Emerging Topics in Statistics and Biostatistics

Andriëtte Bekker
Johannes T. Ferreira
Mohammad Arashi
Ding-Geng Chen *Editors*

Innovations in Multivariate Statistical Modeling

Navigating Theoretical and Multidisciplinary Domains

Springer

Emerging Topics in Statistics and Biostatistics

Andriëtte Bekker · Johannes T. Ferreira ·
Mohammad Arashi · Ding-Geng Chen
Editors

Innovations in Multivariate Statistical Modeling

Navigating Theoretical and Multidisciplinary Domains

Springer

Editors
Andriëtte Bekker
Department of Statistics
University of Pretoria
Pretoria, South Africa

Johannes T. Ferreira
Department of Statistics
University of Pretoria
Pretoria, South Africa

Mohammad Arashi
Department of Statistics
Faculty of Mathematical Sciences
Ferdowsi University of Mashhad
Mashhad, Iran

Ding-Geng Chen
College of Health Solutions
Arizona State University
Phoenix, AZ, USA

ISSN 2524-7735 ISSN 2524-7743 (electronic)
Emerging Topics in Statistics and Biostatistics
ISBN 978-3-031-13970-3 ISBN 978-3-031-13971-0 (eBook)
https://doi.org/10.1007/978-3-031-13971-0

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

Preface

Multivariate statistical analysis has undergone a rich and varied evolution during the latter half of the twentieth century. Academics and practitioners have produced much literature with diverse interests and with varying multidisciplinary knowledge on different topics within the multivariate domain. Due to multivariate algebra being of sustained interest and being a continuously developing field, its appeal breaches laterally across multiple disciplines to act as a catalyst for contemporary advances, with its core inferential genesis remaining in that of statistics.

It is exactly this varied evolution caused by an influx in data production, diffusion and understanding in scientific fields that has blurred many lines between disciplines. The cross-pollination between statistics and biology, engineering, medical science, computer science and even art, has accelerated the vast amount of questions that statistical methodology has to answer and report on. These questions are often multivariate in nature, hoping to elucidate uncertainty on more than one aspect at the same time—and it is here where statistical thinking merges mathematical design with real-life interpretation for understanding this uncertainty.

Statistical advances thus benefit from these algebraic inventions and expansions in the multivariate paradigm. This contributed volume aims to usher novel research emanating from a multivariate statistical foundation into the spotlight, with particular significance in multidisciplinary settings. The overarching spirit of this volume is to highlight current trends, stimulate a focus on and connect multidisciplinary dots from and within multivariate statistical analysis. Guided by these thoughts, a collection of research at the forefront of multivariate statistical thinking is presented here which has been authored by globally recognized subject matter experts.

Outline of This Book Volume

This contributed volume brings together 19 chapters that are organized as follows: Trends in Multi- and Matrix-Variate Analysis (Part One), Aspects of High-Dimensional Methodology and Bayesian Learning (Part Two) and Frontiers in Robust

Analysis and Mixture Modelling (Part Three). All the chapter contributions have undergone a thorough and independent review process.

Part One of this book includes 12 papers focusing on recent trends in multi- and matrix-variate analyses. In Chapter "Association-Based Optimal Subpopulation Selection for Multivariate Data," the authors propose a semiparametric statistical approach for the optimal subpopulation selection based on the patterns of associations in multivariate data. Mattos, Matos and Lachos relaxes the normal assumption for linear mixed-effects models with a censored response by considering the multivariate skew-normal distribution in Chapter "Likelihood-Based Inference for Linear Mixed-Effects Models with Censored Response Using Skew-Normal Distribution". Chapter "Robust Estimation of Multiple Change Points in Multivariate Processes" contains the proposition of a novel likelihood-based technique to identify multiple change points in multivariate processes.

In Chapter "Some Computational Aspects of a Noncentral Dirichlet Family", the authors explore computational issues when considering the estimation of the singly and doubly noncentral Dirichlet distribution. Chapter "Modeling Handwritten Digits Dataset Using the Matrix Variate t Distribution" discusses the implementation of the matrix variate t distribution and computational aspects of the resulting EM algorithm when applied to modelling handwritten digits. Byukusenge and coauthors provide refreshing contributions to matrix residuals of the GMANOVA-MANOVA model in Chapter "On the Identification of Extreme Elements in a Residual for the GMANOVA-MANOVA Model".

In Chapter "Matrix-Variate Smooth Transition Models for Temporal Networks", Billio et al. study matrix-valued panel data characterized by nonlinear dynamics and heavy tails. Chapter "A Flexible Matrix-Valued Response Regression for Skewed Data" sees a contribution of a new flexible family of matrix-variate distributions and is implemented with matrix-variate regression by Baghishani and Ownuk. A nonparametric approach for analysing multivariate functional time series with spectrum analysis is discussed in Chapter "Multivariate Functional Singular Spectrum Analysis: A Nonparametric Approach for Analyzing Multivariate Functional Time Series". Greenacre discusses issues emanating from compositional data analysis when viewed from a matrix-vector representation in Chapter "Compositional Data Analysis—Linear Algebra, Visualization and Interpretation". Alzaatreh, Famoye and Lee define multivariate regression models accounting for positive and negative correlation in Chapter "Multivariate Count Data Regression Models and Their Applications". Finally, in Part One, Chapter "A Generalized Multivariate Gamma Distribution" introduces a multivariate gamma distribution whose marginals are finite mixtures of gamma distributions, by Iranmanesh, Rafiei and Nagar.

Part Two comprises four chapters that highlights aspects of high-dimensional modelling and methodology. In Chapter "A Comparison of Different Clustering Approaches for High-Dimensional Presence-Absence Data", d'Angella and Hennig compare the performance of different clustering methods in high-dimensional presence–absence data environments. Millard et al. utilize the modified elastic net in Chapter "High-Dimensional Feature Selection for Logistic Regression Using Blended Penalty Functions" to deal with high-dimensional problems with highly

correlated predictor variables. Munaweera and coauthors provide an extension of the non-negative garrote method for greater flexibility in ridge regression in the case of unequal shrinkage of regression coefficients in Chapter "A Generalized Quadratic Garrote Approach Towards Ridge Regression Analysis", and Roozbeh eliminates structured noises to improve the prediction performance of the LASSO method in a semiparametric regression framework in Chapter "High Dimensional Nonlinear Optimization Problem in Semiparametric Regression Model".

Part Three includes three chapters that focus on the recent emphasis on robust analysis and mixture modelling. Punzo and Tomarchio apply the eigen decomposition to covariance matrices to design parsimonious finite mixtures of matrix-variate regressions in Chapter "Parsimonious Finite Mixtures of Matrix-Variate Regressions". In Chapter "Robust Multivariate Modelling for Heterogeneous Data Sets with Mixtures of Multivariate Skew Laplace Normal Distributions", the authors consider finite mixtures of multivariate skew Laplace normal distributions with accompanying EM algorithm for implementation. Chapter "Robust Estimation Through Preliminary Testing Based on the LAD-LASSO" concludes the volume with the proposition of a new estimator for sparse and robust regression that improves the preliminary test least absolute deviation estimator.

Acknowledgements

We are grateful to those who have supported the process of creating this book. We thank the contributors to this book for their enthusiastic involvement and their kindness in sharing their professional knowledge and expertise. We thank all the reviewers for providing thoughtful and in-depth evaluations of the papers contained in this book. We gratefully acknowledge the professional support of Laura Briskman, Eva Hirip, Jayanthi Narayanaswamy and Faith Su from Springer who made the publication of this book a reality.

This volume was made possible through funding provided by the following agencies:

1. DST-NRF-SAMRC-SARChI Research Chair in Biostatistics, grant number: 114613;
2. National Research Foundation (NRF), grant ref. SRUG190308422768 number 120839;
3. Centre of Excellence in Mathematics and Statistical Science, University of the Witwatersrand, South Africa; and

Opinions expressed and conclusions arrived at are those of the author(s) and are not necessarily to be attributed to the NRF.

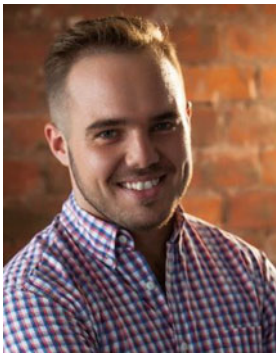
We welcome readers' comments, including notes on typos or other errors, and look forward to receiving suggestions for improvements to future editions. Please send comments and suggestions to any of the editors.

Pretoria, South Africa — Andriëtte Bekker
Pretoria, South Africa — Johannes T. Ferreira
Mashhad, Iran — Mohammad Arashi
Phoenix, USA — Ding-Geng Chen
November 2022

Contents

About the Editors

Andriëtte Bekker is a full professor in Statistics and has been the Head of the Department of Statistics at the Faculty of Natural and Agricultural Sciences, at the University of Pretoria since July 2012 until February 2022. Her expertise lies in distribution theory learning, unsupervised learning, network learning and software learning. She is the academic research leader of the Statistical Theory and Applied Statistics focus area within the Department of Science as well as in the Technology/National Research Foundation (DST-NRF) Centre of Excellence in Mathematical and Statistical Sciences. She is also an elected member of the International Statistical Institute and serves on several national and international, educational and academic research boards. She has published more than 100 peer-reviewed papers in fundamental statistical research. She contributes significantly to the human capacity development of Southern Africa through the multitude of students for whom she acts as supervisor and mentor.



Johannes T. Ferreira is an associate professor in the Department of Statistics at the University of Pretoria, South Africa, and is Junior Focus Area Coordinator for the Statistical Theory and Applied Statistics focus area of the Centre of Excellence in Mathematical and Statistical Science based at the University of the Witwatersrand in Johannesburg. He regularly publishes in accredited peer-reviewed journals and reviews manuscripts for international journals. He is an ASLP 4.1/4.2 fellow of Future Africa and has been identified as one of the Top 200 South Africans under the age of 35 by the Mail & Guardian newspaper in the Education category in 2016. His ramblings can be found on Twitter with the handle @statisafrican.

Mohammad Arashi is an associate professor and Director of the Data Science Laboratory at the Ferdowsi University of Mashhad in Iran and an Extraordinary Professor at the University of Pretoria in South Africa. He is the author of three books with Wiley and an elected member of ISI. He has published more than 150 papers and his major includes shrinkage estimation, variable selection, high-dimensional statistics, graphical models and longitudinal data analysis.

Ding-Geng Chen is an elected fellow of the American Statistical Association and an elected member of the International Statistical Institute. He is currently the executive director and professor in biostatistics at the College of Health Solutions, Arizona State University, USA. He is also the DST-NRF-SAMRC research chair in Biostatistics and an Extraordinary Professor at the Department of Statistics, University of Pretoria, South Africa. He was the Wallace H. Kuralt Distinguished Professor at the University of North Carolina at Chapel Hill, a Professor in Biostatistics at the University of Rochester and the Karl E. Peace endowed eminent scholar chair in biostatistics at Georgia Southern University. He is also a senior consultant for biopharmaceuticals and government agencies with extensive expertise in clinical trial biostatistics and public health statistics. He has written more than 200 refereed publications and co-authored/co-edited 33 books on clinical trial methodology, meta-analysis, causal inference and public health statistics.

Trends in Multi- and Matrix-Variate Analysis

Association-Based Optimal Subpopulation Selection for Multivariate Data

Qing Guo, Xinwei Deng, and Nalini Ravishanker

Abstract In the analysis of multivariate data, a useful problem is to identify a subset of observations for which the variables are strongly associated. One example is in driving safety analytics, where we may wish to identify a subset of drivers with a strong association among their driving behavior characteristics. Other interesting domains include finance, health care, marketing, etc. Existing approaches, such as the Top-k method or the tau-path approach, primarily relate to bivariate data and/or invoke the normality assumption. Directly adapting these methods to the multivariate framework is cumbersome. In this work, we propose a semiparametric statistical approach for the optimal subpopulation selection based on the patterns of associations in multivariate data. The proposed method leverages the concept of general correlation coefficients to enable the optimal selection of a subpopulation for a variety of association patterns. We develop efficient algorithms consisting of sequential inclusion of cases into the subpopulation. We illustrate the performance of the proposed method using simulated data and an interesting real data.

1 Introduction

An interesting and useful problem that arises in the analysis of multivariate data of any size is the identification of a small subset of the entire data, in which the variables are strongly associated. We denote such a problem as *the association-based optimal*

Q. Guo
Department of Statistics, Virginia Tech, Blacksburg, VA, USA
e-mail: qguo0701@vt.edu

X. Deng
Department of Statistics, Virginia Tech, Blacksburg, VA, USA
e-mail: xdeng@vt.edu

N. Ravishanker (✉)
Department of Statistics, University of Connecticut, Mansfield, CT, USA
e-mail: nalini.ravishanker@uconn.edu

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_1

subdata selection. One example is in transportation safety analytics [30], where we may wish to identify a subset of locations with a strong association between their roadway characteristics and traffic accidents. Another example is in epidemiology, where we are interested in finding a subset of health districts whose COVID-19 characteristics (e.g., the number of hospitalizations) are highly correlated. Thus, searching for an optimal subdata in terms of association from a potentially large dataset can be very useful for knowledge discovery, providing data insight, as well as improving the efficiency of data analysis and decision-making.

However, it is not a trivial problem to extract an optimal subset from the whole dataset with multivariate dimensions. First, one needs to define a suitable optimality criterion to quantify the association among variables in the multivariate data. It is well known that the Pearson correlation is mainly for the linear association, while nonparametric rank-based measures of association, such as Kendall's τ or Spearman's ρ can be computationally expensive. Moreover, there can be a large number of pairwise associations as the dimension of the multivariate data increase. It will be very useful to define a good statistic to quantify the overall association. Second, unlike some analyses of multivariate data which correspond to supervised learning, the problem of selecting an optimal subset is in general unsupervised and is important to take scalability into consideration. Third, selecting a subset of data points is a combinatorial optimization issue due to the discrete nature of the problem. Thus, it is crucial to find a suitable algorithm to make the subset selection efficient, with high accuracy.

In the literature, the problem of subdata selection in multivariate analysis has been investigated from different angles. In the *design of experiments* literature [26], the area of optimal design [12] is closely related to the optimal subdata selection. However, the criterion of searching for an optimal design is often based on certain parametric regression models, and often a poll of candidate design points needs to be generated for the search. By contrast, the subdata selection problem does not assume a response variable. In the literature on *data filtering* to reduce the data size, also, the objective is to find a subset of the data which contains as much information as the full data. For example, [11] recently proposed a clustering-based data filtering for big data systems in manufacturing, where subsampling is conducted for data points in each cluster. Along this direction [17, 27], other approaches included probability-based filtering such as random sampling [6, 16] and stratified sampling [13, 25]. Tau-path is another approach in the literature used to identify a subset of the observations, such that the association defined by Kendall's τ between two variables is maximized [21, 29, 30]. However, their approach cannot be directly extended to the multivariate framework, and further it is restricted to Kendall's τ as the measure of the association.

In this work, we propose a semiparametric statistical approach for optimal subdata selection based on the patterns of associations in multivariate data. The proposed method leverages the concept of *general correlation coefficients* [10] to enable the optimal selection of a subpopulation for a variety of association patterns, including the Pearson correlation, Kendall's τ, and Spearman's rank correlation. Specifically, we propose an averaged absolute association (AAA) criterion for finding the association-based optimal subdata from a multivariate dataset. This criterion is

similar in spirit to the root average squared correlation in the experimental design literature [2, 14], where the Pearson correlation is used. Our proposed criterion is more general in the sense that it is not limited to the Pearson correlation, but can be applied to any general association. The use of the absolute value of the general correlation coefficient for each pair of variables can also make the criterion more robust in comparison with the use of the squared values. Moreover, we develop efficient algorithms for the proposed criterion non-convex AAA criterion. Three algorithms, i.e., a genetic algorithm (GA), a forward selection algorithm, and a backward selection algorithm, are developed based on the AAA criterion. Through sequential inclusion and exclusion of data points into the subpopulation, these three algorithms are able to find an optimal subpopulation with different characteristics. We show that the backward algorithm appears to be most efficient computationally, while the genetic algorithm has the advantage of escaping from local optimal solutions. Note that the genetic algorithm has been successfully used in the design of experiments framework [15]. The forward algorithm is slow, but in an exact situation, it can provide better subpopulation solutions than the GA and backward algorithms.

The remainder of the paper is organized as follows. Section 2 details the proposed AAA method. We show a set of simulation-based results to examine the performance of the proposed method in Sect. 3. Section 4 shows a case study on multivariate COVID-19 data in Virginia. We conclude this paper with a discussion in Sect. 5.

2 The Proposed Method

Let $X = (X_1, \ldots, X_p)$ denote a p-dimensional random vector. Suppose that the observed data are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$. Denote the full data matrix as the $n \times p$ matrix $\mathbf{X}_F = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, which is an $n \times p$ matrix. Corresponding to n independent subjects and p exchangeable variables, the goal is to select a subset of the rows of the data with size between m_l and m_u, such that the association of the p variables on the selected subset is maximized.

Averaged Absolute Association (AAA) Criterion

Let $\mathbf{X}_f$ be an $m \times p$ submatrix which is formed by choosing m out of n rows in $\boldsymbol{X}_F$. We formulate the problem as

$$\begin{aligned} &\max_{\mathbf{X}_f \subset \mathbf{X}_F} h\big(A(\mathbf{X_f})\big) \\ &\text{s.t. } m_l \leq \text{size}(\mathbf{X}_f) \leq m_u, \end{aligned} \tag{1}$$

where $A(\mathbf{X_f})$ is a $p \times p$ matrix consisting of the associations among the p variables across the m units, and $h(\cdot)$ is a function that yields a suitable scalar metric of $A(\mathbf{X_f})$.

We describe the idea of the proposed method in the context of a general measure of association.

We first define a criterion for evaluating the associations, by looking at the p variables pairwise. Let (x_i, y_i) and (x_j, y_j) denote observations on any pair of individuals i and j, for $i \neq j, i, j \in \{1, \ldots, n\}$. Let $u_{ij}^{(x)}$ and $u_{ij}^{(y)}$ be the score that we assign on the x and y dimensions, respectively. Then, the general correlation coefficient [10] can be expressed as

$$\tau(x, y) = \frac{\sum_{i,j} u_{ij}^{(x)} u_{ij}^{(y)}}{\sqrt{\sum_{i,j} (u_{ij}^{(x)})^2 \sum_{i,j} (u_{ij}^{(y)})^2}}. \tag{2}$$

Note that various correlations, such as Spearman's rank correlation, Pearson correlation, and Kendall's τ, can be viewed as special cases of the general correlation coefficient. Let r_i and s_i denote the ranks of (x_i, y_i) in the x-dimension and y-dimension, respectively. One can see that the expression in (2) is Kendall's τ coefficient if $u_{ij}^{(x)} = \text{sign}(r_i - r_j)$ and $u_{ij}^{(y)} = \text{sign}(s_i - s_j)$. Also, if $u_{ij}^{(x)} = r_i - r_j$ and $u_{ij}^{(y)} = s_i - s_j$, then the expression in (2) is Spearman's ρ.

Proposition 1 *Consider a set of bivariate vectors* $(x_1, y_1), \ldots, (x_n, y_n)$. *For the general correlation coefficient in (2), we have the following:*

(a) $\tau(x, y)$ *will be Kendall's* τ *correlation coefficient if* $u_{ij}^{(x)} = sign(r_i - r_j)$ *and* $u_{ij}^{(y)} = sign(s_i - s_j)$.

(b) $\tau(x, y)$ *will be Spearman's* ρ *correlation coefficient if* $u_{ij}^{(x)} = r_i - r_j$ *and* $u_{ij}^{(y)} = s_i - s_j$.

(c) $\tau(x, y)$ *will be Pearson's r correlation coefficient if* $u_{ij}^{(x)} = x_i - x_j$ *and* $u_{ij}^{(y)} = y_i - y_j$.

Moreover, even when the entries in $\boldsymbol{X}_F$ are categorical, the general correlation coefficient can still be used to quantify the association among multivariate variables. For example, the ϕ coefficient [7], defined as a measure of association between two binary variables, is a special case of the general correlation coefficient as formalized by the following statement.

Corollary 1 *Consider a set of bivariate binary variables* $(x_1, y_1), \ldots, (x_n, y_n)$ *where* $x_i \in \{0, 1\}$ *and* $y_i \in \{0, 1\}$. *The general correlation coefficient* $\tau(x, y)$ *in (2) with* $u_{ij}^{(x)} = x_i - x_j$ *and* $u_{ij}^{(y)} = y_i - y_j$ *is equivalent to the* ϕ *correlation, i.e.,*

$$\tau(x, y) = \phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}}, \tag{3}$$

where n_{hl} *is the number of points with* (h, l) *for* $h, l = 0, 1$, *and* $n_{l\cdot}$ *is the number of data points with* $x_i = l$, *for* $l = 0, 1$. *The* $n_{1\cdot}$ *is the number of data points with* $x_i = 1$. *Similar definitions apply for* $n_{0\cdot}, n_{\cdot 1}$, *and* $n_{\cdot 0}$.

Next, we construct a suitable scalar metric based on the several core coefficients in order to measure the association among the p variables. Without loss of generality, we assume that the data points in the subpopulation are stacked row by row to form a $m \times p$ matrix $\mathbf{X_f}$. Let the $x_k^{(f)}$ be the kth column of $\mathbf{X_f}$. The association among the p variables can be quantified by

$$A(\mathbf{X}_f) = (\tau_{k,l})_{p\times p} = \begin{pmatrix} 1 & \tau(x_1^{(f)}, x_2^{(f)}) & \dots & \tau(x_1^{(f)}, x_p^{(f)}) \\ \tau(x_2^{(f)}, x_1^{(f)}) & 1 & \dots & \tau(x_2^{(f)}, x_p^{(f)}) \\ \vdots & \vdots & \vdots & \vdots \\ \tau(x_p^{(f)}, x_1^{(f)}) & \dots & \tau(x_p^{(f)}, x_{p-1}^{(f)}) & 1 \end{pmatrix} \tag{4}$$

and we propose

$$h(A(\mathbf{X}_f)) = \frac{2}{p(p-1)} \sum_{k<l} |\tau_{k,l}|. \tag{5}$$

That is, an optimal subdata $\mathbf{X}_f$ is obtained by

$$\begin{aligned} &\max_{\mathbf{X}_f \subset \mathbf{X}_F} \frac{2}{p(p-1)} \sum_{k<l} |\tau_{k,l}| \in [0,1] \\ &\text{s.t. } m_l \le size(\mathbf{X}_f) \le m_u. \end{aligned} \tag{6}$$

We refer to this criterion as the averaged absolute association (AAA) criterion for optimal subpopulation selection. Note that when there are only two variables (i.e., $p = 2$) and Kendall's τ is used for the measure of association, then the tau-path method in the literature [29, 30] can be viewed as a special case of the proposed method.

Efficient Algorithms

Note that the optimization for the AAA criterion in (6) is non-convex and discrete in nature. To address this challenge, we develop three efficient algorithms by taking advantage of the formulation of the general correlation coefficient.

Proposition 2 *Given a set of bivariate vectors* (x_i, y_i), $i = 1, 2, \dots, n$, *the general correlation coefficient in Proposition 1 can be expressed as*

$$\tau(x, y) = \frac{\sum_{i,j} u_{i,j}^{(x)} u_{i,j}^{(y)}}{\sqrt{\sum_{i,j} (u_{i,j}^{(x)})^2} \sqrt{\sum_{i,j} (u_{i,j}^{(y)})^2}} \triangleq \frac{N}{\sqrt{D_1}\sqrt{D_2}} \tag{7}$$

for Kendall's τ and Pearson's ρ. Then we have the following results:

(i) If (x_{n+1}, y_{n+1}) is added, the correlation between $(x_1, x_2, \ldots, x_n, x_{n+1})$ and $(y_1, y_2, \ldots, y_n, y_{n+1})$ can be calculated by

$$\tau(x^{(n+1)}, y^{(n+1)}) = \frac{\overbrace{\sum_{i,j} u_{i,j}^{(x)} u_{i,j}^{(y)}}^{N} + 2\sum_j u_{n+1,j}^{(x)} u_{n+1,j}^{(y)}}{\sqrt{\underbrace{\sum_{i,j}(u_{i,j}^{(x)})^2}_{D_1} + 2\sum_j (u_{n+1,j}^{(x)})^2}\sqrt{\underbrace{\sum_{i,j}(u_{i,j}^{(y)})^2}_{D_2} + 2\sum_j (u_{n+1,j}^{(y)})^2}}. \tag{8}$$

(ii) If an arbitrary (x_k, y_k) is deleted, the correlation between $(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$ and $(y_1, \ldots, y_{k-1}, y_{k+1}, \ldots, y_n)$ can be calculated by

$$\tau(x_{(-k)}, y_{(-k)}) = \frac{\overbrace{\sum_{i,j} u_{i,j}^{(x)} u_{i,j}^{(y)}}^{N} - 2\sum_j u_{k,j}^{(x)} u_{k,j}^{(y)}}{\sqrt{\underbrace{\sum_{i,j}(u_{i,j}^{(x)})^2}_{D_1} - 2\sum_j (u_{k,j}^{(x)})^2}\sqrt{\underbrace{\sum_{i,j}(u_{i,j}^{(y)})^2}_{D_2} - 2\sum_j (u_{k,j}^{(y)})^2}}. \tag{9}$$

The three efficient algorithms, namely the genetic algorithm (GA), forward selection algorithm (FSA), and backward selection algorithm (BSA), are summarized in Algorithms 1, 2, and 3, respectively. In all these algorithms, we utilize the one-step update rule in Proposition 2 to facilitate the computation of the generalized correlations based on their past values, with minimal extra computations.

In the genetic algorithm (Algorithm 1), we start with K candidate subdata sets, and evolve them toward higher association values. Each evolution step involves two phases: the mutation phase (Step 2) and the selection phase (Step 4). In the mutation phase, we construct a new candidate set by either adding or dropping an observation from each of the subdata sets. For each mutation, since only one observation will be included or excluded from the original set, we can efficiently compute the updated correlation coefficient via Proposition 2. In the selection step, we keep a subset of candidate sets with the highest correlation coefficients, as well as a random subset of all the eligible candidates (i.e., the original candidate sets plus the newly mutated candidate sets). Such a strategy involves a trade-off between exploration and exploitation [4, 5], which aims to find the subdata with high association, while keeping the possibility of exploring other subdata sets. To avoid redundant computations, we remove duplicated active sets in each evolution step, and cache all computed state-correlation pairs with a hash table. We stop the algorithm when the

Algorithm 1 Genetic Algorithm for Subpopulation Selection

procedure GENETIC ALGORITHM

Step 1: Randomly Initialize K candidate sets $\mathcal{C}_0 = \{\boldsymbol{X}_{f,1}, \boldsymbol{X}_{f,2}, ..., \boldsymbol{X}_{f,K}\}$.

for g in $1, \cdots, N$ **do**:

Step 2: **Mutation:** Mutate current candidates with

Add: randomly add an observation (add a row in $\boldsymbol{X}_f$)

Drop: randomly drop an observation (drop a row in $\boldsymbol{X}_f$)

Repeat until we have a size C child candidates $\tilde{\mathcal{C}}_g$

Step 3: Compute the AAA criterion scores for all candidate in $\tilde{\mathcal{C}}_g$

Step 4: **Selection:** Add to a new candidate sets $\mathcal{C}_g$

Fittest: T best candidates in $\mathcal{C}_{g-1}$ and $\tilde{\mathcal{C}}_g$

Diversity: randomly pick $K - T$ from the remaining candidates

Step 5: Stop the evolution if the best score has not improved for the last

L iterations (termination rules)

Step 6: Return the final subdata sets.

iteration budget has been exhausted, or the best correlation has not been changed for a pre-defined number of generations.

The forward selection algorithm (FSA, Algorithm 2) starts with all subdata sets of size three (unless the number of candidate sets is excessively large, then we use a random subset). The reason for this is that the subdata with size two will result in the AAA criterion score to be 1. In each step of the FSA, the algorithm adds one observation to each subdata set and selects the one that gives the highest AAA criterion. In the case of a tie among several subdata sets, we keep all of them and continue adding one observation at a time until we break the tie. Similarly, the backward selection algorithm (BSA, Algorithm 3) starts with the full-sized set, and progressively removes one observation at a time in accordance with the optimality criterion. Note that for both FSA and BSA methods, we can take advantage of the one-step update rule in Proposition 2 to enable computational efficiency in the implementation of the algorithms.

Algorithm 2 Forward Selection Algorithm

procedure FORWARD SELECTION

for g in $3, \cdots, m_l$ **do**:

if $g = 3$ **then**

Compute scores for all three possible observation combinations

Select the best combinations

else

Add one observation to existing selected subpopulations

Compute all possible scores and keep the best subpopulation

Return the final subdata sets.

Algorithm 3 Backward Selection Algorithm

procedure BACKWARD SELECTION

for g in $n, n-1, \cdots, m_l$ **do**:
  if $g = n$ **then**
    Compute the score for all observations
  else
    Delete one observation at a time from current selected subpopulation
    Compute all possible scores and keep the best subpopulations
Return the best subdata sets.

It is worth remarking that each of these three algorithms has its own merits and limitations. For the genetic algorithm, it is flexible to add or delete observations. It also can alleviate the problem of being trapped in a local maximum, although it may converge to a local maximum. The forward selection algorithm starts with a small number of observations and adds observations step by step. But, when the sample size n and lower limit m_l are large, it will require maintaining more subset candidates and subsequently need computations to obtain their corresponding AAA criterion scores. The backward selection algorithm starts with the full sets of observations. When the lower limit m_u is large, the backward method will be fast. However, when the sample size n is large, the backward algorithm can be computationally intensive.

As stated earlier, all three algorithms utilize the results in Proposition 2 to perform a one-step update for the calculation of the general correlation. Moreover, these three algorithms are able to obtain diverse solutions since there often exist different subdata sets achieving the same optimal AAA criterion. While existing methods such as the tau-path method can only identify one subdata, our approaches can obtain multiple optimal and near-optimal subdata sets, allowing more flexibility for investigators to probe their data of interest. Another key feature of the proposed methods is the ability to handle high-dimensional data, which is of great importance in machine learning and data science [3, 23]. Our selection of an optimal subset is based on a general criterion which includes not only the traditional linear Pearson correlation but also the non-linear Kendall and Spearman correlation, and other nonparametric association metrics [8].

3 Simulation Study

In this section, we will conduct a set of simulation studies to examine the performance of the proposed method. The code for implementing the proposed method is available at https://github.com/qingguo666/BSA. All simulations are implemented with Python. In section "Evaluation of Selected Subpopulation", we will evaluate the accuracy of the subpopulation selection under different settings. Section "Comparisons

of the Algorithms" compares the three algorithms in terms of selection accuracy and computational time. We also provide a visualization to better understand the characteristics of the selected data points. Section "Comparison with the Tau-Path Method" compares the proposed method with the tau-path algorithm in the literature [29, 30].

Evaluation of Selected Subpopulation

With the loss of generality, we assume that the data are centered with zero means. The simulated population consists of two subpopulations, the first subpopulation (ground truth) is randomly generated from a multivariate normal distribution with covariance matrix $\mathbf{\Sigma}_1$, and the second subpopulation (noise) is randomly generated from a multivariate normal distribution with covariance matrix $\mathbf{\Sigma}_2$. Here, $\mathbf{\Sigma}_1$ has a strong correlation structure and $\mathbf{\Sigma}_2$ is an identity matrix. The goal is to correctly select the correlated subpopulation. Table 1 summarizes the settings for simulation study Table 1. We set the lower bound $m_l = n\alpha$, i.e., the sample size of the first subpopulation. Both the Pearson correlation and Kendall's τ are used as an example of the general correlation in the AAA criterion.

We apply the three algorithms and select the optimal subdata sets with the largest value of the AAA criterion. Note that for the forward selection algorithm and the backward selection algorithm, only the lower bound constraint m_l is needed. However, in the case of the genetic algorithm, both lower and upper constraints are needed since the smaller sample size often leads to higher correlation, we set $m_u = m_l + 3$. To evaluate the performance, we compare the selected subpopulation to the ground truth subpopulation, and report the true positive rate (TPR) [28].

The values of TPR are reported in Fig. 1. The upper and lower panels of Fig. 1 respectively show the selection results for Pearson correlation and Kendall correlation. In general, it is seen that the TPR is sensitive to the proportion of the first subpopulation (i.e., α) and the correlation parameter (ρ). As ρ increases, the TPR values also increase, indicating the higher accuracy of the proposed method. Also, the value of TPR has a strong dependency on the data dimension p. If we fix all other parameters, the proposed method often gives better results for larger p. Further, it

Table 1 Settings for simulation study 1 ($n > p$)

Sample size	$n = 50$ or 100
# of covariates	$p = 5$ or 10
Proportion of the subpopulation	$\alpha = 0.3, 0.5$, or 0.8
Correlation of the first subpopulation	$\rho = 0.3, 0.5$, or 0.8
Covariance matrix of subpopulation	$\mathbf{\Sigma}_1 = \left(\sigma_{ij}\right)_{p\times p}, \sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } i \neq j \end{cases}$ $\mathbf{\Sigma}_2 = \boldsymbol{I}_p$

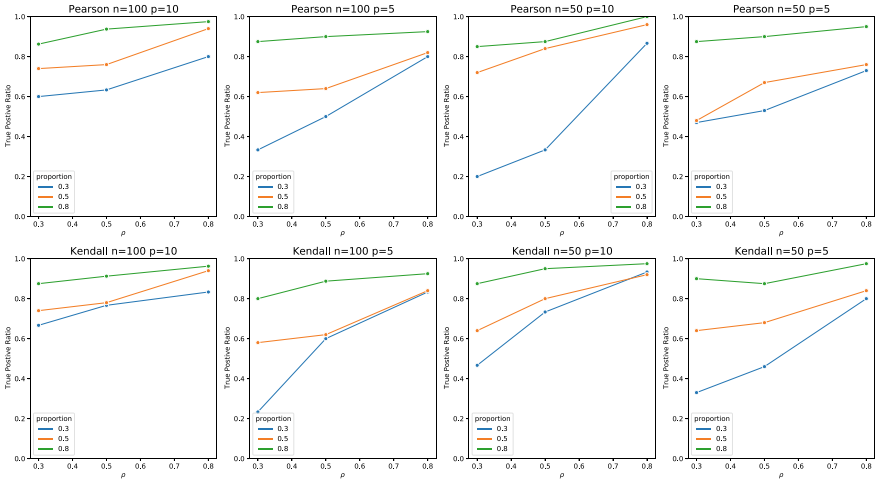


Fig. 1 True positive rate in simulation 1 for Pearson correlation (top row) and Kendall correlation (bottom row)

Table 2 Settings for simulation study 2 when $n < p$

Sample size	$n = 30$
# of covariates	$p = 100$
Proportion of the subpopulation	$\alpha = 0.3$, 0.5, or 0.8
Correlation of the first subpopulation	$\rho =$ 0.3, 0.5, or 0.8
Covariance matrices of the subpopulations	$\boldsymbol{\Sigma}_1 = \left(\sigma_{ij}\right)_{p\times p}, \sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } i \neq j \end{cases}$ $\boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$

appears that the proposed method using the Kendall τ mostly yields better results than that using the Pearson correlation. For $\rho \geq 0.8$, the method using Kendall's τ gives TPR values greater than 0.8. Such a pattern even applies when the proportion of the first subpopulation is small.

We also investigate the performance of the proposed method for the situation when $n < p$, using the backward selection algorithm. The simulation setting is shown in Table 2, and the results are summarized in Table 3. We can see that $n < p$ can estimate the association value (i.e., the AAA score) correctly and identify the subpopulation accurately with a high value of TPR. Compared to the situation when $n > p$, the proposed methods can precisely choose the correct subpopulation better in $n < p$. Under the same proportion of the correlated subpopulation, the true positive rate can be improved with a high correlation coefficient. Even if the proportion of the correlated subpopulation is small, the proposed method can still perform very well when the correlation of the true subpopulation is medium. These simulation results provide convincing support for using the proposed method when $n < p$.

Table 3 Results for simulation study 2 ($n < p$)

Parameter		AAA value		True positive ratio	
		Pearson	Kendall	Pearson	Kendall
$\alpha = 0.3$	$\rho = 0.3$	0.30	0.23	0.78	0.67
	$\rho = 0.5$	0.48	0.28	1.00	0.78
	$\rho = 0.8$	0.77	0.67	1.00	1.00
$\alpha = 0.5$	$\rho = 0.3$	0.23	0.27	0.93	0.87
	$\rho = 0.5$	0.33	0.38	1.00	1.00
	$\rho = 0.8$	0.59	0.64	1.00	1.00
$\alpha = 0.8$	$\rho = 0.3$	0.17	0.20	0.96	0.96
	$\rho = 0.5$	0.34	0.31	1.00	1.00
	$\rho = 0.8$	0.57	0.61	1.00	1.00

Comparisons of the Algorithms

We compare the three algorithms, the genetic algorithm, the forward selection algorithm, and the backward selection algorithm, in terms of computational time and TPR values. For a simulation setting with $n = 50$, and $p = 5$, see Table 1. The results are reported in Table 4. It is seen that the forward selection algorithm (FSA) is time-consuming, because it starts with all possible combinations and there are probably many ties. The backward selection algorithm (BSA) only calculates p possible situations in the first step, thus taking less time. However, the BSA cannot always get the best TPR among the three algorithms, although in general, the TPR for the three algorithms is comparable.

A visualization of the algorithms can help us to further understand their characteristics. Note that our algorithms may not always find the correlated subpopulation,

Table 4 Comparisons of TPR and computing time for the three algorithms when $n = 50$ and $p = 5$

Parameter		True positive ratio			Running time		
		GA	FSA	BSA	GA	FSA	BSA
$\alpha = 0.3$	$\rho = 0.3$	0.2	0.27	**0.47**	28.87	73.97	**0.79**
	$\rho = 0.5$	0.53	0.53	0.53	28.49	77.16	**0.76**
	$\rho = 0.8$	0.67	0.67	0.67	28.97	169.96	**0.81**
$\alpha = 0.5$	$\rho = 0.3$	0.64	0.64	0.64	51.37	77.28	**0.62**
	$\rho = 0.5$	0.68	**0.76**	0.72	51.83	77.34	**0.64**
	$\rho = 0.8$	**0.88**	0.84	**0.88**	51.57	114.04	**0.61**
$\alpha = 0.8$	$\rho = 0.3$	**0.9**	0.78	**0.9**	103.11	72.39	**0.35**
	$\rho = 0.5$	**0.85**	0.825	0.825	102.60	78.92	**0.33**
	$\rho = 0.8$	0.98	0.98	0.98	101.86	250.71	**0.32**

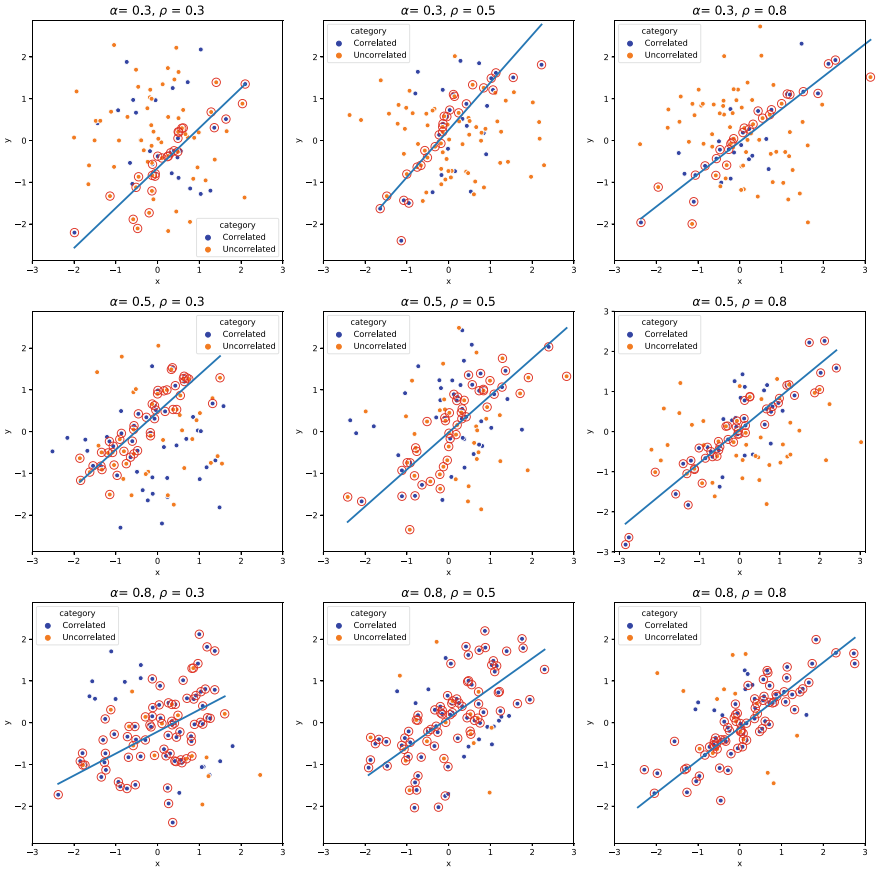


Fig. 2 Visualization of selecting the correlated subpopulation by the BSA when $p = 2$

especially in lower dimensions. To illustrate, let $n = 2$ and $p = 2$ with other settings being as in Table 1. When $p = 2$, we can provide a good 2-D visualization to understand why the methods may have lower accuracy than that in higher dimensions. In Fig. 2, we display the simulated points in a scatter plot and distinguish them using different colors (blue dots for subpopulation 1 and orange dots for subpopulation 2). We use the backward selection algorithm to find an optimal subpopulation, which is denoted by red circles. In addition, we fit a regression line for the selected subpopulation. Recall that the objective is to correctly select the correlated subpopulation (blue dots). The plots show that there are some orange dots and blue dots building a better path than the path built only by the correlated blue points. However, this problem can be alleviated by increasing p. As p becomes larger, it will be difficult to select a data point from the wrong subpopulation (i.e., subpopulation 2). The reason is that a point from the incorrect subpopulation will be difficult to build a better path with correlated points in all dimensions.

Table 5 Results for comparison of Kendall's correlation and run time based on different m_l in the simulation study

		100	200	300	400	500	600	700	800	900	1000
Kendall's correlation	FastBCS	0.962	0.868	0.786	0.696	0.607	0.508	0.390	0.268	0.144	0.001
	BSA	0.962	0.868	0.786	**0.697**	0.607	0.508	0.390	0.268	0.144	0.001
Running time (s) per solution	FastBCS	17.8	17.8	**17.8**	**17.8**	**17.8**	**17.8**	**17.8**	**17.8**	**17.8**	17.8
	BSA	**9.1**	**13.6**	25.9	75.3	97.5	154.0	126.0	93.7	49.5	**4.5**

Comparison with the Tau-Path Method

Tau-path approach [21, 29, 30] deals with subpopulation selection, but only for bivariate data ($p = 2$) using Kendall's τ as the measure of association. Reference [21] discussed the tau-path method for large samples with the new code[1] (up to $n = 10000$). Here, we compare our BSA with the Fast Backward Conditional Search (FastBCS) in [21] for $p = 2$, which can select increasing subsets whose associated tau coefficients become monotonically decreasing. For $n = 1000$, we sample integers without replacement from 1 to 1000 to form a two-dimensional data set. Then, we apply the FastBCS and the BSA to select the optimal subpopulation, where the proposed BSA algorithm uses Kendall's τ as the measure of association. By increasing the set size m_l, we calculate Kendall's correlation of subset chosen by FastBCS and BSA, respectively, with the results reported in Table 5. Generally, these two methods are comparable, giving similar Kendall's coefficients under different settings. It is seen that the value of Kendall's coefficient for the BSA method monotonically decreases as the size of selected subsets increases, which is similar to the pattern from the tau-path method. It is worth noting that the BSA method generally needs more computational time in comparison with the FastBCS method. It is because the BSA method tracks all optimal subsets along the solution path, and explores more possibilities. In the comparison of the average running time per solution, it is found that both methods have their respective merits as shown in Table 5.

4 Case Study

In this section, we illustrate our approach using the COVID-19 data from the Virginia Department of Health. Table 6 gives some samples of the collected data for each week from 05/15/2020 to 02/05/2021 for the 35 districts in the state of Virginia. The data consists of accumulative total cases, accumulative number of people hospitalized,

[1] https://github.com/acaloiaro/topk-taupath.

Table 6 COVID-19 data in VA from 05/15/2020 to 02/05/2021

	Cases Wk 1	Hosp. Wk 1	Deaths Wk 1	PCR Wk 1	...	Cases Wk 39	Hosp. Wk 39	Deaths Wk 39	PCR Wk 39
West Piedmont	69	10	2	NA	...	9862	573	188	163095
Hampton	158	29	3	NA	...	7652	241	69	16729
...	...	...	...	...	...	...	...	...	...
Prince William	4130	398	75	NA	...	41306	1561	366	152350

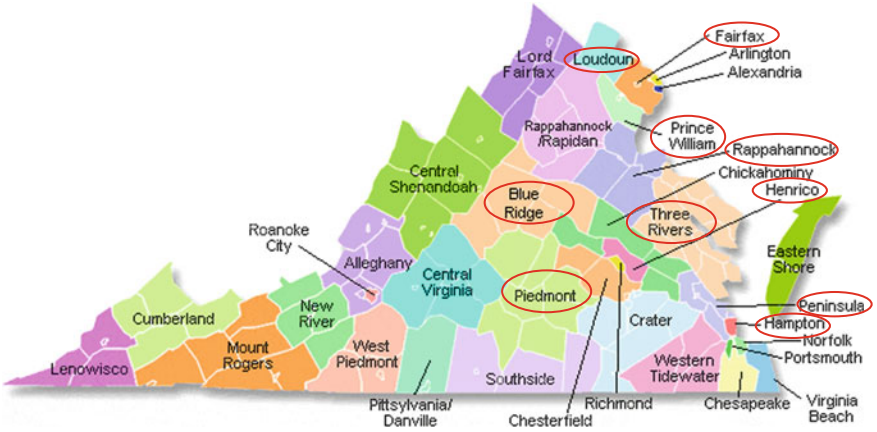


Fig. 3 Ten highly correlated districts in Virginia (circled) selected by the BSA method

accumulative number of deaths, and accumulative number of people taking the PCR test. We consider the districts as observation units and weekly accumulative data as variables. Thus, the data frame contains 35 rows of observations and 156 columns of variables. We deleted columns with missing entries and then we standardized data to have zero mean and unit standard deviation.

To evaluate the performance of the proposed method, we first applied the BSA algorithm to choose 10 highly correlated districts, highlighted in Fig. 3. We can see that the correlated regions are mostly geographically clustered in the northeastern part of the state adjacent to the District of Columbia (Loudoun, Fairfax, Prince William, Rappahannock, Peninsula, and Hampton), slightly extending to the middle of VA (Three Rivers, Henrico, Blue Ridge, and Piedmont). It is known that Northern Virginia is a densely populated region with thriving business activities, which could be a key explanation that these regions share similar characteristics during the pandemic.

Next, we conducted the subselection based on different groups of variables. That is, we conducted the subselection only based on variables of the accumulative total cases, and applied the BSA algorithm to choose 10 highly correlated districts. Similarly, we apply the BSA algorithm to perform the subselection based on the accu-

mulative number of people hospitalized and the accumulative number of deaths, respectively. The results are summarized in Table 7. Note that the solutions of the optimal subpopulation are often not unique. Thus, we report the frequency of districts to be selected in the optimal solutions. From the results in the table, it is seen that the districts selected based on different groups of variables can be different, although there are some overlaps among the selected districts under these three groups of variables. It is worth pointing out that the analysis results based on the number of PCR are not reported here, because two optimal subpopulations obtained by the BSA method have an AAA score of 1 with $m_l = 22$. It implies that almost all districts in Virginia are highly associated with respect to the number of PCR tests.

Table 7 Optimal selection results based on the number of cases, the number of people hospitalized, and the number of deaths, respectively

Districts	**Case** Frequency (out of 2 tied solutions)	**Hospitalization** Frequency (out of 4 tied solutions)	**Deaths** Frequency (out of 2 tied solutions)
West Piedmont	1	0	0
Hampton	0	0	1
Fairfax	1	1	1
Crater	1	0	0
Chickahominy	0	2/4	0
Henrico	1/2	0	1/2
Chesterfield	0	0	1
Arlington	0	2/4	0
Lenowisco	1	0	0
Virginia Beach	0	0	1
Loudoun	1	2/4	0
Three Rivers	0	1	1
Blue Ridge	1	0	0
Southside	0	0	1
Roanoke	0	1	0
Lord Fairfax	0	1	1
Rappahannock Rapidan	0	2/4	0
Rappahannock	1/2	1	0
Richmond	1	0	1
Pittsylvania Danville	1	0	0
Piedmont	0	1	0
Peninsula	0	1	1
Prince William	1	1	1/2

5 Discussion

We propose a new method to select an optimal subpopulation based on association for the multivariate data, together with three efficient algorithms (GA, Backward Selection, and Forward Selection) to achieve this goal. It is possible to combine the backward and forward selection to conduct a stepwise selection procedure, in a spirit similar to stepwise variable selection in regression [20]. However, the stepwise subpopulation selection procedure for subpopulation can be computationally intensive. To mitigate this issue, the genetic algorithm appears to be a good alternative. The GA has similar good properties as a stepwise selection in terms of flexibly adding or dropping observations. Moreover, the GA can start with any number of observations and randomly add or drop cases with limited generations.

The top-k method [9, 22] is commonly used in selecting a subpopulation under the context of ranking lists. The goal is to provide statistical inference on the lengths of informative top-k lists for multiple ranked input lists (full or partial) representing the same set of N objects. The popular tau-path method is to provide an ordered list of observations whose association among variables is gradually decreased. But the existing tau-path methods can only deal with the bivariate data and the analysis is usually limited to Kendall's τ correlation. The angle of our proposed association-based optimal subpopulation selection is different from the top-k method. The proposed backward selection can be used to generalize the tau-path approach for multivariate data since we can track the ordination of observations to be dropped in the algorithm. Moreover, such a generalized tau-path approach is not limited to Kendall's τ correlation, but is also applicable to other association metrics.

The proposed method can also be extended to handle compositional data where the random vector $(X_1, \ldots, X_p)$ has strictly positive components whose sum is constant, which is closely related with the mixture design [24] in the literature of experimental design. It is known that there can be a spurious correlation in the compositional data [19]. To address this challenge, we can consider the additive log-ratio transformation [1, 18] as $Z_j = \log(X_j/X_p)$ to form the vector $(Z_1, \ldots, Z_{p-1})$ and then apply the proposed method based on the transformed data for selecting an optimal subpopulation.

Acknowledgements We are grateful for very useful referee comments that helped us enhance the paper.

Appendix

Proof of Proposition 1
For (c), if $u_{ij}^{(x)} = x_i - x_j$ and $u_{ij}^{(y)} = y_i - y_j$, it is easy to see that

$$\begin{aligned}\sum_{i,j} u_{ij}^{(x)} u_{ij}^{(y)} &= \sum_{i,j}(x_i - x_j)(y_i - y_j) = 2\sum_{i,j} x_i y_i - 2\sum_{i,j} x_i y_j \\ &= 2n\sum_i x_i y_i - 2n^2 \bar{x}\bar{y} \\ &= 2n(\sum_i x_i y_i - n\bar{x}\bar{y}) \\ &= 2n\sum_i (x_i - \bar{x})(y_i - \bar{y}).\end{aligned}$$

Similarly, we can have

$$\begin{aligned}\sum_{i,j}(u_{ij}^{(x)})^2 &= \sum_{i,j}(x_i - x_j)^2 = 2\sum_{i,j} x_i^2 - 2\sum_{i,j} x_i x_j \\ &= 2n\sum_i x_i^2 - 2n^2\bar{x}^2 \\ &= 2n(\sum_i x_i^2 - n\bar{x}^2) \\ &= 2n\sum_i (x_i - \bar{x})^2.\end{aligned}$$

Then the expression in (2) can be written as

$$\begin{aligned}\tau(x, y) &= \frac{\sum_{i,j} u_{ij}^{(x)} u_{ij}^{(y)}}{\sqrt{\sum_{i,j}(u_{ij}^{(x)})^2 \sum_{i,j}(u_{ij}^{(y)})^2}} = \frac{\sum_{i,j}(x_i - x_j)(y_i - y_j)}{\sqrt{\sum_{i,j}(x_i - x_j)^2 \sum_{i,j}(y_i - y_j)}} \\ &= \frac{2n\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{2n\sum_i (x_i - \bar{x})^2 2n\sum_i (y_i - \bar{y})^2}} \\ &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.\end{aligned}$$

Proof of Corollary 1
For binary variables $x \in \{0, 1\}$ and $y \in \{0, 1\}$ and data points (x_i, y_i), $i = 1, \ldots, n$, we can first check the Pearson correlation coefficient. First, we have

$$\begin{aligned}\sum_i (x_i - \bar{x})(y_i - \bar{y}) &= \sum_i x_i y_i - n\bar{x}\bar{y} = n_{11} - n\frac{n_{1\cdot}}{n}\frac{n_{\cdot 1}}{n} \\ &= \frac{1}{n}\left[n_{11}n - n_{1\cdot}n_{\cdot 1}\right] \\ &= \frac{1}{n}\left[n_{11}(n_{11} + n_{10} + n_{01} + n_{00}) - (n_{10} + n_{11})(n_{01} + n_{11})\right] \\ &= \frac{1}{n}\left[n_{11}n_{00} - n_{10}n_{01}\right].\end{aligned}$$

Second, we can get

$$\begin{aligned}\sum_i (x_i - \bar{x})^2 &= \sum_i x_i^2 - n\bar{x}^2 = n_{1\cdot} - n(\frac{n_{1\cdot}}{n})^2 \\ &= \frac{1}{n}\left[n_{1\cdot}n - n_{1\cdot}^2\right] \\ &= \frac{1}{n}\left[n_{1\cdot}n_{0\cdot}\right].\end{aligned}$$

Similarly, we obtain $\sum_i (y_i - \bar{y})^2 = \frac{1}{n}\left[n_{\cdot 1}n_{\cdot 0}\right]$. Thus,

$$\begin{aligned}\tau(x, y) &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \\ &= \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 1}n_{\cdot 0}}}.\end{aligned}$$

References

1. Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological), 44*(2), 139–160.
2. Bingham, D., Sitter, R. R., & Tang, B. (2009). Orthogonal and nearly orthogonal designs for computer experiments. *Biometrika, 96*(1), 51–65.
3. Bühlmann, P., & Van De Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media.
4. Cai, X., Xu, L., Lin, C. D., Hong, Y., & Deng, X. (2021). Sequential design of computer experiments with quantitative and qualitative factors in applications to hpc performance optimization. arXiv:2101.02206
5. Chen, Z., Mak, S., & Wu, C. (2019). A hierarchical expected improvement method for Bayesian optimization. arXiv:1911.07285
6. Clarkson, K. L., & Shor, P. W. (1989). Applications of random sampling in computational geometry, II. *Discrete & Computational Geometry, 4*(5), 387–421.
7. Cramér, H. (2016). *Mathematical Methods of Statistics*, vol. 9, PMS. Princeton University Press.
8. Gibbons, J. D., & Fielden, J. D. G. (1993). *Nonparametric Measures of Association*. SAGE.
9. Hall, P., & Schimek, M. G. (2012). Moderate-deviation-based inference for random degeneration in paired rank lists. *Journal of the American Statistical Association, 107*(498), 661–672.

10. Kendall, M. G. (1948). *Rank Correlation Methods*. Griffin.
11. Li, Y., Deng, X., Jin, R., Ba, S., & Myers, W. (2020). Clustering-based data filtering for manufacturing big data system. *Journal of Quality Technology*.
12. Li, Y., Kang, L., & Deng, X. (2021). A maximin Φ_p-efficient design for multivariate GLM. *Statistica Sinica*, in press.
13. Liberty, E., Lang, K., & Shmakov, K. (2016). Stratified sampling meets machine learning, in *International Conference on Machine Learning*, PMLR, pp. 2320–2329.
14. Lin, C. D., Chien, P., & Deng, X. (2022) Efficient Experimental Design for Regularized Linear Models. In *Advances and Innovations in Statistics and Data Science*.
15. Lin, C. D., Anderson-Cook, C. M., Hamada, M. S., Moore, L. M., & Sitter, R. R. (2015). Using genetic algorithms to design experiments: A review. *Quality and Reliability Engineering International, 31*(2), 155–167.
16. Liu, H., Sadygov, R. G., & Yates, J. R. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry, 76*(14), 4193–4201.
17. Liu, K., Mei, Y., & Shi, J. (2015). An adaptive sampling strategy for online high-dimensional process monitoring. *Technometrics, 57*(3), 305–319.
18. Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Wiley.
19. Pearson, K. (1897). Mathematical contributions to the theory of evolution.-On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London, 60*(359–367), 489–498.
20. Peduzzi, P., Hardy, R., & Holford, T. R. (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics*, 511–516.
21. Sampath, S., Caloiaro, A., Johnson, W., & Verducci, J. S. (2016). The top-K tau-path screen for monotone association in subpopulations. *Wiley Interdisciplinary Reviews: Computational Statistics, 8*(5), 206–218.
22. Schimek, M. G., Budinská, E., Kugler, K. G., Švendová, V., Ding, J., & Lin, S. (2015). TopKLists: A comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Statistical Applications in Genetics and Molecular Biology, 14*(3), 311–316.
23. Shao, J., Wang, Y., Deng, X., Wang, S., et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics, 39*(2), 1241–1265.
24. Shen, S., Kang, L., & Deng, X. (2020). Additive heredity model for the analysis of mixture-of-mixtures experiments. *Technometrics, 62*(2), 265–276.
25. Trost, J. E. (1986). Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies. *Qualitative Sociology, 9*(1), 54–57.
26. Wu, C. J., & Hamada, M. S. (2011). *Experiments: Planning, Analysis, and Optimization*, vol. 552. Wiley.
27. Xian, X., Wang, A., & Liu, K. (2018). A nonparametric adaptive sampling strategy for online monitoring of big data streams. *Technometrics, 60*(1), 14–25.
28. Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports, 1896–1970*, 1432–1449.
29. Yu, L., Verducci, J. S., & Blower, P. E. (2011). The tau-path test for monotone association in an unspecified subpopulation: Application to chemogenomic data mining. *Statistical Methodology, 8*(1), 97–111.
30. Zhang, Y., Ravishanker, N., Ivan, J. N., & Mamun, S. A. (2019). An application of the tau-path method in highway safety. *Journal of the Indian Society for Probability and Statistics, 20*(1), 117–139.

Likelihood-Based Inference for Linear Mixed-Effects Models with Censored Response Using Skew-Normal Distribution

Thalita B. Mattos, Larissa A. Matos, and Victor H. Lachos

Abstract Mixed-effects models are commonly used to fit longitudinal or repeated measures data. A complication arises when the response is censored, for example, due to limits of quantification of the assay used. Although normal distributions are commonly assumed for random effects, such assumptions may be unrealistic, obscuring essential features of among-individual variation. We relax this assumption by considering a likelihood-based inference for linear mixed-effects models with censored response (LMEC) based on the multivariate skew-normal distribution. An ECM algorithm is developed for computing the maximum likelihood estimates for LMEC with the standard errors of the fixed effects and the exact likelihood value as a by-product. The algorithm uses closed-form expressions at the E-step that rely on formulas for the mean and variance of a truncated multivariate skew-normal distribution. The proposed methods are applied to analyze longitudinal HIV viral load data in an AIDS study.

1 Introduction

Longitudinal studies have attracted considerable interest in clinical trials, biological psychology, environmental science, and medical research. They enable the study of change over time of an outcome and the evaluation of determinants of change. Linear and nonlinear mixed-effect (LME/NLME) models are powerful tools for analyzing longitudinal data. In these models, random effects are incorporated to accommodate among-subject variation [9, 20]. The random errors and random effects are routinely

T. B. Mattos · L. A. Matos
Departamento de Estatística, Universidade Estadual de Campinas, Campinas, Brazil
e-mail: thalitadobem@gmail.com

L. A. Matos
e-mail: larissam@unicamp.br

V. H. Lachos (✉)
Department of Statistics, University of Connecticut, Storrs, CT, USA
e-mail: hlachos@uconn.edu

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_2

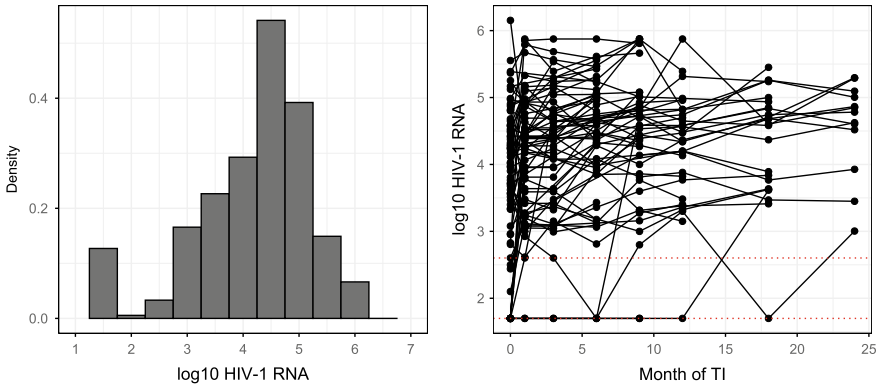


Fig. 1 UTI data. (Left panel) Histogram for HIV viral load (in $\log_{10}$) scale. (Right panel) Individual profiles for HIV viral load (in $\log_{10}$ scale)

assumed to have a normal distribution due to their mathematical tractability and computational convenience.

Although the normality assumption may be reasonable for many situations, a severe normality departure will cause a lack of robustness and subsequently lead to invalid inference and unreasonable estimates [31]. Specially non-normal characteristics such as skewness with heavy right or left tail often appear in virologic responses. For example, Fig. 1 (Left panel) displays the histogram of the viral load measurements for 72 children from an AIDS clinical trial study—UTI (refer to Sect. 4 for details of this dataset). From this figure, it can be seen that the viral load measurements are left-skewed, suggesting that the normality assumption might be inappropriate.

As an alternative to the weakness of unrealistic normality assumptions and eliminating the need for *ad hoc* data transformations, asymmetric distributions can be applied to consider this non-ignorable departure from normality. Lachos et al. [18] proposed a robust generalization of LME, called the skew-normal/independent linear mixed (SNI-LME) model, by assuming a skew-normal/independent (SNI) distribution [7] for the random effects and a normal/independent distribution for the random errors. Ho and Lin [14] proposed a model that provides flexibility in capturing the effects of skewness and heavy tails simultaneously among longitudinal data; they consider an extension of LME assuming a multivariate skew-t distribution for the random effects and a multivariate t distribution for the error terms.

Another complexity of longitudinal studies occurs when the response is censored for some of the observations, which often arises when assay measures are collected over time, and the assay procedure is subject to limits of quantification. As a case in point, the HIV-1 viral load, which is currently the primary marker of HIV infection, has a lower and upper quantification limit, which depends on the type of assay used. The viral load of patients receiving antiretroviral treatment will typically decline and stay for a more extended period below the lower limit of quantification. Figure 1 (Right panel) shows the measures of HIV-1 viral loads of 72 children after unstruc-

tured treatment interruption (UTI). In this case, the HIV-1 RNA measures are subject to censoring below the lower limit of detection of the assay (50 copies/mL or 400 copies/mL). This study presents about 7% of observations below the detection limits of assay quantification and hence are considered to be left-censored. When response observations are below the limits of quantification, a common practice is to impute the censored values by the detection limit or half the detection limit [33]. Such *ad hoc* methods may produce biased results, as pointed out by Hughes [16], Jacqmin-Gadda et al. [17], Matos et al. [24], among others.

In the literature, longitudinal data with censored observations have received considerable attention. Vaida and Liu [29] proposed an exact EM algorithm for LME/NLME with censored response (LMEC/NLMEC), which uses closed-form expressions at the E-step instead of Monte Carlo simulations. Robust extensions of LMEC and NLMEC based on the multivariate-t distribution, named tLMEC and tNLMEC, have been introduced by Matos et al. [23]. On the other hand, under a Bayesian framework, Bandyopadhyay et al. [6] studied LMEC models considering skewness and heavy tails, replacing the Gaussian assumptions with skew-normal/independent (SNI) distribution. However, no previous work has investigated LMEC/NLMEC models based on the skew-normal distributions from a likelihood-based perspective to the best of our knowledge.

In this work, we are devoted to presenting methodological developments of the skew-normal linear mixed model with censored responses (SN-LMEC) from a likelihood-based perspective, which considers the skewness behavior of the random effects. The SN-LMEC is defined by supposing that, for each subject, the random effects follow an SN distribution introduced by Azzalini and Valle [5], while the within-subject errors follow a multivariate normal distribution to prevent identifiability problems. Like Matos et al. [23], we show that the E-step reduces to computing the first two moments of a truncated multivariate SN distribution Galarza Morales et al. [13]. The likelihood function is easily computed as a by-product of the E-step and is used for monitoring convergence and model selection.

The organization of this work is outlined as follows. Section 2 presents the skew-normal distribution (SN) and some of its fundamental properties. Section 3 introduces the model (SN-LMEC) and describes an efficient ECM algorithm for calculating maximum likelihood (ML) estimates of parameters. We also discuss the issues related to empirical Bayes estimates of the random effects and prediction of future responses. A case study of HIV viral load is analyzed in Sect. 4 to evaluate the proposed method. We conclude the article with some discussions in Sect. 5.

2 The Multivariate Skew-Normal Distribution

This section presents the multivariate skew-normal distribution (SN) and multivariate extended skew-normal (ESN) and some of their valuable properties. Some versions, extensions, and unifications of the SN family are carefully surveyed in works such as Azzalini [3] and Arellano-Valle et al. [2].

Definition 1 A random vector $\mathbf{Y}$ has multivariate skew-normal distribution with $p \times 1$ location vector $\boldsymbol{\mu}$, $p \times p$ positive definite dispersion matrix $\boldsymbol{\Sigma}$ and $p \times 1$ skewness parameter vector $\boldsymbol{\lambda}$, and we write $\mathbf{Y} \sim \mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, if its density is given by

$$\mathrm{SN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \tag{1}$$

where $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote, respectively, the probability distribution function (pdf) and the cumulative distribution function (cdf) of the p-variate normal distribution $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We have $\boldsymbol{\Sigma}^{-1/2}$ such that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$. Note that if $\boldsymbol{\lambda} = \mathbf{0}$, then the density of $\mathbf{Y}$ reduces to the $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density.

It is worth mentioning that the multivariate skew-normal distribution is not closed over conditioning. Next, we present its extended version, which holds this property, called the multivariate ESN distribution.

Definition 2 A random vector $\mathbf{Y}$ has multivariate ESN distribution with $p \times 1$ location vector $\boldsymbol{\mu}$, $p \times p$ positive definite dispersion matrix $\boldsymbol{\Sigma}$, $p \times 1$ skewness parameter vector $\boldsymbol{\lambda}$, and shift parameter $\tau \in \mathcal{R}$, denoted by $\mathbf{Y} \sim \mathrm{ESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$, if its density is given by

$$\mathrm{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = \xi^{-1}\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\tau + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \tag{2}$$

with $\xi = \Phi_1(\tau/(1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})^{1/2})$. Note that when $\tau = 0$, we recover the skew-normal distribution defined in (1), that is, $\mathrm{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, 0) = \mathrm{SN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$.

Define,

$$\mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau) = \int_{\mathbf{a}}^{\mathbf{b}} \xi^{-1}\phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\tau + \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}))d\mathbf{y}.$$

So, when $\boldsymbol{\lambda} = 0$ and $\tau = 0$, we recover the multivariate normal case, and then

$$\mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{0}, 0) \equiv L_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbf{a}}^{\mathbf{b}} \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{y}.$$

Proposition 1 *Let* $\mathbf{Y} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ *and* $\mathbf{Y}$ *be partitioned as* $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top)^\top$, *with dimensions* p_1 *and* p_2, $p_1 + p_2 = p$, *respectively. Let*

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top \quad \textit{and} \quad \boldsymbol{\varphi} = (\boldsymbol{\varphi}_1^\top, \boldsymbol{\varphi}_2^\top)^\top$$

be the corresponding partitions of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\lambda}$ *and* $\boldsymbol{\varphi} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\lambda}$. *Then,*

(i) $\mathbf{Y}_1 \sim SN_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, c_{12}\boldsymbol{\Sigma}_{11}^{1/2}\tilde{\boldsymbol{\upsilon}})$; *and*
(ii) $\mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1 \sim ESN_{p_2}(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}, \boldsymbol{\Sigma}_{22.1}^{1/2}\boldsymbol{\varphi}_2, \tau_{2.1})$,

where $c_{12} = (1 + \boldsymbol{\varphi}_2^\top \boldsymbol{\Sigma}_{22.1}\boldsymbol{\varphi}_2)^{-1/2}$, $\tilde{\boldsymbol{\upsilon}} = \boldsymbol{\varphi}_1 + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\varphi}_2$, $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$ *and* $\tau_{2.1} = \tilde{\boldsymbol{\upsilon}}^\top(\mathbf{y}_1 - \boldsymbol{\mu}_1)$.

Proof See Proposition 2 in Galarza Morales et al. [13].

The mean and variance of an ESN random vector are given in the following lemma.

Lemma 1 *Let* $\mathbf{Y} \sim ESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$. *Then,*

(i) $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu} + \eta\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda}$,

(ii) $\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top + \eta\left(\boldsymbol{\mu}\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{1/2} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda}\boldsymbol{\mu}^\top\right) + \eta\tau\boldsymbol{\Sigma}^{1/2}\dfrac{\boldsymbol{\lambda}\boldsymbol{\lambda}^\top}{1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda}}\boldsymbol{\Sigma}^{1/2}$,

(iii) $Var(\mathbf{Y}) = \boldsymbol{\Sigma} - \eta\boldsymbol{\Sigma}^{1/2}\boldsymbol{\lambda}\left(\eta\boldsymbol{\lambda} - \dfrac{\tau}{1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda}}\boldsymbol{\lambda}\right)^\top\boldsymbol{\Sigma}^{1/2}$,

with $\eta = \phi_1(\tau; 0, 1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})/\xi$. *When* $\tau = 0$, *we recover* $\mathbb{E}[\mathbf{Y}]$, $\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top]$ *and* $Var(\mathbf{Y})$ *of the skew-normal distribution.*

An interesting discussion about the parametrization ESN distribution is provided in Castro et al. [8].

Definition 3 Let $\mathbf{X} \sim \text{ESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$ and $\mathbb{P}(\mathbf{a} < \mathbf{X} < \mathbf{b}) > 0$. A random vector $\mathbf{Y}$ has a truncated extended multivariate skew-normal (TESN) distribution in the interval $[\mathbf{a}, \mathbf{b}]$, denoted by $\mathbf{Y} \sim \text{TESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau, [\mathbf{a}, \mathbf{b}])$, if its density is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{\text{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)}{\int_{\mathbf{a}}^{\mathbf{b}} \text{ESN}_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)d\mathbf{y}}, \quad \mathbf{a} \leq \mathbf{y} \leq \mathbf{b}.$$

For the special case of $\tau = 0$, we refer to this distribution as a truncated multivariate skew-normal (TSN) distribution, i.e., $\text{TESN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \mathbf{0}, [\mathbf{a}, \mathbf{b}]) \equiv \text{TSN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, [\mathbf{a}, \mathbf{b}])$.

The following properties of the multivariate truncated ESN distribution are useful for implementing the ECM algorithm in the SN-LMEC model.

Lemma 2 *Let* $\mathbf{Y} \sim TESN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau, [\mathbf{a}, \mathbf{b}])$. *For any measurable function* $g(\cdot)$, *we have that*

$$\mathbb{E}\left[g(\mathbf{Y})\frac{\phi_1(\tau + \boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}))}{\Phi_1(\tau + \boldsymbol{\lambda}^\top\boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}))}\right] = \frac{\eta L}{\mathcal{L}}\mathbb{E}[g(\mathbf{W})],$$

with $\eta = \phi_1(\tau; 0, 1 + \boldsymbol{\lambda}^\top\boldsymbol{\lambda})/\xi$, $L = L_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu} - \boldsymbol{\mu}^*, \boldsymbol{\Psi})$, $\mathcal{L} = \mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \tau)$, $\boldsymbol{\Psi} = \boldsymbol{\Sigma}^{1/2}(\mathbf{I}_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top)^{-1}\boldsymbol{\Sigma}^{1/2}$, $\boldsymbol{\mu}^* = \tau\boldsymbol{\Psi}\boldsymbol{\varphi}$, *and* $\mathbf{W} \sim TN_p(\boldsymbol{\mu} - \boldsymbol{\mu}^*, \boldsymbol{\Psi}, [\mathbf{a}, \mathbf{b}])$.

Proof See Lemma 1 in Galarza et al. [11].

Corollary 1 *Setting* $\tau = 0$*, it follows that* $\mathbf{Y} \sim TSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, [\mathbf{a}, \mathbf{b}])$ *and*

$$\mathbb{E}\left[g(\mathbf{Y}) \frac{\phi_1(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}))}{\Phi_1(\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}))} \right] = \frac{L_0}{\sqrt{\frac{\pi}{2}(1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})}\mathcal{L}_0} \mathbb{E}[g(\mathbf{W}_0)],$$

with $L_0 = L_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Psi})$, $\mathcal{L}_0 = \mathcal{L}_p(\mathbf{a}, \mathbf{b}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, 0)$ *and* $\mathbf{W}_0 \sim TN_p(\boldsymbol{\mu}, \boldsymbol{\Psi}, [\mathbf{a}, \mathbf{b}])$.

Proof The proof is straightforward. Setting $\tau = 0$, it suffices to find that $\boldsymbol{\mu}^* = \mathbf{0}$ and $\eta = \sqrt{2/\pi(1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})}$.

3 The Skew-Normal Linear Mixed-Effects Model with Censored Responses

The Statistical Model

In order to allow symmetric-asymmetric properties in real data sets, the SN-LMEC is defined by extending the normal mixed-effects models presented by Vaida and Liu [29]. The model is specified as follows:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{3}$$

where the subscript i is the subject index, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ is a $n_i \times 1$ vector of observed continuous responses for sample unit i, $\mathbf{X}_i$ is the $n_i \times p$ design matrix corresponding to the $p \times 1$ vector of fixed-effects $\boldsymbol{\beta}$, and $\mathbf{Z}_i$ is the $n_i \times q$ design matrix corresponding to the $q \times 1$ vector of random effects $\mathbf{b}_i$, and $\boldsymbol{\epsilon}_i$ is the $n_i \times 1$ vector of random errors.

In this work, we assume that

$$\begin{pmatrix} \mathbf{b}_i \\ \boldsymbol{\epsilon}_i \end{pmatrix} \overset{\text{ind.}}{\sim} \text{SN}_{q+n_i} \left(\begin{pmatrix} c\boldsymbol{\Delta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_i \end{pmatrix}, \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{0} \end{pmatrix} \right), \tag{4}$$

where $c = -\sqrt{2/\pi}$, $\boldsymbol{\Delta} = \mathbf{D}^{1/2}\boldsymbol{\delta}$, $\boldsymbol{\delta} = \dfrac{\boldsymbol{\lambda}}{(1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda})^{1/2}}$. The value of the location parameter, $c\boldsymbol{\Delta}$, of $\mathbf{b}_i$, is chosen to obtain $\mathbb{E}[\mathbf{b}_i] = 0$, as in the normal model. The dispersion matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ models the between-subjects variability and depends on the unknown and reduced parameter vector $\boldsymbol{\alpha}$ of dimension q. The correlation structure of the error vector is assumed to be $\boldsymbol{\Omega}_i = \sigma^2 \mathbf{E}_i$, where the $n_i \times n_i$ matrix $\mathbf{E}_i$ incorporates a time-dependence structure. Thus, we adopt a DEC structure for $\boldsymbol{\Omega}_i$, as Munoz et al. [26] proposed. This correlation structure allows us to deal with unequally spaced and unbalanced observations and is defined as

$$\boldsymbol{\Omega}_i = \sigma^2 \mathbf{E}_i = \sigma^2 \mathbf{E}_i(\boldsymbol{\phi}; \mathbf{t}_i) = \sigma^2 \left[\phi_1^{|t_{ij} - t_{ik}|^{\phi_2}} \right], \quad i = 1, \ldots, n, \quad j, k = 1, \ldots, n_i,$$

where ϕ_1 is the correlation parameter that describes the autocorrelation between observations separated by the absolute length of two time points; and ϕ_2 is the damping parameter that allows the acceleration of the exponential decay of the autocorrelation function, defining a continuous-time autoregressive model. For practical reasons, the parameter space of ϕ_1 and ϕ_2 is confined within $\boldsymbol{\Phi} = \{(\phi_1, \phi_2) : 0 < \phi_1 < 1, \phi_2 > 0\}$. Particular cases of these correlation structures are: (a) compound symmetry (CS), (b) the first-order autoregressive (AR (1)), and (c) the first-order moving average model (MA(1)). A more detailed discussion of the DEC structure can be found in Munoz et al. [26].

According to Lachos et al. [18], model (3) can be written hierarchically as

$$\begin{aligned} \mathbf{Y}_i|\mathbf{b}_i &\overset{\text{ind.}}{\sim} \mathrm{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Omega}_i), \\ \mathbf{b}_i|T_i = t_i &\overset{\text{ind.}}{\sim} \mathrm{N}_q(\boldsymbol{\Delta} t_i, \boldsymbol{\Gamma}), \\ T_i &\overset{\text{iid.}}{\sim} \mathrm{TN}(c, 1; (c, \infty)), \end{aligned} \tag{5}$$

where $\boldsymbol{\Gamma} = \mathbf{D} - \boldsymbol{\Delta}\boldsymbol{\Delta}^\top$.

It follows from (5) that the density of $\mathbf{Y}_i$ is

$$f(\mathbf{Y}_i) = 2\phi_{n_i}(\mathbf{Y}_i; \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i c\boldsymbol{\Delta}, \boldsymbol{\Sigma}_i)\Phi_1\left(\bar{\boldsymbol{\lambda}}_i^\top \boldsymbol{\Sigma}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta})\right), \tag{6}$$

i.e., $\mathbf{Y}_i \overset{\text{ind.}}{\sim} \mathrm{SN}_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i c\boldsymbol{\Delta}, \boldsymbol{\Sigma}_i, \bar{\boldsymbol{\lambda}}_i), i = 1, \ldots, n,$ where $\boldsymbol{\Sigma}_i = \boldsymbol{\Omega}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top$, $\boldsymbol{\Lambda}_i = (\mathbf{D}^{-1} + \mathbf{Z}_i^\top\boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i)^{-1}$ and $\bar{\boldsymbol{\lambda}}_i = \dfrac{\boldsymbol{\Sigma}_i^{-1/2}\mathbf{Z}_i\mathbf{D}\boldsymbol{\zeta}}{\sqrt{1 + \boldsymbol{\zeta}^\top\boldsymbol{\Lambda}_i\boldsymbol{\zeta}}}$, with $\boldsymbol{\zeta} = \mathbf{D}^{-1/2}\boldsymbol{\lambda}$.

As previously mentioned, the proposed model also considers censored observations, i.e., we assume that the response Y_{ij} is not fully observed for all i, j. Thus, we consider the approach proposed by Vaida and Liu [29] to model the censored responses. Let the observed data for the i-th subject be $(\mathbf{V}_i, \mathbf{C}_i)$, where $\mathbf{V}_i$ represents the vector of uncensored readings ($V_{ij} = V_{0ij}$) or censoring interval (V_{1ij},V_{2ij}), and $\mathbf{C}_i$ is the vector of censoring indicators, such that:

$$C_{ij} = \begin{cases} 1 & \text{if} \quad V_{1ij} \le Y_{ij} \le V_{2ij}, \\ 0 & \text{if} \quad Y_{ij} = V_{0ij}, \end{cases} \tag{7}$$

for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, n_i\}$, i.e., $C_{ij} = 1$ if Y_{ij} is located within a specific interval.

The model defined in (3)–(7) is henceforth called the SN-LMEC model.

The Likelihood Function

To obtain the likelihood function of the SN-LMEC model, first we treat separately the observed and censored components of $\mathbf{Y}_i$, i.e., $\mathbf{Y}_i = (\mathbf{Y}_i^{o\top}, \mathbf{Y}_i^{c\top})^\top$, with $C_{ij} = 0$ for all elements in $\mathbf{Y}_i^o$, and $C_{ij} = 1$ for all elements in $\mathbf{Y}_i^c$. Analogous, we write $\mathbf{V}_i = \text{vec}(\mathbf{V}_i^o, \mathbf{V}_i^c)$, where $\mathbf{V}_i^c = (\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c)$ with

$$\boldsymbol{\mu}_i = (\boldsymbol{\mu}_i^{o\top}, \boldsymbol{\mu}_i^{c\top})^\top,\ \boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_i^{oo} & \boldsymbol{\Sigma}_i^{oc} \\ \boldsymbol{\Sigma}_i^{co} & \boldsymbol{\Sigma}_i^{cc} \end{pmatrix},\ \bar{\boldsymbol{\lambda}}_i = (\bar{\boldsymbol{\lambda}}_i^{o\top}, \bar{\boldsymbol{\lambda}}_i^{c\top})^\top \text{ and } \boldsymbol{\varphi}_i = (\boldsymbol{\varphi}_i^{o\top}, \boldsymbol{\varphi}_i^{c\top})^\top,$$

where $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i c\boldsymbol{\Delta}$ and $\boldsymbol{\varphi}_i = \boldsymbol{\Sigma}_i^{-1/2}\bar{\boldsymbol{\lambda}}_i$. Then, using Proposition 1, we have that

$$\begin{aligned} \mathbf{Y}_i^o &\sim \text{SN}_{n_i^o}(\boldsymbol{\mu}_i^o, \boldsymbol{\Sigma}_i^{oo}, c_i^{co}\boldsymbol{\Sigma}_i^{oo^{1/2}}\tilde{\boldsymbol{\upsilon}}) \quad \text{and} \\ \mathbf{Y}_i^c|\mathbf{Y}_i^o = \mathbf{y}_i^o &\sim \text{ESN}_{n_i^c}(\boldsymbol{\mu}_i^{co}, \mathbf{S}_i, \mathbf{S}_i^{1/2}\boldsymbol{\varphi}_i^c, \tau_i^{co}), \end{aligned}$$

where $\tilde{\boldsymbol{\upsilon}} = \boldsymbol{\varphi}_i^o + \boldsymbol{\Sigma}_i^{oo^{-1}}\boldsymbol{\Sigma}_i^{oc}\boldsymbol{\varphi}_i^c$, $\boldsymbol{\mu}_i^{co} = \boldsymbol{\mu}_i^c + \boldsymbol{\Sigma}_i^{co}\boldsymbol{\Sigma}_i^{oo^{-1}}(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)$, $\mathbf{S}_i = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co}(\boldsymbol{\Sigma}_i^{oo})^{-1}\boldsymbol{\Sigma}_i^{oc}$, $c_i^{co} = (1 + \boldsymbol{\varphi}_i^{c\top}\mathbf{S}_i\boldsymbol{\varphi}_i^c)^{-1/2}$ and $\tau_i^{co} = \tilde{\boldsymbol{\upsilon}}^\top(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o)$.

Thus, the likelihood for the ith subject is given by

$$\begin{aligned} L_i(\boldsymbol{\theta}) = L_i &= f(\mathbf{y}_i|\boldsymbol{\theta}) = f(\mathbf{V}_i|\mathbf{C}_i, \boldsymbol{\theta}) = f(\mathbf{y}_i^o|\boldsymbol{\theta})\mathbb{P}(\mathbf{V}_{1i}^c \le \mathbf{y}_i^c \le \mathbf{V}_{2i}^c|\mathbf{y}_i^o, \boldsymbol{\theta}) \\ &= \text{SN}_{n_i^o}(\boldsymbol{\mu}_i^o, \boldsymbol{\Sigma}_i^{oo}, c_i^{co}\boldsymbol{\Sigma}_i^{oo^{1/2}}\tilde{\boldsymbol{\upsilon}})\mathcal{L}_{n_i^c}(\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c; \boldsymbol{\mu}_i^{co}, \mathbf{S}_i, \mathbf{S}_i^{1/2}\boldsymbol{\varphi}_i^c, \tau_i^{co}), \end{aligned}$$

and the log-likelihood function for the observed data is given by $\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n \log L_i$.

The ECM Algorithm

This section describes how to use the ECM algorithm to compute the Maximum Likelihood (ML) estimation of the SN-LMEC model. The EM algorithm proposed initially by Dempster et al. [10] has several appealing features such as stability of monotone convergence with each iteration increasing the likelihood and simplicity of implementation. Here, we used the ECM algorithm proposed by Meng and Rubin [25]; the ECM replaces the maximization step of EM with a set of conditional maximization steps.

Let $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top)^\top$ and $\mathbf{t} = (t_1, \ldots, t_n)^\top$. Also, let $\mathbf{V} = \text{vec}(\mathbf{V}_1, \ldots, \mathbf{V}_n)$ and $\mathbf{C} = \text{vec}(\mathbf{C}_1, \ldots, \mathbf{C}_n)$, where $(\mathbf{V}_i, \mathbf{C}_i)$ is observed for the ith subject. Treating $\mathbf{y}$, $\mathbf{b}$ and $\mathbf{t}$ as hypothetical missing data, and augmenting with the observed data $\mathbf{V}$, $\mathbf{C}$, we set $\mathbf{y}_c = (\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{y}^\top, \mathbf{b}^\top, \mathbf{t}^\top)^\top$ as the complete data. Hence, it follows from (5) that the complete-data log-likelihood function is of the form

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) &= \sum_{i=1}^{n}\left[\log f(\mathbf{y}_i|\mathbf{b}_i) + \log f(\mathbf{b}_i|t_i) + \log f(t_i)\right] \\ &= \sum_{i=1}^{n}\left\{-\frac{1}{2}\log|\boldsymbol{\Omega}_i| - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top\boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)\right. \\ &\quad \left. - \frac{1}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}(\mathbf{b}_i - t_i\boldsymbol{\Delta})^\top\boldsymbol{\Gamma}^{-1}(\mathbf{b}_i - t_i\boldsymbol{\Delta})\right\} + C,\end{aligned}$$

where C is a constant that is independent of the parameter vector $\boldsymbol{\theta}$.

The E-step of the ECM algorithm computes the expected value of the complete-data log-likelihood function given the observed data $\mathbf{V}$, $\mathbf{C}$ and current values $\widehat{\boldsymbol{\theta}}^{(k)}$, yielding the so-called Q-function

$$Q\left(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}\right) = \mathbb{E}\left[\ell_c(\boldsymbol{\theta};\mathbf{y}_c)\middle|\mathbf{V},\mathbf{C},\widehat{\boldsymbol{\theta}}^{(k)}\right] = \sum_{i=1}^{n} Q_{1i}\left(\boldsymbol{\beta},\sigma^2,\boldsymbol{\phi}|\widehat{\boldsymbol{\theta}}^{(k)}\right) + \sum_{i=1}^{n} Q_{2i}\left(\boldsymbol{\alpha},\boldsymbol{\lambda}|\widehat{\boldsymbol{\theta}}^{(k)}\right),$$

where $\widehat{\boldsymbol{\theta}}^{(k)} = (\widehat{\boldsymbol{\beta}}^{(k)\top}, \widehat{\sigma^2}^{(k)}, \widehat{\boldsymbol{\alpha}}^{(k)\top}, \widehat{\boldsymbol{\phi}}^{(k)\top}, \widehat{\boldsymbol{\lambda}}^{(k)\top})^\top$,

$$\begin{aligned}Q_{1i}\left(\boldsymbol{\beta},\sigma^2,\boldsymbol{\phi}|\widehat{\boldsymbol{\theta}}^{(k)}\right) &= -\frac{n_i}{2}\log\widehat{\sigma^2}^{(k)} - \frac{1}{2}\log|\widehat{\mathbf{E}}_i^{(k)}| - \frac{1}{2\widehat{\sigma^2}^{(k)}}\left[\widehat{\boldsymbol{\beta}}^{(k)^\top}\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1^{(k)}}\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\right. \\ &\quad \left. - 2\widehat{\boldsymbol{\beta}}^{(k)^\top}\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1^{(k)}}\left(\widehat{\mathbf{y}}_i^{(k)} - \mathbf{Z}_i\widehat{\mathbf{b}}_i^{(k)}\right) + \widehat{a}_i^{(k)}(\widehat{\boldsymbol{\phi}}^{(k)})\right], \qquad (8)\end{aligned}$$

$$\begin{aligned}Q_{2i}\left(\boldsymbol{\alpha},\boldsymbol{\lambda}|\widehat{\boldsymbol{\theta}}^{(k)}\right) &= -\frac{1}{2}\log|\widehat{\boldsymbol{\Gamma}}^{(k)}| - \frac{1}{2}\mathrm{tr}\left[\widehat{\boldsymbol{\Gamma}}^{-1^{(k)}}\left(\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(k)} - \widehat{\boldsymbol{\Delta}}^{(k)}\widehat{t_i\mathbf{b}_i^\top}^{(k)} - \widehat{t_i\mathbf{b}_i}^{(k)}\widehat{\boldsymbol{\Delta}}^{(k)^\top}\right.\right. \\ &\quad \left.\left. + \widehat{t_i^2}^{(k)}\widehat{\boldsymbol{\Delta}}^{(k)}\widehat{\boldsymbol{\Delta}}^{(k)^\top}\right)\right], \qquad (9)\end{aligned}$$

with $\widehat{a}_i^{(k)}(\widehat{\boldsymbol{\phi}}^{(k)}) = \mathrm{tr}\left[\widehat{\mathbf{E}}_i^{-1^{(k)}}\left(\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)} - 2\widehat{\mathbf{y}_i\mathbf{b}_i^\top}^{(k)}\mathbf{Z}_i^\top + \mathbf{Z}_i\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(k)}\mathbf{Z}_i^\top\right)\right]$.

The following conditional distributions help obtain the conditional expectations of missing data. From Lachos et al. [18], we have that

$$\begin{aligned}\mathbf{b}_i|T_i = t_i, \mathbf{Y}_i = \mathbf{y}_i &\sim \mathrm{N}_q\left(\mathbf{s}_i t_i + \mathbf{B}_i\mathbf{Z}_i^\top\boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \mathbf{B}_i\right), \\ T_i|\mathbf{Y}_i = \mathbf{y}_i &\sim \mathrm{TN}_1\left(c + m_i, M_i^2; (0,\infty)\right), \\ \mathbf{Y}_i &\sim \mathrm{SN}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \bar{\boldsymbol{\lambda}}_i),\end{aligned}$$

where $M_i = (1 + \boldsymbol{\Delta}^\top\mathbf{Z}_i^\top\boldsymbol{\Upsilon}_i^{-1}\mathbf{Z}_i\boldsymbol{\Delta})^{-1/2}$, $m_i = M_i^2\boldsymbol{\Delta}^\top\mathbf{Z}_i^\top\boldsymbol{\Upsilon}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta})$, $\mathbf{B}_i = (\boldsymbol{\Gamma}^{-1} + \mathbf{Z}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{Z}_i)^{-1}$, $\mathbf{s}_i = (\mathbf{I}_q - \mathbf{B}_i\mathbf{Z}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{Z}_i)\boldsymbol{\Delta}$, $\boldsymbol{\Upsilon}_i = \boldsymbol{\Omega}_i + \mathbf{Z}_i\boldsymbol{\Gamma}\mathbf{Z}_i^\top$.

Therefore, the Q-function is entirely determined by the knowledge of the following expectations:

$$
\begin{aligned}
\widehat{t_i}^{(k)} &= \mathbb{E}\left[T_i \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = c + \widehat{M}_i^{2(k)} \widehat{\boldsymbol{\Delta}}^{(k)\top} \mathbf{Z}_i^\top \widehat{\boldsymbol{\Upsilon}}_i^{-1(k)} \left(\widehat{\mathbf{y}}_i^{(k)} - \mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} - \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}^{(k)}\right) \\
&+ \widehat{M}_i^{(k)} \widehat{\kappa}_i^{(k)}, \\
\widehat{t_i^2}^{(k)} &= \mathbb{E}\left[T_i^2 \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \widehat{M}_i^{4(k)} \widehat{\boldsymbol{\Delta}}^{(k)\top} \mathbf{Z}_i^\top \widehat{\boldsymbol{\Upsilon}}_i^{-1(k)} \widehat{\mathbf{R}}_i^{(k)} \widehat{\boldsymbol{\Upsilon}}_i^{-1(k)} \mathbf{Z}_i \widehat{\boldsymbol{\Delta}}^{(k)} + \widehat{M}_i^{2(k)} + c^2 \\
&+ 2c \widehat{M}_i^{2(k)} \widehat{\boldsymbol{\Delta}}^{(k)\top} \mathbf{Z}_i^\top \widehat{\boldsymbol{\Upsilon}}_i^{-1(k)} \left(\widehat{\mathbf{y}}_i^{(k)} - \mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} - \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}^{(k)}\right) \\
&+ \widehat{M}_i^{3(k)} \widehat{\boldsymbol{\Delta}}^{(k)\top} \mathbf{Z}_i^\top \widehat{\boldsymbol{\Upsilon}}_i^{-1(k)} \left(\widehat{\kappa_i \mathbf{y}_i}^{(k)} - \widehat{\kappa}_i^{(k)} \left(\mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}^{(k)}\right)\right) + 2c \widehat{M}_i^{(k)} \widehat{\kappa}_i^{(k)}, \\
\widehat{t_i \mathbf{y}_i}^{(k)} &= \mathbb{E}\left[T_i \mathbf{Y}_i \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \widehat{M}_i^{2(k)} \left(\widehat{\mathbf{y}_i \mathbf{y}_i^\top}^{(k)} - \widehat{\mathbf{y}}_i^{(k)} \left(\mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}^{(k)}\right)^\top\right) \widehat{\boldsymbol{\Upsilon}}_i^{-1(k)} \mathbf{Z}_i \widehat{\boldsymbol{\Delta}}^{(k)} \\
&+ \widehat{M}_i^{(k)} \widehat{\kappa_i \mathbf{y}_i}^{(k)} + c \widehat{\mathbf{y}}_i^{(k)}, \\
\widehat{\mathbf{b}}_i^{(k)} &= \mathbb{E}\left[\mathbf{b}_i \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \widehat{\mathbf{s}}_i^{(k)} \widehat{t}_i^{(k)} + \widehat{\mathbf{B}}_i^{(k)} \mathbf{Z}_i^\top \widehat{\boldsymbol{\Omega}}_i^{-1(k)} \left(\widehat{\mathbf{y}}_i^{(k)} - \mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)}\right), \\
\widehat{\mathbf{b}_i \mathbf{b}_i^\top}^{(k)} &= \mathbb{E}\left[\mathbf{b}_i \mathbf{b}_i^\top \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \widehat{t_i^2}^{(k)} \widehat{\mathbf{s}}_i^{(k)} \widehat{\mathbf{s}}_i^{(k)\top} + 2 \widehat{\mathbf{B}}_i^{(k)} \mathbf{Z}_i^\top \widehat{\boldsymbol{\Omega}}_i^{-1(k)} \left(\widehat{t_i \mathbf{y}_i}^{(k)} - \widehat{t}_i^{(k)} \mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)}\right) \widehat{\mathbf{s}}_i^{(k)\top} \\
&+ \widehat{\mathbf{B}}_i^{(k)} \mathbf{Z}_i^\top \widehat{\boldsymbol{\Omega}}_i^{-1(k)} \widehat{\mathbf{r}}_i^{(k)} \widehat{\boldsymbol{\Omega}}_i^{-1(k)} \mathbf{Z}_i \widehat{\mathbf{B}}_i^{(k)} + \widehat{\mathbf{B}}_i^{(k)}, \\
\widehat{\mathbf{y}_i \mathbf{b}_i^\top}^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \mathbf{b}_i^\top \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \left(\widehat{\mathbf{y}_i \mathbf{y}_i^\top}^{(k)} - \widehat{\mathbf{y}}_i^{(k)} \widehat{\boldsymbol{\beta}}^{(k)\top} \mathbf{X}_i^\top\right) \widehat{\boldsymbol{\Omega}}_i^{-1(k)} \mathbf{Z}_i \widehat{\mathbf{B}}_i^{(k)} + \widehat{t_i \mathbf{y}_i}^{(k)} \widehat{\mathbf{s}}_i^{(k)\top}, \\
\widehat{t_i \mathbf{b}_i^\top}^{(k)} &= \mathbb{E}\left[T_i \mathbf{b}_i^\top \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \left(\widehat{t_i \mathbf{y}_i^\top}^{(k)} - \widehat{t}_i^{(k)} \widehat{\boldsymbol{\beta}}^{(k)\top} \mathbf{X}_i^\top\right) \widehat{\boldsymbol{\Omega}}_i^{-1(k)} \mathbf{Z}_i \widehat{\mathbf{B}}_i^{(k)} + \widehat{t_i^2}^{(k)} \widehat{\mathbf{s}}_i^{(k)\top},
\end{aligned}
$$

with $\widehat{\mathbf{R}}_i^{(k)} = \widehat{\mathbf{y}_i \mathbf{y}_i^\top}^{(k)} - 2\widehat{\mathbf{y}}_i^{(k)} \left(\mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}^{(k)}\right)^\top + \left(\mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}^{(k)}\right) \left(\mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} + \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}^{(k)}\right)^\top$ and $\widehat{\mathbf{r}}_i^{(k)} = \widehat{\mathbf{y}_i \mathbf{y}_i^\top}^{(k)} - 2\widehat{\mathbf{y}}_i^{(k)} \widehat{\boldsymbol{\beta}}^{(k)\top} \mathbf{X}_i^\top + \mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} \widehat{\boldsymbol{\beta}}^{(k)\top} \mathbf{X}_i^\top$.

It is easy to see that the E-step reduces only to the computation of $\widehat{\mathbf{y}}_i$, $\widehat{\mathbf{y}_i \mathbf{y}_i^\top}$, $\widehat{\kappa}_i$ and $\widehat{\kappa_i \mathbf{y}_i}$. These expected values can be determined in closed form using Lemma 2 and Corollary 1, as follows:

1. If the ith subject has only non-censored components, then

$$
\begin{aligned}
\widehat{\mathbf{y}}_i^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \mathbf{y}_i, \\
\widehat{\mathbf{y}_i \mathbf{y}_i^\top}^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \mathbf{Y}_i^\top \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \mathbf{y}_i \mathbf{y}_i^\top, \\
\widehat{\kappa}_i^{(k)} &= \mathbb{E}\left[\mathrm{W}_\Phi\left(\widehat{\boldsymbol{\lambda}}_i \widehat{\boldsymbol{\Sigma}}_i^{-1/2} \left(\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}} - \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}\right)\right) \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] \\
&= \mathrm{W}_\Phi\left(\widehat{\boldsymbol{\lambda}}_i^{(k)} \widehat{\boldsymbol{\Sigma}}_i^{-1/2(k)} \left(\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}}^{(k)} - \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}^{(k)}\right)\right), \\
\widehat{\kappa_i \mathbf{y}_i}^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \mathrm{W}_\Phi\left(\widehat{\boldsymbol{\lambda}}_i \widehat{\boldsymbol{\Sigma}}_i^{-1/2} \left(\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}} - \mathbf{Z}_i c \widehat{\boldsymbol{\Delta}}\right)\right) \middle| \mathbf{V}_i = \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \mathbf{y}_i \widehat{\kappa}_i^{(k)},
\end{aligned}
$$

   with $\mathrm{W}_\Phi(x) = \phi_1(x)/\Phi(x)$, $x \in \mathbb{R}$.
2. If the ith subject has only censored components, then from Corollary 1,

$$
\begin{aligned}
\widehat{\mathbf{y}}_i^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \mathbb{E}\left[\mathbf{W}_i \middle| \widehat{\boldsymbol{\theta}}^{(k)}\right], \\
\widehat{\mathbf{y}_i \mathbf{y}_i^\top}^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \mathbf{Y}_i^\top \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \mathbb{E}\left[\mathbf{W}_i \mathbf{W}_i^\top \middle| \widehat{\boldsymbol{\theta}}^{(k)}\right], \\
\widehat{\kappa}_i^{(k)} &= \mathbb{E}\left[\mathrm{W}_\Phi\left(\widehat{\tilde{\boldsymbol{\lambda}}}_i \widehat{\boldsymbol{\Sigma}}_i^{-1/2}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} - \mathbf{Z}_i c\widehat{\boldsymbol{\Delta}}\right)\right) \middle| \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] \\
&= \frac{1}{\sqrt{\frac{\pi}{2}\left(1 + \widehat{\tilde{\boldsymbol{\lambda}}}_i^{(k)\top}\widehat{\tilde{\boldsymbol{\lambda}}}_i^{(k)}\right)}} \frac{L_{n_i}\left(\mathbf{V}_{1i}, \mathbf{V}_{2i}; \widehat{\boldsymbol{\mu}}_i^{(k)}, \widehat{\boldsymbol{\Psi}}_i^{(k)}\right)}{\mathcal{L}_{n_i}\left(\mathbf{V}_{1i}, \mathbf{V}_{2i}; \widehat{\boldsymbol{\mu}}_i^{(k)}, \widehat{\boldsymbol{\Sigma}}_i^{(k)}, \widehat{\tilde{\boldsymbol{\lambda}}}_i^{(k)}, 0\right)}, \\
\widehat{\kappa_i \mathbf{y}_i}^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \mathrm{W}_\Phi\left(\widehat{\tilde{\boldsymbol{\lambda}}}_i \widehat{\boldsymbol{\Sigma}}_i^{-1/2}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} - \mathbf{Z}_i c\widehat{\boldsymbol{\Delta}}\right)\right) \middle| \mathbf{V}_i = \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] \\
&= \mathbb{E}\left[\mathbf{W}_{0i} \middle| \widehat{\boldsymbol{\theta}}^{(k)}\right]\widehat{\kappa}_i^{(k)},
\end{aligned}
$$

where $\mathbf{W}_i \sim \mathrm{TSN}_{n_i}\left(\widehat{\boldsymbol{\mu}}_i^{(k)}, \widehat{\boldsymbol{\Sigma}}_i^{(k)}, \widehat{\tilde{\boldsymbol{\lambda}}}_i^{(k)}, [\mathbf{V}_{1i}, \mathbf{V}_{2i}]\right)$, $\mathbf{W}_{0i} \sim \mathrm{TN}_{n_i}\left(\widehat{\boldsymbol{\mu}}_i^{(k)}, \widehat{\boldsymbol{\Psi}}_i^{(k)}, [\mathbf{V}_{1i}, \mathbf{V}_{2i}]\right)$, and $\widehat{\boldsymbol{\Psi}}_i^{(k)} = \widehat{\boldsymbol{\Sigma}}_i^{(k)1/2}\left(\mathbf{I}_{n_i} + \widehat{\tilde{\boldsymbol{\lambda}}}_i^{(k)}\widehat{\tilde{\boldsymbol{\lambda}}}_i^{(k)\top}\right)^{-1}\widehat{\boldsymbol{\Sigma}}_i^{(k)1/2}$.

3. If the ith subject has censored and uncensored components from Proposition 1 and Lemma 2, we have

$$
\begin{aligned}
\widehat{\mathbf{y}}_i^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \middle| \mathbf{V}_i, \mathbf{C}_i, \mathbf{Y}_i^o, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \mathrm{vec}(\mathbf{y}_i^o, \widehat{\mathbf{w}}_i^{(k)}), \\
\widehat{\mathbf{y}_i \mathbf{y}_i^\top}^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \mathbf{Y}_i^\top \middle| \mathbf{V}_i, \mathbf{C}_i, \mathbf{Y}_i^o, \widehat{\boldsymbol{\theta}}^{(k)}\right] = \begin{pmatrix} \mathbf{y}_i^o \mathbf{y}_i^{o\top} & \mathbf{y}_i^o \widehat{\mathbf{w}}_i^{(k)\top} \\ \widehat{\mathbf{w}}_i^{(k)} \mathbf{y}_i^{o\top} & \widehat{\mathbf{w}_i \mathbf{w}_i^\top} \end{pmatrix}, \\
\widehat{\kappa}_i^{(k)} &= \mathbb{E}\left[\mathrm{W}_\Phi\left(\widehat{\tilde{\boldsymbol{\lambda}}}_i \widehat{\boldsymbol{\Sigma}}_i^{-1/2}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} - \mathbf{Z}_i c\widehat{\boldsymbol{\Delta}}\right)\right) \middle| \mathbf{V}_i, \mathbf{C}_i, \mathbf{Y}_i^o, \widehat{\boldsymbol{\theta}}^{(k)}\right] \\
&= \frac{\phi_1\left(\widehat{\tau}_i^{co(k)}; 0, 1 + \widehat{\tilde{\boldsymbol{\lambda}}}_i^{co(k)\top}\widehat{\tilde{\boldsymbol{\lambda}}}_i^{co(k)}\right)}{\Phi_1(\tilde{\tau})} \frac{L_{n_i^c}\left(\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c; \widehat{\tilde{\boldsymbol{\mu}}}_i^{(k)}, \widehat{\tilde{\boldsymbol{\Psi}}}_i^{(k)}\right)}{\mathcal{L}_{n_i^c}\left(\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c; \widehat{\boldsymbol{\mu}}_i^{co(k)}, \widehat{\mathbf{S}}_i^{(k)}, \widehat{\boldsymbol{\lambda}}_i^{co(k)}, \widehat{\tau}_i^{co(k)}\right)}, \\
\widehat{\kappa_i \mathbf{y}_i}^{(k)} &= \mathbb{E}\left[\mathbf{Y}_i \mathrm{W}_\Phi\left(\widehat{\tilde{\boldsymbol{\lambda}}}_i \widehat{\boldsymbol{\Sigma}}_i^{-1/2}\left(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} - \mathbf{Z}_i c\widehat{\boldsymbol{\Delta}}\right)\right) \middle| \mathbf{V}_i = \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\right] \\
&= \mathrm{vec}(\mathbf{y}_i^o, \widehat{\mathbf{w}}_{0i}^{(k)})\widehat{\kappa}_i^{(k)},
\end{aligned}
$$

where

$$
\widehat{\mathbf{w}}_i^{(k)} = \mathbb{E}\left[\mathbf{W}_i \middle| \widehat{\boldsymbol{\theta}}^{(k)}\right], \quad \widehat{\mathbf{w}_i \mathbf{w}_i^\top} = \mathbb{E}\left[\mathbf{W}_i \mathbf{W}_i^\top \middle| \widehat{\boldsymbol{\theta}}^{(k)}\right], \quad \widehat{\mathbf{w}}_{0i}^{(k)} = \mathbb{E}\left[\mathbf{W}_{0i} \middle| \widehat{\boldsymbol{\theta}}^{(k)}\right],
$$

with

$$
\mathbf{W}_i \sim \mathrm{TESN}_{n_i^c}\left(\widehat{\boldsymbol{\mu}}_i^{co(k)}, \widehat{\mathbf{S}}_i^{(k)}, \widehat{\boldsymbol{\lambda}}_i^{co(k)}, \widehat{\tau}_i^{co(k)}, [\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c]\right),
$$

$$
\mathbf{W}_{0i} \sim \mathrm{TN}_{n_i^c}\left(\widehat{\tilde{\boldsymbol{\mu}}}_i^{(k)}, \widehat{\tilde{\boldsymbol{\Psi}}}_i^{(k)}, [\mathbf{V}_{1i}^c, \mathbf{V}_{2i}^c]\right),
$$

and $\tilde{\tau} = \dfrac{\tau_i^{co}}{(1+\boldsymbol{\lambda}_i^{co\top}\boldsymbol{\lambda}_i^{co})^{1/2}}$, $\boldsymbol{\lambda}_i^{co} = \mathbf{S}_i^{1/2}\boldsymbol{\varphi}_i^c$, $\tilde{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i^{co} - \tau_i^{co}\tilde{\boldsymbol{\Psi}}_i\boldsymbol{\varphi}_i^c$, $\tilde{\boldsymbol{\Psi}}_i = \mathbf{S}_i^{1/2}(\mathbf{I}_{n_i^c} + \boldsymbol{\lambda}_i^{co}\boldsymbol{\lambda}_i^{co\top})^{-1}\mathbf{S}_i^{1/2}$.

It can be noted that we need the first and second moments of a TESN distribution. These can be determined in closed form using recurrence relations. For more details on the computation of these moments, we refer to Galarza Morales et al. [13]. These moments can be obtained in the R package `MomTrunc` [12].

The CM-steps then conditionally maximizes $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$ and obtains a new estimate $\widehat{\boldsymbol{\theta}}^{(k+1)}$, as follows:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left(\sum_{i=1}^{n}\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1(k)}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1(k)}\left(\widehat{\mathbf{y}}_i^{(k)} - \mathbf{Z}_i\widehat{\mathbf{b}}_i^{(k)}\right),$$

$$\widehat{\boldsymbol{\Delta}}^{(k+1)} = \frac{\sum_{i=1}^{n}\widehat{t_i\mathbf{b}_i}^{(k)}}{\sum_{i=1}^{n}\widehat{t_i^2}^{(k)}},$$

$$\widehat{\boldsymbol{\Gamma}}^{(k+1)} = \frac{1}{N}\sum_{i=1}^{n}\left(\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(k)} - \widehat{t_i\mathbf{b}_i}^{(k)}\boldsymbol{\Delta}^{(k+1)\top} - \boldsymbol{\Delta}^{(k+1)}\widehat{t_i\mathbf{b}_i}^{(k)\top} + \widehat{t_i^2}^{(k)}\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\Delta}^{(k+1)\top}\right),$$

$$\widehat{\sigma^2}^{(k+1)} = \frac{1}{N}\sum_{i=1}^{n}\left[\widehat{a}_i^{(k)} + \widehat{\boldsymbol{\beta}}^{(k+1)^\top}\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1(k)}\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k+1)} - 2\widehat{\boldsymbol{\beta}}^{(k+1)^\top}\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1(k)}\left(\widehat{\mathbf{y}}_i^{(k)} - \mathbf{Z}_i\widehat{\mathbf{b}}_i^{(k)}\right)\right],$$

$$\widehat{\phi}^{(k+1)} = \underset{\phi\in(0,1)\times\mathcal{R}^+}{\arg\max}\left(-\frac{1}{2}\log(|\mathbf{E}_i|) - \frac{1}{2\widehat{\sigma^2}^{(k+1)}}\left[-2\widehat{\boldsymbol{\beta}}^{(k+1)^\top}\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1(k)}\left(\widehat{\mathbf{y}}_i^{(k)} - \mathbf{Z}_i\widehat{\mathbf{b}}_i^{(k)}\right)\right.\right.$$
$$\left.\left. + \widehat{\boldsymbol{\beta}}^{(k+1)^\top}\mathbf{X}_i^\top\widehat{\mathbf{E}}_i^{-1(k)}\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k+1)} + \widehat{a}_i^{(k)}\right]\right),$$

where $N = \sum_{i=1}^{n} n_i$. The skewness parameter vector, and the parameters of the scale matrix of the random effects $\mathbf{b}$, can be estimated by noting that

$$\widehat{\mathbf{D}}^{(k+1)} = \widehat{\boldsymbol{\Gamma}}^{(k+1)} + \widehat{\boldsymbol{\Delta}}^{(k+1)}\widehat{\boldsymbol{\Delta}}^{(k+1)\top} \text{ and } \widehat{\boldsymbol{\lambda}}^{(k+1)} = \frac{\widehat{\mathbf{D}}^{(k+1)^{-1/2}}\widehat{\boldsymbol{\Delta}}^{(k+1)}}{\left(1 - \widehat{\boldsymbol{\Delta}}^{(k+1)\top}\widehat{\mathbf{D}}^{(k+1)^{-1}}\widehat{\boldsymbol{\Delta}}^{(k+1)}\right)^{1/2}}.$$

This process is iterated until some distance between two successive evaluations of the log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{y})$ in Sect. 3, such as $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})|$ or $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1|$, becomes small enough.

Approximate Standard Errors

In what follows, we reparameterize $\mathbf{D} = \mathbf{F}^2$ for ease of computation and theoretical derivation, where $\mathbf{F}$ is the square root of $\mathbf{D}$, i.e., $\mathbf{F}^{1/2}$, containing $q(q+1)/2$ distinct elements $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{q(q+1)/2})^\top$.

Now, we use the empirical information matrix to compute the asymptotic covariance of the ML estimates. Following the works of Louis [22] and Matos et al. [24], the individual score can be determined as

$$\mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta}) = \frac{\partial \log f(\mathbf{y}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}\left(\frac{\partial \ell_{ic}(\boldsymbol{\theta}|\mathbf{y}_{ic})}{\partial \boldsymbol{\theta}}\Big|\mathbf{V}_i, \mathbf{C}_i.\boldsymbol{\theta}\right),$$

where $\ell_{ic}(\boldsymbol{\theta}|\mathbf{y}_{ic})$ is the complete-data log-likelihood function formed from the complete observation $\mathbf{y}_{ic}$ (for more details, see Louis [22]). As a result, the empirical information matrix $\mathbf{I}_e(\boldsymbol{\theta}|\mathbf{y})$ is reduced to

$$\mathbf{I}_e(\widehat{\boldsymbol{\theta}}|\mathbf{y}) = \sum_{i=1}^{n} \widehat{\mathbf{s}}_i \widehat{\mathbf{s}}_i^{\top},$$

where $\widehat{\mathbf{s}}_i = \left(\widehat{\mathbf{s}}_i(\boldsymbol{\beta})^{\top}, \widehat{\mathbf{s}}_i(\sigma^2), \widehat{\mathbf{s}}_i(\boldsymbol{\alpha})^{\top}, \widehat{\mathbf{s}}_i(\boldsymbol{\phi})^{\top}, \widehat{\mathbf{s}}_i(\boldsymbol{\lambda})^{\top}\right)^{\top}$, has elements given by

$$\begin{aligned}
\widehat{\mathbf{s}}_i(\boldsymbol{\beta}) &= \frac{1}{\widehat{\sigma^2}}\left[\mathbf{X}_i^{\top}\widehat{\mathbf{E}}_i^{-1}\left(\widehat{\mathbf{y}}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} - \mathbf{Z}_i\widehat{\mathbf{b}}_i\right)\right],\\
\widehat{\mathbf{s}}_i(\sigma^2) &= -\frac{n_i}{2\widehat{\sigma^2}} + \frac{1}{2\widehat{\sigma^2}^2}\left[\widehat{a}_i - 2\widehat{\boldsymbol{\beta}}^{\top}\mathbf{X}_i^{\top}\widehat{\mathbf{E}}_i^{-1}\left(\widehat{\mathbf{y}}_i - \mathbf{Z}_i\widehat{\mathbf{b}}_i\right) + \widehat{\boldsymbol{\beta}}^{\top}\mathbf{X}_i^{\top}\widehat{\mathbf{E}}_i^{-1}\mathbf{X}_i\widehat{\boldsymbol{\beta}}\right],\\
\widehat{\mathbf{s}}_i(\boldsymbol{\alpha}) &= \left(\widehat{\mathbf{s}}_i(\alpha_1), \ldots, \widehat{\mathbf{s}}_i(\alpha_{q(q+1)/2})\right)^{\top},\\
\widehat{\mathbf{s}}_i(\boldsymbol{\phi}) &= \left(\widehat{\mathbf{s}}_i(\phi_1), \widehat{\mathbf{s}}_i(\phi_2)\right)^{\top},\\
\widehat{\mathbf{s}}_i(\boldsymbol{\lambda}) &= \left(\widehat{\mathbf{s}}_i(\lambda_1), \ldots, \widehat{\mathbf{s}}_i(\lambda_q)\right)^{\top},
\end{aligned}$$

with $\widehat{a}_i = \operatorname{tr}\left[\widehat{\mathbf{E}}_i^{-1}\left(\widehat{\mathbf{y}_i\mathbf{y}_i^{\top}}^{(k)} - 2\widehat{\mathbf{y}_i\mathbf{b}_i^{\top}}^{(k)}\mathbf{Z}_i^{\top} + \mathbf{Z}_i\widehat{\mathbf{b}_i\mathbf{b}_i^{\top}}^{(k)}\mathbf{Z}_i^{\top}\right)\right]$,

$$\begin{aligned}
\widehat{\mathbf{s}}_i(\alpha_r) &= -\frac{1}{2}\operatorname{tr}\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\alpha_r}\right) + \frac{1}{2}\left\{\operatorname{tr}\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\alpha_r}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\mathbf{b}_i\mathbf{b}_i^{\top}}\right) + \widehat{t_i\mathbf{b}_i^{\top}}\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\mathbf{F}}_r\widehat{\boldsymbol{\delta}} - \widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\alpha_r}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}\right)\right.\\
&\quad + \left.\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\mathbf{F}}_r\widehat{\boldsymbol{\delta}} - \widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\alpha_r}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}\right)^{\top}\widehat{t_i\mathbf{b}_i} - \widehat{t_i^2}\left[\widehat{\boldsymbol{\Delta}}^{\top}\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\mathbf{F}}_r\widehat{\boldsymbol{\delta}} - \widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\alpha_r}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}\right) + \widehat{\boldsymbol{\delta}}^{\top}\dot{\mathbf{F}}_r\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}\right]\right\},\\
\widehat{\mathbf{s}}_i(\phi_s) &= \frac{1}{2\widehat{\sigma^2}}\left[\operatorname{tr}\left(\widehat{\mathbf{y}_i\mathbf{y}_i^{\top}} - 2\widehat{\mathbf{y}_i\mathbf{b}_i^{\top}}\mathbf{Z}_i^{\top} + \mathbf{Z}_i\widehat{\mathbf{b}_i\mathbf{b}_i^{\top}}\mathbf{Z}_i^{\top} - 2\left(\widehat{\mathbf{y}}_i - \mathbf{Z}_i\widehat{\mathbf{b}}_i\right)\widehat{\boldsymbol{\beta}}^{\top}\mathbf{X}_i^{\top} + \mathbf{X}_i\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\top}\mathbf{X}_i^{\top}\right)\widehat{\mathbf{E}}_i^{-1}\dot{\mathbf{E}}_i^{s}\widehat{\mathbf{E}}_i^{-1}\right]\\
&\quad - \frac{1}{2}\operatorname{tr}\left(\widehat{\mathbf{E}}_i^{-1}\dot{\mathbf{E}}_i^{s}\right),\\
\widehat{\mathbf{s}}_i(\lambda_t) &= -\frac{1}{2}\operatorname{tr}\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\lambda_t}\right) + \frac{1}{2}\left\{\operatorname{tr}\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\lambda_t}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\mathbf{b}_i\mathbf{b}_i^{\top}}\right) + \widehat{t_i\mathbf{b}_i^{\top}}\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Delta}}_{\lambda_t} - \widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\lambda_t}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}\right)\right.\\
&\quad + \left.\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Delta}}_{\lambda_t} - \widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\lambda_t}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}\right)^{\top}\widehat{t_i\mathbf{b}_i} - \widehat{t_i^2}\left[\widehat{\boldsymbol{\Delta}}^{\top}\left(\widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Delta}}_{\lambda_t} - \widehat{\boldsymbol{\Gamma}}^{-1}\dot{\boldsymbol{\Gamma}}_{\lambda_t}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}\right) + \dot{\boldsymbol{\Delta}}_{\lambda_t}^{\top}\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}\right]\right\},
\end{aligned}$$

where

$$\dot{\mathbf{\Gamma}}_{\alpha_r} = \frac{\partial \mathbf{\Gamma}}{\partial \alpha_r}\Big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}} = \mathbf{F}\dot{\mathbf{F}}_r + \dot{\mathbf{F}}_r\mathbf{F} - \mathbf{F}\boldsymbol{\delta}\boldsymbol{\delta}^\top\dot{\mathbf{F}}_r - \dot{\mathbf{F}}_r\boldsymbol{\delta}\boldsymbol{\delta}^\top\mathbf{F}, \quad \dot{\mathbf{F}}_r = \frac{\partial \mathbf{F}}{\partial \alpha_r}\Big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}},$$

$$\dot{\mathbf{\Gamma}}_{\lambda_t} = \frac{\partial \mathbf{\Gamma}}{\partial \lambda_t}\Big|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}} = -\mathbf{F}\left(\frac{\dot{\boldsymbol{\lambda}}_{2t}}{1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda}} - \frac{2\lambda_t\boldsymbol{\lambda}\boldsymbol{\lambda}^\top}{(1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda})^2}\right)\mathbf{F},$$

$$\dot{\mathbf{\Delta}}_{\lambda_t} = \frac{\partial \mathbf{\Delta}}{\partial \lambda_t}\Big|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}} = \mathbf{F}\left(\frac{\dot{\boldsymbol{\lambda}}_t}{(1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{1/2}} - \frac{\lambda_t\boldsymbol{\lambda}}{(1+\boldsymbol{\lambda}^\top\boldsymbol{\lambda})^{3/2}}\right),$$

$$\dot{\boldsymbol{\lambda}}_t = \frac{\partial \boldsymbol{\lambda}}{\partial \lambda_t}\Big|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}}, \quad \dot{\boldsymbol{\lambda}}_{2t} = \frac{\partial \boldsymbol{\lambda}\boldsymbol{\lambda}^\top}{\partial \lambda_t}\Big|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}} \text{ and } \dot{\mathbf{E}}_i^s = \frac{\partial \mathbf{E}_i}{\partial \phi_s}\Big|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}},$$

$r = 1, \ldots, q(q+1)/2$, $t = 1, \ldots, q$, and $s = 1, 2$.

For the DEC structure, we have that

$$\frac{\partial \mathbf{E}_i}{\partial \phi_1} = |t_{ij} - t_{ik}|^{\phi_2}\phi_1^{|t_{ij}-t_{ik}|^{\phi_2-1}},$$

$$\frac{\partial \mathbf{E}_i}{\partial \phi_2} = |t_{ij} - t_{ik}|^{\phi_2}\log(|t_{ij} - t_{ik}|)\log(\phi_1)\phi_1^{|t_{ij}-t_{ik}|^{\phi_2}}.$$

Estimation of the Random Effects

This section considers an empirical Bayes inference for the random effects that are useful for interpreting the subject-specific variability. From (3)–(4), it implies that $\mathbf{Y}_i|\mathbf{b}_i \sim \mathrm{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Omega}_i)$ and $\mathbf{b}_i \sim \mathrm{SN}_q(c\boldsymbol{\Delta}, \mathbf{D}, \boldsymbol{\lambda})$. The conditional distribution of $\mathbf{b}_i$ given $\mathbf{Y}_i$ belongs to the extended skew-normal (ESN), and its pdf is

$$f(\mathbf{b}_i|\mathbf{Y}_i) = \frac{f(\mathbf{Y}_i|\mathbf{b}_i)f(\mathbf{b}_i)}{\int f(\mathbf{Y}_i|\mathbf{b}_i)f(\mathbf{b}_i)\mathrm{d}\mathbf{b}_i} = \frac{\phi_q\left(\mathbf{b}_i; c\boldsymbol{\Delta} + \mathbf{D}\mathbf{Z}_i^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta}), \boldsymbol{\Lambda}_i\right)\Phi_1\left(\boldsymbol{\lambda}^\top\mathbf{D}^{-1/2}(\mathbf{b}_i - c\boldsymbol{\Delta})\right)}{\Phi_1\left(\bar{\boldsymbol{\lambda}}_i^\top\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta})\right)},$$

i.e.,

$$\mathbf{b}_i|\mathbf{Y}_i \sim \mathrm{ESN}_q\left(c\boldsymbol{\Delta} + \mathbf{D}\mathbf{Z}_i^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta}), \boldsymbol{\Lambda}_i, \boldsymbol{\Lambda}_i^{1/2}\boldsymbol{\zeta}, \boldsymbol{\zeta}^\top\mathbf{D}\mathbf{Z}_i^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta})\right).$$

Thus, from Lemma 1, it follows that

$$\mathbb{E}\left[\mathbf{b}_i\middle|\mathbf{Y}_i = \mathbf{y}_i, \boldsymbol{\theta}\right] = c\boldsymbol{\Delta} + \mathbf{D}\mathbf{Z}_i^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta}) + \frac{\boldsymbol{\Lambda}_i\boldsymbol{\zeta}}{\sqrt{1+\boldsymbol{\zeta}^\top\boldsymbol{\Lambda}_i\boldsymbol{\zeta}}}\mathrm{W}_\Phi\left(\bar{\boldsymbol{\lambda}}_i\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta})\right).$$

The minimum mean squared error (MSE) estimator of $\mathbf{b}_i$ obtained by the conditional mean of $\mathbf{b}_i$ given $\mathbf{V}_i$ and $\mathbf{C}_i$ is

$$\widehat{\mathbf{b}}_i(\boldsymbol{\theta}) = \mathbb{E}\left[\mathbf{b}_i|\mathbf{V}_i, \mathbf{C}_i\right] = \mathbb{E}\left[\mathbb{E}(\mathbf{b}_i|\mathbf{Y}_i, \boldsymbol{\theta})|\mathbf{V}_i, \mathbf{C}_i\right] \tag{10}$$

$$= c\boldsymbol{\Delta} + \mathbf{D}\mathbf{Z}_i^\top \boldsymbol{\Sigma}_i^{-1}(\widehat{\mathbf{y}}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i c\boldsymbol{\Delta}) + \frac{\boldsymbol{\Lambda}_i\boldsymbol{\zeta}}{\sqrt{1+\boldsymbol{\zeta}^\top\boldsymbol{\Lambda}_i\boldsymbol{\zeta}}}\widehat{\kappa}_i, \tag{11}$$

where $\widehat{\mathbf{y}}_i = \mathbb{E}[\mathbf{Y}_i\big|\mathbf{V}_i, \mathbf{C}_i]$ and $\widehat{\kappa}_i = \mathbb{E}[\mathrm{W}_\Phi(\cdot)\big|\mathbf{V}_i, \mathbf{C}_i]$ depend on the censoring pattern of subject i (see Sect. 3).

The empirical Bayes estimates of random effects are obtained by substituting the ML estimates $\widehat{\boldsymbol{\theta}}$ into $\mathbf{b}_i(\boldsymbol{\theta})$, leading to $\widehat{\mathbf{b}}_i = \mathbf{b}_i(\widehat{\boldsymbol{\theta}})$. In addition, the fitted values of responses can be estimated directly by $\mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i$.

Prediction of Future Observations

The prediction problem for longitudinal data is also of great importance in several practical applications. Rao et al. [27] pointed out that the predictive accuracy of future observations can be taken as an alternative measure of "goodness-of-fit". In order to propose a strategy to generate predicted values from the SN-LMEC model, we use the approach proposed by Wang [32]. Thus, let $\mathbf{y}_{i,\mathrm{obs}}$ be an observed response vector of dimension $n_{i,\mathrm{obs}} \times 1$ for a new subject i over the first portion of time, and $\mathbf{y}_{i,\mathrm{pred}}$ be the corresponding $n_{i,\mathrm{pred}} \times 1$ response vector over the future portion of time. Moreover, let $\mathbf{X}_i^* = (\mathbf{X}_{i,\mathrm{obs}}, \mathbf{X}_{i,\mathrm{pred}})$ and $\mathbf{Z}_i^* = (\mathbf{Z}_{i,\mathrm{obs}}, \mathbf{Z}_{i,\mathrm{pred}})$ denote the $(n_{i,\mathrm{obs}} + n_{i,\mathrm{pred}}) \times p$ and $(n_{i,\mathrm{obs}} + n_{i,\mathrm{pred}}) \times q$ design matrices corresponding to $\bar{\mathbf{y}}_i = (\mathbf{y}_{i,\mathrm{obs}}^\top, \mathbf{y}_{i,\mathrm{pred}}^\top)$.

We use the imputation procedure to deal with the censored values existing in $\mathbf{y}_{i,\mathrm{obs}}$ by replacing the censored values with $\widehat{\mathbf{y}}_i = \mathbb{E}[\mathbf{y}_i|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}]$ obtained from the ECM algorithm. Therefore, a complete dataset, denoted by $\mathbf{y}_{i,\mathrm{obs}^*}$, is obtained when the censored values are imputed. The reason to use the imputation procedure is that it avoids computing truncated conditional expectations of the skew-normal multivariate distribution originated by the censoring scheme. Hence, we have that

$$\bar{\mathbf{y}}_i^* = (\mathbf{y}_{i,\mathrm{obs}^*}^\top, \mathbf{y}_{i,\mathrm{pred}}^\top)^\top \sim \mathrm{SN}_{(n_{i,\mathrm{obs}^*}+n_{i,\mathrm{pred}})}\left(\mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{Z}_i^* c\boldsymbol{\Delta}, \boldsymbol{\Sigma}_i^*, \bar{\boldsymbol{\lambda}}_i^*\right),$$

where $\boldsymbol{\Sigma}_i^* = \begin{pmatrix} \boldsymbol{\Sigma}_i^{\mathrm{obs}^*,\mathrm{obs}^*} & \boldsymbol{\Sigma}_i^{\mathrm{obs}^*,\mathrm{pred}} \\ \boldsymbol{\Sigma}_i^{\mathrm{pred},\mathrm{obs}^*} & \boldsymbol{\Sigma}_i^{\mathrm{pred},\mathrm{pred}} \end{pmatrix}$, $\bar{\boldsymbol{\lambda}}_i^* = \dfrac{\boldsymbol{\Sigma}_i^{*-1/2}\mathbf{Z}_i^*\mathbf{D}\boldsymbol{\zeta}}{\sqrt{1+\boldsymbol{\zeta}^\top\boldsymbol{\Lambda}_i^*\boldsymbol{\zeta}}}$. As mentioned in Wang [32], the best linear predictor of $\mathbf{y}_{i,\mathrm{pred}}$ with respect to the minimum mean squared error (MSE) criterion is the conditional expectation of $\mathbf{y}_{i,\mathrm{pred}}$ given $\mathbf{y}_{i,\mathrm{obs}^*}$, which, from Proposition 1, is given by

$$\widehat{\mathbf{y}}_{i,\text{pred}}(\boldsymbol{\theta}) = \boldsymbol{\mu}^* + \mathrm{W}_{\Phi}\left(\frac{\tau^*}{\sqrt{1+\mathbf{v}_{2i}^{\top}\mathbf{S}_i\mathbf{v}_{2i}}}\right)\frac{\mathbf{S}_i\mathbf{v}_{2i}}{\sqrt{1+\mathbf{v}_{2i}^{\top}\mathbf{S}_i\mathbf{v}_{2i}}}, \tag{12}$$

where $\mathbf{v}_i = (\mathbf{v}_{1i}^{\top}, \mathbf{v}_{2i}^{\top})^{\top} = \boldsymbol{\Sigma}_i^{*-1/2}\bar{\boldsymbol{\lambda}}_i^*$, $\mathbf{S}_i = \boldsymbol{\Sigma}_i^{\text{pred,pred}} - \boldsymbol{\Sigma}_i^{\text{pred,obs}^*}(\boldsymbol{\Sigma}_i^{\text{obs}^*,\text{obs}^*})^{-1}\boldsymbol{\Sigma}_i^{\text{obs}^*,\text{pred}}$, $\boldsymbol{\mu}^* = \mathbf{X}_{i,\text{pred}}\boldsymbol{\beta} + \mathbf{Z}_{i,\text{pred}}c\boldsymbol{\Delta} + \boldsymbol{\Sigma}_i^{\text{pred,obs}^*}(\boldsymbol{\Sigma}_i^{\text{obs}^*,\text{obs}^*})^{-1}(\mathbf{y}_{i,\text{obs}^*} - \mathbf{X}_{i,\text{obs}^*}\boldsymbol{\beta} - \mathbf{Z}_{i,\text{obs}^*}c\boldsymbol{\Delta})$, $\tau^* = \left(\mathbf{v}_{1i} + (\boldsymbol{\Sigma}_i^{\text{obs}^*,\text{obs}^*})^{-1}\boldsymbol{\Sigma}_i^{\text{obs}^*,\text{pred}}\mathbf{v}_{2i}\right)^{\top}\left(\mathbf{y}_{i,\text{obs}^*} - \mathbf{X}_{i,\text{obs}^*}\boldsymbol{\beta} - \mathbf{Z}_{i,\text{obs}^*}c\boldsymbol{\Delta}\right)$.

Therefore, $\mathbf{y}_{i,\text{pred}}$ can be estimated directly by substituting $\widehat{\boldsymbol{\theta}}$ into (12), leading to $\widehat{\mathbf{y}_{i,\text{pred}}} = \widehat{\mathbf{y}}_{i,\text{pred}}(\widehat{\boldsymbol{\theta}})$.

4 Illustrative Example—UTI Data

The UTI data is referred to a study of 72 children and adolescents who had HIV-1 infection and stopped their medications at four academic centers in the United States between January 2000 and September 2004. An unstructured treatment interruption (UTI) is an issue in the adolescent population because the potential alternative of suboptimal adherence can lead to antiretroviral (ARV) resistance and diminished treatment options in the future. More detail about this study can be found in Saitoh et al. [28].

This study aimed to monitor the HIV-1 viral load (RNA) after unstructured treatment interruption. The subjects in the study have taken ARV therapy for at least six months before UTI, and the medication was discontinued for more than three months. The HIV viral loads were studied from the closest time points at 0, 1, 3, 6, 9, 12, 18, 24 months after UTI. The number of observations from baseline (month 0) to month 24 are 71, 62, 58, 57, 43, 34, 24, and 13, respectively. Out of 362 observations, 26(7%) were below the detection limits (50 or 400 copies/mL) and were left-censored at these values.

From a frequentist perspective, this data was previously analyzed by Vaida and Liu [29] and Matos et al. [23] using the LMEC model. In a Bayesian framework, Lachos et al. [19] and Bandyopadhyay et al. [6] analyzed this dataset using normal/independent (NI) distributions and skew-normal/independent (SNI) distributions in the LMEC, respectively. Moreover, this data is available in the `R` package `lmec` [30].

In Fig. 2, we can see the density histogram of random effects obtained after fitting an LMEC considering the normal distribution (NLMEC). The plot reveals the left-skewed nature of subject-specific intercepts at the level of random effects. Therefore, an assumption of symmetric distribution for random effects is not very realistic for the UTI data set.

We revisit the UTI data in order to provide additional inferences for the use of the SN-LMEC. Following Vaida and Liu [29], we consider a profile LME model with random intercepts b_i, given by

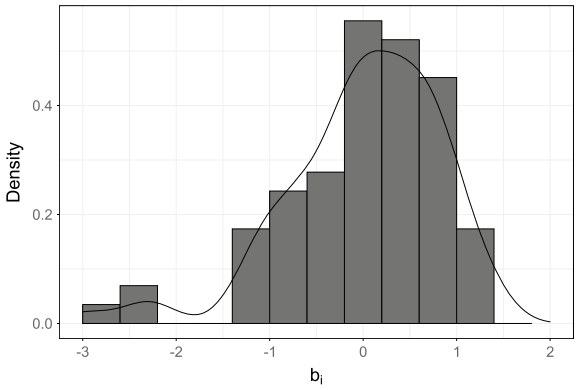


Fig. 2 UTI data. The plot of density histogram of estimated random effects for NLMEC

$$y_{ij} = b_i + \beta_j + \epsilon_{ij}, \tag{13}$$

where y_{ij} is $\log_{10}$(HIV RNA) for subject i at time t_j, $t_1 = 0$, $t_2 = 1$, $t_3 = 3$, $t_4 = 6$, $t_5 = 9$, $t_6 = 12$, $t_7 = 18$, $t_8 = 24$, b_i is the random intercept for the i-th subject, and ϵ_{ij} are random errors. The ML estimates were obtained using the ECM algorithm describes in Sect. 3.

Tables 1 and 2 present the ML estimates and standard errors under the different correlations structures for the SN-LMEC and NLMEC models, respectively. The

Table 1 UTI dataset. Parameter estimates of the SN-LMEC model for UTI dataset under different correlation structures The SE values are estimated as mentioned in Sect. 3

Parameter	AR(1)		CS		DEC		UNC	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	3.6105	0.1183	3.6107	0.1158	3.6063	0.1159	3.6092	0.1106
β_2	4.1739	0.1736	4.1733	0.1613	4.1700	0.1633	4.1725	0.1583
β_3	4.2486	0.1956	4.2479	0.1990	4.2434	0.2011	4.2473	0.1944
β_4	4.3704	0.1828	4.3695	0.1863	4.3635	0.1853	4.3690	0.1829
β_5	4.5756	0.1963	4.5739	0.1997	4.5673	0.2010	4.5743	0.1960
β_6	4.5677	0.2362	4.5666	0.2375	4.5593	0.2412	4.5665	0.2362
β_7	4.6745	0.1999	4.6728	0.2048	4.6642	0.2134	4.6733	0.1995
β_8	4.7850	0.3464	4.7842	0.3494	4.7752	0.3619	4.7838	0.3461
σ^2	0.3414		0.4059		0.392		0.3414	
α	1.9112		1.7273		1.7853		1.9112	
ϕ_1	2e–04		0.1572		0.1541		–	
ϕ_2	–		–		0.0603		–	
λ	–5.3038		–7.877		–7.0462		–5.2288	
Loglik	–406.7471		–406.8604		–406.7113		–406.7474	
AIC	837.4942		837.7209		839.4226		**835.4948**	
BIC	884.1939		884.4206		890.014		**878.3029**	

Table 2 UTI dataset. Parameter estimates of the NLMEC model for UTI dataset under different correlation structures

Parameter	AR(1)		CS		DEC		UNC	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_1	3.6188	0.1253	3.6188	0.1253	3.6194	0.1252	3.6188	0.1253
β_2	4.1815	0.1285	4.1815	0.1285	4.1830	0.1283	4.1815	0.1285
β_3	4.2565	0.1304	4.2565	0.1304	4.2567	0.1303	4.2565	0.1304
β_4	4.3755	0.1307	4.3755	0.1307	4.3740	0.1307	4.3755	0.1307
β_5	4.5815	0.1398	4.5815	0.1398	4.5795	0.1398	4.5815	0.1398
β_6	4.5846	0.1485	4.5846	0.1484	4.5823	0.1487	4.5846	0.1485
β_7	4.6930	0.1646	4.6930	0.1646	4.6892	0.1655	4.6930	0.1646
β_8	4.8092	0.2018	4.8093	0.2017	4.8069	0.2038	4.8093	0.2018
σ^2	0.3414		0.6432		0.4266		0.3414	
α	0.7654		0.4636		0.6786		0.7654	
ϕ_1	2e–04		0.4692		0.225		–	
ϕ_2	–		–		0.0475		–	
Loglik	–412.0448		–412.0421		–411.9436		–412.0421	
AIC	846.0896		846.0842		847.8872		**844.0842**	
BIC	888.8977		888.8923		894.5869		**883.0007**	

values of the log-likelihood function and the AIC and BIC criteria are also presented in these Tables. We can see that the SN-LMEC outperforms the normal consistently in all cases. In particular, the lowest value for both criteria is the one from the SN-LMEC with uncorrelated (UNC) structure, and therefore, this model is selected for further analyses. Considering the SN-LMEC and NLMEC with uncorrelated (UNC) structures, we performed a likelihood ratio test (LRT) to test the hypothesis $H_0 : \boldsymbol{\lambda} = 0$. The LRT statistic was 10.589, resulting in a p-value of 0.0011, i.e., we conclude that the asymmetric model is necessary for modeling the UTI data set.

The mean viral load $\mathbb{E}[y_{ij}] = \beta_j$, for $j = 1, \ldots, 8$, increases gradually throughout 24 months for all models; this is evidence of the negative effect of the antiretroviral therapy interruption on the viral load's levels. For the SN-LMEC under UNC structure, the mean viral load increases from 3.61 at the time of UTI to 4.78 at 24 months. The between-subject variance (α) and within-subject variance (σ^2) estimates are 1.91 and 0.34, respectively. In addition, the estimate of skewness parameter λ is -5.23, indicating left-skewness. Figure 3 (Left panel) shows some individual profiles (in log10 scale) for HIV viral load and estimated trajectories for the SN-LMEC model under UNC structure.

We are also interested in investigating the performance of the prediction for future values described in Sect. 3. We exclude the last two measurements of each individual in the datasets with more than 7 (inclusive) observations (total of 30 individuals), and we compute the predicted values under the SN-LMEC model with UNC correlation structure. Figure 3 (Right panel) shows the comparison between the estimated, the

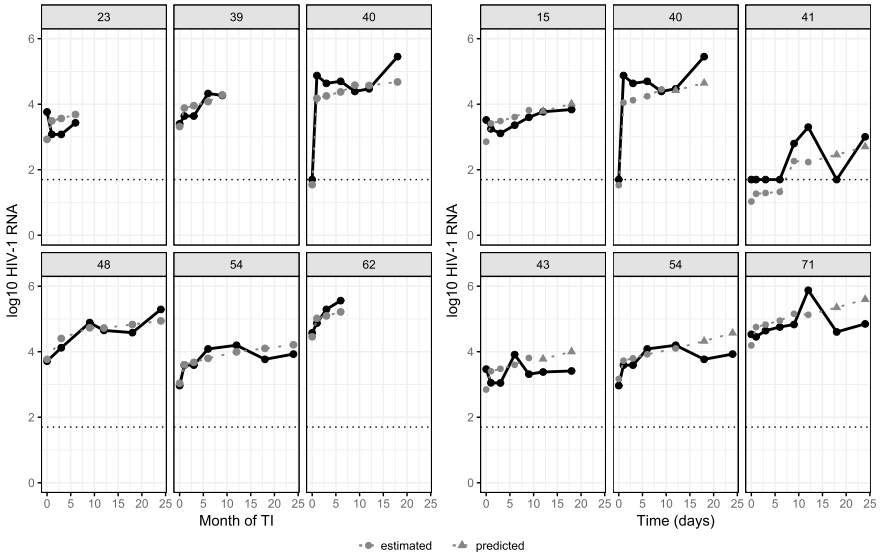


Fig. 3 **UTI data**. (Left panel) Viral loads in $\log_{10}$ scale (black, solid line) for six random subjects and estimated trajectories for the SN-LMEC model under UNC structure. (Right panel) Evaluation of the prediction performance for six random subjects, considering the SN-LMEC model under UNC structure

predicted values, and the real ones, indicating the good performance of the SN-LMEC in terms of prediction.

In conclusion, these results suggest that our proposed model, SN-LMEC, provides precise parameter estimates for our motivating HIV viral load dataset that exhibits departure from the traditional normality assumptions due to skewness.

5 Conclusions

In this work, we have proposed an approach to a linear mixed model with censored responses where the random effects are assumed to have a multivariate skew-normal distribution. We adopted a DEC structure proposed by Munoz et al. [26] to model the autocorrelation existing among irregularly observed measures. The proposed model generalizes previous proposals, such as the SN-LME model proposed by Arellano-Valle et al. [1] (see also, Lin and Lee [21]) and in the context of censored data, the NLMEC model proposed by Vaida and Liu [29] (see also, Matos et al. [24]), which are restricted to left or right censored problems. We developed a computationally tractable ECM algorithm for carrying out ML estimation. The algorithm has a closed-form expression for the E-step, based on formulas for the mean and variance of the truncated extended multivariate skew-normal distribution Galarza Morales et

al. [13]. The computation procedures for estimating random effects and predicting future responses are easy to implement once the ML estimates are obtained. The proposed methods were applied to the AIDS study, providing support for the usefulness and effectiveness of our proposal. All the R code is freely available at the GitHub repository: https://github.com/thalitadobem/snlmec.

Although the SN-LMEC model showed flexibility to model asymmetric data, they can be seriously affected by the presence of outliers. A natural generalization of our method is to extend by considering the skew-*t* distribution [4] or the multivariate skew-elliptical distribution [7]. Another promising avenue for future research is to propose methods that combine both skewness and change-points (see, Huang et al. [15]) within a unified framework in LMEC models, including serial correlation for the random errors.

Acknowledgements The research of Thalita B. Mattos was supported by CAPES. Larissa A. Matos acknowledges support from FAPESP-Brazil (Grant 2020/16713-0).

References

1. Arellano-Valle, R. B., Bolfarine, H., & Lachos, V. H. (2005). Skew-normal linear mixed models. *Journal of Data Science, 3*, 415–438.
2. Arellano-Valle, R. B., Branco, M. D., & Genton, M. G. (2006). A unified view on skewed distributions arising from selections. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique, 34*(4), 581–601. http://www.jstor.org/stable/20445223
3. Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics, 32*(2), 159–188. https://doi.org/10.1111/j.1467-9469.2005.00426.x
4. Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(2), 367–389.
5. Azzalini, A., & Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika, 83*(4), 715–726.
6. Bandyopadhyay, D., Lachos, V. H., Castro, L. M., & Dey, D. K. (2012). Skew-normal/independent linear mixed models for censored responses with applications to HIV viral loads. *Biometrical Journal, 54*(3), 405–425.
7. Branco, M. D., & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis, 79*, 99–113.
8. Castro, L. M., Martín, E. S., & Arellano-Valle, R. B. (2013). A note on the parameterization of multivariate skewed-normal distributions. *Brazilian Journal of Probability and Statistics, 27*(1), 110–115. https://doi.org/10.1214/11-BJPS159
9. Davidian, M., & Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Routledge.
10. Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.
11. Galarza, C. E., Matos, L. A., Lachos, V. H. (2020). Moments of the doubly truncated selection elliptical distributions with emphasis on the unified multivariate skew-*t* distribution. arXiv:2007.14980
12. Galarza, C. E., Kan, R., Lachos, V. H. (2021). MomTrunc: Moments of folded and doubly truncated multivariate distributions. https://CRAN.R-project.org/package=MomTrunc, R package version 5.97

13. Galarza Morales, C. E., Matos, L. A., Dey, D. K., & Lachos, V. H. (2021). On moments of folded and doubly truncated multivariate extended skew-normal distributions. *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2021.2000869(just-accepted)
14. Ho, H. J., & Lin, T. I. (2010). Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Biometrical Journal, 52*(4), 449–469.
15. Huang, Y., Dagne, G. A., Zhou, S., & Wang, Z. (2015). Piecewise mixed-effects models with skew distributions for evaluating viral load changes: a bayesian approach. *Statistical methods in medical research, 24*(6), 730–746.
16. Hughes, J. (1999). Mixed effects models with censored data with application to HIV RNA levels. *Biometrics, 55*, 625–629.
17. Jacqmin-Gadda, H., Thiebaut, R., Chene, G., & Commenges, D. (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics, 1*, 355–368.
18. Lachos, V. H., Ghosh, P., & Arellano-Valle, R. B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica, 20*(1), 303–322.
19. Lachos, V. H., Bandyopadhyay, D., & Dey, D. K. (2011). Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. *Biometrics, 55*, 1304–1318.
20. Laird, N. M., & Ware, J. (1982). Random effects models for longitudinal data. *Biometrics, 38*, 963–974.
21. Lin, T. I., & Lee, J. C. (2008). Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statistics in Medicine, 27*(9), 1490–1507.
22. Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological), 44*(2), 226–233. http://www.jstor.org/stable/2345828
23. Matos, L. A., Prates, M. O., Chen, M. H., & Lachos, V. H. (2013). Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. *Statistica Sinica, 23*, 1323–1345.
24. Matos, L. A., Castro, L. M., & Lachos, V. H. (2016). Censored mixed-effects models for irregularly observed repeated measures with applications to HIV viral loads. *TEST, 25*(4), 627–653. https://doi.org/10.1007/s11749-016-0486-2
25. Meng, X., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 81*, 633–648.
26. Munoz, A., Carey, V., Schouten, J. P., Segal, M., Rosner, B, (1992), A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics, 48*(3), 733–742. http://www.jstor.org/stable/2532340
27. Rao, C. R., et al. (1987). Prediction of future observations in growth curve models. *Statistical Science, 2*(4), 434–447.
28. Saitoh, A., Foca, M., Viani, R. M., Heffernan-Vacca, S., Vaida, F., Lujan-Zilbermann, J., Emmanuel, P. J., Deville, J. G., & Spector, S. A. (2008). Clinical outcomes after an unstructured treatment interruption in children and adolescents with perinatally acquired HIV infection. *Pediatrics, 121*, 513–521.
29. Vaida, F., & Liu, L. (2009). Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics, 18*, 797–817.
30. Vaida, F., Liu, L. (2012). lmec: Linear mixed-effects models with censored responses. https://CRAN.R-project.org/package=lmec, r package version 1.0
31. Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association, 91*(433), 217–221.
32. Wang, W. L. (2013). Multivariate t linear mixed models for irregularly observed multiple repeated measures with missing outcomes. *Biometrical Journal, 55*(4), 554–571.
33. Wu, H., & Ding, A. (1999). Population HIV-1 dynamics in vivo: Applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics, 55*(2), 410–418.

Robust Estimation of Multiple Change Points in Multivariate Processes

Yana Melnykov, Marcus Perry, and Volodymyr Melnykov

Abstract Change point inference is important in various fields of science. Many different procedures have been proposed in the literature but most of them rely on some restrictive assumptions such as the normality of underlying processes or independence of observations. In this paper, a novel likelihood-based technique is proposed for identifying multiple change points in multivariate processes. It provides a way to model various covariance patterns and is robust to skewness observed in data. Through simulation studies, we demonstrate that the proposed procedure is superior over its competitors. The application of the methodology to real-life datasets highlights its usefulness and broad applicability.

1 Introduction

The change point estimation in sequential data has become an important task in many areas of active research. It assumes the existence of at least two different processes observed over some time interval. Since the specific times associated with each process are typically unknown, they have to be estimated along with the processes themselves. The applications of change point estimation procedures can be found in medicine [1], ecology [2], pharmacy [3], engineering [4], finance [5, 6], and many other fields. The problem of process and change point estimation is also known as phase I in statistical process control. Then, phase II would deal with the detection of changes in a process flow based on the already estimated processes.

Researchers have been exploring change point problems for decades but there are still many questions that remain open. One of the earliest papers on the subject was

Y. Melnykov · M. Perry · V. Melnykov (✉)
The University of Alabama, Tuscaloosa AL, 35487, USA
e-mail: vmelnykov@cba.ua.edu

Y. Melnykov
e-mail: ymelnykov@cba.ua.edu

M. Perry
e-mail: mperry@cba.ua.edu

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_3

devoted to the estimation of a change point in means of univariate normal distributions [7]. The problem with a constant mean but possible shift in variance parameters was considered by [8–11]. A generalization of both ideas was considered by [12] who developed a test capable of detecting a change in mean and variance parameters simultaneously.

