{105}{z^9}\right)\phi(z). \tag{10}$$

According to the definition of $\tau_n(x)$ and (10), we first do the Taylor expansion (up to the order b_n^{-4}) for

$$\begin{aligned}\phi(x_n) &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-b_n - x + \frac{x}{2b_n} - \frac{x^2+3x}{4b_n^2} + \frac{2x^2+5x}{8b_n^3}\right)\\ &= \frac{e^{-x}e^{-b_n}}{\sqrt{2\pi}} \cdot \left(1 + \frac{x}{2b_n} - \frac{x^2+6x}{8b_n^2} - \frac{5x^3+6x^2-30x}{48b_n^3} + O(b_n^{-4})\right),\end{aligned}$$

and

$$\begin{aligned}\frac{1}{x_n} &= \left(2b_n + 2x - \frac{x}{b_n} + \frac{x^2+3x}{2b_n^2} - \frac{2x^2+5x}{4b_n^3}\right)^{-1/2}\\ &= \frac{1}{\sqrt{2b_n}}\left(1 - \frac{x}{2b_n} + \frac{3x^2+2x}{8b_n^2} - \frac{5x^3+8x^2+6x}{16b_n^3} + O(b_n^{-4})\right).\end{aligned}$$

Combining the two preceding equations and rearranging the terms, we have

$$\frac{\phi(x_n)}{x_n} = \frac{e^{-x}e^{-b_n}}{\sqrt{4\pi b_n}} \cdot \left(1 - \frac{x}{2b_n^2} - \frac{4x^3+3x^2-6x}{24b_n^3} + O(b_n^{-4})\right).$$

According to (10), we also calculate

$$\begin{aligned}1 - \frac{1}{x_n^2} + \frac{3}{x_n^4} - \frac{15}{x_n^6} &= 1 - \frac{1}{2b_n} + \frac{2x+3}{4b_n^2} - \frac{4x^2+14x+15}{8b_n^3} + O(b_n^{-4})\\ &= \left(1 - \frac{1}{2b_n} + \frac{3}{4b_n^2} - \frac{15}{8b_n^3}\right) \cdot \left(1 + \frac{x}{2b_n^2} - \frac{x^2+3x}{2b_n^3} + O(b_n^{-4})\right).\end{aligned}$$

Recall that b_n is the solution of the equation $1 - \Phi(\sqrt{2b_n}) = 1/n$. According to the approximation to normal probability function in (10), we have

$$\frac{ne^{-b_n}}{\sqrt{4\pi b_n}} = \left(1 - \frac{1}{2b_n} + \frac{3}{4b_n^2} - \frac{15}{8b_n^3} + O(b_n^{-4})\right)^{-1}. \tag{11}$$

Therefore,

$$\begin{aligned}&\left(1 - \frac{1}{x_n^2} + \frac{3}{x_n^4} - \frac{15}{x_n^6}\right)\frac{n\phi(x_n)}{x_n}\\ &\quad = e^{-x} \cdot \left(1 - \frac{x}{2b_n^2} - \frac{4x^3+3x^2-6x}{24b_n^3} + O(b_n^{-4})\right)\\ &\quad\quad \cdot \left(1 - \frac{1}{2b_n} + \frac{3}{4b_n^2} - \frac{15}{8b_n^3} + O(b_n^{-4})\right)^{-1}\end{aligned}$$

$$\cdot\left(1-\frac{1}{2b_n}+\frac{3}{4b_n^2}-\frac{15}{8b_n^3}\right)\cdot\left(1+\frac{x}{2b_n^2}-\frac{x^2+3x}{2b_n^3}+O(b_n^{-4})\right)$$

$$=e^{-x}\left(1-\frac{4x^3+15x^2+30x}{24b_n^3}+O(b_n^{-4})\right).$$

Since $n\phi(x_n)/x_n^9=O(b_n^{-4})$, we have by (10)

$$\tau_n(x)=e^{-x}\left(1-\frac{4x^3+15x^2+30x}{24b_n^3}\right)+O(b_n^{-4}).$$

According to (9), it follows that

$$\begin{aligned}P(Y_n\le x)-G_1(x)&=\exp(-\tau_n(x))+O(n^{-1})-G_1(x)\\&=G_1(x)e^{-x}\cdot\frac{4x^3+15x^2+30x}{24b_n^3}+O(b_n^{-4}).\end{aligned}$$

The proof is complete. □

Using (10) and the Newton–Raphson method, we have the following expansions for b_n:

$$b_n=\log n-\frac{\Delta}{2}+\frac{\Delta-2}{4\log n}+\frac{\Delta^2-6\Delta+14}{16(\log n)^2}+O\left(\frac{(\log\log n)^3}{(2\log n)^3}\right),\tag{12}$$

where

$$\Delta=\log\log n+\log 4\pi.$$

Therefore, Theorem 1 implies that Y_n converges to G_1 with the rate $(\log n)^{-3}$. The detailed derivation of (12) is given in the Appendix.

2.2 Uniform Convergence Rate

In this section, we establish the uniform convergence rate.

Theorem 2 *There exists an absolute constant c_1, such that*

$$\sup_{-\infty<x<\infty}|P(Y_n\le x)-G_1(x)|<\frac{c_1}{(\log n)^3}.$$

We prove Theorem 2 using two lemmas. Recall that $g_n(x)$, defined in (5), is the inverse transform of (4).

Lemma 1 *Let* $\{c_n\}$ *be an increasing sequence of positive integers such that* $c_n^4/b_n \to 0$*, and then*

$$g_n(x) = 2b_n + 2x - \frac{x}{b_n} + \frac{x^2+3x}{2b_n^2} - \frac{2x^2+5x}{4b_n^3} + \frac{d_{1n}(x)}{b_n^3},$$

where $\lim_{n\to\infty} \sup_{-c_n \le x \le c_n} |d_{1n}(x)| = 0$.

Proof According to (5), for $-c_n \le x \le c_n$, we can obtain the following expansion:

$$\begin{aligned} g_n(x) &= 2b_n + 8b_n^2 \cdot \left[\left(1 - \frac{x}{4b_n^2+2b_n+2}\right)^{-1} - 1\right] \\ &= 2b_n + 2x\gamma_n + \frac{x^2\gamma_n^2}{2b_n^2} + \frac{x^3\gamma_n^3}{8b_n^4} \cdot \left(1 - \frac{x\gamma_n}{4b_n^2}\right)^{-1}, \end{aligned} \tag{13}$$

where

$$\gamma_n = \left(1 + \frac{1}{2b_n} - \frac{1}{2b_n^2}\right)^{-1}.$$

When $n \ge 13$, $b_n > 1$, by series expansion of γ_n, we have

$$\begin{aligned} \gamma_n &= 1 - \frac{1}{2b_n} + \frac{3}{4b_n^2} - \frac{5}{8b_n^3} + \frac{e_{1n}}{b_n^4} \\ \gamma_n^2 &= 1 - \frac{1}{b_n} + \frac{e_{2n}}{b_n^2} \\ \gamma_n^3 \left(1 - \frac{x\gamma_n}{4b_n^2}\right)^{-1} &= 1 + e_{3n}. \end{aligned}$$

The following bounds can be easily verified: $|e_{1n}| \le 1$, $|e_{2n}| \le 2$, and $|e_{3n}| \le 1$. Then, by simplifying (13), we have

$$g_n(x) = 2b_n + 2x - \frac{x}{b_n} + \frac{x^2+3x}{2b_n^2} - \frac{2x^2+5x}{4b_n^3} + \frac{16xe_{1n} + 4x^2e_{2n} + x^3(1+e_{3n})}{8b_n^4}.$$

The proof is completed by noting that

$$\sup_{-c_n \le x \le c_n} \left| \frac{16xe_{1n} + 4x^2e_{2n} + x^3(1+e_{3n})}{8b_n} \right| \le \frac{8c_n + 4c_n^2 + c_n^3}{4b_n} \to 0$$

under the condition $c_n^4/b_n \to 0$.

□

Lemma 2 *Let $\{c_n\}$ be the same sequence as used in Lemma 1, and then*

$$\tau_n(x) = e^{-x}\left(1 - \frac{4x^3 + 15x^2 + 30x}{24b_n^3} + \frac{d_{2n}(x)}{b_n^3}\right),$$

where $\lim_{n\to\infty} \sup_{-c_n \le x \le c_n} |d_{2n}(x)| = 0$ *for all* $-c_n \le x \le c_n$.

Proof Recall that $x_n := [g_n(x)]^{1/2}$. Using the normal tail probability bound in (10), we have

$$\left|\tau_n(x) - n\phi(x_n)\left(\frac{1}{x_n} - \frac{1}{x_n^3} + \frac{3}{x_n^5} - \frac{15}{x_n^7}\right)\right| \le \frac{105n\phi(x_n)}{x_n^9}. \tag{14}$$

Write

$$n\phi(x_n)\left(\frac{1}{x_n} - \frac{1}{x_n^3} + \frac{3}{x_n^5} - \frac{15}{x_n^7}\right) = \left(\frac{x_n}{\sqrt{2b_n}}\right)^{-1} \cdot \frac{n\phi(x_n)}{\sqrt{2b_n}} \cdot \left(1 - \frac{1}{x_n^2} + \frac{3}{x_n^4} - \frac{15}{x_n^6}\right). \tag{15}$$

Let

$$x_{1n} := \frac{x}{b_n} - \frac{x}{2b_n^2} + \frac{x^2 + 3x}{4b_n^3} - \frac{2x^2 + 5x}{8b_n^4} + \frac{d_{1n}(x)}{2b_n^4},$$

where $d_{1n}(x)$ is defined in Lemma 1. For the first term on the right-hand side of (15), by Lemma 1,

$$\left(\frac{x_n}{\sqrt{2b_n}}\right)^{-1} = (1 + x_{1n})^{-1/2} = 1 - \frac{x_{1n}}{2} + \frac{3x_{1n}}{8} - \frac{5x_{1n}^3}{16} + R_{1n}(x_{1n}). \tag{16}$$

Under the condition $c_n^4/b_n \to 0$, it holds that $\sup_{-c_n \le x \le c_n} |x_{1n}| \le 5c_n/b_n$, and thus

$$\sup_{-c_n \le x \le c_n} |R_{1n}(x)| = \frac{o(1)}{b_n^3}.$$

The terms on the right-hand side of (16) except for $R_{1n}(x_{1n})$ can be expanded as

$$\left(\frac{x_n}{\sqrt{2b_n}}\right)^{-1} - R_{1n}(x_{1n}) = 1 - \frac{x}{2b_n} + \frac{3x^2 + 2x}{8b_n^2} - \frac{5x^3 + 8x^2 + 6x}{16b_n^3} + \frac{d_{3n}(x)}{b_n^3}.$$

Note that for each fractional term in x_{1n}, the power of x is no greater than that of b_n, and the same claim holds for the series $d_{3n}(x)/b_n^3$. Furthermore, the first term (of the smallest power of x) in the expansion of $d_{3n}(x)$ is x^3/b_n, which goes to 0 uniformly over $-c_n \le x \le c_n$. Therefore, we conclude

$$\lim_{n\to\infty} \sup_{-c_n \le x \le c_n} |d_{3n}(x)| = 0.$$

The other two terms in (15) can be treated similarly:

$$\frac{n\phi(x_n)}{\sqrt{2b_n}} = \frac{ne^{-x}e^{-b_n}}{\sqrt{4\pi b_n}} \cdot \left(1 + \frac{x}{2b_n} - \frac{x^2+6x}{8b_n^2} - \frac{5x^3+6x^2-30x}{48b_n^3} + \frac{d_{4n}(x)}{b_n^3} + R_{2n}(x)\right),$$

$$1 - \frac{1}{x_n^2} + \frac{3}{x_n^4} - \frac{15}{x_n^6} = \left(1 - \frac{1}{2b_n} + \frac{3}{4b_n^2} - \frac{15}{8b_n^3}\right)$$
$$\cdot \left(1 + \frac{x}{2b_n^2} - \frac{x^2+3x}{2b_n^3} + \frac{d_{5n}(x)}{b_n^3} + R_{3n}(x)\right),$$

where

$$\sup_{-c_n \le x \le c_n} |d_{4n}(x)| \to 0 \quad \text{and} \quad |R_{2n}(x)| = \frac{o(1)}{b_n^3},$$
$$\sup_{-c_n \le x \le c_n} |d_{5n}(x)| \to 0 \quad \text{and} \quad |R_{3n}(x)| = \frac{o(1)}{b_n^3}.$$

Combining all the preceding bounds together with (11), we have

$$n\phi(x_n)\left(\frac{1}{x_n} - \frac{1}{x_n^3} + \frac{3}{x_n^5} - \frac{15}{x_n^7}\right) = e^{-x}\left(1 - \frac{4x^3+15x^2+30x}{24b_n^3} + \frac{d_{6n}(x)}{b_n^3}\right).$$

Using similar arguments as those for d_{3n}, we can verify that

$$\lim_{n\to\infty} \sup_{-c_n \le x \le c_n} |d_{6n}(x)| = 0.$$

It is easy to show that $\sup_{-c_n \le x \le c_n} n\phi(x_n)/x_n^9 = o(b_n^{-3})$. So, the proof is complete in view of (14).

□

We are now ready to prove Theorem 2.

Proof (Proof of Theorem 2) Let c_1 be a generic absolute constant which may vary from place to place. We consider three scenarios: $x < -c_n$, $-c_n \le x \le c_n$, and $x > c_n$, with $c_n = 4\log b_n$. Obviously, this choice of c_n satisfies the condition $c_n^4/b_n \to 0$.

