TECHNISCHE UNIVERSITÄT WIEN

DISSERTATION

Smoothed Covariance Estimation for Multi-Source and Spatial Data in the Presence of Outliers

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften unter der Leitung von

Peter Filzmoser
E105 – Institut für Stochastik und Wirtschaftsmathematik, TU Wien

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Patricia Puchhammer
Matrikelnummer: 01426471

Diese Dissertation haben begutachtet:

1. **Univ.Prof. Dipl.Ing. Dr.techn. Peter Filzmoser**
   Institut für Stochastik und Wirtschaftsmathematik, TU Wien
2. **Assoz.Prof. PD Dr.in Bettina Grün**
   Institute für Statistik und Mathematik, WU Wien
3. **Prof. Claudio Agostinelli, PhD**
   Department of Mathematics, University of Trento

Wien, am 30. April 2025

Abstract

Multi-group or multi-source data, in which observations are partitioned into groups by external variables, arise in a wide range of disciplines. Examples include spatial data grouped by proximity, country borders, or geological units; medical data categorized by diagnosis, disease, or age; and temporal data structured by days, months, or years. These groupings are typically associated with continuous variables and reflect inherent relationships among the groups – making separate analysis inappropriate.

Outliers can have a substantial impact on classical, non-robust statistical methods, often distorting results and leading to misleading interpretations if not properly addressed. This issue becomes particularly critical in complex data structures such as multi-group or spatial data, where outliers may remain hidden and bias outcomes more easily. Detecting both classical outliers and those specific to the multi-group or spatial context is essential for producing reliable estimates. Moreover, analyzing these outliers can offer valuable insights, such as the detection of mislabeling or, in the case of geochemical spatial data, the identification of regions of potential mineralization.

This thesis develops and adapts robust statistical methods for application in multi-group settings. Key contributions include the development of a robust, smoothed covariance estimator for spatial and multi-source data – applied to local outlier detection – and its use in geochemical exploration. Furthermore, a sparse multi-group principal component analysis (PCA) framework is proposed, enabling joint analysis of global and group-specific features. Finally, a cellwise robust Gaussian mixture model (GMM) is introduced for the multi-group context, allowing for the detection of transitional group outliers. These theoretical and methodological advances significantly extend the robust statistics toolbox, providing improved analytical frameworks for multi-group data and demonstrating strong performance in both simulation studies and real-world applications.

Kurzfassung

Gruppierte Daten oder Daten aus mehreren Quellen, die aufgrund von externen Variablen eingeteilt werden, gibt es in zahlreichen Disziplinen. Beispiele inkludieren räumliche Daten, die nach geografischer Nähe, Ländergrenzen oder geologischen Einheiten gegliedert sind, medizinische Daten unterteilt nach Diagnosen oder Altersgruppen, und Zeitreihen, die nach Tagen, Monaten oder Jahren gruppiert werden können. Häufig liegt der Gruppierung eine kontinuierliche externe Variable zugrunde, welche die Gruppen inhaltlich miteinander verbindet. Somit können die Gruppen nicht separat betrachtet werden.

Klassische nicht-robuste, statistische Methoden sind anfällig gegenüber Ausreißern, die die Analyse verzerren und irreführende Schlussfolgerungen zulassen. Insbesondere bei komplexeren Daten, die zusätzlich gruppiert oder räumlich korreliert sind, gestaltet sich die Identifikation solcher Anomalien besonders herausfordernd. Sowohl klassische Ausreißer als auch solche mit kontextbezogenen Abweichungen müssen zuverlässig erkannt werden, um die Validität der Ergebnisse sicherzustellen. Darüber hinaus können identifizierte Ausreißer Hinweise auf fehlerhafte Gruppenzuweisungen liefern oder, etwa im geochemischen Kontext, potenziell auf bislang unentdeckte Vererzungen hindeuten.

Im Rahmen dieser Arbeit werden robuste statistische Methoden zur Analyse gruppierter Daten entwickelt und evaluiert. Zentrale Beiträge umfassen die Konstruktion eines robusten, geglätteten Kovarianzschätzers für räumliche und/oder gruppiert strukturierte Datensätze, dessen Einsatz zur Identifikation räumlicher Ausreißer sowie dessen Anwendung in der geochemischen Exploration. Darüber hinaus wird ein Verfahren zur Hauptkomponentenanalyse für gruppierte Daten entwickelt, das eine simultane Analyse wichtiger gruppenspezifischer und gruppenübergreifender Eigenschaften ermöglicht. Abschließend wird ein Gaußsches Mischungsverteilungsmodell für den Multigruppenkontext vorgestellt, das Rückschlüsse auf Transitionsdynamiken von Gruppenausreißern erlaubt. Die theoretischen und methodischen Beiträge erweitern die Intrumente der robusten Statistik und liefern einen Rahmen für die Analyse gruppierter Daten. Simulationsstudien und Echtdatenanwendungen demonstrieren die Stärken der entwickelten Methoden.

Acknowledgement

Writing a thesis is like growing a plant–it takes time, is sometimes unpredictable, but with patience and care the plant eventually takes root and flourishes. I want to thank all the gardeners and plant caretakers in my life who equipped me with the necessary gardening tools and stepped in during stressful times.

First and foremost, I am very grateful to my supervisor Peter for your steady guidance, thought-provoking questions, and kind encouragement, especially during more difficult times. Your dedication of time and effort to supervise and motivate any of your students, including me, is not often encountered and truly makes a difference. I want to thank all my co-authors, but particularly Ines Wilms. Your sharp-mindedness, detail-oriented and kind way inspire me to critically view and improve my work in many aspects, now and in the future.

In addition, I want to thank everyone in the CSTAT group. Liking to go to work and looking forward to meet every single person really enriches everyday life and was probably the best part of the PhD. Special thanks for the good time and valuable support to my PhD colleagues Pia, who I liked to distract with a coffee break, Barbara and Jeremy, with whom I enjoyed sharing and chatting in an office, and Lukas, Marcus and Roman, who probably won the race for the plantiest office (for now). I feel very fortunate to have spent so much time with you and to consider all of you my friends. Also, I want to thank my project partners and colleagues for the good collaboration.

Moreover, a big thank you to my friends and family for supporting me the last three years as well as many many years before that. Many thanks to all my friends from school and from university for the shared time and memories, they are to be continued. Especially to those who also chose to pursue a PhD, the encouraging talks with you did go a long way. To my granny Opi-Omi for always having been there for me and critically encouraging my talents, you are missed. I want to thank my sister, my mum, who always supports me, and my dad, who taught me more practical and maybe handier things than mathematics. Finally, to my partneroni Flo. It is easy to grow and explore life knowing that you will be by my side ready to cheer for me or to provide comfort. Your kind and open heart inspires me every day.

This work is part of the SEMACRET project and is funded by the European Union (Grant Agreement no. 101057741) and UKRI (UK Research and Innovation).

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 30. April 2025

Patricia Puchhammer

Contents

1 Introduction

The ongoing technological progress allows for more complex models and calculations in all fields of research. Especially in robust statistics, calculation-heavy algorithms are common and become applicable for real-life data. This also enables the calculation of new possibly more complex methods considering data whose underlying structure is not homogeneous.

Prominent examples are multi-group (or multi-source) data, where observations are partitioned into groups according to some external criterion. Multi-group data are present in many fields including medicine where groups can be based on diagnoses, spatial data and/or temporal data where underlying groups of interest are connected to country borders or geological units, or time units like days, months or years. Generally, groups are defined based on an underlying continuous process, which provides additional context that should be leveraged in a smooth way.

Although existing methods for homogeneous data could in theory be applied to each group separately, they miss the overall connection between the groups and possibly mask common patterns. They are also not taking advantage of the contextual information of the other groups. Applying a method to all observations combined can lead to some overall valid insights, but all group information is lost. Moreover, the heterogeneity between the groups can lead to spurious conclusions. To capture both the local and group inherent characteristics as well as the group independent and more global patterns, methods targeted towards the multi-group setting that can capture the similarities between the groups need to be considered.

Robust statistics can be used to ensure reliable results. Outliers in the data can obscure the results and in the worst case lead to misleading conclusions since classical methods are easily manipulated by outliers. Robust statistics flagging or reducing the effects of outlying observations need to be applied. Especially for complex data, outliers can heavily distort the statistical analysis in unexpected ways and can easily be masked when classical methods are applied.

In many applications, including geochemistry, observations hold relative information. Thus, they cannot be analyzed with standard statistical methods based on the classic Euclidean space. Compositional data analysis accounts for the relative information with the typical sum restriction and provides a sound framework to deal with compositional data.

This thesis contributes to research by focusing on robust analysis for multi-group data, while addressing the interplay of local and global features. The remainder of the introductory section is structured as follows. Section 1.1 introduces concepts and methods of robust statistics for homogeneous data as well as for different types of data heterogeneity. Section 1.2 describes the basis of the two main algorithms used in this thesis and Section 1.3 introduces the concept of Compositional Data (CoDa). The

remainder of the thesis includes Chapter 2 which introduces a rowwise robust covariance estimator used for local outlier detection. The outlier detection method is then further applied to geochemical data of varying data quality in Chapter 3. Chapter 4 develops a sparse PCA method which is based on robust multi-group covariance estimates. The ideas are then further extended in the cellwise outlier paradigm in Chapter 5, which derives a cellwise robust multi-group Gaussian mixture model to capture smoothness among the groups. The final chapter summarizes the findings and outlines potential directions for future research.

1.1 Robust Statistics

In data analysis, outliers are omnipresent. The goal of robust statistics is to deliver reliable estimates of parameters unaffected by outliers and allow to draw conclusions regarding the main bulk of the data. Hampel et al. (1986) define the field of robust statistics as follows: "In a broad informal sense, robust statistics is a body of knowledge, partly formalized into 'theories of robustness', relating to deviations from idealized assumptions in statistics." As hinted by the quote, assumptions about data generating processes and deviations thereof need to be addressed explicitly. In the following common assumptions for various multivariate data types are discussed further, different notion of outliers are described as well as robust methods to counteract their effects. Also the theoretical concept of the breakdown point is introduced.

1.1.1 Notions of Outlyingness

Typical assumptions in statistics are that observations stem from identically and independently distributed random variables. When inference or the likelihood is of interest the normal distribution is imposed or sometimes more general elliptical distributions. In robust statistics we assume that these assumptions hold for the majority of the data. When observations are not coherent with the assumed statistical properties of a model, they are considered to be outlying and there effect on the analysis should be removed or at least be bounded. However, it depends on the type of data and statistical model, if the mentioned general assumptions need to be fulfilled or whether they need adaptations. For each data and model type discussed below, illustrations of outliers are shown in Figure 1.1.

Classical Multivariate Outliers The typical assumptions on the data generating model like independently and identically distributed observations, possibly from a normal, occur often. Outliers are considered to contaminate the main bulk of the data, which follow the typical assumptions and to deviate from that assumption by originating from another distribution.

An often used model to describe the rowwise contamination scheme is the *Tukey-Huber contamination model* (Tukey, 1962; Huber, 1964). Within this model, we are interested in the distribution of the uncontaminated p-variate random vector $\boldsymbol{Y}$, but

can only observe the possibly contaminated random vector $\boldsymbol{X}$,

$$\boldsymbol{X} = (1 - B)\boldsymbol{Y} + B\boldsymbol{Z},$$

where $B \sim \mathcal{B}(\epsilon)$ is Bernoulli-distributed. Thus, only whole observations can be contaminated. This model serves as basis of many theoretical concepts in robust statistics like the influence function or the breakdown point.

A recent approach is to consider only contaminated cells of observations instead of whole observation rows. Alqallaf et al. (2009) formalize the *fully independent contamination model*

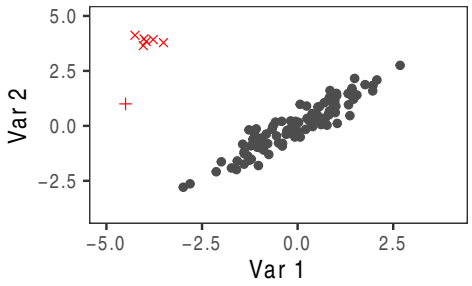
$$\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{Y} + \boldsymbol{B}\boldsymbol{Z},$$

where $\boldsymbol{B} = \text{diag}(B_1, \ldots, B_p)$ and $B_i \sim \mathcal{B}(\epsilon)$ for $i = 1, \ldots, p$ independent from each other. Especially in higher-dimensional settings, more information can be retained. There is no official consensus on when cellwise outliers can be considered rowwise outliers and it seems to depend on the chosen paradigm (Raymaekers and Rousseeuw, 2024a).

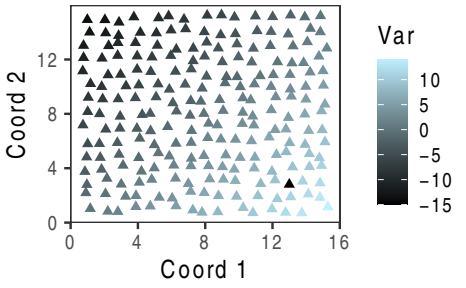
A differentiation between cellwise and rowwise outliers can also be made for the other notions of multivariate outliers discussed below, even though this distinction may not have been thoroughly investigated to date.

Local/Spatial Outliers Spatial data consist of observations with values in the multivariate feature space as well as given spatial coordinates in one to three dimensions or even more (e.g., spatio-temporal data). Temporal data or time-series data can also be seen as a special case of spatial data with one-dimensional coordinates. For spatial data, an assumption often made either explicitly or implicitly is the so-called *Tobler's first law of geography*: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Tobler's Law is the basis of most tools and methods in spatial and geo-statistics (see, e.g., Cressie, 2015) and provides also a guideline on when observations might be outlying in the spatial context. Thus, when an observation differs strongly from spatially close observations in the feature space – contrary to what would be expected based on Tobler's Law – they are spatial (or also local) outliers.

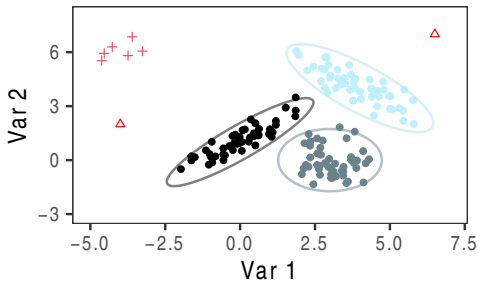
Outliers in Clustering In cluster analysis we assume that we have data, where each observation is drawn independently from one of several, often elliptical, distributions – referred to as clusters. Outlying observations, in this context, are those that do not resemble any of the clusters. A notable distinction from single-distribution frameworks is the possibility of clustered outliers: small groups of outliers that form their own cluster, distant from the main clusters. In addition to estimating the parameters of each cluster, the presence of such outliers also presents challenges in accurately determining the number and structure of clusters.



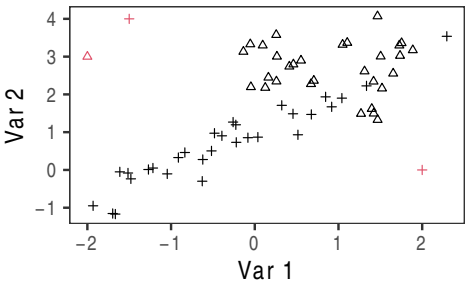
(a) Classical outliers: Uncontaminated, normally distributed observations (dots), cellwise (cross) and rowwise (diagonal cross) outliers.



(b) Local/spatial outliers: The univariate feature value is shown in colors, the spatial locations on the axis. Bottom-right corner contains a spatial outlier (dark).



(c) Cluster outliers: Three regular clusters (dots), clustered outliers (crosses) and isolated outliers (triangles).



(d) Multi-group outliers: Two given groups (crosses and triangles) in black and three outliers in red.

Figure 1.1: Different types of outliers based on the data generating process of uncontaminated data points.

Multi-Group/Multi-Source Outliers Multi-group (also multi-source) data in the context of this thesis describes data sets, where all observations have the same feature space and are partitioned into distinct groups or sources prior to statistical data analysis. The partition can be pre-defined by some external information like medical diagnoses or specified at the beginning of the analysis to emphasize particular interpretative perspectives. Additionally, it is assumed that these groups are related, so a joint analysis provides more insight than an separate analysis for each group or source. Within this framework, outliers can be classified as either global, meaning they do not conform to any group, or local, meaning they are anomalous within their assigned group but not necessarily in others.