Attention has been paid to multivariate settings as well. [13] and [14] considered the framework with a single change point in mean vectors of multivariate normal distributions. Soon after that, the estimation of multiple change points in mean vectors was studied by [15] and [16]. In the same setting of multivariate normal distribution, [17] proposed a procedure for estimating a change in covariance matrices under the assumption of a constant mean vector. Recently, [18] developed a test for estimating change points in mean vectors and covariance matrices simultaneously, thus generalizing the above-listed ideas. Other directions of research in the area of change point estimation include inference for the general exponential family [19, 20], nonparametric methods [21] including probabilistic pruning based on various goodness-of-fit measures [22], and some others.

In this paper, we consider the problem of estimating multiple change points in the framework with multivariate processes. The importance of this problem is rather substantial but the number of existing methods is very limited (e.g., see discussion on this topic in [22]). The most traditional approach taken by the majority of researchers assumes the independence of observations over time as well as their multivariate normality. Unfortunately, both assumptions are often inadequate or unrealistic. Among other alternatives, there are two nonparametric procedures employing probabilistic pruning with Energy statistic [23] and Kolmogorov-Smirnov statistic [24] that are available through the R package ECP [22]. It is worth mentioning that this R package is currently the only one that aims at identifying multiple change points in the multivariate setting. The lack of developments in this important area of change point inference motivates our methodology. Our proposed technique is based on a matrix normal distribution. Due to its form, one can model the covariance structure associated not just with variables (given by matrix rows) or time points (provided by matrix columns), but also the overall covariance structure associated with variables and times. This effectively eliminates some of the common restrictive assumptions such as the independence of observations at different time points. To make the proposed procedure more robust to deviations from normality, we propose incorporating one of several available transformations to near-normality. As a result, the proposed procedure gains robustness features while being capable of accommodating various covariance structures in data.

The rest of the paper is organized as follows below. Section 2 presents the proposed methodology. Section 3 investigates the performance of our procedure and three competitors in various settings. Section 4 applies the developed methods to the analysis of real-life data. The paper concludes with a discussion provided in Sect. 5.

2 Methodology

Matrix Normal Distribution

Let $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T$ be a process observed over T time points with each $\boldsymbol{y}_i$ following a p-variate normal distribution. The entire dataset can be conveniently summarized in the matrix form as shown below

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1T} \\ y_{21} & y_{22} & \cdots & y_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pT} \end{pmatrix}. \tag{1}$$

Here, each row represents a particular variable observed over time, while every column stands for a p-variate measurement at a specific time point. The overall variability associated with $\boldsymbol{Y}$ can often be explained by the variation observed in rows and columns. This leads to the idea of modeling the variability corresponding to p variables separately from that associated with T time points.

One distribution that can be effectively applied in the considered framework is a so-called matrix normal one [25] that has the following probability density function (pdf):

$$\phi_{p\times T}(\boldsymbol{Y}; \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = \frac{(2\pi)^{-\frac{pT}{2}}}{|\boldsymbol{\Sigma}|^{\frac{T}{2}}|\boldsymbol{\Psi}|^{\frac{p}{2}}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{M})^{\top}\right\}\right\}, \tag{2}$$

where $\boldsymbol{Y}$ is the $p \times T$ matrix argument defined in (1) and $\boldsymbol{M}$ is a $p \times T$ mean matrix. The $p \times p$ matrix $\boldsymbol{\Sigma}$ and $T \times T$ matrix $\boldsymbol{\Psi}$ are covariance matrices that model variability associated with rows and columns, respectively. Also, tr$\{\cdot\}$ denotes the trace operator. It can be shown that $\mathrm{vec}(\boldsymbol{Y}) \sim \mathcal{N}_{pT}(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma})$, where vec$(\cdot)$ denotes the vectorization operator that stacks matrix columns on top of each other, $\otimes$ is the Kronecker product, and $\mathcal{N}_{pT}$ is the pT-variate normal distribution with mean vector vec$(\boldsymbol{M})$ and covariance matrix $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$. There is a minor non-identifiability issue caused by the properties of the Kronecker product since $a\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi} \otimes a\boldsymbol{\Sigma}$ for any multiplier $a \in I\!R^{+}$. One simple restriction on $\boldsymbol{\Psi}$ or $\boldsymbol{\Sigma}$ can effectively resolve this problem. The main advantage of taking into account the matrix data structure is the ability to reduce the number of parameters to $T(T+1)/2 + p(p+1)/2 - 1$ from $pT(pT+1)/2$ in the case of the most general covariance matrix. Hence, the proposed model effectively addresses a potential overparameterization issue while still allowing non-zero covariances $Cov(y_{jt}, y_{j't'})$ for any variables j and j' at time points t and t'.

As the specific problem considered in our setting deals with vectors observed over time, matrix $\boldsymbol{\Psi}$ can be conveniently parameterized in terms of a desired time series process. In this paper, we illustrate the methodology based on the autoregressive

process of order 1 (AR(1)). Incorporating moving average or higher order autoregressive processes is very similar as it affects just the covariance matrix $\boldsymbol{\Psi}$. In fact, the AR(1) model has been chosen as an illustration simply because it yields the best results for the application considered in Sect. 4. Under AR(1), the covariance matrix $\boldsymbol{\Psi}$ is given by

$$\boldsymbol{\Psi} = \frac{\delta^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{T-1} \\ \phi & 1 & \phi & \dots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \dots & 1 \end{pmatrix},$$

where ϕ is the correlation coefficient and δ^2 is the variance parameter. Then, one convenient constraint to avoid the non-identifiability issue associated with $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$ is to set $\delta^2 = 1 - \phi^2$. This restriction immediately leads to $\boldsymbol{\Psi} \equiv \boldsymbol{R}_\phi$, where $\boldsymbol{R}_\phi$ denotes the corresponding correlation matrix that relies on a single parameter ϕ. It can be shown that

$$|\boldsymbol{\Psi}| \equiv |\boldsymbol{R}_\phi| = (1-\phi^2)^{T-1} \quad \text{and} \quad \boldsymbol{\Psi}^{-1} \equiv \boldsymbol{R}_\phi^{-1} = \frac{1}{1-\phi^2}(\boldsymbol{I}_T - \phi \boldsymbol{J}_1 + \phi^2 \boldsymbol{J}_2), \tag{3}$$

where $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ are $T \times T$ matrices defined as follows below:

$$\boldsymbol{J}_1 = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{J}_2 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

Expressions in (3) are helpful for speedier maximum likelihood estimation as the potentially time consuming inversion of the $T \times T$ covariance matrix $\boldsymbol{\Psi}$ can be completely avoided.

Change Point Estimation

Consider the problem of estimating change points in the given framework. Let $\boldsymbol{\mu}_0$ be the p-variate mean vector associated with the main process. Suppose, there are K alternative processes with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K$. Then, the mean matrix $\boldsymbol{M}$ can be written as $\boldsymbol{M} = \sum_{k=0}^{K} \boldsymbol{\mu}_k \boldsymbol{m}_k^\top$, where $\boldsymbol{m}_k$ ($k = 0, 1, \dots, K$) is the vector of length T consisting of zeros and ones, with ones being located in those positions where the k^{th} process is observed. From the definition, it follows that $\sum_{k=0}^{K} \boldsymbol{m}_k = \boldsymbol{1}_T$, where $\boldsymbol{1}_T$ is the vector of length T with all elements equal to 1. It can be noted that vectors $\boldsymbol{m}_k$ can present various permutations of zeros and ones. However, in the case of K shift change points at times $t_1, t_2, \dots, t_K$, the mean matrix is given by

$$\boldsymbol{M} = \left(\underbrace{\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_0}_{t_1-1}, \underbrace{\boldsymbol{\mu}_1 \ldots, \boldsymbol{\mu}_1}_{t_2-t_1}, \ldots, \underbrace{\boldsymbol{\mu}_{K-1}, \ldots, \boldsymbol{\mu}_{K-1}}_{t_K-t_{K-1}}, \underbrace{\boldsymbol{\mu}_K, \ldots, \boldsymbol{\mu}_K}_{T-t_K+1} \right).$$

Also, $\boldsymbol{m}_k = \left(\underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{t_k-1}, \underbrace{\mathbf{1}, \ldots, \mathbf{1}}_{t_{k+1}-t_k}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{T-t_{k+1}+1} \right)$ with boundary conditions $t_0 = 1$ and $t_{K+1} = T + 1$. As a result of such parameterization, the mean matrix $\boldsymbol{M}$ involves $p(K + 1)$ parameters.

The log-likelihood function corresponding to Eq. (2) has the following form:

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{Y}; \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = &- \frac{pT}{2} \log(2\pi) - \frac{T}{2} \log |\boldsymbol{\Sigma}| - \frac{p}{2} \log |\boldsymbol{\Psi}| \\ &- \frac{1}{2} \mathrm{tr} \left\{ \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{M}) \boldsymbol{\Psi}^{-1} (\boldsymbol{Y} - \boldsymbol{M})^\top \right\}. \end{aligned}$$

Oftentimes, the normality assumption is not adequate and inference based on such a model may be incorrect or misleading. One possible treatment of such a situation is to employ a transformation to near-normality. Incorporating a transformation into the model makes it considerably more robust to possible violations of the normality assumption. Several immediate candidates include the famous power transformation proposed by [26], alternative families of power transformations as in [27], or the exponential transformation proposed by [28]. Let $\mathcal{T}$ be an invertible and differentiable mapping representing the transformation operator such that $\mathcal{T}(y; \lambda)$ is approximately normally distributed upon the appropriate choice of the transformation parameter λ. In the p-variate setting, the traditional assumption is that the coordinatewise transformation leads to the joint near-normality [29–31], i.e., the p-variate transformation is given by $\mathcal{T}(\boldsymbol{y}; \boldsymbol{\lambda}) = \left(\mathcal{T}(y_1; \lambda_1), \mathcal{T}(y_2; \lambda_2), \ldots, \mathcal{T}(y_p; \lambda_p)\right)^\top$, where the transformation parameter vector is given by $\boldsymbol{\lambda} = \left(\lambda_1, \lambda_2, \ldots, \lambda_p\right)^\top$. This idea can be readily generalized to the matrix framework with $\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda})$ representing data transformed to matrix near-normality based on the p-variate vector $\boldsymbol{\lambda}$.

Taking into account the special forms of $\boldsymbol{\Psi}$ and $\boldsymbol{M}$ and implementing the transformation idea, the log-likelihood function can be further written as

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}, \phi, \boldsymbol{\lambda}) = &- \frac{pT}{2} \log(2\pi) - \frac{T}{2} \log |\boldsymbol{\Sigma}| - \frac{p(T-1)}{2} \log(1 - \phi^2) \\ &- \frac{1}{2(1-\phi^2)} \mathrm{tr} \Big\{ \boldsymbol{\Sigma}^{-1} \big(\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda}) - \sum_{k=0}^{K} \boldsymbol{\mu}_k \boldsymbol{m}_k^\top \big) \left(\boldsymbol{I}_T - \phi \boldsymbol{J}_1 + \phi^2 \boldsymbol{J}_2 \right) \\ &\times \big(\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda}) - \sum_{k=0}^{K} \boldsymbol{\mu}_k \boldsymbol{m}_k^\top \big)^\top \Big\} + \log \Big| \frac{\partial \mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda})}{\partial \boldsymbol{Y}} \Big|, \end{aligned} \tag{4}$$

where the term $\log \left| \frac{\partial \mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda})}{\partial \boldsymbol{Y}} \right|$ represents the log of Jacobian associated with the transformation.

Maximum likelihood estimation leads to the following expressions for $\boldsymbol{\mu}_k$:

$$\boldsymbol{\mu}_k = \left(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}) - \sum_{\substack{k'=0 \\ k' \neq k}}^{K} \boldsymbol{\mu}_{k'} \boldsymbol{m}_{k'}^\top\right) \boldsymbol{R}_\phi^{-1} \boldsymbol{m}_k \left(\boldsymbol{m}_k^\top \boldsymbol{R}_\phi^{-1} \boldsymbol{m}_k\right)^{-1},$$

where $\boldsymbol{R}_\phi^{-1}$ is as in (3). Solving a system of $K+1$ equations leads to the expressions for $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$. Maximum likelihood estimation for $\boldsymbol{\Sigma}_k$ yields the following expression:

$$\boldsymbol{\Sigma} = \frac{(\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}) - \sum_{k=0}^{K} \boldsymbol{\mu}_k \boldsymbol{m}_k^\top) \boldsymbol{R}_\phi^{-1} (\mathcal{T}(\boldsymbol{Y};\boldsymbol{\lambda}) - \sum_{k=0}^{K} \boldsymbol{\mu}_k \boldsymbol{m}_k^\top)^\top}{T}.$$

Substituting expressions for $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ and $\boldsymbol{\Sigma}$ into the log-likelihood function (4) makes the log-likelihood a function of the parameters ϕ and $\boldsymbol{\lambda}$. The maximization with respect to these parameters can be done numerically using one of many available optimization algorithms.

For the purpose of illustration, in this paper we focus on the exponential transformation of Manly given by $\mathcal{T}(y;\lambda) = y^{I(\lambda=0)} \left(\exp\{\lambda y - 1\}\lambda^{-1}\right)^{I(\lambda\neq 0)}$, where $I(\cdot)$ is the indicator function. In this setting, the log of Jacobian in (4) is given by $\boldsymbol{\lambda}^\top \boldsymbol{Y} \mathbf{1}_T$, where $\mathbf{1}_T = (1, 1, \ldots, 1)^\top$ with cardinality $|\boldsymbol{\lambda}| = T$.

The problem of change point estimation requires assessing the number of processes. To avoid potential problems with the adjustment for multiple comparisons, simplify calculations, and avoid testing procedures in general, we employ the variant of the Bayesian Information Criterion (BIC) [32] proposed by [33] specifically for the change point framework. BIC is also an appealing option due to its connection to the Bayes factor commonly used in Bayesian inference for comparing competing models.

As a final note in this section, we would like to remark that the proposed procedure focuses on processes with mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$. In real-life applications, it is possible that just some parts of these vectors will be different while the remaining variables exhibit no change point behavior. The task of detecting changes in specific variables is a challenging standalone problem that is beyond the scope of this work. One practical approach can be to search for such variables after detecting differences in mean vectors first. Such a scenario is considered in Sect. 4.

3 Experiments

In this section, we consider simulation studies devoted to the rigorous evaluation of the proposed methodology. We investigate the performance of the change point estimation procedure in two general settings. In both cases, we assume the existence

Table 1 Parameter values used in the simulation study of Sect. 3

j	$\boldsymbol{\mu}_0$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\Sigma}$			$\boldsymbol{\lambda}$	ϕ
1	1	1.2	1.1	0.133	−0.033	0	3	{0.1, 0.5, 0.9}
2	1.2	1.7	1.5	−0.033	0.067	−0.033	2	
3	−2.3	−2.2	−2.0	0	−0.033	0.033	−0.5	

of three processes observed over 100 time points. In the first case, the first process is observed until the change point at $t_1 = 10$, when the second process starts. Then, the second process runs until the next change point at $t_2 = 20$, when the third process starts and runs for the remaining time. In the second setting, the change points are set to be at times $t_1 = 10$ and $t_2 = 50$. The difference between these two settings is that in the first situation, the first two processes are observed for a relatively short period of time, while the third process is observed for much longer. On the contrary, in the second experiment setting, just the first process is observed for a short period of time as opposed to the other two processes. The parameters used in the simulation study are provided in Table 1.

Various levels of correlation and scaling as reflected by parameters ϕ and $\boldsymbol{\Sigma}$, respectively, are studied. In particular, we consider $\phi = 0.1, 0.5, 0.9$ and $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}/2$, $\boldsymbol{\Sigma}/4$. 250 datasets were simulated for each combination of the covariance matrix and correlation parameter in both considered setting, thus, yielding 4,500 simulated datasets in total. The proposed technique assumes that the exact location of change points is known. The quality of the model fit is assessed by means of BIC. It can be noticed that in the search for the optimal model with K change points, $(T-1)!/(T-1-K)!$ alternatives should be considered. As K is usually rather low, the approach is computationally feasible even for moderate T values. In our experiments, each model could be fitted in under one second. In addition, parallel computing can be readily implemented if the number of models becomes restrictively high.

The illustration of some simulated datasets can be found in Fig. 1. Here, plots (a) and (b) show datasets simulated with $\phi = 0.1$ but with different covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}/4$, respectively. Plots (c) and (d) correspond to the same covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}/4$ but with high correlation of $\phi = 0.9$. The four considered datasets represent the first setting with change points at $t_1 = 10$ and $t_2 = 20$. Within each of the four plots, there are three subplots representing the coordinatewise behavior of the processes reflected by means of the black, blue, and red colors. The top subplot corresponds to the first coordinate, the middle stands for the second one, and the bottom plot represents the third coordinate. Horizontal lines show the true back-transformed values of the corresponding coordinates of vectors $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$.

From examining Fig. 1, it is easy to conclude that the task of change point estimation is far from trivial in these cases. Especially in those cases when the variability is higher (left column of plots), we can observe a number of points that can be mistakenly thought of as change points. Thus, it is fully expected that false change points

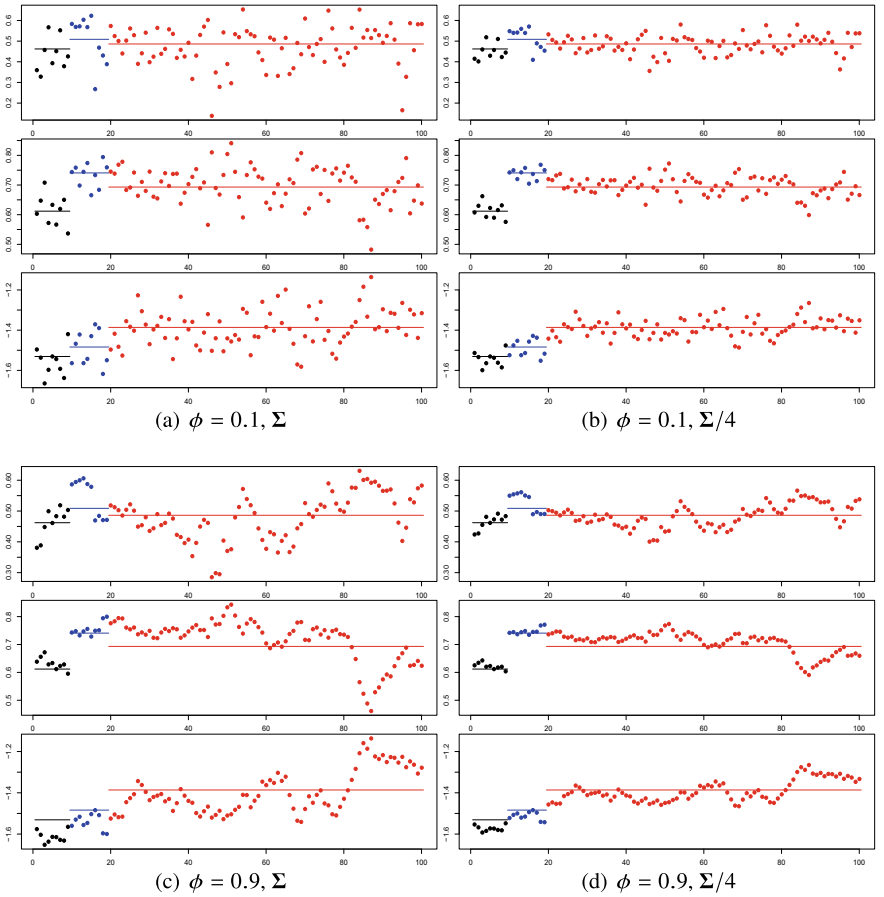


Fig. 1 Datasets generated in the course of the simulation study in Sect. 3 with different scaling (reflected by $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}/4$) and correlation ($\phi = 0.1, 0.9$). Horizontal lines represent true back-transformed values of the corresponding coordinates of parameters $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$

will be found oftentimes. Moreover, we can observe that the first change point should be considerably easier to find than the second one due to the substantial gap in the second coordinate of means related to the first two processes (i.e., between black and blue horizontal lines).

As pointed out by [22], the number of procedures capable of estimating multiple change points in multivariate processes is rather limited. In this section, the developed methodology is compared with one parametric approach that we call naive and two nonparametric procedures available for practitioners through the R package ECP [22]. The naive method is mimicking the most common practical approach with all observations assumed independent and following multivariate normal processes. The two nonparametric procedures are based on probabilistic pruning with

Table 2 Interpretation of notation used in Tables 3 and 4

Notation	Interpretation
$\{t_1, t_2\}$	Both change points are correctly found
$\{t_1, t_2, x\}$	Both change points are correctly identified, but there are false change
	Points found as well
$\{t_1, \tilde{t}_2\}/\{\tilde{t}_1, t_2\}$	One change point is identified correctly, the other one is close by, i.e.
	$\lvert t_k - \hat{t}_k \rvert \leq 3$
$\{t_1\}/\{t_2\}$	one change point is identified correctly and it is the only one found
$\{t_1, !t_2\}/\{!t_1, t_2\}$	One change point is identified correctly, the others are not close, i.e.,
	$\lvert t_k - \hat{t}_k \rvert > 3$

Energy statistic [23, 34] and Kolmogorov-Smirnov statistic [24] used as goodness-of-fit measures. Tables 3 and 4 provide the results of the simulation study in the first ($t_1 = 10$, $t_2 = 20$) and second ($t_1 = 10$, $t_2 = 50$) settings, respectively. The tables include proportions of times various solutions, as per description in Table 2, were found.

As we can observe from Table 3, the proposed method can rather effectively identify change points. Expectedly, the performance of the procedure improves considerably when the variability decreases. For example, in the case with $\phi = 0.9$ and $\mathbf{\Sigma}$, we are able to correctly identify the combination of change points in 14.8% of all cases. The percentage improves to 49.2% and 93.2% for $\mathbf{\Sigma}/2$ and $\mathbf{\Sigma}/4$, respectively. The performance of the procedure somewhat degrades for lower values of parameter ϕ. In particular, the correct setting was found in 63.2% and 55.6% of cases for $\mathbf{\Sigma}/4$ with $\phi = 0.1$ and $\phi = 0.5$, respectively. In the settings with higher variability, the task of estimating both change points correctly is considerably more difficult. It is worth mentioning that in these settings our procedure is capable of identifying at least one change point effectively. In particular, we can notice that there is a relatively low proportion of times when our method identified one point correctly and the other change point estimate was considerably off. Another observation can be made with regard to a low number of false change point detections made by our procedure. In addition, due to a strong penalty carried out by BIC, there is no tendency to overestimate the number of change points as we can see from the line $\{t_1, t_2, x\}$.

From examining Table 3, we can conclude that the closest competitor is the naive procedure. In particular, it demonstrates quite similar results in terms of the proportion of correct solutions for the majority of cases unless $\phi = 0.9$. When ϕ is high, the naive procedure is substantially outperformed by the proposed method in all settings. This observation is not surprising since the cases with lower correlations are more similar to the naive model assuming the independence of observations. Our developed method dramatically outperforms the two nonparametric methods. In the

Table 3 Simulation study from Sect. 3 assuming two change points at times $t_1 = 10$ and $t_2 = 20$. The four methods considered are our proposed procedure, naive procedure, and probabilistic pruning with Energy statistic and Kolmogorov-Smirnov statistic (KS) used as the goodness-of-fit measure. The notation interpretation is provided in Table 2. The bold font highlights the proportion of times the correct combination was found

$K = 2$		$\boldsymbol{\Sigma}$			$\boldsymbol{\Sigma}/2$			$\boldsymbol{\Sigma}/4$		
$t_1 = 10, t_2 = 20$		$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
Method	{10, 20}	**0.060**	**0.032**	**0.148**	**0.332**	**0.168**	**0.492**	**0.632**	**0.556**	**0.932**
	{10, 20, x}	0	0	0	0	0	0	0	0	0
	{10, $\tilde{20}$}/{$\tilde{10}$, 20}	0.200	0.084	0.012	0.336	0.168	0.016	0.304	0.160	0
	{10}/{20}	0.576	0.736	0.692	0.212	0.516	0.424	0.012	0.140	0.040
	{10, !20}/{!10, 20}	0.104	0.112	0.136	0.120	0.148	0.068	0.052	0.144	0.028
Naive	{10, 20}	**0.060**	**0.044**	**0.048**	**0.344**	**0.232**	**0.116**	**0.628**	**0.536**	**0.308**
	{10, 20, x}	0	0	0	0	0	0	0	0	0
	{10, $\tilde{20}$}/{$\tilde{10}$, 20}	0.188	0.192	0.028	0.362	0.224	0.048	0.308	0.252	0.056
	{10}/{20}	0.488	0.108	0	0.136	0.036	0	0.004	0.080	0
	{10, !20}/{!10, 20}	0.212	0.604	0.880	0.152	0.504	0.828	0.060	0.142	0.636
Energy	{10, 20}	**0**	**0**	**0.004**	**0**	**0**	**0.028**	**0.036**	**0.020**	**0.356**
	{10, 20, x}	0	0	0.004	0	0	0.008	0.016	0.008	0.044
	{10, $\tilde{20}$}/{$\tilde{10}$, 20}	0	0	0	0.004	0	0.004	0.028	0.012	0.016
	{10}/{20}	0	0	0	0	0	0.004	0.012	0.004	0.068
	{10, !20}/{!10, 20}	0.024	0.020	0.120	0.080	0.060	0.188	0.192	0.176	0.148
KS	{10, 20}	**0.024**	**0**	**0.004**	**0.020**	**0.016**	**0.012**	**0.044**	**0.028**	**0.024**
	{10, 20, x}	0	0	0	0	0	0	0	0.004	0.004
	{10, $\tilde{20}$}/{$\tilde{10}$, 20}	0.116	0.076	0.044	0.148	0.092	0.052	0.224	0.132	0.076
	{10}/{20}	0.040	0.032	0.020	0.056	0.092	0.040	0.064	0.132	0.060
	{10, !20}/{!10, 20}	0.024	0.016	0.032	0.016	0.024	0.048	0.028	0.020	0.060

easiest case considered with $\phi = 0.9$ and $\boldsymbol{\Sigma}/4$, the probabilistic pruning with Energy statistic is capable of finding the correct combination of change points in 35.6% of cases. In all other cases, both procedures face considerable challenges. One can also notice that nonparametric methods struggle to find even one of the two change points correctly. In the case of $\boldsymbol{\Sigma}/4$, the Kolmogorov-Smirnov statistic (denoted as KS) shows some improvement for $\phi = 0.1$. It is able to estimate one change point correctly and the other one in close proximity to the true change point in 22.4% of all cases.

The inference drawn from Table 4 is mostly similar. In the meantime, we can notice that our method improves the performance in all cases. This happens due to the fact that the number of time points is more evenly distributed among the processes and thus more accurate estimation of parameters is possible. As a result, the difference between the proposed and naive approaches can now be observed for the case with $\boldsymbol{\Sigma}/4$ and $\phi = 0.9$. It is worth mentioning that similar analysis has been repeated for negative parameters $\phi = -0.9, -0.5, -0.1$. The results and findings of

Table 4 Simulation study from Sect. 3 assuming two change points at times $t_1 = 10$ and $t_2 = 50$. The description of the table is similar to that of Table 3

$K = 2$		$\boldsymbol{\Sigma}$			$\boldsymbol{\Sigma}/2$			$\boldsymbol{\Sigma}/4$		
$t_1 = 10, t_2 = 50$		$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 0.1$	$\phi = 0.5$	$\phi = 0.9$
Method	$\{10, 50\}$	**0.232**	**0.116**	**0.216**	**0.384**	**0.324**	**0.576**	**0.632**	**0.624**	**0.948**
	$\{10, 50, x\}$	0	0	0	0	0	0	0	0	0
	$\{10, \tilde{50}\}/\{\tilde{10}, 50\}$	0.368	0.156	0.008	0.460	0.248	0.008	0.336	0.220	0
	$\{10\}/\{50\}$	0.068	0.376	0.600	0	0.096	0.316	0	0.004	0.044
	$\{10, !50\}/\{!10, 50\}$	0.276	0.332	0.168	0.156	0.332	0.100	0.032	0.152	0.008
Naive	$\{10, 50\}$	**0.228**	**0.152**	**0.100**	**0.404**	**0.320**	**0.240**	**0.632**	**0.556**	**0.520**
	$\{10, 50, x\}$	0	0	0	0	0	0	0	0	0
	$\{10, \tilde{50}\}/\{\tilde{10}, 50\}$	0.372	0.256	0.132	0.432	0.284	0.168	0.336	0.284	0.128
	$\{10\}/\{50\}$	0.036	0.008	0	0	0	0	0	0	0
	$\{10, !50\}/\{!10, 50\}$	0.288	0.548	0.696	0.152	0.388	0.588	0.032	0.160	0.352
Energy	$\{10, 50\}$	**0**	**0**	**0**	**0**	**0**	**0.008**	**0.008**	**0.004**	**0.128**
	$\{10, 50, x\}$	0	0	0	0.004	0.004	0	0.008	0	0.064
	$\{10, \tilde{50}\}/\{\tilde{10}, 50\}$	0	0	0.004	0	0	0.012	0	0	0.012
	$\{10\}/\{50\}$	0.068	0.036	0.156	0.128	0.116	0.412	0.296	0.284	0.580
	$\{10, !50\}/\{!10, 50\}$	0.012	0.024	0.076	0.052	0.052	0.088	0.084	0.088	0.152
KS	$\{10, 50\}$	**0**	**0**	**0.004**	**0**	**0.004**	**0.004**	**0.008**	**0.004**	**0**
	$\{10, 50, x\}$	0	0	0	0	0	0	0	0	0
	$\{10, \tilde{50}\}/\{\tilde{10}, 50\}$	0	0	0.004	0.004	0.008	0.008	0.012	0.004	0.016
	$\{10\}/\{50\}$	0.036	0.028	0.016	0.056	0.024	0.044	0.060	0.068	0.076
	$\{10, !50\}/\{!10, 50\}$	0.104	0.056	0.064	0.112	0.096	0.096	0.176	0.104	0.136

these experiments were similar and consistent with those presented in this section. To conclude this section, we can remark that the proposed procedure proves to be a powerful tool for identifying change points.

4 Applications

Illustration on Crime Rates in US Cities

First, we apply the proposed methodology to the US cities crime data obtained from the US Department of Justice, Federal Bureau of Investigation Website (http://www.ucrdatatool.gov/Search/Crime/Crime.cfm). There are seven crime types grouped into two general categories: violent and property crimes. The former includes *Murder*, *Rape*, *Robbery*, and *Aggravated Assault*. The property crimes are *Burglary*, *Larceny Theft*, and *Motor Vehicle Theft*. We focus on crime rates observed between 2000 and 2012. As an example, we choose the data reported by Austin and Cincinnati Police Departments. Figure 2 illustrates violent (left column) and property (right

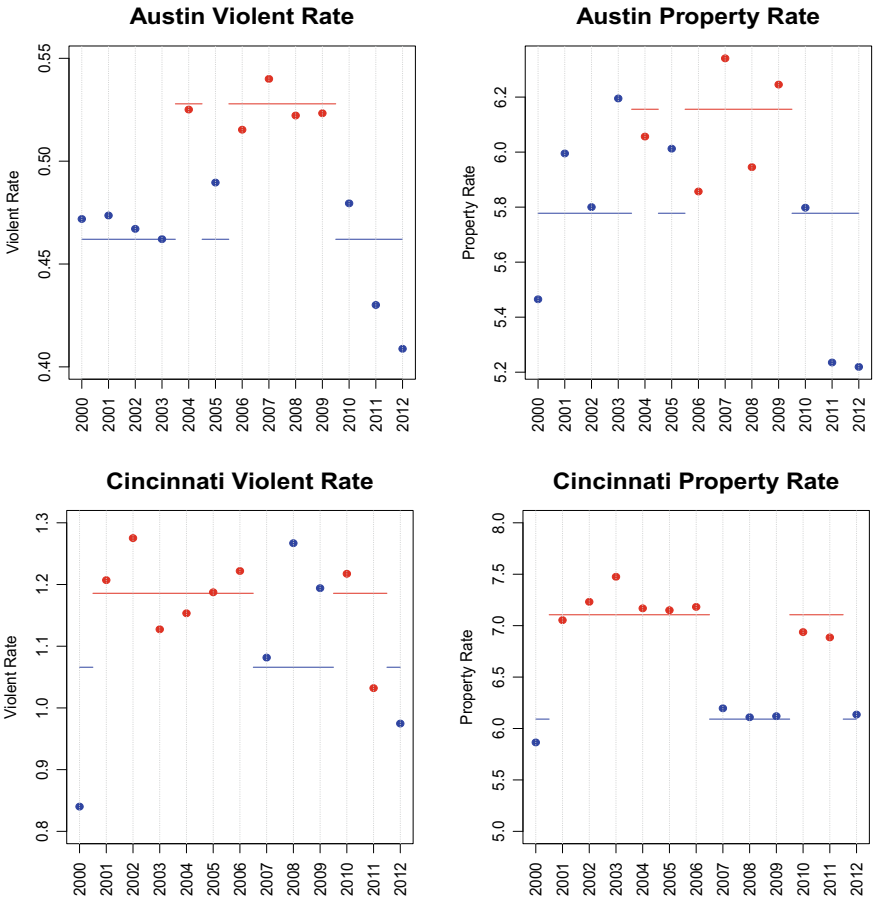


Fig. 2 Violent and Property crime rates in Austin and Cincinnati over the 13-year time period (2000-2012). The blue and red colors represent two processes detected. Horizontal lines stand for the means of the processes

column) crime rates. As the value $T = 13$ is quite low, instead of assuming models with shift-related change points only, we consider all possible orderings of processes.

In the case of Austin, the BIC value associated with a single process (i.e., no change points) is equal to -9.933. After running the developed procedure over all possible orderings of processes, the lowest BIC of -47.081 was found. It is worth mentioning that the naive procedure outlined in Sect. 3 yields BIC -45.225 and the model with the AR(1) structure of $\boldsymbol{\Psi}$ but no transformation parameters produces BIC -44.099. This suggests that even for so few data points as in the considered application, the proposed procedure can be useful. The parameter estimates associated with the model can be found in Table 5. A corresponding illustration is provided in the first row of plots in Fig. 2. Here, the years 2004, 2006, 2007, 2008, and 2009 are associated with

Table 5 Parameter estimates, log-likelihood, and BIC values for Austin and Cincinnati

City	$\hat{\boldsymbol{\mu}}_0$	$\hat{\boldsymbol{\mu}}_1$	$\hat{\boldsymbol{\Sigma}}$		$\hat{\boldsymbol{\lambda}}$	$\hat{\phi}$	$\log \mathcal{L}$	BIC
Austin	168.234	524.023	4422.5	105136.9	17.258	−0.402	39.831	−47.081
	4941.351	8870.934	105136.9	5810, 522	1.548			
Cincinnati	4.130	5.478	1.372	0.004	2.148	0.315	19.693	−6.804
	2.394	2.480	0.004	0.0001	−0.375			

the second process (provided in the red color), while the rest of the years represent the first process (given in the blue color). The horizontal lines reflect back-transformed parameters $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ detected by our methodology. As we can clearly see, the separation into two processes is strongly driven by the variable *Violent Crime*. In the meantime, the variable *Property Crime* demonstrates considerable variability associated with both processes.

The opposite situation is observed for Cincinnati (second row in Fig. 2): the variable *Property Crime* contributes to the separation of the processes more than *Violent Crime*. Model parameters are also provided in Table 5. The BIC value of the best model detected is equal to −6.804 which is considerably better than that of the model with a single process, 19.568. The years 2000, 2007, 2008, 2009, and 2012 are associated with the first process (presented in the blue color), while the rest of the years represent the other process (given in the red color). The BIC value associated with the naive approach is equal to −10.846 suggesting that AR(1) structure of $\boldsymbol{\Psi}$ as well as transformation-related parameters do not bring an improvement to the naive model in this case.

Effect of Colorado Amendment 64

In this section, we demonstrate how our proposed methodology can be applied to the analysis of the effects of public policies. As an example, we focus on studying the effects of the Colorado Amendment 64 which makes the private consumption, production, and possession of marijuana legal. Amendment 64 has been added to the constitution of Colorado in December 2012 but the stores officially opened in January 2014.

The crime rate data have been obtained from the Colorado Bureau of Investigation Department of Public Safety Website (https://www.colorado.gov/pacific/cbi/crime-colorado1) for 10 years: from 2007 to 2016. The same seven variables as described in Sect. 4 have been explored without combining them into the two categories. The goal of our analysis was to check whether the last three years, when the use of

marijuana was legal, were any different from the previous seven years. The value of BIC corresponding to the model with no change points is equal to −996.2, while that related to the model with the change point in 2014 yields BIC equal to −1,006.1. The likelihood ratio test conducted to verify the significance of the change yields P-value 1.47×10^{-6}. As we can see, there is very strong evidence in favor of the change point model based on both BIC and likelihood ratio test.

Figure 3 illustrates the obtained results. The first column consisting of four plots represents violent crimes, while the second column with three plots shows property crimes. The description of individual plots is similar to that of Fig. 2. As we can see, some variables such as *Rape* or *Burglary* seem to contribute substantially to the difference between the two models analyzed. To formalize the analysis, we employed a variable selection procedure. As the number of variables in our experiment is relatively low, we decided to test the model with no change point against the model with the change point at 2014 over all possible combinations of involved variables. The lowest P-value of 1.36×10^{-6} was observed for the combination of variables *Murder*, *Rape*, and *Burglary*. Thus, the most dramatic change in 2014 has been observed for these three variables considered jointly. The corresponding P-value is just marginally lower than the P-value observed for the full model when all seven variables are included, but it gives a good idea about the combination of variables that contribute the most to separating the processes. By examining the contributions of the three variables, we can notice that the crime rate of *Burglary* dropped considerably, while *Rape* and to some extent *Murder* are grown in the last 3 years. Indeed, the proposed analysis does not assume any cause-and-effect conclusions. In fact, we can notice a considerable decrease in *Murder* rates in 2014 and we can also observe that the increase in *Rape* rates began in 2013, i.e., 1 year earlier than when Amendment 64 became effective. Nevertheless, it is obvious that the proposed methodology presents a powerful exploratory tool for studying the effects of public policies.

5 Discussion

In this paper, we developed an efficient method capable of estimating multiple change points in multivariate processes. The proposed technique relies on the matrix normal distribution adjusted by the exponential Manly transformation. Such an adjustment makes the proposed methodology robust to violations of the normality assumption. The matrix setting has an appealing form as rows can represent variables and columns can be associated with time points. Based on the results of challenging simulation studies, we can conclude that the proposed technique is very promising. It outperforms the two nonparametric competitors in all settings. Two applications to crime data considered in the paper demonstrate the usefulness of the developed method.

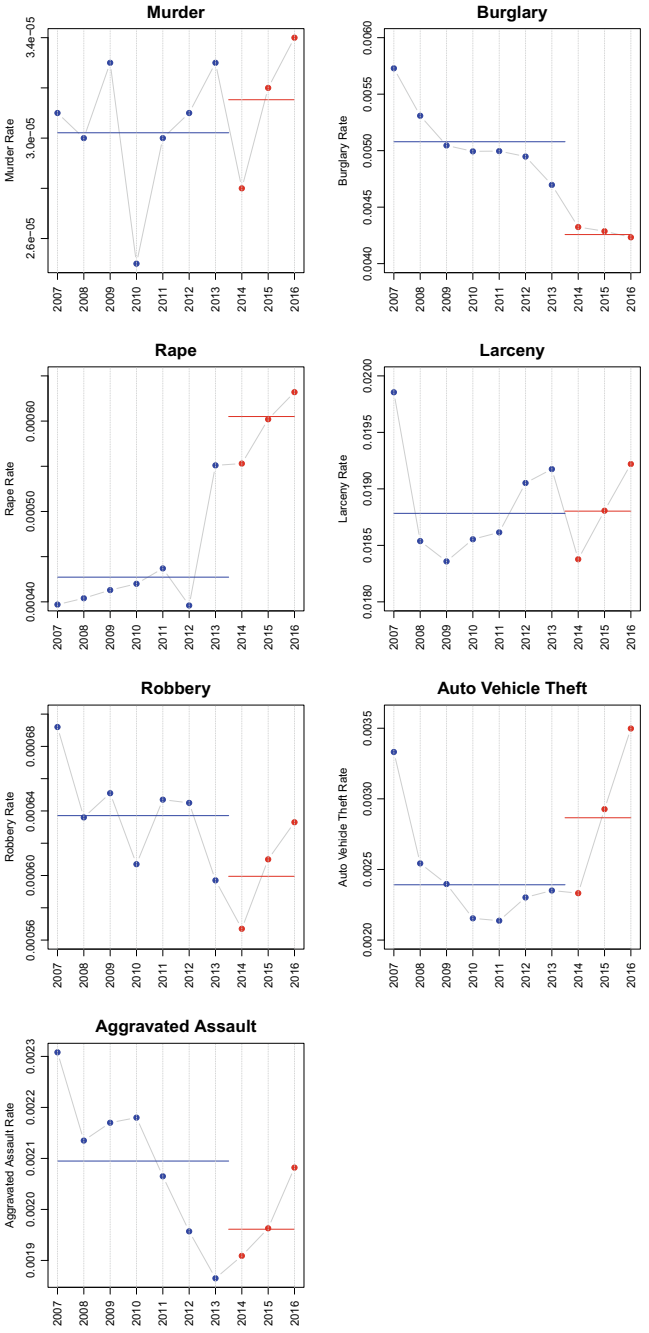


Fig. 3 Crime rates in Colorado over the 10-year time period. The blue and red colors represent two processes. Horizontal lines stand for the back-transformed means of the processes

References

1. Kass-Hout, A. T., Xu, Z., McMurray, P., Park, S., Buckeridge, D., Brownstein, J. S., et al. (2012). Application of change point analysis to daily influenza-like illness emergency department visits. *Journal of the American Medical Informatics Association, 19,* 1075–1081.
2. Patel, S. H., Morreale, S. J., Panagopoulou, A. P., Bailey, H., Robinson, N. J., Paladino, F. V., et al. (2015). Changepoint analysis: A new approach for revealing animal movements and behaviors from satellite telemetry data. *Ecosphere, 6,* 1–13.
3. Baddour, Y., Tholmer, R., & Gavit, P. (2009). Use of change-point analysis for process monitoring and control. *BioPharm international* (Vol. 22).
4. Nigro, M. B., Pakzad, S. N., & Dorvash, S. (2014). Localized structural damage detection: A change point analysis. *Computer-Aided civil and infrastructure engineering, 29,* 416–432.
5. Lenardon, M. J., & Amirdjanove, A. (2006). Interaction between stock indices via changepoint analysis. *Applied Stochastic Models in Business and Industry, 22,* 573–586.
6. Pepelyshev, A., & Polunchenko, A. S. (2015). Real-time financial surveillance via quickest change-point detection methods. *Statistics and its interface* (Vol. 0, pp. 1–14).
7. Page, E. S. (1957). On problem in which a change in parameter occurs at an unknown points. *Biometrika, 42,* 248–252.
8. Hsu, D. A. (1977). Tests for variance shifts at an unknown time point. *Applied Statistics, 26,* 279–284.
9. Davis, W. W. (1979). Robust methods for detection of shifts of the innovation variance of a time series. *Technometrics, 21,* 313–320.
10. Inclán, C. (1993). Detection of multiple changes of variance using posterior odds. *Journal of Business and Economics Statistics, 11,* 189–300.
11. Chen, J., & Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *Jornal of the American Statistical Association, 92,* 739–747.
12. Horváth, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of Statistics, 21,* 671–680.
13. Sen, A. K., & Srivastava, M. S. (1973). On multivariate tests for detecting change in mean. *Sankhyá, A35,* 173–186.
14. Srivastava, M. S., & Worsley, K. J. (1986). Likelihood ratio tests for a chance in the multivariate normal mean. *Journal of the American Statistical Association, 81,* 199–204.
15. Zhao, L. C., Krishnaiah, P. R., & Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis, 20,* 1–25.
16. Zhao, L. C., Krishnaiah, P. R., & Bai, Z. D. (1986). On detection of the number of signals when the noise covariance matrix is arbitrary. *Journal of Multivariate Analysis, 20,* 26–49.
17. Chen, J., & Gupta, A. K. (2004). Statistical inference of covariance change points in Gaussian model. *Journal of Theoretical and Applied Statistics, 38,* 17–28.
18. Chen, J., & Gupta, A. K. (2011). *Parametric statistical change point analysis*, 2nd ed. Springer.
19. Perry, M. B., & Pignatiello, J. J. (2008). A change point model for the location parameter of exponential family densities. *IIE Transactions, 40,* 947–956.
20. Nyambura, S., Mundai, S., & Waititu, A. (2016). Estimation of change point in Poisson random variable using the maximum likelihood method. *American Journal of Theoretical and Applied Statistics, 5,* 219–224.
21. Pettitt, A. N. (1979). A non-parametric approach to the change point problem. *Journal of the American Statistical Association, 28,* 126–135.
22. James, N. A., & Matteson, D. S. (2014). ECP: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software, 62,* 1–25.
23. Rizzo, M., & Szekely, G. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification, 22,* 151–183.
24. Kifer, D., Ben-David, S., & Gehrke, J. (2004). Detecting change in data streams. *International Conference on Very Large Data Bases, 30,* 180–191.
25. Krzanowski, W. J., & Marriott, F. H. C. (1994). *Multivariate analysis, part 1: Distributions, ordination and inference*. Wiley.

26. Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26*(2), 211–252.
27. Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika, 87,* 954–959.
28. Manly, B. F. J. (1976). Exponential data transformations. *Journal of the Royal Statistical Society, Series D, 25*(1), 37–42.
29. Andrews, D. F., Gnanadesikan, R., & Warner, J. L. (1971). Transformations of multivariate data. *Biometrics, 27*(4), 825–840.
30. Lindsey, C., & Sheather, S. (2010). Power transformation via multivariate Box-Cox. *The Stata Journal, 10*(1), 69–81.
31. Zhu, X., & Melnykov, V. (2018). Manly transformation in finite mixture modeling. *Computational Statistics and Data Analysis, 121,* 190–208.
32. Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6*(2), 461–464.
33. Shen, G., & Ghosh, J. (2011). Developing a new BIC for detecting change-points. *Journal of Statistical Planning & Inference, 141,* 1436–1447.
34. Rizzo, M., & Szekely, G. (2010). Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics, 4,* 1034–1055.

Some Computational Aspects of a Noncentral Dirichlet Family

Tanita Botha, Johannes T. Ferreira, and Andriette Bekker

Abstract The Dirichlet distribution is arguably the most well-known multivariate distribution for implementation on the unitary simplex. Different generalizations exist, one of which is the noncentral counterpart. The noncentral representation depends on the noncentrality parameters through the confluent hypergeometric function of several variables and admits both singly and doubly noncentral representations. This chapter explores the computational aspects when the estimation of this singly and doubly noncentral Dirichlet is of interest. It investigates to what degree the additional parameter(s) and their effect on the doubly noncentral Dirichlet, compared to the singly alternative, affects the practical implementation of the model. Real data examples are used for this investigation by using maximum likelihood estimation for the parameters and further strengthened by simulation studies.

Keywords Singly · Doubly · Likelihood Ratio · Household expenditure · Pekin duckling

1 Introduction

The Dirichlet distribution is well known when working with data on the unitary simplex (0, 1) and is a multivariate generalization of the beta distribution. This distribution has many applications in fields such as biological, financial, and cyber

T. Botha (✉) · J. T. Ferreira · A. Bekker
Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa
e-mail: tanita.botha@up.ac.za

J. T. Ferreira
e-mail: johan.ferreira@up.ac.za

A. Bekker
e-mail: andriette.bekker@up.ac.za

J. T. Ferreira · A. Bekker
Centre of Excellence in Mathematical and Statistical Sciences, Johannesburg, South Africa

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_4

analytics where compositional data and measurements apply and is also a conjugate prior for the multinomial distribution in Bayesian statistics [5].

This distribution is known for its mathematical properties and ease of parameter interpretation [13] but is poorly parameterized and cannot model many different types of dependence patterns [2]. Several generalizations of the Dirichlet distribution have been developed, one of which is the noncentral Dirichlet distribution. A recent study from [16] illustrated how the noncentral Dirichlet has seen more applications during the last few years, including [5, 18].

Reference [5] used a noncentral Dirichlet distribution as a previously unconsidered prior when estimating Tsallis entropy in a Bayesian framework for a multinomial likelihood. Here the noncentral Dirichlet distribution was constructed via the use of Poisson mass function weights, inspired by [7]. They showed how this construction of the pdf (probability density function) isolated the noncentrality parameters by retaining them as parameters in independent Poisson mass functions. Reference [4] extended this work by introducing new bivariate noncentral gamma distributions emanating from a scale mixture of the normal class by also showcasing the noncentrality as infinite Poisson mass function mixtures. Reference [5] focussed on the multivariate analogue of the singly and doubly noncentral beta distribution of [14], but the investigation did not focus on the parameter estimation as such. For the interest of the reader, [18] discussed a new noncentral generalization of the beta distribution, the doubly noncentral beta (DNCB) distribution, which is presented as a new non-negative matrix factorization model for (0, 1) bounded-support data.

This chapter explores inferential aspects between the singly and doubly noncentral Dirichlet distributions and investigates whether the additional parameter, which distinguishes the singly from the doubly, has a significant effect when fitting a noncentral Dirichlet distribution to data. These aspects include (1) parameter estimation, (2) goodness-of-fit measures (using the Akaike information criteria (AIC) and the Bayesian information criteria (BIC)), (3) the Likelihood ratio (LR) test statistic, and (4) runtime. Since the noncentral distribution exhibits an infinite sum construction, an additional truncation investigation is done to show how the estimated results converge once the truncation of infinite summations become sufficiently large, thus eliminating the computational challenge of implementing infinite sums.

Reference [3] has shown that flexible distributions are powerful models for data fitting but there are challenges when estimating the parameters. This chapter considers maximum likelihood estimation (MLE) and a two step estimation adaption of the MLE. In this context, the aim of the chapter is to deal with some computational aspects of MLE, particularly in a noncentral Dirichlet family with a specific focus on the doubly and singly constructions. This multivariate investigation, emanating from a data-driven approach, is the pathway to understand and investigate this conventionally computationally challenging multivariate distribution.

The chapter is outlined as follows. Section 2 discusses the foundation of the Dirichlet family. In Section 3, the methodology used within the chapter is discussed. Section 4 reports the results of two data examples as well as two simulation studies. Section 5 contains the discussion and final thoughts.

2 Foundations of the Dirichlet

This section is intended to give a brief overview of the components which will be used in the rest of the chapter. The Pochhammer coefficient is defined as

$$(s)_k = \frac{\Gamma(s+k)}{\Gamma(s)} = s(s+1)(s+2)\cdots(s+k-1); s > 0. \tag{1}$$

The generalized hypergeometric function with r upper and q lower parameters can be defined as (see [11])

$${}_rF_q(a_1,\ldots,a_r;b_1,\ldots,b_q;x) = \sum_{i=1}^{\infty} \frac{(a_1)_i\cdots(a_r)_i}{(b_1)_i\cdots(b_q)_i}\frac{x^i}{i!}, x \in \mathbb{R}. \tag{2}$$

The Dirichlet distribution can be derived by considering ratios of functions of independent gamma random variables. A random variable X has a gamma distribution with parameter ($\pi > 0$) and ($\theta > 0$), denoted by $X \sim Ga(\pi,\theta)$, if the pdf is given by

$$f(x) = \frac{\exp\left(-\frac{x}{\theta}\right)x^{\pi-1}}{\theta^{\pi}\Gamma(\pi)}; x > 0 \tag{3}$$

where $\Gamma(\cdot)$ denotes the usual gamma function. Let $X_1,\ldots,X_{K+1}$ be independent random variables, $X_i \sim Ga(\pi_i,\theta)$, $i = 1,\ldots,K+1$, and let

$$p_i = \frac{X_i}{\sum_{i=1}^{K} X_i + X_{K+1}}, i = 1,\ldots,K \text{ and } 0 < p_i < 1. \tag{4}$$

The joint pdf of $\mathbf{p} = (p_1,\ldots,p_K)$ (4) will then have a Dirichlet distribution (of type 1, see [9, 17]) of order $K \geq 2$ and parameters $\Pi = (\pi_1,\pi_2,\ldots,\pi_{K+1})$ for $\pi_i > 0, i = 1,\ldots,k+1$, with pdf with respect to the Lebesgue measure on the Euclidean space $\mathbb{R}^K$ given by

$$h(\mathbf{p};\Pi) = \frac{\Gamma(\pi^+)}{\prod_{i=1}^{k+1}\Gamma(\pi_i)}\prod_{i=1}^{K+1} p_i^{\pi_i-1} \tag{5}$$

where $\pi^+ = \sum_{i=1}^{K+1}\pi_i$, on the K dimensional simplex defined by

$$\begin{aligned} p_1, p_2, \ldots, p_K &> 0 \\ p_1 + p_2 + \cdots + p_K &< 1 \\ p_{K+1} = 1 - p_1 - \cdots - p_K&, \end{aligned}$$

and 0 otherwise.

The *noncentral* Dirichlet distribution can be obtained by using noncentral gamma variables. A random variable X has a noncentral gamma distribution with parameters $(\pi > 0)$, $(\theta > 0)$ and $(\lambda \geq 0)$, denoted by $X \sim Ga(\pi, \theta, \lambda)$ and the pdf is given by [8]

$$f(x) = \sum_{j=0}^{\infty} \frac{\exp\left(\frac{-\lambda}{2}\right)\left(\frac{\lambda}{2}\right)^j}{j!} \frac{\exp\left(\frac{-x}{\theta}\right) x^{\pi+j-1}}{\theta^{\pi+j}\Gamma(\pi+j)}; x > 0. \tag{6}$$

If (6), instead of (3) is considered for the construction in (4) then the *doubly noncentral Dirichlet distribution* can be represented as

$$\begin{aligned} &h(\mathbf{p};\Pi,\Lambda) \\ &= \sum_{j_1=0}^{\infty} \cdots \sum_{j_{K+1}=0}^{\infty} \frac{\exp\left(\frac{-\lambda_1}{2}\right)\left(\frac{\lambda_1}{2}\right)^{j_1}}{j_1!} \cdots \frac{\exp\left(\frac{-\lambda_K}{2}\right)\left(\frac{\lambda_K}{2}\right)^{j_K}}{j_K!} \frac{\exp\left(\frac{-\lambda_{K+1}}{2}\right)\left(\frac{\lambda_{K+1}}{2}\right)^{j_{K+1}}}{j_{K+1}!} \\ &\times \frac{\Gamma(\pi_1+j_1+\cdots+\pi_K+j_K+\pi_{K+1}+j_{K+1})}{\Gamma(\pi_1+j_1)\Gamma(\pi_K+j_K)\Gamma(\pi_{K+1}+j_{K+1})} p_1^{\pi_1+j_1-1} \cdots p_K^{\pi_K+j_K-1} \left(1-\sum_{i=1}^{K} p_i\right)^{\pi_{K+1}+j_{K+1}-1} \end{aligned} \tag{7}$$

where Λ denotes the vector of noncentral parameters $(\lambda_1, \ldots, \lambda_K, \lambda_{K+1})$ with $\lambda_i > 0 \, \forall \, i$.

The pdf in Eq. (7) reflects a parametrization of the noncentral Dirichlet distribution of [17] and can be represented via the confluent hypergeometric function of several variables where $h(\mathbf{p};\Pi)$ denotes the (unconditional) Dirichlet distribution (see (5)) with parameter Π:

$$\begin{aligned} &h(\mathbf{p};\Pi,\Lambda) \\ &= h(\mathbf{p};\Pi)\exp\left(-\sum_{i=1}^{K+1}\frac{\lambda_i}{2}\right) \\ &\times \sum_{\phi} \frac{(\pi^+)_{j_1+\cdots+j_{K+1}}}{(\pi_1)_{j_1}\cdots(\pi_{K+1})_{j_{K+1}} j_1!\cdots j_{K+1}!} \left(\frac{\lambda_1}{2}p_1\right)^{j_1} \cdots \left(\frac{\lambda_K}{2}p_K\right)^{j_K} \left(\frac{\lambda_{K+1}}{2}\left(1-\sum_{i=1}^{K}p_i\right)\right)^{j_{K+1}} \\ &= h(\mathbf{p};\Pi)\exp\left(-\sum_{i=1}^{K+1}\frac{\lambda_i}{2}\right)\Psi_2^{(K+1)}\left(\pi^+;\pi_1,\ldots,\pi_{K+1};\frac{\lambda_1}{2}p_1,\ldots,\frac{\lambda_K}{2}p_K,\frac{\lambda_{K+1}}{2}\left(1-\sum_{i=1}^{K}p_i\right)\right) \end{aligned} \tag{8}$$

and where $\sum_{\phi} = \sum_{j_1=0}^{\infty} \cdots \sum_{j_{K+1}=0}^{\infty}$ and $\boldsymbol{\Psi}_2^{(K+1)}(\cdot)$ is the confluent hypergeometric function of several variables.

The distribution in (7) and (8) is thus the multivariate analogue of the *doubly* noncentral beta distribution (see [14]). This representation paves the way for the multivariate analogue of the *singly* noncentral beta distribution of [14] where $\boldsymbol{\Psi}_2^{(K)}(\sum_{i=1}^{K}\pi_i;\pi_1,\ldots,\pi_K;\frac{\lambda_1}{2}p_1,\ldots,\frac{\lambda_K}{2}p_K)$ (i.e., k variate confluent hypergeometric function of several variables instead of $K+1$). The pdf is given by

$$h(\mathbf{p};\Pi,\Lambda)$$
$$= h(\mathbf{p};\Pi)\exp\left(-\sum_{i=1}^{K}\frac{\lambda_i}{2}\right)\Psi_2^{(K)}\left(\sum_{i=1}^{K}\pi_i;\pi_1,\ldots,\pi_K;\frac{\lambda_1}{2}p_1,\ldots,\frac{\lambda_K}{2}p_K\right) \quad (9)$$

where

$$\Psi_2^{(K)}\left(\sum_{i=1}^{K}\pi_i;\pi_1,\ldots,\pi_K;\frac{\lambda_1}{2}p_1,\ldots,\frac{\lambda_K}{2}p_K\right)$$
$$=\sum_{\phi}\frac{(\pi_1+\cdots+\pi_{K+1})_{j_1+\cdots+j_K}}{(\pi_1)_{j_1}\ldots(\pi_K)_{j_K}j_1!\ldots j_K!}\left(\frac{\lambda_1}{2}p_1\right)^{j_1}\cdots\left(\frac{\lambda_K}{2}p_K\right)^{j_K}$$

and $\sum_{\phi}=\sum_{j_1=0}^{\infty}\cdots\sum_{j_K=0}^{\infty}$.

The doubly (8) and singly (9) construction are infinite multivariate mixtures of the Poisson mass function on the conditional Dirichlet distribution and it can be seen that the major difference between the *singly* (9) and *doubly* (8) construction is the additional parameter λ_{K+1}. It is of interest to investigate what effect, and if this effect is significant, this additional parameter has when considering parameter estimation. This needs to be explored as the parameter λ_{K+1} is the representative for the $(1-\sum_{i=1}^{K}p_i)$ component in (8) and may be expected to have a larger influence than any other individual λ_i for $i=1,2,\ldots,K$ and has meaningful computational implications.

3 Methods and Approach

MLE was used in order to find the estimates of the parameters of interest. This method maximizes the log-likelihood functions for the distributions of interest and are investigated in the following sections for $K=2$.

Log-Likelihood

Doubly Noncentral Dirichlet Distribution

Suppose N vector observations $p_1,\ldots,p_N$ of dimension Kx1 are drawn independently from the *doubly* noncentral Dirichlet distribution given in (8), then the log-likelihood will take the following form [15]:

$$
\begin{aligned}
&LogL(\mathbf{p};\,\mathbf{\Pi},\,\mathbf{\Lambda})_{Doubly}\\
&= N\log\Gamma(\pi^{+}) - N(\log\Gamma(\pi_1) + \log\Gamma(\pi_2) + \log\Gamma(\pi_3))\\
&+ N\left((\pi_1 - 1)\log\bar{p}_1 + (\pi_2 - 1)\log\bar{p}_2 + (\pi_3 - 1)(1 - \log\bar{p}_1 - \log\bar{p}_2)\right)\\
&- \frac{N(\lambda_1 + \lambda_2 + \lambda_3)}{2} + \sum_{i=1}^{N}\log\mathbf{\Psi}_2^{(3)}(\cdot)
\end{aligned}
\tag{10}
$$

where $log\,\bar{p}_k = \frac{1}{N}\sum_{i=1}^{N}\log p_{ik}$ and $\mathbf{\Psi}_2^{(3)}(\cdot)$ can be expressed as

$$
\begin{aligned}
&\mathbf{\Psi}_2^{(3)}[\pi^{+};\pi_1,\pi_2,\pi_3;\,\tfrac{\lambda_1}{2}p_1,\,\tfrac{\lambda_2}{2}p_2,\,\tfrac{\lambda_3}{2}(1-p_1-p_2)]\\
&= \sum_{j_1=0}^{\infty}\frac{(\pi^{+})_{j_1}}{(\pi_1)_{j_1}}\frac{\left(\frac{\lambda_1}{2}p_1\right)^{j_1}}{j_1!}\mathbf{\Psi}_2^{(2)}[\pi^{+}+j_1;\pi_2,\pi_3;\,\tfrac{\lambda_2}{2}p_2,\,\tfrac{\lambda_3}{2}(1-p_1-p_2)]\\
&= \sum_{j_1=0}^{\infty}\frac{(\pi^{+})_{j_1}}{(\pi_1)_{j_1}}\frac{\left(\frac{\lambda_1}{2}p_1\right)^{j_1}}{j_1!}\sum_{j_2=0}^{\infty}\frac{(\pi^{+}+j_1)_{j_2}}{(\pi_2)_{j_2}}\frac{\left(\frac{\lambda_2}{2}p_2\right)^{j_2}}{j_2!}\mathbf{\Psi}_2^{(1)}[\pi^{+}+j_1+j_2;\pi_3;\,\tfrac{\lambda_3}{2}(1-p_1-p_2)]\\
&= \sum_{j_1=0}^{\infty}\frac{(\pi^{+})_{j_1}}{(\pi_1)_{j_1}}\frac{\left(\frac{\lambda_1}{2}p_1\right)^{j_1}}{j_1!}\sum_{j_2=0}^{\infty}\frac{(\pi^{+}+j_1)_{j_2}}{(\pi_2)_{j_2}}\frac{\left(\frac{\lambda_2}{2}p_2\right)^{j_2}}{j_2!}\sum_{j_3=0}^{\infty}\frac{(\pi^{+}+j_1+j_2)_{j_3}}{(\pi_3)_{j_3}}\frac{\left(\frac{\lambda_3}{2}(1-p_1-p_2)\right)^{j_3}}{j_3!}\\
&= \sum_{j_1=0}^{\infty}\frac{(\pi^{+})_{j_1}}{(\pi_1)_{j_1}}\frac{\left(\frac{\lambda_1}{2}p_1\right)^{j_1}}{j_1!}\sum_{j_2=0}^{\infty}\frac{(\pi^{+}+j_1)_{j_2}}{(\pi_2)_{j_2}}\frac{\left(\frac{\lambda_2}{2}p_2\right)^{j_2}}{j_2!}\,{}_1F_1[\pi^{+}+j_1+j_2;\pi_3;\,\tfrac{\lambda_3}{2}(1-p_1-p_2)]
\end{aligned}
$$

where ${}_1F_1(\cdot)$ represents the confluent hypergeometric function. This function (10) is used in the MLE estimation of the parameters in the applications that follow.

Singly Noncentral Dirichlet Distribution

Suppose N vector observations $p_1, \ldots, p_N$ of dimension Kx1 are drawn independently from the *singly* noncentral Dirichlet distribution given in (9), the log-likelihood will be represented by

$$
\begin{aligned}
&LogL(\mathbf{p};\,\mathbf{\Pi},\,\mathbf{\Lambda})_{Singly}\\
&= N\log\Gamma(\pi^{+}) - N(\log\Gamma(\pi_1) + \log\Gamma(\pi_2) + \log\Gamma(\pi_3))\\
&+ N\left((\pi_1 - 1)\bar{p}_1 + (\pi_2 - 1)\bar{p}_2 + (\pi_3 - 1)(1 - \bar{p}_1 - \bar{p}_2)\right)\\
&- \frac{N(\lambda_1 + \lambda_2)}{2} + \sum_{i=1}^{N}\log\mathbf{\Psi}_2^{(2)}(\cdot)
\end{aligned}
\tag{11}
$$

where $\mathbf{\Psi}_2^{(2)}(\cdot)$ can be expressed as

$$
\begin{aligned}
&\Psi_2^{(2)}[\pi^+;\pi_1,\pi_2;\frac{\lambda_1}{2}p_1,\frac{\lambda_2}{2}p_2]\\
&=\sum_{j_1=0}^{\infty}\frac{(\pi^+)_{j_1}}{(\pi_1)_{j_1}}\frac{\left(\frac{\lambda_1}{2}p_1\right)^{j_1}}{j_1!}\Psi_2^{(1)}[\pi^+ + j_1;\pi_2;\frac{\lambda_2}{2}p_2]\\
&=\sum_{j_1=0}^{\infty}\frac{(\pi^+)_{j_1}}{(\pi_1)_{j_1}}\frac{\left(\frac{\lambda_1}{2}p_1\right)^{j_1}}{j_1!}\sum_{j_2=0}^{\infty}\frac{(\pi^+ + j_1)_{j_2}}{(\pi_2)_{j_2}}\frac{\left(\frac{\lambda_2}{2}p_2\right)^{j_2}}{j_2!}\\
&=\sum_{j_1=0}^{\infty}\frac{(\pi^+)_{j_1}}{(\pi_1)_{j_1}}\frac{\left(\frac{\lambda_1}{2}p_1\right)^{j_1}}{j_1!}{}_1F_1[\pi^+ + j_1;\pi_2;\frac{\lambda_2}{2}p_2]
\end{aligned}
$$

where ${}_1F_1(\cdot)$ represents the confluent hypergeometric function. This function (11) is used in the MLE estimation of the parameters in the applications that follow.

Method for Investigating λ_3

The effect of λ_3 will be evaluated by fitting both the *singly* (9) and *doubly* (8) noncentral Dirichlet distributions to the same datasets and investigating the differences in results using these two distributions. The model performance will be assessed using different goodness-of-fit measures, namely, AIC and BIC, defined as

$$AIC = 2k - 2LogL_{max} \text{ and } BIC = k\log N - 2LogL_{max}$$

where k is the number of free parameters, N is the number of observations in the data and $LogL_{max}$ is the maximized log-likelihood value. Models with lower AIC and BIC are considered models with better performance. Since the models being compared consist of a different number of parameters (when estimating the full set of parameters compared to the two alternative two-step approaches which chooses the $\boldsymbol{\pi}$ to be fixed hence only the estimates of $\boldsymbol{\Lambda}$ are required), the LR test will be used to compare the goodness-of-fit between the two different models in order to determine if the restricted models performed adequately.

The LR test, having an asymptotic Chi-squared distribution (with the number of degrees of freedom equal to the number of parameters fixed by H_0 [15]), can be used when comparing two statistical models which differ in the number of parameters, the complex model containing more parameters than the restricted model. The interest would then be to evaluate if the additional parameters in the full/unrestricted model makes the model significantly more accurate.

The LR test statistic is given by $-2(LogL_{max_{restricted}} - LogL_{max_{unrestricted}})$ where $LogL_{max_{restricted}}$ would be the maximum log-likelihood of the model under the null hypothesis (which is the restricted model or model with less parameters) and

$LogL_{max_{unrestricted}}$ is the log-likelihood of the full/unrestricted model (which would be the model with more parameters).

- H_o: The restricted model should be used.
- H_a : The full/unrestricted model yielded significantly more accurate results.

This means that rejection of the null hypothesis would indicate that the full model (estimating all the parameters of interest) is significantly more accurate than the restricted model. For this investigation, the restricted models will consist of two cases: (1) the $\boldsymbol{\pi}$s will be fixed to be equal to 1, then only the $\boldsymbol{\Lambda}$ would need estimation, and (2) the $\boldsymbol{\pi}$s would be assigned the values of the estimates obtained when estimating the Dirichlet distribution (5) and once again only the $\boldsymbol{\Lambda}$ would need to be estimated.

A graphical guide is also considered in order to display the behaviour of the deviance function. This is expressed as

$$D(\boldsymbol{\pi}; \boldsymbol{\Lambda}) = 2\left(LogL_{max} - LogL_{profile}\right) \tag{12}$$

where $LogL_{max}$ is the log-likelihood of the model with the estimates which maximized the relevant log-likelihood and $LogL_{profile}$ is the profile log-likelihood function obtained by varying the values of π_1 and λ_1 while the values of π_2; π_3; λ_2 and λ_3 in the doubly (8) construction and the values of π_2; π_3; λ_2 in the singly (9) construction remain fixed to the parameter estimates. The combination of π_2 and λ_2 was not considered, as the outcome will remain the same due to the symmetry of the distribution.

Initial Parameters for MLE Search

Due to the potential complexity of the distributions, a wide range of starting values, via the use of a grid search, were considered in order to determine the best possible starting values. An additional two-step estimation process was also investigated of which the steps are explained below. This method explored the use of the Dirichlet estimates (for the $\boldsymbol{\pi}$) as well as making the $\boldsymbol{\pi} = \mathbf{1}$ and then only estimating the values of the $\boldsymbol{\Lambda}$. For the two data examples, as well as the simulations, three estimation approaches were investigated:

1. Estimation of all parameters (π_1; π_2 and π_3 and λ_i where $i = 1, 2, 3$ for the doubly (8) construction and $i = 1, 2$ for the singly (9) construction) by using MLE.
2. Two-step estimation approaches
    a. Setting the $\boldsymbol{\pi} = \mathbf{1}$ and estimating only the λ_is (where $i = 1, 2, 3$ for the doubly (8) construction and $i = 1, 2$ for the singly (9)) using MLE;
    b. Setting the $\boldsymbol{\pi}$ equal to the Dirichlet (5) estimates (π_1; π_2 and π_3) and estimating only the λ_is (where $i = 1, 2, 3$ for the doubly (8) construction and $i = 1, 2$ for the singly (9)) using MLE.

The choice of $\boldsymbol{\pi} = \mathbf{1}$ was made as this will ensure that the focus will be on estimating $\boldsymbol{\Lambda}$ rather than $\boldsymbol{\pi}$. Together with obtaining the estimates which will yield the best model fit, the run time was also investigated as the most accurate estimates calculated in a quick and simple way would be the optimal process.

4 Data Fitting

For this section, two different datasets were identified and the parameters estimated. The investigation was strengthened by including two simulation studies as well. The MLE for the log-likelihoods discussed above, were determined by using the `optim` function in R (statistical software). For the first simulation study, data was generated from each of the three distributions of interest (the Dirichlet (5) distribution, the doubly noncentral Dirichlet (8) distribution and the singly (9) noncentral Dirichlet distribution), while the data generated for the second simulation consisted of a single dataset on which all three distributions were fit. All estimated parameters were reported together with the standard error (SE).

Simulation Study 1

200 bivariate observations from the Dirichlet distribution with parameters ($\pi_1 = 3.2154$; π_2=20.3818; $\pi_3 = 21.6859$), for the doubly noncentral Dirichlet distribution ($\pi_1 = 3.2154$; π_2=20.3818; $\pi_3 = 21.6859$; $\lambda_1 = 7.8564$, $\lambda_2 = 26.6966$; $\lambda_3 = 3.596$) and for the singly noncentral Dirichlet distribution ($\pi_1 = 3.2154$; $\pi_2 = 20.3818$; $\pi_3 = 21.6859$; $\lambda_1 = 7.8564$, $\lambda_2 = 26.6966$), were simulated according to the Acceptance Rejection algorithm (as adapted, see [5]).

Table 1 reports the results for both the singly (9) and doubly (8) constructions, Table 2 reports the LR test results and Fig. 1 overlays the scatter plot and contour plots for the simulated and fitted results.

It can be seen that in all three cases, the models where all the relevant parameters were estimated (the full/unrestricted models) yielded the best AIC results when compared to the restricted models although not being significantly better in all cases.

The LR test showed that, for the singly (9) instance, both sets of constraints (when $\boldsymbol{\pi}$ are set equal to 1 as well as when $\boldsymbol{\pi}$ are set to the Dirichlet estimates) resulted in a rejection of the null hypothesis. This shows that the unrestricted model yielded estimates which had a significantly better fit. For the doubly (8) construction, the null hypothesis could be rejected for the restricted model where $\boldsymbol{\pi} = \mathbf{1}$ (showing the unrestricted model yielded a significantly improved model fit), while for the case where $\boldsymbol{\pi}$ was chosen as the Dirichlet estimates, the unrestricted model did not perform significantly better.

Figure 2 illustrates how the deviance increases as the values of π_1 and λ_1 increases with the white dot indicating the estimated parameter values. The other parameters

Table 1 Parameter estimates and standard errors for the Dirichlet (5), doubly noncentral Dirichlet (8) and the singly noncentral Dirichlet (9) distributions together with the goodness-of-fit measures

	Dirichlet (5) estimates	Doubly (8) estimates	Singly (9) estimates
	Estimate all parameters		
$\hat{\pi}_1$	3,2608 (0,0958)	0,0061 (0,0864)	0,0026 (0,1831)
$\hat{\pi}_2$	22,8908 (0,7421)	10,8783 (0,0271)	38,2250 (0,2227)
$\hat{\pi}_3$	23,7218 (0,7795)	13,3265 (0,1714)	24,7172 (0,1849)
$\hat{\lambda}_1$	n/a	11,6468 (0,0593)	17,1634 (0,2385)
$\hat{\lambda}_2$	n/a	27,5281 (0,2997)	4,5135 (0,3634)
$\hat{\lambda}_3$	n/a	6,9626 (0,1710)	n/a
AIC	–655,7234	–617,5583	–664,2661
BIC	–647,9079	–609,7428	–656,4506
Time (sec)	0,17	43,64	0,65
	Two step estimation where $\boldsymbol{\pi} = \mathbf{1}$		
$\hat{\lambda}_1$	n/a	5,8477 (0,1506)	0 (0,6563)
$\hat{\lambda}_2$	n/a	30,5153 (0,5269)	3,9142 (0,4525)
$\hat{\lambda}_3$	n/a	21,7476 (0,3507)	n/a
AIC	–655,7234	–549,163	–218,1457
BIC	–647,9079	–541,3475	–210,3302
Time (sec)	0,17	17,14	0,37
	Two step estimation where $\boldsymbol{\pi}$ are the Dirichlet estimates		
$\hat{\lambda}_1$	n/a	8,9649 (0,2723)	9,9267 (0,5070)
$\hat{\lambda}_2$	n/a	25,3784 (0,8008)	27,6297 (1,6249)
$\hat{\lambda}_3$	n/a	1,1000 (0,4844)	n/a
AIC	–655,7234	–614,7539	–616,51
BIC	–647,9079	–606,9384	–608,6945
Time (sec)	0,17	7.79	0,42

Table 2 LR test results for simulation study 1

LR test results			
Full doubly model compared to the restricted models	dof	Test statistic	p-value
Doubly ($\boldsymbol{\pi} = \mathbf{1}$)	3	68,3953	<0,0001
Doubly ($\hat{\boldsymbol{\pi}}$ is set to the Dirichlet estimates)	3	2,8043	0,4228
Full singly model compared to the restricted models			
Singly ($\boldsymbol{\pi} = \mathbf{1}$)	3	446,1203	<0,0001
Singly ($\hat{\boldsymbol{\pi}}$ is set to the Dirichlet estimates)	3	47,7561	<0,0001

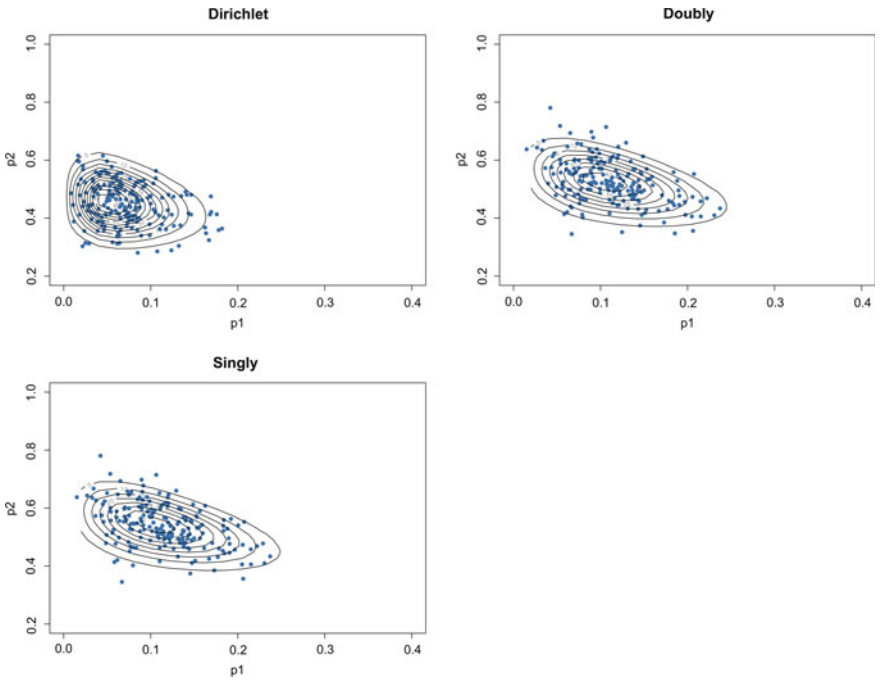


Fig. 1 Scatter plots of the simulation study 1 dataset and contour plots of the fitted results

were kept equal to their estimated values while π_1 and λ_1 were varied. It can be seen that, as these parameters move away from the optimal estimates, the deviance is originally not that sensitive, but becomes more sensitive as the parameters become larger.

Simulation Study 2

For this illustration, a single dataset was simulated, and all three models (the Dirichlet (5), the doubly (8) noncentral Dirichlet as well as the singly (9) noncentral Dirichlet distributions) were fitted to this single dataset.

This data was simulated using the following steps:

1. Generate 100 random variables from a Weibull distribution $X_i \sim Weibull(k_i, \theta_i)$ for $i = 1, 2, 3$ where $k_1 = 2, k_2 = 7, k_3 = 5$ and $\theta_1 = 0.2, \theta_2 = 0.9, \theta_3 = 1$.
2. Define $\boldsymbol{p} = (p_1, p_2, p_3)$ as $p_i = \frac{X_i}{\sum_{i=1}^{3} X_i}$ for $i = 1, 2, 3$ and then generate the random sample.
3. The Dirichlet, noncentral doubly Dirichlet, and noncentral singly Dirichlet distributions are then fitted to $\boldsymbol{p}$.

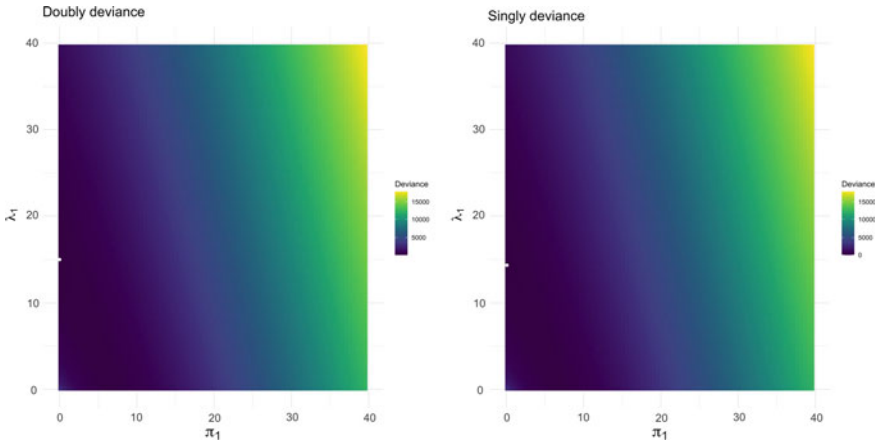


Fig. 2 Deviance function for simulation study 1

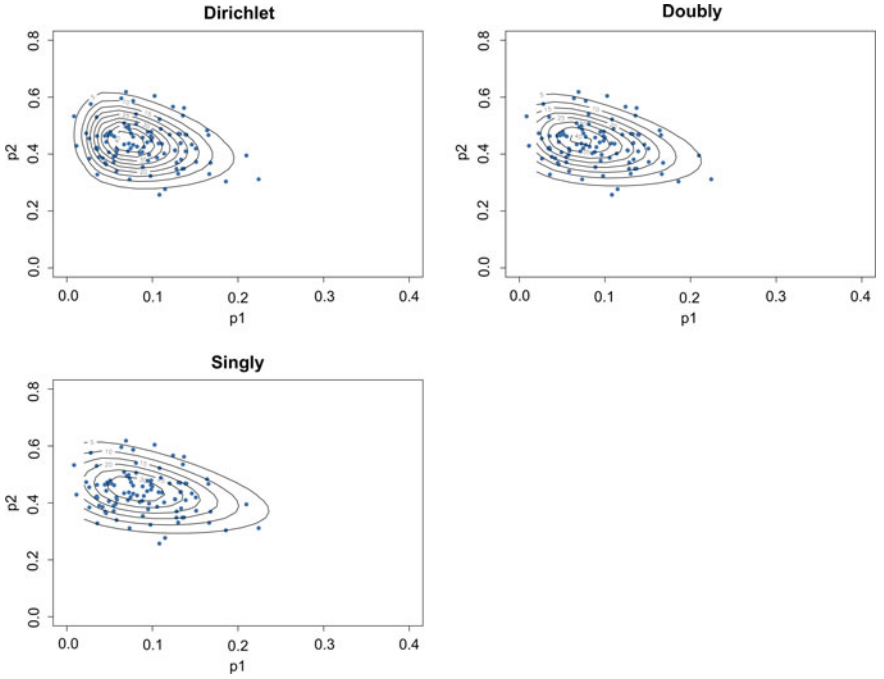


Fig. 3 Scatter plots of the simulation study 2 dataset and contour plots of the fitted results

Table 3 reports the results for both the singly (9) and doubly (8) constructions, Table 4 reports the LR test results and Fig. 3 overlays the scatter plot and contour plots for the generated and fitted results.

Table 3 Parameter estimates and standard errors for the Dirichlet (5), doubly noncentral Dirichlet (8) and the singly noncentral Dirichlet (9) distributions together with the goodness-of-fit measures

	Dirichlet (5) estimates	Doubly (8) estimates	Singly (9) estimates
	Estimate all parameters		
$\hat{\pi}_1$	3,6491 (0,0931)	0,0519 (0,1400)	0,0019 (0,0897)
$\hat{\pi}_2$	17,7015 (0,5129)	13,8592 (0,0783)	18,1034 (0,0274)
$\hat{\pi}_3$	19,3169 (0,5694)	24,6011 (0,1267)	19,8102 (0,0326)
$\hat{\lambda}_1$	n/a	9,9120 (0,1477)	8,7957 (0,1534)
$\hat{\lambda}_2$	n/a	19,9680 (0,2434)	1,4466 (0,0618)
$\hat{\lambda}_3$	n/a	2,0128 (0,1546)	n/a
AIC	–586,5146	–656,1447	–629,1464
BIC	–578,6991	–648,3292	–621,3309
Time (sec)	0,24	37,57	1,18
	Two step estimation where $\boldsymbol{\pi} = \mathbf{1}$		
$\hat{\lambda}_1$	n/a	5,9679 (0,1880)	0 (0,0721)
$\hat{\lambda}_2$	n/a	30,6960 (0,6569)	1,7340 (0,0514)
$\hat{\lambda}_3$	n/a	35,8423 (0,9089)	n/a
AIC	–586,5146	–601,7124	–154,7021
BIC	–578,6991	–593,8969	–146,8865
Time (sec)	0,24	31,95	0,4
	Two step estimation where $\boldsymbol{\pi}$ are the Dirichlet estimates		
$\hat{\lambda}_1$	n/a	6,5269 (0,2026)	0,1405 (0,1023)
$\hat{\lambda}_2$	n/a	26,8695 (0,8908)	0 (0,1207)
$\hat{\lambda}_3$	n/a	30,7942 (0,9817)	n/a
AIC	–586,5146	–625,1907	–586,6108
BIC	–578,6991	–617,3752	–578,7953
Time (sec)	0,24	18,01	0,44

Table 4 LR test results for simulation study 2

LR test results			
Full doubly model compared to the restricted models	dof	Test statistic	p-value
Doubly ($\boldsymbol{\pi} = \mathbf{1}$)	3	54,4323	<0,0001
Doubly ($\hat{\boldsymbol{\pi}}$ is set to the Dirichlet estimates)	3	30,9540	<0,0001
Full singly model compared to the restricted models			
Singly ($\boldsymbol{\pi} = \mathbf{1}$)	3	474,4443	<0,0001
Singly ($\hat{\boldsymbol{\pi}}$ is set to the Dirichlet estimates)	3	42,5356	<0,0001

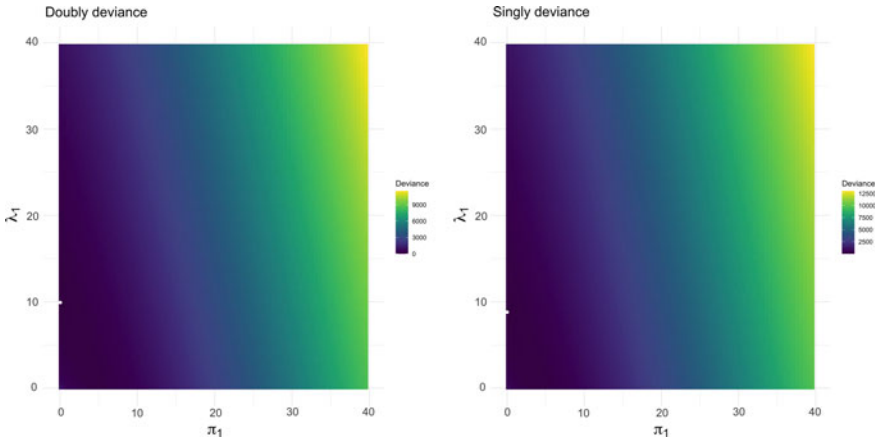


Fig. 4 Deviance function for simulation study 2

When estimates for the parameters were determined for the same dataset, it can be seen that the full/unrestricted models (estimating all parameters) yielded models with better performance than the restricted cases (for both the doubly (8) and the singly (9) constructions). It can also be seen that the noncentral Dirichlet distributions resulted in more flexible fits to the simulated data. Both the singly (9) and the doubly (8) fitted results seem to fit the simulated data shapes.

Similar to simulation study 1, Fig. 4 illustrates how the deviance increases as the values of π_1 and λ_1 increases with the white dot indicating the estimated parameter values.

Dataset 1—Household Expenditure Data

The first dataset that was considered, obtained from [6], was collected through household budget surveys aimed at studying consumer demand. This dataset was also used in studies such as [13] and reports the household expenditures (in Hong Kong Dollars) on two commodity groups of a sample of 40 individuals. The variables considered are the proportions spent on housing (including fuel and lights) (p_1), foodstuffs (including alcohol and tobacco) (p_2), and the rest are classified as services and other goods (including transport and vehicles, clothing, footwear, and durable goods). Table 5 reports the results for both the singly (9) and doubly (8) constructions and Fig. 5 overlays the scatter plot and contour plots for the simulated and fitted results.

From Table 5, it can be seen that the full/unrestricted models yielded the best AIC results with the LR test in Table 6 indicating that for both the doubly (8) and singly (9) constructions, the null hypothesis was rejected (reporting that the full model with no restrictions has significantly better model performance).

Table 5 Parameter estimates and standard errors for the Dirichlet (5), doubly noncentral Dirichlet (8) and the singly noncentral Dirichlet (9) distributions together with the goodness-of-fit measures

	Dirichlet (5) estimates	Doubly (8) estimates	Singly (9) estimates
	Estimate all parameters		
$\hat{\pi}_1$	2,6284 (0,0459)	1,4574 (0,1189)	3,3934 (0,0082)
$\hat{\pi}_2$	1,1749 (0,0152)	0,0004 (0,0715)	0,0021 (0,0213)
$\hat{\pi}_3$	2,3097 (0,0375)	3,1973 (0,1783)	4,7542 (0,0109)
$\hat{\lambda}_1$	n/a	5,8467 (0,0151)	4,5663 (0,0159)
$\hat{\lambda}_2$	n/a	4,6404 (0,0151)	6,5546 (0,0125)
$\hat{\lambda}_3$	n/a	1,8552 (0,0151)	n/a
AIC	–84,1573	–128,1886	–124,9050
BIC	–79,0907	–123,1220	–119,8384
Time (sec)	0,14	40,89	2,45
	Two step estimation where $\pi = 1$		
$\hat{\lambda}_1$	n/a	3,1755 (0,0835)	0 (0,0048)
$\hat{\lambda}_2$	n/a	2,3016 (0,1289)	0,3404 (0,0342)
$\hat{\lambda}_3$	n/a	8,5647 (0.3159)	n/a
AIC	–84,1573	–83,4674	–84,5856
BIC	–79,0907	–78,4008	–79,5189
Time (sec)	0,14	5,84	0,49
	Two step estimation where π are the Dirichlet estimates		
$\hat{\lambda}_1$	n/a	3,0546 (0,1480)	0 (0,0048)
$\hat{\lambda}_2$	n/a	3,4754 (0,1536)	0,3404 (0,0342)
$\hat{\lambda}_3$	n/a	1,7723 (0,0759)	n/a
AIC	–84,1573	–89,6995	–84,5856
BIC	–79,0907	–84,6328	–79,5189
Time (sec)	0,14	10,32	0,26

A consideration when determining the estimates of the parameters is the fact that the log-likelihood function contains multiple infinite summations, inherited from $\boldsymbol{\Psi}_2^{(3)}(\cdot)$, which complicates the computation of the parameters. We show that since those summations stem from Poisson distribution, one can expect convergence as the number of summations increases. Table 7 shows that from 15 summations onward, the results remain equal to 4 decimal places. This confirms that the use of 20 summations, as was used in the Household data example, yields the required results.

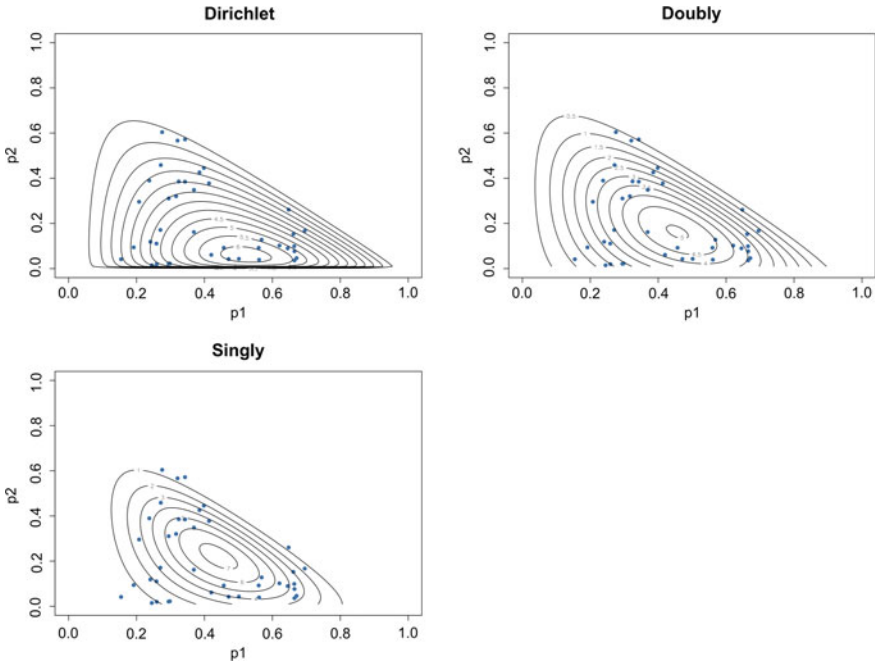


Fig. 5 Scatter plots of the Household dataset and contour plots of the fitted results

Table 6 LR test results of the Household dataset

LR test results			
Full doubly model compared to the restricted models	dof	Test statistic	p-value
Doubly ($\boldsymbol{\pi} = \mathbf{1}$)	3	44.7212	<0.0001
Doubly ($\hat{\boldsymbol{\pi}}$ is set to the Dirichlet estimates)	3	38.4892	< 0.0001
Full singly model compared to the restricted models			
Singly ($\boldsymbol{\pi} = \mathbf{1}$)	3	40.3194	< 0.0001
Singly ($\hat{\boldsymbol{\pi}}$ is set to the Dirichlet estimates)	3	40.3194	< 0.0001

Dataset 2—Pekin Duckling Data

The second dataset that was considered was obtained from [12], referenced in studies like [1], and reported the blood serum proportions (pre-albumin, albumin, and globulin) in 3-week old Pekin ducklings. Table 8 reports the results for both the singly (9)

Table 7 Summation truncation confirmation for the Household dataset

	Singly (9)				
Summation truncation	10	15	20	30	50
$\hat{\pi}_1$	3,2246	3,3934	3,3934	3,3934	3,3934
$\hat{\pi}_2$	0,0004	0,0021	0,0021	0,0021	0,0021
$\hat{\pi}_3$	4,4448	4,7542	4,7542	4,7542	4,7542
$\hat{\lambda}_1$	4,4521	4,5663	4,5663	4,5663	4,5663
$\hat{\lambda}_2$	6,1203	6,5546	6,5546	6,5546	6,5546
Time (sec)	1	1,68	2,19	4,23	11,13
AIC	–124,51	–124,91	–124,91	–124,91	–124,91
BIC	–123,82	–124,21	–124,21	–124,21	–124,21
	Doubly (8)				
Summation truncation	10	15	20	30	50
$\hat{\pi}_1$	3,7347	1,4574	1,4574	1,4574	1,4574
$\hat{\pi}_2$	0,0034	0,0004	0,0004	0,0004	0,0004
$\hat{\pi}_3$	3,3343	3,1973	3,1973	3,1973	3,1973
$\hat{\lambda}_1$	2,5902	5,8467	5,8467	5,8467	5,8467
$\hat{\lambda}_2$	6,11	4,6404	4,6404	4,6404	4,6404
$\hat{\lambda}_3$	2,1889	1,8552	1,8552	1,8552	1,8552
Time (sec)	8,79	17,05	29,18	49,31	64,998
AIC	–130,5	–128,19	–128,19	–128,19	–128,19
BIC	–129,81	–127,5	–127,5	–127,5	–127,5

and doubly (8) constructions, Table 9 report the LR test results and Fig. 6 overlays the scatter plot and contour plots for the simulated and fitted results.

For the Pekin Ducklings dataset, it can be seen that the full parameter models had the highest AIC values. The LR test results in Table 9 indicate that for the doubly (8) construction, the null hypothesis was rejected when $\boldsymbol{\pi} = \mathbf{1}$ (showing that the full unrestricted model yielded a better fit), while for the $\boldsymbol{\pi}$ which was set to the Dirichlet estimates did not have enough evidence to reject the null hypothesis (indicating that the restricted model performance was sufficiently close to the full model). For the singly (9) construction in both the restricted cases, the null hypothesis was rejected indicating that the restricted models resulted in much less impressive performance.

5 Final Thoughts and Future Directions

This chapter aimed to investigate the impact that the additional parameter for $K + 1$ of the doubly (8) construction has on parameter estimation and run time. From the above data examples and simulations, it can be seen that the additional parameter (and

Table 8 Parameter estimates and standard errors for the Dirichlet (5), doubly noncentral Dirichlet (8) and the singly noncentral Dirichlet (9) distributions together with the goodness-of-fit measures

	Dirichlet (5) estimates	Doubly (8) estimates	Singly (9) estimates
	Estimate all parameters		
$\hat{\pi}_1$	3,2154 (0,0805)	0,0092 (0,0439)	0,0005 (0,0314)
$\hat{\pi}_2$	20,3818 (0,5674)	10,9633 (0,0824)	38,5011 (0,2122)
$\hat{\pi}_3$	21,6859 (0,6206)	23,9699 (0,0247)	46,2162 (0,2600)
$\hat{\lambda}_1$	n/a	7,8564 (0,0973)	14,4756 (0,1142)
$\hat{\lambda}_2$	n/a	26,6966 (0,2008)	12,1713 (0,1117)
$\hat{\lambda}_3$	n/a	3,5696 (0,0152)	n/a
AIC	–140,2500	–154,2719	–159,5879
BIC	–136,8435	–150,8654	–156,1814
Time (sec)	0,17	33,95	2,43
	Two step estimation where $\boldsymbol{\pi} = \mathbf{1}$		
$\hat{\lambda}_1$	n/a	4,1783 (0,1965)	0.000 (0,0631)
$\hat{\lambda}_2$	n/a	30,5914 (0,0846)	2,0289 (0,0635)
$\hat{\lambda}_3$	n/a	34,4858 (0,1331)	n/a
AIC	–140,2500	–137,1889	–32,6149
BIC	–136,8435	–133,7824	–29,2084
Time (sec)	0,17	24,56	0,48
	Two step estimation where $\hat{\pi}$ are the Dirichlet estimates		
$\hat{\lambda}_1$	n/a	4,9133 (0,0522)	0,0951 (0,0404)
$\hat{\lambda}_2$	n/a	26,3713 (0,2957)	0,2613 (0,0562)
$\hat{\lambda}_3$	n/a	28,8659 (0,3402)	n/a
AIC	–140,2500	–147,3002	–140,2650
BIC	–136,8435	–143,8937	–136,8585
Time (sec)	0,17	19,63	0,34

Table 9 LR test results of the Pekin Ducklings dataset

LR test results			
Full doubly model compared to the restricted models	dof	Test statistic	p-value
Doubly ($\boldsymbol{\pi} = \mathbf{1}$)	3	18.0830	<0.0001
Doubly ($\hat{\boldsymbol{\pi}}$ is set to the Dirichlet estimates)	3	6.9717	0.0728
Full singly model compared to the restricted models			
Singly ($\boldsymbol{\pi} = \mathbf{1}$)	3	126.9730	<0.0001
Singly ($\hat{\boldsymbol{\pi}}$ is set to the Dirichlet estimates)	3	19.3228	<0.0001

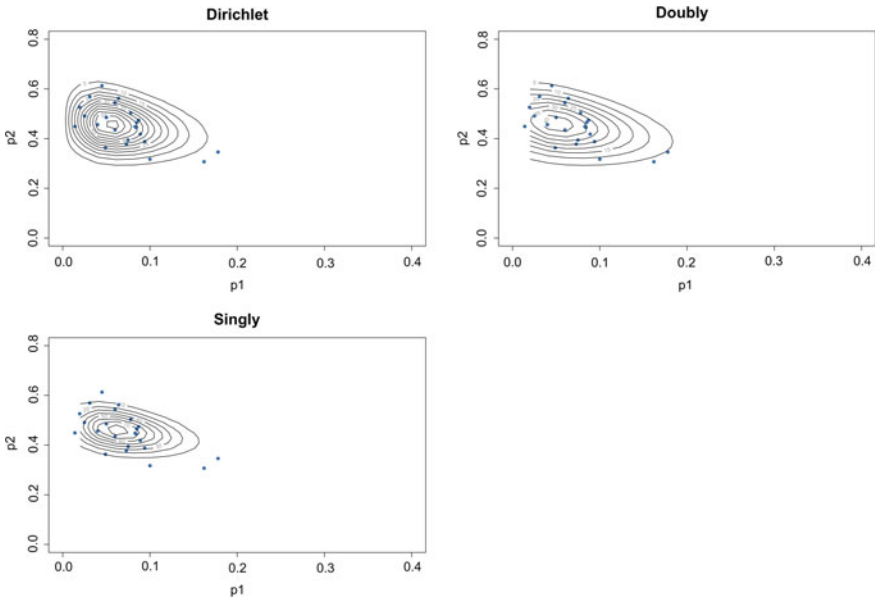


Fig. 6 Scatter plots of the Pekin Ducklings dataset and contour plots of the fitted results

additional model complexity) did yield more accurate models. Model complexity should be carefully considered as the data shape is likely to have an effect on the parameter estimation results.

The overall run time of the doubly models was always higher but still short enough to imply reasonably quick estimation. A wide range of starting values, via the use of a grid search, had to be considered to ensure the best model fit (more so for the doubly (8) than for the singly (9) constructions) and further investigations included two additional estimation approaches using a two-step estimation proses. Figures 5 and 6 showed that the doubly construct was able to better capture the outliers. Reference [10] investigated improved Dirichlet models with respect to outlier detections in more detail. It was also shown that, although the noncentral Dirichlet has an infinite sum construction, the summations converge quickly, which makes this form usable when estimating parameters.

Through the parameter estimation study, it was observed that the flexibility of the noncentral Dirichlet (5) resulted in cases where there could be more than one set of parameters that yielded valid results. This was also seen for the Skewed t-distribution investigation by [3]. This means that even though the simulated example was produced using a single set of parameters, several models exist which yielded good fits, but with different parameters. Investigations showed that this persists even for larger samples.

A possible future consideration includes investigating the new noncentral generalization of the beta distribution introduced by [18] and considering, implementing, and investigating the corresponding noncentral Dirichlet distribution.

Acknowledgements This work is based on the research supported in part by the National Research Foundation of South Africa (SARChI Research Chair- UID: 71199; Grant ref. SRUG190308422768 nr. 120839; and Grant ref. RA201125576565, nr 145681), the Research Development Programme at the University of Pretoria 296/2022 as well as the Centre of Excellence in Mathematical and Statistical Sciences Grant ref. #2022-047-STA, based at the University of the Witwatersrand, South Africa. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Appendix

The following functions were used within the `optim` function in the R, to find the MLE of the doubly noncentral Dirichlet distribution.

```
loglik <- function(par,X) {

# Data info
log.pbar <- colMeans(log(X))
pbar <- colMeans((X))
N <- nrow(X)
K <- ncol(X)

# Parameters
pi1 = par[1]
pi2 = par[2]
pi3 = par[3]

lambda1 = par[4]
lambda2 = par[5]
lambda3 = par[6]

LogLA <- N*(lgamma(pi1+pi2+pi3))
LogLB <- - N*(lgamma(pi1)+lgamma(pi2)+lgamma(pi3))
LogLC <- N*((pi1 - 1)*log.pbar[1] + (pi2 - 1)*log.pbar[2] +
 (pi3 - 1)*log.pbar[3])

LogLD <- -N*(lambda1/2 + lambda2/2 + lambda3/2)

itterations = 20
sumall = 0
for(j1 in 0:itterations)
for(j2 in 0:itterations)
{

A = ((gamma(pi1+pi2+pi3+j1))/(gamma(pi1+pi2+pi3)))/
      ((gamma(pi1+j1))/(gamma(pi1)))
B = (((lambda1/2)*pbar[1])^j1)/factorial(j1)
C = ((gamma(pi1+pi2+pi3+j1+j2))/(gamma(pi1+pi2+pi3+j1)))/
```

```
       ((gamma(pi2+j2))/(gamma(pi2)))
D = (((lambda2/2)*pbar[2])^j2)/factorial(j2)
E = genhypergeo(pi1+pi2+pi3+j1+j2,pi3,(lambda3/2)*(pbar[3]))

sum = A*B*C*D*E
sumall = sumall + sum
}

LogLE <- N*log(sumall)

LL <- LogLA + LogLB + LogLC + LogLD + LogLE

return(-LL)
}
```

References

1. Arashi, M., Bekker, A., de Waal, D., & Makgai, S. (2020). Constructing multivariate distributions via the Dirichlet generator. In *Computational and methodological statistics and biostatistics* (pp. 159–186). Springer.
2. Ascari, R., Migliorati, S., & Ongaro, A. (2021). The double flexible Dirichlet: A structured mixture model for compositional data. *Applied Modeling Techniques and Data Analysis 2: Financial, Demographic, Stochastic and Statistical Models and Methods, 18*, 135–152 (2021).
3. Azzalini, A., & Salehi, M. (2020). Some computational aspects of maximum likelihood estimation of the skew-t distribution. In *Computational and methodological statistics and biostatistics* (pp. 3–28). Springer.
4. Bekker, A., & Ferreira, J. T. (2018). Bivariate gamma type distributions for modeling wireless performance metrics. *Statistics, Optimization & Information Computing, 6*(3), 335–353.
5. Botha, T., Ferreira, J. T., & Bekker, A. (2021). Alternative Dirichlet priors for estimating entropy via a power sum functional. *Mathematics, 9*(13), 1493.
6. Cox, D. R., Hinkley, D. V., Rubin, D., & Silverman, B. W. (1984). Monographs on statistics and applied probability. Springer.
7. Ferreira, J. T., Bekker, A., & Arashi, M. (2016). Bivariate noncentral distributions: An approach via the compounding method. *South African Statistical Journal, 50*(1), 103–122.
8. Knüsel, L., & Bablok, B. (1996). Computation of the noncentral gamma distribution. *SIAM Journal on Scientific Computing, 17*(5), 1224–1231.
9. Kotz, S., Balakrishnan, N., & Johnson, N. L. (2004). Continuous multivariate distributions, Models and applications (vol. 1). Wiley.
10. Makgai, S., Bekker, A., & Arashi, M. (2021). Compositional data modeling through Dirichlet innovations. *Mathematics, 9*(19), 2477.
11. Mathai, A. M., Saxena, R. K., & Haubold, H. J. (2009). The H-function: Theory and applications. Springer Science & Business Media.
12. Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika, 49*(1/2), 65–82.
13. Ongaro, A., & Migliorati, S. (2013). A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis, 114*, 412–426.
14. Ongaro, A., & Orsi, C. (2015). Some results on non-central beta distributions. *Statistica, 75*(1), 85–100.
15. Orsi, C. (2021). A new class of non-central Dirichlet distributions. arXiv:2107.14392
16. Orsi, C. (2021). A novel approach to handling the non-central Dirichlet distribution. arXiv:2108.08947

17. Sánchez, L. E., Nagar, D., & Gupta, A. (2006). Properties of noncentral Dirichlet distributions. *Computers & Mathematics with Applications, 52*(12), 1671–1682.
18. Schein, A., Nagulpally, A., Wallach, H., Flaherty, P. (2021). Doubly non-central beta matrix factorization for DNA methylation data. arXiv:2106.06691

Modeling Handwritten Digits Dataset Using the Matrix Variate t Distribution

Y. Murat Bulut and Olcay Arslan

Abstract In this paper, we consider matrix variate t distribution and explore some of its distributional properties to model handwritten digits dataset. In particular, we show that the marginal and the conditional distributions are also matrix variate t distribution. We provide parameter estimation of the matrix variate t distribution using the EM algorithm. We give a small simulation study to show the performance of the proposed EM algorithm for finding the estimates for the parameters of interests. A real data example illustrates that the matrix variate t distribution can be used as a robust alternative to the matrix variate normal distribution for modeling matrix variate datasets with some atypical observations.

1 Introduction

After data collection techniques have developed rapidly, in many applied areas such as medical research, data has been started to store as three-way data. For example, functional Magnetic Resonance Imaging (fMRI) data, electroencephalograph data, and Nuclear Magnetic Resonance (NMR) data contain matrix variate structure (Niu et al. [15]). The importance of the matrix variate distributions has started to increase to model the data mentioned above. Also, parallel to the rapid developments in computer technology, estimation of the parameters of the matrix variate distributions has become more accessible. Due to these reasons, defining new matrix variate distributions has been taken much more interest. Especially in robust statistical analysis, heavy-tailed distributions are essential such as slash and t distributions. For this reason, univariate and multivariate slash and t distributions are studied extensively. Some examples of these works can be given as follows. Joarder [12] was interested

Y. M. Bulut (✉)
Eskişehir Osmangazi University, Eskişehir, Turkey
e-mail: ymbulut@ogu.edu.tr

O. Arslan
Ankara University, Ankara, Turkey
e-mail: oarslan@ankara.edu.tr

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_5

in the expectation of the Wishart distribution based on the multivariate t distribution. Nadarajah [14] proposed an alternative t distribution. Also, the conditional distribution of the multivariate t distribution is studied by Ding [4]. Recently, Doğru et al. [6] obtained the maximum Lq (MLq) estimator of the multivariate t distribution. In the last decades, the applications of the multivariate t distribution have increased. As an example of the applications, Pesevski et al. [16] used the multivariate t distribution in the clustering. For further details on the multivariate t distribution, the reader can look into the book of Kotz and Nadarajah [13]. However, in the matrix variate case, the literature is very limited. Dickey [3] proposed matrix variate t distribution, but, in his work, estimation of the parameters of the distribution has not been studied. After the work of Dickey [3], Javier and Gupta [11] looked at distributional properties of the matrix variate t distribution. Also, Gupta et al. [10] introduced a matrix variate Pearson type VII distribution, and the matrix variate t distribution can be obtained as a particular case of this distribution. Thompson et al. [20] used the matrix variate t distribution in classification problem. Bulut [1] redefined matrix variate t distribution as a scale mixture of matrix variate normal and gamma distribution to estimate the parameters of the matrix variate t distribution using the Expectation-Maximization (EM) algorithm. Then, Doğru et al. [5] introduced the finite mixtures of matrix variate t distributions. Recently, Bulut and Arslan [2] have defined the matrix variate slash distribution as a heavily tailed alternative of the matrix variate normal distribution and studied some statistical properties of it. The main aim of this study is to model the handwritten digits dataset via matrix variate t distribution. The dataset has been used by Dryden and Mardia [7] and S*á*nchez et al. [19].

The paper is organized as follows. In Sect. 2, we first define the matrix variate t distribution as the scale mixture of the matrix variate normal and gamma distributions. We study some distributional properties, such as expectation, variance, characteristic function, and marginal and conditional distributions of the matrix variate t distribution in the same section. In Sect. 3, we first consider the ML estimation and then show that the ML estimators can be obtained using the EM algorithm. In Sect. 4, we provide a small simulation study to illustrate the performance of the proposed EM algorithm for finding the ML estimates for the parameters of the matrix variate t distribution. We also give a real data example to demonstrate the robustness capability of the matrix variate t distribution for handling outliers in the matrix variate datasets.

2 Matrix Variate t Distribution

A $n \times p$-variate random matrix X is said to have a matrix variate t distribution with mean matrix M of size $n \times p$ and variance-covariance matrices Σ and Ψ of size $n \times n$ and $p \times p$, respectively, denoted by $Mt_{n,p}(M, \Sigma, \Psi, m)$, if it has the probability density function given by

$$f(X) = \frac{|\Sigma|^{-\frac{p}{2}} |\Psi|^{-\frac{n}{2}} \Gamma\left(\frac{np+m}{2}\right)}{(\pi m)^{\frac{np}{2}} \Gamma\left(\frac{m}{2}\right)} \left[1 + \frac{\delta_X(M, \Sigma, \Psi)}{m}\right]^{-\frac{np+m}{2}}, \quad (1)$$

where

$$\delta_X(M, \Sigma, \Psi) = tr\left\{\Sigma^{-1}(X - M)\Psi^{-1}(X - M)'\right\}$$

which is the Mahalanobis distance from X to the center M with respect to Σ and Ψ.

Matrix variate t distribution can be obtained as a scale mixture of matrix variate normal and gamma distributions [10]. The scale mixture representation of matrix variate t distribution is given in the following definition. Also, Gupta and Nagar [9] gave details of matrix variate t distribution using Wishart distribution as a scale distribution.

Definition 1 Let $Z \sim N_{n,p}(0, I_n, I_p)$ and $Y \sim Gamma\left(\frac{m}{2}, \frac{m}{2}\right)$ be two independent random matrix and random variable and let X be a new random matrix defined as

$$X = M + \Sigma^{\frac{1}{2}} Z \Psi^{\frac{1}{2}} Y^{-\frac{1}{2}},$$

where $M \epsilon R^{n \times p}$, Σ, and Ψ are positive definite symmetric matrices and $\Sigma^{\frac{1}{2}}$, $\Psi^{\frac{1}{2}}$ are the positive definite square root of Σ and Ψ, respectively. Then X has matrix variate t distribution with pdf given in (1).

The matrix variate t distribution given in the above definition is a special case of matrix variate Pearson type VII distribution defined by [10], when q is taken as $q = \frac{np+m}{2}$. When $m = 1$, we obtain matrix variate Cauchy distribution, which is given by [18]. Note that if $p = 1$ and $\Psi = I$ then our distribution reduces to the multivariate t distribution.

In the following propositions we will give some properties of the matrix variate t distribution.

Proposition 1 *Let $X \sim Mt_{n,p}(M, \Sigma, \Psi, m)$ then the expectation and the covariance of X are given by*

$$E(X) = M \quad (2)$$

$$Cov(X) = \frac{m}{m-2}(\Sigma \otimes \Psi), \; m > 2. \quad (3)$$

Proof It can be easily seen that conditional distribution of X given Y is $N_{n,p}\left(M, y^{-1}\Sigma, \Psi\right)$.

$$E(X) = M + \Sigma^{\frac{1}{2}} E(Z) \Psi^{\frac{1}{2}} E(Y^{-\frac{1}{2}}) = M$$

$$\begin{aligned} Cov(X) &= E_Y\left(Cov\left(X \mid Y\right)\right) \\ &= E_Y\left(Y^{-1}\left(\Sigma \otimes \Psi\right)\right) \\ &= E_Y\left(Y^{-1}\right)\left(\Sigma \otimes \Psi\right) \\ &= \frac{m}{m-2}\left(\Sigma \otimes \Psi\right). \end{aligned}$$

Proposition 2 *If* $X \sim Mt_{n,p}\left(M, \Sigma, \Psi, m\right)$, *then the characteristic function of* X *is*

$$\phi_X\left(T\right) = etr\left(iT'M\right)\frac{2\left(\frac{m}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)}\left(\frac{tr\left(T'\Sigma T\Psi\right)}{m}\right)^{\frac{m-2}{4}} K_{\frac{m}{2}}\left(\sqrt{m \times tr\left(T'\Sigma T\Psi\right)}\right), \tag{4}$$

where $K_\lambda\left(s\right) = \frac{1}{2}\int_0^\infty u^{\lambda-1}\exp\left\{-\frac{1}{2}s\left(u+u^{-1}\right)\right\}du$, $s > 0$ *is the modified Bessel function of the third kind (or MacDonald function).*

Proof We know conditional expectation of X given Y and characteristic function of matrix variate normal distribution. So we can obtain characteristic function of the matrix variate t distribution using these two properties as follows:

$$\begin{aligned} \phi_X\left(T\right) &= E\left[etr\left(iXT'\right)\right] = E_Y\left[E_{X|Y}\left[etr\left(iXT'\right)\right]\right] \\ &= E_Y\left[etr\left(iT'M - \frac{1}{2}y^{-1}T'\Sigma T\Psi\right)\right] \\ &= E_Y\left[etr\left(iT'M\right)etr\left(-\frac{1}{2}y^{-1}T'\Sigma T\Psi\right)\right] \\ &= etr\left(iT'M\right)E_Y\left[etr\left(-\frac{1}{2}y^{-1}T'\Sigma T\Psi\right)\right] \\ &= etr\left(iT'M\right)\int_0^\infty \exp\left\{-\frac{y^{-1}}{2}tr\left(T'\Sigma T\Psi\right)\right\}\frac{\left(\frac{m}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)}y^{\frac{m}{2}-1}\exp\left\{-\frac{m}{2}y\right\}dy \\ &= etr\left(iT'M\right)\frac{\left(\frac{m}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)}\int_0^\infty y^{\frac{m}{2}-1}\exp\left\{-\frac{y^{-1}}{2}tr\left(T'\Sigma T\Psi\right) - \frac{m}{2}y\right\}dy \\ &= etr\left(iT'M\right)\frac{\left(\frac{m}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)}\left(\frac{tr\left(T'\Sigma T\Psi\right)}{m}\right)^{\frac{m-2}{4}} K_{\frac{m}{2}}\left(\sqrt{m \times tr\left(T'\Sigma T\Psi\right)}\right). \end{aligned}$$

Proposition 3 *Let* $X \sim Mt_{n,p}\left(M, \Sigma, \Psi, m\right)$ *and* $V = A + BXC$ *where* A *is an* $n \times p$ *matrix,* B *and* C *are* $n \times n$ *and* $p \times p$ *non-singular matrices, respectively. Then* $V \sim Mt_{n,p}\left(A + BMC, B\Sigma B', C'\Psi C, m\right)$.

Proof This can be easily proved using the characteristics function as follows:

$$\begin{aligned}
\phi_V(T) &= E\left[etr\left(iVT'\right)\right] \\
&= E\left[etr\left(i(A+BXC)T'\right)\right] \\
&= etr\left(iAT'\right) E\left[etr\left(XCT'B\right)\right] \\
&= etr\left(iAT'\right) E\left[etr\left(XT_1'\right)\right] \\
&= etr\left(iT'(A+BMC)\right) \frac{2\left(\frac{m}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} \left(\frac{tr\left(T_1'\Sigma T_1\Psi\right)}{m}\right)^{\frac{m-2}{4}} \\
&\quad \times K_{\frac{m}{2}}\left(\sqrt{m \times tr\left(T_1'\Sigma T_1\Psi\right)}\right) \\
&= etr\left(iT'(A+BMC)\right) \frac{2\left(\frac{m}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} \left(\frac{tr\left(T'(B\Sigma B')T(C'\Psi C)\right.}{m}\right)^{\frac{m-2}{4}} \\
&\quad \times K_{\frac{m}{2}}\left(\sqrt{m \times tr\left(T'(B\Sigma B')T(C'\Psi C)\right)}\right).
\end{aligned}$$

So, $V \sim Mt_{n,p}\left(A+BMC, B\Sigma B', C'\Psi C, m\right)$.

Proposition 4 *Let* $X \sim Mt_{n,p}(M, \Sigma, \Psi, m)$, *and partition* X, M, Σ *and* Ψ *as* $X = \binom{X_{1r}}{X_{2r}}$, $M = \binom{M_{1r}}{M_{2r}}$, $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$ *and* $\Psi = \begin{pmatrix}\Psi_{11} & \Psi_{12}\\ \Psi_{21} & \Psi_{22}\end{pmatrix}$ *where* X_{1r} *and* M_{1r} *are* $s \times p$ *matrices,* Σ_{11} *is* $s \times s$ *matrix.*

Then (i) $X_{1r} \sim Mt_{s,p}(M_{1r}, \Sigma_{11}, \Psi, m)$.

(ii) $X_{2r} \mid X_{1r} \sim Mt_{(n-s),p}\left(M_{2|1}, \Sigma_{2|1}, \Psi, m\right)$, *where*

$$\begin{aligned}
M_{2|1} &= M_{2r} + \Sigma_{21}\Sigma_{11}^{-1}(X_{1r} - M_{1r}) \\
\Delta_1^2 &= tr\left\{\Sigma_{11}^{-1}(X_{1r} - M_{1r})\Psi^{-1}(X_{1r} - M_{1r})'\right\} \\
\Sigma_{22.1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
\Sigma_{2|1} &= \frac{m + \Delta_1^2}{ps + m}\Sigma_{22.1}.
\end{aligned}$$

Proof (i) In Proposition 3, when we take as $A = 0$, $B = (I_s \quad 0)$ and $C' = \begin{pmatrix} I_p \\ 0 \end{pmatrix}$, we get density of X_{1r}.

(ii) We can sketch the proof of part (ii) as follows. Firstly,

$$\begin{aligned}
(X-M)'\Sigma^{-1}(X-M) = &(X_{1r} - M_{1r})'\Sigma_{11}^{-1}(X_{1r} - M_{1r}) \\
&+ \left(X_{2r} - M_{2|1}\right)'\Sigma_{2.1}^{-1}\left(X_{2r} - M_{2|1}\right).
\end{aligned}$$

Then,

$$
\begin{aligned}
P\left(X_{2r} \mid X_{1r}\right) &= \frac{\frac{|\Sigma|^{-\frac{p}{2}}|\Psi|^{-\frac{n}{2}}\Gamma\left(\frac{pn+m}{2}\right)}{(\pi m)^{\frac{pn}{2}}\Gamma\left(\frac{m}{2}\right)}\left[1+\frac{tr\left\{\Sigma^{-1}(X-M)\Psi^{-1}(X-M)'\right\}}{m}\right]^{-\frac{pn+m}{2}}}{\frac{|\Sigma_{11}|^{-\frac{p}{2}}|\Psi|^{-\frac{s}{2}}\Gamma\left(\frac{ps+m}{2}\right)}{(\pi m)^{\frac{ps}{2}}\Gamma\left(\frac{m}{2}\right)}\left[1+\frac{tr\left\{\Sigma_{11}^{-1}(X_{1r}-M_{1r})\Psi^{-1}(X_{1r}-M_{1r})'\right\}}{m}\right]^{-\frac{ps+m}{2}}} \\
&= \frac{|\Sigma_{22.1}|^{-\frac{p}{2}}|\Psi|^{-\frac{n-s}{2}}\Gamma\left(\frac{p(n-s)+\alpha}{2}\right)}{(\pi m)^{\frac{p(n-s)}{2}}\Gamma\left(\frac{\alpha}{2}\right)}\left[1+\frac{\Delta_1^2+\Delta_2^2}{m}\right]^{-\frac{pn+m}{2}}\left[1+\frac{\Delta_1^2}{m}\right]^{\frac{ps+m}{2}} \\
&= \frac{|\Sigma_{22.1}|^{-\frac{p}{2}}|\Psi|^{-\frac{n-s}{2}}\Gamma\left(\frac{p(n-s)+\alpha}{2}\right)}{\left(\pi\left(m+\Delta_1^2\right)\right)^{\frac{p(n-s)}{2}}\Gamma\left(\frac{\alpha}{2}\right)}\left[1+\frac{\Delta_2^2}{m+\Delta_1^2}\right]^{-\frac{p(n-s)+\alpha}{2}} \\
&= \frac{\left|\Sigma_{2|1}\right|^{-\frac{p}{2}}|\Psi|^{-\frac{n-s}{2}}\Gamma\left(\frac{p(n-s)+\alpha}{2}\right)}{(\pi\alpha)^{\frac{p(n-s)}{2}}\Gamma\left(\frac{\alpha}{2}\right)} \\
&\quad\times\left[1+\frac{tr\left\{\Sigma_{2|1}^{-1}\left(X_{2r}-M_{2r}\right)\Psi^{-1}\right\}\left(X_{2r}-M_{2r}\right)'}{\alpha}\right]^{-\frac{p(n-s)+\alpha}{2}},
\end{aligned}
$$

where

$$\alpha = ps + m.$$

Therefore, $X_{2r} \mid X_{1r} \sim Mt_{n-s,p}\left(M_{2|1}, \Sigma_{2|1}, \Psi, \alpha\right)$.

3 Parameter Estimation

In this section, we discuss parameter estimation of the matrix variate t distribution. We use maximum likelihood estimation method and EM algorithm to estimate parameters of interest. We show that the maximum likelihood estimators are equivalent to the estimators obtained from the EM algorithm.

Maximum Likelihood Estimation

We assume that we have i.i.d. data matrices $X_1,\ X_2,\ ...,\ X_l$ in $R^{n\times p}$ and model these dataset with matrix variate t distribution with parameters M, Σ, and Ψ. The log-likelihood function is obtained as follows:

$$l(M, \Sigma, \Psi; X) = \frac{pl}{2} \ln \left|\Sigma^{-1}\right| + \frac{nl}{2} \ln \left|\Psi^{-1}\right| + l \ln \left\{\Gamma\left(\frac{np+m}{2}\right)\right\} \tag{5}$$
$$-\frac{npl}{2} \ln (\pi m) - l \ln \left\{\Gamma\left(\frac{m}{2}\right)\right\}$$
$$-\frac{np+m}{2} \sum_{i=1}^{l} \ln \left\{1 + \frac{\delta_{X_i}(M, \Sigma, \Psi)}{m}\right\}.$$

When we take the derivatives of the log-likelihood function with respect to the parameters M, Σ, Ψ, and setting them to zero yield the following estimating equations:

$$\widehat{M} = \frac{\sum_{i=1}^{l} w_i X_i}{\sum_{i=1}^{l} w_i} \tag{6}$$

$$\widehat{\Sigma} = \frac{1}{pl} \sum_{i=1}^{l} w_i \left(X_i - \widehat{M}\right) \widehat{\Psi}^{-1} \left(X_i - \widehat{M}\right)' \tag{7}$$

$$\widehat{\Psi} = \frac{1}{nl} \sum_{i=1}^{l} w_i \left(X_i - \widehat{M}\right)' \widehat{\Sigma}^{-1} \left(X_i - \widehat{M}\right), \tag{8}$$

where

$$w_i = \frac{np+m}{m + tr\left\{\widehat{\Sigma}^{-1} \left(X_i - \widehat{M}\right) \widehat{\Psi}^{-1} \left(X_i - \widehat{M}\right)'\right\}}, i = 1, 2, ..., l \tag{9}$$

are the weights. Since the weights are dependent on the estimators, Eqs. (6)–(8) are not the explicit forms of the estimators. The numerical methods should be used to solve these equations. Further, if we want to estimate the degrees of freedom (m), we have to solve the following equation along with the equations given above:

$$\frac{l}{2} \Upsilon\left(\frac{np+m}{2}\right) - \frac{npl}{2m} - \frac{l}{2} \Upsilon\left(\frac{m}{2}\right) - \frac{1}{2} \sum_{i=1}^{l} \ln \left\{1 + \frac{\delta_{X_i}(M, \Sigma, \Psi)}{m}\right\}$$
$$+ \frac{np+m}{2m} \sum_{i=1}^{l} \frac{\delta_{X_i}(M, \Sigma, \Psi)}{m + \delta_{X_i}(M, \Sigma, \Psi)} = 0. \tag{10}$$

When we rewrite this equation, we get

$$\widehat{m} = \frac{np\left(l - \sum_{i=1}^{l} \frac{\delta_{X_i}(M,\Sigma,\Psi)}{\widehat{m}+\delta_{X_i}(M,\Sigma,\Psi)}\right)}{l\Upsilon\left(\frac{np+\widehat{m}}{2}\right) - l\Upsilon\left(\frac{\widehat{m}}{2}\right) - \sum_{i=1}^{l} \ln\left(1 + \frac{\delta_{X_i}(M,\Sigma,\Psi)}{\widehat{m}}\right) + \sum_{i=1}^{l} \frac{\delta_{X_i}(M,\Sigma,\Psi)}{\widehat{m}+\delta_{X_i}(M,\Sigma,\Psi)}}, \quad (11)$$

where Υ is the Digamma function. For the sake of robustness the degrees of freedom m will be taken as fixed. The reason for this can be explain as follows. When $\delta_{X_i}(M, \Sigma, \Psi)$ tends to ∞, Eq. (10) also tends to ∞, which means that the observations with larger Mahalanobis distance will ruin the estimates of the m. That is, from the robustness point of view, the influence function will be unbounded when we estimate the degrees of freedom along with the other parameters. This is not a desired case in robustness, therefore, from now on, we will assume that the degrees of freedom m is fixed and will use it as the robustness tuning constant. Since when m tends to infinity the matrix variate t distribution tends to matrix variate normal distribution. We will prefer to take small values of m to maintain robustness.

Estimation via EM Algorithm

In this section, the EM algorithm to obtain the ML estimates of (M, Σ, Ψ) is given. To use EM algorithm, we will use the scale mixture representation of the matrix variate t distribution given in definition (1). We take Y's as missing and X's as observed data and form the complete data as (X,Y). The log-likelihood function for the complete data (X_i, Y_i), for $i = 1, 2, ..., l$ can be obtained as

$$\begin{aligned} L(M, \Sigma, \Psi) = &-\frac{npl}{2}\ln(2\pi) - \frac{pl}{2}\ln|\Sigma| - \frac{nl}{2}\ln|\Psi| \\ &+\frac{ml}{2}\ln(m) - \frac{ml}{2}\ln(2) - l\ln\left\{\Gamma\left(\frac{m}{2}\right)\right\} \\ &+\left(\frac{np+m}{2} - 1\right)\sum_{i=1}^{l}\ln y_i - \frac{1}{2}\sum_{i=1}^{l} y_i\left[m + \delta_{X_i}(M, \Sigma, \Psi)\right]. \end{aligned} \quad (12)$$

We can easily maximize the complete data log-likelihood function but Y is a latent variable and we cannot observe it. Maximization of this function depends on Y. Since we cannot observe Y, we cannot use them. To handle this problem, we will take conditional expectation of $L(M, \Sigma, \Psi)$ given the observed data X_i and the current estimates of parameters. When we take conditional expectation of $L(M, \Sigma, \Psi)$, we get

$$
\begin{aligned}
Q\left(M,\Sigma,\Psi\right) &= E\left(L\left(M,\Sigma,\Psi\right)|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right) \\
&= -\frac{npl}{2}\ln\left(2\pi\right)-\frac{pl}{2}\ln\left|\Sigma\right|-\frac{nl}{2}\ln\left|\Psi\right| \\
&\quad +\frac{ml}{2}\ln\left(m\right)-\frac{ml}{2}\ln\left(2\right)-l\ln\left\{\Gamma\left(\frac{m}{2}\right)\right\} \\
&\quad +\left(\frac{np+m}{2}-1\right)\sum_{i=1}^{l}E\left(\ln Y_i|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right) \\
&\quad -\frac{1}{2}\sum_{i=1}^{l}E\left(Y_i|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right)\left[m+\delta_{X_i}\left(M,\Sigma,\Psi\right)\right],
\end{aligned}
\tag{13}
$$

where $E\left(\ln Y_i|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right)$ and $E\left(Y_i|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right)$ are conditional expectations of $\ln Y_i$ and Y_i given the observed data X_i and current estimates of parameters. To find this conditional expectation, we need conditional distribution of Y given X. After some straightforward algebra, we obtain conditional distribution that has gamma distribution with parameters $\frac{np+m}{2}$, $\frac{m+\delta_X(M,\Sigma,\Psi)}{2}$ as follows:

$$
f_{Y|X}\left(y\right)=\frac{\left(\frac{m+tr\left\{\Sigma^{-1}(X-M)\Psi^{-1}(X-M)'\right\}}{2}\right)^{\frac{np+m}{2}}}{\Gamma\left(\frac{np+m}{2}\right)}y^{\frac{np+m}{2}-1}\exp\left\{-\frac{m+\delta_X\left(M,\Sigma,\Psi\right)}{2}y\right\}.
\tag{14}
$$

Using this conditional distribution, we get

$$
\eta_i=E\left(\ln Y_i|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right)=\int_0^{\infty}\ln y_i f\left(y_i|X_i\right)dy_i
\tag{15}
$$

and

$$
w_i=E\left(Y_i|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right)=\frac{np+m}{m+\delta_{X_i}\left(M,\Sigma,\Psi\right)}.
\tag{16}
$$

It should be noticed that the weight functions w_i are the same with the weights given in the ML estimation. If the conditional expectations $E\left(\ln Y_i|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right)$ and $E\left(Y_i|X_i,\widehat{M},\widehat{\Sigma},\widehat{\Psi}\right)$ are replaced by η_i and w_i we get the following objective function to be maximized:

$$
\begin{aligned}
Q\left(M,\Sigma,\Psi\right) &= C-\frac{pl}{2}\ln\left|\Sigma\right|-\frac{nl}{2}\ln\left|\Psi\right| \\
&\quad +\left(\frac{np+m}{2}-1\right)\sum_{i=1}^{l}\eta_i-\frac{1}{2}\sum_{i=1}^{l}w_i\left[m+\delta_{X_i}\left(M,\Sigma,\Psi\right)\right].
\end{aligned}
\tag{17}
$$

Taking the derivatives of $Q\left(M,\Sigma,\Psi\right)$ with respect to the parameters M, Σ and Ψ then setting them equal to zero yield the following estimators:

$$\widehat{M} = \frac{ave\left\{w_i X_i\right\}}{ave\left\{w_i\right\}} \tag{18}$$

$$\widehat{\Sigma} = \frac{1}{p} ave\left\{w_i \left(X_i - \widehat{M}\right) \widehat{\Psi}^{-1} \left(X_i - \widehat{M}\right)'\right\} \tag{19}$$

and

$$\widehat{\Psi} = \frac{1}{n} ave\left\{w_i \left(X_i - \widehat{M}\right)' \widehat{\Sigma}^{-1} \left(X_i - \widehat{M}\right)\right\}, \tag{20}$$

where "ave" stands for the arithmetic average over $i = 1,\ 2, ..., l$. The solutions of these equations can be obtained using the following iteratively reweighting algorithm.

Iteratively reweighting algorithm

1 Set iteration number k = 1 and select initial estimates for the parameters M, Σ and Ψ.
2 Using the current estimates $M^{(k)}$, $\Sigma^{(k)}$ and $\Psi^{(k)}$ for $k = 1, 2, 3, ...$, and the equation given in (16) calculate the weight $w_i^{(k)}$ for $i = 1, 2, 3, ..., l$, and find the averages $ave\left\{w_i^{(k)}\right\}$ and $ave\left\{w_i^{(k)} X_i\right\}$.
3 Use the following updating equations to calculate the new estimates: $M^{(k+1)}$, $\Sigma^{(k+1)}$ and $\Psi^{(k+1)}$

$$M^{(k+1)} = \frac{ave\left\{w_i^{(k)} X_i\right\}}{ave\left\{w_i^{(k)}\right\}}$$

$$\Sigma^{(k+1)} = \frac{1}{pl} \sum_{i=1}^{l} w_i^{(k)} (X_i - M^{(k+1)}) \left(\Psi^{(k+1)}\right)^{-1} (X_i - M^{(k+1)})'$$

$$\Psi^{(k+1)} = \frac{1}{nl} \sum_{i=1}^{l} w_i^{(k)} (X_i - M^{(k+1)})' \left(\Sigma^{(k+1)}\right)^{-1} (X_i - M^{(k+1)}).$$

4 Repeat these steps until convergence. The stopping rule is taken as 10^{-10}.

4 Simulation Study and Real Data Example

In this section, we give a small simulation study to show performance of the proposed algorithm. Also, a real data example is given to show that our proposed method can be applied to the matrix variate datasets in real world.

Simulation Study

In this part of the study, we give a simulation study to show the robustness property of the matrix variate t distribution. We compare the matrix variate t distribution with the matrix variate normal distribution when the dataset contains outlier(s). In simulation study, we take $n = 2, 3, 4$, $p = 2, 4$, and $m = 3, 5$. The data are generated using scale mixture representation given in Definition 1. Also, we randomly choose positive definite variance-covariance matrices. We compute the mean Euclidean distance given in [8] and [2] to compare the performance of the estimates. To illustrate the superiority of the matrix variate t distribution toward the outliers, we add one and two outliers to the dataset. The iteratively reweighting algorithm given in Sect. 3 is used to obtain the estimates. The results are shown in Tables 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18. In these tables, we give the mean iteration numbers with standard error and the mean Euclidean distances for the parameters. All simulations are done in the R software program [17].

We give the results when there is no outliers in the dataset in Tables 1, 2, 3, 4, 5, and 6, the results are given when dataset includes one outlier in Tables 7, 8, 9, 10, 11, and 12. The results for the case, which includes two outliers, are given in Tables 13, 14, 15, 16, 17, and 18. For all simulation designs, the mean Euclidean distances for the matrix variate t distribution's parameters are smaller than that of the matrix variate normal distribution.

When we look primarily at the mean Euclidean distances for the Σ parameter, the matrix variate normal distribution distances are larger than the distances for the matrix variate t distribution.

For example, when we analyze the estimations of the location matrices M, the estimations using the matrix variate t distribution are closer to the true parameter matrices than the estimations with the matrix variate normal distribution. We give the true parameter matrices and the obtained estimates with matrix variate normal and matrix variate t distributions as follows:

- 2×2 case without outlier ($m = 3$):

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \qquad \widehat{M}_{MVt} = \begin{bmatrix} 1.02579 & 1.99764 \\ 2.07785 & 3.07007 \end{bmatrix} \qquad \widehat{M}_{MVN} = \begin{bmatrix} 1.05434 & 2.07219 \\ 2.07839 & 3.08469 \end{bmatrix}.$$

- 2×2 case with one outlier ($m = 3$):

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \qquad \widehat{M}_{MVt} = \begin{bmatrix} 1.04607 & 2.07781 \\ 2.16318 & 3.08066 \end{bmatrix} \qquad \widehat{M}_{MVN} = \begin{bmatrix} 1.13911 & 2.27243 \\ 1.71526 & 2.47729 \end{bmatrix}.$$

- 2×2 case with two outliers ($m = 3$):

Table 1 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 2×2 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	1.53237	10.88280	1.94936	17.61(1.31908)	1.21797	1.32396	1.51361	24.72(0.53769)
20	1.18893	10.4495	1.93466	11.97(0.93889)	0.95300	1.22603	1.30925	18.74(0.96011)
50	0.76955	10.07860	1.92757	8.90(0.38599)	0.55281	1.10976	1.18488	15.52(0.13594)
100	0.61493	10.00100	1.88982	8.19(0.52507)	0.40804	1.00638	1.14758	14.30(0.10493)

Table 2 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 2×2 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	1.31135	5.67594	1.99442	13.35(0.73187)	1.16800	1.18285	1.56003	21.71(0.55891)
20	0.88369	5.55367	1.89179	9.34(0.29481)	0.78524	1.13332	1.27444	16.60(0.22383)
50	0.66616	5.52565	1.89017	6.89(0.15301)	0.55281	1.10976	1.18488	15.52(0.13594)
100	0.45524	5.49319	1.85726	6.08(0.09917)	0.39709	1.0291	1.10328	12.15(0.11135)

Table 3 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 3×2 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	4.78136	67.11050	1.90553	19.16(0.58529)	3.47081	6.06784	2.12486	31.51(0.47979)
20	3.50388	55.54650	1.89685	13.85(0.54502)	2.52649	5.71896	1.77348	24.89(0.20345)
50	2.21740	53.14490	1.86865	9.89(0.25737)	1.51121	5.24169	1.82630	23.54(0.11584)
100	1.87734	52.4362	1.83690	9.33(0.26364)	1.05382	5.27909	1.72196	23.00(0.08040)

Table 4 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 3×2 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	4.08908	31.5536	1.91308	16.98(0.42091)	3.49790	5.93789	2.23945	26.95(0.52231)
20	2.85034	30.82850	1.87306	11.35(0.31087)	2.44390	5.49936	1.95885	18.93(0.20313)
50	1.87483	29.85800	1.86298	8.41(0.13341)	1.59983	5.29687	1.77588	15.57(0.10565)
100	1.26602	27.10200	1.84024	7.18(0.10860)	1.05693	5.29533	1.68642	14.36(0.06439)

Table 5 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 4×4 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	11.94530	142.53600	7.06758	25.05(1.19978)	8.47687	4.66531	4.83432	62.16(0.80876)
20	9.32115	137.52800	6.93982	18.39(1.45386)	5.68134	3.30332	3.15771	56.18(0.29142)
50	5.46061	130.44300	6.89875	12.58(0.42359)	3.55748	2.02918	2.07763	54.32(0.12050)
100	5.06928	119.19600	6.83008	14.61(2.69402)	2.46003	1.46633	1.51339	54.35(0.11044)

Table 6 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 4×4 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	10.16880	94.54790	6.98385	20.30(0.82001)	8.26667	4.91365	4.61107	38.99(0.35746)
20	6.89187	83.09010	6.88661	13.03(0.22449)	5.49452	2.98050	3.17875	35.33(0.16333)
50	4.29206	79.611	6.85099	9.67(0.22566)	3.46958	2.10986	1.85238	34.17(0.08415)
100	3.06737	76.61960	6.8349	7.99(0.10298)	2.62371	1.46563	1.42781	33.74(0.05794)

Table 7 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 2×2 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	2.18727	17.2975	1.88016	25.61(2.33938)	1.36905	1.76541	1.30950	26.27(0.65518)
20	1.56593	17.24030	1.94617	14.86(1.11944)	0.93722	1.37685	1.19880	19.28(0.26018)
50	0.89778	13.96740	1.88840	9.26(0.52831)	0.59472	1.09853	1.10231	15.8(0.14839)
100	0.58364	11.86260	1.89749	7.54(0.18879)	0.40260	1.04750	1.09249	14.46(0.09036)

Table 8 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 3×2 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	6.27684	121.649	1.92856	20.87(0.57096)	3.61334	6.07881	2.61482	33.33(0.54976)
20	3.82555	76.86360	1.88381	14.88(0.50418)	2.47253	5.37650	2.1294	25.77(0.21313)
50	2.61031	75.2623	1.85132	11.52(0.38677)	1.61722	5.21727	1.88113	23.91(0.13714)
100	1.78150	68.16320	1.83695	9.38(0.36210)	1.13345	5.19366	1.74295	23.26(0.09059)

Table 9 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 4×4 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	14.70456	230.77700	6.97649	30.61(1.54390)	8.55757	4.84936	5.15138	65.52(1.02341)
20	11.34840	203.12200	6.95982	22.96(1.44606)	6.14856	3.46047	3.39355	57.24(0.29717)
50	5.75301	161.73400	6.91232	12.76(0.52688)	3.61144	2.06692	2.05314	54.59(0.12317)
100	4.25004	150.18600	6.85427	11.74(0.76616)	2.48383	1.48713	1.45781	54.28(0.08998)

Table 10 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 2×2 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	2.01314	14.43470	1.91552	21.43(1.85430)	1.27855	1.86455	1.28172	23.87(0.69423)
20	1.20700	9.80097	1.89145	11.75(0.47020)	0.87824	1.30199	1.20849	17.01(0.32208)
50	0.70878	7.42444	1.87195	7.65(0.17603)	0.56506	1.13337	1.12709	13.47(0.14665)
100	0.49123	6.28141	1.83493	6.40(0.11282)	0.41604	1.02832	1.09677	12.33(0.10923)

Table 11 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 3×2 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	5.64158	63.64490	1.90579	20.16(0.53290)	4.01874	5.66999	2.88791	27.48(0.53134)
20	3.53555	47.72030	1.84252	13.47(0.35858)	2.52044	5.39133	2.07890	19.12(0.23019)
50	2.01804	36.22790	1.83647	9.17(0.17059)	1.54069	5.24319	1.80490	15.58(0.10462)
100	1.39809	33.71810	1.81699	7.44(0.11748)	1.10455	5.16384	1.70371	14.42(0.05538)

Table 12 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 4×4 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	13.24570	155.44400	6.92639	28.09(1.09619)	8.71037	5.15590	5.12549	40.90(0.37349)
20	8.78697	139.15100	6.90079	17.94(0.76868)	5.83730	3.25174	3.34487	35.96(0.18308)
50	4.60099	105.20800	6.84502	10.96(0.33452)	3.44264	2.01139	1.99864	34.27(0.08147)
100	3.27489	88.46030	6.81577	8.79(0.28152)	2.61521	1.42120	1.46098	33.87(0.05801)

Table 13 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 2×2 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	2.90817	29.21220	2.00677	26.37(2.28847)	1.45373	2.16479	1.37877	26.96(0.51383)
20	1.81940	22.36340	1.95108	17.66(2.26788)	0.95471	1.47266	1.21067	20.07(0.33006)
50	0.93785	14.27650	1.89780	9.74(0.40939)	0.54762	1.13665	1.11164	15.61(0.14135)
100	0.66672	14.05740	1.84960	8.39(0.33782)	0.37422	1.04377	1.10616	14.57(0.09771)

Table 14 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 3×2 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	9.82594	320.32400	1.87242	24.50(0.79766)	4.00437	6.37362	3.26314	34.36(0.52369)
20	5.14815	158.32400	1.85230	16.88(0.56520)	2.87443	6.16002	2.45727	26.89(0.25698)
50	2.59307	76.77820	1.84002	11.25(0.36636)	1.53412	5.18393	1.89400	24.10(0.12753)
100	1.82908	76.61300	1.81621	9.63(0.40643)	1.05388	5.14385	1.78087	23.28(0.08885)

Table 15 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 4×4 matrix with $m = 3$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	19.65080	357.10000	7.00335	33.04(1.48364)	9.60454	5.13721	5.59402	68.82(1.01338)
20	12.61360	319.15600	6.91620	21.82(1.94435)	6.22262	3.34243	3.67310	57.89(0.29777)
50	6.44658	199.61700	6.87067	14.48(0.82468)	3.58040	2.00048	2.28111	55.41(0.15182)
100	4.13812	159.36700	6.84432	10.87(0.33894)	2.61339	1.44061	1.50266	54.30(0.08933)

Table 16 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 2×2 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	2.79052	20.17480	1.91165	24.46(1.85654)	1.48302	2.29779	1.33598	24.22(0.69857)
20	1.66260	14.01260	1.87709	13.38(0.66024)	0.95643	1.63151	1.08406	17.03(0.24963)
50	0.74747	9.14411	1.87636	8.37(0.25491)	0.48363	1.20668	1.07402	13.82(0.15465)
100	0.47954	7.71265	1.84479	6.69(0.16859)	0.35208	1.11195	1.05436	12.27(0.08973)

Table 17 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 3×2 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	7.32243	113.39800	1.88167	22.13(0.58597)	4.27117	5.96972	3.32009	28.63(0.53232)
20	4.33854	76.10910	1.87482	14.85(0.38439)	2.71232	5.13454	2.44000	20.01(0.23333)
50	2.29875	43.4556	1.86963	9.93(0.18654)	1.58115	5.10503	1.89792	16.15(0.10481)
100	1.47028	37.69870	1.85242	7.76(0.13342)	1.06506	4.93999	1.79238	14.65(0.06416)

Table 18 Mean Euclidean distance and mean number of iteration ($\mp$ standard error) for 4×4 matrix with $m = 5$

1	MVN				MVt			
l	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. Num.	$Dist.for M$	$Dist.for \Sigma$	$Dist.for \Psi$	It. num.
10	19.19320	313.418	6.95041	33.32(1.49706)	9.13528	5.91771	5.98180	43.11(0.35359)
20	9.88829	174.07400	6.84235	17.91(0.78201)	6.42038	3.29534	3.79361	36.79(0.15718)
50	5.06966	110.72700	6.80395	11.12(0.18162)	3.69997	2.08250	2.10726	34.70(0.07588)
100	3.29944	98.06360	6.77235	9.08(0.24066)	2.51498	1.31089	1.49056	33.95(0.05389)

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \qquad \widehat{M}_{MVt} = \begin{bmatrix} 1.06798 & 2.05100 \\ 2.16018 & 3.19864 \end{bmatrix} \qquad \widehat{M}_{MVN} = \begin{bmatrix} 1.32406 & 2.53885 \\ 2.76579 & 4.08112 \end{bmatrix}.$$

- 3×2 case without outlier ($m = 3$):

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix} \qquad \widehat{M}_{MVt} = \begin{bmatrix} 1.05280 & 1.91374 \\ 2.21235 & 2.99623 \\ 3.19964 & 4.07000 \end{bmatrix} \qquad \widehat{M}_{MVN} = \begin{bmatrix} 0.80346 & 1.80379 \\ 2.20842 & 3.08979 \\ 3.20990 & 4.09955 \end{bmatrix}.$$

- 3×2 case with one outlier ($m = 3$):

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix} \qquad \widehat{M}_{MVt} = \begin{bmatrix} 1.17342 & 2.03351 \\ 2.26554 & 3.23642 \\ 3.05007 & 3.96752 \end{bmatrix} \qquad \widehat{M}_{MVN} = \begin{bmatrix} 1.38366 & 2.52874 \\ 2.36595 & 3.69411 \\ 3.32959 & 4.34463 \end{bmatrix}.$$

- 3×2 case with two outliers ($m = 3$):

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix} \qquad \widehat{M}_{MVt} = \begin{bmatrix} 1.03132 & 2.12592 \\ 2.09868 & 3.13250 \\ 3.17662 & 4.44383 \end{bmatrix} \qquad \widehat{M}_{MVN} = \begin{bmatrix} 0.89290 & 2.74668 \\ 2.18187 & 3.98364 \\ 3.98512 & 6.40407 \end{bmatrix}.$$

- 4×4 case without outlier ($m = 3$):

$$M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{bmatrix} \quad \widehat{M}_{MVt} = \begin{bmatrix} 0.89207 & 1.95076 & 2.97223 & 4.03787 \\ 2.14901 & 3.11992 & 3.85708 & 4.93792 \\ 3.13713 & 3.91302 & 4.99331 & 5.93357 \\ 4.10254 & 5.39211 & 6.07825 & 7.35972 \end{bmatrix} \quad \widehat{M}_{MVN} = \begin{bmatrix} 1.35678 & 2.03718 & 2.76801 & 4.11028 \\ 2.56443 & 3.00045 & 3.87414 & 5.02325 \\ 3.04378 & 4.00131 & 4.65277 & 6.12111 \\ 4.22871 & 5.45834 & 5.92372 & 7.68975 \end{bmatrix}.$$

- 4×4 case with one outlier ($m = 3$):

$$M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{bmatrix} \quad \widehat{M}_{MVt} = \begin{bmatrix} 0.73220 & 1.95560 & 3.17691 & 3.86809 \\ 1.97699 & 2.92470 & 3.73882 & 5.17241 \\ 2.95812 & 3.61036 & 5.23018 & 6.15853 \\ 4.21818 & 5.22513 & 6.04554 & 7.11775 \end{bmatrix} \quad \widehat{M}_{MVN} = \begin{bmatrix} 1.33423 & 2.25872 & 3.82733 & 4.65344 \\ 1.95235 & 3.49446 & 4.71045 & 6.02251 \\ 3.39677 & 4.39333 & 6.27633 & 7.28391 \\ 5.22891 & 6.47515 & 6.82254 & 8.21866 \end{bmatrix}.$$

- 4×4 case with two outliers ($m = 3$):

$$M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{bmatrix} \quad \widehat{M}_{MVt} = \begin{bmatrix} 1.05188 & 1.77309 & 3.31074 & 4.04982 \\ 1.91885 & 2.87254 & 3.67961 & 4.89278 \\ 3.19680 & 4.25647 & 5.43013 & 6.16060 \\ 4.31925 & 5.40456 & 6.43534 & 7.32990 \end{bmatrix} \quad \widehat{M}_{MVN} = \begin{bmatrix} 1.27886 & 2.29372 & 3.66465 & 5.49613 \\ 1.94104 & 3.83083 & 4.60531 & 6.77915 \\ 4.12300 & 5.60599 & 6.74433 & 8.37796 \\ 5.15879 & 6.50417 & 7.87118 & 9.38572 \end{bmatrix}.$$

Table 19 Loglike values for real data

Model	Loglike without outlier	Loglike with outlier
Normal	−1128.66	−1667.11
$t(3)$	−1470.31	−1642.61
$t(5)$	−1465.32	−1648.01

Real Data Example

In this section, we give a real data example to show that parameter estimation procedure is working as claimed. The dataset, which is used in this study, is used by [7] and [19]. The description of the dataset is as follows which is taken from [7]. *There are* $n = 13$ *landmarks and* $k = 2$ *dimensions for* $m = 30$ *handwritten records of digit* 3. *A landmark is a point of correspondence on each object. The first landmark* 1 *is at the extreme bottom left, landmark* 4 *is at the maximum curvature of the bottom arc, landmark* 7 *is at the extreme of the central protrusion, landmark* 10 *is at the maximum curvature of the top arc, and landmark* 13 *is the extreme top left point. The other landmarks are pseudo-landmarks, localized at approximately equal intervals between the previous landmarks.*

We will use the dataset to estimate mean and variance-covariance matrices by using proposed algorithm. We take different degrees of freedom to compare which model is the best for the dataset. We use log-likelihood value to select the best model (Table 19).