We begin with the situation $-c_n \le x \le c_n$. By (7), it holds that

$$\left| P(Y_n \le x) - \left(1 - \frac{\tau_n(x)}{n}\right)^n \right| \le 2^{-n}.$$

By (8) and Lemma 2, we have

$$|P(Y_n \le x) - G_1(x)| \le 2G_1(x)e^{-x}\left(\frac{|4x^3 + 15x^2 + 30x|}{24b_n^3} + \frac{|d_{2n}(x)|}{b_n^3}\right) + \frac{1}{2^n} + \frac{1}{n},$$

when n is large enough. Since $\sup_{-c_n \le x \le c_n} |d_{2n}(x)| \to 0$, it suffices to show that

$$\sup_{-c_n \le x \le c_n} \left| G_1(x)e^{-x}(4x^3 + 15x^2 + 30x) \right| < \infty.$$

Numerical evaluations show that

$$\sup_{-\infty < x < \infty} \left| G_1(x)e^{-x}(4x^3 + 15x^2 + 30x) \right| < 20.$$

Therefore, we have

$$\sup_{-c_n < x < c_n} |P(Y_n \le x) - G_1(x)| < \frac{c_1}{(\log n)^3}.$$

Now, we consider the second scenario $x > c_n$. We will show that both $G_1(x)$ and $P(Y_n \le x)$ are close to 1, and their differences from 1 are of the order $1/(\log n)^3$. Since $x > c_n = 4 \log b_n$,

$$G_1(x) = \exp(-e^{-x}) > \exp(-b_n^4) \ge 1 - 1/b_n^4. \tag{17}$$

On the other hand, recall the definition of $g(\cdot)$ in (5)

$$\begin{aligned} 1 - P(Y_n \le x) \le P(Y_n \ge 4 \log b_n) &= P\left[M_n^2 \ge g(4 \log b_n)\right] \\ &\le P\left(M_n^2 \ge 2b_n + 4 \log b_n \cdot \frac{8b_n^2}{4b_n^2 + 2b_n - 2}\right). \end{aligned}$$

Note that $8b_n^2/(4b_n^2 + 2b_n - 2) > 1.5$ for $n \ge 33$. Let $y_n^2 = 2b_n + 6 \log b_n$, and then

$$P(M_n^2 \ge y_n^2) \le P(M_n \ge y_n) + 1/2^n.$$

Let $\tau_n = n[1 - \Phi(y_n)]$. Using the normal tail probability bounds (10), we have

$$\begin{aligned}\tau_n &\le \frac{n}{\sqrt{2\pi}}(2b_n + 6\log b_n)^{-1/2} \cdot \exp(-b_n - 3\log b_n)\\ &= \frac{ne^{-b_n}}{\sqrt{3\pi b_n}}\left(1 + \frac{3\log b_n}{b_n}\right)^{-1/2} \cdot \exp(-3\log b_n).\end{aligned}$$

Recall $1 - \Phi(\sqrt{2b_n}) = 1/n$, so that by (10),

$$\frac{ne^{-b_n}}{\sqrt{4\pi b_n}}\left(1 - \frac{1}{2b_n}\right) < 1.$$

When $n \ge 33$, we have

$$\left(1 + \frac{3\log b_n}{b_n}\right)^{-1/2} \cdot \left(1 - \frac{1}{2b_n}\right)^{-1} < 1,$$

and it follows that

$$\tau_n < \exp(-3\log b_n) = 1/b_n^3.$$

Using (8), we deduce that when n is large enough,

$$P(M_n \ge y_n) = 1 - (\Phi(y_n))^n = 1 - \left(1 - \frac{\tau_n}{n}\right)^n \le 1 - e^{-\tau_n} + \frac{1}{n-1} \le \tau_n + \frac{1}{n-1}.$$

Therefore, we conclude

$$1 - P(Y_n \le x) < \frac{1}{b_n^3} + \frac{1}{n-1} + \frac{1}{2^n} < \frac{c_1}{(\log n)^3},$$

for some absolute constant c_1. The preceding inequality, together with (17), completes the proof for $x > c_n$.

Finally, we consider $x < -c_n$ by showing that both $G_1(x)$ and $P(Y_n \le x)$ converge to 0 faster than $1/(\log n)^3$. Using the definition of b_n, we have when $n \ge 33$ and $x < -c_n = -4\log b_n$,

$$G_1(x) = \exp(-e^{-x}) < \exp(-b_n^4) < 1/b_n^4.$$

On the other hand, when $x \le -4\log b_n$,

$$P(Y_n \le x) \le P[M_n^2 \le g(-4\log b_n)] \le P\left(M_n^2 \le 2b_n - 4\log b_n \cdot \frac{8b_n^2}{4b_n^2 + 2b_n - 2}\right).$$

Again since $8b_n^2/(4b_n^2+2b_n-2) > 1.5$ when $n \geq 33$, if we let ${y'_n}^2 = 2b_n - 6\log b_n$, then

$$P(Y_n \leq x) \leq P(M_n \leq y_n).$$

Let $\tau'_n = n[1-\Phi(y'_n)]$, and we have by (10),

$$\begin{aligned}\exp(-\tau'_n) &< \exp\left\{-\frac{ne^{-b_n}}{\sqrt{4\pi b_n}}\left(1-\frac{3\log b_n}{b_n}\right)^{-1/2}\right.\\ &\qquad \left.\cdot\left(1-\frac{1}{(2b_n-6\log b_n)^2}\right)\cdot\exp(3\log b_n)\right\}\\ &< \exp\{-\exp(3\log b_n)\}\\ &< 1/b_n^3,\end{aligned}$$

when n is large enough. We conclude by (8)

$$P(Y_n \leq x) < \frac{1}{b_n^3} + \frac{1}{n} < \frac{c_1}{(\log n)^3},$$

which completes the proof. □

2.3 Comparisons of Different Convergence Rates

The best uniform convergence rate that can be obtained for M_n^2, if only centering and rescaling is allowed, is $(\log n)^{-2}$. We will give a summary of the progression in the literature. We also explain why the transformed M_n^2 can have a faster convergence rate $(\log n)^{-3}$.

In order for M_n^2 to have the limiting distribution G_1, the simplest option is to choose

$$b_{n1} = \log n - \log(\log n)/2 - \log(4\pi)/2;$$

then as a counterpart of (1), it holds that $\frac{1}{2}(M_n^2 - 2b_{n1}) \Rightarrow G_1$, where we use $\Rightarrow$ to denote the convergence in distribution. Using similar arguments as given in Hall (1979), it can be shown that the convergence rate is $(\log\log n)^2/\log n$. Similarly as (2), if b_{n1} is the solution of the equation

$$4\pi b_{n2}\exp(2b_{n2}) = n^2$$

and M_n^2 is centered by b_{n2}, then the rate of convergence is analogous to (3)

$$\frac{C_1}{\log n} < \sup_{-\infty<x<\infty} \left| P\left[\tfrac{1}{2}(M_n^2 - 2b_{n2}) \le x \right] - G_1(x) \right| < \frac{C_2}{\log n}. \tag{18}$$

Again, (18) can be established following the proof in Hall (1979).

We note that $\sqrt{2b_{n1}}$ is an approximation of the $(1-1/n)$-th quantile of standard normal distribution obtained by using the following approximation of the tail probability:

$$1 - \Phi\left(\sqrt{2b_{n1}}\right) \approx \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\log n}} \cdot \exp(-b_{n1}) = \frac{1}{n},$$

and b_{n2} is obtained by the following approximation of $1 - \Phi(\sqrt{2b_{n2}})$:

$$1 - \Phi\left(\sqrt{2b_{n2}}\right) \approx \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2b_{n2}}} \cdot \exp(-b_{n2}) = \frac{1}{n}.$$

If we choose b_{n3} through a more precise approximation of $1 - \Phi(\sqrt{2b_{n3}})$

$$1 - \Phi\left(\sqrt{2b_{n3}}\right) \approx \frac{1}{\sqrt{4\pi b_{n3}}}\left(1 - \frac{1}{2b_{n3}}\right) \exp\left(-b_{n3}\right) = \frac{1}{n},$$

and set $a_{n3} = 2 - 1/b_{n3}$, then $a_{n3}^{-1}(M_n^2 - 2b_{n3}) \Rightarrow G_1$ with the convergence rate

$$\frac{C_1}{(\log n)^2} < \sup_{-\infty<x<\infty} \left| P\left[a_{n3}^{-1}(M_n^2 - 2b_{n3}) \le x \right] - G_1(x) \right| < \frac{C_2}{(\log n)^2}. \tag{19}$$

The way we represent the preceding result is slightly different from the original one given by Hall (1980). The choices of a_{n3} and b_{n3} differ from those in Hall (1980) by smaller order terms, which do not affect the convergence rates. We choose the current formulation in order to have a better comparison with our main result.

To achieve a better rate of convergence, we first choose b_n precisely through $1 - \Phi(\sqrt{2b_n}) = 1/n$. Second, observe that the events in (18) and (19) can be written as

$$M_n^2 \le 2b_{n2} + 2x$$
$$M_n^2 \le 2b_{n3} + 2x - x/b_{n3},$$

respectively. According to (6), the event $[Y_n \le x]$ implies that

$$M_n^2 \le 2b_n + 2x - \frac{x}{b_n} + \frac{x^2 + 3x}{2b_n^2} + O\left(b_n^{-3}\right). \tag{20}$$

We see that a term of order $O(b_n^{-2})$ is needed on the right-hand side to achieve the convergence rate $(\log n)^{-3}$ in Theorem 1. In fact, it is this expansion which motivates the proposed nonlinear transform Y_n.

2.4 k-th Maxima

In this section, we present pointwise and uniform convergence rates for the k-th maxima $M_{n,k}$, defined as the k-th largest among the first n variables $\{X_1, X_2, \ldots, X_n\}$. These results follow from almost the same arguments as those for the maxima, so we state them without proofs.

Theorem 3 *Let b_n be the solution of the equation $1 - \Phi(\sqrt{2b_n}) = 1/n$. For a given positive integer k, define*

$$Y_{n,k} := \left[1 - \left(1 + \frac{M_{n,k}^2 - 2b_n}{8b_n^2}\right)^{-1}\right]\left(4b_n^2 + 2b_n - 2\right).$$

(i) For each fixed $-\infty < x < \infty$, it holds that

$$P(Y_{n,k} \le x) - G_k(x) = G_1(x)\frac{e^{-kx}}{(k-1)!} \cdot \frac{4x^3 + 15x^2 + 30x}{24b_n^3} + O(b_n^{-4}),$$

where $G_k(x) := G_1(x)\sum_{j=0}^{k-1} e^{-jx}/j!$.

(ii) There exists a constant $c_2 > 0$, such that

$$\sup_{-\infty<x<\infty} |P(Y_{n,k} \le x) - G_k(x)| < \frac{c_2}{(\log n)^3}.$$

3 Applications and Numerical Comparisons

3.1 Numerical Comparisons

In this section, we numerically compare the convergence rates of different versions of the normalized M_n^2, introduced in Sect. 2.3. Specifically, we compare with $G_1(x)$, the CDF of $Y_{n1} := \frac{1}{2}(M_n^2 - 2b_{n1})$, $Y_{n2} := \frac{1}{2}(M_n^2 - 2b_{n2})$, $Y_{n3} := (2 - 1/b_{n3})^{-1}(M_n^2 - 2b_{n3})$, and Y_n, labeled by b_{n1}, b_{n2}, b_{n3}, and b_n, respectively, in Fig. 1. The vertical lines mark 90%, 95%, and 99% quantiles of the Gumbel distribution. We see that the distribution of Y_n (blue curve) is uniformly closer to $G_1(x)$, no matter what the

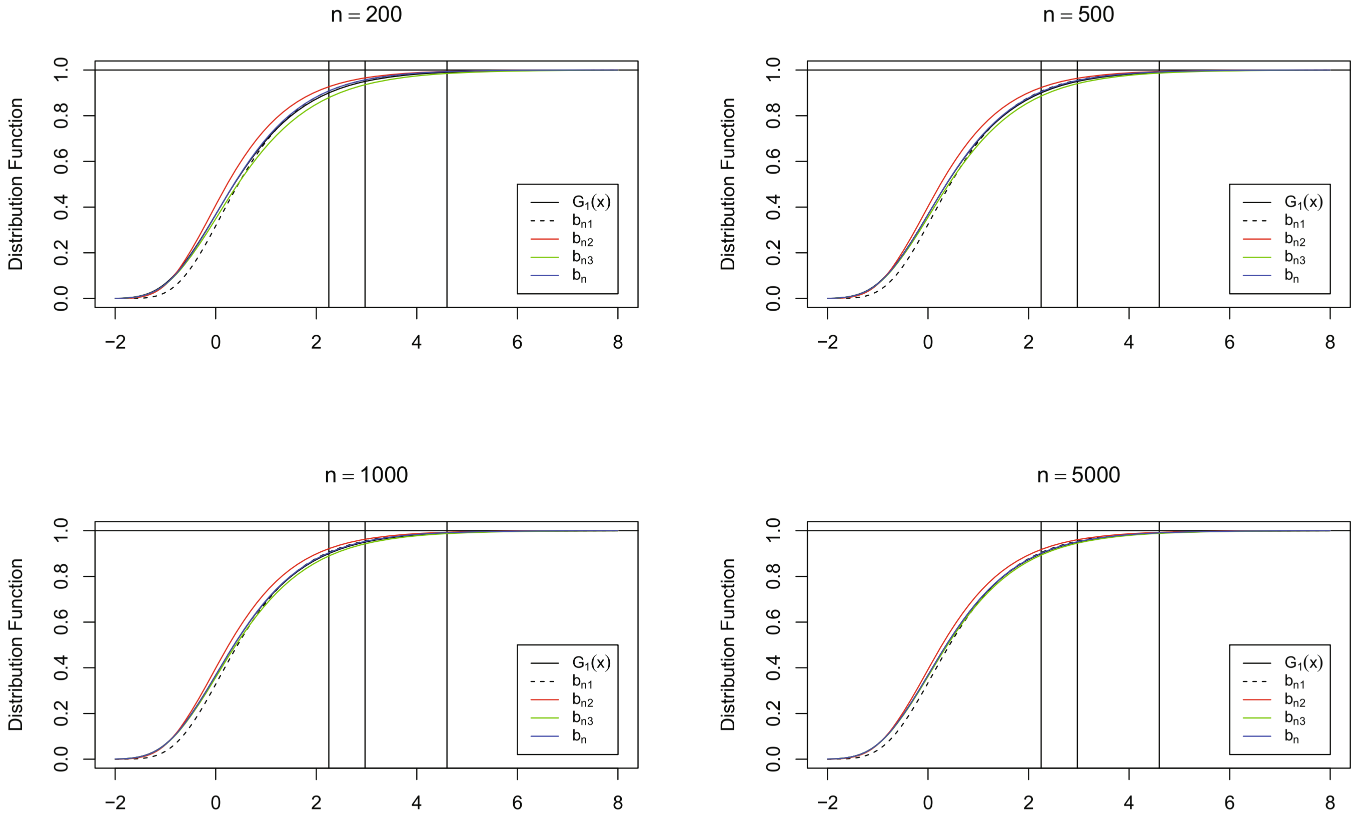


Fig. 1 Comparison of the CDFs. The black line is the true CDF of the Gumbel distribution. The dashed, red, and green (labeled by b_{n1}, b_{n2}, and b_{n3}) curves are the empirical CDFs, corresponding to the convergence rates $(\log\log n)^2/\log n$, $(\log n)^{-1}$, and $(\log n)^{-2}$, respectively. The blue line depicts the empirical CDF of the proposed Y_n, of convergence rate $(\log n)^{-3}$

sample size is. Figure 2 zooms into the upper tail for a clearer visualization. An interesting finding is that the faster theoretical convergence rates of Y_{n2} and Y_{n3} over Y_{n1} are not reflected through the plots for Y_{n2} even when the sample size is as large as 10^5. The distribution of Y_{n3} starts to be closer to $G_1(x)$ in the upper tail when $n = 10^5$. We remark that the inferior performances of Y_{n2} and Y_{n3} need not necessarily contradict the theoretical convergence rates: from Fig. 1, it is seen that the convergence of Y_{n1} is much slower in the left tail. On the other hand, in Fig. 2, it is more clearly seen that Y_n always has a faster convergence rate, compared with the rest. Furthermore, the CDF of Y_n lies above $G_1(x)$, indicating that if a hypothesis test is based on the maximum type statistic, then it is guaranteed to be conservative by using Y_n. This is in contrast to Y_{n3}, which is always below $G_1(x)$. Similar patterns are observed for the second maxima in Fig. 3. Two additional figures for the 3rd and 4th maxima are given in the Appendix.