Limitations of Non-Robust Methods

While outliers themselves also provide valuable insights into the data-generating process, including them in a non-robust estimation of parameters like location and covariance can severely distort the results, leading to estimates that no longer accurately reflect the

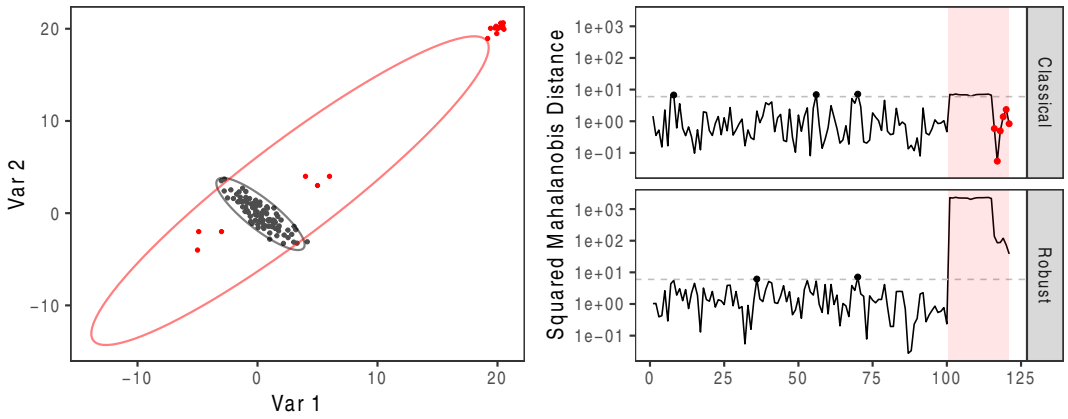


Figure 1.2: Classical versus robust parameter estimation. Left panel: 95% tolerance ellipses of classical (red) and robust (black) estimates based on the MCD-estimator applied to data with outliers (red dots). Right panel: Squared MD for classical (top) and robust (bottom) estimators. Red dots are masked outliers and black dots are non-outlying observations above the χ^2-threshold (dashed grey line) indicating masking and swamping, respectively.

main bulk of the data. Classical non-robust estimates, such as sample covariance and mean, are heavily influenced by outliers and in extreme cases one single extreme point can be sufficient to distort them arbitrarily strong (see also Section 1.1.4). The left panel of Figure 1.2 illustrates the tolerance ellipses for the sample mean and covariance (red) and robust estimates (black) for bivariate contaminated data points. Compared to 100 homogeneous observations a small number of 21 outliers lead to extreme distortions of the sample estimates. Although outliers can be visually identified in the bivariate case, such detection becomes infeasible in higher dimensions.

When detecting outliers during or after estimating procedures, robustness in all steps of a method is crucial to mitigate against masking and swamping effects. Masking of outliers occurs when they cannot be clearly distinguished from the majority of non-outlying observations due to biased or distorted estimates. Conversely, swamping refers to the incorrect classification of regular observations as outliers. An commonly used metric for identifying outliers is the Mahalanobis Distance (MD) of an observation $\boldsymbol{x} \in \mathbb{R}^p$,

$$\mathrm{MD}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}$$

which relies on location and covariance estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. If observations are normally distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then $\mathrm{MD}^2 \sim \chi^2_p$. Therefore, an often used threshold for flagging observations as outliers is the 95%-th quantile of the χ^2_p-distribution. As illustrated in the right panel of Figure 1.2, all outliers are correctly identified by the robust estimator whereas only a subset is correctly detected by the classical sample estimates demonstrating their vulnerability to masking effects.

Masking and swamping effects can also occur for heterogeneous data, as illustrated in Figure 1.3. Two regular clusters are contaminated with isolated and clustered outliers.

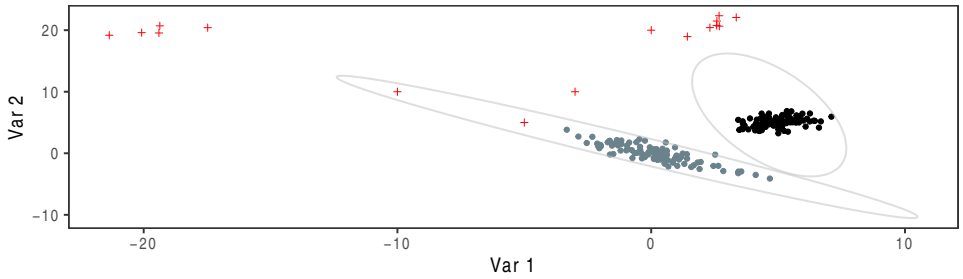


Figure 1.3: Clustering: Masking and swamping effect for 8% of outliers (crosses) and `mclust` with two clusters, a non-robust Gaussian Mixture model implementation (Fraley et al., 2024).

The non-robust clustering method returns two expanded covariance matrices shown as 95%-tolerance ellipse, where two outliers are masked and two regular observations are falsely flagged as outlying. In cluster analysis it is possible to increase the number of clusters until all outliers have their own cluster, however this is often not optimal and robust methods for clustering are preferable (García-Escudero et al., 2010).

1.1.2 General Strategies for Robustness

There are different strategies to get outlier-robust estimates for covariance and location. While outliers or outlying cells can be identified during the estimation procedure and fully removed from the resulting estimates, another approach is to bound or down-weight extreme effects of any observation or cell. Selected methods are described to illustrate main concepts in the large literature body of robust statistics.

Methods Based on Down-Weighting Extreme Effects

One approach to construct rowwise robust estimators of location and covariance is to generalize the concept of Maximum-Likelihood estimators (MLE). The large group of M-estimators (Huber, 1964; Maronna, 1976) use reweighting of observations in the univariate and of Mahalanobis Distances in the multivariate case. A non-decreasing weighting function is used which determines the degree of robustness against outliers. M-estimators are a large class of estimators that include, e.g., the median and the Least-Squares estimator. Another robust estimator is the S-estimator (Davies, 1987), which uses a robust M-estimator for scale. The Stahel-Donoho estimator (Stahel, 1981; Donoho, 1982) projects the data onto many different 1-dimensional subspaces, for which outlying measures are calculated. These are then used to down-weight outlying observations for parameter estimation.

Some of the rowwise robust methods are recently extended to the cellwise paradigm (see, e.g., Raymaekers and Rousseeuw, 2024a). The cellwise Stahel-Donoho estimator (Van Aelst et al., 2011) considers cellwise instead of rowwise weights. An extension of the S-estimator was developed by Agostinelli et al. (2015), called the two-step generalized S-estimator, which was further adapted by Leung et al. (2017). Both

consist of first filtering extreme outlying cells before applying a Generalized-S-estimator in the second step.

Methods Based on Removing Outliers

Robust methods that flag outlying observations and fully remove them are, for example, the Minimum Covariance Determinant (MCD) estimator developed by Rousseeuw (1985) that selects a subset of non-outlying observations to minimize the determinant of the corresponding sample covariance matrix. This is equivalent to choosing a subset of observations to maximize their Gaussian likelihood (Raymaekers and Rousseeuw, 2023). Its use is suggested for $n > 5p$. For high-dimensional data, the Minimum Regularized Covariance Determinant (MRCD) estimator (Boudt et al., 2020) minimizes the determinant of a regularized covariance matrix of a selected outlier-free subset. Both provide rowwise robust location and covariance estimates.

Cellwise robust methods that remove outlying cells are for example cellHandler (Raymaekers and Rousseeuw, 2021), which iterates between setting outlying cells to missing and estimating the parameters in a missing value scenario, and cellMCD (Raymaekers and Rousseeuw, 2023) which unifies the flagging of outliers and parameter estimation in one objective function. A slightly different but successful algorithm is the Detecting Deviating Data Cells (DDC) algorithm (Rousseeuw and Bossche, 2018), which flags and further imputes outlying cells. However, it does not inherently provide a location or covariance estimate.

1.1.3 Robustness for Complex Data

There are many robust methods for non-homogeneous data. Here, we focus on some well-known examples for specific areas of research.

Spatial Data

Local outlier detection in spatial data as well as robust spatial estimation methods are very diverse. A large literature body exists on geographically weighted methods that are based on a moving-window and weighting approach and also include extensions regarding outlier robustness. For example, Harris et al. (2013) use the MCD estimator applied to spatially weighted observations to robustify the estimation of local covariance matrices and detect local outliers. Other statistical methods detect spatial outliers on the basis of high pairwise MD between two observations $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\mathrm{MD}_{pair}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{y})) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})' \boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{y})^{-1} (\boldsymbol{x} - \boldsymbol{y})},$$

where $\boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{y})$ is a spatially appropriate robust covariance matrix, whose estimation is the main challenge. Here, often MCD or MRCD based estimates are chosen for robustness (Filzmoser et al., 2013; Ernst and Haesbroeck, 2016).

Another more algorithmic approach is the local outlier factor (Breunig et al., 2000, LOF), where the denseness of an observation in the multivariate space is compared to the denseness of its k-nearest neighbors in the multivariate space. Here, local refers

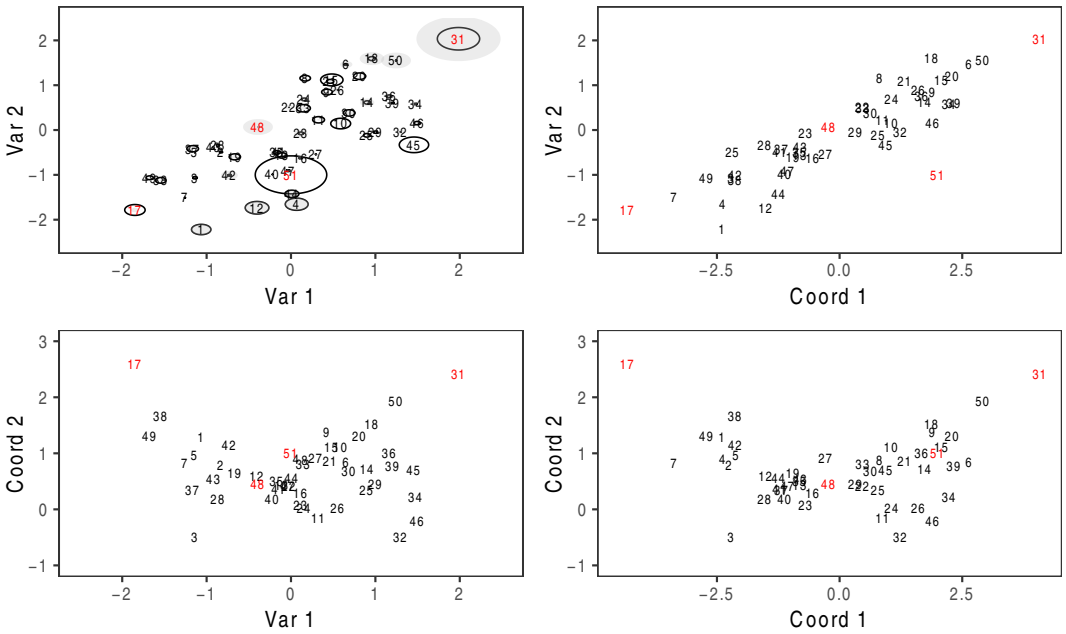


Figure 1.4: LOF for bivariate spatial data: Top left panel shows the multivariate feature space and the standard LOF (area of grey area) compared to the spatial LOF (area of black circle). The bottom right panel shows the spatial locations. The panels on the off-diagonal show the relationships of coordinates and features. One prominent local outlier is observation 51.

to the multivariate feature space. A spatial adaptation is provided by Schubert et al. (2012) by comparing the denseness in the multivariate space to the nearest neighbors in the coordinate space (see Figure 1.4).

Heterogeneous Data - Clustering

There is a large literature on clustering, where trimming approaches are used to robustify clustering methods against rowwise outliers (García-Escudero et al., 2010). A robust version of the well-known k-means method is trimmed k-means (Cuesta-Albertos et al., 1997). An extension of the MCD to clustering is introduced by Gallegos (2002); Gallegos and Ritter (2005), which was extended to allow for more flexible covariance shapes by García-Escudero et al. (2008), a method known as TCLUST. For mixture models, one approach is also to trim the likelihood (Markatou, 2000; Neykov et al., 2007).

Regarding cellwise outliers, current approaches are sclust (Farcomeni, 2014a) for Gaussian mixture models. Trimming-based methods are Farcomeni (2014b), who introduces snipping, and the cellwise-trimming approach of García-Escudero et al. (2021). Recently, a cellwise robust Gaussian mixture model extension of the cellMCD was introduced by Zaccaria et al. (2024).

Another direction is taken by Peel and McLachlan (2000) who address outliers

by mixtures of heavy-tailed t-distributions. However, as shown in Hennig (2004), heavy-tailed mixtures are not more robust against extreme outliers than Gaussian mixtures.

1.1.4 Breakdown Point

One of the theoretical concepts to assess robustness of an estimator is the breakdown point (BP). Maronna et al. (2006) characterize the rowwise breakdown point of an estimator as the largest amount of contamination (proportion of atypical points) that the data may contain such that the estimator still gives some information about the real parameter, i.e., about the distribution of the "typical" points.

Formally, an estimator $\hat{\theta}$ taking values in a parameter space Θ should remain bounded as well as bounded away from the boundaries of Θ. In the multivariate setting this implies that location estimates should remain bounded. A distinction is commonly made between two types of breakdown points: the explosion BP, which occurs when the largest eigenvalue becomes unbounded, and the implosion BP, which arises when the smallest eigenvalue approaches zero, leading to singularity of the estimated covariance matrix.

Moreover, one can distinguish between the asymptotic contamination BP, a concept based on contaminated distributions, and the finite-sample contamination BP (Donoho and Huber, 1983), which is defined for a finite sample of contaminated observation. For the finite-sample BP, outliers can either be added to an uncontaminated sample (addition BP) or a subset of uncontaminated observations is replaced by arbitrary values (replacement BP). The rest of this section focuses on the finite-sample breakdown point.

Rowwise Finite-Sample Breakdown Point

Given n fixed uncontaminated observations $\boldsymbol{X}_n$, the addition BP of an estimator $\hat{\theta}$ is defined as

$$\delta_n^*(\hat{\theta}, \boldsymbol{X}_n) = \min\left\{\frac{m}{m+n} : \hat{\theta}(\boldsymbol{X}_n \cup \boldsymbol{Y}_m) \text{ breaks down}, \boldsymbol{Y}_m \in \mathbb{R}^{m\times p}\right\},$$

where $\boldsymbol{Y}_m$ denotes m arbitrary observations added to $\boldsymbol{X}_n$. The replacement BP is defined as

$$\delta_n^*(\hat{\theta}, \boldsymbol{X}_n^m) = \min\left\{\frac{m}{n} : \hat{\theta}(\boldsymbol{X}_n^m) \text{ breaks down}\right\},$$

where $\boldsymbol{X}_n^m$ denotes the contaminated copy of $\boldsymbol{X}_n$ with up to m observations replaced by arbitrary values. Although both concepts depend on the matrix $\boldsymbol{X}_n$, many statistical methods have BPs that only depend on n. For methods with data-independent BP and where the BPs have a certain (and common) expression, Zuo (2001) shows a direct relation between addition and replacement BP.

Cellwise Finite-Sample Breakdown Point

For the cellwise paradigm, the finite-sample replacement BP is defined as

$$\epsilon_n^*(\hat{\theta}, \boldsymbol{X}_n^m) = \min\left\{\frac{m}{n} : \hat{\theta}(\boldsymbol{X}_n^m) \text{ breaks down}\right\},$$

where $\boldsymbol{X}_n^m$ denotes the contaminated copy of $\boldsymbol{X}_n$ where up to m cells per variable are replaced by arbitrary values (Raymaekers and Rousseeuw, 2023). In their work it is shown that the cellwise BP is always lower than the rowwise (replacement) BP,

$$\delta_n^*(\hat{\theta}, \boldsymbol{X}_n^m) \geq \epsilon_n^*(\hat{\theta}, \boldsymbol{X}_n^m).$$

Finite-Sample Breakdown Point for Clustering

Regarding clustering, the above defined BPs, especially the replacement BP, are typically data dependent, i.e., dependent on $\boldsymbol{X}_n$ and the underlying cluster structure. Moreover, there is variability in the literature regarding the definition of breakdown with respect to the number of affected estimates. According to Garcia-Escudero and Gordaliza (1999) an estimator for clustering breaks down if one of the cluster estimates breaks down, for Gallegos (2003) all estimates need to break down simultaneously. Hennig (2004) propose the r-components parameter breakdown point in a univariate mixture model setting and also accounts for the estimated mixture proportions. Moreover, he proposes the classification BP, which is applicable to more general clustering methods without classical parameter estimation.

To reduce the issue of data dependency of the BP, Hennig (2004) propose a setting of ideally well-clustered data, which was extended by Cuesta-Albertos et al. (2008) to the multivariate setting. Basic assumptions are that clusters are infinitely far apart from each other and from outliers. Further, outliers are infinitely far away from each other and thus do not form clusters of their own (for more details see also Chapter 5). Other approaches consist of a new notion of breakdown that are cluster dependent. Hennig (2008) introduces the dissolution point as well as isolation robustness to capture the stability of detected clusters for general clustering methods.