We estimate the mean and variance-covariance matrices for this data using the matrix variate normal and matrix variate t distributions. The estimation procedure described in Sect. 3 is used to estimate the parameters. We observe that results are similar. They both give similar estimates for the mean matrix and similar covariance estimates for Σ and Ψ. To see the effect of the outliers we change the location of two observations to make them as potential outliers and then we again estimate the mean and variance-covariance matrices using matrix variate normal and matrix variate t distributions. Table 19 summarizes the log-likelihood values obtained from original dataset and the dataset with outliers. From this table, we can see that the log-likelihood values for the matrix variate normal distribution are smaller than the log-likelihood value for matrix variate t distribution when there are no outliers. On the other hand, for the dataset with outliers, the log-likelihood value for matrix variate t is getting smaller compared to the log-likelihood value for the matrix variate normal distribution. Figure 1 displays the scatter plot of the dataset without and with outliers along with the estimated mean matrices obtained from matrix variate normal and matrix variate t distributions. We can observe that without outliers the figures are very similar (first column of Fig. 1). However, we can observe that, unlike the matrix variate t distribution case, the estimated mean matrix obtained from the matrix variate normal distribution is affected by the outliers. Therefore, the matrix variate

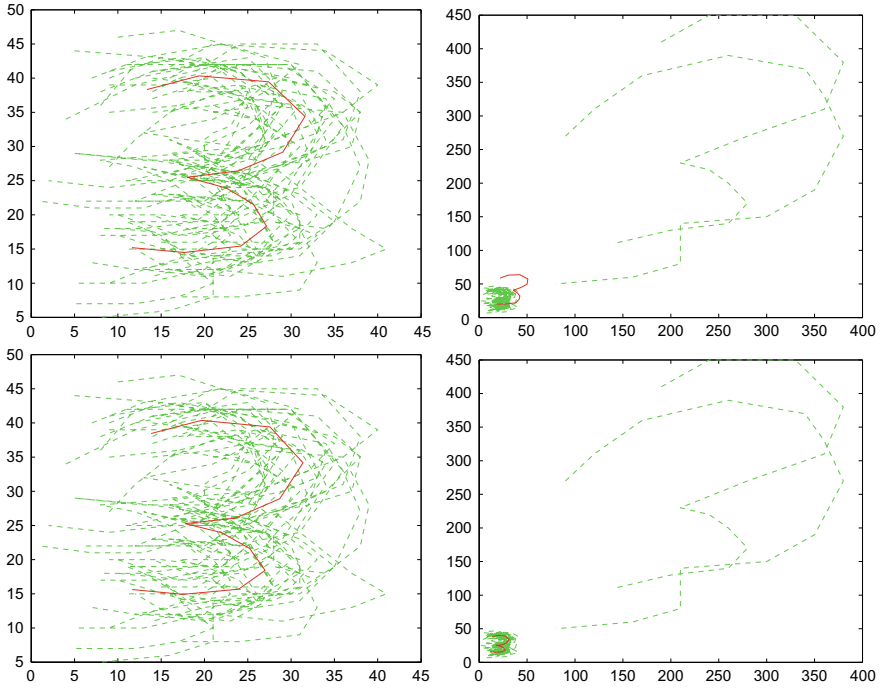


Fig. 1 Data and estimated mean matrix

t distribution can be used to estimate the location and scatter matrices of the matrix variate data to obtain robust estimators.

5 Conclusions

In this paper, the matrix variate t distribution, which is redefined as the scale mixture of the matrix variate normal distribution, using the univariate gamma distribution, is discussed, and some of its distributional properties are given. Parameter estimation is conducted with the maximum likelihood estimation method, and the EM algorithm is given to compute the estimates of the parameters. It is shown that ML estimators of the location and the scatter matrices are alternative robust estimators to the sample mean and the sample covariance estimators. A simulation study and a real data application are given to compare the modeling performances of the matrix variate distribution over the matrix variable normal distribution.

The main purpose of this article is to provide an alternative model for modeling matrix variate datasets with possible outliers. The results of the simulation study and the real data example show that the matrix variate t distribution can be a good

alternative model to the matrix variate normal distribution in case of a possible outlier problem.

Acknowledgements We thank the Associate Editor and the anonymous referees for their valuable comments and suggestions that have greatly improved the paper.

References

1. Bulut, Y. M. (2015). Matrix variate distributions and applications in robust statistical analysis (Unpublished doctoral dissertation), Eskişehir, Turkey.
2. Bulut, Y. M., & Arslan, O. (2015). Matrix variate slash distribution. *Journal of Multivariate Analysis, 137*, 173–178.
3. Dickey, J. M. (1967). Matricvariate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *The Annals of Mathematical Statistics, 38*, 511–518.
4. Ding, P. (2016). On the Conditional Distribution of the Multivariate t Distribution. *The American Statistician, 70*(3), 293–295.
5. Doğru, F. Z., Bulut, Y. M., & Arslan, O. (2016). Finite mixtures of matrix variate t distributions. *Gazi University Journal of Science, 29*(2), 335–341.
6. Doğru, F. Z., Bulut, Y. M., & Arslan, O. (2018). Doubly Reweighted Estimators for the parameters of the multivariate t distribution. *Communication in Statistics-Theory and Methods, 47*(19), 4751–4771.
7. Dryden, I. L., & Mardia, K. V. (1998). *Statistical shape analysis*. Chichester: Wiley.
8. Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation, 64,* 105–123.
9. Gupta, A.K., & Nagar, D. K. (2000). *Matrix variate distributions*. Boca Raton: Chapman & Hall/CRC.
10. Gupta, A. K., Varga, T., & Bodnar, T (2013). *Elliptically contoured models in statistics and portfolio theory*. New York: Springer.
11. Javier, W. R., & Gupta, A. K. (1985). On matric variate-t distribution. *Communication in Statistics-Theory and Methods, 14*(6), 1413–1425.
12. Joarder, A. H. (1998). Some useful Wishart expectations based on the multivariate t-model. *Statistical Papers, 39*, 223–229.
13. Kotz, S., & Nadarajah, S. (2004). *Multivariate t distribution and their applications*. Cambridge: Cambridge University Press.
14. Nadarajah, S. (2009). The product t density distribution arising from the product of two Student's t PDFs. *Statistical Papers, 50*, 605–615.
15. Niu, L., Liu, X., & Zhao, J. (2020). Robust estimator of the correlation matrix with sparse Kronecker structure for a high-dimensional matrix-variate. *Journal of Multivariate Analysis, 177*.
16. Pesevski, A., Franczak, B. C., & McNicholas, P. D. (2018). Subspace clustering with the multivariate t distribution. *Pattern Recognition, 112*, 297–302.
17. R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
18. Sarr, A., & Gupta, A. K. (2011). Exponential scale mixture of matrix variate Cauchy distribution. *Proceedings of the American Mathematical Society, 139*(4), 1483–1494.
19. Sánchez, L., Leiva, V., Caro-Lopera, F. J., & Cysneiros, F. J. (2015). On matrix variate Birnbaum-Saunders distributions and their estimation and application. *Brazilian Probability and Statistics, 29*(4), 790–812.
20. Thompson, G. Z., Maitra, R., Meeker, W. Q., & Bastawros, A. F. (2020). Classification with the matrix-variate-t distribution. *Journal of Computational and Graphical Statistics, 29*(3), 668–674.

On the Identification of Extreme Elements in a Residual for the GMANOVA-MANOVA Model

Béatrice Byukusenge, Dietrich von Rosen, and Martin Singull

Abstract Two different matrix residuals in a special GMANOVA-MANOVA model have previously been established (see Byukusenge et al., 2021, "On residual analysis in the GMANOVA-MANOVA model"). The residual that is studied in this article is constructed via the difference of the observed group means and the estimated mean structure. The residual provides information about the appropriateness of the model assumptions concerning the mean structure. The aim of this paper is to study the distribution of the largest elements (by absolute value) of the residual via two data sets. Parametric bootstrap is used to identify thresholds so that extreme elements of the residuals can be identified.

1 Introduction

In statistics, when we have a random sample or dataset reflecting a practical phenomenon, usually the aim is to interpret estimators or to build a predictive model. The model presented in this paper belongs to the class of multivariate models called the GMANOVA-MANOVA model. References to the MANOVA model can be found in many textbooks on multivariate analysis, for example [1] or [15]. The GMANOVA

B. Byukusenge · D. von Rosen · M. Singull (✉)
Department of Mathematics, Linköping University, Linköping, Sweden
e-mail: martin.singull@liu.se

B. Byukusenge
e-mail: beatrice.byukusenge@liu.se

D. von Rosen
e-mail: dietrich.von.rosen@slu.se

B. Byukusenge
Department of Mathematics, University of Rwanda, Kigali, Rwanda

D. von Rosen
Department of Energy and Technology, Swedish University of Agricultural Sciences, Uppsala, Sweden

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_6

is a generalization of the MANOVA (see, e.g., [12, 14]), and the GMANOVA-MANOVA model is a special case of the Extended Growth Curve model (EGCM). The model was introduced by [5], and particularly, it can be used for the analysis of balanced repeated measurements with covariates.

Residuals usually play an important role in the model validation process and are often based on the difference between the predicted and observed values. Among other things, they can be used to check if model assumptions are satisfied.

Residuals for the GMANOVA model were introduced by von Rosen [13] and have been studied by several authors. In [13], some basic properties of the residuals were derived. Hamid and von Rosen [10] studied the residuals in the Extended Growth Curve model.

In the GMANOVA-MANOVA model, the mean follows a bilinear structure, e.g., see [14], and hence the residuals are more structured than for the MANOVA model, which is a linear model. Moreover, the GMANOVA model and its extension belong to the curved exponential family whereas the MANOVA model belongs to the exponential family. This makes the assessment of the model validation more difficult compared to, for example, univariate linear models. Recently, Byukusenge et al. [2, 3] have established two important residuals for handling the GMANOVA-MANOVA model.

Particularly, it has been shown by [2] that one residual follows a matrix normal distribution whereas the second residual obeys a complicated distribution, which only can be approximated. The second residual matrix contains the difference between the group means and the estimated mean structure and is the residual, which is considered in this article.

The distribution of the "extreme" elements of the residual matrix is approximated via parametric bootstrap samples. We are interested to decide if "extreme" elements of the residual lie in the tails of their distributions. Thus cut-off points for being "extreme" are needed to be established.

The following notations will be used. For any matrix $\boldsymbol{A}$, we denote by $\boldsymbol{A}'$ the transpose of $\boldsymbol{A}$. The column vector space generated by the columns of $\boldsymbol{A}$ is written $\mathcal{C}(\boldsymbol{A})$ and its orthogonal complement $\mathcal{C}(\boldsymbol{A})^{\perp}$ (always assuming a standard inner product). Furthermore, let $\mathcal{C}(\boldsymbol{A}) \boxplus \mathcal{C}(\boldsymbol{B})$ denotes the orthogonal sum of the two vector spaces. Let $\boldsymbol{V}$ be a positive definite matrix. For projectors, the following notation is used: if $\boldsymbol{V} = \boldsymbol{I}$, i.e., the identity matrix, $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}\left(\boldsymbol{A}'\boldsymbol{A}\right)^{-}\boldsymbol{A}'$ and otherwise $\boldsymbol{P}_{\boldsymbol{A},\boldsymbol{V}} = \boldsymbol{A}\left(\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A}\right)^{-}\boldsymbol{A}'\boldsymbol{V}^{-1}$. Here $^{-}$ denotes an arbitrary generalized inverse. Furthermore, $\boldsymbol{Q}_{\boldsymbol{A}} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}}$, i.e., is the orthogonal projection on $\mathcal{C}(\boldsymbol{A})^{\perp}$. Lastly, in this paper, $\mathbf{1}_a$ stands for the column vector of a ones whereas $\mathbf{0}_a$ stands for the column vector of a zeroes.

This paper is organized as follows: in Sect. 2, the model discussed in this work is introduced, and the two residuals useful for evaluating the GMANOVA-MANOVA model are presented. Data analysis in Sect. 3 shows and motivates the idea that the residual, which is studied, is essential for the understanding of the model. Some concluding remarks are given in Sect. 4.

2 Background

This section comprises the model under consideration, together with the maximum likelihood estimators of the unknown parameters and expressions for residuals. The parametric bootstrap technique for obtaining simulated distributions of the "extreme" elements of the residual is also presented.

Residuals in the GMANOVA-MANOVA Model

The model considered in this paper is defined as follows:

Definition 1 (*GMANOVA-MANOVA model*) The model for the observation matrix $\boldsymbol{X}$: $p \times n$ is defined via

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}_1\boldsymbol{C}_1 + \boldsymbol{B}_2\boldsymbol{C}_2 + \boldsymbol{E}, \qquad \boldsymbol{E} \sim N_{p,n}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_n), \tag{1}$$

where $N_{p,n}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_n)$ the matrix normal distribution with independently distributed columns, $\boldsymbol{A}$: $p \times q$, is the within-individuals design matrix, $\boldsymbol{C}_1$: $k_1 \times n$, the between-individuals design matrix, $\boldsymbol{C}_2$: $k_2 \times n$, is a matrix including covariate information, continuous or discrete, $\boldsymbol{B}_i$, $i \in \{1, 2\}$, consist of the unknown mean parameters and $\boldsymbol{\Sigma}$ is an unknown positive definite dispersion matrix.

Remark 1 The expression $\boldsymbol{B}_2\boldsymbol{C}_2$ is the MANOVA part of the model and $\boldsymbol{A}\boldsymbol{B}_1\boldsymbol{C}_1$ is the GMANOVA part.

In this paper, we focus on the use of residual matrices for the GMANOVA-MANOVA model. The derivation and the interpretation of these residuals are detailed in [2, 3]. The expression of the residuals matrices follows from

$$\boldsymbol{X} - \boldsymbol{A}\widehat{\boldsymbol{B}}_1\boldsymbol{C}_1 - \widehat{\boldsymbol{B}}_2\boldsymbol{C}_2, \tag{2}$$

i.e., these residuals are obtained in the usual way as for example in linear models. Under full rank conditions on $\boldsymbol{A}$, $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ that is $r(\boldsymbol{A}) = q$, $r(\boldsymbol{C}_1) = k_1$ and $r(\boldsymbol{C}_2) = k_2$ and under the assumption that $\mathcal{C}(\boldsymbol{C}_1') \cap \mathcal{C}(\boldsymbol{C}_2') = \{\boldsymbol{0}\}$ the parameter estimators $\widehat{\boldsymbol{B}}_1$ and $\widehat{\boldsymbol{B}}_2$ in (2), are uniquely expressed as follows:

$$\widehat{\boldsymbol{B}}_1 = (\boldsymbol{A}'\boldsymbol{S}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{S}^{-1}\boldsymbol{X}\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'(\boldsymbol{C}_1\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1')^{-1}, \tag{3}$$

$$\widehat{\boldsymbol{B}}_2 = (\boldsymbol{X} - \boldsymbol{A}\widehat{\boldsymbol{B}}_1\boldsymbol{C}_1)\boldsymbol{C}_2'(\boldsymbol{C}_2\boldsymbol{C}_2')^{-1}, \tag{4}$$

where $\boldsymbol{S}$ is given by

$$\boldsymbol{S} = \boldsymbol{X}\boldsymbol{Q}_{\boldsymbol{C}_2'}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'})\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{X}', \tag{5}$$

and $\boldsymbol{Q}_{\boldsymbol{C}_2'} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{C}_2'}$. It is assumed that the inverse of $\boldsymbol{S}$ exists, which is true with probability 1 if parameters are estimable.

The estimated covariance matrix equals

$$n\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} + (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}})\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P}'_{\boldsymbol{A},\boldsymbol{S}}). \tag{6}$$

Moreover, the expression for the estimated model $\widehat{\boldsymbol{X}}$ in the GMANOVA-MANOVA model (1) is presented in Theorem 1.

Theorem 1 *Let $\widehat{\boldsymbol{B}}_1$ and $\widehat{\boldsymbol{B}}_2$ be the MLEs given in (3) and (4), respectively. Then the predicted values equal*

$$\widehat{\boldsymbol{X}} = \boldsymbol{A}_1\widehat{\boldsymbol{B}}_1\boldsymbol{C}_1 + \widehat{\boldsymbol{B}}_2\boldsymbol{C}_2 = \boldsymbol{X}\boldsymbol{P}_{\boldsymbol{C}_2'} + \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'}. \tag{7}$$

Remark 2 The relation in (7) holds also without full rank conditions.

It should be noted that

$$\boldsymbol{Q}_{\boldsymbol{C}_2'} = \boldsymbol{Q}_{\boldsymbol{C}_1':\boldsymbol{C}_2'} + \boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'}$$

which follows from the space decomposition $\mathcal{C}(\boldsymbol{C}_2')^{\perp} = \mathcal{C}(\boldsymbol{C}_1' : \boldsymbol{C}_2')^{\perp} \boxplus \mathcal{C}(\boldsymbol{C}_1' : \boldsymbol{C}_2') \cap \mathcal{C}(\boldsymbol{C}_2')^{\perp}$, and where $\mathcal{C}(\boldsymbol{C}_1' : \boldsymbol{C}_2') \cap \mathcal{C}(\boldsymbol{C}_2')^{\perp}$ is identical to $\mathcal{C}(\boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'})$. Moreover, (2) equals

$$\boldsymbol{X}\boldsymbol{Q}_{\boldsymbol{C}_1':\boldsymbol{C}_2'} + \boldsymbol{X}\boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'} - \boldsymbol{A}\widehat{\boldsymbol{B}}_1\boldsymbol{C}_1\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1' = \boldsymbol{X}\boldsymbol{Q}_{\boldsymbol{C}_1':\boldsymbol{C}_2'} + (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}})\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'}$$

and the next definition concerning residuals makes sense.

Definition 2 (*see* [2]) For the GMANOVA-MANOVA model given in (1), the following residuals are defined:

$$\begin{aligned}\boldsymbol{R}_1 &= \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{C}_1':\boldsymbol{C}_2'}) = \boldsymbol{R}_{11} + \boldsymbol{R}_{12},\\ \boldsymbol{R}_2 &= (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}})\boldsymbol{X}\boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'},\end{aligned}$$

where $\boldsymbol{R}_{11} = \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{C}_1':\boldsymbol{C}_2'})$ and $\boldsymbol{R}_{12} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}})\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{C}_1':\boldsymbol{C}_2'})$.

The residuals $\boldsymbol{R}_1$, $\boldsymbol{R}_{11}$, $\boldsymbol{R}_{12}$ and $\boldsymbol{R}_2$ are illustrated in Fig. 1 where a decomposition of the tensor space connected to the model is shown and illustrates the model and its residuals. Note that since $\boldsymbol{S}$ is included in the projection $\boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}$ the distribution for $\boldsymbol{R}_2$ it is complicated and is not available in a simple form.

Residual $\boldsymbol{R}_1$ can be utilized to indicate if individual observations deviate from the rest of the observations and will not be considered in this work. Instead the focus will be on residual $\boldsymbol{R}_2$, which is suitable for checking assumption about the mean structure (the model). Moreover, it is possible to calculate the mean and dispersion

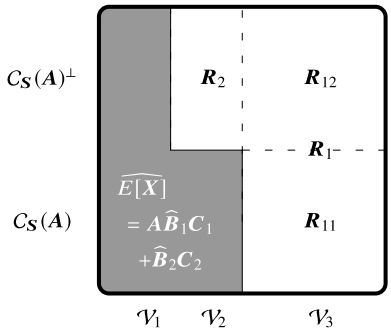


Fig. 1 The model in Definition 1 is illustrated where the mean and the residual spaces are shown. Spaces connected to the between-individual decomposition are given by, $\mathcal{V}_1 = \mathcal{C}(\boldsymbol{C}_2')$, $\mathcal{V}_2 = \mathcal{C}(\boldsymbol{C}_2')^\perp \cap \mathcal{C}(\boldsymbol{C}_1' : \boldsymbol{C}_2')$ and $\mathcal{V}_3 = \mathcal{C}(\boldsymbol{C}_1' : \boldsymbol{C}_2')^\perp$. The residuals are introduced in Definition 2

for the residuals. It can also be noted that the residuals are uncorrelated with the estimated mean, $\widehat{\boldsymbol{X}} = \boldsymbol{A}\widehat{\boldsymbol{B}}_1\boldsymbol{C}_1 + \widehat{\boldsymbol{B}}_2\boldsymbol{C}_2$, but the residuals are not independently distributed with the estimated mean. For details about these facts see [2]. To evaluate the distribution for "large" elements of $\boldsymbol{R}_2$, we are suggesting to generate parametric bootstrap samples to study the distribution of the "largest" elements of the residual.

The GMANOVA-MANOVA Model and the Parametric Bootstrap Technique

The bootstrap philosophy was introduced in 1979 by Efron [7]. Over the years, many authors have continued to develop the approach. For nice introductions and overviews, see for example [6, 8]. In this paper, we use a technic often called parametric bootstrap to provide the distributions for the "largest" elements of the residual $\boldsymbol{R}_2$ in the GMANOVA-MANOVA model. Parametric bootstrap has been applied in many non-standard situations (see, e.g., [4]). When studying the "largest" elements of $\boldsymbol{R}_2$, the problem is that we study extreme values of dependent and non-identical observations.

For parametric bootstrap, the underlying distribution assumption for the data is taken into account. This is not the case for the usual bootstrap approach that is completely based on the observations at hand. In fact, parametric bootstrap assumes that the data come from a known class of distributions. In our case, as it is in many applications, the distribution involves unknown parameters which have to be estimated from a given data set. The "estimated distribution" is then used to simulate the "bootstrap" samples.

In this paper, we use the parametric bootstrap technique to obtain an approximation of the distribution for the three "largest" values in $\boldsymbol{R}_2$ and define cut off points so that with high probability a certain value is extreme. The approach runs as follows: Given a data set and using (3)–(6) maximum likelihood estimates $\widehat{\boldsymbol{B}}_1$, $\widehat{\boldsymbol{B}}_2$ and $\widehat{\boldsymbol{\Sigma}}$ are obtained; then random numbers which are elements of $\boldsymbol{E}^* \sim N_{p,n}(\boldsymbol{0}, \widehat{\boldsymbol{\Sigma}}, \boldsymbol{I}_n)$ are generated m times (m is large) and for each m,

$$\boldsymbol{X}^* = \boldsymbol{A}\widehat{\boldsymbol{B}}_1\boldsymbol{C}_1 + \widehat{\boldsymbol{B}}_2\boldsymbol{C}_2 + \boldsymbol{E}^* \tag{8}$$

is calculated; thereafter, using Definition 2, $\boldsymbol{R}_2^* = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}})\boldsymbol{X}^*\boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'}$ is constructed. To indicate that the process is repeated m times it is written

$$\boldsymbol{R}_{2i}^* = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}})\boldsymbol{X}^*\boldsymbol{P}_{\boldsymbol{Q}_{\boldsymbol{C}_2'}\boldsymbol{C}_1'}, \quad i \in \{1, \ldots, m\}. \tag{9}$$

We are interested in "large" values of the residual and the distributions of the three "largest" residuals are of interest.

Definition 3 Define the largest by the absolute value of $\boldsymbol{R}_2$ by ${}_1R_2$, the second "largest" by ${}_2R_2$ and the third "largest" by ${}_3R_2$, and the largest by absolute value of $\boldsymbol{R}_{2i}^*$ by ${}_1R_{2i}^*$, the second "largest" by ${}_2R_{2i}^*$ and the third "largest" by ${}_3R_{2i}^*$.

Moreover, cut-off points for identifying extreme residuals in these statistics are obtained through quantiles (95%, 99%) of the estimated distribution.

The proposed parametric bootstrap approach is described in the below given algorithm.

Algorithm 1 Parametric bootstrap procedure

1: Set m to be the number of required bootstrap samples.
2: For a given data set compute the parameter estimates $\widehat{\boldsymbol{B}}_1$, $\widehat{\boldsymbol{B}}_2$ and $\widehat{\boldsymbol{\Sigma}}$, presented in (3)–(6).
3: For $i \in \{1, \ldots, m\}$ sample $\boldsymbol{E}^*$ from $N_{p,n}(\boldsymbol{o}, \widehat{\boldsymbol{\Sigma}}, \boldsymbol{I}_n)$ and compute $\boldsymbol{X}^*$ according to (8).
4: Compute the residual components ${}_jR_{2i}^*$, $j \in \{1, 2, 3\}$, $i \in \{1, \ldots, m\}$, via (9).

3 Data Analysis

In this section, we will study the distribution of the three largest residuals by absolute value of $\boldsymbol{R}_2$ for a realistic but artificial data set, which consists of 69 patients suffering from a disease. The data set, given in the Appendix (see Table 3)[1], is also considered by [3] and it is similar to some real studies which have taken place (see [9, 11]).

To initiate the parametric bootstrap, it is needed to estimate $\boldsymbol{B}_1$, $\boldsymbol{B}_2$ and $\boldsymbol{\Sigma}$ when the GMANOVA-MANOVA model is applied, and in the study, participants were

[1] Both data and code for this data analysis can be found at https://gitlab.liu.se/maroh70/on-the-identification-of-extreme-elements-in-a-residual-for-the-gmanova-manova-model.

assumed to be randomly assigned to one out of two different treatment groups. All participants in Group 1 received treatment one whereas those in Group 2 received treatment two. For each participant, a measure of estimated autoimmunity (EA) was obtained at each clinical visits. EA is the response that is to be analyzed. The clinical visits took place at baseline (time 0, initiation of the treatment) 6, 12 and 18 months after initiation. The disease is expected to have a negative effect on EA. For each participant, it was recorded whether the patient had been previously treated for the disease. Participants in Group 1 who were not treated previously constitute Group 1a, whereas the remaining ones constituted Group 1b. Similarly, we have Group 2a and Group 2b. Thus, there are four groups of individuals that are considered. To analyze the data with the GMANOVA-MANOVA model given in (1), the following design matrices will be used:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 1 & 6 \\ 1 & 12 \\ 1 & 18 \end{pmatrix}, \quad \boldsymbol{C}_1 = \begin{pmatrix} \mathbf{1}'_{35} & \mathbf{0}'_{34} \\ \mathbf{0}'_{35} & \mathbf{1}'_{34} \end{pmatrix}, \quad \boldsymbol{C}_2 = \begin{pmatrix} \mathbf{0}'_{13} & \mathbf{1}'_{22} & \mathbf{0}'_{10} & \mathbf{1}'_{24} \end{pmatrix}. \tag{10}$$

The covariate effect $\boldsymbol{B}_2\boldsymbol{C}_2$ with the above choice of $\boldsymbol{C}_2$ implies that the effect from previous treatment is the same for both treatment groups. Other choices can also be used, e.g.,

$$\boldsymbol{C}_2 = \left(\mathbf{1}'_{13} \otimes \begin{pmatrix} -1 \\ 0 \end{pmatrix} \mathbf{1}'_{22} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mathbf{1}'_{10} \otimes \begin{pmatrix} 0 \\ -1 \end{pmatrix} \mathbf{1}'_{24} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right),$$

which yields a different interpretation of the covariate effect than when using $\boldsymbol{C}_2$ in (10). However, the main purpose with this article is to discuss the residual $\boldsymbol{R}_2$ and not different choices of $\boldsymbol{C}_2$, which has been discussed in [3]. Therefore only the design matrices in (10) will be considered.

Using the data given in the Appendix, Table 3, the maximum likelihood estimates (3)–(6) equal

$$\widehat{\boldsymbol{B}}_1 = \begin{pmatrix} 12.6 & 12.5 \\ -0.17 & -0.22 \end{pmatrix}, \quad \widehat{\boldsymbol{B}}_2 = \begin{pmatrix} 0.93 \\ 1.12 \\ 1.00 \\ 1.24 \end{pmatrix},$$

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 4.71 & 1.78 & 1.93 & 1.53 \\ & 4.84 & 2.98 & 1.55 \\ & & 5.44 & 2.12 \\ & & & 4.80 \end{pmatrix}. \tag{11}$$

Fitted "growth curves" can be presented through $\boldsymbol{A}\widehat{\boldsymbol{B}}_1\boldsymbol{c}_1 + \widehat{\boldsymbol{B}}_2 c_2$ where $\boldsymbol{c}_1$ is a vector which equals either $\binom{1}{0}$ or $\binom{0}{1}$ and c_2 equals either 1 or 0. By choosing $\boldsymbol{c}_1$ and c_2 the "growth curves" can be expressed for each of the four groups. Fitted "growth

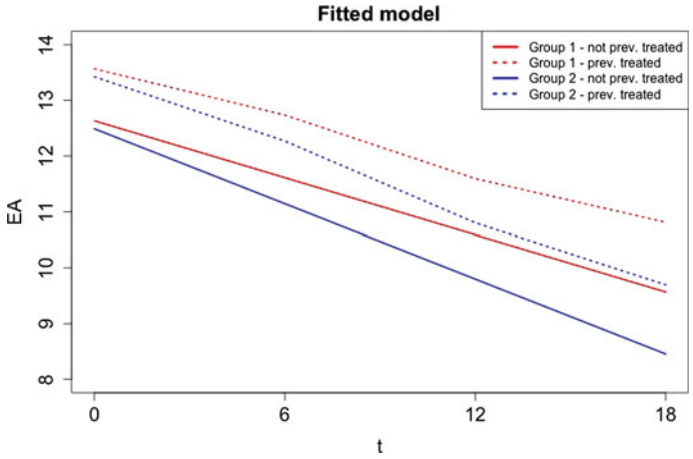


Fig. 2 Fitted growth curves for the GMANOVA-MANOVA model defined via (10). The previously untreated subgroups are modeled as baselines with linear growths given by $\widehat{\boldsymbol{B}}_1$ in (11), and the previously treated subgroups are equally shifted from the baselines through $\widehat{\boldsymbol{B}}_2$ presented in (11)

curves" for the four groups are shown in Fig. 2. The broken lines show the pretreated groups of individuals and a small effect of pre-treatment can be seen. In the model, the previously untreated subgroups act as baselines with a linear growth $\widehat{\boldsymbol{B}}_1$ and the previously treated subgroups are shifted from the baseline groups by $\widehat{\boldsymbol{B}}_2$. There is a clear decreasing trend of EA in each subgroup.

The residual $\boldsymbol{R}_1$, given in Definition 2, is shown in Fig. 3. This figure shows that no one of the individuals is far away of the group means. Thus, there is no presence of "outliers" in the given data.

The main focus in this paper is, however, on the residual elements in $\boldsymbol{R}_2$. Numerically, the residual $\boldsymbol{R}_2$ equals

$$\boldsymbol{R}_2 = \left(\mathbf{1}'_{13} \otimes \begin{pmatrix} 0.074 \\ -0.12 \\ -0.014 \\ 0.032 \end{pmatrix} : \mathbf{1}'_{22} \otimes \begin{pmatrix} -0.088 \\ 0.14 \\ -0.051 \\ -0.01161 \end{pmatrix} : \mathbf{1}'_{10} \otimes \begin{pmatrix} 0.24 \\ -0.39 \\ 0.083 \\ 0.055 \end{pmatrix} : \mathbf{1}'_{24} \otimes \begin{pmatrix} 0.081 \\ -0.13 \\ 0.046 \\ 0.011 \end{pmatrix} \right). \tag{12}$$

The residuals $\boldsymbol{R}_2$ are shown in Fig. 4, and all residual values are between -0.39 and 0.24. The main question is at what level a deviating residual in Fig. 4 can be acceptable. To answer the question, the residual distribution of the three "largest" elements of $\boldsymbol{R}_2$, i.e., ${}_j R_{2i}$, $j \in \{1, 2, 3\}$, given in Definition 3, is explored.

Our approach is to investigate the distribution of the three largest elements of the matrix $\boldsymbol{R}_2$ through 10,000 parametric bootstrap samples obtained using Algorithm 1 in Sect. 2. The complete marginal bootstrap distributions of the "largest" elements in

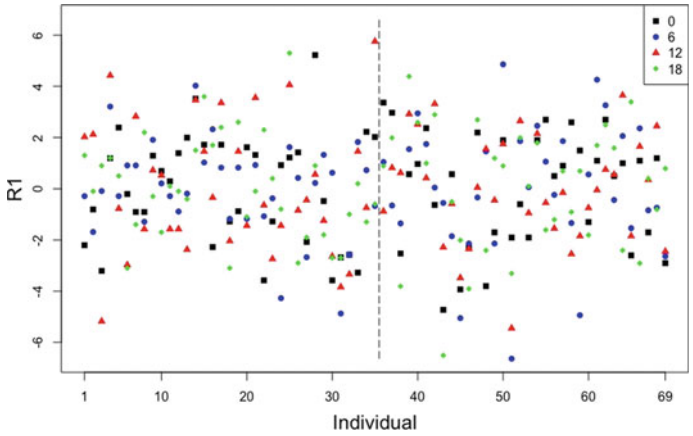


Fig. 3 The residuals $\boldsymbol{R}_1$, introduced in Definition 2, are presented for the model defined via (10). For each individual, the four time points are presented in different colors

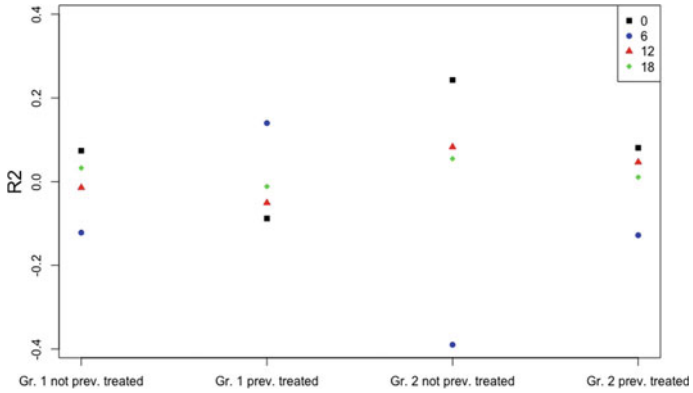


Fig. 4 All elements in residual $\boldsymbol{R}_2$ given in (12) are presented. Note that the elements are identical within each subgroup

$\boldsymbol{R}_2$ are presented in Fig. 5. Moreover, in Table 1, the 95% and 99% quantiles for the marginal parametric bootstrap estimated distribution of the "largest" elements in $\boldsymbol{R}_2$ are presented. It follows that ${}_1R_2 = 0.39$, ${}_2R_2 = 0.24$ and ${}_3R_2 = 0.14$ are below the thresholds defined via the 95% and 99% percentiles of the bootstrap distributions. Thus, there are no indications that the model does not fit.

Finally, it will be shown that the presented approach has the potential to indicate when a model does not fit data. To see this, a synthetic data set is generated that is presented in the Appendix, Table 4. Moreover, Fig. 6 illustrates the mean structure used to generate the data, and it can be seen that Group 1 follows a quadratic mean model. All groups will now be modeled under the assumption of a linear mean

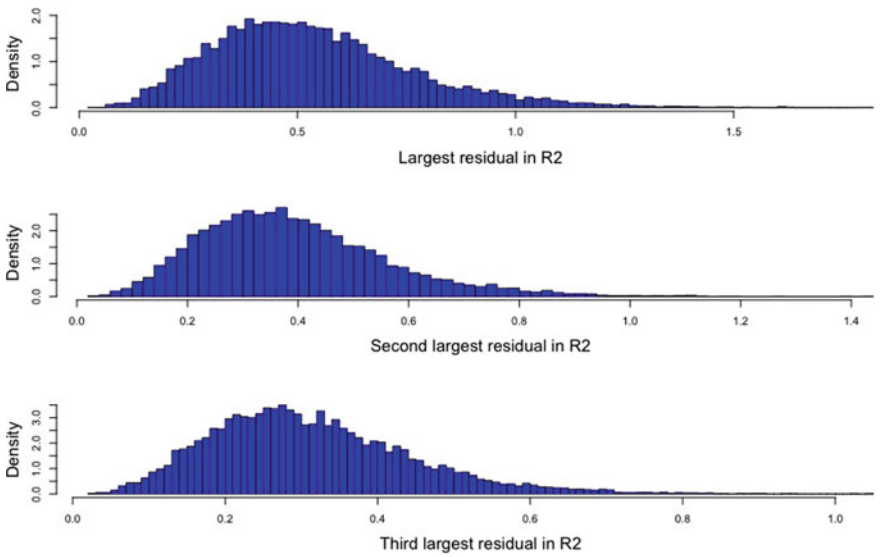


Fig. 5 The approximative distribution for the three "largest" elements for $\boldsymbol{R}_2$ in (12), obtained when 10,000 bootstrap samples have been generated

Table 1 The estimated percentiles (95% and 99%) for the three largest residual elements (in absolute values) in the residual $\boldsymbol{R}_2$

Percentile (%)	${}_1R^*_{2i}$	${}_2R^*_{2i}$	${}_3\boldsymbol{R}^*_{2i}$
99	1.2042	0.8775	0.6950
95	0.9641	0.7035	0.5514

structure, where the design matrices are given in (10). The following maximum likelihood estimates are obtained:

$$\widehat{\boldsymbol{B}}_1 = \begin{pmatrix} 15.9 & 12.71 \\ -0.34 & -0.24 \end{pmatrix}, \quad \widehat{\boldsymbol{B}}_2 = \begin{pmatrix} 0.09 \\ -1.3 \\ -1.2 \\ 2.18 \end{pmatrix},$$
$$\widehat{\Sigma} = \begin{pmatrix} 3.42 & -1.19 & -0.61 & 0.42 \\ & 14.83 & 11.05 & -1.23 \\ & & 12.65 & 0.03 \\ & & & 4.20 \end{pmatrix}. \tag{13}$$

Moreover, the residual $\boldsymbol{R}_2$ is given by

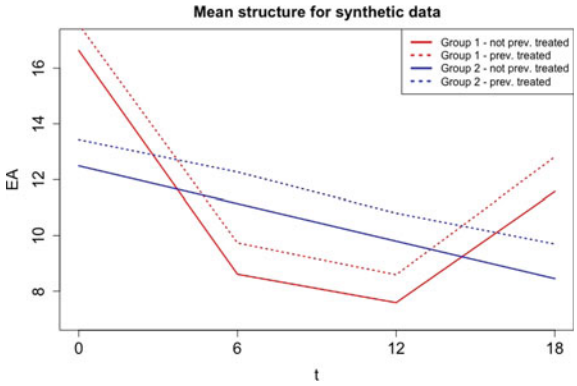


Fig. 6 Illustration of the GMANOVA-MANOVA model where Group 1 follows a quadratic mean model and Group 2 follows a linear model

$$\boldsymbol{R}_2 = \left(\mathbf{1}'_{13} \otimes \begin{pmatrix} 1.26 \\ -5.77 \\ -4.37 \\ 1.09 \end{pmatrix} : \mathbf{1}'_{22} \otimes \begin{pmatrix} 0.46 \\ -2.33 \\ -2.05 \\ 0.49 \end{pmatrix} : \mathbf{1}'_{10} \otimes \begin{pmatrix} 0.38 \\ -1.30 \\ -0.45 \\ 0.16 \end{pmatrix} : \mathbf{1}'_{24} \otimes \begin{pmatrix} -0.42 \\ 2.14 \\ 1.88 \\ -0.45 \end{pmatrix} \right). \tag{14}$$

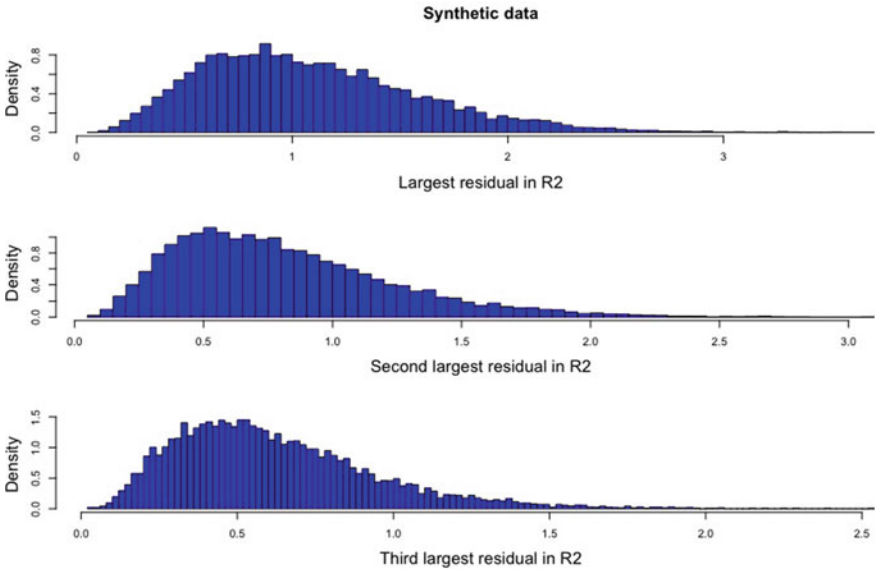


Fig. 7 The distribution for the three "largest" elements for $\boldsymbol{R}_2$ obtained after 10,000 bootstrap samples when Group 1 follows a quadratic mean

Table 2 The estimated percentiles (95% and 99%) for the three largest residual elements (in absolute values) of the residual $\boldsymbol{R}_2$ in (14)

Percentile (%)	${}_1R^*_{2i}$	${}_2R^*_{2i}$	${}_3\boldsymbol{R}^*_{2i}$
99	2.4923	2.0744	1.6025
95	2.0222	1.6272	1.2276

With the help of the estimates in (13) for ${}_jR_2$, $j \in \{1, 2, 3\}$, a bootstrap sample of 10, 000 observations is generated. The bootstrap distributions are presented in Fig. 7. Correspondingly to Table 1, in Table 2, the 95% and 99% quantiles for the parametric bootstrap estimated marginal distribution of the "largest" elements in $\boldsymbol{R}_2$ are given.

From (14), it follows that ${}_1R_2 = 5.77$, ${}_2R_2 = 4.37$ and ${}_3R_2 = 2.33$. Hence since all these values are above the given quantiles in Table 2, this indicates that the linear model that is used in the analysis is not correct, which is exactly how it should be.

4 Concluding Remarks

The elements of the residual $\boldsymbol{R}_2$ in the GMANOVA-MANOVA model under consideration are complicated functions in the observations. Since the interest is in the largest (by absolute value) elements of $\boldsymbol{R}_2$, it is impossible to find the exact distributions for these elements. Therefore, the parametric bootstrap approach has been applied. It was chosen to consider the three "largest" elements in $\boldsymbol{R}_2$. This is of course an arbitrary approach and most often only the "largest" value is considered. The three values provide more information than only one value but we would come in an unpleasant situation if only the second "largest" observation would be deemed to be an outlier but not the "largest" element of $\boldsymbol{R}_2$. Finally, it is noted that the residual $\boldsymbol{R}_2$ showed to be a useful matrix, i.e., the values of $\boldsymbol{R}_2$ became large only when the model in the analysis deviated from the true model.

Acknowledgements The authors would like to thank the two anonymous referees for the comments that improved the quality and presentation of this paper.
The research of Béatrice Byukusenge has been supported by the sub-program of Applied Mathematics and Statistics under the Sida-funded bilateral program, The University of Rwanda-Sweden Programme for Research, Higher Education and Institutional Advancement. Dietrich von Rosen is supported by the Swedish Research Council (2017-03003).

Appendix

In the Appendix, the data sets that are used in the analysis are presented (Tables 3 and 4).

Table 3 Repeated measurements for individuals, together with the covariates 'Previously treated' (see also [3])

	Clinical visits					Previously
Id	0	6	12	18	Group	Treated
1	10.50	11.20	12.60	10.90	1	0
2	11.90	9.80	12.70	9.50	1	0
3	9.50	11.40	5.40	10.50	1	0
4	13.90	14.70	15.00	10.80	1	0
5	15.10	11.20	9.80	10.10	1	0
6	12.50	12.40	7.60	6.50	1	0
7	11.80	12.40	13.40	8.20	1	0
8	11.80	10.20	9.00	11.80	1	0
9	14.00	13.40	11.30	9.30	1	0
10	13.40	11.70	11.10	7.90	1	0
11	13.00	11.20	9.00	9.70	1	0
12	14.10	10.60	9.00	9.50	1	0
13	14.70	11.30	8.20	9.20	1	0
14	17.00	16.90	15.00	12.30	1	1
15	15.20	13.90	13.00	14.40	1	1
16	11.20	15.20	11.20	12.50	1	1
17	15.20	13.70	14.90	13.20	1	1
18	12.20	11.70	9.50	7.70	1	1
19	12.60	13.70	13.00	13.40	1	1
20	15.10	11.70	10.10	9.70	1	1
21	14.80	13.80	15.10	10.70	1	1
22	9.90	11.80	10.90	13.10	1	1
23	12.20	12.50	8.80	11.20	1	1
24	14.40	8.60	10.10	10.00	1	1
25	14.70	14.50	15.60	16.10	1	1
26	14.90	13.30	10.70	7.90	1	1
27	11.40	10.20	11.10	8.90	1	1
28	18.70	13.10	12.10	11.70	1	1
29	13.00	14.20	10.30	9.00	1	1
30	9.90	13.50	8.90	8.10	1	1
31	10.80	8.00	7.70	8.10	1	1
32	10.90	10.30	8.20	9.80	1	1
33	10.20	14.70	13.00	11.00	1	1
34	15.70	13.60	10.80	9.50	1	1
35	15.50	12.20	17.30	10.20	1	1

(continued)

Table 3 (continued)

	Clinical visits					Previously
Id	0	6	12	18	Group	Treated
36	16.10	11.80	9.00	9.40	2	0
37	15.70	10.10	10.70	10.50	2	0
38	10.20	9.40	10.50	4.70	2	0
39	13.30	12.30	12.80	12.90	2	0
40	13.70	13.70	12.40	11.10	2	0
41	15.10	12.50	10.30	9.50	2	0
42	12.10	10.80	13.20	11.40	2	0
43	8.00	10.20	7.60	2.00	2	0
44	13.30	8.90	9.30	8.00	2	0
45	8.80	5.70	6.40	6.50	2	0
46	11.20	10.00	8.50	5.80	2	1
47	15.70	11.80	10.90	12.40	2	1
48	9.70	13.60	12.40	7.30	2	1
49	11.80	10.00	10.40	10.90	2	1
50	15.40	17.00	12.60	10.60	2	1
51	11.60	5.50	5.40	6.40	2	1
52	12.90	14.00	13.50	11.70	2	1
53	11.60	12.20	9.90	9.80	2	1
54	15.40	14.60	13.00	11.50	2	1
55	16.20	13.20	10.30	8.10	2	1
56	14.00	11.90	9.30	8.50	2	1
57	14.40	14.00	10.70	10.40	2	1
58	16.10	10.80	8.30	8.80	2	1
59	15.00	7.20	9.00	10.40	2	1
60	12.20	12.70	10.10	7.90	2	1
61	14.60	16.40	10.80	11.40	2	1
62	16.20	15.40	11.60	12.20	2	1
63	14.00	11.70	11.40	11.30	2	1
64	14.50	14.20	14.50	7.30	2	1
65	10.90	10.60	9.00	13.10	2	1
66	14.60	14.50	12.50	6.80	2	1
67	11.80	11.30	11.20	10.10	2	1
68	14.70	11.40	13.30	8.90	2	1
69	10.60	9.50	8.40	10.50	2	1

Table 4 Repeated measurements for individuals, together with the covariates 'Previously treated'. Data for Group 1 follows a quadratic model

	Clinical visits					Previously
Id	0	6	12	18	Group	Treated
1	19.86	7.96	9.81	11.30	1	0
2	18.32	9.18	12.51	11.98	1	0
3	18.32	8.57	6.15	12.73	1	0
4	18.11	5.69	5.96	12.22	1	0
5	16.65	9.63	10.30	9.72	1	0
6	15.41	4.81	2.93	12.52	1	0
7	15.57	6.75	5.24	12.30	1	0
8	19.30	11.21	6.98	11.52	1	0
9	16.53	10.24	8.38	8.42	1	0
10	16.47	8.07	8.27	9.49	1	0
11	15.21	9.98	7.09	12.19	1	0
12	14.62	5.75	6.31	10.86	1	0
13	16.86	7.81	7.96	11.62	1	0
14	16.95	13.28	12.27	10.97	1	1
15	16.29	10.84	13.73	15.80	1	1
16	18.72	11.45	8.92	12.53	1	1
17	17.26	6.93	7.63	9.46	1	1
18	13.34	11.19	8.82	11.69	1	1
19	16.57	12.15	8.99	14.85	1	1
20	16.53	13.21	8.81	9.37	1	1
21	19.64	13.84	10.15	13.30	1	1
22	18.34	7.68	3.43	14.14	1	1
23	16.43	11.32	10.15	12.52	1	1
24	17.47	10.80	6.15	11.01	1	1
25	17.27	9.10	6.01	10.26	1	1
26	19.12	10.94	6.88	13.16	1	1
27	13.97	7.83	5.65	13.10	1	1
28	14.64	15.29	12.41	12.81	1	1
29	15.77	8.10	8.20	15.79	1	1
30	16.61	8.88	9.45	10.51	1	1
31	15.25	7.17	3.28	6.49	1	1
32	16.99	10.56	10.59	13.20	1	1
33	15.54	7.61	7.19	12.43	1	1
34	18.10	8.74	6.53	9.76	1	1
35	14.84	7.65	9.72	11.27	1	1

(continued)

Table 4 (continued)

	Clinical visits					Previously
Id	0	6	12	18	Group	Treated
36	12.47	12.95	12.42	11.35	2	0
37	13.74	5.05	7.51	8.77	2	0
38	14.62	9.04	5.90	8.55	2	0
39	13.58	12.86	11.43	5.90	2	0
40	12.82	9.50	6.26	7.13	2	0
41	17.17	12.13	11.26	8.15	2	0
42	13.34	11.07	11.20	9.74	2	0
43	12.64	6.99	6.32	4.05	2	0
44	12.62	11.33	11.66	8.83	2	0
45	10.56	7.89	7.33	4.84	2	0
46	12.19	10.21	6.26	7.68	2	1
47	12.49	12.32	7.81	9.14	2	1
48	8.90	10.09	9.94	11.04	2	1
49	11.14	9.51	7.87	8.48	2	1
50	13.92	10.90	10.80	10.87	2	1
51	15.30	10.20	17.32	9.61	2	1
52	9.99	8.47	8.62	10.14	2	1
53	10.23	13.11	10.64	8.37	2	1
54	13.31	11.39	10.53	8.95	2	1
55	11.77	12.25	12.63	14.19	2	1
56	10.43	13.32	12.03	11.56	2	1
57	12.33	11.21	8.75	9.45	2	1
58	14.57	12.80	12.35	8.40	2	1
59	13.32	12.82	8.54	9.46	2	1
60	13.22	12.81	10.50	11.50	2	1
61	10.74	11.30	7.02	7.63	2	1
62	9.51	14.12	9.53	10.63	2	1
63	13.93	7.57	7.98	10.93	2	1
64	10.87	14.14	12.93	12.75	2	1
65	9.35	11.45	9.41	12.47	2	1
66	13.68	14.80	9.45	10.79	2	1
67	14.18	16.47	12.95	9.86	2	1
68	15.93	16.45	14.02	9.19	2	1
69	13.54	12.88	13.95	14.42	2	1

References

1. Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.
2. Byukusenge, B., von Rosen, D. & Singull, M. (2021). On residual analysis in the GMANOVA-MANOVA model. (submitted).
3. Byukusenge, B., von Rosen, D. & Singull, M. (2021). On an important residual in the GMANOVA-MANOVA model. (submitted).
4. Cheng, R. (2017). *Non-Standard Parametric Statistical Inference*. Oxford: Oxford University Press.
5. Chinchilli, V. M., & Elswick, R. K. (1985). A mixture of the MANOVA and GMANOVA models. *Communications in Statistics-Theory and Methods, 14,* 3075–3089.
6. Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
7. Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals Statistics, 7,* 1–26.
8. Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability (Vol. 57). New York: Chapman and Hall.
9. Ellison, G. W., Myers, L. W., Mickey, M. R., Graves, M. C., Tourtellotte, W. W., Syndulko, K., et al. (1989). A placebo-controlled, randomized, double-masked, variable dosage, clinical trial of azathioprine with and without methylprednisolone in multiple sclerosis. *Neurology, 39,* 1018–1026.
10. Hamid, J. S., & von Rosen, D. (2006). Residuals in the extended growth curve model. *Scandinavian Journal of Statistics, 33,* 121–138.
11. Heitjan, D. F. (1991). Nonlinear modeling of serial immunologic data: A case study. *Journal of the American Statistical Association, 86,* 891–898.
12. Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika, 51,* 313–326.
13. von Rosen, D. (1995). Residuals in the Growth Curve model. *Annals of the Institute of Statistical Mathematics, 47,* 129–136.
14. von Rosen, D. (2018). *Bilinear Regression Analysis: An Introduction*. Lecture Notes Statistics (Vol. 220). New York: Springer.
15. Srivastava, M. S., & Khatri, C. G. (1979). *An Introduction to Multivariate Statistics*. New York: North-Holland.

Matrix-variate Smooth Transition Models for Temporal Networks

Monica Billio, Roberto Casarin, Michele Costola, and Matteo Iacopini

Abstract In many fields, network analysis is used to investigate complex relationships. The increased availability of temporal network data opens the way to the statistical analysis of the network topology and its dynamics. In addition network data are subject to measurement errors and random fluctuations. This calls for realistic time series models which account for relevant features of the data. In this chapter, we propose a new modeling and inference framework for studying matrix-valued panel data characterized by nonlinear dynamics and heavy tails. We assume a smooth transition model for the dynamics and a matrix-variate t distribution for the error term and show how the model can be used in temporal network analysis. Some properties of the model including the close-form expression for the predictor are given. We adopt a Bayesian approach to inference and design an efficient Markov chain Monte Carlo algorithm for approximating the posterior distribution. We apply the proposed model to a volatility network among European firms and an international oil production network and show its ability to account for structural changes. Our framework is motivated by temporal network data, nevertheless, it is general and can be of interest to all researchers interested in the analysis of matrix-variate time series.

M. Billio (✉) · R. Casarin · M. Costola
Ca' Foscari University of Venice, Fondamenta San Giobbe, 873, 30121 Venice, Italy
e-mail: billio@unive.it

R. Casarin
e-mail: r.casarin@unive.it

M. Costola
e-mail: michele.costola@unive.it

M. Iacopini
Queen Mary University of London, 327 Mile End Road, London E1 4NS, UK
e-mail: m.iacopini@qmul.ac.uk

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_7

1 Introduction

Scientific applications more and more involve datasets where the sampling unit is in the form of a matrix instead of a vector. Examples of matrix-valued time series include financial variables, such as balance sheet data for a set of firms; economic indicators, including GDP, industrial production, and unemployment rate, for a group of countries. The two motivating examples considered in this chapter are from the financial and commodity literature, where time series of networks are used to encode the connectivity structure of the markets. The sequence of networks can be represented as a collection of adjacency matrices. In the financial application, an element of a matrix describes a link between two financial entities, such as firms, institutions, or financial assets. Instead, in the energy application, an element of an adjacency matrix describes a link between two oil producers, such as countries or private companies.

Matrix-valued data has received particular attention in statistics and, more recently, in econometrics. An excellent review of the most frequently used matrix-variate distributions and their properties is provided in [45]. Regression models for matrix-valued data focused on particular aspects including temporal persistence [25], fat-tails innovation [11], missing data [73], hierarchical models [52], matrix factor models [24, 74, 77], and graphical models [17, 53, 75]. It is widely recognized that economic and financial time series exhibit abrupt changes, especially during downturns and crisis periods. To account for these stylized facts, nonlinear models with time-varying parameters, such as smooth transition and regime-switching dynamics, have been introduced [31, 46]. Most of the existing approaches focus on vector-valued data or vectorized matrix data.

In this chapter, we propose a new class of nonlinear models for matrix-valued panel data. We assume smooth transition (STR) dynamics in the conditional mean of a matrix regression model (STR-MAR) to account for regimes and structural changes in the data. In addition, we assume heavy-tailed innovations to account for outliers and departure from normality in the data due to excess kurtosis. STR models involve the change from one linear regime to the other [7]. The STR models have been successfully applied in several fields, including economics and finance, to deal with nonlinearities and parameter instability [e.g. 55, 70]. Several extensions have been provided for investigating panel data [43], and to account for dependence on past realizations, leading to smooth transition autoregressive models (STAR) [27, 33, 69]. See also [31, 72] for a review. Other extensions have been proposed to allow for a smooth transition in higher order conditional moments [4, 20, 21, 37, 44, 58, 60]. STR models have been also extended to the multivariate setting with many applications in macroeconomics and finance [16, 29, 59, 66, 67]. Our STR-MAR models extend multivariate smooth transition models to the matrix-variate setup. We show how STR-MAR models can be used in network modeling and provide some theoretical properties such as analytical expression for the optimal predictor of the network degree.

We adopt a Bayesian approach to inference since it accounts for extra-sample information and for soft parameter restrictions through the specification of the prior distribution as argued in [40, 54, 56, 57, 63] and in the STAR models literature [22, 42]. Also, the Bayesian approach allows us to make the likelihood function more tractable by following a data augmentation approach. We exploit the scale mixture representation of the matrix t distribution [45] and rely on data augmentation to obtain closed-form full conditional distributions for the locations and scale matrices and to design an efficient Markov chain Monte Carlo algorithm for approximating the posterior distribution.

As an illustration, we provide two applications where we investigate the dynamics of networks arising in two areas of economics, that are commodity and financial markets. The number of theoretical [e.g., 1, 34] and empirical studies [e.g., 12, 30] on financial networks has increased over the last decade. We apply the proposed STR-MAR model to investigate the volatility spill-over effects in the stock returns of the 50 most capitalized European firms. Starting with the global financial crisis and the European sovereign debt crisis, researchers and policy authorities have focused on the study of the (in)stability within the financial and banking sectors. During the outbreak of COVID-19, central banks re-introduced asset purchase programs to support the stability of the financial system. The analysis of the financial connectedness has provided insights into policymakers on the effect of COVID-19 on the system [10].

In the application to the commodity market, we consider the international oil production of 17 major oil-producing countries. As argued in [3, 18] the analysis of linkages among oil-producing countries allows for unveiling the micro-economic relationships between producers and the developments in the global oil production. Our model can be used to identify and predict the regimes of connectivity and, consequently, to hedge geographical or geopolitical risks affecting global oil production.

The proposed model detects structural changes in both applications in correspondence to well-defined events (i.e., the outbreak of COVID-19 and the oil cut production). We remark that our approach is general and can be applied in any analysis involving matrix-variate data.

The remaining of the chapter is structured as follows. Section 2 introduces a novel econometric framework for matrix-valued panel data. Section 3 presents a Bayesian inference procedure. Section 4 illustrates the empirical analysis and the main results. Section 5 concludes.

2 A Smooth Transition Matrix Model

Transition Mechanisms

Let $Y_t \in \mathbb{R}^{n\times m}$, $t \in \mathbb{N}$, be a sequence of $n \times m$ random matrices with real-valued entries and let $\Phi_j : \mathbb{R}^k \times D \mapsto (0,1)^{n\times m}$, $j = 1, \ldots, J$, be a sequence of smooth transition matrix functions $\Phi_j(\mathbf{x}; \mathbf{d}) \in (0,1)^{n\times m}$ with transition variables $\mathbf{x} \in \mathbb{R}^k$

and transition parameters $\mathbf{d} \in D$, satisfying $\Phi_1(\mathbf{x}; \mathbf{d}) + \ldots + \Phi_J(\mathbf{x}; \mathbf{d}) = \boldsymbol{\iota}_n \boldsymbol{\iota}_m'$ for all $\mathbf{x} \in \mathbb{R}^k$ and $\mathbf{d} \in D$, where $\boldsymbol{\iota}_r = (1, \ldots, 1)' \in \mathbb{R}^r$ denotes the unit vector.

Our STR-MAR model with J regimes is defined as follows:

$$Y_t = \sum_{j=1}^{J} \Phi_j(\mathbf{x}_t; \mathbf{d}) \circ B_j + E_t, t \in \mathbb{N} \tag{1}$$

where $B_1, \ldots, B_J \in \mathbb{R}^{n \times m}$ is a sequence of regime-specific matrices, $\circ$ denotes the element-by-element Hadamard's product and E_t is a matrix-variate error term. The class of models is very general and includes the following transition functions as a special case:

(a) row-specific transition, where the transition functions are the same across the columns of Y_t and differ cross rows, that is the transition functions are $\Phi_j(\mathbf{x}_t; \mathbf{d}) = \boldsymbol{\phi}_j(\mathbf{x}_t; \mathbf{d}) \otimes \boldsymbol{\iota}_m'$, $j = 1, \ldots, J$, where $\otimes$ denotes the Kronecker product and $\boldsymbol{\phi}_j(\mathbf{x}_t; \mathbf{d}) = (\phi_{j1}(\mathbf{x}_t; \mathbf{d}_1), \ldots, \phi_{jn}(\mathbf{x}_t; \mathbf{d}_n))'$ is a n-dimensional vector of transition functions.
(b) column-specific transition, where the transition functions are the same across the rows of Y_t and differ cross columns, that is the transition functions are $\Phi_j(\mathbf{x}_t; \mathbf{d}) = \boldsymbol{\iota}_n \otimes \boldsymbol{\phi}_j(\mathbf{x}_t; \mathbf{d})'$, j=1, …, J, where $\boldsymbol{\phi}_j(\mathbf{x}_t; \mathbf{d}) = (\phi_{j1}(\mathbf{x}_t; \mathbf{d}_1), \ldots, \phi_{jm}(\mathbf{x}_t; \mathbf{d}_m))'$ is a m-dimensional vector of transition functions.
(c) common transition, where the transition function is the same for all the elements of Y_t, that is the transition functions are $\Phi_j(\mathbf{x}_t; \mathbf{d}) = \phi_j(\mathbf{x}_t; \mathbf{d}) \boldsymbol{\iota}_m \boldsymbol{\iota}_n'$, $j = 1, \ldots, J$, where $\phi_j(\mathbf{x}_t; \mathbf{d})$ is a univariate transition function.

Motivated by the application and the type of research questions we will introduce later on in this paper, we choose the common transition specification with one covariate $\mathbf{x}_t = x_t$ and transition parameter vector $\mathbf{d}_j = (\alpha_j, \delta_j)'$. The resulting STR-MAR model is

$$Y_t = \sum_{j=1}^{J} \phi_j(x_t; \alpha_j, \delta_j) B_j + E_t, \tag{2}$$

where the collection of regime-specific parameters is $\boldsymbol{\psi} = (B_1, \ldots, B_J, \alpha_1, \ldots, \alpha_J, \delta_1, \ldots, \delta_J)$ with B_j, $j = 1, \ldots, J$, a sequence of $m \times n$ matrices. The conditional mean of the process Y_t is denoted by $M_t = g(x_t; \boldsymbol{\psi})$ and is a nonlinear function of the covariate x_t. The functional form of the smooth transition $\phi_j(x_t; \alpha_j, \delta_j)$ is discussed here below.

For the sake of exposition and without loss of generality, let us consider a 2-regimes model, that is $J = 2$, denote $\phi(\cdot) = \phi_1(\cdot)$, and assume the second regime is the reference one, that is $\phi_2(\cdot) = 1 - \phi(\cdot)$. Also, we define $\alpha = \alpha_j$ and $\delta = \delta_j$, for $j = 1, 2$. Some common choices for the smooth transition function are:

(a) logistic (LT)

$$\phi(x;\alpha,\delta) = \frac{1}{1+\exp(-\alpha(x-\delta))}, \qquad \alpha > 0,\ \delta \in \mathbb{R}$$

(b) exponential (ET)

$$\phi(x;\alpha,\delta) = 1-\exp(-\alpha(x-\delta)^2), \qquad \alpha > 0,\ \delta \in \mathbb{R}$$

(c) absolute logistic (ALT)

$$\phi(x;\alpha,\delta) = \frac{1-\exp(-\alpha|x-\delta|)}{1+\exp(-\alpha|x-\delta|)}, \qquad \alpha > 0,\ \delta \in \mathbb{R}$$

(d) quadratic logistic (QLT)

$$\phi(x;\alpha,\delta_1,\delta_2) = \frac{1}{1+\exp(-\alpha(x-\delta_1)(x-\delta_2))}, \qquad \alpha > 0,\ \delta_1 \leq \delta_2,\ \delta_1,\delta_2 \in \mathbb{R}$$

The *slope parameter* $\alpha > 0$ determines the speed of the transition between two regimes, and the *threshold parameter* δ is the threshold to the impact of the *transition variable*, x, on the transition function.

In the logistic transition, the function $\phi(x;\alpha,\delta)$ is monotonically increasing with inflection point at δ. For very large values of the slope parameter, the function converges to a step-wise function that gives a sudden shift between regimes that is $\phi(x;\alpha,\delta) \to H(x)$ for $\alpha \to +\infty$, where $H(x)$ is the Heaviside function. When $\alpha = 0$ the STR function becomes constant and the transition mechanism disappears that is $\phi(x;\alpha,\delta) \to 1$ for $\alpha \to 0$. In the other transition models, the function $\phi(x;\alpha,\delta)$ attains the minimum value at δ and is monotonically decreasing for $x < \delta$ and increasing for $x > \delta$. In Fig. 1, the transition functions are drawn for $\delta = 1$ and for different values of the slope parameter ($\alpha = 1, 3, 5, 10$). For all transition functions, the transition speeds up in a neighborhood of δ and is relatively slow for low values of α, and faster for larger values of α. In the following, we choose the absolute logistic transition, since we expect a symmetric behavior of the regime-switching dynamics in our applications. Nevertheless, a model comparison procedure can be applied to select the best transition model for the dataset considered following, for example, a Bayes factor [64, Chap. 5] or the DIC criterion [19].

Furthermore, we propose a robust model with t-distributed errors. [78] introduced Student's t errors in Bayesian linear models since the Student's t has the Cauchy and normal distributions as special cases and allows for outliers and departure from normality in the data due to excess of kurtosis. This assumption has been employed in many papers [e.g., see 41, 68] and extended along different directions to allow for more flexible and robust models [e.g., see 26, 61, 62]. Recently, [11] showed that

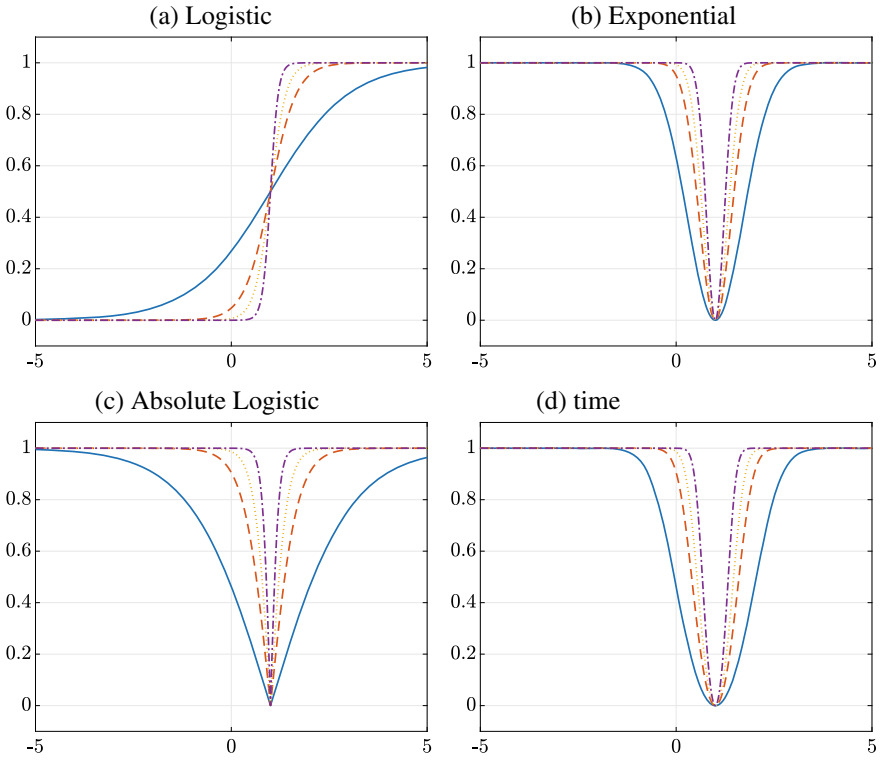


Fig. 1 Transition functions $\phi(x; \alpha, \delta)$ for different transition parameter values. *Note* The threshold parameter is taken as $\delta = 1$. Functions are drawn for different values of the slope parameter $\alpha = 1$ (solid), $\alpha = 3$ (dashed), $\alpha = 5$ (dotted), $\alpha = 10$ (dashed-dotted)

matrix-variate t errors can be useful in modeling network-valued data to filter out outliers due to measurement and network estimation errors. In our model, we assume the error term E_t follows a matrix-variate t distribution. A random matrix $X \in \mathbb{R}^{n \times m}$ follows a matrix-variate t distribution [see 45, Chap. 4], $X \sim t_{n,m}(\nu, M, \Sigma_1, \Sigma_2)$, if it has probability density function

$$P(X|\nu, M, \Sigma_1, \Sigma_2) = \kappa^{-1} |\Sigma_1|^{-\frac{m}{2}} |\Sigma_2|^{-\frac{n}{2}} \left| I_n + \Sigma_1^{-1} (X - M) \Sigma_2^{-1} (X - M)' \right|^{-\frac{\nu+m+n-1}{2}} \tag{3}$$

where κ is the normalizing constant

$$\kappa^{-1} = \frac{\Gamma_n\left(\frac{\nu+m+n-1}{2}\right)}{\pi^{\frac{mn}{2}} \Gamma_n\left(\frac{\nu+n-1}{2}\right)}$$

and $\Gamma_n(\cdot)$ is the multivariate gamma function and $|\cdot|$ denotes the matrix determinant. The matrix $M \in \mathbb{R}^{n\times m}$ is the location parameter, $\nu > 0$ is the degrees of freedom parameter, and the positive definite matrices $\Sigma_1 \in \mathbb{R}^{n\times n}$ and $\Sigma_2 \in \mathbb{R}^{m\times m}$ are scale parameters driving the covariances between each of the n rows and the m columns of X, respectively.

In summary, the STR-MAR model with ALT transition considered in the application on square matrices (i.e., $n = m$) is

$$\begin{aligned} Y_t &= M_t + E_t \qquad\qquad E_t \sim t_{n,n}(0, \Sigma_1, \Sigma_2) \\ M_t &= B_1\phi(x_t; \alpha, \delta) + B_2(1 - \phi(x_t; \alpha, \delta)) \\ \phi(x_t; \alpha, \delta) &= \frac{1 - \exp(-\alpha\left|x_t - \delta\right|)}{1 + \exp(-\alpha|x_t - \delta|)} \end{aligned} \tag{4}$$

with $\alpha > 0$, $\delta \in \mathbb{R}$, and $x_t \in \mathbb{R}$ is a predetermined variable. Let us define the information set as the sigma-algebra $\mathcal{I}_t = \sigma(\{Y_u, x_u\}_{u\leq t})$ generated by the past values of the dependent variable Y_u and the transition variable x_u. Thanks to the properties of the matrix-variate t distribution, for any choice of the transition functions and any number of regimes, the conditional mean and variance of Y_t are $\mathbb{E}(Y_t|\mathcal{I}_{t-1}) = M_t$, if $\nu > 1$, and $\mathbb{V}ar(Y_t|\mathcal{I}_{t-1}) = \Sigma_2 \otimes \Sigma_1/(\nu - 2)$, if $\nu > 2$.

In smooth transition models, the parameters have continuous dynamics, and the transition function can describe different types of behavior including sudden shifts. The model class can be extended along with different directions. Besides the direct extension to $J > 2$ regimes [e.g., following the strategy in 60], the proposed model can be generalized to account for multiple covariates and edge-specific transitions, as discussed in Sect. 2.

Latent variable models for the underlying network M_t represent an alternative framework. Continuously varying random parameters (e.g., parameters with random walk dynamics) can be more flexible, nevertheless, they are less tractable from the inference perspective and pose computational challenges. We leave this topic for future research.

Nonlinear Network Models

An emerging field in econometrics is the analysis of network-valued time series data. Since a network can be represented by means of a square matrix, called adjacency matrix, our model (2) can be directly applied to investigate the dynamic effects in time series of networks. In this chapter, we introduce a model for weighted networks, which can find direct application to signed weighted graphs [e.g., see 9].

Let us denote with $\mathcal{G} = (G_t)_{t\in\mathbb{N}}$ a temporal network, where $G_t = (V, E_t, Z_t)$ is a weighted network [13], $V = \{1, \ldots, n\}$ the vertex set, $E_t \subset V \times V$ the edge set, and $Z_t \in \mathbb{R}_+^{n\times n}$ the weighted adjacency matrix, such that $Z_{ij,t} = 0$ for each $(i, j) \notin E_t$ and $Z_{ij,t} \in \mathbb{R}_+$, otherwise.

The general class of models in eq. (2) can be used to fit and predict temporal networks, that is sequences of weighted adjacency matrices $Z_t, t \in \mathbb{N}$. When the edge weights are real-valued, STR-MAR applies directly to Z_t, whereas, for edge weights constrained to a subset of $\mathbb{R}$, the model can be applied on a suitable transformation of the weighted adjacency matrix. The transformations commonly used are logarithmic, when the edge weights are positive, that is,

$$Y_{ij,t} = \log(Z_{ij,t}), \quad Z_{ij,t} \in \mathbb{R}_+, \; i, j = 1, \ldots, n \tag{5}$$

and the logistic, when the edge weights belong to the unit interval, that is,

$$Y_{ij,t} = \text{logit}(Z_{ij,t}), \quad Z_{ij,t} \in (0, 1), \; i, j = 1, \ldots, n. \tag{6}$$

Network statistics are often used to describe the dynamics of the network topology; thus, the researcher is usually interested in forecasting them. In addition, in temporal networks, the lagged values of the network statistics can be used to capture persistence effects in the structure of the network, which means they can be used as transition variables to build a parsimonious STR autoregressive model.

In network analysis, the most frequently used statistics include direct connectivity measures, such as network degree, and indirect connectivity measures, such as clustering coefficient and eigenvector centrality. See [9] for a review on connectivity measures. In this chapter, these statistics are obtained from the binary adjacency matrix representation of the network, A_t, which in our notation is defined by setting a threshold on the edge weights, $A_{ij,t} = \mathbb{I}(Z_{ij,t} < 0.05)$. We define the degree of the node i as $k_{it} = \sum_{j=1}^{n} A_{ij,t}$. Then, the total network degree (d_t), the weighted degree (d_t^W), the average clustering coefficient (c_t), and the average eigenvector centrality (e_t) are defined as

$$\begin{aligned}
d_t &= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij,t} \qquad d_t^W = \sum_{i=1}^{n}\sum_{j=1}^{n} Z_{ij,t} \\
c_t &= \frac{1}{n}\sum_{i=1}^{n} C_{i,t} \qquad C_{i,t} = \frac{1}{k_{it}(k_{it}-1)}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{ij,t}A_{jk,t}A_{ki,t} \\
e_t &= \frac{1}{n}\sum_{i=1}^{n} E_{i,t} \qquad E_{i,t} = \frac{1}{k_{it}(k_{it}-1)}\sum_{j=1}^{n}\sum_{k=1}^{n} A_{ij,t}A_{jk,t}A_{ki,t},
\end{aligned} \tag{7}$$

where $C_{i,t}$ and $E_{i,t}$ are the clustering coefficient and eigenvector centrality, respectively, of node i at time t. In this chapter, we consider mainly the total network degree and the average clustering coefficient.

Our STR-MAR model can be used to forecast the network structure and the statistics. In the logit transform case, which will be used later on in this chapter, it is possible to find an analytical expression for the expected network degree. The predicted total degree of the graph G_{t+1} given $\mathcal{I}_t$ is

$$\begin{aligned}\mathbb{E}(d_{t+1}|\mathcal{I}_t) &= \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{P}(Y_{ij,t} > 0.05) \\ &= \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(\frac{1}{2} - \frac{1}{2}\Phi\left(\frac{\text{logit}^{-1}(0.05) - \mu_{ijt}}{\sqrt{2\sigma_{ij}^2}}\right)\right) \end{aligned} \tag{8}$$

where $\Phi(\cdot)$ denotes the cumulative density function (CDF) of the standard normal distribution.

Following the results in [48] on the moment of the logit-normal distribution and by applying Proposition 1 in [47], the expected weighted total degree can be written as

$$\mathbb{E}(d_{t+1}^W|\mathcal{I}_t) = \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}(Y_{ij}) = \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(\frac{1}{2} + \frac{b_{ijt} + c_{ijt}}{a_{ijt}}\right) \tag{9}$$

where

$$a_{ijt} = 1 + 2\sum_{r=1}^{\infty}\exp(-\sigma_{ij}^2 r^2/2)\cosh(r\mu_{ijt}) \tag{10}$$

$$b_{ijt} = \sum_{r=1}^{\infty}\exp(-\sigma_{ij}^2 r^2/2)\sinh(r\mu_{ijt})\tanh(\sigma_{ij}^2 r/2)) \tag{11}$$

$$c_{ijt} = \frac{2\pi}{\sigma_{ij}^2}\sum_{r=1}^{\infty}\frac{\exp(-(2r-1)^2\pi^2/2\sigma_{ij}^2)\sin((2r-1)\pi\mu_{ijt}/\sigma_{ij}^2)}{\sinh((2r-1)\pi^2/\sigma_{ij}^2)} \tag{12}$$

where $\mu_{ijt} = B_{1ij}\phi(x_t;\alpha,\delta) + B_{2ij}(1-\phi(x_t;\alpha,\delta))$ is the time-varying parameter, $\sigma_{ij}^2 = \Sigma_{1ij}\Sigma_{2ij}/(\nu-2)$ is the scale parameter, and $\sinh(\cdot)$, $\cosh(\cdot)$, and $\tanh(\cdot)$ are the hyperbolic sine, cosine, and tangent functions, respectively. The coefficients a_{ijt}, b_{ijt}, and c_{ijt} are evaluated by applying the truncation rule in [47], that is, all terms are included in the summation until $r = r^*$, where $r^* = \min\{|\mu_{ijt}/\sigma_{ij}^2 - \sqrt{2\log(1/\epsilon)}/\sigma|, |\mu_{ijt}/\sigma_{ij}^2 + \sqrt{2\log(1/\epsilon)}/\sigma_{ij}|, \lceil 1/2 + \sigma_{ij}\sqrt{\log(2/\epsilon)/2}/\pi\rceil\}$ and $\epsilon = 10^{-8}$.

Figure 2 shows the expected total degree $\mathbb{E}(d_{t+1}^W|\mathcal{I}_t)$ as a function of the previous period degree, that is $x_t = d_t^W$, for the following parameter setting: $B_{ij,1} \sim \mathcal{N}(-1,1)$ and $B_{ij,2} \sim \mathcal{N}(3,1)$ $i.i.d$ for all $i, j = 1,\ldots,50$ and $\sigma_{ij}^2 = 1.3$. When the transition variable d_t^W converges to the value of the threshold parameter, a large proportion of edge weights converge to one and the total degree increases. For smaller values of α, the peak in the expected degree is less concentrated about the threshold parameter value.

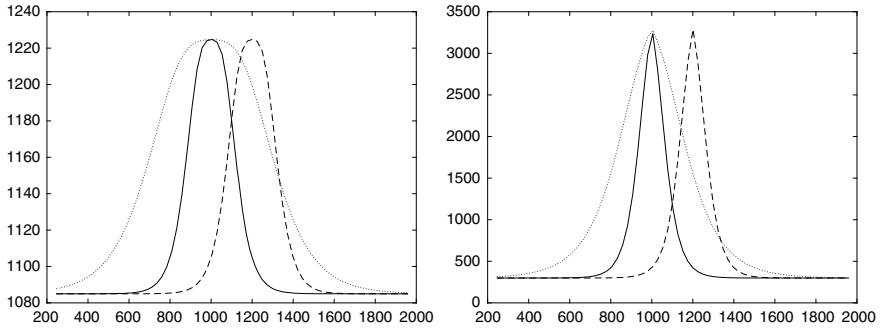


Fig. 2 Expected total degree $\mathbb{E}(d_{t+1}|\mathcal{I}_t)$ (left) and weighted total degree $\mathbb{E}(d^W_{t+1}|\mathcal{I}_t)$ (right) for a 50 nodes network as a function of the previous period degree d_t (horizontal axis). In each plot, the expected degree for $\alpha = 0.02$, $\delta = 1000$ (solid line), $\alpha = 0.02$, $\delta = 1200$ (dash line), and $\alpha = 0.008$, $\delta = 1000$ (dotted line)

Extensions

The model presented in Eq. (1) includes many special cases which can be used to extend the common transition model given in Eq. (2). In the following, we provide three directions to follow to extend the common-transition model.

1. *Smooth transition in the t-distribution.* The regime transition may involve also the degrees of freedom and/or the scale matrices in the t-distribution. Therefore, the smooth transition function $\phi_j(x; d)$ can be included in those elements, as follows:

$$\nu = \sum_{j=1}^{J} \bar{\nu}_j \phi_j(x; \mathbf{d})$$

$$\Sigma_\ell = \sum_{j=1}^{J} \Sigma_{\ell,j} \phi_j(x; \mathbf{d}), \qquad \ell = 1, 2.$$

2. *Modelling the smooth transition function using a set of covariates.* The smooth transition function can be expressed according to a set of covariates $\mathbf{x} = (x_1, \ldots, x_k)' \in \mathbb{R}^k$ that may exert an impact on the change of the regime. We provide here below an example of three specifications for the absolute logistic function (ALT),

$$\phi(\mathbf{x};\mathbf{d}) = \frac{2}{1+\exp(-\alpha\prod_{i=1}^{k}|x_i|-\delta_i)} - 1, \tag{13}$$

$$\phi(\mathbf{x};\mathbf{d}) = \frac{2}{1+\exp(-\alpha\sum_{i=1}^{k}|x_i|-\delta_i)} - 1, \tag{14}$$

$$\phi(\mathbf{x};\mathbf{d}) = \frac{2}{1+\exp(-\alpha|\mathbf{x}|'\boldsymbol{\beta}-\delta)} - 1, \tag{15}$$

where $\mathbf{d} = (\alpha, \delta_1, \ldots, \delta_k)$ in Eq. (13) and (14), and $\mathbf{d} = (\alpha, \delta_1, \ldots, \delta_k, \beta)$ in Eq. (15).

3. *Entry-specific smooth transition* As discussed in the presentation of our general STR-MAR model, a smooth transition function specific for each entry of the matrix Y_t can be specified. In this setting, the number of parameters to estimate increases and either parsimonious parametrization strategies or suitable inference procedures should be considered.

3 Bayesian Inference

Prior Specification

In this section, we present the prior distributions for the model parameters. Bayesian inference for STR and STAR models for scalar response has been developed in [22, 42, 63], then [15] proposed an extension to the multivariate case. In this chapter, we provide a further extension to the matrix-variate case, following the prior specification strategy given in [11] for matrix-variate models.

For the coefficient matrices B_j, with $j = 1, 2$, we assume a matrix normal distribution

$$B_j \sim \mathcal{MN}_{n,n}(0, \underline{\Omega}_1, \underline{\Omega}_2), \tag{16}$$

where $\underline{\Omega}_1 = \underline{\omega}_1 I_n$, and $\underline{\Omega}_2 = \underline{\omega}_2 I_n$, with $\underline{\omega}_1 > 0$ and $\underline{\omega}_2 > 0$ fixed. A $(n \times p)$ random matrix X is said to be distributed as a matrix normal [45, Chap. 2] with mean M and covariance matrices Σ_1, Σ_2, with Σ_1 of size $(n \times n)$ and Σ_1 of size $(p \times p)$, written $X \sim \mathcal{MN}_{n,p}(M, \Sigma_1, \Sigma_2)$, if its probability density function is

$$P(X|M, \Sigma_1, \Sigma_2) = (2\pi)^{-\frac{np}{2}}|\Sigma_2|^{-\frac{p}{2}}|\Sigma_1|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\operatorname{tr}\left(\Sigma_2^{-1}(X-M)'\Sigma_1^{-1}(X-M)\right)\right). \tag{17}$$

The distribution is equivalent to a multivariate normal distribution with a product-separable covariance structure, that is, $X \sim \mathcal{MN}_{n,p}(M, \Sigma_1, \Sigma_2)$ is equivalent to $\operatorname{vec}(X) \sim \mathcal{N}_{np}(\operatorname{vec}(M), \Sigma_2 \otimes \Sigma_1)$, where $\operatorname{vec}(\cdot)$ is a vectorization operator that

stacks all the columns of a matrix into a column vector. Since $\Sigma_2 \otimes \Sigma_1 = (\Sigma_2/a) \otimes (a\Sigma_1)$ for any $a \neq 0$, the noise covariance matrices of the matrix-variate t distribution, Σ_1, Σ_2, are not identifiable. We address this issue by restricting the trace of each covariance matrix to a given value, that is by imposing $\text{tr}(\Sigma_j) = \varrho_j$, for $j = 1, 2$. For the noise covariances, Σ_1 and Σ_2, we assume the following independent prior distributions:

$$\Sigma_1 \sim \mathcal{W}_n^*(\underline{\Psi}_1^{-1}, \underline{\kappa}_1, \underline{S}_1), \qquad \Sigma_2 \sim \mathcal{IW}_n^*(\underline{\Psi}_2, \underline{\kappa}_2, \underline{S}_2), \tag{18}$$

where $\mathcal{W}_n^*(\Psi, \kappa, S)$ and $\mathcal{IW}_n^*(\underline{\Psi}, \underline{\kappa}, \underline{S})$ denote the truncated Wishart and inverse Wishart distributions, respectively, with densities proportional to

$$\mathcal{W}_n^*(\Sigma_1|\underline{\Psi}_1^{-1}, \underline{\kappa}_1, \underline{S}_1) \propto \mathcal{W}_n(\Sigma_1|\underline{\Psi}_1^{-1}, \underline{\kappa}_1)\mathbb{I}_{\underline{S}_1}(\Sigma_1), \quad \underline{S}_1 = \{\Sigma_1 : \text{tr}(\Sigma_1) = \varrho_1\},$$
$$\mathcal{IW}_n^*(\Sigma_2|\underline{\Psi}_2^{-1}, \underline{\kappa}_2, \underline{S}_2) \propto \mathcal{IW}_n(\Sigma_2|\underline{\Psi}_2^{-1}, \underline{\kappa}_2)\mathbb{I}_{\underline{S}_1}(\Sigma_2), \quad \underline{S}_2 = \{\Sigma_2 : \text{tr}(\Sigma_2) = \varrho_2\},$$

where $\mathcal{W}_n$ and $\mathcal{IW}_n$ denote the Wishart and inverse Wishart distributions, respectively. We use the scale parametrization for the Wishart and inverse Wishart distributions [see 39, Appendix A, p.577], where $\underline{\Psi}_1^{-1}$ and $\underline{\Psi}_2$ are the scale matrices and $\underline{\kappa}_1$ and $\underline{\kappa}_2$ are the degrees of freedom hyper-parameters.

Since the variance of Y_t is defined only for $\nu > 2$, we assume the following gamma prior distribution truncated on the interval $(2, +\infty)$:

$$\nu \sim \mathcal{TG}a(\underline{a}_\nu, \underline{b}_\nu; 2, +\infty). \tag{19}$$

The gamma prior distribution for the degrees of freedom parameter has been previously considered, for example, in [42, 76]. For the use of an improper prior, see [36]. Owing to the use of proper prior distributions for B_1 and B_2, their posterior distributions are well defined for $\nu > 0$ [42], whereas the constraint $\nu > 2$ is required when using improper prior distributions.

Regarding the slope parameter, α, and the threshold parameter, δ, [56] show that the posterior estimates are insensitive to the prior specification according to the DIC criterion. Nevertheless, the RMSE and MAE are significantly higher when both prior distributions are uniform. Thus, we assume the following non-uniform and independent prior distributions:

$$\delta \sim \mathcal{N}(\underline{\mu}_\delta, \underline{\sigma}_\delta^2), \qquad \alpha \sim \mathcal{G}a(\underline{a}_\alpha, \underline{b}_\alpha). \tag{20}$$

We summarize our Bayesian model in the Directed Acyclic Graph representation of Fig. 3.

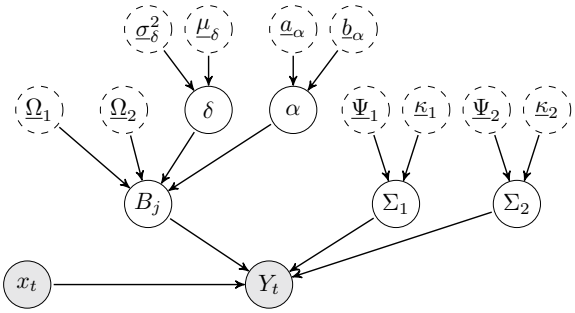


Fig. 3 Directed Acyclic Graph of the proposed Bayesian STR-MAR model for temporal networks. It exhibits the conditional independence structure of the observation model for Y_t with covariate x_t (gray circles), the parameters B_j, Σ_1, Σ_2, α, δ (white solid circles) and the fixed hyperparameters $\underline{a}_\alpha, \underline{b}_\alpha, \underline{\mu}_\delta, \underline{\sigma}_\delta^2, \underline{\Psi}_1, \underline{\kappa}_1, \underline{\Psi}_2, \underline{\kappa}_2$ (white dashed circles). The directed arrows show the causal dependence structure of the model

Posterior Approximation

Denote the collection of parameters with $\boldsymbol{\theta} = (B_1, B_2, \Sigma_1, \Sigma_2, \nu, \alpha, \delta)$, let $\mathbf{x} = (x_1, \ldots, x_T)$ be the collection of covariates, and let $\mathbf{Y} = (Y_1, \ldots, Y_T)$ be the collection of all observed networks. The likelihood of the model in Eq. (4) is

$$P(\mathbf{Y}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{t=1}^{T} c \cdot \left|I_n + \Sigma_1^{-1}\right|(Y_t - M_t)\Sigma_2^{-1}(Y_t - M_t)'^{-\frac{\nu+2n-1}{2}}, \qquad (21)$$

where

$$c = \frac{\Gamma_n\left(\frac{\nu+2n-1}{2}\right)}{\pi^{\frac{n^2}{2}}\Gamma_n\left(\frac{\nu+n-1}{2}\right)}\left|\Sigma_1\right|^{-\frac{n}{2}}\left|\Sigma_2\right|^{-\frac{n}{2}}.$$

Since the joint posterior distribution implied by the prior distributions in Eq. (16)–(18)–(19) and the likelihood in Eq. (21) is not tractable, we follow a data augmentation approach [65, Chap. 9]. We exploit the representation of the matrix t distribution as a scale mixture of matrix normal distributions, with Wishart mixing distribution [71]. From Theorem 4.3.1 in [45], if $S \sim \mathcal{W}_p(\Sigma_1^{-1}, \nu + p - 1)$ and $X|S \sim \mathcal{MN}_{p,m}(M, S^{-1}, \Sigma_2)$, then $X \sim t_{p,m}(\nu, M, \Sigma_1, \Sigma_2)$. By assuming th parametrization of the inverse Wishart proposed in [39], we obtain the equivalent representation $W = S^{-1} \sim \mathcal{IW}_p(\Sigma_1, \nu + p - 1)$ and $X|W \sim \mathcal{MN}_{p,m}(M, W, \Sigma_2)$. We apply this result to $Y_t \sim t_{n,n}(\nu, M_t, \Sigma_1, \Sigma_2)$ and obtain the complete data likelihood

$$P(\mathbf{Y}, \mathbf{W}|\mathbf{x}, \boldsymbol{\theta}) = \prod_{t=1}^{T} \Big[(2\pi)^{-\frac{n^2}{2}} |W_t|^{-\frac{n}{2}} |\Sigma_2|^{-\frac{n}{2}} \exp\Big(-\frac{1}{2} \operatorname{tr}\left(\Sigma_2^{-1}(Y_t - M_t)' W_t^{-1}(Y_t - M_t)\right)\Big) \cdot \frac{|\Sigma_1|^{\frac{\nu+n-1}{2}}}{2^{\frac{(\nu+n-1)n}{2}} \Gamma_n(\frac{\nu+n-1}{2})} |W_t|^{-\frac{\nu+2n}{2}} \exp\Big(-\frac{1}{2} \operatorname{tr}(\Sigma_1 W_t^{-1})\Big)\Big], \tag{22}$$

where $\mathbf{W} = (W_1, \ldots, W_T)$ is the collection of auxiliary variables, with $W_t \sim \mathcal{IW}_n(\Sigma_1, \nu + n - 1)$.

The data augmentation approach combined with our prior assumptions allows us to derive analytically the full conditional distributions of B_1, B_2, Σ_1, Σ_2, and $\mathbf{W}$. Since the joint posterior distribution is not tractable, we implement an MCMC approach based on a Gibbs sampling algorithm to sample from the posterior distribution and to approximate all posterior quantities of interest. The Gibbs sampler generates iteratively from the full conditional distributions of the parameters. See Appendix for the details of the approximated inference procedure.

4 Empirical Analysis

In this section, we first describe the network extraction method and then illustrate the inference results for our model to oil and financial market datasets.

We estimate networks among statistical units such as firms and oil-producing countries following the pairwise Granger-causality approach proposed in [12]. In the proposed framework, the network G_t is directed, each node denotes a firm and an edge from node i to node j indicated that i Granger causes j at time t. The weighted adjacency matrix Z_t contains the p-values of the pairwise Granger test.

We apply a rolling window approach, with a window length of τ observations and estimate the following VAR(m) model for the time series of interest:

$$\begin{aligned} x_{i,t} &= \sum_{l=1}^{m} b_{11l} x_{i,t-l} + \sum_{l=1}^{m} b_{12l} x_{j,t-l} + \varepsilon_{it} \\ x_{j,t} &= \sum_{l=1}^{m} b_{21l} x_{i,t-l} + \sum_{l=1}^{m} b_{22l} x_{j,t-l} + \varepsilon_{jt} \end{aligned} \tag{23}$$

where $x_{i,t} = \hat{\sigma}_{i,t}$ and $i, j = 1, \ldots, N$. The entry $Z_{ij,t}$, $i \neq j$, of the weighted adjacency matrix is defined as the p-value of the F-statistic under the null hypothesis $b_{ij,1} = b_{ij,2} = \ldots = b_{ij,m} = 0$, of the Granger test. Therefore, the element $Z_{ij,t} \in (0, 1)$ represents the probability that the relationship between $x_{i,t}$ and $x_{j,t}$ is statistically significant.

Remark 1 The Granger-causality (GC) approach is a simple and flexible approach in dealing with large dimensional data, but it may introduce estimation errors, also due to spurious relationships [e.g., see the discussion in 2]. Our proposed model

accounts for those issues via the specification of a noise term, E_t, with matrix-variate t distribution. We expect our model can be used for other types of network data. For example, tail networks can be considered especially when focusing on the returns dimension [i.e., quantile regression approaches, as in 23]. We leave this application to future extensions of the proposed STR-MAR model.

Volatility Networks

We investigate the dynamics of financial risk spillovers by considering a network of volatility linkages among the following companies.[1] The data sample includes the 50 European firms with the largest market capitalization and ranges from 4 January 2016 to 30 September 2020, at the weekly frequency (Friday–Friday), thus including the period before and after the outbreak of COVID-19. The analyzed firms (22 German, 24 French, 4 Italian) belong to 11 GICS sectors: Financials (7 firms), Communication Services (4 firms), Consumer Discretionary (11 firm), Consumer Staples (5 firms), Health Care (6 firms), Energy (2 firms), Industrials (5 firms), Information Technology (3 firms), Materials (2 firms), Real Estate (1 firms), Utilities (3 firms), and not classified in a specific GICS sector (1 firm).[2]

The weekly volatility for firm i, $\hat{\sigma}^2_{i,t}$, is computed using the estimator of the variance proposed by [38]:

$$\begin{aligned}\hat{\sigma}^2_{i,t} &= 0.511(H_{i,t} - L_{i,t})^2 - 0.383(C_{i,t} - O_{i,t})^2 \\ &\quad - 0.019[(C_{i,t} - O_{i,t})(H_{i,t} + L_{i,t} - 2O_{i,t}) - 2(H_{i,t} - O_{i,t})(L_{i,t} - O_{i,t})],\end{aligned} \tag{24}$$

$i = 1, \ldots, n, t = 1, \ldots, T^*$, where $H_{i,t}$ is the weekly logarithmic high price, $L_{i,t}$ is the weekly logarithmic low price, $O_{i,t}$ is the weekly logarithmic opening price, and $C_{i,t}$ is the logarithmic closing price. The weekly prices have been obtained by taking in a given week the maximum among the daily high prices (weekly High Price), the minimum among the daily low prices (weekly Low Price), the opening price of the first available day in a week (weekly Opening Price), and the closing prices of the last available day in a week (weekly Closing Price).

We use a rolling window length of $\tau = 104$ weeks (2 years) and estimate a sequence of Granger-causal networks from the collection of weekly time series of length $T^* = 248$. In the proposed framework, the nodes represent the institutions and the graph encodes the Granger-causality relationships among them. As our causal financial networks are directed, each adjacency matrix is asymmetric.

[1] The study of financial spillovers can also be performed by focusing on tail networks obtained from return quantile regressions [14]. We leave this application to future extensions of the proposed STR-MAR model.

[2] The list of the firms, the countries, and the information about their GICS sectors and industries are available upon request to the Authors.

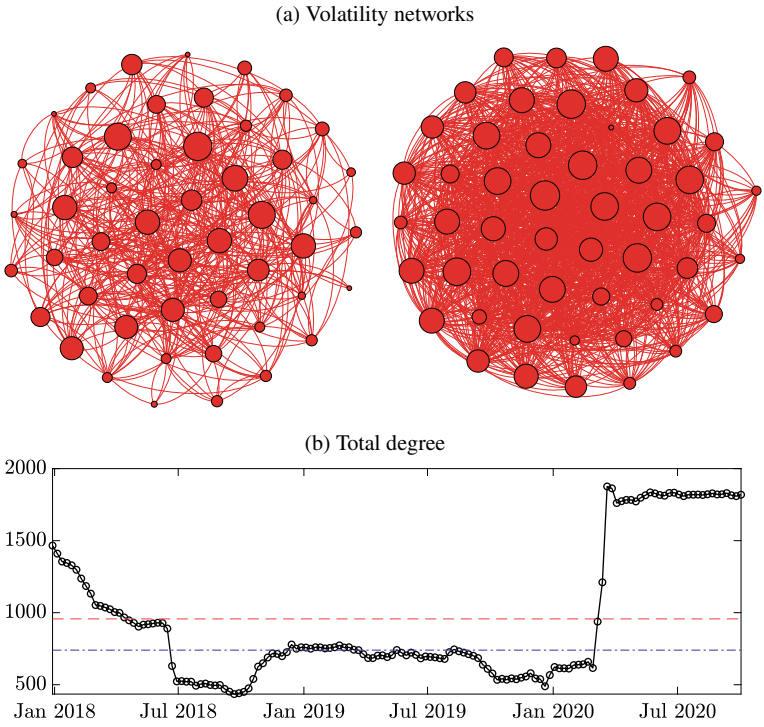


Fig. 4 Top: volatility networks for 50 large capitalization European firms (nodes) on 20 December 2019 (left) and 20 March 2020 (right). We show the pairwise directed edges extracted with pairwise Granger procedure. The edges are clockwise-oriented and the node size is based on the node degree centrality. Bottom: network total degree (black, solid line), together with sample average (red, dashed line), and sample median (blue, dash-dotted line). *Note* Network visualization have been made with Gephi software [8]

We consider a logistic link function $Y_{ij,t} = \text{logit}(Z_{ij,t})$ to obtain the edges of a volatility network. It represents the connectivity patterns among the stock volatility of a set of European institutions. We get from the rolling window analysis a sample of $T = 145$ adjacency matrices, covering the period from 29 December 2017 to 20 October 2020.[3]

As exogenous variable, we consider (i) the degree of the network d_t (see Eq. 7); (ii) the average clustering coefficient c_t; and (iii) the implied volatility on the Euro STOXX 50 index (V2X). All the covariates are lagged by one period.

Panel a in Fig. 4 shows the volatility network on 20 December 2019 (left) and 20 March 2020 (right), and Panel b, the evolution of the total degree over time (black, solid line), together with sample average (red, dashed line), and sample median (blue,

[3] The estimation algorithm has been parallelized and implemented in MATLAB on two nodes at the High-Performance Computing (HPC) cluster at VERA, Ca' Foscari University of Venice. Each node has 2 CPUs Intel Xeon with 20 cores 2.4 GHz and 768 GB of RAM.

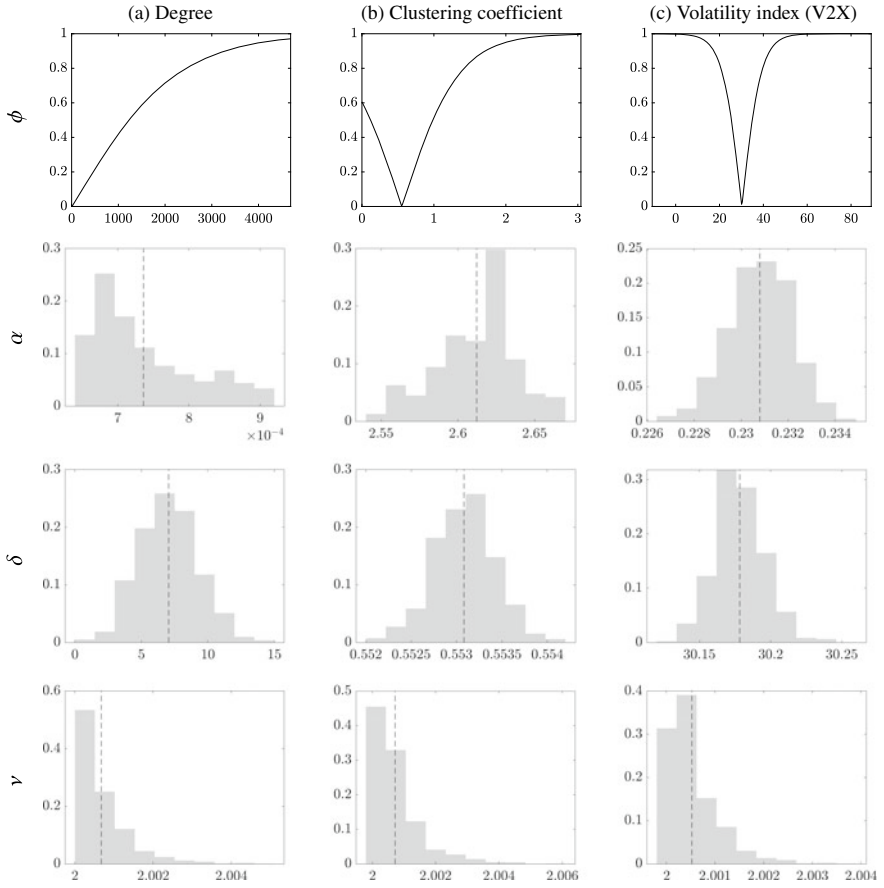


Fig. 5 First row: transition function $\phi(x; \hat{\alpha}, \hat{\delta})$, evaluated on a grid of values for x, where $\hat{\alpha}$ and $\hat{\delta}$ are the posterior means of α, δ, respectively. Other rows: posterior distribution (histograms) and posterior mean (dashed line) of the transition, threshold, and degrees of freedom parameters. *Note* in the columns the results for different choices of the transition variable

dash-dotted line). We observe a change in the network structure in correspondence with the outbreak of COVID-19 in March 2020. This suggests the adoption of non-linear models that accounts for multiple regimes.

Figure 5 includes the transition function for the degree, clustering coefficient, and the volatility index (first row) using the Bayesian estimates of the parameters, that are the posterior mean of α (second row), δ (third row), and ν (last row). The posterior distribution of the degrees of freedom parameter, ν, is concentrated on values right above 2, which provides evidence of the excess of kurtosis in the distribution of the error term. It follows that the Student's t represents a more suitable alternative with respect to the Gaussian case. The posterior distribution of α concentrates on values far away from zero, which supports the assumption of a smooth transition. When the

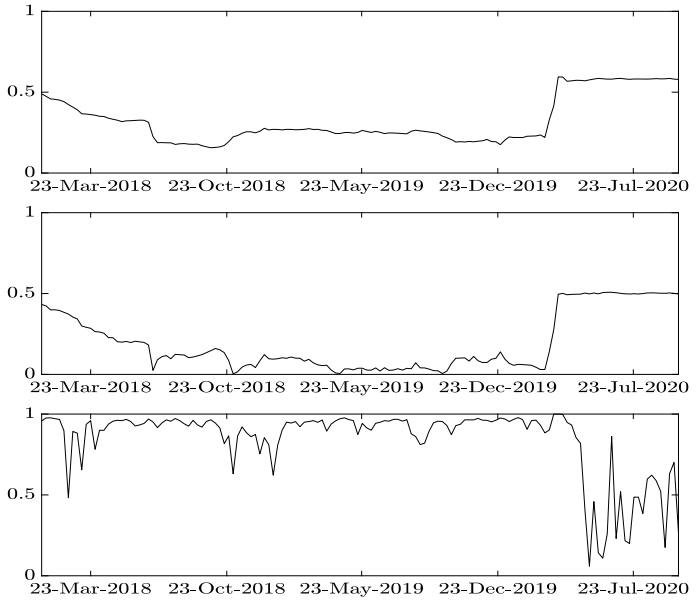


Fig. 6 Posterior mean of the transition function $\phi(x_t; \hat{\alpha}, \hat{\delta})$ for degree (top panel), clustering coefficient (middle panel), and the volatility index (bottom panel)

V2X is used as covariate, the observed network is close to the configuration B_2 when the volatility level approaches to $\hat{\delta} = 30.15$ which is far above the average volatility level. Conversely, in the clustering coefficient case, the estimated threshold (0.55) is below the average value of the covariate (0.67). Comparing these two cases, we also find that the estimated slope parameter, $\hat{\alpha}$, is substantially higher for the Clustering Coefficient (2.61) than for the V2X (0.23), suggesting that transitions occur faster for the volatility index. Figure 6 shows the posterior mean of the transition function for the degree (top panel), the clustering coefficient (medium panel), and the volatility index (bottom panel). Interestingly, the dynamic for the degree and clustering coefficient shows a structural change in the correspondence of the outbreak of COVID-19 in March 2020. This highlights the network topology has suddenly changed and has been persistently affected due to the spread of the pandemic. Differently, the volatility index indicates a structural change after the spread of COVID-19 but with a mean-reverting behavior towards the previous regime.

Finally, with reference to the model using the volatility index as covariate, Fig. 7 shows the networks implied by the matrices B_1 and B_2. Specifically, the networks are obtained by applying a threshold of 0.05 to the matrix of p-values, and the node size is proportional to the total node degree. The different topology of the two networks provides evidence of a change in the dependence structure across the states. Therefore, we conclude that the proposed model can successfully describe time-

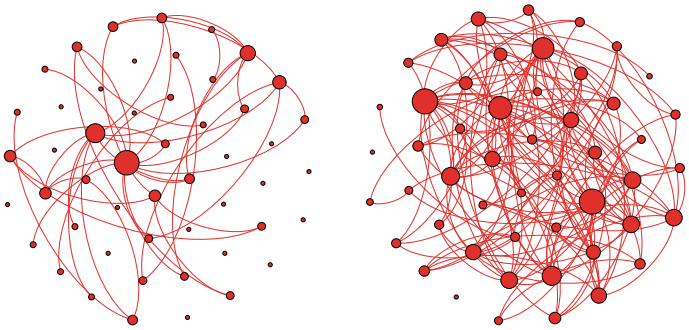


Fig. 7 Networks implied by the matrices B_1 (left) and B_2 (right) for the model with the economic activity index as covariate. The networks are obtained by applying a threshold of 0.05 to the matrix of p-values. The node size is proportional to the total node degree

varying network sequences that evolve smoothly between different states. Similar results have been obtained when using the degree or the clustering coefficient.

Oil Production Networks

Our model can be used to study the dynamics of the linkages between oil-producing countries in the international oil market. We aim at modeling the dependence among the 17 biggest producers: Algeria (AL), Angola (AN), Brazil (BR), Canada (CA), Colombia (CO), India (IN), Kuwait (KU), Mexico (ME), Nigeria (NI), Norway (NO), Oman (OM), Qatar (QA), Saudi Arabia (SA), the UAE (UA), the UK (UK), the USA (US), and Venezuela (VE). This study analyzes monthly data from March 1998 to January 2015 and has been downloaded from the Monthly Oil Data Service (MODS) database provided by the International Energy Agency (IEA). We estimate the Granger-causal networks using a rolling window of length $\tau = 60$ months, and obtain a monthly time series of $T = 143$ networks.

As for the previous application, we include the degree and the clustering coefficient as exogenous variables. Given the relationship between the oil production and the global business cycle, we also include the index of global real economic activity in industrial commodity markets [32, 35, 49, 50] which measures the changes in the global demand and is based on ocean transportation fares. All the covariates are lagged by one period.

Panel a in Fig. 8 shows the oil network between June 2006 (left) and March 2009 (right), while Panel b reports the evolution of the total degree over time. Interestingly, we observe an increasing trend in the connectivity level from the beginning of 2007 to 2010. Our proposed model is able to capture this smooth transition.

Figure 9 includes the transition function for the degree, clustering coefficient, and the economic activity index (first row) using the posterior mean of α (second row), δ

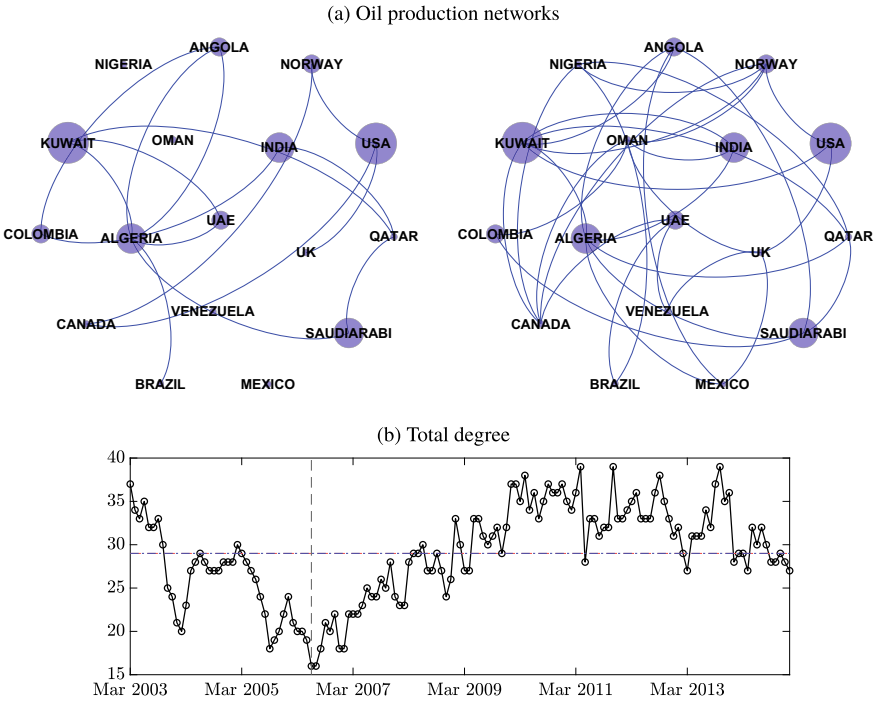


Fig. 8 Top: oil production networks for 17 countries (nodes) before the outbreak of the 2008 financial crisis (June 2006, left) and after it (January 2009, right). We show the pairwise directed edges extracted with the pairwise Granger procedure. The edges are clockwise-oriented and the node size is proportional to the node total degree centrality. Bottom: network total degree (black, solid line), together with sample average (red, dashed line), and sample median (blue, dash-dotted line). The vertical dashed line indicates an episode of a very low connectivity network (June 2006). *Note* Network visualization have been made with Gephi software [8]

(third row), and ν (last row). The posterior distribution for the degrees of freedom, ν, concentrates around 2, which provides evidence of fat tails for the noise distribution and confirms that the Student's t represents a proper choice of the model with respect to the standard Gaussian case. The estimated value of α supports the smooth transition hypothesis. For all covariates, the posterior mean of the threshold parameter, $\hat{\delta}$, is above the sample average of the variable. Figure 10 shows the posterior mean of the transition function (i.e., $\phi(\cdot)$) for degree (top panel), clustering coefficient (medium panel), and the economic activity index (bottom panel). Interestingly, the dynamic for degree shows a transition in the correspondence of March 2004 when OPEC members unanimously agree to implement a cut on oil production. This indicates that the connectivity of the oil production network exhibits a structural change during that episode. Another peak in the transition function for degree is found around October 2005 after the outbreak of Hurricanes Katrina and Rita in August of the same year. The extreme weather events shut down oil and gas production in the Gulf of Mexico

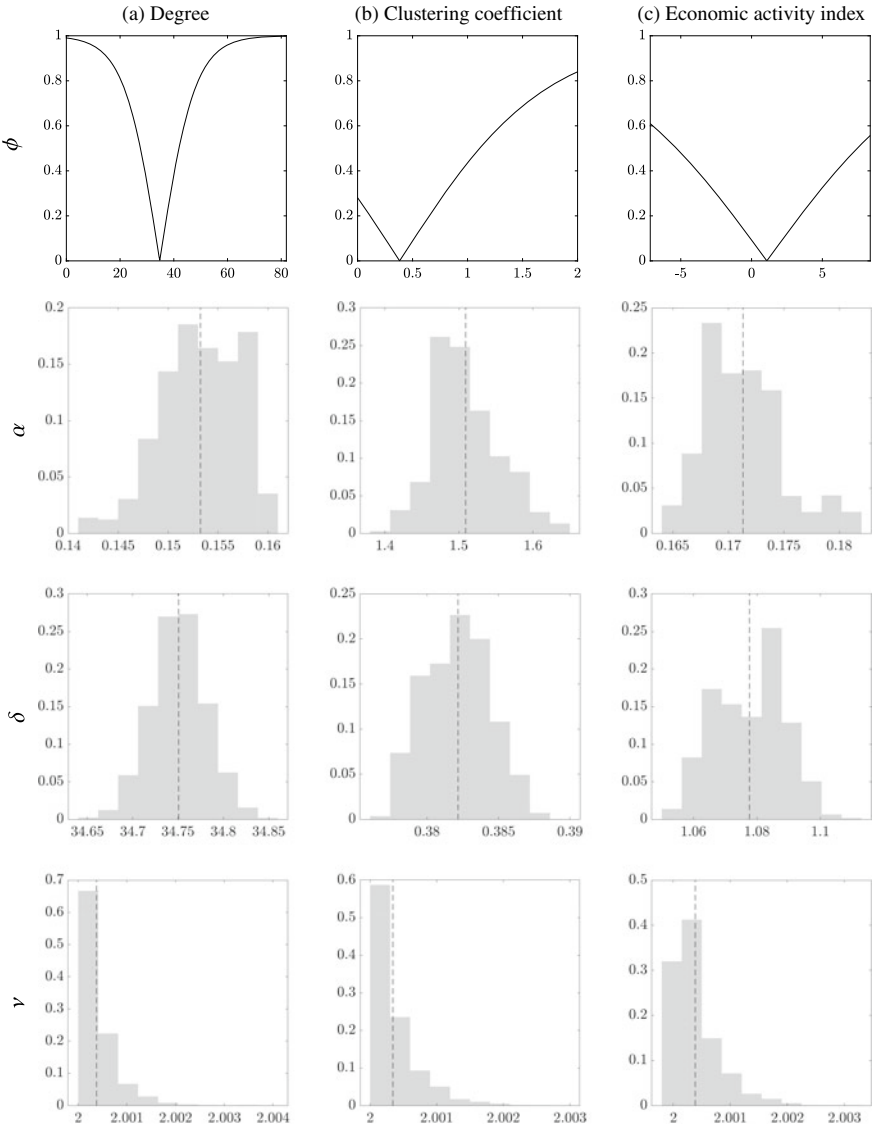


Fig. 9 First row: transition function $\phi(x; \hat{\alpha}, \hat{\delta})$, evaluated on a grid of values for x, where $\hat{\alpha}$ and $\hat{\delta}$ are the posterior means of α, δ, respectively. Other rows: posterior distribution of the transition, threshold, and degrees of freedom parameters. Other rows: posterior distribution (histograms) and posterior mean (dashed line) of the transition, threshold, and degrees of freedom parameters. *Note* in the columns the results for different choices of the transition variable

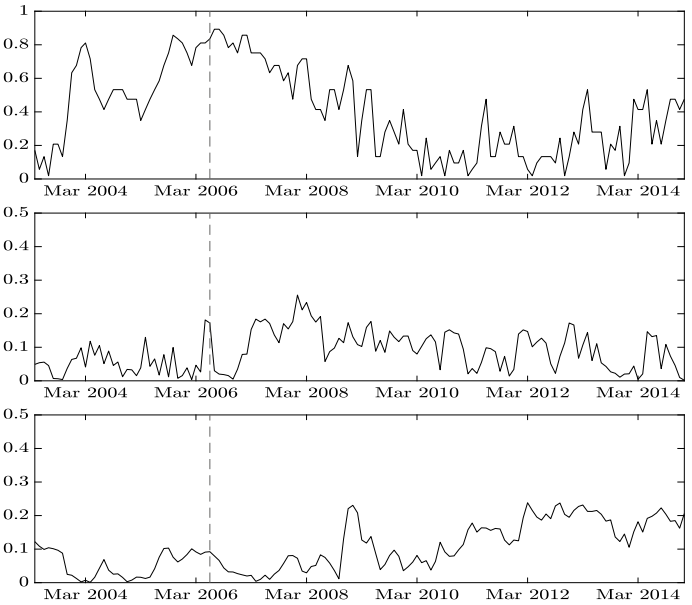


Fig. 10 Posterior mean of the transition function $\phi(x_t; \hat{\alpha}, \hat{\delta})$ for degree (top panel), clustering coefficient (middle panel), and the economic activity index (bottom panel). The vertical dashed line indicates an episode of very low connectivity network (June 2006)

which represented 25% of US crude oil production and 20% of natural gas [51]. The disruption continued until June 2006 representing a total cut in the oil production close to 450,000 bbl/d for the Gulf of Mexico [28]. Other subsequent episodes have occurred during the Global and European financial crises but have involved lower values for the smooth transition function. Regarding the clustering coefficient and the economic activity index, the patterns of the smooth transition functions show that there are no particular structural changes over the considered period thus indicating that those variables maintain a linear relationship with the oil network over time.

Figure 11 plots the networks implied by the matrices B_1 and B_2, estimated from the model with the economic activity index as covariate. We find a remarkable difference between the two network structures; however, we stress that our smooth transition model represents an observed network sequence as a convex combination of two latent ones, perturbed by random noise. Therefore, the empty network implied by B_2 should not be interpreted as a state of the world that characterizes some periods of the sample (as would be in Markov-switching models).

There are few cases in the sample where the matrix B_2 is assigned a combination weight close to one. It is possible to see that those cases correspond to very low connectivity network episodes. For example, the estimated STR-MAR model assigns weights close to zero and one to $\hat{B}_1$ and $\hat{B}_2$, respectively, in correspondence of June 2006 (see the dashed vertical line in Figs. 8 and 10). This date corresponds to a

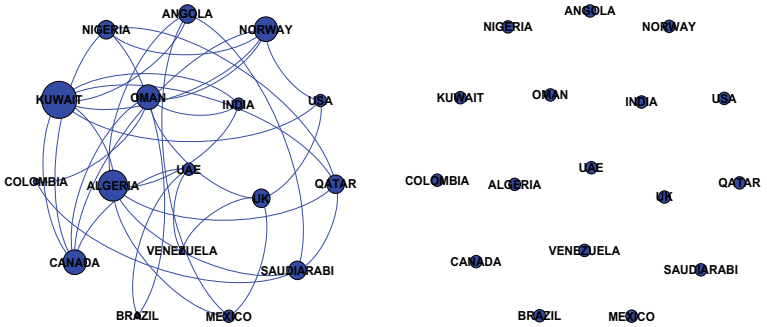


Fig. 11 Networks implied by the matrices B_1 (left) and B_2 (right) for the model with the economic activity index as covariate. The networks are obtained by applying a threshold of 0.05 to the matrix of p-values. The node size is proportional to the total node degree

network of our dataset with low connectivity level. This structure is associated to the stagnation of oil supply and the corresponding increase in oil price that led to a low interdependence among oil-producing countries.

5 Conclusion

In this chapter, we have introduced a smooth transition model for matrix-valued panel data to account for nonlinearity in time series and heavy tails in the innovations, and we have provided the relevant properties of the model including closed-form expression for the predictors. The adopted inference approach is within a Bayesian framework and an MCMC algorithm is proposed for the approximation of the posterior distribution. In the empirical analysis, we have provided applications to financial volatility and oil production networks showing the ability of the model in detecting the change in the network structure. The proposed framework is general and can be applied to several fields of data science that involve matrix-variate data.

Acknowledgements This research used the SCSCF and the HPC multiprocessor cluster systems provided by the Venice Centre for Risk Analytics (VERA) at the Ca' Foscari University of Venice. MB, RC, and MC acknowledge financial support from the Italian Ministry MIUR under the PRIN project Hi-Di NET—Econometric Analysis of High-Dimensional Models with Network Structures in Macroeconomics and Finance (grant agreement no. 2017TA7TYC). MI acknowledges financial support from the Marie Skłodowska-Curie Actions, European Union, Seventh Framework Program HORIZON 2020 under REA grant agreement no. 887220.

Appendix—Posterior approximation

The Gibbs sampler iterates the following steps:

1. Draw $(\nu, \mathbf{W})$ from the joint posterior distribution $P(\nu, \mathbf{W}|\mathbf{Y}, \mathbf{x}, B_1, B_2, \Sigma_1, \Sigma_2, \alpha, \delta)$ with a collapsed-Gibbs step that first samples $\nu \sim P(\nu|\mathbf{Y}, B_1, B_2, \Sigma_1, \Sigma_2, \alpha, \delta)$ and then $\mathbf{W} \sim P(\mathbf{W}|\mathbf{Y}, \mathbf{x}, B_1, B_2, \Sigma_1, \Sigma_2, \nu, \alpha, \delta)$.
2. Draw $\text{vec}(B_j)$ from the multivariate normal distribution $P(\text{vec}(B_j)|\mathbf{Y}, \mathbf{x}, \mathbf{W}, \Sigma_1, \Sigma_2, \alpha, \delta)$.
3. Draw Σ_1 from the Wishart distribution $P(\Sigma_1|\mathbf{W}, \nu)$.
4. Draw Σ_2 from the inverse Wishart distribution $P(\Sigma_2|\mathbf{Y}, \mathbf{x}, B_1, B_2, \mathbf{W}, \alpha, \delta)$.
5. Draw α and δ from their joint full conditional distribution $P(\alpha, \delta|\mathbf{Y}, \mathbf{x}, B_1, B_2, \mathbf{W}, \Sigma_2)$ using an adaptive Metropolis–Hastings algorithm.

In the following, we provide the derivation of the full conditional distributions used in the Gibbs sampler and discuss the sampling methods.

Sampling B_1. Let $\tilde{Y}_t = Y_t - (1 - \phi(x_t; \alpha, \delta))B_2$. The coefficient matrix B_1 is drawn from the posterior full conditional distribution

$$
\begin{aligned}
&P(B_1|\mathbf{Y}, \mathbf{W}, \mathbf{x}, B_2, \Sigma_1, \Sigma_2) \propto P(B_1)P(\mathbf{Y}, \mathbf{W}|\mathbf{x}, B_1, B_2, \Sigma_1, \Sigma_2, \alpha, \delta) \\
&\propto \exp\Big(-\frac{1}{2}\,\text{tr}\,\big(\text{vec}\big(B_1\big)'(\underline{\Omega}_2 \otimes \underline{\Omega}_1)^{-1}\text{vec}\big(B_1\big)\big) \\
&\quad - \frac{1}{2}\sum_{t=1}^{T}\text{tr}\,\big(\text{vec}\big(Y\big)_t - \phi(\tilde{x_t}; \alpha, \delta)B_1{}'(\Sigma_2 \otimes W_t)^{-1}\text{vec}\big(Y\big)_t - \phi(\tilde{x_t}; \alpha, \delta)B_1\big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\,\text{tr}\,\big(\text{vec}\big(B_1\big)'\big[(\underline{\Omega}_2 \otimes \underline{\Omega}_1)^{-1} + \sum_{t=1}^{T}(\Sigma_2 \otimes W_t)^{-1}\phi(x_t; \alpha, \delta)^2\big]\text{vec}\big(B_1\big) \\
&\quad - 2\sum_{t=1}^{T}\text{vec}\big(\tilde{Y}\big)_t{}'(\Sigma_2 \otimes W_t)^{-1}\phi(x_t; \alpha, \delta)\text{vec}\big(B_1\big)\big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}\,\text{tr}\,\big(\text{vec}\big(B_1\big)'\overline{\Omega}_1^{-1}\text{vec}\big(B_1\big) - 2\text{vec}\big(\overline{M}\big)_1'\overline{\Omega}_1^{-1}\text{vec}\big(B_1\big)\big)\Big),
\end{aligned}
$$

where

$$
\overline{\Omega}_1 = \Big[(\underline{\Omega}_2 \otimes \underline{\Omega}_1)^{-1} + \sum_{t=1}^{T}(\Sigma_2 \otimes W_t)^{-1}\phi(x_t; \alpha, \delta)^2\Big]^{-1},
$$

$$
\text{vec}\big(\overline{M}\big)_1 = \overline{\Omega}_1\sum_{t=1}^{T}(\Sigma_2 \otimes W_t)^{-1}\phi(x_t; \alpha, \delta)\text{vec}\big(\tilde{Y}\big)_t
$$

meaning that $\text{vec}\big(B_1\big) \sim \mathcal{N}_{n^2}(\text{vec}\big(\overline{M}\big)_1, \overline{\Omega}_1)$.

Sampling B_2. Let $\bar{Y}_t = Y_t - \phi(x_t;\alpha,\delta)B_1$. The coefficient matrix B_2 is drawn from the posterior full conditional distribution

$$
\begin{aligned}
&P(B_2|\mathbf{Y},\mathbf{W},\mathbf{x},B_1,\Sigma_1,\Sigma_2,\alpha,\delta) \propto P(B_2)P(\mathbf{Y},\mathbf{W}|\mathbf{x},B_1,B_2,\Sigma_1,\Sigma_2,\alpha,\delta)\\
&\propto \exp\Big(-\frac{1}{2}\operatorname{tr}\big(\operatorname{vec}(B_2)'(\underline{\Omega}_2\otimes\underline{\Omega}_1)^{-1}\operatorname{vec}(B_2)\big)\\
&\quad-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\big(\operatorname{vec}(Y)_t-(1-\bar{\phi}(x_t;\alpha,\delta))B_2{}'(\Sigma_2\otimes W_t)^{-1}\operatorname{vec}(Y)_t-(1-\bar{\phi}(x_t;\alpha,\delta))B_2\big)\Big)\\
&\propto \exp\Big(-\frac{1}{2}\operatorname{tr}\big(\operatorname{vec}(B_2)'\big[(\underline{\Omega}_2\otimes\underline{\Omega}_1)^{-1}+\sum_{t=1}^{T}(\Sigma_2\otimes W_t)^{-1}(1-\phi(x_t;\alpha,\delta))^2\big]\operatorname{vec}(B_2)\\
&\quad-2\sum_{t=1}^{T}\operatorname{vec}(\bar{Y})_t{}'(\Sigma_2\otimes W_t)^{-1}(1-\phi(x_t;\alpha,\delta))\operatorname{vec}(B_2)\big)\Big)\\
&\propto \exp\Big(-\frac{1}{2}\operatorname{tr}\big(\operatorname{vec}(B_2)'\overline{\Omega}_2^{-1}\operatorname{vec}(B_2)-2\operatorname{vec}(\overline{M})_2'\overline{\Omega}_2^{-1}\operatorname{vec}(B_2)\big)\Big),
\end{aligned}
$$

where

$$
\begin{aligned}
\overline{\Omega}_2 &= \Big[(\underline{\Omega}_2\otimes\underline{\Omega}_1)^{-1}+\sum_{t=1}^{T}(\Sigma_2\otimes W_t)^{-1}(1-\phi(x_t;\alpha,\delta))^2\Big]^{-1},\\
\operatorname{vec}(\overline{M})_2 &= \overline{\Omega}_2\sum_{t=1}^{T}(\Sigma_2\otimes W_t)^{-1}(1-\phi(x_t;\alpha,\delta))\operatorname{vec}(\bar{Y})_t
\end{aligned}
$$

meaning that $\operatorname{vec}(B_2)\sim\mathcal{N}_{n^2}(\operatorname{vec}(\overline{M})_2,\overline{\Omega}_2)$.

Sampling W_t. The auxiliary covariance matrices W_t, $t=1,\ldots,T$, are drawn from the posterior full conditional distribution

$$
\begin{aligned}
&P(W_t|\mathbf{Y},\mathbf{x},B_1,B_2,\Sigma_1,\Sigma_2,\nu,\alpha,\delta)\propto P(\mathbf{Y},\mathbf{W}|\mathbf{x},B_1,B_2,\Sigma_1,\Sigma_2,\nu,\alpha,\delta)\\
&\quad\propto |W_t|^{-\frac{n+\nu+n+1+n-1}{2}}\exp\Big(-\frac{1}{2}\operatorname{tr}\big((Y_t-M_t)\Sigma_2^{-1}(Y_t-M_t)'W_t^{-1}\big)-\frac{1}{2}\operatorname{tr}\big(\Sigma_1W_t^{-1}\big)\Big)\\
&\quad\propto |W_t|^{-\frac{(\nu+2n-1)+n+1}{2}}\exp\Big(-\frac{1}{2}\operatorname{tr}\big(\Sigma_1+(Y_t-M_t)\Sigma_2^{-1}(Y_t-M_t)'W_t^{-1}\big)\Big)\\
&\quad\propto \mathcal{IW}_n(\overline{W},\overline{\nu})
\end{aligned}
$$

where

$$
\overline{W}=\Sigma_1+(Y_t-M_t)\Sigma_2^{-1}(Y_t-M_t)',\qquad \overline{\nu}=\nu+2n-1.
$$

Sampling Σ_1. Given a Wishart prior, the posterior full conditional distribution for Σ_1 is conjugate and obtained as follows:

$$
\begin{aligned}
P(\Sigma_1|\mathbf{W},\nu) &\propto P(\Sigma_1|\gamma)P(\mathbf{W}|\Sigma_1,\nu)\\
&\propto \left|\Sigma_1\right|^{\frac{\underline{\kappa}_1-n-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\underline{\Psi}_1\Sigma_1\right)\right)\prod_{t=1}^{T}\left|\Sigma_1\right|^{\frac{\nu+n-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\Sigma_1 W_t^{-1}\right)\right)\\
&\propto \left|\Sigma_1\right|^{\frac{(\underline{\kappa}_1+T(\nu+n-1))-n-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\underline{\Psi}_1\Sigma_1+\sum_{t=1}^{T}[W_t^{-1}]\Sigma_1\right)\right)\\
&\propto \mathcal{W}_n(\overline{\Psi}_1,\overline{\kappa}_1)
\end{aligned}
$$

where

$$
\overline{\Psi}_1 = \left(\underline{\Psi}_1+\sum_{t=1}^{T} W_t^{-1}\right)^{-1}, \qquad \overline{\kappa}_1 = \underline{\kappa}_1 + T(\nu+n-1).
$$

Sampling Σ_2. Given an inverse Wishart prior, the posterior full conditional distribution for Σ_2 is conjugate. Using the properties of the Kronecker product and of the vectorization and trace operators, we obtain

$$
\begin{aligned}
P(\Sigma_2|\mathbf{Y},\mathbf{x},B_1,B_2,\mathbf{W},\alpha,\delta) &\propto P(\Sigma_2)P(\mathbf{Y},\mathbf{W}|\mathbf{x},B_1,B_2,\Sigma_2,\alpha,\delta)\\
&\propto \left|\Sigma_2\right|^{-\frac{\underline{\kappa}_2+Tn+n+1}{2}} \exp\left(-\frac{1}{2}\Big(\operatorname{tr}\left(\underline{\Psi}_2\Sigma_2^{-1}\right)+\operatorname{tr}\Big(\sum_{t=1}^{T}(Y_t-M_t)'W_t^{-1}(Y_t-M_t)\Sigma_2^{-1}\Big)\Big)\right)\\
&\propto \left|\Sigma_2\right|^{-\frac{\underline{\kappa}_2+Tn+n+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\underline{\Psi}_2\Sigma_2^{-1}+S_2\Sigma_2^{-1}\right)\right)\\
&\propto \mathcal{IW}_n(\overline{\Psi}_2,\overline{\kappa}_2)
\end{aligned}
$$

where we defined $S_2 = \sum_{t=1}^{T}(Y_t-M_t)'W_t^{-1}(Y_t-M_t)$ and

$$
\overline{\Psi}_2 = \underline{\Psi}_2 + S_2, \qquad \overline{\kappa}_2 = \underline{\kappa}_2 + Tn.
$$

Sampling ν. Combining the prior distribution and the likelihood in Eqs. (19)–(21), one gets

$$
\begin{aligned}
&P(\nu|\mathbf{Y},\mathbf{x},B_1,B_2,\Sigma_1,\Sigma_2,\alpha,\delta) \propto P(\nu)P(\mathbf{Y}|\mathbf{x},B_1,B_2,\Sigma_1,\Sigma_2,\nu,\alpha,\delta)\\
&\propto \nu^{\underline{a}_\nu-1}e^{-\nu/\underline{b}_\nu}\prod_{t=1}^{T}\frac{\Gamma_n\left(\frac{\nu+2n-1}{2}\right)}{\Gamma_n\left(\frac{\nu+n-1}{2}\right)}\left|I_n+\Sigma_1^{-1}(Y_t-M_t)\Sigma_2^{-1}(Y_t-M_t)'\right|^{-\frac{\nu+2n-1}{2}}\mathbb{I}_{(2,+\infty)}(\nu)\\
&\propto \nu^{\underline{a}_\nu-1}e^{-\nu/\underline{b}_\nu}\left[\frac{\Gamma_n\left(\frac{\nu+2n-1}{2}\right)}{\Gamma_n\left(\frac{\nu+n-1}{2}\right)}\right]^T\left[\prod_{t=1}^{T}\left|I_n+\Sigma_1^{-1}(Y_t-M_t)\Sigma_2^{-1}(Y_t-M_t)'\right|\right]^{-\frac{\nu+2n-1}{2}}\mathbb{I}_{(2,+\infty)}(\nu)\\
&\propto \mathcal{TG}a(\overline{a}_\nu,\overline{b}_\nu;2,+\infty)\left[\frac{\Gamma_n\left(\frac{\nu+2n-1}{2}\right)}{\Gamma_n\left(\frac{\nu+n-1}{2}\right)}\right]^T
\end{aligned}
$$

where

$$\overline{a}_v = \underline{a}_v \qquad \overline{b}_v = \left[\frac{1}{\underline{b}_v} + \frac{1}{2}\sum_{t=1}^{T} \log\left(\left|I_n + \Sigma_1^{-1}(Y_t - M_t)\Sigma_2^{-1}(Y_t - M_t)'\right|\right)\right]^{-1}.$$

We sample from this distribution using an adaptive RWMH step with truncated lognormal proposal distribution [5, 6].
Sampling α and δ. Combining the prior distribution and the likelihood in Eqs. (20)–(21) one gets

$$\begin{aligned}
P(\alpha, \delta|\mathbf{Y}, \mathbf{x}, B_1, B_2, \mathbf{W}, \Sigma_2) &\propto P(\alpha)P(\delta)P(\mathbf{Y}|\mathbf{x}, B_1, B_2, \mathbf{W}, \Sigma_2, \delta, \alpha) \\
&\propto \alpha^{\underline{a}_\alpha - 1}\exp(-\alpha/\underline{b}_\alpha)\exp(-(\delta - \underline{\mu}_\delta)^2(2\underline{\sigma}_\delta^2)^{-1}) \\
&\quad \cdot \prod_{t=1}^{T}\exp\left(-\frac{1}{2}\operatorname{tr}\left(\Sigma_2^{-1}(Y_t - M_t)'W_t^{-1}(Y_t - M_t)\right)\right) \\
&\propto \alpha^{\underline{a}_\alpha - 1}\exp(-\alpha/\underline{b}_\alpha)\exp(-(\delta - \underline{\mu}_\delta)^2(2\underline{\sigma}_\delta^2)^{-1}) \\
&\quad \cdot \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left(\Sigma_2^{-1}\big(Y_t - B_1\phi(x_t;\alpha,\delta) - B_2(1 - \phi(x_t;\alpha,\delta))\big)'W_t^{-1}\right.\right. \\
&\quad \left.\left. \cdot \big(Y_t - B_1\phi(x_t;\alpha,\delta) - B_2(1 - \phi(x_t;\alpha,\delta))\big)\right)\right)
\end{aligned}$$

We sample from this distribution using an adaptive RWMH step [5, 6, see] with lognormal proposal for the parameter α. Let $\boldsymbol{\eta} = (\log\alpha, \delta)'$ the value of the MCMC chain the iteration $(j-1)$, the candidate is generated as follows:

$$\boldsymbol{\eta}^* = \boldsymbol{\eta}^{(j-1)} + \lambda^{(j)}\text{Chol}(S^{(j-1)})\boldsymbol{\varepsilon}^{(j)}, \qquad \boldsymbol{\varepsilon}^{(j)} \sim \mathcal{N}(\mathbf{0}, I_2) \tag{25}$$

and $\boldsymbol{\eta}^{(j)} = \boldsymbol{\eta}^*$ with probability $\rho^{(j)} = \min(1, r^{(j)})$, where $r^{(j)} =$ and $\boldsymbol{\eta}^{(j)} = \boldsymbol{\eta}^{(j-1)}$ with probability $1 - \rho^{(j)}$. The updating scheme for the adaptation parameters is as follows

$$\log\lambda^{(j)} = \log\lambda^{(j-1)} + \gamma^{(j)}(\rho^{(j)} - \rho^*) \tag{26}$$

$$\boldsymbol{\mu}^{(j)} = \boldsymbol{\mu}^{(j-1)} + \gamma^{(j)}(\boldsymbol{\eta} - \boldsymbol{\mu}^{(j)}) \tag{27}$$

$$\Upsilon^{(j)} = \Upsilon^{(j-1)} + \gamma^{(j)}((\boldsymbol{\eta} - \boldsymbol{\mu}^{(j)})(\boldsymbol{\eta} - \boldsymbol{\mu}^{(j)})' - \Upsilon^{(j-1)}) \tag{28}$$

where $\gamma^{(j)} = C/j^\kappa$ with $\kappa = 0.65$, and the target acceptance rate is $\rho^* = 0.30$.

References

1. Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review, 105*(2), 564–608.
2. Acharya, V., Engle, R., & Richardson, M. (2012). Capital shortfall: A new approach to ranking and regulating systemic risks. *American Economic Review, 102*(3), 59–64.
3. Al Rousan, S. ., Sbia, R. ., & Onur Tas, B. K. (2018). Analysis of OPEC and non-OPEC members. A dynamic network analysis of the world oil market. *Energy Economics, 75,* 28–41.
4. Anderson, H. M., Nam, K., & Vahid, F. (1999). Asymmetric nonlinear smooth transition GARCH models. In *Nonlinear time series analysis of economic and financial data* (pp. 191–207). Springer.
5. Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing, 18*(4), 343–373.
6. Atchadé, Y. F., & Rosenthal, J. S. et al. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli, 11*(5), 815–828.
7. Bacon, D. W., & Watts, D. G. (1971). Estimating the transition between two intersecting straight lines. *Biometrika, 58*(3), 525–534.
8. Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third International AAAI Conference on Weblogs and Social Media*.
9. Billio, M., Casarin, R., Costola, M., & Frattarolo, L. (2019). Opinion dynamics and disagreements on financial networks. *Advances in Decision Sciences, 23*(4), 1–27.
10. Billio, M., Casarin, R., Costola, M., & Iacopini, M. (2021). COVID-19 spreading in financial networks: A semiparametric matrix regression model. *Econometrics and Statistics*, forthcoming.
11. Billio, M., Casarin, R., Costola, M., & Iacopini, M. (2021). A matrix-variate t model for networks. *Frontiers in Artificial Intelligence, 4,* 49.
12. Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics, 104*(3), 535–559.
13. Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C., Gómez-Gardenes, J., Romance, M., et al. (2014). The structure and dynamics of multilayer networks. *Physics Reports, 544*(1), 1–122.
14. Bonaccolto, G., Caporin, M., & Panzica, R. (2019). Estimation and model-based combination of causality networks among large us banks and insurance companies. *Journal of Empirical Finance, 54,* 1–21.
15. Calderón, S. A. V., & Nieto, F. H. (2017). Bayesian analysis of multivariate threshold autoregressive models with missing data. *Communications in Statistics-Theory and Methods, 46*(1), 296–318.
16. Camacho, M. (2004). Vector smooth transition regression models for US GDP and the composite index of leading indicators. *Journal of Forecasting, 23*(3), 173–196.
17. Carvalho, C. M., & West, M. et al. (2007). Dynamic matrix-variate graphical models. *Bayesian Analysis, 2*(1), 69–97.
18. Casarin, R., Iacopini, M., Molina, G., ter Horst, E., Espinasa, R., Sucre, C., & Rigobon, R. (2020). Multilayer network analysis of oil linkages. *The Econometrics Journal, 23*(2), 269–296.
19. Celeux, G., Forbes, F., Robert, C. P., & Mike, D. (2006). Titterington. Deviance information criteria for missing data models. *Bayesian Analysis, 1*(4), 651–673.
20. Chan, F., & McAleer, M. (2002). Maximum likelihood estimation of STAR and STAR-GARCH models: Theory and Monte Carlo evidence. *Journal of Applied Econometrics, 17*(5), 509–534.
21. Chan, F., & McAleer, M. (2003). Estimating smooth transition autoregressive models with GARCH errors in the presence of extreme observations and outliers. *Applied Financial Economics, 13*(8), 581–592.
22. Chen, C. W. S., & Lee, J. C. (1995). Bayesian inference of threshold autoregressive models. *Journal of Time Series Analysis, 16*(5), 483–492.

23. Chen, C. Y.-H., Härdle, W. K., & Okhrin, Y. (2019). Tail event driven networks of sifis. *Journal of Econometrics, 208*(1), 282–298.
24. Chen, E. Y., Tsay, R. S., & Chen, R. (2020). Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association, 115*(530), 775–793.
25. Chen, R., Xiao, H., & Yang, D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics, 222*(1), 539–560.
26. S., Tiwari, R. C., & Jammalamadaka, S. R. (1988). Bayes prediction in regressions with elliptical errors. *Journal of Econometrics, 38*(3), 349–360.
27. Chini, E. Z. (2018). Forecasting dynamically asymmetric fluctuations of the US business cycle. *International Journal of Forecasting, 34*(4), 711–732.
28. Cohen, M. (2010). STEO supplement: Why are oil prices so high.
29. Davis, G. M., & Ensor, K. B. (2007). Multivariate time-series analysis with categorical and continuous variables in an lstr model. *Journal of Time Series Analysis, 28*(6), 867–885.
30. Diebold, F. X., & Yılmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics, 182*(1), 119–134.
31. van Dijk, D., Teräsvirta, T., & Franses, P. H. (2002). Smooth transition autoregressive models-a survey of recent developments. *Econometric Reviews, 21*(1), 1–47.
32. Economou, E., Agnolucci, P., Fattouh, B., & De Lipis, V. (2017). A structural model of the world oil market: The role of investment dynamics and capacity constraints. *Oxford Institute for Energy Studies.–2017.–WPM*, 75:40.
33. Eitrheim, Ø., & Teräsvirta, T. (1996). Testing the adequacy of smooth transition autoregressive models. *Journal of Econometrics, 74*(1), 59–75.
34. Elliott, M., Golub, B., & Jackson, M. O. (2014). Financial networks and contagion. *American Economic Review, 104*(10), 3115–53.
35. Espinasa, R., Horst, E. T., Reyes, S. G., Manzano, O., Molina, G., & Rigobon, R. (2017). A micro-based model for world oil market. *Energy Economics, 66*, 431–449.
36. Fonseca, T. C., Ferreira, M. A. R., & Migon, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika, 95*(2), 325–333.
37. Franses, P. H., Neele, J., van Dijk, D. et al. (1998). Forecasting volatility with switching persistence GARCH models. Technical report, Econometric Institute, Erasmus University Rotterdam.
38. Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business, 53*(1), 67–78.
39. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. CRC Press.
40. Gerlach, R., & Chen, C. W. S. (2008). Bayesian inference and model comparison for asymmetric smooth transition heteroskedastic models. *Statistics and Computing, 18*(4), 391.
41. Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics, 8*(S1), S19–S40.
42. Geweke, J., & Terui, N. (1993). Bayesian threshold autoregressive models for nonlinear time series. *Journal of Time Series Analysis, 14*(5), 441–454.
43. González, A., Teräsvirta, T., & van Dijk, D. (2005). Panel smooth transition regression models. Technical report, SSE/EFI Working Paper Series in Economics and Finance.
44. González-Rivera, G. (1998). Smooth-transition GARCH models. *Studies in Nonlinear Dynamics & Econometrics, 3*(2).
45. Gupta, A. K., & Nagar, D. K. (1999). *Matrix variate distributions*. CRC Press.
46. Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics, 45*(1–2), 39–700.
47. Holmes, J. B., & Schofield, M. R. (2020). Moments of the logit-normal distribution. *Communications in statistics—Theory and methods*, forthcoming.
48. Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika, 36*(1–2), 149–176.
49. Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review, 99*(3), 1053–1069.

50. Kilian, L. (2019). Measuring global real economic activity: Do recent critiques hold up to scrutiny? *Economics Letters, 178,* 106–110.
51. Kumins, L. C., & Bamberger, R. (2005). Oil and gas disruption from hurricanes Katrina and Rita. Congressional Research Service, Library of Congress.
52. Landim, F., & Gamerman, D. (2000). Dynamic hierarchical models: An extension to matrix-variate observations. *Computational Statistics & Data Analysis, 35*(1), 11–42.
53. Leng, C., & Tang, C. Y. (2012). Sparse matrix graphical models. *Journal of the American Statistical Association, 107*(499), 1187–1200.
54. Li, K., Fang, L., & Tao, L. (2019). Bayesian panel smooth transition model with spatial correlation. *Plos One, 14*(3), e0211467.
55. Lin, C.-F. J., & Teräsvirta, T. (1994). Testing the constancy of regression parameters against continuous structural change. *Journal of Econometrics, 62*(2), 211–228.
56. Livingston, Jr. G., & Nur, D. (2017). Bayesian inference for smooth transition autoregressive (STAR) model: A prior sensitivity analysis. *Communications in Statistics-Simulation and Computation, 46*(7), 5440–5461.
57. Lubrano, M. (2000). Bayesian analysis of nonlinear time-series models with a threshold. In W. A. Barnett, D. F. Hendry, S. Hylleberg, T. Terasvirta, D. Tjostheimand, & A. W. Wurtz, (Eds.), *Nonlinear Econometric Modeling in Time Series: Proceedings of the Eleventh International Symposium in Economic Theory*. Cambridge University Press.
58. Lundbergh, S., & Teräsvirta, T. et al. (1999). *Modelling economic high-frequency time series with STAR-STGARCH models*. Tinbergen Institute.
59. Lütkepohl, H., & Netšunajev, A. (2017). Structural vector autoregressions with smooth transition in variances. *Journal of Economic Dynamics and Control, 84,* 43–57.
60. McAleer, M., & Medeiros, M. C. (2008). A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. *Journal of Econometrics, 147*(1), 104–119.
61. Osiewalski, J., & Steel, M. F. J. (1993). Bayesian marginal equivalence of elliptical regression models. *Journal of Econometrics, 59*(3), 391–403.
62. Osiewalski, J., & Steel, M. F. J. (1993). Robust Bayesian inference in elliptical regression models. *Journal of Econometrics, 57*(1), 345–363.
63. Peguin-Feissolle, A. (1994). Bayesian estimation and forecasting in non-linear models application to an LSTAR model. *Economics Letters, 46*(3), 187–194.
64. Robert, C. P. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation*. Springer.
65. Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. Springer.
66. Silvennoinen, A., & Teräsvirta, T. (2009). Modeling multivariate autoregressive conditional heteroskedasticity with the double smooth transition conditional correlation GARCH model. *Journal of Financial Econometrics, 7*(4), 373–411.
67. Silvennoinen, A., & Teräsvirta, T. (2021). Consistency and asymptotic normality of maximum likelihood estimators of a multiplicative time-varying smooth transition correlation garch model. *Econometrics and Statistics*, forthcoming.
68. Steel, M. F. J. (1998). Bayesian analysis of stochastic volatility models with flexible tails. *Econometric Reviews, 17*(2), 109–143.
69. Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association, 89*(425), 208–218.
70. Teräsvirta, T. (1998). Modeling economic relationships with smooth transition regressions. In A. Ullah & D. E. A. Giles (Eds.), *Handbook of applied economic statistics*. CRC Press.
71. Thompson, G. Z., Maitra, R., Meeker, W. Q., & Bastawros, A. F. (2020). Classification with the matrix-variate-t distribution. *Journal of Computational and Graphical Statistics, 29*(3), 668–674.
72. Tong, H. (2011). Threshold models in time series analysis-30 years on. *Statistics and its Interface, 4*(2), 107–118.
73. Triantafyllopoulos, K. (2008). Missing observation analysis for matrix-variate time series data. *Statistics & Probability Letters, 78*(16), 2647–2653.