Let c_α be the $(1-\alpha)$-th quantile of $G_1(x)$. We find the smallest sample size n such that $P(Y_n > c_\alpha)$ reaches $\pm 10\%$ of α. The results are summarized in Table 1 for all of Y_{ni}, $i = 1, 2, 3$ and Y_n. Overall Y_n needs much smaller sample sizes. Such sizes do not exist for Y_{n2} when $n \leq 10^6$, so we choose not to report them.

3.2 An Example

In this section, we consider an example on testing the covariance structure. Suppose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ is a sequence of independent and identically distributed p-dimensional random vectors. Let $R = \{\rho_{ij}\}_{1\leq i,j\leq p}$ be the correlation matrix of $\boldsymbol{x}_1$. Consider the hypothesis testing problem:

$$H_0 : R = I_p \quad \text{vs} \quad H_1 : R \neq I_p.$$

Jiang et al. (2004) proposed to use the maximum absolute sample correlation $L_N = \max_{1\leq i<j\leq p} |\hat{\rho}_{ij}|$ as the test statistic and proved that $\frac{1}{2}(NL_N^2 - 2b_{n1})$ converges in distribution to G_1, where $n = p(p-1)$. We consider the test statistics T_{Ni}, $i = 1, 2, 3$, and T_N, which are defined in the same way as Y_{ni} and Y_n in Sect. 3.1, but replacing M_n^2 therein by NL_N^2. The p-values are calculated by comparing the test statistics with the Gumbel distribution G_1. By treating the sample correlations $N\hat{\rho}_{ij}$ as iid standard normal random variables, we obtain another approximation of the p-value, given by $1 - [\Phi(NL_N^2)]^n$. The test done this way is named as T_0.

For the asymptotic tests considered here, two approximations are involved: (i) Gaussian approximation of $\hat{\rho}_{ij}$ and (ii) approximation of the maximum by the Gumbel distribution. It has been understood that Gaussian approximation usually has a much higher convergence rate, especially in view of the recent development on the topic (see, for example, Chernozhukov et al. 2013, and a series of follow-up works). Therefore, the bottleneck is the convergence rate of the maximum to the Gumbel distribution. We report the empirical rejection probabilities based on 5000 repetitions in Table 2 and Table 3, where $\boldsymbol{x}_i \sim N(\boldsymbol{0}, I_p)$, and $\boldsymbol{x}_i$ has iid

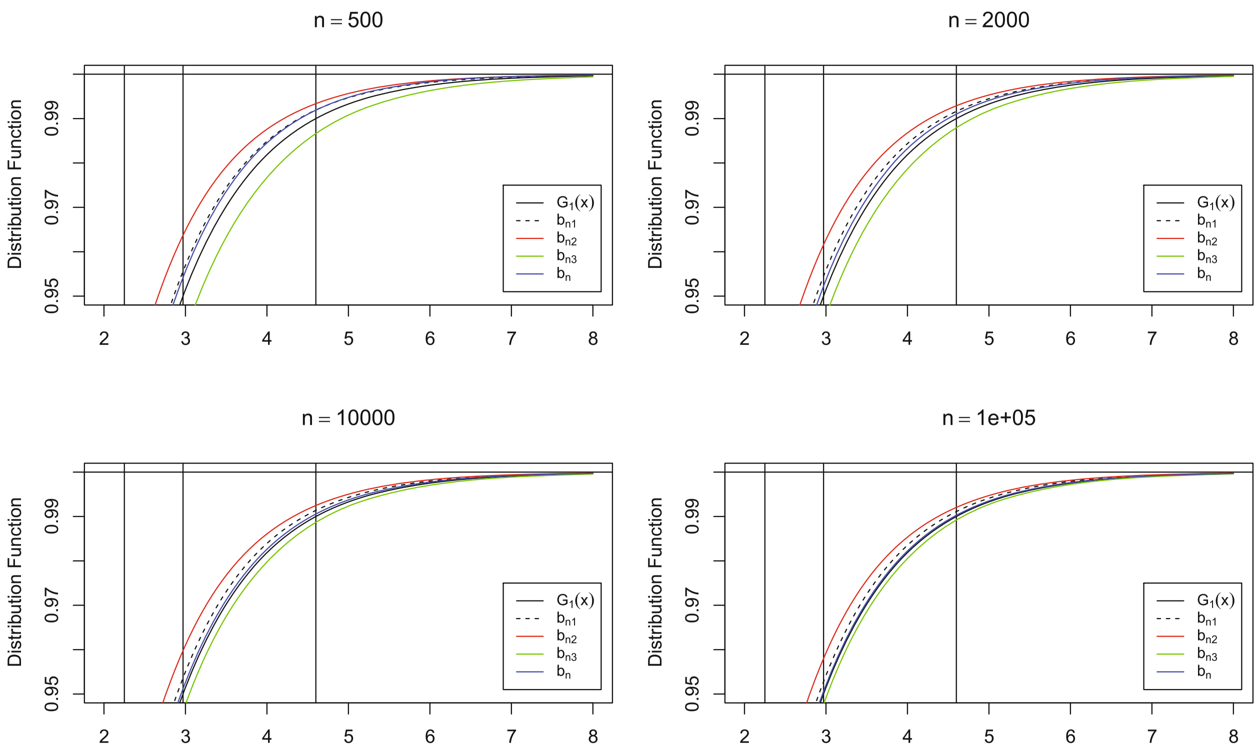


Fig. 2 Comparison of the CDFs in the upper tail. The black line is the true CDF of the Gumbel distribution. The vertical lines mark 90%, 95% and 99% quantiles of the Gumbel distribution. The dashed, red, and green (labeled by b_{n1}, b_{n2}, and b_{n3}) curves are the empirical CDFs, corresponding to the convergence rates $(\log\log n)^2/\log n$, $(\log n)^{-1}$ and $(\log n)^{-2}$, respectively. The blue line depicts the empirical CDF of the proposed Y_n, of convergence rate $(\log n)^{-3}$

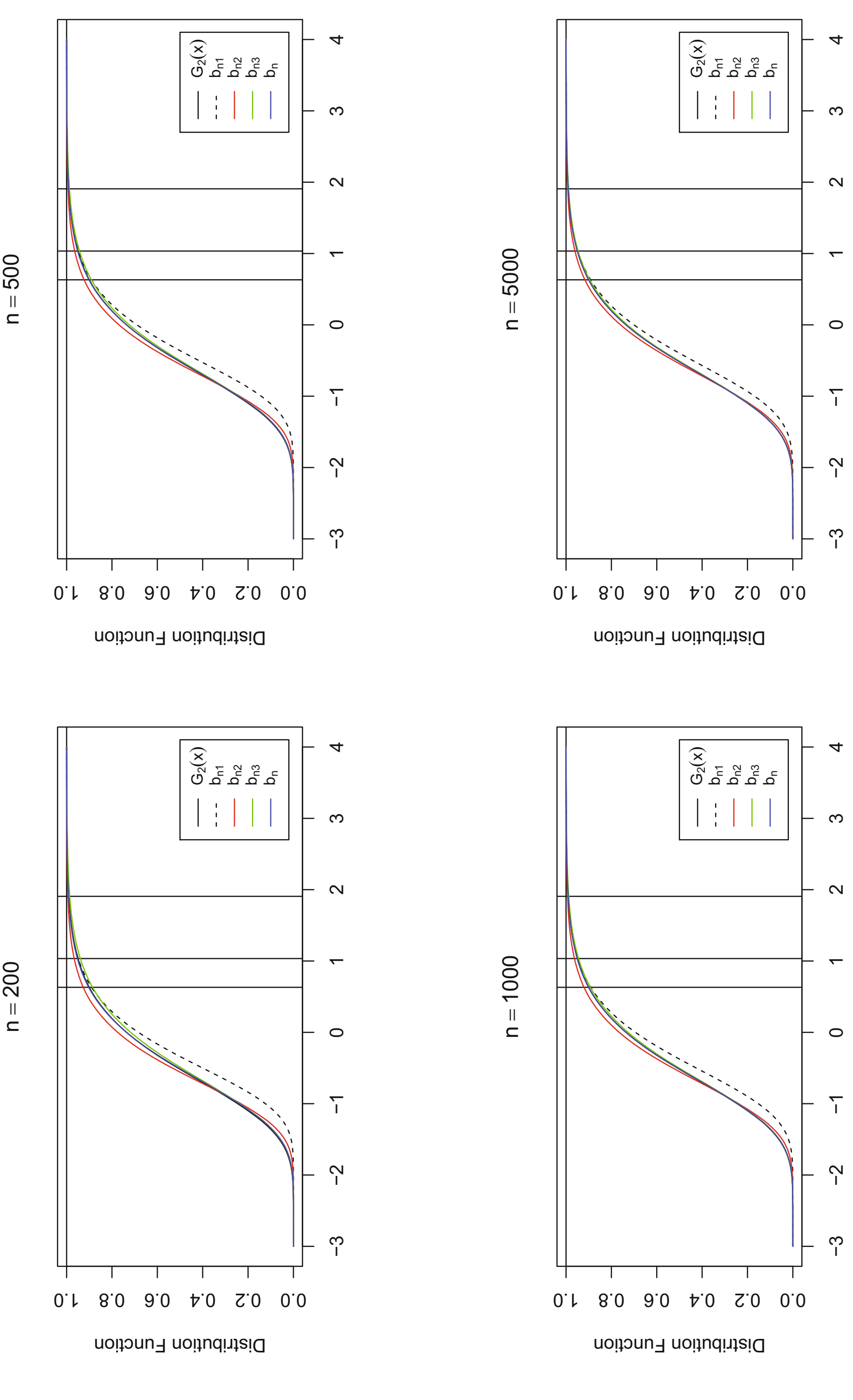


Fig. 3 Comparison of CDFs for second maxima

Table 1 Smallest sample size to reach ±10% of the nominal level

α	Y_{n1}	Y_{n2}	Y_{n3}	Y_n
10%	92	–	1230	293
5%	995	–	3639	686
1%	359,965	–	38,208	4126

Table 2 The empirical rejection probabilities (%) when $\boldsymbol{x}_i$ is $\mathbb{N}(0, I_p)$

		$n = 256$			$n = 512$			$n = 1024$		
p	Test	10%	5%	1%	10%	5%	1%	10%	5%	1%
32	T_0	8.96	4.42	0.72	9.64	4.86	0.94	10.74	5.60	1.28
	T_{N1}	8.62	4.02	0.62	8.94	4.48	0.78	10.28	5.02	1.10
	T_{N2}	7.10	3.36	0.48	7.72	3.80	0.54	8.48	4.18	0.80
	T_{N3}	10.06	5.12	1.02	10.64	5.46	1.16	11.92	6.42	1.62
	T_N	8.94	4.28	0.66	9.46	4.72	0.88	10.64	5.36	1.20
64	T_0	7.68	3.78	0.80	9.94	5.34	0.80	9.42	4.74	1.00
	T_{N1}	7.48	3.30	0.66	9.46	4.72	0.70	9.02	4.44	0.84
	T_{N2}	6.26	2.76	0.62	8.42	3.96	0.66	7.88	3.84	0.70
	T_{N3}	8.44	4.10	0.96	10.60	5.88	0.90	10.06	5.12	1.10
	T_N	7.68	3.76	0.72	9.88	5.24	0.80	9.36	4.70	0.96
128	T_0	7.60	3.32	0.62	9.30	4.86	0.80	9.86	4.82	0.98
	T_{N1}	7.34	3.12	0.60	8.90	4.52	0.72	9.56	4.58	0.82
	T_{N2}	6.26	2.72	0.60	7.86	3.82	0.66	8.20	4.00	0.68
	T_{N3}	8.16	3.82	0.66	9.78	5.14	0.90	10.14	5.30	1.14
	T_N	7.60	3.32	0.62	9.30	4.78	0.78	9.86	4.76	0.92
256	T_0	6.44	2.94	0.34	8.64	3.96	0.62	8.54	4.22	0.74
	T_{N1}	6.08	2.70	0.28	8.46	3.78	0.58	8.38	3.98	0.64
	T_{N2}	5.40	2.38	0.24	7.42	3.28	0.52	7.48	3.64	0.42
	T_{N3}	6.80	3.10	0.42	8.92	4.26	0.72	9.16	4.42	0.80
	T_N	6.44	2.92	0.34	8.66	3.94	0.60	8.54	4.20	0.70

t_7 entries, respectively. We see that the empirical sizes of T_N, T_{N1}, and T_0 are in general close to the nominal ones, and their performances are stable across different sample sizes and dimensions. The results are also consistent with our findings in Sect. 3.1. More extensive simulations, covering more sample sizes and dimensions, continue to support the observations above. These results are omitted for the sake of space.