1.2 Algorithms

In the following sections, two algorithms utilized in Chapters 4 and 5 are described in greater detail. Section 1.2.1 is based on the work of Boyd et al. (2011) on the Alternating Direction Method of Multipliers (ADMM), while Section 1.2.2 builds on the comprehensive descriptions of the Expectation-Maximization (EM) algorithm by McLachlan and Krishnan (2008).

1.2.1 Alternating Direction Method of Multipliers

The Alternating Direction Method of Multipliers (ADMM) is an optimization algorithm and is based on the Dual Ascent method and the Method of Multipliers.

The Dual Ascent methods is applied to a constrained convex optimization problem (primal problem) of the form

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

with $\boldsymbol{x} \in \mathbb{R}^p, \boldsymbol{A} \in \mathbb{R}^{q \times p}, \boldsymbol{b} \in \mathbb{R}^q$ and f convex with values in $\mathbb{R}$. The corresponding dual problem involves maximizing the dual function

$$g(\boldsymbol{y}) = \inf_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{y})$$

with respect to $\boldsymbol{y}$, where the Lagragian is defined as $L(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}'(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$. Assuming strong duality and a unique minimizer, the optimal solutions to the primal and dual problem can be recovered from each other. The Dual Ascent method then uses gradient ascent to solve the dual problem assuming that the dual function g is differentiable.

The Method of Multipliers modifies the standard Lagrangian by introducing a quadratic penalty term, resulting in the augmented Lagrangian

$$L_p(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}'(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}||_2$$

with penalty parameter $\rho > 0$. This formulation is equivalent to the original constrained problem. Then, dual ascent is applied to solve the dual problem. Applying Dual Ascent to the augmented Lagrangian results in improved convergence properties: the dual function is differentiable under less restrictive conditions, and convergence is guaranteed even if the primal problem is not strictly convex or does not always have finite values. Optimality conditions that need to be met for an optimum $\boldsymbol{x}^*, \boldsymbol{y}^*$ are primal feasibility, $\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b} = 0$, and dual feasibility, $\nabla f(\boldsymbol{x}^*) + \boldsymbol{A}'\boldsymbol{y}^* = 0$.

The standard optimization problem for the ADMM is formalized as

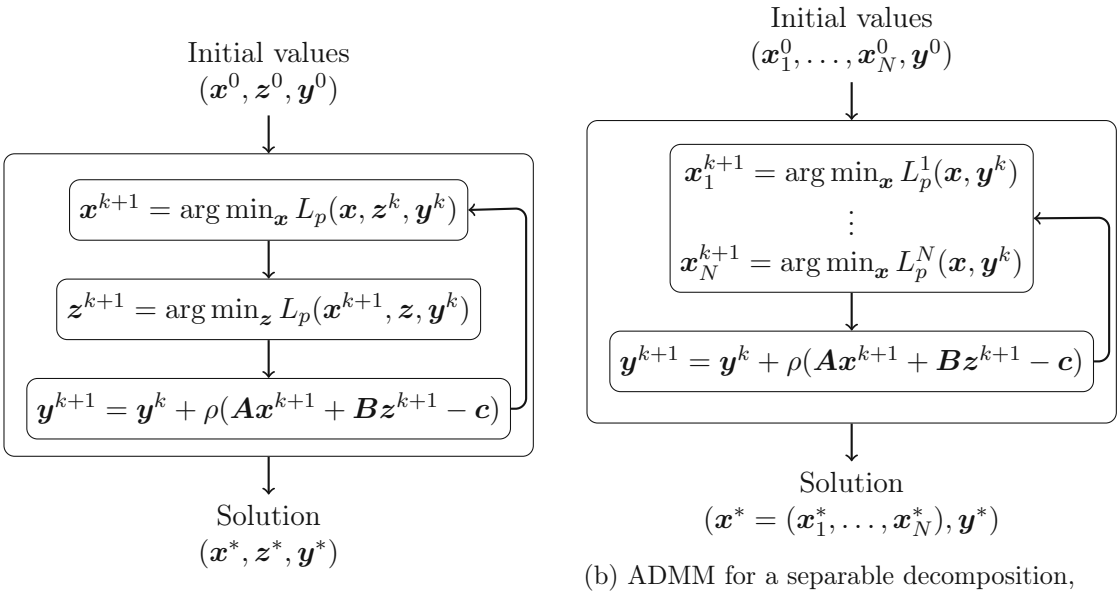
$$\min_{\boldsymbol{x}, \boldsymbol{z}} f(\boldsymbol{x}) + g(\boldsymbol{z}) \tag{1.1}$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} = \boldsymbol{c},$$

with $\boldsymbol{x} \in \mathbb{R}^p, \boldsymbol{z} \in \mathbb{R}^q, \boldsymbol{A} \in \mathbb{R}^{r \times p}, \boldsymbol{B} \in \mathbb{R}^{r \times q}, \boldsymbol{c} \in \mathbb{R}^r$ and f and g convex. The ADMM then minimizes the augmented Lagrangian function L_p

$$L_p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) = f(\boldsymbol{x}) + g(\boldsymbol{z}) + \boldsymbol{y}'(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}) + \frac{\rho}{2}||\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{z} - \boldsymbol{c}||_2$$

from the Method of Multipliers iteratively (see also Figure 1.5),

$$\boldsymbol{x}^{k+1} = \arg\min_{\boldsymbol{x}} L_p(\boldsymbol{x}, \boldsymbol{z}^k, \boldsymbol{y}^k),$$
$$\boldsymbol{z}^{k+1} = \arg\min_{\boldsymbol{z}} L_p(\boldsymbol{x}^{k+1}, \boldsymbol{z}, \boldsymbol{y}^k),$$
$$\boldsymbol{y}^{k+1} = \boldsymbol{y}^k + \rho(\boldsymbol{A}\boldsymbol{x}^{k+1} + \boldsymbol{B}\boldsymbol{z}^{k+1} - \boldsymbol{c}).$$



(a) ADMM for the standard optimization problem in Equation (1.1).

(b) ADMM for a separable decomposition, i.e., Equation (1.2) with parallelization.

Figure 1.5: Visualization of the ADMM iteration structure for different problem settings.

Theoretical results regarding convergence exist when f and g fulfill certain convexity assumptions and if the Lagrangian has a saddle point. Optimality conditions from the Method of Multipliers can be transferred similarly and the algorithm can be stopped, if primal and dual feasibility conditions are fulfilled. For non-convex problems there is no guarantee for convergence of any kind.

There are numerous extensions of ADMM, ranging from algorithmic enhancements to theoretical considerations. On the algorithmic side, improvements include variations in the penalty parameter ρ during iterations, inexact minimization steps for $\mathbf{x}$ and $\mathbf{z}$, and alternative update orders – all aimed at accelerating convergence. Beyond these, more structural considerations involve different decompositions of the problem or reformulations that are equivalent to the original formulation but allow for an application of ADMM.

A particularly useful case arises when the objective function f is additively separable for a partition of $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_1) + \ldots + f(\boldsymbol{x}_N). \tag{1.2}$$

Together with the separability of the quadratic term $||\boldsymbol{A}\boldsymbol{x}||_2^2$ corresponding to such partition, the ADMM can be leveraged, as the minimization steps can be carried out in parallel (see also Figure 1.5).

1.2.2 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is a widely known technique in statistics for maximizing likelihood functions, particularly in situations where some data is unobserved. Common examples include datasets with missing values or truncated observations. Less obvious examples are mixture models, where each observation is assumed to originate from one of several distributions, but the specific origin is unknown.

In such settings, the likelihood function for the observed (incomplete) data is typically complex and difficult to optimize directly. In contrast, the complete-data likelihood – which includes the unobserved or latent variables – is often more tractable. The EM algorithm addresses this by iteratively alternating between two steps:

- The Expectation step (E-step) estimates the missing or latent data given the current parameter estimates.
- The Maximization step (M-step) then maximizes the expected complete-data log-likelihood obtained in the E-step.

An important extension of the EM algorithm is the Generalized EM (GEM) algorithm, which relaxes the requirement of fully maximizing the complete-data likelihood in the M-step. Instead, it only requires that the new parameter estimate increases the likelihood compared to the previous estimate.

Applications of the EM algorithm in the context of mixture models include Gaussian mixture models with missing data in the component distributions (Eirola et al., 2014), as well as more general settings involving elliptical distributions (Mouret et al., 2023).

1.3 Compositional Data

The following introductory summary to Compositional Data Analysis (CoDA) is based on the work of Filzmoser et al. (2018).

The concept of compositional data was introduced by Aitchison (1982). Compositional data are characterized by their relative nature: the essential information lies not in the absolute magnitudes of the components but in their proportions relative to each other. This inherently implies scale invariance, meaning that multiplying all components by a constant does not change the compositional information. Typical forms of compositional data include percentages, proportions, and concentration measurements such as parts per million (ppm) or milligrams per kilogram (mg/kg).

As a consequence, compositional data can always be represented with the constraint of a constant sum. For instance, in a composition expressed in percentages, the sum of all components must equal 100%; in the context of material composition, the total mass of subcomponents cannot exceed the overall sample mass (e.g., 1 kg). As a result, the data reside in a simplex, a constrained sample space, rather than in unconstrained Euclidean space. This violates assumptions commonly made in standard statistical methods, which often implicitly or explicitly rely on the standard Euclidean geometry.

1.3.1 Aitchison Geometry and the Simplex

The D-part simplex is the sample space of compositional data and is defined as

$$S^D = \left\{ \boldsymbol{x} = (x_1, \ldots, x_D)' \in \mathbb{R}^D : x_i > 0, \sum_{i=1}^{p} x_i = \kappa \right\}.$$

Due to scale invariance, κ can be replaced by 1. The geometry that equips the simplex to be a vector space is not Euclidean but the so-called Aitchison geometry. The perturbation and the powering operations are defined as

$$\boldsymbol{x} \oplus \boldsymbol{y} = (x_1 y_1, \ldots, x_D y_D)' \qquad \alpha \odot \boldsymbol{x} = (x_1^\alpha, \ldots, x_D^\alpha)',$$

respectively. The inner product is defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j},$$

and together with the Aitchison geometry an Euclidean linear vector space structure with dimension $D-1$ is obtained. Working within this vector space ensures compositional coherence.

1.3.2 Transformations

Many methods are developed in the standard Euclidean space. Thus, it would be convenient to transform the data from the Aitchison space to the Euclidean space and then applying the statistical methods at hand to the new coordinates. This is possible with some of the transformations described below, however, obstacles regarding the interpretation of results occur. Three often used transformations based on log-ratios are described in the remainder of Section 1.3 and visualized in Figure 1.6.

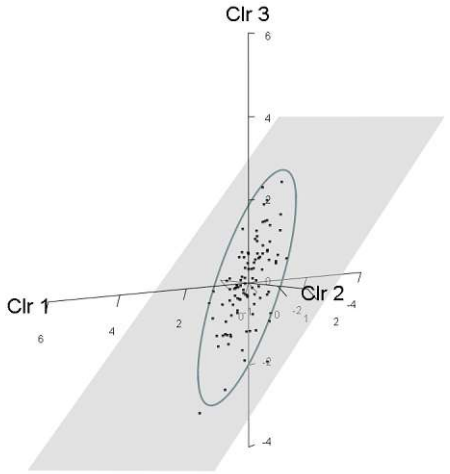
Additive Log-Ratio Coordinates (alr) The first transformation is additive log-ratios with respect to one variable j,

$$\operatorname{alr}_j(\boldsymbol{x}) = \left(\ln \frac{x_1}{x_j}, \ldots, \ln \frac{x_{j-1}}{x_j}, \ln \frac{x_{j+1}}{x_j}, \ldots, \ln \frac{x_D}{x_j} \right)'.$$

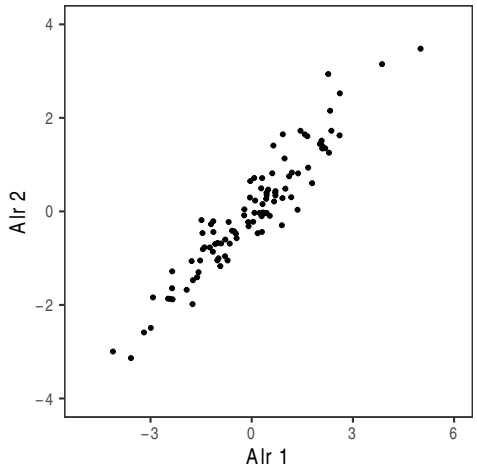
Additive log-ratio coordinates are compatible with powering and perturbation,

$$\operatorname{alr}_j(\boldsymbol{x} \oplus \boldsymbol{y}) = \operatorname{alr}_j(\boldsymbol{x}) + \operatorname{alr}_j(\boldsymbol{y}), \qquad \operatorname{alr}_j(\alpha \odot \boldsymbol{x}) = \alpha \cdot \operatorname{alr}_j(\boldsymbol{x}).$$

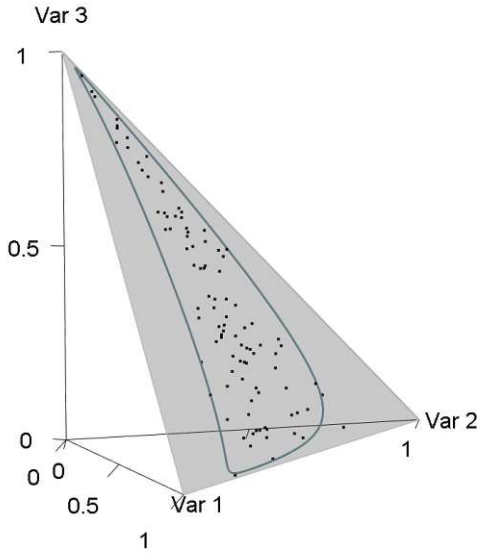
One disadvantage of the alr-transformation is the dependence on j, which can be chosen arbitrarily. Moreover, the non-orthogonality of the resulting coordinate system implies that the Aitchison inner product is not the Euclidean inner product applied to the transformed values and thus, the transformation is not isometric.



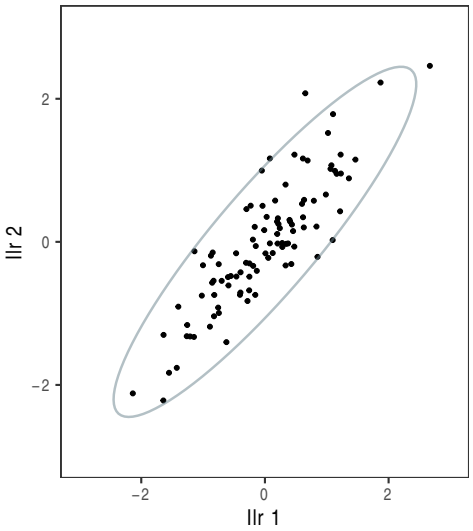
(a) Values and 95% tolerance ellipse transformed from ilr to clr-coefficients.



(b) Values transformed to alr-coordinates.



(c) Values and 95% tolerance ellipse transformed from ilr to the simplex.



(d) Normally distributed ilr-coordinates with 95%-tolerance ellipse.

Figure 1.6: Different transformation spaces: On the bottom left panel the simplex and in the bottom right panel ilr-transformed values. On the upper left panel clr-coefficients are shown and on the upper right panel alr-coordinates with Var 1 as basis. The tolerance ellipse is based on the covariance and mean used to construct normally distributed ilr-coordinates.

Centered Log-Ratio Coefficients (clr) Here, each compositional element of the simplex is geometrically centered,

$$\operatorname{clr}(\boldsymbol{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{k=1}^{D} x_k}}, \dots \ln \frac{x_D}{\sqrt[D]{\prod_{k=1}^{D} x_k}} \right)'.$$

Advantages are that there is no need for the subjective choice of a denominator and the clr-transformation is compatible with powering and perturbation as well as the inner product,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \langle \operatorname{clr}(\boldsymbol{x}), \operatorname{clr}(\boldsymbol{y}) \rangle.$$

Thus, the clr-transformation is isometric implying that distance-based statistical methods can easily be applied to clr-transformed data. However, the entries sum up to zero and thus, the coefficients are based on a linearly dependent generating system with D vectors. On the one hand, this is problematic for univariate methods and interpretations since the coefficients cannot be considered separately. On the other, many multivariate methods rely on the inversion of covariance matrices, that end up being singular.

Isometric Log-Ratio Coordinates (ilr) The idea is to form an orthonormal basis in the space spanned by the clr-coefficients. Only $D - 1$ vectors are necessary for a basis. There are infinitely many options to choose such a basis – one particular constructive basis is called pivot coordinates. Compatibility with powering and perturbation and isometry hold for all ilr-coordinates, however interpretation of ilr-coordinates with respect to certain original compositional parts is difficult or not feasible.