74. Wang, D., Liu, X., & Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics, 208*(1), 231–248.
75. Wang, H., & West, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika, 96*(4), 821–834.
76. Wang, J. J. J., Chan, J. S. K., & Boris Choy, S. T. (2011). Stochastic volatility models with leverage and heavy-tailed distributions: A Bayesian approach using scale mixtures. *Computational Statistics & Data Analysis, 55*(1), 852–862.
77. Yu, L., He, Y., Kong, X., & Zhang, X. (2021). Projected estimation for large-dimensional matrix factor models. *Journal of Econometrics*, forthcoming.
78. Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. *Journal of the American Statistical Association, 71*(354), 400–405.

A Flexible Matrix-Valued Response Regression for Skewed Data

Hossein Baghishani and Jamil Ownuk

Abstract Newly applied situations generate data with complex structures in which both response and explanatory variables are three-way/matrix-valued. In real applications, outliers usually contaminate matrix-valued data. This chapter introduces a new flexible family of matrix-variate distributions that includes the matrix normal distribution as a particular member. By considering the introduced distribution for the error term, we develop a regression model for skewed matrix-valued responses with covariates that can be either a scalar, a vector, or a matrix. We extend the proposed matrix-variate regression using the envelope methodology to construct a more parsimonious parameterized model. The model fit is illustrated and compared with a new matrix-variate skew-normal model, as well as a matrix-variate normal model, on both simulated and real examples.

1 Introduction

In the framework of regression models, matrix-valued (three-way) responses are commonly observed in various applications when multivariate responses are measured on different occasions. Examples include:

- Data from crossover designs;
- Longitudinal multivariate responses;
- Multivariate growth curve data;
- Imaging data
- Multivariate temporal and spatial data, as well as spatio-temporal data.

The application we are considering here is related to a study on twin crossover bioassay [31], where the response can be treated as a matrix-valued random variable.

H. Baghishani (✉) · J. Ownuk
Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran
e-mail: hbaghishani@shahroodut.ac.ir

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_8

Modeling matrix-valued data has received a lot of attention recently. Considering applications with only matrix-valued covariates, we can refer to [9–11, 24, 27, 32]. For dealing with matrix-valued responses, we can mention the works of [8, 12, 22, 23, 30].

In most of the proposed models, it is assumed that the error term follows a matrix-variate normal (MatN) distribution. However, such an assumption is often violated due to the non-symmetric/skewed errors and the presence of outliers in the data. There are a few works on modeling skewed matrix data. For example, [15, 18] developed four skewed matrix-variate distributions. References [16, 17] also proposed using the skewed distributions in a class of mixture models for clustering and classification. Other related works include [1, 21, 33].

Introducing relevant regression models for dealing with skewed matrix-valued data is relatively limited in the literature. To fill this gap, we propose a regression model for skewed matrix-valued responses with covariates that can be either a scalar, a vector, or a matrix. To this end, we first develop a new flexible matrix-variate distribution by extending the recently introduced univariate distribution of [26], which is named unimodal-bimodal normal (UBN). The matrix-variate UBN (MatUBN) distribution incorporates skewness as well as other flexible features such as bimodality; it also includes the MatN distribution as a particular member. Following [12], we also develop an envelope extension of the regression model. By applying the envelope method, when the matrix-variate dimension is large, we can achieve dimension reduction in the analysis by excluding redundant information, leading to substantial efficiency gains in estimation.

The remainder of this chapter is laid out as follows. In Sect. 2, we present a detailed background as well as construct the MatUBN distribution. Section 3 introduces the new class of matrix variate regression model incorporating a skewed error term with a MatUBN distribution. Furthermore, the development of envelope methods for the proposed matrix-variate regression is exhibited. A simulation study and a real data application displaying the benefit of the proposed model are discussed in Sects. 4 and 5. Finally, Sect. 6 achieves some conclusions and possible avenues for future work.

2 Background

Generally, we suppose that the data takes the form of a matrix with r rows (representing occasions) and m columns (representing variables) for each subject. Therefore, we have an $r \times m$ observed matrix for each subject.

Matrix-variate Normal Distribution

The MatN distribution is the most well known among the matrix-variate distributions, and its different properties, as well as parameter estimation, are thoroughly

reviewed in the literature ([13, 25]). A random matrix $\boldsymbol{\mathcal{X}}$ of dimension $r \times m$, with the realization $\boldsymbol{X}$, follows a MatN distribution with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ if its probability density function is as

$$f(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \frac{1}{(2\pi)^{\frac{rm}{2}} (\det\boldsymbol{\Sigma}_1)^{\frac{m}{2}} (\det\boldsymbol{\Sigma}_2)^{\frac{r}{2}}} e^{-\frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{X}-\boldsymbol{\mu})'\}},$$

where $\boldsymbol{\mu}$ is an $r \times m$ location (mean) matrix, $\boldsymbol{\Sigma}_1$ is an $r \times r$ symmetric and positive definite scale matrix for the rows of $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\Sigma}_2$ is an $m \times m$ symmetric and positive definite scale matrix for the columns of $\boldsymbol{\mathcal{X}}$ and det as well as tr$(\cdot)$ mean the determinant and trace operations, respectively. We denote the MatN distribution by $\mathrm{N}_{r,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$. The MatN distribution of dimension $r \times m$ is a special case of the rm-dimensional normal distribution when its covariance matrix is separable in the form $\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$ in which $\otimes$ denotes the Kronecker product. Therefore, it can be confirmed that

$$\mathrm{vec}(\boldsymbol{\mathcal{X}}) \sim \mathrm{N}_{rm}(\mathrm{vec}(\boldsymbol{\mu}), \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1)$$

iff $\boldsymbol{\mathcal{X}} \sim \mathrm{N}_{r,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$, where vec$(\cdot)$ denotes the vectorization operator.

If $\boldsymbol{\mathcal{X}} \sim \mathrm{N}_{r,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ and $\mathbf{A}$ is any constant matrix with proper dimension, then some useful properties of the MatN distribution, as given in [20], are as follows:

(1) $\boldsymbol{\mathcal{X}}' \sim \mathrm{N}_{m,r}\left(\boldsymbol{\mu}', \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_1\right)$.
(2) Let $\mathrm{cov}_c(\boldsymbol{\mathcal{X}}) = \mathrm{E}[(\boldsymbol{\mathcal{X}} - \mathrm{E}[\boldsymbol{\mathcal{X}}])(\boldsymbol{\mathcal{X}} - \mathrm{E}[\boldsymbol{\mathcal{X}}])']$ be the covariance matrix over the columns of $\boldsymbol{\mathcal{X}}$ and let $\mathrm{cov}_r(\boldsymbol{\mathcal{X}}) = \mathrm{E}[(\boldsymbol{\mathcal{X}} - \mathrm{E}[\boldsymbol{\mathcal{X}}])'(\boldsymbol{\mathcal{X}} - \mathrm{E}[\boldsymbol{\mathcal{X}}])]$ be the covariance matrix over the rows of $\boldsymbol{\mathcal{X}}$. Hence, if cov(vec$(\boldsymbol{\mathcal{X}})$) has the Kronecker structure, then

$$\mathrm{cov}_c(\boldsymbol{\mathcal{X}}) = \mathrm{tr}\left(\boldsymbol{\Sigma}_2\right)\boldsymbol{\Sigma}_1, \quad \mathrm{cov}_r(\boldsymbol{\mathcal{X}}) = \mathrm{tr}\left(\boldsymbol{\Sigma}_1\right)\boldsymbol{\Sigma}_2.$$

Consequently, we also call $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ the column and row covariance matrices of $\boldsymbol{\mathcal{X}}$, respectively.
(3) $\mathrm{E}[(\boldsymbol{\mathcal{X}} - \mathrm{E}[\boldsymbol{\mathcal{X}}])\,\mathbf{A}\,(\boldsymbol{\mathcal{X}} - \mathrm{E}[\boldsymbol{\mathcal{X}}])'] = \mathrm{tr}\left(\mathbf{A}'\boldsymbol{\Sigma}_2\right)\boldsymbol{\Sigma}_1, \mathbf{A} \in \mathbb{R}^{m\times m}$
(4) $\mathrm{E}[(\boldsymbol{\mathcal{X}} - \mathrm{E}[\boldsymbol{\mathcal{X}}])'\,\mathbf{A}\,(\boldsymbol{\mathcal{X}} - \mathrm{E}[\boldsymbol{\mathcal{X}}])] = \mathrm{tr}\left(\boldsymbol{\Sigma}_1\mathbf{A}'\right)\boldsymbol{\Sigma}_2, \mathbf{A} \in \mathbb{R}^{r\times r}$.

In the context of matrix-valued data analysis, the MatN distribution is extensively applied. However, there is a lack of skewed matrix-variate methodology in the regression framework. In the following, we will develop a flexible skewed matrix-variate distribution based on the UBN distribution. We first briefly introduce the univariate UBN distribution.

Unimodal–bimodal Normal (UBN) Distribution

The UBN distribution, introduced by [26], is a generalized version of the normal distribution that can account for skewness and bimodality. This distribution is more flex-

ible than the normal distribution and many of its generalizations. A univariate random variable Z follows a standard UBN distribution, denoted by $Z \sim \text{UBN}(0, 1, k, a)$, when its probability density function could be written as

$$f(z|k,a) = c_{k,a} e^{k|z|} \phi(z-a) \qquad z \in \mathbb{R},$$

where $k \in \mathbb{R}$ and $a \in \mathbb{R}$ are the shape parameters, $c_{k,a}^{-1} = e^{\frac{k^2}{2}} \left(e^{ka}\Phi(k+a) + e^{-ka}\Phi(k-a)\right)$, and $\phi(\cdot)$ as well as $\Phi(\cdot)$ denote the density and distribution functions of standard normal distribution, respectively. The first and second moments of Z are given by

$$\begin{aligned}
\text{E}(Z) &= e_{k,a} = \frac{(k+a)\,\text{e}^{ka}\Phi(k+a) - (k-a)\,\text{e}^{-ka}\Phi(k-a)}{\text{e}^{ka}\Phi(k+a) + \text{e}^{-ka}\Phi(k-a)} \\
\text{E}(Z^2) &= e_{k,a}^2 = \frac{(1+(k+a)^2)\text{e}^{ka}\Phi(k+a) + (1+(k-a)^2)\text{e}^{-ka}\Phi(k-a) + 2k\text{e}^{-ka}\phi(k-a)}{\text{e}^{ka}\Phi(k+a) + \text{e}^{-ka}\Phi(k-a)}.
\end{aligned}$$

When $a \neq 0$, the distribution is skewed; for $k = a = 0$, we will have a normal distribution.

Skewed Matrix-Variate UBN (MatUBN) Distribution

To construct the matrix-variate UBN distribution, we assume that rm independent random variables Z_{ij} follow an identical $\text{UBN}(0, 1, k, a)$ distribution. Therefore, the random matrix $\mathcal{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_m)$ of dimension $r \times m$ has the probability density function as

$$f(\mathbf{Z}) = \Pi_{i=1}^{r} \Pi_{j=1}^{m} f(z_{ij}),$$

where, for $j = 1, \ldots, m$, $\mathbf{Z}_j = (Z_{1j}, Z_{2j}, \ldots, Z_{rj})'$. In a matrix notation, it could be shown that

$$f(\mathbf{Z}) = c_{k,a}^{rm} (2\pi)^{-\frac{rm}{2}} e^{k\text{tr}\left\{\mathbf{J}'_{r\times m}|\mathbf{Z}|\right\}} e^{-\frac{1}{2}\text{tr}\{(\mathbf{Z}-a\mathbf{J}_{r\times m})(\mathbf{Z}-a\mathbf{J}_{r\times m})'\}}, \tag{1}$$

where k and a are the shape parameters, $\mathbf{J}_{r\times m}$ is an $r \times m$ matrix of ones, and $|\cdot|$ represents the elementwise absolute value of a matrix or a vector.

The proof of the following proposition is easily achieved by using the same arguments for the given results of the MatN distribution.

Proposition 1 *If the random matrix $\mathcal{Z}$ has the probability density given in (1), for any constant matrix $\mathbf{A}$ with proper dimension, we have*

$$
\begin{aligned}
\mathrm{E}[\mathcal{Z}] &= e_{k,a}\mathbf{J}_{r\times m}\\
\mathrm{cov}_c(\mathcal{Z}) &= v_{k,a}\,\mathrm{tr}\left(\mathbf{I}_{m\times m}\right)\mathbf{I}_{r\times r}\\
\mathrm{cov}_r(\mathcal{Z}) &= v_{k,a}\,\mathrm{tr}\left(\mathbf{I}_{r\times r}\right)\mathbf{I}_{m\times m}\\
\mathrm{E}[(\mathcal{Z}-\mathrm{E}[\mathcal{Z}])\,\mathbf{A}\,(\mathcal{Z}-\mathrm{E}[\mathcal{Z}])'] &= v_{k,a}\,\mathrm{tr}\left(\mathbf{A}'\right)\mathbf{I}_{r\times r},\ \mathbf{A}\in\mathbb{R}^{m\times m}\\
\mathrm{E}[(\mathcal{Z}-\mathrm{E}[\mathcal{Z}])'\,\mathbf{A}\,(\mathcal{Z}-\mathrm{E}[\mathcal{Z}])] &= v_{k,a}\,\mathrm{tr}\left(\mathbf{A}'\right)\mathbf{I}_{m\times m},\ \mathbf{A}\in\mathbb{R}^{r\times r},
\end{aligned}
$$

where $v_{k,a}=e_{k,a}^2-(e_{k,a})^2$ *and* $\mathbf{I}_{r\times r}$ *is an* $r\times r$ *identity matrix.*

Now, we can construct the MatUBN distribution. Let $\boldsymbol{\mathcal{X}}=\boldsymbol{\Sigma}_1^{\frac{1}{2}}\mathcal{Z}\boldsymbol{\Sigma}_2^{\frac{1}{2}}+\boldsymbol{\mu}$, where $\boldsymbol{\mu}\in\mathbb{R}^{r\times m}$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are $r\times r$ and $m\times m$ positive definite scale matrices, respectively, and $\boldsymbol{\Sigma}^{\frac{1}{2}}$ means the square root of matrix $\boldsymbol{\Sigma}$. Hence, $\boldsymbol{\mathcal{X}}$ follows a MatUBN distribution and has the following probability density:

$$
\begin{aligned}
f(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2) = {} & \frac{c_{k,a}^{rm}e^{k\mathrm{tr}\left\{\mathbf{J}'_{r\times m}\left|\boldsymbol{\Sigma}_1^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\right|\right\}}}{(2\pi)^{\frac{rm}{2}}\,(\det\boldsymbol{\Sigma}_1)^{\frac{m}{2}}\,(\det\boldsymbol{\Sigma}_2)^{\frac{r}{2}}}\times\\
& e^{-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{X}-\boldsymbol{\mu}-a\boldsymbol{\Sigma}_1^{\frac{1}{2}}\mathbf{J}_{r\times m}\boldsymbol{\Sigma}_2^{\frac{1}{2}}\right)\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{X}-\boldsymbol{\mu}-a\boldsymbol{\Sigma}_1^{\frac{1}{2}}\mathbf{J}_{r\times m}\boldsymbol{\Sigma}_2^{\frac{1}{2}}\right)'\right\}},
\end{aligned}
\tag{2}
$$

where k and a are the shape parameters and $\boldsymbol{\mu}$ is the location matrix. Furthermore, according to Proposition 2 in the following, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ could be interpreted as column and row covariance matrices. For $a=0$, the distribution is symmetric and asymmetric otherwise. The MatN distribution is a special case of the MatUBN family when $k=a=0$. We denote this distribution by MatUBN$(\boldsymbol{\mu},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,k,a)$.

Proposition 2 *For the random matrix* $\boldsymbol{\mathcal{X}}\sim$ MatUBN $(\boldsymbol{\mu},\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,k,a)$ *and any constant matrix* $\mathbf{A}$ *with proper dimension, we have*

$$
\begin{aligned}
\mathrm{E}[\boldsymbol{\mathcal{X}}] &= e_{k,a}\boldsymbol{\Sigma}_1^{\frac{1}{2}}\mathbf{J}_{r\times m}\boldsymbol{\Sigma}_2^{\frac{1}{2}}+\boldsymbol{\mu}\\
\mathrm{cov}_c(\boldsymbol{\mathcal{X}}) &= v_{k,a}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_2\right)\boldsymbol{\Sigma}_1\\
\mathrm{cov}_r(\boldsymbol{\mathcal{X}}) &= v_{k,a}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_1\right)\boldsymbol{\Sigma}_2\\
\mathrm{E}[(\boldsymbol{\mathcal{X}}-\mathrm{E}[\boldsymbol{\mathcal{X}}])\,\mathbf{A}\,(\boldsymbol{\mathcal{X}}-\mathrm{E}[\boldsymbol{\mathcal{X}}])'] &= v_{k,a}\,\mathrm{tr}\left(\mathbf{A}'\boldsymbol{\Sigma}_2\right)\boldsymbol{\Sigma}_1,\ \mathbf{A}\in\mathbb{R}^{m\times m}\\
\mathrm{E}[(\boldsymbol{\mathcal{X}}-\mathrm{E}[\boldsymbol{\mathcal{X}}])'\,\mathbf{A}\,(\boldsymbol{\mathcal{X}}-\mathrm{E}[\boldsymbol{\mathcal{X}}])] &= v_{k,a}\,\mathrm{tr}\left(\boldsymbol{\Sigma}_1\mathbf{A}'\right)\boldsymbol{\Sigma}_2,\ \mathbf{A}\in\mathbb{R}^{r\times r}.
\end{aligned}
$$

Interestingly, a multivariate version of UBN (MUBN) distribution could be obtained when $m=1$. In this case, the corresponding probability density of vector $\boldsymbol{\mathcal{X}}$ is

$$
f(\boldsymbol{x})=\frac{c_{k,a}^{r}e^{k\mathbf{J}'_{r\times 1}\left|\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})\right|}}{(2\pi)^{\frac{r}{2}}\,(\det\boldsymbol{\Sigma})^{\frac{1}{2}}}e^{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}-a\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{J}_{r\times 1}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}-a\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{J}_{r\times 1}\right)},
\tag{3}
$$

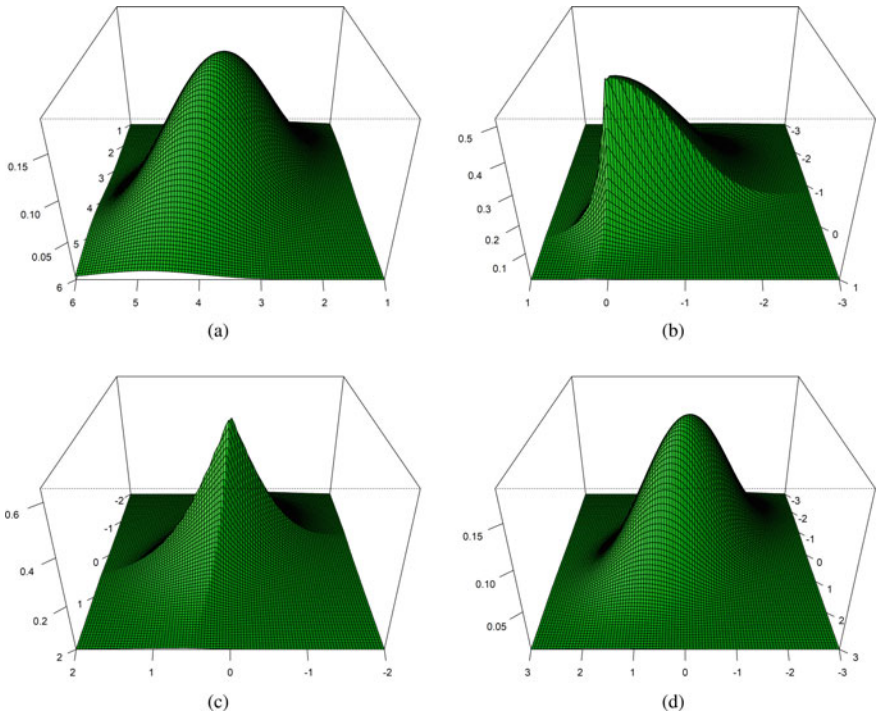


Fig. 1 Plots of the MUBN density for $\boldsymbol{\mu} = \mathbf{0}$, vech $(\boldsymbol{\Sigma}) = (1, 0.5, 1)'$ and some selected values for shape parameters: **(a)** $k = -1$ and $a = 4$; **(b)** $k = -2$ and $a = -2$; **(c)** $k = -2$ and $a = 0$; **(d)** $k = 0$ and $a = 0$

where k and a are the shape parameters, $\boldsymbol{\mu}$ is the location vector of length r and $\boldsymbol{\Sigma}$ is the $r \times r$ positive definite scale matrix. We denote it by MUBN$(\boldsymbol{\mu}, \boldsymbol{\Sigma}, k, a)$.

Proposition 3 *For* $\boldsymbol{X} \sim$ MUBN $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, k, a)$ *we have*

$$\mathrm{E}[\boldsymbol{X}] = e_{k,a} \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{J}_{r\times 1} + \boldsymbol{\mu}$$
$$\mathrm{E}[(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])\,(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])'] = v_{k,a} \boldsymbol{\Sigma}.$$

Figure 1 shows some of the possible surfaces of a bivariate UBN density for selected values of the parameters. The figure illustrates the appropriate flexibility of the distribution. Similar to the MatN distribution, an attractive property of the MatUBN distribution is its close relationship to the MUBN, presented in the following proposition.

Proposition 4 *If the random matrix* $\boldsymbol{\mathcal{X}} \sim$ MatUBN $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, k, a)$*, we have*

$$\mathrm{vec}(\boldsymbol{\mathcal{X}}) \sim \mathrm{MUBN}(\mathrm{vec}(\boldsymbol{\mu}), \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1, k, a).$$

3 Proposed Regression Model

Our goal here is to develop a regression model for skewed matrix-valued responses and a set of covariates. Let $\boldsymbol{Y}_i \in \mathbb{R}^{r\times m}$ be the observed matrix for each subject, with $i = 1, \ldots, n$. The non-stochastic covariates can be matrix-valued $\boldsymbol{X}_i \in \mathbb{R}^{p_1\times p_2}$, vector-valued $\boldsymbol{X}_i \in \mathbb{R}^p$, or scalar. The first approach that could come to mind is using the vec operator to stack the columns of a matrix and introduce a standard vector-variate regression:

$$\text{vec}(\boldsymbol{\mathcal{Y}}_i) = \boldsymbol{\mu} + \boldsymbol{\gamma}\,\text{vec}(\boldsymbol{X}_i) + \boldsymbol{\epsilon}_i, \tag{4}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{rm}$, $\boldsymbol{\gamma} \in \mathbb{R}^{rm\times p_1 p_2}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^{rm}$ is the error vector with covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{rm\times rm}$. When imposing a sparsity assumption is not applicable, such a model may be over-parameterized and hard to interpret. [12] proposed a comparably parsimonious model that reflects the underlying structure of the matrix-variate response and covariates as follows:

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mu} + \boldsymbol{\beta}_1 \boldsymbol{X} \boldsymbol{\beta}_2' + \boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{r\times m}$ is the overall mean, $\boldsymbol{\beta}_1 \in \mathbb{R}^{r\times p_1}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{m\times p_2}$ are the row and column coefficient matrices, which are uniquely defined up to a proportionality constant, and $\boldsymbol{\epsilon} \in \mathbb{R}^{r\times m}$ is the zero-mean matrix-variate normally-distibuted random error with the covariance matrix $\text{cov}(\text{vec}(\boldsymbol{\epsilon})) = \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$, that is assumed be independent of $\boldsymbol{X}$.

The proposed model of [12] provides a general framework for regression modeling, including simple regression, multiple regression, and multivariate multiple regression as special cases. It also has significant advantages over the vectorized model (4):

- We will lose dependency information of a matrix-variate response by vectorizing the response or modeling the row or column vectors separately. Therefore, we will fail to take the underlying data structure into account. Model (5) avoids the issue by applying the matrix-variate structure of the data.
- The covariances of the column vectors of $\boldsymbol{\epsilon}$ are all proportional to $\boldsymbol{\Sigma}_1$ and the covariances of the row vectors of $\boldsymbol{\epsilon}$ are all proportional to $\boldsymbol{\Sigma}_2$. These relations are desirable for matrix-valued variables, especially for multivariate repeated measures and multivariate longitudinal data.
- An essential advantage, especially when the matrix dimensions of response and covariates are relatively high, is reducing the number of parameters from $rmp_1p_2 + \frac{rm(rm+1)}{2}$ to $rp_1 + mp_2 - 1 + (\frac{r(r+1)}{2} + \frac{m(m+1)}{2} - 1)$. This results in a significant efficiency gain.
- Model (5) can be implemented when the sample size n is smaller than the dimension of $\text{vec}(\boldsymbol{X})$, while model (4) is not useful in such a case without forcing additional constraints.

It is common to assume that the error term is zero-mean; however, this assumption may not be appropriate for skewed data. Accordingly, we propose a new setting proper for the skewed matrix-valued response by incorporating a MatUBN error term in model (5).

Model Formulation

To develop a skewed matrix-variate regression model, we assume that the error term in (5) has a MatUBN distribution with zero location $\boldsymbol{\mu} = \mathbf{0}$, scale matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, and shape parameters k and a. Then, $\boldsymbol{\epsilon}$ has covariance matrix $\text{cov}(\text{vec}(\boldsymbol{\epsilon})) = \upsilon_{k,a}\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$, where as aforementioned, $\boldsymbol{\Sigma}_1 = \frac{\text{cov}_c(\boldsymbol{\epsilon})}{\upsilon_{k,a}\text{tr}(\boldsymbol{\Sigma}_2)}$ and $\boldsymbol{\Sigma}_2 = \frac{\text{cov}_r(\boldsymbol{\epsilon})}{\upsilon_{k,a}\text{tr}(\boldsymbol{\Sigma}_1)}$ are called the column and row covariance matrices of $\boldsymbol{\epsilon}$. Note that $\text{E}(\boldsymbol{\epsilon}) = e_{k,a}\boldsymbol{\Sigma}_1^{\frac{1}{2}}\mathbf{J}_{r\times m}\boldsymbol{\Sigma}_2^{\frac{1}{2}}$, which is not generally zero.

The column coefficient matrix $\boldsymbol{\beta}_2$ is restricted to have unit Frobenius norm and positive element in its first row and column for identifiability. As [12] have noticed, various constraints essentially change the scale of the parameter matrices and do not affect model fitting or prediction. Moreover, we need $\boldsymbol{\Sigma}_2$ to have unit Frobenius norm and positive first diagonal element to identify the two covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ uniquely. Note that if no constraints are given on $\boldsymbol{\beta}_2$ and $\boldsymbol{\Sigma}_2$, the Kronecker products $\boldsymbol{\beta}_2 \otimes \boldsymbol{\beta}_1$ and $\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$ are still identifiable.

For fitting the model, we use the maximum likelihood (ML) approach. The log-likelihood of the model is given by

$$\begin{aligned}\boldsymbol{L} = {} & rm\log(c_{k,a}) + k\text{tr}\left\{\mathbf{J}'_{r\times m}\left|\boldsymbol{\Sigma}_1^{-\frac{1}{2}}\left(\boldsymbol{Y}-\boldsymbol{\mu}-\boldsymbol{\beta}_1\boldsymbol{X}\boldsymbol{\beta}_2'\right)\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\right|\right\} \\ & -\frac{rm}{2}\log(2\pi) - \frac{m}{2}\log\left(\det\boldsymbol{\Sigma}_1\right) - \frac{r}{2}\log\left(\det\boldsymbol{\Sigma}_2\right) \\ & -\frac{1}{2}\text{tr}\left\{\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}-\boldsymbol{\beta}_1\boldsymbol{X}\boldsymbol{\beta}_2' - a\boldsymbol{\Sigma}_1^{\frac{1}{2}}\mathbf{J}_{r\times m}\boldsymbol{\Sigma}_2^{\frac{1}{2}}\right)\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{Y}-\boldsymbol{\mu}-\boldsymbol{\beta}_1\boldsymbol{X}\boldsymbol{\beta}_2' - a\boldsymbol{\Sigma}_1^{\frac{1}{2}}\mathbf{J}_{r\times m}\boldsymbol{\Sigma}_2^{\frac{1}{2}}\right)'\right\}.\end{aligned}$$

Unfortunately, closed-form ML estimators of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$ do generally not exist. [12] proposed a two-step iterative algorithm to compute the ML estimators for their model. However, it is not possible to develop such an algorithm for our model. Specifically, for the known $\boldsymbol{\beta}_2$ and $\boldsymbol{\Sigma}_2$ in the MatN model, closed-form estimators for the remaining parameters could be determined. It is not the case for the MatUBN model due to its two additional shape parameters as well as the presence of the absolute value function in the likelihood. The total number of parameters in our model is $\mathcal{T} = rm + rp_1 + mp_2 + \frac{r(r+1)}{2} + \frac{m(m+1)}{2} + 2$. When the dimensions of parameters are not large, we can directly maximize the log-likelihood of the model, supporting the normalization constraints on $\boldsymbol{\beta}_2$ and $\boldsymbol{\Sigma}_2$. To do this, we use the BFGS quasi-Newton numerical optimization method (proposed by [2, 14, 19, 29] independently), implemented in the `optim` function of `R` package ([28]). Moreover, to approximate the standard errors of ML estimates of parameters, we can obtain them

by taking the square roots of the diagonal elements of the observed information matrix evaluated at ML estimates. Such a matrix can also be obtained numerically using software packages such as the `optim` function. However, these estimates of standard errors could not be efficient due to propagating errors in numerical approximation of second derivatives. Hence, we use a bootstrap approach by treating the covariates as non-stochastic and resampling from the residuals of the fitted regression model to compute standard errors of the ML estimates of parameters.

The existence of the ML estimators with the MatN error term requires mild conditions [12], which implies the model can be directly applied to the high-dimensional setting where $n < \dim(\mathrm{vec}(\boldsymbol{X}))$. Given the straightforward generalized structure of our model, it would be easy to show that a similar result is established for the skewed matrix-variate regression model.

With high dimension $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\mathcal{T}$ is large, and, consequently, we will lose the efficiency in fitting the model. To address the issue, we apply the envelope approach for the introduced model.

Extending the Model Using Envelope Formulation

The idea of envelope methodology for parsimonious parameterizations and producing substantial gains in efficiency was introduced by [5] in a multivariate linear regression model. Furthermore, they illustrated that these gains would influence other tasks such as prediction. In effect, envelopes separate the relevant and irrelevant parts of the responses and then improve efficiency. As a few related works, we can suggest studying [3, 4, 6, 7].

As inherited from the multivariate linear regression model, the motivation for enveloping in the matrix-variate regression stems from the fact that there may be linear combinations of the rows or columns of $\boldsymbol{\mathcal{Y}}$ whose distribution is invariant to changes in $\boldsymbol{X}$. [5] called such linear combinations $\boldsymbol{X}$-invariants. To establish the envelope structure for the skewed matrix-variate regression model, we will use the similar notations and definitions employed by [12].

Let $\mathrm{span}(\mathbf{B})$ denote the subspace of $\mathbb{R}^r$ that is spanned by the columns of matrix $\mathbf{B} \in \mathbb{R}^{r\times u}$. Let also $\mathbf{P}_{\mathbf{B}} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{\dagger}\mathbf{B}'$ is the projection onto $\mathrm{span}(\mathbf{B})$ and $\mathbf{Q}_{\mathbf{B}} = \mathbf{I}_r - \mathbf{P}_{\mathbf{B}}$ is the orthogonal projection, where "†" denotes the Moore–Penrose inverse. It is said that a subspace $\mathcal{S} \subset \mathbb{R}^r$ is a reducing subspace of $\mathbf{M} \in \mathbb{R}^{r\times r}$ whenever $\mathbf{M}$ is decomposed by $\mathcal{S}$ as $\mathbf{M} = \mathbf{P}_{\mathcal{S}}\mathbf{M}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{M}\mathbf{Q}_{\mathcal{S}}$, where $\mathbf{P}_{\mathcal{S}}$ is the projection onto $\mathcal{S}$ and $\mathbf{Q}_{\mathcal{S}}$ is the orthogonal projection. In this case, we say that $\mathcal{S}$ reduces $\mathbf{M}$.

Ding and Cook [12] extended enveloping method to the matrix-variate regression by allowing the possibility that there are $\boldsymbol{X}$-invariants in both the rows and the columns of $\boldsymbol{\mathcal{Y}}$. To this end, they suppose that there are subspaces $\mathcal{S}_L \subset \mathbb{R}^r$ and $\mathcal{S}_R \subset \mathbb{R}^m$ so that

$$(a)\ \mathbf{Q}_{\mathcal{S}_L}\boldsymbol{\mathcal{Y}}|\boldsymbol{X} \sim \mathbf{Q}_{\mathcal{S}_L}\boldsymbol{\mathcal{Y}} \quad \& \quad (b)\ \mathrm{cov}_c(\mathbf{P}_{\mathcal{S}_L}\boldsymbol{\mathcal{Y}}, \mathbf{Q}_{\mathcal{S}_L}\boldsymbol{\mathcal{Y}}|\boldsymbol{X}) = \mathbf{0} \tag{6}$$

$$(a)\ \boldsymbol{\mathcal{Y}}\mathbf{Q}_{\mathcal{S}_R}|\boldsymbol{X} \sim \boldsymbol{\mathcal{Y}}\mathbf{Q}_{\mathcal{S}_R} \quad \& \quad (b)\ \mathrm{cov}_r(\boldsymbol{\mathcal{Y}}\mathbf{P}_{\mathcal{S}_R}, \boldsymbol{\mathcal{Y}}\mathbf{Q}_{\mathcal{S}_R}|\boldsymbol{X}) = \mathbf{0}. \tag{7}$$

Let $\mathcal{B}_1 = \mathrm{span}(\boldsymbol{\beta}_1)$ and $\mathcal{B}_2 = \mathrm{span}(\boldsymbol{\beta}_2)$. Conditions (6a) and (7a) result in the marginal distributions of $\mathbf{Q}_{\mathcal{S}_L}\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Y}}\mathbf{Q}_{\mathcal{S}_R}$ are not dependent on $\boldsymbol{X}$. This is equivalent to $\mathcal{B}_1 \subset \mathcal{S}_L$ and $\mathcal{B}_2 \subset \mathcal{S}_R$. Conditions (6b) and (7b) imply that $\mathbf{Q}_{\mathcal{S}_L}\boldsymbol{\mathcal{Y}}$ is not affected by changes in $\boldsymbol{X}$ through a linear association with $\mathbf{P}_{\mathcal{S}_L}\boldsymbol{\mathcal{Y}}$ and that changes in $\boldsymbol{X}$ do not change $\boldsymbol{\mathcal{Y}}\mathbf{Q}_{\mathcal{S}_R}$ through a linear association with $\boldsymbol{\mathcal{Y}}\mathbf{P}_{\mathcal{S}_R}$. Moreover, condition (6b) holds iff $\mathbf{P}_{\mathcal{S}_L}\boldsymbol{\Sigma}_1\mathbf{Q}_{\mathcal{S}_L} = \mathbf{0}$, which is equivalent to requiring that $\mathcal{S}_L$ reduces $\boldsymbol{\Sigma}_1$, and hence, $\boldsymbol{\Sigma}_1 = \mathbf{P}_{\mathcal{S}_L}\boldsymbol{\Sigma}_1\mathbf{P}_{\mathcal{S}_L} + \mathbf{Q}_{\mathcal{S}_L}\boldsymbol{\Sigma}_1\mathbf{Q}_{\mathcal{S}_L}$. Similarly, $\mathcal{S}_R$ reduces $\boldsymbol{\Sigma}_2$; therefore, $\boldsymbol{\Sigma}_2 = \mathbf{P}_{\mathcal{S}_R}\boldsymbol{\Sigma}_2\mathbf{P}_{\mathcal{S}_R} + \mathbf{Q}_{\mathcal{S}_R}\boldsymbol{\Sigma}_2\mathbf{Q}_{\mathcal{S}_R}$.

The intersection of all reducing subspaces $\mathcal{S}_L$ of $\boldsymbol{\Sigma}_1$ that contain $\mathcal{B}_1$ is called the $\boldsymbol{\Sigma}_1$-envelope of $\mathcal{B}_1$ and denoted by $\mathcal{E}_{\boldsymbol{\Sigma}_1}(\mathcal{B}_1) \equiv \mathcal{E}_1$. As well, the intersection of all reducing subspaces $\mathcal{S}_R$ of $\boldsymbol{\Sigma}_2$ that contain $\mathcal{B}_2$ is the $\boldsymbol{\Sigma}_2$-envelope of $\mathcal{B}_2$, denoted as $\mathcal{E}_{\boldsymbol{\Sigma}_2}(\mathcal{B}_2) \equiv \mathcal{E}_2$. The subspaces $\mathcal{E}_{\boldsymbol{\Sigma}_1}(\mathcal{B}_1)$ and $\mathcal{E}_{\boldsymbol{\Sigma}_2}(\mathcal{B}_2)$ are the fundamental constructs that provide row and column reduction of $\boldsymbol{\mathcal{Y}}$ and are uniquely defined. Indeed, they indicate that changes in $\boldsymbol{X}$ change $\boldsymbol{\mathcal{Y}}$ only through $\mathbf{P}_{\mathcal{E}_1}\boldsymbol{\mathcal{Y}}\mathbf{P}_{\mathcal{E}_2}$, and they take the promise of a much better estimation of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

We can now use $\mathcal{E}_{\boldsymbol{\Sigma}_1}(\mathcal{B}_1)$ and $\mathcal{E}_{\boldsymbol{\Sigma}_2}(\mathcal{B}_2)$ to apply the envelope methodology to reparametrize the skewed version of model (5). Let $\mathbf{L} \in \mathbb{R}^{r\times u_1}(u_1 \le r)$ and $\mathbf{R} \in \mathbb{R}^{m\times u_2}(u_2 \le m)$ be semiorthogonal bases of $\mathcal{E}_{\boldsymbol{\Sigma}_1}(\mathcal{B}_1)$ and $\mathcal{E}_{\boldsymbol{\Sigma}_2}(\mathcal{B}_2)$, respectively, where u_1 and u_2 are the known dimensions of the corresponding row and column envelopes. There are two coordinate matrices $\boldsymbol{\eta}_1 \in \mathbb{R}^{u_1\times p_1}$ and $\boldsymbol{\eta}_2 \in \mathbb{R}^{u_2\times p_2}$ such that $\boldsymbol{\beta}_1 = \mathbf{L}\boldsymbol{\eta}_1$ and $\boldsymbol{\beta}_2 = \mathbf{R}\boldsymbol{\eta}_2$. By letting $(\mathbf{L}, \mathbf{L}_0)$ and $(\mathbf{R}, \mathbf{R}_0)$ to be orthogonal matrices, the model (5) with the MatUBN error term can be reparameterized as the following envelope model:

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\eta}_1 X \boldsymbol{\eta}_2'\mathbf{R}' + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{\Sigma}_1 = \mathbf{L}\boldsymbol{\Omega}_1\mathbf{L}' + \mathbf{L}_0\boldsymbol{\Omega}_{10}\mathbf{L}_0'$, $\boldsymbol{\Sigma}_2 = \mathbf{R}\boldsymbol{\Omega}_2\mathbf{R}' + \mathbf{R}_0\boldsymbol{\Omega}_{20}\mathbf{R}_0'$, and $\boldsymbol{\Omega}_j > 0$ and $\boldsymbol{\Omega}_{j0} > 0$, for $j = 1, 2$, are unknown. A similar reparameterization is used for $\boldsymbol{\Sigma}_j$. The total number of parameters in model (8) is $rm + u_1p_1 + u_2p_2 + \frac{r(r+1)}{2} + \frac{m(m+1)}{2}$. We can see that the envelope model (8) further reduces $(r-u_1)p_1 + (m-u_2)p_2 + 2$ parameters from model (5).

To fit the proposed envelope model, we apply the same approach performed for fitting model (5).

4 Simulation Study

To assess the performance of the proposed skewed matrix-variate regression (5) and its extended envelope matrix-variate regression (8) in the presence of outliers/skewed matrix-valued data, in this section, we designed a simulated example. We compared the produced results with the results of the regression model under the MatN error

term. Moreover, to have a more relevant comparison with a competing model that could take into account the skewness of the data, we introduced a new matrix-variate skew-normal (MatSN) distribution, denoted by MatSN($\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{b}$), with the following probability density function:

$$f(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{b}) = 2\phi_{r\times m}(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)\Phi(\boldsymbol{b}'\boldsymbol{\Sigma}_1^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{\Sigma}_2^{-\frac{1}{2}}\mathbf{1}_m),$$

where $\boldsymbol{\mu}$ is an $r \times m$ location matrix, $\boldsymbol{\Sigma}_1$ is an $r \times r$ symmetric and positive definite scale matrix for the rows of $\boldsymbol{X}$, $\boldsymbol{\Sigma}_2$ is an $m \times m$ symmetric and positive definite scale matrix for the columns of $\boldsymbol{X}$, $\boldsymbol{b}$ is an $r \times 1$ skewness vector, $\phi_{r\times m}(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ represents the probability density function of a MatN random variable, $\Phi(\cdot)$ is the distribution function of the standard normal distribuation, and $\mathbf{1}_m$ is an $m \times 1$ vector of ones. For developing related matrix-variate skew-normal regression and its extended envelope models, we also assumed that the error term in (5) and (8) has a MatSN($\mathbf{0}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{b}$).

Similar to [12], we simulated data based on model (8), considering a MatUBN error term, using the following settings:

$$\begin{aligned}
&r = m = p_1 = p_2 = 5, \quad u_1 = u_2 = 2, \quad k = -1, \quad a = 4\\
&\boldsymbol{\Omega}_1 = \sigma^2\mathbf{I}_{u_1}, \quad \boldsymbol{\Omega}_{10} = \sigma_0^2\mathbf{I}_{r-u_1}, \quad \boldsymbol{\Omega}_2 = \sigma^2\mathbf{I}_{u_2}, \quad \boldsymbol{\Omega}_{20} = \sigma_0^2\mathbf{I}_{m-u_2}\\
&\sigma^2 = 0.5, \quad \sigma_0^2 = 2.5.
\end{aligned}$$

We generated the semiorthogonal matrices $\mathbf{L}$ and $\mathbf{R}$ using orthogonalizing matrices of independent standard uniform random variables. We also generated the elements in $\boldsymbol{\mu}$, $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ as well as covariates from independent standard normal distributions. Finally, we considered three different sample sizes $n = 50, 100, 200$, and, for each sample size, 100 replicates were simulated.

We fitted six models to the data based on (5) and (8): two assuming the error term has the MatN distribution, two assuming the error term follows the MatSN distribution, and two allowing the error term has the MatUBN distribution. We set the envelope dimensions to the true dimensions $u_1 = u_2 = 2$ for model (8). To evaluate the estimation accuracy of the coefficient parameters as well as the covariance estimators in each model, we computed the following criteria using the Frobenius norm:

$$\|\hat{\boldsymbol{\beta}}_1 \otimes \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2\|_{\mathrm{F}}, \quad \|\hat{\boldsymbol{\Sigma}}_1 \otimes \hat{\boldsymbol{\Sigma}}_2 - \boldsymbol{\Sigma}_1 \otimes \boldsymbol{\Sigma}_2\|_{\mathrm{F}}.$$

The average estimation errors were calculated over the 100 random samples for each sample size under each model and are depicted in Fig. 2. Panels (a) and (b) of the figure show the average estimation errors for $\boldsymbol{\beta}_2 \otimes \boldsymbol{\beta}_1$ and $\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$, respectively. The envelope approach, excluding the MatN model, improved estimation accuracy in both regression coefficients and covariances for all sample sizes, compared to the matrix-variate regression model without enveloping, by effectively removing $\boldsymbol{X}$-invariants from estimation. The MatN model without enveloping has a better performance

in estimating only regression coefficients. Skewed MatSN and MatUBN envelope models are more efficient in parameter estimation than the envelope regression with MatN error term. This efficiency gain is more evident in estimating covariances. The same results hold for modeling without enveloping in estimating covariances. The improvement of the envelope estimation in the Kronecker product of covariances is substantial, while it is moderate for the regression coefficients. Furthermore, the MatUBN model is superior to the MatSN model in estimating covariances, while the result is reversed for the regression coefficients.

5 Applications

We applied our proposed models to analyze multivariate bioassay data regarding a cross-over assay of insulin based on rabbit blood sugar concentration ([31]). We also compared the results with the ones from the MatN regression. A thorough analysis of these data is reported in the Supplementary file of [12] under their proposed models.

There are four groups of animals with nine rabbits in each group and different treatments assigned to the groups. So the sample size is 36. Let K_1 and K_2 express the low and high dose levels, 0.75 units and 1.5 units of the standard treatment; let also T_1 and T_2 are the same two dose levels of the test treatment. The treatment assignments are given in Table 18 of Supplementary of [12]. The blood sugar of each rabbit was measured at 0, 1, 2, 3, 4, and 5 h after injection of the insulin dose each day. Similar to [12], we also considered the percentage decreases of the blood sugar concentration at 1, 2, 3, 4, and 5 h relative to the initial concentration at 0 h. Therefore, for each rabbit, we have a measured matrix $\boldsymbol{Y} \in \mathbb{R}^{5\times 2}$. The two columns show the percentage decreases under two different treatments on the first and second days, and the rows show the hourly percentage decreases each day. Treatment effects and dose levels also indicate covariates in the data, defined as a matrix $\boldsymbol{X} \in \mathbb{R}^{2\times 2}$ for each rabbit, whose rows represent standard and the test treatments, and columns show day 1 and day 2.

We fitted the two following models on the data

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mu} + \boldsymbol{\beta}_1 \boldsymbol{X} \boldsymbol{\beta}_2' + \boldsymbol{\epsilon} \tag{9}$$

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{\eta}_1 \boldsymbol{X} \boldsymbol{\beta}_2' + \boldsymbol{\epsilon}, \tag{10}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{5\times 2}$, $\boldsymbol{\beta}_1 \in \mathbb{R}^{5\times 2}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{2\times 2}$, $\mathbf{L} \in \mathbb{R}^{5\times 1}$, and $\boldsymbol{\eta}_1 \in \mathbb{R}^{1\times 2}$. We assumed that the error term $\boldsymbol{\epsilon} \in \mathbb{R}^{5\times 2}$ is distributed as a MatUBN, MatSN or MatN with two covariance matrices $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{\Sigma}_1 = \mathbf{L}\boldsymbol{\Omega}_1\mathbf{L}' + \mathbf{L}_0\boldsymbol{\Omega}_{10}\mathbf{L}_0' \in \mathbb{R}^{5\times 5}$ for both models, in which $\boldsymbol{\Omega}_2 > 0$ and $\boldsymbol{\Omega}_{10} > 0$ are unknown. The elements in the upper-right corner of $\boldsymbol{\beta}_2$ were set to zero because blood sugar concentration levels observed on the first day cannot be affected by the treatment received on the second day. Tables 1 and 2 describes the estimation results for models (9) and (10), respectively. We reported estimates (ESs) and estimated standard errors (SEs) for each element in

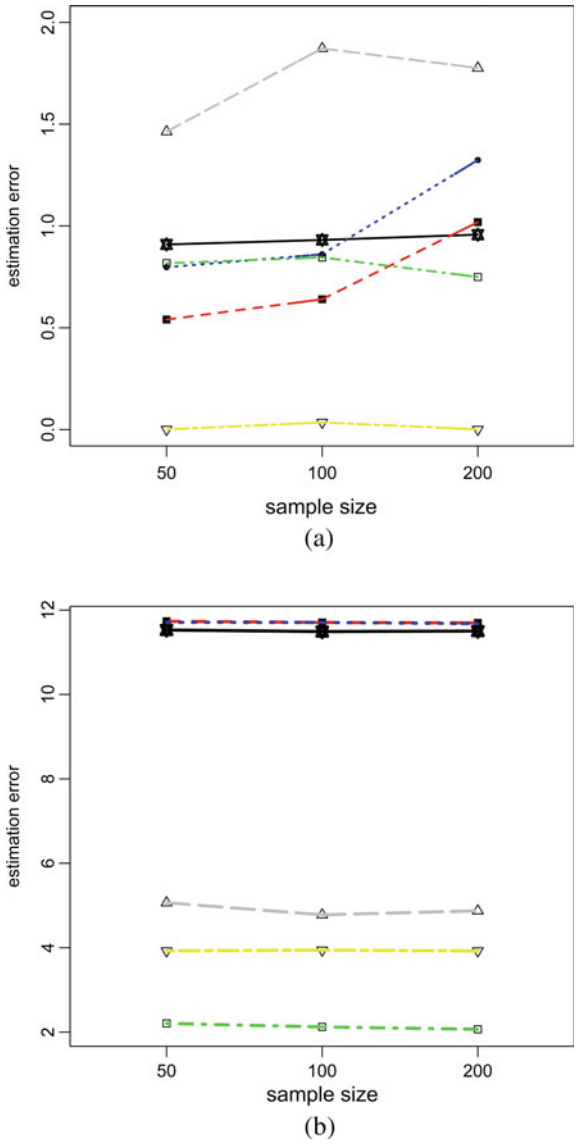


Fig. 2 Average estimation errors of **(a)** $\boldsymbol{\beta}_2 \otimes \boldsymbol{\beta}_1$ and **(b)** $\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$ in models (5) with MatUBN (_____), MatN (- - - - -) MatSN (.....) and (8) with MatUBN (-.-.-.-), MatN (— — —) and MatSN (- — - —)

Table 1 ES and SE of the parameters in model (9) considering MatUBN, MatN, or MatSN distributed errors

	MatUBN		MatSN		MatN	
	ES	SE	ES	SE	ES	SE
$\text{vec}(\hat{\boldsymbol{\beta}}_1)$	−5.451	2.404	−9.196	3.346	−7.677	5.848
	−12.837	2.220	−19.304	2.936	−16.232	9.069
	−19.918	1.897	−29.182	2.281	−24.809	21.173
	−13.721	3.044	−23.632	3.006	−19.850	18.855
	−3.858	2.553	−8.581	3.734	−7.406	8.443
	−5.499	2.423	−9.031	3.320	−7.468	5.765
	−11.520	2.256	−19.077	2.990	−15.867	9.304
	−18.189	2.050	−29.533	2.378	−25.212	21.845
	−11.495	3.010	−23.401	3.009	−19.546	18.967
	−1.069	2.451	−7.949	3.701	−6.593	8.160
$\text{vec}(\hat{\boldsymbol{\beta}}_2)$	−2.219	0.276	−9.588	0.431	−7.285	8.353
	−1.074	0.300	−8.933	0.452	−6.505	9.282
	−1.206	0.354	−0.743	0.215	−0.861	1.714
$\text{vech}(\hat{\boldsymbol{\Sigma}}_1)$	19.960	4.313	12.328	2.351	8.248	5.103
	16.337	3.785	11.375	1.966	7.600	4.471
	13.316	3.336	9.368	1.856	6.300	3.654
	7.213	3.266	4.476	2.073	3.067	2.287
	2.166	2.103	2.021	1.698	1.408	1.423
	22.690	4.792	17.480	2.361	11.644	6.586
	19.742	4.316	15.384	1.736	10.296	5.668
	11.109	3.410	8.277	2.010	5.562	3.383
	2.680	2.550	4.366	1.855	2.935	2.016
	35.735	6.589	23.956	1.957	16.080	8.998
	28.892	4.597	17.866	1.888	11.997	6.948
	12.600	2.978	10.675	1.818	7.150	4.093
	45.541	7.434	29.101	2.724	19.511	11.694
	22.623	4.383	18.584	2.342	12.467	7.235
	19.872	4.442	19.643	2.816	13.186	8.062
$\text{vech}(\hat{\boldsymbol{\Sigma}}_2)$	55.881	11.118	4.838	1.418	7.226	6.559
	11.095	4.778	1.271	0.690	1.902	2.094
	55.185	11.130	4.764	1.386	7.135	6.057
AIC	2458.685		2485.43		2483.74	

the tables for all proposed regression models. For computing SEs, we considered 100 bootstrap samples. In each bootstrap sample, observations are constructed as $\boldsymbol{Y}^* = \hat{\boldsymbol{Y}} + \boldsymbol{E}^*$, where $\hat{\boldsymbol{Y}}$ are the fitted values from the original model, and $\boldsymbol{E}^*$ are the resampled residuals. Direct computing of SEs of estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\Sigma}_1$ in model (10) is one of the advantages of the bootstrap method. According to the AIC criterion, the

Table 2 ES and SE of the parameters in model (10) considering MatUBN, MatN, or MatSN distributed errors

	MatUBN		MatSN		MatN	
	ES	SE	ES	SE	ES	SE
$\text{vec}(\hat{\boldsymbol{\beta}}_1)$	19.701	4.278	13.170	3.550	12.018	3.203
	32.370	5.507	22.075	5.365	20.145	4.761
	48.998	7.674	31.862	7.807	29.077	6.870
	39.093	11.044	29.055	8.051	26.515	7.149
	14.138	8.338	16.011	5.130	14.611	4.715
	17.944	4.426	12.821	3.512	11.700	3.246
	29.483	5.681	21.491	5.248	19.612	4.835
	44.628	7.965	31.019	7.694	28.307	7.001
	35.607	10.822	28.285	7.870	25.812	7.138
	12.877	8.041	15.587	5.081	14.224	4.709
$\text{vec}(\hat{\boldsymbol{\beta}}_2)$	−3.345	0.290	0.918	1.381	1.005	1.608
	0.582	0.286	0.453	1.409	0.495	1.634
	0.619	0.276	0.706	0.332	0.773	0.305
$\text{vech}(\hat{\boldsymbol{\Sigma}}_1)$	20.422	8.407	8.013	2.270	8.098	2.452
	18.678	8.263	7.671	2.123	7.753	2.303
	14.968	7.538	6.340	2.115	6.408	2.238
	5.756	4.756	2.625	1.965	2.653	2.015
	0.169	2.522	0.548	1.442	0.554	1.462
	29.957	12.343	12.423	3.363	12.555	3.702
	28.082	12.049	11.571	3.317	11.694	3.611
	12.812	7.234	5.808	2.757	5.869	2.925
	1.325	3.366	1.827	1.898	1.847	1.926
	48.525	19.672	18.268	5.391	18.463	5.942
	30.798	13.862	12.286	4.609	12.417	5.003
	7.683	5.756	5.093	2.890	5.148	3.021
	41.124	18.035	18.061	5.944	18.254	6.564
	18.389	8.425	10.038	3.883	10.145	4.243
	19.142	7.849	11.346	4.054	11.467	4.523
$\text{vech}(\hat{\boldsymbol{\Sigma}}_2)$	25.611	9.313	7.653	2.174	7.572	2.385
	5.517	2.510	1.694	0.894	1.677	0.912
	23.267	8.492	7.025	2.110	6.951	2.327
AIC	2472.689		2489.454		2487.454	

best-selected model is (9) with the MatUBN error term. Furthermore, in comparison to the fit from the model without enveloping, the envelope model (10) shows smaller standard errors for the most elements (Tables 1 and 2).

6 Concluding

In this manuscript, we introduced the skewed MatUBN regression that could be more efficient than the matrix regression with a normal-distributed error when the data are skewed or have outliers. Then, we applied the envelope methodology to construct a further parsimonious parametrized model for skewed matrix-valued responses. We also presented a new matrix-variate skew-normal model as a relevant competing model. Both proposed models are flexible for modeling skewed data, while the MatUBN model could also be applied to bimodal responses. Computed numerical results explain that the proposed skewed models are preferred.

However, our MatUBN model has disadvantages: (i) Due to the absolute function in the kernel of MatUBN density, designing an efficient fitting algorithm similar to the one proposed by [12] in the normal matrix-variate regression is difficult if not possible. Therefore, developing an efficient estimation method for the model could be interesting. (ii) We did not investigate the asymptotic properties of the proposed estimators. It could open avenues for future work.

Acknowledgements The authors wish to thank two anonymous reviewers for careful reading and constructive suggestions, leading to this improved version of the book chapter.

References

1. Akdemir, D., & Gupta, A. K. (2010). A matrix-variate skew distribution. *European Journal of Pure and Applied Mathematics, 3*(2), 128–140.
2. Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA Journal of Applied Mathematics, 6*, 76–90.
3. Cook, R. D. (2018). *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. Hoboken, NJ: Wiley.
4. Cook, R. D., & Zhang, X. (2015). Simultaneous envelopes for multivariate linear regression. *Technometrics, 57*(1), 11–25.
5. Cook, R. D., Li, B., & Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica, 20*, 927–1010.
6. Cook, R. D., Helland, I. S., & Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B Statistical Methodology, 75*, 851–877.
7. Cook, R. D., Forzani, L., & Su, Z. (2016). A note on fast envelope estimation. *The Journal of Multivariate Analysis, 150*, 42–54.
8. Ding, S. (2014). Sufficient dimension reduction for complex data structures, Ph.D. thesis: Retrieved from the University of Minnesota Digital Conservancy. Retrieved from https://hdl.handle.net/11299/164799.
9. Ding, S., & Cook, R. D. (2014). Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica, 24*, 463–492.
10. Ding, S., & Cook, R. D. (2015). Tensor sliced inverse regression. *The Journal of Multivariate Analysis, 133*, 216–231.
11. Ding, S., & Cook, R. D. (2015). Higher-order sliced inversion regression. *Wiley Interdisciplinary Reviews: Computational Statistics, 7*(4), 249–257.
12. Ding, S., & Cook, R. D. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series B Statistical Methodology, 80*(2), 387–408.

13. Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation, 64*(2), 105–123.
14. Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal, 13*, 317–322.
15. Gallaugher, M. P. B., & McNicholas, P. D. (2017). A matrix variate skew-t distribution. *Stat., 6*(1), 160–170.
16. Gallaugher, M. P. B., & McNicholas, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition, 80*, 83–93.
17. Gallaugher, M. P. B., & McNicholas, P. D. (2019). Mixtures of skewed matrix variate bilinear factor analyzers. *Advances in Data Analysis and Classification, 14*, 415–434.
18. Gallaugher, M. P. B., & McNicholas, P. D. (2019). Three skewed matrix variate distributions. *Statistics and Probability Letters, 145*, 103–109.
19. Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation, 24*, 23–26.
20. Gupta, A. K., & Nagar, D. K. (1999). *Matrix Variate Distributions* (1st ed.). Boca Raton: Chapman and Hall/CRC.
21. Harrar, S. W., & Gupta, A. K. (2008). *On matrix variate skew-normal distributions. Stats., 42*(2), 179–194.
22. Hu, W., Pan, T., Kong, D., & Shen, W. (2020). Nonparametric matrix response regression with application to brain imaging data analysis. *Biometrics*. https://doi.org/10.1111/biom.13362.
23. Kong, D., An, B., Zhang, J., & Zhu, H. (2020). L2RM: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association, 115*, 403–424.
24. Li, B., Kim, K. M., & Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *Annals of Statistics, 38*(2), 1094–1121.
25. Nguyen, T. T. (1997). A note on matrix variate normal distribution. *The Journal of Multivariate Analysis, 60*(1), 148–153.
26. Ownuk, J., Nezakati, A., & Baghishani, H. (2021). Developing flexible classes of distributions to account for both skewness and bimodality. Retrieved fromhttps://arxiv.org/abs/2106.16022.
27. Pfeiffer, R. M., Forzani, L., & Bura, E. (2012). Sufficient dimension reduction for longitudinally measured predictors. *Statistics in Medicine, 31*, 2414–2427.
28. R-Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
29. Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation, 24*, 647–656.
30. Viroli, C. (2012). On matrix-variate regression analysis. *The Journal of Multivariate Analysis, 111*, 296–309.
31. Vølund, A. (1980). Multivariate bioassay. *Biometrics, 36*(2), 225–236.
32. Xue, Y., & Yin, X. (2014). Sufficient dimension folding for regression mean function. *Journal of Computational and Graphical Statistics, 23*(4), 1028–1043.
33. Zhang, H., Zhu, F., & Li, S. (2016). Robust matrix regression. Retrieved from arXiv:1611.04686.

Multivariate Functional Singular Spectrum Analysis: A Nonparametric Approach for Analyzing Multivariate Functional Time Series

Jordan Trinka, Hossein Haghbin, and Mehdi Maadooliat

Abstract In this chapter, we develop multivariate functional singular spectrum analysis (MFSSA) over different dimensional domains with the goal of decomposing a multivariate functional time series (MFTS) into interpretable partitions such as mean, periodic, and trend components. The approach is flexible in the sense that the MFTS signal may be composed of functional observations such as curves and surfaces and the offered flexibility can lead to richer signal extraction results for correlated data. In the following, we first review (a) the singular spectrum analysis (SSA) algorithm, (b) the functional extension of SSA (FSSA), and (c) the multivariate extension of SSA (MSSA) where these methods serve as decomposition routines for scalar, univariate functional, and multivariate scalar time series data, respectively. Second, we use these backgrounds to derive the MFSSA algorithm. Third, we discuss two other approaches to MFSSA known as horizontal MFSSA (HMFSSA) and vertical MFSSA (VMFSSA) and we find that VMFSSA gives the same decomposition results as MFSSA due to isomorphic vector spaces. Fourth, we showcase the superior signal extraction results of MFSSA as compared to other approaches of MFTS decomposition by way of a simulation study and real data study applications of remote sensing and temperature data. We also developed a user-friendly R package called **Rfssa** used to implement the proposed method and we developed an R Shiny web application that can be used to further explore the algorithm.

Hossein Haghbin contributed equally as much as Jordan Trinka.

J. Trinka · M. Maadooliat (✉)
Marquette University, 1250 W. Wisconsin Ave, Milwaukee, WI 53233, USA
e-mail: mehdi.maadooliat@mu.edu

H. Haghbin
Persian Gulf University, Bushehr, Bushehr Province, Iran
e-mail: haghbin@pgu.ac.ir

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_9

1 Background

One of the popular approaches in the decomposition of time series data is accomplished using rates of change. In this approach, the observed time series is partitioned (decomposed) into informative trends plus potential seasonal (cyclical) and noise (irregular) components. Aligned with this principle, singular spectrum analysis-based approaches are model-free procedures that are commonly used as a nonparametric techniques in analyzing time series. The methods are intrinsically motivated as exploratory and model building tools rather than confirmatory procedures [6]. These algorithms do not require restrictive assumptions such as stationarity, linearity, and normality. They can be used for a wide range of purposes such as trend and periodic component detection and extraction, smoothing, forecasting, change-point detection, gap filling, causality, and so on (see, e.g., [6, 7, 17, 19–21, 24]). The goal of this chapter is to establish an SSA-based procedure to perform signal extraction on time-dependent, multivariate functional data. In order to motivate the proposed approach, we first present the univariate singular spectrum analysis (SSA) algorithm, it's functional extension, and the multivariate extension (not functional) and we find that these routines serve as the building blocks of our novel method.

General Scheme of Univariate Singular Spectrum Analysis

The goal of SSA is to extract out modes of variation in time series data. It provides a representation of the given time series in terms of rank one approximations generated from a singular value decomposition (SVD) of a so-called trajectory matrix [1]. These rank one approximations are built from so-called eigentriples where an eigentriple is defined in SSA literature as a set that contains a singular value, corresponding left singular vector, and corresponding right singular vector that are found using the SVD technique [6, 13]. Presently, many studies have been published with extensions and applications of SSA. The extension to multivariate SSA (MSSA) is straightforward [4, 5, 8, 13, 14], and an extension of SSA to a two-dimensional setting can be found in Golyandina et al. [4] and references therein. For now, we continue on with our review of the SSA algorithm.

Univariate Singular Spectrum Analysis Algorithm

Throughout this subsection, we consider y_i's as elements of the Euclidean space, $\mathbb{R}$. Suppose that $\mathbf{y}_N = [y_1, y_2, \ldots, y_N]^\top$ is a realization of size N from a time series. The basic SSA algorithm consists of four steps: Embedding, Decomposition, Grouping, and Reconstruction. The purpose of the first step of embedding is to generate a trajectory matrix that corresponds with the time series where the matrix can be decomposed using SVD.

Step 1. Embedding

This step generates a matrix by tracking a moving window of size L over the original time series, where L is an integer called the window length parameter and $1 < L < N/2$. Embedding can be regarded as a mapping operator, $\mathcal{T}$, that transfers the series, $\mathbf{y}_N$, into a so-called trajectory matrix, $\mathbf{X}$, of dimension $L \times K$, defined by

$$\mathbf{X} = \mathcal{T}(\mathbf{y}_N) = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_K],$$

where $K = N - L + 1$ and $\boldsymbol{x}_j = \left[y_j, y_{j+1}, \ldots, y_{j+L-1}\right]^\top \in \mathbb{R}^L$ is known as the j^{th}, L-lagged vector for $j = 1, \ldots, K$. Note that the trajectory matrix, $\mathbf{X}$, is a Hankel matrix which means that the elements along the anti-diagonals are equal. Indeed, the embedding operator, $\mathcal{T}$, is a one-to-one mapping from $\mathbb{R}^N$ into $\mathbb{R}_H^{L\times K}$ where $\mathbb{R}_H^{L\times K}$ is the space of all $L \times K$ Hankel matrices in the space of $L \times K$ real-valued matrices. In addition, we also have that the columns of $\mathbf{X}$ describe time series behavior over sub-intervals of time and we now extract out the dominant modes of variation shared between these L-lagged vectors in the following step.

Step 2. Decomposition

In this step, an SVD for the trajectory matrix, $\mathbf{X}$, is computed as

$$\mathbf{X} = \sum_{i=1}^{r} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^\top = \sum_{i=1}^{r} \mathbf{X}_i, \tag{1}$$

where r is the rank of $\mathbf{X}$, $\sqrt{\lambda_i}$ is the i^{th} singular value, $\mathbf{u}_i \in \mathbb{R}^L$ is the associated left singular vector, and $\mathbf{v}_i \in \mathbb{R}^K$ is the associated right singular vector of $\mathbf{X}$. In addition, we define the set, $\{\sqrt{\lambda_i}, \mathbf{u}_i, \mathbf{v}_i\}$, as the i^{th} eigentriple of $\mathbf{X}$, and we use the elements of this eigentriple to construct the i^{th} rank one elementary matrix of dimension $L \times K$ given by $\mathbf{X}_i = \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^\top$. In the next step, we form disjoint groups of eigentriples where the eigentriples in each group describe similar time series behavior. We then add rank one elementary matrices together within each group where each elementary matrix is built using the corresponding eigentriple.

Step 3. Grouping

Consider a partition of the set of indices, $\{1, 2, \ldots, r\}$, into m disjoint subsets, $\{I_1, I_2, \ldots, I_m\}$. For any positive integer q, i.e., $1 \leq q \leq m$, the $L \times K$ matrix, $\mathbf{X}_{I_q}$, is defined as $\mathbf{X}_{I_q} = \sum_{i\in I_q} \mathbf{X}_i$. Thus, by the expansion of Eq. (1), we have the grouped matrix decomposition

$$\mathbf{X} = \mathbf{X}_{I_1} + \mathbf{X}_{I_2} + \cdots + \mathbf{X}_{I_m}. \tag{2}$$

Each group in Eq. (2) should correspond to a component in the time series decomposition such as mean, seasonality, trend, etc. Exploratory plots of eigentriple elements (such as plots of singular values, left/right singular vectors, etc.) can be leveraged as helpful tools when making decisions on how to form the disjoint groups [6]. In the final step of SSA, we extract out a time series from each $\mathbf{X}_{I_q}$.

Step 4. Reconstruction

The resulting matrices, $\mathbf{X}_{I_q}$, are transformed back into the form of the original series, $\mathbf{y}_N$, by the inverse operator $\mathcal{T}^{-1} : \mathbb{R}_H^{L\times K} \to \mathbb{R}^N$. In order to do this, it is necessary that each $\mathbf{X}_{I_q}$ is approximated by a matrix in $\mathbb{R}_H^{L\times K}$. This approximation is performed optimally in the sense of orthogonal projection of $\mathbf{X}_{I_q}$ onto $\mathbb{R}_H^{L\times K}$ with respect to the Frobenius norm. Denote this projection by $\Pi_H : \mathbb{R}^{L\times K} \to \mathbb{R}_H^{L\times K}$. It is shown that applying Π_H to some $L \times K$ matrix is the same as averaging of the matrix elements over the anti-diagonals (where $i + j =$ constant). As such, we may define the $L \times K$ matrix, $\widetilde{\mathbf{X}}_{I_q} = \Pi_H \mathbf{X}_{I_q}$, as the projection of $\mathbf{X}_{I_q}$ onto $\mathbb{R}_H^{L\times K}$ and from here, we define the q^{th} reconstructed time series of length N as $\tilde{\mathbf{y}}_N^q = \mathcal{T}^{-1}\left(\widetilde{\mathbf{X}}_{I_q}\right)$ for $q = 1, \ldots, m$. We have that each $\tilde{\mathbf{y}}_N^q$ describes a mode of variation present in the original series, $\mathbf{y}_N$ such as mean, periodicity, etc.

It is well known that SSA does not require restrictive assumptions; however, it is ideal to have a time series with additive components. If a time series is composed of such additive components, then we call the series separable. Therefore, we present tools to measure the separability of components in the following.

Separability

Let $\mathbf{y}_N = [y_1, \ldots, y_N]^\top$ and $\mathbf{z}_N = [z_1, \ldots, z_N]^\top$ be two time series. The weighted-correlation (w-correlation) between $\mathbf{y}_N$ and $\mathbf{z}_N$ is defined as

$$\rho^{(w)}(\mathbf{y}_N, \mathbf{z}_N) = \frac{\langle \mathbf{y}_N, \mathbf{z}_N \rangle_w}{\|\mathbf{y}_N\|_w \|\mathbf{z}_N\|_w}, \tag{3}$$

where $\langle \mathbf{y}_N, \mathbf{z}_N \rangle_w = \sum_{i=1}^N w_i y_i z_i$, $w_i = \min\{i, L, N - i + 1\}$, and $\|\mathbf{y}_N\|_w = \sqrt{\langle \mathbf{y}_N, \mathbf{y}_N \rangle_w}$. We call $\mathbf{y}_N$ and $\mathbf{z}_N$ w-orthogonal if $\rho^{(w)}(\mathbf{y}_N, \mathbf{z}_N) = 0$ for appropriate values of L (see the next paragraph for more details). Note that $\tilde{\mathbf{y}}_N^q, q = 1, \ldots, r$, is the reconstructed time series produced by the group, I_q, and the matrix, $\boldsymbol{\rho}^{(w)} = \left[\rho^{(w)}(\tilde{\mathbf{y}}_N^i, \tilde{\mathbf{y}}_N^j)\right]_{i,j=1}^r$, is called the w-correlation matrix [6]. If $\tilde{\mathbf{y}}_N^i$ and $\tilde{\mathbf{y}}_N^j$ are approximately w-orthogonal, then it is recommended that we perform the grouping stage of SSA in such a way where i and j belong to different groups of indices so that we do not mix together dissimilar modes of variation in our reconstructions. It is clear that the w-correlation matrix can help in determining how to perform the grouping stage

of SSA. The determination of L and the grouping parameters is essential in the SSA algorithm and we now discuss techniques used to select values for these parameters.

Parameter Selection

There are two basic parameters in the SSA procedure: window length (L) and grouping parameters. Choosing improper values for these parameters yields an incomplete reconstruction and misleading results in subsequent analysis. In spite of the importance of choosing L and grouping parameters for SSA, no ideal solution has been proposed yet. A thorough review of the problem shows that the optimal choice of the parameters depends on the intrinsic structure of the data and the purposes of the study [6, 8]. However, there are several recommendations and rules that work well for a wide range of scenarios. It is recommended to select the window length parameter, L, to be a large integer that is multiple of the periodicities of the time series, but not larger than $N/2$.

In addition, there are several methods for effective grouping. These techniques include analyzing the periodogram, paired plot of the singular vectors, scree plot of the singular values, and w-correlation plot; see Golyandina et al. [6] for more details. We also note that these techniques used in parameter selection for the SSA routine hold true for all forms of the algorithm including the functional, multivariate, and multivariate functional versions to be discussed later. From here, we continue on with a review of the functional extension to the SSA algorithm.

General Scheme of Functional Singular Spectrum Analysis

Functional data analysis (FDA) embodies the evaluation and exploration of data that is comprised of functions such as curves or surfaces [23]. Functional principal component analysis (FPCA) is a technique that is used to find the most informative directions in a time-independent collection of functional subjects [23]. Functional SSA (FSSA) was developed by Haghbin et al. [10] as a novel technique that is used to decompose a time-dependent collection of functional subjects, known as a functional time series (FTS), into mean, seasonal, trend, and noise components. FSSA works to decompose a FTS in a similar fashion as SSA using a functional SVD (FSVD). This method was compared with other techniques of dimension reduction of a FTS, including dynamic FPCA (DFPCA) [16], and it was found that FSSA is the ideal approach in terms of reconstruction accuracy of the true signal assumed to be present in the data. Now we present the FSSA methodology.

From hereafter, we consider $\mathbf{y}_N = (y_1, \ldots, y_N)$ as an FTS of length N such that each function vector, $y_i : [0, 1] \to \mathbb{R}$, belongs to the Hilbert space $\mathbb{H} = L^2([0, 1])$. We define the inner product equipped to $\mathbb{H}$ as $\langle x, y \rangle_{\mathbb{H}} = \int_0^1 x(s)y(s)ds$, where $x, y \in \mathbb{H}$. For a positive integer k, the Hilbert space, $\mathbb{H}^k$, is defined as the Carte-

sian product of k copies of $\mathbb{H}$, i.e., a length k function vector, $\boldsymbol{x} \in \mathbb{H}^k$, evaluated at a point $\mathbf{s} = [s_1, s_2, \ldots, s_k]^\top \in [0,1]^k$ is a vector of length k denoted by $\boldsymbol{x}(\mathbf{s}) = [x_1(s_1), x_2(s_2), \ldots, x_k(s_k)]^\top$, where $x_i \in \mathbb{H}$. We define the inner product equipped to $\mathbb{H}^k$ as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{H}^k} = \sum_{i=1}^k \langle x_i, y_i \rangle_{\mathbb{H}}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}^k$. The norms are denoted by $\|\cdot\|_{\mathbb{H}}$ and $\|\cdot\|_{\mathbb{H}^k}$ in the spaces $\mathbb{H}$ and $\mathbb{H}^k$, respectively. For function vectors $x \in \mathbb{H}_1$, and $y \in \mathbb{H}_2$, where $\mathbb{H}_1$ and $\mathbb{H}_2$ are two Hilbert spaces, we define the tensor (outer) product corresponding to the operator $x \otimes y : \mathbb{H}_1 \to \mathbb{H}_2$, as $(x \otimes y)h = \langle x, h \rangle_{\mathbb{H}_1} y$, where $h \in \mathbb{H}_1$ and $\langle \cdot, \cdot \rangle_{\mathbb{H}_1}$ is the inner product equipped to $\mathbb{H}_1$.

For positive integers L and K, we denote $\mathbb{H}^{L \times K}$ as the linear space spanned by operators $\boldsymbol{\mathcal{Z}} : \mathbb{R}^K \to \mathbb{H}^L$, specified by function vectors $(z_{i,j})_{i=1,\ldots,L}^{j=1,\ldots,K}$, where

$$\boldsymbol{\mathcal{Z}}\mathbf{a} = \left(\sum_{j=1}^K a_j z_{1,j}, \ldots, \sum_{j=1}^K a_j z_{L,j} \right), \ z_{i,j} \in \mathbb{H}, \text{ and } \mathbf{a} = [a_1, \ldots, a_K]^\top \in \mathbb{R}^K.$$

We call an operator $\widetilde{\boldsymbol{\mathcal{Z}}} = \left(\tilde{z}_{i,j}\right) \in \mathbb{H}^{L \times K}$ Hankel if for all $i = 1, \ldots, L$, and $j = 1, \ldots, K$, we have $\|\tilde{z}_{i,j} - g_s\|_{\mathbb{H}} = 0$ for some $g_s \in \mathbb{H}$ where $s = i + j$. The space of such Hankel operators is denoted by $\mathbb{H}_H^{L \times K}$. For two given operators $\boldsymbol{\mathcal{Z}}_1 = \left(z_{i,j}^{(1)}\right)$ and $\boldsymbol{\mathcal{Z}}_2 = \left(z_{i,j}^{(2)}\right)$ in $\mathbb{H}^{L \times K}$, define

$$\langle \boldsymbol{\mathcal{Z}}_1, \boldsymbol{\mathcal{Z}}_2 \rangle_F = \sum_{i=1}^L \sum_{j=1}^K \langle z_{i,j}^{(1)}, z_{i,j}^{(2)} \rangle_{\mathbb{H}}.$$

It follows immediately that $\langle \cdot, \cdot \rangle_F$, which defines an inner product on $\mathbb{H}^{L \times K}$. We call $\langle \cdot, \cdot \rangle_F$ as the Frobenius inner product of two operators in $\mathbb{H}^{L \times K}$. The associated Frobenius norm is $\|\boldsymbol{\mathcal{Z}}\|_F = \sqrt{\langle \boldsymbol{\mathcal{Z}}, \boldsymbol{\mathcal{Z}} \rangle_F}$. Before discussing the FSSA algorithm, we present a lemma that is used in the last step of the method.

Lemma 1 *Let $x_i \in \mathbb{H}$ for $i = 1, \ldots, N$. If $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$, then*

$$\sum_{i=1}^N \|x_i - \bar{x}\|_{\mathbb{H}}^2 \leq \sum_{i=1}^N \|x_i - y\|_{\mathbb{H}}^2,$$

for all $y \in \mathbb{H}$.

Proof We add and subtract $\bar{x}$ to obtain

$$\sum_{i=1}^N \|x_i - y\|_{\mathbb{H}}^2 = \sum_{i=1}^N \|x_i - \bar{x}\|_{\mathbb{H}}^2 + 2 \sum_{i=1}^N \langle x_i - \bar{x}, \bar{x} - y \rangle_{\mathbb{H}} + N \|\bar{x} - y\|_{\mathbb{H}}^2.$$

Notice that $\sum_{i=1}^N \langle x_i - \bar{x}, \bar{x} - y \rangle_{\mathbb{H}} = 0$, then we have

$$\sum_{i=1}^{N} \|x_i - y\|_{\mathbb{H}}^2 = \sum_{i=1}^{N} \|x_i - \bar{x}\|_{\mathbb{H}}^2 + N\|\bar{x} - y\|_{\mathbb{H}}^2 \geq \sum_{i=1}^{N} \|x_i - \bar{x}\|_{\mathbb{H}}^2$$

which completes the proof. □

Now we present the steps of the FSSA methodology.

Functional Singular Spectrum Analysis Algorithm

For an integer $1 < L < N/2$, let $K = N - L + 1$ and define

$$\boldsymbol{x}_j = \left(y_j, y_{j+1}, \ldots, y_{j+L-1}\right) \tag{4}$$

to be the j^{th}, L-lagged function vector in $\mathbb{H}^L$ corresponding to FTS, $\mathbf{y}_N$, for $j = 1, \ldots, K$. The following provides the FSSA procedure where the FSVD technique is applied to an operator whose range is defined by the linear span of the $\boldsymbol{x}_j$'s where such L-lagged function vectors capture FTS behavior over sub-intervals of time.

Step 1. Embedding

Define the operator $\boldsymbol{\mathcal{X}} : \mathbb{R}^K \to \mathbb{H}^L$ with

$$\boldsymbol{\mathcal{X}}\mathbf{a} = \sum_{j=1}^{K} a_j \boldsymbol{x}_j, \qquad \mathbf{a} = [a_1, \ldots, a_K]^\top \in \mathbb{R}^K. \tag{5}$$

We call $\boldsymbol{\mathcal{X}}$ the trajectory operator. It is easy to see that $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}}\mathbf{y}_N$, where $\boldsymbol{\mathcal{T}}$ is an invertible operator from $\mathbb{H}^N$ to $\mathbb{H}_H^{L\times K}$. Evaluating $\boldsymbol{\mathcal{X}}\mathbf{a}$ at a given L-dimensional vector, $\mathbf{s} \in [0, 1]^L$, is the same as the matrix product $\mathbf{X}(\mathbf{s})\mathbf{a}$, where $\mathbf{X}(\mathbf{s})$ is an $L \times K$ Hankel matrix given by

$$\mathbf{X}(\mathbf{s}) = [\boldsymbol{x}_1(\mathbf{s}) \cdots \boldsymbol{x}_K(\mathbf{s})].$$

Now we define the adjoint of $\boldsymbol{\mathcal{X}}$ in the following proposition.

Proposition 1 *The operator, $\boldsymbol{\mathcal{X}}$, is a bounded linear operator. If we define the operator $\boldsymbol{\mathcal{X}}^* : \mathbb{H}^L \to \mathbb{R}^K$, given by the vector*

$$\boldsymbol{\mathcal{X}}^* z = [\langle \boldsymbol{x}_1, z\rangle_{\mathbb{H}^L}, \langle \boldsymbol{x}_2, z\rangle_{\mathbb{H}^L}, \ldots, \langle \boldsymbol{x}_K, z\rangle_{\mathbb{H}^L}]^\top, \ z \in \mathbb{H}^L, \tag{6}$$

then $\boldsymbol{\mathcal{X}}^*$ is the adjoint operator for $\boldsymbol{\mathcal{X}}$.

Proof The boundedness and linearity of $\boldsymbol{\mathcal{X}}$ is straightforward from the definition of the operator. Now we show the form of the adjoint of $\boldsymbol{\mathcal{X}}$. For a given vector, $\mathbf{a} \in \mathbb{R}^K$, and length L function vector, $z \in \mathbb{H}^L$, we have

$$\langle \mathcal{X}\mathbf{a}, \boldsymbol{z}\rangle_{\mathbb{H}^L} = \sum_{j=1}^{K} a_j \langle \boldsymbol{x}_j, \boldsymbol{z}\rangle_{\mathbb{H}^L} = \sum_{i=1}^{L}\sum_{j=1}^{K} a_j \langle y_{i+j-1}, z_i\rangle_{\mathbb{H}}$$
$$= \left\langle \begin{bmatrix} a_1 \\ \vdots \\ a_K \end{bmatrix}, \begin{bmatrix} \sum_{i=1}^{L} \langle y_i, z_i\rangle_{\mathbb{H}} \\ \vdots \\ \sum_{i=1}^{L} \langle y_{i+K-1}, z_i\rangle_{\mathbb{H}} \end{bmatrix} \right\rangle_{\mathbb{R}^K} = \langle \mathbf{a}, \mathcal{X}^*\boldsymbol{z}\rangle_{\mathbb{R}^K}$$

which completes the proof. □

From here, we enter into step two with the goal of extracting out time-dependent modes of variation from $\mathcal{X}$.

Step 2. Decomposition

In this step, we decompose the trajectory operator, $\mathcal{X}$, into a set of rank one operators. To this end, we denote the range of $\mathcal{X}$ by $R(\mathcal{X})$. Clearly, $R(\mathcal{X}) = \text{span}\left\{\boldsymbol{x}_j\right\}_{j=1}^{K}$ is $r-$dimensional, where r is a positive integer less than or equal to K. Therefore $\mathcal{X}$ is a finite-rank (r-dimensional) operator. The following theorem provides us with an FSVD of $\mathcal{X}$ and $\mathcal{X}^*$.

Theorem 1 *Consider the trajectory operator, $\mathcal{X}$, in Eq. (5). There exists orthonormal elements, $\{\boldsymbol{\psi}_i\}_{i=1}^{r}$, from $\mathbb{H}^L$ and orthonormal vectors, $\{\mathbf{v}_i\}_{i=1}^{r}$, from $\mathbb{R}^K$ such that*

$$\mathcal{X}\mathbf{a} = \sum_{i=1}^{r} \sqrt{\lambda_i}\langle \mathbf{v}_i, \mathbf{a}\rangle_{\mathbb{R}^K} \boldsymbol{\psi}_i, \quad for\ all\ \mathbf{a} \in \mathbb{R}^K, \tag{7}$$

where the λ_i's are non-ascending positive scalars. We also obtain

$$\mathcal{X}^*\boldsymbol{z} = \sum_{i=1}^{r} \sqrt{\lambda_i}\langle \boldsymbol{\psi}_i, \boldsymbol{z}\rangle_{\mathbb{H}^L} \mathbf{v}_i, \quad for\ all\ \boldsymbol{z} \in \mathbb{H}^L. \tag{8}$$

Proof Since $\mathcal{X}$ and $\mathcal{X}^*$ are finite-rank operators, they are also compact and as such, the proof immediately follows from Theorem 7.6 of Weidmann [30]. □

We call $\sqrt{\lambda_i}$ as the i^{th} singular value, $\boldsymbol{\psi}_i$ as the i^{th} left singular function, and $\mathbf{v}_i$ as the i^{th} right singular vector of the trajectory operator, $\mathcal{X}$. The following proposition expands further on the properties of the elements found in an FSVD of $\mathcal{X}$.

Proposition 2 *In Theorem 1, the set, $\{\boldsymbol{\psi}_i\}_{i=1}^{r}$, forms a basis system for $R(\mathcal{X})$, and each vector, $\mathbf{v}_i$, can be written as*

$$\mathbf{v}_i = \frac{\mathcal{X}^*\boldsymbol{\psi}_i}{\sqrt{\lambda_i}} = \left[\frac{\langle \boldsymbol{\psi}_i, \boldsymbol{x}_1\rangle_{\mathbb{H}^L}}{\sqrt{\lambda_i}}, \ldots, \frac{\langle \boldsymbol{\psi}_i, \boldsymbol{x}_K\rangle_{\mathbb{H}^L}}{\sqrt{\lambda_i}}\right]^\top, \qquad i = 1, \ldots, r. \tag{9}$$

Proof We first derive the form of the right singular vectors and then we show that the left singular functions span $R(\mathcal{X})$. We use Eqs. (7) and (8) to obtain $\mathcal{X}\mathbf{v}_i = \sqrt{\lambda_i}\boldsymbol{\psi}_i$ and $\mathcal{X}^*\boldsymbol{\psi}_i = \sqrt{\lambda_i}\mathbf{v}_i$. We use Eqs. (7), (8), and (6) to obtain Eq. (9). Now by substituting Eq. (9) into Eq. (7), we obtain

$$\boldsymbol{x}_j = \sum_{i=1}^{r} \langle \boldsymbol{x}_j, \boldsymbol{\psi}_i \rangle_{\mathbb{H}^L} \boldsymbol{\psi}_i.$$

Since $R(\mathcal{X}) = \text{span}\left\{\boldsymbol{x}_j\right\}_{j=1}^{K}$, it follows that $\{\boldsymbol{\psi}_i\}_{i=1}^{r}$ is an orthonormal basis for $R(\mathcal{X})$. □

The collection $\{\sqrt{\lambda_i}, \boldsymbol{\psi}_i, \mathbf{v}_i\}$ is defined as the i^{th} eigentriple of $\mathcal{X}$, and the right singular vectors, $\mathbf{v}_i$'s, can be used to produce paired plots similar to paired plots seen in SSA literature [6]. Let $\mathcal{X}_i : \mathbb{R}^K \to \mathbb{H}^L$ be the rank one operator defined by $\mathcal{X}_i = \sqrt{\lambda_i}\mathbf{v}_i \otimes \boldsymbol{\psi}_i$, for $i = 1, \ldots, r$. We use Eq. (7) and decompose $\mathcal{X}$ as

$$\mathcal{X} = \sum_{i=1}^{r} \mathcal{X}_i. \tag{10}$$

Similar to SSA, we group eigentriples together in disjoint sets to separate out sources of variation and then we add the corresponding rank one operators together within groups. This routine is given in the following step.

Step 3. Grouping

The grouping step consists of rearranging and partitioning the elementary operators, $\mathcal{X}_i$'s, in Eq. (10). To do this, we mimic the approaches used in the grouping step of Sect. 1 for the univariate SSA and implement the equivalent function version of those in Haghbin et al. [11]. We consider a partition, $\{I_1, I_2, \ldots, I_m\}$ of the set of indices, $\{1, \ldots, r\}$, and we define the operator, $\mathcal{X}_{I_q} : \mathbb{R}^K \to \mathbb{H}^L$, given by $\mathcal{X}_{I_q} = \sum_{i \in I_q} \mathcal{X}_i$, such that we have the expansion

$$\mathcal{X} = \mathcal{X}_{I_1} + \mathcal{X}_{I_2} + \cdots + \mathcal{X}_{I_m}. \tag{11}$$

The grouping should be done so that each operator, $\mathcal{X}_{I_q} \in \mathbb{H}^{L \times K}$, captures a unique component of variation present in the original FTS such as mean, periodic, or trend behaviors for $q = 1, \ldots, m$. In step four, we have the goal of extracting out an FTS that corresponds with each $\mathcal{X}_{I_q}$.

Step 4. Reconstruction

We use the operator $\mathcal{T}^{-1} : \mathbb{H}_H^{L \times K} \to \mathbb{H}^N$ to transform back each operator, $\mathcal{X}_{I_q}$, seen in Eq. (11), to an FTS, $\tilde{\mathbf{y}}_N^q$. Since $\mathcal{X}_{I_q} \in \mathbb{H}^{L \times K}$ is not necessarily Hankel, we first

project $\boldsymbol{\mathcal{X}}_{I_q}$ to $\mathbb{H}_H^{L\times K}$. Note that $\mathbb{H}_H^{L\times K}$ is a closed subspace of $\mathbb{H}^{L\times K}$, then by Theorem 3.2.3 of Shalit [25], there exists a unique operator, $\widetilde{\boldsymbol{\mathcal{X}}}_{I_q} \in \mathbb{H}_H^{L\times K}$, such that

$$\|\boldsymbol{\mathcal{X}}_{I_q} - \widetilde{\boldsymbol{\mathcal{X}}}_{I_q}\|_F^2 \leq \|\boldsymbol{\mathcal{X}}_{I_q} - \widetilde{\boldsymbol{\mathcal{Z}}}\|_F^2, \text{ for any operator } \widetilde{\boldsymbol{\mathcal{Z}}} \in \mathbb{H}_H^{L\times K}.$$

To specify $\widetilde{\boldsymbol{\mathcal{X}}}_{I_q}$, we denote the elements of $\boldsymbol{\mathcal{X}}_{I_q}$ and $\widetilde{\boldsymbol{\mathcal{X}}}_{I_q}$ by the function vectors, $\left(x_{i,j}^q\right)$ and $\left(\tilde{x}_{i,j}^q\right)$, respectively. Using Lemma 1, it is easy to extend the diagonal averaging approach in Golyandina et al. [6] to $\mathbb{H}^{L\times K}$ and obtain $\tilde{x}_{i,j}^q$'s in the following:

$$\tilde{x}_{i,j}^q = \frac{1}{n_s} \sum_{(l,k):l+k=s} x_{l,k}^q, \tag{12}$$

where $s = i + j$ and n_s stands for the number of (l, k) pairs such that $l + k = s$. This diagonal averaging technique may be viewed as an orthogonal projection operator, $\boldsymbol{\Pi}_H : \mathbb{H}^{L\times K} \to \mathbb{H}_H^{L\times K}$, and we find that $\widetilde{\boldsymbol{\mathcal{X}}}_{I_q} = \boldsymbol{\Pi}_H \boldsymbol{\mathcal{X}}_{I_q}$. Now we define $\tilde{\mathbf{y}}_N^q = \boldsymbol{\mathcal{T}}^{-1}\widetilde{\boldsymbol{\mathcal{X}}}_{I_q}$ as a reconstructed FTS. Similar to SSA, the goal of FSSA is to separate out an FTS into additive components. The extension of separability to the FTS realm is easily made based on the SSA discussion of separability. From here, we give a review of the MSSA algorithm.

General Scheme of Multivariate Singular Spectrum Analysis

Often times, many variables are observed as a result of a single stochastic process and investigation of time series components can be made richer by performing a multivariate analysis of these vector observations. The MSSA algorithm is a technique that has seen success over SSA in decomposing a multidimensional time series into interpretable components if the covariates are moderately correlated [4]. MSSA also has been broken up into two approaches of vertical MSSA (VMSSA) and horizontal MSSA (HMSSA) where VMSSA involves the vertical stacking of univariate Hankel trajectory matrices while HMSSA works with the horizontal stacking of the same elements [14]. These two approaches to MSSA have both been extended to allow for prediction of multidimensional time series via recurrent and vector forecasting algorithms where the extensions for the horizontal approach are presented in Golyandina et al. [4] and extensions for both approaches are presented in Hassani and Mahmoudvand [14]. Over the course of the last 15 years MSSA has seen significant success in various areas of application see [4, 9, 15, 27]. From here, we offer the two approaches of VMSSA and HMSSA.

Multivariate Singular Spectrum Analysis Algorithm

Given p univariate time series of length N, $\{y_i^{(j)}\}_{i=1,\ldots,N}^{j=1,\ldots,p}$, a multivariate time series can be considered as a series of length N of p-dimensional vectors, $\vec{y}_i = \left[y_i^{(1)}, \ldots, y_i^{(p)}\right]^\top$. This allows us to write the multivariate time series in the form of the $N \times p$ matrix, $\mathbf{y}_N = \left[\vec{y}_1 \cdots \vec{y}_N\right]^\top \in \mathbb{R}^{N\times p}$. In the first step of MSSA, we form a matrix that we may apply an SVD to with the goal of extracting out time-dependent modes of variation and the manner in how we perform the first step gives rise to VMSSA or HMSSA.

Step 1. Embedding

We choose an integer L, where $1 < L < N/2$, set $K = N - L + 1$, and create the set of $L \times K$, univariate trajectory matrices, $\{\mathbf{X}^{(j)}\}_{j=1}^p$. These trajectory matrices have the form

$$\mathbf{X}^{(j)} = \left[\mathbf{x}_1^{(j)} \cdots \mathbf{x}_K^{(j)}\right], \tag{13}$$

where $\mathbf{x}_k^{(j)} = \left[y_k^{(j)}, \ldots, y_{k+L-1}^{(j)}\right]^\top \in \mathbb{R}^L$ is referred to as the k^{th}, L-lagged vector of $\mathbf{y}_N$, associated with variable j. In HMSSA, we concatenate the univariate trajectory matrices horizontally to obtain an $L \times pK$ multivariate trajectory matrix

$$\mathbf{X} = \left[\mathbf{X}^{(1)} \cdots \mathbf{X}^{(p)}\right], \tag{14}$$

whereas in the VMSSA approach, we concatenate those univariate trajectory matrices vertically to obtain the associated $pL \times K$ multivariate trajectory matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(p)} \end{bmatrix}. \tag{15}$$

We note that since each univariate trajectory matrix, $\mathbf{X}^{(j)}$, is Hankel (antidiagonal elements are all equal), the multivariate trajectory matrices generated by the embedding step of HMSSA and VMSSA are block Hankel.

As shown in Sect. 3, there is an interchangeable relationship between the functional extension of VMSSA and multivariate FSSA (MFSSA). Without loss of generality, in the remainder of this section we review VMSSA and so hereafter, we define the $pL \times K$ multivariate trajectory matrix, $\mathbf{X} : \mathbb{R}^K \to \mathbb{R}^{pL}$, to be that of Eq. (15). Often times, this embedding step is viewed as applying an invertible transformation $\mathcal{T} : \mathbb{R}^{N\times p} \to \mathbb{R}_H^{pL\times K}$ such that

$$\mathbf{X} = \mathcal{T}(\mathbf{y}_N),$$

where $\mathbb{R}_H^{pL\times K}$ is the space of all $pL \times K$ block Hankel matrices. In the next step, we extract out time-dependent modes of variation from $\mathbf{X}$.

Step 2. Decomposition

We obtain an SVD of the rank $r \in \mathbb{N}$ $(1 \leq r \leq K)$ multivariate trajectory matrix, $\mathbf{X}$, given by

$$\mathbf{X} = \sum_{i=1}^{r} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^\top = \sum_{i=1}^{r} \mathbf{X}_i,$$

where $\{\sqrt{\lambda_i}\}_{i=1}^r$ are the singular values of $\mathbf{X}$, $\{\mathbf{u}_i\}_{i=1}^r$ are the orthonormal left singular vectors in $\mathbb{R}^{pL}$, and $\{\mathbf{v}_i\}_{i=1}^r$ are the orthonormal right singular vectors in $\mathbb{R}^K$. We define the i^{th} eigentriple of $\mathbf{X}$ as the set $\{\sqrt{\lambda_i}, \mathbf{u}_i, \mathbf{v}_i\}$ and from each eigentriple, we calculate a rank one $pL \times K$ elementary matrix, $\mathbf{X}_i = \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^\top$, to be used in grouping. In step three, we group eigentriples together that describe similar multivariate time series behavior such as mean, periodicity, trend, etc.

Step 3. Grouping

For grouping we partition the set of indices, $\{1, 2, \ldots, r\}$, into m disjoint subsets, $\{I_1, I_2, \ldots, I_m\}$, such that for any positive integer $q = 1, \ldots, m$, the $pL \times K$ matrix, $\mathbf{X}_{I_q}$, is defined as $\mathbf{X}_{I_q} = \sum_{i\in I_q} \mathbf{X}_i$. This allows us to write the original multivariate trajectory matrix, $\mathbf{X}$, as

$$\mathbf{X} = \mathbf{X}_{I_1} + \mathbf{X}_{I_2} + \cdots + \mathbf{X}_{I_m}. \tag{16}$$

The grouping should be done so that each $\mathbf{X}_{I_q}$ describes a different feature of the original multivariate time series such as trend or seasonality. The grouping can be achieved by looking at exploratory plots like paired plots or scree plots [6, 14]. In step four, we extract out a multivariate time series from each $\mathbf{X}_{I_q}$ that captures a different component of variation present in $\mathbf{y}_N$.

Step 4. Reconstruction

Note that the $\mathbf{X}_{I_q}$'s $(q = 1, \ldots, m)$, given in Eq. (16), are not necessarily block Hankel, and therefore we cannot use the $\mathcal{T}^{-1}$ transformation to extract out a multivariate time series from each $\mathbf{X}_{I_q}$. A popular remedy in literature is to use the orthogonal projection approach and approximate the $\mathbf{X}_{I_q}$'s with appropriate block Hankel matrices.

The matrix, $\mathbf{X}_{I_q}$, can be written in the block form:

$$\mathbf{X}_{I_q} = \begin{bmatrix} \mathbf{X}_{I_q}^{(1)} \\ \vdots \\ \mathbf{X}_{I_q}^{(p)} \end{bmatrix},$$

where $\mathbf{X}_{I_q}^{(j)}$ is an $L \times K$ matrix for $j = 1, \ldots, p$. The orthogonal projection of $\mathbf{X}_{I_q}$ onto the space of block Hankel matrices is done by averaging the antidiagonal elements of each $\mathbf{X}_{I_q}^{(j)}$. We denote this approximated $pL \times K$ block Hankel matrix as $\widetilde{\mathbf{X}}_{I_q}$, and use the inverse transformation, $\mathcal{T}^{-1} : \mathbb{R}_H^{pL \times K} \to \mathbb{R}^{N \times p}$, to obtain the q^{th} reconstructed multivariate time series

$$\tilde{\mathbf{y}}_N^q = \mathcal{T}^{-1}(\widetilde{\mathbf{X}}_{I_q}) \in \mathbb{R}^{N \times p}.$$

We define the w-correlation of two multidimensional time series as $\langle \mathbf{y}_N, \mathbf{z}_N \rangle = \sum_{j=1}^{p} \sum_{i=1}^{N} w_i y_i^{(j)} z_i^{(j)}$ and from here, we find that the extension of separability for a multivariate time series is straightforward from the discussion of separability offered in the review of SSA.

Given the discussions on FSSA and MSSA, it is natural to combine ideas from both of these methods to develop an algorithm that can perform the decomposition process on time-dependent multivariate functional data. Multivariate functional data are observed when a stochastic process gives rise to multiple different functions over possibly different dimensional domains. Multivariate FPCA (MFPCA) was developed so that more than one variable of functional subjects could be included in the decomposition routine. Chiou et al. [2] extended MFPCA to include a normalized approach which accounts for differences in degrees of variability in the covariates as well as differences in units. MFPCA was further extended by Happ and Greven [12] to account for different dimensional domains so that one could perform dimension reduction on multivariate functional data that might be comprised of curves, surfaces, or any other finite dimensional domain altogether. A primary assumption of MFPCA is that the functional data are independent of time. With the goal of performing decomposition on a multivariate FTS (MFTS), one might conjecture to use FSSA on the covariates independently of one another but such an approach fails to capture any cross-correlations between variables. MFSSA provides us a way to perform decomposition of a MFTS while capturing these cross-correlations to further enrich analysis and strengthen reconstruction accuracy of the true signal. In addition, MFSSA is developed in the following to handle functions taken over any finite dimensional domain. This can allow the user to explore relationships between time-dependent curves, images, or any other function whose domain is compact.

The rest of the chapter is organized as follows. In Sect. 2, we derive the MFSSA algorithm and we give the computer implementation strategy of the proposed method. In Sect. 3, we derive the functional extensions to VMSSA and HMSSA and we find that the decompositions obtained from the functional extension to VMSSA and MFSSA are equivalent due to isomorphic vector spaces. Next in Sect. 4 we give a simulation study that shows the MFSSA algorithm outperforms all other approaches of MFTS signal extraction. Also in Sect. 4, we offer two real data studies where the first is of smoothed remote sensing images and intraday temperature curves that showcase how MFSSA enriches the signal extraction process for correlated data observed over different dimensional domains. The second real data study uses remotely sensed near-infrared (NIR) and shortwave infrared (SWIR) images of surface reflectance

(SR) from which kernel density estimates (KDE's) are extracted and analyzed using MFSSA. The second real data study shows that when MFTS variables share the same domain and have similar function values, the signal extraction is strengthened for correlated data. Finally, in Sect. 5, we finish with a discussion on the method and the results of the algorithm.

2 General Scheme of Multivariate Functional Singular Spectrum Analysis

The mathematical foundations in the following are used throughout the chapter and form the theoretical backbone of the MFSSA algorithm.

Preliminaries and Notations

For each $j = 1, \ldots, p$, consider an m_j-dimensional domain, T_j, to be a compact subset of $\mathbb{R}^{m_j}$, and let $\mathbb{F}_j = L^2\left(T_j\right)$ be the Hilbert space of square integrable function vectors defined on T_j. From hereafter, we define the Hilbert space, $\mathbb{H}$, to be the Cartesian product vector space given by $\mathbb{H} = \mathbb{F}_1 \times \cdots \times \mathbb{F}_p$, where each multivariate function vector, $\vec{x} \in \mathbb{H}$, can be denoted by the p-tuple $\left(x^{(1)}, \ldots, x^{(p)}\right)$. Note that $\mathbb{H}$ is equipped with inner product given by

$$\langle \vec{x}, \vec{y} \rangle_{\mathbb{H}} = \sum_{j=1}^{p} \left\langle x^{(j)}, y^{(j)} \right\rangle_{\mathbb{F}_j} = \sum_{j=1}^{p} \int_{T_j} x^{(j)}\left(s_j\right) y^{(j)}\left(s_j\right) ds_j, \quad s_j \in T_j,$$

for some multivariate function vectors, $\vec{x}, \vec{y} \in \mathbb{H}$. From hereafter, we specify an MFTS of length N as $\mathbf{y}_N = (\vec{y}_1, \ldots, \vec{y}_N)$, where $\vec{y}_i \in \mathbb{H}$.

Similarly, for a given $L \in \mathbb{N}$, $\mathbb{H}^L$ is a function vector space that is built from the Cartesian product of L copies of $\mathbb{H}$, and each length L multivariate function vector, $\boldsymbol{x} \in \mathbb{H}^L$, can be denoted by the L-tuple $(\vec{x}_1, ..., \vec{x}_L)$. Clearly $\mathbb{H}^L$ is a Hilbert space with respect to the inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{H}^L} = \sum_{i=1}^{L} \langle \vec{x}_i, \vec{y}_i \rangle_{\mathbb{H}}, \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}^L.$$

Next we define $\mathbb{H}^{L \times K}$ to be the space spanned by linear operators, $\boldsymbol{\mathcal{V}} : \mathbb{R}^K \to \mathbb{H}^L$, specified by multivariate function vectors, $\left(\vec{v}_{i,k}\right)_{i=1,\ldots,L}^{k=1,\ldots,K} \in \mathbb{H}$, as

$$\boldsymbol{\mathcal{V}}(\mathbf{a}) = \left(\sum_{k=1}^{K} a_k \overrightarrow{v}_{1,k}, \ldots, \sum_{k=1}^{K} a_k \overrightarrow{v}_{L,k} \right), \qquad \mathbf{a} = [a_1, a_2, \ldots, a_K]^\top \in \mathbb{R}^K.$$

For two operators $\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{Z}} \in \mathbb{H}^{L \times K}$, the Frobenius inner product can be defined as

$$\langle \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{Z}} \rangle_F = \sum_{i=1}^{L} \sum_{k=1}^{K} \left\langle \overrightarrow{v}_{i,k}, \overrightarrow{z}_{i,k} \right\rangle_{\mathbb{H}},$$

which induces the Frobenius norm given by $\|\boldsymbol{\mathcal{V}}\|_F = \sqrt{\langle \boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{V}} \rangle_F}$. We denote $\mathbb{H}_H^{L \times K}$ to be the Hankel subspace of $\mathbb{H}^{L \times K}$ such that for any operator, $\widetilde{\boldsymbol{\mathcal{V}}} = \left(\overrightarrow{\tilde{v}}_{i,k} \right) \in \mathbb{H}_H^{L \times K}$, there exists a multivariate function vector, $\overrightarrow{g}_u \in \mathbb{H}$, such that $\| \overrightarrow{\tilde{v}}_{i,k} - \overrightarrow{g}_u \|_{\mathbb{H}} = 0$ where $u = i + k$.

Multivariate Functional Singular Spectrum Analysis Algorithm

Similar to other SSA-based algorithms, MFSSA consists of four steps: Embedding, Decomposition, Grouping, and Reconstruction. In the first step of embedding, we establish a so-called multivariate trajectory operator whose range is built from elements that describe MFTS behavior over sub-intervals of time.

Step 1. Embedding

As one may note, the columns of a univariate trajectory matrix, as given in Eq. (13), are the corresponding L-lagged vectors for the j^{th} variable. Therefore, a trajectory matrix can be seen as a linear operator from $\mathbb{R}^K$ to the space of linear combinations of the L-lagged vectors. We use this as a motivation to introduce the trajectory operator for FSSA.

In a similar fashion, we define the L-lagged multivariate function vectors in $\mathbb{H}^L$ of the form

$$\boldsymbol{x}_k = (\overrightarrow{y}_k, \overrightarrow{y}_{k+1}, \ldots, \overrightarrow{y}_{k+L-1}), \quad k = 1, \ldots, K \tag{17}$$

which correspond with the MFTS, $\mathbf{y}_N$. We define a linear operator, specified with $\boldsymbol{x}_k$'s, to obtain the multivariate trajectory operator, $\boldsymbol{\mathcal{X}} : \mathbb{R}^K \to \mathbb{H}^L$. As such, for some vector $\mathbf{a} = [a_1, a_2, \ldots, a_K]^\top \in \mathbb{R}^K$, we have

$$\boldsymbol{\mathcal{X}}(\mathbf{a}) = \sum_{k=1}^{K} a_k \boldsymbol{x}_k. \tag{18}$$

We define $R(\boldsymbol{\mathcal{X}}) = \text{span}\{\boldsymbol{x}_k\}_{k=1}^K$ to be the range of the operator $\boldsymbol{\mathcal{X}}$, and clearly the rank of the multivariate trajectory operator is, $r \in \mathbb{N}$, where $1 \leq r \leq K$. This step of embedding can also be viewed as applying the invertible transformation, $\boldsymbol{\mathcal{T}} : \mathbb{H}^N \to \mathbb{H}_H^{L\times K}$, such that

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}}(\mathbf{y}_N). \tag{19}$$

In the following proposition, we establish some useful properties of $\boldsymbol{\mathcal{X}}$ and we define it's adjoint.

Proposition 3 *The operator given in Eq.* (18) *is a bounded and linear operator with adjoint* $\boldsymbol{\mathcal{X}}^* : \mathbb{H}^L \to \mathbb{R}^K$ *where for some length L multivariate function vector,* $\boldsymbol{z} \in \mathbb{H}^L$*, we obtain the following K-dimensional vector*

$$\boldsymbol{\mathcal{X}}^*\boldsymbol{z} = \left[\langle \boldsymbol{x}_1, \boldsymbol{z}\rangle_{\mathbb{H}^L}, \langle \boldsymbol{x}_2, \boldsymbol{z}\rangle_{\mathbb{H}^L}, \ldots, \langle \boldsymbol{x}_K, \boldsymbol{z}\rangle_{\mathbb{H}^L}\right]^\top \in \mathbb{R}^K.$$

Proof Notice that $\boldsymbol{\mathcal{X}}$ is a finite-rank operator, and thus $\boldsymbol{\mathcal{X}}$ is bounded and linear as a result. Now let $\boldsymbol{z} \in \mathbb{H}^L$, then

$$\langle \boldsymbol{\mathcal{X}}\mathbf{a}, \boldsymbol{z}\rangle_{\mathbb{H}^L} = \sum_{k=1}^K a_k \langle \boldsymbol{x}_k, \boldsymbol{z}\rangle_{\mathbb{H}^L} = \mathbf{a}^\top \boldsymbol{\mathcal{X}}^*\boldsymbol{z} = \left\langle \mathbf{a}, \boldsymbol{\mathcal{X}}^*\boldsymbol{z}\right\rangle_{\mathbb{R}^K}, \quad \mathbf{a} \in \mathbb{R}^K$$

and we have that $\boldsymbol{\mathcal{X}}^*$ is the adjoint of $\boldsymbol{\mathcal{X}}$. □

Now we discuss the second step of MFSSA where we extract out time-dependent modes of variation from $\boldsymbol{\mathcal{X}}$.

Step 2. Decomposition

Since $\boldsymbol{\mathcal{X}}$ is a finite-rank operator, one may employ Theorem 7.6 of Weidmann [30] and obtain a multivariate FSVD (MFSVD) for $\boldsymbol{\mathcal{X}}$ as

$$\boldsymbol{\mathcal{X}}(\mathbf{a}) = \sum_{i=1}^r \sqrt{\lambda_i}\langle \mathbf{v}_i, \mathbf{a}\rangle_{\mathbb{R}^K} \boldsymbol{\psi}_i = \sum_{i=1}^r \sqrt{\lambda_i}\mathbf{v}_i \otimes \boldsymbol{\psi}_i(\mathbf{a}), \tag{20}$$

where $\otimes$ stands for the tensor (outer) product. From hereafter, $\{\sqrt{\lambda_i}\}_{i=1}^r$ are the singular values, $\{\boldsymbol{\psi}_i\}_{i=1}^r$ are the orthonormal left singular functions spanning an r-dimensional subspace of $\mathbb{H}^L$, and $\{\mathbf{v}_i\}_{i=1}^r$ are the orthonormal right singular vectors spanning an r-dimensional subspace of $\mathbb{R}^K$. We call the set, $\{\sqrt{\lambda_i}, \boldsymbol{\psi}_i, \mathbf{v}_i\}$, as the i^{th} eigentriple of $\boldsymbol{\mathcal{X}}$. Now we define the rank one elementary operators built from the eigentriples, $\boldsymbol{\mathcal{X}}_i : \mathbb{R}^K \to \mathbb{H}^L$, given by $\boldsymbol{\mathcal{X}}_i = \sqrt{\lambda_i}\mathbf{v}_i \otimes \boldsymbol{\psi}_i$ for $i = 1, \ldots, r$. It is easy to see that $\boldsymbol{\mathcal{X}} = \sum_{i=1}^r \boldsymbol{\mathcal{X}}_i$. The following proposition establishes equalities between $\boldsymbol{\mathcal{X}}$ and eigentriple elements.

Proposition 4 *Let* $\{\sqrt{\lambda_i}, \boldsymbol{\psi}_i, \mathbf{v}_i\}$ *be the* i^{th} *eigentriple of* $\boldsymbol{\mathcal{X}}$*,* $i = 1, \ldots, r$*. The following holds:*

$$\boldsymbol{\psi}_i = \left(\sqrt{\lambda_i}\right)^{-1} \boldsymbol{\mathcal{X}} \mathbf{v}_i, \quad \mathbf{v}_i = \left(\sqrt{\lambda_i}\right)^{-1} \boldsymbol{\mathcal{X}}^* \boldsymbol{\psi}_i.$$

Proof Recall that an MFSVD of $\boldsymbol{\mathcal{X}}$ gives us a set of orthonormal left singular functions, $\{\boldsymbol{\psi}_i\}_{i=1}^r$, and a set of orthonormal vectors, $\{\mathbf{v}_i\}_{i=1}^r$. We find that

$$\boldsymbol{\mathcal{X}}(\mathbf{v}_i) = \sum_{j=1}^{r} \sqrt{\lambda_j} \left(\mathbf{v}_i^\top \mathbf{v}_j\right) \boldsymbol{\psi}_j = \sqrt{\lambda_i} \boldsymbol{\psi}_i.$$

From here, we have that $\boldsymbol{\psi}_i = \left(\sqrt{\lambda_i}\right)^{-1} \boldsymbol{\mathcal{X}}(\mathbf{v}_i)$. By Theorem 7.6 of Weidmann [30], we have that the operator, $\boldsymbol{\mathcal{X}}^*$, has an MFSVD with the same eigentriples as $\boldsymbol{\mathcal{X}}$ and we obtain the following:

$$\boldsymbol{\mathcal{X}}^*(\boldsymbol{\psi}_i) = \sum_{j=1}^{r} \sqrt{\lambda_j} \left\langle \boldsymbol{\psi}_i, \boldsymbol{\psi}_j \right\rangle_{\mathbb{H}^L} \mathbf{v}_j = \sqrt{\lambda_i} \mathbf{v}_i$$

which implies that $\mathbf{v}_i = \left(\sqrt{\lambda_i}\right)^{-1} \boldsymbol{\mathcal{X}}^*(\boldsymbol{\psi}_i)$. □

Now we present step three of the MFSSA algorithm where we group together eigentriples that describe similar MFTS behavior.

Step 3. Grouping

The grouping step of MFSSA follows the same manner as the grouping step of MSSA. We partition the set of indices, $\{1, 2, \ldots, r\}$, into m disjoint subsets, $\{I_1, I_2, \ldots, I_m\}$, such that for $q = 1, \ldots m$, the operator, $\boldsymbol{\mathcal{X}}_{I_q} : \mathbb{R}^K \to \mathbb{H}^L$, is defined by $\boldsymbol{\mathcal{X}}_{I_q} = \sum_{i \in I_q} \boldsymbol{\mathcal{X}}_i$. As such, we have that

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}_{I_1} + \boldsymbol{\mathcal{X}}_{I_2} + \cdots + \boldsymbol{\mathcal{X}}_{I_m}.$$

Similar to other forms of SSA, exploratory plots, such as scree plots, paired plots, w-correlation plots, and others can be developed to determine how to obtain the m disjoint groups. In addition, we also have that each operator, $\boldsymbol{\mathcal{X}}_{I_q}$, corresponds with some component of the original MFTS such as mean, periodic, trend, or other behaviors. In the next step, our goal is to extract an MFTS from each $\boldsymbol{\mathcal{X}}_{I_q}$ that describes one of these modes of variation.

Step 4. Reconstruction

Let $\boldsymbol{\mathcal{Y}} \in \mathbb{H}^{L \times K}$ be an operator defined by $\boldsymbol{\mathcal{Y}} = \left(\vec{y}_{i,j}\right)$. Since $\mathbb{H}_H^{L \times K}$ is a closed subspace of $\mathbb{H}^{L \times K}$, we have by Theorem 3.2.3 of Shalit [25] that there exists a unique Hankel operator, $\widetilde{\boldsymbol{\mathcal{Y}}} = \left(\vec{\tilde{y}}_{i,j}\right) \in \mathbb{H}_H^{L \times K}$, such that

$$\|\boldsymbol{\mathcal{Y}} - \widetilde{\boldsymbol{\mathcal{Y}}}\|_F^2 \leq \|\boldsymbol{\mathcal{Y}} - \widetilde{\boldsymbol{\mathcal{Z}}}\|_F^2,$$

for any Hankel operator, $\widetilde{\boldsymbol{\mathcal{Z}}} \in \mathbb{H}_H^{L\times K}$. Define the orthogonal projection, $\boldsymbol{\Pi}_H : \mathbb{H}^{L\times K} \to \mathbb{H}_H^{L\times K}$, such that we have $\widetilde{\boldsymbol{\mathcal{Y}}} = \boldsymbol{\Pi}_H \boldsymbol{\mathcal{Y}}$. We achieve the projection, $\boldsymbol{\Pi}_H$, by using Lemma 1 and the resulting diagonal averaging technique given by

$$\vec{\tilde{y}}_{i,k} = \frac{1}{n_u} \sum_{(n,m):n+m=u} \vec{y}_{n,m}, \tag{21}$$

where n_u is the number of (n, m) pairs such that $n + m = u$. With $\boldsymbol{\Pi}_H$, we have that $\widetilde{\boldsymbol{\mathcal{X}}}_{I_q} = \boldsymbol{\Pi}_H \boldsymbol{\mathcal{X}}_{I_q}$ for $q = 1, \ldots, m$. We then employ the inverse of $\boldsymbol{\mathcal{T}} : \mathbb{H}^N \to \mathbb{H}_H^{L\times K}$, from Eq. (19), to obtain the following formula for the reconstruction:

$$\tilde{\mathbf{y}}_N^q = \boldsymbol{\mathcal{T}}^{-1} \widetilde{\boldsymbol{\mathcal{X}}}_{I_q}.$$

Similar to other types of SSA, we have that each $\tilde{\mathbf{y}}_N^q$ is an MFTS that describes a component of the original signal, $\mathbf{y}_N$, such as mean behavior, periodicity, trend, etc. Even though MFSSA is a model-free procedure, it is ideal that $\mathbf{y}_N$ have additive components which implies that the MFTS is separable. In the following, we introduce the idea of separability for MFTS.

Separability

Let $\mathbf{x}_N = \mathbf{y}_N + \mathbf{z}_N$ where $\mathbf{y}_N$ and $\mathbf{z}_N$ are MFTS of length N. We define the weighted-covariance between the MFTS as

$$\langle \mathbf{y}_N, \mathbf{z}_N \rangle_w = \sum_{j=1}^{p} \sum_{i=1}^{N} w_i \left\langle y_i^{(j)}, z_i^{(j)} \right\rangle_{\mathbb{F}_j},$$

where $w_i = \min\{i, L, N - i + 1\}$. The definition of the norm, $\|\mathbf{y}_N\|_w$, the discussion on w-orthogonality of $\mathbf{y}_N$ and $\mathbf{z}_N$, and the definition of the w-correlation matrix is straightforward from the FSSA discussion on separability and we may use these tools to help in the grouping stage of MFSSA. Now we discuss the computer implementation strategy.

Computer Implementation Strategy

In the following, we describe the basis that is used to represent a multivariate function vector in $\mathbb{H}$. We then use the basis for the multivariate function vectors to construct the basis used to build the L-lagged multivariate function vectors in $\mathbb{H}^L$, and the implementation of the multivariate trajectory operator, $\boldsymbol{\mathcal{X}}$.

Let $\{v_i^{(j)}\}_{i\in\mathbb{N}}$ be the collection of linearly independent basis function vectors in the function vector space, $\mathbb{F}_j$ for $j = 1, ..., p$. Each observation in $\mathbb{F}_j$ can be projected onto the subspace $F_j = \text{span}\{v_i^{(j)}\}_{i=1}^{d_j}$ where d_j is a positive integer that can be determined by a variety of techniques like cross-validation. To this end, each function vector, $y_i^{(j)} \in F_j$, can be uniquely expressed as

$$y_i^{(j)} = \sum_{k=1}^{d_j} c_{i,k}^{(j)} v_k^{(j)}, \quad i = 1, ..., N, \quad c_{i,k}^{(j)} \in \mathbb{R}.$$

Now we set $d_0 = 0$, $d = \sum_{j=0}^{p} d_j$, and define the Hilbert space, $\mathbb{H}_d = F_1 \times \ldots \times F_p \subset \mathbb{H}$.

For each $q \in \{1, \ldots, d\}$, there exist a unique $j_q \in \{1, \ldots, p\}$ such that $\sum_{i=0}^{j_q-1} d_i + 1 \leq q \leq \sum_{i=0}^{j_q} d_i$. Now consider the multivariate function vector, $\vec{v}_q \in \mathbb{H}_d$, which is the zero function in all components except the j_q^{th} element, which is $v_{\ell_q}^{(j_q)}$, where $\ell_q = q - \sum_{i=0}^{j_q-1} d_i$. We expand on some interesting properties of the multivariate function vectors, $\vec{v}_q$'s, in the following Lemma.

Lemma 2 *The following holds:*

(i) Each multivariate function vector, $\vec{y}_i \in \mathbb{H}_d$, can be uniquely represented as a linear combination of $\vec{v}_q$'s such that

$$\vec{y}_i = \sum_{q=1}^{d} c_{i,\ell_q}^{(j_q)} \vec{v}_q, \quad i = 1, ..., N.$$

(ii) The set $\{\vec{v}_q\}_{q=1}^{d}$ is a basis system of $\mathbb{H}_d$.

Proof We prove the two parts of this lemma in the following.

(i) Let $M_{j_q} = \sum_{i=0}^{j_q} d_i$, then we obtain the following multivariate function vectors of $\mathbb{H}_d$.

$$\vec{y}_i^{(1)} = \left(y_i^{(1)}, 0, \ldots, 0\right) = \sum_{q=1}^{d_1} c_{i,\ell_q}^{(1)} \vec{v}_q$$

$$\vec{y}_i^{(2)} = \left(0, y_i^{(2)}, 0, \ldots, 0\right) = \sum_{q=d_1+1}^{d_1+d_2} c_{i,\ell_q}^{(2)} \vec{v}_q$$

$$\vdots$$

$$\vec{y}_i^{(j_q)} = \left(0, \ldots, 0, y_i^{(j_q)}, 0, \ldots, 0\right) = \sum_{q=M_{j_q-1}+1}^{M_{j_q}} c_{i,\ell_q}^{(j_q)} \vec{v}_q$$

$$\vdots$$

$$\vec{y}_i^{(p)} = \left(0, \ldots, 0, y_i^{(p)}\right) = \sum_{q=M_{p-1}+1}^{d} c_{i,\ell_q}^{(p)} \vec{v}_q.$$

From this, we find that any multivariate function vector, $\vec{y}_i \in \mathbb{H}_d$, can be expressed as

$$\begin{aligned}\vec{y}_i &= \left(y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(j_q)}, \ldots, y_i^{(p)}\right) \\ &= \vec{y}_i^{(1)} + \vec{y}_i^{(2)} + \cdots + \vec{y}_i^{(j_q)} + \cdots + \vec{y}_i^{(p)} = \sum_{q=1}^{d} c_{i,\ell_q}^{(j_q)} \vec{v}_q.\end{aligned}$$

(ii) This part of the proof is a direct consequence of the proof of part *(i)*. □

Now for each $k \in \{1, \ldots, Ld\}$, one can see that there exists unique $q_k \in \{1, \ldots, d\}$ and $r_k \in \{1, \ldots, L\}$ such that $k = (q_k - 1)L + r_k$. Consider $\boldsymbol{\phi}_k$ as a length L multivariate function vector that is zero in all components, except the r_k-th component, which is $\vec{v}_{q_k}$. The following lemma now defines a linearly independent basis for the L-lagged multivariate function vectors, $\boldsymbol{x}_j$'s, where the basis is defined using the $\vec{v}_q$'s.

Lemma 3 *The sequence, $\{\boldsymbol{\phi}_k\}_{k=1}^{Ld}$, is a basis system for $\mathbb{H}_d^L$, where $\mathbb{H}_d^L$ is the Cartesian product of L copies of $\mathbb{H}_d$.*

Proof The proof will be divided into two steps. In the first step, it will be shown that $\mathbb{H}_d^L = \text{span}\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{Ld}\}$. In the second step, it will be proved that $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{Ld}$ are linearly independent. Let $\boldsymbol{z} = (\vec{z}_1, \ldots, \vec{z}_L)$ be a length L multivariate function vector in $\mathbb{H}_d^L$. By definition, each component of $\boldsymbol{z}$ admits the basis expansions $\vec{z}_j = \sum_{q=1}^{d} c_{j,\ell_q}^{(j_q)} \vec{v}_q$, $j = 1, \ldots, L$. Therefore

$$\boldsymbol{z} = \begin{pmatrix} \vec{z}_1 \\ \vec{z}_2 \\ \vdots \\ \vec{z}_L \end{pmatrix} = \begin{pmatrix} \sum_{q=1}^{d} c_{1,\ell_q}^{(j_q)} \vec{v}_q(s) \\ \sum_{q=1}^{d} c_{2,\ell_q}^{(j_q)} \vec{v}_q(s) \\ \vdots \\ \sum_{q=1}^{d} c_{L,\ell_q}^{(j_q)} \vec{v}_q(s) \end{pmatrix}$$

$$= c_{1,\ell_1}^{(j_1)} \begin{pmatrix} \vec{\nu}_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + c_{L,\ell_1}^{(j_1)} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vec{\nu}_1 \end{pmatrix} + c_{1,\ell_2}^{(j_2)} \begin{pmatrix} \vec{\nu}_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$+ \cdots + c_{L,\ell_2}^{(j_2)} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vec{\nu}_2 \end{pmatrix} + \cdots + c_{1,\ell_d}^{(j_d)} \begin{pmatrix} \vec{\nu}_d \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + c_{L,\ell_d}^{(j_d)} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vec{\nu}_d \end{pmatrix} = \sum_{k=1}^{Ld} c_{r_k,\ell_{q_k}}^{(j_{q_k})} \boldsymbol{\phi}_k,$$

which implies the first step. To prove linear independency, if $\sum_{k=1}^{Ld} a_k \boldsymbol{\phi}_k = 0$, then

$$\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = a_1 \begin{pmatrix} \vec{\nu}_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + a_L \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vec{\nu}_1 \end{pmatrix} + a_{L+1} \begin{pmatrix} \vec{\nu}_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + a_{2L} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vec{\nu}_2 \end{pmatrix}$$

$$+ \cdots + a_{(d-1)L+1} \begin{pmatrix} \vec{\nu}_d \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + a_{dL} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vec{\nu}_d \end{pmatrix}$$

$$= \begin{pmatrix} a_1 \vec{\nu}_1 + a_{L+1} \vec{\nu}_2 + \cdots + a_{(d-1)L+1} \vec{\nu}_d \\ a_2 \vec{\nu}_1 + a_{L+2} \vec{\nu}_2 + \cdots + a_{(d-1)L+2} \vec{\nu}_d \\ \vdots \\ a_L \vec{\nu}_1 + a_{2L} \vec{\nu}_2 + \cdots + a_{dL} \vec{\nu}_d \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{d} a_{(j-1)L+1} \vec{\nu}_j \\ \sum_{j=1}^{d} a_{(j-1)L+2} \vec{\nu}_j \\ \vdots \\ \sum_{j=1}^{d} a_{jL} \vec{\nu}_j \end{pmatrix}.$$

This means $\sum_{j=1}^{d} a_{(j-1)L+i} \vec{\nu}_j = 0$ for all $i = 1, \ldots, L$, and consequently each scalar $a_{(j-1)L+i} = 0$, for $j = 1, \ldots, d$, since $\{\vec{\nu}_j\}_{j=1}^{d}$ are linearly independent. □

Using Lemma 3, we define a linear operator $\boldsymbol{\mathcal{P}} : \mathbb{R}^{Ld} \to \mathbb{H}_d^L$, specified with $\boldsymbol{\phi}_k$'s, such that any length L multivariate function vector, $\boldsymbol{x} \in \mathbb{H}_d^L$, can be written as

$$\boldsymbol{x} = \sum_{i=1}^{Ld} b_i \boldsymbol{\phi}_i = \boldsymbol{\mathcal{P}}(\mathbf{b}),$$

where $\mathbf{b} \in \mathbb{R}^{Ld}$. Similar to Eq. (17), we define the L-lagged multivariate function vectors for the MFTS, $\mathbf{y}_N$, as $\boldsymbol{x}_k = (\vec{y}_k, \vec{y}_{k+1}, \ldots, \vec{y}_{k+L-1}) \in \mathbb{H}_d^L$, where $k = 1, \ldots, K$. Therefore, the associated multivariate trajectory operator, given in Eq. (18), is defined as $\boldsymbol{\mathcal{X}} : \mathbb{R}^K \to \mathbb{H}_d^L$. The next lemma gives the tools necessary to implement the L-lagged multivariate function vectors and $\boldsymbol{\mathcal{X}}$ using the $\boldsymbol{\phi}_i$'s.

Lemma 4 *The following holds:*

(i) The corresponding coefficient vector of the k^{th} L-lagged multivariate function vector, $\boldsymbol{x}_k$, with respect to the operator, $\boldsymbol{\mathcal{P}}$, is

$$\mathbf{b}_k = \left[c_{k,\ell_1}^{(j_1)}, \ldots, c_{k+L-1,\ell_1}^{(j_1)}, c_{k,\ell_2}^{(j_2)}, \ldots, c_{k+L-1,\ell_2}^{(j_2)}, \ldots, c_{k+L-1,\ell_d}^{(j_d)}\right]^\top \in \mathbb{R}^{Ld}.$$

(ii) For any vector $\mathbf{a} \in \mathbb{R}^K$, we have that $\boldsymbol{\mathcal{X}}(\mathbf{a}) = \boldsymbol{\mathcal{P}}(\mathbf{Ba})$, where $\mathbf{B} = \left[b_{i,k}\right]_{i=1,\ldots,Ld}^{k=1,\ldots,K} = [\mathbf{b}_1\mathbf{b}_2\cdots\mathbf{b}_K]_{Ld\times K}$ is a matrix, and $b_{i,k}$ is the i^{th} component of the Ld-dimensional vector, $\mathbf{b}_k$.

Proof We prove the two parts of this lemma in the following:

(i) Let $M_{j_q} = \sum_{i=0}^{j_q} d_i$, then we obtain the following length L multivariate function vectors of $\mathbb{H}_d^L$.

$$\boldsymbol{x}_k^{(1)} = \begin{pmatrix} \overrightarrow{y}_k^{(1)} \\ \overrightarrow{y}_{k+1}^{(1)} \\ \vdots \\ \overrightarrow{y}_{k+L-1}^{(1)} \end{pmatrix} = \begin{pmatrix} \overrightarrow{y}_k^{(1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \overrightarrow{y}_{k+1}^{(1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \overrightarrow{y}_{k+L-1}^{(1)} \end{pmatrix} = \sum_{i=1}^{Ld_1} b_{i,k}\boldsymbol{\phi}_i$$

$$= \sum_{q_i=1}^{d_1}\sum_{r_i=1}^{L} c_{k+r_i-1,\ell_{q_i}}^{(1)}\boldsymbol{\phi}_i$$

$$\boldsymbol{x}_k^{(2)} = \begin{pmatrix} \overrightarrow{y}_k^{(2)} \\ \overrightarrow{y}_{k+1}^{(2)} \\ \vdots \\ \overrightarrow{y}_{k+L-1}^{(2)} \end{pmatrix} = \begin{pmatrix} \overrightarrow{y}_k^{(2)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \overrightarrow{y}_{k+1}^{(2)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \overrightarrow{y}_{k+L-1}^{(2)} \end{pmatrix} = \sum_{i=Ld_1+1}^{L(d_1+d_2)} b_{i,k}\boldsymbol{\phi}_i$$

$$= \sum_{q_i=d_1+1}^{d_1+d_2}\sum_{r_i=1}^{L} c_{k+r_i-1,\ell_{q_i}}^{(2)}\boldsymbol{\phi}_i$$

$$\vdots$$

$$\boldsymbol{x}_k^{(p)} = \begin{pmatrix} \overrightarrow{y}_k^{(p)} \\ \overrightarrow{y}_{k+1}^{(p)} \\ \vdots \\ \overrightarrow{y}_{k+L-1}^{(p)} \end{pmatrix} = \begin{pmatrix} \overrightarrow{y}_k^{(p)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \overrightarrow{y}_{k+1}^{(p)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \overrightarrow{y}_{k+L-1}^{(p)} \end{pmatrix} = \sum_{i=LM_{p-1}+1}^{Ld} b_{i,k}\boldsymbol{\phi}_i$$

$$= \sum_{q_i=M_{p-1}+1}^{d} \sum_{r_i=1}^{L} c_{k+r_i-1,\ell_{q_i}}^{(p)} \boldsymbol{\phi}_i.$$

As a result, we find that the k^{th} L-lagged multivariate function vector, $\boldsymbol{x}_k$, is given by $\boldsymbol{x}_k = \boldsymbol{x}_k^{(1)} + \boldsymbol{x}_k^{(2)} + \cdots + \boldsymbol{x}_k^{(p)} = \sum_{i=1}^{Ld} b_{i,k}\boldsymbol{\phi}_i$ and the components of $\mathbf{b}_k$ are found in Lemma 4.

(ii) Now notice that

$$\boldsymbol{\mathcal{X}}(\mathbf{a}) = \sum_{k=1}^{K} a_k \boldsymbol{x}_k = \sum_{i=1}^{Ld} \left(\sum_{k=1}^{K} b_{i,k} a_k \right) \boldsymbol{\phi}_i = \boldsymbol{\mathcal{P}}(\mathbf{Ba}), \quad \mathbf{a} \in \mathbb{R}^K, \tag{22}$$

and the proof is completed. □

Now that we have a way to represent our L-lagged multivariate function vectors and multivariate trajectory operator using the $\boldsymbol{\phi}_i$'s, we leverage the following theorem to obtain the eigentriples of $\boldsymbol{\mathcal{X}}$.

Theorem 2 *Define the $Ld \times K$ matrix, $\mathbf{X} = \mathbf{G}^{1/2}\mathbf{B}$, where $\mathbf{G} = \left[\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle_{\mathbb{H}^L}\right]_{i,j=1}^{Ld}$ is the $Ld \times Ld$ Gram matrix. Denote the collection $\{\sqrt{\lambda_i}, \mathbf{u}_i, \mathbf{v}_i\}$ as the i^{th} eigentriple of $\mathbf{X}$. Now define $\mathbf{w}_i = \mathbf{G}^{-1/2}\mathbf{u}_i$ and $\boldsymbol{\psi}_i = \boldsymbol{\mathcal{P}}(\mathbf{w}_i)$. The following holds:*

(i) $\boldsymbol{\mathcal{X}}^*\boldsymbol{\psi}_i = \sqrt{\lambda_i}\mathbf{v}_i$.
(ii) $\boldsymbol{\mathcal{X}}\mathbf{v}_i = \sqrt{\lambda_i}\boldsymbol{\psi}_i$.
(iii) The collection, $\{\boldsymbol{\psi}_i\}_{i=1}^r$, forms an orthonormal basis for $R(\boldsymbol{\mathcal{X}})$.

Proof We prove each part of this theorem in succession.

(i) For any two length L multivariate function vectors, $\boldsymbol{x} = \boldsymbol{\mathcal{P}}\mathbf{b}$ and $\boldsymbol{y} = \boldsymbol{\mathcal{P}}\mathbf{c}$, in $\mathbb{H}_d^L$, it is easy to see that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{H}^L} = \mathbf{b}^\top \mathbf{G}\mathbf{c}, \quad \mathbf{b}, \mathbf{c} \in \mathbb{R}^{Ld}. \tag{23}$$

Therefore,

$$\boldsymbol{\mathcal{X}}^*\boldsymbol{\psi}_i = \left[\langle \boldsymbol{x}_1, \boldsymbol{\psi}_i \rangle_{\mathbb{H}^L}, \ldots, \langle \boldsymbol{x}_K, \boldsymbol{\psi}_i \rangle_{\mathbb{H}^L}\right]^\top = \left[\mathbf{w}_i^\top \mathbf{G}\mathbf{b}_1, \ldots, \mathbf{w}_i^\top \mathbf{G}\mathbf{b}_k\right]^\top$$
$$= \mathbf{B}^\top \mathbf{G}\mathbf{w}_i = \mathbf{B}^\top \mathbf{G}^{1/2}\mathbf{u}_i = \mathbf{X}^\top \mathbf{u}_i = \sqrt{\lambda_i}\mathbf{v}_i.$$

(ii) Using Eq. (22) gives

$$\begin{aligned}\mathcal{X}\mathbf{v}_i &= \mathcal{P}\mathbf{B}\mathbf{v}_i = \mathcal{P}\mathbf{G}^{-1/2}\mathbf{G}^{1/2}\mathbf{B}\mathbf{v}_i = \mathcal{P}\mathbf{G}^{-1/2}\mathbf{X}\mathbf{v}_i \\ &= \sqrt{\lambda_i}\mathcal{P}\mathbf{G}^{-1/2}\mathbf{u}_i = \sqrt{\lambda_i}\mathcal{P}\mathbf{w}_i = \sqrt{\lambda_i}\boldsymbol{\psi}_i.\end{aligned}$$

(iii) First we prove that the left singular functions, $\boldsymbol{\psi}_i'$s, belong to $R(\mathcal{X})$. From (ii) and Eq. (5), we have

$$\boldsymbol{\psi}_i = \frac{\mathcal{X}\mathbf{v}_i}{\sqrt{\lambda_i}} = \sum_{j=1}^{K} \frac{v_{ij}}{\sqrt{\lambda_i}}\boldsymbol{x}_j \in R(\mathcal{X}), \qquad \text{where } \mathbf{v}_i = [v_{i1}, \ldots, v_{iK}]^\top.$$

Next, to show orthonormality, for any $i, j = 1, \ldots, r$ we have

$$\langle \boldsymbol{\psi}_i, \boldsymbol{\psi}_j \rangle_{\mathbb{H}^L} = \mathbf{w}_i^\top \mathbf{G}\mathbf{w}_j = \mathbf{u}_i^\top \mathbf{u}_j = \delta_{i,j}.$$

We note that $\sum_{i=1}^{r} \mathbf{u}_i\mathbf{u}_i^\top = \mathbf{I}_r$, where $\mathbf{I}_r$ is the $r \times r$ identity matrix and therefore $\sum_{i=1}^{r} \mathbf{w}_i\mathbf{w}_i^\top = \mathbf{G}^{-1}$. We use this result to show that any length L multivariate function vector, $\boldsymbol{x} \in R(\mathcal{X})$, can be written as a linear combination of $\boldsymbol{\psi}_i$'s. To this end, note that for all $\boldsymbol{x} \in R(\mathcal{X})$, there exists a vector $\mathbf{a} \in \mathbb{R}^K$, such that $\boldsymbol{x} = \mathcal{X}\mathbf{a}$. Using Eqs. (22) and (23) we have

$$\begin{aligned}\boldsymbol{x} &= \mathcal{P}\mathbf{B}\mathbf{a} = \mathcal{P}\mathbf{G}^{-1}\mathbf{G}\mathbf{B}\mathbf{a} = \mathcal{P}\left(\sum_{i=1}^{r} \mathbf{w}_i\mathbf{w}_i^\top\right)\mathbf{G}\mathbf{B}\mathbf{a} = \sum_{i=1}^{r} \mathcal{P}\mathbf{w}_i\mathbf{w}_i^\top\mathbf{G}\mathbf{B}\mathbf{a} \\ &= \sum_{i=1}^{r} \boldsymbol{\psi}_i\mathbf{w}_i^\top\mathbf{G}\mathbf{B}\mathbf{a} = \sum_{i=1}^{r} \langle \boldsymbol{x}, \boldsymbol{\psi}_i \rangle_{\mathbb{H}^L}\boldsymbol{\psi}_i.\end{aligned}$$

It remains to be proven that the $\boldsymbol{\psi}_i'$s are linearly independent. Note that

$$\sum_{i=1}^{r} c_i\boldsymbol{\psi}_i = \sum_{i=1}^{r} c_i\mathcal{P}\mathbf{w}_i = \mathcal{P}\sum_{i=1}^{r} c_i\mathbf{G}^{-1/2}\mathbf{u}_i = \mathcal{P}\mathbf{G}^{-1/2}\sum_{i=1}^{r} c_i\mathbf{u}_i.$$

Therefore $\sum_{i=1}^{r} c_i\boldsymbol{\psi}_i = 0$ implies $\sum_{i=1}^{r} c_i\mathbf{u}_i = 0$, and the linear independency of $\mathbf{u}_i$'s implies $c_i = 0$ for all $i = 1, \cdots, r$, which completes the proof. □

The following corollary formalizes the eigentriples of $\mathcal{X}$.

Corollary 1 *The collection of eigentriples,* $\{\sqrt{\lambda_i}, \mathbf{v}_i, \boldsymbol{\psi}_i\}_{i=1}^{r}$, *defines an MFSVD of* $\mathcal{X}$.

3 Generalizing Multivariate Singular Spectrum Analysis to Multivariate Functional Singular Spectrum Analysis

We note that a key step in extending different SSA-based approaches is how to obtain the trajectory matrix (operator) in the embedding step. Despite the fact that in SSA, the trajectory matrix is a linear combination of the associated L-lagged vectors, this is not the case for MSSA.

In Sect. 2, we obtain MFSSA by generalizing FSSA, where we introduce the multivariate trajectory operator as a linear combination of L-lagged multivariate function vectors that correspond to an MFTS. Alternatively, we may mimic the approach of MSSA algorithms (HMSSA or VMSSA) and develop new trajectory operators that are not necessarily based on L-lagged function vectors. The following subsections extend HMSSA and VMSSA to obtain the functional versions, respectively.

From Horizontal Multivariate Singular Spectrum Analysis to Horizontal Multivariate Functional Singular Spectrum Analysis

We may view the column vectors of matrix $\mathbf{X}^{(j)}$ in Eq. (13), $\mathbf{x}_k^{(j)}$'s, as the univariate L-lagged vectors for the j^{th} variable. We see that $\mathbf{X}^{(j)}$ is an operator from $\mathbb{R}^K \to \mathbb{R}^L$, which can be seen as a linear combination of these L-lagged vectors:

$$\mathbf{X}^{(j)}\mathbf{a}^{(j)} = \sum_{k=1}^{K} a_k^{(j)}\mathbf{x}_k^{(j)}, \qquad \mathbf{a}^{(j)} = \left[a_1^{(j)}, \ldots, a_K^{(j)}\right]^\top \in \mathbb{R}^K.$$

In the embedding step of HMSSA, the trajectory matrix, given in Eq. (14), can be seen as a linear operator, $\mathbf{X} : \mathbb{R}^{pK} \to \mathbb{R}^L$, where

$$\mathbf{X}\mathbf{a} = \sum_{j=1}^{p}\sum_{k=1}^{K} a_k^{(j)}\mathbf{x}_k^{(j)}, \qquad \mathbf{a} = \left[\left(\mathbf{a}^{(1)}\right)^\top \cdots \left(\mathbf{a}^{(p)}\right)^\top\right]^\top \in \mathbb{R}^{pK}. \tag{24}$$

In order to make the extension to the function space, we need to assume that the L-lagged function vectors (not multivariate) in HMFSSA, denoted with $\boldsymbol{x}_k^{(j)}$, are in the space $\mathbb{F}_j^L$, for $j = 1, \ldots, p$ where $\mathbb{F}_j^L$ is the Cartesian product of L copies of $\mathbb{F}_j$. But the linear combination of $\boldsymbol{x}_k^{(j)}$'s are well defined if and only if $\mathbb{F}_1^L = \cdots = \mathbb{F}_p^L$, or equivalently $T_1 = \cdots = T_p$. This special type of SSA is known as Horizontal MFSSA (HMFSSA) and given the fact that it is restrictive and not as useful in terms of signal extraction of an MFTS signal (see Sect. 4), we do not present the full algorithm here.

From Vertical Multivariate Singular Spectrum Analysis to Vertical Multivariate Functional Singular Spectrum Analysis

In the embedding step of VMSSA, the trajectory matrix, given in Eq. (15), can be seen as a linear operator, $\mathbf{X} : \mathbb{R}^K \to \mathbb{R}^{pL}$, with

$$\mathbf{X}\mathbf{a} = \sum_{k=1}^{K} a_j \underline{\mathbf{x}}_k, \qquad \mathbf{a} = [a_1, \ldots, a_K]^\top \in \mathbb{R}^K \quad \text{and } \underline{\mathbf{x}}_k = \begin{bmatrix} \mathbf{x}_k^{(1)} \\ \vdots \\ \mathbf{x}_k^{(p)} \end{bmatrix} \in \mathbb{R}^{pL}.$$

To develop VMFSSA, we extend this operator to the function space, i.e., $\underline{\mathbf{x}}_k$ should belong to a new unfolded Hilbert space, $\mathbb{H}^{p,L} = \underbrace{\mathbb{F}_1 \times \cdots \times \mathbb{F}_1}_{L \text{ times}} \times \ldots \times \underbrace{\mathbb{F}_p \times \cdots \times \mathbb{F}_p}_{L \text{ times}}$. Here, each length pL function vector, $\underline{\boldsymbol{x}} \in \mathbb{H}^{p,L}$, is denoted by $\underline{\boldsymbol{x}} = \left(x_1^{(1)}, \ldots, x_L^{(1)}, \ldots, x_1^{(p)}, \ldots, x_L^{(p)}\right)$. It is easy to see that $\mathbb{H}^{p,L}$ is a Hilbert space equipped with inner product

$$\left\langle \underline{\boldsymbol{x}}, \underline{\boldsymbol{y}} \right\rangle_{\mathbb{H}^{p,L}} = \sum_{i=1}^{L} \sum_{j=1}^{p} \left\langle x_i^{(j)}, y_i^{(j)} \right\rangle_{\mathbb{F}_j} = \sum_{i=1}^{L} \langle \vec{x}_i, \vec{y}_i \rangle_{\mathbb{H}} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{H}^L}.$$

Therefore, there exists an isomorphic map, $\boldsymbol{\mathcal{U}} : \mathbb{H}^L \to \mathbb{H}^{p,L}$, defined by $\boldsymbol{\mathcal{U}}(\boldsymbol{x}) = \underline{\boldsymbol{x}}$, where $\boldsymbol{x} \in \mathbb{H}^L$ is a length L multivariate function vector. Now one may define the linear operator $\underline{\boldsymbol{\mathcal{X}}} : \mathbb{R}^K \to \mathbb{H}^{p,L}$, specified with $\underline{\boldsymbol{x}}_k$'s, as

$$\underline{\boldsymbol{\mathcal{X}}}\mathbf{a} = \sum_{k=1}^{K} a_j \underline{\boldsymbol{x}}_k, \qquad \mathbf{a} \in \mathbb{R}^K \text{ and } \underline{\boldsymbol{x}}_k \in \mathbb{H}^{p,L}.$$

The following theorem illustrates the equivalency between the results of the decomposition process applied in MFSSA and VMFSSA.

Theorem 3 *Let* $\{\sqrt{\lambda_i}, \mathbf{v}_i, \boldsymbol{\psi}_i\}$ *be the* i^{th} *eigentriple of* $\boldsymbol{\mathcal{X}}$. *The following holds:*

(i) $\underline{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{U}}\boldsymbol{\mathcal{X}}$.

(ii) $\underline{\boldsymbol{\mathcal{X}}}$ *is a rank* r *operator with the* i^{th} *eigentriple denoted by* $\{\sqrt{\lambda_i}, \mathbf{v}_i, \underline{\boldsymbol{\psi}}_i\}$, *where* $\underline{\boldsymbol{\psi}}_i = \boldsymbol{\mathcal{U}}\boldsymbol{\psi}_i$.

Proof We prove both parts of this theorem in the following:

(i) Let $\mathbf{a} \in \mathbb{R}^K$, then we have that

$$\mathcal{U}\mathcal{X}(\mathbf{a}) = \sum_{k=1}^{K} a_k \mathcal{U}\boldsymbol{x}_k = \sum_{k=1}^{K} a_k \underline{\boldsymbol{x}}_k = \underline{\mathcal{X}}(\mathbf{a})$$

and as such, we have that $\underline{\mathcal{X}} = \mathcal{U}\mathcal{X}$.

(ii) We again, let $\mathbf{a} \in \mathbb{R}^K$, then we have

$$\mathcal{U}\mathcal{X}(\mathbf{a}) = \sum_{i=1}^{r} \sqrt{\lambda_i}\mathbf{a}^\top \mathbf{v}_i \mathcal{U}\boldsymbol{\psi}_i = \sum_{i=1}^{r} \sqrt{\lambda_i}\mathbf{a}^\top \mathbf{v}_i \underline{\boldsymbol{\psi}}_i = \underline{\mathcal{X}}(\mathbf{a}).$$

This implies that the i^{th} eigentriple of $\underline{\mathcal{X}}$ is $\left(\sqrt{\lambda_i}, \mathbf{v}_i, \underline{\boldsymbol{\psi}}_i\right)$ and that $\underline{\boldsymbol{\psi}}_i = \mathcal{U}\boldsymbol{\psi}_i$.