Table 3 The empirical rejection probabilities (%) when $\boldsymbol{x}_i$ has iid t_7 entries

		$n = 256$			$n = 512$			$n = 1024$		
p	Test	10%	5%	1%	10%	5%	1%	10%	5%	1%
32	T_0	9.84	4.82	1.02	9.18	4.70	1.24	10.22	5.28	1.12
	T_{N1}	9.28	4.36	0.76	8.82	4.34	1.06	9.58	4.66	0.92
	T_{N2}	7.84	3.74	0.66	7.28	3.72	0.92	8.26	3.90	0.68
	T_{N3}	10.74	5.70	1.36	10.04	5.32	1.40	11.38	6.12	1.20
	T_N	9.70	4.66	0.84	9.14	4.44	1.16	10.02	5.02	0.98
64	T_0	9.28	4.28	0.94	10.18	5.02	0.82	9.02	4.78	1.00
	T_{N1}	8.96	3.88	0.70	9.80	4.50	0.68	8.46	4.54	0.78
	T_{N2}	7.70	3.44	0.52	8.42	3.94	0.56	7.44	3.82	0.70
	T_{N3}	9.72	4.74	1.04	10.76	5.44	0.98	9.82	5.28	1.14
	T_N	9.24	4.22	0.84	10.18	4.94	0.78	9.00	4.74	0.88
128	T_0	9.14	4.64	0.82	9.74	4.90	1.30	9.96	4.76	0.94
	T_{N1}	8.90	4.32	0.76	9.32	4.60	1.14	9.58	4.44	0.84
	T_{N2}	7.82	3.90	0.56	8.10	3.76	0.84	8.26	3.84	0.70
	T_{N3}	9.64	4.84	0.90	10.20	5.16	1.50	10.32	5.10	1.08
	T_N	9.14	4.58	0.80	9.74	4.80	1.20	9.96	4.66	0.90
256	T_0	9.08	4.32	0.94	10.20	4.98	0.98	9.98	5.36	1.30
	T_{N1}	8.80	4.02	0.88	9.96	4.74	0.84	9.68	5.02	1.18
	T_{N2}	7.92	3.50	0.86	8.82	4.18	0.80	8.84	4.52	1.10
	T_{N3}	9.36	4.72	1.04	10.58	5.30	1.04	10.48	5.62	1.38
	T_N	9.08	4.30	0.94	10.20	4.98	0.94	9.98	5.30	1.20

4 Conclusion

We propose a monotone transform of the squared normal extreme and prove that its pointwise and uniform convergence rates are both of the order $(\log n)^{-3}$, which improves the existing results in the literature. The theoretical improvements are also demonstrated and supported numerically.

Acknowledgments The research is supported in part by National Science Foundation grant DMS-2027855 and DMS-2052949. The authors thank the Editors and three anonymous referees for their insightful comments which have led to significant improvement of the manuscript.

Appendix

Expansion of b_n

Recall b_n is the solution of the equation $1 - \Phi(\sqrt{2b_n}) = 1/n$. We use the following approximation to the normal density:

$$1 - \Phi(z) = \left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5} - \frac{15}{z^7}\right)\phi(z).$$

Then $\sqrt{2b_n}$ is the solution of the following equation:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7}\right) = 1/n. \tag{21}$$

Our goal is to use three consecutive applications of the Newton-Raphson approximation method to obtain the solution of (21) and then calculate b_n accordingly. Let

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7}\right).$$

then the derivative of $f(x)$ is:

$$f'(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(-1 + \frac{105}{x^8}\right).$$

We start from

$$x_0 = \sqrt{2\log n} - \frac{\Delta}{2\sqrt{2\log n}},$$

where

$$\Delta = \log\log n + \log 4\pi.$$

By Newton-Rhapson approximation method,

$$f(x_0) + f'(x_0)(x_1 - x_0) = 1/n.$$

Then we can obtain:

$$x_1 = \sqrt{2\log n} - \frac{\Delta}{2\sqrt{2\log n}} - \frac{\Delta^2 - 4\Delta + 8}{8(2\log n)^{3/2}}.$$

Repeat this procedure for two more times, we have

$$x_2 = \sqrt{2\log n} - \frac{\Delta}{2\sqrt{2\log n}} - \frac{\Delta^2 - 4\Delta + 8}{8(2\log n)^{3/2}} - \frac{\Delta^3 - 8\Delta^2 + 32\Delta - 56}{16(2\log n)^{5/2}}.$$

$$\begin{aligned} x_3 =& \sqrt{2\log n} - \frac{\Delta}{2\sqrt{2\log n}} - \frac{\Delta^2 - 4\Delta + 8}{8(2\log n)^{3/2}} - \frac{\Delta^3 - 8\Delta^2 + 32\Delta - 56}{16(2\log n)^{5/2}} \\ & - \frac{15\Delta^4 - 184\Delta^3 + 1152\Delta^2 - 4128\Delta + 7040}{384(2\log n)^{7/2}}. \end{aligned}$$

Then by $b_n = x_3^2/2$, it can be easily calculated:

$$b_n = \log n - \frac{\Delta}{2} + \frac{\Delta - 2}{4\log n} + \frac{\Delta^2 - 6\Delta + 14}{16(\log n)^2} + O\left(\frac{(\log\log n)^3}{(2\log n)^3}\right).$$

Additional Figures

In this section we provide comparisons of the CDFs of the third and fourth maxima (Figs. 4 and 5).

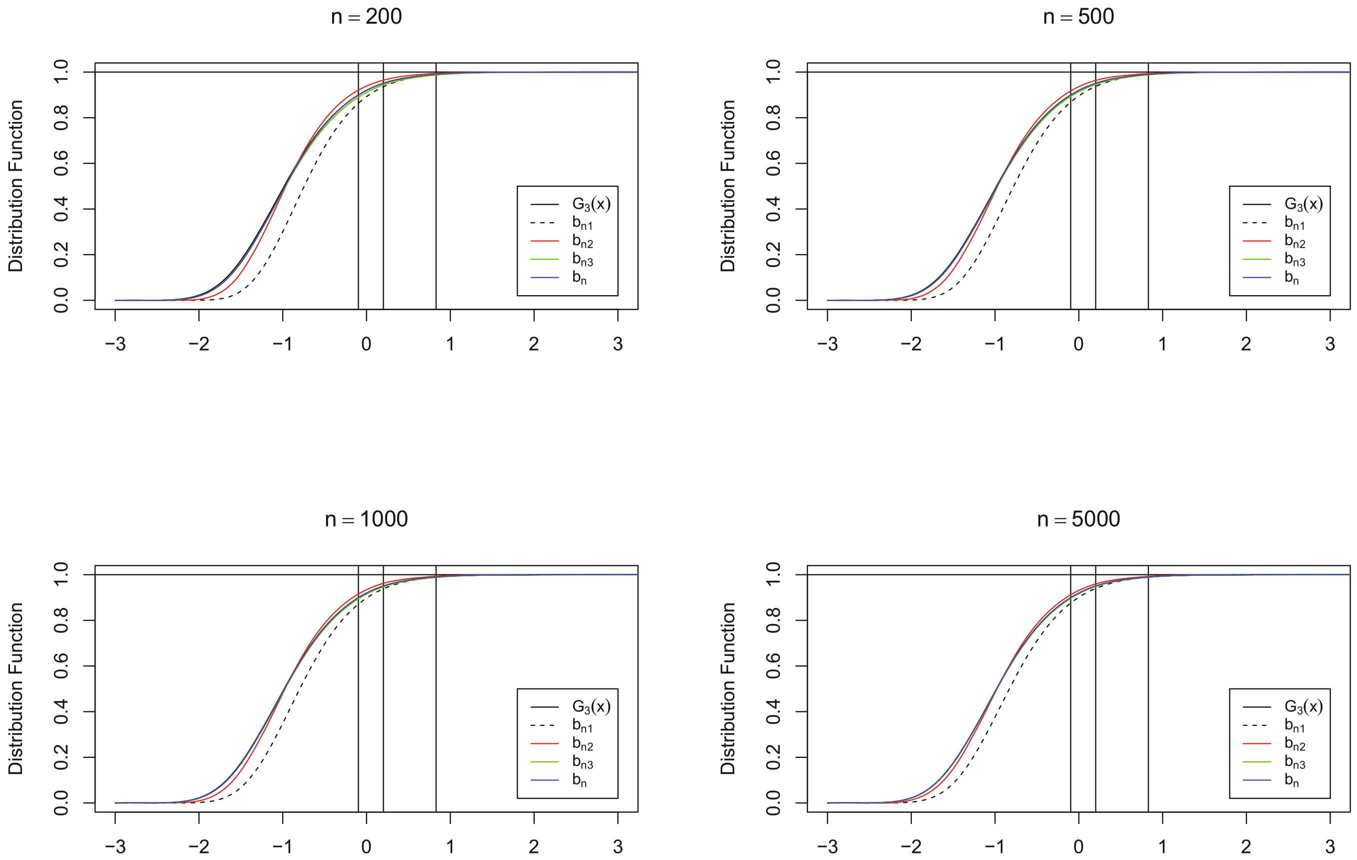


Fig. 4 Comparison of the CDFs for third maxima

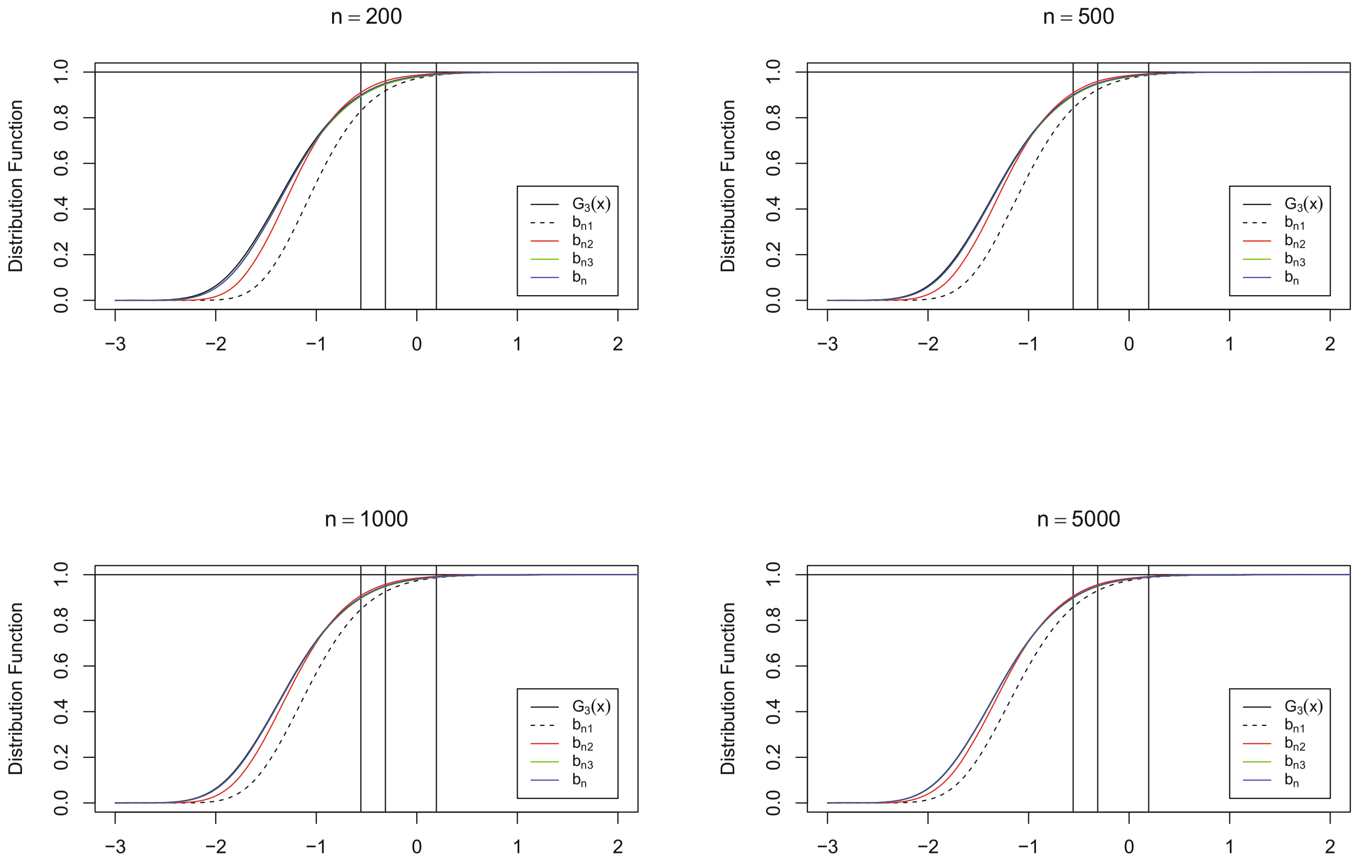


Fig. 5 Comparison of the CDFs for fourth maxima

References

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications.

Cai, T. T., Liu, W., & Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*(2), 349–372.

Chernozhukov, V., Chetverikov, D., & Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics, 41*(6), 2786–2819.

Cohen, J. P. (1982a). Convergence rates for the ultimate and pentultimate approximations in extreme-value theory. *Advances in Applied Probability, 14*(4), 833–854.

Cohen, J. P. (1982b). The penultimate form of approximation to normal extremes. *Advances in Applied Probability, 14*(2), 324–339.

Daniels, H. (1982). A transformation for normal extremes. *Journal of Applied Probability, 19*(A), 201–206.

de Haan, L., & Resnick, S. (1996). Second-order regular variation and rates of convergence in extreme-value theory. *Annals of Probability, 24*(1), 97–124.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society* (vol. 24, pp. 180–190). Cambridge University Press.

Hall, P. (1979). On the rate of convergence of normal extremes. *Journal of Applied Probability, 16*(2), 433–439.

Hall, P. (1980). Estimating probabilities for normal extremes. *Advances in Applied Probability, 12*(2), 491–500.

Hall, P. (1991). On convergence rates of suprema. *Probabability Theory Related Fields, 89*(4), 447–455.

Hall, W. J., & Wellner, J. A. (1979). The rate of convergence in law of the maximum of an exponential sample. *Statistica Neerlandica, 33*(3), 151–154.

Jiang, T., et al. (2004). The asymptotic distributions of the largest entries of sample correlation matrices. *The Annals of Applied Probability, 14*(2), 865–880.

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. New York, Berlin: Springer-Verlag.

Peng, Z., Saralees, N., and Lin, F. (2010). Convergence rate of extremes for the general error distribution. *Journal of Applied Probability, 47*(3), 668–679.

Romano, J. P., & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association, 100*(469), 94–108.

Rootzén, H. (1983). The rate of convergence of extremes of stationary normal sequences. *Advances in Applied Probability, 15*(1), 54–80.

Smith, R. L. (1982). Uniform rates of convergence in extreme-value theory. *Advances in Applied Probability, 14*(3), 600–622.

Local Spectral Analysis of Qualitative Sequences via Minimum Description Length

David S. Stoffer

Abstract The idea of signal detection in the frequency domain for qualitative-valued time series was developed in Stoffer et al. (Biometrika 80(3):611–622, 1993) under the assumption of homogeneity. The tool is called the spectral envelope and is related to the concept of scaling qualitative data. After reviewing the basic ideas, we present a method for fitting a local spectral envelope to heterogeneous sequences based on a minimum description length (MDL) criterion for choosing the best fitting model based on parsimony. Inherent in the methodology is the detection of breakpoints in long sequences. Because the search space is immense, optimization is accomplished via a genetic algorithm (GA) to effectively tackle the problem. Numerical examples are given using sleep state data and DNA sequences.