1.4 Overview

Chapter 2 is based on the article Puchhammer, P. and Filzmoser, P. (2024). Spatially smoothed robust covariance estimation for local outlier detection. *Journal of Computational and Graphical Statistics*, 33(3):928–940. It presents a rowwise robust covariance estimator for spatially grouped data, developed along the lines of the MRCD framework, and applied to local outlier detection. P. Puchhammer conceived the idea and methodology, derived the theoretical results, and implemented the algorithm in `R` R Core Team (2024) and `C++`, made available via the CRAN package `ssMRCD` (Puchhammer and Filzmoser, 2023). She further evaluated the method through simulation studies and a data example, and contributed to the writing and editing of the manuscript.

Chapter 3 applies the local outlier detection method developed in Puchhammer and Filzmoser (2024) to geochemical data in the context of mineral exploration. The suitability of local outlier detection methods for data of varying quality and scale is discussed. This chapter is based on Puchhammer, P., Kalubowila, C., Braus, L., Pospiech, S., Sarala, P., and Filzmoser, P. (2024a). A performance study of local outlier detection methods for mineral exploration with geochemical compositional

data. *Journal of Geochemical Exploration*, 258:107392. P. Puchhammer participated in conceptual discussions, co-developed the study design, extended the data analysis and results initially carried out by L. Braus, and contributed to the writing and editing of the manuscript.

Chapter 4 presents work based on Puchhammer, P., Wilms, I., and Filzmoser, P. (2024b). Sparse outlier-robust PCA for multi-source data. *arXiv preprint arXiv:2407.16299*. It focuses on sparse PCA methods for multi-group data. The development emphasizes joint sparsity patterns and covariance estimation, offering improved interpretability and insight into global and local patterns. P. Puchhammer contributed to the conceptual development in collaboration with her co-authors, implemented the algorithm in `R`, and tested the methodology through simulation studies and real data examples. She also contributed to the writing and editing of the manuscript.

Chapter 5 is based on the work Puchhammer, P., Wilms, I., and Filzmoser, P. (2025). A smooth multi-group Gaussian Mixture Model for cellwise robust covariance estimation. *arXiv preprint arXiv:2504.02547*. It introduces a cellwise robust Gaussian mixture model (GMM) for multi-group data, in which theoretical concepts are formalized and breakdown points are proven. The method also enables exploration of transition dynamics between groups by adapting the flexibility of the grouping structure. P. Puchhammer developed the methodology within the GMM framework, extended robust clustering theory to the multi-group context, implemented the method in `R`, conducted simulation studies, and contributed to the manuscript's writing and editing.

Chapter 7 concludes the thesis by summarizing the main findings and providing an outlook on future research directions.

2 Spatially Smoothed Robust Covariance Estimation for Local Outlier Detection

This chapter was published as Puchhammer, P. and Filzmoser, P. (2024). Spatially smoothed robust covariance estimation for local outlier detection. *Journal of Computational and Graphical Statistics*, 33(3):928–940. DOI: 10.1080/10618600.2023.2277875.

2.1 Introduction

The identification of multivariate outliers is probably one of the most important tasks in multivariate data analysis. A need to find outliers in order to make further analyses more reliable, or the direct interest in the outliers themselves motivate the numerous approaches available for multivariate outlier detection. The identified outliers are supposed to deviate to a certain extent from the main trend or structure of the data majority, and thus they are also called "global outliers" (Filzmoser et al., 2013). In contrast, the term "local outliers" refers to a setting where additional information regarding some kind of neighborhood is available, for example provided by spatial coordinates of the observations. Then, local outliers are observations which clearly differ from the multivariate measurements of their spatial neighbors indicating local anomalies that spark interest and make further analysis essential. Nevertheless, the values themselves might still be in an ordinary range of the data set, and thus the observation would not be outlying in a global sense.

Existing statistical approaches for multivariate local outlier detection are often based on a distance measure and neighborhood structure. A neighborhood a is defined as a subset of the set of observation indexes, say $\{1, \ldots, n\}$. A p-variate observation $\boldsymbol{x}_i$, for $i \in \{1, \ldots, n\}$, is defined to be in neighborhood a if and only if $i \in a$. The decision if some observation $\boldsymbol{x}$ is in a neighborhood is typically based on its spatial coordinates $s(\boldsymbol{x})$. One way to construct the spatial neighborhood is to take a spatial k-nearest-neighborhood of each point $\boldsymbol{x}$, where $k \in \mathbb{N}$. For a fixed $\boldsymbol{x}$, the spatial distance to another point $\boldsymbol{y}$ is defined as the Euclidean distance

$$d_{\boldsymbol{x}}(\boldsymbol{y}) = \|s(\boldsymbol{x}) - s(\boldsymbol{y})\| = \left[(s(\boldsymbol{x}) - s(\boldsymbol{y}))' (s(\boldsymbol{x}) - s(\boldsymbol{y}))\right]^{1/2}.$$

A spatial neighborhood $a_k(\boldsymbol{x})$ can then be defined as the set of the k many spatial nearest observations,

$$a_k(\boldsymbol{x}) = \{\boldsymbol{x}_j : \ d_{\boldsymbol{x}}(\boldsymbol{x}_j) \leq d_{\boldsymbol{x}(k)}\}, \tag{2.1}$$

where $d_{\boldsymbol{x}(1)} \leq d_{\boldsymbol{x}(2)} \leq \ldots \leq d_{\boldsymbol{x}(k)} \leq d_{\boldsymbol{x}(n)}$ are the sorted distances to all observations $\boldsymbol{x}_i, i = 1, \ldots, n$. Regarding local outlier detection, a distance measure often used to evaluate outlyingness is the pairwise Mahalanobis distance (MD). For a neighborhood a with covariance $\boldsymbol{\Sigma}_a$ and an observation $\boldsymbol{x}_i$ in neighborhood a, the pairwise MD is defined as

$$\mathrm{MD}_{\boldsymbol{\Sigma}_a}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left[(\boldsymbol{x}_i - \boldsymbol{x}_j)' \boldsymbol{\Sigma}_a^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)\right]^{1/2} \quad \text{for all } j \in a.$$

In general, the MD describes the distance between two observations, where the Euclidean distance in the feature space is adapted according to local distribution properties.

The estimation method used to determine $\boldsymbol{\Sigma}_a$ for all neighborhoods is key to good and reliable results. It is essential that outlying observations themselves are not affecting the estimation, since this could possibly mask outliers, leaving them undetected. Thus, a robust covariance estimation on the neighborhood level is necessary. One of the most widely used robust estimators for the covariance is the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1984, 1985), where one has to identify the h sub-sample of observations (where h is fixed e.g. to half of the observations) that minimize the determinant of its sample covariance. The MCD covariance estimator is then given by the sample covariance of the h subset, multiplied by a consistency factor (Croux and Haesbroeck, 1999). For its computation, the fastMCD algorithm developed by Rousseeuw and Driessen (1999) introduces an iterative concentration step, so-called C-step, that guarantees a decrease of the objective function until convergence to a (local) minimum, making the MCD estimator faster and even more popular. The global minimum is then approximated by iterating for a number of random starting values and choosing the smallest local minimum. By selecting a small number of good deterministic starting values for the fastMCD, the detMCD algorithm from Hubert et al. (2012) improves run-time even more. In spite of these excellent features of the MCD estimator, as well as affine equivariance and high robustness, one drawback is that the concept is not applicable in case of singularity of the sample covariance matrix of the h subset, which can easily occur. As for many methods, regularity of the estimated covariance is also needed to compute Mahalanobis distances. Especially in a setting where we are restricted to local neighborhoods consisting of a possibly small subset of observations, we might have a situation where regularity cannot be achieved and thus an inversion of the local covariance matrix is not possible. One solution is to base the local estimation on regularized robust covariance estimators, such as the recently developed Minimum Regularized Covariance Determinant (MRCD) estimator from Boudt et al. (2020) (or also on the Fritsch estimator, Fritsch et al. (2012)). One of the many attractive properties of the MRCD is that a slightly adapted fastMCD algorithm based on C-steps is also applicable.

Existing methods for local outlier detection have different ways to define the covariance matrix in order to ensure regularity. The method of Filzmoser et al. (2013) is dealing with regularity issues by using the MCD estimator calculated on the whole data set, i.e., $\boldsymbol{\Sigma}_a = \boldsymbol{\Sigma}$, thus imposing a global covariance structure. They have shown that for i.i.d. Gaussian random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the conditional distribution of the pairwise squared $\mathrm{MD}_{\boldsymbol{\Sigma}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$, for $j = 1, \ldots, n$,

given $\boldsymbol{x}_i$, is a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\mathrm{MD}^2_{\boldsymbol{\Sigma}}(\boldsymbol{x}_i, \boldsymbol{\mu})$. Instead of a fixed cut-off value for the pairwise MD, different sophisticated visual approaches are used. By plotting a degree of isolation based on the pairwise MD and quantiles of the non-central chi-square distribution, suspicious and highly isolated observations can be discovered and analyzed in more detail. Quite contrary to Filzmoser et al. (2013), the method of Ernst and Haesbroeck (2016) uses a very local covariance estimation by taking individual k-nearest-neighbor (kNN) neighborhoods for each point separately into account. To tackle the regularity issues, they use a regularized covariance estimation (originally the estimator from Fritsch et al. (2012), for the MRCD see also the adaptation made in Bellino et al. (2019)) for each individual kNN neighborhood. Additionally, they introduce the concept of the "next distance", which is also MD based, and use the upper fence of the adjusted boxplot of Hubert and Vandervieren (2008) of all next distances as a non-parametric cut-off value for detecting outliers.

Since it is not necessary to use a MD concept to find local outliers, a short detour to machine learning techniques might be interesting. One of the most prominent approaches for detecting multivariate local outliers in machine learning is the local outlier factor (LOF) developed by Breunig et al. (2000). Initially, locality refers to multivariate values and Euclidean distances in the feature space but this method can also be canonically adapted to spatial local outlier detection. In Schubert et al. (2012) this adaptation and further LOF-based approaches are discussed. Interestingly, also LOF and its variants are based on a concept of distance and neighborhoods.

The existing methods have shortcomings in various ways that have not yet been properly addressed. The rather global nature of the method of Filzmoser et al. (2013) leads to a reliable and robust estimation of the covariance. Nevertheless, it is somewhat questionable if the estimated covariance is applicable and representative for the covariance structure on the local level. Trying to solve this issue of missing locality in the estimation, Ernst and Haesbroeck (2016) resort to a very local approach by recalculating the covariance matrix for each observation separately. Although more locality is achieved, the method is not taking into account that covariance matrices are not likely to change abruptly from one neighborhood to a next one. Also, the number of estimated parameters is extremely high and based on rather few observations, even if the local covariance structure is abruptly changing. A more global estimation of the local covariance matrices might be more stable and reliable and might also avoid repetitive calculations. It seems that until now there are only two extremes regarding locality of the covariance estimation available.

We bridge the gap between the fully global and the fully local approach by providing a covariance estimator based on the MRCD that addresses the missing locality on the one hand and the missing spatial smoothness on the other. By providing the possibility to set the amount of spatial smoothing and the size of the neighborhoods we get a generalization of the two detection methods, with the goal that good outlier detection properties based on the new local covariance estimations are achieved. Moreover, the covariance estimate can be seen as a generalization of the MRCD when the data set has additional sub-structures.

The paper is organized as follows. In Section 2 we introduce the new covariance

estimator, derive its properties as well as properties of the original MRCD and establish the methodology to detect local outliers. An algorithm and the derivation of a generalized C-step are discussed in Section 3. Section 4 provides simulation results regarding run time, convergence and outlier detection, while in Section 5, a real data set is analyzed using the newly developed local outlier detection method. Finally, the main results are presented and summarized in the conclusions.

2.2 Methodology

2.2.1 Spatially Smoothed MRCD Estimators

Assume that the p-dimensional observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, for $i = 1, \ldots, n$, are arranged as rows in the $n \times p$ matrix $\boldsymbol{X}$ with $n > 2p$. Furthermore, each observation has spatial information available, e.g. spatial coordinates, and is assigned to one of N many neighborhoods, defined by the index sets $a_1, \ldots, a_N$ of size $n_1, \ldots, n_N$.

The goal of the proposed method is to obtain local covariance estimates for each neighborhood that are suitable for calculating the pairwise MDs and to some extent smooth among nearby neighborhoods. Since the MCD estimator requires at least $n > 2p$ to provide a regular covariance matrix, which is a severe limitation especially for small neighborhoods, we focus on its regularized extension, the MRCD estimator. Instead of minimizing the determinant of the sample covariance matrix as in the MCD case, the minimization objective is the determinant of a convex combination of the sample covariance and a symmetric and positive definite matrix, the so-called target matrix $\boldsymbol{T}$. Boudt et al. (2020) suggest a data-driven approach based on the condition number of the covariance matrix to set the degree of regularization ρ which is used in the convex combination. Regarding the target matrix $\boldsymbol{T}$, it is sensible to choose a robust and regular covariance, e.g. a diagonal matrix based on univariate robust scale estimates.

In the following, we adapt the idea of the MRCD estimator to our setting of local and smooth covariance estimation. Let $\mathcal{H} = (H_1, \ldots, H_N)$ be subsets of the index sets $a_1, \ldots, a_N$ defining the neighborhoods. The size of each subset H_i is $h_i = |H_i| = \lceil \alpha n_i \rceil$, for $i = 1, \ldots, N$, where α is selected in the interval $[0.5, 1]$, and $\lceil \cdot \rceil$ is the ceiling function, rounding up to the next integer. A smaller value of α will result in more robustness against outliers, and it would also be possible to adjust this value to each neighborhood individually. The observations of subset H_i are written as matrix $\boldsymbol{X}_{H_i}$, with dimensionality $h_i \times p$. Let the neighborhood specific MRCD-based covariance matrix $\boldsymbol{K}_i(\mathcal{H})$, for $i = 1, \ldots, N$, be defined as

$$\boldsymbol{K}_i(\mathcal{H}) = \rho_i \boldsymbol{T} + (1 - \rho_i) c_\alpha Cov(\boldsymbol{X}_{H_i}), \tag{2.2}$$

with $Cov(\boldsymbol{Y})$ being the sample covariance matrix of $\boldsymbol{Y}$, and c_α a consistency factor for the proportion α (see Croux and Haesbroeck, 1999). The regularization parameter ρ_i is set individually for each neighborhood, and it could also be chosen as zero if the estimated covariance matrix is already invertible. Finally, since we want to smooth the covariance matrices, it seems counter intuitive to choose neighborhood specific target

matrices, which would also require more parameter estimations. Therefore, we assume a global target matrix $\boldsymbol{T}$, taken as a robust and regular covariance matrix estimated based on the full data set $\boldsymbol{X}$. Since we assume $n > 2p$ we propose to use the MCD estimator for $\boldsymbol{X}$ as target matrix.

We want to find the combination of subsets in $\mathcal{H}$ that minimizes the objective function

$$f(\mathcal{H}) = \sum_{i=1}^{N} \det \left((1-\lambda)\boldsymbol{K}_i(\mathcal{H}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}) \right). \tag{2.3}$$

The tuning parameter $\lambda \in [0, 1]$ is used to balance the influence of an individual local neighborhood and the remaining neighborhoods in the covariance estimations. In case of $\lambda = 0$, there is no spatial influence at all which is equivalent to the estimation of the MRCD for each neighborhood separately while using a global target matrix. For the other extreme $\lambda = 1$, the covariance matrix in a specific neighborhood is an average over the surrounding covariance estimates without adding local information from the neighborhood itself. Moreover, it is possible to interpret the second part in the determinant as a penalization term. Due to the minimization of the determinant, observations from a_i that match well with the main trend of observations in neighborhoods with positive weights ω_{ij} are more likely to be in the optimal H-set if λ is increased. The weights ω_{ij} are supposed to be non-negative, and we set $\omega_{ii} = 0$. The elements of the weight vector $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{iN})'$ indicate the relative influence that the estimated covariances of other neighborhoods have on the covariance estimation of the i-th neighborhood. Also, each weight vector has to sum up to one, $\sum_{j=1}^{N} \omega_{ij} = 1$ for all $i = 1, \ldots, N$. All these weights need to be pre-specified, for example based on inverse geographical distances, and are collected as rows in the weighting matrix $\boldsymbol{W} \in \mathbb{R}^{N \times N}$.

Note that for the objective function (2.3) a global minimum $\mathcal{H}^* = (H_i^*)_{i=1,\ldots,N}$ exists since its domain consists of a finite number of subset combinations. For this global minimum, the estimated covariance matrix for each neighborhood a_i is

$$\hat{\boldsymbol{\Sigma}}_{SSM,i} = (1-\lambda)\boldsymbol{K}_i(\mathcal{H}^*) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^*), \tag{2.4}$$

and the location estimate $\hat{\boldsymbol{\mu}}_{SSM,i}$ is the sample mean of the selected observations $\boldsymbol{X}_{H_i^*}$. We call these estimators the spatially smoothed MRCD (ssMRCD) location and covariance estimators.