□

Therefore, the decompositions obtained via MFSSA and VMFSSA are interchangeable and subsequently the respective groupings and reconstructions are equivalent.

4 Numerical Studies

In order to explore the capabilities of MFSSA and HMFSSA, we implement a simulation study where we compare our two novel algorithms to other approaches of signal extraction for MFTS data. We also present applications to remote sensing data which is used to further illustrate the interesting qualities of MFTS data that are discovered by MFSSA.

Simulation Study

For the simulation, we generate a bivariate FTS of lengths $N = \{100, 200\}$ by projecting the following discrete observations, sampled over regular intervals in i and t, onto a B-spline basis with 15 basis elements

$$Y_t^{(1)}(s_i) = y_t^{(1)}(s_i) + X_t^{(1)}(s_i)$$
$$Y_t^{(2)}(s_i) = y_t^{(2)}(s_i) + X_t^{(2)}(s_i), \quad s_i \in [0, 1], \quad i = 1, \ldots, 100, \quad t = 1, \ldots, N.$$

We have that $y_t^{(1)} = \mu_t + \delta_t^{(1)}$ and $y_t^{(2)} = \delta_t^{(2)}$ are nonrandom, true signal terms. We take $\mu_t = \kappa t$ as an increasing trend component with $\kappa = \{0.00, 0.02\}$, $\delta_t^{(j)}$ are taken as seasonal components with expressions given as

$$\begin{aligned}\delta_t^{(1)} &= e^{s_i^2}\cos(2\pi\omega_1 t) + e^{1-s_i^2}\cos(2\pi\omega_2 t) - \sin(2\pi\omega_1 t)\cos(4\pi s_i) \\ &\quad + \sin(2\pi\omega_2 t)\sin(\pi s_i) \\ \delta_t^{(2)} &= e^{s_i^2}\sin(2\pi\omega_1 t) + \cos(2\pi\omega_1 t)\cos(4\pi s_i),\end{aligned}$$

where $\omega_1 = \{0.1, 0.5\}$, $\omega_2 = \{0, 0.25\}$, and $X_t^{(j)}$ are error terms for $j = 1, 2$. The error terms follow four models drawn directly from Haghbin et al. [10], one being a Gaussian white noise and the other three coming from a functional autoregressive model of order 1 (FAR1) given by

$$X_t(s) = \Psi X_{t-1}(s) + \epsilon_t(s),$$

where the collection $\{\epsilon_t(s)\}_{t=1}^N$ are taken as independent functions of Brownian motion over the unit interval and Ψ is an integral operator with kernel

$$\psi(s,u) = \gamma_0\left(2 - (2s-1)^2 - (2u-1)^2\right).$$

We choose γ_0 such that the norm of Ψ, given as

$$\|\Psi\|_{\mathcal{S}}^2 = \int_0^1\int_0^1 |\psi(s,u)|^2 ds du,$$

takes on values of 0, 0.5, or 0.9 in order to obtain our autoregressive models. Due to the presence of a trend component and two frequencies, we require five components in all algorithms to reconstruct the true structures which is due to the fact that each of the two frequencies is expressed in a sine and a cosine term. We compare reconstruction results of MFSSA to the results of HMFSSA where each algorithm is ran on the simulated bivariate FTS. We also compare the MFSSA and HMFSSA results to that of FSSA, MSSA, and DFPCA ran on each simulated covariate independently of one another. For MSSA we specify that the $200 \times N$ data matrix, $\mathbf{Q}$, follows the form

$$\mathbf{Q} = \begin{bmatrix}\mathbf{Q}_1 \\ \mathbf{Q}_2\end{bmatrix},$$

such that each $100 \times N$ matrix, $\mathbf{Q}_j$, follows the form $\mathbf{Q}_j = \left[Y_t^{(j)}(s_i)\right]_{i=1,\ldots,100}^{t=1,\ldots,N}$ for $j = 1, 2$.

For all of the SSA-based algorithms we set $L = \{20, 40\}$ and for all algorithms, we measure the error of each reconstruction with the following root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N \times n \times p}\sum_{j=1}^{p}\sum_{t=1}^{N}\sum_{i=1}^{n}\left(y_t^{(j)}(s_i) - \hat{y}_t^{(j)}(s_i)\right)^2},$$

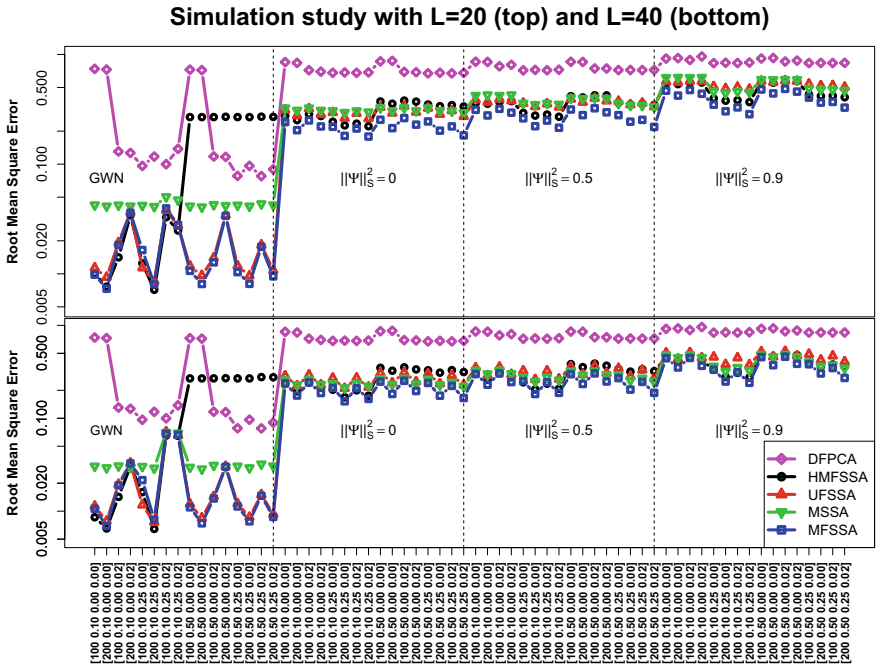


Fig. 1 Simulation study where the vertical axes give outputs of average root mean square error for 100 replications of each simulation setup and each entry on the horizontal axis follows the form $[N, \omega_1, \omega_2, \kappa]$. The top plot is when $L = 20$ and the bottom plot is when $L = 40$

where $\hat{y}_t^{(j)}(s_i)$ is the reconstruction of covariate j, at time point t, evaluated at point s_i. For every unique combination of parameters and error terms, we repeat the simulation 100 times and report the mean of the RMSE's in Fig. 1 whose vertical axes are taken over a log scale.

We see in the top plot that $L = 20$ and in the bottom plot, $L = 40$, while the vertical lines separate out the simulated data by noise models and in addition, each tick mark on the horizontal should be read as $[N, \omega_1, \omega_2, \kappa]$. From these two subfigures, we find that MFSSA almost always outperforms other techniques of signal extraction of an MFTS while HMFSSA also outperforms other techniques occasionally.

Application to NDVI Images and Intraday Temperature Data

Most researchers believe that the amount of vegetation present within a region is correlated to the region's temperature. Researchers can use this correlation to get a better understanding of how the vegetation and temperature in a region changes over time. Data that tracks the intraday hourly mean temperature, in Celsius, for a variety

of United States weather stations is available for download from Diamond et al. [3]. In addition, satellite images of varying resolutions, regions, time periods, spectral bands, and their variants have been made available for download and analyzed using various techniques [29]. The NDVI measure, recorded in the satellite images, is bounded between zero and one and is used to track changes in vegetation over time [18]. It was determined that using MSSA over SSA can lead to richer analysis of correlated data [4]. If a variable with strong seasonality components and another variable with strong mean components are included together in an MSSA analysis, we expect to find strong seasonality and mean component reflected in the exploratory plots. Similarly, we expect MFSSA applied to correlated functional variables will strengthen signal extraction results and now we give a real data study illustrating the improvement.

We introduce a bivariate example of intraday hourly mean temperature curves and NDVI images of a parallelogram shaped region just east of Glacier National Park in Montana, U.S.A. located between longitudes of 113.30°W − 113.56°W and latitudes of 48.71°N − 48.78°N. The curves and images were recorded every 16 days starting January 1, 2008 and ending September 30, 2013. We start by applying FSSA with a lag of 45 (to capture yearly periodic behavior in temperature and vegetative processes) to the function curves and smoothed images separately from one another and we obtain the following plots of the singular vectors.

It is clear from Fig. 2b that there exists a strong seasonality component in the intraday temperature curves of (a) accounting for 54.72% of the variation in the data while a mean behavior component accounts for 15.38% of the variation in the data. We also see from Fig. 2d that the mean component captures 65.07% of the variation of the NDVI image data while the seasonality components only account for 28.56% of the variation of the data. We normalize the intraday temperature curves by dividing each sampling point by the standard deviation of all the sampling points (dividing by 10.13) since the NDVI images have values that are significantly smaller. We now apply MFSSA with a lag of 45 to the normalized intraday temperature curves and NDVI images in a bivariate analysis to obtain the following plots.

Analyzing the plot of singular values and the w-correlation blocks in Fig. 3a and b suggests that component one should be grouped by itself, components two and three should be grouped together, and four and five should be grouped together. Figure 3c shows that in the bivariate analysis, the mean component becomes dominant with the seasonal components taking on the second and third main sources of variation. This shows that combining the temperature curves and NDVI images functional data into a bivariate analysis reveals a stronger mean component as opposed to the weaker mean component seen in Fig. 2b.

Application to Remote Sensing Density Curves

To further show that MFSSA enriches data analysis of correlated variables, we use another bivariate example of near-infrared (NIR) and shortwave infrared (SWIR)

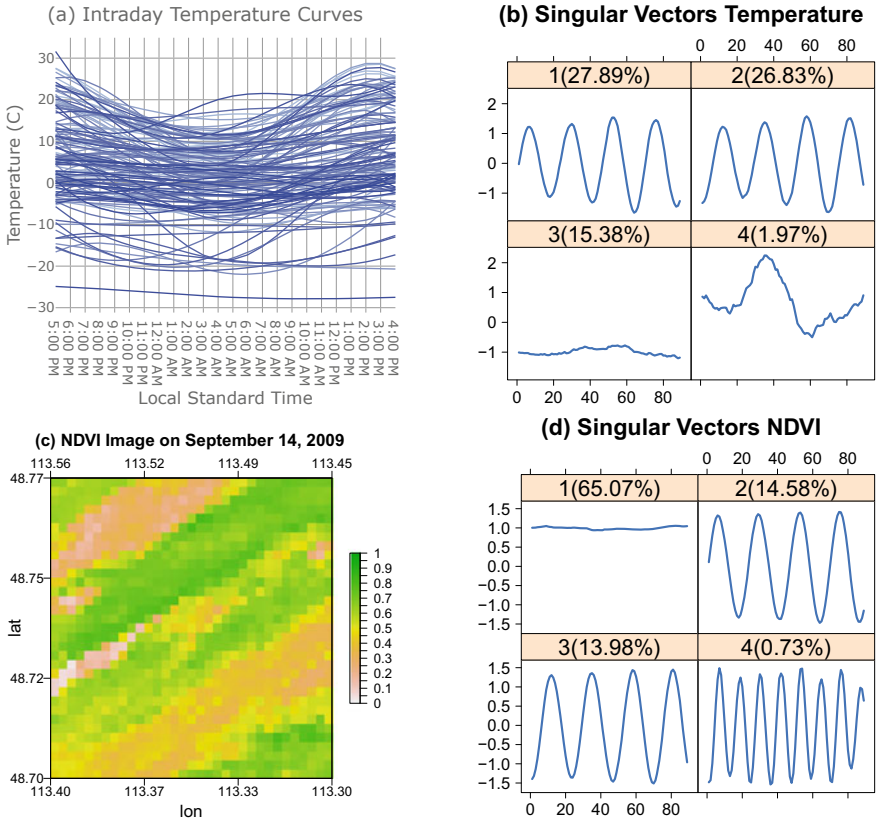


Fig. 2 **a** Intraday temperature curves recorded in Saint Mary, Montana, U.S.A.; **b** Right singular vectors from FSSA analysis applied to intraday temperature curves; **c** An NDVI image of Saint Mary, Montana, U.S.A.; **d** Right singular vectors from FSSA analysis applied to NDVI images

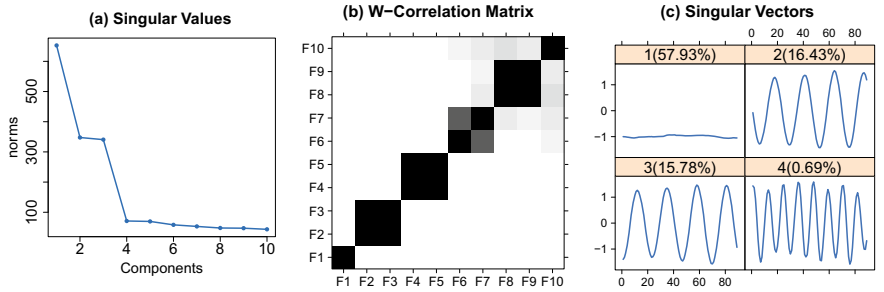


Fig. 3 **a** Plot of singular values from MFSSA applied to intraday temperature curves and NDVI images from Saint Mary, Montana, U.S.A.; **b** Plot of the w-correlation matrix generated from MFSSA; **c** Plot of right singular vectors from the MFSSA analysis

images taken every 8 days of a region just outside of the city of Jambi, Indonesia between 103.61°E − 103.68°E and 1.67°S − 1.60°S over the timeline of February 18, 2000 and November 25, 2019. The wavelength of the NIR images ranges from 841 to 876 nanometers (nm) and the wavelength of the SWIR images are within the values of 2105–2155 nm. NIR light can be used for imaging vegetation and it is used in the calculation of the NDVI measure [18] while shortwave infrared is often used for imaging the moisture content in soil where a lower surface reflectance (SR) corresponds to higher moisture content [26]. As mentioned in Prasetyo et al. [22], it appears that this particular part of the Jambi province was a hot spot for controlled fires between 2001 and 2015 and the loss of vegetation over the course of about a decade will be reflected in lower NIR and higher SWIR SR values as time moves on. In this data study, we consider a new type of MFTS where both variables share the same domain and take on similar values in the output of each function. To this end, we obtain the kernel density estimates (KDEs) corresponding to both of the NIR and the SWIR SR values in the images, using Silverman's rule of thumb [28]. In addition, we replaced outliers in the SWIR densities with the average of densities from the preceding and proceeding days. Similar results, as compared to the following, still

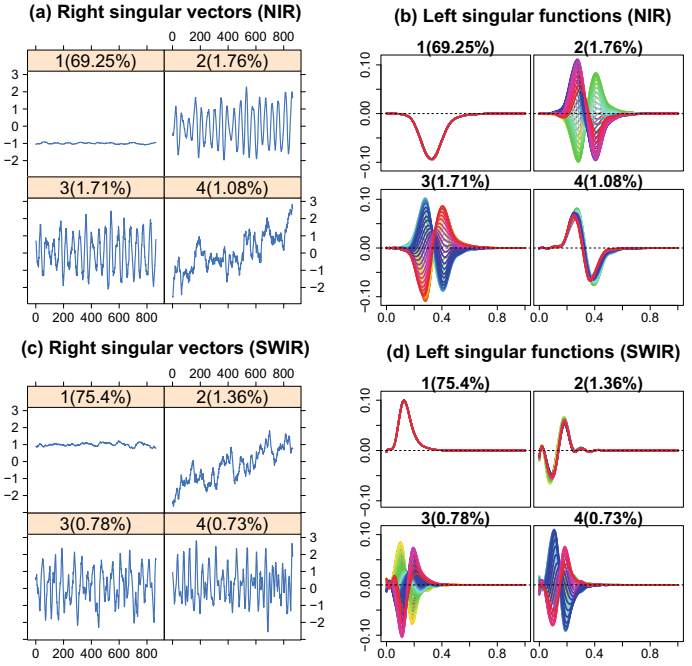


Fig. 4 **a** The right singular vectors that are obtained when applying FSSA with a lag of $L = 45$ to the NIR densities data; **b** The left singular functions corresponding to the NIR densities data; **c** The right singular vectors that are obtained when applying FSSA with a lag of $L = 45$ to the SWIR densities data; **d** The left singular functions corresponding to the SWIR densities data

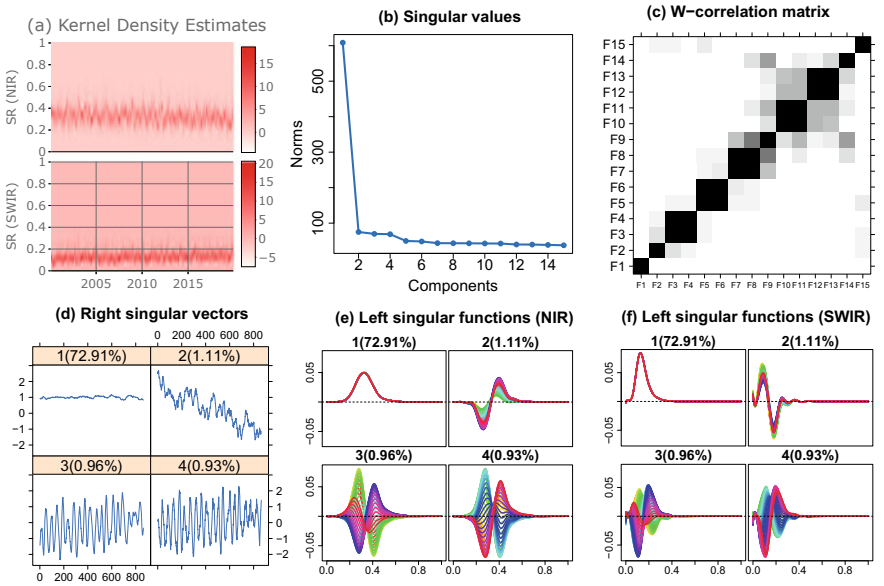


Fig. 5 **a** Heatplots of the NIR and SWIR densities bivariate FTS; **b** The plot of the leading singular values that is generated from MFSSA applied with a lag of 45 to the bivariate FTS shown in (**a**); **c** The plot of the w-correlation matrix; **d** The plot of the first four right singular vectors; **e** The plot of the first four left singular functions that correspond with NIR densities; **f** The plot of the first four left singular functions that correspond with SWIR densities

hold even if the outliers are not removed. Applying FSSA with a lag of 45 to the NIR and SWIR densities separately, where this choice of lag approximately captures yearly periodic behavior, gives the following exploratory plots.

Figure 4a and b gives us the right singular vectors and left singular functions of the NIR densities while Fig. 4c and d is the right singular vectors and left singular functions of the SWIR densities. We find that applying FSSA to the NIR densities captures seasonality in the second and third components while trend is present in the fourth component similar to the FSSA, NDVI real data study of Haghbin et al. [10]. Applying FSSA to the SWIR densities shows that trend is a more dominant behavior captured in the second component as compared to the seasonal behaviors captured in components three and four. Applying MFSSA decomposition with a lag of 45 to the bivariate NIR/SWIR example gives the following exploratory plots.

The bivariate FTS can be found in Fig. 5a while b and c are plots of singular values and w-correlation, respectively. See that Fig. 5d gives us our MFSSA right singular vectors which showcases the weights that are multiplied by the left singular functions shown in Fig. 5e and f. Since we are performing MFSSA, we obtain 45 functions per left singular function that corresponds to the NIR densities as well as another set of 45 functions per left singular function that corresponds to the SWIR densities. Notice the trend behavior for the NIR densities is present in component two

as according to Fig. 5e which indicates that adding SWIR densities into the analysis with the NIR densities created a more pronounced trend result as compared with Fig. 4b. To this end, we find that performing a bivariate analysis on the NIR/SWIR densities enriched our data analysis as expected.

5 Discussion

Throughout this chapter, we presented MFSSA as a novel technique of decomposition of an MFTS. Although VMSSA and HMSSA can be extended to the respective functional versions (HMFSSA and VMFSSA), we found that HMFSSA is a restrictive and weaker type of signal extraction of an MFTS signal and that VMFSSA solves the same problem as MFSSA. That is why we proposed to use the isomorphic property of two Hilbert spaces in order to extend FSSA and develop its multivariate version. Using remote sensing data examples we showed that the proposed MFSSA algorithm is able to handle functions taken over different dimensional domains and is able to uncover more dominant modes of variation present in MFTS signals where these same types of variability might not be extracted from univariate approaches. We also developed an R package, called **Rfssa** (available on CRAN), that implements the MFSSA algorithm and it also includes the NDVI dataset [11]. Furthermore, a shiny web application is also included in the package, and it is available at http://sctc.mscs.mu.edu/mfssa.htm for reproducing the results of this chapter or analyzing any other MFTS.

References

1. Alexandrov, T. (2009). A method of trend extraction using singular spectrum analysis. *RevStat, 7*(1), 1–22.
2. Chiou, J.-M., Chen, Y.-T., & Yang, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica, 24*(4), 1571–1596.
3. Diamond, H. J., Karl, T., Palecki, M. A., Baker, C. B., Bell, J. E., Leeper, R. D., Easterling, D. R., Lawrimore, J. H., Meyers, T. P., Helfert, M. R., Goodge, G., & Thorne, P. W. (2013). U.S. climate reference network after one decade of operations: Status and assessment. Retrieved April 2020, from https://www.ncdc.noaa.gov/crn/qcdatasets.html.
4. Golyandina, N., Korobeynikov, A., Shlemov, A., & Usevich, K. (2015). Multivariate and 2-D extensions of singular spectrum analysis with the Rssa package. *Journal of Statistical Software, 67*(2), 1–78.
5. Golyandina, N., Korobeynikov, A., & Zhigljavsky, A. (2018). *Singular spectrum analysis with R*. Berlin: Springer.
6. Golyandina, N., Nekrutkin, V., & Zhigljavsky, A. A. (2001). *Analysis of time series structure: SSA and related techniques*. New York: Chapman and Hall/CRC.
7. Golyandina, N., & Osipov, E. (2007). The "Caterpillar"-SSA method for analysis of time series with missing values. *Journal of Statistical Planning and Inference, 137*(8), 2642–2653.
8. Golyandina, N., & Zhigljavsky, A. (2013). *Singular spectrum analysis for time series*. Berlin: Springer.

9. Groth, A., & Ghil, M. (2011). Multivariate singular spectrum analysis and the road to phase synchronization. *Physical Review E: Statistical, Nonlinear & Soft Matter Physics, 84*(3–2), 036206–1–036206–10.
10. Haghbin, H., Morteza Najibi, S., Mahmoudvand, R., Trinka, J., & Maadooliat, M. (2021). Functional singular spectrum analysis. *Stat,* e330 STAT-20-0240.R1.
11. Haghbin, H., Najibi, S. M., Trinka, J., & Maadooliat, M. (2019). *Rfssa: Functional singular spectrum analysis*. R package version 1.0.0.
12. Happ, C., & Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association, 113*(522), 649–659.
13. Hassani, H., & Mahmoudvand, R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics, 1*(1), 55–83.
14. Hassani, H., & Mahmoudvand, R. (2018). *Singular spectrum analysis: Using R*. London: Palgrave Macmillan.
15. Hassani, H., Rua, A., Silva, E. S., & Thomakos, D. (2019). Monthly forecasting of GDP with mixed-frequency multivariate singular spectrum analysis. *International Journal of Forecasting, 35*(4), 1263–1272.
16. Hörmann, S., Kidziński, Ł, & Hallin, M. (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77*(2), 319–348.
17. Kondrashov, D., Shprits, Y., & Ghil, M. (2010). Gap filling of solar wind data by singular spectrum analysis. *Geophysical Research Letters, 37*(15).
18. Lambin, E. F. (1999). Monitoring forest degradation in tropical regions by remote sensing: Some methodological issues. *Global Ecology and Biogeography, 8*(3–4), 191–198.
19. Mahmoudvand, R., & Rodrigues, P. C. (2016). Missing value imputation in time series using singular spectrum analysis. *International Journal of Energy and Statistics, 4*(1), 1650005.
20. Mohammad, Y., & Nishida, T. (2011). On comparing SSA-based change point discovery algorithms. In *2011 IEEE/SICE International Symposium on System Integration (SII)* (pp. 938–945). IEEE.
21. Moskvina, V., & Zhigljavsky, A. (2003). An algorithm based on singular spectrum analysis for change-point detection. *Communications in Statistics-Simulation and Computation, 32*(2), 319–352.
22. Prasetyo, L. B., Dharmawan, A. H., Nasdian, F. T., & Ramdhoni, S. (2016). Historical forest fire occurrence analysis in Jambi Province during the period of 2000–2015: Its distribution & land cover trajectories. *Procedia Environmental Sciences, 33*, 450–459.
23. Ramsay, J. O. & Silverman, B. W. (2005). *Functional data analysis*. Springer series in statistics. New York: Springer.
24. Rodrigues, P. C., & Mahmoudvand, R. (2016). Correlation analysis in contaminated data by singular spectrum analysis. *Quality and Reliability Engineering International, 32*(6), 2127–2137.
25. Shalit, O. M. (2017). *A first course in functional analysis*. Boca Raton: Chapman and Hall/CRC.
26. Shin, H., Yu, J., Jeong, Y., Wang, L., & Yang, D. (2017). Case-based regression models defining the relationships between moisture content and shortwave infrared reflectance of beach sands. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(10), 4512–4521.
27. Silva, E. S., Hassani, H., & Heravi, S. (2018). Modeling european industrial production with multivariate singular spectrum analysis: A cross-industry analysis. *Journal of Forecasting, 37*(3), 371–384.
28. Silverman, B. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.
29. Tuck, S. L., Phillips, H. R., Hintzen, R. E., Scharlemann, J. P., Purvis, A., & Hudson, L. N. (2014). MODISTools - downloading and processing MODIS remotely sensed data in R. *Ecology and Evolution, 4*(24), 4658–4668.
30. Weidmann, J. (1980). *Linear operators in Hilbert spaces*. Graduate texts in mathematics. New York: Springer.

Compositional Data Analysis—Linear Algebra, Visualization and Interpretation

Michael Greenacre

Abstract Compositional data analysis is concerned with multivariate data that have a constant sum, usually 1 or 100%. These are data often found not only in biochemistry and geochemistry, but also in the social sciences, when relative values are of interest rather than the raw values. Recent applications are in the area of very high-dimensional "omics" data. Logratios are frequently used for this type of data, i.e. the logarithms of ratios of the components of the data vectors. These ratios raise interesting issues in matrix-vector representation, computation and interpretation, which will be dealt with in this chapter.

1 Introduction

Consider the table in Fig. 1a: these are amounts spent in 2019 on four different budget items in the European Union (EU), showing only six EU countries with the addition of Iceland, Norway and Switzerland. The full data set consists of 30 countries. The amounts are in millions of local currency and are clearly not comparable across the countries. Converting all the amounts to the same currency, for example euros, alleviates but does not solve the data coding problem, since some countries are small and others are large. It is thus the relative amounts spent on the different items that are of interest, and so it seems that comparability across countries is assured by simply expressing the four amounts in each row as percentages, as in Fig. 1b. The rows of data in Fig. 1b are called *compositions*: nonnegative multivariate data with the constant sum constraint, 100% in this case. The act of transforming the original monetary values to proportions, or percentages, is called in this context *normalization* or *closure*.

But is the coding problem really solved? If additional budget items were added to the table in Fig. 1a, such as public services, social protection and culture, the relative amounts in Fig. 1b would change, necessarily decreasing, and they would

M. Greenacre (✉)
Universitat Pompeu Fabra, Barcelona, Spain
e-mail: michael.greenacre@upf.edu

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_10

(a)

	Defence	Economy	Health	Education	Totals
Belgium	3765.800	31783.500	36132.600	29323.500	101005.400
Bulgaria	1466.795	7751.240	5970.827	4646.598	19835.460
Czechia	50695.000	352358.000	437283.000	281190.000	1121526.000
Denmark	26212.000	71315.000	192406.000	146505.000	436438.000
Germany	36745.000	114879.000	254006.000	149201.000	554831.000
⋮	⋮	⋮	⋮	⋮	⋮
Sweden	61811.000	221853.000	351586.000	348457.000	983707.000
Iceland	2220.328	148780.390	237157.803	213419.858	601578.378
Norway	67114.000	214366.000	310060.000	199598.000	791138.000
Switzerland	5927.060	28174.984	15391.838	39531.782	89025.662

(b)

	Defence	Economy	Health	Education
Belgium	3.7%	31.5%	35.8%	29.0%
Bulgaria	7.4%	39.1%	30.1%	23.4%
Czechia	4.5%	31.4%	39.0%	25.1%
Denmark	6.0%	16.3%	44.1%	33.6%
Germany	6.6%	20.7%	45.8%	26.9%
⋮	⋮	⋮	⋮	⋮
Sweden	6.3%	22.6%	35.7%	35.4%
Iceland	0.4%	24.7%	39.4%	35.5%
Norway	8.5%	27.1%	39.2%	25.2%
Switzerland	6.7%	31.6%	17.3%	44.4%

Fig. 1 Expenditure in European Union countries plus Iceland, Norway and Switzerland on four budget items, in 2019. In **a**, the amounts are in millions of local currency. In **b**, the amounts are expressed as percentages relative to the totals

reduce by different proportions since the additional budget amounts would not be in proportion to the totals of the four budget items shown in Fig. 1a. This is what makes compositional data different from any other multivariate data in Statistics—the values of each component in the table depend on the values of the other components. It would make no sense, for example, to compute correlations on such a data matrix, since there are necessarily many negative correlations created by the constant sum constraint, and the correlation between health and education would be different in Fig. 1b from the correlation between health and education in an expanded table of budget items also expressed as compositions.

Aitchison [1, 2] showed that using ratios of the components was a solution to the data coding issue. Ratios remain constant between two components irrespective of adding components to or removing components from a composition—they are said to be *subcompositionally coherent*. Furthermore, Aitchison proposed that ratios be logarithmically transformed: for example, if X_j and X_k are two components then the *logratio* transformation is $\log(X_j/X_k) = \log(X_j) - \log(X_k)$, i.e. the difference in their logarithms, which is a linear transformation on the log scale.

The objective of this chapter is to express the theory of logratio transformations in linear algebra terms, and in the process give a flavor of the visualization possibilities of logratios and their interpretation. In Sect. 2, basic definitions and results are given in the matrix-vector form, as well as inverse logratio transformations, including the important topic of log-contrasts. Section 3 shows how logratios can be visualized through cluster analysis and biplots, and Sect. 4 concludes with a discussion.

2 Basic Algebraic Definitions and Results

The practical aspects of compositional data analysis are given in two recent books, by Greenacre [12] and Filzmoser et al. [8], and a comprehensive review is given by Greenacre [15]. In this section, the algebra of compositional data analysis is given in matrix-vector notation.

Logratio Transformations and Associated Pattern Matrices

Suppose $\mathbf{x}\,(J \times 1)$ is a J-component compositional vector of positive values with sum 1: i.e. $\mathbf{1}^{\mathsf{T}}\mathbf{x} = 1$, where $\mathbf{1}$ is a vector of ones of appropriate order, in this case a vector of J ones. If $\log(\mathbf{x})$ denotes the vector of log-transformed values, then almost all logratio transformations are defined by a linear transformation of the form $\mathbf{P}\log(\mathbf{x})$, where $\mathbf{P}$ is called the *logratio pattern matrix*. If $\mathbf{X}\,(I \times J)$ denotes the data set, with sampling units as rows and components as columns, then the compositions are in the rows and the constant row sums are defined by post-multiplication by $\mathbf{1}$: $\mathbf{X1} = \mathbf{1}$, where the $\mathbf{1}$ on the right is $J \times 1$. Similarly, if $\mathbf{L} = \log(\mathbf{X})$ denotes the $I \times J$ matrix of logarithms of $\mathbf{X}$, then the application of the pattern matrix to the rows of $\mathbf{L}$ implies post-multiplication on the right by the transpose of the pattern matrix: $\mathbf{LP}^{\mathsf{T}}$. In the following, the matrix $\mathbf{P}$ will be subscripted by the type of logratio transformation.

The pairwise logratio pattern matrix $\mathbf{P}_{\text{LR}}$ $(\frac{1}{2}J(J-1) \times J)$ is defined as

$$\mathbf{P}_{\text{LR}} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \cdots 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \cdots 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \cdots 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \ddots \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \cdots 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 \cdots 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \cdots 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 \cdots 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \ddots \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 \cdots 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \cdots 0 & 1 & -1 \end{bmatrix} \tag{1}$$

so that the $I \times \frac{1}{2}J(J-1)$ matrix of pairwise logratios is $\mathbf{LP}_{\text{LR}}^{\mathsf{T}}$. Each row of $\mathbf{P}_{\text{LR}}$ defines a pairwise logratio (LR), when applied to the logarithm of a composition, $\log(\mathbf{x})$. For example, the first row of $\mathbf{P}_{\text{LR}}$ would engender the logratio $\log(x_1) - \log(x_2)$ corresponding to the pair $(1, 2)$. Notice the lexicographic ordering in the rows, corresponding to the ratio pairs (12), (13), (14), $\cdots$, ($1J$), (23), (24), (25), $\cdots$, (J–2, J), (J–1, J). We will abbreviate the term "pairwise logratio" by LR throughout the rest of this chapter.

The matrix of LRs is of rank $J-1$, assuming $I \geq J$, otherwise it is of rank $I-1$ (we shall assume for ease of description that there are at least as many rows as columns in the compositional data matrix). This can be seen easily using a result by Greenacre [12, 13] that a connected directed acyclic graph (DAG) consisting of $J-1$ LRs generates all the $\frac{1}{2}J(J-1)$ pairwise logratios through linear combinations. For example, for the four-component example of Fig. 1a with six LRs, a possible connected DAG is shown by the three solid arrows in Fig. 2, representing the ratios Health/Economy, Health/Education and Defence/Education (the arrow points toward the numerator component). The other three logratios, indicated by dashed arrows,

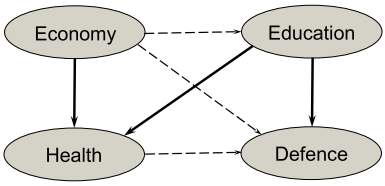


Fig. 2 Directed acyclic graph (DAG) that connects the four components of Fig. 1a, indicated by the solid arrows. The dashed arrows indicate the other three pairwise logratios that can be obtained from those of the DAG

can be obtained from those in the DAG as ratios combined either through multiplication or division depending on the direction of the arrows: Education/Economy = (Health/Economy)/(Health/Education)), or in linearized logratio form using addition or subtraction: log(Education) – log(Economy = [log(Health)–log(Economy)] – [log(Health)–log(Education)].

Assuming the four components above are in the order {Defence, Economy, Health, Education} as in the table in Fig. 1, the pattern matrix associated with the three logratios (solid arrows) in the DAG above has this form:

$$\mathbf{P} = \begin{bmatrix} 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & -1 \end{bmatrix} \tag{2}$$

A special case of the LRs is the *additive logratio* (ALR) transformation, the set of $J - 1$ LRs where the denominator component is the same, called the *reference* component. For example, if the last component is the reference, then the ALR pattern matrix $\mathbf{P}_{\text{ALR}}$ ($(J - 1) \times J$) is

$$\mathbf{P}_{\text{ALR}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \tag{3}$$

The matrix of ALRs, $\mathbf{L}\mathbf{P}_{\text{ALR}}^{\mathsf{T}}$ ($I \times (J - 1)$), is of rank $J - 1$. For example, if for the four components in Fig. 1 the third component Health was chosen as the reference part, then (3) would be a 3×4 matrix with the -1s down the third column and a 1 in each row in columns 1, 2 and 4. Figure 3 shows the DAG associated with this ALR transformation, where Health is placed in the center and arrows emanate outwards to the other three components.

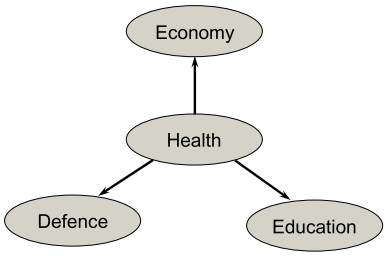


Fig. 3 DAG corresponding to the ALR transformation of the four components in Fig. 1, where Health is the reference part

The next most important logratio transformation is the *centered logratio* (CLR) transformation, the ratio of each component divided by the geometric mean of all the components. The usual unweighted definition is the following, for a row $[x_1, x_2, \ldots, x_J]$ of $\mathbf{X}$:

$$\mathrm{CLR}(j) = \log\left(\frac{x_j}{\left(\prod_k x_k\right)^{1/J}}\right) = \log(x_j) - \frac{1}{J}\sum_k \log(x_k) \quad j = 1, \ldots, J \tag{4}$$

but it is preferred here to give a more general weighted definition assuming positive weights c_j $(j = 1, \ldots, J)$ for the components, where $\sum_j c_j = 1$, and thus a weighted geometric mean in the denominator:

$$\mathrm{CLR}(j) = \log\left(\frac{x_j}{\prod_k x_k^{c_k}}\right) = \log(x_j) - \sum_k c_k \log(x_k) \quad j = 1, \ldots, J \tag{5}$$

Hence, (4) is the special case of (5) with equal weights $1/J$ for all the J components.

The CLR pattern matrix $\mathbf{P}_{\mathrm{CLR}}$ $(J \times J)$ for the general case is

$$\mathbf{P}_{\mathrm{CLR}} = \begin{bmatrix} 1-c_1 & -c_2 & -c_3 & \cdots & -c_J \\ -c_1 & 1-c_2 & -c_3 & \cdots & -c_J \\ -c_1 & -c_2 & 1-c_3 & \cdots & -c_J \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -c_1 & -c_2 & -c_3 & \cdots & 1-c_J \end{bmatrix} \tag{6}$$

Notice that $\mathbf{P}_{\mathrm{CLR}}$ is just the idempotent centering matrix $\mathbf{I} - \mathbf{1c}^{\mathsf{T}}$, where $\mathbf{c}^{\mathsf{T}} = [c_1 \; c_2 \cdots c_J]$. The rows of $\mathbf{P}_{\mathrm{CLR}}$ sum to 0, and the matrix of CLRs, $\mathbf{LP}_{\mathrm{CLR}}^{\mathsf{T}}$, has rank $J - 1$, just like the matrix of LRs, $\mathbf{LP}_{\mathrm{LR}}$, and the matrix of ALRs, $\mathbf{LP}_{\mathrm{ALR}}$.

More complex logratio transformations are the *isometric logratios* (ILRs) [7] and their slightly simpler special case, the *pivot logratios* (PLRs) [8]—these are often called "balances" although the term can be misleading [14, 17]. ILRs are log-transformed ratios of geometric means of groups of components. In PLRs, one of these groups is a single component and a linear independent set of $J - 1$ PLRs takes the components in a fixed order and the numerator of each ratio is a single component and the denominator is the geometric mean of the others "to the right of" the

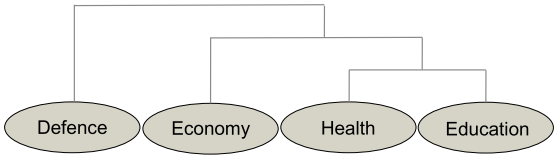


Fig. 4 Dendrogram graph associated with three PLRs corresponding to the order of the components Defence, Economy, Health, and Education

numerator component. For both ILRs and PLRs, there is a scalar constant involved, which is omitted here for simplicity—see Greenacre [12] for the exact (unweighted) definition. Again, it is preferred to give the more general weighted definition here, which will be useful in practice when parts are considered differentially weighted. The pattern matrix $\mathbf{P}_{\text{PLR}}$ ($(J-1)\times J$) for the weighted case is

$$\mathbf{P}_{\text{PLR}} = \begin{bmatrix} 1 & -\frac{c_2}{c_2+\cdots+c_J} & -\frac{c_3}{c_2+\cdots+c_J} & \cdots & -\frac{c_{J-2}}{c_2+\cdots+c_J} & -\frac{c_{J-1}}{c_2+\cdots+c_J} & -\frac{c_J}{c_2+\cdots+c_J} \\ 0 & 1 & -\frac{c_3}{c_3+\cdots+c_J} & \cdots & -\frac{c_{J-2}}{c_3+\cdots+c_J} & -\frac{c_{J-1}}{c_3+\cdots+c_J} & -\frac{c_J}{c_3+\cdots+c_J} \\ 0 & 0 & 1 & \cdots & -\frac{c_{J-2}}{c_4+\cdots+c_J} & -\frac{c_{J-1}}{c_4+\cdots+c_J} & -\frac{c_J}{c_4+\cdots+c_J} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\frac{c_{J-1}}{c_{J-1}+c_J} & -\frac{c_J}{c_{J-1}+c_J} \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \quad (7)$$

Like the pattern matrices before, (1), (3) and (6), the row sums of $\mathbf{P}_{\text{PLR}}$ are zero and the matrix $\mathbf{L}\mathbf{P}_{\text{PLR}}^{\mathsf{T}}$ of PLRs has rank $J-1$. There are $J!$ permutations of the J components, hence $J!$ sets of PLRs possible, depending on the ordering of the J components. For the four components of Fig. 1, in the order given, a graph of the associated PLRs is in the form of the binary dendrogram of Fig. 4. Since there are $4! = 24$ ordered permutations possible for this example, there are 24 different sets of PLRs possible. Notice that the last member of a set of PLRs is a simple LR.

The ILRs involving more general ratios of geometric means are usually defined according to a dendrogram graph, and the number of possible dendrograms increases even more rapidly with the number of components. The ILR pattern matrix becomes more difficult to express in general, so we will just give the special case associated with the dendrogram in Fig. 5, assuming the order shown of the components.
The specific pattern 3×4 matrix for this four-component example corresponding to Fig. 5, again with rows summing to 1, would thus be

$$\mathbf{P}_{\text{ILR}} = \begin{bmatrix} \frac{c_1}{c_1+c_2} & \frac{c_2}{c_1+c_2} & -\frac{c_3}{c_3+c_4} & -\frac{c_4}{c_3+c_4} \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \quad (8)$$

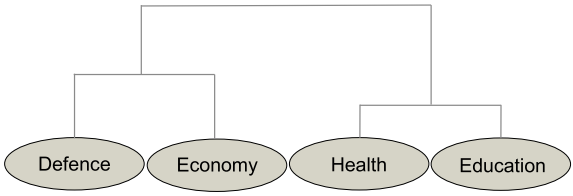


Fig. 5 Dendrogram graph associated with three ILRs corresponding to the three ratios (moving down the tree): (1) geometric mean of Defence and Economy/geometric mean of Health and Education, (2) Defence/Economy and (3) Health/Education

Finally, there is a class of nonlinear transformations called *amalgamation* (or *summated*) *logratios*, abbreviated as SLRs. Like the LRs, these are true balances between the components as they use sums rather than geometric means when combining components in the ratios [14]. For example, the dendrogram in Fig. 5 would translate to the following three SLRs, the last two of which are regular LRs:

$$(1)\ \log\left(\frac{\text{Defence+Economy}}{\text{Health+Education}}\right) \quad (2)\ \log\left(\frac{\text{Defence}}{\text{Economy}}\right) \quad (3)\ \log\left(\frac{\text{Health}}{\text{Education}}\right) \tag{9}$$

The first SLR above cannot be written as a linear function of the logarithms, but has the advantage of being more easily interpreted in a practical application. It is still isomorphic, however, as shown in the next subsection.

Inverting Logratio Transformations

All the logratio transformations defined in the previous subsection are isomorphic; that is, there is a one-to-one mapping of the J-part compositions to the corresponding $J-1$ logratios. And in each case, the logratios can be inverted back to the original compositions (i.e. back-transformed), including the nonlinear SLRs. Each transformation in turn relies on a square *inversion pattern matrix*, denoted by $\mathbf{Q}$, closely related to the respective logratio pattern matrix $\mathbf{P}$. The back-transformation in each case involves solving a system of J linear equations of the form $\mathbf{Qx} = \mathbf{e}$, i.e. $\mathbf{x} = \mathbf{Q}^{-1}\mathbf{e}$, where $\mathbf{Q}$ involves elements that are functions of the logratios, and $\mathbf{e}$ is the vector $[\,0\ \ 0\ \cdots\ 0\ \ 1\,]^{\mathsf{T}}$.

Starting with the ALRs, suppose that $\mathbf{y} = \mathbf{P}_{\text{ALR}}\mathbf{x}$ is the vector of $J-1$ ALRs, where $\mathbf{x}$ is a $J \times 1$ composition and $\mathbf{P}_{\text{ALR}}$ is given by (3). Then the inverse operation of finding $\mathbf{x}$ from $\mathbf{y}$ is the solution of the following equation:

$$\mathbf{Q}_{\text{ALR}}\mathbf{x} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -e^{y_1} \\ 0 & 1 & \cdots & 0 & -e^{y_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -e^{y_{J-1}} \\ 1 & 1 & \cdots & 1 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{10}$$

Notice that the -1s in the pattern matrix, i.e. in the denominator positions, are substituted by $-e^{y_1}, -e^{y_2}, \ldots, -e^{y_{J-1}}$ in rows $1, 2, \ldots, J-1$ respectively, and then a row of 1s is added, which with the last 1 in the right-hand vector imposes the constraint that the x_js in the solution sum to 1 [14]. This is equivalent to the simpler calculation of exponentiating the $J-1$ logratios and expanding them with a 1, i.e. $[\, e^{y_1}\ e^{y_2} \cdots\ e^{y_{J-1}}\ \ 1\,]$ and then normalizing the result to sum to 1. So, although using the linear equations approach by solving (10) is actually an inefficient way of back-transforming the compositions, it is enlightening because it provides a way to deal with any set of $J-1$ independent LRs, e.g. the three LRs defined in the DAG of Fig. 2.

The way of inverting a set of $J-1$ independent pairwise logratios (LRs) is a simple generalization of the system of equations (10) above. The $(J-1) \times J$ pattern matrix $\mathbf{P}_{\text{LR}}$ for the transformation to the LRs now has each of the $J-1$ rows corresponding to a specific LR, $\log(x_j/x_{j'})$, with a 1 in column j of the numerator part and -1 in column j' of the denominator part. To obtain the inversion pattern matrix, the -1 is again replaced in row k by $-e^{y_k}$ and then a row of 1s is added as the last row, as before. For example, for the DAG in Fig. 2, the logratio pattern matrix and corresponding inversion pattern matrix, with their matrix equations are, in the order Defence, Economy, Health, Education, and LRs Health/Economy, Health/Education and Defence/Education, for $\mathbf{x}$ (4×1) a compositional vector, $\mathbf{y}$ (3×1) the logratio transformation, and $\mathbf{e} = [\,0\ 0\ 0\ 1\,]^{\mathsf{T}}$:

$$\mathbf{P}_{\text{LR}} = \begin{bmatrix} 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & -1 \end{bmatrix}, \text{ i.e. } \mathbf{y} = \mathbf{P}_{\text{LR}} \log(\mathbf{x}) \quad \mathbf{Q}_{\text{LR}} = \begin{bmatrix} 0 & -e^{y_1} & 1 & 0 \\ 0 & 0 & 1 & -e^{y_2} \\ 1 & 0 & 0 & -e^{y_3} \\ 1 & 1 & 1 & 1 \end{bmatrix}, \text{ i.e. } \mathbf{x} = \mathbf{Q}_{\text{LR}}^{-1}\mathbf{e} \tag{11}$$

The same approach can be used to invert a set of $J-1$ independent amalgamation logratio balances (SLRs), which must involve each part at least once. The SLRs have a pattern matrix $\mathbf{P}_{\text{SLR}}$ indicating which parts are in the numerator and the denominator, but this is not the matrix of the transformation, which is not linear in $\log(\mathbf{x})$, hence not a contrast matrix. For example, the pattern matrix for the SLRs in (9) as well as the corresponding inversion pattern matrix are as follows:

$$\mathbf{P}_{\text{SLR}} = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \qquad \mathbf{Q}_{\text{SLR}} = \begin{bmatrix} 1 & 1 & -e^{y_1} & -e^{y_1} \\ 1 & -e^{y_2} & 0 & 0 \\ 0 & 0 & 1 & -e^{y_3} \\ 1 & 1 & 1 & 1 \end{bmatrix} \tag{12}$$

Each row of $\mathbf{P}_{\text{SLR}}$ corresponding to an SLR has 1s in the columns of the numerator parts, and -1s in the columns of the denominator parts. Then, as before, replace all the -1s in the k-th row of $\mathbf{Q}_{\text{SLR}}$ by $-e^{y_k}$ and add a row of 1s as the last row, which again automatically closes the parts in the solution. The compositional vector can then be recovered by $\mathbf{x} = \mathbf{Q}_{\text{LSR}}^{-1}\mathbf{e}$. Notice that both the LR and ALR inverse transformations are special cases of the SLR one, where LRs and ALRs have only one numerator and one denominator part. For specific examples of inverse transforms of sets of LRs and sets of SLRs, see Greenacre [14].

Log-Contrasts

A log-contrast is a linear combination of logarithms of all the components of composition, with the condition that the coefficients sum to 0:

$$\sum_j a_j \log(x_j), \quad \text{where } \sum_j a_j = 0 \tag{13}$$

The logratio pattern matrices $\mathbf{P}_{\text{ALR}}$ in (3), $\mathbf{P}_{\text{CLR}}$ in (6) and $\mathbf{P}_{\text{PLR}}$ in (7), as well as a particular pattern matrix associated with a DAG such as (2), transform the vector $\log(\mathbf{x})$ to the corresponding set of $J-1$ logratios, or J logratios in the case of the CLR transformation. For any one of these sets of logratios, denoted in general by $\ell_1, \ell_2, \ldots$, the coefficients $\mathbf{c}$ of a linear combination of them $c_1\ell_1 + c_2\ell_2 + \cdots$ can be converted to the coefficients of the log-contrast simply by pre-multiplying $\mathbf{c}$ by the transpose of the pattern matrix. For example, for a linear combination of ALRs, the J coefficients of the log-contrast are $\mathbf{a} = \mathbf{P}_{\text{ALR}}^{\mathsf{T}}\mathbf{c}$. This result is useful when a linear combination of logratios, used as explanatory variables in a generalized linear model, is estimated in explaining/predicting a response variable [5].

This result is illustrated for four different transformations, for the linear modeling of a response variable in the form of the proportion of total budget in the same 30 countries devoted to Housing and Community Amenities, logarithmically transformed, denoted by $\log(y)$. For example, if one defines the ALR with Health as the reference component, then the estimated regression model is, along with p-values in parentheses and proportion of explained variance, R^2:

$$\log(y) = \underset{(p=0.004)}{-0.482} + \underset{(p=0.030)}{0.029\log\left(\frac{\text{Defence}}{\text{Health}}\right)} + \underset{(p=0.83)}{1.032\log\left(\frac{\text{Economy}}{\text{Health}}\right)} - \underset{R^2=0.295}{0.904\log\left(\frac{\text{Education}}{\text{Health}}\right)}$$

The coefficients of the log-contrast, using the corresponding ALR pattern matrix, are then (always remembering the order of the components):

$$\mathbf{a} = \mathbf{P}_{\text{ALR}}^{\mathsf{T}}\mathbf{c} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.029 \\ 1.032 \\ -0.904 \end{bmatrix} = \begin{bmatrix} 0.029 \\ 1.032 \\ -0.157 \\ -0.904 \end{bmatrix}$$

It can be verified that $\mathbf{1}^\mathsf{T}\mathbf{a} = 1$. Thus, the regression model can be written as the constant plus the log-contrast:

$$\log(y) = -0.482 + 0.029\log(\text{Defence}) + 1.032\log(\text{Economy}) - 0.157\log(\text{Health}) - 0.904\log(\text{Education})$$

The coefficients of the components in the log-contrast are interpreted as the additive effect on the mean of the dependent variable $\log(y)$ of increasing each component in turn while decreasing all others by a common factor. Exponentiating the coefficients would give the multiplicative effects on y itself.

From the form of $\mathbf{P}_{\text{ALR}}^\mathsf{T}$ and the fact that any ALR transformation will give the same log-contrast, it is clear that the coefficient of the reference part is the one that will change. For example, if Education is the reference, the regression coefficients turn out to be $[\ 0.029\ \ 1.032\ \ -0.157\ ]^\mathsf{T}$ and the log-contrast is shown to be identical:

$$\mathbf{a} = \mathbf{P}_{\text{ALR}}^\mathsf{T}\mathbf{c} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 0.029 \\ 1.032 \\ -0.157 \end{bmatrix} = \begin{bmatrix} 0.029 \\ 1.032 \\ -0.157 \\ -0.904 \end{bmatrix}$$

For any set of pairwise logratios, once again the same log-contrast is obtained. For example, here is the regression model using the logratios in the DAG of Fig. 2:

$$\log(y) = -0.482 - 1.032\log\left(\frac{\text{Health}}{\text{Economy}}\right) + 0.875\log\left(\frac{\text{Health}}{\text{Education}}\right) + 0.029\log\left(\frac{\text{Defence}}{\text{Education}}\right)$$
$$(p = 0.004) \qquad (p = 0.036) \qquad (p = 0.83) \qquad R^2 = 0.295$$

Then, using the pattern matrix $\mathbf{P}$ in (2):

$$\mathbf{a} = \mathbf{P}^\mathsf{T}\mathbf{c} = \begin{bmatrix} 0 & 0 & 1 \\ -1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix} \begin{bmatrix} -1.032 \\ 0.875 \\ 0.029 \end{bmatrix} = \begin{bmatrix} 0.029 \\ 1.032 \\ -0.157 \\ -0.904 \end{bmatrix}$$

The same result is obtained for the CLRs, as well as any set of ILRs or PLRs. One difference with the CLRs is that the pattern matrix is 4×4 and only three coefficients are obtained in a regression, so the fourth one has to be set to zero. In all cases, the constant as well as the R^2 and the p-value for the whole model (which is $p = 0.026$) are identical across the variations, as they all reduce to the same log-contrast.

Often, some type of variable selection is made to arrive at a more parsimonious model. Selecting fewer explanatory variables implies forming a subcomposition of the parts. A statistical criterion is needed to make the selection and there are many possible ways to achieve this. For example, one could do a permutation test on the coefficients of the log-contrast. Using the CLR transformation, and randomizing the order of the response variable 999 times, the p-values for each log-contrast coefficient were estimated as

Defence: $p = 0.86$, Economy: $p = 0.003$, Health: $p = 0.66$, Education: $p = 0.037$.

It seems that only one logratio, that of Economy/Education, can be used as a predictor, and gives the following result:

$$\log(y) = -0.562 + 1.010 \log\left(\frac{\text{Economy}}{\text{Eduction}}\right) \qquad R^2 = 0.286$$
$$(p = 0.002)$$

This leads to the trivial log-contrast of the two components in the model:

$$\log(y) = -0.562 + 1.010 \log(\text{Economy}) - 1.010 \log(\text{Education})$$

For more details about logratios used as predictors in linear modeling, see [5, 6].

3 Logratio Visualization

In this section, we look at various ways of visualizing a compositional data set. Basically, once a logratio transformation is made, any of the various well-known multivariate visualization methods can be implemented, such as cluster analysis and dimension-reduced component methods. Care has to be taken in the interpretation because of the unit sum constraint on the original data. Since these methods rely on interpoint distances, the first thing to do is to define the distance measures between rows and between columns of the data matrix.

If the matrix $\mathbf{Z} = [z_{i,jj'}]$ $(I \times \frac{1}{2}J(J-1))$ denotes the matrix of LRs $\log(x_{ij}/x_{ij'})$, and $\mathbf{Y} = [y_{ij}]$ $(I \times J)$ the CLR-transformed data set, then the *logratio distance* $d_{ii'}$ between samples i and i' can be defined in two equivalent forms, shown for the weighted and unweighted versions in (14) and (15), respectively [12]:

$$d_{ii'} = \sqrt{\sum\sum_{j<j'} c_j c_{j'} (z_{i,jj'} - z_{i',jj'})^2} = \sqrt{\sum_j c_j (y_{ij} - y_{ij'})^2} \tag{14}$$

$$d_{ii'} = \sqrt{\frac{1}{J^2}\sum\sum_{j<j'} (z_{i,jj'} - z_{i',jj'})^2} = \sqrt{\frac{1}{J}\sum_j (y_{ij} - y_{ij'})^2} \tag{15}$$

The advantage of the versions using CLRs is the use of a much narrower matrix, but—as will be emphasized repeatedly—the results should always be interpreted in terms of pairwise logratios. The CLRs as such have no inherent interpretation as representing the components and simply act as a shortcut to analyzing all the LRs. The above holds when the structure of a composition is being investigated internally, that is, in an unsupervised learning mode. However, when logratios are used in a supervised mode as predictors in a generalized linear model, there is less interest in their relationship to one another, and attention is focused on their effect on the response variable, where an increase in one component is at the expense of decreasing others.

There are results similar to (14) and (15) for the columns, by transposing the data set, renormalizing and performing the same operations. The samples are usu-

	Public services	Defence	Public safety	Economy	Environment	Housing	Health	Culture & Religion	Education	Social protection
Belgium	13.3%	1.5%	3.2%	12.8%	2.5%	0.7%	14.6%	2.4%	11.8%	37.2%
Bulgaria	7.8%	3.4%	7.6%	17.8%	1.9%	3.4%	13.7%	2.0%	10.7%	31.8%
Czechia	10.6%	2.1%	4.6%	14.8%	2.0%	1.6%	18.4%	3.5%	11.8%	30.5%
Denmark	12.2%	2.3%	1.9%	6.2%	0.8%	0.4%	16.7%	3.2%	12.7%	43.5%
Germany	12.5%	2.4%	3.6%	7.4%	1.3%	1.0%	16.3%	2.3%	9.6%	43.7%
Estonia	9.0%	5.3%	4.6%	10.1%	1.7%	1.0%	13.7%	5.2%	15.5%	33.9%
Ireland	11.0%	0.9%	3.8%	9.3%	1.6%	2.7%	19.3%	2.2%	12.8%	36.3%
Greece	16.7%	4.2%	4.4%	8.4%	2.9%	0.4%	11.2%	1.7%	8.3%	41.7%
Spain*	13.0%	2.0%	4.3%	9.5%	2.1%	1.0%	14.5%	2.7%	9.5%	41.3%
France*	10.0%	3.1%	3.0%	10.8%	1.8%	1.9%	14.5%	2.6%	9.5%	42.9%
Croatia	12.0%	2.2%	4.6%	17.5%	1.4%	4.3%	13.5%	3.3%	10.2%	31.0%
Italy	15.3%	2.6%	3.8%	8.3%	1.8%	1.0%	14.0%	1.6%	8.0%	43.5%
Cyprus	18.6%	4.4%	4.3%	11.5%	0.7%	3.9%	9.9%	2.3%	13.4%	31.0%
Latvia	9.8%	5.0%	5.8%	13.8%	1.5%	2.7%	11.0%	3.8%	15.0%	31.4%
Lithuania	10.1%	4.7%	4.0%	8.6%	1.0%	1.4%	18.0%	3.4%	13.3%	35.4%
Luxembourg	11.8%	0.9%	2.8%	12.3%	2.2%	1.4%	12.0%	3.0%	11.0%	42.7%
Hungary	17.9%	2.2%	4.7%	17.6%	1.0%	1.7%	10.0%	6.6%	10.3%	27.9%
Malta	15.0%	1.9%	3.4%	13.8%	3.9%	1.1%	14.7%	2.9%	14.2%	29.1%
Netherlands	9.8%	3.0%	4.4%	9.0%	3.3%	0.8%	18.3%	2.8%	11.8%	36.7%
Austria	11.8%	1.2%	2.8%	11.9%	0.8%	0.6%	17.1%	2.4%	9.9%	41.6%
Poland	10.1%	3.9%	5.0%	11.6%	1.3%	1.3%	11.7%	3.2%	12.0%	40.0%
Portugal**	15.7%	1.8%	4.0%	8.5%	1.4%	1.1%	15.4%	2.1%	10.3%	39.7%
Romania	11.7%	4.6%	6.2%	13.1%	2.0%	2.9%	13.8%	2.9%	10.1%	32.7%
Slovenia	12.0%	2.3%	3.7%	10.5%	1.3%	1.0%	15.4%	3.1%	12.6%	38.1%
Slovakia*	12.5%	2.6%	5.5%	12.0%	1.9%	1.2%	18.0%	2.9%	9.8%	33.6%
Finland	14.9%	2.3%	2.2%	7.9%	0.4%	0.6%	13.4%	2.8%	10.6%	45.1%
Sweden	14.0%	2.5%	2.7%	9.0%	1.0%	1.4%	14.2%	2.6%	14.1%	38.6%
Iceland	16.5%	0.2%	3.4%	11.2%	1.4%	1.3%	17.9%	6.9%	16.1%	25.1%
Norway	9.3%	3.7%	2.4%	11.7%	1.8%	1.6%	16.9%	3.6%	10.9%	38.2%
Switzerland	12.9%	2.5%	5.0%	11.8%	1.7%	0.6%	6.5%	3.0%	16.6%	39.4%

Fig. 6 Expenditure in European Union countries plus Iceland, Norway and Switzerland on ten budget items, in 2019. The data are expressed as percentages of the expenditure on these items (i.e. row sums are 100%). Some column names have been slightly abbreviated—see the original longer names in Fig. 7

ally unweighted (i.e. with weights $1/I$) but can also be differentially weighted if required—see Greenacre [12]. Later for the definition of the biplot, the theory is presented in complete generality with weights on the rows and the columns, where equal weighting on the columns is referred to as the unweighted analysis.

Figure 7 shows the clustering of the rows and columns respectively of an extended data set of country budget items (Fig. 6), using the `easyCODA` package [12] in R [19], and Ward clustering in each case [21]. In Ward clustering, the successive clusters are built up by averaging the CLRs of objects contributing to a particular node. For example, at the top of Fig. 7a the group of countries Malta, Belgium, Luxembourg and Ireland are joined together by averaging their respective CLRs across the items. Similarly, at the top of Fig. 7b the group of items Social protection, General Public services and Health are joined together by averaging their respective CLRs across the countries, in this case by weighted averaging. The weights are introduced due to the higher ratios created by low-percentage items and lower ratios created by high-percentage items [12, 16]. In both cases, the between-group logratio variance of the individual leaves at the start of the clustering is equal to the total logratio variance, and it reduces to zero when all objects are clustered. The sum of the node values is thus equal to the total logratio variance, in this case including item weighting in both clusterings.

A completely different way of performing the clustering of the components is by successively amalgamating the components [14], shown in Fig. 8. Amalgamation clustering aggregates the components by simple summation, which is a more natural way of combining them. For example, the initial two items in Fig. 8a, Social protection

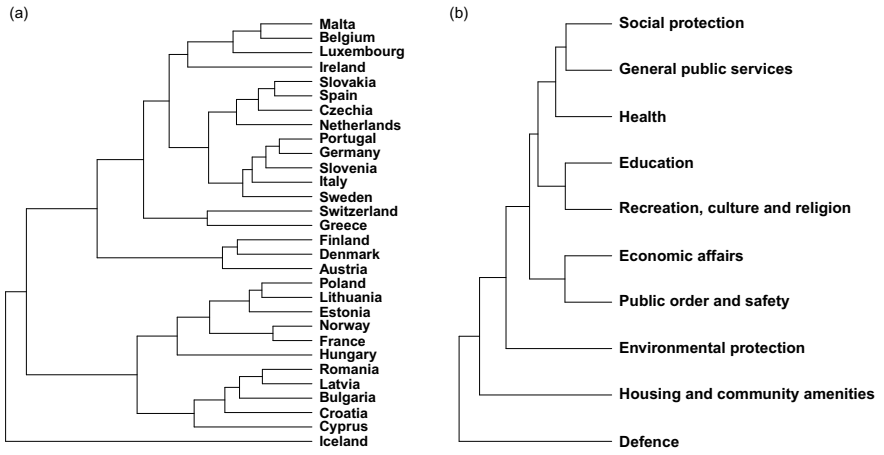


Fig. 7 Clustering of **a** countries and **b** budget items, using Ward clustering of the logratio distances. The budget items are weighted, as in (14), using weights proportional to their average marginal proportions

and Education, are combined by summing their percentages, and in the next step of the clustering the percentages of Economic affairs are similarly added to the two before. The cluster heights are defined by the amount (or percentage) of explained logratio variance—at the start, the full set of budget items explains 100% of the variance. When Social protection and Education are combined by summing, there is a part of explained variance that is lost, but these two combined are minimizing that loss. The procedure ends when all the items are combined into a constant vector, which explains no variance at all. Thus, the values at the nodes similarly decompose the total logratio variance.

Both unweighted and weighted versions are shown in Fig. 8, giving different results, and different from the Ward clustering in Fig. 7b. The unweighted version turns out to be identical to the graph structure of a set of PLRs. The weighted version, which decomposes weighted logratio variance, de-emphasizes the role of Defence, and shows that there are three pairs of items that could easily be merged: Housing and community amenities with Economic affairs, Environmental protection with Public order and safety, and Education with Recreation, culture and religion.

Next, a logratio biplot is shown in Fig. 9, using logratio analysis (LRA). This biplot is obtained using the singular value decomposition (SVD) of the double-centered matrix of the log-transformed compositional data set, $\log(\mathbf{X})$. An unweighted or weighted version of LRA is possible, and here weighted LRA is used, where weights are imposed on the budget items equal to the average proportions of the items across the countries, as before. This weighting will reduce the influence of some budget items with low average proportions but high logratio variance. The configuration of the countries is approximating the weighted logratio distances in (14). This biplot and the clusterings of Fig. 7 are using the same weighted logratio distances, hence the

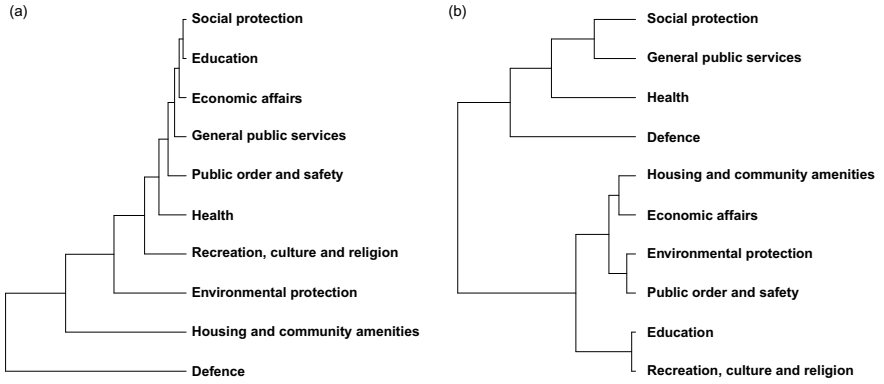


Fig. 8 Clustering of budget items using amalgamation clustering: **a** unweighted and **b** weighted

outlying position of Defence, but notice that only 55.8% of the interpoint distance variance is explained by the two-dimensional biplot, whereas the cluster analysis is performed in the full space of the objects (countries and budget items, respectively).

Using matrix-vector notation, the sequence of steps to perform LRA and arrive at the biplot in Fig. 9 is as follows, starting from the matrix of log-transformed compositional data, $\log(\mathbf{X})$. Note that the most general row- and column-weighted version is given here, with row and column weights $\mathbf{r}$ and $\mathbf{c}$, respectively. Usually, but not necessarily, the rows are equally weighted: $\mathbf{r} = (1/I)\mathbf{1}$, and the description weighted or unweighted LRA refers to the nature of the column weights $\mathbf{c}$, which can be different (e.g. by default equal to the marginal average proportions of the components [16]) or equal: $\mathbf{c} = (1/J)\mathbf{1}$ [4].

$$\text{Double-center the matrix } \log(\mathbf{X}): \quad \mathbf{Z} = (\mathbf{I} - \mathbf{1}\mathbf{r}^{\mathsf{T}})\log(\mathbf{X})(\mathbf{I} - \mathbf{1}\mathbf{c}^{\mathsf{T}})^{\mathsf{T}} \tag{16}$$

$$\text{Apply weights to rows and columns:} \quad \mathbf{S} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{Z}\mathbf{D}_c^{\frac{1}{2}} \tag{17}$$

$$\text{Perform the SVD:} \quad \mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\mathsf{T}} \tag{18}$$

$$\text{Principal coordinates of rows:} \quad \mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\alpha \tag{19}$$

$$\text{Standard coordinates of columns:} \quad \Gamma = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V} \tag{20}$$

$$\text{Contribution coordinates of columns:} \quad \Gamma^* = \mathbf{D}_c^{\frac{1}{2}}\Gamma = \mathbf{V} \tag{21}$$

Two-dimensional biplots of the results are given by plotting the row principal coordinates in the first two columns of $\mathbf{F}$ with either the corresponding column standard coordinates in the first two columns of Γ (*asymmetric biplot*) or those of the contribution coordinates in Γ^* (*contribution biplot*) [11].

The two-dimensional *symmetric map* of the LRA is the plotting of the first two columns of the row principal coordinates $\mathbf{F}$ in (19) jointly with those of the column principal coordinates $\mathbf{G}$:

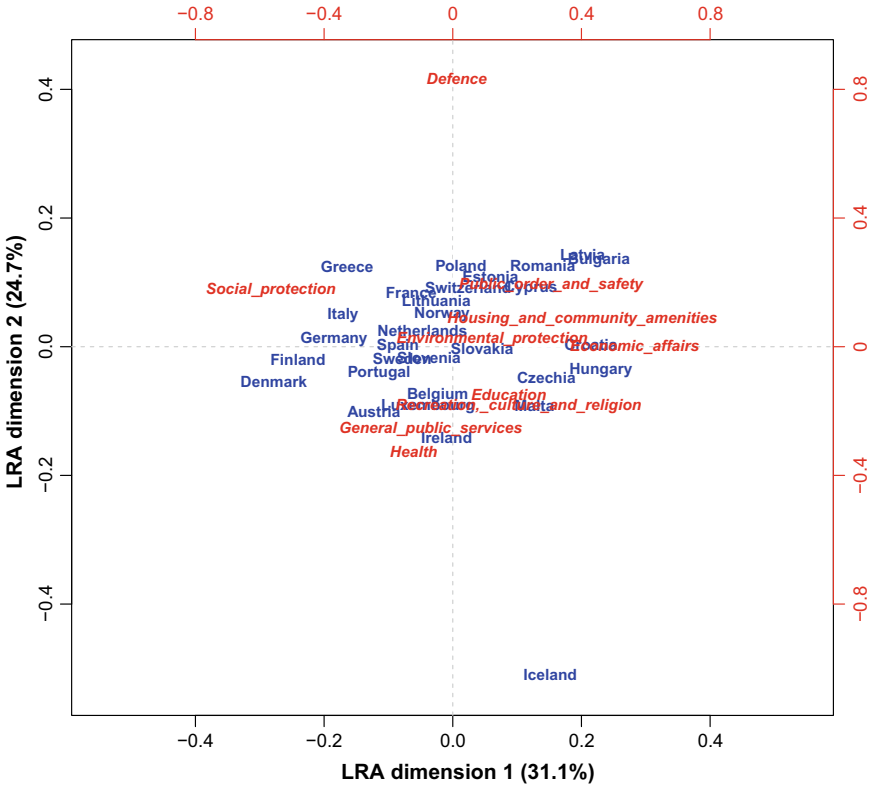


Fig. 9 Weighted logratio biplot of the data in Fig. 6, with asymmetric biplot scaling: rows in principal coordinates, columns in standard coordinates

Principal coordinates of columns: $$\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\alpha \tag{22}$$

This is not a biplot, strictly speaking, but has the practical advantage that the row and column points are equally scaled along the principal axes, their weighted variances both being equal to the amount of variance explained on the axes, i.e. the eigenvalues or squared singular values α_k^2 on axis k, $k = 1, 2$.

The sum of the squared singular values, $\sum_k \alpha_k^2$ (i.e. sum of eigenvalues), equals the total logratio variance and α_k^2 is the part of variance explained by axis k, usually expressed as a percentage of this total, as shown on the axes in Fig. 9.

The steps (16)–(21) are equivalent to performing a principal component analysis (PCA) on the matrix of (weighted) CLRs, because the CLRs are the row-centered $\log(\mathbf{X})$ (i.e. $\log(\mathbf{X})(\mathbf{I} - \mathbf{1}\mathbf{c}^\mathsf{T})^\mathsf{T}$ in (16)) and PCA automatically performs column-centering (i.e. the centering $(\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})$ in (16), with equal weights in $\mathbf{r}$), hence the double-centering. The steps from (17) onwards define the PCA with its variations of display coordinates. Because of the double-centering of $\log(\mathbf{X})$, the definition of

each principal component as a linear combination of the CLRs turns out to be a log-contrast.

The interpretation of Fig. 9 is not the same as a regular PCA, however. The CLRs as variables are not interpretable *per se*, but rather the differences between pairs of CLR points, which depict the pairwise logratios themselves. Aitchison and Greenacre [4] show that the LRA biplot optimizes the display of these pairwise logratios, which is not the case in a regular PCA (i.e. in a regular PCA, the optimization of the variables is not equivalent to the optimization of the differences between pairs of variables). Thus, the horizontal dispersion in Fig. 9 is due to logratios such as Public Order and Safety divided by Social Protection, while the vertical dispersion is dominated by the logratio of Defence relative to Health, for example, where it is clear that Iceland's ratios of Defence relative to the other budget items are low.

The equivalence between the PCA of the CLRs and the PCA of the LRs (i.e. LRA in both cases) can be neatly shown using the respective logratio pattern matrices defined earlier. The proof follows the one given for unweighted logratios in Appendix of [4], but is more elegantly defined for the general weighted case. The double-centered matrix $\mathbf{Z}$ in (16) is the matrix $\mathbf{L} = \log(\mathbf{X})$ post-multiplied by the transposed column-centering matrix, which is identical to the transposed CLR pattern matrix in (6), and pre-multiplied by the row-centering matrix, i.e. $\mathbf{Z} = (\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{L}\mathbf{P}_{\text{CLR}}^\mathsf{T}$. The corresponding result for the matrix of pairwise LRs $\mathbf{L}\mathbf{P}_{\text{LR}}^\mathsf{T}$ is the row-centered matrix $\mathbf{Y} = (\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{L}\mathbf{P}_{\text{LR}}^\mathsf{T}$. The weights in the two respective cases are $c_1, c_2, \ldots, c_J$ for the J CLRs and $c_1c_2, c_1c_3, \ldots, c_{J-1}c_J$ for the $\frac{1}{2}J(J-1)$ LRs, gathered in the diagonals of the diagonal matrices $\mathbf{D}_c$ and $\mathbf{D}_{cc}$, respectively. The matrix of weighted scalar products for the rows of $\mathbf{Z}$ and $\mathbf{Y}$ are then identical. These $I \times I$ matrices are called the *form matrices*:

$$\mathbf{Z}\mathbf{D}_c\mathbf{Z}^\mathsf{T} = \mathbf{Y}\mathbf{D}_{cc}\mathbf{Y}^\mathsf{T} \tag{23}$$

thanks to the result $\mathbf{P}_{\text{CLR}}^\mathsf{T}\mathbf{D}_c\mathbf{P}_{\text{CLR}} = \mathbf{P}_{\text{LR}}^\mathsf{T}\mathbf{D}_{cc}\mathbf{P}_{\text{LR}}$. This proves that the configuration of the samples is identical. The distances in (14) correspond exactly to the form matrices in (23).

Notice that the equivalence of the PCA of the CLRs and the PCA of the LRs is a particular result, thanks to the definition of the CLRs. In regular PCA, the low-dimensional result is not optimal for differences between variables. For example, Gabriel [9] gives examples of difference vectors in biplots but makes no statements about their optimality, whereas [10] shows specifically how difference vectors can be optimally displayed.

4 Summary and Discussion

To capture the relative values in a compositional data set, the essential step is to perform one of the available logratio transformations. The complete set of pairwise logratios contains all the logratio variance, but only a subset of them, one less than

the number of components, is required to explain the totality of this variance in a regression sense. For multivariate analysis that requires the complete information in a compositional data set, for example cluster analysis or the reduction to principal dimensions, the centered logratios (CLRs) are sufficient. The CLRs are equivalent to analyzing all pairwise logratios, but they are not linearly independent, so if the inverse of their singular covariance matrix is required, for example for discriminant analysis or regression, then the generalized inverse needs to be used. Additive logratios (ALRs) have been shown to be a satisfactory substitute for a compositional data set of high dimensionality, where the challenge is to find the reference component that leads to the transformation that best captures the logratio geometry [18].

If pairwise logratios are used as predictors in a generalized linear model or any other supervised learning procedure, for example generalized additive modeling or classification and regression trees, then only those that are related to the response variable would be chosen using some variable selection method. In this case, it is more relevant to ascertain how well the logratios explain the response variable (i.e. supervised learning objective), not how well they explain structure in the compositional data set itself (i.e. unsupervised learning objective). The interpretation of logratios as predictors in a linear model is best made via the log-contrast form, which is equivalent to the linear combination of logratios. But notice that when logratios are used as explanatory variables (see Sect. 2.6), there are important issues of effect-size interpretation, discussed by [5].

As John Aitchison said, "Compositional data analysis is simple" [3]. The basic concept is the logratio transformation, after which statistical analysis proceeds very much as before. But care needs to be taken in the interpretation: the analyst has to think in terms of pairwise logratios and realize the implications of the results of the initial normalization of the data to have constant sums.

The one real snag of compositional data analysis, what I have called its "Achilles heel", is how to handle zero values, which often abound in a compositional data set. Several proposals have been made to substitute data zeros with small positive values so that logratio transformations can be performed—see [8, 15] for an overview of these. This is a continuing issue which so far has not been fully resolved.

Another issue is that of using amalgamated components in logratios; that is, if A, B and C are components, is it valid to define a logratio as $\log(A/(B+C))$? Although this might seem like the most natural thing to do if the combining of components B and C makes substantive sense, this practice is surprisingly opposed on mathematical grounds by several authors who favor isometric logratios (ILRs), for example [7]. In the alternative clustering algorithm using amalgamations (see Fig. 8), the groupings of components by simple summations are easy to understand, compared to the groupings by geometric means in ILRs and PLRs. The use of amalgamation logratios (or summated logratios, abbreviated as SLRs) is defended in several publications, notably by Greenacre et al. [17], Greenacre [14], Wood and Greenacre [22] and Smithson and Broomell [20], where it is maintained that any summing of components that fits the research question is permitted.

References

1. Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society: Series B, 44*, 139–77.
2. Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman & Hall.
3. Aitchison, J. (1997). The one-hour course in compositional data analysis, or compositional data analysis is simple. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97* (pp. 3–35). International Association for Mathematical Geology.
4. Aitchison, J., & Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 51*, 375–92.
5. Coenders, G., & Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT, 44*, 201–20.
6. Coenders, G., & Greenacre, M. (2022). Three approaches to supervised learning for compositional data with pairwise logratios. *Journal of Applied Statistics*. https://doi.org/10.1080/02664763.2022.2108007.
7. Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geosciences, 37*, 795–828.
8. Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis*. Oxford: Oxford University Press.
9. Gabriel, K. R. (1972). Analysis of meteorological data by means of canonical decomposition and biplots. *Journal of Applied Meteorology and Climatology, 11*, 1077–7.
10. Greenacre, M. (2003). Singular value decomposition of matched matrices. *Journal of Applied Statistics, 30*(10), 1101–13.
11. Greenacre, M. J. (2013). Contribution biplots. *Journal of Computational Statistics and Graphics*, 22: 107–122.
12. Greenacre, M. (2018). *Compositional data analysis in practice*. Boca Raton: Chapman & Hall / CRC Press.
13. Greenacre, M. (2019). Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences, 51*, 649–82.
14. Greenacre, M. (2020). Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. *Applied Computing and Geosciences, 5,* 100017. https://doi.org/10.1016/j.acags.2019.100017.
15. Greenacre, M. (2021). Compositional data analysis. *Annual Review of Statistics and Its Application, 8*, 271–99.
16. Greenacre, M., & Lewi, P. (2009). Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of Classification, 26*, 29–54.
17. Greenacre, M., Grunsky, E., & Bacon-Shone, J. (2020). A comparison of amalgamation and isometric logratios in compositional data analysis. *Computers & Geosciences, 148*, 104621.
18. Greenacre, M., Martínez-Álvaro, M., & Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: A validation of the additive logratio transformation. *Frontiers in Microbiology, 12,* 2625. ISSN 1664-302X. https://doi.org/10.3389/fmicb.2021.727398. https://www.frontiersin.org/article/10.3389/fmicb.2021.727398.
19. R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
20. Smithson, M., & Broomell, S. (2021). Compositional data analysis tutorial. *Psychological Methods, 26* (In revision).
21. Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.
22. Wood, J., & Greenacre, M. (2021). Making the most of expert knowledge to analyse archaeological data: A case study on parthian and sasanian glazed pottery. *Archaeological and Anthropological Sciences, 13*, 110. https://doi.org/10.1007/s12520-021-01341-0.

Multivariate Count Data Regression Models and Their Applications

Ayman Alzaatreh, Felix Famoye, and Carl Lee

Abstract Multivariate regression models based on multivariate discrete distributions will be defined and studied. Multivariate discrete distributions including some distributions generated from the T-$R\{W\}$ method will be defined. Models that allow both positive and negative correlation between any pair of response variables will be considered. The model parameters will be estimated by using the method of maximum likelihood estimation. The application of these regression models will be illustrated by using two numerical datasets.

Keywords T-X family · Sarmanov family · Nested models · Non-nested models

1 Introduction

Many univariate count data regression models have been defined and studied. Some of these models have been extended to bivariate, and a few of them have been extended to multivariate count data regression models. Examples can be found in the books by [6, 25] and the references therein.

The major limitation of the standard Poisson regression model is that it can be used to model response count data in which the mean and the variance are about equal, the case of equi-dispersion. Quite often, a count data may be over-dispersed (variance > mean) or under-dispersed (variance < mean). Few univariate count data regression models in the literature allow for over- and under-dispersion. Winkelmann [25] (pp. 45–56) discussed three such models. Sellers and Shmueli [22] gave

A. Alzaatreh
Department of Math & Statistics, American University of Sharjah, Sharjah, UAE
e-mail: aalzaatreh@aus.edu

F. Famoye (✉) · C. Lee
STAD Department, Central Michigan University, Mt. Pleasant, MI, USA
e-mail: felix.famoye@cmich.edu

C. Lee
e-mail: carl.lee@cmich.edu

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_11

COM-Poisson regression model. Sun and Ong [23] discussed the generalized inverse trinomial distribution. These regression models have some drawbacks that include probability mass functions in complicated forms and/or some of the means and variances of the response variables are not available in closed forms.

There are some methods to extend the univariate count data regression models to bivariate or multivariate count data regression models. One method is the trivariate reduction that leads to a bivariate regression model (see [14], Chap. 1). The drawback of this technique is that the correlation between the two variates is always positive. Modeling multivariate count data using copulas is another common technique. Copula-based models allow for flexible dependence structure and marginal distributions. However, the choice of copulas and the marginal distributions is not an easy task. Nikoloulopoulos and Karlis [19] considered multivariate Archimedean copulas, partially symmetric copulas, and mixtures of max-id bivariate copulas. Some other types of copula-based multivariate models for count data were reviewed by [18] and the references therein. Other techniques to obtain bivariate models and their drawbacks were discussed by [9].

In this paper, we define a family of multivariate distributions using the T-$R\{W\}$ framework. The multivariate count data regression models based on the T-$R\{W\}$ family of discrete distributions are defined. Some important characteristics of the multivariate count data regression models include the following: (i) they allow any type of correlation between any two or more variates, (ii) they allow for both under- and over-dispersion for each variate, (iii) they allow correlations and dispersions to be determined independently, and (iv) the cumulative distribution function (CDF) of each marginal distribution is in closed form. The means and variances of the marginal distributions may not be in closed form, but they are easy to compute.

Another motivation for this work is from the paper by [11] on the bivariate exponentiated-exponential-geometric regression (BEEGR) model. A simulation study was carried out to compare the BEEGR model with the bivariate generalized Poisson regression (BGPR) model studied by [13]. Data were generated from each of the two regression models, both models were fitted to the data, and the proportion of times one model is significantly better than the other was recorded. P_{ee} denotes the proportion of times BEEGR is better than the BGPR while P_{gp} denotes the proportion of times BGPR is better than the BEEGR. When the data were simulated from BEEGR model, one would expect $P_{ee} > P_{gp}$. One would expect the opposite if the data were simulated from the BGPR model. The results presented by [11] showed that P_{ee} exceeded P_{gp} in all cases when the data were generated from the BEEGR model. On the other hand, P_{gp} did not exceed P_{ee} in all cases when the data were simulated from the BGPR model. Based on the conclusion in the paper, the BEEGR model seems to have advantages over the BGPR model.

In Sect. 2, we briefly review the T-$R\{W\}$ family of distributions. In the section, we focus on the T-$R\{W\}$ family of discrete distributions. Some sub-families of the T-$R\{W\}$ discrete distributions are presented with some specific examples. Some examples of generalized geometric distributions are provided. Bivariate and multivariate generalized geometric distributions are defined in Sect. 3. Multivariate generalized geometric regression model is defined in Sect. 3. In Sect. 4, we propose the maximum

likelihood method for estimating the model parameters and some tests that include goodness-of-fit statistics. In Sect. 5, application of some multivariate generalized geometric regression models is illustrated with two numerical data sets. In Sect. 6, we provide some summary and concluding remarks.

2 Review of *T-R*{*W*} Family of Distributions

There are many methods proposed in the literature for generating statistical distributions. See, for example, a review by [16] and the references therein. [4] proposed the *T-X*(*U*) family of distributions by connecting the random variable *T* with support (*a*, *b*) and *X* using the transformation *U*. The CDF of the *T-X*(*U*) family of distributions is given by

$$G(x) = \int_a^{U(F(x))} r(t)dt = R\{U(F(x))\}, \tag{1}$$

where $U(.) : (0, 1) \rightarrow (a, b)$ satisfies the following two conditions: (i) $U(.)$ is monotonically non-decreasing and absolutely continuous and (ii) $U(0) \rightarrow a$ and $U(1) \rightarrow b$. The CDFs of *T* and *X* are $R(t)$ and $F(x)$, respectively. The corresponding probability density function (PDF) for the continuous random variable *X* in (1) is given by

$$g(x) = f(x)u(F(x))r\{U(F(x))\}, \text{ where } u(x) = \tfrac{d}{dx}U(x). \tag{2}$$

Giving a function *U* and a random variable *T*, the resulting *T-X*(*U*) distribution is a generalized distribution of *X*.

Aljarrah et al. [2] considered the function $U(F(x))$ in (1) to be the quantile function $Q_W(F(x))$ of a random variable *W* and defined the *T-R*{*W*} framework. By using the unified notation in [3], we denote the CDFs of the random variables *T*, *R,* and *W* by $F_T(y) = P(T \leq y)$, $F_R(y) = P(R \leq y)$ and $F_W(y) = P(W \leq y)$, respectively. The corresponding quantile functions are $Q_T(p)$, $Q_R(p)$ and $Q_W(p)$, where $Q_Z(p) = \inf\{z : F_Z(z) \geq p\}$. If the densities exist, they can be denoted by $f_T(y)$, $f_R(y)$, and $f_W(y)$. The CDF of the random variable *Y* in the *T-R*{*W*} family of distributions is defined as

$$F_Y(y) = \int_a^{Q_W(F_R(y))} f_T(t)dt = F_T\{Q_W(F_R(y))\}. \tag{3}$$

It is assumed the supports of the random variables *T* and *W* are (*a*, *b*) and (*c*, *d*), respectively, where $(a, b) \subset (c, d)$ for $-\infty \leq a < b \leq \infty$ and $-\infty \leq c < d \leq \infty$. The PDF corresponding to the CDF in (3) when *R* is continuous can be written as

Table 1 Examples of *T-R{W}* distributions based on different choices of random variable *W**

W	$Q_W(p)$	$G(y)$
(a) Uniform	p	$F_T[F_R(y)]$
(b) Exponential	$-\log(1-p)$	$F_T\left[-\log(1-F_R(y))\right]$
(c) Log-logistic	$p/(1-p)$	$F_T[F_R(y)/(1-F_R(y))]$
(d) Logistic	$\log(p/(1-p))$	$F_T\left[\log\{F_R(y)/(1-F_R(y))\}\right]$

*Standard random variable W

$$f_Y(y) = f_R(y) \times \frac{f_T(Q_W(F_R(y)))}{f_W(Q_W(F_R(y)))}. \tag{4}$$

When the support (a, b) of the random variable T is a subset of the support (c, d) of the random variable W, [2] gave the support for the random variable Y in (3) when R is continuous. In this paper, we assume that the support of the random variable T is the same as that of the random variable W. Given the quantile function of random variable W and random variable T, the resulting *T-R{W}* is a generalized distribution of the random variable R. For example, Table 1 shows various *T-R{W}* families based on different choices of Q_W. To simplify the notation, we will use $G(y)$ to replace $F_Y(y)$. Different *T-R{W}* families can be defined using (3), interested readers are referred to [1–3] and the references therein.

If R is a discrete random variable with support $\mathbb{N}_* = \mathbb{N} \cup \{0\}$, then (3) is the CDF of a discrete *T-R{W}* family with the corresponding probability mass function (PMF)

$$g(y) = F_T(Q_W(F_R(y))) - F_T(Q_W(F_R(y-1))), \quad y \in \mathbb{N}_*, \tag{5}$$

where $F(-1) = 0$. An interesting special case of (5) was studied by [5]. They studied the case T-geometric{standard exponential} which defines a discrete analogue to the non-negative continuous random variable T.

Note that the family of discrete distributions in (3) is well defined only if the random variables T and W have the same support. Next, several sub-families of discrete *T-R{W}* are defined.

Sub-Families of Discrete T-R{W} *Distributions*

In this sub-section, several sub-families of the discrete *T-R{W}* will be defined. We study the cases when T follows exponential, exponentiated-exponential, Weibull, and logistic distributions and W follows standard forms of exponential, log-logistic, and logistic distributions.

(a) Discrete exponential-*R{W}* distributions: If a random variable T follows the exponential distribution with parameter λ, then $f_T(y) = \lambda e^{-\lambda t}$, $y > 0$, $\lambda >$

0. From (5), the CDF and the PMF of the discrete exponential-$R\{W\}$ are, respectively, given by

$$G(y) = 1 - \exp\{-\lambda Q_W(F_R(\lfloor y \rfloor))\}, \ \ y \geq 0, \tag{6}$$

$$g(y) = \exp\{-\lambda Q_W(F_R(\lfloor y - 1 \rfloor))\} - \exp\{-\lambda Q_W(F_R(\lfloor y \rfloor))\}, \ \ y \in \mathbb{N}_*, \tag{7}$$

where $\lfloor y \rfloor = \max\{m \in \mathbb{Z} | m \leq y\}$ is the floor function.

Example 1 (Discrete exponential-$R\{$standard exponential$\}$ distributions, denoted as E-$R\{$Es$\}$): If Q_W is the quantile function of the exponential distribution given in Table 1, then the CDF of the discrete E-$R\{$Es$\}$ is given by

$$G(y) = 1 - (1 - F_R(\lfloor y \rfloor))^{\lambda}, \ \ y \geq 0. \tag{8}$$

If $\lambda = 1$, $G(y)$ in (8) reduces to $F_R(y)$.

Example 2 (Discrete exponential-$R\{$standard log-logistic$\}$ distributions, denoted as E-$R\{$LLs$\}$): If Q_W is the quantile function of the standard log-logistic distribution given in Table 1, then the CDF of the discrete E-$R\{$LLs$\}$ is given by

$$G(y) = 1 - \exp\left(\frac{-\lambda F_R(\lfloor y \rfloor)}{1 - F_R(\lfloor y \rfloor)}\right), \ \ y \geq 0. \tag{9}$$

b. Discrete exponentiated-exponential-$R\{W\}$ distributions: If a random variable T follows the exponentiated-exponential distribution defined by [12] with PDF, $f_T(y) = \alpha\lambda(1 - e^{-\lambda y})^{\alpha - 1}e^{-\lambda y}, \ \ y > 0, \ \alpha, \lambda > 0$, then the CDF and the PMF of the discrete exponentiated-exponential-$R\{W\}$ are, respectively, given by

$$G(y) = (1 - \exp\{-\lambda Q_W(F_R(\lfloor y \rfloor))\})^{\alpha}, \ \ y \geq 0, \ \text{and} \tag{10}$$

$$\begin{aligned} g(y) = &(1 - \exp\{-\lambda Q_W(F_R(\lfloor y - 1 \rfloor))\})^{\alpha} \\ &- (1 - \exp\{-\lambda Q_W(F_R(\lfloor y \rfloor))\})^{\alpha}, \ \ y \in \mathbb{N}_*. \end{aligned} \tag{11}$$

Example 3 (Discrete exponentiated-exponential-$R\{$standard exponential$\}$ distributions, denoted as EE-$R\{$Es$\}$): If Q_W is the quantile function of the standard exponential distribution given in Table 1, then the CDF of the discrete EE-$R\{$Es$\}$ is given by

$$G(y) = \left\{1 - (1 - F_R(\lfloor y \rfloor))^{\lambda}\right\}^{\alpha}, \ \ y \geq 0. \tag{12}$$

If $\alpha = 1$, $G(y)$ in (12) reduces to the CDF of the discrete exponential-$R\{$standard exponential$\}$ in (8).

Example 4 (Discrete exponentiated-exponential-$R\{$standard log-logistic$\}$ distributions, denoted as EE-$R\{$LLs$\}$): If Q_W is the quantile function of the standard log-logistic distribution given in Table 1, then the CDF of the discrete EE-$R\{$LLs$\}$ is given by

$$G(y) = \left\{1 - \exp\left(\frac{-\lambda F_R(\lfloor y \rfloor)}{1 - F_R(\lfloor y \rfloor)}\right)\right\}^{\alpha}, \quad y \geq 0. \tag{13}$$

If $\alpha = 1$, $G(y)$ in (13) reduces to the CDF of the discrete exponential-$R\{$log-logistic$\}$ in (9).

c. Discrete Weibull-$R\{W\}$ distributions: If a random variable T follows the Weibull distribution with parameters c and λ, then $f_T(y) = c\lambda^c y^{c-1} e^{-(\lambda y)^c}$, $y \geq 0$. The CDF and the PMF of the discrete Weibull-$R\{W\}$ are, respectively, given by

$$G(y) = 1 - \exp\left\{-(\lambda Q_W(F_R(\lfloor y \rfloor)))^c\right\}, \quad y \geq 0 \tag{14}$$

$$g(y) = \exp\left\{-(\lambda Q_W(F_R(\lfloor y - 1 \rfloor)))^c\right\} - \exp\left\{-(\lambda Q_W(F_R(\lfloor y \rfloor)))^c\right\}, \quad y \in \mathbb{N}_*. \tag{15}$$

Example 5 (Discrete Weibull-$R\{$standard exponential$\}$ distributions, denoted as W-$R\{$Es$\}$): If Q_W is the quantile function of the standard exponential distribution given in Table 1, then the CDF of the discrete W-$R\{$Es$\}$ is given by

$$G(y) = 1 - \exp\left\{-(-\lambda \log(1 - F_R(\lfloor y \rfloor)))^c\right\}, \quad y \geq 0. \tag{16}$$

If $c = 1$, $G(y)$ in (16) reduces to the CDF of E-$R\{$Es$\}$ in (8).

Example 6 (Discrete Weibull-$R\{$standard log-logistic$\}$ distributions, denoted as W-$R\{$LLs$\}$): If Q_W is the quantile function of the standard log-logistic distribution given in Table 1, then the CDF of the discrete W-$R\{$LLs$\}$ is given by

$$G(y) = 1 - \exp\left\{-\left(\frac{\lambda F_R(\lfloor y \rfloor)}{1 - F_R(\lfloor y \rfloor)}\right)^c\right\}, \quad y \geq 0. \tag{17}$$

If $c = 1$, $G(y)$ in (17) reduces to the CDF of the discrete E-$R\{$LLs$\}$ in (9).

d. Discrete logistic-$R\{W\}$ distributions: If a random variable T follows the logistic distribution with parameter λ, then $f_T(y) = \lambda e^{-\lambda y}(1 + e^{-\lambda y})^{-2}$, $-\infty < y < \infty$. The CDF and the PMF of the discrete L-$R\{W\}$ are, respectively, given by

$$G(y) = \frac{1}{1 + \exp\{-\lambda Q_W(F_R(\lfloor y \rfloor))\}}, \quad y \geq 0 \text{ and} \tag{18}$$

$$g(y) = \frac{1}{1+\exp\{-\lambda Q_W(F_R(\lfloor y \rfloor))\}} - \frac{1}{1+\exp\{-\lambda Q_W(F_R(\lfloor y-1 \rfloor))\}}, \; y \in \mathbb{N}_*. \tag{19}$$

Example 7 (Discrete logistic-$R\{\text{standard logistic}\}$ distributions, denoted as L-$R\{\text{Ls}\}$): If Q_W is the quantile function of the logistic distribution given in Table 1, then the CDF of the discrete L-$R\{\text{Ls}\}$ family is given by

$$G(y) = \frac{F_R^{\lambda}(\lfloor y \rfloor)}{F_R^{\lambda}(\lfloor y \rfloor) + (1 - F_R(\lfloor y \rfloor))^{\lambda}}, \; y \geq 0. \tag{20}$$

If $\lambda = 1$, $G(y)$ in (20) reduces to the baseline CDF $F_R(y)$.

The Family of Generalized Geometric Distributions

In this sub-section, we propose a family of generalized geometric distributions by taking the random variable R in the discrete T-$R\{W\}$ framework to follow the geometric distribution with the CDF $F_R(y) = 1 - \varphi^{y+1}$, $0 < \varphi < 1$, $y = 0, 1, 2, \ldots$. The random variable Y denotes the number of failures to obtain the first success and φ is the probability of a failure. Table 2 provides the CDF of some examples of T-geometric$\{W\}$ families. The corresponding PMFs can be found using Eqs. (7), (11), (15), or (19).

Remark 1

(a) Note that the random variable Y takes values $y = 0, 1, 2, \ldots$ for all distributions in Table 2.

(b) Exponential-geometric{standard exponential} in Table 2(i) is just the geometric distribution and exponentiated-exponential-geometric{standard exponential} in Table 3(iii) is the exponentiated-exponential-geometric distribution defined by [5]. By using (8), we obtain $G(y) = 1 - (1 - F_R(\lfloor y \rfloor))^{\lambda} = 1 - \varphi^{\lambda(y+1)} = 1 -$

Table 2 T-geometric$\{W\}$ families based on different choices for the random variables W and T

T-geometric$\{W\}$	CDF $G(y)$
(i) E-geometric{Es}	$G(y) = 1 - \theta^{y+1}$; $\theta \in (0, 1)$
(ii) E-geometric{LLs}	$G(y) = 1 - \exp\left\{-\lambda\left(\theta^{-(y+1)} - 1\right)\right\}$; $\lambda > 0$, $\theta \in (0, 1)$
(iii) EE-geometric{Es}	$G(y) = (1 - \theta^{y+1})^{\alpha}$; $\alpha > 0, \theta \in (0, 1)$
(iv) EE-geometric{LLs}	$G(y) = \left(1 - \exp\left\{-\lambda\left(\theta^{-(y+1)} - 1\right)\right\}\right)^{\alpha}$; $\lambda, \alpha > 0, \theta \in (0, 1)$
(v) W-geometric{Es}	$G(y) = 1 - \theta^{(y+1)^c}$; $c > 0$, $\theta \in (0, 1)$
(vi) W-geometric{LLs}	$G(y) = 1 - \exp\left\{-\left(\lambda(\theta^{-(y+1)} - 1)\right)^c\right\}$; $\lambda, c > 0, \theta \in (0, 1)$
(vii) L-geometric{Ls}	$G(y) = (1 - \theta^{y+1})^{\lambda}/[(1 - \theta^{y+1})^{\lambda} + \theta^{\lambda(y+1)}]$; $\lambda > 0, \theta \in (0, 1)$

Table 3 Bivariate T-geometric$\{W\}$ families based on different choices for variables W and T

Bivariate T-geometric$\{W\}$	Joint probability mass function $h(y_1, y_2)$
(i) BE-geometric{Es}	$\prod_{t=1}^{2}\left[\theta_t^{y_t}(1-\theta_t)\right]\times\{1+\delta(y_1,y_2)\}$
(ii) BE-geometric{LLs}	$\prod_{t=1}^{2}\left[e^{-\lambda_t\left(\theta_t^{-y_t}-1\right)}-e^{-\lambda_t\left(\theta^{-(y_t+1)}-1\right)}\right]\times\{1+\delta(y_1,y_2)\}$
(iii) BEE-geometric{Es}	$\prod_{t=1}^{2}\left[\left(1-\theta_t^{y_t+1}\right)^{\alpha_t}-\left(1-\theta_t^{y_t}\right)^{\alpha_t}\right]\times\{1+\delta(y_1,y_2)\}$
(iv) BEE-geometric{LLs}	$\prod_{t=1}^{2}\left[\left(1-e^{-\lambda_t\left(\theta_t^{-(y_t+1)}-1\right)}\right)^{\alpha_t}-\left(1-e^{-\lambda_t\left(\theta_t^{-y_t}-1\right)}\right)^{\alpha_t}\right]\times\{1+\delta(y_1,y_2)\}$
(v) BW-geometric{Es}	$\prod_{t=1}^{2}\left[\theta_t^{(y_t+1)c_t}-\theta_t^{y_t c_t}\right]\times\{1+\delta(y_1,y_2)\}$
(vi) BW-geometric{LLs}	$\prod_{t=1}^{2}\left[e^{-\left(\lambda_t(\theta_t^{-y_t}-1)\right)^{c_t}}-e^{-\left(\lambda_t(\theta_t^{-(y_t+1)}-1)\right)^{c_t}}\right]\times\{1+\delta(y_1,y_2)\}$
(vii) BL-geometric{Ls}	$\prod_{t=1}^{2}\left[\frac{\left(1-\theta_t^{y_t+1}\right)^{\lambda_t}}{\left(1-\theta_t^{y_t+1}\right)^{\lambda_t}+\theta_t^{\lambda_t(y_t+1)}}-\frac{\left(1-\theta_t^{y_t}\right)^{\lambda_t}}{\left(1-\theta_t^{y_t}\right)^{\lambda_t}+\theta_t^{\lambda_t y_t}}\right]\times\{1+\delta(y_1,y_2)\}$

θ^{y+1}, where $\theta=\varphi^{\lambda}$. See Proposition 1 in [15], where it was shown that the CDF of T-geometric family is characterized by $\theta^{\xi(y)}$ for $\theta\in(0,1)$, where ξ is a function of y. For the Weibull-geometric{standard exponential}, we use (16) to obtain $G(y)=1-\exp\left\{-\left(-\lambda\log\varphi^{y+1}\right)^{c}\right\}=1-\exp\{-(-\lambda\log\varphi)^{c}\times(y+1)^{c}\}=1-[e^{-(-\lambda\log\varphi)^{c}}]^{(y+1)^{c}}=1-\theta^{(y+1)^{c}}$, where $\theta=e^{-(-\lambda\log\varphi)^{c}}$. Some of the probability mass functions in Table 2 are derived in Chapter Appendix.

(c) All the families of discrete distributions generated in this section can also be generated by using any other discrete random variable R with CDF $F_R(y)$. For example, the random variable R can be replaced by the Poisson, binomial, and negative binomial distributions. The only difference is that the geometric distribution has a closed-form CDF. An option is to use approximate functions (e.g., incomplete gamma function for Poisson) to calculate the CDF.

(d) By using the exponentiated Weibull distribution in place of the Weibull distribution, one will obtain the exponentiated Weibull-geometric distribution studied by [11].

(e) These new discrete distributions are characterized by at least one or more new parameters. For example, in Table 2, all the distributions in 2(ii) through 2(vii) are generalization of the geometric distribution characterized by one parameter. Among the new parameters is a shape parameter or a scale parameter. Only the distribution in 2(ii) has new scale parameter. All others from 2(iii) through 2(vii) have either one new shape parameter or have one new shape parameter and one new scale parameter. Of course, one can have more parameters if the

standard quantile functions are not used. These generalized new distributions are derived to have one or two more parameters than the commonly used classic distributions. Thus, the additional parameters are practically useful for capturing the shape and/or scale of the empirical data which could not be properly fitted by the classic distributions. This is one way to reduce the number of parameters and provide enough flexibility for fitting empirical data in the derived new distributions. Furthermore, one can apply test statistics given in Sect. 4 to test if the new parameters are statistically significant to select the most appropriate distribution for fitting a given data.

Bivariate and multivariate extensions of T-geometric$\{W\}$ family are considered in the next section.

3 Bivariate and Multivariate T-geometric$\{W\}$ Families

In this section, we use the technique discussed in [21] to propose bivariate and multivariate extensions of univariate T-geometric$\{W\}$ family. [17] extended bivariate Sarmanov distribution to multivariate case with several parameters measuring covariances of order 2, 3, …, for m-variate distribution.

Sarmanov Family of Bivariate and Multivariate Distributions

In this sub-section, we define the bivariate and multivariate Sarmanov family of distributions.

Definition 1 (*Bivariate Sarmanov family of distributions*) Let Y_1 and Y_2 be two random variables with their probability density (or mass) functions $f_1(y_1)$ and $f_2(y_2)$. Let $\phi_t(u),\ t = 1, 2$ be bounded nonconstant functions such that $E[\phi_t(Y_t)] = 0$. Then the bivariate Sarmanov family is defined as

$$h(y_1, y_2) = f_1(y_1) f_2(y_2)\{1 + \omega\phi_1(y_1)\phi_2(y_2)\},\quad (y_1, y_2) \in \mathbb{R} \times \mathbb{R}, \tag{21}$$

where $\omega \in \mathbb{R}$ satisfies the condition $1 + \omega\phi_1(y_1)\phi_2(y_2) \geq 0$ for all y_1 and y_2.

Theorem 1 *Assume* (Y_1, Y_2) *follows the Sarmanov family in (21) with* μ_t *and* σ_t^2 *the mean and variance for* $Y_t,\ t = 1, 2$. *Then*

(i) *The marginal probability mass functions in (21) are* $f_1(y_1)$ *and* $f_2(y_2)$.
(ii) *If* $\omega = 0$, *then* Y_1 *and* Y_2 *are independent*.
(iii) $E(Y_1Y_2) = \mu_1\mu_2 + \omega v_1 v_2$, *where* $v_t = E[Y_t\phi_t(Y_t)],\ t = 1, 2$.
(iv) $E(Y_2|Y_1 = y_1) = \mu_2 + \omega v_2\phi_1(y_1)$.
(v) $\rho = corr(Y_1, Y_2) = \frac{\omega v_1 v_2}{\sigma_1\sigma_2}$.

Proof Straightforward from (21). For more details refer to [17].

Remark 2 Let Y_1 and Y_2 be two random variables defined on $[0, \infty)$ with moment generating functions $M_t(u) = E\left(e^{uY_t}\right)$, $t = 1, 2$. Define $\phi_t(y_t) = e^{-y_t} - M_t(-1)$. Then $h(y_1, y_2)$ in (21) is a valid joint probability density (or mass) function.

Definition 2 (*Multivariate Sarmanov family of distributions*) Let $Y_1, Y_2, \ldots, Y_d$ be random variables with their probability density or mass functions $f_1, f_2, \ldots, f_d$. Let $\phi_t(y)$, $t = 1, 2, \ldots, d$ be bounded nonconstant functions such that $E[\phi_t(Y_t)] = 0$. Then the multivariate Sarmanov family is defined as

$$h(y_1, y_2, \ldots, y_d) = \left\{\prod_{t=1}^{d} f_t(y_t)\right\} \left(1 + R_{\phi_1,\phi_2,\ldots,\phi_d,\Omega_d}(y_1, y_2, \ldots, y_d)\right), \quad (y_1, y_2, \ldots, y_d) \in \mathbb{R}^d, \tag{22}$$

where $R_{\phi_1,\phi_2,\ldots,\phi_d,\Omega_d}(y_1, y_2, \ldots, y_d) = \sum_{j_1<j_2}^{d} \omega_{j_1,j_2}\phi_{j_1}\phi_{j_2} + \sum_{j_1<j_2<j_3}^{d} \omega_{j_1,j_2,j_3}\phi_{j_1}\phi_{j_2}\phi_{j_3}$ $+ \ldots + \omega_{1,2,\ldots,d}\prod_{t=1}^{d}\phi_t$ and Put space after and $\Omega_d = \left\{\omega_{j_1,j_2}, \omega_{j_1,j_2,j_3} \ldots, \omega_{1,2,\ldots,d}\right\}$ is chosen so that $R_{\phi_1,\phi_2,\ldots,\phi_d,\Omega_d}(y_1, y_2, \ldots, y_d) \geq -1$.

Theorem 1 can be extended to the multivariate case. For more details, see [17]. Next, bivariate and multivariate extensions of the *T*-geometric{*W*} family are proposed.

Bivariate and Multivariate T-*geometric{*W*} Families*

Let Y_1 and Y_2 follow the discrete *T*-*R*{*W*} family in (5) with PDF $f_1(y_1)$ and $f_2(y_2)$, respectively. Define $c_t = M_t(-1)$ where $M_t(u)$ is the moment generating function of Y_t, $t = 1, 2$. Then by Remark 2, the bivariate discrete *T*-*R*{*W*} family can be written as

$$h(y_1, y_2) = f_1(y_1)f_2(y_2)\{1 + \omega(e^{-y_1} - c_1)(e^{-y_2} - c_2)\}, \quad (y_1, y_2) \in \mathbb{N}_* \times \mathbb{N}_*. \tag{23}$$

Table 3 presents several examples of bivariate discrete *T*-*R*{*W*} family. In Table 3 $\delta(y_1, y_2) = \omega(e^{-y_1} - c_1)(e^{-y_2} - c_2)$.

A multivariate extension of the discrete *T*-*R*{*W*} family can be defined using (21). For simplicity, let $R_{\phi_1,\phi_2,\ldots,\phi_d,\Omega_d}(y_1, y_2, \ldots, y_d) = \sum_{j_1<j_2}^{d} \omega_{j_1,j_2}\phi_{j_1}\phi_{j_2}$, where $\phi_t(y_t) = e^{-y_t} - c_t$ and $c_t = M_t(-1)$, $t = 1, 2, \ldots, d$. Then, from (22) and (5), the multivariate

discrete T-$R\{W\}$ family is given by (24)

$$h(y_1, y_2, \ldots, y_d) = \prod_{t=1}^{d} [f_t(y_t)] \times \left\{ 1 + \sum_{t<v}^{d} \omega_{tv}(e^{-y_t} - c_t)(e^{-y_v} - c_v) \right\}, \quad (y_1, y_2, \ldots, y_d) \in \mathbb{N}_*^d. \tag{24}$$

Table 4 presents several examples of multivariate T-geometric$\{W\}$ family. In Table 4 $\delta(x_1, x_2, \ldots, x_d) = \sum_{t<v}^{d} \omega_{tv}(e^{-x_t} - c_t)(e^{-x_v} - c_v)$. Note that a multivariate version with more than pairwise associations between the response variables can be defined.

Multivariate T-*geometric{W} Regression Model*

Let the count response variable Y_{it}, $i = 1, 2, \ldots, n$; $t = 1, 2 \ldots, d$ where n is the sample size and let $\mathbf{x}_i = (x_{i0} = 1, x_{i1}, x_{i2}, \ldots, x_{i(k-1)})_i^T$ be a vector of $k - 1$ covariates. We assume that the same covariates affect each count response $\mathbf{Y}_t$, thus $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \ldots = \mathbf{x}_{id} = \mathbf{x}_i$. Assume that the conditional distribution of $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{id})^T$ for any given $\mathbf{x}_i = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{i(k-1)})^T$ follows the multivariate T-geometric$\{W\}$ family with $E(Y_{it}|x_{ij}) = \theta_t(x_{ij}) = \gamma(x_{ij}, \beta_{tj})$, $t = 1, 2, \ldots, d$, and $j = 0, 1, \ldots, (k-1)$, where $0 < \gamma(\mathbf{x}_i, \boldsymbol{\beta}_t) < 1$ is a known function of x_{ij}, and a vector $\boldsymbol{\beta}_t = (\beta_{t0}, \beta_{t1}, \ldots, \beta_{t(k-1)})$ of regression parameters. Now, let the function $\gamma(\mathbf{x}_i, \boldsymbol{\beta}_t)$ be the logit function. Therefore,

$$\theta_t(\mathbf{x}_i) = \theta_t = \gamma(\mathbf{x}_i, \boldsymbol{\beta}_t) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_t)}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}_t)}. \tag{25}$$

This implies that the bivariate and multivariate T-geometric$\{W\}$ regression models can be written, respectively, as

$$h(y_{i1}, y_{i2}|\mathbf{x}_i) = f_1(y_{i1}; \theta_1) f_2(y_{i2}; \theta_2)\{1 + \omega(e^{-y_{i1}} - c_1)(e^{-y_{i2}} - c_2)\}, \tag{26}$$

$$h(y_{i1}, y_{i2}, \ldots, y_{id}) = \prod_{t=1}^{d} [f_t(y_{it}; \theta_t)] \times \left\{ 1 + \sum_{t<v}^{d} \omega_{tv}(e^{-y_{it}} - c_t)(e^{-y_{iv}} - c_v) \right\}, \tag{27}$$

where θ_t is given in (25) and $f_t(y_{it}; \theta_t)$ follows T-geometric$\{W\}$ family for $t = 1, 2, \ldots, d$. In the next sections, we will discuss the statistical inference for the

Table 4 Multivariate T-geometric$\{W\}$ families based on different choices for variables W and T

Multivariate T-geometric$\{W\}$	Joint probability mass function $h(y_1, y_2, \ldots, y_d)$
(i) ME-geometric{Es}	$\prod_{t=1}^{d} \left[\theta_t^{y_t}(1-\theta_t)\right] \times \{1 + \delta(y_1, y_2, \ldots, y_d)\}$
(ii) ME-geometric{LLs}	$\prod_{t=1}^{d} \left[e^{-\lambda_t\left(\theta_t^{-y_t}-1\right)} - e^{-\lambda_t\left(\theta^{-(y_t+1)}-1\right)}\right] \times \{1 + \delta(y_1, y_2, \ldots, y_d)\}$
(iii) MEE-geometric{Es}	$\prod_{t=1}^{d} \left[\left(1-\theta_t^{y_t+1}\right)^{\alpha_t} - \left(1-\theta_t^{y_t}\right)^{\alpha_t}\right] \times \{1 + \delta(y_1, y_2, \ldots, y_d)\}$
(iv) MEE-geometric{LLs}	$\prod_{t=1}^{d} \left[\left(1-e^{-\lambda_t\left(\theta_t^{-(y_t+1)}-1\right)}\right)^{\alpha_t} - \left(1-e^{-\lambda_t\left(\theta_t^{-y_t}-1\right)}\right)^{\alpha_t}\right] \times \{1 + \delta(y_1, y_2, \ldots, y_d)\}$
(v) MW-geometric{Es}	$\prod_{t=1}^{d} \left[\theta_t^{(y_t+1)^{c_t}} - \theta_t^{y_t c_t}\right] \times \{1 + \delta(y_1, y_2, \ldots, y_d)\}$
(vi) MW-geometric{LLs}	$\prod_{t=1}^{d} \left[e^{-\left(\lambda_t(\theta_t^{-y_t}-1)\right)^{c_t}} - e^{-\left(\lambda_t(\theta_t^{-(y_t+1)}-1)\right)^{c_t}}\right] \times \{1 + \delta(y_1, y_2, \ldots, y_d)\}$
(vii) ML-geometric{Ls}	$\prod_{t=1}^{d} \left[\frac{\left(1-\theta_t^{y_t+1}\right)^{\lambda_t}}{\left(1-\theta_t^{y_t+1}\right)^{\lambda_t} + \theta_t^{\lambda_t(y_t+1)}} - \frac{\left(1-\theta_t^{y_t}\right)^{\lambda_t}}{\left(1-\theta_t^{y_t}\right)^{\lambda_t} + \theta_t^{\lambda_t y_t}}\right] \{1 + \delta(y_1, \ldots, y_d)\}$

bivariate T-geometric$\{W\}$ regression model and application to real-life data sets. All results in these sections can be extended to the multivariate case in (27).

4 Inference on Bivariate and Multivariate T-geometric$\{W\}$ Regression Models

The maximum likelihood estimation method can be used to estimate the unknown parameters in the bivariate T-geometric$\{W\}$ regression model in (26). If a random sample of size n is taken from the bivariate T-geometric$\{W\}$ regression model in (26), then the log-likelihood function is given by

$$\log L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \log h(y_{i1}, y_{i2}|\mathbf{x}_i)$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{2} \log f_t(y_{it}; \theta_t) + \sum_{i=1}^{n} \log\{1 + \omega(e^{-y_{i1}} - c_1)(e^{-y_{i2}} - c_2)\}, \quad (28)$$

where $\boldsymbol{\beta}$ is the vector of all unknown regression parameters and $\boldsymbol{\alpha}$ is the vector of parameters from the marginal PMFs in $h(y_{i1}, y_{i2}|\mathbf{x}_i)$. In next sub-sections, we discuss some statistical tests and goodness-of-fit measures.

By using the discrete exponentiated-exponential-geometric distribution in (11) with φ^{λ} set to parameter θ, we obtain the multivariate exponentiated-exponential-geometric regression (MEEGR) model as (29)

$$P(y_1, y_2, \ldots, y_d) = \prod_{t=1}^{d}\left[(1-\theta_t^{y_t+1})^{b_t} - (1-\theta_t^{y_t})^{b_t}\right]$$

$$\left[1 + \sum_{t<v}^{d} \omega_{tv}(e^{-y_t} - c_t)(e^{-y_v} - c_v)\right] \quad (29)$$

where $\theta_t = \theta_t(x_i) = \exp(\beta_{t0}+\beta_{t1}x_{i1})/[1+\exp(\beta_{t0}+\beta_{t1}x_{i1})] = 1/\{1+\exp[-(\beta_{t0}+\beta_{t1}x_{i1})]\}$. The marginal PMF for each y_t is $P(y_t) = (1-\theta_t^{y_t+1})^{b_t} - (1-\theta_t^{y_t})^{b_t}$. Suppose $d = 3$, the conditional PMF of y_2 given y_1 is

$$P(y_2|y_1) = [(1-\theta_2^{y_2+1})^{b_2} - (1-\theta_2^{y_2})^{b_2}][1 + \omega_{12}(e^{-y_1} - c_1)(e^{-y_2} - c_2)]/y_2! \quad (30)$$

The conditional PMF of y_3 given y_1 and y_2 is given by (31)

$$P(y_3|y_1, y_2) = [(1-\theta_3^{y_3+1})^{b_3} - (1-\theta_3^{y_3})^{b_3}]$$

$$[1 + A/\{1 + \omega_{12}(e^{-y_1} - c_1)(e^{-y_2} - c_2)\}], \tag{31}$$

where $A = \omega_{13}(e^{-y_1} - c_1)(e^{-y_3} - c_3) + \omega_{23}(e^{-y_2} - c_2)(e^{-y_3} - c_3)$. These results may be used when simulating from the multivariate MEEG distribution. Other multivariate generalized geometric distributions can be derived by using the different methods suggested in Table 4. When the dispersion parameters $b_t, t = 1, 2, \ldots, d$ are all set to 1, the MEEGR in (29) reduces to the multivariate geometric regression (MGR) model. The initial estimates for the MEEGR model can be taken as the final estimates from the univariate EEGR models for the responses. The initial estimates of the correlations can be the sample correlations.

Test for Independence

A pair of random variables Y_t and Y_ν are independent when the parameter $\omega_{t\nu}$ (or $\rho_{t\nu}$) is equal to zero. For a multivariate regression model $h(y_{i1}, y_{i2}, \ldots, y_{id})$, we test the null hypothesis

$H_0 : \omega_{t\nu} = 0$ against $H_a : \omega_{t\nu} \neq 0$, where $t < \nu$, $t = 1, 2, 3, \ldots, d$, and $\nu = 2, 3, \ldots, d$.

Suppose L_{ind} is the likelihood function under H_0 and L_a is the likelihood function under H_a. The test statistic is given by $\chi^2_{ind} = -2\log(L_{ind}/L_a)$, which has a chi-square distribution with $d(d-1)/2$ degrees of freedom.

Test for Dispersion

The MEEGR model in (29) reduces to the MGR model when the dispersion parameters $b_t = 1$, for $t = 1, 2, \ldots, d$. To carry out this test, we test if the $b_t = 1$. This gives a situation where all the response variables are equi-dispersed. Thus, we test the hypotheses.

$H_0 : b_t = 1$ for $t = 1, 2, \ldots, d$ against $H_a : H_0$ is not true.

If L_{dis} is the likelihood function when H_0 is true and L_a is the likelihood function when H_0 is false, the test statistic is given by $\chi^2_{dis} = -2\log(L_{dis}/L_a)$, which has an approximate chi-square distribution with d degrees of freedom.

Test to Compare Nested and Non-nested Models

In this sub-section, we use the notation $\underline{y}_i = y_{i1}, y_{i2}, \ldots, y_{id}$ and compare the model $f(\underline{y}_i)$ (e.g., MEEGR) and model $g(\underline{y}_i)$ (e.g., MGPR) defined by [10]. Given the two regression models, we consider the following hypotheses: (33) and (34).

$$H_0 :, \text{Model } f(\underline{y}_i) \text{ and Model } g(\underline{y}_i) \text{ are equivalent}$$

against

$$H_f : f(\underline{y}_i) \text{ is better than } f(\underline{y}_i) \text{ or } H_g : g(\underline{y}_i) \text{ is better than } f(\underline{y}_i). \tag{32}$$

The likelihood ratio statistic for testing model $f(\underline{y}_i)$ against model $g(\underline{y}_i)$ is defined as

$$L_* = \sum_{i=1}^{n} \log\left(\frac{f(\underline{y}_i)}{g(\underline{y}_i)}\right). \tag{33}$$

If the two models $f(\underline{y}_i)$ and $g(\underline{y}_i)$ are nested, the statistic in (33) has a chi-square distribution. If the two models are not nested, the statistic in (33) is not chi-square distributed.

Vuong [24] used the Kullback-Leibler information criterion to discriminate between two non-nested models. To test the null hypothesis H_0 in (32), Vuong proposed the test statistic

$$Z_* = \frac{L_*}{\hat{w}\sqrt{n}}, \text{ where } w^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{f(\underline{y}_i)}{g(\underline{y}_i)}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(\underline{y}_i)}{g(\underline{y}_i)}\right)\right]^2 \tag{34}$$

is an estimate of the variance of $L_*/\sqrt{n}$. For a non-nested model, Vuong showed that Z_* is approximately standard normal distributed under H_0 in (32). At significant level α, H_0 is rejected in favor of H_g if $Z_* < -z_{\alpha/2}$, and H_0 is rejected in favor of H_f if $Z_* > z_{\alpha/2}$. If $|Z_*| \leq z_{\alpha/2}$, we fail to reject H_0. Thus, both models are equivalent.

Goodness-Of-Fit Statistics

The log-likelihood defined in (26) for the bivariate regression model can be used as a goodness-of-fit statistics for the model. Other alternative measures of goodness of fit are the AIC and the BIC. These statistics are based on the log-likelihood statistic and they are defined as follows: AIC $= -2LL + 2p$, BIC $= -2LL + p\log(n)$, where n is the sample size, p is the number of estimated parameters, and LL is the log-likelihood statistic. Both the AIC and BIC are better than the log-likelihood because they consider the number of parameters in the regression model to control over-parametrization. These measures are usually provided by the SAS NLMIXED procedure. The smaller the measures, the better the model.

5 Application

In this section, we illustrate the applications of MEEGR model to two real-life data sets. The results from MEEGR are compared with the results from the MGPR model. The first data set is on the number of sex partners and the second data set is on profiling inmates. For each application, we assume the marginal mean of each Y_{it} in the MGPR model has a log-linear relationship with the covariates x_{itj} ($j = 1, 2, \ldots, k$) through.

$\log[E(Y_{it})] = \beta_{t0} + \beta_{t1}x_{it1} + \cdots + \beta_{tk}x_{itk}$, where $i = 1, 2, \ldots, n$ and we have k covariates.

The probability $\theta_t = \theta_t(x_{itj})$ in the MEEGR is defined by the logit function.

$\theta_t(x_{it}) = 1/\{1 + \exp[-(\beta_{t0} + \beta_{t1}x_{it1} + \cdots + \beta_{tk}x_{itk})]\}$, where $i = 1, 2, \ldots, n$ for k covariates.

Sex Partners Data

The number of sex partners data was used by [10] to illustrate the multivariate generalized Poisson regression model. See the reference for more information about the data. [10] considered the following three response variables: y_1 = number of male sex partners in the past year, y_2 = number of male oral sex partners in the past year, and y_3 = number of sex partners in the past year who are 5 years older than the respondent. The descriptive statistics and the pairwise correlations between the response variables are presented in Table 5. Both variables y_1 and y_2 are under-dispersed since the sample variances are smaller than their corresponding sample means. The predictor variables used by [10] were race (1 = white, 0 = others), educational level (range from 1 = below 9th grade to 5 = college graduate), marital status (1 = married, 0 = others), and age.

The data is fitted to the MEEGR model and the model fit is compared with that of MGPR model. The results are presented in Table 6. The log-likelihood values from the MEEGR and MGPR models are, respectively, –2926.21 and –3532.24. We compared the MEEGR model with the MGPR by using the Vuong statistics in (34) and obtained the test statistic Z_* in Sub-Sect. 4.2 to be 23.24 (p-value < 0.0001), which is highly significant. This indicates that MEEGR model provides

Table 5 Pairwise correlations, mean, and variance for response variables in sex partners' data

	Correlation		Mean	Variance
	Y_2	Y_3		
Y_1	0.1345*	0.3581*	1.3133	1.0847
Y_2	–	0.1078*	0.3719	0.3495
Y_3	–	–	0.3672	0.4108

*Significant at 5% level (n = 1280)

a better fit than the MGPR model. The effects of the explanatory variables on the response variables for MGPR model are the same for the MEEGR model. Thus, the response variables y_1 and y_3 are negatively affected by marital status and age. The response variable y_2 is negatively affected by race, education, and age. The older respondents tend to have smaller number of male sex partners, smaller number of oral sex partners, and smaller number of sex partners 5 years older during the past year. On the relationship between the response variables, ordinary correlation shows that they are all positively correlated and significant. The MGPR and the MEEGR models show that variables y_1 and y_3 have significant positive correlation.

Table 6 Parameter estimates (standard errors in parentheses) for sex partners' data

	y_1		Y_2	
Variable	MEEGR model	MGPR model	MEEGR model	MGPR model
Constant (x_0)	−1.515 (0.174)*	0.735 (0.108)*	– 1.271 (0.263)*	0.232 (0.206)
Race (x_1)	0.006 (0.088)	0.005 (0.049)	– 0.478 (0.105)*	– 0.433 (0.104)*
Education (x_2)	0.031 (0.033)	0.022 (0.021)	– 0.141 (0.040)*	– 0.121 (0.039)*
Marital (x_3)	−0.577 (0.097)	– 0.383 (0.050)*	– 0.041 (0.103)	– 0.070 (0.103)
Age (x_4)	−0.013 (0.003)*	– 0.009 (0.002)*	– 0.020 (0.005)*	– 0.018 (0.005)*
$\hat{b}_t$	32.411 (3.962)*	– 0.028 (0.007)*	6.454 (1.143)*	– 0.044 (0.027)
	Y_3			
Variable	MEEGR model	MGPR model		
Constant (x_0)	– 0.931 (0.267)*	– 0.160 (0.267)		
Race (x_1)	0.018 (0.111)	0.021 (0.120)		
Education (x_2)	– 0.051 (0.045)	– 0.047 (0.051)		
Marital (x_3)	– 0.646 (0.115)*	– 0.774 (0.132)*		
Age (x_4)	– 0.014 (0.005)*	– 0.017 (0.006)*		
$\hat{b}_t$	2.831 (0.402)*	0.142 (0.062)*		
$\hat{\rho}_{tv}$ for MEEGR	$\hat{\rho}_{12} = 0.0272$ (0.0360)			
	$\hat{\rho}_{13} = 0.2552$ (0.0378)*			
	$\hat{\rho}_{23} = 0.0241$ (0.0282)			
$\hat{\rho}_{tv}$ for MGPR	$\hat{\rho}_{12} = 0.1009$ (0.0735)			
	$\hat{\rho}_{13} = 0.5574$ (0.0843)*			
	$\hat{\rho}_{23} = 0.0352$ (0.0330)			
Log-likelihood	For MEEGR: –2926.21			
	For MGPR: –3520.63			

*Significant at 5% level

Inmates Profiling Data

By 1996, the Los Angeles (LA) County jails faced a serious overcrowding problem. Two suggested solutions to the problem were to build more jail capacity or to divert a greater number of incoming inmates to community-based, intermediate sanctions. A research team was formed to review a profile of inmates in the jail system and to determine how many of them were suitable candidates for intermediate sanctions. A random sample of 1000 un-convicted inmates were selected. Demographic variables include age, race, current offense severity, and gender. The sample size for the collected data set is 931. Among the other variables collected are prior arrests (y_1), number of subsequent arrests (y_2), number of LA custodies (y_3), and number of CDC custodies (y_4). See the reference [20] for information about the data.

In the collected data, the race variable consists of Black, Hispanic, White, and others with a frequency of 30. We define two dummy variables x_4 and x_5 to represent the groups Black and Hispanic, respectively. Thus, the White group is the reference category. The other explanatory variables are x_1, x_2, and x_3 representing age, rank (current offense severity), and gender (where male is represented by 1), respectively. After eliminating any missing case for the four response variables and the five explanatory variables, we are left with a total of $n = 781$. In this sample, the percentages of male, black, Hispanic, and white are, respectively, 94, 40, 45, and 15%, respectively. Table 7 contains the descriptive statistics for the response variables and their bivariate correlations.

Both the MEEGR and MGPR models are fitted to the data set and the results are presented in Table 8. From the log-likelihood statistics, the MEEGR model performed better than the MGPR model. We carried out the test in Eq. (33) and we obtained the test statistic $Z_* = 10.46$ with p-value <0.0001. Thus, the MEEGR outperforms the MGPR in fitting the data. All the results about the dispersion parameters are the same for both regression models. They showed that all response variables are over-dispersed. Both models show that the number of arrests is positively affected by age and there is a significant difference between Black and White. The number of subsequent arrests is positively associated with the current offence severity in both models. Only the MEEGR model showed that female inmates have more subsequent arrests than male inmates. While age and Black positively influence number

Table 7 Pairwise correlation, mean, and variance of response variables in inmates profiling data

	Correlation			Mean	Variance
	Y_2	Y_3	Y_4		
Y_1	0.0608	0.4993*	0.41421*	10.9373	85.4614
Y_2	–	0.1200*	−0.0903*	0.7414	1.2458
Y_3	–	–	0.2319*	4.6095	23.0024
Y_4	–	–	–	2.2228	6.3477

*Significant at 5% level ($n = 781$)

of LA custodies, the current offence severity negatively influences the number of LA custodies in both models. On the other hand, while there is a difference between Hispanic and White in the number of LA custodies under MEEGR model, there is no such difference under the MGPR model. The results from both models are about the same for the explanatory variables for the number of CDC custodies.

From Table 7, all bivariate correlations are significant at 5% level with the exception between y_1 and y_2. All significant correlations are positive with the exception of correlation between variables y_2 and y_4. Table 7 shows that all pairwise correlations are significant under the MEEGR model with the exception of correlation between variables y_2 and y_3. Both models give only one significant negative correlation between y_2 and y_4. Also, both models give the correlation between y_2 and y_3 to be insignificant.

Table 8 Parameter estimates (standard errors in parentheses) for inmate profiling data

Variable	y_1		y_2	
	MEEGR model	MGPR model	MEEGR model	MGPR model
Constant (x_0)	0.428 (0.254)*	0.910 (0.221)*	0.401 (0.364)	0.130 (0.343)
Age (x_1)	0.039 (0.004)*	0.029 (0.004)*	– 0.010 (0.006)	– 0.003 (0.006)
Rank (x_2)	0.0002 (0.003)	0.001 (0.003)	0.015 (0.004)*	0.013 (0.004)*
Gender (x_3)	0.159 (0.140)	0.175 (0.134)	– 0.537 (0.213)*	– 0.392 (0.206)
Black (x_4)	0.315 (0.096)*	0.428 (0.099)*	– 0.132 (0.173)	– 0.059 (0.165)
Hispanic (x_5)	0.095 (0.098)	0.150 (0.098)	– 0.056 (0.172)	– 0.109 (0.163)
$\hat{b}_t$	1.508 (0.103)*	0.178 (0.009)*	0.889 (0.091)*	0.373 (0.058)*
–	y_3		y_4	
–	MEEGR model	MGPR model	MEEGR model	MGPR model
Constant (x_0)	0.640 (0.168)*	0.555 (0.227)*	– 0.769 (0.258)*	– 0.528 (0.270) +
Age (x_1)	0.008 (0.003)*	0.014 (0.004)*	0.011 (0.004)*	0.014 (0.005)*
Rank (x_2)	– 0.007 (0.003)*	– 0.008 (0.003)*	0.001 (0.003)	0.002 (0.003)
Gender (x_3)	0.115 (0.085)	0.262 (0.143)	0.676 (0.175)*	0.740 (0.174)*
Black (x_4)	0.394 (0.081)*	0.551 (0.110)*	0.310 (0.115)*	0.280 (0.123)*
Hispanic (x_5)	0.297 (0.085)*	0.165 (0.109)	– 0.013 (0.117)	– 0.112 (0.120)
$\hat{b}_t$	1.214 (0.053)*	0.235 (0.014)*	1.439 (0.094)*	0.240 (0.021)*
$\hat{\rho}_{tv}$ for MEEGR	$\hat{\rho}_{12} = 0.1829$ (0.0218)*; $\hat{\rho}_{13} = 0.7146$ (0.0406)*; $\hat{\rho}_{14} = 0.2684$ (0.0269)*			
	$\hat{\rho}_{23} = 0.0227$ (0.0436); $\hat{\rho}_{24} = -0.1049$ (0.0215)*; $\hat{\rho}_{34} = 0.2605$ (0.0265)*			
$\hat{\rho}_{tv}$ for MGPR	$\hat{\rho}_{12} = 0.0314$ (0.0066)*; $\hat{\rho}_{13} = 0.0177$ (0.0105); $\hat{\rho}_{14} = 0.0380$ (0.0130)*			
	$\hat{\rho}_{23} = -0.0452$ (0.0304); $\hat{\rho}_{24} = -0.1909$ (0.0236)*; $\hat{\rho}_{34} = 0.1402$ (0.0294)*			
Log-likelihood	For MEEGR: –6779.20 For MGPR: –6978.17			

*Significant at 5% level

6 Summary and Conclusions

Famoye [10] stated some advantages of the MGPR model when compared to the multivariate negative binomial regression model and the multivariate Poisson-lognormal regression model. Only the MGPR model is capable of modeling both over-dispersion and under-dispersion. The other two models are applicable to over-dispersed response count data. The MGPR model allows for correlations of any sign among count response variables and the correlations are independent of the dispersion parameters. Several multivariate models that can be used to fit several dependent response variables are proposed in this paper. Many of these multivariate models can be used for over-dispersed or under-dispersed count response variables. The models allow correlations of any sign.

In this paper, we define families of bivariate and multivariate generalized geometric regression models. These are based on the T-$R\{W\}$ framework. Many examples are given in the paper. The population means and variances of the response variables are not in closed form, but they are easy to compute because the CDF of the distributions is in closed form. It is easy to compute both the CDF and PMF for the distributions.

Two numerical data sets are used to illustrate the MEEGR model and it is compared with the MGPR model (see [10]). In both data sets, the MEEGR model outperformed the MGPR model by using the goodness-of-fit test. The MEEGR model has closed-form PMF and CDF while the MGPR model has closed-form PMF and its mean and variance are in closed form. Both models can be used to fit data that exhibit over- or under-dispersion. The MGPR has a negative dispersion parameter for the under-dispersed situation while the dispersion parameter for the MEEGR model is always positive. Care must be taken when estimating the MGPR model for under-dispersed situation because the model gets truncated above, (see [8], p. 4 or [7], p. 165) about the truncation error in the generalized Poisson distribution. This error is negligible when the number of nonzero probability classes is at least 5. In both numerical examples in Sect. 5, the response variable with the smallest range has seven nonzero probability classes.

In Sect. 4, we provide materials that can be used to compare multivariate count data regression models. This is illustrated in the application section (Sect. 5), where two generalized multivariate distributions and models are compared, and variables are selected using some of the test statistics described in Sect. 4. In this chapter, a family of multivariate count data regression models is defined. Future work can include further inferences and a comparison between some of the models that were defined in the chapter. New models can be defined and compared to the models in the chapter.

7 Appendix

In this Appendix, we will use the following notation: $\lfloor y \rfloor = \dot{y}$.

Table 2(i): Exponential-geometric{standard exponential}.
By using Eqs. (6) and (8), we have

$$G(y) = F_Y(y) = F_T\{\lambda Q_W(F_R(\dot{y})\} = 1 - \exp\{\lambda \log(1 - F_R(\dot{y}))\}$$
$$= 1 - (1 - F_R(\dot{y}))^{\lambda} = 1 - \varphi^{\lambda(\dot{y}+1)} = 1 - \theta^{\dot{y}+1} \text{ where } 0 < \theta = \varphi^{\lambda} < 1.$$

Table 2(iii): Exponentiated-exponential-geometric{standard exponential}.
By using Eqs. (10) and (12), we have

$$G(y) = F_Y(y) = F_T[Q_W(F_R(\dot{y})] = \left[1 - \exp\{\lambda \log(1 - F_R(\dot{y}))\}\right]^{\alpha}$$
$$= \left[1 - \exp\{\log \varphi^{\lambda(\dot{y}+1)}\}\right]^{\alpha} = \left[1 - (\exp[\lambda \log \varphi])^{\dot{y}+1}\right]^{\alpha}$$
$$= \left[1 - \theta^{\dot{y}+1}\right]^{\alpha} \text{ where } 0 < \theta = \exp[\lambda \log \varphi] < 1,\ \alpha > 0.$$

Table 2(iv): Exponentiated-exponential-geometric{standard log-logistic}.
By using Eq. (13), we have

$$G(y) = F_Y(y) = F_T[Q_W(F_R(\dot{y})] = \left[1 - \exp\{-\lambda F_R(\dot{y})/[1 - F_R(\dot{y})]\}\right]^{\alpha}$$
$$= \left[1 - \exp\left\{-\lambda(1 - \varphi^{\dot{y}+1})/\varphi^{\dot{y}+1}\right\}\right]^{\alpha} = \left[1 - \exp\left\{-\lambda(\theta^{-\dot{y}-1} - 1)\right\}\right]^{\alpha},$$

where $0 < \theta = \varphi < 1,\ \lambda > 0,\ \alpha > 0$.

Table 2(v): Weibull-geometric{standard exponential}.
By using Eq. (16), we have

$$G(y) = F_Y(y) = F_T[\lambda Q_W(F_R(\dot{y})] = 1 - \exp\{-\lambda F_R(\dot{y})/[1 - F_R(\dot{y})]\}$$
$$= 1 - \exp\left\{-(\lambda Q_W(F_R(\dot{y})))^c\right\} = 1 - \exp\left\{-(-\lambda \log[1 - F_R(\dot{y})])^c\right\}$$
$$= 1 - \exp\left\{-\left(-\lambda \log \varphi^{\dot{y}+1}\right)^c\right\} = 1 - \exp\left\{-([-\lambda \log \varphi](\dot{y} + 1))^c\right\}$$
$$= 1 - \exp\left\{-(-\lambda \log \varphi)(\dot{y} + 1)^c\right\} = 1 - \theta^{(\dot{y}+1)^c},$$
$$\text{where } 0 < \theta = e^{-(-\lambda \log \varphi)} < 1,\ c > 0.$$

Table 2(vii): Logistic-geometric{standard logistic}.
By using Eq. (18), we have

$$G(y) = F_Y(y) = F_T[\lambda Q_W(F_R(\dot{y})] = \left[1 + \exp\{-\lambda Q_W(F_R(\dot{y}))]\}\right]^{-1}$$
$$= \left[1 + \exp\left\{\log(F_R(\dot{y})/(1 - F_R(\dot{y}))^{-\lambda}\right\}\right]^{-1}$$
$$= \left[1 + (F_R(\dot{y})/(1 - F_R(\dot{y}))^{-\lambda}\right]^{-1}$$
$$= F_R^{\lambda}(\dot{y}) \Big/ \left[F_R^{\lambda}(\dot{y}) + \left(1 - F_R^{\lambda}(\dot{y})\right)^{\lambda}\right]$$

$$= \left(1-\varphi^{\dot{y}+1}\right)^{\lambda} \Big/ \left[\left(1-\varphi^{\dot{y}+1}\right)^{\lambda} + \varphi^{\lambda(\dot{y}+1)}\right], \text{ where } 0 < \theta = \varphi < 1,\ \lambda > 0.$$

Acknowledgements The authors are grateful to the Editor and the Reviewer. Their comments and suggestions have greatly improved the chapter.

References

1. Aldeni, M., Lee, C. & Famoye, F. (2017). Families of distributions arising from the quantile of generalized lambda distribution. *Journal of Statistical Distributions and Applications*, *4*(25).
2. Aljarrah, M., Lee, C. & Famoye, F. (2014). On generating *T-X* family of distributions using quantile functions. *Journal of Statistical Distributions and Applications*, *1*(2).
3. Alzaatreh, A., Lee, C. & Famoye, F. (2014). *T*-normal family of distributions: A new approach to generalize the normal distribution. *Journal of Statistical Distributions and Applications*, *1*(16).
4. Alzaatreh, A., Lee, C., & Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron, 71*(1), 63–79.
5. Alzaatreh, A., Lee, C., & Famoye, F. (2012). On the discrete analogues of continuous distributions. *Statistical Methodology, 9*, 589–603.
6. Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.
7. Consul, P.C. & Famoye, F. (2006). *Lagrangian probability distributions*. Boston: Birkhäuser Inc.
8. Consul, P.C. (1989). *Generalized Poisson distributions, 99*, Marcel Dekker Inc.
9. Famoye, F. (2010). On the bivariate negative binomial regression model. *Journal of Applied Statistics, 37*(6), 969–981.
10. Famoye, F. (2015). A multivariate generalized poisson regression model. *Communications in Statistics-Theory and Methods, 44*, 497–511.
11. Famoye, F. (2019). Bivariate exponentiated-exponential geometric regression model. *Statistica Neerlandica, 73*, 434–450.
12. Gupta, R. D., & Kundu, D. (2001). Exponentiated-exponential family: An alternative to gamma and weibull distributions. *Biometrical Journal, 43*, 117–130.
13. Hofer, V., & Leitner, J. (2012). A bivariate sarmanov regression model for count data with generalized poisson marginal. *Journal of Applied Statistics, 39*(12), 2599–2617.
14. Kocherlakota, S., & Kocherlakota, K. (1992). *Bivariate discrete distributions*. Marcel Dekker Inc.
15. Lee, C., Famoye, F., & Akinsete, A. (2021). Generalized count data regression models and their applications to health care data. *Annals of Data Science, 8*(2), 367–386.
16. Lee, C., Famoye, F., & Alzaatreh, A. (2013). Methods for generating families of univariate continuous distributions in the recent decades. *WIREs Computational Statistics, 5*, 219–238.
17. Lee, M.-L.T. (1996). Properties and applications of the sarmanov family of bivariate distributions. *Communications in Statistics-Theory and Methods, 25*, 1207–1222.
18. Nikoloulopoulos, A.K. (2013). A survey on multivariate copula-based models for multivariate discrete response data. In P. Jaworski, F. Durante & W. K. Härdle (Eds.), *Copulae in Mathematical and Quantitative Finance, Lecture Notes in Statistics: Proceedings of the Workshop Held in Cracow, 10−11 July 2012* (pp. 231−249). Berlin, Heidelberg: Springer.
19. Nikoloulopoulos, A. K., & Karlis, D. (2010). Modeling multivariate count data using copulas. *Communications in Statistics-Simulation and Computation, 39*, 172–187.
20. Petersilia, J., Turner, S., & Fain, T. (1996). Profiling inmates in the Los Aangeles county Jail, 1996–1998. [distributor], 2006–03–30. https://doi.org/10.3886/ICPSR03271.v2.

21. Sarmanov, O. V. (1966). Generalized normal correlation and two-dimensional Fréchet classes. *Doklady Akademii Nauk, 168*, 32–35.
22. Sellers, K. F., & Shmueli, G. (2010). A flexible regression model for count data. *Annals of Applied Statistics, 4*(2), 943–961.
23. Sun, S. Z., & Ong, S. H. (2016). A generalized inverse trinomial distribution with application. *Statistical Methodology, 33*, 217–233.
24. Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*(2), 307–333.
25. Winkelmann, R. (2008). *Econometric analysis of count data* (5th ed.). Springer.

A Generalized Multivariate Gamma Distribution

Anis Iranmanesh, Maryam Rafiei, and Daya Krishna Nagar

Abstract In this chapter, we introduce a multivariate gamma distribution whose marginals are finite mixtures of gamma distributions and correlation between any pair of variables is negative. Several of its properties such as joint moments, correlation coefficients, moment generating function, Rényi and Shannon entropies have been derived. Simulation study have been conducted to evaluate the performance of the maximum likelihood method.

1 Introduction

Gamma distribution is an important continuous distribution in probability and statistics. Several distributions such as exponential, Erlang, and chi-square are special cases of this distribution. Several univariate generalizations of gamma distribution have also been studied. Gamma distribution and its variants have been applied in different disciplines to model continuous variables that are positive and have skewed distributions. Gamma distribution has been used to model amounts of daily rainfall (Aksoy [1]) and in neuroscience this model is often used to describe the distribution of inter-spike intervals (Robson and Troy [26]). The gamma distribution is widely used as a conjugate prior in Bayesian statistics. It also plays an important role in actuarial sciences (Furman [9]).

Several multivariate generalizations of univariate gamma distributions are also available in the literature. Mathai and Moschopoulos [20, 21] introduced two multivariate gamma models as the joint distribution of certain linear combinations/partial sums of independent three parameter gamma variables. All the components of their multivariate gamma vectors are positively correlated and have three parameter

A. Iranmanesh (✉) · M. Rafiei
Department of Mathematics and Statistics, Mashhad Branch, Islamic Azad University, Mashhad, Iran
e-mail: iranmanesh0030@mshdiau.ac.ir; anisiranmanesh@yahoo.com

D. K. Nagar
Instituto de Matemáticas, Universidad de Antioquia, Medellín, Colombia

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_12

gamma distributions. They have also indicated that their models have potential applications in stochastic processes and reliability. Furman [9] used the multivariate reduction technique to derive a multivariate probability model possessing a dependence structure and gamma marginals. Kowalczyk and Tyrcha [17] used a re-parameterized form of the gamma distribution to define a multivariate gamma vector and studied a number of properties of their distribution. Recently, Semenikhine, Furman and Su [28] introduced a multiplicative multivariate gamma distribution with gamma marginals and applied their results in actuarial science. They proved that the correlation coefficient between any pair of variables is positive and belongs to $(0, 1/2)$. Multivariate gamma distributions have been used in diverse fields like hydrology, space (wind modeling), reliability, traffic modeling, and finance. For further results on multivariate gamma distribution, the reader may consult articles by Balakrishnan and Ristić [4], Carpenter and Diawara [5], Dussauchoy and Berland [6], Gaver [10], Krishnaiah and Rao [18], Marcus [19], Pepas et al.[23], Royen [27], Vaidyanathan and Lakshmi [33], and an excellent text by Kotz, Balakrishnan and Johnson [16]. For a good review on bivariate gamma distributions, see Balakrishnan and Lai [3], Arnold, Castillo and Sarabia [2], Hutchinson and Lai [14], and Kotz, Balakrishnan and Johnson [16]. For a review on some recent work and applications the reader is referred to Rafiei, Iranmanesh, and Nagar [24] and references therein.

In this chapter, we introduce a multivariate gamma distribution whose marginals are finite mixtures of gamma distributions and correlation between any pair of variables is negative. We organize our work as follows: In Sect. 2, we introduce the new multivariate gamma distribution. In Sects. 3 and 4, results on marginal distributions and factorizations of the multivariate gamma distribution are derived. Sections 5–8 deal with properties such as joint moments, correlation, moment generating function, entropies and estimations. In Sect. 9, simulations of the new distribution are performed in different ways, and the results are provided to evaluate the performance of the maximum likelihood method. Section 10 contains the conclusion. Finally, the Appendix lists a number of results used in this chapter.

2 The Multivariate Gamma Distribution

Recently, Rafiei, Iranmanesh, and Nagar [24] have defined a bivariate gamma distribution with parameters α, β and k and the pdf

$$f(x_1, x_2; \alpha, \beta, k) = \frac{\Gamma(2\alpha)}{\beta^{2\alpha+k}\Gamma^2(\alpha)\,\Gamma(2\alpha+k)}(x_1x_2)^{\alpha-1}(x_1+x_2)^k \exp\left[-\frac{1}{\beta}(x_1+x_2)\right],$$

where $x_1 > 0, x_2 > 0, \alpha > 0, \beta > 0$, and $k \in \mathbb{N}_0$. A natural multivariate generalization of this distribution can be given as follows.

Definition 1 The random variables $X_1, \ldots, X_n$ are said to have a generalized multivariate gamma distribution, denoted as $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$, if their joint pdf is given by

$$f(x_1, \ldots, x_n; \alpha_1, \ldots, \alpha_n; \beta, k) = C(\alpha_1, \ldots, \alpha_n; \beta, k) \prod_{i=1}^{n} x_i^{\alpha_i - 1} \left(\sum_{i=1}^{n} x_i \right)^k \times \exp\left(-\frac{1}{\beta} \sum_{i=1}^{n} x_i \right), \; x_i > 0, \; i = 1, \ldots, n, \quad (1)$$

where $\alpha_1 > 0, \ldots, \alpha_n > 0, \beta > 0, k \in \mathbb{N}_0$ and $C(\alpha_1, \ldots, \alpha_n; \beta, k)$ is the normalizing constant.