Keywords Breakpoint detection · Categorical time series · Genetic Algorithm · MDL · Nonstationary processes · Spectral envelope

1 Introduction

Qualitative-valued time series are frequently encountered in diverse applications such as economics, medicine, psychology, geophysics, and genomics, to mention a few. The fact that the data are categorical-valued does not preclude the need to detect signals in the same way that is done with quantitative-valued time series. Here, we explore an approach based on scaling and the spectral envelope, which was introduced in Stoffer, Tyler, & McDougall (1993).

First we discuss the concept of scaling categorical variables, and then we use the idea to develop spectral analysis qualitative-valued processes. In doing so, the spectral envelope and optimal scaling are introduced, and their properties are discussed. The spectral envelope and the corresponding optimal scaling is a popu-

D. S. Stoffer (✉)
Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA
e-mail: stoffer@pitt.edu

M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*,
https://doi.org/10.1007/978-3-031-22687-8_22

Table 1 Per Minute Infant EEG Sleep States (read down and across)

REM	NR2	NR4	NR2	NR1	NR2	NR3	NR4	NR1	NR1	REM
REM	REM	NR4	NR1	NR1	NR2	NR4	NR4	NR1	NR1	REM
REM	REM	NR4	NR1	NR1	REM	NR4	NR4	NR1	NR1	REM
REM	NR3	NR4	NR1	REM	REM	NR4	NR4	NR1	NR1	REM
REM	NR4	NR4	NR1	REM	REM	NR4	NR4	NR1	NR1	REM
REM	NR4	NR4	NR1	REM	REM	NR4	NR4	NR1	NR1	REM
REM	NR4	NR4	NR2	REM	NR2	NR4	NR4	NR1	NR1	NR2
REM	NR4	NR4	REM	REM	NR2	NR4	NR4	NR1	REM	
NR2	NR4	NR4	NR1	REM	NR2	NR4	NR4	NR1	REM	
REM	NR2	NR4	NR1	REM	NR3	NR4	NR2	NR1	REM	

lation idea. We also discuss efficient estimation in the homogeneous case. Pertinent theoretical results are also summarized. Examples of using the methodology on sleep state sequences and DNA sequences, which are typically heterogeneous, are given. The examples include an analysis of a gene in the Epstein–Barr virus and Herpesvirus saimiri. The main contribution is the development of a local procedure using minimum description length (MDL) coupled with optimization via a genetic algorithm (GA).

Our work on the spectral envelope was motivated by collaborations with neurologists who performed sleep studies on neonates with an interest in sleep cycles. For example, Table 1 shows the per minute sleep state of an infant taken from a study on the effects of prenatal exposure to alcohol. Details can be found in Stoffer et al. (1988), but briefly, an electroencephalographic (EEG) sleep recording of approximately 2 h is obtained on a full term infant 24–36 hours after birth, and the recording is scored by a pediatric neurologist for sleep state. There are two main types of sleep, Non-Rapid Eye Movement (Non-REM), also known as *quiet sleep* and Rapid Eye Movement (REM), also known as *active sleep*. In addition, there are four stages of Non-REM (NR1 —NR4), with NR1 being the “most active” of the four states, and finally awake (AW), which naturally occurs briefly through the night. This particular infant was never awake during the study.

Neurologists usually order sleep states by brain activity, however, the idea of ordering sleep states is somewhat tenuous. For example, for a typical normal healthy adult, sleep begins in stage NR1 and progresses into stages NR2, NR3, and NR4. Sleep moves through these stages repeatedly before entering REM sleep. But sleep does not progress through these stages in sequence. Typically, sleep transitions between REM and stage NR2 so that one can move between the states without passing through other sleep states. Finally, there is no evidence to support that distance between say, NR4 and NR3 is the same as between NR2 and NR1 (in addition, state NR2 is considered a transitional state rather than an actual state of sleep).

However, it is not too difficult to notice a pattern in the data if one concentrates on REM versus Non-REM sleep states. But, it would be difficult to try to assess patterns

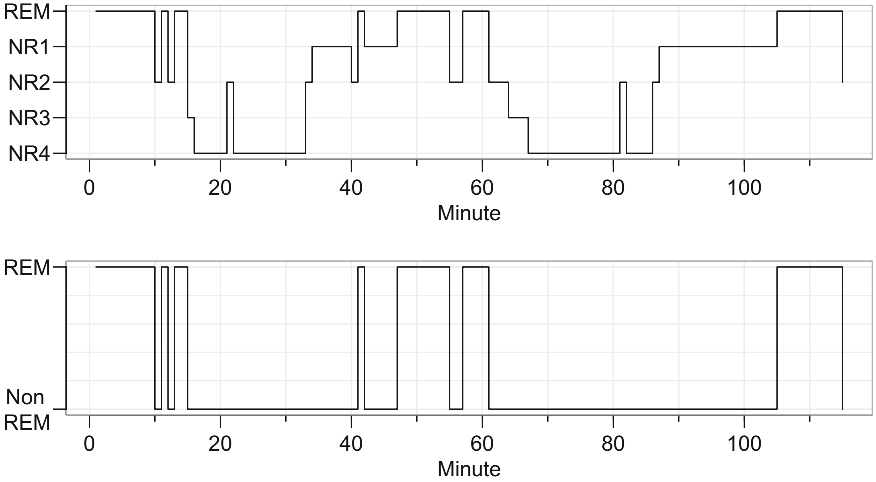


Fig. 1 Time plot of the EEG sleep state data in Table 1 using the scaling in (1) [TOP] and using the scaling in (2) [BOTTOM]

in a longer sequence–or if there were more categories–without some graphical aid. One simple method would be to *scale* the data, that is, *assign numerical values to the categories* and then draw a time plot of the scales. One obvious scaling that is frequently used by neurologists is:

$$\texttt{NR4} = 1, \quad \texttt{NR3} = 2, \quad \texttt{NR2} = 3, \quad \texttt{NR1} = 4, \quad \texttt{REM} = 5, \quad \texttt{AW} = 6, \tag{1}$$

and the top of Fig. 1 (often referred to as a *hypnogram*) shows the time plot using this scaling. Another interesting scaling might be to combine the quiet states and the active states:

$$\texttt{NR4} = \texttt{NR3} = \texttt{NR2} = \texttt{NR1} = 0, \quad \texttt{REM} = \texttt{AW} = 1\,. \tag{2}$$

The time plot using scalings (1) and (2) shown Fig. 1 is similar and we notice the general cyclic (in and out of REM sleep) behavior of this infant's sleep pattern. Figure 2 shows the estimated spectrum of the sleep data using the scalings in both (1) and (2). Note that there is a large peak at the frequency corresponding to 1 cycle every 60 min using either scaling. Most of us would feel comfortable with this analysis even though we made arbitrary and ad hoc choices about the particular scaling. It is evident from the data (without any scaling) that if the interest is in infant sleep cycling, this particular sleep study indicates that the infant cycles between REM and Non-REM sleep at a rate of about one cycle per hour.

The intuition used in the previous example is lost when one considers a long nucleotide DNA sequence. Briefly, a DNA strand can be viewed as a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five

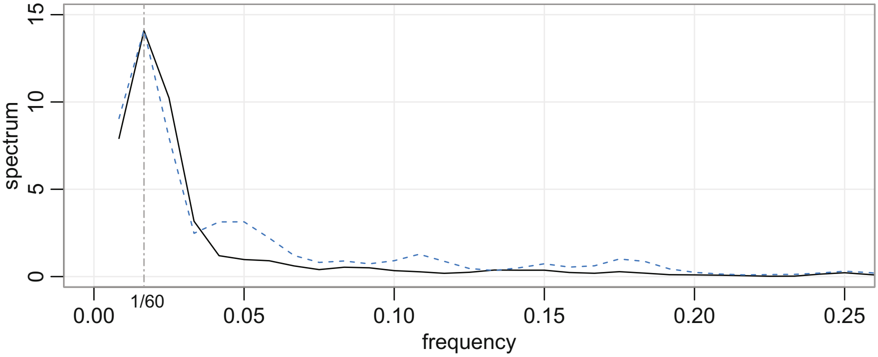


Fig. 2 Estimated spectrum of the EEG sleep state data in Table 1 based on the scaling in (1) [solid line] and on the scaling in (2) [dashed line]. The peaks in each correspond to a frequency of one cycle every 60 min

carbon sugar, and a phosphate group where four different bases can be grouped by size, the pyrimidines, thymine (T) and cytosine (C), and the purines, adenine (A) and guanine (G). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the $5'$ carbon of one sugar linked to the $3'$ carbon of the next, giving the string direction. DNA molecules occur naturally as a double helix composed of polynucleotide strands with the bases facing inwards. The two strands are complementary, so it is sufficient to represent a DNA molecule by a sequence of bases on a single strand. Thus, a strand of DNA can be represented as a sequence of letters, termed base pairs (*bp*), from the finite alphabet {A, C, G, T}. The order of the nucleotides contains the genetic information specific to the organism. Expression of information stored in these molecules is a complex multistage process. One important task is to translate the information stored in the protein-coding sequences (CDS) of the DNA. A common problem in analyzing long DNA sequence data is in identifying CDS that are dispersed throughout the sequence and separated by regions of noncoding (which makes up most of the DNA). Table 2 shows part of the Herpesvirus saimiri (HVS) DNA sequence. The entire sequence consists of approximately 157,000 bp.

One could try scaling according to the purine-pyrimidine alphabet, $\mathtt{A} = \mathtt{G} = 0$ and $\mathtt{C} = \mathtt{T} = 1$, but this is not necessarily of interest for every nucleotide sequence. There are numerous other alphabets of interest, for example, one might focus on the strong-weak hydrogen bonding alphabet $\mathtt{S} = \{\mathtt{C}, \mathtt{G}\} = 0$ and $\mathtt{W} = \{\mathtt{A}, \mathtt{T}\} = 1$; the bp C and G have a strong hydrogen bonding interaction whereas A and T have a weak bonding. In addition, there is no compelling theory that states that any reduction alphabet should be considered. While model calculations as well as experimental data strongly agree that some kind of periodic signal exists in certain DNA sequences, there is a large disagreement about the exact type of periodicity. In addition, there is disagreement about which nucleotide alphabets are involved in the signals; for example, compare Ioshikhes et al. (1996) with Satchwell et al. (1986).

Table 2 Part of the Herpesvirus saimiri (HVS) DNA Sequence(read across and down)

GGTCGCGAGG	GTCTAGCGCC	TCGAAACCGG	CTCGGAGCAC	AAGCAGACTC	TAGCCCCCTC
CCCTAGTACA	CAGAGCCCAG	CAGGCAGCTA	CAGCCGCTCA	ACGCGAGTCC	CTCCCCTTGC
TCAAGCTCTT	TAGTACACTT	TTTGTCTTTT	ATACAATAGT	TTTATTACTG	CATAGTATAA
GACATTTACT	GCAGCACTAT	GTGATTCACT	TTGATTCTTT	TACATTTTTT	TAAACATAAT
TACTAGCATT	AAACCAATTA	TGATTAATAG	CAAAACAATA	ATAACTAGCA	GCAATAGGAT
AGTTACAGAA	CAGTCTGTGC	ATTTGTCACC	TTCTTGCTCG	TGTTCACTGT	GCAGGCTTCC
GACTTCTGCG	TAGACATGTT	CTTCACTTCC	TGCTCCTCCG	CAGCCACTGA	CACGTACTGC
TGATAAGCCT	ACTGGGGTGC	TTAAATGTGA	TGAGCTCCGT	GAGCCAGATG	GTGTTGGTAA
GCCTACTGCT	CCCGATAGTG	CTGTTGGTCT	TCCTGGGCAT	CCGCTTTCTT	GCACTGGGTG
GCCAAGCAAG	CAGTAGGGAT	TATAAGGCCC	AAAGGGCCCT	GCATTTAAAA	GCGTTACAGG
TAAGTATGGT	GTAGGTCCAT	CATCTCCATC	ACTTCTTTCA	TCAGTATTGT	GTGGAGGATC
TCCGTTGCTT	TCATCGTTTT	CTTGTGGGTC	TCCTTCACCT	AGACCTCTTG	CCATTTTCTT
ACACGTCTAA	GCTTCAGTTT	GTTTAGCTGA	TTCTTGTAGT	GTTGTCTGTC	TTGCTAATTC

If we consider the naive approach of arbitrarily assigning numerical values (scales) to the categories and then proceeding with a spectral analysis, the result will depend on the particular assignment of numerical values. The obvious problem of being arbitrary is illustrated as follows: Suppose we observe the sequence ATCTACATG..., then setting $\texttt{A} = \texttt{G} = 0$ and $\texttt{C} = \texttt{T} = 1$ yields the numerical sequence 011101010..., which is not very interesting. However, if we used the strong-weak bonding alphabet, $\texttt{W} = \{\texttt{A}, \texttt{T}\} = 0$ and $\texttt{S} = \{\texttt{C}, \texttt{G}\} = 1$, then the sequence becomes 001001001 ..., which is very interesting.

In addition, if one considers the sequence $\{\texttt{G}, \texttt{A}, \texttt{T}, \texttt{A}, \texttt{G}, \texttt{A}, \texttt{T}, \texttt{A}, \ldots\}$, it is repeating every four bp (GATA ...). But, the sequence is also repeating every two bp if we consider the sequence in terms of not-A [$\bar{\texttt{A}}$] and A, ($\bar{\texttt{A}}\,\texttt{A}\,\bar{\texttt{A}}\,\texttt{A}\ \ldots$). It should be clear then, that one does not want to focus on only one scaling. Instead, the focus should be on finding scalings that bring out all of the interesting features in the data. Rather than choose values arbitrarily, the spectral envelope approach selects scales that help emphasize any periodic feature that exists in a categorical time series of virtually any length in a quick and automated fashion. In addition, the technique can help in determining whether a sequence is merely a random assignment of categories.

2 Spectral Envelope

As a general description, the spectral envelope is a frequency based, principal components technique applied to a multivariate time series. In this section we will focus on the basic concept and its use in the analysis of categorical time series. Technical details can be found in Stoffer, Tyler, & McDougall (1993). In addition, various extensions and applications may be found in McDougall, Stoffer, & Tyler (1997) and Stoffer, Tyler, & Wendt (2000).

In establishing the spectral envelope for categorical time series, we addressed the basic question of how to efficiently discover periodic components in categorical time series. Let $\{X_t;\ t = 0, \pm 1, \pm 2, \dots\}$ be a categorical-valued time series with finite state-space $C = \{c_1, c_2, \dots, c_{k+1}\}$. Assume that X_t is stationary and $p_j = \Pr\{X_t = c_j\} > 0$ for $j = 1, 2, \dots, k+1$. For $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_{k+1})' \in \mathbb{R}^{k+1}$, denote by $X_t(\boldsymbol{\beta})$ the real-valued stationary time series corresponding to the scaling that assigns the category c_j the numerical value β_j, for $j = 1, 2, \dots, k+1$. Our goal was to find scalings $\boldsymbol{\beta}$ so that the spectral density, $f(\omega; \boldsymbol{\beta})$ assuming it exists, of the scaled process is in some sense interesting, and to summarize the spectral information by what we called the spectral envelope.