Although the neighborhood structure and the value of λ will often depend on the data at hand, there are some sensible and natural choices for $\boldsymbol{W}$. If we have a neighborhood structure that has no further meaning and might just be used to divide the spatial space into subsets, an inverse-distance based weight matrix, also used in Sections 2.4 and 2.5, might be a good choice. Other possibilities include binary matrices with ones if neighborhoods share a border and zero otherwise, with rows scaled appropriately. Moreover, the regularity parameters can be set by default. For a neighborhood a_i we suggest to set the regularization parameter ρ_i as the data driven value that is proposed

by the MRCD algorithm in Boudt et al. (2020), when interpreting the neighborhood as its own data set.

2.2.2 Theoretical Properties

In the following we will show that the spatially smoothed MRCD estimators proposed here are – in contrast to the original MRCD estimator – affine equivariant, and we derive its breakdown point. For a data set $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, a location and covariance estimator are called affine equivariant for all neighborhoods if for any non-singular matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, any vector $\boldsymbol{b} \in \mathbb{R}^p$, and all $i = 1, 2, \ldots, N$, it holds that

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{SSM,i}(\boldsymbol{X}\boldsymbol{A}' + \mathbf{1}_n\boldsymbol{b}') &= \hat{\boldsymbol{\mu}}_{SSM,i}(\boldsymbol{X})\boldsymbol{A}' + \boldsymbol{b}', \\ \hat{\boldsymbol{\Sigma}}_{SSM,i}(\boldsymbol{X}\boldsymbol{A}' + \mathbf{1}_n\boldsymbol{b}') &= \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{SSM,i}(\boldsymbol{X})\boldsymbol{A}'. \end{aligned} \tag{2.5}$$

Theorem 2.2.2.1 (Affine equivariance)**.** *Let $\boldsymbol{T}$ be any robust, regular and affine equivariant estimate of the covariance for the data set $\boldsymbol{X}$, here denoted as $\boldsymbol{T}(\boldsymbol{X})$. Then, the spatially smoothed MRCD estimators of location and covariance with target matrix $\boldsymbol{T}(\boldsymbol{X})$ are affine equivariant.*

Proof. The proof is given in the appendix. □

The assumptions of Theorem 2.2.2.1 for $\boldsymbol{T}(\boldsymbol{X})$ can be fulfilled by the MCD applied to the full data set $\boldsymbol{X}$ for $n > 2p$. Nevertheless, any robust estimator satisfying the assumptions can be used, e.g. S-estimators (Rousseeuw and Leroy, 1987). Note that for local outlier detection tasks we typically have enough observations globally to get regularity with standard robust estimators. As a remedy if regularity is not achievable, the MRCD (or e.g. the OGK estimator of Maronna and Zamar (2002)) can be used. Assuming that the target matrix can be estimated in a robust, regular and affine equivariant way representing a covariance, the MRCD would also be affine equivariant. However, this might question the usefulness of the MRCD since we already have a robust, regular and affine equivariant covariance estimator available, namely $\boldsymbol{T}(\boldsymbol{X})$. Therefore, in typical application scenarios, the ssMRCD is affine equivariant whereas the original MRCD is not. Anyhow, in the case of global high-dimensionality, i.e. $n \leq p + 1$, affine equivariance can only be achieved by the non-robust sample mean and covariance (Tyler, 2010). Neither the MRCD nor the ssMRCD can then be affine equivariant.

Another important property of robust estimators is the finite sample breakdown point, which is defined as the minimal fraction of points that need to be exchanged in order to make the estimators useless. Before considering the spatially smoothed MRCD we have to derive the breakdown point of the original MRCD without prior scaling of the observations, from now on called *raw MRCD*. The breakdown point of a location estimator $\hat{\boldsymbol{\mu}}_n$ is formally defined as

$$\epsilon_n^*(\hat{\boldsymbol{\mu}}_n; \boldsymbol{X}_n) = \frac{1}{n}\min\{m : \sup ||\hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_{n,m}) - \hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_n)|| = +\infty\},$$

where $\boldsymbol{X}_{n,m}$ is the data matrix $\boldsymbol{X}_n$ with m-many observations exchanged with arbitrary values (Maronna et al., 2006).

For the covariance estimate $\hat{\boldsymbol{\Sigma}}_n$ the finite sample breakdown point is defined as

$$\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_n; \boldsymbol{X}_n) = \frac{1}{n} \min\{m : \sup \max_j |\ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))) - \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_n)))| = +\infty\},$$

with $\lambda_1(\boldsymbol{\Sigma}), \ldots, \lambda_p(\boldsymbol{\Sigma})$ denoting the eigenvalues of a matrix $\boldsymbol{\Sigma}$ in decreasing order. Since the eigenvalues are sorted, we only have to consider the biggest eigenvalue $\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m})))$ which might explode when exchanging observations with arbitrary values (*explosion breakdown point*) and the smallest eigenvalue $\lambda_p(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m})))$ which might become zero (*implosion breakdown point*) and thus implies singularity (Maronna et al., 2006).

Theorem 2.2.2.2. *Consider the raw MRCD estimator with fixed $\rho > 0$, regular and fixed $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ and the data matrix $\boldsymbol{X}_n$. Then, the following statements hold:*

a. The MRCD location estimator $\hat{\boldsymbol{\mu}}_n$ has the finite sample breakdown point $\min(h, n-h+1)/n$.

b. The MRCD covariance estimator $\hat{\boldsymbol{\Sigma}}_n$ has the finite sample explosion breakdown point $(n-h+1)/n$.

c. The MRCD covariance estimator $\hat{\boldsymbol{\Sigma}}_n$ has the finite sample implosion breakdown point 1.

Proof. The proof is given in the appendix. □

Regarding the finite sample breakdown point of location and covariance estimators with multiple estimators let us define the finite sample breakdown point ϵ_n^* as the minimal fraction of points of the same arbitrary neighborhood that need to be exchanged in order to make at least one of the estimators useless. For the location estimators $\hat{\boldsymbol{\mu}}_{SSM,n,i}$, $i = 1, \ldots, N$, the formal definition is

$$\epsilon_n^*((\hat{\boldsymbol{\mu}}_{SSM,n,i})_{i=1}^N; \boldsymbol{X}_n) = \min_{i=1,\ldots,N} \frac{1}{n_i} \min\{m : \sup ||\hat{\boldsymbol{\mu}}_{SSM,n,i}(\boldsymbol{X}_{n,m}^i) - \hat{\boldsymbol{\mu}}_{SSM,n,i}(\boldsymbol{X}_n)|| = +\infty\},$$

where $\boldsymbol{X}_{n,m}^i$ is the matrix $\boldsymbol{X}_n$ with m-many observations of neighborhood a_i exchanged with arbitrary values.

For the breakdown points of the ssMRCD estimator, we restrict the values of the parameters $\boldsymbol{W}$ and λ to exclude the following special case. In general it is possible that all entries of one column of the weight matrix are zero for (at least) one neighborhood a_i ($\omega_{ij} = 0$ for all $j = 1, \ldots, N$) meaning that neighborhood a_i does not contribute to spatial smoothing for any other neighborhood. If additionally $\lambda = 1$, the observations and the MRCD-based covariance matrix of neighborhood a_i (see Equation (2.2)) do also not contribute to the estimation of the ssMRCD-covariance estimate of neighborhood a_i itself. Thus, the values of the observations in neighborhood a_i do not affect the estimation in any way implying that it is not sensible to include this neighborhood in the calculation of the breakdown point. Therefore, in the following theorems of this section we assume that there is at least one non-zero entry per column of $\boldsymbol{W}$ if $\lambda = 1$.

Theorem 2.2.2.3. *The location estimators* $(\hat{\boldsymbol{\mu}}_{SSM,n,i})_{i=1}^{N}$ *of the spatially smoothed MRCD have the finite sample breakdown point*

$$\epsilon_n^*((\hat{\boldsymbol{\mu}}_{SSM,n,i})_{i=1}^{N}; \boldsymbol{X}_n) = \min_{i=1,\dots,N} \min(n_i - h_i + 1, h_i)/n_i.$$

Proof. The proof is given in the appendix. □

For the covariance estimates $\hat{\boldsymbol{\Sigma}}_{SSM,n,i}$, $i = 1, \dots, N$, the finite sample breakdown point is defined accordingly,

$$\epsilon_n^* \left(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i} \right)_{i=1}^{N}; \boldsymbol{X}_n \right) =$$
$$\min_{i=1,\dots,N} \frac{1}{n_i} \min\{m : \sup_j \max | \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}(\boldsymbol{X}_{n,m}^i))) - \ln(\lambda_j(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}(\boldsymbol{X}_n)))| = +\infty\}.$$

Again, we can differentiate between explosion and implosion breakdown point.

Theorem 2.2.2.4. *Given a fixed and regular target matrix* $\boldsymbol{T}$*, the finite sample implosion breakdown point of the spatially smoothed MRCD covariance estimators* $\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i} \right)_{i=1}^{N}$ *is equal to*

$$\epsilon_n^*(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i} \right)_{i=1}^{N}; \boldsymbol{X}_n) = 1.$$

The finite sample explosion breakdown point is

$$\epsilon_n^*(\left(\hat{\boldsymbol{\Sigma}}_{SSM,n,i} \right)_{i=1}^{N}; \boldsymbol{X}_n) = \min_{i=1,\dots,N} (n_i - h_i + 1)/n_i.$$

Proof. The proof is given in the appendix. □

Note that for all the breakdown properties of the original and the spatially smoothed MRCD, the target matrix $\boldsymbol{T}$ is assumed to be regular and fixed. In applications the target matrix would be some covariance estimator $\boldsymbol{T}(\boldsymbol{X})$ with its own breakdown point. Then, the explosion breakdown point of the ssMRCD with estimated target matrix is the minimum of the two breakdown points, and the implosion breakdown point is 1 (see online appendix, after proof of Theorem 2.2.2.4).

2.2.3 Local Outlier Detection

The final step for detecting outliers is to decide how the spatially smoothed covariances available for neighborhoods a_i, $i = 1, \dots, N$, will be linked to the pairwise MD.

The method is based on Ernst and Haesbroeck (2016). In order to compare each observation $\boldsymbol{x}$ with its local neighbors we need a second neighborhood structure that provides spatially close neighbors in contrast to the structural neighborhoods a_i that are used for the smoothed covariance estimation. Thus, we select some $k \in \mathbb{N}$ and calculate the spatial k nearest neighbors, $b_k(\boldsymbol{x})$, see also Definition (2.1), where a

typical value might be $k = 10$. However, when applying the method, the spatial structure of the data set should also be considered.

For each observation $\boldsymbol{x} \in a_i$, the *next distance* is defined as

$$d(\boldsymbol{x}) = min_{\boldsymbol{y} \in b_k(\boldsymbol{x})} \left[(\boldsymbol{x} - \boldsymbol{y})' \hat{\boldsymbol{\Sigma}}^{-1}_{SSM,i} (\boldsymbol{x} - \boldsymbol{y}) \right]^{1/2},$$

which is the minimum of all pairwise MDs based on the covariance matrix $\hat{\boldsymbol{\Sigma}}_{SSM,i}$ with all observations in the spatial k nearest neighborhood $b_k(\boldsymbol{x})$. The neighborhood $b_k(\boldsymbol{x})$ is not necessarily a subset of a_i. However, due to the spatial smoothing of the covariance matrices and the spatial correlation of regular observations, an observation $\boldsymbol{y}$ spatially close to $\boldsymbol{x}$ should be similar enough to not be classified as outlier even if the covariance matrix of another but close neighborhood a_i is used. In the case of a strong difference between $\boldsymbol{x}$ and $\boldsymbol{y}$, the observation $\boldsymbol{y}$ would still be classified as outlying.

The next distance $d(\boldsymbol{x})$ is used as a measure of outlyingness. If the next distance is high, none of the observations in the spatial k nearest neighborhood is similar to the observation $\boldsymbol{x}$, which means that $\boldsymbol{x}$ would be flagged as a local outlier. As a notion what values for the next distance are considered as high, a non-parametric cut-off value can be used based on the upper fence of the adjusted boxplot (Hubert and Vandervieren, 2008) using all next distances from all neighborhoods a_i, $i = 1, \ldots, N$. Observations above the cut-off value are considered to be local outliers.

Possible further extensions like the restriction to homogeneous neighborhoods as suggested in Ernst and Haesbroeck (2016) are not included but are interest of future research.

2.3 Algorithm and C-Step

For computing the spatially smoothed MRCD location and covariance estimators we need to minimize the objective function (2.3). However, since the number of possible combinations of subsets is comparable with the MCD or the MRCD, it is in general not feasible to just calculate the value of the objective function for all these combinations and select the best one. Instead, the strategy of C-steps (concentration steps) introduced for the MCD estimator by Rousseeuw and Driessen (1999) will be adapted to this problem setting. Given an index set H_1 corresponding to h observations of the data matrix $\boldsymbol{X}$, the C-step chooses the subsequent subset H_2 where the h observations with the smallest Mahalanobis distances, based on the arithmetic mean and sample covariance matrix of the observations from H_1, are taken. Rousseeuw and Driessen (1999) have shown that this procedure converges to a local minimum. The C-step idea has also been extended for the MRCD (Boudt et al., 2020), and we will now adapt the generalized C-step to our setting.

Theorem 2.3.0.1. *For each $j = 1, \ldots, N$, let $\rho_j \in (0, 1)$ and H^0_j be any starting subset of a_j of respective size $h_j \in (n_j/2, n_j)$. Let α_j be the corresponding fraction of the observations used, $\alpha_j = h_j/n_j$. Let $\mathcal{H}^0 = (H^0_1, \ldots, H^0_N)$ be the combination of the subsets and $\lambda \in [0, 1)$ fixed. The target matrix $\boldsymbol{T}(\boldsymbol{X})$ is assumed to be positive definite, symmetric and fixed, and $\boldsymbol{K}_j(\mathcal{H}^0)$ is defined as in Equation (2.2) with $\boldsymbol{T} =$*

$\boldsymbol{T}(\boldsymbol{X})$. *Calculate the sample mean for each neighborhood* $a_j, j = 1, \ldots, N$, *based on the respective subset,* $\boldsymbol{m}_j(\mathcal{H}^0) = \frac{1}{h_j} \sum_{k \in H_j^0} \boldsymbol{x}_k$.

Fix neighborhood a_i. *For* $\boldsymbol{x} \in a_i$, *let the MD-based measure with the subset given by* $\mathcal{H}^0$ *be defined as*

$$d(\boldsymbol{x}; \mathcal{H}^0) = (\boldsymbol{x} - \boldsymbol{m}_i(\mathcal{H}^0))' \left[(1-\lambda)\boldsymbol{K}_i(\mathcal{H}^0) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right]^{-1} (\boldsymbol{x} - \boldsymbol{m}_i(\mathcal{H}^0)).$$

For a new subset H_i^1 *of* a_i *of size* h_i *with*

$$\sum_{k \in H_i^1} d(\boldsymbol{x}_k; \mathcal{H}^0) \leq \sum_{k \in H_i^0} d(\boldsymbol{x}_k; \mathcal{H}^0),$$

denote $\tilde{\mathcal{H}} = (H_1^0, \ldots, H_i^1, \ldots, H_N^0)$ *(note that* $\boldsymbol{K}_j(\tilde{\mathcal{H}}) = \boldsymbol{K}_j(\mathcal{H}^0)$ *for* $j \neq i$*). Then,*

$$\det \left((1-\lambda)\boldsymbol{K}_i(\tilde{\mathcal{H}}) + \lambda \sum_{\substack{j=1, \\ j\neq i}}^{N} \omega_{ij} \boldsymbol{K}_j(\tilde{\mathcal{H}}) \right) \leq \det \left((1-\lambda)\boldsymbol{K}_i(\mathcal{H}^0) + \lambda \sum_{\substack{j=1, \\ j\neq i}}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right)$$

with equality if and only if $\boldsymbol{K}_i(\tilde{\mathcal{H}}) = \boldsymbol{K}_i(\mathcal{H}^0)$ *and* $\boldsymbol{m}_i(\tilde{\mathcal{H}}) = \boldsymbol{m}_i(\mathcal{H}^0)$.

Proof. The proof is given in the appendix. □

The C-step theorem states that the objective function will decrease with every step as long as the other covariance estimators stay fixed. In the implemented algorithm described below, this will in general not be the case. However, the theorem and its proof should be sufficient to motivate and provide a reason for the algorithm proposed in the following.

The algorithm makes use of the C-step to solve the minimization problem based on Boudt et al. (2020). Suppose that we can estimate the target matrix $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ by the affine equivariant MCD estimator, then we also obtain affine equivariance for the spatially smoothed MRCD. Thus, we can ignore the scaling step in Boudt et al. (2020) and reduce the number of parameter estimations. Using an eigen-decomposition of $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$, with $\boldsymbol{Q}$ containing the eigenvectors as columns, and $\boldsymbol{\Lambda}$ being the diagonal matrix of positive eigenvalues, we transform the observations,

$$\boldsymbol{z}_i = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}'\boldsymbol{x}_i, \tag{2.6}$$

for $i = 1, \ldots, n$. Thus, in the next steps we can use $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ as data matrix and $\boldsymbol{T} = \boldsymbol{I}_p$ as target matrix, which numerically simplifies the data-driven selection procedure for ρ_j .