By integrating the joint density of $X_1, \ldots, X_n$ over its support set, the normalizing constant is derived as

$$\begin{aligned} [C(\alpha_1, \ldots, \alpha_n; \beta, k)]^{-1} &= \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{n} x_i^{\alpha_i - 1} \left(\sum_{i=1}^{n} x_i \right)^k \exp\left(-\frac{1}{\beta} \sum_{i=1}^{n} x_i \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \beta^{\sum_{i=1}^{n} \alpha_i + k} \left[\prod_{i=1}^{n} \Gamma(\alpha_i) \right] (\alpha_1 + \cdots + \alpha_n)_k \,, \end{aligned}$$

where the last line has been obtained by using Lemma 2. Finally, from the above expression

$$C(\alpha_1, \ldots, \alpha_n; \beta, k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\beta^{\sum_{i=1}^{n} \alpha_i + k} \left[\prod_{i=1}^{n} \Gamma(\alpha_i) \right] \Gamma(\alpha_1 + \cdots + \alpha_n + k)}. \quad (2)$$

For $k = 0$, the multivariate gamma density simplifies to the product of n independent univariate gamma densities with common scale parameter β. For $k = 1$, the multivariate gamma density can be written as

$$\sum_{j=1}^{n} \left(\frac{\alpha_j}{\sum_{i=1}^{n} \alpha_i} \right) \frac{x_j^{\alpha_j} \exp(-x_j/\beta)}{\beta^{\alpha_j + 1} \Gamma(\alpha_j + 1)} \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{x_i^{\alpha_i - 1} \exp(-x_i/\beta)}{\beta^{\alpha_i} \Gamma(\alpha_i)}, \; x_1 > 0, \ldots, x_n > 0. \quad (3)$$

For $n = 2$ in (1), the bivariate gamma density is obtained as

$$C(\alpha_1, \alpha_2; \beta, k) x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} (x_1 + x_2)^k \exp\left[-\frac{1}{\beta}(x_1 + x_2) \right], \; x_1 > 0, \; x_2 > 0, \quad (4)$$

where

$$C(\alpha_1, \alpha_2; \beta, k) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\beta^{-(\alpha_1 + \alpha_2 + k)}}{\Gamma(\alpha_1 + \alpha_2 + k)}.$$

In a recent article, Rafiei, Iranmanesh, and Nagar [24] have studied the above distribution for $\alpha_1 = \alpha_2$. Substituting $n = 2$ in (3) or $k = 1$ in (4), the generalized bivariate gamma density takes the form

$$\frac{\alpha_1}{\alpha_1+\alpha_2}\frac{x_1^{\alpha_1}x_2^{\alpha_2-1}\exp[-(x_1+x_2)/\beta]}{\beta^{\alpha_1+\alpha_2+1}\Gamma(\alpha_1+1)\Gamma(\alpha_2)} + \frac{\alpha_2}{\alpha_1+\alpha_2}\frac{x_1^{\alpha_1-1}x_2^{\alpha_2}\exp[-(x_1+x_2)/\beta]}{\beta^{\alpha_1+\alpha_2+1}\Gamma(\alpha_1)\Gamma(\alpha_2+1)},\ x_1>0,\ x_2>0,$$

which yields the marginal density of X_1 as

$$\frac{\alpha_1}{\alpha_1+\alpha_2}\frac{x_1^{\alpha_1}\exp(-x_1/\beta)}{\beta^{\alpha_1+1}\Gamma(\alpha_1+1)} + \frac{\alpha_2}{\alpha_1+\alpha_2}\frac{x_1^{\alpha_1-1}\exp(-x_1/\beta)}{\beta^{\alpha_1}\Gamma(\alpha_2+1)},\ x_1>0.$$

Clearly, the marginal density of X_1 is a mixture of two gamma densities indicating that, in general, marginal density of any subset of $X_1, \ldots, X_n$ is not a generalized multivariate gamma.

It may be noted here that the multivariate gamma distribution defined above belongs to the Liouville family of distributions (Sivazlian [30], Gupta and Song [12], Gupta, and Richards [13], Song and Gupta [31]). Because of mathematical tractability, this distribution further enriches the class of multivariate Liouville distributions and may serve as an alternative to many existing distributions belonging to this class.

3 Marginal Distributions

In this section, we derive results on marginal distributions of the generalized multivariate gamma distribution defined in this chapter. By using multinomial expansion of $\left(\sum_{i=1}^n x_i\right)^k$, namely,

$$\left(\sum_{i=1}^n x_i\right)^k = \sum_{k_1+k_2+\cdots+k_n=k}\binom{k}{k_1,k_2,\ldots,k_n}x_1^{k_1}x_2^{k_2}\cdots x_n^{k_n}$$

in (1), the joint density of $X_1, \ldots, X_n$ can be restated as

$$C(\alpha_1,\ldots,\alpha_n;\beta,k)\sum_{k_1+k_2+\cdots+k_n=k}\binom{k}{k_1,k_2,\ldots,k_n}\prod_{i=1}^n x_i^{\alpha_i+k_i-1}\exp\left(-\frac{1}{\beta}\sum_{i=1}^n x_i\right),$$

where $x_i > 0, i = 1, 2, \ldots, n$. Thus, the generalized multivariate gamma distribution is a finite mixture of product of independent gamma densities.

In the remaining part of this section and the next section, we derive marginal distributions, distribution of partial sums and several factorizations of the generalized multivariate gamma distribution.

Theorem 1 *Let* $(X_1, \ldots, X_n) \sim \mathrm{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Then, for* $1 \le s \le n-1$, *the marginal density of* $X_1, \ldots, X_s$ *is given by*

$$C\left(\alpha_1, \ldots, \alpha_s, \sum_{i=s+1}^{n} \alpha_i; \beta, k\right) \prod_{i=1}^{s} x_i^{\alpha_i - 1} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{s} x_i\right) \left(\sum_{i=1}^{s} x_i\right)^k$$
$$\times \beta^{\sum_{i=s+1}^{n} \alpha_i} \sum_{j=0}^{k} \binom{k}{j} \Gamma\left(\sum_{i=s+1}^{n} \alpha_i + j\right) \left(\frac{\sum_{i=1}^{s} x_i}{\beta}\right)^{-j}, \; x_i > 0, \; i = 1, \ldots, s.$$

Proof Integrating out $x_{s+1}, \ldots, x_n$ in (1), the marginal density of $X_1, \ldots, X_s$ is derived as

$$C(\alpha_1, \ldots, \alpha_n; \beta, k) \prod_{i=1}^{s} x_i^{\alpha_i - 1} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{s} x_i\right)$$
$$\times \int_0^\infty \cdots \int_0^\infty \prod_{i=s+1}^{n} x_i^{\alpha_i - 1} \left(\sum_{i=1}^{s} x_i + \sum_{i=s+1}^{n} x_i\right)^k \exp\left(-\frac{1}{\beta} \sum_{i=s+1}^{n} x_i\right) \prod_{i=s+1}^{n} \mathrm{d}x_i$$
$$= C(\alpha_1, \ldots, \alpha_n; \beta, k) \prod_{i=1}^{s} x_i^{\alpha_i - 1} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{s} x_i\right)$$
$$\times \frac{\prod_{i=s+1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum_{i=s+1}^{n} \alpha_i)} \int_0^\infty x^{\sum_{i=s+1}^{n} \alpha_i - 1} \left(\sum_{i=1}^{s} x_i + x\right)^k \exp\left(-\frac{1}{\beta} x\right) \mathrm{d}x \tag{5}$$

where the last line has been obtained by using (16). Substituting $x / \sum_{i=1}^{s} x_i = z$ in (5), the marginal density of $X_1, \ldots, X_s$ is rewritten as

$$C\left(\alpha_1, \ldots, \alpha_s, \sum_{i=s+1}^{n} \alpha_i; \beta, k\right) \prod_{i=1}^{s} x_i^{\alpha_i - 1} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{s} x_i\right) \left(\sum_{i=1}^{s} x_i\right)^{\sum_{i=s+1}^{n} \alpha_i + k}$$
$$\times \int_0^\infty z^{\sum_{i=s+1}^{n} \alpha_i - 1} (1+z)^k \exp\left[-\frac{1}{\beta} \left(\sum_{i=1}^{s} x_i\right) z\right] \mathrm{d}z. \tag{6}$$

Now, writing $(1+z)^k$ using binomial theorem and integrating z in (6), the marginal density of $X_1, \ldots, X_s$ is derived. □

Alternately, the density of $X_1, \ldots, X_s$ given in (7) can be written as

$$C\left(\alpha_1, \ldots, \alpha_s, \sum_{i=s+1}^{n} \alpha_i; \beta, k\right) \beta^{\sum_{i=s+1}^{n} \alpha_i + k} \prod_{i=1}^{s} x_i^{\alpha_i - 1} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{s} x_i\right)$$
$$\times \sum_{j=0}^{k} \binom{k}{j} \Gamma\left(\sum_{i=s+1}^{n} \alpha_i + k - j\right) \left(\frac{\sum_{i=1}^{s} x_i}{\beta}\right)^j, \; x_i > 0, \; i = 1, \ldots, s.$$

Corollary 1 *The marginal density of X_1 is given by*

$$C\left(\alpha_1, \sum_{i=2}^{n} \alpha_i; \beta, k\right) \beta^{\sum_{i=2}^{n} \alpha_i + k} x_1^{\alpha_1 - 1} \exp\left(-\frac{1}{\beta} x_1\right)$$
$$\times \sum_{j=0}^{k} \binom{k}{j} \Gamma\left(\sum_{i=2}^{n} \alpha_i + k - j\right) \left(\frac{x_1}{\beta}\right)^j, \; x_1 > 0.$$

Corollary 2 *The marginal density of X_1 and X_2 is given by*

$$C\left(\alpha_1, \alpha_2, \sum_{i=3}^{n} \alpha_i; \beta, k\right) \beta^{\sum_{i=3}^{n} \alpha_i + k} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \exp\left[-\frac{1}{\beta}(x_1 + x_2)\right]$$
$$\times \sum_{j=0}^{k} \binom{k}{j} \Gamma\left(\sum_{i=3}^{n} \alpha_i + k - j\right) \left(\frac{x_1 + x_2}{\beta}\right)^j, \; x_1 > 0, x_2 > 0.$$

Substituting $u = z/(1+z)$ with $dz = (1-u)^{-2} du$ in (6), one gets

$$C\left(\alpha_1, \ldots, \alpha_s, \sum_{i=s+1}^{n} \alpha_i; \beta, k\right) \prod_{i=1}^{s} x_i^{\alpha_i - 1} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{s} x_i\right) \left(\sum_{i=1}^{s} x_i\right)^{\sum_{i=s+1}^{n} \alpha_i + k}$$
$$\times \int_0^1 u^{\sum_{i=s+1}^{n} \alpha_i - 1} (1-u)^{-(\sum_{i=s+1}^{n} \alpha_i + k + 1)} \exp\left[-\frac{\left(\sum_{i=1}^{s} x_i\right) u}{\beta(1-u)}\right] \mathrm{d}u. \tag{7}$$

Now, writing

$$(1-u)^{-(\sum_{i=s+1}^{n} \alpha_i + k + 1)} \exp\left[-\frac{\left(\sum_{i=1}^{s} x_i\right) u}{\beta(1-u)}\right] = \sum_{j=0}^{\infty} u^j L_j^{(\sum_{i=s+1}^{n} \alpha_i + k)}\left(\frac{\sum_{i=1}^{s} x_i}{\beta}\right)$$

in (7) and integrating u, the marginal density of $X_1, \ldots, X_s$, in series involving generalized Laguerre polynomials, is derived as

$$C\left(\alpha_1,\ldots,\alpha_s,\sum_{i=s+1}^{n}\alpha_i;\beta,k\right)\prod_{i=1}^{s}x_i^{\alpha_i-1}\exp\left(-\frac{1}{\beta}\sum_{i=1}^{s}x_i\right)\left(\sum_{i=1}^{s}x_i\right)^{\sum_{i=s+1}^{n}\alpha_i+k}$$
$$\times\sum_{j=0}^{\infty}\frac{1}{\sum_{i=s+1}^{n}\alpha_i+j}L_j^{(\sum_{i=s+1}^{n}\alpha_i+k)}\left(\frac{\sum_{i=1}^{s}x_i}{\beta}\right).$$

Theorem 2 *Let* $(X_1,\ldots,X_n)\sim \mathrm{GMG}(\alpha_1,\ldots,\alpha_n;\beta,k)$. *Then, for* $2\le r\le n$, *the marginal density of* $X_r,\ldots,X_n$ *is given by*

$$C\left(\sum_{i=1}^{r-1}\alpha_i,\alpha_r,\ldots,\alpha_n;\beta,k\right)\beta^{\sum_{i=1}^{r-1}\alpha_i+k}\prod_{i=r}^{n}x_i^{\alpha_i-1}\exp\left(-\frac{1}{\beta}\sum_{i=r}^{n}x_i\right)$$
$$\times\sum_{\ell=0}^{k}\binom{k}{\ell}\Gamma\left(\sum_{i=1}^{r-1}\alpha_i+k-\ell\right)\left(\frac{\sum_{i=r}^{n}x_i}{\beta}\right)^{\ell},\ x_i>0,\ i=r,\ldots,n.$$

Proof Similar to the proof of Theorem 1. □

Corollary 3 *The marginal density of* X_n *is given by*

$$C\left(\sum_{i=1}^{n-1}\alpha_i,\alpha_n;\beta,k\right)\beta^{\sum_{i=1}^{n-1}\alpha_i+k}x_n^{\alpha_n-1}\exp\left(-\frac{1}{\beta}x_n\right)$$
$$\times\sum_{\ell=0}^{k}\binom{k}{\ell}\Gamma\left(\sum_{i=1}^{n-1}\alpha_i+k-\ell\right)\left(\frac{x_n}{\beta}\right)^{\ell},\ x_n>0.$$

Theorem 3 *Let* $(X_1,\ldots,X_n)\sim \mathrm{GMG}(\alpha_1,\ldots,\alpha_n;\beta,k)$. *Then, for* $r=1,\ldots,n$, *the marginal density of* X_r *is given by*

$$C\left(\sum_{i(\neq r)=1}^{n}\alpha_i,\alpha_r;\beta,k\right)\beta^{\sum_{i(\neq r)=1}^{n}\alpha_i+k}x_r^{\alpha_r+k-1}\exp\left(-\frac{x_r}{\beta}\right)$$
$$\times\sum_{j=0}^{k}\binom{k}{j}\Gamma\left(\sum_{i(\neq r)=1}^{n}\alpha_i+k-j\right)\left(\frac{x_r}{\beta}\right)^{j},\ x_r>0.$$

4 Factorizations

This section deals with several factorizations of the multivariate gamma distribution defined in Sect. 2.

In the next theorem, we give the joint distribution of partial sums of random variables distributed jointly as generalized multivariate gamma.

Let $n_1, \ldots, n_\ell$ be non-negative integers such that $\sum_{i=1}^{\ell} n_i = n$ and define

$$\alpha_{(i)} = \sum_{j=n_{i-1}^*+1}^{n_i^*} \alpha_j, \quad n_0^* = 0, \quad n_i^* = \sum_{j=1}^{i} n_j, \quad i = 1, \ldots, \ell.$$

Theorem 4 *Let* $(X_1, \ldots, X_n) \sim \mathrm{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $Z_j = X_j / X_{(i)}$, $j = n_{i-1}^* + 1, \ldots, n_i^* - 1$ *and* $X_{(i)} = \sum_{j=n_{i-1}^*+1}^{n_i^*} X_j$, $i = 1, \ldots, \ell$. *Then,*

(i) $(X_{(1)}, \ldots, X_{(\ell)})$ *and* $(Z_{n_{i-1}^*+1}, \ldots, Z_{n_i^*-1})$, $i = 1, \ldots, \ell$, *are independently distributed,*

(ii) $(Z_{n_{i-1}^*+1}, \ldots, Z_{n_i^*-1}) \sim \mathrm{D1}(\alpha_{n_{i-1}^*+1}, \ldots, \alpha_{n_i^*-1}; \alpha_{n_i^*})$, $i = 1, \ldots, \ell$, *and*

(iii) $(X_{(1)}, \ldots, X_{(\ell)}) \sim \mathrm{GMG}(\alpha_{(1)}, \ldots, \alpha_{(\ell)}; \beta, k)$.

Proof Substituting $x_{(i)} = \sum_{j=n_{i-1}^*+1}^{n_i^*} x_j$ and $z_j = x_j / x_{(i)}$, $j = n_{i-1}^* + 1, \ldots, n_i^* - 1$, $i = 1, \ldots, \ell$ with the Jacobian

$$\begin{aligned} &J(x_1, \ldots, x_n \to z_1, \ldots, z_{n_1-1}, x_{(1)}, \ldots, z_{n_{\ell-1}^*+1}, \ldots, z_{n-1}, x_{(\ell)}) \\ &= \prod_{i=1}^{\ell} J(x_{n_{i-1}^*+1}, \ldots, x_{n_i^*} \to z_{n_{i-1}^*+1}, \ldots, z_{n_i^*-1}, x_{(i)}) \\ &= \prod_{i=1}^{\ell} x_{(i)}^{n_i - 1}. \end{aligned}$$

in the density of $(X_1, \ldots, X_n)$ given by (1), we get the joint density of $Z_{n_{i-1}^*+1}, \ldots, Z_{n_i^*-1}, X_{(i)}$, $i = 1, \ldots, \ell$ as

$$\begin{aligned} &C(\alpha_1, \ldots, \alpha_n; \alpha, \beta) \prod_{i=1}^{\ell} x_{(i)}^{\alpha_{(i)}-1} \left(\sum_{i=1}^{\ell} x_{(i)} \right)^k \exp\left(-\frac{1}{\beta} \sum_{i=1}^{\ell} x_{(i)} \right) \\ &\times \prod_{i=1}^{\ell} \left[\left(\prod_{j=n_{i-1}^*+1}^{n_i^*-1} z_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=n_{i-1}^*+1}^{n_i^*-1} z_j \right)^{\alpha_{n_i^*}-1} \right], \end{aligned} \tag{8}$$

where $x_{(i)} > 0$, $i = 1, \ldots, \ell$, $z_j > 0$, $j = n_{i-1}^* + 1, \ldots, n_i^* - 1$, $\sum_{j=n_{i-1}^*+1}^{n_i^*-1} z_j < 1$, $i = 1, \ldots, \ell$. From the factorization in (8), it is easy to see that $(X_{(1)}, \ldots, X_{(\ell)})$ and $(Z_{n_{i-1}^*+1}, \ldots, Z_{n_i^*-1})$, $i = 1, \ldots, \ell$, are independently distributed. Further $(X_{(1)}, \ldots, X_{(\ell)}) \sim \mathrm{GMG}(\alpha_{(1)}, \ldots, \alpha_{(\ell)}; \beta, k)$ and $(Z_{n_{i-1}^*+1}, \ldots, Z_{n_i^*-1}) \sim \mathrm{D1}(\alpha_{n_{i-1}^*+1}, \ldots, \alpha_{n_i^*-1}; \alpha_{n_i^*})$, $i = 1, \ldots, \ell$. □

Corollary 4 *Let* $(X_1, \ldots, X_n) \sim \mathrm{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $Z_i = X_i / Z$, $i = 1, \ldots, n-1$, *and* $Z = \sum_{j=1}^{n} X_j$. *Then,* $(Z_1, \ldots, Z_{n-1})$ *and* Z *are independent,* $(Z_1, \ldots, Z_{n-1}) \sim \mathrm{D1}(\alpha_1, \ldots, \alpha_{n-1}; \alpha_n)$ *and* $Z \sim \mathrm{G}\left(\sum_{i=1}^{n} \alpha_i + k, \beta\right)$.

Corollary 5 *If* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$, *then* $\sum_{j=1}^{n} X_j$ *and* $\frac{\sum_{i=1}^{s} X_i}{\sum_{i=1}^{n} X_i}$ *are independent. Further*

$$\frac{\sum_{i=1}^{s} X_i}{\sum_{i=1}^{n} X_i} \sim \text{B1}\left(\sum_{i=1}^{s} \alpha_i, \sum_{i=s+1}^{n} \alpha_i\right), \ s < n.$$

Theorem 5 *Let* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $W_j = X_j / X_{n_i^*}$, $j = n_{i-1}^* + 1, \ldots, n_i^* - 1$ *and* $X_{(i)} = \sum_{j=n_{i-1}^*+1}^{n_i^*} X_j$, $i = 1, \ldots, \ell$. *Then,*

(i) $(X_{(1)}, \ldots, X_{(\ell)})$ *and* $(W_{n_{i-1}^*+1}, \ldots, W_{n_i^*-1})$, $i = 1, \ldots, \ell$, *are independently distributed,*

(ii) $(W_{n_{i-1}^*+1}, \ldots, W_{n_i^*-1}) \sim \text{D2}(\alpha_{n_{i-1}^*+1}, \ldots, \alpha_{n_i^*-1}; \alpha_{n_i^*})$, $i = 1, \ldots, \ell$, *and*

(iii) $(X_{(1)}, \ldots, X_{(\ell)}) \sim \text{GMG}(\alpha_{(1)}, \ldots, \alpha_{(\ell)}; \beta, k)$.

Corollary 6 *Let* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $W_i = X_i / X_n$, $i = 1, \ldots, n-1$ *and* $Z = \sum_{j=1}^{n} X_j$. *Then,* $(W_1, \ldots, W_{n-1})$ *and* Z *are independent,* $(W_1, \ldots, W_{n-1}) \sim \text{D2}(\alpha_1, \ldots, \alpha_{n-1}; \alpha_n)$ *and* $Z \sim \text{G}(\sum_{i=1}^{n} \alpha_i, \beta, k)$.

Corollary 7 *If* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$, *then* $\sum_{j=1}^{n} X_j$ *and* $\frac{\sum_{i=1}^{s} X_i}{\sum_{i=s+1}^{n} X_i}$ *are independent. Further*

$$\frac{\sum_{i=1}^{s} X_i}{\sum_{i=s+1}^{n} X_i} \sim \text{B2}\left(\sum_{i=1}^{s} \alpha_i, \sum_{i=s+1}^{n} \alpha_i\right), \ s < n.$$

In next six theorems, we give several factorizations of the generalized multivariate gamma density.

Theorem 6 *Let* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $Y_n = \sum_{j=1}^{n} X_j$ *and* $Y_i = \sum_{j=1}^{i} X_j / \sum_{j=1}^{i+1} X_j$, $i = 1, \ldots, n-1$. *Then,* $Y_1, \ldots, Y_n$ *are independent,* $Y_i \sim \text{B1}(\sum_{j=1}^{i} \alpha_j, \alpha_{i+1})$, $i = 1, \ldots, n-1$, *and* $Y_n \sim \text{G}(\sum_{i=1}^{n} \alpha_i + k, \beta)$.

Proof Substituting $x_1 = y_n \prod_{i=1}^{n-1} y_i, x_2 = y_n(1 - y_1) \prod_{i=2}^{n-1} y_i, \ldots, x_{n-1} = y_n (1 - y_{n-2}) y_{n-1}$ and $x_n = y_n(1 - y_{n-1})$ with the Jacobian $J(x_1, \ldots, x_n \to y_1, \ldots, y_n) = \prod_{i=2}^{n} y_i^{i-1}$ in (1) we get the desired result. □

Theorem 7 *Let* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $Z_n = \sum_{j=1}^{n} X_j$ *and* $Z_i = X_{i+1} / \sum_{j=1}^{i} X_j$, $i = 1, \ldots, n-1$. *Then,* $Z_1, \ldots, Z_n$ *are independent,* $Z_i \sim \text{B2}(\alpha_{i+1}, \sum_{j=1}^{i} \alpha_j)$, $i = 1, \ldots, n-1$, *and* $Z_n \sim \text{G}(\sum_{j=1}^{n} \alpha_j + k, \beta)$.

Proof The desired result follows from Theorem 6 by noting that $(1 - Y_i)/Y_i \sim \text{B2}(\alpha_{i+1}, \sum_{j=1}^{i} \alpha_j)$. □

Theorem 8 *Let* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $W_n = \sum_{j=1}^{n} X_j$ *and* $W_i = \sum_{j=1}^{i} X_j / X_{i+1}$, $i = 1, \ldots, n-1$. *Then,* $W_1, \ldots, W_n$ *are independent,* $W_i \sim \text{B2}(\sum_{j=1}^{i} \alpha_j, \alpha_{i+1})$, $i = 1, \ldots, n-1$, *and* $W_n \sim \text{G}(\sum_{i=1}^{n} \alpha_i + k, \beta, k)$.

Proof The result follows from Theorem 7 by noting that $1/Z_i \sim \text{B2}(\sum_{j=1}^{i} \alpha_j, \alpha_{i+1})$. □

Theorem 9 *Let* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $Y_n = \sum_{j=1}^{n} X_j$ *and* $Y_i = X_i / \sum_{j=i}^{n} X_j$, $i = 1, \ldots, n-1$. *Then,* $Y_1, \ldots, Y_n$ *are independent,* $Y_i \sim \text{B1}(\alpha_i, \sum_{j=i+1}^{n} \alpha_j)$, $i = 1, \ldots, n-1$, *and* $Y_n \sim \text{G}(\sum_{i=1}^{n} \alpha_i + k, \beta)$.

Proof Substituting $x_1 = y_n y_1, x_2 = y_n y_2(1-y_1), \ldots, x_{n-1} = y_n y_{n-1}(1-y_1)\cdots(1-y_{n-2})$, and $x_n = y_n(1-y_1)\cdots(1-y_{n-1})$ with the Jacobian $J(x_1, \ldots, x_n \to y_1, \ldots, y_n) = y_n^{n-1} \prod_{i=1}^{n-2}(1-y_i)^{n-i-1}$ in (1), we get the desired result. □

Theorem 10 *Let* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $Z_n = \sum_{j=1}^{n} X_j$ *and* $Z_i = X_i / \sum_{j=i+1}^{n} X_j$, $i = 1, \ldots, n-1$. *Then,* $Z_1, \ldots, Z_n$ *are independent,* $Z_i \sim \text{B2}(\alpha_i, \sum_{j=i+1}^{n} \alpha_j)$, $i = 1, \ldots, n-1$, *and* $Z_n \sim \text{G}(\sum_{i=1}^{n} \alpha_i + k, \beta)$.

Proof The result follows from Theorem 9 by observing that $Y_i/(1-Y_i) \sim \text{B2}(\alpha_i, \sum_{j=i+1}^{n} \alpha_j)$. □

Theorem 11 *Let* $(X_1, \ldots, X_n) \sim \text{GMG}(\alpha_1, \ldots, \alpha_n; \beta, k)$. *Define* $W_n = \sum_{j=1}^{n} X_j$ *and* $W_i = \sum_{j=i+1}^{n} X_j / X_i$, $i = 1, \ldots, n-1$. *Then,* $W_1, \ldots, W_n$ *are independent,* $W_i \sim \text{B2}(\sum_{j=i+1}^{n} \alpha_j, \alpha_i)$, $i = 1, \ldots, n-1$, *and* $W_n \sim \text{G}(\sum_{i=1}^{n} \alpha_i + k, \beta)$.

Proof The result follows from Theorem 10 by noting that $1/W_i \sim \text{B2}(\sum_{j=i+1}^{n} \alpha_j, \alpha_i)$. □

5 Joint Moments

By definition

$$E(X_1^{r_1} \cdots X_n^{r_n}) = C(\alpha_1, \ldots, \alpha_n; \beta, k) \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{n} x_i^{\alpha_i + r_i - 1} \left(\sum_{i=1}^{n} x_i \right)^k \times \exp\left(-\frac{1}{\beta} \sum_{i=1}^{n} x_i \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n = \frac{C(\alpha_1, \ldots, \alpha_n; \beta, k)}{C(\alpha_1 + r_1, \ldots, \alpha_n + r_n; \beta, k)}.$$

Now, simplifying the above expression by using (2), one gets

$$E(X_1^{r_1} \cdots X_n^{r_n}) = \beta^r \frac{\Gamma(\alpha)\Gamma(\alpha + r + k)}{\Gamma(\alpha + k)\Gamma(\alpha + r)} \prod_{i=1}^{n} \frac{\Gamma(\alpha_i + r_i)}{\Gamma(\alpha_i)},$$

where $\alpha = \sum_{i=1}^{n} \alpha_i$ and $r = \sum_{i=1}^{n} r_i$.

Further, substituting appropriately in the above expression, one gets

$$E(X_\ell^{r_\ell} X_m^{r_m}) = \beta^{r_\ell + r_m} \frac{\Gamma(\alpha)\Gamma(\alpha + r_\ell + r_m + k)}{\Gamma(\alpha + k)\Gamma(\alpha + r_\ell + r_m)} \frac{\Gamma(\alpha_\ell + r_\ell)\Gamma(\alpha_m + r_m)}{\Gamma(\alpha_\ell)\Gamma(\alpha_m)},$$

$$E(X_\ell X_m) = \beta^2 \frac{\alpha_\ell \alpha_m (\alpha + k)(\alpha + k + 1)}{\alpha(\alpha + 1)},$$

$$E(X_j) = \beta \frac{\alpha_j (\alpha + k)}{\alpha},$$

and

$$E(X_j^2) = \beta^2 \frac{\alpha_j (\alpha_j + 1)(\alpha + k)(\alpha + k + 1)}{\alpha(\alpha + 1)}.$$

Finally, by using appropriate definitions, we get

$$\text{var}(X_j) = \beta^2 \frac{\alpha_j (\alpha + k)[\alpha(\alpha + 1) + (\alpha - \alpha_j)k]}{\alpha^2(\alpha + 1)},$$

$$\text{cov}(X_\ell, X_m) = -k\beta^2 \frac{\alpha_\ell \alpha_m (\alpha + k)}{\alpha^2(\alpha + 1)},$$

$$\text{corr}(X_\ell, X_m) = -k \sqrt{\frac{\alpha_\ell \alpha_m}{[\alpha(\alpha + 1) + (\alpha - \alpha_\ell)k][\alpha(\alpha + 1) + (\alpha - \alpha_m)k]}}.$$

6 Moment Generating Function

By definition, the joint mgf of $X_1, \ldots, X_n$ is given by

$$M_{X_1,\ldots,X_n}(t_1, \ldots, t_n) = C(\alpha_1, \ldots, \alpha_n; \beta, k) \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^n x_i^{\alpha_i - 1} \left(\sum_{i=1}^n x_i \right)^k \times \exp\left(\sum_{i=1}^n t_i x_i - \frac{1}{\beta} \sum_{i=1}^n x_i \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n. \tag{9}$$

Substituting $x_1 = r_1 s, \ldots x_{n-1} = r_{n-1} s$ and $x_n = s(1 - \sum_{i=1}^{n-1} r_i)$ in (9) with the Jacobian $J(x_1, \ldots, x_{n-1}, x_n \to r_1, \ldots, r_{n-1}, s) = s^{n-1}$ and integrating s, we get

$$
\begin{aligned}
&M_{X_1,\ldots,X_n}(t_1,\ldots,t_n)\\
&= C(\alpha_1,\ldots,\alpha_n;\beta,k)\beta^{\sum_{i=1}^n \alpha_i+k}\Gamma\left(\sum_{i=1}^n \alpha_i + k\right)\int_{\substack{r_1+\cdots+r_{n-1}<1\\ 0<r_i,\, i=1,\ldots,n-1}}\cdots\int \prod_{i=1}^{n-1} r_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n-1} r_i\right)^{\alpha_n-1}\\
&\quad\times\left[\sum_{i=1}^{n-1}(1-\beta t_i)r_i + (1-\beta t_n)\left(1-\sum_{i=1}^{n-1} r_i\right)\right]^{-(\sum_{i=1}^n \alpha_i+k)} \mathrm{d}r_1\cdots\mathrm{d}r_{n-1}, \qquad (10)
\end{aligned}
$$

where $1-t_i\beta>0$, $i=1,\ldots,n$. Now, writing

$$
\left[\sum_{i=1}^{n-1}(1-\beta t_i)r_i + (1-\beta t_n)\left(1-\sum_{i=1}^{n-1} r_i\right)\right]^{-(\sum_{i=1}^n \alpha_i+k)}
$$

$$
= (1-t_n\beta)^{-(\sum_{i=1}^n \alpha_i+k)}\left[1-\sum_{i=1}^{n-1} r_i\left(1-\frac{1-t_i\beta}{1-\beta t_n}\right)\right]^{-(\sum_{i=1}^n \alpha_i+k)},
$$

$$
\frac{1-t_i\beta}{1-t_n\beta}<1, \quad i=1,\ldots,n-1
$$

in (10) and integrating r, we get

$$
\begin{aligned}
&M_{X_1,\ldots,X_n}(t_1,\ldots,t_n)\\
&= C(\alpha_1,\ldots,\alpha_n;\beta,k)\beta^{\sum_{i=1}^n \alpha_i+k}\Gamma\left(\sum_{i=1}^n \alpha_i + k\right)(1-t_n\beta)^{-(\sum_{i=1}^n \alpha_i+k)}\\
&\quad\times\int_{\substack{r_1+\cdots+r_{n-1}<1\\ 0<r_i,i=1,\ldots,n-1}}\cdots\int \prod_{i=1}^{n-1} r_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n-1} r_i\right)^{\alpha_n-1}\left[1-\sum_{i=1}^{n-1} r_i\frac{\beta(t_i-t_n)}{1-\beta t_n}\right]^{-(\sum_{i=1}^n \alpha_i+k)}\mathrm{d}r_1\cdots\mathrm{d}r_{n-1}\\
&= C(\alpha_1,\ldots,\alpha_n;\beta,k)\beta^{\sum_{i=1}^n \alpha_i+k}\Gamma\left(\sum_{i=1}^n \alpha_i + k\right)(1-t_n\beta)^{-(\sum_{i=1}^n \alpha_i+k)}\frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)}\\
&\quad\times F_D^{(n-1)}\left(\sum_{i=1}^n \alpha_i + k, \alpha_1,\ldots,\alpha_{n-1};\sum_{i=1}^n \alpha_i;\frac{\beta(t_1-t_n)}{1-\beta t_n},\ldots,\frac{\beta(t_{n-1}-t_n)}{1-\beta t_n}\right),
\end{aligned}
$$

where the last line has been obtained by using the integral representation of the fourth hypergeometric function of Lauricella given in (15). Finally, substituting for $C(\alpha_1,\ldots,\alpha_n;\beta,k)$ and simplifying, we get

$$M_{X_1,\ldots,X_n}(t_1,\ldots,t_n) = (1-t_n\beta)^{-(\sum_{i=1}^n \alpha_i + k)} \times F_D^{(n-1)}\left(\sum_{i=1}^n \alpha_i + k, \alpha_1, \ldots, \alpha_{n-1}; \sum_{i=1}^n \alpha_i; \frac{\beta(t_1-t_n)}{1-\beta t_n}, \ldots, \frac{\beta(t_{n-1}-t_n)}{1-\beta t_n}\right).$$

For $t_1 = \cdots = t_n = t$, we have

$$M_{X_1,\ldots,X_n}(t,\ldots,t) = M_{X_1+\cdots+X_n}(t) = (1-t\beta)^{-(\sum_{i=1}^n \alpha_i + k)}$$

which is the mgf of a gamma random variable with shape parameter $\sum_{i=1}^n \alpha_i + k$ and scale parameter β.

7 Entropies

In this section, exact forms of Rényi and Shannon entropies are derived for the multivariate gamma distribution defined in this article.

Let $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ be a probability space. Consider a pdf f associated with $\mathcal{P}$, dominated by $\sigma-$finite measure μ on $\mathcal{X}$. Denote by $H_{SH}(f)$ the well-known Shannon entropy introduced in Shannon [29]. It is define by

$$H_{SH}(f) = -\int_{\mathcal{X}} f(x)\log f(x)\,\mathrm{d}\mu. \tag{11}$$

One of the main extensions of the Shannon entropy was defined by Rényi [25]. This generalized entropy measure is given by

$$H_R(\eta, f) = \frac{\log G(\eta)}{1-\eta} \qquad (\text{for } \eta > 0 \text{ and } \eta \neq 1), \tag{12}$$

where

$$G(\eta) = \int_{\mathcal{X}} f^{\eta} d\mu.$$

The additional parameter η is used to describe complex behavior in probability models and the associated process under study. Rényi entropy is monotonically decreasing in η, while Shannon entropy (11) is obtained from (12) for $\eta \uparrow 1$. For details see Nadarajah and Zografos [22], Zografos and Nadarajah [36] and Zografos [35].

Theorem 12 *For the generalized multivariate gamma distribution defined by the pdf (1), the Rényi and the Shannon entropies are given by*

$$H_R(\eta, f) = \frac{1}{1-\eta}\Bigg[\eta \ln C(\alpha_1, \ldots, \alpha_n; \beta, k) + \Big[\eta \sum_{i=1}^{n}(\alpha_i - 1) + n + \eta k\Big] \ln\left(\frac{\beta}{\eta}\right) + \sum_{i=1}^{n} \ln \Gamma[\eta(\alpha_i - 1) + 1] + \ln \Gamma\left[\eta \sum_{i=1}^{n}(\alpha_i - 1) + n + \eta k\right] - \ln \Gamma\left[\eta \sum_{i=1}^{n}(\alpha_i - 1) + n\right]$$

and

$$H_{SH}(f) = -\ln C(\alpha_1, \ldots, \alpha_n; \beta, k) - \Bigg[\Big(\sum_{i=1}^{n}\alpha_i + k - n\Big)\ln\beta - \Big(\sum_{i=1}^{n}\alpha_i + k\Big) + \sum_{i=1}^{n}(\alpha_i - 1)\psi(\alpha_i) + \Big(\sum_{i=1}^{n}\alpha_i + k - n\Big)\psi\Big(\sum_{i=1}^{n}\alpha_i + k\Big) - \Big(\sum_{i=1}^{n}\alpha_i - n\Big)\psi\Big(\sum_{i=1}^{n}\alpha_i\Big)\Bigg],$$

respectively, where $\psi(z) = \frac{d}{dz}\ln\Gamma(z) = \frac{1}{\Gamma(z)}\frac{d}{dz}\Gamma(z)$ *is the digamma function.*

Proof For $\eta > 0$ and $\eta \neq 1$, using the joint density of $X_1, \ldots, X_n$ given by (1), we have

$$\begin{aligned} G(\eta) &= \int_0^\infty \cdots \int_0^\infty f^\eta(x_1, \ldots, x_n; \alpha_1, \ldots, \alpha_n; \beta, k) \prod_{i=1}^{n} \mathrm{d}x_i \\ &= [C(\alpha_1, \ldots, \alpha_n; \beta, k)]^\eta \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^{n} x_i^{\eta(\alpha_i - 1)} \left(\sum_{i=1}^{n} x_i\right)^{\eta k} \\ &\quad \times \exp\left(-\frac{\eta}{\beta}\sum_{i=1}^{n} x_i\right) \prod_{i=1}^{n} \mathrm{d}x_i \\ &= [C(\alpha_1, \ldots, \alpha_n; \beta, k)]^\eta \frac{\prod_{i=1}^{n}\Gamma[\eta(\alpha_i - 1) + 1]}{\Gamma[\eta\sum_{i=1}^{n}(\alpha_i - 1) + n]} \\ &\quad \times \int_0^\infty x^{\eta\sum_{i=1}^{n}(\alpha_i - 1) + n + \eta k - 1} \exp\left(-\frac{\eta x}{\beta}\right) \mathrm{d}x, \end{aligned}$$

where the last line has been obtained by using (16). Finally, evaluating the above integral by using gamma integral and simplifying the resulting expression, we get

$$G(\eta) = [C(\alpha_1, \ldots, \alpha_n; \beta, k)]^{\eta} \frac{\prod_{i=1}^{n} \Gamma[\eta(\alpha_i - 1) + 1]}{\Gamma[\eta \sum_{i=1}^{n} (\alpha_i - 1) + n]} \Gamma\left[\eta \sum_{i=1}^{n} (\alpha_i - 1) + n + \eta k\right] \times \left(\frac{\beta}{\eta}\right)^{\eta \sum_{i=1}^{n} (\alpha_i - 1) + n + \eta k}.$$

Now, taking logarithm of $G(\eta)$ and using (12) we get $H_R(\eta, f)$. The Shannon entropy is obtained from $H_R(\eta, f)$ by taking $\eta \uparrow 1$ and using L'Hopital's rule. □

8 Estimation

Let $(X_{11}, \ldots, X_{1n}), \ldots, (X_{N1}, \ldots, X_{Nn})$ be a random sample from GMG$(\alpha_1, \ldots, \alpha_n; \beta, k)$. The log-likelihood function, denoted by $l(\alpha_1, \ldots, \alpha_n; \beta)$, is given by

$$l(\alpha_1, \ldots, \alpha_n; \beta) = N \left[\ln \Gamma(\alpha) - (\alpha + k) \ln \beta - \sum_{i=1}^{n} \ln \Gamma(\alpha_i) - \ln \Gamma(\alpha + k)\right] + \sum_{h=1}^{N} \sum_{i=1}^{n} (\alpha_i - 1) \ln x_{hi} + k \sum_{h=1}^{N} \ln \left(\sum_{i=1}^{n} x_{hi}\right) - \frac{1}{\beta} \sum_{h=1}^{N} \sum_{i=1}^{n} x_{hi},$$

where $\alpha = \sum_{i=1}^{n} \alpha_i$. Now, differentiating $l(\alpha_1, \ldots, \alpha_n; \beta)$ w.r.t. α_i, we get

$$\frac{\partial l(\alpha_1, \ldots, \alpha_n; \beta)}{\partial \alpha_i} = N [\psi(\alpha) - \ln \beta - \psi(\alpha_i) - \psi(\alpha + k)] + \sum_{h=1}^{N} \ln x_{hi}.$$

Further,

$$\frac{\partial l(\alpha_1, \ldots, \alpha_n; \beta)}{\partial \beta} = -\frac{N(\alpha + k)}{\beta} + \frac{1}{\beta^2} \sum_{h=1}^{N} \sum_{i=1}^{n} x_{hi},$$

$$\frac{\partial^2 l(\alpha_1, \ldots, \alpha_n; \beta)}{\partial \alpha_i \partial \alpha_\ell} = N [\psi_1(\alpha) - \psi_1(\alpha + k)], \quad 1 \le i \ne \ell \le n,$$

$$\frac{\partial^2 l(\alpha_1, \ldots, \alpha_n; \beta)}{\partial \alpha_i^2} = N[\psi_1(\alpha) - \psi_1(\alpha_i) - \psi_1(\alpha + k)],$$

where $\psi_1(z)$ is the trigamma function defined as the derivative of the digamma function, $\psi_1(z) = \frac{d}{dz} \psi(z)$,

$$\frac{\partial^2 l(\alpha_1, \ldots, \alpha_n; \beta)}{\partial \alpha_i \partial \beta} = -\frac{N}{\beta},$$

$$\frac{\partial^2 l(\alpha_1,\ldots,\alpha_n;\beta)}{\partial\beta^2} = \frac{N(\alpha+k)}{\beta^2} - \frac{2}{\beta^3}\sum_{h=1}^{N}\sum_{i=1}^{n} x_{hi}.$$

Now, noting that $\sum_{i=1}^{n} X_i \sim \mathrm{G}(\alpha+k,\beta)$ and the expected value of a constant is the constant itself, we obtain

$$\theta_{i\ell} = \theta_{\ell i} = E\left[\frac{\partial^2 l(\alpha_1,\ldots,\alpha_n;\beta)}{\partial\alpha_i\partial\alpha_\ell}\right] = N\psi_1(\alpha) - N\psi_1(\alpha+k),\ 1 \le i \ne \ell \le n,$$

$$\theta_{i\,n+1} = \theta_{n+1\,i} = E\left[\frac{\partial l(\alpha_1,\ldots,\alpha_n;\beta)}{\partial\alpha_i\partial\beta}\right] = -\frac{N}{\beta},\ 1 \le i \le n,$$

$$\theta_{ii} = E\left[\frac{\partial^2 l(\alpha_1,\ldots,\alpha_n;\beta)}{\partial\alpha_i^2}\right] = N\psi_1(\alpha) - N\psi_1(\alpha_i) - N\psi_1(\alpha+k),\ 1 \le i \le n,$$

$$\theta_{n+1\,n+1} = E\left[\frac{\partial^2 l(\alpha_1,\ldots,\alpha_n;\beta)}{\partial\beta^2}\right] = -\frac{N(\alpha+k)}{\beta^2}.$$

The Fisher information matrix for the multivariate gamma distribution given by the density (1) is defined as

$$-\begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1n} & \theta_{1\,n+1} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{1n} & \theta_{2\,n+1} \\ \vdots & & & & \vdots \\ \theta_{n1} & \theta_{n2} & \cdots & \theta_{nn} & \theta_{n\,n+1} \\ \theta_{n+1\,1} & \theta_{n+1\,2} & \cdots & \theta_{n+1\,n} & \theta_{n+1\,n+1} \end{pmatrix}.$$

Further

$$\frac{\partial l(\alpha_1,\ldots,\alpha_n;\beta)}{\partial\beta} = -\frac{N(\alpha+k)}{\beta} + \frac{1}{\beta^2}\sum_{h=1}^{N}\sum_{i=1}^{n} x_{hi} = 0$$

gives

$$(\alpha+k)\beta = \sum_{i=1}^{n} \bar{x}_i \tag{13}$$

and

$$\frac{\partial l(\alpha_1,\ldots,\alpha_n;\beta)}{\partial\alpha_i} = N\left[\psi(\alpha) - \ln\beta - \psi\left(\alpha_i\right) - \psi(\alpha+k)\right] + \sum_{h=1}^{N}\ln x_{hi} = 0$$

gives

$$\psi(\alpha + k) - \psi(\alpha) + \ln \beta + \psi(\alpha_i) = \ln \tilde{x}_i,\ i = 1, \ldots, n,$$

where $\tilde{x}_i = \prod_{h=1}^{N} x_{hi}^{1/N}$, $i = 1, 2, \ldots, n$. Further, using

$$[\psi(z + m) - \psi(z) = \sum_{j=0}^{m-1} \frac{1}{z + j}$$

we have

$$\sum_{j=0}^{k-1} \frac{1}{\alpha + j} + \ln \beta + \psi(\alpha_i) = \ln \tilde{x}_i, \quad i = 1, \ldots, n. \tag{14}$$

Thus, by solving numerically (13) and (14), the MLEs of α_i and β can be obtained.

9 Simulation

In this section, a simulation study for $p = 3$ is conducted to evaluate the performance of maximum likelihood method. For $p = 2$, see Rafiei, Iranmanesh, and Nagar [24]. Samples of size $n = 50, 200, 500$ from Equation (1) for selected values of parameters are generated by MCMC methods (Gibbs Metropolise, Markov Chain Monte Carlo Metropolise, Metropolise, Metropolise gaussian, random walk Metropolise and Metropolise-Hastings). We have performed the simulation for particular values of parameters, namely, $\alpha_1 = 1$, $\alpha_2 = 2$, $\alpha_3 = 3$, $\beta = 2$, $k = 4, 8$, and $\alpha_1 = 2$, $\alpha_2 = 2$, $\alpha_3 = 1$, $\beta = 2$, $k = 4, 8$. The results were similar for other choices. MLEs for parameters based on the numerical procedures were computed. This procedures was repeated five hundred times and $(\widehat{\alpha_1}, \widehat{\alpha_2}, \widehat{\alpha_3}, \widehat{\beta})$, the average of biases (Ab) and the mean squared errors (MSE) were obtained by using Monte Carlo methods (the parameter k is an integer and the derivative method is not used to calculate its MLE).

Different packages such as MCMC, MCMCpack, gibbs.met, LearnBayes, MHadaptive, MetroHastings and walkMetropolis in *R* were used for simulation. After performing simulation using the above methods and comparing results, it was observed that the Gibbs sampling method provides better results. Therefore, the output of Gibbs method is presented in Tables 1, 2, 3, 4 and Figs. 1, 2, 3, 4 and 5. The MLEs of parameters and correlation cofficients are reported in Tables 1 and 2. The DEoptim package in *R* was used to calculate the MLEs. The average of biases and the mean squared errors of all the estimators are reported in Tables 3 and 4. In particular, biases for the maximum likelihood estimators of α_1, α_2, α_3 and β are close to 0 and the mean squared errors of all estimators always decrease with increasing n.

Figure 1 shows 3D scatter plot of the simulation data for $\alpha_1 = 1, \alpha_2 = 2, \alpha_3 = 3$, $\beta = 2, k = 4, n = 50, 500$. Figure 2 shows 3D plot of the simulation data for $\alpha_1 = 1$, $\alpha_2 = 2, \alpha_3 = 3, \beta = 2, k = 4$. Figs. 3 and 4 show pairs style of the simulation data for $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 2, \beta = 2, k = 8, n = 50$ and $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 1, \beta = 2,$

Table 1 MLEs of parameters and correlation cofficients

α_1	α_2	α_3	β	k	n	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\beta}$	$Corr(X_1, X_2)$	$Corr(X_1, X_3)$	$Corr(X_2, X_3)$
1	2	3	2	4	50	1.2315	2.6113	3.9812	1.5833	−0.3135	−0.3761	−0.1916
					200	1.0519	2.1541	3.1909	1.8805	0.0315	−0.2767	−0.3253
					500	1.0200	2.0488	3.0889	1.9368	−0.1351	−0.0087	−0.2359
1	2	3	2	8	50	1.4617	3.1957	4.7645	1.5884	0.3833	−0.2242	−0.6268
					200	1.0911	2.2651	3.4170	1.8736	−0.3302	−0.5510	−0.5064
					500	1.0332	2.0814	3.1645	1.9528	−0.1628	−0.1630	−0.3739

Table 2 MLEs of parameters and correlation cofficients

α_1	α_2	α_3	β	k	n	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\beta}$	$Corr(X_1, X_2)$	$Corr(X_1, X_3)$	$Corr(X_2, X_3)$
2	2	1	2	4	50	2.5823	2.6287	1.2518	1.6702	0.0371	−0.3001	−0.0493
					200	2.2695	2.2684	1.1074	1.8370	−0.2307	−0.5114	0.1883
					500	2.0589	2.0384	1.0200	1.9507	−0.0746	−0.2110	−0.5513
2	2	1	2	8	50	3.1594	3.0801	1.4560	1.5814	−0.3203	−0.6072	−0.1092
					200	2.2567	2.3256	1.1060	1.9016	−0.3175	−0.2600	−0.1340
					500	1.0386	2.0907	3.1642	1.9729	−0.2865	−0.2239	−0.2340

Table 3 The average of biases and the mean squared errors of estimators

α_1	α_2	α_3	β	k	n	$Ab(\hat{\alpha}_1)$	$Ab(\hat{\alpha}_2)$	Ab($\hat{\alpha}_3$)	$Ab(\hat{\beta})$	$MSE(\hat{\alpha}_1)$	$MSE(\hat{\alpha}_2)$	$MSE(\hat{\alpha}_3)$	$MSE(\hat{\beta})$
1	2	3	2	4	50	0.2315	0.6112	0.9812	−0.4167	0.1969	1.3238	4.0505	0.1994
					200	0.0519	0.1541	0.1910	−0.1195	0.0276	0.2070	0.4197	0.0301
					500	0.0200	0.0488	0.0889	−0.0632	0.0094	0.0643	0.1793	0.0104
1	2	3	2	8	50	0.4617	1.1957	1.7635	−0.4116	0.6210	4.1107	8.8690	0.6040
					200	0.0911	0.2650	0.4170	−0.1264	0.0465	0.3430	1.0176	0.0458
					500	0.0332	0.0814	0.1644	−0.0472	0.0163	0.0930	0.3178	0.0160

Table 4 The average of biases and the mean squared errors of estimators

α_1	α_2	α_3	β	k	n	$Ab(\hat{\alpha}_1)$	$Ab(\hat{\alpha}_2)$	Ab($\hat{\alpha}_3$)	$Ab(\hat{\beta})$	$MSE(\hat{\alpha}_1)$	$MSE(\hat{\alpha}_2)$	$MSE(\hat{\alpha}_3)$	$MSE(\hat{\beta})$
2	2	1	2	4	50	0.5823	0.6287	0.2518	−0.3298	1.5689	1.6406	0.2319	1.4540
					200	0.2695	0.2684	0.1074	−0.1630	0.4725	0.4779	0.0697	0.4350
					500	0.0589	0.0384	0.0200	−0.0493	0.0640	0.0654	0.0108	0.0581
2	2	1	2	8	50	1.1594	1.0801	0.45560	−0.4186	4.2375	3.5325	1.0332	4.0178
				8	200	0.2567	0.3256	0.1060	−0.0984	0.3755	0.5124	0.0606	0.3522
				8	500	0.0386	0.0907	0.1642	−0.0271	0.0174	0.1017	0.3396	0.0173

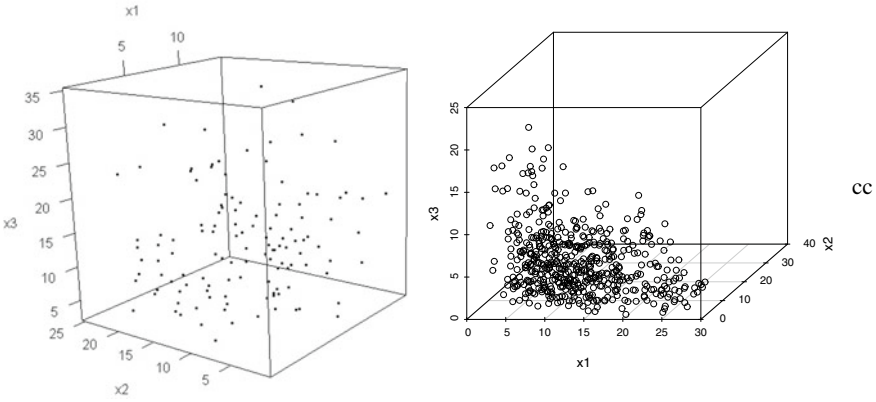


Fig. 1 3D scatter plot of simulation data with $\alpha_1 = 1, \alpha_2 = 2, \alpha_3 = 3, \beta = 2, k = 4, n = 50{,}500$

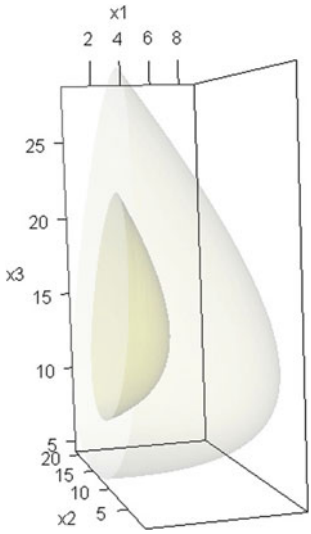


Fig. 2 3D plot for simulation data, $\alpha_1 = 1$, $\alpha_2 = 2, \alpha_3 = 3, \beta = 2$, $k = 4$

$k = 8, n = 500$, respectively. Figure 5 shows Trace plot for $\alpha_1 = 1, \alpha_2 = 2, \alpha_3 = 3$, $\beta = 2, k = 8, n = 500$. Finally, simulation points and 3D contour plot for different selected values of parameters are shown in Figs. 6, 7, 8 and 9.

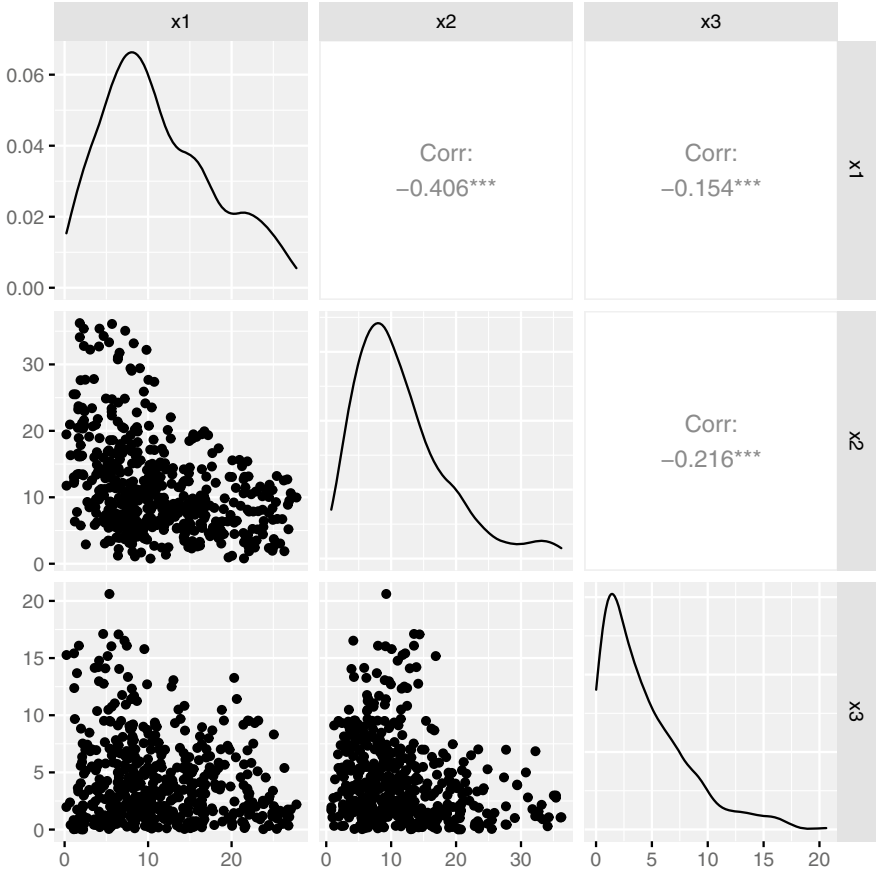


Fig. 3 Pairs plot for $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 1, \beta = 2, k = 8, n = 50$

10 Conclusion

In this chapter, a new multivariate gamma distribution whose marginals are finite mixtures of gamma distributions is defined. It is shown that the correlation between any pair of variables is negative. Therefore, the newly introduced distribution could be suitable for fitting multivariate data with negative correlations. Several of its properties such as joint moments, correlation coefficients, moment generating function, Rényi and Shannon entropies have been derived. In Sect. 8, the method of MLE has been applied to estimate the parameters. Because the resulting likelihood equations are nonlinear, numerical methods have been used to solve them. Simulation studies have been conducted to evaluate the performance of the maximum likelihood method. Moreover, various tables and figures have been provided to confirm a proper simulation and results of the MLE method for estimating the parameters.

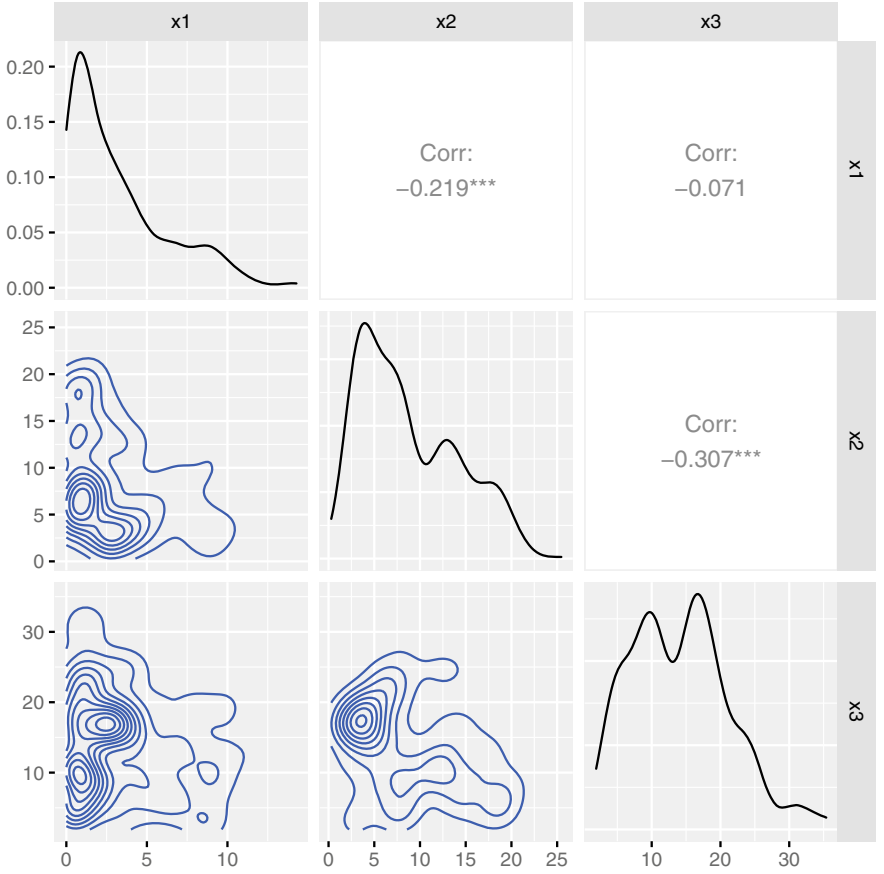


Fig. 4 Pairs plot for $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 1, \beta = 2, k = 8, n = 500$

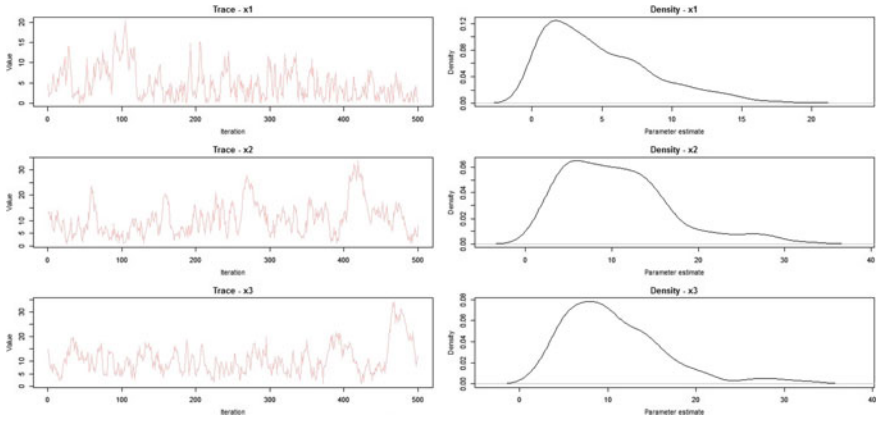


Fig. 5 Trace plots for $\alpha_1 = 1, \alpha_2 = 2, \alpha_3 = 3, \beta = 2, k = 8, n = 500$

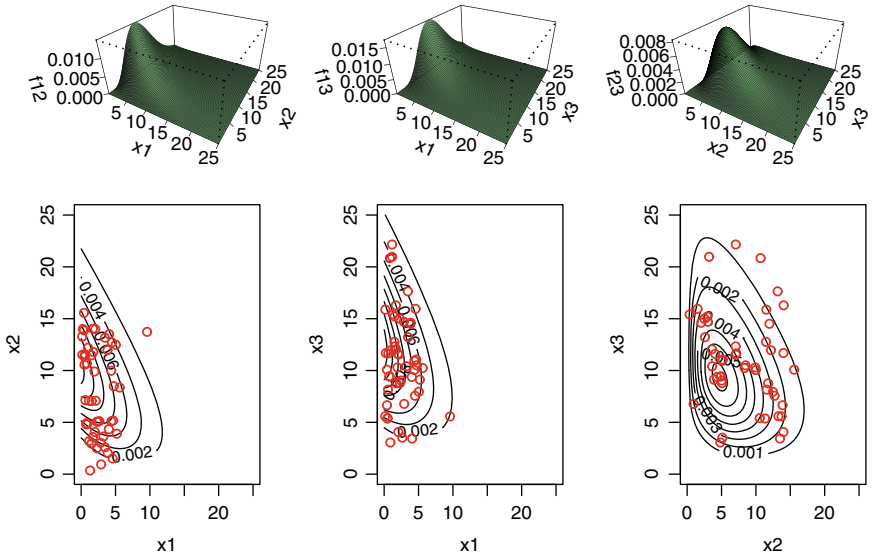
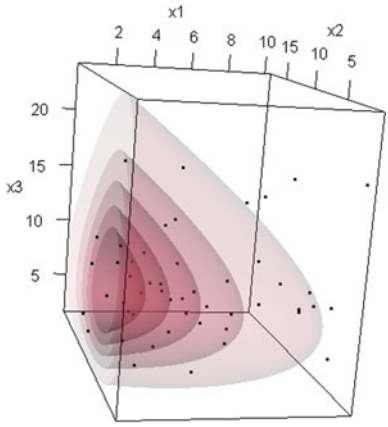


Fig. 6 Simulation points and 3D contour plot for $\alpha_1 = 1, \alpha_2 = 2, \alpha_3 = 3, \beta = 2, k = 4, n = 50$

Fig. 7 Simulation points and contour plot for $\alpha_1 = 1$, $\alpha_2 = 2, \alpha_3 = 3, \beta = 2$, $k = 4, n = 50$



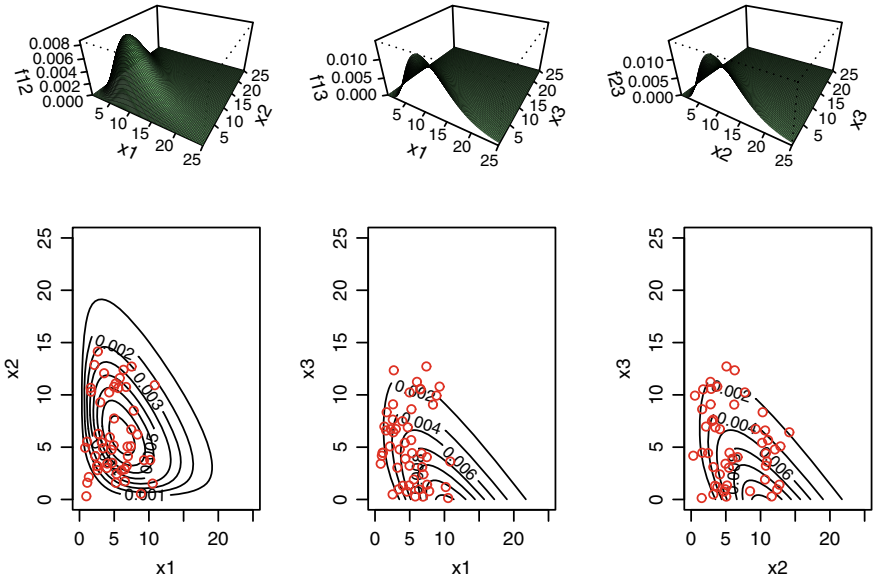


Fig. 8 Simulation points and contour plot for $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 1, \beta = 2, k = 4, n = 50$

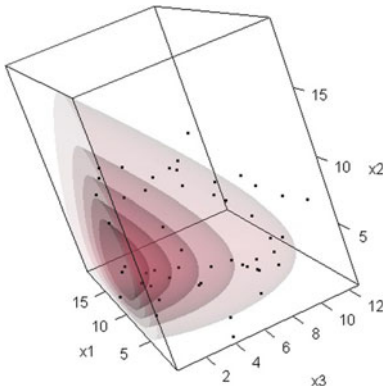


Fig. 9 Simulation points and 3D contour plot for $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 1, \beta = 2, k = 4, n = 50$

Acknowledgements Authors are grateful to the worthy reviewers for their constructive and helpful comments and suggestions.

Appendix

In this section, we give definitions and results that will be used in subsequent sections. Throughout this work we will use the Pochhammer symbol $(a)_n$ defined by $(a)_n = a(a+1)\cdots(a+n-1) = (a)_{n-1}(a+n-1)$ for $n = 1, 2, \ldots$, and $(a)_0 = 1$.

The fourth hypergeometric function of Lauricella, denoted by $F_D^{(n)}$, in n variables $z_1, \ldots, z_n$ is defined by

$$F_D^{(n)}(a, b_1, \ldots, b_n; c; z_1, \ldots, z_n) = \sum_{j_1, \ldots, j_n = 0}^{\infty} \frac{(a)_{j_1+\cdots+j_n}(b_1)_{j_1}\cdots(b_n)_{j_n} z_1^{j_1}\cdots z_n^{j_n}}{(c)_{j_1+\cdots+j_n} j_1!\cdots j_n!},$$

where $|z_i| < 1$, $i = 1, \ldots, n$. An integral representation of $F_D^{(n)}$ in Exton [7, p. 49, Eq. (2.3.5)] is given as

$$\begin{aligned} &F_D^{(m)}(a, b_1, \ldots, b_m; c; z_1, \ldots, z_m) \\ &= \frac{\Gamma(c)}{\prod_{i=1}^n \Gamma(b_i)\Gamma(c - \sum_{i=1}^n b_i)} \\ &\quad \times \int_{\substack{\sum_{i=1}^n x_i < 1 \\ 0 < x_i,\, i=1,\ldots,n}} \cdots \int \frac{\prod_{i=1}^n x_i^{b_i - 1}\left(1 - \sum_{i=1}^n x_i\right)^{c - \sum_{i=1}^n b_i - 1}}{\left(1 - \sum_{i=1}^n z_i t_i\right)^a} \mathrm{d}x_1 \cdots \mathrm{d}x_n. \end{aligned} \tag{15}$$

For further results and properties of this function the reader is referred to Exton [7] and Srivastava and Karlsson [32].

Let $f(\cdot)$ be a continuous function and $\alpha_i > 0$, $i = 1, \ldots, n$. The integral

$$D_n(\alpha_1, \ldots, \alpha_n; f) = \int_0^{\infty} \cdots \int_0^{\infty} \prod_{i=1}^n x_i^{\alpha_i - 1} f\left(\sum_{i=1}^n x_i\right) \prod_{i=1}^n \mathrm{d}x_i$$

is known as the Liouville-Dirichlet integral. Substituting $y_i = x_i/x$, $i = 1, \ldots, n-1$ and $x = \sum_{i=1}^n x_i$ with the Jacobian $J(x_1, \ldots, x_{n-1}, x_n \to y_1, \ldots, y_{n-1}, x) = x^{n-1}$ it is easy to see that

$$D_n(\alpha_1, \ldots, \alpha_n; f) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)} \int_0^{\infty} x^{\sum_{i=1}^n \alpha_i - 1} f(x)\, \mathrm{d}x. \tag{16}$$

Finally, we define the beta type 1, beta type 2 and Dirichlet type 1 distributions. These definitions can be found in Wilks [34], Fang, Kotz and Ng [8], Johnson, Kotz and Balakrishnan [15], and Kotz, Balakrishnan and Johnson [16].

Definition 2 A random variable X is said to have the beta type I distribution with parameters (a, b), $a > 0$, $b > 0$, denoted as $X \sim \mathrm{B1}(a, b)$, if its pdf is given by

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},\ 0<x<1.$$

Definition 3 A random variable X is said to have the beta type II distribution with parameters (a, b), denoted as $X \sim \mathrm{B2}(a, b)$, $a > 0$, $b > 0$, if its pdf is given by

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1+x)^{-(a+b)},\ x>0.$$

Definition 4 The random variables $U_1, \ldots, U_n$ are said to have a Dirichlet type 1 distribution with parameters $\alpha_1, \ldots, \alpha_n$ and α_{n+1}, denoted by $(U_1, \ldots, U_n) \sim \mathrm{D1}(\alpha_1, \ldots, \alpha_n; \alpha_{n+1})$, if their joint pdf is given by

$$\frac{\Gamma(\sum_{i=1}^{n+1}\alpha_i)}{\prod_{i=1}^{n+1}\Gamma(\alpha_i)}\prod_{i=1}^{n}u_i^{\alpha_i-1}\left(1-\sum_{i=1}^{n}u_i\right)^{\alpha_{n+1}-1},$$
$$0<u_i,\ i=1,\ldots,n,\ \sum_{i=1}^{n}u_i<1, \tag{17}$$

where $\alpha_i > 0$, $i = 1, \ldots, n+1$.

The Dirichlet type 1 distribution, which is a multivariate generalization of the beta type 1 distribution, has been considered by several authors and is well known in the scientific literature. By making the transformation $V_j = U_j/(1-\sum_{i=1}^{n}U_i)$, $j = 1, \ldots, n$, in (17), the Dirichelt type 2 density, which is a multivariate generalization of beta type 2 density, is obtained as

$$\frac{\Gamma(\sum_{i=1}^{n+1}\alpha_i)}{\prod_{i=1}^{n+1}\Gamma(\alpha_i)}\prod_{i=1}^{n}u_i^{\alpha_i-1}\left(1+\sum_{i=1}^{n}u_i\right)^{-\sum_{i=1}^{n+1}\alpha_i},\quad v_i>0,\ i=1,\ldots,n. \tag{18}$$

We will write $(V_1, \ldots, V_n) \sim \mathrm{D2}(\alpha_1, \ldots, \alpha_n; \alpha_{n+1})$ if the joint density of $V_1, \ldots, V_n$ is given by (18).

The matrix variate generalizations of beta type 1, beta type 2 and Dirichlet type 1 distributions have been defined and studied extensively. For example, see Gupta and Nagar [11].

Definition 5 Multinomial Theorem: For a positive integer k and a non-negative integer m,

$$(z_1+\cdots+z_m)^k=\sum_{k_1+\cdots+k_m=k}\binom{k}{k_1,\ldots,k_m}z_1^{k_1}\cdots z_m^{k_m},$$

where

$$\binom{k}{k_1,\ldots,k_m}=\frac{k!}{k_1!\cdots k_m!}.$$

The numbers appearing in the theorem are the multinomial coefficients. They can be expressed in numerous ways, including as a product of binomial coefficients of factorials:

$$\binom{k}{k_1, k_2, \ldots, k_m} = \frac{k!}{k_1!\,k_2!\cdots k_m!} = \binom{k_1}{k_1}\binom{k_1+k_2}{k_2}\cdots\binom{k_1+k_2+\cdots+k_m}{k_m}$$

Lemma 1 *For $a_1 > 0, \ldots, a_m > 0$ and $k \in \mathbb{N}$, we have*

$$k! \sum_{k_1+\cdots+k_m=k} \frac{(a_1)_{k_1}\ldots(a_m)_{k_m}}{k_1!\cdots k_m!} = (a_1+\cdots+a_m)_k = \frac{\Gamma(a_1+\cdots+a_m+k)}{\Gamma(a_1+\cdots+a_m)}.$$

Proof Writing $(1-\theta)^{-(a_1+\cdots+a_m)}$ as $(1-\theta)^{-a_1}\cdots(1-\theta)^{-a_m}$ and using power series expansion, for $0 < \theta < 1$, we get

$$(1-\theta)^{-a_1}\cdots(1-\theta)^{-a_m} = \sum_{k_1=0}^{\infty}\cdots\sum_{k_m=0}^{\infty}\frac{(a_1)_{k_1}\ldots(a_m)_{k_m}}{k_1!\cdots k_m!}\theta^{k_1+\cdots+k_n} = \sum_{k=0}^{\infty}\theta^k \sum_{k_1+\cdots+k_m=k}\frac{(a_1)_{k_1}\ldots(a_m)_{k_m}}{k_1!\cdots k_m!}$$

and

$$(1-\theta)^{-(a_1+\cdots+a_m)} = \sum_{k=0}^{\infty}\frac{(a_1+\cdots+a_m)_k}{k!}\theta^k.$$

Now, comparing coefficients of θ^k, we get the desired result. □

Lemma 2 *Let*

$$g(a_1,\ldots,a_m;\beta,k) = \int_0^{\infty}\cdots\int_0^{\infty}\prod_{i=1}^{m} z_i^{a_i-1}\left(\sum_{i=1}^{m} z_i\right)^k \exp\left(-\frac{1}{\beta}\sum_{i=1}^{m} z_i\right) dz_1\cdots dz_m, \tag{19}$$

where $a_1 > 0, \ldots, a_m > 0$ and $k \in \mathbb{N}$. Then

$$g(a_1,\ldots,a_m;\beta,k) = \beta^{\sum_{i=1}^{m} a_i+k}\left[\prod_{i=1}^{m}\Gamma(a_i)\right](a_1+\cdots+a_m)_k$$

Proof Expanding $\left(\sum_{i=1}^{m} z_i\right)^k$ in (19) by using multinomial theorem and integrating $z_1, \ldots, z_m$, we obtain

$$g(a_1,\ldots,a_m;\beta,k) = \sum_{k_1+\cdots+k_m=k} \binom{k}{k_1,\ldots,k_m} \prod_{i=1}^{m} \int_0^{\infty} z_i^{z_i+k_i-1} \exp\left(-\frac{1}{\beta} z_i\right) dz_i$$

$$= \beta^{\sum_{i=1}^{m}\alpha_i+k} \sum_{k_1+\cdots+k_m=k} \binom{k}{k_1,\ldots,k_m} \prod_{i=1}^{m} \Gamma(a_i+k_i).$$

Now, using Lemma 1, we get the desired result. □

References

1. Aksoy, H. (2000). Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences, 24,* 419–428.
2. Arnold, B. C., Castillo, E., & Sarabia, J. M. (1999). *Conditional specification of statistical models*. New York: Springer.
3. Balakrishnan, N., & C.-D, L. (2009). *Continuous bivariate distributions* (2nd ed.). Dordrecht: Springer.
4. Balakrishnan, N., & Ristić, M. M. (2016). Multivariate families of gamma-generated distributions with finite or infinite support above or below the diagonal. *Journal of Multivariate Analysis, 143,* 194–207.
5. Carpenter, M., & Diawara, N. (2007). A multivariate gamma distribution and its characterizations. *American Journal of Mathematical and Management Sciences, 27*(3–4), 499–507.
6. Dussauchoy, A., & Berland, R. (1975). A multivariate gamma type distribution whose marginal laws are gamma, and which has a property similar to a characteristic property of the normal case. In G. P. Patil, S. Kotz & J. K. Ord (Eds.), *Statistical Distributions in Scientific Work, Vol. 1: Models and Structures*, (pp. 319–328).
7. Exton, H. (1976). *Multiple hypergeometric functions and applications*. Chichester: Ellis Horwood.
8. Fang, K. T. D., Kotz, S., & Ng, K. W., *Symmetric multivariate and realated distributions*. London, New York: Chapman and Hall.
9. Furman, E. (2008). On a multivariate gamma distribution. *Statistics and Probability Letters, 78*(15), 2353–2360.
10. Gaver, D. P., Jr. (1970). Multivariate gamma distributions generated by mixture. *Sankhya: The Indian Journal of Statistics Series A, 32*(1), 123–126.
11. Gupta, A. K., & Nagar, D. K. (2000). *Matrix variate distributions*. Boca Raton: Chapman & Hall/CRC.
12. Gupta, A. K., & Song, D. (1996). Generalized Liouville distribution. *Computters & Mathematics with Applications, 32*(2), 103–109.
13. Gupta, R. D., & Richards, D. S. P. (2001). The history of Dirichlet and Liouville distributions. *International Statistical Review*, *69*(3), 433–446.
14. Hutchinson, T. P., & Lai, C. D. (1991). *The engineering statistician's guide to continuous bivariate distributions*. Adelaide: Rumsby Scientific Publishing.
15. Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2, 2nd Edn.). New York: Wiley.
16. Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous multivariate distributions-1* (2nd ed.). New York: Wiley.
17. Kowalczyk, T., & Tyrcha, J. (1989). Multivariate gamma distributions-properties and shape estimation. *Statistics: A Journal of Theoretical and Applied Statistics, 20*(3) , 465–474.
18. Krishnaiah, P. R., & Rao, M. M. (1961). Remarks on a multivariate gamma distribution. *The American Mathematical Monthly, 68*(4), 342–346.

19. Marcus, M. (2014). Multivariate gamma distributions. *Electronic Communications in Probability, 19,* 1–10.
20. Mathai, A. M., & Moschopoulos, P. G. (1991). On a multivariate gamma. *Journal of Multivariate Analysis, 39*(1), 135–153.
21. Mathai, A. M., & Moschopoulos, P. G. (1992). A form of multivariate gamma distribution. *Annals of the Institute of Statistical Mathematics, 44*(1), 97–106.
22. Nadarajah, S., & Zografos, K. (2005). Expressions for Rényi and Shannon entropies for bivariate distributions. *Information Sciences*, *170*(2–4), 173–189.
23. Peppas, K. P., Alexandropoulos, G. C., Datsikas, C. K., & Lazarakis, F. I. (2011). Multivariate gamma-gamma distribution with exponential correlation and its applications in radio frequency and optical wireless communications. *IET Microwaves Antennas & Propagation, 5*(3), 364–371.
24. Rafiei, M., Iranmanesh, A., & Nagar, D. K. (2020). A bivariate gamma distribution whose marginals are finite mixtures of gamma distributions. *Statistics, Optimization & Information Computing, 8*(4), 950–971.
25. Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. I, pp. 547–561). Berkeley, CA: University of California Press.
26. Robson, J. G., & Troy, J. B. (1987). Nature of the maintained discharge of Q, X, and Y retinal ganglion cells of the cat. *Journal of the Optical Society of America, A, 4,* 2301–2307.
27. Royen, T. (2007). Integral representations and approximations for multivariate gamma distributions. *Annals of the Institute of Statistical Mathematics, 59*(3), 499–513.
28. Semenikhine, V., Furman, E., & Su, J. (2018). On a multiplicative multivariate gamma distribution with applications in insurance. *Risks*, *6*(3), 79.
29. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(379–423), 623–656.
30. Sivazlian, B. D. (1981). On a multivariate extension of the gamma and beta distributions. *Siam Journal on Applied Mathematics*, *41*(2), 205–209.
31. Song, D., & Gupta, A. K. (1997). Properties of generalized Liouville distribution. *Random Operators and Stochastic Equations*, *5*(4), 337–348.
32. Srivastava, H. M., & Karlsson, P. W. (1985). *Multiple gaussian hypergeometric series*. New York: Wiley.
33. Vaidyanathan, V. S., & Lakshmi, R. V. (2015). Parameter estimation in multivariate gamma distribution. *Statistics, Optimization and Information Computing, 3,* 147–159.
34. Wilks, S. S. (1962). *Mathematical Statistics*. New York: Wiley.
35. Zografos, K. (1999). On maximum entropy characterization of Pearson's type II and VII multivariate distributions. *Journal of Multivariate Analysis, 71*(1), 67–75.
36. Zografos, K., & Nadarajah, S. (2005). Expressions for Rényi and Shannon entropies for multivariate distributions. *Statistics & Probability Letters, 71*(1), 71–84.

Aspects of High-Dimensional Methodology and Bayesian Learning

A Comparison of Different Clustering Approaches for High-Dimensional Presence-Absence Data

Gabriele d'Angella and Christian Hennig

Abstract Presence-absence data is defined by vectors or matrices of zeroes and ones, where the ones usually indicate a "presence" in a certain place. Presence-absence data occur, for example, when investigating geographical species distributions, genetic information, or the occurrence of certain terms in texts. There are many applications for clustering such data; one example is to find so-called biotic elements, i.e., groups of species that tend to occur together geographically. Presence-absence data can be clustered in various ways, namely, using a latent class mixture approach with local independence, distance-based hierarchical clustering with the Jaccard distance, K-modes, a density-based approach, or also using clustering methods for continuous data on a multidimensional scaling representation of the distances. These methods are conceptually very different from each other, and can therefore not easily be compared theoretically. We compare their performance with a comprehensive simulation study based on models for species distributions.

Keywords Multidimensional scaling · Biogeography · Cluster analysis · Simulation study · Benchmarking · Jaccard's distance

1 Introduction

Presence-absence data comprises observations that are vectors or matrices of zeroes and ones, where the ones usually indicate a "presence" in a certain place. Presence-absence data are often high-dimensional (the number of "places" may be large compared to the number of observations), and occur in a large range of applications, for example when investigating geographical species distributions, genetic information, or the occurrence of certain terms in texts.

G. d'Angella · C. Hennig (✉)
Alma Mater Studiorum - Università di Bologna, Via Zamboni, 33, 40126 Bologna, Italy
e-mail: christian.hennig@unibo.it

G. d'Angella
e-mail: gabriele.dangella2@unibo.it

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_13

Often there is an interest in clustering presence-absence data. An example for this is the search for "biotic elements", which are groups of species showing very similar distribution areas that can constitute evidence for the existence of areas of endemism generated by the formation of barriers over geologic time periods [15, 16].

Here we present a simulation study for comparing different approaches to clustering such data. This is based on a model for generating artificial but realistic presence-absence data with known clusters, the recovery of which can be compared. Not only are we interested in ranking methods, we also investigate characteristics of the data (such as degree of cluster overlap) that drive the methods' performances; the comparison of methods may strongly depend on such characteristics. We will also explain in detail why some methods do not perform that well. See [33] for guidelines on benchmark studies to compare clustering methods.

There are various approaches to clustering presence-absence data. Two major approaches are (1) the use of distance-based clustering methods such as average linkage or partitioning around medoids [13] using distance measures between the observations [8, 30], and (2) latent class mixture models [25]. Compared distance-based and latent class clustering omn categorical data in a simulation study [1]. Some of the most popular clustering methods such as K-means [22] and model-based clustering based on Gaussian mixtures [29] require Euclidean data and cannot directly be applied to presence-absence data. As approach (3), Hausdorf and Hennig [16, 17] have proposed to use methods for Euclidean data on the output of a Multidimensional Scaling (MDS) [5] specifically with presence-absence data for biotic element analysis. Furthermore, [3] generalized their density-based clustering in this way to categorical data.

MDS is a set of techniques that generate a Euclidean representation so that the resulting Euclidean distances approximate a given usually non-Euclidean distance structure. Low dimensionality of the representation is often desirable. Such a representation is often used for visualization (for which low dimensionality is needed), but can also be used for making distance data accessible to more elaborate statistical methodology that requires Euclidean input, as is done here for cluster analysis.

A disadvantage of this approach is that the MDS generally will lose some information in the original data (the lower the dimension, the larger the loss). A key interest in the present study is whether this approach can compete with the more direct approaches (1), and (2).

The relationship between clustering and scaling techniques has also been discussed in de Leeuw and Heiser [10] and in Kruskal [24]. Simultaneous use of MDS and clustering is treated by Desarbo [12] and Oh and Raftery [27]. Some biogeographical work related to our study has been done by Vavrik [34] (comparison of clustering methods for fossil data) and Ulrich and Gotelli [32] (null models for simulating presence-absence data).

Section 2 has a basic description of the data, distance, and MDS. Section 3 introduces the involved clustering methods. Section 4 describes in detail how the data for the simulations were generated. Section 5 discusses the results of the study, and Sect. 6 concludes the paper.

2 Data and Preprocessing

Data here are m-dimensional binary: $\mathbf{x}_i = (x_{i1}, \ldots, x_{im}),\ i = 1, \ldots, n$, for all j : $x_{ij} \in \{0, 1\}$. The data are modelled with a specific application area in mind, which is analyzing presence (1) and absence (0) of species in geographical regions, where i indexes the species and j indexes the region. Clusters are called "biotic elements" and are of interest because they provide insight into natural history [15, 17]. Our results may however also be informative for presence-absence data in other applications.

Methods are compared regarding their capability to retrieve true clusters in the data as set in the simulation. The cluster membership of the observations is denoted as $\mathbf{c} = (c_1, \ldots, c_n)$, where $c_i = k$ denotes that observation i is in cluster $k \in \{1, \ldots, K\}$.

Some of the compared clustering methods operate on distances. As distance we use the popular Jaccard distance [8, 11, 21], for discussion and alternatives see [8, 13, 18, 30]:

$$d_J(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{\sum_{j=1}^{m} \mathbb{1}(x_{1j} = 1 \wedge x_{2j} = 1)}{\sum_{j=1}^{m} \mathbb{1}(x_{1j} = 1 \vee x_{2j} = 1)} \tag{1}$$

where $\mathbb{1}$ denotes the indicator function.

A topic of key interest here is the performance of clustering techniques for Euclidean data that use an MDS representation of the distances as input. In this study, two MDS methods were applied: classical scaling [31] as implemented in the R-function `cmdscale`, and ratio MDS [5, 6] as implemented in the R-package `smacof`, function `mds`. Two and three-dimensional MDS solutions were used in the simulation study. Although this seems low, the results (as provided in Sect. 5) were good enough that there is not much room for improvement by higher dimensional solutions (in [27], lower dimensional proxies yielded clustering solutions almost as good as those exploiting higher dimensional configurations). An advantage of using a low-dimensional MDS is the ease of visualization of the data. However, in more general applications, using a higher dimensional MDS solution may be helpful in case that too much information is lost by a low-dimensional solution.

However, there are no indications against the setting of $p > 3$, p being the dimension of the MDS solution, when it comes to the application of techniques like K-means or Gaussian mixture models and this should be considered in case the comparison with other clustering recovery methods shows that too small a p might lose too much information. In this project we deal with multivariate binary data: not only does MDS constitute a powerful visualization tool, but it also enables us to resort to methods requiring continuous data input for our clusters recovery purposes. These methods are introduced in the next subsection.

3 Clustering Methods

We used the following clustering methods.

Latent Class Analysis

Latent class analysis operates directly on the presence-absence data. The method used here is the simplest of those that are referred to in the literature as "latent class analysis" [25, Chap. 3]. It models the data as a mixture of locally independent Bernoulli distributions, i.e., the different variables/regions are assumed independent within clusters, although they can depend on the whole dataset. Data are modelled as i.i.d. according to the density [4, Chap. 6]:

$$f_\eta(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{m} \theta_{jk}^{x_{ij}} (1-\theta_{jk})^{1-x_{ij}}, \tag{2}$$

where θ_{jk} is the probability of positive response for variable j in cluster k, namely the probability that a species in the k^{th} group inhabits the j^{th} cell , and π_k is the proportion of cluster k. The parameter vector $\eta = ((\pi_1, \theta_{11}, \ldots, \theta_{1m}), \ldots, (\pi_K, \theta_{1K}, \ldots, \theta_{mK}))$ stores mixture proportions and probabilities for presence per variable and cluster. The parameters can be fitted by Maximum Likelihood, using the EM-algorithm [35]. Observations can then be assigned to clusters by maximizing the estimated posterior probability of the observations belonging to the clusters [4, 25]. The Bayesian Information Criterion (BIC, [28]) can be used to estimate the number of clusters, but in our study the number of clusters was taken as fixed and known. In our study we used the R-function `poLCA` in the R-package with the same name for computing this kind of latent class analysis.

Methods Operating on Distances

Some methods were used that take the Jaccard distance matrix as input, namely, standard Single, Average, and Complete Linkage clustering [13, Chap. 4] with the dendrogram cut in such a way that the required number of clusters is produced. These were computed using the R-function `hclust`. Furthermore we used Partitioning Around Medoids (PAM; [23]) as computed by the R-function `pam`, and the K-modes algorithm [19], which tries to optimize the PAM objective function using the simple matching distance as computed by R-function `kmodes` in package `klaR` with parameter `fast` set to FALSE to avoid error messages.

Methods Operating on Euclidean Data

For clustering the Euclidean MDS-output, we used two of the most popular clustering methods, namely K-means [22] and the Gaussian mixture model [13, Chap. 6]. K-means was computed by the R-function `kmeans` (using parameters `nstart = 100, iter.max = 100` in order to allow for a more stable performance than granted by the default values), the Gaussian mixture model was fitted by the function `Mclust` in package `mclust` [29]. More precisely, data is assumed to be generated by a model with density

$$f_\eta(\mathbf{x}_i^{\mathtt{mds}}) = \sum_{k=1}^{K} \pi_k \phi_{\mathbf{a}_k, \Sigma_k}(\mathbf{x}_i^{\mathtt{mds}}), \tag{3}$$

where K is the number of mixture components, $\phi_{\mathbf{a},\Sigma}$ is the density of the multivariate Gaussian distribution with mean vector $\mathbf{a}$ and covariance matrix Σ, $\pi_1, \ldots, \pi_K$ are the mixture proportions, and η is a vector collecting all parameters. These parameters can be estimated by Maximum Likelihood as implemented in the EM-algorithm. `mclust` provides several constrained covariance matrix models (e.g., all covariance matrices equal) besides a fully flexible model, and the `mclust`—software uses the Bayesian Information Criterion (BIC) to select the best one [29]. Observations are assigned to clusters by estimating the probability that an observation was generated by cluster k:

$$\hat{P}(c_i = k|\mathbf{x}_i^{\mathtt{mds}}) = \frac{\hat{\pi}_k \phi_{\hat{\mathbf{a}}_k \hat{\Sigma}_k}(\mathbf{x}_i^{\mathtt{mds}})}{f_{\hat{\eta}}(\mathbf{x}_i^{\mathtt{mds}})}, \tag{4}$$

where the hat denotes estimated parameters, and estimating the cluster membership labels as follows:

$$\hat{c}_i = \underset{k\in\{1,\ldots,K\}}{\arg\max}\, \hat{P}(c_i = k|\mathbf{x}_i^{\mathtt{mds}}). \tag{5}$$

The BIC is also often used to estimate the number of clusters, but this is treated as fixed and known here, which also means that there is no need to use one of the many available methods (that often lead to contradictory results) to decide the number of clusters for the other methods [14].

Furthermore, [3] have proposed to run the density-based clustering method `pdfCluster` (R-function and package name) [2] on MDS-output in order to handle categorical or binary data. The method estimates the density by kernel methods. Clustering is then performed by finding density level sets using neighborhood graphs. As opposed to the other methods in the study, pdfCluster will implicitly determine the number of clusters rather than allowing for it to be fixed. In order to avoid errors, we used the parameter settings `graphtype="pairs"` and `n.grid` equal to the number of observations.

K-means, the Gaussian mixture model, and pdfCluster are combined with the different MDS techniques and two dimensionalities of the MDS output. The combinations are called kc2, ks2, kc3, ks3, gc2, gs2, gc3, gs3, pdfc2, pdfs2, pdfc3, pdfs3 in the results section. "k", "g", "pdf" refer to K-means, the Gaussian mixture, and pdfCluster, respectively. "c" and "s" denote `cmsscale`, and `smacof`'s `mds`, respectively. The number in the end is the number of MDS dimensions used.

4 The Simulation

In order to compare the methods, we simulate datasets of n species, each of which is represented by a presence-absence vector of length m, and cluster them.

Simulations enable us to investigate the features of the data that drive the performance of the clustering methods. Real biogeographical data sets can make this assessment difficult, as it is hard to separate results due to singularities in data from those due to methods properties [7]. Among the plethora of aspects that can be considered, in this project we examined the level of clusters' overlap, the cluster size, the width of the cluster specific areas and the number of clusters: in particular, we compared a situation with three clusters endowed with their specific geographic area with a situation where on top of these three groups a cluster of universal spreaders is added, as often exist in reality [9]. Such a cluster consists of species that are widespread across the whole map. These species are not informative regarding biotic elements, however they qualify as clusters in a data analytic sense, namely as a set of observations that behaves in a certain way. In the following, we will call the clusters made up of species that are *not* universal spreaders "proper", because these species inhabit a specific group of cells and can signal the existence of a biotic element.

The 24 parameter combinations for the simulated data are described below, after the presentation of the algorithm that was used to generate this kind of data.

Data Generation

The algorithm implemented to simulate each species resorts repeatedly to a random draw from a distribution that chooses one category of the m possible ones, each of which has its own probability of being picked. At each step, one category has to be picked, and therefore these probabilities sum to one. This distribution is called here "categorical distribution". Let $j = 1, \ldots, m$ be the categories. Each category has a probability p_j of being picked and $\sum_{j=1}^{m} p_j = 1$. The probability mass function of our categorical distribution is:

$$f_{\mathbf{p}}(y) = \prod_{j=1}^{m} {p_j}^{[y=j]} \tag{6}$$

where $[y = j]$ is equal to 1 if $y = j$ and 0 otherwise.

The species in our data set inhabit the geographic units (cells) of a map. To each of the m categories corresponds a cell: at every drawing step, a category is picked and its associated binary indicator is set to 1, meaning that the species is present in this geographic unit of the map. A value of 0 means that the cell has not been drawn and the species is absent.

Each cluster will have its specific area on the map, consisting of a set of cells whose probability to be inhabited by the species of the cluster will be larger than elsewhere. To simplify, there will be a unique probability inside these areas (p_{in}) and a unique probability outside of them (p_{out}), with $p_{in} > p_{out}$. Once the cluster specific areas have been determined, the probability vectors specific to each cluster can be produced. Species belonging to a given cluster will mainly inhabit the cluster specific area, while their occurrence outside of it will be rarer. The rareness of species outside of their specific cluster area will depend on the level of overlap ω, see below.

We define the following notation:

(a) observations $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$ taking values in the product set $\{0, 1\}^m$;
(b) cluster labels $k = 1, \ldots, K$;
(c) species range $r_i = \sum_{j=1}^{m} x_{ij}$, for $i = 1, \ldots, n$;
(d) vector of species cluster labels $\mathbf{c} = (c_1, \ldots, c_n)$;
(e) vector of cell cluster labels $\mathbf{d} = (d_1, \ldots, d_m)$. A cell is given the label k when it belongs to the specific area of a cluster, i.e., the set of cells with elevated probability for cluster k. The probability to be inhabited by a species belonging to that cluster is p_{in}. Each cell can belong to at most one (exactly one in this study) cluster specific area.

The following parameters of the simulation are fixed in advance:

- the number of species n and the number of cells m. n was either 300 or 400, see below. m was 60;
- the level of overlap $\omega = \frac{p_{out}}{p_{in}}$. Three levels of overlap are used, i.e., low ($\omega = 0.05$), medium ($\omega = 0.2$) and high ($\omega = 0.4$);
- the number of clusters K. This was either 3 (with $n = 300$), or 4 (with $n = 400$). There were always 3 proper clusters, and a cluster of universal spreaders was added or not;
- the number of species belonging to each cluster k, i.e., $|\{i : c_i = k\}|$ for $k = 1, \ldots, K$. Two possibilities are investigated: one with all clusters having 100 observations and one with clusters of various sizes (a small clusters with 50 species, one with 100 and one with 150 species). If a group of universal spreaders is present, it consists of 100 observations;
- the number of cells constituting each cluster specific area on the map, i.e., $|\{j : d_j = k\}|$ for $k = 1, \ldots, K$. Two possibilities are investigated, one with equal-sized specific areas (all areas 20 cells) and one with heterogeneous specific areas (10, 20, and 30 cells). With unequal area sizes and unequal cluster sizes, cluster sizes were chosen in line with the area sizes, i.e., the larger the cluster specific area, the larger the cluster.

The range r_i of the i^{th} species is instead randomly drawn. If the species belongs to a proper cluster, it is uniformly drawn between 1 and the size of the cluster specific area. If it is a universal spreader ($u.s.$), its range is at least as large as the widest cluster specific area on the map, namely

$$r_{u.s.} = \max_k |\{j : d_j = k\}|,$$

and the maximum is m. As a consequence, a universal spreader will always cover a region of the map that involves more cells than the habitat of any proper species. These choices have been made inspired by real datasets of this kind that we have seen; particularly, species ranges are highly variable.

Here is the algorithm for the i^{th} species draw, provided that it is not a universal spreader:

1. pick r_i according to a discrete uniform distribution with domain $\{1, \ldots, |\{j : c_i = d_j\}|\}$;
2. set $\mathbf{x}_i^0 = \mathbf{0}_m$, namely the observation is initialized as a vector of zeros of length m;
3. for $t = 0, \ldots, r_i - 1$:
   a. compute p^t_{in,c_i} and p^t_{out,c_i} such that:
   $$\begin{cases} \frac{p^t_{out,c_i}}{p^t_{in,c_i}} = \omega \\ |\{j : d_j = c_i \wedge x^t_{ij} = 0\}| \cdot p^t_{in,c_i} + |\{j : d_j \neq c_i \wedge x^t_{ij} = 0\}| \cdot p^t_{out,c_i} = 1 \end{cases}$$
   b. specify the vector $\mathbf{p}_i^t$ according to the following rule:
   $$p^t_{ij} = \begin{cases} p^t_{in,c_i}, & \text{if } c_i = d_j \wedge x^t_{ij} = 0 \\ p^t_{out,c_i}, & \text{if } c_i \neq d_j \wedge x^t_{ij} = 0 \\ 0 & \text{otherwise} \end{cases}$$
   c. draw j from a categorical distribution (Eq. 6) with probability vector $\mathbf{p}_i^t$;
   d. set $x^{t+1}_{ij} = 1$.
4. set the i^{th} observation in the data set equal to $\mathbf{x}_i^{r_i}$.

Note that the probability vector $\mathbf{p}_i^t$ is updated after each draw in such a way that every category can only be drawn once. Step 3 makes sure that ω (i.e., the chosen ratio between the probability to draw a cell without and within the cluster specific area) is respected by the probability vector $\mathbf{p}_i^t$. Species generated in this way may well have presences outside the cluster specific area, and absences within it, although the probability to be present within it is larger. This reflects realistically observed patterns.

The algorithm for the draw of the i^{th} observation being a universal spreader is slightly different:

1. set $\mathbf{x}_i^0 = \mathbf{0}_m$, namely the observation is initialized as a vector of zeros of length m;
2. pick r_i according to a discrete uniform distribution with domain $\{r_{u.s.}, \ldots, m\}$;

Table 1 The 24 simulation scenarios resulting from the combination of the investigated data parameters

Scenario	ω	K	Sizes	Areas	Case	ω	K	Sizes	Areas
1	0.05	3	=	=	13	0.05	3	=	$\neq$
2	0.20	3	=	=	14	0.20	3	=	$\neq$
3	0.40	3	=	=	15	0.40	3	=	$\neq$
4	0.05	3+$u.s.$	=	=	16	0.05	3+$u.s.$	=	$\neq$
5	0.20	3+$u.s.$	=	=	17	0.20	3+$u.s.$	=	$\neq$
6	0.40	3+$u.s.$	=	=	18	0.40	3+$u.s.$	=	$\neq$
7	0.05	3	$\neq$	=	19	0.05	3	$\neq$	$\neq$
8	0.20	3	$\neq$	=	20	0.20	3	$\neq$	$\neq$
9	0.40	3	$\neq$	=	21	0.40	3	$\neq$	$\neq$
10	0.05	3+$u.s.$	$\neq$	=	22	0.05	3+$u.s.$	$\neq$	$\neq$
11	0.20	3+$u.s.$	$\neq$	=	23	0.20	3+$u.s.$	$\neq$	$\neq$
12	0.40	3+$u.s.$	$\neq$	=	24	0.40	3+$u.s.$	$\neq$	$\neq$

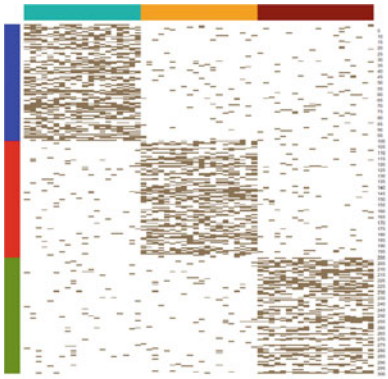
3. for $t = 0, \ldots, r_i - 1$:
   a. pick j according to a discrete uniform distribution with domain $\{1, \ldots, m\} \setminus \{j : x_{ij}^t = 1\}$;
   b. set $x_{ij}^{t+1} = 1$.
4. set the i^{th} observation in the data set equal to $\mathbf{x}_i^{r_i}$.

Scenarios

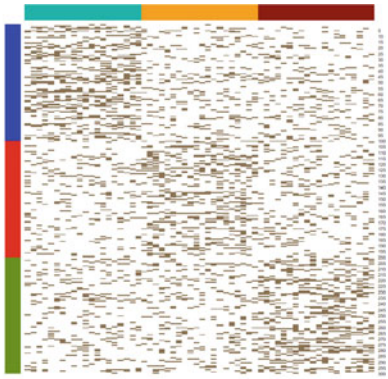
24 simulation scenarios have been used, which are listed in Table 1.

Equal (=) group sizes imply that there are 100 observations per cluster (be it a proper one or a group of universal spreaders). When the group sizes are different ($\neq$), there is a cluster with 50 species, one with 100, and one with 150 (there are always 100 $u.s.$ species). When the areas are equal (=), the 60 cells are evenly divided across the clusters; when the areas differ ($\neq$), the smallest cluster specific area is made up of 10 cells/variables and the biggest one consists of 30 cells/variables.

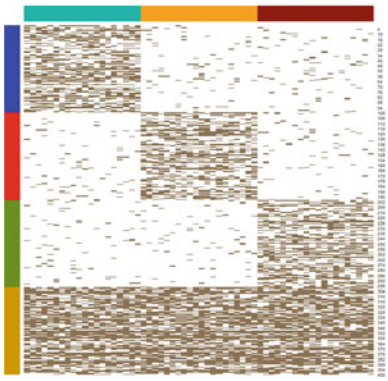
We visualize the spread of the species in four exemplary cases using heatmaps. They represent the species in the rows and the cells in the columns: in each row, grey units represent cells where the species is present. In Fig. 1a the overlap is low and all groups have the same number of members and cluster specific cells. In Fig. 1b the high level of overlap makes it rather tough to distinguish the cluster specific areas. Case 4 (Fig. 1c) differs from Case 1 in that it includes a group of universal spreaders. In Case 13 (Fig. 1d), clusters have specific areas of differing sizes, therefore species belonging to the green cluster will be able to inhabit up to thirty cells.



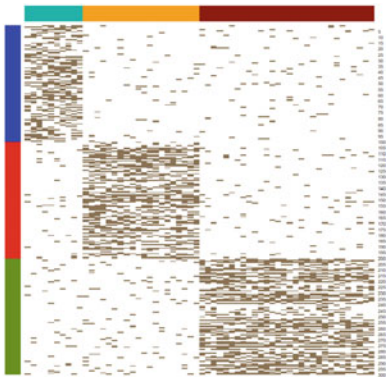
(a) Heatmap for case 1: no $u.s.$ group, equal sizes, equal areas, low overlap.



(b) Heatmap for case 3: no $u.s.$ group, equal sizes, equal areas, high overlap



(c) Heatmap for case 4: with $u.s.$ group, equal sizes, equal areas, low overlap.



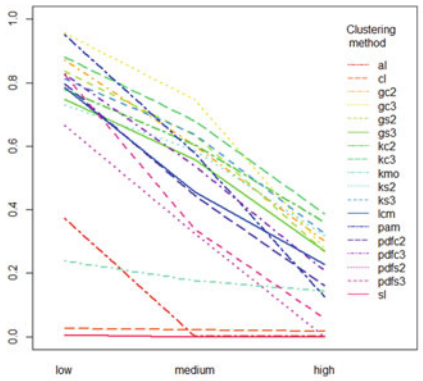
(d) Heatmap for case 13: no $u.s.$ group, equal sizes, different areas, low overlap.

Fig. 1 Presence-absence heatmaps for cases 1, 3, 4 and 13. Rows correspond to species, columns correspond to cells. Colored bars represent species clusters (rows) and cluster specific areas (columns)

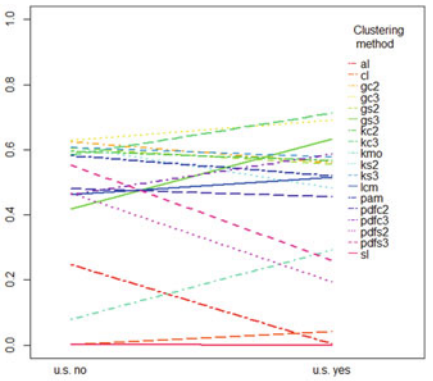
5 Results

General Results

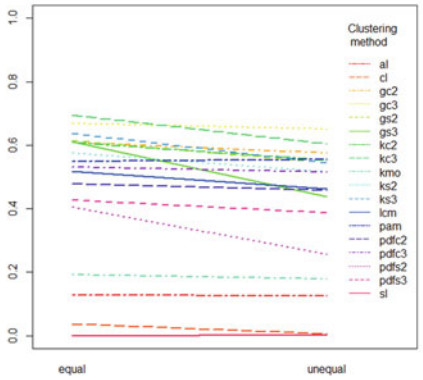
The simulations were evaluated using the Adjusted Rand Index [20] in order to compare clusterings generated by the methods to the true clusterings, a measure of similarity between data clusterings that ranges from -1 to 1: two independent random partitions are expected to return an ARI of zero, whereas an ARI equal to 1 implies identical clusterings. Not only did we try to investigate what methods led



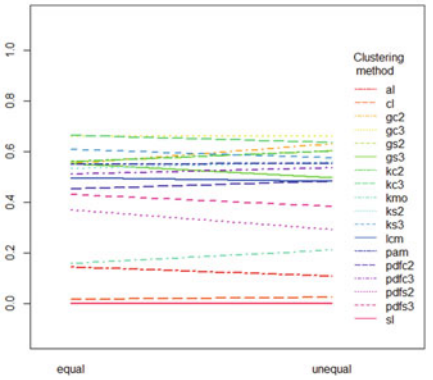
(a) Interaction plot showing the average ARI values (vertical axis) for the three **levels of overlap** (ω).



(b) Interaction plot showing the average ARI values (vertical axis) depending on the presence (yes) or absence (no) of **universal spreaders** in the simulation setup.



(c) Interaction plot showing the average ARI values (vertical axis) with equal or unequal **cluster sizes**.



(d) Interaction plot showing the average ARI values (vertical axis) with equal or unequal **cluster specific area widths**.

Fig. 2 Interaction plots of the average ARI values in the 24 cases (each simulated ten times). The lines referring to each of the twelve clustering recovery methods are indicated in the legends

to a satisfactory clustering recovery, but we also checked what data features had a significant impact on the methods' performance.

Each of the 24 scenarios reported in Table 1 was simulated ten times. This configuration generated a data set with 3120 rows: thirteen methods applied on 24 scenarios, each simulated ten times.

Figure 2 shows interaction plots of the results. They show the ARI means on the vertical axis and the levels of the various data features on the horizontal axis. The different lines in the plots refer to the clustering methods. Figure 3 shows boxplots

of the ARI values achieved by the methods by overlap and presence of universal spreaders.

Unsurprisingly, methods are generally better with lower overlap, while some methods are helped whereas others are harmed by universal spreaders. Overall patterns regarding cluster sizes and cluster specific area widths are not that striking.

The performance of hierarchical clustering is generally rather bad, and clearly dominated by the other methods. Average linkage performs somewhat better than the other two methods for low overlap and without universal spreaders, but it is still worse than all the non-hierarchical methods; in fact the interaction plots could be interpreted as showing two clusters of methods, namely the hierarchical methods achieving low ARIs, and most other methods achieving better ARIs with K-modes and pdfs2 in between. We explore the reasons for this in Sect. 5.

K-modes performs a bit better than the hierarchical methods, but worse than the other methods. It becomes mostly better with universal spreaders. In particular, the use of the simple matching distance seems to be a worse choice for this kind of data than Jaccard for PAM.

The latent class analysis belongs to the cluster of better methods, but it cannot compete with K-means and the Gaussian mixture after MDS, as can be seen in particular in Fig. 2d. It does absolutely and relatively better with low overlap. PAM is among the best methods with low overlap, but deteriorates even stronger than the latent class analysis with more overlap.

Out of the different variants of K-means and the Gaussian mixture model, regarding the overall average ARI those based on a 3-dimensional classical MDS (kc3, gc3) perform best, as can be seen in Fig. 2c, d. There is considerable variation over the different experimental factor levels though. Figure 3 shows that gc2 is excellent with low overlap and without universal spreaders. kc3 performs much better with universal spreaders and does not stand out without them. Using the ratio (`smacof`) MDS is mostly worse than using classical MDS. Particularly it does not work well with the Gaussian mixture model and 3-dimensional MDS output with low and medium overlap and no universal spreaders, although universal spreaders help that method, particularly with medium overlap. Unequal sizes of cluster specific areas generally favor the classical MDS. The combination of medium overlap and universal spreaders generally favours the 3-dimensional MDS strongly, which otherwise has a small but not so substantial advantage. The K-means methods are overall competitive, but relatively weaker compared to the Gaussian mixture model for unequal cluster sizes. High overlap favours the K-means methods. pdfCluster is overall not quite as good as K-means and the Gaussian mixture. Overall it has a somewhat similar performance to the latent class analysis, with pdfc3 doing mostly better and the other three versions doing mostly worse. The impossibility to fix the number of clusters for pdfCluster puts it at a slight disadvantage, as the correct number of clusters was fixed for the other methods. pdfCluster mostly found the correct number of clusters, but occasionally (particularly with the `smacof`-MDS) it would put everything together in a single cluster.

Figure 3 gives some information about the variability of the results. pam and gc3 have most stable results for low overlap (without universal spreaders also gc2),

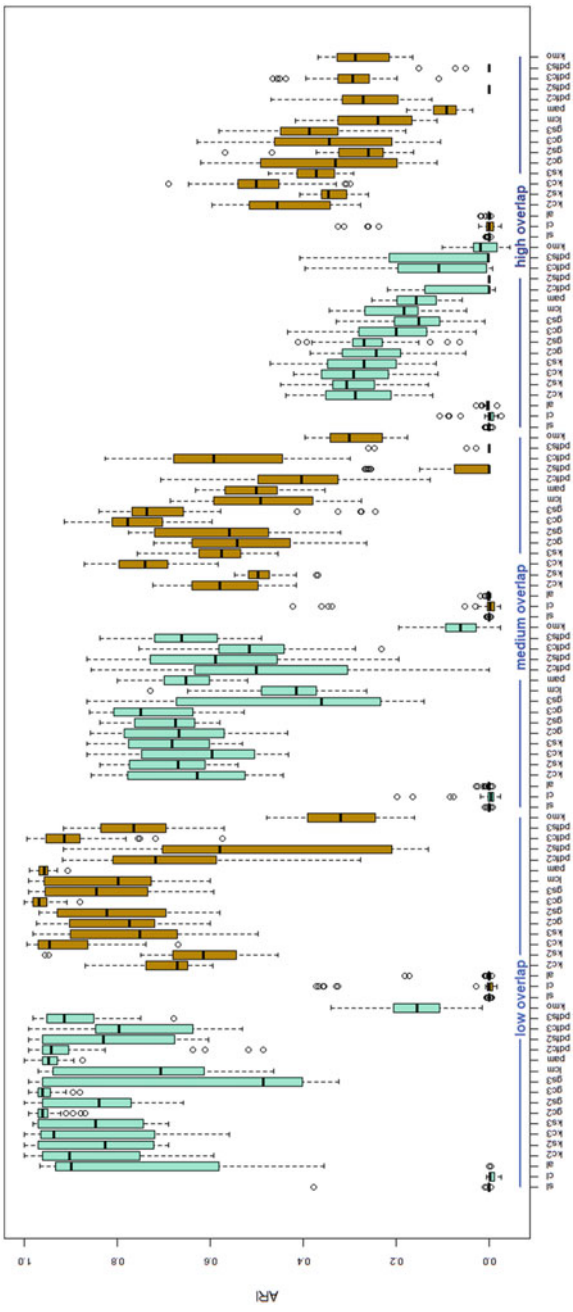


Fig. 3 Boxplots of the ARIs achieved by the clustering methods, grouped by level of overlap and presence of a cluster of universal spreaders (sea-green boxplots: no *u.s.* group; golden boxplots: with *u.s.* group)

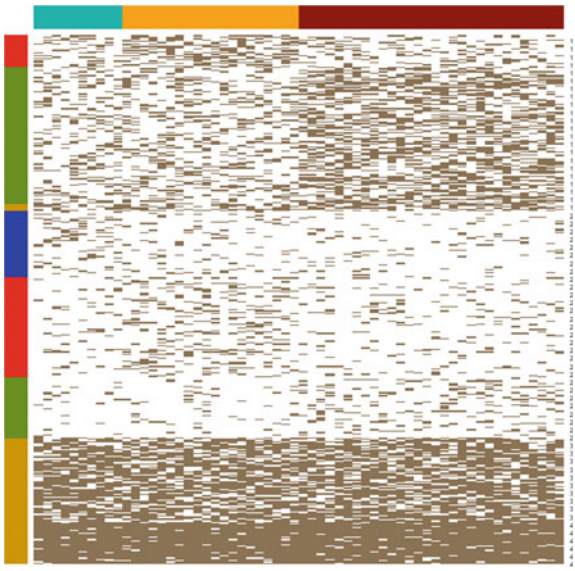


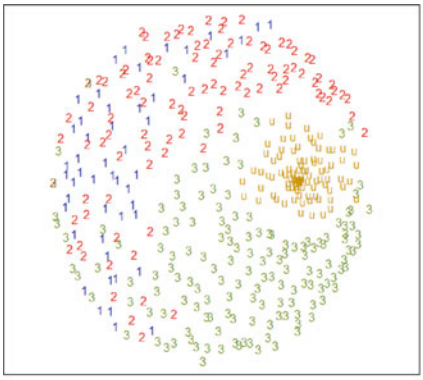
Fig. 4 Heatmap from a simulation of scenario 24: high overlap, with *u.s.* group, various sizes and areas. ARI = 0.3538. Rows sorted by LCA clustering. True cluster membership on the left (blue = 1, red = 2, green = 3, gold = *u.s.*), cluster specific areas on top (light-blue = 1, orange = 2, dark red = 3), LCA cluster membership on the right (numbers)

whereas other methods, gs3 in particular, have substantial variation. Otherwise stability can mostly be found among bad results (e.g., the linkage methods).

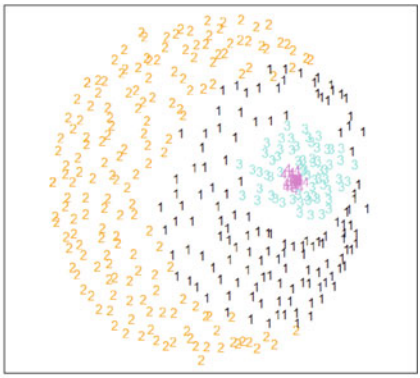
More Detailed Insight

In order to explore behavior of some methods in more detail and to understand their weaknesses, we use some exemplary visualizations.

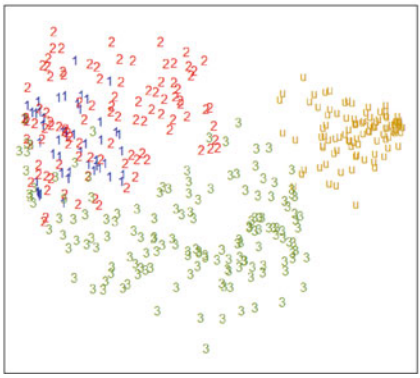
For examining the performance of latent class analysis, we consider a simulation from scenario 24 of Table 1 with high overlap and different cluster sizes. The heatmap in Fig. 4 shows the four cluster latent class solution indirectly by the order of rows; colors on the side denote the true clusters, colors on top denote the cluster specific areas. A typical behavior of latent class analysis with large overlap was to build a cluster with sparse species, i.e., with small ranges, regardless to which original cluster they belonged. The smaller true clusters were hard to identify. Latent class analysis treats presences and absences symmetrically, which leads to a tendency to group sparse species together; also species with large range are easily grouped together, which is a good thing with a universal spreader cluster. Therefore its existence helps the latent class analysis to often achieve a larger ARI. It can also happen, however, that the universal spreaders cluster is split up into two clusters (relatively larger and smaller ranges), as happened in Fig. 4. The symmetric treatment of presences and absences is not good for identifying the proper clusters, at least not with large overlap. These rely on presences much more than on absences, and therefore the focus of the Jaccard distance on joint presences rather than joint absences will help.



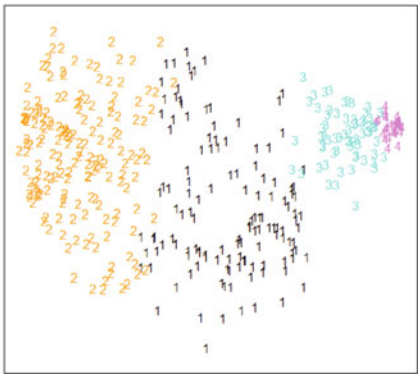
(a) Ratio `smacof` 2D mapping colored according to **true cluster membership**. The `u`'s are universal spreaders.



(b) Ratio `smacof` 2D mapping colored according **LCA clustering**.



(c) Classical scaling 2D configuration colored according to **true cluster membership**. The `u`'s are universal spreaders.



(d) Classical scaling 2D configuration colored according to **LCA clustering**.

Fig. 5 Ratio `smacof` MDS and classical scaling, based on the Jaccard's distance matrix computed on the data set simulated under scenario 24 (heatmap in Fig. 4), colored by true clustering (left) and LCA clustering (right)

Figure 5 shows two-dimensional MDS maps of these data, highlighting the true cluster memberships ((a) and (c)) and the latent class clusters ((b) and (d)), for classical and ratio (`smacof`) MDS. Although not based on the MDS input, the latent class solution looks geometrically sensible in terms of both MDS visualizations, actually more so than the true clusters. However, latent class clusters 1 and 2 are very heterogeneous and involve large distances, merging observations from two or more true clusters together. This is in line with the fact that for large overlap with universal

spreaders, starting from the MDS output, K-means, which prioritizes within-cluster homogeneity over separation and geometrically visible cluster shapes, does better than the Gaussian mixture model (Fig. 3). The true clusters 1 and 2 are apparently hard to separate for any method, but they are still more homogeneous than latent class cluster 2, which holds the majority of both of them.

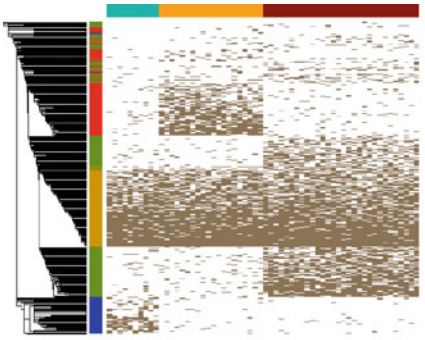
Figure 5 also serves for understanding why classical scaling performs better in our experiments. Geometrically, the ratio MDS is not worse at arranging the true clusters into a sensible shape, but both K-means and the Gaussian mixture model are connected to multivariate Gaussian distributions, which are characterized by linear relations between the variables, whereas the ratio MDS produces nonlinear cluster shapes and boundaries between true clusters. The ratio MDS in itself does not fail, but it does not serve as well as input for the Euclidean cluster methods. The classical MDS also separates the universal spreaders on one side of the plot, whereas in the ratio MDS they are, although also nicely separated, surrounded by observations from the proper clusters. Presence of universal spreaders helped some methods to achieve better results simply because they are often more easily "visible" as clusters to the methods.

Regarding the weak performance of the hierarchical methods, Fig. 6 shows three heatmaps from data from scenario 22, with low overlap, and universal spreaders, and clusters have different sizes and different widths of their specific areas. These are shown together with the dendrograms form the three hierarchical clusterings.

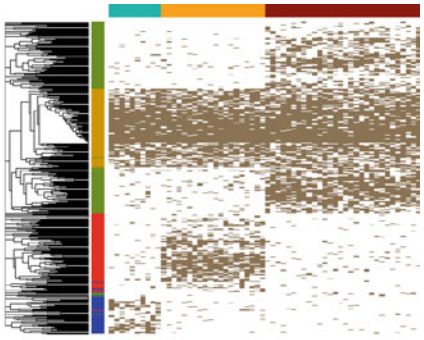
Single linkage (Fig. 6a) clearly suffers from its well known chaining issue [13, Chap. 4]. Fixing the number of clusters at 4, all observations are merged, and only a few very sparse species with a large Jaccard distance to all the other species are isolated as one-point clusters. Looking at the colors indicating the true clusters, the dendrogram as a whole is not perfect, but to some extent in line with the true clusters. A single clustering obtained by cutting the dendrogram at a certain height would however require a very large number of clusters to have at least the bigger ones of them in some agreement with the true clusters. The same phenomenon occurred with most other datasets, causing ARIs generally close to zero.

Complete linkage is not affected by chaining, but Fig. 6b shows a different problem. The Jaccard distance assigns its maximum value 1 to any two species that do not have joint presences. As occurs often in real data, the simulated data have several species with a small range that easily have no overlap. Complete linkage can't therefore join them in the same cluster, and they are still separated at the top level of the dendrogram. The dendrogram as a whole can be seen as even better in line with the true clustering, but in order to reflect this in a clustering with a fixed number of clusters, the method would need to integrate some sparse species with some clusters that include species to which they have the maximum Jaccard distance, and complete linkage cannot do that.

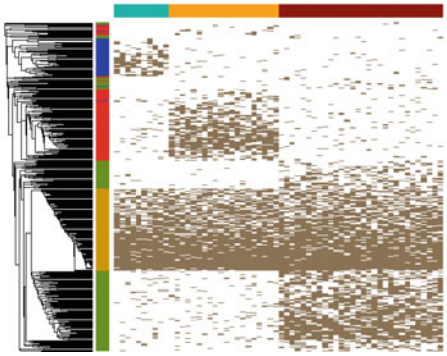
Unfortunately, average linkage (Fig. 6c) cannot solve both of these problems at the same time, but is rather also affected by chaining, isolating some sparse species when cutting for lower numbers of clusters. Once more, some useful agreement with the true clustering can be found in the dendrogram as a whole, but with a fixed



(a) Single linkage clustering. ARI= −0.0039.



(b) Complete linkage clustering. ARI= −0 0051



(c) Average linkage clustering. ARI= 0.0004.

Fig. 6 Heatmaps of a simulation of scenario 22 ($\omega = 0.05$, with *u.s.* cluster, various cluster sizes and areas), sorted according to the three hierarchical clustering solutions. Colors on the left side indicate true cluster membership of the species, while those on top refer to the cluster specific area of the cells

number of clusters 4, the ARI is about zero, and a too large number of clusters would be needed to achieve a substantially better value.

The best ARIs that could have been achieved by optimal cutting are 0.427 for single linkage (181 clusters), 0.469 for complete linkage (35 clusters), and 0.522 for average linkage (81 clusters). These are still lower than those for the MDS-based methods, latent class analysis, and PAM, largely due to the many small clusters involved. Still it shows that the results of the hierarchical methods are ultimately also connected to the true clusterings, though this requires looking deeper into the dendrogram.

A major lesson to learn from these insights regards the role of the Jaccard distance. On one hand, it is advantageous by treating presences and absences in an asymmetric manner. The latent class analysis suffers from treating them symmetrically, compared to the MDS-based methods. On the other hand, it also creates problems, particularly

for the hierarchical methods, by assigning a maximum distance to pairs of sparse species that have no cell in common. Hausdorf and Hennig [16, 17] have preferred the Kulczynski dissimilarity, which is not a metric, but this also assigns maximum dissimilarity if there is no presence in common, although it may help by assessing sparse species as closer to species with a larger range of which they are a subset. Hennig and Hausdorf [18] propose to involve geographical distances into the distance computation, and this may indeed improve matters, if only in a way that depends strongly on the geography of the specific real data, which will make setting up general models for simulation a more difficult task.

6 Conclusions

We have run a simulation study in order to evaluate the clustering performance, on multivariate presence-absence data, of a combination of multidimensional scaling and clustering methods for quantitative data (K-means, Gaussian mixture models, and pdfCluster), compared with methods that either operate directly on the data (such as the latent class analysis), or on a Jaccard or simple matching distance matrix computed from the data (hierarchical methods, PAM, and K-modes). Two different MDS methods and MDS outputs of dimensionality 2 and 3 were involved.

The results suggest that the MDS-based techniques can be a valuable tool to cluster such data, returning by and large better ARI values than the competitors. The hierarchical methods did particularly badly, mainly due to the fact that cutting the dendrogram at the correct number of clusters proved inadequate; however it would be hard to repair this problem by any automatic method for deciding the number of clusters, because the required number of clusters for achieving a better ARI would be very large, and therefore potentially undesirable by researchers who want a simple and interpretable clustering solution. K-modes worked slightly better than the hierarchical methods.

Latent class analysis proved to be competitive when the data features were not demanding, but its performance deteriorated faster than that of MDS-based methods in more complicated setups. Similarly, PAM performed very well for low overlap but markedly worse, also in relative terms, with higher overlap.

For the MDS-based methods, classical MDS did a better job than `smacof`'s ratio MDS, due to the fact that classical MDS arranged the data in such a way that clusters could be separated more linearly, which makes the job of K-means and the Gaussian mixture model easier. However, without any linearity requirement, pdfCluster shared this behavior. Cluster methods such as spectral clustering [26] that do not involve linearity could also be tried with ratio MD. 3 dimensions worked overall slightly better than 2, but there is no reason to believe that increasing the dimensionality above 3 will improve results substantially. Comparing Gaussian mixture models and K-means, there is no clear winner. Both of these approaches worked well in some situations, and both worked mostly better than pdfCluster; the fact that the latter estimates the number of clusters automatically may have affected its results though.

Regarding characteristics of the data, unsurprisingly, increasing levels of overlap among clusters turned out to be the most important factor, making separation of the true clusters more difficult for all methods.

Interesting extensions for future research could be involving other distance measures and clustering methods, with possible alternatives suggested above, and in particular involving methods to estimate the number of clusters, although this is a hard task due to the large number of possible combinations between clustering methods for fixed K and methods to choose K.

References

1. Anderlucci, L., & Hennig, C. (2014). The clustering of categorical data: A comparison of a model-based and a distance-based approach. *Communications in Statistics - Theory and Methods, 43*, 704–721.
2. Azzalini, A., & Torelli, N. (2007). Clustering via nonparametric density estimation. *Statistics and Computing, 17*, 71–80.
3. Azzalini, A., & Menardi, G. (2016). Density-based clustering with non-continuous data. *Computational Statistics, 31*, 771–798.
4. Bartholomew, D. J., Moustaki, I., & Knott, M. (2011). *Latent variable models and factor analysis: A unified approach*. New York: Wiley.
5. Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. Berlin: Springer.
6. Borg, I., Groenen, P., & Mair, P. (2013). *Applied multidimensional scaling*. Berlin: Springer.
7. Casagranda, M. D., Taher, L., & Szumik, C. A. (2012). Endemicity analysis, parsimony and biotic elements: A formal comparison using hypothetical distributions. *Cladistics, 28*, 645–654.
8. Cheetham, A., & Hazel, J. (1969). Binary (presence-absence) similarity coefficients. *Journal of Paleontology, 43*, 1130–1136.
9. Chen, Y. (2013). Biotic element analysis of reptiles of China: A test of vicariance model. *Current Zoology, 59*, 449–457.
10. de Leeuw, J., & Heiser, W. (1982). Theory of multidimensional scaling. In P. R. Krishnaiah & I. N. Kanal (Eds.), *Handbook of statistics* (Vol. 2, pp. 285–316). Amsterdam: North Holland.
11. Dennis, R. L. H., Williams, W. R., & Shreeve, T. G. (1998). Faunal structures among European butterflies: Evolutionary implications of bias for geography, endemism and taxonomic affiliation. *Ecography, 21*, 181–203.
12. Desarbo, W., Howard, D., & Jedidi, K. (1991). Multiclus: A new method for simultaneously performing multidimensional scaling and cluster analysis. *Psychometrika, 56*, 121–136.
13. Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*. New York: Wiley.
14. Halkidi, M., Vazirgiannis, M., & Hennig, C. (2015). Method-independent indices for cluster validation and estimating the number of clusters. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of Cluster Analysis* (pp. 595–618). Boca Raton: CRC Press.
15. Hausdorf, B. (2002). Units in biogeography. *Systematic Biology, 51*, 648–652.
16. Hausdorf, B., & Hennig, C. (2003). Biotic element analysis in biogeography. *Systematic Biology, 52*, 717–723.
17. Hennig, C., & Hausdorf, B. (2004). Distance-based parametric bootstrap tests for clustering of species ranges. *Computational Statistics and Data Analysis, 45*, 875–895.
18. Hennig, C., & Hausdorf, B. (2006). Design of dissimilarity measures: A new dissimilarity between species distribution areas. In V. Batagelj, H. Bock, A. Ferligoj, & A. Ziberna (Eds.), *Data Science and Classification* (pp. 29–37). Berlin: Springer.

19. Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. In H. Lu, H. Motoda, & H. Luu (Eds.), *KDD: techniques and applications* (pp. 21–34). Singapore: World Scientific.
20. Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.
21. Jaccard, P. (1901). Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles, 37*, 241–272.
22. Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters, 31*, 651–666.
23. Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.
24. Kruskal, J. (1977). The relationship between multidimensional scaling and clustering. In J. Van Ryzin (Ed.), *Classification and clustering* (pp. 17–44). New York: Academic Press.
25. McCutcheon, A. L., & Hagenaars, J. A. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.
26. Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (14th ed., pp. 849–856). Cambridge: MIT Press.
27. Oh, M.-S., & Raftery, A. E. (2007). Model-based clustering with dissimilarities: A Bayesian approach. *Journal of Computational and Graphical Statistics, 16*, 559–585.
28. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.
29. Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*, 289–317.
30. Shi, G. R. (1993). Multivariate data analysis in palaeoecology and palaeobiogeography - A review. *Palaeogeography, Palaeoclimatology, 105*, 199–234.
31. Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
32. Ulrich, W., & Gotelli, N. J. (2012). A null model algorithm for presence-absence matrices based on proportional resampling. *Ecological Modelling, 244*, 20–27.
33. Van Mechelen, I., Boulesteix, A.-L., Dangl, R., Dean, N., Guyon, I., Hennig, C., Leisch, F., & Steinley, D. (2018). Benchmarking in cluster analysis: A white paper. arXiv:1809.10496.
34. Vavrek, M. J. (2016). A comparison of clustering methods for biogeography with fossil datasets. *PeerJ, 4*, e1720.
35. Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In A. L. McCutcheon & J. A. Hagenaars (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge: Cambridge University Press.

High-Dimensional Feature Selection for Logistic Regression Using Blended Penalty Functions

Salomi Millard, Mohammad Arashi, and Gaonyalelwe Maribe

Abstract The datasets analysed, in a Biostatistics environment, are frequently high-dimensional and multicollinearity is expected due to the nature of the features. However, many traditional approaches to statistical analysis and feature selection cease to be useful in the presence of high-dimensionality and multicollinearity. Penalised regression methods have proved to be practical and attractive for dealing with these problems. In this chapter, we propose a new penalised approach, the modified elastic-net (MEnet), for statistical analysis and feature selection using a combination of the ridge and bridge penalties. This method is designed to deal with high-dimensional problems with highly correlated predictor variables. Furthermore, it has a closed-form solution, unlike the most frequently used penalised techniques, which makes it simple to implement on high-dimensional data. We show how this approach can be used to analyse high-dimensional data with binary responses, e.g. microarray data, and simultaneously select significant features. A simulation study demonstrates the properties and practical aspects of the proposed method.

1 Introduction

While the origins of statistical learning preceded some of the first scientific journals, the largest development in statistics occurred during the twentieth century. Practical problems that motivated the development of new statistical theory during this period were low dimensional. These problems consisted of a small number of features,

S. Millard · M. Arashi (✉) · G. Maribe
Department of Statistics, University of Pretoria, Pretoria, South Africa
e-mail: arashi@um.ac.ir

S. Millard
e-mail: salomi.millard@up.ac.za

G. Maribe
e-mail: g.maribe@up.ac.za

M. Arashi
Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_14

measured on a much larger number of experimental units. If n denotes the number of experimental units and p the number of features, then for the most part traditional theory considered "small p, large n" scenarios [9].

However, the rapid advancement in data acquisition technologies and data processing capabilities have changed the data analytics environment drastically over the last decades. These developments resulted in applications for which the number of features far exceeds the number of experimental units; some examples include image analysis, microarray analysis, and text analysis. Although the problems that arise in these respective frontiers of science differ, they share a mutual objective: knowledge acquisition from massive and high-frequency data [3].

According to [4], there are three important pillars of any statistical procedure, namely statistical accuracy, model interpretability, and computational complexity. In traditional methods, where the number of experimental units far exceeds the number of features, the efficiency of one aspect does not come at the cost of others. However, traditional methods experience considerable difficulties when the number of features is comparable to or larger than the number of experimental units. Amongst these difficulties are designing statistical procedures that are efficient in inference, deriving asymptotic or non-asymptotic theory, interpreting the estimated models, and creating computationally efficient and robust statistical procedures.

Traditionally, good statistical practice requires the number of experimental units to exceed the number of estimable parameters by some solid margin. Reference [6] suggests that a plausible rule of thumb is $\frac{n}{p} \geq 5$. To be more specific, let $\rho = \frac{p}{n}$. Then, $\rho < 1$ is considered as a low-dimensional case, while $\rho \in (1, \infty)$ accounts for a high-dimensional problem. The case $\rho = 1$, i.e. $p = n$ can be either high or low-dimensional depending on the value of p. Also note, p can be smaller than n, but still very large. The above explanation does not provide a clear understanding about the relation of p and n. Thus, to better understand low/high-dimensional regimes, let $p = p_n$ denote the growth of the number of features as n grows. Existing low/high-dimensional problems then chase the following relations between n and p_n. For a low-dimensional case, i.e. $p_n < n$ we have $p_n = \left[\frac{n}{2}\right]$, $p_n = \left[4.5n^{\frac{1}{4}}\right]$, or $p_n = \left[\frac{n}{b\log(n)}\right]$ where $b > 1$, while for a high-dimensional case, i.e. $p_n > n$ we may have $\log(p_n) = O_p\left(n^b\right)$, where $0 < b < 1$. Table 1 gives an illustration of low- and high-dimensional regime for these cases.

As stated by [9], "the 'large p, small n' world would therefore seem to depend on a certain statistical alchemy—the computational transformation of ignorance into parameter estimates by fearless specification and fitting of high-dimensional models". Nevertheless, "large p, small n" models and methods have been applied successfully in a wide variety of disciplines. In a Biostatistics environment, measurements are typically obtained for a large number of candidate predictor variables in an attempt to avoid overseeing a significant relationship between a predictive factor and the response variable. Furthermore, the datasets to be analysed, frequently exhibits multicollinearity due to the nature of the features.

Most traditional approaches to statistical analysis and feature selection cease to be useful in the presence of high-dimensionality and multicollinearity, due to compu-

Table 1 Examples of low/high-dimensional regime (n, p_n)

	Low-dimensional				High-dimensional	
np_n	$\frac{n}{2}$	$4.5n^{\frac{1}{4}}$	$\frac{n}{1.5\log(n)}$	$\frac{n}{2\log(n)}$	$e^{n^{0.4}}$	$e^{n^{0.5}}$
25	13	10	5	4	27	148
50	25	12	9	6	70	1177
100	50	14	14	11	244	22026
250	125	18	30	23	2238	7358659
500	250	21	54	40	21338	5141855148

tational restrictions as well as its inability to identify a feasible model. As a result, feature selection techniques for high-dimensional data that exhibits multicollinearity have become a fundamental challenge in statistics. Penalised regression has become a popular method for analysing these types of data. Penalised regression is a promising class of regression models for continuous shrinkage and feature selection in the presence of multicollinearity and/or high-dimensionality. It trades off a small increase in the bias of the regression parameters in exchange for a significant decrease in the variance of the regression parameters, improving the model accuracy and interpretability. Furthermore, some penalised regression models yield a sparse solution, thereby addressing the curse of dimensionality.

Let $\left\{\left(\mathbf{x_i}^\top, y_i\right)\right\}_{i=1}^n$ be a set of observations following a generalised linear model (GLM), with predictors, $\mathbf{x_i} = \left(x_{i1}, \ldots, x_{ip}\right)^\top \in \mathbb{R}^p$, and regression coefficients, $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^\top \in \mathbb{R}^{p+1}$. The penalised GLM has the following log-likelihood function:

$$l_\lambda(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = l\left(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}\right) - P_\lambda\left(\boldsymbol{\beta}\right)$$

where $l\left(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}\right)$ is the log-likelihood of the GLM of interest, $\mathbf{X} = \left(\mathbf{x_1}, \ldots, \mathbf{x_n}\right)^\top$, $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, P_λ is some penalty function and λ is the tuning parameter. The frequent occurrence of high-dimensional datasets that exhibits multicollinearity emphasises the need for penalty functions that perform well in high-dimensional regression problems with highly correlated predictor variables. It is known that the ridge penalty [7] outperforms many other penalties in the presence of multicollinearity. However, one of the major shortcomings of the ridge penalty, is that it does not yield a sparse solution. Some penalties, namely the elastic-net (Enet) [11], Mnet [8], and combined penalty (CP) [10] has been proposed to deal with high-dimensional regression problems in the presence of multicollinearity. These penalties are a linear combination of the ridge penalty and some penalty that has desirable properties, such as the ability to yield a sparse solution or the oracle properties as defined by [2]. In this chapter, we propose the novel modified elastic-net (MEnet) as a linear combination of the ridge and bridge [5] penalties. This penalty is designed to deal with high-dimensionality and multicollinearity. The chapter is structured as follows: Sect. 2

introduces the novel modified elastic-net (MEnet) as a blended penalty function for the GLM and derives a computationally efficient estimation algorithm. Section 3 illustrates the use of the MEnet penalty by means of a simulation study. Section 4 applies the MEnet penalty to a well-known colon cancer classification problem, and finally Sect. 5 concludes the chapter and provides some suggestions for future work.

2 Penalised GLM with the MEnet Penalty

In this section, we propose the novel modified elastic-net (MEnet) for the GLM. A computationally efficient estimation routine is given in Sect. 2. We consider the case of the one parameter natural exponential family with

$$f\left(y|\eta\right) = h\left(y, \phi\right) \exp \frac{\left(\eta y - A\left(\eta\right)\right)}{\phi}. \tag{1}$$

Modified Elastic-Net Penalty

The proposed MEnet penalty, as a linear combination of the ridge and bridge [5] penalties, with regression coefficients, $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_p\right)^\top$, is given by

$$P\left(\boldsymbol{\beta}\right)_{\lambda,\alpha}^{MEnet} = \lambda \left[\left(\frac{1-\alpha}{2} \right) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|^\gamma \right], \tag{2}$$

where $\gamma > 0$ and $\gamma \neq 2$. Note that this is equivalent to

$$P\left(\boldsymbol{\beta}\right)_{\lambda_1,\lambda_2}^{MEnet} = \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^\gamma, \tag{3}$$

with $\lambda_1 = \frac{\lambda(1-\alpha)}{2}$ and $\lambda_2 = \lambda\alpha$. The penalty is inspired by the elastic-net penalty and include as special cases; the ridge ($\alpha = 0$), LASSO ($\alpha = 1, \gamma = 1$), bridge ($\alpha = 1$), and elastic-net ($\alpha \neq 0, \gamma = 1$) penalties. The MEnet penalty is designed to deal with both high-dimensionality and multicollinearity.

Penalised Likelihood Function

Based on the random sample, $\left\{\left(\mathbf{x_i}^\top, y_i\right)\right\}_{i=1}^n$, the log-likelihood of the GLM is given by

$$l(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \frac{y_i \mathbf{x}_i^\top \boldsymbol{\beta}}{\phi} - \sum_{i=1}^{n} \frac{A\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{\phi} + \sum_{i=1}^{n} \log h\left(y_i, \phi\right). \tag{4}$$

In order to obtain the penalised log-likelihood, the MEnet penalty as given by (3) is subtracted from (4), resulting in

$$l(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})_{\lambda_1,\lambda_2}^{MEnet} = l(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) - \lambda_1 \sum_{j=1}^{p} |\beta_j|^2 - \lambda_2 \sum_{j=1}^{p} |\beta_j|^\gamma. \tag{5}$$

The bridge penalty is nonconvex if $\gamma < 1$, which makes the optimisation of the MEnet penalty challenging for such values of γ. An approximation to the MEnet penalty is utilised to address this challenge. This approximation results in closed-form solutions that greatly simplifies the computation of the MEnet regression estimates.

Reforming of the MEnet Penalty Term

The local quadratic approximation of [2] is used to derive an approximation to the MEnet penalty. Let $P(|\beta_j|)_\lambda$ denote the penalty function on β_j, indexed by the tuning parameter $\lambda > 0$ for controlling the extent of penalisation. Each penalty term has the form $\sum_{j=1}^{p} P(|\beta_j|)_\lambda$. For simplicity we assume $P(|\beta_j|)_\lambda = \lambda P(|\beta_j|)$. Given that

$$\frac{d}{d\beta_j} P\left(|\beta_j|\right)_\lambda = P\left(|\beta_j|\right)_\lambda' = P\left(|\beta_j|\right)_\lambda' \frac{\beta_j}{|\beta_j|}, \tag{6}$$

we can approximate (6), around the local point $\boldsymbol{\beta}^o = (\beta_1^o, \cdots, \beta_p^o)^\top$, by

$$\frac{d}{d\beta_j} P(|\beta_j|)_\lambda \approx P(|\beta_j^o|)_\lambda' \frac{\beta_j}{|\beta_j^o|}. \tag{7}$$

Integrating (7) from β_j^o to β_j yields

$$P(|\beta_j|)_\lambda \approx P(|\beta_j^o|)_\lambda + \frac{1}{2} \frac{P(|\beta_j^o|)_\lambda'}{|\beta_j^o|} \left(\beta_j^2 - \beta_j^{o2}\right). \tag{8}$$

Hence, using (8), we approximate the penalty term by

$$\sum_{j=1}^{p} P(|\beta_j|)_\lambda \approx \sum_{j=1}^{p}\left[P(|\beta_j^o|)_\lambda + \frac{1}{2}\frac{P(|\beta_j^o|)'_\lambda}{|\beta_j^o|}\left(\beta_j^2 - \beta_j^{o2}\right)\right]$$
$$= \left[\sum_{j=1}^{p} P(|\beta_j^o|)_\lambda - \frac{1}{2}\sum_{j=1}^{p}\frac{P(|\beta_j^o|)'_\lambda}{|\beta_j^o|}\beta_j^{o2}\right] + \frac{1}{2}\sum_{j=1}^{p}\frac{P(|\beta_j^o|)'_\lambda}{|\beta_j^o|}\beta_j^2$$
$$= c + \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{Q}_\lambda \boldsymbol{\beta}, \tag{9}$$

where c is a function of known values β_j^o and

$$\mathbf{Q}_\lambda = \text{Diag}\left(\frac{P'_\lambda(|\beta_1^o|)}{|\beta_1^o|}, \cdots, \frac{P'_\lambda(|\beta_p^o|)}{|\beta_p^o|}\right). \tag{10}$$

After some algebraic manipulation, the MEnet penalty term can be approximated, using (9) and (10), by

$$P(\boldsymbol{\beta})^{MEnet}_{\lambda_1,\lambda_2} = \sum_{j=1}^{p} P_{\lambda_1}(|\beta_j|) + \sum_{j=1}^{p} P_{\lambda_2}(|\beta_j|)$$
$$= \lambda_1 \sum_{j=1}^{p} |\beta_j|^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^\gamma$$
$$\approx c_1 + \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{Q}_{\lambda_1}\boldsymbol{\beta} + c_2 + \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{Q}_{\lambda_2}\boldsymbol{\beta}, \tag{11}$$

where

$$\mathbf{Q}_{\lambda_1} = \text{Diag}\left(\frac{P'_{\lambda_1}(|\beta_1^o|)}{|\beta_1^o|}, \cdots, \frac{P'_{\lambda_1}(|\beta_p^o|)}{|\beta_p^o|}\right) = 2\mathcal{I}_p, \tag{12}$$

with $\mathbf{I}_p$ the p-dimensional identity matrix, and

$$\mathbf{Q}_{\lambda_2} = \text{Diag}\left(\frac{P'_{\lambda_2}(|\beta_1^o|)}{|\beta_1^o|}, \ldots, \frac{P'_{\lambda_2}(|\beta_p^o|)}{|\beta_p^o|}\right)$$
$$= \text{Diag}\left(\gamma|\beta_1^o|^{\gamma-2}, \ldots, \gamma|\beta_p^o|^{\gamma-2}\right). \tag{13}$$

Parameter Estimation

From (5) and (11), the maximum likelihood estimate for β^{MEnet} is obtained by maximising

$$l(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})^{MEnet}_{\lambda_1,\lambda_2} \approx \sum_{i=1}^{n} \frac{y_i \mathbf{x}_i^\top \boldsymbol{\beta}}{\phi} - \sum_{i=1}^{n} \frac{A\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{\phi} + \sum_{i=1}^{n} \log h\left(y_i, \phi\right) - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{Q}_{\lambda_1}\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{Q}_{\lambda_2}\boldsymbol{\beta}. \tag{14}$$

The Newton Raphson algorithm (NRA) is used to maximise (14). The gradient vector of the NRA is given by

$$\begin{aligned} \nabla_\phi\left(\boldsymbol{\beta}^{MEnet}\right) &= \frac{\partial l(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})}{\partial \boldsymbol{\beta}} - \frac{\partial P\left(\boldsymbol{\beta}\right)^{MEnet}_{\lambda_1,\lambda_2}}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{\phi}\mathbf{X}^\top\left(\mathbf{y}-\mu\right) - \lambda_1 \mathbf{Q}_{\lambda_1}\boldsymbol{\beta} - \lambda_2 \mathbf{Q}_{\lambda_2}\boldsymbol{\beta}, \end{aligned} \tag{15}$$

with $(\mathbf{y}-\mu)^\top = ((y_1-\mu_1)\ (y_2-\mu_2)\ \ldots\ (y_n-\mu_n))^\top$ and $\boldsymbol{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$. Furthermore, the Hessian is given by

$$\begin{aligned} H_\phi\left(\boldsymbol{\beta}^{MEnet}\right) &= \frac{\partial l(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} - \frac{\partial P\left(\boldsymbol{\beta}\right)^{MEnet}_{\lambda_1,\lambda_2}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} \\ &= -\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X} - \lambda_1 \boldsymbol{Q}_{\lambda_1} - \lambda_2 \boldsymbol{Q}_{\lambda_2}, \end{aligned} \tag{16}$$

where $\boldsymbol{W} = \frac{1}{\phi^2}\text{Diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right)$ and $\sigma_i^2 = \text{Var}(Y_i|\boldsymbol{x}_i)$. For convenience we will use $\boldsymbol{\beta}$ instead of $\boldsymbol{\beta}^{MEnet}$. From (15) and (16) the Newton Raphson update rule is

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{Old} + \left(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X} + \lambda_1 \boldsymbol{Q}_{\lambda_1} + \lambda_2 \boldsymbol{Q}_{\lambda_2}\right)^{-1} \times \left(\frac{1}{\phi}\boldsymbol{X}^\top\left(\boldsymbol{y}-\boldsymbol{\mu}\right) - \lambda_1 \boldsymbol{Q}_{\lambda_1}\boldsymbol{\beta}^{old} - \lambda_2 \boldsymbol{Q}_{\lambda_2}\boldsymbol{\beta}^{old}\right). \tag{17}$$

The NRA for the GLM with the MEnet penalty is given in Algorithm 1.