We chose $\boldsymbol{\beta}$ to maximize the power (variance) at each frequency ω, across frequencies $\omega \in (-1/2,\ 1/2]$, relative to the total power $\sigma^2(\boldsymbol{\beta}) = \mathrm{var}\{X_t(\boldsymbol{\beta})\}$. That is, we chose $\boldsymbol{\beta}(\omega)$, at each ω of interest, so that

$$\lambda(\omega) = \sup_{\boldsymbol{\beta} \not\propto \mathbb{1}} \left\{ \frac{f(\omega; \boldsymbol{\beta})}{\sigma^2(\boldsymbol{\beta})} \right\}, \tag{3}$$

where $\mathbb{1}$ is the $(k+1) \times 1$ vector of ones. Note that $\lambda(\omega)$ is not defined if $\boldsymbol{\beta} \propto \mathbb{1}$ because such scalings correspond to assigning each category the same value; in this case $f(\omega; \boldsymbol{\beta}) \equiv 0$ and $\sigma^2(\boldsymbol{\beta}) = 0$. The optimality criterion $\lambda(\omega)$ possesses the desirable property of being invariant under location and scale changes of $\boldsymbol{\beta}$.

As in most scaling problems for categorical data, it was useful to represent the categories in terms of the vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_{k+1}$, where $\boldsymbol{e}_j$ represents the $(k+1)\times 1$ vector with a one in the j-th row, and zeros elsewhere. We then defined a $(k+1)$-dimensional stationary time series $\boldsymbol{Y}_t$ by $\boldsymbol{Y}_t = \boldsymbol{e}_j$ when $X_t = c_j$. The time series $X_t(\boldsymbol{\beta})$ can be obtained from the $\boldsymbol{Y}_t$ time series by the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_t$. Assume that the vector process $\boldsymbol{Y}_t$ has a continuous spectral density matrix denoted by $\boldsymbol{f}_Y(\omega)$. For each ω, $\boldsymbol{f}_Y(\omega)$ is a $(k+1)\times(k+1)$ complex-valued Hermitian matrix. Note that the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_t$ implies that $\boldsymbol{f}_Y(\omega; \boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{f}_Y(\omega)\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{f}_Y^{re}(\omega)\boldsymbol{\beta}$, where $\boldsymbol{f}_Y^{re}(\omega)$ denotes the real part of $\boldsymbol{f}_Y(\omega)$. The optimality criterion can thus be expressed as

$$\lambda(\omega) = \sup_{\boldsymbol{\beta} \not\propto \mathbb{1}} \left\{ \frac{\boldsymbol{\beta}'\boldsymbol{f}_Y^{re}(\omega)\boldsymbol{\beta}}{\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}}, \right\} \tag{4}$$

where $\boldsymbol{V}$ is the variance-covariance matrix of $\boldsymbol{Y}_t$. The resulting scaling $\boldsymbol{\beta}(\omega)$ is called the optimal scaling.

In this case, $\boldsymbol{Y}_t$ is a multivariate point process and any particular component of $\boldsymbol{Y}_t$ is the individual point process for the corresponding state (for example, the first component of $\boldsymbol{Y}_t$ indicates whether or not the process is in state c_1 at time t). For any fixed t, $\boldsymbol{Y}_t$ represents a single observation from a simple multinomial sampling scheme. It readily follows that $\boldsymbol{V} = \boldsymbol{D} - \boldsymbol{p}\ \boldsymbol{p}'$, where $\boldsymbol{p} = (p_1, \dots, p_{k+1})'$, and $\boldsymbol{D}$

is the diagonal matrix $\boldsymbol{D} = \text{diag}\{p_1, \ldots, p_{k+1}\}$. Since, by assumption, $p_j > 0$ for $j = 1, 2, \ldots, k + 1$, it follows that $\text{rank}(\boldsymbol{V}) = k$ with the null space of $\boldsymbol{V}$ being spanned by $\mathbb{1}$. For any $(k + 1) \times k$ full rank matrix $\boldsymbol{Q}$ whose columns are linearly independent of $\mathbb{1}$, $\boldsymbol{Q}'\boldsymbol{V}\boldsymbol{Q}$ is a $k \times k$ positive definite symmetric matrix.

With the matrix $\boldsymbol{Q}$ as previously defined, and for $-1/2 < \omega \le 1/2$, define $\lambda(\omega)$ to be the largest eigenvalue of the determinantal equation

$$|\boldsymbol{Q}'\boldsymbol{f}_Y^{re}(\omega)\boldsymbol{Q} - \lambda\boldsymbol{Q}'\boldsymbol{V}\boldsymbol{Q}| = 0,$$

and let $\boldsymbol{b}(\omega) \in \mathbb{R}^k$ be any corresponding eigenvector, that is,

$$\boldsymbol{Q}'\boldsymbol{f}_Y^{re}(\omega)\boldsymbol{Q}\boldsymbol{b}(\omega) = \lambda(\omega)\boldsymbol{Q}'\boldsymbol{V}\boldsymbol{Q}\boldsymbol{b}(\omega).$$

The eigenvalue $\lambda(\omega) \ge 0$ does not depend on the choice of $\boldsymbol{Q}$ because for any two such matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, there exists a nonsingular matrix $\boldsymbol{M}$ such that $\boldsymbol{Q}_1\boldsymbol{M} = \boldsymbol{Q}_2\boldsymbol{M}$. Although the eigenvector $\boldsymbol{b}(\omega)$ depends on the particular choice of $\boldsymbol{Q}$, the equivalence class of scalings associated with $\boldsymbol{\beta}(\omega) = \boldsymbol{Q}\boldsymbol{b}(\omega)$ does not depend on $\boldsymbol{Q}$. A convenient choice of $\boldsymbol{Q}$ is $\boldsymbol{Q} = [\mathbf{I} \mid \mathbf{0}\,]'$, where $\mathbf{I}$ is the $k \times k$ identity matrix and $\mathbf{0}$ is the $k \times 1$ vector of zeros. For this choice, $\boldsymbol{Q}'\boldsymbol{f}_Y^{re}(\omega)\boldsymbol{Q}$ and $\boldsymbol{Q}'\boldsymbol{V}\boldsymbol{Q}$ are the upper $k \times k$ blocks of $\boldsymbol{f}_Y^{re}(\omega)$ and $\boldsymbol{V}$, respectively. This choice corresponds to setting the last component of $\boldsymbol{\beta}(\omega)$ to zero.

The value $\lambda(\omega)$ itself has a useful interpretation; specifically, $\lambda(\omega)d\omega$ represents the largest proportion of the total power that can be attributed to the frequencies $\omega d\omega$ for any particular scaled process $X_t(\boldsymbol{\beta})$, with the maximum being achieved by the scaling $\boldsymbol{\beta}(\omega)$. Because of its central role, $\lambda(\omega)$ was defined to be the *spectral envelope* of a stationary categorical time series.

The name spectral envelope is appropriate since $\lambda(\omega)$ envelopes the standardized spectrum of any scaled process. That is, given any $\boldsymbol{\beta}$ normalized so that $X_t(\boldsymbol{\beta})$ has total power one, $f(\omega; \boldsymbol{\beta}) \le \lambda(\omega)$ with equality if and only if $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\beta}(\omega)$.

Although the law of the process $X_t(\boldsymbol{\beta})$ for any one-to-one scaling $\boldsymbol{\beta}$ completely determines the law of the categorical process X_t, information is lost when one restricts attention to the spectrum of $X_t(\boldsymbol{\beta})$. Less information is lost when one considers the spectrum of $\boldsymbol{Y}_t$. Dealing directly with the spectral density $\boldsymbol{f}_Y(\omega)$ itself is somewhat cumbersome since it is a function into the set of complex Hermitian matrices. Alternatively, one can view the spectral envelope as an easily understood, parsimonious tool for exploring the periodic nature of a categorical time series with a minimal loss of information.

The constraint that $\boldsymbol{\beta}(\omega)$ is real-valued leads to restricting attention to the real part of the spectrum as seen in (4). If we allow complex-valued scalings, then we would concentrate on the latent roots and vectors of the complex-valued spectral matrix function, $\boldsymbol{f}_Y(\omega)$. In this case, the problem is related to principal component analysis or canonical analysis of time series in the spectral domain as discussed in Brillinger (2001, Ch. 9, 10). Although Brillinger formulates the problem in terms of data compression, the problems are similar and the relationship is discussed in more detail in Stoffer, Tyler, & Wendt (2000, Sec. 7) As a note, we mention that this technique is not restricted to the use of sinusoids. In Stoffer et al. (1988), the use of the Walsh basis of square-waves functions that take only the values ±1 only, is described.

If we observe a finite realization of the stationary categorical time series X_t, or equivalently, the multinomial point process $\boldsymbol{Y}_t$, for $t = 1, \dots, n$, the theory for estimating the spectral density of a multivariate, real-valued time series is well established and can be applied to estimating $\boldsymbol{f}_Y(\omega)$, the spectral density matrix of $\boldsymbol{Y}_t$. Given an estimate $\widehat{\boldsymbol{f}}_Y(\omega)$ of $\boldsymbol{f}_Y(\omega)$, estimates $\widehat{\lambda}(\omega)$ and $\widehat{\boldsymbol{\beta}}(\omega)$ of the spectral envelope, $\lambda(\omega)$, and the corresponding scalings, $\boldsymbol{\beta}(\omega)$, can then be obtained. Estimation is discussed briefly in the next section.

2.1 Estimation

In view of the dimension reduction mentioned in the previous section, the easiest way to estimate the spectral envelope is to fix the scale of the last state at 0, and then select the indicator vectors to be k-dimensional. That is, to estimate the spectral envelope and the optimal scalings given a stationary categorical sequence, $\{X_t;\ t = 1, \dots, n\}$, with state-space $C = \{c_1, \dots, c_{k+1}\}$, perform the following tasks.

(i) Form $k \times 1$ vectors $\{\boldsymbol{Y}_t,\ t = 1, \dots, n\}$ as follows.

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{\epsilon}_j \qquad \text{if} \quad X_t = c_j\,, \quad j = 1, \dots, k\,; \\ \boldsymbol{Y}_t &= \boldsymbol{0} \qquad \text{if} \quad X_t = c_{k+1}\,, \end{aligned}$$

where $\boldsymbol{\epsilon}_j$ is a $k \times 1$ vector with a 1 in the j-th position as zeros elsewhere, and $\boldsymbol{0}$ is the $k \times 1$ vector of zeros.

(ii) Calculate the (fast) Fourier transform of the data,

$$\boldsymbol{d}(\omega_j) = n^{-1/2} \sum_{t=1}^{n} \boldsymbol{Y}_t \exp(-2\pi i t j/n)\,.$$

Note that $\boldsymbol{d}(\omega_j)$ is a $k \times 1$ complex-valued vector. Calculate the periodogram,

$$\boldsymbol{I}_n(\omega_j) = \boldsymbol{d}(\omega_j)\boldsymbol{d}^*(\omega_j)\,,$$

for $j = 1, \dots, \lfloor n/2 \rfloor$, where * denotes conjugate transpose.

(iii) Smooth the periodogram as preferred to obtain $\widehat{\boldsymbol{f}}_Y(\omega)$, a consistent estimator, and retain the real part of the spectral matrix estimate. Time series texts such as Shumway & Stoffer (2017) that cover the spectral domain will have an extensive discussion on consistent estimation of a spectral density. Most spectral density estimators can be written in the form

$$\widehat{\boldsymbol{f}}_Y(\omega) = \int_{-1/2}^{1/2} K_n(\omega - \lambda)\boldsymbol{I}_n(\lambda)d\lambda, \tag{5}$$

where $K_n(\omega)$ is the spectral window. The integral is typically approximated by a sum, and the amount of smoothing is controlled by the bandwidth (b_n) of the window.

(iv) Calculate the $k \times k$ covariance matrix of the data, $\boldsymbol{S} = n^{-1}\sum_{t=1}^{n}(\boldsymbol{Y}_t - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_t - \overline{\boldsymbol{Y}})'$, where $\overline{\boldsymbol{Y}}$ is the sample mean of the data.

(v) For each $\omega_j = j/n$, for $j = 1, \ldots, \lfloor n/2 \rfloor$, determine the largest eigenvalue and the corresponding eigenvector of the matrix $2n^{-1}\boldsymbol{S}^{-1/2}\widehat{\boldsymbol{f}}_Y^{re}(\omega_j)\boldsymbol{S}^{-1/2}$. Note that $\boldsymbol{S}^{-1/2}$ is the inverse of the unique square root matrix of $\boldsymbol{S}$.

(vi) The sample spectral envelope $\widehat{\lambda}(\omega_j)$ is the eigenvalue obtained in the previous step. If $\boldsymbol{b}(\omega_j)$ denotes the eigenvector obtained in the previous step, the optimal sample scaling is $\widehat{\boldsymbol{\beta}}(\omega_j) = \boldsymbol{S}^{-1/2}\boldsymbol{b}(\omega_j)$; this will result in k values, the $k+1$-st value being held fixed at zero.

Any standard programming language can be used to do the calculations. The R package `astsa` (Stoffer 2021), which supports the text (Shumway & Stoffer 2017), includes a script that will calculate the spectral envelope as well as additional scripts to handle various types of data files.

Under the conditions for which $\widehat{\boldsymbol{f}}_Y(\omega)$ has an asymptotic distribution (e.g., see Brillinger 2001, Thm. 7.4.4), if $\lambda(\omega)$ is a distinct root (which implies that $\lambda(\omega) > 0$), then, independently, for any collection of Fourier frequencies $\{\omega_i;\ i = 1, \ldots, M\}$, M fixed, and for large n and $\nu_n^2 \sim n\, b_n$ ($n, \nu_n \to \infty$ but $b_n \to 0$),

$$\nu_n \frac{\widehat{\lambda}(\omega_i) - \lambda(\omega_i)}{\lambda(\omega_i)} \sim \text{AN}(0, 1)\,. \tag{6}$$

For example, when estimation is accomplished by a symmetric moving average of the periodogram,

$$\widehat{\boldsymbol{f}}_Y(\omega) = \sum_{q=-r_n}^{r_n} h_q \boldsymbol{I}_n(\omega_{j+q})\,, \tag{7}$$

where $\{\omega_{j+q};\ q = 0, \pm 1, \ldots, \pm r_n\}$ is a band of frequencies and ω_j is the fundamental frequency closet to ω, and such that the weights satisfy $h_q = h_{-q} \geq 0$ and $\sum_{q=-r_n}^{r_n} h_q = 1$, then

$$\nu_n^{-2} = \sum_{q=-r_n}^{r_n} h_q^2\,.$$

If a simple average is used, $h_q = 1/(2r_n + 1)$, then $\nu_n^2 = (2r_n + 1)$ and the bandwidth is $b_n = \nu_n^2/n$. Based on these results, asymptotic normal confidence intervals and tests for $\lambda(\omega)$ can be readily constructed.