In order to start the iteration process by making use of the C-steps, we need starting values, which should be robust and regular covariance estimates for each neighborhood. As proposed for the original MRCD estimator, we will also use the deterministic MCD algorithm of Hubert et al. (2012) for each neighborhood separately. This approach

results in six estimates for location and scatter for each neighborhood, which show especially good convergence properties. Furthermore, for each neighborhood a_j, the value of ρ_j is calculated using the data-driven selection procedure based on a maximal condition number according to steps 3.2 to 3.4 from Boudt et al. (2020). The set of six deterministic starting values for each neighborhood is restricted to those with a sufficiently small condition number (for more details see step 3.5. in Boudt et al., 2020).

One new issue that arises is the number of possible combinations of initial subsets: for N neighborhoods we end up with up to 6^N subset combinations as possible starting values. Since a high number of starting values might not be essential for a good approximation, we will consider an upper limit of $6N$ starting values in the following, and they will be selected at random out of the possible combinations that are left after the ρ-selection step. While accuracy can be increased with additional starting values, the restriction to $6N$ leads to computational feasibility for the algorithm and still provides reliable estimates as can be seen in the performance of local outlier detection in Section 2.4.

Suppose now that we start the procedure with an initial subset $\mathcal{H}^0 = (H_1^0, \ldots, H_N^0)$, then we apply the C-step for each neighborhood a_i and obtain a new subset H_1^i. The combination of these subsets $\mathcal{H}^1 = (H_1^1, \ldots, H_N^1)$ is used as starting point for the next iteration step, etc. After there is no change in the subsets, the iteration process stops (see also Figure A.8 in the online supplements). After applying the C-step iterations for all starting values, we choose the subset combination with the smallest objective function value as the final subset combination $\mathcal{H}^* = (H_1^*, \ldots, H_N^*)$.

Although Theorem 2.3.0.1 is not proving that the objective function decreases with every step due to the additional covariance matrices being adapted separately for each neighborhood after each iteration step, simulation results show that the algorithm provides stable results and good monotonic behavior in most cases (see Section 2.4).

After receiving the final combination of subsets $\mathcal{H}^*$ for each neighborhood, the matrices $\boldsymbol{K}_i^Z$ are back-transformed to

$$\boldsymbol{K}_i^* = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}\left[\boldsymbol{K}_i^Z(\mathcal{H}^*)\right]\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}'. \tag{2.7}$$

The covariance estimate for neighborhood a_i is then

$$\hat{\boldsymbol{\Sigma}}_{SSM,i} = (1-\lambda)\boldsymbol{K}_i^* + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij}\boldsymbol{K}_j^*, \tag{2.8}$$

and the location estimate is the arithmetic mean of the optimal subset $\boldsymbol{X}_{\mathcal{H}_i^*}$. The detailed numerical procedure is summarized as pseudo-code in Algorithm 1.

Regarding the tuning parameter λ there is no standard procedure to get a good value that is similarly automated like the calculation procedure for the regularization parameters ρ_i. However, there are multiple possibilities demonstrated in Section 2.5 that can be used to choose λ in applications.

Algorithm 1 Algorithm for the spatially smoothed MRCD estimator.

Step 1.1: Calculation of target matrix $\boldsymbol{T}$ using the MCD estimator on $\boldsymbol{X}$
Step 1.2: Eigen-decomposition of $\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ and transform observations according to Equation (2.6)
Step 2: Initialization step according to steps 3.1 to 3.5 (without C-step) from Boudt et al. (2020) for each neighborhood
for $i = 1, \ldots, N$ **do**
    Fix neighborhood a_i
    Get 6 initial deterministic sets of h_i observations from a_i according to Hubert et al. (2012)
    Calculate 6 initial covariance matrices and mean estimates
    Select neighborhood-specific ρ_i via data-driven approach
    Filter subset of initial starting estimates according to condition number (step 3.5 from Boudt et al., 2020)
end for
Select set of initial h-set combinations as starting values at random
Step 3.1: C-step: For each initial combination of subsets, iterate until convergence
Step 3.2: Select best combination of subset based on objective function value
Step 4: Calculate $\boldsymbol{K}_i(\mathcal{H}^*)$ $\forall i$ and use Equations (2.7) and (2.8) to get final estimates

2.4 Numerical Simulations

In order to test the new method, two simulation setups are constructed which also incorporate the neighborhood structures necessary for the covariance estimation and outlier detection. For both setups we simulate covariance matrices which depend on a parameter δ, denoted by $\boldsymbol{\Sigma}(\delta)$, where the entries are defined as $(\boldsymbol{\Sigma}(\delta))_{jk} = \delta^{(j-k)}$, for $j, k \in \{1, \ldots, p\}$, leading to positive definiteness and symmetry.

Setup 1: The first setup is inspired by the original idea of covariance matrices smoothly transforming over space. We start by setting up the (two-dimensional) coordinates (s_i^1, s_i^2) of the observations $\boldsymbol{x}_i$ with $n_{sim} = 41$ observations per coordinate axis, evenly spread between 0 and 20, resulting in $n = n_{sim}^2 = 1681$ data points in total.

For N_{sim} many areas a_{lm}, $l, m \in \{1, \ldots, \sqrt{N_{sim}}\}$, of similar observations we construct a second spatial grid with each cell consisting of n_{sim}/N_{sim} observations on average. The borders of the areas for the first coordinate are defined as $b_l^1 = l\frac{20}{\sqrt{N_{sim}}}$ for $l = 0, \ldots, \sqrt{N_{sim}}$. Analogously, the borders for the second coordinate are defined as $b_m^2 = m\frac{20}{\sqrt{N_{sim}}}$ for $m = 0, \ldots, \sqrt{N_{sim}}$, leading the an evenly spaced grid. Thus, the area a_{lm} consists of observations $\{\boldsymbol{x}_i : s_i^1 \in (b_{l-1}^1, b_l^1] \cap s_i^2 \in (b_{m-1}^2, b_m^2]\}$ for $l, m \neq 0$. For either l or m equal to 1, the left edge of the interval (which would be zero) is included. The coordinate centers of the area a_{lm} are defined as $c_{lm}^1 = \frac{b_{l-1}^1 + b_l^1}{2}$ and $c_{lm}^2 = \frac{b_{m-1}^2 + b_m^2}{2}$.

The observed values of observations in area a_{lm} are then randomly drawn from a p-

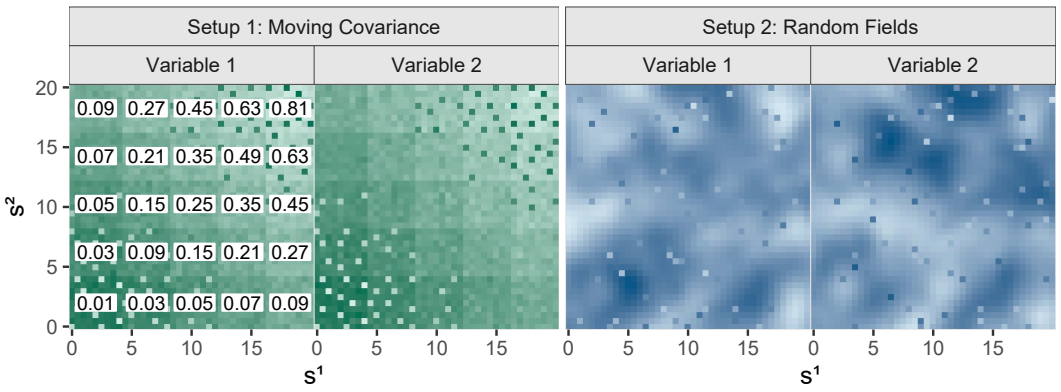


Figure 2.1: Simulation scenarios with $p = 2$ and a 5% contamination rate. On the left hand side the simulation setup 1 is presented with contamination achieved through the swapping process described in Ernst and Haesbroeck (2016), $N_{sim} = 25$ and $n_{sim} = 41$. The values printed on the left-most panel are corresponding to the parameter δ_{lm}. On the right hand side, setup 2 with $\nu = 3$ is shown with completely random swapping.

dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_{lm}, \boldsymbol{\Sigma}(\delta_{lm}))$, where $\boldsymbol{\mu}_{lm} := ((c^1_{lm}+c^2_{lm})/2, \ldots, (c^1_{lm}+c^2_{lm})/2) \in \mathbb{R}^{p\times 1}$, thus having entry values between 0 and 20. For the covariance matrix $\boldsymbol{\Sigma}(\delta_{lm})$ for areas a_{lm} we use the structure described above with parameter δ_{lm} defined as

$$\delta_{lm} = \left(0.1 + l\frac{0.9-0.1}{\sqrt{N_{sim}}}\right)\left(0.1 + m\frac{0.9-0.1}{\sqrt{N_{sim}}}\right) \in [0.01, 0.81],$$

increasing smoothly from the left bottom to the right upper corner. A simulated data set with $p = 2$ is presented in Figure 2.1 (left) where the grid structure and the change of the mean are clearly visible. The resulting δ_{lm} for each area is shown as well.

Setup 2: To get a more flexible simulation setup, a random field, specifically the parsimonious multivariate Matérn model (see Gneiting et al., 2010; Ernst and Haesbroeck, 2016), is used as suggested in Ernst and Haesbroeck (2016) and Harris et al. (2013). Instead of the constructed matrices used in Ernst and Haesbroeck (2016) and Harris et al. (2013) we again choose a matrix structure $\boldsymbol{\Sigma}(\delta)$ as described above with $\delta = 0.7$, since it can be extended for higher dimensions. For δ being set to 0.7 we get a range of high to low correlations reflecting different relationships between variables. The spatial smoothness is assumed to be the same for all variables, and it is regulated by one smoothness parameter ν, which is taking values in $\{0.5, 1.5, 3\}$. A higher ν leads to more spatial smoothness and in general more distinct outliers after contamination. The spatial scale parameter a of the Matérn model is set to one. A grid structure for the coordinates of the observations with values ranging from 0 to 20 with grid size 0.5 is imposed, leading to $41^2 = 1681$ observations overall, similar to the standard setting in setup 1.

Lastly, contamination with outliers is achieved by swapping coordinates of observations with each other. Filzmoser et al. (2013) exchange observations that are completely randomly chosen, whereas Ernst and Haesbroeck (2016) propose to swap the most

extreme observations regarding the first score of the global robust principal components. In order to avoid the problem of exchanging whole areas of observations with each other due to high spatial correlation, once observations are swapped, their 15 closest neighbors are removed from the swapping process. This leads to a clear distinction of outlying observations without the possibility of other outliers being close (see also Figure 2.1). Thus, swapping according to Ernst and Haesbroeck (2016) should in general result in a better performance for all considered methods. Both swapping approaches in both setups will be analyzed with a varying contamination level β between 1% and 15%.

For the ssMRCD covariance estimation we impose a grid based neighborhood structure. Similar to the description of setup 1 we use evenly spaced borders and assign the observations to neighborhoods n_i, for $i = 1, \ldots, N$. Thus, the case $N = N_{sim}$ in setup 1 depicts a perfect match of the neighborhoods selected for the ssMRCD and the real underlying covariance structure. For $N_{sim} > N$ the ssMRCD uses less neighborhoods leading to more smoothness of the covariance estimation. The weighting matrix $\boldsymbol{W}$ is based on the inverse (Euclidean) distance of the centers, that are defined equivalently to setup 1.

We will focus on the suitability of the ssMRCD covariances for local outlier detection and refer to the online supplement for additional analysis regarding computational efficiency and convergence behavior of the algorithm described in Section 2.3. We compare its performance to the local outlier detection methods of Filzmoser et al. (2013) (F), Ernst and Haesbroeck (2016)(EH) and the local outlier factor methodology of Breunig et al. (2000), canonically adapted to spatial neighborhoods as described in Schubert et al. (2012) (LOF). Both simulation setups vary in the parameters p, ν and N_{sim}, respectively, and both swapping processes are used, each combination is simulated 100 times. All methods considered compare each observation to k many of its neighboring observations which we will assume to be equal to 10 for all methods. For the ssMRCD we will assume the default values ($\lambda = 0.5$, $N = 25$, and $\boldsymbol{W}$ based on inverse-distances) arrived from the parameter sensitivity analysis included in the online supplement.

The outlier classification method of Filzmoser et al. (2013) has a parameter β_F which is the percentage of neighboring observations that a local outlier is allowed to be similar to. Here we use the value $\beta_F = 0.1$, as proposed in Ernst and Haesbroeck (2016), meaning that $0.1k = 1$ observation is allowed to be similar within the 10 nearest neighbors. For inliers the expected value of the isolation degree is β_F. If the actual degree of isolation is higher than the expected value, this signals local outlyingness. As cutoff for classifying an observation as local outlier we use twice the value of β_F, so 20%. This cutoff is less strict than in the simulation setup from Ernst and Haesbroeck (2016) who take a cutoff of three times the expected value, i.e. 30%. Note that the methodology of Filzmoser et al. (2013) mostly focuses on visual outlier detection tools, so the cutoff value chosen here might not be optimal.

For the regularized spatial detection technique by Ernst and Haesbroeck (2016), the parameter β_{EH}, which gives the fraction of the most homogeneous neighborhoods included in the outlier detection procedure, is set to one. In the simulations we are interested in all outliers, and the heterogeneity in the simulated data should be

comparable for all of the observations. Thus, only considering a fraction leads to non comparable results. Moreover, the simulation results of Ernst and Haesbroeck (2016) show that over all considered setups, $\beta_{EH} = 1$ is also optimal. As regular covariance matrix estimator we use the MRCD with the default target matrix (equi-correlated target matrix) and $\alpha = 75\%$.

Last but not least, the non-parametric LOF, which calculates a local outlier factor for each observation based on a comparison of the so-called "local reachability density" with its k-nearest neighbors, needs a cutoff value. Since there is no fixed rule on how to choose a cutoff value, a local outlier factor above 1.5 determines an outlier. This value is also used in the original paper of Breunig et al. (2000).

The results for both simulation setups and the completely random switching are shown in Figures 2.2 and 2.3. For the results regarding the swapping method of Ernst and Haesbroeck (2016) shown in Figures A.13 and A.14 as well as a comparison of the four methods in terms of computational efficiency we refer to the online supplement.

Starting with the false positive rate (FPR), we see that the method of Ernst and Haesbroeck (2016) has some issues with classifying too many normal observations as outliers in nearly all settings. This is likely due to the very local covariance estimation which might be too strict in general leading to a strong swamping effect, especially in settings where there is no strong spatial correlation of the observed values. Interestingly, the behavior of the FPR for the method of Filzmoser et al. (2013) depends on the data simulation setup. For the moving matrix scenarios, the FPR is rather high, for random fields it is very low. The LOF and the ssMRCD-based outlier detection method have reliable low FPR for all scenarios.

The outcome for the false negative rate (FNR) is quite different. While for Ernst and Haesbroeck (2016) the FNR is in many settings below all other methods, this might just be due to the high FPR. The method of Filzmoser et al. (2013) has a very high FNR in most scenarios even in those with a high FPR. Regarding LOF, the simulation scenario has a strong effect on its performance. For the moving matrix setup we see a rather good performance compared to the other methods, while for random fields the FNR can hardly keep up with the other methods except the method of Filzmoser et al. (2013). The ssMRCD method is somewhere in between. Although the corresponding FNRs for the moving matrix setup with completely random swapping are not overwhelming, they are still in a reasonable range for the switching method of Ernst and Haesbroeck (2016). Moreover, the FNRs for the ssMRCD outlier detection technique in the other scenarios are compellingly low.

Comparing the F1-scores of the different methods, the best method to use in general depends on the scenario. While the three selected methods seem to have pitfalls in at least one simulation scenario, the ssMRCD-based method is consistently showing reliable results and is mostly among the two best methods. Moreover, less extreme behavior occurs when it comes to the FNR and the FPR. Thus, the ssMRCD-based local outlier detection method could be the method of choice for standard outlier detection tasks. However, note that with increasing contamination it might become more beneficial to use a method that is generally more prone to flag observations as outliers, e.g. the technique of Ernst and Haesbroeck (2016), or to adapt the parameters of the ssMRCD to allow for more locality.