Algorithm 1 The Newton Raphson algorithm for the penalised GLM with the MEnet penalty.

Step 1. Calculate appropriate fixed values to be used in Q_{λ_1} and Q_{λ_2}.
Step 2. Choose initial values for $\boldsymbol{\beta}^{Old}$.
Step 3. Update $\boldsymbol{\beta}^{New}$ as given in (17) for the current values of $\boldsymbol{\beta}^{Old}$.
Step 4. Set $\boldsymbol{\beta}^{Old} = \boldsymbol{\beta}^{New}$.
Step 5. Repeat Steps 3 and 4 until convergence.

3 Simulation Study

The performance of the penalised logistic regression model with the proposed MEnet penalty is assessed in this section. The binary logistic regression model is used to model the probability of a certain event. It assumes a linear relationship between a set of predictor variables, $\boldsymbol{X}$, and the log odds of a response variable, $\boldsymbol{Y}$. Consider a set of observations, $\left\{\left(\mathbf{x_i}^\top, y_i\right)\right\}_{i=1}^{n}$, following a logistic regression model, with predictors, $\mathbf{x_i} = \left(x_{i1}, \ldots, x_{ip}\right)^\top \in \mathbb{R}^p$, and regression coefficients, $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^\top \in \mathbb{R}^{p+1}$. The logistic regression model, with $\pi\left(\mathbf{x}_i; \boldsymbol{\beta}\right) = P\left(Y_i = 1 | \mathbf{x}_i\right)$, is given by

$$\log\left(\frac{\pi\left(\mathbf{x}_i; \boldsymbol{\beta}\right)}{1 - \pi\left(\mathbf{x}_i; \boldsymbol{\beta}\right)}\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i.$$

The log-likelihood of the penalised logistic regression model with the MEnet is

$$l(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})_{\alpha,\lambda}^{MEnet} = \sum_{i=1}^{n}[y_i \log \pi(\mathbf{x_i}; \boldsymbol{\beta}) + (1 - y_i)\log(1 - \pi(\mathbf{x_i}; \boldsymbol{\beta}))] - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{Q}_{\lambda_1}\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{Q}_{\lambda_2}\boldsymbol{\beta}.$$

A high-dimensional scenario with varying sample sizes is considered. The predictor variables consist of four independent blocks and are generated from a $N\left(\mathbf{0}, \Sigma\right)$ distribution with unit variance. As illustrated in Table 2, two of the blocks exhibit high multicollinearity and two of the blocks have nonzero regression coefficients. The correlation structure in the four blocks are $\Sigma_{Block1} \sim u(0.75, 0.85)$, $\Sigma_{Block3} \sim u(0.75, 0.85)$, $\Sigma_{Block2} \sim u(0.3, 0.4)$ and $\Sigma_{Block4} \sim u(0.3, 0.4)$. Furthermore $\boldsymbol{\beta}_{Block1,2} \sim u(0.08, 0.12)$ and $\boldsymbol{\beta}_{Block3,4} = \mathbf{0}$. Blocks 1 and 2 each consists of 15 significant features such that a total of 30 significant features is considered in the simulation study.

The number of features under consideration is $p = 500$. Furthermore, three sample size, $n \in \{100, 200, 300\}$, are considered. The simulation is repeated 500 times and consists of a training dataset and an independent testing dataset for assessing the prediction capability of the resulting model. Each training dataset contains $n \in \{100, 200, 300\}$ independent observations while the testing dataset contains 100

Table 2 Properties of the design matrix

Nonzero regression coefficients	Multicollinearity	
	Yes	No
Yes	Block 1	Block 2
No	Block 3	Block 4

independent observations. Algorithm 2 is used to obtain the optimal model. Many other criteria of the model performance such as the area under the ROC curve or mean prediction error may be utilised in steps 4 and 5 of Algorithm 2. For the purpose of comparison, the values of α and γ are $\alpha \in \{0.25, 0.5, 0.75\}$ and $\gamma \in \{0.25, 0.5, 0.75, 1, 1.5\}$. The averages of the following performance measures are used to evaluate each penalty: number of selected features, average prediction error on the testing dataset, accuracy on the testing dataset when using a prediction threshold of 0.5 and sensitivity (that is, the proportion of nonzero regression coefficients that were correctly estimated as nonzero).

Algorithm 2 Optimal model selection for the GLM with the MEnet penalty.

Step 1. Determine the optimal tuning parameter, λ, using k-fold cross-validation on the training dataset.
Step 2. Using the optimal tuning parameter, λ, estimate the regression coefficients on the whole training dataset.
Step 3. Apply a range of appropriate feature selection thresholds of which some regression coefficient are set to zero.
Step 4. For each feature selection threshold, determine the accuracy on the testing dataset.
Step 5. The optimal model is selected such that the accuracy obtained with the corresponding feature selection threshold is maximised.

The goal is to determine whether the MEnet penalty has the ability to correctly identify the significant features as well as to determine the prediction performance when compared to the elastic-net penalty, as the main competitor. Table 3 gives the results for $n = 300$. It should be noted that for a fixed value of γ, the effect of α on the average number of nonzero regression coefficients estimated by the MEnet penalty is marginal. In Table 3 it is clear that, for the Enet, Mnet and CP the true effect of α is in the average number of nonzero regression coefficients and not in the performance measures such as the average prediction error or accuracy. As stated in Algorithm 2, a feature selection threshold based on the prediction accuracy is used to obtain the nonzero regression coefficients. For this reason, the effect of α, for a fixed value of γ, is not the same as the effect on the Enet, Mnet, and CP.

The CP and Mnet penalties select much fewer features than what is simulated and as a result, the average sensitivity of these penalties is low. However, it is interesting to note that with so few nonzero regression coefficients, the average prediction errors and accuracies are quite competitive to both the Enet and MEnet penalties, but at the cost of an incorrect model specification. The remainder of the discussion will focus on the comparison of the MEnet and Enet penalty since the Enet penalty is considered as the main competitor of the MEnet penalty. The average prediction error and the average prediction accuracy obtained by the MEnet penalty, irrespective of γ, is better than that obtained by the Enet penalty. Furthermore the performance of these two measures improves as γ increases in the range (0,1]. Although the average prediction accuracy is lower for $\gamma > 1$ than for $\gamma \leq 1$, it is still higher than that obtained by the Enet penalty. Considering the behaviour of the average number of selected features, the average prediction error and the average prediction accuracy,

Table 3 Performance of the MEnet, Enet, Mnet, and CP. Performance measures included are: average number of selected features, average prediction error, the average prediction accuracy, and the proportion of nonzero regression coefficients that were correctly estimated as nonzero

Penalty	Hyperparameters	Number of selected features	Prediction error	Prediction accuracy	Sensitivity
MEnet	$\alpha = 0.25$	34	0.406	0.739	0.703
with	$\alpha = 0.5$	33	0.405	0.739	0.701
$\gamma = 0.25$	$\alpha = 0.75$	33	0.405	0.739	0.702
MEnet	$\alpha = 0.25$	35	0.404	0.742	0.706
with	$\alpha = 0.5$	34	0.407	0.742	0.677
$\gamma = 0.5$	$\alpha = 0.75$	35	0.402	0.742	0.705
MEnet	$\alpha = 0.25$	38	0.395	0.746	0.688
with	$\alpha = 0.5$	37	0.395	0.746	0.678
$\gamma = 0.75$	$\alpha = 0.75$	36	0.395	0.746	0.675
MEnet	$\alpha = 0.25$	40	0.389	0.748	0.627
with	$\alpha = 0.5$	41	0.387	0.748	0.626
$\gamma = 1$	$\alpha = 0.75$	40	0.387	0.748	0.619
MEnet	$\alpha = 0.25$	51	0.374	0.739	0.535
with	$\alpha = 0.5$	51	0.370	0.739	0.518
$\gamma = 1.5$	$\alpha = 0.75$	50	0.370	0.738	0.507
Enet	$\alpha = 0.25$	49	0.411	0.714	0.674
	$\alpha = 0.5$	32	0.412	0.710	0.561
	$\alpha = 0.75$	26	0.413	0.706	0.496
Mnet	$\alpha = 0.25$	12	0.394	0.702	0.317
	$\alpha = 0.5$	9	0.394	0.691	0.230
	$\alpha = 0.75$	8	0.393	0.690	0.218
CP	$\alpha = 0.25$	21	0.392	0.715	0.488
	$\alpha = 0.5$	16	0.395	0.707	0.417
	$\alpha = 0.75$	15	0.395	0.704	0.385

a model with $\gamma < 1$ is preferable. The average sensitivity of the MEnet penalty is the highest for smaller γ. The MEnet penalty with $\gamma < 1$ is able to obtain a higher sensitivity than the Enet penalty. Furthermore, when the Enet penalty with $\alpha = 0.25$ is considered, this is achieved whilst also selecting fewer features.

As mentioned before, sample sizes $n \in \{100, 200, 300\}$ were considered. Figure 1a shows that the median number of selected features are quite similar for $n \in \{200, 300\}$, but slightly less than the median number of selected features in the $n = 100$ case. Furthermore, the median proportion of nonzero regression coefficients that are estimated as nonzero (sensitivity) tends to increase with the sample

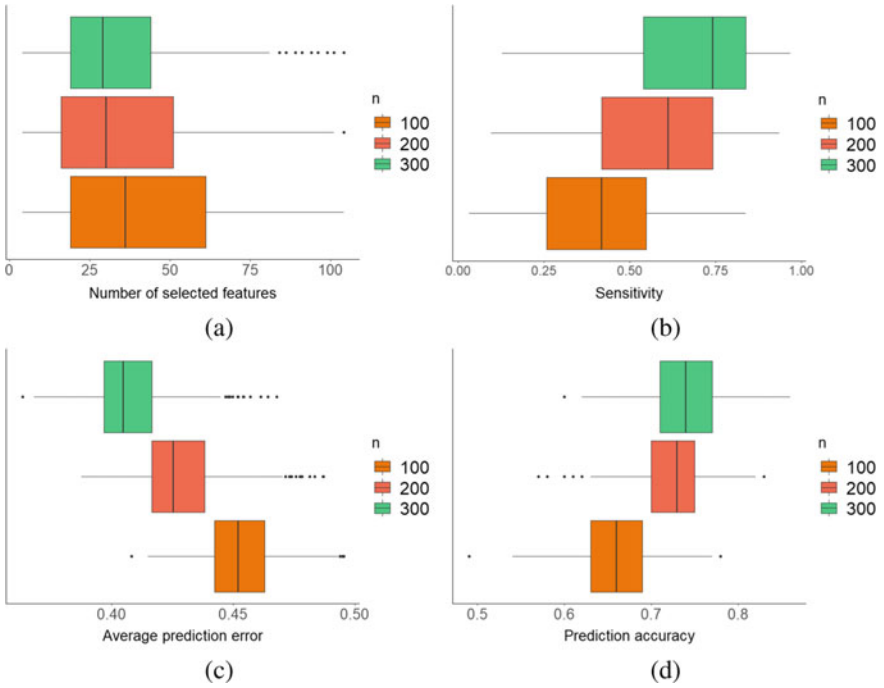


Fig. 1 Effect of varying sample sizes on the evaluation criteria when using the MEnet penalty with $\gamma = 0.5, \alpha = 0.5$ on the simulated data: **a** Number of selected features, **b** The proportion of nonzero regression coefficients that were correctly estimated as nonzero (sensitivity), **c** Average prediction error on the testing dataset, **d** Prediction accuracy when using a threshold of 0.5 on the testing dataset

size as illustrated in Fig. 1b. As expected, Fig. 1c and d confirms that the average prediction error and accuracy obtained on the testing datasets improves as the sample size increases.

4 Colon Cancer Classification

In this section, the colon cancer dataset of [1] is used to illustrate the performance of the MEnet penalty. This dataset contains 2000 gene expression levels on 62 colon tissues such that 40 of the tissues are malicious. This example is classified as an ultra high-dimensional case. A training dataset with 70% of the original sample is randomly selected from the colon cancer dataset. In all cases, fourfold cross-validation on the training data is used to determine the optimal value of λ. For consistency, Algorithm 2 is used to obtain the optimal model when the MEnet penalty is used. This is repeated on 50 testing and training datasets. The averages of the

Table 4 Performance of the MEnet, bridge, Enet, Mnet, and CP on the colon cancer dataset. The performance measures are: the average number of selected features, the average prediction error on the testing dataset, and the average prediction accuracy on the testing dataset

Penalty	Hyperparameters	Number of selected features	Prediction error	Prediction accuracy
MEnet	$\gamma = 0.5$	26	0.268	0.810
	$\gamma = 0.75$	32	0.307	0.757
Bridge	$\gamma = 0.5$	23	0.396	0.761
	$\gamma = 1$ (LASSO)	9	0.338	0.760
	$\gamma = 2$ (ridge)	2000	0.358	0.753
Enet	$\alpha = 0.25$	44	0.359	0.756
	$\alpha = 0.5$	27	0.330	0.767
	$\alpha = 0.75$	15	0.336	0.778
Mnet	$\alpha = 0.25$	9	0.301	0.793
	$\alpha = 0.5$	5	0.313	0.780
	$\alpha = 0.75$	4	0.309	0.800
	$\alpha = 1$ (MCP)	4	0.314	0.787
CP	$\alpha = 0.25$	19	0.290	0.805
	$\alpha = 0.5$	15	0.299	0.812
	$\alpha = 0.75$	12	0.295	0.805
	$\alpha = 1$ (SCAD)	11	0.312	0.792

following performance measures are used to evaluate each penalty: the number of selected features, the average prediction error on the testing dataset, the prediction accuracy on the testing dataset when using a threshold of 0.5. The results obtained on the colon cancer dataset are given in Table 4. In this application, the MEnet penalty with $\gamma \in \{0.5, 0.75\}$ and $\alpha = 0.5$ is considered. As illustrated in the simulation study, when γ is fixed, the effect of α on the results is marginal. Hence, for illustrative purposes, only a single value of α is considered. The results obtained by the MEnet penalty are compared to that obtained by the bridge, Enet, CP, and Mnet penalties. Figure 2 displays the evaluation criteria boxplots when using the penalties under consideration.

The average number of features selected by the LASSO and Mnet penalties is lowest, with all variations yielding at most 9 significant features. The Enet with $\alpha = 0.75$ and variations of the CP results in 11 to 19 significant features. The MEnet, Enet with $\alpha \in \{0.25, 0.5\}$, and bridge with $\gamma = 0.5$, results in an average number of significant features varying from 23 to 44. As before, the CP and Mnet penalties are able to obtain highly competitive average prediction errors and accuracies with only a few selected features. The distribution of the number of features selected by the various penalties is given in Fig. 2a. The average prediction error obtained by the bridge and Enet penalties are higher than those obtained by the Mnet, Cp, and MEnet penalties. The MEnet penalty with $\gamma = 0.5$ has the lowest average prediction error of all penalties under consideration. However, some variations on the CP and Mnet penalties yield a smaller average prediction error than the MEnet penalty with

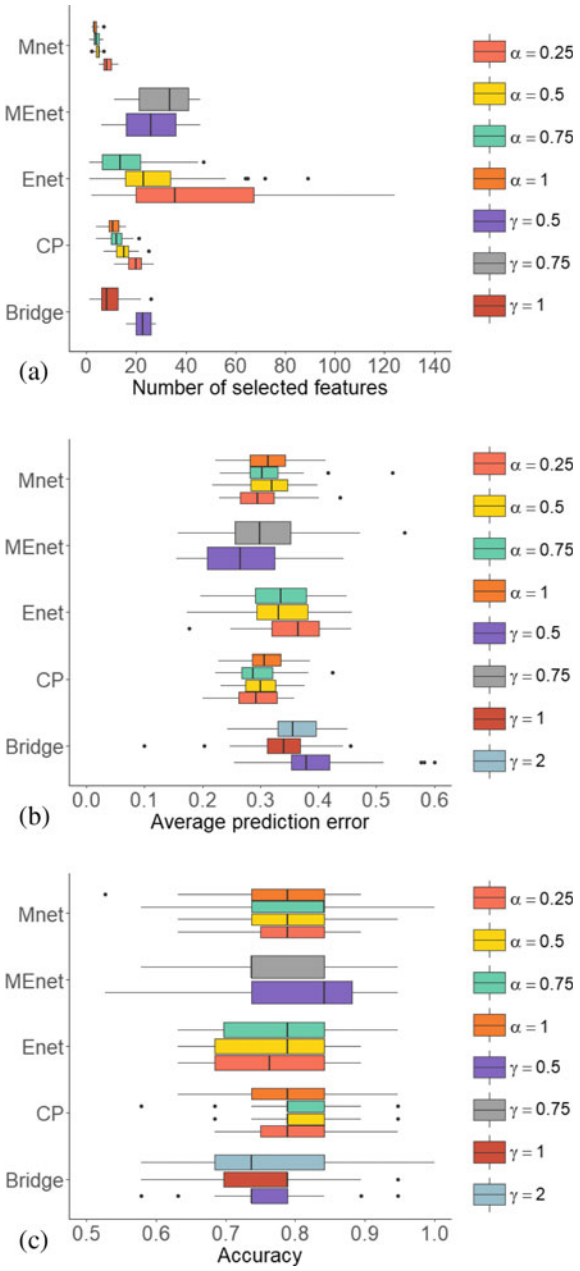


Fig. 2 Evaluation criteria boxplots when using the MEnet, Enet, Mnet, Cp, and bridge penalties on the colon cancer dataset: **a** Number of selected features, **b** Average prediction error on the testing dataset, **c** Accuracy when using a threshold of 0.5 on the testing dataset

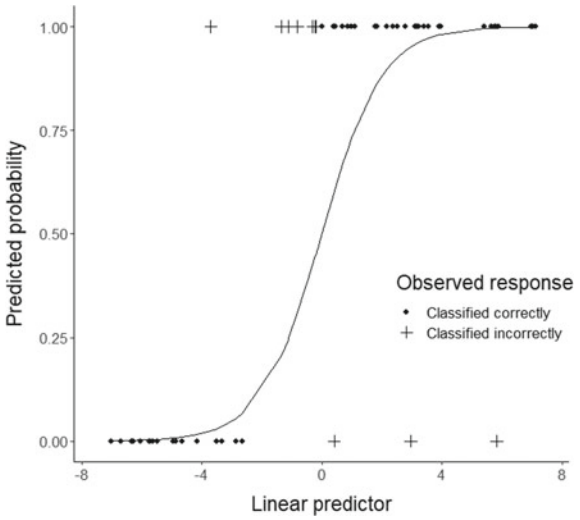


Fig. 3 Classification of the colon cancer dataset when using the MEnet penalty with $\gamma = 0.5$

$\gamma = 0.75$. The average accuracy obtained by the bridge, Enet, and MEnet with $\gamma = 0.75$, is slightly lower than that obtained by the CP, Mnet, and MEnet with $\gamma = 0.5$. The MEnet with $\gamma = 0.5$ yields an average accuracy, that is outperformed by only the CP with $\alpha = 0.5$. The distribution of the average prediction error and accuracy on the testing datasets is given in Fig. 2b and c.

Figure 3 illustrates the adequacy of the MEnet penalty with $\gamma = 0.5$ for classification of the colon cancer dataset. It is clear that the majority of the observation is classified correctly as also indicated in Table 4.

5 Conclusion and Future Work

In this chapter, the modified elastic-net (MEnet) penalty, as a linear combination of the ridge and bridge penalties, was proposed to address the curse of dimensionality and multicollinearity in the GLM. This penalty includes the ridge, LASSO, bridge, and elastic-net as special cases. A computationally efficient estimation routine was derived and the proposed MEnet penalty has been evaluated with its adequacy confirmed through a simulation study and a cancer data classification application. The results obtained find the performance of the newly proposed MEnet penalty to be highly competitive with the performance of the elastic-net penalty. This chapter gave the first steps towards the extension of the MEnet penalty for the GLM. According to this study, the following extensions are envisaged: further simulation studies should be conducted in order to evaluate the performance of the MEnet penalty under different multicollinearity and dimensionality regimes; the asymptotic properties of the estimator in (17) are desirable in order to conduct inference on the results obtained;

instead of using the MEnet penalty, intuitively, we can conduct two-step analysis by first estimating the regression coefficients using the ridge estimator and then implement bridge penalised estimation using the residuals in the refitted model, extending the proposed penalty for longitudinal data analysis and mixtures of GLMs.

Acknowledgements We would like to thank two anonymous reviewers for their comments which improved the presentation. This work was based on research supported in part by the National Research Foundation (NRF) of South Africa, SARChI Research Chair UID: 71199; Ref.: IFR170227223754 grant No. 109214, STATOMET at the Department of Statistics at the University of Pretoria, South Africa. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily attributed to the NRF.

References

1. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences, 96*(12), 6745–6750.
2. Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.
3. Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality. In *Proceedings of the International Congress of Mathematicians*. Madrid, Spain.
4. Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica, 20*(1), 101.
5. Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35*(2), 109–135.
6. Hamilton, W. C. (1970). The revolution in crystallography. *Science, 169*(3941), 133–141.
7. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.
8. Huang, J., Breheny, P. J., Lee, S., Ma, S., & Zhang, C. H. (2016). The Mnet method for variable selection. *Statistica Sinica*, 903–923.
9. Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A, 367*(1906), 4237–4253.
10. Wang, X., Park, T., & Carriere, K. (2010). Variable selection via combined penalization for high-dimensional data analysis. *Computational Statistics & Data Analysis, 54*(10), 2230–2243.
11. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.

A Generalized Quadratic Garrote Approach Towards Ridge Regression Analysis

Inesh Munaweera, Saman Muthukumarana, and Mohammad Jafari Jozani

Abstract Ridge regression is widely used in multiple linear regression analysis to address the prevalent multicollinearity issue in high-dimensional settings. In the standard form of ridge regression analysis, all model coefficients are shrunken towards zero at a similar rate regardless of the importance of each variable. In this paper, we provide an extension of the non-negative garrote method to give more flexibility to the ridge regression approach for unequal shrinkage of regression coefficients. We show that this approach is capable of shrinking smaller coefficients even faster than the adaptive lasso while keeping the larger coefficients almost untouched. Our generalized quadratic garrote approach enables practitioners to have more control over the amount of shrinkage on each regression coefficient estimate. We study the theoretical properties of our generalized quadratic Garrote regression estimators. Finally, we provide extensive numerical studies involving sparse, nearly sparse, and high dimensional settings and illustrate the practical use of the suggested shrinkage approach with the Boston Housing Dataset.

1 Introduction

The standard multiple linear regression model can be written as

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, 2, ..., n, \tag{1}$$

I. Munaweera · S. Muthukumarana · M. Jafari Jozani (✉)
University of Manitoba, 66 Chancellors Cir, Winnipeg, MB R3T 2N2, Canada
e-mail: m_jafari_jozani@umanitoba.ca

I. Munaweera
e-mail: ineshpma@myumanitoba.ca

S. Muthukumarana
e-mail: saman.muthukumarana@umanitoba.ca

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_15

where, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\top$ is the vector of unknown coefficients, y_i is the response value for the ith observation and $x_{i1}, ..., x_{ip}$ are corresponding values at p explanatory (predictor) variables. We assume ϵ_i's to be independent and identically distributed (iid) with mean $E(\epsilon_i) = 0$ and variance $Var(\epsilon_i) = \sigma^2$. In matrix form one can write the multiple linear regression model as follows.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where, $\boldsymbol{y}$ is the response vector and $\boldsymbol{X}$ is the $n \times (p + 1)$ design matrix where the first column consists of 1's associated with β_0 in (1). Furthermore, $\boldsymbol{\epsilon}$ is the vector of random errors with $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. The most widely used method for estimating β_j's is the ordinary least squares (OLS) technique which minimizes the sum of squared errors (SSE) given by

$$\text{SSE} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{3}$$

where, $\boldsymbol{A}^\top$ denotes the transpose of $\boldsymbol{A}$. Given that $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ exists, the solution to the above minimization problem can be obtained as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y},$$

where $\hat{\boldsymbol{\beta}}$ is the vector of least squares estimates. As pointed out by [1], OLS estimates might not be the best choice when multicollinearity is present in data. As a consequence of the multicollinearity problem, OLS estimates tend to have large standard errors which lead to highly unstable coefficients. Hence, a small change in the dataset will result in a large change in coefficient estimates. Large standard errors will lead to statistically non-significant coefficient estimates, and the corresponding confidence intervals for the true regression coefficients will be wide. Furthermore, OLS estimates can have unexpected signs and magnitudes which can lead to misleading conclusions [2].

One approach to deal with the multicollinearity issue is to allow for bias and develop bias estimators that are more stable than OLS estimators. Regularized least-squares estimation (Shrinkage method) provides a class of such methods that do the coefficient estimation in a regularized manner as a remedy for the inflation and the instability associated with coefficient estimates. As a result of the regularization, parameter estimates are shrunken towards zero.

Ridge regression is one of the most popular shrinkage methods, which was first suggested by [1] and further illustrated in [3]. The idea is to obtain more precise estimates of model parameters by simply adding a small positive quantity to each diagonal element of $\boldsymbol{X}^\top \boldsymbol{X}$. This will add some bias to coefficient estimates. The ridge estimator $\hat{\boldsymbol{\beta}}^R$, for β is given by

$$\hat{\boldsymbol{\beta}}^R = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{4}$$

Here, $\lambda > 0$ is called the tuning parameter which defines the amount of shrinkage on the estimated coefficients. Ridge estimator can be obtained as the solution to the following minimization problem:

$$\hat{\boldsymbol{\beta}}^R = \underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \text{ subject to } \sum_{j=1}^{p} {\beta_j}^2 \leq s. \quad (5)$$

That is, simply adding a penalty on the size of regression coefficients measured by the L_2 norm. $\sum_{j=1}^{p} {\beta_j}^2 = ||\boldsymbol{\beta}||_2^2$ as the squared distance of the coefficient vector $\boldsymbol{\beta}$ from the origin.

The concept of non-negative garrote (NNG) which was originally suggested in [4] as a better subset selection, is interesting. The lasso [5], which is one of the most widely used shrinkage methods, was motivated by the non-negative garrote. Most of the shrinkage methods such as ridge regression [1], the lasso [5] and the elastic net [6], do not involve OLS estimates in the model estimation process. There are indeed certain situations that we cannot rely on OLS estimates or we cannot obtain them at all. However, OLS estimates possess many statistically desirable characteristics. Hence, it makes sense if someone does not want to ignore OLS estimates when estimating the model parameters. Instead of completely avoiding OLS estimates, we can use the NNG estimation method to adjust OLS estimates to achieve a higher prediction accuracy. This is done by adjusting each OLS estimates $\hat{\beta}_j$. To this end, [4] proposed to find optimum positive constants $c_j's$ such that $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} c_j x_{ij}\hat{\beta}_j$ is a better model than OLS. In order to avoid drastic changes to the OLS model he imposed an intuitively sound constrained $\sum_{j=1}^{p} c_j \leq p$. In other words, he formulated his problem as

$$\text{minimizing} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} c_j\hat{\beta}_j x_{ij})^2, \text{ subject to } \sum_{j=1}^{p} c_j \leq s. \quad (6)$$

The NNG problem can be re-written with Lagrangian multiplier as the problem of finding

$$(\hat{c}_0, \hat{c}_1, \ldots, \hat{c}_p) = \underset{c_0,c_1,\ldots,c_p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} c_j\hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p} c_j.$$

Here as well, $\lambda > 0$ is the tuning parameter that can be obtained using cross-validation or as a value that results in the least prediction error within a sequence of values. Further details on selecting the tuning parameter can be found in [7]. Once we have estimated c_j's, the NNG estimates are obtained as $\hat{\beta}_j^{NNG} = \hat{c}_j\hat{\beta}_j$. One of the desirable properties of the NNG method is that the coefficient estimates do not depend on the scale. Hence, there is no need to scale the variables before the analysis.

The most commonly used regularized regression methods, including ridge regression and the famous lasso have a common weakness. That is, they do not consider the importance of each variable when applying the shrinkage, and all the model coefficients will be shrunken towards zero at a similar rate. As an example, applying ridge regression to estimate a model to predict the severity of heart disease, will result in a similar shrinkage on least important coefficients and most important coefficients. The user has no control over the amount of shrinkage on any coefficient. What if the researcher wants less shrinkage on the larger coefficients, and instead he wants to apply more shrinkage on the least important ones. Suppose due to theoretical and/or practical justifications, we do not want to shrink a set of variables at all. For instance, assume the researcher who deals with heart disease data knows that the exercise level and stress level of the individual have direct impacts on the severity of the heart disease, and he does not want to shrink the effect of those variables. This cannot be achieved with any of the aforementioned regularized regression methods.

There are several other shrinkage methods that were developed to address some of the above mentioned concerns. As an example, the generalized ridge regression approach suggested in [3] can be used to apply unequal shrinkage on different coefficients. However, the way that the generalized ridge approach defines the weights on coefficients does not provide any flexibility for the user to decide which variables should be shrunk more or less. The adaptive lasso suggested by [8], as well as the group lasso [9] are more user-friendly methods in which the user can define the amount of shrinkage on each or a group of coefficients. The adaptive lasso usually penalizes the coefficients inversely proportional to their size. That is, it applies less shrinkage on larger coefficients while applying more shrinkage on smaller coefficients. Because of the subset selection property of the adaptive lasso, smaller coefficients will be set to exactly zero than the regular lasso approach. However, the adaptive lasso is not appropriate if someone wants to retain all the variables in the model. One can use Group lasso to apply shrinkage on groups of variables [9]. Hence, all the parameters in the same group will be either non-zero or zero together. Group lasso is especially useful for factor selection where the factors can be represented by a group of input variables.

Following the non-negative garrote idea, [4] suggests an approach (quadratic garrote) that can shrink coefficients unequally, and at the same time retains all the variables in the model. By observing the nature of the quadratic garrote penalty, we see that the quadratic garrote is capable of shrinking smaller coefficients even faster than the adaptive lasso while keeping the larger coefficients almost untouched. However, we do not find any publication in the literature which has further implemented the idea. Hence, in this paper, we implement the quadratic garrote idea and generalize the quadratic garrote so that it gives the flexibility for the user to decide the level of shrinkage on each variable directly, based on his experience or prior knowledge.

The outline of the paper is as follows. In Sect. 2, we study the quadratic garrote method suggested by [4] and further generalize it to obtain a more practically sound shrinkage approach where the user has some control over the amount of shrinkage on each regression coefficient estimate. There, we study the theoretical properties of the suggested estimators. In Sect. 3, we compare the performance of the quadratic

garrote with the other shrinkage methods using a simulation study under different settings which resemble real-life problems. In Sect. 4, we further illustrate the use of the generalized garrote and the quadratic garrote with the Boston housing dataset. Finally, in Sect. 5, we provide concluding remarks followed by conclusions and a discussion.

2 Quadratic Garrote

Consider the optimization problem of estimating $\{c_j\}_{j=1}^{p}$ to minimize

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} c_j \hat{\beta}_j x_{ij})^2 \text{ subject to } \sum_{j=1}^{p} c_j^2 \leq s, \tag{7}$$

where, $s > 0$ and $\hat{\beta}_j$'s are the OLS estimates. For future reference, we name this approach to be the quadratic garrote (QG). Then the jth QG estimator is obtained as $\hat{\beta}_j^{\,Q} = c_j \hat{\beta}_j$. Equivalently, we can re-write the problem using the Lagrange multiplier as the following minimization problem in terms of c_j's.

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} c_j \hat{\beta}_j x_{ij})^2 + \lambda \sum_{j=1}^{p} c_j^2, \tag{8}$$

where, λ is the tuning parameter. Even though [4] suggested the approach, they did not implement the idea. Nevertheless, the author expects quadratic garrote to be uniformly more accurate than ridge regression and to be almost as stable as the ridge regression.

For the two predictor case, the nature of the QG penalty can be seen in Fig. 1 considering $\hat{\boldsymbol{\beta}} = (1, 2)^\top$. The QG penalty is elliptical in shape. Hence, unlike the ridge penalty, it shrinks the two OLS coefficient estimate differently. In this case, QG method shrinks β_2 two times than β_1. As other shrinkage methods, QG approach also restricts the parameter space of the coefficients by imposing the QG penalty and then selecting the best estimates in the restricted space.

To understand the idea, let's consider an example with two predictors x_1 and x_2, which has been plotted in Fig. 2. In Fig. 2, the observations are generated from a multivariate normal distribution with mean vector $\begin{pmatrix} 5 \\ 10 \end{pmatrix}$ and the covariance matrix $\begin{pmatrix} 2 & 1.8 \\ 1.8 & 3 \end{pmatrix}$. The response y was generated from the model $Y = 10Z_1 + 20Z_2 + \epsilon$, with $\epsilon \sim N(0, 15)$ where Z_1, Z_2 are the standardized predictors.

The contour plots of SSE($\boldsymbol{\beta}$) for the ridge and QG methods are given in Figs. 3 and 4 along with the OLS estimates and the ridge and QG solution for a fixed s. The ridge and QG estimates on the plots were evaluated with the best s using 10-fold

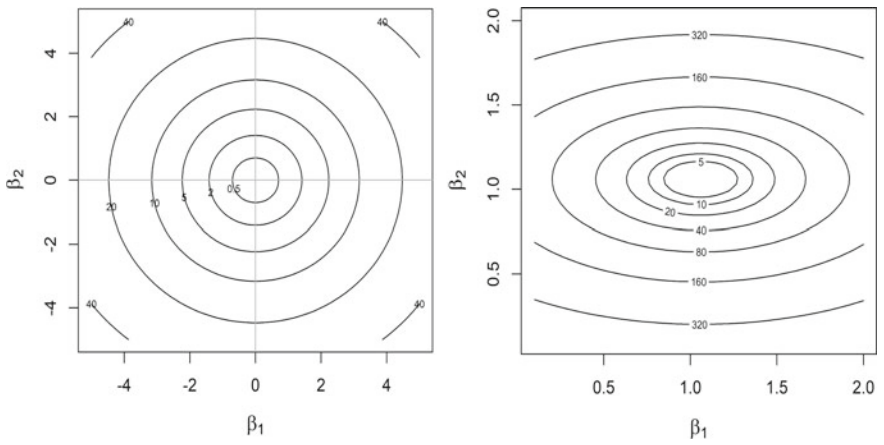


Fig. 1 Ridge constraints for different s values (Left). Each circle represents those $(\beta_1, \beta_2) \in \mathbb{R}$ such that $\beta_1^2 + \beta_2^2 = s$. QG constraints for different s values (Right). Each circle represents those $(\beta_1, \beta_2) \in \mathbb{R}$ such that $\beta_1^2/\hat{\beta}_1^2 + \beta_2^2/\hat{\beta}_2^2 = s$

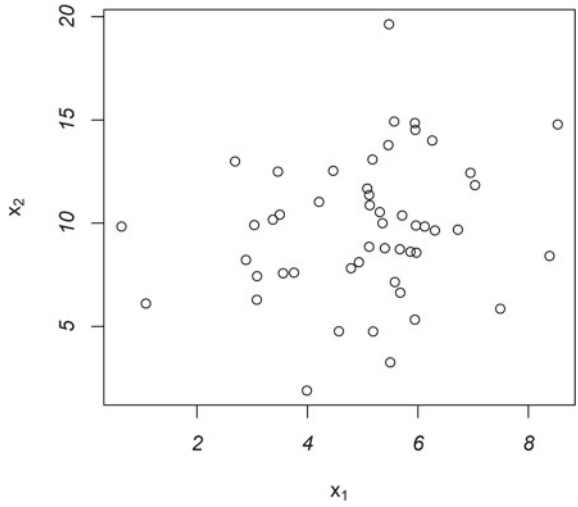


Fig. 2 Scatter plot of a example of size $n = 50$ on two predictors x_1 and x_2

cross-validation errors. As we observe, the ridge and QG estimates are much smaller than the least squares estimates and they lie on the boundary of the penalty functions shown by the circle for ridge and ellipse for QG. That is, for a given s (or λ), the ridge or QG estimates can be found at the point where the first contours of SSE($\boldsymbol{\beta}$) (ellipse) touches the boundary of the penalty region. Figure 4 shows the solution for the QG garrote estimate for the example dataset along with the QG penalty at $s = 0.5$ and the contours of SSE. When we compare the QG solution with the ridge solution

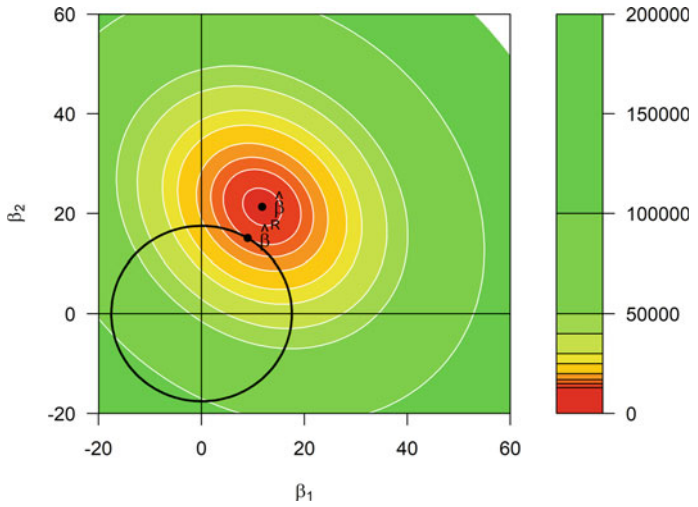


Fig. 3 Contour plot of SSE with the ridge penalty and the solution. Here, $\hat{\boldsymbol{\beta}}$ denotes the OLS estimate of $\boldsymbol{\beta}$ while $\hat{\boldsymbol{\beta}}^{\boldsymbol{R}}$ is the corresponding ridge estimate for specific value of λ

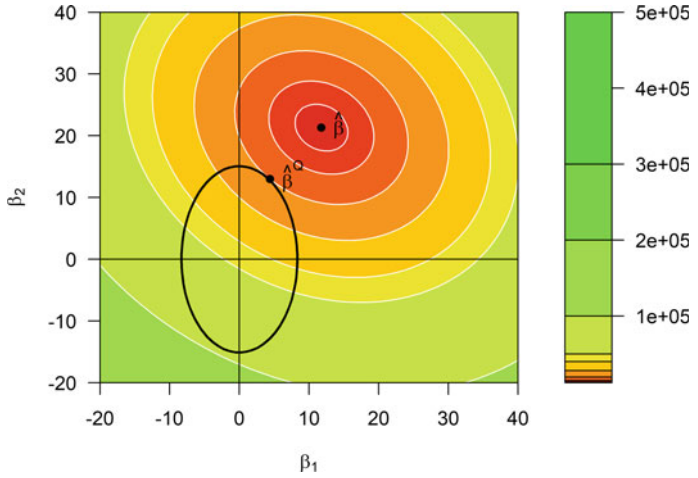


Fig. 4 Contour plot of SSE with the QG penalty and the QG solution at $s = 0.5$

in Fig. 3 where the penalty is a circle, we see that using the quadratic garrote, the larger coefficient ($\hat{\beta}_1$) has been shrunk less than the smaller one.

We can further generalize the quadratic garrote to obtain a more flexible shrinkage approach which has a wide range of practical applications. The following theorem defines the generalized quadratic garrote estimation problem.

Theorem 1 *Consider the multiple linear regression model* $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i; i = 1, \ldots, n$, *where* ϵ_i *are iid with* $E(\epsilon_i) = 0$ *and* $V(\epsilon_i) = \sigma^2$. *Let* $\tilde{\boldsymbol{\beta}}$ *be the generalized quadratic garrote estimate given by*

$$\tilde{\boldsymbol{\beta}} = \underset{\beta_0, \beta_1, \ldots, \beta_p}{argmin} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p} d_j^2 \beta_j^2 \leq s, \tag{9}$$

where, d_j^2*'s are some positive quantities (shrinking factors) which can depend on* $\hat{\beta}_j$ *or they can be fixed constants. Then, in usual matrix notation, the solution for* (9) *is obtained as*

$$\tilde{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{B})^{-1} \boldsymbol{X}^\top \boldsymbol{y},$$

where $\boldsymbol{B}$ *is a* $p \times p$ *diagonal matrix with diagonal elements* d_j^2, *and* λ *is determined such that* $\sum_{j=1}^{p} d_j^2 \beta_j^2 = s$.

We notice that the minimization problem in (9) has some similarity with the generalized ridge regression approach suggested in [3]. However, there are some major differences between the two approaches. We can rewrite minimization problem in (9) with the Lagrange multiplier as

$$Q(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} d_j^2 \beta_j^2. \tag{10}$$

Here, $\lambda \sum_{j=1}^{p} d_j^2 \beta_j^2$ is the penalty term. For the comparison purpose, assume d_j^2s to be unknown and let $\lambda d_j^2 = \lambda_j$. Then we have

$$Q(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \sum_{j=1}^{p} \lambda_j \beta_j^2, \tag{11}$$

which is the generalized ridge regression problem.

Since we assume d_j^2 to be known in the generalized garrote approach, the size of the penalty is defined with the single tuning parameter λ. Contrary, in the generalized QG there are p number of tuning parameters. Reference [3] defined the generalized ridge regression in the eigenvector space. Hence, prior to applying the generalized ridge method, we have to project the design matrix into the eigenspace. As a result, the tuning parameter λ_j does not correspond to the level of shrinkage on jth coefficient estimate. However, in the generalized garrote approach, λd_j^2 directly defines the amount of shrinkage on the corresponding coefficient estimate.

We can easily derive the ridge estimator and the quadratic garrote estimator with Theorem 1.

Example 1: Ridge regression

Setting all d_j^2's to be 1 in (10), we have the classical ridge regression problem. Then, $\boldsymbol{B}$ is the $p \times p$ identity matrix and we have the solution, $\hat{\boldsymbol{\beta}}^R = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Example 2: Quadratic garrote

The quadratic garrote minimization problem in (9) can be rewritten with the Lagrange multiplier as

$$Q(c_1, c_2, \ldots, c_p, \lambda) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} c_j \hat{\beta}_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{k} c_j^2. \tag{12}$$

Let $\hat{\beta}_j^{\,Q} = c_j \hat{\beta}_j$ be the new quadratic garrote coefficients with $c_j = \hat{\beta}_j^Q / \hat{\beta}_j$. Substituting c_j in (12) we obtain the quadratic garrote estimates as

$$\hat{\boldsymbol{\beta}}^Q = \underset{\beta_0, \beta_1, \ldots, \beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \left(\frac{1}{\hat{\beta}_j^2}\right) \beta_j^2. \tag{13}$$

Applying Theorem 1 with $d_j^2 = 1/\hat{\beta}_j^{\,2}$ we derive the vector of the quadratic garrote estimator as

$$\hat{\boldsymbol{\beta}}^{(Q)}(\lambda) = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{B})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

where $\boldsymbol{B} = \operatorname{diag}(1/\hat{\beta}_j^{\,2})$ and $\hat{\beta}_j$'s are the OLS estimators.

The generalized quadratic garrote estimator possesses a practical advantage over other shrinkage methods. Imagine one wants to keep a specific set of variables unshrunk or with minimum shrinkage while applying more shrinkage on some other variables. With the quadratic garrote, we can arbitrarily decide the level of shrinkage on each variable while maintaining a reduction in MSE. For two variables case, Fig. 5 illustrates the idea visually. Contours represent the SSE of OLS estimation and the minimum SSE is achieved at $\hat{\boldsymbol{\beta}}$. Shaded regions represent the generalized quadratic garrote constraints for different d_j vectors. We see that the nature of the constraint changes with d_j. As we see in Fig. 5a, when each $d_j = 1$, the constraint is the same as the ridge constraint, and we get with ridge regression solution. Observe, Fig. 5b, where $d_1^2 < d_2^2$. In this case, β_2 is twice constrained than β_1. Hence, estimated coefficient for β_2 is shrunken more towards zero than β_1. This setting is good if we want less shrinkage on β_1. This story is the opposite for Fig. 5c. Figure 5d shows the case when there is no penalty on β_1. In this case, only β_2 will be constrained and β_1 can take any value. In this specific example, the estimate of β_1 is less than the OLS estimate. However, this is not always true. $\hat{\beta}_1^Q$ can be even larger than the corresponding OLS estimate since it is not constrained.

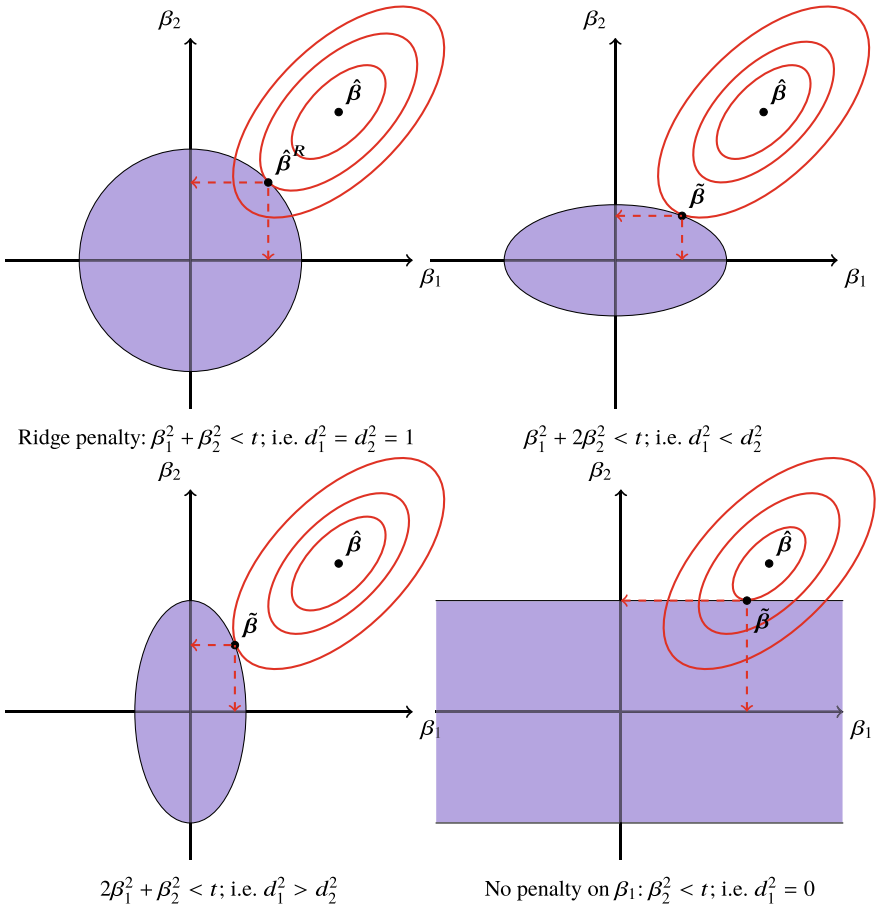


Fig. 5 Nature of the penalty on parameters with different d_j^2s for two predictor case

Variance and Bias

When $\boldsymbol{B}$ is independent of $\boldsymbol{y}$, we can easily obtain the expectation and variance of $\tilde{\boldsymbol{\beta}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}}$. Since $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$,

$$E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{Z}\boldsymbol{\beta}, \tag{14}$$

and

$$Var(\tilde{\boldsymbol{\beta}}) = \boldsymbol{Z}Var(\boldsymbol{\beta})\boldsymbol{Z}^\top = \sigma^2\boldsymbol{Z}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{Z}^\top, \tag{15}$$

where $\boldsymbol{Z} = (\boldsymbol{I}_p + \lambda\boldsymbol{B}(\boldsymbol{X}^\top\boldsymbol{X})^{-1})^{-1}$.

It is obvious that the quadratic garrote estimator is biased ($E(\tilde{\boldsymbol{\beta}}) \neq \boldsymbol{\beta}$). Consider the mean squared error of $\tilde{\boldsymbol{\beta}}$ given by

$$\text{MSE}(\tilde{\boldsymbol{\beta}}) = E\left[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]. \tag{16}$$

This can be decomposed as

$$\begin{aligned}\text{MSE}(\tilde{\boldsymbol{\beta}}) &= \sigma^2 \text{Trace}\left[\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{B})^{-2}\right] + \lambda^2 \boldsymbol{\beta}^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{B} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{B})^{-1} \boldsymbol{\beta}, \\ &= \gamma_1(\lambda) + \gamma_2(\lambda), \end{aligned} \tag{17}$$

where, $\gamma_1(\lambda)$ is the sum of the total variation of $\tilde{\boldsymbol{\beta}}$ vector or simply the sum of the diagonal elements of $Var(\tilde{\boldsymbol{\beta}})$ in (15). $\gamma_2(\lambda)$ is the squared bias of $\tilde{\boldsymbol{\beta}}$. It is easy to see that $\gamma_1(\lambda)$ is monotonically decreasing in λ. We can show that

$$\lim_{\lambda \to 0^+} \gamma_1(\lambda) = \sigma^2 \text{Trace}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right], \text{ and } \quad \lim_{\lambda \to \infty} \gamma_1(\lambda) = 0.$$

Furthermore, we show that $\gamma_2(\lambda)$ is a monotonically increasing function in λ and

$$\lim_{\lambda \to 0^+} \gamma_2(\lambda) = 0, \quad \text{and} \quad \lim_{\lambda \to \infty} \gamma_2(\lambda) = \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

That is, $\gamma_2(\lambda)$ is bounded above by the squared length of $\boldsymbol{\beta}$. When the design matrix $\boldsymbol{X}$ is orthonormal,

$$\gamma_1^*(\lambda) = \sigma^2 \sum_{j=1}^{p} \frac{1}{(1 + \lambda d_j^2)^2} \quad \text{and} \quad \gamma_2^*(\lambda) = \sum_{j=1}^{p} \frac{d_j^4 \beta_j^2}{(\frac{1}{\lambda} + d_j^2)^2}. \tag{18}$$

We see that $\gamma_1^*(\lambda)$ is a monotonically decreasing function in λ with

$$\lim_{\lambda \to 0^+} \gamma_1^*(\lambda) = \sigma^2 p, \text{ and } \lim_{\lambda \to \infty} \gamma_1^*(\lambda) = 0,$$

and $\gamma_2^*(\lambda)$ is a monotonically increasing function in λ with the limits as below:

$$\lim_{\lambda \to 0^+} \gamma_2^*(\lambda) = 0, \text{ and } \lim_{\lambda \to \infty} \gamma_2^*(\lambda) = \sum_{i=1}^{p} \beta_j^2 = \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

That is, $\gamma_2^*(\lambda)$ is bounded above by the length of $\boldsymbol{\beta}$. Hence, the limits of $\text{MSE}(\tilde{\boldsymbol{\beta}})^*$ are obtained as

$$\lim_{\lambda \to 0^+} \text{MSE}(\tilde{\boldsymbol{\beta}})^* = \sigma^2 p, \tag{19}$$

which is the MSE($\hat{\boldsymbol{\beta}}$), where $\hat{\boldsymbol{\beta}}$ is the vector of OLS estimators, and

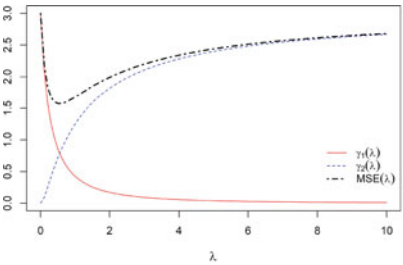
$$\lim_{\lambda \to \infty} \text{MSE}(\tilde{\boldsymbol{\beta}})^* = \boldsymbol{\beta}^\top \boldsymbol{\beta}. \tag{20}$$

It is easy to show that, MSE($\tilde{\boldsymbol{\beta}}$)* is not a monotone function in λ. However, we can show that MSE($\tilde{\boldsymbol{\beta}}$)* first goes through a minimum before it increases. We can obtain an upper bound for λ as

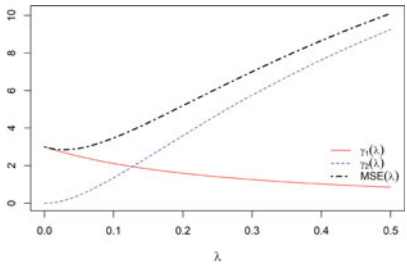
$$0 < \lambda < \sigma^2 / \max(\beta_j^2 d_j^4). \tag{21}$$

Hence, MSE($\tilde{\boldsymbol{\beta}}$)* will be decreasing for some $\lambda < \sigma^2 / \max(\beta_j^2 d_j^4)$. So, there exist some $\tilde{\boldsymbol{\beta}}$, such that MSE($\tilde{\boldsymbol{\beta}}$)* is lower than $\sigma^2 p = \text{MSE}(\hat{\boldsymbol{\beta}})$.

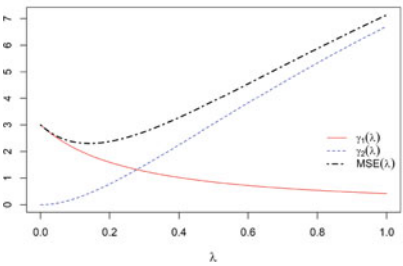
Figure 6a–d show the behavior of these error components for different $\{d_j\}$ vectors and different $\boldsymbol{\beta}$ vectors. In all the examples, we see the expected theoretical behavior for the three curves. Also we notice that, when large coefficients get larger weights, the minimum of MSE($\tilde{\boldsymbol{\beta}}$)* achieves quickly and we see a drastic drop in MSE($\tilde{\boldsymbol{\beta}}$)*. In contrast, when larger coefficients get smaller weights, minimum of MSE($\tilde{\boldsymbol{\beta}}$)* achieves slowly and also we do not observe a significant drop in MSE($\tilde{\boldsymbol{\beta}}$)*. This



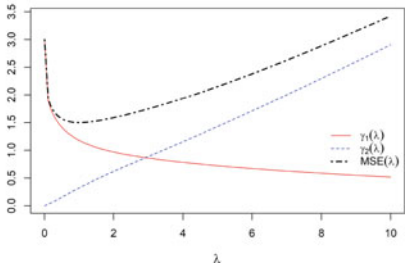
$\boldsymbol{\beta} = (2, 2, 2)^\top$, $\boldsymbol{B} = \text{diag}\{1, 2, 3\}$ and $\sigma = 1$



$\boldsymbol{\beta} = (0.2, 2, 5)^\top$, $\boldsymbol{B} = \text{diag}\{1, 2, 3\}$ and $\sigma = 1$.



$\boldsymbol{\beta} = (5, 2, 0.2)^\top$, $\boldsymbol{B} = \text{diag}\{1, 2, 3\}$ and $\sigma = 1$.



$\boldsymbol{\beta} = (0.2, 2, 5)^\top$,
$\boldsymbol{B} = \text{diag}\{\left(\frac{1}{0.2^2}\right)^2, \left(\frac{1}{2^2}\right), \left(\frac{1}{5^2}\right)\}$ and $\sigma = 1$.

Fig. 6 MSE of generalized quadratic garrote estimator

is an indicator that the idea of the quadratic garrote might be a better substitute for the ridge regression since it defines d_j^2's inversely proportional to the squared OLS estimates. A detailed study of the behavior of the generalized quadratic garrote coefficients under orthonormal design assumption can be found in [7].

When B depends on $\boldsymbol{y}$, we cannot obtain a closed form solution for the expectation and variance of $\tilde{\boldsymbol{\beta}}$. However, by applying the Taylor series expansion for $\boldsymbol{B}$, we can obtain approximate solutions. Let $d_j^2 = f_j(\hat{\beta}_j)$, then it is easy to show that

$$\tilde{\boldsymbol{\beta}} \approx (\boldsymbol{I}_p + \lambda \boldsymbol{D}(\boldsymbol{X}^\top \boldsymbol{X})^{-1})^{-1} \hat{\boldsymbol{\beta}}, \tag{22}$$

where, $\boldsymbol{D} = \text{diag}(f_j(\beta_j))$. Even if $(\boldsymbol{I}_p + \lambda \boldsymbol{D}(\boldsymbol{X}^\top \boldsymbol{X})^{-1})^{-1} \hat{\boldsymbol{\beta}}$ is an approximation for $\tilde{\boldsymbol{\beta}}$, we have the advantage of $\boldsymbol{D}$ being independent of $\boldsymbol{y}$. Hence, we can use all the results that we derived above, where $\boldsymbol{B}$ was independent of $\boldsymbol{y}$.

3 Simulation Study

In this section, we perform a simulation study considering different settings which resemble real-life problems such as multicollinearity issue, high dimensional setting, etc. The performance of quadratic garrote will be compared with ridge and lasso estimation methods in each setting. OLS estimation will be used as the benchmark. Consider the following three settings:

1. Sparse setting,
2. Nearly-sparse setting,
3. High dimensional setting.

In each setting, data were generated using the model $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, ..., n$. Note that we can write the linear regression model as $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$. $\boldsymbol{x}_i$'s were generated from a $MVN(\boldsymbol{0}, \Sigma)$, where ijth entry of Σ is $\rho^{|i-j|}$ and $\rho = 0.5$. The sample size was 100 in each setting. 10-fold cross-validation method was used to select the best tuning parameter for each of the shrinkage method. Finally, mean-squared prediction error was used to compare the prediction accuracy between models with 5-fold cross-validation.

Sparse Setting

This setting has been used in [5] to compare the performance of the lasso with other shrinkage methods. We intend to see how the quadratic garrote model performs when there are a considerable fraction of zeros in $\boldsymbol{\beta}$. We set the true population coefficient vector to be $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$. The random error was set to have $\sigma = 3$. The signal to noise ratio of the sample data was approximately 2. Consider an additive model $Y = f(X) + \epsilon$. Then, according to [10] the signal to noise ratio is defined as,

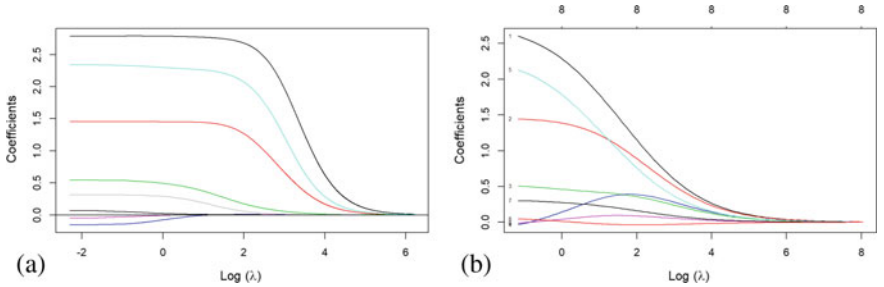


Fig. 7 Trace plot of quadratic garrote estimators (**a**) and ridge estimators (**b**) for the sparse setting

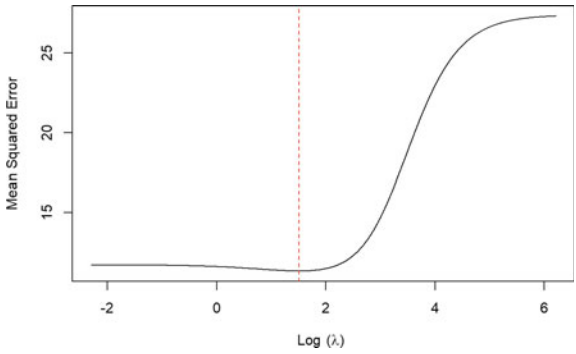


Fig. 8 Mean squared error for quadratic garrote under the sparse setting

$$\text{Signal to noise ratio} = \frac{Var(f(X))}{Var(\epsilon)}. \tag{23}$$

The signal to noise ratio for the multiple linear model is given by $Var(X^\top \beta)/Var(\epsilon)$. The given value of the signal to noise ratio of 2, is the estimated signal to noise ratio of the sample data, calculated as the ratio of the sample variance of $\boldsymbol{X\beta}$ vector to the variance of the residual error vector.

Figure 7 shows the trace plots of quadratic garrote estimates and ridge estimates. We can see an interesting behavior of quadratic garrote estimators (Fig. 7a) compared to ridge estimators (Fig. 7b). In the quadratic garrote trace plot, we notice that the larger coefficients do not shrink until the tuning parameter (λ) is very large. However, the smaller coefficients approach zero very fast even for a very small λ with almost no effect on large coefficients. This property cannot be seen with ridge regression. The ridge regression method shrinks all coefficients from the beginning.

Figure 8 shows the mean squared error for quadratic garrote against λ. The best coefficients were selected to have minimum MSE (Vertical line on Fig. 8 represents the location of minimum MSE). The estimates are given in the Table 1. Quadratic garrote estimates for the true parameters 3 and 2 are very close to their OLS estimates than the ridge estimates. On the other hand, the estimated coefficients for zero-valued parameters are very small in the quadratic garrote than the ridge estimates.

Table 1 Estimated coefficients under the sparse setting

	β_j	OLS	QG	Ridge	Lasso
X1	3.00	2.7865	2.7456	2.5871	2.6667
X2	1.50	1.4571	1.3995	1.4421	1.4136
X3	0.00	0.5471	0.2717	0.5042	0.3809
X4	0.00	−0.1553	0.0038	−0.0231	0
X5	2.00	2.3489	2.2097	2.1141	2.1432
X6	0.00	−0.0522	0.0008	−0.0075	0
X7	0.00	0.3161	0.1126	0.3000	0.2130
X8	0.00	0.0707	0.0030	0.0449	0

Table 2 Mean squared errors of the models under sparse setting

	MSE
OLS	12.2485
Quadratic garrote	11.7609
Ridge	12.1020
Lasso	11.7974

Table 2 summarizes the MSEs of each of our models. For the sparse setting, quadratic garrote gives the least prediction errors than all the other methods. It even has better prediction accuracy than the lasso, which has the second best MSE.

Nearly-Sparse Setting

In this setting, we set the population coefficients vector to have a few large values and others are set to have values closer to zero. Let $\boldsymbol{\beta} = (3, 1.5, z, z, 2, z, z, z, z, z)^\top$, where $z \sim Unif(0, b)$, $0 < b \leq 1$. We can set z coefficients to have smaller values by selecting a very small b. The setting was repeated for $b \in \{0.1, 0.5, 1\}$, and the random error σ was selected to be $\sqrt{2}$ such that the signal to noise ratio is around 10 for each case. A similar setting has been suggested in [11].

The trace plots for the quadratic garrote coefficient estimates and the ridge estimates for different b values are shown in Fig. 9. The behavior of trace plots for both quadratic garrote and ridge is similar to their behavior under the sparse setting. In all values of b, we see that the quadratic garrote does not shrink the larger coefficients unless λ is very large. On the other hand, smaller coefficients shrink towards zero very fast even for a small λ.

Estimated quadratic garrote coefficients at the best λ which was determined by cross-validation for each scenario, are presented in Table 3. When $b = 1$, we do not observe much shrinkage on each coefficient in any shrinkage method except

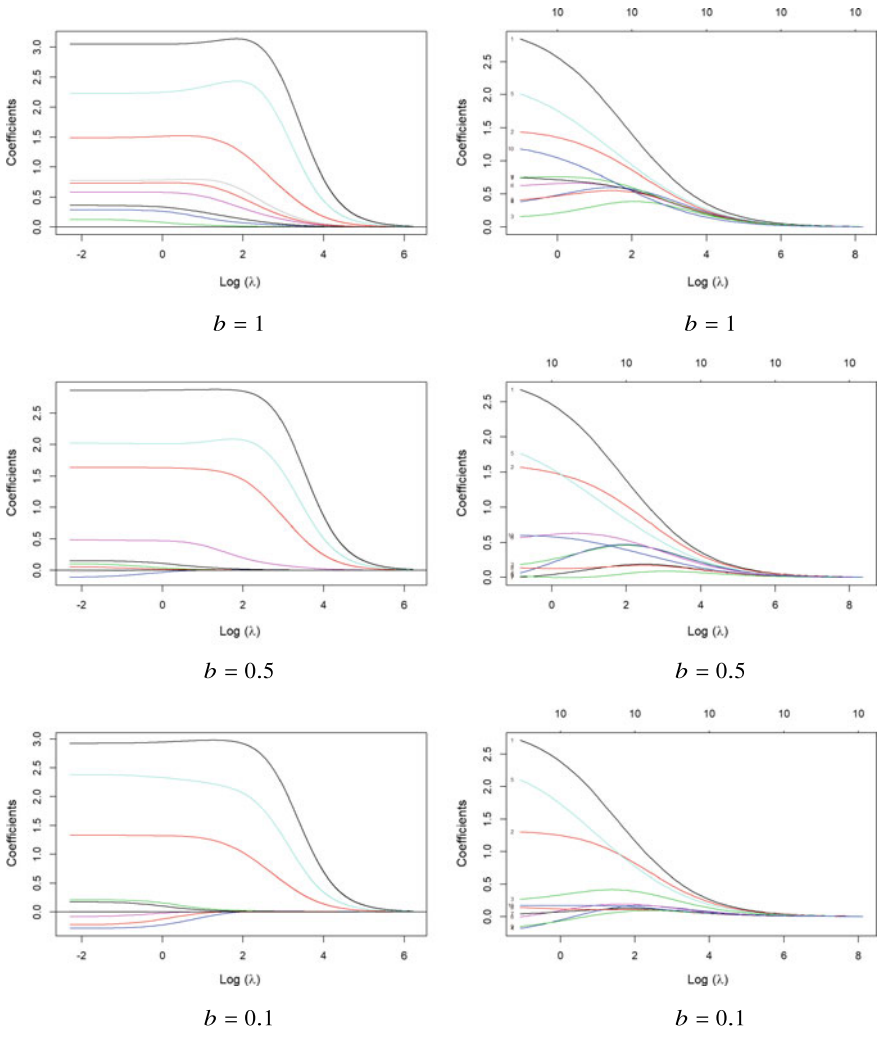


Fig. 9 Trace plot of quadratic garrote estimates (Left) and ridge estimates (Right) of the nearly sparse setting

for smaller coefficients. However, when $b = 0.5$ and $b = 0.1$, the coefficients are shrunk in all the models. In these scenarios, unlike ridge regression, the quadratic garrote shrinks the smaller coefficient almost to zero with no shrinkage on the larger coefficients.

Mean squared cross-validation errors for each approach are summarized in Table 4. We see that, for all values of b that we used, the quadratic garrote performs better than the ridge regression approach. As b becomes smaller, quadratic garrote does better prediction than all the other methods. In all the cases, we see

Table 3 Estimated QG coefficients and prediction errors under the nearly-sparse setting

	Beta	OLS	QNNG	Ridge	Lasso
(a) $b = 1$					
X1	3.0000	2.9778	2.9776	2.8094	2.9419
X2	1.5000	1.7939	1.7751	1.7103	1.7455
X3	0.1966	−0.0158	−0.0001	0.0879	0
X4	0.7164	0.8819	0.8559	0.8474	0.8375
X5	2.0000	2.1116	2.1510	1.9420	2.1165
X6	0.3621	0.3048	0.2457	0.4195	0.2757
X7	0.3911	0.6722	0.6691	0.6455	0.6427
X8	0.8133	0.6658	0.6467	0.6483	0.6372
X9	0.4280	0.5192	0.5016	0.5416	0.5003
X10	0.9592	0.8133	0.8130	0.7347	0.7780
(b) $b = 0.5$					
X1	3.0000	2.8638	2.8721	2.6576	2.8328
X2	1.5000	1.6368	1.6167	1.5636	1.6081
X3	0.1539	0.1109	0.0162	0.1905	0.0537
X4	0.1288	−0.1171	−0.0019	0.0785	0
X5	2.0000	2.0214	2.0270	1.7472	1.9712
X6	0.2762	0.4946	0.4278	0.5760	0.4568
X7	0.0282	−0.0369	0.0024	0.0062	0
X8	0.2343	0.1597	0.0629	0.1342	0.1467
X9	0.2419	0.0654	0.0081	0.0188	0.0044
X10	0.4062	0.5987	0.5916	0.5993	0.5940
(c) $b = 0.1$					
X1	3.0000	2.9236	2.9779	2.6933	2.8817
X2	1.5000	1.3315	1.2467	1.3001	1.2416
X3	0.0336	0.2110	0.0464	0.2706	0.0631
X4	0.0464	−0.2883	−0.0655	−0.1734	0
X5	2.0000	2.3813	2.2207	2.0770	2.0929
X6	0.0061	−0.0905	0.0000	0.0034	0
X7	0.0197	0.0183	0.0001	0.0442	0
X8	0.0474	0.1739	0.0228	0.1394	0
X9	0.0301	−0.2275	−0.0123	−0.1473	0
X10	0.0607	0.1711	0.0346	0.1691	0.0521

that the ridge regression approach gives the highest prediction errors. As b becomes smaller, prediction errors of the ridge regression approach are even higher than the OLS approach.

Table 4 Mean squared errors of the models under nearly sparse setting

	MSE		
	$b = 1$	$b = 0.5$	$b = 0.1$
OLS	3.0830	2.3372	2.1892
Quadratic garrote	3.0643	2.2154	2.1114
Ridge	3.0753	2.3774	2.2375
Lasso	3.0330	2.2203	2.1419

High Dimensional Setting

As we discussed in Sect. 1, the problems associated with the OLS estimator in the high dimensional setting are more serious than they are in the lower dimensions. Hence, it is important to examine how the suggested model performs in the high dimensional setting. To this end, we consider three scenarios.

In the first scenario, the coefficient vector consists of the majority of very small coefficients and some relatively large coefficients which represent the important predictors. Let's take

$$\boldsymbol{\beta} = (z, z, ..., z, 1, 1, 1, 1, 1, z, z, ..., z, 4, 4, 4, 4, 4)^{\top},$$

where z's are small positive coefficients independently generated from a $Gamma(1, 10)$ distribution. There are 30 parameters in four blocks, where the first and third blocks contain 10 z values in each block. σ was set to be 5, which gave approximately a signal to noise ratio of 8 in the sample.

In the second scenario, we add some more moderately large coefficients which have the same number of smaller coefficients. Here we consider

$$\boldsymbol{\beta} = (3, 3, 3, 3, 3, z, z, ..., z, 1, 1, ..., 1, z, z, ..., z, 2, 2, 2, 2, 2)^{\top},$$

where z's are iid from a $Gamma(1, 10)$ distribution. There are 40 parameters in five blocks, where the second to fourth blocks contain 10 values in each block. σ was set to be 5, and the signal to noise ratio was approximately 11.

In the third scenario, coefficient vector contains many smaller coefficients, and some moderately large coefficients with few very large coefficients. We set

$$\boldsymbol{\beta} = (z, z, ..., z, 5, 5, 5, z, z, ..., z, 1, 1, 1, 1, 1, z, z, ..., z, 10, 10)^{\top},$$

where z's are iid from a $Gamma(1, 10)$ distribution. It contain 40 parameters in six blocks, where each of the z block contain 10 values. $\sigma = 7$ and the signal noise ratio was approximately 10.

The trace plot of each scenario is shown in Fig. 10. The dotted vertical line represents the λ with respect to the model with minimum MSE obtained using 10-

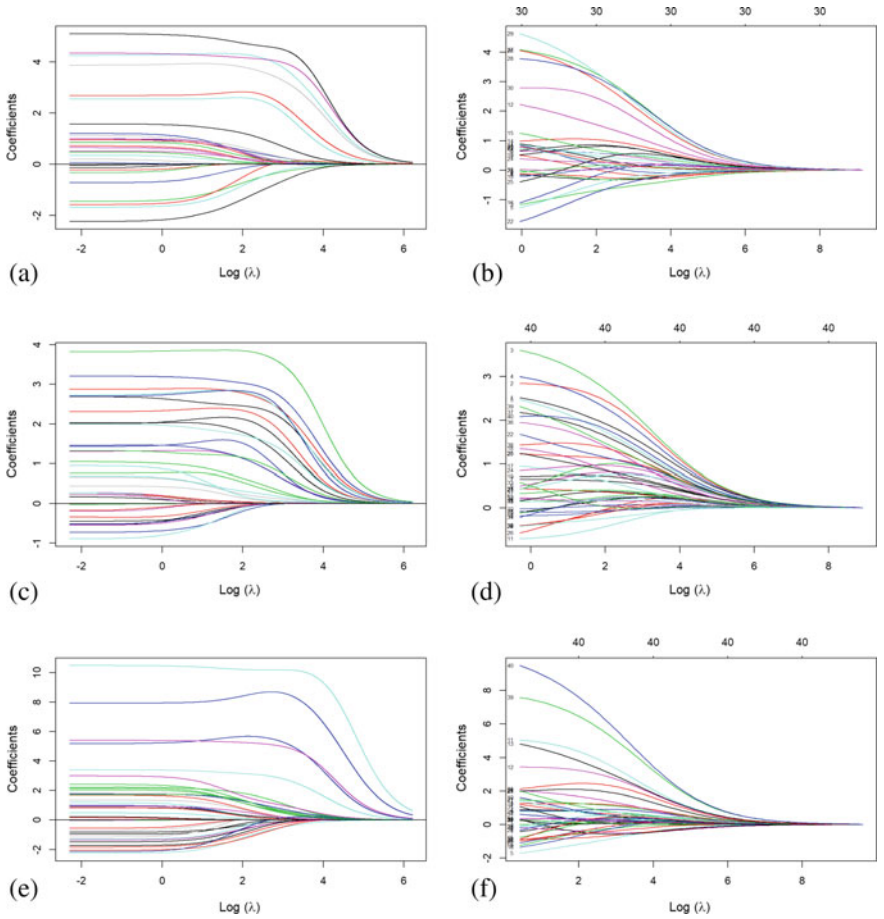


Fig. 10 Trace plot of quadratic garrote estimates (Left) and ridge estimates (Right) of the high dimensional setting where **a**, **b**—Scenario 1, **c**, **d**—Scenario 2, **e**, **f**—Scenario 3 (Dotted line indicates the best lambda w.r.t. minimum MSE)

fold cross-validation. Even though the quadratic garrote does not do subset selection, but when we observe the trace plots of quadratic garrote estimates (Fig. 10a, c, e), it can be seen that at the best λ, a large number of coefficients has been shrunken towards zero with a very little or almost no effect on the large coefficients. This is an important property in the high dimensional setting because we can have a better prediction while keeping the most important variables unchanged. But in the ridge trace plots (Fig. 10b, d, f), we can see that there all the coefficients shrink towards zero similarly.

The prediction error of each model is presented in Table 5. In the first scenario, where there are many small coefficients with few moderately large coefficients, the ridge regression does slightly better than all the other approaches. But in Scenario

Table 5 Mean squared errors of the models under high dimensional setting

	MSE		
	Scenario 1	Scenario 2	Scenario 3
OLS	33.8628	35.5502	101.8177
Quadratic garrote	32.5943	25.5129	70.9068
Ridge	29.6387	25.8657	84.4298
Lasso	30.6148	21.5234	68.9327

2, where there is the same number of small coefficients and moderately large coefficients, quadratic garrote performs slightly better than the ridge. However, in Scenario 3, where there are a large number of small coefficients with a few large coefficients and a few very large coefficients, the quadratic garrote performs well with a significantly low prediction error compared to the ridge regression model. In this case, the quadratic garrote model is even competitive with the lasso. However, the lasso outperforms both the quadratic garrote and the ridge regression especially in the high dimensional setting in terms of the prediction error.

4 Example: The Boston Housing Dataset

In this section, we further illustrate the use of the generalized garrote and the quadratic garrote with the Boston housing dataset. This is a famous dataset that is readily available in the MASS library in R. The dataset consists of housing information in suburbs of Boston in 1970, which has been first cited in [12]. The dataset contains 506 observations with 14 attributes namely,

- crim: per capita crime rate by town.
- zn: proportion of residential land zoned for lots over 25,000 sq.ft.
- indus: proportion of non-retail business acres per town.
- chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- nox: nitrogen oxides concentration (parts per 10 million).
- rm: average number of rooms per dwelling.
- age: proportion of owner-occupied units built prior to 1940.
- dis: weighted mean of distances to five Boston employment centers.
- rad: index of accessibility to radial highways.
- tax: full-value property-tax rate per $10,000.
- ptratio: pupil-teacher ratio by town.
- black: $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town.
- lstat: percentage of lower status of the population.
- medv: median value of owner-occupied homes in $1000s.

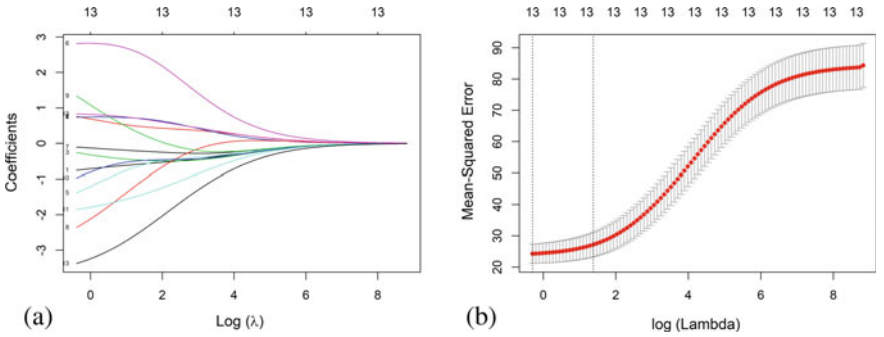


Fig. 11 **a** Solution path of ridge regression approach for the Boston housing dataset. **b** Cross validation error of ridge regression for the Boston housing dataset

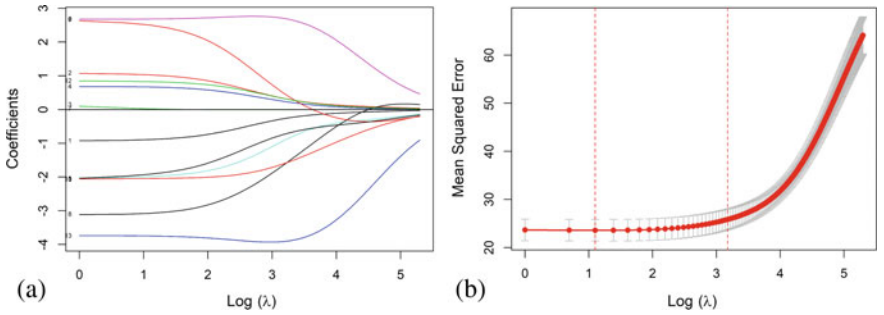


Fig. 12 **a** Solution path of quadratic garrote for Boston dataset. **b** Cross validation error of quadratic garrote for Boston dataset

Here as well suppose that we want to build a model to predict medv with all the other variables as predictors. The solution path for the quadratic garrote can be found in Fig. 12a. In the ridge regression solution path in Fig. 11a, we noticed that, as we increase λ, all the coefficients shrink towards zero at a similar rate from the beginning. On the other hand, quadratic garrote shrinks the smaller coefficient faster than larger coefficients. The largest coefficients shrink only for large values of λ and resist shrinking otherwise. The 10-fold cross-validation error plot of quadratic garrote for the Boston dataset can be found in Fig. 12b. Two vertical dotted lines represent the λ values with respect to the minimum prediction error and prediction error with the one standard error rule. The error bars represent the standard deviation of the mean squared error values at each λ.

Suppose the researcher who conducts the study is told that according to a previous study the crime rate, distance to work, and the number of rooms in the dwelling are the key factors that determine the house price and he should not shrink the effect of those variables in the model. Also, he wants only a moderate shrinkage on the variables which represent air pollution, the age of the dwelling, and accessibility to highways. For this situation, the researcher can use the gen-

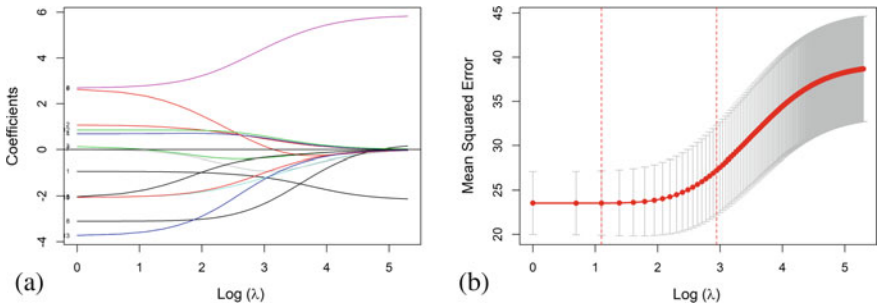


Fig. 13 **a** Solution path of generalized quadratic garrote for Boston dataset. **b** Cross validation error of generalized quadratic garrote for Boston dataset

eralized quadratic garrote with a user-defined vector of shrinking factors, such as $\{d_j\} = (0, 1, 1, 1, 0.5, 0, 0.5, 0, 0.5, 1, 1, 1, 1)^\top$. The shrinking factor $d_j = 0$ avoids imposing any penalty on the corresponding coefficient estimate. Higher the value of d_j, the higher the shrinkage on the corresponding coefficient. The solution path of the generalized quadratic garrote with the above defined shrinking factors can be found in Fig. 13a and the prediction error is presented in Fig. 13b. We observe that as we increase λ, some model coefficients actually increase instead of shrinking towards zero. Those are the parameters that we omit from shrinking by setting the corresponding d_j's to be exactly zero. One can confirm those variables which do not shrink by observing the estimated model coefficients in Table 6.

Table 6 summarizes the results of two quadratic garrote models along with OLS, the ridge, and the lasso results. All models have been evaluated at the λ which gives the minimum prediction error. Compared to the ridge regression and the lasso, the quadratic garrote does minimum shrinkage on the larger coefficients, and on the other hand, it shrinks smaller coefficients by a larger factor than the ridge or the lasso. Generalized quadratic garrote, in which we arbitrarily defined shrinking factors, does what we intended. It does not shrink the coefficients of the variables crim, rm, and dis. Those are the variables corresponding to the d_j's that we set to be zero. It applies minimum shrinkage on the coefficients of nox, age, and rad since we set a smaller shrinking factor $d_j = 0.5$ for them. However, other coefficients where we set a larger d_j values have been shrunk towards zero by a larger factor compared to other shrinkage methods. One of the most important things that we notice is, that the prediction errors of both of the suggested approaches are slightly smaller than the ridge regression.

In Sect. 3, we saw that the quadratic garrote uniformly performs well under different simulation settings. In this section, we further confirmed that the QG method and the generalized QG approach also performed well with real data. The prediction accuracy of suggested methods was better than the ridge regression for most of the cases and, they were competitive to lasso as well.

Table 6 Coefficient estimates and cross validation errors for each shrinkage method

	OLS	QG	GQG (d_j^2)	Ridge	Lasso
crim	−0.9291	−0.8927	−0.9373 (0.00)	−0.7441	−0.8600
zn	1.0826	1.0272	1.0205 (1.00)	0.7450	0.9834
indus	0.1410	0.0221	0.0008 (1.00)	−0.2765	0.0000
chas	0.6824	0.6688	0.6960 (1.00)	0.7381	0.6836
nox	−2.0588	−1.9845	−2.0121 (0.25)	−1.3396	−1.9101
rm	2.6769	2.6836	2.7907 (0.00)	2.8215	2.7096
age	0.0195	0.0001	−0.0718 (0.25)	−0.1113	0.0000
dis	−3.1071	−3.0787	−3.0971 (0.00)	−2.3029	−2.9734
rad	2.6649	2.5045	2.3369 (0.25)	1.2771	2.2667
tax	−2.0788	−1.9181	−1.7312 (1.00)	−0.9274	−1.7098
ptratio	−2.0626	−2.0454	−2.0110 (1.00)	−1.8367	−2.0193
black	0.8501	0.8277	0.8506 (1.00)	0.8258	0.8281
lstat	−3.7473	−3.7486	−3.5660 (1.00)	−3.3445	−3.7313
MSE	23.1516	23.0843	23.1137	23.5036	23.1050

5 Discussion

In this paper, we developed the quadratic garrote method suggested by [4], and we further extended the idea to obtain the generalized quadratic garrote which is more flexible in the sense that the user can control the amount of shrinkage on each coefficient estimate. We derived a closed-form solution for the quadratic garrote problem and studied the theoretical properties of the suggested estimator such as variance, expectation, and bias. In addition, through simulation studies under different settings and with an example, we showed that the quadratic garrote is a worthy substitute for ridge regression. Furthermore, with the Boston housing dataset, we illustrated how to use the generalized ridge regression with a predefined amount of shrinkage on each coefficient based on one's experience or prior knowledge.

In high dimensional setting where the majority of population regression coefficients are small in size and few of the coefficients are considerably larger than the rest, we saw that the quadratic garrote is capable of shrinking the smaller coefficients while keeping the larger coefficients almost unchanged. Another most important observation is, in the aforementioned setting, the presented methods showed much lower prediction errors compared to the ridge regression or OLS models.

To find the tuning parameter, one can use different methods such as k-fold cross-validation and the little bootstrap procedure. Non-negative garrote with the little bootstrap method is recommended when explanatory variables are fixed [13], and k-fold cross-validation is used when explanatory variables are random [4]. However, for high dimensional data, the k-fold cross-validation method is less computationally intensive than the bootstrap method and also cross-validation is commonly used for calculating the prediction error to select the best tuning parameter in many shrinkage

methods such as the ridge regression and lasso [14]. Hence, in this work, we used the k-fold cross-validation procedure to evaluate the tuning parameters for all the models.

Even though the suggested shrinkage method meets our objectives, we can point out some limitations. First, since we developed quadratic garrote following the idea behind the non-negative garrote, the regression coefficient estimates of each method depend on OLS estimates. Hence, the quadratic garrote fails when OLS estimates are infeasible. Also, quadratic garrote does not perform subset selection. Another concern with the generalized quadratic garrote is its subjectiveness. As we saw, in generalized quadratic garrote, the user can control the amount of shrinkage on each coefficient by defining the appropriate shrinking factors. This might incorporate some additional subjectiveness to the coefficient estimates.

Acknowledgements Authors acknowledge the partial research support from the Natural Sciences and Engineering Research Council of Canada (NSERC). We acknowledge the reviews received by the referees.

References

1. Horel, A. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, *58,* 54–59.
2. Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q., & Lillard Jr, J. (2014). A study of effects of multicollinearity in the multivariable analysis. *International Journal Of Applied Science And Technology*, *4*, 9.
3. Arthur, E., & Robert, W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67. https://doi.org/10.1080/00401706.1970.10488634. http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.
4. Breiman, L. (1995). Better subset regression using the nonnegative garrote. In *Technometrics*, pp. 373–384.
5. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. In *Journal Of The Royal Statistical Society. Series B (Methodological)*, pp. 267–288.
6. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal Of The Royal Statistical Society: Series B (Statistical Methodology),67,* 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
7. Munaweera, I. (2018). Shrinkage Estimators under Generalized Garrote and LINEX Loss Functions for Regression Analysis. (University of Manitoba).
8. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal Of The American Statistical Association*, *101*, 1418–1429.
9. Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal Of The Royal Statistical Society: Series B (Statistical Methodology)*, *68*, 49–67. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x.
10. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York: Springer. https://books.google.ca/books?id=tVIjmNS3Ob8C.
11. Zhang, K., Yin, F., & Xiong, S. (2014). Comparisons of penalized least squares methods by simulations. arXiv:1405.1796.
12. Harrison, D., & Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *Journal Of Environmental Economics And Management*, *5*, 81–102.

13. Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal Of The American Statistical Association*, *87*, 738–754. http://www.jstor.org/stable/2290212.
14. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

High-Dimensional Nonlinear Optimization Problem in Semiparametric Regression Model

Mahdi Roozbeh

Abstract By evolving science, knowledge, and technology, new and precise methods for measuring, collecting, and recording information have been innovated, which have been resulted in the appearance and developing of high-dimensional data, in which the number of explanatory variables is much larger than the number of observations. Analysis and modeling the high-dimensional data is one of the most challenging problems faced by the world today. Interpreting such data is not easy and needs to use the modern methods. Penalized methods are one of the most popular ways to analyze the high-dimensional data. Semiparametric models, which a combination of both parametric and nonparametric models, are very flexible models. They are useful when the model contains both parametric and nonparametric elements in the data set. As known, the LASSO approach is a popular technique for variable selection in high-dimensional sparse regression models. Here, we show that the prediction performance of the LASSO method can be improved by eliminating the structured noises. The main purpose of this research is to introduce a modified variable selection or estimation method for a high-dimensional semiparametric regression model through a nonlinear mixed-integer programming technique. Finally, the performance of the proposed method is examined through a real-data analysis about the production of vitamin B2 and some Monte-Carlo simulation studies.

1 Introduction

Let $(y_1, \boldsymbol{x}_1, t_1), \ldots, (y_n, \boldsymbol{x}_n, t_n)$ be observations that follow the semiparametric regression model (SRM)

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + f(t_i) + \epsilon_i, \quad i = 1, ..., n, \tag{1.1}$$

M. Roozbeh (✉)
Department of Statistics, Faculty of Mathematics, Statistics and Computer Sciences, Semnan University, P.O. Box 35195-363 Semnan, Iran
e-mail: mahdi.roozbeh@semnan.ac.ir

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_16

where $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is p-dimensional vector of observed predictors or explanatory variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is a p-dimensional vector of unknown parameters, the t_i's are known and non-random in some bounded domain $D \subset \mathbb{R}$, $f(t_i)$ is an unknown smooth function and ϵ_i's are independent and identically distributed random errors with zero mean and variance σ^2, which are independent of $(\boldsymbol{x}_i, t_i)$. It is remarked that semiparametric regression models are more flexible than standard linear models since they combine both linear and nonlinear components because it is believed that the response variable y linearly depends on x, but nonlinearly related to t via unknown function $f(\cdot)$.

Most of the approaches for the semiparametric regression model are based on the different types of nonparametric estimation procedures. There have been several approaches to estimating $\boldsymbol{\beta}$ and $f(\cdot)$. An extensive study regarding the estimation and application of the model (1.1) can be found in the monograph of [13]. An alternative approach to the nonparametric procedure is differencing methodology. This incoming uses differences to remove the trend in the data that arises from the function $f(\cdot)$ and does not require an estimator of the function $f(\cdot)$ and is often called difference-based procedure. Provided that $f(\cdot)$ is differentiable and the t ordinates are closely spaced, it is possible to remove the effect of the function $f(\cdot)$ by differencing the data appropriately. In model (1.1), [25] concentrated on the estimation of the linear component and used differencing to eliminate bias induced by the presence of the nonparametric component. The difference-based estimation procedure is optimal in the sense that the estimator of the linear component is asymptotically efficient and the estimator of the nonparametric component is asymptotically minimax rate optimal for the semiparametric model ([24]). Thus, differencing allows one to perform inference on $\boldsymbol{\beta}$ as if there were no nonparametric component $f(\cdot)$ in the model (1.1). Once $\boldsymbol{\beta}$ is estimated, a variety of nonparametric techniques could be applied to estimate $f(\cdot)$ as if $\boldsymbol{\beta}$ were known. Reference [24] used higher order differences for optimal efficiency in estimating the linear part by using a special class of difference sequences.

Nowadays, many real-world data problems carry structures in which the number of predictors p may considerably exceed the sample size n, i.e., $p \gg n$. Called high-dimensional problems, for such situations several studies have been pursued addressing the prediction of a new response variable, estimation of an underlying vector parameter, and variables selection approach. Note that classic statistical methods cannot be used for estimating of the model (1.1) when $p > n$, because they would overfit the data, besides severe the identifiability issues. As a rudimentary study made by [23], the least absolute shrinkage and selection operator (LASSO) is a penalization technique for variable selection and estimation in high-dimensional sparse linear models. Another reasonable way out of the ill-posedness of estimation in model (1.1) can be given by assuming a sparse structure in the sense of typically saying that only a few of the components of $\boldsymbol{\beta}$ are nonzero. Based on this fact, estimation of full parametric regression model in the case of $p > n$ and statistical inference afterward has been initiated about a decade ago. Sample attempts include the studies made by [4, 9, 18, 19, 27–29].

In this research, it can be shown that high-dimensional semiparametric regression models can be estimated under some levels of sparsity based on a nonlinear mixed-integer programming approach.

This chapter is organized as follows: Sect. 2 contains a nutshell of the difference-based methodology. The sparse restricted semiparametric regression model and its estimation are considered in Sect. 3. In Sect. 4, we introduce a penalized LASSO method. We deal with the estimation of SRM based on the mixed-integer nonlinear programming approach in Sect. 5. Section 6 is devoted to some Monte-Carlo simulation studies along with an application in riboflavin vitamin B2 production. Finally, we conclude our approach by giving some remarks in Sect. 7.

2 Differencing Approach to Approximate the Model

In this section we use a difference-based technique to estimate the linear regression coefficient vector $\boldsymbol{\beta}$. This technique has been used to remove the nonlinear component in semiparametric regression model by various authors (e.g., [1, 2, 25, 26] and [3]). Consider the following semiparametric regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f}(t) + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, ..., y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^\top$ is an $n \times p$ matrix, $\boldsymbol{f}(t) = (f(t_1), ..., f(t_n))^\top$ and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$.

We assume that in general, $\boldsymbol{\epsilon}$ is a vector of disturbances, with a multivariate normal distribution, $N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{V})$, where $\boldsymbol{V}$ is a symmetric, positive definite known matrix and σ^2 is an unknown parameter.

Reference [25] suggested estimating $\boldsymbol{\beta}$ on the basis of the mth order differencing equation when $\boldsymbol{V} = \boldsymbol{I}_n$ as

$$\sum_{j=0}^{m} d_j y_{i-j} = \Big(\sum_{j=0}^{m} d_j \boldsymbol{x}_{i-j}\Big)\boldsymbol{\beta} + \sum_{j=0}^{m} d_j f(t_{i-j}) + \sum_{j=0}^{m} d_j \epsilon_{i-j}, \tag{2.2}$$

where $d_0, d_1, ..., d_m$ are differencing weights.

How Does the Approximation Work?

Suppose t_i are equally spaced on the unit interval and $f'(\cdot) \leq L$. By the mean value theorem, for some $t_i^* \in [t_{i-1}, t_i]$ we have

$$f(t_i) - f(t_{i-1}) = f'(t_i^*)(t_i - t_{i-1}) \leq \frac{L}{n}.$$

Note that with $m = p = 1$ from (2.2) we have

$$\begin{aligned} y_i - y_{i-1} &= (x_i - x_{i-1})\beta + f(t_i) - f(t_{i-1}) + \epsilon_i - \epsilon_{i-1} \\ &= (x_i - x_{i-1})\beta + O\big(\frac{1}{n}\big) + \epsilon_i - \epsilon_{i-1} \\ &\cong (x_i - x_{i-1})\beta + \epsilon_i - \epsilon_{i-1}. \end{aligned}$$

We then estimate the linear regression coefficient β by the ordinary least-squares estimator based on the differences. Then we obtain the least-squares estimate $\hat{\beta}_{diff} = \frac{\sum_{i=2}^{n}(x_i - x_{i-1})(y_i - y_{i-1})}{\sum_{i=2}^{n}(x_i - x_{i-1})^2}$.

Now let $\boldsymbol{d} = (d_0, ..., d_m)$ be a $(m+1)$-vector, where m is the order of differencing and $d_0, d_1, ..., d_m$ are differencing weights minimizing the variance of linear estimators i.e.,

$$min_{d_0,...,d_m} \sum_{l=1}^{m} \Big(\sum_{j=0}^{m} d_j d_{l+j} \Big)^2,$$

satisfying the conditions

$$\sum_{j=0}^{m} d_j = 0, \quad \sum_{j=0}^{m} d_j^2 = 1. \tag{2.3}$$

The role of constraints (2.3) is now evident. The first condition ensures that, as the t's become close, the nonparametric effect is removed and the second one ensures that the variance of the sum of weighted residuals remains equals to σ^2 in (2.2).

Now, we define the $(n-m) \times n$ differencing matrix $\boldsymbol{D}$ whose elements satisfy (2.3) as

$$\boldsymbol{D} = \begin{pmatrix} d_0 & d_1 & ... & d_m & 0 & 0 & ... & 0 \\ 0 & d_0 & d_1 & ... & d_m & 0 & ... & 0 \\ \vdots & \ddots & & & & & & \vdots \\ 0 & 0 & ... & 0 & d_0 & d_1 & ... & d_m \end{pmatrix}.$$

This and related matrices are given, for example, in [26].

Applying the differencing matrix to model (2.1) permits direct estimation of the parametric effect. As a result of developments in [22] it is known that the parameter vector $\boldsymbol{\beta}$ in (2.1) can be estimated with parametric efficiency. We now show the difference-based estimators that can be used for this purpose. Since the data have been ordered so that the values of the nonparametric variable(s) are close, the application of the differencing matrix $\boldsymbol{D}$ in model (2.1) removes the nonparametric effect in large samples. If $f(\cdot)$ is an unknown function that is the inferential object and has a bounded first derivative, then $\boldsymbol{D}\boldsymbol{f}(t)$ is close to $\boldsymbol{0}$, so that by applying the differencing matrix, we may rewrite (2.1) as

$$\boldsymbol{D}\boldsymbol{y} \doteq \boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\epsilon},$$

or

$$\boldsymbol{y_D} \doteq \boldsymbol{X_D}\boldsymbol{\beta} + \boldsymbol{\epsilon_D}, \tag{2.4}$$

where $\boldsymbol{y_D} = \boldsymbol{D}\boldsymbol{y}$, $\boldsymbol{X_D} = \boldsymbol{D}\boldsymbol{X}$, $\boldsymbol{\epsilon_D} = \boldsymbol{D}\boldsymbol{\epsilon}$.

So, $\boldsymbol{\epsilon_D}$ is a $(n-m)$-vector of disturbances distributed with $E(\boldsymbol{\epsilon_D}) = \boldsymbol{0}$ and $E(\boldsymbol{\epsilon_D}\boldsymbol{\epsilon_D}^\top) = \sigma^2\boldsymbol{V_D}$ where $\boldsymbol{V_D} = \boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^\top \neq \boldsymbol{I}_{n-m}$.

For arbitrary differencing coefficients satisfying (2.3), [25] defines a simple differencing estimator of the parameter $\boldsymbol{\beta}$ in a semiparametric regression model when $\boldsymbol{V} = \boldsymbol{I}_n$ as

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{D}} = (\boldsymbol{X_D}^\top\boldsymbol{X_D})^{-1}\boldsymbol{X_D}^\top\boldsymbol{y_D}. \tag{2.5}$$

Thus, differencing allows one to perform inferences on $\boldsymbol{\beta}$ as if there were no nonparametric component $f(\cdot)$ in the model (2.1) [26]. Once $\boldsymbol{\beta}$ is estimated, a variety of nonparametric techniques could be applied to estimate $f(\cdot)$ as if $\boldsymbol{\beta}$ were known.

We can estimate the linear parameter $\boldsymbol{\beta}$ in (2.4) under the assumption $\text{cov}(\boldsymbol{\epsilon_D}) = \sigma^2\boldsymbol{V_D}$, by minimizing the generalized sum of squared errors

$$SS(\boldsymbol{D}, \boldsymbol{\beta}) = (\boldsymbol{y_D} - \boldsymbol{X_D}\boldsymbol{\beta})^\top\boldsymbol{V_D}^{-1}(\boldsymbol{y_D} - \boldsymbol{X_D}\boldsymbol{\beta}). \tag{2.6}$$

The unique minimizer of (2.6) is the generalized difference-based estimator (GDE) given by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GD}} = \text{argmin}_{\boldsymbol{\beta}}\, SS(\boldsymbol{D}, \boldsymbol{\beta}) = \boldsymbol{C}(\boldsymbol{D})^{-1}\boldsymbol{X_D}^\top\boldsymbol{V_D}^{-1}\boldsymbol{y_D}, \quad \boldsymbol{C}(\boldsymbol{D}) = \boldsymbol{X_D}^\top\boldsymbol{V_D}^{-1}\boldsymbol{X_D}. \tag{2.7}$$

Motivated by [5], we partition the regression parameter $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, where the subvector $\boldsymbol{\beta}_i$ has dimension $p_i, i = 1, 2$ and $p_1 + p_2 = p$. Thus the underlying model has form

$$\boldsymbol{y_D} = \boldsymbol{X_{D1}}\boldsymbol{\beta}_1 + \boldsymbol{X_{D2}}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon_D}, \tag{2.8}$$

where $\boldsymbol{X_D}$ is partitioned according to $(\boldsymbol{X_{D1}}, \boldsymbol{X_{D2}})$ in such a way that $\boldsymbol{X_{Di}}$ is a $n \times p_i$ submatrix, $i = 1, 2$. With respect to this partitioning, the GDEs of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are respectively given by

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\boldsymbol{GD1}} &= \boldsymbol{S}_1(\boldsymbol{D})^{-1}\boldsymbol{X_{D_1}}^\top\boldsymbol{\Sigma}_2(\boldsymbol{D})^{-1}\boldsymbol{y_D}, \quad \boldsymbol{S}_1(\boldsymbol{D}) = \boldsymbol{X_{D_1}}^\top\boldsymbol{\Sigma}_2(\boldsymbol{D})^{-1}\boldsymbol{X_{D1}} \\
\hat{\boldsymbol{\beta}}_{\boldsymbol{GD2}} &= \boldsymbol{S}_2(\boldsymbol{D})^{-1}\boldsymbol{X_{D_2}}^\top\boldsymbol{\Sigma}_1(\boldsymbol{D})^{-1}\boldsymbol{y_D}, \quad \boldsymbol{S}_2(\boldsymbol{D}) = \boldsymbol{X_{D_2}}^\top\boldsymbol{\Sigma}_1(\boldsymbol{D})^{-1}\boldsymbol{X_{D2}}
\end{aligned} \tag{2.9}$$

where

$$\boldsymbol{\Sigma}_i(\boldsymbol{D})^{-1} = \boldsymbol{V_D}^{-1} - \boldsymbol{V_D}^{-1}\boldsymbol{X_{Di}}(\boldsymbol{X_{D_i}}^\top\boldsymbol{V}^{-1}\boldsymbol{X_{Di}})^{-1}\boldsymbol{X_{D_i}}^\top\boldsymbol{V_D}^{-1}, \quad i = 1, 2. \tag{2.10}$$

The sparse model is defined when $\mathcal{H}_o : \boldsymbol{\beta}_2 = 0$ is true. In this paper, we refer the restricted semiparametric regression model (RSRM) to the sparse model.

For the RSRM, the generalized difference-based restricted estimator (GDRE) takes the form of

$$\hat{\boldsymbol{\beta}}_{GDR1}(\boldsymbol{D}) = \boldsymbol{C}_1(\boldsymbol{D})^{-1}\boldsymbol{X}_{\boldsymbol{D}_1}^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}\boldsymbol{y}_{\boldsymbol{D}}, \quad \boldsymbol{C}_1(\boldsymbol{D}) = \boldsymbol{X}_{\boldsymbol{D}_1}^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}\boldsymbol{X}_{\boldsymbol{D}1}. \tag{2.11}$$

According to [21], the GDRE performs better than GDE when model is sparse. However, the former estimator performs poorly as $\boldsymbol{\beta}_2$ deviates from the origin. The following result provides the relation between the submodel and full model estimators of $\boldsymbol{\beta}_1$.

Proposition 2.1 *Under the assumptions in Eqs.* (2.9) *and* (2.11)*, we have*

$$\hat{\boldsymbol{\beta}}_{GD1} = \hat{\boldsymbol{\beta}}_{GDR1} - \boldsymbol{C}_1(\boldsymbol{D})^{-1}\boldsymbol{X}_{\boldsymbol{D}_1}^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}\boldsymbol{X}_{\boldsymbol{D}2}\hat{\boldsymbol{\beta}}_{GD2}.$$

3 Ridge Estimation of Sparse Semiparametric Regression Model

Under situations in which the matrix $\boldsymbol{C}(\boldsymbol{D})$ is ill-conditioned due to linear relationship among the covariates of $\boldsymbol{X}_{\boldsymbol{D}}$ matrix (as in multicollinearity) or the number of independent variables (p) is larger than the sample size (n), the proposed estimators in the previous section are not applicable, because, we always find a linear combination of the columns in $\boldsymbol{X}_{\boldsymbol{D}}$ which is exactly equal to one other column. Mathematically, the design matrix is not full rank, $rank(\boldsymbol{X}_{\boldsymbol{D}}) \leq \min(n, p) < p$ for $p > n$, and one may write $\boldsymbol{X}_{\boldsymbol{D}}\boldsymbol{\beta} = \boldsymbol{X}_{\boldsymbol{D}}(\boldsymbol{\beta} + \boldsymbol{\zeta})$ for every $\boldsymbol{\zeta}$ in the null space of $\boldsymbol{X}_{\boldsymbol{D}}$. Therefore, without further assumptions, it is impossible to infer or estimate $\boldsymbol{\beta}$ from data. We note that this issue is closely related to the classical setting with $p < n$ but with $rank(\boldsymbol{X}_{\boldsymbol{D}}) < p$ (due to linear dependence among covariables) or ill-conditioned design leading to difficulties with respect to identifiability. We note, however, that for prediction or estimation of $\boldsymbol{X}_{\boldsymbol{D}}\boldsymbol{\beta}$ (that is the underlying semiparametric regression surface), identifiability of the parameters is not necessarily needed. From a practical point of view, high empirical correlations among two or a few other covariables lead to unstable results for estimating $\boldsymbol{\beta}$ or for pursuing variable selection. To overcome this problem, we follow [14] and [16] and obtain the restricted ridge estimator by minimizing the sum of squared partial residuals with a spherical restriction and a linear restriction $\boldsymbol{\beta}_2 = 0$, i.e., the RSRM is transformed into an optimal problem with two restrictions:

$$\min_{\boldsymbol{\beta}} \ (\boldsymbol{y}_{\boldsymbol{D}} - \boldsymbol{X}_{\boldsymbol{D}}\boldsymbol{\beta})^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}(\boldsymbol{y}_{\boldsymbol{D}} - \boldsymbol{X}_{\boldsymbol{D}}\boldsymbol{\beta}) \quad \text{subject to} \quad \boldsymbol{\beta}^{\top}\boldsymbol{\beta} \leq \phi^2 \quad \text{and} \quad \boldsymbol{\beta}_2 = \boldsymbol{0}.$$

The resulting estimator is a generalized difference-based restricted ridge estimator (GDRRE), given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GDR1}(k_n) &= \boldsymbol{C}_1^{-1}(\boldsymbol{D}, k_n)\boldsymbol{X}_{\boldsymbol{D}1}^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}\boldsymbol{y}_{\boldsymbol{D}} \\ &= \overbrace{\left(\boldsymbol{I}_{p_1} + k_n(\boldsymbol{X}_{\boldsymbol{D}1}^{\top}\boldsymbol{V}^{-1}\boldsymbol{X}_{\boldsymbol{D}1})^{-1}\right)^{-1}}^{\boldsymbol{T}_1(\boldsymbol{D},k_n)}\hat{\boldsymbol{\beta}}_{GDR1} \\ &= \boldsymbol{T}_1(\boldsymbol{D}, k_n)\hat{\boldsymbol{\beta}}_{GDR1}, \end{aligned} \tag{3.1}$$

where $k_n \geq 0$ is the ridge parameter as a function of sample size n and $\boldsymbol{C}_1(\boldsymbol{D}, k_n) = \boldsymbol{X}_{\boldsymbol{D}1}^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}\boldsymbol{X}_{\boldsymbol{D}1} + k_n\boldsymbol{I}_{p_1}$.

In a similar manner shown previously, the generalized difference-based unrestricted ridge estimators (GDUREs) of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively have forms

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GD1}(k_n) &= \boldsymbol{S}_1^{-1}(\boldsymbol{D}, k_n)\boldsymbol{X}_{\boldsymbol{D}1}^{\top}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{D}, k_n)\boldsymbol{y}_{\boldsymbol{D}} \\ &= \overbrace{\left(\boldsymbol{I}_{p_1} + k_n(\boldsymbol{X}_{\boldsymbol{D}1}^{\top}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{D}, k_n)\boldsymbol{X}_{\boldsymbol{D}1})^{-1}\right)^{-1}}^{\boldsymbol{R}_1(\boldsymbol{D},k_n)}\hat{\boldsymbol{\beta}}_{GD1} \\ &= \boldsymbol{R}_1(\boldsymbol{D}, k_n)\hat{\boldsymbol{\beta}}_{GD1}, \end{aligned} \tag{3.2}$$

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{GD2}(k_n) &= \boldsymbol{S}_2^{-1}(\boldsymbol{D}, k_n)\boldsymbol{X}_{\boldsymbol{D}2}^{\top}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{D}, k_n)\boldsymbol{y}_{\boldsymbol{D}} \\ &= \overbrace{\left(\boldsymbol{I}_{p_2} + k_n(\boldsymbol{X}_{\boldsymbol{D}2}^{\top}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{D}, k_n)\boldsymbol{X}_{\boldsymbol{D}2})^{-1}\right)^{-1}}^{\boldsymbol{R}_2(\boldsymbol{D},k_n)}\hat{\boldsymbol{\beta}}_{GD2} \\ &= \boldsymbol{R}_2(\boldsymbol{D}, k_n)\hat{\boldsymbol{\beta}}_{GD2}, \end{aligned} \tag{3.3}$$

where $\boldsymbol{S}_1(\boldsymbol{D}, k_n) = \boldsymbol{X}_{\boldsymbol{D}1}^{\top}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{D}, k_n)\boldsymbol{X}_{\boldsymbol{D}1} + k_n\boldsymbol{I}_{p_1}$, $\boldsymbol{S}_2(\boldsymbol{D}, k_n) = \boldsymbol{X}_{\boldsymbol{D}2}^{\top}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{D}, k_n)\boldsymbol{X}_{\boldsymbol{D}2} + k_n\boldsymbol{I}_{p_2}$ and

$$\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{D}, k_n) = \boldsymbol{V}_{\boldsymbol{D}}^{-1} - \boldsymbol{V}_{\boldsymbol{D}}^{-1}\boldsymbol{X}_{\boldsymbol{D}i}(\boldsymbol{X}_{\boldsymbol{D}i}^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}\boldsymbol{X}_{\boldsymbol{D}i} + k_n\boldsymbol{I}_{p_i})^{-1}\boldsymbol{X}_{\boldsymbol{D}i}^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}, \quad i = 1, 2. \tag{3.4}$$

Similar to Proposition 2.1, we have the following result without proof.

Proposition 3.1 *The generalized difference-based restricted and unrestricted ridge estimators of* $\boldsymbol{\beta}_1$ *have the following relation*

$$\hat{\boldsymbol{\beta}}_{GD1}(k_n) = \hat{\boldsymbol{\beta}}_{GDR1}(k_n) - \boldsymbol{C}_1^{-1}(\boldsymbol{D}, k_n)\boldsymbol{X}_{\boldsymbol{D}1}^{\top}\boldsymbol{V}_{\boldsymbol{D}}^{-1}\boldsymbol{X}_{\boldsymbol{D}2}\hat{\boldsymbol{\beta}}_{GD2}(k_n).$$

Following [5–7, 21], we use the following test statistic for testing the sparsity hypothesis $\mathcal{H}_o : \boldsymbol{\beta}_2 = 0$

$$\pounds_n(\boldsymbol{D}, k_n) = \frac{\hat{\boldsymbol{\beta}}_{GD2}^{\top}(k_n)\boldsymbol{S}_2(\boldsymbol{D}, k_n)\hat{\boldsymbol{\beta}}_{GD2}(k_n)}{(n - p_1)s^2(\boldsymbol{D}, k_n)}, \tag{3.5}$$

where,

$$s^2(\boldsymbol{D}, k_n) = \frac{1}{n - p_1}\big(\boldsymbol{y_D} - \boldsymbol{X}_{\boldsymbol{D}1}\hat{\boldsymbol{\beta}}_{\boldsymbol{GD}1}(k_n)\big)^\top \boldsymbol{V}_{\boldsymbol{D}}^{-1}\big(\boldsymbol{y_D} - \boldsymbol{X}_{\boldsymbol{D}1}\hat{\boldsymbol{\beta}}_{\boldsymbol{GD}1}(k_n)\big). \quad (3.6)$$

The test statistic $\pounds_n(\boldsymbol{D}, k_n)$ has an asymptotic non-central chi-square distribution with p_2 degrees of freedom and non-centrality parameter $\frac{1}{2}\Delta^*$ (see [5] for more details).

As known, for any particular estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ the risk function under square error loss is measured by

$$\mathrm{R}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = E\big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big].$$

So, the following lemma is now immediate.

Lemma 3.1 ([16]) *The bias, covariance matrix, and risk functions of GDRRE can be evaluated as follows:*

$$\begin{aligned}
\mathrm{bias}\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{GDR}1}(k_n)\big) &= E\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{GDR}1}(k_n) - \boldsymbol{\beta}\big) \\
&= -k_n \boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-1}\boldsymbol{\beta}, \\
\mathrm{Cov}\big(\hat{\boldsymbol{\beta}}_{R1}(k_n)\big) &= \sigma^2 \boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-1}\boldsymbol{C}_1(\boldsymbol{D})\boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-1}, \\
\mathrm{R}\big(\hat{\boldsymbol{\beta}}_{R1}(k_n), \boldsymbol{\beta}\big) &= \sigma^2 tr\big(\boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-1}\boldsymbol{C}_1(\boldsymbol{D})\boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-1}\big) + k_n^2 \boldsymbol{\beta}^\top \boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-2}\boldsymbol{\beta},
\end{aligned}$$

where $\boldsymbol{S}_1 = \boldsymbol{X}_1^\top \boldsymbol{X}_1$.

The spectral decomposition of the (symmetric) positive definite matrix $\boldsymbol{C}_1(\boldsymbol{D})$ can be given by

$$\boldsymbol{C}_1(\boldsymbol{D}) = \boldsymbol{\Gamma}_1 \boldsymbol{\Lambda}_1 \boldsymbol{\Gamma}_1^\top, \ \boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_{p_1}),$$

where the columns of $\boldsymbol{\Gamma}_1$ are eigenvectors of the matrix $\boldsymbol{C}_1(\boldsymbol{D})$ and the scalers λ_1,..., λ_{p_1} are its eigenvalues, satisfying

$$\lambda_1 \geq \ldots \geq \lambda_{p_1} > 0,$$

without loss of generality. Now, we can establish the following result about the risk of GDRRE.

Theorem 3.1 *The risk function of GDRRE can be given by*

$$\mathrm{R}\big(\hat{\boldsymbol{\beta}}_{\boldsymbol{GDR}1}(k_n), \boldsymbol{\beta}\big) = \sigma^2 \sum_{i=1}^{p_1} \frac{\lambda_i}{(\lambda_i + k_n)^2} + k_n^2 \sum_{i=1}^{p_1} \frac{\alpha_i^2}{(\lambda_i + k_n)^2},$$

where α_i *is the* i*th entry of* $\boldsymbol{\alpha}_1 = \boldsymbol{\Gamma}_1^\top \boldsymbol{\beta}_1 = (\alpha_1, \ldots, \alpha_{p_1})^\top$.

Proof. Since $\boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-1} = \boldsymbol{\Gamma}_1(\boldsymbol{\Lambda} + k_n \boldsymbol{I}_{p_1})^{-1}\boldsymbol{\Gamma}_1^\top$, we have

$$\begin{aligned}
\mathrm{tr}\Big(\mathrm{Cov}\big(\hat{\boldsymbol{\beta}}_{GDR1}(k_n)\big)\Big) &= \sigma^2 \mathrm{tr}\left(\boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-1}\boldsymbol{C}_1(\boldsymbol{D})\boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-1}\right) \\
&= \sigma^2 \mathrm{tr}\left(\boldsymbol{\Gamma}_1(\boldsymbol{\Lambda}_1 + k_n \boldsymbol{I}_{p_1})^{-1}\boldsymbol{\Gamma}_1^\top \boldsymbol{\Gamma}_1 \boldsymbol{\Lambda}_1 \boldsymbol{\Gamma}_1^\top \boldsymbol{\Gamma}_1 (\boldsymbol{\Lambda}_1 + k_n \boldsymbol{I}_{p_1})^{-1}\boldsymbol{\Gamma}_1^\top\right) \\
&= \sigma^2 \mathrm{tr}\left(\boldsymbol{\Lambda}_1(\boldsymbol{\Lambda}_1 + k_n \boldsymbol{I}_{p_1})^{-2}\right) \\
&= \sigma^2 \sum_{i=1}^{p_1} \frac{\lambda_i}{(\lambda_i + k_n)^2}. \qquad (3.7)
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{QB}\big(\hat{\boldsymbol{\beta}}_{GDR1}(k_n)\big) &= k_n^2 \boldsymbol{\beta}_1^\top \boldsymbol{C}_1(\boldsymbol{D}, k_n)^{-2}\boldsymbol{\beta} \\
&= k_n^2 \boldsymbol{\alpha}_1^\top \boldsymbol{\Gamma}_1^\top \boldsymbol{\Gamma}_1 (\boldsymbol{\Lambda}_1 + k_n \boldsymbol{I}_{p_1})^{-1}\boldsymbol{\Gamma}_1^\top \boldsymbol{\Gamma}_1 (\boldsymbol{\Lambda} + k_n \boldsymbol{I}_{p_1})^{-1}\boldsymbol{\Gamma}_1^\top \boldsymbol{\Gamma}_1 \boldsymbol{\alpha}_1 \\
&= k_n^2 \boldsymbol{\alpha}_1^\top (\boldsymbol{\Lambda}_1 + k_n \boldsymbol{I}_{p_1})^{-1}(\boldsymbol{\Lambda}_1 + k_n \boldsymbol{I}_{p_1})^{-1}\boldsymbol{\alpha}_1 \\
&= \boldsymbol{\alpha}_1^\top \mathrm{diag}\Big((\lambda_1 + k_n)^{-2}, \ldots, (\lambda_{p_1} + k_n)^{-2}\Big)\,\boldsymbol{\alpha}_1 \\
&= k_n^2 \sum_{i=1}^{p_1} \frac{\alpha_i^2}{(\lambda_i + k_n)^2}, \qquad (3.8)
\end{aligned}$$

where QB(.) is the quadratic bias of an estimator. By adding (3.7) and (3.8), the proof is complete. □

4 Least Absolute Shrinkage and Selection Operator Approach

The amount of data we are faced with keeps growing. From around the late 1990s, wide data sets emerged, in which the number of variables far exceeds the number of observations. This was mainly due to our increasing ability to measure a large amount of information automatically (see [10] for more details).

Penalized regression can perform variable selection and prediction in a 'Big Data' environment more effectively and efficiently in contrast to the other methods. Initially proposed by [23], the LASSO (least absolute shrinkage and selection operator) is based on minimizing mean squared error, which is based on balancing the opposing factors of bias and variance to build the most predictive model. In fact, LASSO shrinks the regression coefficients toward zero by penalizing the regression model with a ℓ_1-norm penalty term, i.e., sum of the absolute value of the coefficients. LASSO regression is a simple technique to reduce model complexity and usually prevent over-fitting which may result from simple linear regression.

In the case of LASSO regression, the penalty term is embedded to force the coefficient estimates with minor contributions to the model to be exactly set to zero.

This means that LASSO can be also seen as an alternative to the subset selection methods for performing variable selection in order to reduce the complexity of the model.

LASSO is an extension of the ordinary least squares (OLS) regression which adds a penalty to the residual sum of squares (RSS), being equal to the sum of the absolute values of the non-intercept beta coefficients multiplied by the parameter λ that slows (when $\lambda < 1$) or accelerates (when $\lambda > 1$) the penalty. Therefore, the following optimization problem should be solved based on LASSO problem:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{D}, \lambda_n) = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y_D} - \boldsymbol{X_D}\boldsymbol{\beta})^\top (\boldsymbol{y_D} - \boldsymbol{X_D}\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} |\beta_j| \right\},$$

Figure 1 shows the constraint area of the LASSO method for $p = 2$, in which elliptical contours of the function are shown by the full. They are centered at the OLSE. The constraint region is the rotated square. LASSO solution is the first place that the contours touch the square, and this will sometimes occur at a corner, corresponding to a zero coefficient. LASSO is frequently used in practice since the ℓ_1 penalty allows us to shrink some coefficients to zero, that is, to produce sparse estimation models that are highly interpretable.

It is notable that increasing λ_n will increase bias and decrease variance. Likewise, decreasing λ_n reduces bias and increases variance. A big part of the building, the best models in LASSO deal with the bias-variance trade-off. Bias refers to how correct (or incorrect) the model is. A very simple model that makes a lot of mistakes is said to have a high bias. A very complicated model that does well on its training data is said to have a low bias. There are several ways to choose the optimal λ_n, such as AIC,

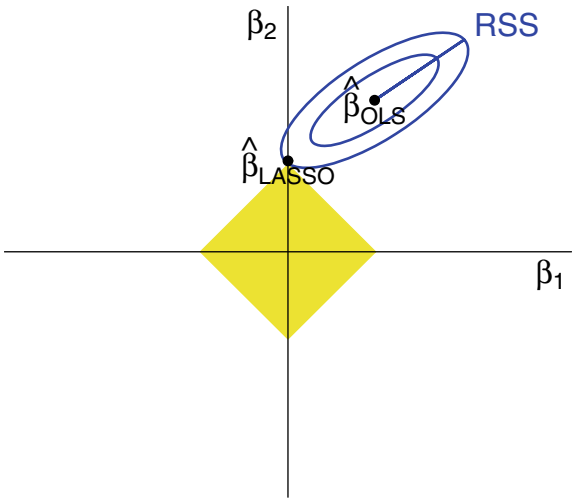


Fig. 1 Constraint region of LASSO

BIC, C_p. For this purpose, one of the most popular methods is the cross-validation (CV) method.

5 A Mathematical Heuristic Algorithm for Estimation of High-Dimensional SRM

Here, we suggest a mixed-integer programming model based on the research work of [8] and [20] for finding the effective explanatory variables to be used in a high-dimensional SRM.

To proceed, at first it is notable that we need to discover the p_1 fittest explanatory variables among all of the p explanatory variables of the data set ($p_1 < n$). Assume that $\widetilde{\boldsymbol{x}}_j$, $j = 1, 2, ..., p$, is the jth column of the design matrix $\boldsymbol{X_D}$. For each column $\widetilde{\boldsymbol{x}}_j$ of $\boldsymbol{X_D}$ we define a binary (indicator) variable z_j which is equal to one when $\widetilde{\boldsymbol{x}}_j$ is used in the regression and is equal to zero, otherwise. Now, we need to define the following essential constraint in our model:

$$\sum_{j=1}^{p} z_j = p_1,\ z_j \in \{0, 1\},\ j = 1, 2, ..., p.$$

On the other hand, if $z_j = 0$, then $\widetilde{\boldsymbol{x}}_j$ is not considered in the regression and so, we should set $\boldsymbol{\beta}_j = 0$. This fact can be guaranteed by the following set of inequalities:

$$-Mz_j \le \boldsymbol{\beta}_j \le Mz_j,\ j = 1, 2, ..., p,$$

where M is an enough large positive constant, being an upper bound for the set $\{|\boldsymbol{\beta}_j|\}_{j=1}^{p}$. Based on these preliminaries, now we are in a position to propose the initial form of our optimization model as follows:

$$\begin{aligned} &\min_{\boldsymbol{\beta}}\ (\boldsymbol{y_D} - \boldsymbol{X_D}\boldsymbol{\beta})^{\top}(\boldsymbol{y_D} - \boldsymbol{X_D}\boldsymbol{\beta}),\\ &s.t.\ \sum_{j=1}^{p} z_j = p_1,\\ &\quad -Mz_j \le \boldsymbol{\beta}_j \le Mz_j, \qquad j = 1, 2, ..., p,\\ &\quad z_j \in \{0, 1\}, \qquad j = 1, 2, ..., p. \end{aligned} \tag{5.1}$$

It is worth noting that although the problem (5.1) is an *NP*-hard problem, it can be effectively solved by the metaheuristic algorithms to achieve an approximate solution (see [8]. Moreover, nowadays there exist powerful solvers such as CPLEX which are able to handle such problems even in large-scale cases. Making reasonable simplifications in the models may considerably enhance the efficiency of the solvers in the solution process. Based on this fact, we simplify the objective function of the

model (5.1) by transferring some complicated terms to the constraints. That is, we suggest the following modified version of (5.1):

$$\begin{aligned} &\min_{\boldsymbol{\Psi}} \ (\boldsymbol{y_D} - \boldsymbol{\Psi})^\top (\boldsymbol{y_D} - \boldsymbol{\Psi}), \\ &s.t. \ \sum_{j=1}^{p} z_j = p_1, \\ &\qquad -Mz_j \le \boldsymbol{\beta}_j \le Mz_j, \quad j = 1, 2, ..., p, \\ &\qquad \boldsymbol{\Psi} = \boldsymbol{X_D}\boldsymbol{\beta}, \\ &\qquad z_j \in \{0, 1\}, \qquad\qquad j = 1, 2, ..., p, \end{aligned} \tag{5.2}$$

with a simpler structure of the objective function Hessian in contrast to the model (5.1). Solution of the optimization problem (5.2) is called the generalized difference-based nonlinear mixed-integer programming estimator (GDNMPE).

6 Numerical studies

In order to test the performance of our suggested approach, here we analyze a real world and some simulated data sets in the SRM for the high-dimensional cases. We examine the efficiency of the given method in the real-world examples concerning the riboflavin (vitamin B2) production data set as well as some simulated data sets.

Application to Riboflavin Production Data Set

In this section, we test the proposed estimation method on the riboflavin (vitamin B2) production data set in Bacillus subtilis ([9]), being a classical real example in high-dimensional SRM which can be found in R package "hdi." The single real-valued response variable is the logarithm of the riboflavin production rate. The effect of $p = 4088$ predictor variables which are the logarithm of the expression levels of 4088 genes on the response variable would be modeled using the linear regression. There is one rather homogeneous data set from $n = 71$ samples that were hybridized repeatedly during a fed-batch fermentation process where different engineered strains and strains growth were analyzed under different fermentation conditions. Based on 10-fold cross-validation, the LASSO shrinks 4022 parameters to zero and remains $p_1 = 66$ significant explanatory variables.

The ridge trace plot, a plot of estimated coefficients against a shrinkage parameter which is a common graphical adjunct to help determine a favorable trade-off of the bias against the precision (inverse variance) of the estimates, is shown in Fig. 2. We

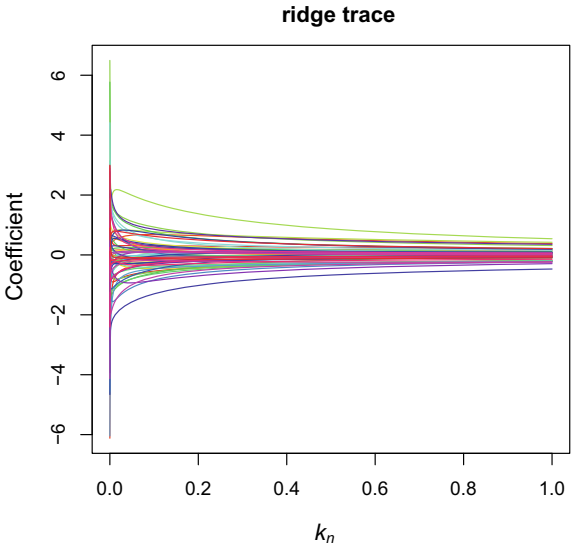


Fig. 2 The diagram of ridge trace for the riboflavin data set

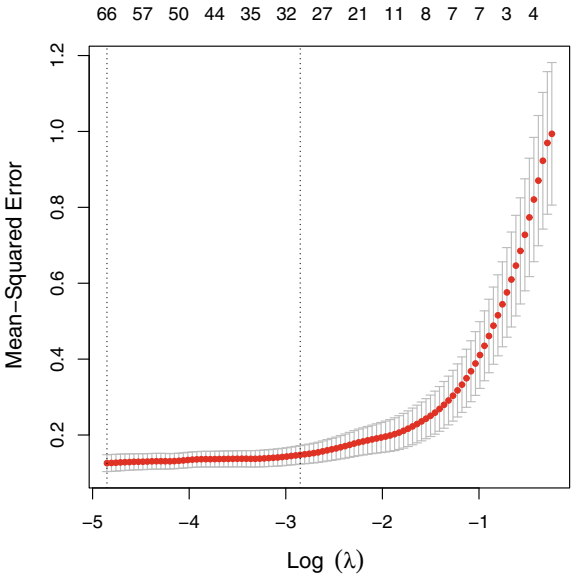


Fig. 3 The diagram of cross-validation curve (the red dotted line), and upper and lower standard deviation curves along the sequence $\{\lambda_n\}$ for the riboflavin data set

plotted Fig. 3 to find the best value of the LASSO parameter (λ_n) which minimizes the cross-validation criterion. As seen in Fig. 3, the minimal mean squared error is achieved at $\lambda_n = 0.0078$.

The parametric part of the considered Model with 65 effective explanatory variables, i.e., $\boldsymbol{\beta}_1$, is estimated by a first-order differencing coefficients, $d_0 = 0.7071$ and $d_1 = -0.7071$ in which $m = 1$, and then, the nonparametric part is estimated

by kernel methodology and cross-validation criteria. Optimal differencing weights do not have analytic expressions but may be calculated easily using an optimization routine. Reference [12] presented weights to order $m = 10$. These contain some minor errors.

To detect the nonparametric part of the model, we calculate

$$s_i^2(\boldsymbol{D}) = \frac{1}{n - p_1}\big(\boldsymbol{y_D} - \boldsymbol{X_{D1}}[,-i]\hat{\boldsymbol{\beta}}_{GD1}\big)^\top \boldsymbol{V_D^{-1}}\big(\boldsymbol{y_D} - \boldsymbol{X_{D1}}[,-i]\hat{\boldsymbol{\beta}}_{GD1}\big),\ i = 1, ..., 66,$$

where $\boldsymbol{X_{D1}}[,-i]$ is obtained by deleting the ith column of matrix $\boldsymbol{X}_1$. Among all 66 effective genes, "$YYBG\_at$" had minimum $s_i^2(\boldsymbol{D})$ value and so this can be considered as a nonparametric part. We also use the added-variable plots to identify the parametric and nonparametric components of the model. Added-variable plots enable us to visually assess the effect of each predictor, having adjusted for the effects of the other predictors. By looking at added-variable plot (Fig. 4), we consider "$YYBG\_at$" as a nonparametric part. As it can be seen from this figure, the nonlinear relation between "$YYBG\_at$" and the response variable seems to have a better fit than linear relation after removing the effects of other predictors, and so, the specification of the sparse semiparametric regression model can be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{f}(t) + \boldsymbol{\epsilon}, \quad t = YYBG\_at \tag{6.1}$$

where $p_1 = 65$ and $p_2 = 4022$.

For estimating the nonparametric part of the model, $f(t)$, we use the Gaussian kernel method (see [17] for more details).

To use the approach (5.2), we set $M = 1$. Also, the method has been applied by CPLEX version 12.6.1. In Table 1, we present the estimates of the coefficients of the effective parametric genes in SRM based on the proposed estimators. We numerically calculated RSS $= (\boldsymbol{y_D} - \boldsymbol{X_D}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y_D} - \boldsymbol{X_D}\hat{\boldsymbol{\beta}})$ for GDURE, GDRRE and GDNMPE. It is a measure of the error or goodness of the prediction. As seen, the mixed-integer programming model fits the data better than the others. Finally, we estimated the nonparametric effect ($f(YYBG\_at)$) after estimating the linear part by GDURE, GDRRE, and GDNMPE using kernel regression method in Figure 5, i.e., we used kernel fit to regress $\boldsymbol{Z} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ on $YYBG\_at$.

Some Simulation Studies

In this subsection, based on some simulation studies, we examine the performance of the three different proposed methods. We also evaluate the efficiency of the proposed estimator in the high-dimensional cases.

To achieve different degrees of collinearity, following [15] and [11] the predictor variables were generated for $n = 50$ and $p = \{150, 180\}$ (high-dimensional) from the following model:

Table 1 Estimation of the coefficients of the effective genes based on the proposed methods for the riboflavin data set

Method	GDURE	GDRRE		GDNMPE
Label of the gene	Coefficient estimation	Coefficient estimation	Label of the gene	Coefficient estimation
ALST_at	-1.217731e-04	-0.544607	*ACCA_at*	0.727057
AMYX_at	9.259859e-05	0.919408	*BIRA_s_at*	2.738694
ARAA_at	−2.423074e-05	−2.251603	*ARGH_at*	−6.137601
ARAM_at	1.038319e-05	0.154134	*FLIF_at*	−0.147270
ARGF_at	9.567998e-06	−0.231887	*YQAG_at*	−0.009113
BIOB_at	2.006911e-05	0.275029	*GGAA_at*	1.310371
COTJC_at	-8.900675e-05	0.097685	*GUAB_at*	1.261439
DEGA_at	-7.891485e-05	0.284943	*LCTE_at*	−1.044405
LCTP_at	-2.722333e-05	−0.461230	*LICA_at*	−1.629179
LEVF_at	−9.970859e-05	−0.192662	*NTH_at*	-2.594287
LYSC_at	−2.017575e-04	−0.175473	*OPPC_at*	−1.381273
METK_at	5.225437e-05	−0.112374	*PHRK_at*	0.997843
PURR_at	9.816026e-06	1.145270	*PURF_at*	2.101418
SACB_at	−1.660888e-05	−0.198362	*RAPK_at*	−1.624683
sigM_at	2.683243e-05	0.680754	*RPLJ_at*	0.459227
SIPU_at	4.434275e-05	1.133496	*RPLR_at*	−1.015745
SPOIIAA_at	1.663782e-04	2.417561	*SPOVAC_at*	6.547311
SPOIIIAB_at	6.611147e-05	-0.469733	*YAAQ_at*	0.236426
SPOVAA_at	7.485906e-05	0.052439	*YCGO_at*	0.341390
SSPF_at	−7.522401e-05	0.942005	$GAP136 - F\_at$	2.266223
XLYA_at	−7.588471e-06	−0.317887	*YDDD_at*	−0.183998
YCGO_at	-7.146751e-05	−0.066030	*YFKS_at*	−2.356648
YCLC_at	−4.752549e-05	−1.841358	*YJBV_at*	−3.275892
YDAO_at	7.311426e-05	2.493456	*YFHE_r_at*	5.763074
YDDH_at	1.498316e-04	0.329809	*YKAA_at*	0.997701
YDDK_at	−8.059497e-05	−1.389042	*YLMC_at*	−1.977877
YEBC_at	1.710338e-04	-0.949894	*YMFB_at*	−0.959574
YESV_at	1.118750e-04	0.204742	*YOCH_at*	3.111592
YETH_at	1.128877e-04	−0.377397	*YOMM_at*	1.795015
YFNA_at	7.938866e-05	-0.065452	*YPUF_at*	−0.265969
YHDS_r_at	−1.199883e-05	−0.452141	*YQBT_at*	−1.365757
YKBA_at	2.273325e-06	1.872915	*YQID_at*	2.578543
YKNV_at	-2.156316e-04	−0.932007	*YRKL_at*	−4.557744
YLXH_at	3.120245e-05	2.204903	*YSMA_at*	4.428141
YMAH_i_at	6.830865e-05	0.840069	*YTIA_at*	2.820209

(continued)

Table 1 (continued)

Method	GDURE	GDRRE		GDNMPE
Label of the gene	Coefficient estimation	Coefficient estimation	Label of the gene	Coefficient estimation
YNEF_at	3.127307e-05	−1.468254	*YURM_at*	−3.312196
YOAB_at	2.563135e-05	−2.816168	*YURX_at*	−5.979373
YOAC_at	7.150370e-05	−1.254557	*YUSF_at*	0.745406
YOBR_at	1.002874e-04	0.741075	*YUXO_at*	−3.287791
YOMT_at	4.739930e-05	−0.858392	*YVAV_at*	1.191750
YPGA_at	3.997361e-06	0.573591	*YVSG_at*	1.353822
YQAE_at	−9.920698e-05	−0.135255	*YWGB_at*	−3.995705
YQJS_at	1.296992e-04	0.486399	*YXIE_at*	2.759543
YQJT_at	−5.189563e-05	−0.206190	*YXLC_at*	−0.336443
YQJU_at	−2.620583e-05	−1.199482	*YXLE_at*	−4.824330
YRVJ_at	−1.465138e-04	0.260066	*YCNF_at*	2.096664
YTGB_at	7.351869e-06	0.723246	*YCZF_at*	0.884965
YTIP_at	−2.278199e-05	−2.044283	*YDGK_at*	−0.828611
YUID_at	5.021464e-05	0.070762	*YEEC_at*	−0.119783
YULC_at	−2.068166e-04	1.784028	*ACOA_at*	−2.877613
YURQ_at	−9.043668e-05	−0.964982	*YKBA_at*	−2.137831
YUSJ_at	3.517705e-05	−0.125736	*YKNW_at*	1.305234
YUSL_at	4.404325e-05	−0.432172	*YKOU_at*	−0.103588
YVFM_at	4.438915e-05	0.385502	*YKUL_at*	2.830268
YVFO_at	−5.643278e-05	−0.552383	*YLOH_at*	−4.062187
YVOA_at	−2.689035e-05	−1.884981	*YOJL_at*	−1.293150
YVRG_at	−1.404958e-06	−0.162089	*YOMT_at*	1.942282
YWRO_at	−2.906734e-05	2.064755	*YONV_at*	2.903451
YXEH_at	4.383370e-05	1.285049	*YPUG_at*	0.879176
YXIB_at	1.732469e-06	−0.691739	*YQKE_at*	0.291981
YXKI_at	−8.541844e-05	-0.403996	*YRBA_at*	−2.288893
YXLD_at	−-9.640283e-06	−0.413108	*YWPF_at*	−0.208449
YXLE_at	4.286699e-05	−0.103899	*YXIT_at*	0.118111
YYBI_at	−1.379380e-05	0.584797	*YXLG_at*	0.884451
YYCP_at	−1.072422e-04	0.571022	*YYZE_at*	3.072341
RSS	69.41865	2.019622	RSS	0.367467

$$x_{ij} = (1-\gamma^2)^{\frac{1}{2}} z_{ij} + \gamma z_{ip}, \qquad i = 1, 2, ..., n, \quad j = 1, 2, \ldots, p, \tag{6.2}$$

where z_{ij}'s are independent standard normal pseudo-random numbers and γ is specified so that the correlation between any two explanatory variables is given by $\gamma^2 = 0.75$. These variables are then standardized so that $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$ are in correlation forms. The observations for the dependent variable are determined by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f}(t) + \boldsymbol{\epsilon}, \tag{6.3}$$

where

$$\boldsymbol{\beta} = \left(\underbrace{0, \ldots, 0}_{20}, \underbrace{2, \ldots, 2}_{20}, \underbrace{0, \ldots, 0}_{20}, \underbrace{-2, \ldots, 2}_{20}, \underbrace{0, \ldots, 0}_{p-80} \right)^{\top}$$

and

$$f(t) = exp\big(sin(t) * cos(3 * t)\big) * exp(sqrtt), \ t \in [0, 3].$$

Let again $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^{\top}$, where $\boldsymbol{\beta}_1$ denotes the coefficients of effective variables of $\boldsymbol{\beta}$ while $\boldsymbol{\beta}_2$ denotes the coefficients of non-effective variables. Also, we considered $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n), \ \sigma^2 = 0.64$.

We use a first-order differencing coefficients to estimate the linear parameters of model (6.3). We also use the following criteria and ([29]) to assess the numerical performance:

- MME presents the median of the model error (ME) measure of an estimator, where $\text{ME} = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\top}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$;
- C shows the average number of zero coefficients correctly estimated to be zero;

Table 2 Properties of the proposed estimators for the simulated data sets

	p	150	180
GDURE	MME	161.79214	167.50962
	C	1.00000	0.66667
	IC	0.00000	0.00000
	U–fit	0.00000	0.00000
	C–fit	0.30978	0.29609
	O–fit	0.69022	0.70391
GDRRE	MME	113.41205	123.10328
	C	16.00000	29.33333
	IC	0.03031	0.02857
	U–fit	0.06383	0.07826
	C–fit	0.26894	0.28406
	O–fit	0.66723	0.63768
GDNMPE	MME	157.99010	154.36350
	C	105.33333	138.00000
	IC	38.00000	37.33333
	U–fit	0.55556	0.53617
	C–fit	0.41975	0.44893
	O–fit	0.02469	0.01490

- IC shows the average number of nonzero coefficients incorrectly estimated to be zero;
- U–fit (Under fit) shows the proportion of excluding any significant variables;
- C–fit (Correct fit) presents the proportion of selecting the exact subset model;
- O–fit (Over fit) shows the proportion of including all three significant variables and some noise variables.

Table 2 displays a summary of the results obtained from the simulation studies. According to this table, the number of explanatory variables and degree of collinearity affect the performances of GDURE and GDRRE in the sense of MME criterion, i.e., as they increase the MME of GDURE and GDRRE often increases, whereas this issue is not observed for GDNMPE. GDRRE may only reduce ME criterion and does not reduce model complexity (MC) criteria since it does not shrink the coefficients of non-effective explanatory variables to zero. GDNMPE has an advantage compared to the GDRRE since it shrinks the coefficients of non-effective explanatory variables to zero even if it has a larger ME compared to the GDRRE. In other words, GDNMPE performs better than the other estimators in the sense of MC criteria.

7 Summary and Conclusions

In this chapter, under sparsity assumption on some elements of $\boldsymbol{\beta}$, we proposed a penalized estimator based on nonlinear mixed-integer programming in a semiparametric regression model. Finally, a real-data example and some simulated data sets

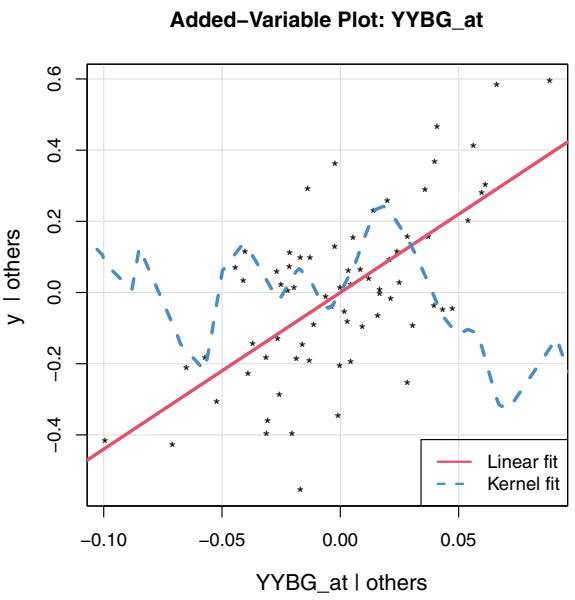


Fig. 4 Added-variable plot of explanatory variables $YYBG\_at$ vs. dependent variable, linear fit (red solid line) and kernel fit (blue dashed line)

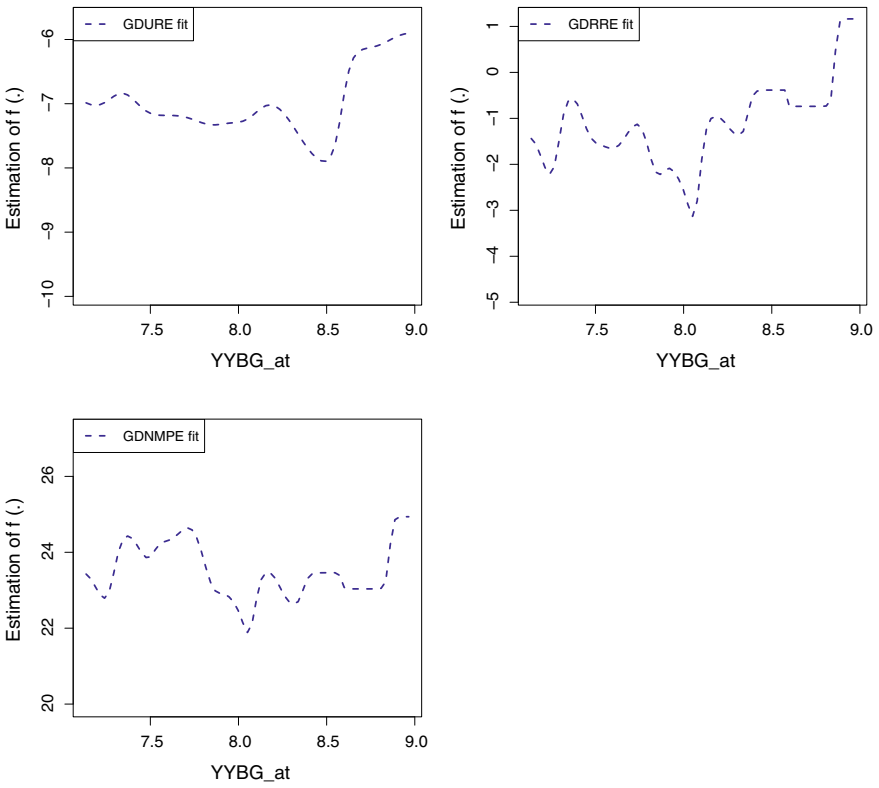


Fig. 5 Fitted curves of nonparametric part of model (6.1) by proposed methods for the riboflavin data set

were analyzed to evaluate the performance of the proposed estimator numerically. In the real example study, as it can be seen from Fig. 4, the nonlinear relation between the dependent variable and YYBG_at can be detected and so, the pure parametric model does not fit to the data and semiparametric regression model fits more significantly. Further, from Table 1 and Fig. 5, it can be deduced that GDNMPE is efficient in the sense of RSS criterion. In the simulated data sets, from Table 2, it can be deduced that GDNMPE is quite efficient in the sense of MC criteria. Moreover, because of the sparsity of the data, GDURE was the worst estimator for the parametric part in this examples.

Acknowledgements The author thank the anonymous Reviewers and the Associate Editor for their valuable comments and suggestions that helped to improve the quality of this work.

References

1. Akdeniz, F., & Roozbeh, M. (2017). Efficiency of the generalized difference-based weighted mixed almost unbiased two parameter estimator in Partially Linear Model. *Communications in Statistics - Theory & Methods, 46*, 12259–12280.
2. Akdeniz, F., & Roozbeh, M. (2019). Generalized difference-based weighted mixed almost unbiased ridge estimator in partially linear models. *Statistical Papers, 60*, 1717–1739.
3. Akdeniz Duran, E., Akdeniz, F., & Roozbeh, M. (2018). A new difference-based weighted mixed Liu estimator in partially linear models. *Statistics: A Journal of Theoretical and Applied Statistics, 52*, 1309–1327.
4. Amini, M., & Roozbeh, M. (2019). Improving the prediction performance of the LASSO by subtracting the additive structural noises. *Computational Statistics, 34*, 415–432.
5. Arashi, M., & Roozbeh, M. (2019). Some improved estimation strategies in high-dimensional semiparametric regression models with application to riboflavin production data. *Statistical Papers, 60*, 317–336.
6. Arashi, M. (2012). Preliminary test and Stein estimators in simultaneous linear equations. *Linear Algebra and its Applications, 436*(5), 1195–1211.
7. Arashi, M., & Tabatabaey, S. M. M. (2009). Improved variance estimation under sub-space restriction. *Journal of Multivariate Analysis, 100*, 1752–1760.
8. Bertsimas, D., & Tsitsiklis, J. N. (1997). *Introduction to linear optimization*. Massachusetts: Athena Scientific.
9. Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and Its Application, 1*, 255–278.
10. Efron, B., & Hastie, T. (2017). *Computer age statistical inference*. Cambridge: Cambridge University Press.
11. Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association, 76*, 131–139.
12. Hall, P., Kay, J. W., & Titterington, D. M. (1990). On estimation of noise variance in two-dimensional signal processing. *Advances in Applied Probability, 23*, 476–495.
13. Härdle, W., Liang, H., & Gao, J. (2000). *Partially linear models*. Heidelberg: Physika Verlag.
14. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics, 12,* 55–67.
15. McDonald, G. C., & Galarneau, D. I. (1975). A monte carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association, 70*, 407–416.
16. Roozbeh, M. (2015). Shrinkage ridge estimators in semiparametric regression models. *Journal of Multivariate Analysis, 136*, 56–74.
17. Roozbeh, M. (2018). Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion. *Computational Statistics & Data Analysis, 117*, 45–61.
18. Roozbeh, M., Maanavi, M., & Babaie-Kafaki, S. (2020). Robust high-dimensional semiparametric regression using optimized differencing method applied to vitamin B2 production data. *Iranian Journal of Health Sciences, 8*, 9–22.
19. Roozbeh, M., Babaie-Kafaki, S., & Aminifard, Z. (2021). A nonlinear mixed-integer programming approach for variable selection in linear regression model. *Communications in Statistics - Simulation & Computation*. https://doi.org/10.1080/03610918.2021.1990323
20. Roozbeh, M., Babaie-Kafaki, S., & Maanavi, M. (2021). A heuristic algorithm to combat outliers and multicollinearity in regression model analysis. *Iranian Journal of Numerical Analysis and Optimization*. https://doi.org/10.22067/IJNAO.2021.68160.1008.
21. Saleh, A. KMd. E. (2006). *Theory of preliminary test and stein-type estimation with applications*. New York: Wiley.
22. Speckman, P. (1988). Kernel somoothing in partial linear models. *Journal of the Royal Statistical Society, Series B, 50,* 413–436.
23. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

24. Wang, L., Brown, L. D., & Cai, T. T. (2011). A difference based approach to semiparametric partial linear model. *Electronic Journal of Statistics, 5*, 619–641.
25. Yatchew, A. (1997). An elemantary estimator of the partial linear model. *Economics Letters, 57*, 135–143.
26. Yatchew, A. (2003). *Semiparametric regression for the applied econometrican*. Cambridge: Cambridge University Press.
27. Yüzbaşi, B., & Arashi, M. (2019). Double shrunken selection operator. *Communications in Statistics - Simulation & Computation, 48*, 666–674.
28. Yüzbaşi, B., Arashi, M., & Ahmed, S. E. (2020). Shrinkage estimation strategies in generalized ridge regression models under low/high-dimension regime. *International Statistical Review, 88*, 229–251.
29. Yüzbaşi, B., Arashi, M., & Akdeniz, F. (2021). Penalized regression via the restricted bridge estimator. *Soft Computing, 25*, 8401–8416.