Significance thresholds for consistent spectral envelope estimates can easily be computed using the following approximations. Using a first order Taylor expansion we have

$$\log \widehat{\lambda}(\omega) \approx \log \lambda(\omega) + \frac{\widehat{\lambda}(\omega) - \lambda(\omega)}{\lambda(\omega)},$$

so that $(n, \nu_n \to \infty, b_n \to 0)$

$$\nu_n[\log \widehat{\lambda}(\omega) - \log \lambda(\omega)] \sim \text{AN}(0, 1). \tag{8}$$

It also follows that $\mathbf{E}[\log \widehat{\lambda}(\omega)] \approx \log \lambda(\omega)$ and $\text{var}[\log \widehat{\lambda}(\omega)] \approx \nu_n^{-2}$. If there is no signal present in a sequence of length n, we expect $\lambda(j/n) \approx 2/n$ for $1 < j < n/2$, and hence approximately $(1 - \alpha) \times 100\%$ of the time, $\log \widehat{\lambda}(\omega)$ will be less than $\log(2/n) + (z_\alpha/\nu_n)$ where z_α is the $(1 - \alpha)$ upper tail cutoff of the standard normal distribution. Although this method is a bit crude, from our experience, thresholding at very small α-levels (say, $\alpha = 10^{-4}$ to 10^{-6}, depending on the size of n) works well. Finally, we mention that inference for estimators of the scaling vectors $\boldsymbol{\beta}(\omega)$ is discussed extensively in Stoffer et al. (1993, Theorems 3.1–3.3).

2.2 An Example

As a simple example of the kind of analysis that can be accomplished, we consider the gene BNRF1 (3741 bp long) from Herpesvirus saimiri (HVS). Since we are considering the nucleotide sequence consisting of four bp, we use the following indicator vectors to represent the data:

$$\boldsymbol{Y}_t = (1, 0, 0)' \text{ if } X_t = \texttt{A}; \qquad \boldsymbol{Y}_t = (0, 1, 0)' \text{ if } X_t = \texttt{C};$$
$$\boldsymbol{Y}_t = (0, 0, 1)' \text{ if } X_t = \texttt{G}; \qquad \boldsymbol{Y}_t = (0, 0, 0)' \text{ if } X_t = \texttt{T},$$

so that the scale for the thymine nucleotide, T, is set to zero. Figure 3 shows the spectral envelope estimate of the entire coding sequence. The figure also shows a strong signal at frequency 1/3; the corresponding optimal scaling is A = 0.27, C = 0.56, G = 0.79, T = 0.0, which indicates the signal is not in terms of any alphabet that collapses the nucleotides such as the purine-pyrimidine (0–1) alphabet, which biotechnologists tend to use, would lead to wrong conclusions. To establish the heterogeneity of the gene, the bottom of Fig. 3 shows the spectral envelope for the first, second and third 1000 bp of the sequence. Clearly each segment has different

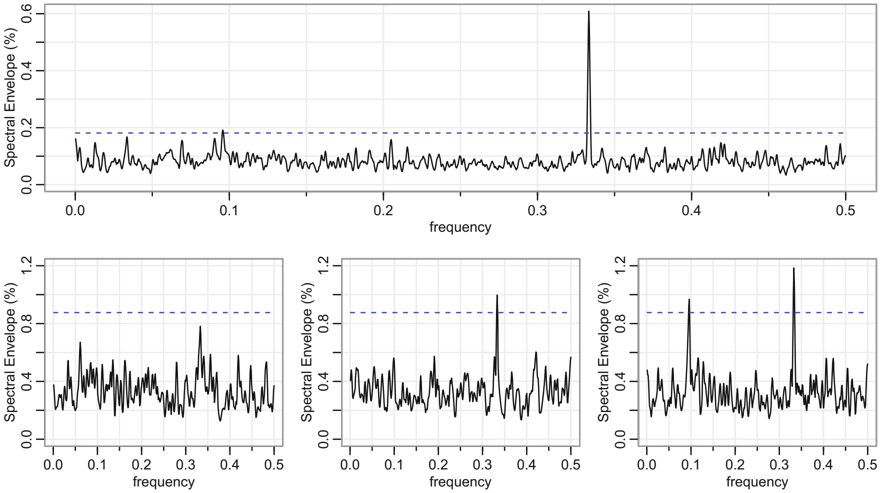


Fig. 3 TOP: Sample spectral envelope of the gene BNRF1 (3741 bp long) from the Herpesvirus saimiri. BOTTOM: Sample spectral envelope for the first, second and third 1000 bp of the gene

results; the first segment may be noise, the second segment has a peak at the period of 3 bp, and the third segment shows peaks at the 3 and 10 bp periods.

The period of 10 by may be attributed to *histones*, which are proteins that act as spools around which DNA winds and play a role in gene regulation. Bending occurs at an approximate period of 10 bp. The idea of rotational signals for nucleosome positioning is based on the fact that nucleosomal DNA is tightly wrapped around its protein core. The bending of the wound DNA requires compression of the grooves that face toward the core and a corresponding widening of the grooves facing the outside. Since, depending on the nucleotide sequence, DNA bends more easily in one plane than another, Trifonov & Sussman (1980) proposed that the association between the DNA sequence and its preferred bending direction might facilitate the necessary folding around the core particle.

Figure 4 shows a dynamic spectral envelope with a block size of 500. Evidently, even within small segments of the gene, it is not homogeneous. There is, however, a basic cyclic pattern that exists through most of the gene as evidenced by the peak at $\omega = 1/3$ except at the end of the gene. Table 3 shows the optimal scalings at the one-third frequency and we note that the corresponding alphabets can vary significantly. In addition, there are some parts of the gene where the 10 bp cycle exists. In the next section, we will develop a less ad hoc method.

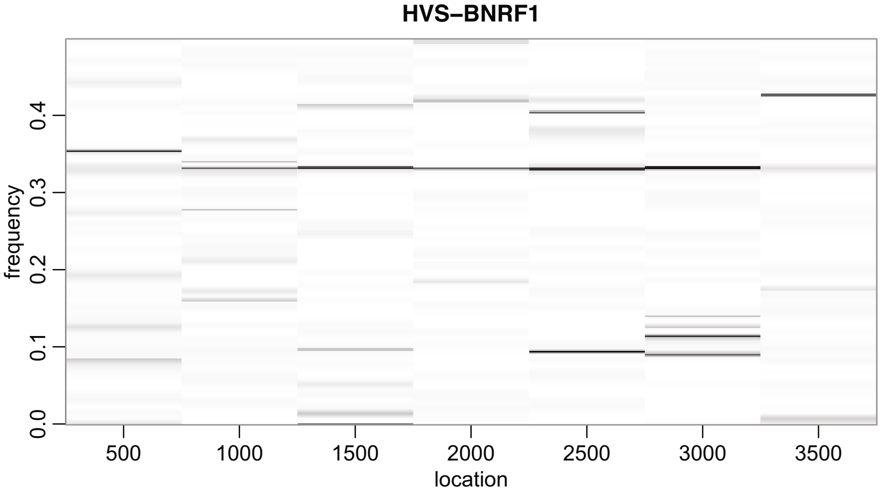


Fig. 4 Dynamic spectral envelope estimates for the BNRF1 gene of the Herpesvirus saimiri based on blocks of 500 bp. The horizontal axis indicates the location of the 500 bp blocks used to calculate the spectral envelope. Darker regions indicate larger values of the spectral envelope

Table 3 Blockwise (500 bp) optimal scaling, $\widehat{\boldsymbol{\beta}}(\frac{1}{3})$, for HVS-BNRF1

Block	A	C	G	T
$1^\dagger$	0.32	0.90	0.31	0
2	0.05	0.88	−0.47	0
3	0.14	0.59	0.80	0
4	0.33	0.62	0.71	0
5	0.49	−0.33	0.81	0
6	0.02	0.67	0.74	0
$7^\dagger$	0.33	0.42	0.85	0

† $\hat{\lambda}(\frac{1}{3})$ is not significant in this block

3 Local Analysis

Let a categorical-valued time series $\{X_t;\ t = 1,\dots,n\}$ consist of an unknown number of segments, m, and let ξ_j be the unknown location of the end of the jth segment, $j = 0, 1, \dots, m$, with $\xi_0 = 0$ and $\xi_m = n$. Then conditional on m and $\boldsymbol{\xi} = (\xi_0, \dots, \xi_m)'$, assume that the process $\{X_t\}$ is piecewise stationary. That is,

$$X_t = \sum_{j=1}^{m} X_{t,j}\, \delta_{t,j}\ , \tag{9}$$

where for $j = 1, \dots, m$, the indicator processes $\boldsymbol{Y}_{t,j}$ corresponding to $X_{t,j}$ have spectral density $\boldsymbol{f}_j(\omega)$ that may depend on parameters, and $\delta_{t,j} = 1$ if

$t \in [\xi_{j-1}+1, \xi_j]$ and 0 otherwise. The piecewise assumption is not very restrictive because slowly varying series may be approximated by a piecewise process (e.g., Adak 1998). Estimation is discussed in the next three subsections and an example is given in the fourth subsection.

3.1 Local Whittle Likelihood

An essential part of the local procedure is the calculation of the local likelihood. In our case, the estimation of the spectral matrix is done nonparametrically via kernel smoothing. Hence, Whittle's form of the likelihood (Whittle 1957) suits our analysis because it depends only on the Fourier transform of the data.

Consider a realization $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from process (9), where the breakpoints are known. Let n_j be the number of observations in the jth segment. We assume that the spectra are positive definite, and that each n_j is large enough for the local Whittle likelihood to provide a good approximation. Given a partition of the time series $\boldsymbol{x}$, the jth segment consists of the observations $\boldsymbol{x}_j = \{x_t : \xi_{j-1}+1 \leq t \leq \xi_j\}$, $j = 1, \ldots, m$, with underlying spectral densities $\boldsymbol{f}_j$ and Fourier transforms $\boldsymbol{d}_j$ evaluated at frequencies $\omega_{k_j} = k_j/n_j$, for $0 \leq k_j \leq n_j - 1$. For a given partition $\boldsymbol{\xi}$, the approximate likelihood of the time series is given by

$$L(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m \mid \boldsymbol{x}, \boldsymbol{\xi}) \approx \prod_{j=1}^{m} (2\pi)^{-n_j/2}$$

$$\times \prod_{k_j=0}^{n_j-1} \exp\left\{-\frac{1}{2}\left[\log|\boldsymbol{f}_j(\omega_{k_j})| + \boldsymbol{d}_j^*(\omega_{k_j})\boldsymbol{f}_j^{-1}(\omega_{k_j})\boldsymbol{d}_j(\omega_{k_j})\right]\right\}, \qquad (10)$$

where $|\cdot|$ denotes determinant. Conditional on the breakpoints, the local spectral envelope functions can be defined in terms of the local spectral matrix functions in an obvious manner.

3.2 Minimum Description Length

Here, we derive a minimum description length (MDL) criterion for choosing the best fitting model, where "best" is defined as the model that enables the best compression of the observed series $\boldsymbol{x} = \{x_1, \ldots, x_n\}$.

There are various versions of the minimum description length principle as put forth by Rissanen (1978, 1989) and the version adopted here is a two-part code. Let $\mathcal{F}$ denote the model and $\hat{\mathcal{F}}$ the fitted model. In this case, the first part, denoted by $\hat{\mathcal{C}}$, represents the *complexity* of the fitted model $\hat{\mathcal{F}}$, and the second part, denoted by $\hat{\mathcal{A}}$, represents the *accuracy* of the fitted model $\mathcal{F}$. The idea of the minimum description length principle is to find the best pair of $\hat{C}$ and $\hat{\mathcal{A}}$ so that via encoding (or compressing), $\boldsymbol{x}$ can be transmitted (or stored) with the least amount of codelength (or memory). To quantify this idea, let $\mathrm{CL}_{\mathcal{F}}(\cdot)$ denote the codelength of an object based on model $\mathcal{F}$. Then we have the decomposition

$$\mathrm{CL}_{\mathcal{F}}(\boldsymbol{x}) = \mathrm{CL}_{\mathcal{F}}(\hat{C}) + \mathrm{CL}_{\mathcal{F}}(\hat{\mathcal{A}} \mid \hat{C}) \tag{11}$$

for the data $\boldsymbol{x}$. This approach leads to familiar concepts such as BIC (Schwarz 1978) where model accuracy is measured by the negative of the log-likelihood evaluated at the estimated parameters, and complexity is based on the number of parameters in the model and the sample size.

For the complexity term in (11), we must consider the various parameters of the model, which includes the number of segments, m, the change points, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)$, and the individual bands in each segment, $B_1, \dots, B_m$ where $B_j \sim n_j b_{n_j}$ for $j = 1, \dots, m$, is the number of frequencies included in the smoothing band for segment j as described in (7). In this case we have

$$\mathrm{CL}_{\mathcal{F}}(\hat{C}) = \mathrm{CL}_{\mathcal{F}}(m) + \mathrm{CL}_{\mathcal{F}}(\xi_1, \dots, \xi_m \mid m) + \mathrm{CL}_{\mathcal{F}}(B_1, \dots, B_m \mid m, \boldsymbol{\xi}) \,. \tag{12}$$

The values B_j determine the number of distinct bands of frequencies for which $\boldsymbol{f}_j(\omega)$ is estimated. We mention that Hannan & Rissanen (1988) presented a method based on Rissanen (1978) to choose the bandwidth via stochastic complexity and minimum description length (MDL) in the case of stationarity. However, their approach is rarely used because it is overly complex and involves putting a prior on the value of spectral density in each band, which in turn depends on the bandwidth and leads to a somewhat circular argument.