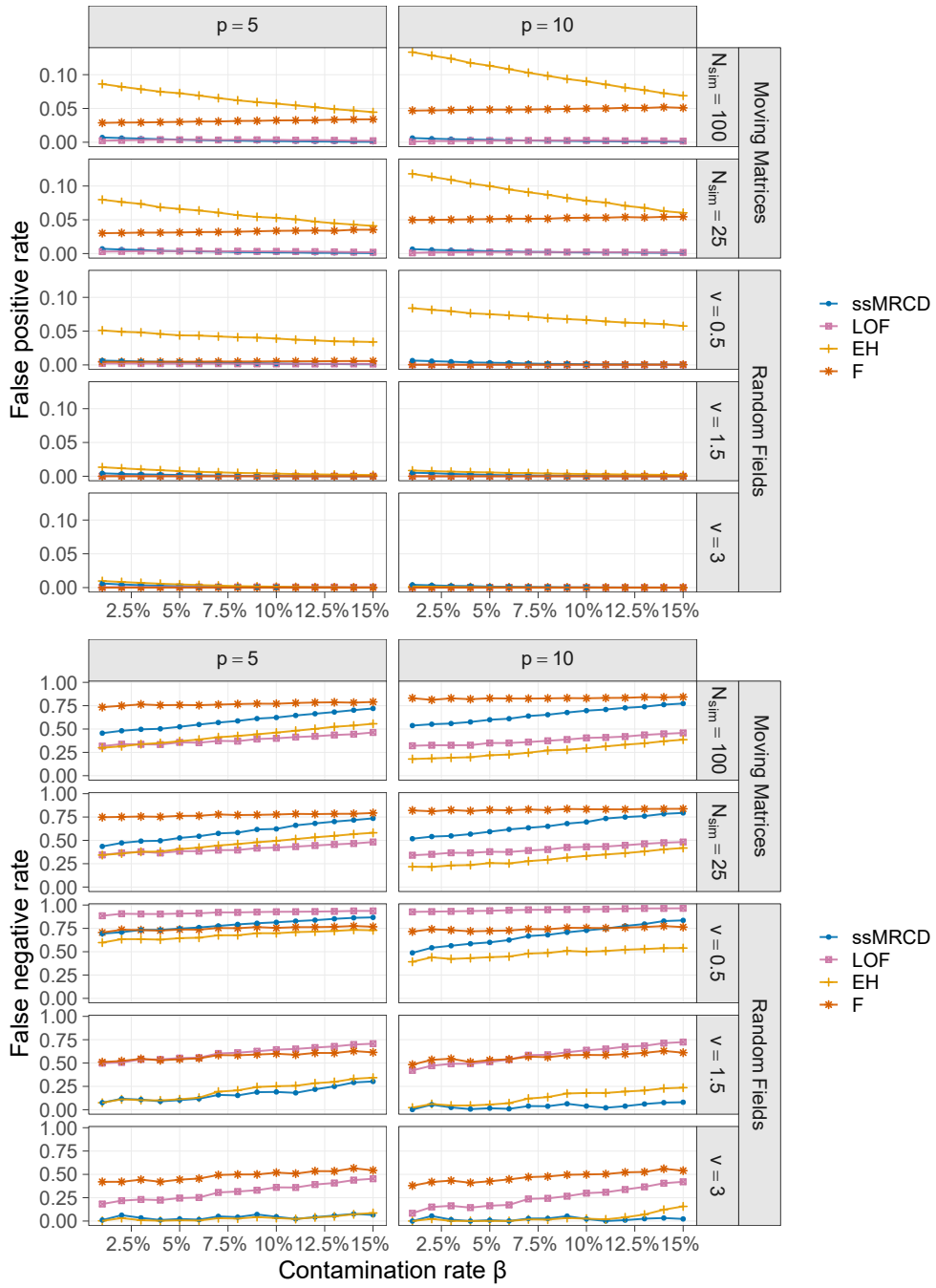


Figure 2.2: False positive and false negative rate for all four outlier detection methods with varying contamination levels achieved through completely random switching for different scenarios. Each point represents the mean of 100 repetitions.

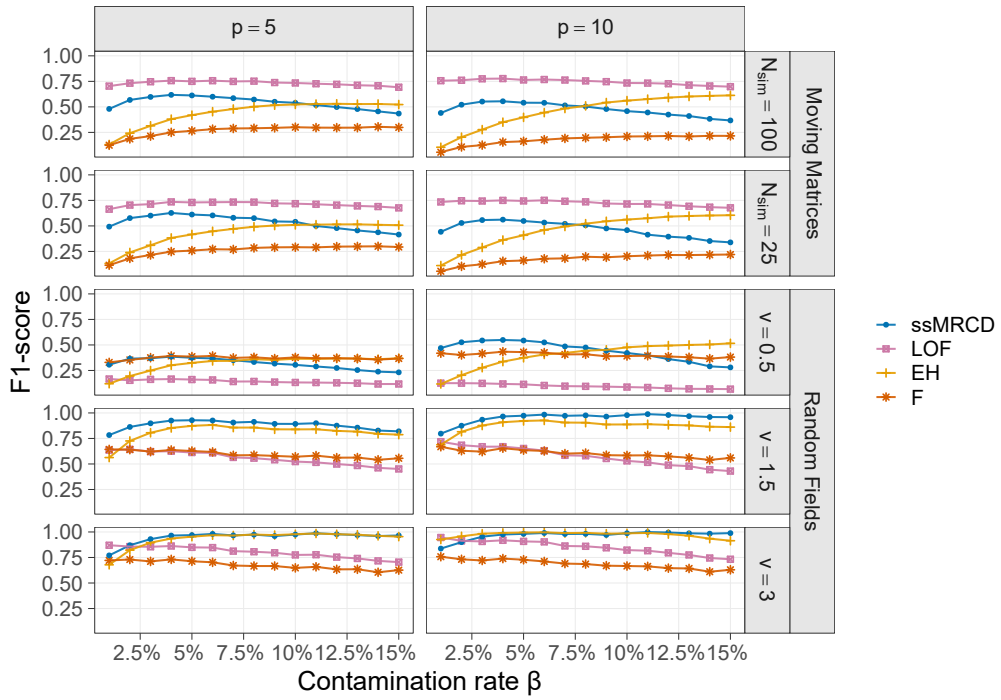


Figure 2.3: F1-Score for all four outlier detection methods with varying contamination levels achieved through completely random switching for different scenarios. Each point represents the mean of 100 repetitions.

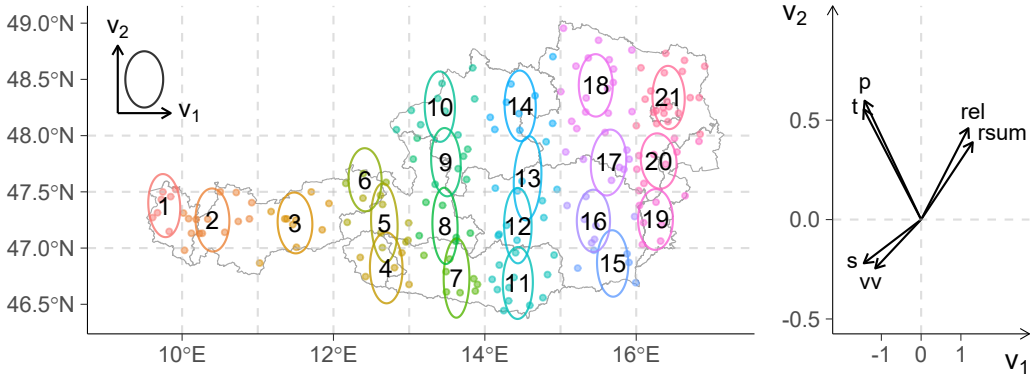


Figure 2.4: On the left hand side, the contours of Austria and its districts are shown together with the imposed grid structure for the ssMRCD neighborhoods. Here, singular points are assigned to a neighboring neighborhood. For each neighborhood a_i the index i is placed at the center, and the tolerance ellipses of corresponding correlation matrices based on the ssMRCD estimator are plotted along the first and second eigenvector coordinate of $\boldsymbol{T}$, which can be seen in the upper left corner as reference. The biplot of $\boldsymbol{T}$ is shown on the right hand side.

2.5 Example

In this section we consider a data set provided by GeoSphere Austria (2024). It consists of monthly weather data for Austrian weather stations and is used to test and compare the different methods. The data set contains measurements of air pressure [hPa] (p), relative air humidity [%] (rel), the monthly sum of sunshine duration [h] (s), wind velocity [m/s] (vv), air temperature in 2 meters above the ground [°C] (t), and the average daily sum of precipitation [mm] (rsum), averaged over all months in 2021, for $n = 183$ weather stations distributed all over Austria (see also Figure 2.4 or 2.7). The coordinates used for all methods are given in latitude and longitude.

We set $k = 10$ for all methods, i.e. we want to compare one observation with its ten closest neighbors, independent of the methodology used. Although for the method of Ernst and Haesbroeck (2016) it is possible to remove observations with comparably high levels of heterogeneity among the neighbors, we want to include all observations, thus setting $\beta = 1$. Even if there is increased heterogeneity among the neighbors, an observation might still be an interesting outlier clearly visible with the naked eye. Moreover, as mentioned in the prior section, the simulation results in Ernst and Haesbroeck (2016) show the best performance for high β. For the methodology of Filzmoser et al. (2013) in accordance with the simulation setup we allow for one of the ten neighbors to be similar to the local outlier. The cutoff value is again set to 0.2. We use the same cutoff value of 1.5 for the LOF as in the simulations.

For the ssMRCD local outlier detection method we use a grid based neighborhood structure for the covariance estimation. Due to the Alpine landscape especially in

the Western parts of Austria we aim at a rather local covariance structure, thus choosing a rather fine grid with $N = 21$ neighborhoods and $n_i \approx 8.7$ observations per neighborhood on average. Other possible options could be based on underlying structures, e.g. due to historical or political reasons, or on other classifying methods like clustering of the spatial coordinates. Furthermore, we use inverse-distance weights for the weighting matrix $\boldsymbol{W}$ between neighborhoods based on their center and select the default smoothing degree of $\lambda = 0.5$ to gain enough smoothing but still keep the locality of the fine grid structure.

As an alternative to using the default value of $\lambda = 0.5$ we can set up a simulation procedure. Assuming that the real data is uncontaminated, we can swap observations similar to the simulation studies in Section 2.4 and define them as local outliers. We can apply the outlier detection technique with the ssMRCD and different choices of λ (this can also be applied to other parameter settings of the ssMRCD), and then analyze the fraction of found outliers and the total number of outliers. Since only focusing on the known FNR for the found outliers leads to an increased false positive rate, it is sensible to also take the total number of found outliers into account. A good value of λ is a trade-off between a low FNR and a comparatively low number of found outliers overall. Interestingly, this procedure endorses the choice of $\lambda = 0.5$ in this data set.

The resulting ssMRCD correlation matrices for each neighborhood can be seen in Figure 2.4. The observations and the tolerance ellipses of the ssMRCD correlation matrices are colorized according to their neighborhoods. Since we have dimension $p = 6$, we reduce the dimensionality to the first two eigenvectors v_1 and v_2 of the global MCD correlation matrix $\boldsymbol{T}$, which is displayed at the upper left corner, hoping to depict most of the relevant variance. Moreover, a biplot of $\boldsymbol{T}$ is added at the right hand side to link the correlation matrices to our weather data.

Applying all four outlier detection methods leads to 24 observations in total classified as outliers. The most outliers (21) are classified by the method of Ernst and Haesbroeck (2016), the least (3) by Filzmoser et al. (2013), which is consistent with the simulation results regarding FPR and FNR, especially for the random fields setup. The distances which are used for outlier detection for each method and observation are shown in Figure 2.5. For further comparison of the results, the upper part of Figure 2.6 shows all 24 classified outliers with the corresponding ratio of distance value to cut-off value. Ratios above one are outliers. We can see that there are multiple weather stations that are classified as outliers only by the method of Ernst and Haesbroeck (2016) which lends itself to a notion consistent with the simulation results that there are some false positives among these weather stations. One example for a false positive could be panel b) in in the lower part of Figure 2.6. The station Feuerkogel (panel a)) was not detected by the method of Filzmoser et al. (2013), also consistent to the simulation results for the random fields setup and the generally high FNR. Interestingly, also LOF seems to have drawbacks and fails for example for the weather station Patscherkofel (panel c)), which was not detected as outlier. Nevertheless, the weather station Schoeckl (panel d)) was detected by all of them.

When looking at Figure 2.7 we can find some explanation for the local outlyingness for two of the three local outlier stations and why panel b) might not be outlying. While the stations Schoeckl and Feuerkogel are rather exposed on higher altitudes than

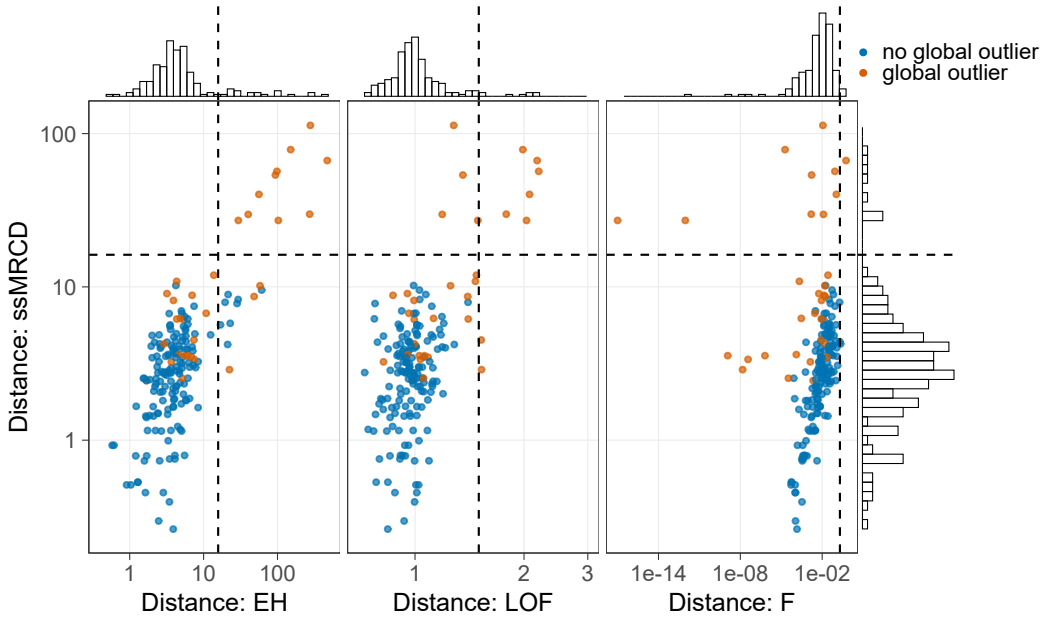


Figure 2.5: Distance-distance plots with the outlyingness scores of EH (next distance), of LOF (local outlier factor) and of F (isolation degree) against the next distance of the ssMRCD-based method. Observations are separated into global outliers based on the robust MD with the MCD as covariance estimator. At the margins the distribution of the different outlyingness scores are depicted by histograms.

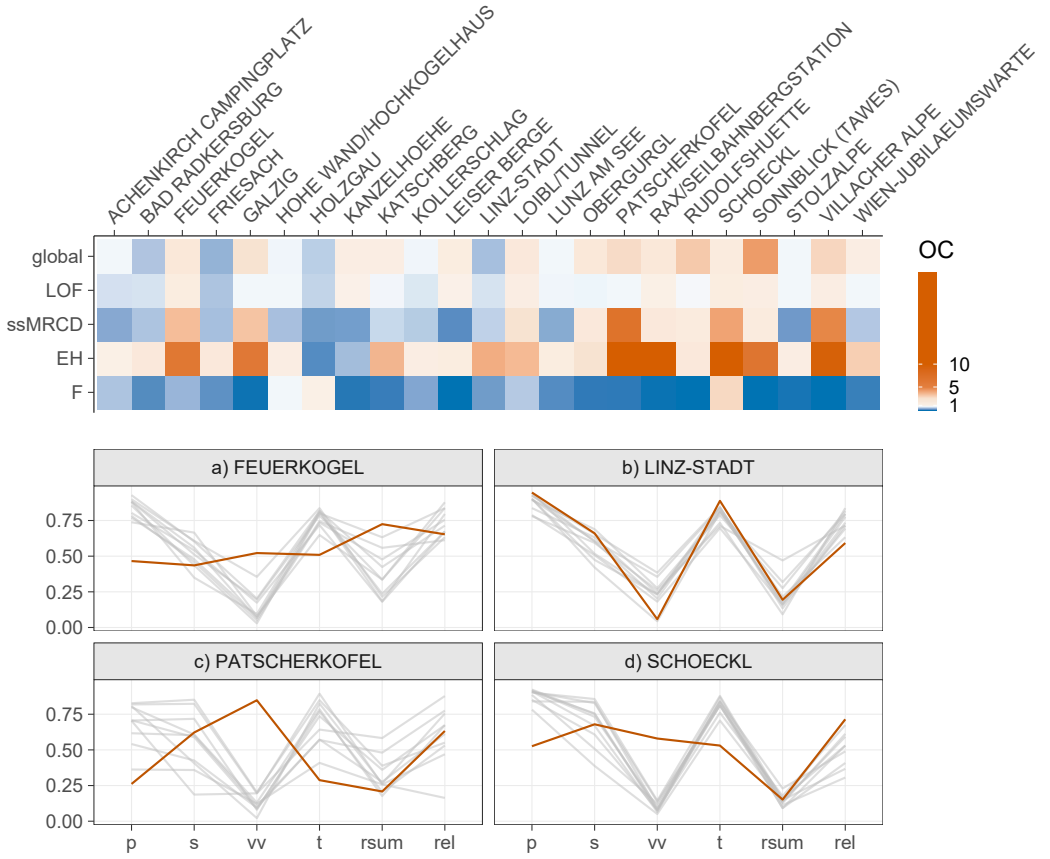


Figure 2.6: Upper part: Weather stations classified as outliers colorized according to their outlyingness score in relation to the cut-off value (OC) for all four methods and their global outlyingness. Lower part: in panels a)-d) four exemplary weather stations are selected to show differences on methodologies. Each variable is scaled to range $[0, 1]$. The values of the selected outliers are emphasized against the corresponding 10 nearest weather stations.