Frontiers in Robust Analysis and Mixture Modelling

Parsimonious Finite Mixtures of Matrix-Variate Regressions

Antonio Punzo and Salvatore D. Tomarchio

Abstract Over the years, there has been an increased interest in the analysis of matrix-variate data. In the model-based clustering literature, finite mixtures of matrix-variate regressions have been recently introduced. However, a serious concern about this model is the excessive number of parameters associated with the two covariance matrices, related to the responses, for each mixture component. To attain parsimony, the well-known eigen-decomposition is applied to the covariance matrices, yielding a family of 98 different parsimonious mixture models. Parameter estimation, under the maximum likelihood paradigm, is carried out via an expectation-conditional maximization (ECM) algorithm. Our family of models is applied to real data with the aim to assess their clustering performance and for analyzing their behavior with respect to other parsimonious mixture models.

1 Introduction

The importance of finite mixture models in statistical data analyses is underlined by the high volume of articles about mixture applications present in the statistical and general scientific literature. Because of their flexibility, mixture models are a convenient statistical tool for modeling a wide range of phenomena characterized by unobserved heterogeneity, and constitute a powerful device for clustering and classification (for more details, see, e.g., [1]). In many analyses, mixture models are used for modeling one or more variables, and they are referred to as unconditional finite mixture models. However, when covariates are available, useful clustering insights can be gained by accounting for functional dependencies of the responses on the covariates [2]. To this purpose, finite mixtures of regressions (FMR) have been proposed in the literature (see, e.g., [3, 4]). Finite mixtures of regressions with concomitant

A. Punzo (✉) · S. D. Tomarchio
Dipartimento di Economia e Impresa, Università degli Studi di Catania, Catania, Italy
e-mail: antonio.punzo@unict.it

S. D. Tomarchio
e-mail: daniele.tomarchio@unict.it

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_17

variables ([5]) are an extension of the FMR obtained by allowing the mixing weights to depend on some concomitant variables (which often coincide with the covariates) via a multinomial logistic regression. This model is also known as the mixture of experts (MoE). Both types of models have been extensively investigated in the univariate and multivariate model-based clustering literature (for recent contributions see, e.g., [2, 6–9]). However, in the last decade there has been an increased interest in the application of finite mixture models to matrix-variate (or three-way) data. This data structure can occur in several and different application domains, such as multivariate longitudinal data, multivariate spatio-temporal data, multivariate repeated measures, multivariate spatial data, and multivariate time-series [10]. In all these cases we observe p variables measured in r different situations on N observations, so that the data can be arranged in a three-way array characterized by the following three dimensions: variables (rows), situations (columns) and observations (layers). In other terms, we have a $p \times r$ matrix for each statistical observation.

The main contributions in this field have focused on unconditional finite mixture models (see, e.g., [10–17]). The first attempt of extending the FMR to the matrix-variate framework is due to [18], who consider mixtures of matrix regression time series for modeling crime data. Another matrix-variate regression-based proposal has been recently introduced by [19]. These models, in addition to make possible an assessment of the functional relationship among a set of $p \times r$ responses and a set of $q \times r$ covariates, can implicitly introduce parsimony in the mean matrices if q is smaller than r. Nevertheless, the potential overparameterization issue caused by the matrix-variate data structure remains a crucial problem, especially when the sample size is small, the dimension of the matrices is high or both aspects occur. Thus, in this work we extend this branch of literature by proposing parsimonious finite mixtures of matrix-variate regressions (FMMVR). Specifically, we introduce parsimony via the eigen-decomposition of the component covariance matrices, in the fashion of [14, 20, 21], producing a family of 98 parsimonious FMMVR (for another approach used in the matrix-variate framework to attain parsimony in the component covariance matrices, see [22]). We present our family of parsimonious models in Sect. 2, along with an expectation-conditional maximization (ECM) algorithm for parameter estimation. In Sect. 3, we first assess model selection using the Bayesian information criterion (BIC; [23]) and the integrated completed likelihood (ICL; [24]) on simulated data. The classification performance of the best fitting models according to BIC and ICL is also evaluated. Then, we apply our parsimonious models to a real dataset concerning students careers' indicators measured for families of degree courses among the non-telematics Italian universities. The computational times required to fit our parsimonious FMMVR are reported. Our parsimonious FMMVRs are also compared to other existing parsimonious competitors. Firstly, we consider an unconditional baseline clustering method for the responses, based on parsimonious matrix-variate normal mixtures (MVNM) [14]. Furthermore, on the pr-dimensional vectorized data, we fit the four families of parsimonious multivariate models discussed in [25], which include

- multivariate normal mixtures on the responses (MNM);
- multivariate FMR (MFMR);
- multivariate gating network MoE (MgMoE);
- multivariate full MoE (MfMoE).

Thus, we are considering both multivariate unconditional and regression-based families of parsimonious models. Finally, we give conclusions and avenues for further research in Sect. 4.

2 Methodology

Parsimonious Matrix-Variate FMR

Let $\boldsymbol{Y} \in \mathbb{R}^{p\times r}$ be a random matrix containing p responses measured in r occasions. Moreover, suppose we observe a set of q covariates for each occasion, in the form of a matrix $\boldsymbol{X} \in \mathbb{R}^{q\times r}$. A generic matrix-variate regression model [26] for $\boldsymbol{Y}$ has the form

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}^* + \boldsymbol{U}, \tag{1}$$

where $\boldsymbol{B}$ is the $p \times (1+q)$ matrix of regression coefficients, $\boldsymbol{X}^*$ is the $(1+q) \times r$ matrix containing $\boldsymbol{X}$ and information about the intercept, and $\boldsymbol{U}$ is the error term matrix of dimension $p \times r$. In this paper we assume, as usual for FMMVR models, that $\boldsymbol{Y}|\boldsymbol{X}$ has a MVN distribution with $p \times r$ mean matrix $\boldsymbol{B}\boldsymbol{X}^*$, $p \times p$ row covariance matrix $\boldsymbol{\Sigma}$, and $r \times r$ column covariance matrix $\boldsymbol{\Psi}$; for notational convenience we write $\boldsymbol{Y}|\boldsymbol{X} \sim \mathcal{N}_{p\times r}(\boldsymbol{B}\boldsymbol{X}^*, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$. Therefore, the probability density function (pdf) of $\boldsymbol{Y}|\boldsymbol{X}$ is

$$f_{\text{MVN}}(\boldsymbol{Y}; \boldsymbol{B}\boldsymbol{X}^*, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = (2\pi)^{-\frac{pr}{2}} |\boldsymbol{\Sigma}|^{-\frac{r}{2}} |\boldsymbol{\Psi}|^{-\frac{p}{2}} \exp\left\{-\tfrac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{X}^*)\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{B}\boldsymbol{X}^*)'\right]\right\}. \tag{2}$$

Suppose that there exist K disjoint clusters in the data. Thus, the conditional distribution of $\boldsymbol{Y}|\boldsymbol{X}$, according to a FMMVR with K components, can be written as

$$p(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_{\text{MVN}}\left(\boldsymbol{Y}; \boldsymbol{B}_k\boldsymbol{X}^*, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k\right), \tag{3}$$

where $\pi_k > 0$ is the mixing weight, with $\sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\Theta} = \{\pi_k, \boldsymbol{B}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k\}_{k=1}^{K}$.

One concern related to model (3) is the possible high number of parameters to be estimated. To address this issue, we consider the eigen-decomposition of the component covariance matrices. In detail, a $t \times t$ covariance matrix $\boldsymbol{\Phi}_k$ can be decomposed as

$$\boldsymbol{\Phi}_k = \lambda_k \boldsymbol{\Gamma}_k \boldsymbol{\Delta}_k \boldsymbol{\Gamma}_k^{\top}, \tag{4}$$

Table 1 Nomenclature, covariance matrix structure, and number of free parameters in $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_K$ for the parsimonious models obtained via the eigen-decomposition of the component covariance matrices

Family	Model	Type	Volume	Shape	Orientation	# of free parameters in $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_K$
Spherical	EII	$\lambda \boldsymbol{I}$	Equal	Spherical	–	1
Spherical	VII	$\lambda_k \boldsymbol{I}$	Variable	Spherical	–	K
Diagonal	EEI	$\lambda \boldsymbol{\Delta}$	Equal	Equal	Axis-Aligned	t
Diagonal	VEI	$\lambda_k \boldsymbol{\Delta}$	Variable	Equal	Axis-Aligned	$K + t - 1$
Diagonal	EVI	$\lambda \boldsymbol{\Delta}_k$	Equal	Variable	Axis-Aligned	$K(t-1) + 1$
Diagonal	VVI	$\lambda_k \boldsymbol{\Delta}_k$	Variable	Variable	Axis-Aligned	Kt
General	EEE	$\lambda \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Gamma}^\top$	Equal	Equal	Equal	$t(t+1)/2$
General	VEE	$\lambda_k \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Gamma}^\top$	Variable	Equal	Equal	$t(t+1)/2 + K - 1$
General	EVE	$\lambda \boldsymbol{\Gamma} \boldsymbol{\Delta}_k \boldsymbol{\Gamma}^\top$	Equal	Variable	Equal	$t(t-1)/2 + K(t-1) + 1$
General	VVE	$\lambda_k \boldsymbol{\Gamma} \boldsymbol{\Delta}_k \boldsymbol{\Gamma}^\top$	Variable	Variable	Equal	$t(t-1)/2 + Kt$
General	EEV	$\lambda \boldsymbol{\Gamma}_k \boldsymbol{\Delta} \boldsymbol{\Gamma}_k^\top$	Equal	Equal	Variable	$Kt(t-1)/2 + t$
General	VEV	$\lambda_k \boldsymbol{\Gamma}_k \boldsymbol{\Delta} \boldsymbol{\Gamma}_k^\top$	Variable	Equal	Variable	$Kt(t-1)/2 + K + t - 1$
General	EVV	$\lambda \boldsymbol{\Gamma}_k \boldsymbol{\Delta}_k \boldsymbol{\Gamma}_k^\top$	Equal	Variable	Variable	$Kt(t+1)/2 - K + 1$
General	VVV	$\lambda_k \boldsymbol{\Gamma}_k \boldsymbol{\Delta}_k \boldsymbol{\Gamma}_k^\top$	Variable	Variable	Variable	$Kt(t+1)/2$

where $\lambda_k = |\boldsymbol{\Phi}_k|^{1/t}$, $\boldsymbol{\Gamma}_k$ is a $t \times t$ orthogonal matrix whose columns are the normalized eigenvectors of $\boldsymbol{\Phi}_k$, and $\boldsymbol{\Delta}_k$ is the scaled ($|\boldsymbol{\Delta}_k| = 1$) diagonal matrix of the eigenvalues of $\boldsymbol{\Phi}_k$. Geometrically, λ_k determines the volume, $\boldsymbol{\Gamma}_k$ indicates the orientation, and $\boldsymbol{\Delta}_k$ determines the shape of the kth cluster. By imposing constraints on the three components of (4), the fourteen parsimonious models of Table 1 are obtained.

Considering that in (3) we have two covariance matrices for each mixture component, this would yield to $14 \times 14 = 196$ parsimonious FMMVR. However, there is a

non-identifiability issue, since $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi}^* \otimes \boldsymbol{\Sigma}^*$ if $\boldsymbol{\Sigma}^* = a\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}^* = a^{-1}\boldsymbol{\Psi}$. As a result, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are identifiable up to a multiplicative constant a [14]. To avoid this issue, the column covariance matrix $\boldsymbol{\Psi}$ is restricted to have $|\boldsymbol{\Psi}| = 1$, implying that in (4) the parameter λ_k is unnecessary. This reduces the number of models related to $\boldsymbol{\Psi}$ from 14 to 7, i.e., $\boldsymbol{I}$, $\boldsymbol{\Delta}$, $\boldsymbol{\Delta}_k$, $\boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}^\top$, $\boldsymbol{\Gamma}\boldsymbol{\Delta}_k\boldsymbol{\Gamma}^\top$, $\boldsymbol{\Gamma}_k\boldsymbol{\Delta}\boldsymbol{\Gamma}_k^\top$, $\boldsymbol{\Gamma}_k\boldsymbol{\Delta}_k\boldsymbol{\Gamma}_k^\top$. Therefore, we introduce $14 \times 7 = 98$ parsimonious FMMVR.

Maximum Likelihood Estimation

Maximum likelihood (ML) is the traditional approach to estimate the parameters of model (3). ML estimates can be computationally obtained by using the expectation-conditional maximization (ECM) algorithm [27]. The ECM algorithm differs from the well-known expectation-maximization (EM) algorithm [28] because the M-step is replaced by a sequence of simpler and computationally convenient CM-steps.

Let $\mathcal{S} = \{(\boldsymbol{Y}_i, \boldsymbol{X}_i)\}_{i=1}^N$ be a random sample of N independent observations from model (3). In the FMR framework, $\mathcal{S}$ is viewed as being incomplete because, for each observation, we do not know its component membership. To govern this source of incompleteness, we use an indicator vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$, where $z_{ik} = 1$ if observation i is in group k, and $z_{ik} = 0$ otherwise. Now, the complete-data are $\mathcal{S}_c = \{(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{z}_i)\}_{i=1}^N$ and the complete-data log-likelihood is

$$l_c(\boldsymbol{\Theta}|\mathcal{S}_c) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \ln(\pi_k) + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[-\frac{pr}{2}\ln(2\pi) - \frac{r}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{p}{2}\ln|\boldsymbol{\Psi}_k| - \frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_i - \boldsymbol{B}_k\boldsymbol{X}_i^*)\boldsymbol{\Psi}_k^{-1}(\boldsymbol{Y}_i - \boldsymbol{B}_k\boldsymbol{X}_i^*)'\right]\right]. \quad (5)$$

In the following, the quantities marked with one dot correspond to the updates at the previous iteration and those marked with two dots represent the updates at the current iteration. After initialization, done by implementing the approach discussed in [16], the ECM algorithm proceeds as follows.

E-Step

The E-step requires calculation of the conditional expectation of (5), given the observed data and the current estimate of the parameters $\dot{\boldsymbol{\Theta}}$. To do this, we need to calculate

$$\ddot{z}_{ik} = \mathbb{E}_{\dot{\boldsymbol{\Theta}}}[Z_{ik}|\boldsymbol{X}_i, \boldsymbol{Y}_i] = \frac{\dot{\pi}_k f_{\text{MVN}}\left(\boldsymbol{Y}_i; \dot{\boldsymbol{B}}_k\boldsymbol{X}_i^*, \dot{\boldsymbol{\Sigma}}_k, \dot{\boldsymbol{\Psi}}_k\right)}{\sum_{j=1}^{K}\dot{\pi}_j f_{\text{MVN}}\left(\boldsymbol{Y}_i; \dot{\boldsymbol{B}}_j\boldsymbol{X}_i^*, \dot{\boldsymbol{\Sigma}}_j, \dot{\boldsymbol{\Psi}}_j\right)}, \quad (6)$$

which corresponds to the posterior probability that the unlabeled observation $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ belongs to the kth component of the mixture.

CM-Step 1

Consider $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$, where $\boldsymbol{\Theta}_1 = \{\pi_k, \boldsymbol{B}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ and $\boldsymbol{\Theta}_2 = \{\boldsymbol{\Psi}_k\}_{k=1}^K$. At the first CM-step, we maximize the expectation of the complete-data log-likelihood with respect to $\boldsymbol{\Theta}_1$, fixing $\boldsymbol{\Theta}_2$ at $\dot{\boldsymbol{\Theta}}_2$. In particular, we obtain

$$\ddot{\pi}_k = \frac{\sum_{i=1}^N \ddot{z}_{ik}}{N} \quad \text{and} \quad \ddot{\boldsymbol{B}}_k = \left[\sum_{i=1}^N \ddot{z}_{ik} \boldsymbol{Y}_i \dot{\boldsymbol{\Psi}}_k^{-1} \boldsymbol{X}_i^{*\top}\right] \left[\sum_{i=1}^N \ddot{z}_{ik} \boldsymbol{X}_i^* \dot{\boldsymbol{\Psi}}_k^{-1} \boldsymbol{X}_i^{*\top}\right]^{-1}. \quad (7)$$

The update for $\boldsymbol{\Sigma}_k$ depends on the parsimonious structure considered. For notational simplicity, let $\ddot{\boldsymbol{V}} = \sum_{k=1}^K \ddot{\boldsymbol{V}}_k$ be the update of the within component row scatter matrix, where $\ddot{\boldsymbol{V}}_k = \sum_{i=1}^N \ddot{z}_{ik} \left(\boldsymbol{Y}_i - \ddot{\boldsymbol{B}}_k \boldsymbol{X}_i^*\right) \dot{\boldsymbol{\Psi}}_k^{-1} \left(\boldsymbol{Y}_i - \ddot{\boldsymbol{B}}_k \boldsymbol{X}_i^*\right)'$ is the update of the row scatter matrix related to the kth component. The updates for the 14 parsimonious structures of $\boldsymbol{\Sigma}_k$ are:

- Model EII [$\boldsymbol{\Sigma}_k = \lambda \boldsymbol{I}$]

$$\ddot{\lambda} = \frac{\text{tr}\left\{\ddot{\boldsymbol{V}}\right\}}{prN};$$

- Model VII [$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{I}$]

$$\ddot{\lambda}_k = \frac{\text{tr}\left\{\ddot{\boldsymbol{V}}_k\right\}}{pr \sum_{i=1}^N \ddot{z}_{ik}};$$

- Model EEI [$\boldsymbol{\Sigma}_k = \lambda \boldsymbol{\Delta}$]

$$\ddot{\boldsymbol{\Delta}} = \frac{\text{diag}\left(\ddot{\boldsymbol{V}}\right)}{\left|\text{diag}\left(\ddot{\boldsymbol{V}}\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\left|\text{diag}\left(\ddot{\boldsymbol{V}}\right)\right|^{\frac{1}{p}}}{rN};$$

- Model VEI [$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{\Delta}$]

$$\ddot{\boldsymbol{\Delta}} = \frac{\text{diag}\left(\sum_{k=1}^K \dot{\lambda}_k^{-1} \ddot{\boldsymbol{V}}_k\right)}{\left|\text{diag}\left(\sum_{k=1}^K \dot{\lambda}_k^{-1} \ddot{\boldsymbol{V}}_k\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_k = \frac{\text{tr}\left\{\ddot{\boldsymbol{\Delta}}^{-1} \ddot{\boldsymbol{V}}_k\right\}}{pr \sum_{i=1}^N \ddot{z}_{ik}};$$

- Model EVI [$\boldsymbol{\Sigma}_k = \lambda \boldsymbol{\Delta}_k$]

$$\ddot{\boldsymbol{\Delta}}_k = \frac{\operatorname{diag}\left(\ddot{\boldsymbol{V}}_k\right)}{\left|\operatorname{diag}\left(\ddot{\boldsymbol{V}}_k\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\sum_{k=1}^{K}\left|\operatorname{diag}\left(\ddot{\boldsymbol{V}}_k\right)\right|^{\frac{1}{p}}}{rN};$$

- Model VVI [$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{\Delta}_k$]

$$\ddot{\boldsymbol{\Delta}}_k = \frac{\operatorname{diag}\left(\ddot{\boldsymbol{V}}_k\right)}{\left|\operatorname{diag}\left(\ddot{\boldsymbol{V}}_k\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_k = \frac{\left|\operatorname{diag}\left(\ddot{\boldsymbol{V}}_k\right)\right|^{\frac{1}{p}}}{r\sum_{i=1}^{N}\ddot{z}_{ik}};$$

- Model EEE [$\boldsymbol{\Sigma}_k = \lambda \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}^\top$]

$$\ddot{\boldsymbol{\Sigma}} = \frac{\ddot{\boldsymbol{V}}}{rN};$$

- Model VEE [$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}^\top$]

$$\ddot{\boldsymbol{\Gamma}}\ddot{\boldsymbol{\Delta}}\ddot{\boldsymbol{\Gamma}}^\top = \frac{\sum_{k=1}^{K}\dot{\lambda}_k^{-1}\ddot{\boldsymbol{V}}_k}{\left|\sum_{k=1}^{K}\dot{\lambda}_k^{-1}\ddot{\boldsymbol{V}}_k\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_k = \frac{\operatorname{tr}\left\{(\ddot{\boldsymbol{\Gamma}}\ddot{\boldsymbol{\Delta}}\ddot{\boldsymbol{\Gamma}}^\top)^{-1}\ddot{\boldsymbol{V}}_k\right\}}{pr\sum_{i=1}^{N}\ddot{z}_{ik}};$$

- Model EVE [$\boldsymbol{\Sigma}_k = \lambda \boldsymbol{\Gamma}\boldsymbol{\Delta}_k\boldsymbol{\Gamma}^\top$] There is no analytical solution for $\boldsymbol{\Gamma}$. Therefore, we implement an iterative minorization-maximization (MM) algorithm [29]. Specifically, a surrogate function can be constructed as

$$f(\boldsymbol{\Gamma}) = \sum_{k=1}^{K}\operatorname{tr}\left\{\boldsymbol{V}_k\boldsymbol{\Gamma}\boldsymbol{\Delta}_k^{-1}\boldsymbol{\Gamma}^\top\right\} \leq S + \operatorname{tr}\left\{\dot{\boldsymbol{F}}\boldsymbol{\Gamma}\right\},$$

where S is a constant and $\dot{\boldsymbol{F}} = \sum_{k=1}^{K}\left(\boldsymbol{\Delta}_k^{-1}\dot{\boldsymbol{\Gamma}}^\top\boldsymbol{V}_k - e_k\boldsymbol{\Delta}_k^{-1}\dot{\boldsymbol{\Gamma}}^\top\right)$, with e_k being the largest eigenvalue of $\boldsymbol{V}_k$. The update of $\boldsymbol{\Gamma}$ is given by $\ddot{\boldsymbol{\Gamma}} = \dot{\boldsymbol{G}}\dot{\boldsymbol{H}}^\top$, where $\dot{\boldsymbol{G}}$ and $\dot{\boldsymbol{H}}$ are obtained from the singular value decomposition of $\dot{\boldsymbol{F}}$. This process is repeated until a specified convergence criterion is met and the estimate $\ddot{\boldsymbol{\Gamma}}$ is obtained from the last iteration. Then, we obtain

$$\ddot{\boldsymbol{\Delta}}_k = \frac{\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}^\top\ddot{\boldsymbol{V}}_k\ddot{\boldsymbol{\Gamma}}\right)}{\left|\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}^\top\ddot{\boldsymbol{V}}_k\ddot{\boldsymbol{\Gamma}}\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\sum_{k=1}^{K}\operatorname{tr}\left(\ddot{\boldsymbol{\Gamma}}\ddot{\boldsymbol{\Delta}}_k^{-1}\ddot{\boldsymbol{\Gamma}}^\top\ddot{\boldsymbol{V}}_k\right)}{prN};$$

- Model VVE [$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{\Gamma}\boldsymbol{\Delta}_k\boldsymbol{\Gamma}^\top$] Similarly to the EVE case, there is no analytical solution for $\boldsymbol{\Gamma}$, and its update is obtained by employing the MM algorithm described above. Then, we have

$$\ddot{\boldsymbol{\Delta}}_k = \frac{\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}^\top \ddot{\boldsymbol{V}}_k \ddot{\boldsymbol{\Gamma}}\right)}{\left|\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}^\top \ddot{\boldsymbol{V}}_k \ddot{\boldsymbol{\Gamma}}\right)\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_k = \frac{\left|\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}^\top \ddot{\boldsymbol{V}}_k \ddot{\boldsymbol{\Gamma}}\right)\right|^{\frac{1}{p}}}{r \sum_{i=1}^N \ddot{z}_{ik}};$$

- Model EEV [$\boldsymbol{\Sigma}_k = \lambda \boldsymbol{\Gamma}_k \boldsymbol{\Delta} \boldsymbol{\Gamma}_k^\top$] An algorithm similar to the one proposed by [20] is employed. First of all, consider the eigen-decomposition $\boldsymbol{V}_k = \boldsymbol{L}_k \boldsymbol{\Omega}_k \boldsymbol{L}_k^\top$, with eigenvalues in the diagonal matrix $\boldsymbol{\Omega}_k$ following descending order and orthogonal matrix $\boldsymbol{L}_k$ composed of the corresponding eigenvectors. Then, we obtain

$$\ddot{\boldsymbol{\Gamma}}_k = \ddot{\boldsymbol{L}}_k, \quad \ddot{\boldsymbol{\Delta}} = \frac{\sum_{k=1}^K \ddot{\boldsymbol{\Omega}}_k}{\left|\sum_{k=1}^K \ddot{\boldsymbol{\Omega}}_k\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\left|\sum_{k=1}^K \ddot{\boldsymbol{\Omega}}_k\right|^{\frac{1}{p}}}{rN};$$

- Model VEV [$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{\Gamma}_k \boldsymbol{\Delta} \boldsymbol{\Gamma}_k^\top$] By using the same algorithm applied for the EEV model, we have

$$\ddot{\boldsymbol{\Gamma}}_k = \ddot{\boldsymbol{L}}_k, \quad \ddot{\boldsymbol{\Delta}} = \frac{\sum_{k=1}^K \lambda_k^{-1} \ddot{\boldsymbol{\Omega}}_k}{\left|\sum_{k=1}^K \lambda_k^{-1} \ddot{\boldsymbol{\Omega}}_k\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda}_k = \frac{\operatorname{tr}\left\{\ddot{\boldsymbol{\Omega}}_k \ddot{\boldsymbol{\Delta}}^{-1}\right\}}{pr \sum_{i=1}^N \ddot{z}_{ik}};$$

- Model EVV [$\boldsymbol{\Sigma}_k = \lambda \boldsymbol{\Gamma}_k \boldsymbol{\Delta}_k \boldsymbol{\Gamma}_k^\top$]

$$\ddot{\boldsymbol{\Gamma}}_k \boldsymbol{\Delta}_k \boldsymbol{\Gamma}_k^\top = \frac{\ddot{\boldsymbol{V}}_k}{\left|\ddot{\boldsymbol{V}}_k\right|^{\frac{1}{p}}} \quad \text{and} \quad \ddot{\lambda} = \frac{\sum_{k=1}^K \left|\ddot{\boldsymbol{V}}_k\right|^{\frac{1}{p}}}{rN};$$

- Model VVV [$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{\Gamma}_k \boldsymbol{\Delta}_k \boldsymbol{\Gamma}_k^\top$]

$$\ddot{\boldsymbol{\Sigma}}_k = \frac{\ddot{\boldsymbol{V}}_k}{r \sum_{i=1}^N \ddot{z}_{ik}}.$$

CM-Step 2

At the second CM-step, we maximize the expectation of the complete-data log-likelihood with respect to $\boldsymbol{\Theta}_2$, keeping $\boldsymbol{\Theta}_1$ fixed at $\ddot{\boldsymbol{\Theta}}_1$. The update for $\boldsymbol{\Psi}_k$ depends on which of the 7 parsimonious structures is considered. For notational purposes, let $\ddot{\boldsymbol{W}} = \sum_{k=1}^K \ddot{\boldsymbol{W}}_k$ be the update of the within component column scatter matrix, where $\ddot{\boldsymbol{W}}_k = \sum_{i=1}^N \ddot{z}_{ik} \left(\boldsymbol{Y}_i - \ddot{\boldsymbol{B}}_k \boldsymbol{X}_i^*\right)' \ddot{\boldsymbol{\Sigma}}_k^{-1} \left(\boldsymbol{Y}_i - \ddot{\boldsymbol{B}}_k \boldsymbol{X}_i^*\right)$ is the update of the column

scatter matrix related to the kth component. With the exclusion of the II model, for which no parameters need to be estimated, we have:

- Model EI [$\boldsymbol{\Psi}_k = \boldsymbol{\Delta}$]

$$\ddot{\boldsymbol{\Delta}} = \frac{\operatorname{diag}\left(\ddot{\boldsymbol{W}}\right)}{\left|\operatorname{diag}\left(\ddot{\boldsymbol{W}}\right)\right|^{\frac{1}{r}}};$$

- Model VI [$\boldsymbol{\Psi}_k = \boldsymbol{\Delta}_k$]

$$\ddot{\boldsymbol{\Delta}}_k = \frac{\operatorname{diag}\left(\ddot{\boldsymbol{W}}_k\right)}{\left|\operatorname{diag}\left(\ddot{\boldsymbol{W}}_k\right)\right|^{\frac{1}{r}}};$$

- Model EE [$\boldsymbol{\Psi}_k = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}^\top$]

$$\ddot{\boldsymbol{\Psi}} = \frac{\ddot{\boldsymbol{W}}}{\left|\ddot{\boldsymbol{W}}\right|^{\frac{1}{r}}};$$

- Model VE [$\boldsymbol{\Psi}_k = \boldsymbol{\Gamma}\boldsymbol{\Delta}_k\boldsymbol{\Gamma}^\top$] Similarly to the EVE and VVE models in CM-Step 1, there is no analytical solution for $\boldsymbol{\Gamma}$. Thus, we implement an MM algorithm by following the same procedure explained for the EVE model in CM-Step 1, but replacing $\boldsymbol{V}$ with $\boldsymbol{W}$. Then, we have

$$\ddot{\boldsymbol{\Delta}}_k = \frac{\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}^\top \ddot{\boldsymbol{W}}_k \ddot{\boldsymbol{\Gamma}}\right)}{\left|\operatorname{diag}\left(\ddot{\boldsymbol{\Gamma}}^\top \ddot{\boldsymbol{W}}_k \ddot{\boldsymbol{\Gamma}}\right)\right|^{\frac{1}{r}}};$$

- Model EV [$\boldsymbol{\Psi}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Delta}\boldsymbol{\Gamma}_k^\top$] By using the same approach of the EEV and VEV models in CM-Step 1, but changing $\ddot{\boldsymbol{V}}$ with $\ddot{\boldsymbol{W}}$, we have

$$\ddot{\boldsymbol{\Gamma}}_k = \ddot{\boldsymbol{L}}_k \quad \text{and} \quad \ddot{\boldsymbol{\Delta}} = \frac{\sum_{k=1}^{K} \ddot{\boldsymbol{\Omega}}_k}{\left|\sum_{k=1}^{K} \ddot{\boldsymbol{\Omega}}_k\right|^{\frac{1}{r}}};$$

- Model VV [$\boldsymbol{\Psi}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Delta}_k\boldsymbol{\Gamma}_k^\top$]

$$\ddot{\boldsymbol{\Psi}}_k = \frac{\ddot{\boldsymbol{W}}_k}{\left|\ddot{\boldsymbol{W}}_k\right|^{\frac{1}{r}}}.$$

Computational and Operative Details

Computation is performed on a Windows 10 PC, with AMD Ryzen 7 3700x CPU, 16.0 GB RAM, using the R 64-bit statistical software. Given the high number of parsimonious models, we consider parallel computing using 14 cores.

As commonly done in the model-based clustering literature, to select the number of groups K we use the BIC and the ICL. Furthermore, to assess the clustering performance of the parsimonious models, we use the well-known adjusted Rand index (ARI; [30]), that calculates the agreement between the true classification and the one predicted by the considered model. An ARI of 1 indicates perfect agreement between the two partitions, whereas the expected value of the ARI under a random classification is 0.

3 Data Analyses

In this section, we investigate several aspects related to our models by using both simulated and real data.

Simulated Data

Here, we investigate the capability of BIC and ICL in detecting the data generating model, and the classifications produced by the best fitting models according to both the information criteria. Considering the high number of models, we only generate data from the EEE–EE FMMVR for the sake of simplicity. We set $p = 2$, $r = 10$, $q = 3$, $N = 250$, and $K = 2$. We generate 50 datasets. On each dataset, all the 98 parsimonious FMMVR models are fitted for $K \in \{1, 2, 3\}$, and the best fitting model according to BIC and ICL is considered.

Overall, we obtained the same results by using both information criteria. The first aspect we noticed is that the correct K is selected 49 times out of 50. The only exception is a case where $K = 3$ is chosen. As concerns the recovering of the parsimonious covariance structure, 49 times out of 50 the best fitting model has an EEE–EE structure, with only one case where the VEE–VE structure is preferred. Lastly, in terms of classification, we obtained an average ARI of 0.99 indicating a practically perfect clustering of the data.

Real Data

Data Description

The National Agency for the Evaluation of Universities and Research Institutes (ANVUR) plays a fundamental role in the Italian higher education system. Among

its tasks, this agency maintains data on Italian universities' quantitative indicators concerning the academic careers of the students as well as the results of their teaching activities. Such data are considered in the following. Specifically, $N = 75$ degree families in the non-telematics Italian universities are analyzed. According to Italian law, each family represents a set of degrees sharing a similar topic and having identical legal value. There are $K = 2$ groups in the data: the first is composed of 33 families of bachelor's degrees while the second is constituted by 42 families of master's degrees. Considering the relatively small sample size, it might be useful to consider parsimonious models.

For this application, we consider the following $p = 2$ responses that measure the level of completion of studies by students:

1. the percentage of students that graduate within T years,
2. the percentage of students that drop after $T + 1$ years,

where T is the normal duration of the study program. Then, the following $q = 3$ covariates are taken into account:

1. the percentage of course credits earned in the first year over the total to be achieved,
2. the percentage of first year students that have earned at least 1/3 of the first year course credits,
3. the percentage of students that have earned at least 40 course credits during the solar year.

All these indicators are measured over $r = 3$ years. Every family of degrees is measured at the national level, i.e., the value of each indicator corresponds to the average of all the degrees belonging to the same family, across the country, for the reference year.

Results

We fit the competing parsimonious models to the data for $K \in \{1, 2, 3\}$. Relatedly, to fit the multivariate models mentioned in Sect. 1, we use the **MoEClust** package [31]. Before showing the clustering results, we report in Table 2 information on the computational times taken by fitting the 98 parsimonious FMMVR sequentially (default in `R`) and by parallelizing them on 14 cores. As we can see, the computational burden is greatly reduced, and all the models can be fitted in a reasonable fast way. More in detail, by analyzing the sequential times, the fastest fitted model (0.09 s.) is the EVI–VI FMMVR with $K = 1$, whereas the slowest (4.88 s.) is the EII–EE FMMVR with $K = 3$.

Table 3 presents the parsimonious structure, number of groups K, ARI and number of estimated parameters of the best models selected by BIC and ICL (which provide the same results), for each family of models.

By starting with the analysis of the parsimonious matrix-variate models, we can see that for both MVNM and FMMVR the best model has an EVI–EE parsimonious structure and $K = 2$ groups. However, differently from the EVI–EE MVNM, the

Table 2 Computational time (in seconds) for fitting our 98 parsimonious FMMVR sequentially or via parallel computing, for each value of K

K	Sequential	Parallel
1	32.40	5.83
2	154.72	16.17
3	176.71	16.64

Table 3 Parsimonious configuration, number of clusters (K), ARI and number of estimated parameters of the best models selected by BIC and ICL, for each family of models

Model	Parsimonious configuration	K	ARI	# of parameters
MVNM	EVI–EE	2	0.84	21
FMMVR	EVI–EE	2	1.00	25
MNM	EEV	2	0.94	49
MFMR	VVV	1	0.00	45
MgMoE	VEE	3	0.64	49
MfMoE	VVV	1	0.00	45

EVI-EE FMMVR produces a perfect data classification (ARI = 1). Thus, in this case, taking into account a regression structure aided in better clustering performance, at the cost of only four additional parameters.

When the multivariate models are considered, we notice that the best clustering result is provided by the EEV MNM with $K = 2$. It is interesting to note that this is the only multivariate model for which the true number of groups is selected; moreover, its classification is better than the one from its matrix-variate counterpart. Nevertheless, it produces a worse classification compared to the EVI–EE FMMVR. Furthermore, it should be also noticed the increased number of parameters obtained by working with multivariate models on the vectorized data. This is because, with the exclusion of the spherical cases that do not depend on the data dimensionality (see Table 1), matrix-variate models have the desirable feature of reducing the number of free covariance parameters [15, 16], adding further parsimony with respect to the multivariate models.

4 Conclusions

A family of 98 parsimonious FMMVR has been introduced. Parsimony has been attained by using the eigen-decomposition of the two matrix-variate covariance matrices (of the responses) for each mixture component. An expectation-conditional maximization algorithm has been illustrated and used for parameter estimation. Our

family of models has been fitted to simulated data and to a real dataset concerning Italian universities' quantitative indicators along with matrix-variate and multivariate parsimonious competitors. We first illustrated that BIC and ICL correctly recover the data generating model on simulated data. Then, on the real dataset, we reported the computational times for fitting our family of models, that by exploiting parallel computing are quite low. Furthermore, we have investigated the clustering performance of all the competing models, obtaining a perfect data classification only when the best among our models is considered. The disadvantages of considering the multivariate models (on the vectorized data) with respect to matrix-variate models, in terms of number of estimated parameters, have been also illustrated.

Possible directions for future work in this area include the use of other matrix-variate distributions for the error term of each mixture component. For example, to accommodate skewness, families of transformations leading to near-normality [13, 18, 32] or skewed matrix-variate distributions [33, 34] could be considered. Another possible avenue could be to extend our models in a tensor-variate framework, in line with the work of [35] who introduce tensor-variate mixtures.

References

1. McNicholas, P. D. (2016). *Mixture model-based classification*. Boca Raton: Chapman and Hall/CRC Press.
2. Murphy, K., & Murphy, T. B. (2020). Gaussian parsimonious clustering models with covariates and a noise component. *Advances in Data Analysis and Classification, 14*, 293–325.
3. DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification, 5*(2), 249–282.
4. Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer Science & Business Media.
5. Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83*(401), 173–178.
6. Chamroukhi, F. (2017). Skew t mixture of experts. *Neurocomputing, 266*, 390–408.
7. Doğru, F. Z., & Arslan, O. (2017). Parameter estimation for mixtures of skew Laplace normal distributions and application in mixture regression modeling. *Communications in Statistics-Theory and Methods, 46*(21), 10879–10896.
8. Mazza, A., Battisti, M., Ingrassia, S., & Punzo, A. (2019). Modeling return to education in heterogeneous populations: An application to Italy. In I. Greselin, L. Deldossi, L. Bagnato, & M. Vichi (Eds.), *Statistical Learning of Complex Data, Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 121–131). Switzerland: Springer International Publishing.
9. Mazza, A., & Punzo, A. (2020). Mixtures of multivariate contaminated normal regression models. *Statistical Papers, 61*(2), 787–822.
10. Viroli, C. (2011). Model based clustering for three-way data structures. *Bayesian Analysis, 6*(4), 573–602.
11. Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing, 21*(4), 511–522.
12. Gallaugher, M. P. B., & McNicholas, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition, 80*, 83–93.
13. Melnykov, V., Zhu, X., P. D. (2018). On model-based clustering of skewed matrix data. *Journal of Multivariate Analysis, 167,* 181–194.

14. Sarkar, S., Zhu, X., Melnykov, V., & Ingrassia, S. (2020). On parsimonious models for modeling matrix data. *Computational Statistics & Data Analysis, 142*, 106822.
15. Gallaugher, M. P. B., & McNicholas, P. D. (2020). Mixtures of skewed matrix variate bilinear factor analyzers. *Advances in Data Analysis and Classification, 14*, 415–434.
16. Tomarchio, S. D., Punzo, A., & Bagnato, L. (2020). Two new matrix-variate distributions with application in model-based clustering. *Computational Statistics & Data Analysis, 152*, 107050.
17. Tomarchio, S. D., Gallaugher, M. P. B., Punzo, A., & McNicholas, P. D. (2022). Mixtures of matrix-variate contaminated normal distributions. *Journal of Computational and Graphical Statistics, 31*(2), 413–421.
18. Melnykov, V., & Zhu, X. (2019). Studying crime trends in the USA over the years 2000–2012. *Advances in Data Analysis and Classification, 13*(1), 325–341.
19. Tomarchio, S. D., McNicholas, P. D., & Punzo, A. (2021). Matrix normal cluster-weighted models. *Journal of Classification, 38*(3), 556–575.
20. Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition, 28*(5), 781–793.
21. Tomarchio, S. D., Punzo, A., & Maruotti, A. (2022). Parsimonious Hidden Markov Models for Matrix-Variate Longitudinal Data. *Statistics and Computing, 32*(3), 1–18.
22. Gallaugher, M. P. B., & McNicholas, P. D. (2020). Parsimonious mixtures of matrix variate bilinear factor analyzers. *Advanced Studies in Behaviormetrics and Data Science: Essays in honor of Akinori Okada* (pp. 177–196).
23. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.
24. Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(7), 719–725.
25. Murphy, K., & Murphy, T. B. (2020). Gaussian parsimonious clustering models with covariates and a noise component. *Advances in Data Analysis and Classification, 14*, 293–325.
26. Viroli, C. (2012). On matrix-variate regression analysis. *Journal of Multivariate Analysis, 111*, 296–309.
27. Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*(2), 267–278.
28. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.
29. Browne, R. P., & McNicholas, P. D. (2014). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification, 8*(2), 217–226.
30. Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.
31. Murphy, K., & Murphy, T. B. (2020). MoEClust: Gaussian Parsimonious Clustering Models with Covariates and a Noise Component. https://cran.r-project.org/package=MoEClust, R package version 1.3.3.
32. Zhu, X., Sarkar, S., & Melnykov, V. (2022). MatTransMix: An R package for matrix model-based clustering and parsimonious mixture modeling. *Journal of Classification, 39*(1), 147–170.
33. Gallaugher, M. P. B., & McNicholas, P. D. (2017). A matrix variate skew-t distribution. *Statistics, 6*(1), 160–170.
34. Gallaugher, M. P. B., & McNicholas, P. D. (2019). Three skewed matrix variate distributions. *Statistics & Probability Letters, 145*, 103–109.
35. Sarkar, S., Melnykov, V., & Zhu, X. (2021). Tensor-variate finite mixture modeling for the analysis of university professor remuneration. *The Annals of Applied Statistics, 15*(2), 1017–1036.

Robust Multivariate Modelling for Heterogeneous Data Sets with Mixtures of Multivariate Skew Laplace Normal Distributions

Fatma Zehra Doğru and Olcay Arslan

Abstract Modelling multivariate heterogeneous data with taking into account skewness and thick-tailedness is a challenging problem. Finite mixture model of multivariate normal distributions is often considered for modelling heterogeneous data sets in multivariate settings. However, in real-life problems, besides heterogeneity of the data sets, they may have skewed and/or heavy-tailed empirical distributions so that modelling with finite mixture of normal distributions may not provide an adequate model to represent all the features of data. Finite mixtures of more flexible multivariate distributions have been introduced in the literature to simultaneously overcome heterogeneity, skewness and heavy-tailedness in multivariate data sets. In this study, using the favourable properties of the multivariate skew Laplace normal (MSLN) distribution proposed by [13, 14], we introduce finite mixtures of MSLN distributions that can be considered as an alternative mixture model to deal with skewness and heavy-tailedness simultaneously in multivariate heterogeneous data sets. We will propose the maximum likelihood (ML) estimation method to estimate the parameters of finite mixtures of MSLN distributions via the expectation-maximization (EM) algorithm proposed by [10]. We will provide a simulation study and a real data example to demonstrate the performance and the applicability of proposed mixture model.

1 Introduction

Finite mixture models provide a wide range of applicability in many fields such as classification, cluster and latent class analysis, density estimation, data mining, image analysis, genetics, medicine, pattern recognition, etc. Further, these models are very useful because of their huge flexibility and approximation ability of com-

F. Z. Doğru (✉)
Department of Statistics, Faculty of Arts and Sciences, Giresun University, 28200 Giresun, Turkey
e-mail: fatma.dogru@giresun.edu.tr

O. Arslan
Department of Statistics, Faculty of Science, Ankara University, 06100 Ankara, Turkey
e-mail: oarslan@ankara.edu.tr

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_18

plex densities in modelling heterogeneous data set with multimodality, skewness and heavy-tailedness simultaneously. One can find broad studies on finite mixture models and applications, for instance, the monographs by [9, 17, 20, 30, 31, 33, 44], the edited volume of [34] and also the papers by [7, 19].

Some flexible distributions have been proposed to model multivariate skew and/or heavy-tailed data sets. For instance, the multivariate skew-normal (MSN) distribution was proposed by [4, 23] for modelling skewness in multivariate settings deviating from normality. However, the MSN distribution fails in modelling heavy-tailedness. Further, the multivariate skew and heavy-tailed data sets are modelled using multivariate skew-t (MST) distribution proposed by [28, 40] and multivariate skew-t-normal (MSTN) distribution recently proposed by [29]. However, these two multivariate skew and heavy-tailed distributions have been generated using multivariate t-distribution and it is known that the t-distribution has an extra parameter called degrees of freedom which usually causes computational intensity. Additionally, multivariate skew generalized Laplace normal (MSGLN) distribution is proposed by [45] for modelling skew and heavy-tailed multivariate data settings. Both MSTN and MSGLN distributions can be used for modelling skew and heavy-tailed multivariate heterogeneous data settings. Alternatively, the multivariate skew Laplace normal (MSLN) distribution was proposed by [13, 14] for modelling both skewness and heavy-tailedness in multivariate data sets which is a special case of the MSGLN distribution. We note that if the parameter α in the MSGLN distribution equals $(p+1)/2$ the distribution will be the MSLN distribution. The MSLN distribution is also more practicable than the MSN distribution owing to its extensive range of skewness and heavy-tailedness for modelling multivariate data sets. In spite of their usefulness for modelling skew and heavy-tailed multivariate data sets, the MST and MSTN distributions have an extra degrees of freedom parameter and the MSGLN distribution includes an extra shape parameter, which represents a very broad form of these families with several parameters. These extra parameters cause more computational time in the estimation procedure. On the other hand, for the MSLN case, there are only three parameters to deal with and this enables computational easiness. Because of this important advantage, the MSLN distribution will be an alternative to recently proposed MSTN and MSGLN distributions to model both skewness and heavy-tailedness in the multivariate data settings.

Furthermore, concerning heterogeneous multivariate skew and/or heavy-tailed data sets, [37] proposed finite mixtures of multivariate t distributions to model heterogeneous multivariate heavy-tailed data sets as an extension of finite mixtures of multivariate normal distributions. Also, finite mixtures of MSN distributions were proposed by Lin (2009) to model heterogeneous multivariate skew data sets as an alternative to the finite mixtures of multivariate normal distributions. However, both multivariate t and MSN distributions lack the flexibility and robustness for modelling skewness and heavy-tailedness simultaneously. For this purpose, in heterogeneous multivariate data sets, finite mixtures of MST distributions by [28, 40] and finite

mixtures of MSTN distributions by [29] have been introduced for modelling both skewness and heavy-tailedness. Recently, finite mixtures of multivariate scale-shape mixtures of skew-normal distributions have been proposed by [45]. To gain both modelling skewness and heavy-tailedness simultaneously in multivariate heterogeneous data sets and more superiority in computation tractability than given models in literature, we generalize the MSLN distribution to finite mixtures of MSLN distributions as a competitive alternative to finite mixtures of MSTN distributions and finite mixtures of MSGLN distributions.

The remainder of the paper is designed as follows. Section 2 describes some notations and properties of the MSLN distribution. Section 3 introduces finite mixtures of MSLN distributions and offers an EM algorithm for the ML estimators of the proposed model. Also, in the same section, the initial values for the EM algorithm and empirical information matrix of finite mixtures of MSLN distributions to calculate the standard errors of the proposed model are given. Section 4 tests the applicability of the proposed model with a simulation study and a real data example. Section 5 gives some conclusions.

2 The MSLN Distribution

A random vector $\boldsymbol{Y} \in R^p$ is said to have the MSLN distribution proposed by [13, 14] ($\boldsymbol{Y} \sim MSLN\left(\boldsymbol{\mu}, \Sigma, \boldsymbol{\lambda}\right)$) if its stochastic representation is given as follows:

$$\boldsymbol{Y} = \boldsymbol{\mu} + \Sigma^{1/2}\left[\frac{\boldsymbol{\lambda}|U_1|}{\sqrt{V\left(V + \boldsymbol{\lambda}^T\boldsymbol{\lambda}\right)}} + \left(V I_p + \boldsymbol{\lambda}\boldsymbol{\lambda}^T\right)^{-1/2}\boldsymbol{U}_2\right], \tag{1}$$

where $\boldsymbol{\mu} \in R^p$, Σ is a positive definite matrix, $U_1 \sim N\left(0, 1\right)$, $\boldsymbol{U}_2 \sim N_p\left(0, I_p\right)$ and V has the inverse gamma distribution with the following probability density function (pdf):

$$g\left(v\right) = \frac{1}{\Gamma\left(\frac{p+1}{2}\right)2^{\frac{p+1}{2}}} v^{-\left(\frac{p+1}{2}+1\right)} e^{-\frac{1}{2v}}, \ v > 0. \tag{2}$$

Here, the random variables $U_1, \boldsymbol{U}_2$ and V are mutually independent. Note that the stochastic representation given in (1) will be useful for generating random numbers and implementing the steps of the EM algorithm in the ML estimation. Further, let $\gamma = \sqrt{V^{-1}\left(V + \boldsymbol{\lambda}^T\boldsymbol{\lambda}\right)}|U_1|$. Then, the hierarchical representation of MSLN distribution is given below:

$$\begin{aligned} \boldsymbol{Y}|\left(\gamma, v\right) &\sim N_p\left(\boldsymbol{\mu}+\frac{\Sigma^{\frac{1}{2}}\boldsymbol{\lambda}\gamma}{v+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}, \Sigma^{\frac{1}{2}}\left(vI_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right)^{-1}\Sigma^{\frac{1}{2}}\right), \\ \gamma|v &\sim TN\left(0, \frac{v+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}{v}; (0,\infty)\right), \\ v &\sim g\left(v\right), \end{aligned} \tag{3}$$

where $TN\left(\cdot\right)$ denotes the truncated normal distribution. Using this hierarchical representation we can give the following joint pdf of $\boldsymbol{Y}$, γ, and V:

$$f\left(\boldsymbol{y}, \gamma, v\right) = \frac{1}{2^p\pi^{\frac{p+1}{2}}|\Sigma|^{\frac{1}{2}}}\frac{v^{-\frac{3}{2}}e^{-\frac{1}{2v}}}{\Gamma\left(\frac{p+1}{2}\right)}\exp\left\{-\frac{1}{2}\left(v\boldsymbol{u}^T\boldsymbol{u}+\left(\gamma-\boldsymbol{\lambda}^T\boldsymbol{u}\right)^2\right)\right\}, \tag{4}$$

where $\boldsymbol{u}=\Sigma^{-\frac{1}{2}}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)$. Taking the integral of (4) over γ gives the following joint density of $(\boldsymbol{Y}, V)$:

$$f\left(\boldsymbol{y}, v\right) = \frac{1}{2^{p-\frac{1}{2}}\pi^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\frac{v^{-\frac{3}{2}}e^{-\frac{1}{2v}}}{\Gamma\left(\frac{p+1}{2}\right)}\exp\left\{-\frac{v\boldsymbol{u}^T\boldsymbol{u}}{2}\right\}\Phi\left(\boldsymbol{\lambda}^T\boldsymbol{u}\right). \tag{5}$$

Then, the marginal pdf of $\boldsymbol{Y}$ will be obtained by integrating (5) over v:

$$f_{MSLN}\left(\boldsymbol{y};\boldsymbol{\mu},\Sigma,\boldsymbol{\lambda}\right) = 2f_{MLap}\left(\boldsymbol{y};\boldsymbol{\mu},\Sigma\right)\Phi\left(\boldsymbol{\lambda}^T\Sigma^{-\frac{1}{2}}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right), \tag{6}$$

where $\boldsymbol{\mu}\in R^p$ is a location parameter, $\lambda\in R^p$ is a skewness parameter, Σ is the scatter matrix and $f_{MLap}\left(\cdot;\boldsymbol{\mu},\Sigma\right)$ denotes the pdf of multivariate Laplace (MLap) distribution with the location vector $\boldsymbol{\mu}$ and the scatter matrix Σ. Here, the MLap distribution is a special case of the multivariate Kotz-type distribution studied by [36, 38] that its pdf is given by

$$f_{MLap}\left(\boldsymbol{y};\boldsymbol{\mu},\Sigma\right) = \frac{|\Sigma|^{-\frac{1}{2}}}{2^p\pi^{\frac{p-1}{2}}\Gamma\left(\frac{p+1}{2}\right)}e^{-\sqrt{(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}}, \quad \boldsymbol{y}\in R^p, p\geq 1, \tag{7}$$

(one can see the papers [16, 18, 22, 26, 35] for further information). We note that there are other extensions of the univariate Laplace distribution called MLap distribution such as studied by [2, 25]. Additionally, the pdf given in (7) can be obtained as a scale mixture of the multivariate normal (MN) distribution (e.g. see [21, 26]). As it was stated by [3], using the scale mixture approach provides MLap distribution has a longer tail than the MN distribution which enables a wider study field in robust modelling. However, the MLap distribution is not well enough for modelling skewness in the data despite being a heavy-tailed alternative to the MN distribution. For this reason, a skew extension of the MLap distribution that is the MSLN distribution can be used neediness of modelling skewness and heavy-tailedness simultaneously. To have

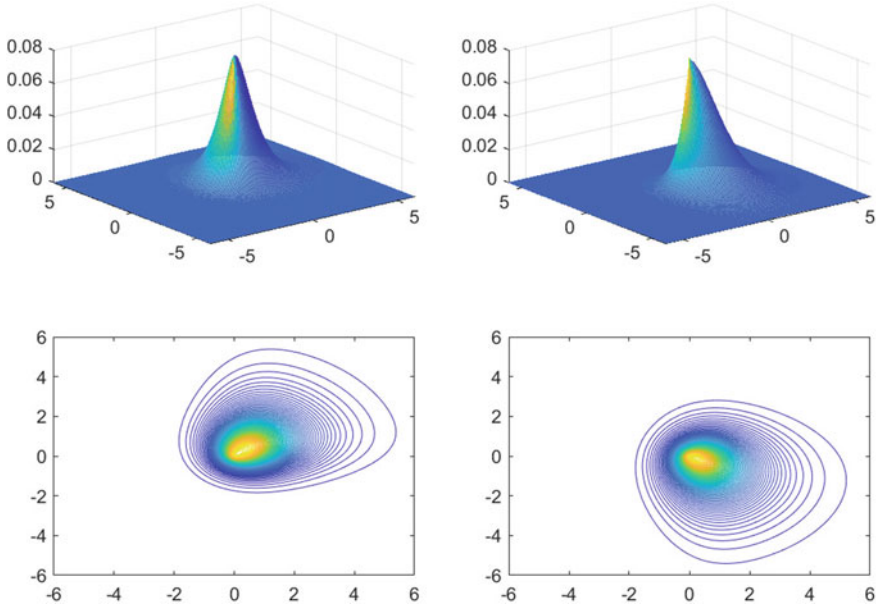


Fig. 1 Two examples of MSLN densities for $\boldsymbol{\mu} = (0, 0)^T$, $\Sigma = I$, $\boldsymbol{\lambda} = (1, 1)^T$ (left side) and $\boldsymbol{\mu} = (0, 0)^T$, $\Sigma = I$, $\boldsymbol{\lambda} = (1, -0.5)^T$ (right side)

some ideas about the shape of the MSLN distribution, we plot its pdf along with the contour plots for two-dimensional case. Figure 1 shows the plots for $\boldsymbol{\lambda} = (1, 1)^T$ and $\boldsymbol{\lambda} = (1, -0.5)^T$. These plots depict the peakedness, heavy-tailedness and skewness of the MSLN distribution.

Further, dividing (4) by (5) yields the following conditional density function of γ given $\boldsymbol{Y}$ and V:

$$f(\gamma|\boldsymbol{y}, v) = \frac{1}{\sqrt{2\pi}\,\Phi\left(\boldsymbol{\lambda}^T\boldsymbol{u}\right)} \exp\left\{-\frac{1}{2}\left(\gamma - \boldsymbol{\lambda}^T\boldsymbol{u}\right)^2\right\}. \tag{8}$$

According to this density function, it is obvious that γ and V are conditionally independent; therefore, the distribution of $\gamma|\boldsymbol{Y}$ is given by

$$\gamma|\boldsymbol{y} \sim TN\left(\boldsymbol{\lambda}^T\boldsymbol{u}, 1; (0, \infty)\right). \tag{9}$$

Also, dividing (5) by (6) we have the following conditional density function:

$$f(v|\boldsymbol{y}) = \frac{1}{\sqrt{2\pi}} v^{-\frac{3}{2}} \exp\left\{\frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} - \frac{1}{2}\left(v\boldsymbol{u}^T\boldsymbol{u} + v^{-1}\right)\right\}. \tag{10}$$

Thus, the following is another scale mixture representation of the MSLN distribution

$$\begin{aligned} \boldsymbol{Y}|v &\sim MSN_p\left(\boldsymbol{\mu}, v^{-1}\Sigma, v^{-\frac{1}{2}}\boldsymbol{\lambda}\right), \\ v &\sim g\left(v\right). \end{aligned} \tag{11}$$

Let $\boldsymbol{Y} \sim MSLN_p\left(\boldsymbol{\mu}, \Sigma, \boldsymbol{\lambda}\right)$. Then, the expectation and variance of $\boldsymbol{Y}$ are

$$\begin{aligned} E\left(\boldsymbol{Y}\right) &= \mu + \sqrt{\frac{2}{\pi}}\,\Sigma^{\frac{1}{2}}\boldsymbol{\lambda}\zeta, \\ Cov\left(\boldsymbol{Y}\right) &= \left(p+1\right)\Sigma - \frac{2}{\pi}\Sigma^{\frac{1}{2}}\boldsymbol{\lambda}\boldsymbol{\lambda}^T\Sigma^{\frac{1}{2}}\zeta^2, \end{aligned}$$

where $\zeta = E\left(\frac{v^{-\frac{1}{2}}}{\sqrt{v+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}\right)$ can be calculated using numerical methods.

Proposition 1 *By the help of the hierarchical representation given in (3), we obtain the following conditional expectations:*

$$\begin{aligned} E\left(V|\boldsymbol{y}\right) &= \left(\left(\boldsymbol{y}-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right)^{-\frac{1}{2}}, \\ E\left(\gamma|\boldsymbol{y}\right) &= \boldsymbol{\lambda}^T\boldsymbol{u} + \frac{\phi\left(\boldsymbol{\lambda}^T\boldsymbol{u}\right)}{\Phi\left(\boldsymbol{\lambda}^T\boldsymbol{u}\right)}, \end{aligned} \tag{12}$$

where $\boldsymbol{u} = \Sigma^{-\frac{1}{2}}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)$.

Note that these conditional expectations will be used in the EM algorithm given in the next section.

3 Finite Mixtures of the MSLN Distributions

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be a p-dimensional random sample observed from a g-component mixture of MSLN distributions. The pdf of a g-component finite mixtures of MSLN distributions is given below:

$$f\left(\boldsymbol{y}|\boldsymbol{\Theta}\right) = \sum_{i=1}^{g} w_i\, f_i\left(\boldsymbol{y}; \boldsymbol{\mu}_i, \Sigma_i, \boldsymbol{\lambda}_i\right), \tag{13}$$

where w_i is the mixing probability with $\sum_{i=1}^{g} w_i = 1$, $0 \leq \pi_i \leq 1$, $f_i\left(\boldsymbol{y}; \boldsymbol{\mu}_i, \Sigma_i, \boldsymbol{\lambda}_i\right)$ shows the pdf of the ith component (pdf of the MSLN distribution) given in (6) and $\boldsymbol{\Theta} = \left(w_1, \ldots, w_g, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \Sigma_1, \ldots, \Sigma_g, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_g\right)^T$ is the unknown parameter vector.

ML Estimation

We can find the ML estimator of $\boldsymbol{\Theta}$ by maximizing the following log-likelihood function:

$$\ell\left(\boldsymbol{\Theta}\right) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} w_i f\left(\boldsymbol{y}; \boldsymbol{\mu}_i, \Sigma_i, \boldsymbol{\lambda}_i\right)\right). \tag{14}$$

However, since there is not an explicit maximizer of (14), we need a numerical algorithm to solve the equation given above. Thus, the EM algorithm proposed by [10] can be used to obtain the ML estimator of $\boldsymbol{\Theta}$. Here, we will carry out the following EM algorithm:

Let $\boldsymbol{Z}_j = \left(Z_{1j}, \ldots, Z_{gj}\right)^T$ be the latent variables with

$$Z_{ij} = \begin{cases} 1, & \text{if } jth \text{ observation is coming from the } i^{th} \text{ component} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where $j = 1, \ldots, n$ and $i = 1, \ldots, g$. To perform the steps of the EM algorithm, we will take advantage of the stochastic representation of the MSLN distribution given in (1). Then, the hierarchical representation for finite mixtures of MSLN distributions will be

$$\begin{aligned} \boldsymbol{Y}_j|\gamma_j, v_j, z_{ij} = 1 &\sim N_p\left(\boldsymbol{\mu}_i + \frac{\Sigma_i^{\frac{1}{2}}\boldsymbol{\lambda}_i\gamma_j}{v_j + \boldsymbol{\lambda}_i^T\boldsymbol{\lambda}_i}, \Sigma_i^{\frac{1}{2}}\left(v_j I_p + \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^T\right)^{-1}\Sigma_i^{\frac{1}{2}}\right), \\ \gamma_j|v_j, z_{ij} = 1 &\sim TN\left(0, \frac{v_j + \boldsymbol{\lambda}_i^T\boldsymbol{\lambda}_i}{v_j}; (0, \infty)\right), \\ v_j|z_{ij} = 1 &\sim g\left(v_j\right). \end{aligned} \tag{16}$$

Let $(\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{v}, \boldsymbol{z})$ be the complete data, where $\boldsymbol{y} = \left(\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T\right)^T$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$, $\boldsymbol{v} = (v_1, \ldots, v_n)$ and $\boldsymbol{z} = (z_1, \ldots, z_n)^T$. Using the hierarchical representation given above and ignoring the constants, the complete data log-likelihood function can be obtained as

$$\begin{aligned} \ell_c\left(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{v}, \boldsymbol{z}\right) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij} \Bigg\{ &\log w_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\left(3\log v_j + {v_j}^{-1}\right) \\ &- \frac{1}{2} v_j \left(\boldsymbol{y}_j - \boldsymbol{\mu}_i\right)^T {\Sigma_i}^{-1}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_i\right) \\ &- \frac{1}{2}\left[\gamma_j^2 - 2\gamma_j\boldsymbol{\beta}_i^T\left(\boldsymbol{y}_j - \boldsymbol{\mu}_i\right) + \boldsymbol{\beta}_i^T\left(\boldsymbol{y}_j - \boldsymbol{\mu}_i\right)\left(\boldsymbol{y}_j - \boldsymbol{\mu}_i\right)^T\boldsymbol{\beta}_i\right]\Bigg\}. \end{aligned} \tag{17}$$

Now we can find the ML estimators by maximizing the complete data log-likelihood function given in (16); however, this function includes latent variables. To get rid of this latency problem, we should take the conditional expectation of the complete data log-likelihood function given the observed data $\boldsymbol{y}_j$

$$E\left(\ell_c\left(\boldsymbol{\Theta};\boldsymbol{y},\boldsymbol{\gamma},v,z\right)|\boldsymbol{y}_j\right)=\sum_{j=1}^{n}\sum_{i=1}^{g}E\left(Z_{ij}|\boldsymbol{y}_j\right)\left\{\log w_i-\frac{1}{2}\log|\Sigma_i|-\frac{3}{2}E\left(\log V_j|\boldsymbol{y}_j\right)\right.$$
$$-\frac{1}{2}E\left({V_j}^{-1}|\boldsymbol{y}_j\right)-\frac{1}{2}E\left(V_j|\boldsymbol{y}_j\right)\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)^T{\Sigma_i}^{-1}\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)-\frac{1}{2}E\left(\gamma_j^2|\boldsymbol{y}_j\right)$$
$$\left.+E\left(\gamma_j|\boldsymbol{y}_j\right)\boldsymbol{\beta}_i^T\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)-\frac{1}{2}\boldsymbol{\beta}_i^T\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)^T\boldsymbol{\beta}_i\right\}. \tag{18}$$

Since some conditional expectations are not related to the parameters, we only compute the conditional expectations $E\left(V_i|\boldsymbol{y}_i\right)$ and $E\left(\gamma_i|\boldsymbol{y}_i\right)$ with the help of Proposition 1. On the other hand, the conditional expectation $E\left(Z_{ij}|\boldsymbol{y}_j\right)$ can be calculated using the classical theory of mixture modelling.

Now, we can perform the EM algorithm using the following steps:

EM algorithm:
1. Take initial parameter estimate $\boldsymbol{\Theta}^{(0)}$ and a stopping rule Δ.
2. E-Step: Calculate the following conditional expectations for $k=0,1,2,\ldots$ iteration

$$\hat{z}_{ij}^{(k)}=E\left(Z_{ij}|y_j,\hat{\boldsymbol{\Theta}}^{(k)}\right)=\frac{\hat{\pi}_i^{(k)}f\left(y_j;\hat{\boldsymbol{\mu}}_i^{(k)},\hat{\Sigma}_i^{(k)},\hat{\boldsymbol{\lambda}}_i^{(k)}\right)}{f\left(y_j|\hat{\boldsymbol{\Theta}}^{(k)}\right)}, \tag{19}$$

$$\hat{v}_{ij}^{(k)}=E\left(V_j|y_j\right)=\left(\left(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_i^{(k)}\right)^T\hat{\Sigma}_i^{(k)-1}\left(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_i^{(k)}\right)\right)^{-\frac{1}{2}}, \tag{20}$$

$$\hat{\gamma}_{ij}^{(k)}=E\left(\gamma_j|\boldsymbol{y}_j\right)=\hat{\boldsymbol{\lambda}}_i^{(k)T}\boldsymbol{u}_{ij}+\frac{\phi\left(\hat{\boldsymbol{\lambda}}_i^{(k)T}\boldsymbol{u}_{ij}\right)}{\Phi\left(\hat{\boldsymbol{\lambda}}_i^{(k)T}\boldsymbol{u}_{ij}\right)}, \tag{21}$$

where $\boldsymbol{u}_{ij}=\hat{\Sigma}_i^{(k)-\frac{1}{2}}\left(\boldsymbol{y}_j-\hat{\boldsymbol{\mu}}_i^{(k)}\right)$. Using these conditional expectations, the objective function can be formed as follows:

$$Q\left(\boldsymbol{\Theta};\hat{\boldsymbol{\Theta}}^{(k)}\right)=\sum_{j=1}^{n}\sum_{i=1}^{g}\hat{z}_{ij}^{(k)}\left\{\log w_i-\frac{1}{2}\log|\Sigma_i|-\frac{1}{2}\hat{v}_{ij}^{(k)}\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)^T{\Sigma_i}^{-1}\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)\right.$$
$$\left.+\hat{\gamma}_{ij}^{(k)}\boldsymbol{\beta}_i^T\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)-\frac{1}{2}{\boldsymbol{\beta}_i}^T\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)^T\boldsymbol{\beta}_i\right\}. \tag{22}$$

3. M-step: Maximize the $Q\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(k)}\right)$ concerning $\boldsymbol{\Theta}$ to get the $(k+1)\,th$ parameter estimates for the parameters of interest. This maximization produces the following updating equations:

$$\hat{w}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}}{n}\,, \tag{23}$$

$$\hat{\boldsymbol{\mu}}_i^{(k+1)} = \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \left(\hat{v}_{ij}^{(k)} \hat{\Sigma}_i^{(k)-1} + \hat{\boldsymbol{\beta}}_i^{(k)} \hat{\boldsymbol{\beta}}_i^{(k)T}\right)\right)^{-1} \times \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \left(\hat{v}_{ij}^{(k)} \hat{\Sigma}_i^{(k)-1} \boldsymbol{y}_j - \hat{\gamma}_{ij}^{(k)} \hat{\boldsymbol{\beta}}_i^{(k)} + \hat{\boldsymbol{\beta}}_i^{(k)} \hat{\boldsymbol{\beta}}_i^{(k)T} \boldsymbol{y}_j\right)\right), \tag{24}$$

$$\hat{\Sigma}_i^{(k+1)} = \frac{1}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}} \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{v}_{ij}^{(k)} \left(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k)}\right)\left(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k)}\right)^T\right), \tag{25}$$

$$\hat{\boldsymbol{\beta}}_i^{(k+1)} = \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k)})(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k)})^T\right)^{-1} \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \hat{\gamma}_{ij}^{(k)} (\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(k)})\right), \tag{26}$$

$$\hat{\boldsymbol{\lambda}}_i^{(k+1)} = \left(\hat{\Sigma}_i^{(k+1)}\right)^{\frac{1}{2}} \hat{\boldsymbol{\beta}}_i^{(k+1)}. \tag{27}$$

4. Repeat E and M steps until the convergence rule $\|\hat{\boldsymbol{\Theta}}^{(k+1)} - \hat{\boldsymbol{\Theta}}^{(k)}\| < \Delta$ is satisfied. Otherwise, the absolute difference of the actual log-likelihood $\|\ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)})\| < \Delta$ or $\|\ell(\hat{\boldsymbol{\Theta}}^{(k+1)})/\ell(\hat{\boldsymbol{\Theta}}^{(k)}) - 1\| < \Delta$ can be used as a stopping criterion (see [12]). Otherwise, the Aitken acceleration-based criterion given in [32] $\|\ell^{(k+1)} - \ell_\infty^{(k+1)}\| < \Delta$ can also be used as a stopping criterion, where $\ell^{(k+1)}$ is the observed log-likelihood for $\hat{\boldsymbol{\Theta}}^{(k+1)}$ and $\ell_\infty^{(k+1)}$ is the asymptotic estimate of the log-likelihood at $k+1$ iteration

$$\ell_\infty^{(k+1)} = \ell^{(k)} + \frac{1}{1-\varepsilon^{(k)}} \left(\ell^{(k+1)} - \ell^{(k)}\right).$$

Here, $\varepsilon^{(k)}$ shows the Aitken's acceleration at k iteration, where $\varepsilon^{(k)} = \frac{\ell^{(k+1)} - \ell^{(k)}}{\ell^{(k)} - \ell^{(k-1)}}$.

Note that estimators obtained in the EM algorithm will be robust against the outliers in the data. For instance, weight $\hat{v}_{ij}$ depends on Mahalanobis distance between $\boldsymbol{y}_i$ and $\boldsymbol{\mu}$ for each component and this weight function will be a decreasing function of $\left(\boldsymbol{y}_j - \boldsymbol{\mu}_i\right)^T \Sigma_i{}^{-1} \left(\boldsymbol{y}_j - \boldsymbol{\mu}_i\right)$. Therefore, outliers are down-weighted by the related weights and this makes estimators robust in terms of outliers.

Initial Values

It is a challenging problem to provide convergence in the EM algorithm. There is a way based on the k-means clustering algorithm given by [24] summarized as follows. This method is used by [27–29] and [15].

Steps of initialization:
(i) Perform the K-means clustering algorithm and partition into g groups.
(ii) Set the component labels $\hat{z}_j^{(0)} = \left\{z_{ij}\right\}_{i=1}^{g}$ according to the K-means clustering results.
(iii) The initial values of mixing probabilities, component locations and component scale variances can be initialized as

$$\hat{w}_i^{(0)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(0)}}{n}, \qquad \hat{\mu}_i^{(0)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(0)} y_j}{\sum_{j=1}^{n} \hat{z}_{ij}^{(0)}},$$

$$\hat{\Sigma}_i^{(0)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(0)} \left(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(0)}\right)\left(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_i^{(0)}\right)^T}{\sum_{j=1}^{n} \hat{z}_{ij}^{(0)}}.$$

(iv) For the skewness parameters, use the skewness coefficient vector of each clusters.

The Empirical Information Matrix

The standard error of an estimator has a huge importance to explore statistical inferences. The standard errors of ML estimators of mixture model parameters can be obtained based on the information-based method given by [8]. Hence, we will obtain the inverse of the empirical information matrix to approximate the asymptotic covariance matrix of estimators. This information matrix can be given as

$$\hat{I}_e = \sum_{j=1}^{n} \hat{\boldsymbol{s}}_j \hat{\boldsymbol{s}}_j^T. \tag{28}$$

Here, the individual scores are computed by $\hat{\boldsymbol{s}}_j = E_{\hat{\boldsymbol{\Theta}}}\left(\frac{\partial \ell_{cj}(\boldsymbol{\Theta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_j, \boldsymbol{v}_j, \boldsymbol{z}_j)}{\partial \boldsymbol{\Theta}} | \boldsymbol{y}_j\right)$, $j = 1, \ldots, n$ and the complete data log-likelihood function for the jth observation is represented as $\ell_{cj}\left(\boldsymbol{\Theta}; \boldsymbol{y}_j, \boldsymbol{\gamma}_j, \boldsymbol{v}_j, \boldsymbol{z}_j\right)$. Let define $S = \Sigma^{\frac{1}{2}}$ and $\boldsymbol{\Theta} = \left(w_i, \boldsymbol{\mu}_i, S_i, \boldsymbol{\lambda}_i\right)$ be the parameter vector. After ignoring constants, we obtain as

$$\ell_{cj}\left(\boldsymbol{\Theta};\boldsymbol{y}_j,\boldsymbol{\gamma}_j,\boldsymbol{v}_j,\boldsymbol{z}_j\right)=\sum_{i=1}^{g}z_{ij}\left\{\log w_i-\log|S_i|-\frac{1}{2}\left(v_j\boldsymbol{u}_{ij}^T\boldsymbol{u}_{ij}+\left(\gamma_j-\boldsymbol{\lambda}_i^T\boldsymbol{u}_{ij}\right)^2\right)\right\}, \tag{29}$$

where $\boldsymbol{u}_{ij}=S_i^{-1}\left(\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)\right)$. The individual score vector for jth observation $\hat{\boldsymbol{s}}_j=\left(\hat{\boldsymbol{s}}_{j,w_1},\ldots,\hat{\boldsymbol{s}}_{j,w_{g-1}},\hat{\boldsymbol{s}}_{j,\boldsymbol{\mu}_1},\ldots,\hat{\boldsymbol{s}}_{j,\boldsymbol{\mu}_g},\hat{s}_{j,s_1},\ldots,\hat{\boldsymbol{s}}_{j,s_g},\hat{\boldsymbol{s}}_{j,\boldsymbol{\lambda}_1},\ldots,\hat{s}_{j,\boldsymbol{\lambda}_g}\right)^T$, where $s_i=vech\left(S_i\right)$. To obtain the closed forms for $\hat{s}_j$, we take the derivation of ℓ_{cj} according to the parameters and the derivation yields as

$$\hat{\boldsymbol{s}}_{j,w_r}=\frac{\hat{z}_{rj}}{\hat{w}_r}-\frac{\hat{z}_{gj}}{\hat{w}_g},\ r=1,\ldots,g-1, \tag{30}$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\mu}_i}=\hat{z}_{ij}\left(\hat{v}_{ij}\hat{S}_i^{-1}\hat{\boldsymbol{u}}_{ij}-\left(\hat{\gamma}_{ij}-\hat{\boldsymbol{\lambda}}_i^T\hat{\boldsymbol{u}}_{ij}\right)\hat{S}_i^{-1}\hat{\boldsymbol{\lambda}}_i\right), \tag{31}$$

$$\begin{aligned}\hat{\boldsymbol{s}}_{j,s_i}=&\,vech\left\{-\hat{z}_{ij}\left(2\hat{S}_i^{-1}-Diag\left(\hat{S}_i^{-1}\right)\right)+\hat{z}_{ij}\hat{v}_{ij}\left(\hat{B}_{ij}+\hat{B}_{ij}^T-Diag\left(\hat{B}_{ij}\right)\right)\right.\\&\left.-\,\hat{z}_{ij}\left(\hat{\gamma}_{ij}-\hat{\lambda}_i^T\hat{u}_{ij}\right)\left(\hat{C}_{ij}+\hat{C}_{ij}^T-Diag\left(\hat{C}_{ij}\right)\right)\right\},\end{aligned} \tag{32}$$

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\lambda}_i}=\hat{z}_{ij}\left(\hat{\gamma}_{ij}-\hat{\boldsymbol{\lambda}}_i^T\hat{\boldsymbol{u}}_{ij}\right)\hat{\boldsymbol{u}}_{ij}, \tag{33}$$

where $\hat{S}_i=\hat{\Sigma}_i^{\frac{1}{2}}$, $\hat{u}_{ij}=\hat{S}_i^{-1}\left(\left(y_j-\hat{\mu}_i\right)\right)$, $\hat{B}_{ij}=\hat{S}_i^{-1}\hat{u}_{ij}\hat{u}_{ij}^T$, $\hat{C}_{ij}=\hat{u}_{ij}\hat{\lambda}_i^T\hat{S}_i^{-1}$ and $\hat{z}_{ij}$, $\hat{v}_{ij}$ and $\hat{\gamma}_{ij}$ can be calculated by using the equations given in (19)–(21) evaluated at $\hat{\boldsymbol{\Theta}}$. At this point, the standard errors of the estimates can be found by using the square roots of the diagonal elements of $\hat{I}_e^{-1}$.

4 Applications

This part comprises a simulation study and a real data example to demonstrate the performance of estimators for parameters of the proposed multivariate mixture model. The EM algorithm given in Sect. 3 is used to compute parameter estimates. The computational details can be summarized as follows:

Notes on computation:

(i) The simulation study and real data example are conducted by MATLAB R2017b software.

(ii) The stopping rule is set as 10^{-6} for all numerical computations.

(iii) For the simulation study, the replication number is taken as $N=500$ and the sample sizes (n) are, respectively, taken as 250, 500, 1000, and 2000 for all simulation scenarios.

(iii) A random sample from MSLN distribution can be generated using the following steps:

–Sample $U_1\sim N\left(0,1\right)$, $\boldsymbol{U}_2\sim N_p\left(0,I_p\right)$ and V from the inverse gamma distribu-

tion given in (2) independently. Note that a random sample from the inverse gamma distribution can be generated by applying the relation $\frac{1}{V} \sim Gamma\left(\frac{(p+1)}{2}, 2\right)$. Then, $\boldsymbol{Y}=\boldsymbol{\mu}+\Sigma^{1/2}\left[\frac{\boldsymbol{\lambda}|U_1|}{\sqrt{V\left(V+\boldsymbol{\lambda}^T\boldsymbol{\lambda}\right)}}+\left(VI_p+\boldsymbol{\lambda}\boldsymbol{\lambda}^T\right)^{-\frac{1}{2}}\boldsymbol{U}_2\right]$ generates the sample from the $MSLN_p\left(\boldsymbol{\mu},\ \Sigma, \boldsymbol{\lambda}\right)$ distribution.

(iv) We note that in the EM algorithm numerical uncertainty could occur when the weights $(\hat{z}_{ij}^{(k)})$ go to zero. One can put a hard threshold on $\hat{z}_{ij}^{(k)}$ for kth iteration, which was offered by [47], to prevent overflow in the computation of $\hat{z}_{ij}^{(k)}$. Let ϵ be a predetermined small value. Hence, if $\hat{z}_{ij}^{(k)} > \epsilon$, $\hat{z}_{ij}^{(k)}$ will be used for the next iteration; else, ϵ will be used instead of $\hat{z}_{ij}^{(k)}$. In the computation, we will take 10^{-6} for ϵ.

Simulation Study

This simulation study is performed to illustrate the performance of finite mixtures of MSLN distributions. The performance of estimators will be measured by the bias, standard errors (SEs) and the mean Euclidean distance values. The formula of bias can be given with the following formula:

$$\widehat{bias}\left(\hat{\theta}\right) = \overline{\theta} - \theta,$$

where θ is the true parameter value, $\overline{\theta} = \frac{1}{N}\sum_{j=1}^{N}\hat{\theta}_j$ and $\hat{\theta}_j$ is the estimate of θ for the jth simulated data. The mean Euclidean distances of the estimators are calculated using the average of the Euclidean norm between the estimates and the true parameter values. Such as, for the mean Euclidean distance of $\hat{\boldsymbol{\mu}}_i$ will be given as

$$\|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\| = \frac{1}{N}\left(\sum_{j=1}^{N}\left(\hat{\mu}_{ij} - \mu_{ij}\right)^2\right)^{\frac{1}{2}}.$$

Furthermore, the other mean Euclidean distances for the other estimators are also computed similarly. Otherwise, for the mixing probability estimator $(\hat{w})$ the distance will be the mean squared error (MSE) which is given with the following formula:

$$\widehat{MSE}\left(\hat{w}\right) = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{w}_j - w\right)^2,$$

where w is the true parameter value and $\hat{w}_j$ is the estimate of w for the jth simulated data. Note that the SEs of estimates are computed by using the empirical information matrix of the proposed mixture model given in Sect. 3.

This section pays attention to model heterogeneity in multivariate settings diversely using two simulation scenarios.

Scenario 1. We generate the data from the following two-component mixtures of MSLN distributions:

$$f(\boldsymbol{y}_j|\boldsymbol{\Theta}) = w_1 f_1(\boldsymbol{y}_j; \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\lambda}_1) + (1 - w_1) f_2(\boldsymbol{y}_j; \boldsymbol{\mu}_2, \Sigma_2, \boldsymbol{\lambda}_2), \tag{34}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2})^T$, $\Sigma_i = \begin{bmatrix} \sigma_{i,11} & \sigma_{i,12} \\ \sigma_{i,21} & \sigma_{i,22} \end{bmatrix}$, $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2})^T$, $i = 1, 2$ with the parameter values

$$\boldsymbol{\mu}_1 = (2, 3)^T , \ \boldsymbol{\mu}_2 = (0, 2)^T , \ \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\lambda}_1 = (1, 1)^T , \ \boldsymbol{\lambda}_2 = (1, -1)^T , \ \pi_1 = 0.6.$$

The scatter-contour plot of the simulated data set example with 1000 random samples generated from the model given in (34) is displayed in Fig. 2. Different colours indicate the separate groups of clusters in this figure. This scenario scheme is an example of two moderately separated clusters for the two-component mixture model case.

Table 1 summarizes the simulation results of Scenario 1 for the sample sizes 250, 500, 1000 and 2000. This table includes the true parameter values, bias, standard errors and the mean Euclidean distance values of estimates. We can observe from this table that the proposed EM algorithm is working accurately to estimate all parameters. Moreover, the mean Euclidean distances for all parameter estimates are getting smaller when the sample size n increases. In addition to that, the centre,

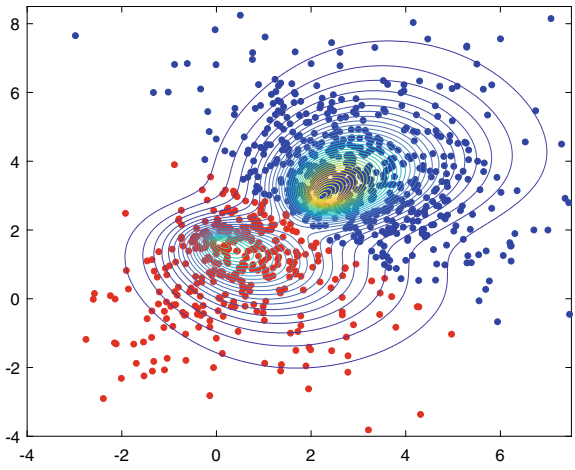


Fig. 2 Scatter-contour plot of the simulated data set with $n = 1000$ generated from model (34) for Scenario 1

Table 1 Bias, SEs and mean Euclidean distance values of estimates for n = 250, 500, 1000 and 2000

n	Parameter	Component 1				Component 2			
		True	Bias	SE	Distance	True	Bias	SE	Distance
250	w_i	0.6	0.0318	0.0550	0.0246	0.4	–	–	–
	μ_{i1}	2	0.2003	0.2307	0.4619	0	0.0513	0.3403	0.6232
	μ_{i2}	3	−0.0834	0.2073		2	−0.2208	0.2253	
	$\sigma_{i,11}$	1	−0.0211	0.1527		1	0.3555	0.2081	
	$\sigma_{i,12}$	0	0.1602	0.0822	0.3691	0	−0.1948	0.1231	0.6012
	$\sigma_{i,22}$	1	0.1350	0.1313		1	−0.1138	0.1409	
	λ_{i1}	1	−0.3733	0.4961	1.0662	1	0.2649	1.2829	1.1148
	λ_{i2}	1	0.2053	0.5359		−1	−0.3465	0.7758	
500	w_i	0.6	0.0464	0.0389	0.0167	0.4	–	–	–
	μ_{i1}	2	0.2015	0.1597	0.3765	0	0.0285	0.2324	0.5023
	μ_{i2}	3	−0.1136	0.1382		2	−0.2197	0.1470	
	$\sigma_{i,11}$	1	−0.0573	0.1060		1	0.2672	0.2081	
	$\sigma_{i,12}$	0	0.1219	0.0554	0.2646	0	−0.2113	0.0813	0.4577
	$\sigma_{i,22}$	1	0.1163	0.0876		1	−0.1215	0.0955	
	λ_{i1}	1	−0.4901	0.2532	0.8795	1	0.1396	0.6691	0.8499
	λ_{i2}	1	0.0812	0.2917		−1	−0.3259	0.4227	
1000	w_i	0.6	0.0463	0.0264	0.0080	0.4	–	–	–
	μ_{i1}	2	0.1714	0.1032	0.3102	0	0.0122	0.1558	0.3448
	μ_{i2}	3	−0.1489	0.0893		2	−0.2077	0.0918	
	$\sigma_{i,11}$	1	−0.0856	0.0624		1	0.2203	0.0828	
	$\sigma_{i,12}$	0	0.1105	0.0345	0.2090	0	−0.2163	0.0512	0.3586
	$\sigma_{i,22}$	1	0.1170	0.0580		1	−0.1436	0.0616	
	λ_{i1}	1	−0.5157	0.1303	0.7116	1	−0.0064	0.3272	0.5468
	λ_{i2}	1	0.1095	0.1443		−1	−0.2551	0.1977	
2000	w_i	0.6	0.0420	0.0181	0.0037	0.4	–	–	–
	μ_{i1}	2	0.1549	0.0689	0.2731	0	0.0397	0.1054	0.2725
	μ_{i2}	3	−0.1854	0.0603		2	−0.1975	0.0593	
	$\sigma_{i,11}$	1	−0.1002	0.0428		1	0.2116	0.0545	
	$\sigma_{i,12}$	0	0.1081	0.0237	0.1885	0	−0.1980	0.0333	0.3168
	$\sigma_{i,22}$	1	0.1224	0.0402		1	−0.1615	0.0402	
	λ_{i1}	1	−0.5093	0.0771	0.6129	1	−0.1137	0.1757	0.3843
	λ_{i2}	1	0.1368	0.0864		−1	−0.1898	0.0924	

scale and skewness of all components are adequately captured by the finite mixture of MSLN distributions.

Scenario 2. We generate the data from the following three-component mixtures of MSLN distributions:

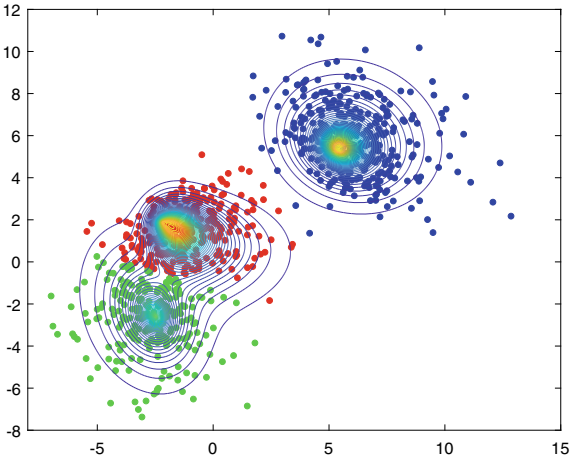


Fig. 3 Scatter-contour plot of the simulated data set with $n = 1000$ generated from model (35) for Scenario 2

$$f(\boldsymbol{y}_j|\boldsymbol{\Theta}) = w_1 f_1(\boldsymbol{y}_j; \boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\lambda}_1) + w_2 f_2(\boldsymbol{y}_j; \boldsymbol{\mu}_2, \Sigma_2, \boldsymbol{\lambda}_2) + (1 - (w_1 + w_2)) f_3(\boldsymbol{y}_j; \boldsymbol{\mu}_3, \Sigma_3, \boldsymbol{\lambda}_3), \tag{35}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2})^T$, $\Sigma_i = \begin{bmatrix} \sigma_{i,11} & \sigma_{i,12} \\ \sigma_{i,21} & \sigma_{i,22} \end{bmatrix}$, $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2})^T$, $i = 1, 2, 3$ with the parameter values

$$\boldsymbol{\mu}_1 = (5, 5)^T , \ \boldsymbol{\mu}_2 = (-2, 2)^T , \ \boldsymbol{\mu}_3 = (-2, -2)^T , \ \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\lambda}_1 = (1, 1)^T , \ \boldsymbol{\lambda}_2 = (1, -1)^T , \ \boldsymbol{\lambda}_3 = (-1, -1)^T , \ \pi_1 = 0.4, \ \pi_2 = 0.3.$$

The scatter-contour plot of the simulated data set example with 1000 random samples generated from the model given in (35) is demonstrated in Fig. 3. This figure displays the separated groups of clusters in different colours. For the three-component mixture model case, this scenario is an example of two moderately separated clusters and a well-separated cluster.

The simulation results of Scenario 2 are given in Table 2 for the sample sizes 250, 500, 1000 and 2000. This table consists of the true parameter values, bias, standard errors and the mean Euclidean distance values of estimates. According to this table, the proposed EM algorithm can be used for estimating all parameters. Furthermore, the mean Euclidean distances for all parameter estimates are decreasing while the sample size n are increasing. Similar to Scenario 1, the centre, scale and skewness of all three components are successfully catch.

Table 2 Bias, SEs and mean Euclidean distance values of estimates for n = 250, 500, 1000 and 2000

n	Parameter	Component 1				Component 2				Component 3			
		True	Bias	SE	Dist.	True	Bias	SE	Dist.	True	Bias	SE	Dist.
250	w_i	0.4	0.0116	0.0235	0.0011	0.3	0.0076	0.0253	0.0036	0.3	–	–	–
	μ_{i1}	5	0.4768	0.2713	0.6683	−2	0.0399	0.2594	0.4726	−2	−0.4537	0.3318	0.8936
	μ_{i2}	5	0.2904	0.2751		2	0.1358	0.2868		−2	−0.1657	0.4741	
	$\sigma_{i,11}$	1	0.0561	0.1309		1	0.1752	0.2099		1	−0.2429	0.1612	
	$\sigma_{i,12}$	0	−0.1788	0.0773	0.2831	0	−0.3472	0.1381	0.7902	0	−0.1134	0.1130	0.6335
	$\sigma_{i,22}$	1	0.0006	0.1230		1	0.1865	0.1979		1	0.4830	0.2247	
	λ_{i1}	1	−0.8520	0.2263	1.1380	1	0.9621	1.6858	1.5505	−1	0.8124	0.3813	1.4238
	λ_{i2}	1	−0.6944	0.2705		−1	−0.6545	1.4930		−1	1.0967	0.3969	
500	w_i	0.4	0.0115	0.0163	0.0006	0.3	0.0165	0.0180	0.0020	0.3	–	–	–
	μ_{i1}	5	0.4481	0.1828	0.6104	−2	−0.0053	0.1663	0.3656	−2	−0.4544	0.2215	0.7338
	μ_{i2}	5	0.2937	0.1865		2	0.0900	0.1913		−2	−0.2757	0.3080	
	$\sigma_{i,11}$	1	0.0463	0.0888		1	0.1557	0.1410		1	−0.2405	0.1082	
	$\sigma_{i,12}$	0	− 0.1726	0.0517	0.2302	0	−0.3262	0.0900	0.5940	0	−0.1442	0.0738	0.5008
	$\sigma_{i,22}$	1	−0.0078	0.0819		1	0.0686	0.1286		1	0.3980	0.1463	
	λ_{i1}	1	−0.8564	0.1336	1.1330	1	0.8219	0.8612	1.2087	−1	0.8168	0.2096	1.3940
	λ_{i2}	1	−0.7166	0.1627		−1	−0.4228	0.7658		−1	1.1044	0.2023	

(continued)

Table 2 (continued)

n	Parameter	Component 1				Component 2				Component 3			
		True	Bias	SE	Dist.	True	Bias	SE	Dist.	True	Bias	SE	Dist.
1000	w_i	0.4	0.016	0.0114	0.0004	0.3	0.0242	0.0127	0.0014	0.3	–	–	–
	μ_{i1}	5	0.4448	0.1260	0.5712	−2	−0.0144	0.1127	0.3007	−2	−0.4555	0.1527	0.6447
	μ_{i2}	5	0.3030	0.1277		2	0.0897	0.1287		−2	−0.3364	0.2114	
	$\sigma_{i,11}$	1	0.0465	0.0617		1	0.1276	0.0946		1	−0.2358	0.0749	
	$\sigma_{i,12}$	0	− 0.1756	0.0350	0.2105	0	−0.3228	0.0611	0.4972	0	−0.1766	0.0511	0.4422
	$\sigma_{i,22}$	1	−0.0132	0.0555		1	0.0344	0.0886		1	0.3589	0.1011	
	λ_{i1}	1	−0.8505	0.0845	1.1317	1	0.7039	0.5082	0.9845	−1	0.8191	0.1334	1.3814
	λ_{i2}	1	−0.7351	0.1016		−1	−0.3311	0.4419		−1	1.1018	0.1289	
2000	w_i	0.4	0.0117	0.0080	0.0002	0.3	0.0284	0.0089	0.0011	0.3	–	–	–
	μ_{i1}	5	0.4330	0.0871	0.5424	−2	−0.0752	0.0773	0.2162	−2	−0.4451	0.1043	0.6186
	μ_{i2}	5	0.3045	0.0875		2	0.0221	0.0864		−2	−0.3846	0.1440	
	$\sigma_{i,11}$	1	0.0346	0.0428		1	0.1782	0.0676		1	−0.2182	0.0521	
	$\sigma_{i,12}$	0	− 0.1732	0.0242	0.1955	0	−0.3002	0.0413	0.4339	0	−0.1930	0.0353	0.3990
	$\sigma_{i,22}$	1	−0.0060	0.0385		1	−0.0551	0.0585		1	0.3169	0.0698	
	λ_{i1}	1	−0.8491	0.0557	1.1307	1	0.6934	0.3254	0.8346	−1	0.8167	0.0839	1.3701
	λ_{i2}	1	−0.7429	0.0655		−1	−0.1375	0.2468		−1	1.0942	0.0837	

An Illustrative Real Data Example: Old Faithful Geyser Data Set

The geyser eruptions data sets in different USA parks have been collected by park geologists. The first national park in Wyoming, USA was the Yellowstone National Park formed in 1872. The Old Faithful and a collection of the world's most splendid geysers and hot water resources are located in the Yellowstone National Park. The Old Faithful geyser data set includes 272 pairs of measurements, referring to the time interval between the starts of successive eruptions and the duration of the subsequent eruption. This data set was analysed by [5, 11, 43] and many others. One can see [41] for the details of the Old Faithful data set. Additionally, this data set was used by [39] to present the applicability of the mixsmsn package in R for the multivariate case. This data set can be accessed by using data ("faithful") in R. We consider the Old Faithful geyser data set as a real data example to show the applicability of finite mixtures of MSLN distributions (FM-MSLN) and compare results with finite mixtures of MSTN distributions (FM-MSTN) and finite mixtures of MSGLN distributions (FM-MSGLN). This comparison will be done with the following information criteria:

$$-2\ell(\hat{\boldsymbol{\Theta}}) + mc_n,$$

where $\ell(\cdot)$ represents the maximized log-likelihood, m is the number of free parameters to be estimated in the model and c_n is the penalty term. Here, it is taken as $c_n = 2$ for the Akaike information criteria (AIC) [1], $c_n = \log(n)$ for the Bayesian information criteria (BIC) [42] and $c_n = 0.2\sqrt{n}$ for the efficient determination criteria (EDC) [6].

The real data example fitting results for FM-MSGLN, FM-MSTN and FM-MSLN are summarized in Table 3. This table involves estimates, standard errors (SEs), log-likelihood, values of AIC, BIC and EDC. Here, the SEs of estimators are computed with the help of the Fisher information-based method given by [8], one can see Sect. 3 for more information about the computation steps of the SE for FM-MSLN. It can be seen from Table 3 that FM-MSLN enables the best fit as a consequence of owning the highest log-likelihood value and the lowest AIC, BIC and EDC values. Furthermore, CPU times (CT) in seconds are outlined in Table 4 to compare the computation times for FM-MSGLN, FM-MSTN and FM-MSLN. We notice that the computational time of FM-MSLN is less than the computational times of FM-MSTN and FM-MSGLN. Figure 4 displays scatter-contour plots of the data of the fitted two-component FM-MSGLN, FM-MSTN and FM-MSLN models. It can be observed from Fig. 4 that the proposed mixture model, FM-MSLN, captures bimodality and asymmetry perfectly and provides the best adequate fit to the Old Faithful geyser data set.

Table 3 ML estimation results of the Old Faithful geyser data set for FM-MSTN, FM-MSGLN and FM-MSLN

Component		FM-MSGLN		FM-MSTN		FM-MSLN	
		Estimate	SE	Estimate	SE	Estimate	SE
1	w_1	0.6559	0.3210	0.6198	0.0425	0.6415	0.0804
	μ_{11}	4.5330	0.0410	4.5801	0.1541	4.7950	0.0637
	μ_{12}	82.0000	0.4571	79.3601	3.2761	83.9052	0.9567
	$\sigma_{1,11}$	0.1554	0.3074	0.1814	0.1829	0.2711	0.1739
	$\sigma_{1,12}$	0.9985	0.1674	0.2700	0.1391	2.1047	0.1158
	$\sigma_{1,22}$	28.1861	2.6003	27.7556	0.7121	34.4311	1.0911
	λ_{11}	−0.6171	0.3745	−1.1448	0.8787	−1.8785	0.8926
	λ_{12}	−0.4691	0.3742	0.3490	1.1930	−1.8564	0.9841
	ν_1	–	–	15.7829	26.4154	–	–
	α_1	0.9949	0.0630	–	–	–	–
2	μ_{21}	1.8330	0.0224	1.7300	0.0337	1.7189	0.0393
	μ_{22}	54.0000	0.5820	51.2491	1.0709	50.7770	1.3785
	$\sigma_{2,11}$	0.0565	0.1158	0.1666	0.0954	0.1707	0.1461
	$\sigma_{2,12}$	0.2227	0.1251	1.5611	0.0724	1.9321	0.1154
	$\sigma_{2,22}$	26.5696	0.3989	38.3925	0.9161	45.7395	1.6290
	λ_{21}	1.3627	0.9609	4.3879	1.9344	4.1350	2.4470
	λ_{22}	0.0954	0.8491	3.7501	2.2122	4.3050	3.5075
	ν_2	–	–	4.4866	3.1462	–	–
	α_2	0.8614	0.0031	–	–	–	–
Information Criteria	$-\ell(\hat{\boldsymbol{\Theta}})$	−1305.0165		−1302.4555		**−1299.53845**	
	BIC	2705.3317		2700.2097		**2683.1639**	
	AIC	2644.0330		2638.9111		**2629.0769**	
	EDC	2666.1073		2660.9853		**2648.5542**	

Table 4 CPU time (CT) in seconds of the Old Faithful geyser data set for FM-MSGLN, FM-MSTN and FM-MSLN

		FM-MSGLN	FM-MSTN	FM-MSLN
CPU times in seconds	CT	290.2023	192.2511	**178.5986**

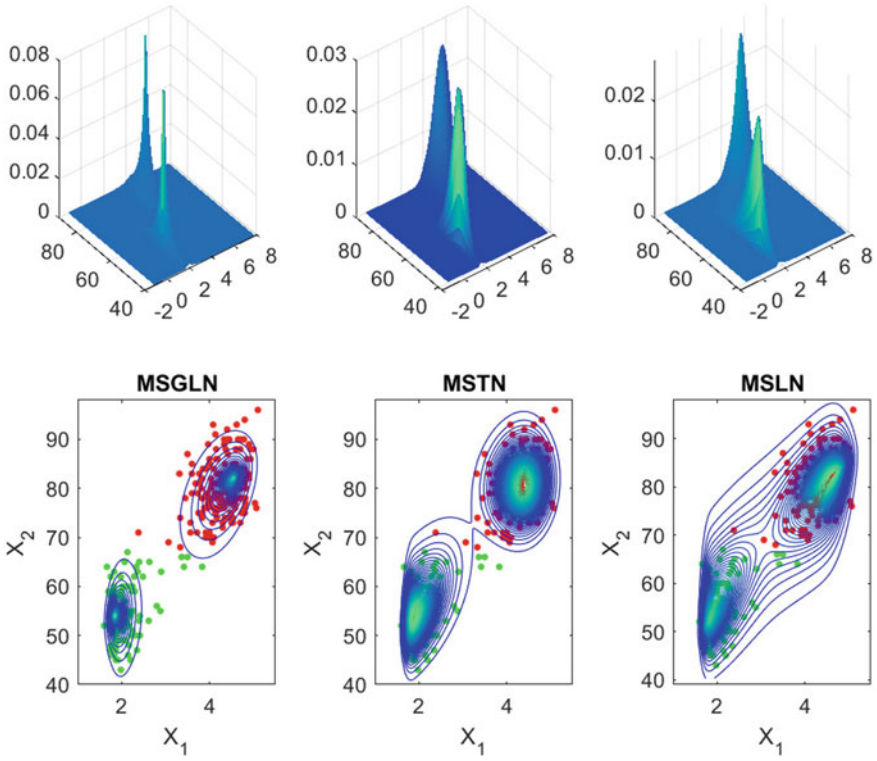


Fig. 4 Scatter-contour plots of the Old Faithful geyser data set of the fitted two-component FM-MSGLN, FM-MSTN and FM-MSLN models

5 Conclusions

This paper has proposed FM-MSLN, which can be used for modelling skewness and heavy-tailedness in multivariate heterogeneous data sets. We have proposed an EM algorithm to estimate the ML estimates for parameters of FM-MSLN. A simulation study has been conducted to assess the performance of the proposed mixture model and the results have proved that the proposed EM algorithm works accurately to estimate all parameters. In addition to that, to model both skewness and heavy-tailedness in the data, a real data example has been provided to indicate the applicability of the proposed mixture model and compare it with FM-MSTN and FM-MSGLN. It is seen that FM-MSLN has better performance than FM-MSTN and FM-MSGLN according to the information criteria. Also, the proposed mixture model has less computational time comparing with the other mixture models. Therefore, we can say that the FM-MSLN can be used as an alternative multivariate mixture model for heterogeneous skew and heavy-tailed data sets with the advantage of less computational time.

There is a significant number of topics that earn attention in future studies. As per one of the Reviewer's recommendations, accommodating missing values ([46]) in the FM-MSLN can be considered an interesting modelling approach for robust clustering of high-dimensional data settings. Further, selecting the optimal number of components can be one of the interesting future works. Another approach to our current study is to find the optimal number of components g adopting a model choice criteria instead of assuming the number of components is known.

Acknowledgements This study is supported by "Giresun University" (Grant number:FEN-BAP-A-270220-16) as a Type A project of Scientific Research and Development Projects. We thank two anonymous referees and the Editor, whose comments and recommendations have helped us to improve this paper extensively.

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. & Caski, F. (Eds.) *Proceeding of the Second International Symposium on Information Theory* (pp. 267–281) Budapest: Akademiai Kiado.
2. Anderson, D. N. (1992). A multivariate linnik distribution. *Statistics & Probability Letters, 14,* 333–336.
3. Arslan, O. (2010). An alternative multivariate skew Laplace distribution: Properties and estimation. *Statistical Papers, 51*(4), 865–887.
4. Azzalini, A., & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika, 83,* 715–726.
5. Azzalini, A., & Bowman, A. W. (1990). A look at some data on the Old Faithful Geyser. *Journal of the Royal Statistical Society: Series C Applied Statistics, 39,* 357–365.
6. Bai, Z. D., Krishnaiah, P. R., & Zhao, L. C. (1989). On rates of convergence of efficient detection criteria in signal processing with white noise. *IEEE Transactions on Information Theory, 35,* 380–388.
7. Banfield, J. D., & Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49,* 803–821.
8. Basford, K. E., Greenway, D. R., McLachlan, G. J., & Peel, D. (1997). Standard errors of fitted means under normal mixture. *Computational Statistics, 12,* 1–17.
9. Böhning, D. (2000). *Computer assisted analysis of mixtures and applications: Meta-analysis, disease mapping, and others*. London: Chapman and Hall/CRC.
10. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B Statistical Methodology, 39*(1), 1–38.
11. Denby, L., & Pregibon, D. (1987). An example of the use of graphics in regression. *American Statistician, 41,* 33–38.
12. Dias, J. G., & Wedel, M. (2004). An empirical comparison of EM, SEM and MCMC performance for problematic gaussian mixture likelihoods. *Statistics and Computing, 14,* 323–332.
13. Doğru, F. Z., & Arslan, O. (2019). Multivariate skew Laplace normal distribution: Properties and applications. In *Y-BIS Conference 2019-Recent Advances in Data Science and Business Analytics*, İstanbul/Turkey.
14. Doğru, F. Z., & Arslan, O. (2022). Multivariate skew Laplace normal distribution for modeling skewness and heavy-tailedness in multivariate data sets. *Statistics and Its Interface, 15*(4), 475–485.

15. Doğru, F. Z., & Arslan, O. (2021). Finite mixtures of multivariate skew Laplace distributions. *Revstat Statistical Journal, 19*(1), 35–46.
16. Ernst, M. D. (1998). A multivariate generalized Laplace distribution. *Computational Statistics, 13,* 227–232.
17. Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman and Hall.
18. Fang, K. T., Kotz, S., N.g., & K. W. (1990). *Symmetric multivariate and related distributions*. London: Chapman and Hall.
19. Fraley, C., & Raftery, A. E. (1999). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578–588 (1999)
20. Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. London: Springer.
21. Gómez-Sánchez-Manzano, E., Gómez-Villegas, M. A., & Marín, J. M. (2008). Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications. *Communications in Statistics - Theory and Methods, 37*(6), 972–985.
22. Gómez, E., Gómez-Villegas, M. A., & Marín, J. M. (1998). A multivariate generalization of the power exponential family of Distributions. *Communications in Statistics - Theory and Methods, 27,* 589–600.
23. Gupta, A. K., González-Farías, G., & Domínguez-Molina, J. A. (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis, 89*(1), 181–190.
24. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C Applied Statistics, 28*(1), 100–108.
25. Kotz, S., Kozubowski, T. J., & Podgórski, K. (2003). An asymmetric multivariate Laplace Distribution. Working paper.
26. Lange, K., & Sinsheimer, J. S. (1993). Normal/Independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics, 2,* 175–198.
27. Lin, T. I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis, 100*(2), 257–65.
28. Lin, T. I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing, 20,* 343–356.
29. Lin, T. I., Ho, H. J., & Lee, C. R. (2014). Flexible mixture modelling using the multivariate skew-t-normal distribution. *Statistics and Computing, 24,* 531–546.
30. Lindsay, B. G. (1995). *Mixture models: Theory, geometry, and applications*. NSF-CBMS Regional Conference Series in probability and Statistics (Vol. 5). Hayward, CA: Institute of Mathematical Statistics.
31. McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications*. New York: Marcel Dekker.
32. McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions*. New York: Wiley.
33. McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
34. Mengersen, K., Robert, C., & Titterington, M. (2011). *Mixtures: Estimation and applications*. New York: Wiley.
35. Nadarajah, S. (2003). The Kotz-type distribution with application. *Statistics, 37,* 341–358.
36. Naik, D. N., & Plungpongpun, K. (2006). A Kotz type distribution for multivariate statistical inference. In N. Balakrishnan, E. Castillo, & J. M. Sarabia (Eds.), *Advances in distribution theory, order statistics, and inference* (pp. 111–124). Boston: Birkhauser.
37. Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing, 10*(4), 339–348.
38. Plungpongpun, K. (2003). Analysis of multivariate data using Kotz type distributions. Ph.D. Dissertation, Computation and Applied Mathematics, Old Dominion University, USA.
39. Prates, M. O., Lachos, V. H., & Cabral, C. (2013). mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software, 54*(12), 1–20.
40. Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L., et al. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National academy of Sciences of the United States of America, 106,* 8519–8524.
41. Rinehart, J. S. (1969). Thermal and seismic indications of Old Faithful Geyser's inner working. *Journal of Geophysical Research, 74,* 566–573.

42. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.
43. Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B Statistical Methodology, 47,* 1–52.
44. Titterington, D. M., Smith, A. F. M., & Markov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.
45. Wang, W. L., Jamalizadeh, A., & Lin, T. I. (2020). Finite mixtures of multivariate scale-shape mixtures of skew-normal distributions. *Statistical Papers, 61,* 2643–2670.
46. Wang, W. L., & Lin, T. I. (2021). Robust clustering via mixtures of t factor analyzers with incomplete data. *Advances in Data Analysis and Classification,* 1–32.
47. Wei, Y. (2012). Robust mixture regression models using t-distribution. Master Report, Department of Statistics, Kansas State University, USA.

Robust Estimation Through Preliminary Testing Based on the LAD-LASSO

M. Norouzirad, M. Arashi, F. J. Marques, and F. Esmaeili

Abstract The least absolute deviation (LAD) estimator is an alternative to the ordinary least squares estimator when some outliers exist, or the error term in the regression model has a heavy-tailed distribution. The gist of this chapter is to present a new estimator for sparse and robust linear regression that improves the preliminary test LAD estimator, an estimator which depends on a test decision. Our strategy is to apply auxiliary information in the estimation obtained from employing the LAD-LASSO operator to find the null hypothesis, building the preliminary test estimator and its improvement. A Monte-Carlo simulation study shows that this new estimator is better than others. Moreover, an objective data analysis confirms that our proposed estimator performs better in the prediction error sense than the LAD, LAD-LASSO, and preliminary test estimators.

M. Norouzirad (✉)
Center for Mathematics and Applications (NovaMath), FCT NOVA, Caparica, Portugal
e-mail: mina.norouzirad@gmail.com; m.norouzirad@fct.unl.pt

M. Arashi (✉)
Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: arashi@um.ac.ir; m_arashi_stat@yahoo.com

Department of Statistics, University of Pretoria, Pretoria, South Africa

F. J. Marques
Center for Mathematics and Applications (NovaMath), FCT NOVA, Caparica, Portugal

Department of Mathematics, FCT NOVA, Caparica, Portugal
e-mail: fjm@fct.unl.pt

F. Esmaeili
Department of Statistics, Faculty of Statistics, Mathematics and Computer, Allameh Tabataba'i University, Tehran, Iran
e-mail: foadesmaeili5@gmail.com

A. Bekker et al. (eds.), *Innovations in Multivariate Statistical Modeling*,
Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-031-13971-0_19

1 Introduction

A regression model can be accommodated with a suitable prediction property if the input variables are correctly selected. The ordinary least squares (OLS) estimator, the minimizer of the sum of squared errors, provides the best prediction if all the input variables are significant and the underlying regression assumptions are met. In case of violation of the assumptions, such as multicollinearity, outliers, or/and heavy-tailed responses, the OLS fails to provide a high level of prediction accuracy. Furthermore, when there are many variables, the interpretation of the estimates is challenging.

To achieve a sparse model, the least absolute shrinkage and selection operator (LASSO), a frequently used method for simultaneously selecting and estimating variables, is proposed by [23]. The LASSO estimator retains the good features of the classical methods, both subset selection for selecting important parameters and the ridge regression proposed by [9], which stabilizes the estimates by restricting the coefficients and then shrinks the coefficients. On the other hand, the OLS estimates can be distorted when the error/response has a heavy-tailed distribution or outliers among the data. It is well known that the OLS estimate is not robust to even a single outlier, and the breakdown point is $1/n$ for a sample of size n. One of the most commonly used techniques to overcome this problem is the least absolute deviation (LAD) estimation strategy that has $\sqrt{n}$-consistency and asymptotic normality [17].

The LASSO estimator is obtained by minimizing the sum of squared residuals. Then, it will be significantly degraded in having noise. While sparse and robust linear regression is still a developing field of research, and several approaches have been published, which are sparse Least trimmed squares (LTS), a sparse version of the LTS estimator [1], a robust and sparse elastic net least trimmed squares (enetLTS) [14], a robust version of sparse partial least squares (SPRM) [10], etc., and a good review of these methods can be found in [8]; here we used the strategy proposed by [24] which is the least absolute deviation, shrinkage, and selection method (LAD-LASSO).

Suppose we have a set of covariates to fit a regression model for predicting the response variable. If a priori is known or suspected that a subset of the covariates does not significantly contribute to the overall prediction of the average response, they may be left aside, and a model without these covariates may be considered. In such situations, a subset of the covariates may be viewed as a nuisance, such that they are not of primary interest to the researcher, but they cannot be completely ignored either. Their effect must be considered when estimating the remaining regression parameters in such cases. Now, the challenge is to decide whether the full or subset model is the best. A plausible solution is the *preliminary test estimator*.

A preliminary test estimator is an estimator whose value depends on a test for statistical significance. In other words, firstly, we conduct a test, and on the other hand, based on the test result, we decide about the final estimator. Thus, the estimate is an observed value of an estimator whose expression is based on a test decision. In the statistical literature, the preliminary test estimation of parameters was introduced by [2–4] to estimate the parameters of a model when it is suspected that some "uncertain

prior information" (UPI) on the parameter(s) of interest are available. The method involves a statistical test of UPI based on the appropriate statistics and a decision on whether the model parameters should be accepted.

In some studies, the authors use "testimator" instead of the preliminary test estimator. The term "testimator" was firstly introduced by [18]. They described estimators based on the inferences derived from a preliminary test(s) and applied the method to estimate the mean of the multi-normal distribution. For more bibliography and detailed discussion about preliminary test estimation in regression models, refer to [22].

However, the preliminary test estimator suffers from being discontinuous and highly depends on the level of significance of the test. This motivated researchers to improve the preliminary test estimator by incorporating the shrinkage strategy. We only refer to [20] for a detailed and extensive study and overview of the methods. However, our goal here is slightly different from what has been done in the literature, e.g., [15] and [16]. If the LAD-LASSO is well motivated to be used, we also define the improved preliminary test estimator by incorporating the idea of shrinkage. Thus, this paper aims to strengthen the preliminary test estimator defined by combining a model obtained from the LAD-LASSO technique (the sub-model) with the full model using the shrinkage approach.

A review of the LAD-LASSO technique can be found in Sect. 2. Our strategy will give a shrinkage-type estimator with a lower prediction error than the LAD and LAD-LASSO estimators. The improved estimator is defined in Sect. 3. The performance of the proposed estimator is analyzed using a simulation study and a real dataset in Sect. 4. The information about R code can be find in Sect. 5, while we summarize our findings in Sect. 6.

2 LAD-LASSO Estimator

Consider the conventional regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is the response vector, $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)$ is an $n \times p$ fixed design matrix, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is the vector of unobservant random errors that has a cumulative distribution function $F(\boldsymbol{\epsilon})$ with median zero.

Reference [23] proposed the LASSO given by

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2 + n\lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2}$$

where $\lambda > 0$ is the tuning parameter. It shrinks the coefficient to zero and sets some coefficients to exactly zero, based on the value of λ. Therefore, the procedure com-

bines variable selection and shrinking of the coefficients of a penalized regression. The finite-dimensional performance of the LASSO estimator under standard errors was shown by [23]. The statistical properties of the LASSO estimator were studied by [7, 11, 19, 25].

When errors in (1) are distributed in a heavy-tailed manner, the performance of the LASSO becomes weaker due to the OLS estimator; it is sensitive to the heavy-tailed error distributions and outliers. Due to this sensitivity, [24] proposed a robust regression shrinkage and selection method that can do regression shrinkage and selection (like LASSO), and it is also resistant to the presence of outliers in the response variable or heavy-tailed errors (like LAD). Indeed, they combined the usual LAD criterion and the LASSO-type penalty to produce the LAD-LASSO estimator that belongs to the consistent class of model selection criteria. On the other hand, if the true model is of finite dimension and is included in a set of candidate models, then the LAD-LASSO estimator can identify the true model consistently. The obtained LAD-LASSO estimator is successful in simultaneously estimating robust regression and selecting variables.

Comparing the LAD-LASSO with the LAD, the LAD-LASSO performs parameter estimation while selecting the model. Furthermore, the LAD-LASSO is more resistant to heavy-tailed distributions and outliers than the LASSO.

Since the LASSO estimator uses the same tuning parameter for all regression coefficients, the resulting estimators may suffer an appropriate bias [7]. Knowing this, [24] modified the objective function of LASSO estimation to allow for different tuning parameters for each coefficient. Thus, the LAD-LASSO estimator is defined as follows:

$$\hat{\boldsymbol{\beta}}^{\mathrm{LAD-LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{p} \left| y_i - \boldsymbol{X}_i^{\top} \boldsymbol{\beta} \right| + n \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}, \tag{3}$$

where $\lambda_j > 0$ are the different tuning parameters for $j = 1, \ldots, p$. [24] proposed a method for calculating the LAD-LASSO estimator.

3 Improvement Strategy on LAD

By decomposing the unknown regression parameters as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\top}, \boldsymbol{\beta}_2^{\top})^{\top}$, where $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{p_1})^{\top}$ is the main covariates and $\boldsymbol{\beta}_2 = (\beta_{p_1+1}, \ldots, \beta_p)^{\top} = (\beta_1, \ldots, \beta_{p_2})^{\top}$ is the p_2 nuisance covariates that we are primarily interested in the assumption that it will be "close to zero". So that Eq. (1) may also be written as

$$\boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are assumed to have dimensions $n \times p_1$ and $n \times p_2$, respectively.

Our ultimate goal is the estimation of $\boldsymbol{\beta}_1$ when it is suspected that other variables are eliminated (i.e., $\boldsymbol{\beta}_2 = \mathbf{0}$). To test the nested linear models, the null hypothesis may be written as

$$\mathcal{H}_o : \boldsymbol{\beta}_2 = \mathbf{0}, \tag{5}$$

and it is common to consider the linear alternative hypothesis, $\mathcal{H}_A : \boldsymbol{\beta}_2 \neq \mathbf{0}$.

The LAD estimator of $\boldsymbol{\beta}$ is defined as $\tilde{\boldsymbol{\beta}}^{\text{LAD}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_1 \right\}$, where $\|\boldsymbol{v}\|_1 = \sum_{i=1}^{p} |v_i|$ is the L_1 norm, for $\boldsymbol{v} = (v_1, \ldots, v_t)^\top$. Denote the LAD estimator of $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}}^{\text{LAD}} = \left(\tilde{\boldsymbol{\beta}}_1^{\text{LAD}}, \tilde{\boldsymbol{\beta}}_2^{\text{LAD}} \right)^\top$. Thus, $\tilde{\boldsymbol{\beta}}_1^{\text{LAD}}$ is a LAD estimator of $\boldsymbol{\beta}_1$.

Under the null hypothesis $\boldsymbol{\beta}_2 = \mathbf{0}$ (ignore nuisance covariates and identify sub-model), the model in Eq. (4) reduces to $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$. So that the LAD estimator of sub-model $\boldsymbol{\beta}_1$ is defined as

$$\hat{\boldsymbol{\beta}}_1^{\text{LAD}} = \arg\min_{\boldsymbol{\beta}_1} \left\{ \left\| \boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 \right\|_1 \right\}. \tag{6}$$

Notice that $\tilde{\boldsymbol{\beta}}^{\text{LAD}}$ refers to the full model, $\tilde{\boldsymbol{\beta}}_1^{\text{LAD}}$ is the significant component of the full-model estimator. Accordingly, the notation $\hat{\boldsymbol{\beta}}_1^{\text{LAD}}$ is used for sub-model estimation.

Similar to [17], we consider the following assumptions to ensure $\sqrt{n}$-consistency and also the asymptotic properties of LAD estimators.

Assumption A. $F(\boldsymbol{\epsilon})$ is continuous and has continuous positive density $f(\boldsymbol{\epsilon})$ at the median,

Assumption B. For some positive definite (p.d.) matrix $\boldsymbol{C}$, $\lim_{n\to\infty} n^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{C}$, meaning $n^{-1} \max_{1\leq i\leq n} \boldsymbol{X}_i^\top \boldsymbol{X}_i \to 0$.

To test the null hypothesis, we use the following result.

Theorem 3.1 ([6]) The test statistic for testing $\mathcal{H}_o$ against $\mathcal{H}_A$ is given by

$$\mathcal{L}_n = \|\boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1^{\text{LAD}}\|_1 - \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{LAD}}\|_1. \tag{7}$$

Under assumptions A & B and the null hypothesis (5), as $n \to \infty$

$$\mathcal{L}_n \xrightarrow{\mathcal{D}} \frac{1}{4f(0)}\chi^2_{p_1}, \tag{8}$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution and χ^2_ν is the chi-squared distribution with ν degrees of freedom.

According to Theorem 3.1, the asymptotic distribution of $\mathcal{L}_n$ depends on the distribution function of errors. To avoid density estimation, following [6], we propose a new method for distributional approximation. We summarize this method as below.

Algorithm [6]
The algorithm is carried out in three steps:

Step 1. Given the dataset $\{(x_i, y_i) : i = 1, \ldots, n\}$, generate a random sample of size n, independent of the dataset, from a distribution with mean and variance both equal to 1. e.g., the standard exponential distribution. The random sample is denoted by $(\omega_1, \ldots, \omega_n)$. Combining the dataset and random sample gives the data-embedded sample (DES),

$$\left\{(x_i, y_i, \omega_j), i, j \in \{1, \ldots, n\}\right\}.$$

Step 2. Under the sub (null hypothesis) and full models, respectively let

$$\hat{\boldsymbol{\beta}}_1^{\omega\text{LAD}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\omega}(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1)\|_1$$
$$\tilde{\boldsymbol{\beta}}^{\omega\text{LAD}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\omega}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_1$$

Step 3. Define

$$\mathcal{L}_n^* = \left(\sum_{i=1}^n \omega_i \left|y_i - \boldsymbol{X}_{1i}\hat{\boldsymbol{\beta}}_1^{\omega\text{LAD}}\right| - \sum_{i=1}^n \omega_i \left|y_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}^{\omega\text{LAD}}\right|\right) - \left(\sum_{i=1}^n \omega_i \left|y_i - \boldsymbol{X}_{1i}\hat{\boldsymbol{\beta}}_1^{\text{LAD}}\right| - \sum_{i=1}^n \omega_i \left|y_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}^{\text{LAD}}\right|\right). \quad (9)$$

□

The statistic $\mathcal{L}_n^*$ is used to approximate the distribution of $\mathcal{L}_n$. Reference [6] established the validity of using this method to approximate the distribution $\mathcal{L}_n$, if assumption (C) given below holds.

Assumption C. The random weights $\omega_1, \omega_2, \ldots$ are i.i.d. non-negative random variables such that $\mathrm{E}(\omega_1) = \mathrm{Var}(\omega_1) = 1$, and the sequences $\{\omega_i\}$ and $\{\boldsymbol{X}_i, e_i\}$ are independent.

Remark 3.1 It should be noted in the DES scheme, a random component $\{\omega_1, \ldots, \omega_n\}$ is embedded in the dataset each time. This method is not a resampling nor bootstrap.

When $\boldsymbol{\beta}_2 = \mathbf{0}$ is true, the sub-model LAD has smaller asymptotic dispersion than the full-model LAD estimator. However, for $\boldsymbol{\beta}_2 \neq \mathbf{0}$, the sub-model LAD estimator may be biased and inconsistent in many cases. For this reason, it is plausible to consider

a preliminary test LAD estimator by taking $\tilde{\boldsymbol{\beta}}_1^{\text{LAD}}$ or $\hat{\boldsymbol{\beta}}_1^{\text{LAD}}$ depending on whether the null hypothesis is rejected or not. We denote it by $\hat{\boldsymbol{\beta}}_1^{\text{ptLAD}}$. It is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1^{\text{ptLAD}} &= \tilde{\boldsymbol{\beta}}_1^{\text{LAD}} I\left(\mathcal{L}_n \geq \mathcal{L}_\alpha\right) + \hat{\boldsymbol{\beta}}_1^{\text{LAD}} I\left(\mathcal{L}_n < \mathcal{L}_\alpha\right) \\ &= \tilde{\boldsymbol{\beta}}_1^{\text{LAD}} - (\tilde{\boldsymbol{\beta}}_1^{\text{LAD}} - \hat{\boldsymbol{\beta}}_1^{\text{LAD}}) I(\mathcal{L}_n < \mathcal{L}_\alpha), \end{aligned} \tag{10}$$

where $I(A)$ is an indicator function which is 1, if A is true and zero, otherwise; and $\mathcal{L}_\alpha$ is the upper $100\alpha\%$ percentile of the null distribution of $\mathcal{L}_n$, which can be approximated by $(1-\alpha)$-percent of 10000 values of $\mathcal{L}_n^*$ defined in Eq. (9).

The proposed PTE is not a practical estimator for two reasons:
(1) It has a discrete nature, i.e., it results in either $\tilde{\boldsymbol{\beta}}_1^{\text{LAD}}$ or $\hat{\boldsymbol{\beta}}_1^{\text{LAD}}$, and therefore, not both contribute to the information gain of the final estimate.
(2) It heavily depends on the level of significance of the test. It causes different levels, and we get different results.

Hence, we combine the best of preliminary test estimator and Stein-type shrinkage strategies to propose the improved preliminary test LAD estimator as

$$\hat{\boldsymbol{\beta}}_1^{\text{iptLAD}} = \hat{\boldsymbol{\beta}}_1^{\text{LAD}} - (1 - c\mathcal{L}_n^{-1})^+ \left(\tilde{\boldsymbol{\beta}}_1^{\text{LAD}} - \hat{\boldsymbol{\beta}}_1^{\text{LAD}}\right), \quad c = p_2 - 2,\ p_2 \geq 3, \tag{11}$$

where $(1 - c\mathcal{L}_n^{-1})^+ = \max\{0, 1 - c\mathcal{L}_n^{-1}\}$.

The amount of factor $(1 - c\mathcal{L}_n^{-1})$ depends on the value of the test statistic $\mathcal{L}_n$ and the constant c. If $c\mathcal{L}_n^{-1} \geq 1$, this factor will be negative. Thus, some coefficient signs can be reversed. This over-shrinkage phenomenon does not affect the risk performance of the estimator when using the $(1 - c\mathcal{L}_n^{-1})$ as a factor but causes the interpretation of coefficients to be more difficult. From a practical point of view, the sign change may confuse the analyst, especially when compared with the preliminary test estimator. Thus, we use $(1 - c\mathcal{L}_n^{-1})^+$ as a factor to not change the sign of covariates.

4 Numerical Study

In this section, we compare the proposed improved preliminary test LAD estimator ($\hat{\boldsymbol{\beta}}_1^{\text{iptLAD}}$) with the preliminary test LAD estimator ($\hat{\boldsymbol{\beta}}_1^{\text{ptLAD}}$), and the LAD-LASSO estimator ($\hat{\boldsymbol{\beta}}^{\text{LAD-LASSO}}$) numerically.

Synthetic Data Analysis

Here, we carry out a study to numerically compare the performance of the proposed estimators mentioned above. The response variable is generated according to

$$y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{12}$$

where ϵ_i is generated from the heavy-tailed standard double exponential (Laplace) distribution.

The variable $\boldsymbol{X}_i$ is generated from the standard p-variate normal distribution. As in this paper, the hypothesis is considered as

$$\mathcal{H}_o : \beta_i = 0, \qquad \forall i = p_1 + 1, \ldots, p. \tag{13}$$

Therefore, the parameters are partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top = (\boldsymbol{\beta}_1^\top, \mathbf{0})^\top$ and consider $\boldsymbol{\beta}_1 = (1, \ldots, 1)^\top$, a p_1 vector of 1s, and $\mathbf{0}$ a p_2 vector of 0s.

To evaluate the stable results, repeat the process 1000 times. The relative Mean Square Error (rel.MSE) criterion is used to compare the performance of the estimators.

Since the purpose of this paper is to improve the preliminary test estimator, we define the relative MSE (rel.MSE) as

$$\text{rel.MSE}(\hat{\boldsymbol{\beta}}_1^*; \hat{\boldsymbol{\beta}}_1^{\text{ptLAD}}) = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}_1^{\text{ptLAD}})}{\text{MSE}(\hat{\boldsymbol{\beta}}_1^*)},$$

where $\hat{\boldsymbol{\beta}}_1^*$ is any of the two estimators under our study: the improved preliminary test and LAD-LASSO estimator. The value by which the RMSE is larger than the unity indicates the degree of superiority of the estimator $\hat{\boldsymbol{\beta}}_1^*$ over $\hat{\boldsymbol{\beta}}_1^{\text{ptLAD}}$. The RMSEs for estimators under study are computed for $n \in \{30, 50, 100\}$ and $(p_1, p_2) = (3, 5), (3, 7), (6, 4), (8, 12), (14, 6)$. Table 1 gives the results of simulations. To obtain a better realization of the MSE behavior, Fig. 1 shows their distribution for some values of (p_1, p_2).

Based on Table 1, the following results were obtained:

(1) The LAD-LASSO estimator is better only in one situation: $n = 30$, $p = 20$, ($p_1 = 8$, and $p_2 = 12$). It means that the number of observations is less than others ($n = 50, 100$), and the model is sparse here (more parameters are zero). Thus, it can be stated that for the sparse model, when the number of observations is relatively low, the LAD-LASSO behaves better than the preliminary test estimator. If we see the improvement's value, it is ignorable. Thus, it is nearly true if we conclude that the preliminary test estimator always behaves better than the LAD-LASSO estimator from the perspective of low MSE.
With an increase of n, there is no evidence for the stable behavior of this estimator.

Table 1 The relative MSE values for the candidate estimators in the synthetic data

n	p_1	p_2	ptLAD	iptLAD	LAD-LASSO
30	3	5	1.0000	1.0270	0.9450
	3	7	1.0000	1.9922	0.4791
	6	4	1.0000	1.2327	0.6462
	8	12	1.0000	3.0833	1.0050
	14	6	1.0000	0.9370	0.6737
50	3	5	1.0000	1.0375	0.8784
	3	7	1.0000	1.4972	0.4143
	6	4	1.0000	1.1621	0.6518
	8	12	1.0000	1.6684	0.5785
	14	6	1.0000	1.1720	0.6521
100	3	5	1.0000	1.0206	0.8393
	3	7	1.0000	1.3959	0.3808
	6	4	1.0000	1.1524	0.5952
	8	12	1.0000	1.3281	0.4704
	14	6	1.0000	1.1707	0.7153

(2) If we fix p_1, the number of components of $\boldsymbol{\beta}_1$, and increase p_2, the size of $\boldsymbol{\beta}_2$, we conclude that for all values of n, the performance of the improved preliminary test estimator will be better.
(3) Let fix p. Two cases will happen: (i) p_2 is large e.g., for $p = 10$, ($p_1 = 3$, and $p_2 = 7$) and for $p = 20$, ($p_1 = 8$, and $q = 12$), (ii) p_2 is small e.g., for $p = 10$ ($p_1 = 6$, and $p_2 = 4$) and for $p = 20$ ($p_1 = 14$, and $p_2 = 6$). The rel. MSE of the improved preliminary test estimator, in case (i) is larger than its value in case (ii) for small and medium n. For large n, it is vice versa. Thus, it can be stated that the improved estimator performs better for the sparse model, especially when the sample size is small.

Figure 1 shows the distribution of the MSEs for the preliminary test, its improvement, and LAD-LASSO estimators for some values of (p_1, p_2) in Table 3. The results obtained from this figure are:

(1) Based on the distribution of MSEs, the values of LAD-LASSO estimator are greater than the other competitors. Table 1 shows that the performance of this estimator is better than the preliminary test estimator for small n, $p = 20$ ($p_1 = 8$ and $p_2 = 12$). If we check the related figure, here, $n = 30$, ($p_1 = 8$ and $p_2 = 12$), the figure confirms this result since some values of MSE of the preliminary test estimator are larger than that one of the LAD-LASSO estimator.
(2) More values of MSEs are around the mean, since the width of violin plots is large. There are more values of MSE around its mean for the improved preliminary test compared to the preliminary test estimator.
(3) The values in the x-axis show that the MSE values decrease with an increase of n.

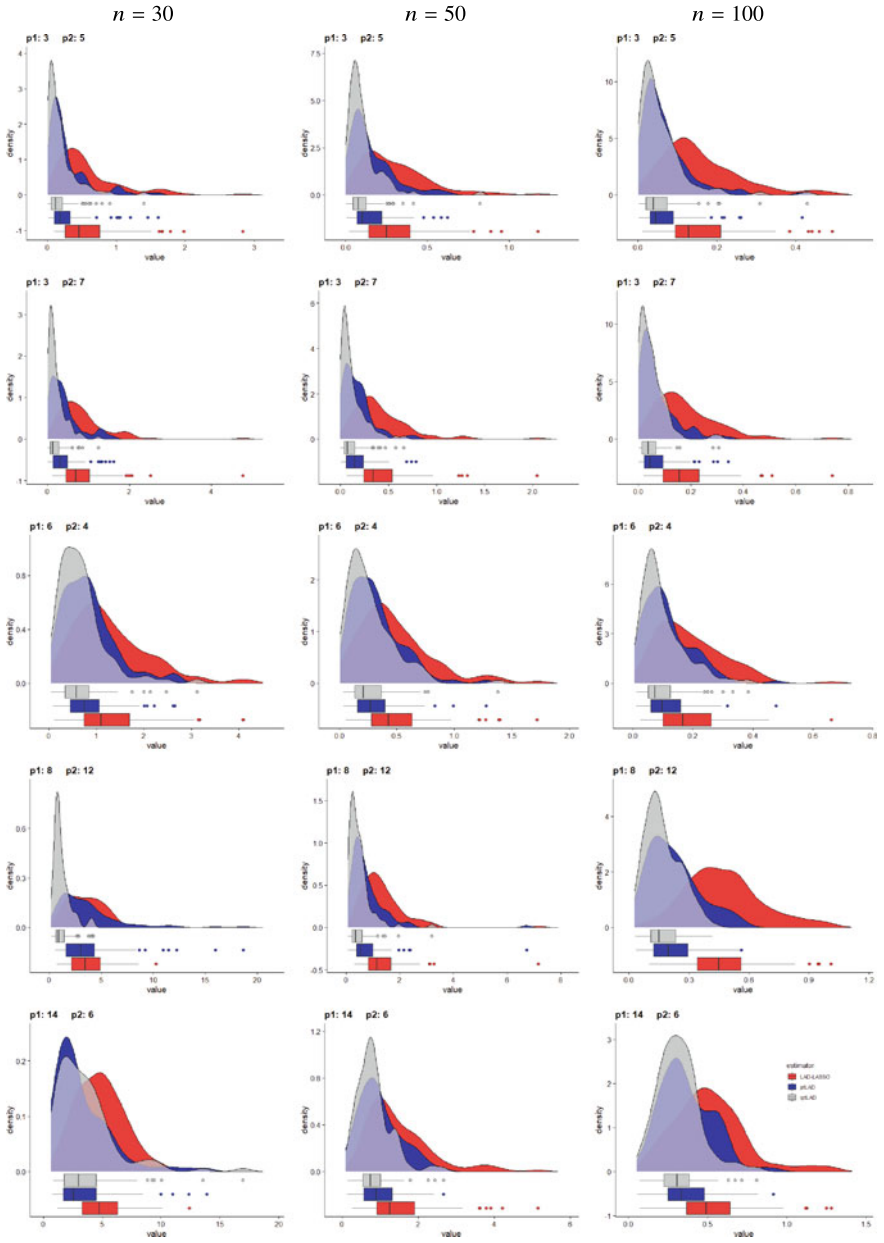


Fig. 1 Distributions and box-plots of the MSEs for the three estimators in the synthetic data

Gross Domestic Product Data Analysis

This section uses a real data example to illustrate the application of the improved LAD-preliminary test estimation.

The Barro dataset of [5] is used and consists of 161 observations on national growth rates for periods 1960–1985, i.e., 1960, 1965, 1970, 1975, and 1985, or for an average of five years subperiod over 1960–1985 in the world. Reference [13] also studies this dataset to describe a brief empirical foray into models of international economic growth designed via a quantile regression application. This dataset is available in *quantreg* package [12] in R software.

This data set is obtained to find a model between
Y.NET: Annual Change Per Capita gross domestic product (GDP),
and the covariates:
MSE2: Male Secondary Education, FSE2: Female Secondary Education, FHE2: Female Higher Education, MHE2: Male Higher Education, LEXP2: Life Expectancy, INTR2: Human Capital, GEDY2: Education/GDP, IY2: Investment/GDP, GCONY2: Public Consumption/GDP, LBLAK2: Black Market Premium, POL2: Political Instability, TTRAD2: Growth rate Terms Trade. The primary goal is to determine those variables that significantly contribute to the annual change in GDP and consequently, identify the sub-model that results in better prediction.

The summary statistics of all variables are given in Table 2.

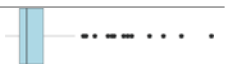
It can be realized from Fig. 2 that the observations of Australia, Bangladesh, Botswana, Ghana, and Venezuela in 1985 may be outliers. The result of applying an outlier test function by using the car package in R confirms that the observation of Bangladesh in 1980 is also an outlier.

Thus, this dataset suffers from the problem of outliers, and it is convenient to use the LAD estimation. The LAD-LASSO model returns mse2, fhe2, Intr2, gedy2, and pol2 as the best covariates to be included. We considered the dropped variables to be a nuisance and hypothesized that their coefficients were zero. Thus, we have two models: the full model with all covariates and the sub-model, excluding the nuisance variables.

In the following, we use K-fold cross-validation to estimate the model's prediction errors. In K-fold cross-validation, the dataset is randomly divided into K subsets of roughly equal size. One subset is left aside and termed the test set, while the reaming $K - 1$ subsets, called the training set, are used to fit the model. The fitted model is then used to predict the responses of the test data set. Finally, the prediction errors are obtained using the squared deviation of the observed and predicted values in the test set. The process is repeated for all K subsets, and the prediction errors are combined.

Since cross-validation is a random procedure, the estimated prediction error varies across runs and for different values of K. To account for the random variation, we repeat the process 2000 times and estimate the average prediction errors along with this number, as no noticeable variations in the standard deviations were observed for the higher number of repetitions.

Table 2 Summary statistics for the gross domestic product dataset

Parameter	Variables	Mean	sd	Min	Max	Boxplot
	Y.NET	0.019	0.025	−0.056	0.081	
β_1	LGDP2	7.790	0.954	5.820	9.508	
β_2	MSE2	0.967	0.857	0.024	4.227	
β_3	FSE2	0.712	0.833	0.000	0.712	
β_4	FHE2	0.079	0.122	0.000	0.712	
β_5	MHE2	0.158	0.175	0.000	1.011	
β_6	LEXP2	4.044	0.203	3.555	4.315	
β_7	INTR2	1.462	2.549	−1.122	15.042	
β_8	GEDY2	0.036	0.014	0.008	0.072	
β_9	IY2	0.201	0.088	0.017	0.404	
β_{10}	GCONY2	0.091	0.062	0.005	0.280	
β_{11}	LBLAK2	0.191	0.307	0.000	2.117	
β_{12}	POL2	0.168	0.241	0.000	0.802	
β_{13}	TTRAD2	−0.006	0.037	−0.118	0.141	

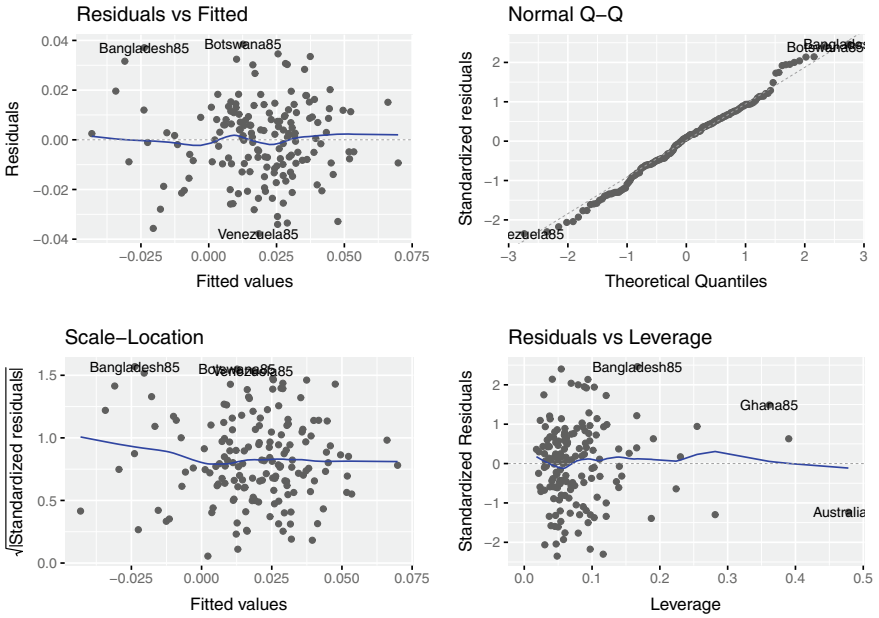


Fig. 2 Diagnostic plots for identifying outliers in the gross domestic product dataset

We measured the prediction error performance using the relative mean prediction error (rel.MPE) criterion. rel. MPE of an estimator $\hat{\boldsymbol{\beta}}_1^*$ is defined as

$$\text{rel.MPE}(\tilde{\boldsymbol{\beta}}_1^{\text{LAD}}; \hat{\boldsymbol{\beta}}_1^*) = \frac{\text{MPE}(\tilde{\boldsymbol{\beta}}_1^{\text{LAD}})}{\text{MPE}(\hat{\boldsymbol{\beta}}_1^*)}, \tag{14}$$

where $\hat{\boldsymbol{\beta}}_1^*$ is one of the improved LAD preliminary test and LAD-LASSO estimators.

We obtain five significant covariates based on running LAD-LASSO as FSE2, FHE2, MHE2, and GEDY2. Thus, the hypothesis may be considered as

$$\boldsymbol{\beta}_2 = (\beta_3, \beta_4, \beta_5, \beta_8)^\top = (0, 0, 0, 0)^\top.$$

The shrinkage estimates were obtained in two steps: first, a sub-model was selected based on the LAD-LASSO procedure. In the second step, the sub-model obtained in the first step was considered the restricted model, and then shrinkage estimates were obtained.

Table 3 shows RMPE, and their standard deviations for the improved LAD-preliminary test estimator and LAD-LASSO estimators based on K-fold cross-validation repeated 2000 times.

We find that the improved preliminary test LAD estimator has a smaller average prediction error than the LAD-LASSO estimator; Fig. 3 confirms this result.

Table 3 Relative prediction errors based on the K-fold CV repeated 2000 times for the gross domestic product data

Estimator	$K = 5$	$K = 10$
Preliminary test LAD estimator	1.0000	1.0000
Improved preliminary test estimator	1.2293	1.2363
LAD-LASSO	0.4898	0.4943

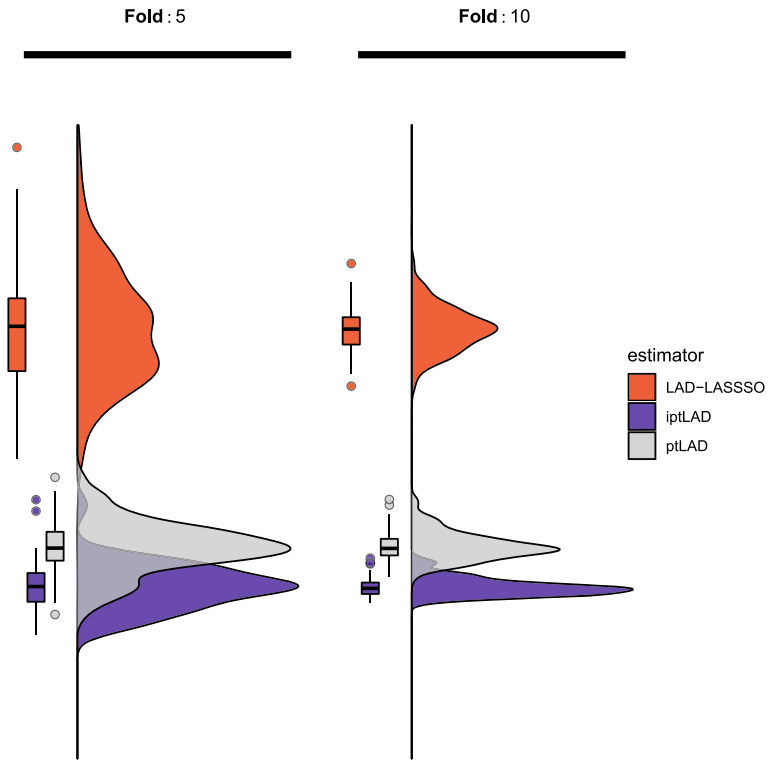


Fig. 3 Prediction error values of the estimators

Following, we propose an algorithm to find the best estimator for the response prediction. (1) determine the significant variables with the LAD-LASSO estimator, then (2) compute the preliminary test LAD estimator, and (3) calculate the improved preliminary test LAD estimator. The values of these estimators are given in Table 4.

Table 4 Estimate values for the gross domestic product data

Parameter	Variable	LAD-LASSO	ptLAD	iptLAD
β_1	lgdp2	−0.0218	−0.0267	−0.0247
β_2	mse2	0.0102	0.0108	0.0048
β_3	fse2	0.0000	0.0000	0.0000
β_4	fhe2	0.0000	0.0000	0.0000
β_5	mhe2	0.0000	0.0000	0.0000
β_6	lexp2	0.0408	0.0550	0.0513
β_7	lintr2	−0.0018	0.0000	0.0000
β_8	gedy2	0.0000	0.0000	0.0000
β_9	Iy2	0.0794	0.0851	0.0942
β_{10}	gcony2	−0.1072	−0.1063	−0.1070
β_{11}	lblakp2	−0.0255	−0.0256	−0.0299
β_{12}	pol2	−0.0254	−0.0306	−0.0251
β_{13}	ttrad2	0.1315	0.1696	0.1826

5 Codes

The codes are available at https://github.com/mnrzrad/iptLAD.

6 Conclusion

If uncertain prior information about the parameters is available, the estimators can be obtained by incorporating them via the preliminary test estimator. It first tests that information as a null hypothesis, then estimates based on the test result. However, variable selection techniques may be used to find the candidate model when there is no information. In this paper, we combined the information obtained from this method with the full-model LAD regression to achieve a preliminary test estimator and proposed the improved preliminary test LAD estimator. Based on the simulation study and real data analysis, the MSE and prediction error of the proposed estimator are less than those of the LAD-LASSO estimator and the LAD-preliminary test estimator, respectively. It appears that the performance of the preliminary test LAD estimator is better than the LAD-LASSO estimator.

Acknowledgements M. Norouzirad and F. J. Marques wish to acknowledge funding provided by the National Funds through the FCT—Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications).
M. Arashi's work was based upon research supported in part by the National Research Foundation (NRF) of South Africa, SARChI Research Chair UID: 71199; Ref.: IFR170227223754 grant No.

109214. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily attributed to the NRF.

References

1. Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics, 7*(1), 226–248.
2. Bancroft, T. A. (1944). On biases in estimating due to use of preliminary tests of significance. *The Annals of Mathematical Statistics, 15*(2), 195–204.
3. Bancroft, T. A. (1964). Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. *Biometrics, 20*(3), 427–442.
4. Bancroft, T. A. (1965). Inference for incompletely specified models in the physical sciences (with discussion). *Bulletin ISI, Proceedings 35th Session, 41*(1), 497–515.
5. Barro, R., & Lee, J. W. (1994). Dataset for a panel of 138 countries, http://admin.nber.org/pub/barro.lee.
6. Chen, K., Ying, Z., Zhang, H., & Zhao, L. (2008). Analysis of least absolute deviation. *Biometrika, 95*(1), 107–122.
7. Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96,* 1348–1360.
8. Filzmoser, P., Serneels, S., Maronna, R., & Croux, C. (2009). Robust multivariate methods in chemometrics. In Brown, S. D., Tauler, R., & Walczak, B. (Eds.), *Comprehensive chemometrics: Chemical and biochemical data analysis* (pp. 663–722). Elsevier. http://hdl.handle.net/20.500.12708/26445.
9. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12,* 186–199.
10. Hoffmann, I., Serneels, S., Filzmoser, P., & Croux, C. (2015). Sparse partial robust M regression. *Chemometrics and Intelligent Laboratory Systems, 149,* 50–59.
11. Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics, 28*(5), 1356–1378.
12. Koenker, R. (2015). quantreg: Quantile regression, R package version 5.19.
13. Koenker, R., & Machado, J. A. F. (1999). Goodness of fit and related inference process for quantile regression. *Journal of the American Statistical Association, 94,* 1296–1310.
14. 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.
15. Norouzirad, M., & Arashi, M. (2019). Preliminary test and Stein-type shrinkage ridge estimators in robust regression. *Statistical Papers, 60,* 1849–1882.
16. Norouzirad, M., Hossain, S., & Arashi, M. (2018). Shrinkage and penalized estimators in weighted least absolute deviations regression models. *Journal of Statistical Computation and Simulation, 88*(8), 1557–1575.
17. Pollard, D. (1991). Asymptotic for least absolute deviation regression estimators. *Econometric Theory, 7*(2), 186–199.
18. Rahman, M., & Gokhale, D. V. (1996). Testimation in regression parameter estimation. *Biometrical Journal, 38*(7), 809–817.
19. Rosset, S., & Zhu, J. (2004). Discussion of "Least angle regression" by Efron et al. *The Annals of Statistics, 32*(2), 469–475.
20. Saleh, A. K. Md. E. (2006). *Theory of preliminary test and stein-type estimation with applications*. New York: Wiley.
21. Saleh, A. KMd. E., Arashi, M., & Tabatabaey, S. M. M. (2014). *Statistical inference for models with multivariate t-distributed errors*. New Jersy: Wiley.

22. Sclove, S. L., Morris, C., & Radhakrishnan, R. (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics, 43,* 1481–1490.
23. Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society B, 58,* 267–288.
24. Wang, H., Li. G., & Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-LASSO. *Journal of Business & Economic Statistics, 25*(3), 347–355.
25. Zhao, P., & Yu, B. (2006). On model selection consistency of LASSO. *Journal of Machine Learning Research, 7,* 2541–2563.