To evaluate (12), the codelength for an integer m is $\log_2 m$ bits. For the second term, we note that knowledge of the breakpoints, ξ_j, is equivalent to knowledge of the number of observations in segment j, namely n_j. Noting that the n_j are bounded by the number of observations, n, we have a bound, $\mathrm{CL}_{\mathcal{F}}(n_j) = \log_2 n$ so that

$$\mathrm{CL}_{\mathcal{F}}(\xi_1, \dots, \xi_m \mid m) = \mathrm{CL}_{\mathcal{F}}(n_1, \dots, n_m \mid m) = m \log_2 n \,.$$

Each bandwidth value will cost about $\log_2 B_j$ bits. In addition, the bandwidth in each segment $j = 1, \dots, m$ is determined by maximizing the likelihood based on the segment data of n_j observations. For this, we can use a result of Rissanen that states a maximum likelihood estimate of p parameters computed from n_j observations can be effectively encoded with $\frac{1}{2} p \log_2 n_j$ bits, making the third term

$$\mathrm{CL}_{\mathcal{F}}(B_1, \dots, B_m \mid m, \boldsymbol{\xi}) = \log_2 B_j + \frac{p}{2} \sum_{j=1}^{m} \log_2 n_j \,,$$

where in this case, p is the number of parameters in the spectral matrix. For a k-dimensional spectral matrix, there are k real-valued parameters on the diagonal and $k(k-1)/2$ complex-valued parameters on the lower off-diagonals, each with one real and one imaginary part (the upper off-diagonals are the conjugates); hence,

$$p = k + k(k-1) = k^2 .$$

For the second term in (11), it is shown in Rissanen (1989) that the codelength of the accuracy term, $\hat{\mathscr{A}}$, is the negative of the $\log_2$ likelihood of the fitted model $\hat{C}$. In our case, we use the Whittle likelihood approximation given in (10).

Combining the results and working with natural log instead of base 2, we obtain an approximation to the MDL of the model,

$$\begin{aligned} \mathrm{MDL} = {} & \log m + m \log n + \sum_{j=1}^{m} \log B_j + \frac{k^2}{2} \sum_{j=1}^{m} \log n_j \\ & + \sum_{j=1}^{m} \left\{ \frac{n_j}{2} \log(2\pi) + \frac{1}{2} \sum_{k_j=0}^{n_j-1} \left[\log |\boldsymbol{f}_j(\omega_{k_j})| + \boldsymbol{d}_j^*(\omega_{k_j}) \boldsymbol{f}_j^{-1}(\omega_{k_j}) \boldsymbol{d}_j(\omega_{k_j}) \right] \right\} . \end{aligned} \tag{13}$$

3.3 Optimization via Genetic Algorithm

Because the search space is enormous, optimization is a nontrivial task, we use a genetic algorithm (GA) to effectively tackle the problem. A tutorial may be found in Whitley (1994). In addition, Matlab has a toolbox with supporting videos demonstrating GAs that are also good references (Mathworks 2021). Our GA is similar to the one specified in Davis et al. (2006) who used it to fit local autoregressions to nonstationary univariate time series.

Briefly, genetic algorithms are a class of iterative optimization methods that use the principles of evolutionary biology. The algorithm typically begins with some initial randomly chosen population and each generation afterwards produces an offspring population using genetic operators. Genetic operators include selection, recombination or crossover, and mutation, which are based on the principle of natural selection to find the best solution while using the principle of diversity to avoid convergence to a local minima.

There are many variations of a genetic algorithm (GA). For example, parallel implementations can be applied to speed up the convergence rate as well as to reduce the chance of converging to suboptimal solutions (Alba et al. 1999). We implement an Island model, where instead of running only one search in one giant population, we simultaneously runs NI (Number-of-Islands) canonical GAs in NI different sub-populations. The key feature is that a number of individuals are migrated among the islands according to some migration policy. The migration can be implemented in numerous ways (e.g., Alba & Troya 2002) and here we adopt the migration policy that after every M_i generations, the worst M_N chromosomes from the j-th island are replaced by the best M_N chromosomes from the $(j-1)$-st island, for $j = 1, \dots, NI$. For $j = 1$, the best M_N chromosomes are migrated from the NI-th island. In our examples, we used $NI = 40$, $M_i = 5$, $M_N = 2$, and a sub-population size of 40.

Chromosome Representation The performance of a genetic algorithm depends on how a possible solution is represented as a chromosome. For our problem, a chromosome should carry complete information for any model $\mathcal{F}$; i.e., the number of segments m, the breakpoints ξ_j, and the segment bands B_j. Once these parameters are specified, the Whittle likelihood is uniquely determined. For our problem, a chromosome $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$ is of length n with gene values δ_t defined as $\delta_t = -1$ if there is not a breakpoint at position t, and $\delta_t = B_j$ if $t = \xi_{j-1}$ and the band of the j-th piece is B_j. Furthermore, any band size, B_j, is limited to $P_0 = 2r_0 + 1 = 21$ (or 10 fundamental frequencies on either side of the center frequency) and a minimum span on the n_j, ranging from 30 to 70 is specified depending on the size of the band.

Initial Population Generation The GAs start with an initial population of random chromosomes, and the following strategy was used to generate each of them. First, select a value for $B_1 \in \{0, \dots, P_0\}$ with equal probabilities and set $\delta_1 = B_1$. Then the next $n_{j_1} - 1$ genes $\delta_2, \dots, \delta_{n_{j_1}}$ are set to -1 so that the minimum span constraint is imposed for this first piece. The next gene $\delta_{n_{j_1}+1}$ in line will either be initialized as a breakpoint with probability π, or it will be assigned -1 with probability $1-\pi$. If it is to be initialized as a breakpoint, then we set $\delta_{n_{j_1}} = r_2$, where r_2 is randomly drawn from $\{0, \dots, P_0\}$. Otherwise, if $\delta_{n_{n_1}}$ is to be assigned -1, the initialization process will move to the next gene in line and decide if this gene should be a breakpoint gene. This process continues in a similar fashion, and a random chromosome is generated when the process hits the last gene δ_n. In the example, we set $\pi = 10/n$ where n is the length of the sequence.

Crossover and Mutation Once a set of initial random chromosomes is generated, new chromosomes are generated by either a crossover or a mutation operation. In our implementation we set the probability for conducting a crossover operation as $1 - \pi$. For the crossover operation, two parent chromosomes are chosen from the current population. The parents are chosen with probabilities inversely proportional to their ranks sorted by their MDL values so that chromosomes having smaller MDL values have a higher chance of being selected. From these two parents, the gene values δ_t of the child chromosome is inherited as follows. First, δ_1 will take on the corresponding value from either the first or second parent with equal probabilities. If the value is -1, then the same gene-inheriting process will be repeated for the next gene in line. Otherwise, the bandwidth is that of the current piece with the minimum span constraint imposed. The same gene-inheriting process will be applied to the next available δ_t.

For mutation, one child is reproduced from one parent. The process starts with $t = 1$ and every δ_t can take on one of the following three values: (i) with probability π_r it will take the corresponding δ_t value from the parent, (ii) with probability π_N it will take the value -1, or (iii) with probability $1 - \pi_r - \pi_N$, it will take a randomly generated bandwidth (subject to the constraints). In our example in the next section, we set $\pi_r = \pi_N = 0.3$.

Declaration of Convergence In our example, we use the Island Model in which migration is allowed for every $M_i = 5$ generations. At the end of each migration the overall best chromosome is noted. If this best chromosome does not change for 10 consecutive migrations, or the total number of migrations exceeds 20, this best chromosome is taken as the solution to this optimization problem.

3.4 Another Example

Here, we focus on an analysis of the CDS BNRF1 of the Epstein–Barr virus (EBV), which is roughly 4000 bp long. We selected a subsequence of length $n = 1000$ starting at bp 2500 of the CDS. We chose this section because we know from previous experience (Stoffer, Tyler, & Wendt 2000) that, while most of the CDS contains a signal of period 3, there is a part that appears to be noise. We kept the choice of kernel and corresponding bandwidth simple in that we used the modified Daniell kernel (which is the default in R) with two passes, but allowing the bandwidth to grow. The Daniell kernel corresponds to simple averaging. The modified version simply puts half weights at the ends. For example, if $r = 1$ in (7), the modified Daniell weights are (1/4, 1/2, 1/4). Passing those again yields weights (1/16, 4/16, 6/16, 4/16, 1/16). If one thinks of the first set of weights as a discrete distribution of a random variable with support $\{-1, 0, 1\}$, then the second pass is the distribution of the sum of two independent random variables with that distribution. Thus, in the GA, the value of r in a segment is allow to grow. In this case we used an approximation suggested by Tukey (1950) to obtain the band of the kernel in segment j as $B_j = \nu^2_{r_j}$ in the notation below (7), where $2\,r_j + 1$ is the width of the band is segment j. In the case of simple averaging, the value is $B_j = 2r_j + 1$.

The GA found two breakpoints at $t = 456$ and 851. Figure 5 shows the estimated spectral envelope for each of the three segments as well as for the entire sequence along with significance thresholds. The segment locations appear in the upper right of each plot and the significance threshold used in the figure is 0.0001 for all plots. The first segment shows the typical 3 bp cycle, which was seen in the first example using HVS-BNRF1 and again in the spectral envelope of the entire sequence shown at the bottom right (marked ‘ALL’). The large values near the zero frequency appear to indicate fractional noise, which is not unusual for DNA sequences; e.g., see Voss (1993). The second and third segments appear to be noise, but the variability in the third segment is slightly larger than the second; in addition, there appears to be fractional noise in the third segment.

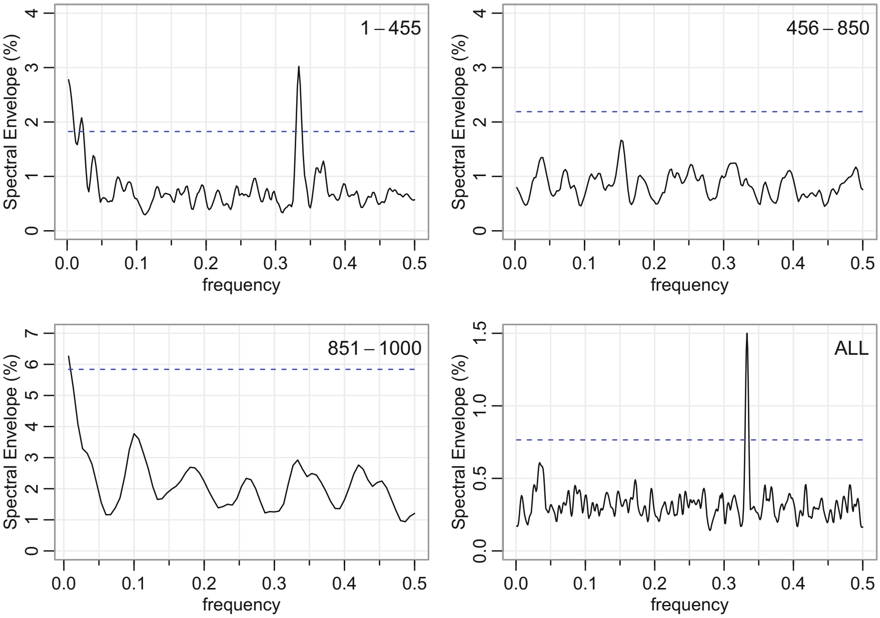


Fig. 5 The estimated spectral envelopes for the various segments found by the genetic algorithm in a section of 1000 bp of EBV-BNRF1. The values in the upper right corner are the locations of the segments. The horizontal dashed lines are 0.0001 significance threshold as discussed after (8). The graphic on the bottom right is the spectral envelope for the entire 1000 bp and the corresponding threshold is the 0.0001 level

Data Availability

The sleep data used in the introduction is included in the R package `astsa` (Stoffer 2021) and it is from the first subject included in the data frame `sleep1`. The DNA sequences used throughout this manuscript may be found online at the National Center for Biotechnology Information (NCBI). The Epstein–Barr virus sequence may be found at NCBI (2021a) and the Herpesvirus saimiri sequence may be found at NCBI (2021b). The EBV sequence is also included as a data set in `astsa` as is the CDS BNRF1 of each virus (as `bnrf1ebv` and `bnrf1hvs`).

References

Adak, S. (1998). Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association, 93*(444), 1488–1501.

Alba, E., & Troya, J. M. (2002). Improving flexibility and efficiency by adding parallelism to genetic algorithms. *Statistics and Computing, 12*(2), 91–114.

Alba, E., Troya, J. M., et al. (1999). A survey of parallel distributed genetic algorithms. *Complexity, 4*(4), 31–52.

Brillinger, D. (2001). *Time Series: Data Analysis and Theory* (vol. 36). New York: Society for Industrial Mathematics.

Davis, R., Lee, T., & Rodriguez-Yam, G. (2006). Structural breaks estimation for nonstationary time series models. *Journal of the American Statistical Association, 101*, 223–239.

Hannan, E., & Rissanen, J. (1988). The width of a spectral window. *Journal of Applied Probability, 25*, 301–307.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *Journal of Molecular Biology, 262*(2), 129–139.

Mathworks (2021). MATLAB Global Optimization Toolbox. https://www.mathworks.com/videos/what-is-a-genetic-algorithm-100904.html.

McDougall, A., Stoffer, D., & Tyler, D. (1997). Optimal transformations and the spectral envelope for real-valued time series. *Journal of Statistical Planning and Inference, 57*(2), 195–214.

NCBI (2021a). Epstein-Barr virus (EBV) genome, strain B95-8 - Nucleotide - National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/nuccore/V01555.

NCBI (2021b). Saimiriine herpesvirus 2 complete genome - Nucleotide - National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/nuccore/NC_001350.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465–471.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry* (vol. 15). Singapore: World Scientific.

Satchwell, S. C., Drew, H. R., & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology, 191*(4), 659–675.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Shumway, R., & Stoffer, D. (2017). *Time Series Analysis and Its Applications: With R Examples*, 4th ed.. New York: Springer.

Stoffer, D. S. (2021). *astsa: Applied Statistical Time Series Analysis*. R package version 1.14.3. https://github.com/nickpoison/astsa

Stoffer, D. S., Scher, M. S., Richardson, G. A., Day, N. L., & Coble, P. A. (1988). A Walsh–Fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. *Journal of the American Statistical Association, 83*(404), 954–963.

Stoffer, D. S., Tyler, D. E., & McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika, 80*(3), 611–622.

Stoffer, D. S., Tyler, D. E., & Wendt, D. A. (2000). The spectral envelope and its applications. *Statistical Science 15*(3), 224–253.

Trifonov, E. N., & Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences, 77*(7), 3816–3820.

Tukey, J. (1950). The sampling theory of power spectrum estimates. In *Symposium on Applications of Autocorrelation Analysis to Physical Problems: US Office of Naval Research* (pp. 47–67).

Voss, R. F. (1993). $1/f$ Noise and Fractals in DNA-base Sequences. In *Applications of Fractals and Chaos* (pp. 7–20). Berlin: Springer.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing, 4*(2), 65–85.

Whittle, P. (1957). Curve and periodogram smoothing. *Journal of the Royal Statistical Society B, 19*, 38–47.