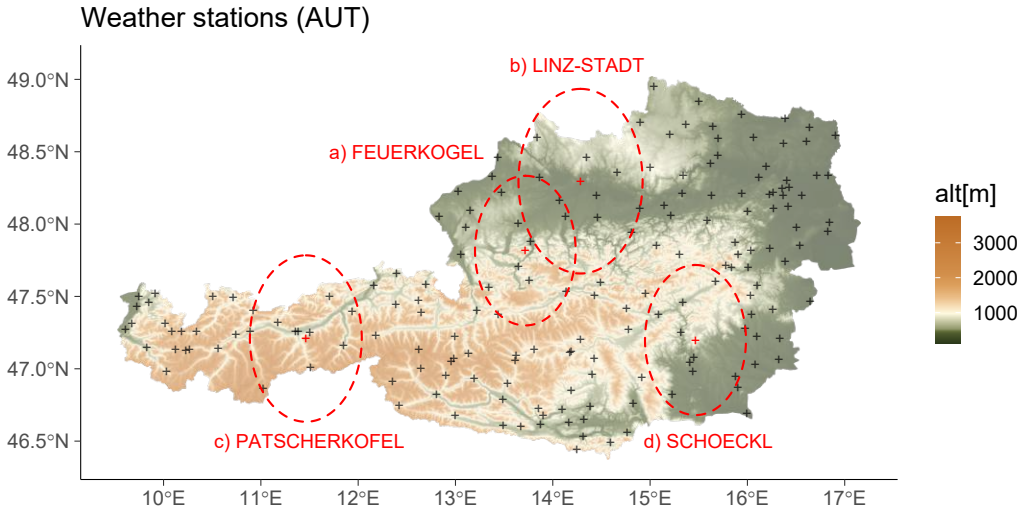


Figure 2.7: Altitude map of Austria with all weather stations and four selected outliers. Each dashed ellipse indicates the k nearest neighbors with whom the corresponding outlier is compared.

most of the surrounding k-nearest stations which can easily lead to different patterns regarding weather, the station Linz-Stadt (panel b)) is in a rather flat area similar to its neighbors. The station Patscherkofel is already deep in the Alpine area and is surrounded by other stations in valleys but also on mountains. Although from panel c) in Figure 2.6 it is evident that Patscherkofel differs significantly in wind velocity, it is not clear why it differs so much also from stations with similar altitude and exposure.

2.6 Conclusions

In this paper we enhance the limited toolbox for multivariate local outlier detection by extending the approaches of Filzmoser et al. (2013) and Ernst and Haesbroeck (2016). The developed ssMRCD based on the MRCD (Boudt et al., 2020) bridges the gap between fully local and fully global covariance matrices used in the pairwise MD by exchanging the extremely local covariance matrices used in Ernst and Haesbroeck (2016) with spatially smooth estimates.

We define the ssMRCD by means of a minimization problem and prove theoretical properties of the estimator, such as equivariance and breakdown point. A heuristic is provided for the stable convergence property of the proposed algorithm under reasonable spatial changes in underlying covariance matrices. Moreover, the methods of Filzmoser et al. (2013), Ernst and Haesbroeck (2016) and the ssMRCD outlier detection method are compared with the local outlier factor adapted for local outliers (Schubert et al., 2012) regarding outlier detection performance and computational efficiency for simulated data and real world data from Austrian weather stations.

While we support the conclusion of Ernst and Haesbroeck (2016) that it is difficult

to select the "best" method for outlier detection techniques, the ssMRCD-based outlier detection technique seems to be the only method providing reliable (but still improvable) results over all analyzed simulation scenarios. Note, that there might be non-analyzed scenarios where the ssMRCD-based outlier detection technique is not performing satisfactorily enough. Additionally, it is able to compete with the other methods regarding runtime even though the computation is quite complex. However, for a thorough real data analysis it is still preferable to use different outlier detection methods and compare the results in order to exploit all possible advantages of the available methods. Comparing results of multiple methodologies provides more insight in the data and significant local outliers can be classified with more reliability overall.

The ssMRCD covariance structure can be exploited also beyond local outlier detection. All covariance based methods that are sensible to adapt to spatial data can be extended by using the ssMRCD instead, e.g. spatial principal component analysis. A special case for the application of the ssMRCD might also be spatial data with structural breaks that need to be considered in the analysis. Finally, the presented ideas could also be transferred to a time series context, where the spatial dependency is replaced by the temporal dependency of multivariate time series, and the dependence structure could change over time. Such settings are usually quite challenging for outlier detection.

Software Availability

An implementation of the methodology and the Austrian weather data is available in the R-package `ssMRCD` on CRAN.

Appendix A

A.1 Proofs

Proofs of Section 2.2

Proof of Theorem 2.2.2.1. Let i be fixed, $\boldsymbol{T}(\boldsymbol{X})$ be an estimator of covariance as described above and $\boldsymbol{Y} := \boldsymbol{X}\boldsymbol{A}' + \mathbf{1}_n\boldsymbol{b}'$ be the transformed data matrix. Then, for any subset combination $\mathcal{H}$ and for all $j = 1, \ldots, N$ it holds that

$$\begin{aligned}
\boldsymbol{K}_j^Y(\mathcal{H}) &= \rho_j \boldsymbol{T}(\boldsymbol{Y}) + (1-\rho_j) Cov(\boldsymbol{Y}_{H_j}) \\
&= \rho_j \boldsymbol{T}(\boldsymbol{X}\boldsymbol{A}' + \mathbf{1}_n\boldsymbol{b}) + (1-\rho_j) Cov(\boldsymbol{X}_{H_j}\boldsymbol{A}' + \mathbf{1}_n\boldsymbol{b}) \\
&= \rho_j \boldsymbol{A}\boldsymbol{T}(\boldsymbol{X})\boldsymbol{A}' + (1-\rho_j)\boldsymbol{A} Cov(\boldsymbol{X}_{H_j})\boldsymbol{A}' \\
&= \boldsymbol{A}\left[\rho_j \boldsymbol{T}(\boldsymbol{X}) + (1-\rho_j) Cov(\boldsymbol{X}_{H_j})\right]\boldsymbol{A}' \\
&= \boldsymbol{A}\boldsymbol{K}_j^X(\mathcal{H})\boldsymbol{A}'.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\left((1-\lambda)\boldsymbol{K}_i^Y(\mathcal{H}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j^Y(\mathcal{H})\right) = \\
&= \left((1-\lambda)\boldsymbol{A}\boldsymbol{K}_i^X(\mathcal{H})\boldsymbol{A}' + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{A}\boldsymbol{K}_j^X(\mathcal{H})\boldsymbol{A}'\right) \\
&= \boldsymbol{A}\left((1-\lambda)\boldsymbol{K}_i^X(\mathcal{H}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j^X(\mathcal{H})\right)\boldsymbol{A}'. \qquad \text{(A.9)}
\end{aligned}$$

By using the multiplicative property of the determinant and $\det(\boldsymbol{A}) \neq 0$ we see that $\boldsymbol{A}$ is only a constant in the minimization problem and is not affecting the choice of the optimal combination of subsets,

$$\begin{aligned}
f(\mathcal{H}) &= \sum_{i=1}^{N} \det\left((1-\lambda)\boldsymbol{K}_i^Y(\mathcal{H}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j^Y(\mathcal{H})\right) \\
&= \det(\boldsymbol{A})^2 \sum_{i=1}^{N} \det\left((1-\lambda)\boldsymbol{K}_i^X(\mathcal{H}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j^X(\mathcal{H})\right).
\end{aligned}$$

Together with Equation (A.9), affine equivariance is proven for the covariance estimator. Since the location estimator is defined as the arithmetic mean, which is affine equivariant, the property stated in Equation (2.5) is also fulfilled. □

Proof of Theorem 2.2.2.2. **c.** Regarding notation see also Boudt et al. (2020). The eigenvalues for the transformed covariance matrix $\boldsymbol{K_W} = \rho\boldsymbol{I} + (1-\rho)c_\alpha \boldsymbol{S_W}(H)$, where $\boldsymbol{S_W}(H)$ is the covariance matrix of a subset H of $\boldsymbol{X}$, are bounded below by $\rho > 0$. Thus, $\lambda_i(\boldsymbol{K_W^{-1}}) = \lambda_i(\boldsymbol{K_W})^{-1} \leq 1/\rho$ and $\left\|\boldsymbol{K_W^{-1}}\right\|_2 \leq 1/\rho$ where $\|.\|_2$ denotes the spectral matrix norm. For covariance matrices the spectral norm is equal to its biggest eigenvalue. It follows that

$$\begin{aligned}\left\|\left(\boldsymbol{Q\Lambda}^{1/2}\boldsymbol{K_W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q'}\right)^{-1}\right\|_2 &= \left\|\boldsymbol{Q\Lambda}^{-1/2}\boldsymbol{K_W^{-1}}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q'}\right\|_2 \\ &\leq \left\|\boldsymbol{Q'\Lambda}^{-1/2}\right\|_2^2 \left\|\boldsymbol{K_W^{-1}}\right\|_2 \\ &\leq c/\rho,\end{aligned}$$

for $c > 0$, since $\boldsymbol{Q'\Lambda}^{-1/2}$ is also regular and fixed. This implies that

$$\lambda_i\left(\boldsymbol{Q\Lambda}^{1/2}\boldsymbol{K_W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q'}\right)^{-1} = \lambda_i\left((\boldsymbol{Q\Lambda}^{1/2}\boldsymbol{K_W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q'})^{-1}\right) \leq c/\rho$$

for all $i = 1, \ldots, p$. It follows that

$$\lambda_i(\boldsymbol{Q\Lambda}^{1/2}\boldsymbol{K_W}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q'}) \geq \rho/c > 0 \quad \forall i = 1, \ldots, p,$$

for all subsets H, specifically for the optimal subset H^*. Thus, the eigenvalues of $\hat{\boldsymbol{\Sigma}}_n = \boldsymbol{Q\Lambda}^{1/2}\boldsymbol{K_W^*}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q'}$ are also bounded away from zero, and it follows that the implosion breakdown point is 1.

b. It is clear that $\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_n; \boldsymbol{X}_n) \leq (n-h+1)/n$ since in this case there would always be at least one observation in the selected subset independent of its value spoiling the estimation. We need to show that $\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_n; \boldsymbol{X}_n) > (n-h)/n$.

Suppose $\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_n; \boldsymbol{X}_n) \leq (n-h)/n$. We can change $m \leq n-h$ observations arbitrarily and denote the resulting matrix as $\boldsymbol{X}_{n,m} = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^*, \boldsymbol{x}_{m+1}, \ldots, \boldsymbol{x}_n)'$, where $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^*$ are the exchanged observations (w.l.o.g. placed in the first m rows). The supremum being infinite is equivalent to

$$\forall C > 0 \; \exists \boldsymbol{X}_{n,m}: \quad |\ln(\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))) - \ln(\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_n)))| > C. \tag{A.10}$$

Since $\ln(\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_n))$ is constant and ln is monotonously increasing and unrestricted we can w.l.o.g. assume that the value inside of the absolute value is non-negative. Additionally, moving the constant $\ln(\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_n)))$ to the right hand side, Equation (A.10) is equivalent to the unboundedness of the biggest eigenvalue $\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}))$,

$$\forall C > 0 \; \exists \; \boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^*: \quad \lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m})) > C. \tag{A.11}$$

Note that $\det(\boldsymbol{X}) = \prod_{i=1}^p \lambda_i(\boldsymbol{X})$ for any p-dimensional matrix $\boldsymbol{X}$ and that $\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m}) = (1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;H^*}) + \rho\boldsymbol{T}$, where H^* is the subset of observations of $\boldsymbol{X}_{n,m}$ that minimize $\det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;H}) + \rho\boldsymbol{T})$ over all subsets H. As shown above, the eigenvalues of $(1-\rho)c_\alpha Cov(\boldsymbol{X}) + \rho\boldsymbol{T}$ are bounded away from zero for all $\boldsymbol{X}$ and $\rho > 0$,

$$\lambda_i((1-\rho)c_\alpha Cov(\boldsymbol{X}) + \rho\boldsymbol{T}) \geq c > 0, \quad \forall i = 1, \ldots, p.$$

For the matrix $\boldsymbol{X}_{n,m;H^*}$, it follows that

$$\prod_{i=2}^{p} \lambda_i(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m})) \geq c^{p-1} > 0.$$

Let the constant $\tilde{C}$ be defined as

$$\tilde{C} = \frac{\det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}}) + \rho \boldsymbol{T})}{c^{p-1}},$$

where $\tilde{H} = m+1, \ldots, m+h$ denote h indices of fixed and unchanged observations of $\boldsymbol{X}_{n,m}$, which exist due to $m \leq n-h$. Then, due to condition (A.11) for $\tilde{C}$ there exists $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^*$ such that $\lambda_1(\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{X}_{n,m})) > \tilde{C}$ which leads to

$$\det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;H^*}) + \rho \boldsymbol{T}) > \det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}}) + \rho \boldsymbol{T}).$$

This contradicts the minimization of the determinant property of the selected subset H^*. Thus, $\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_n; \boldsymbol{X}_n) > (n-h)/n$.

a. Using the same argument as before, it is clear that $\epsilon_n^*(\hat{\boldsymbol{\mu}}_n; \boldsymbol{X}_n) \leq (n-h+1)/n$. Let us show that $\epsilon_n^*(\hat{\boldsymbol{\mu}}_n; \boldsymbol{X}_n) \leq h/n$. Again we can argue that $\sup ||\hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_{n,m}) - \hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_n)|| = +\infty$ is equivalent to

$$\forall C > 0 \; \exists \; \boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^* : \; \|\hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_{n,m})\| > C. \tag{A.12}$$

We have to find $m = h$ many exchanged data points in a way that the norm of the location estimator is unbounded but the determinant of the covariance matrix is still minimal. For the fixed data set $\boldsymbol{X}_n$, we obtain the optimal subset H^* of observations and add a fixed but arbitrarily large number $L > 0$ to the first coordinate of these $m = h$ observations,

$$\forall i \in H^* : \tilde{x}_{i1} = x_{i1} + L \text{ and } \tilde{x}_{ij} = x_{ij} \; \forall j = 2, \ldots, p.$$

Thus, the sample mean of the first coordinate of the selected subset $\boldsymbol{X}_{n,m;H^*}$ is equal to the original mean of the first coordinate of $\boldsymbol{X}_{n;H^*}$ plus L. Similarly, the sample covariance is the same as before given that we take the same subset H^*, since it is independent of constant shifts applied to all used observations. This implies that also the regularized covariance and its determinant are the same which was minimal for all other subsets of $\boldsymbol{X}_n$. In order to show minimality of the subset H^* for the new data matrix $\boldsymbol{X}_{n,m}$ it follows that we only have to consider the subsets that have both original and exchanged (arbitrarily large) observations.

Regarding the sample mean of the first coordinate of one of these subsets $\tilde{H}$, it is

$$\begin{aligned}
\tilde{M}_1 &= \frac{1}{h}\left(\sum_{j=1}^{k} x_{i_j 1} + \sum_{j=k+1}^{h} \tilde{x}_{i_j 1}\right) \\
&= \frac{1}{h}\left(\sum_{j=1}^{k} x_{i_j 1} + \sum_{j=k+1}^{h} (x_{i_j 1} + L)\right) \\
&= M_1 + L - \frac{k+1}{h} L,
\end{aligned}$$

where M_1 is the (fixed) mean of the first coordinate of the subset $\boldsymbol{X}_{n;\tilde{H}}$ and both sums are not empty. Regarding the variance of the first coordinate, which is the first diagonal entry of the sample covariance matrix, we see

$$\begin{aligned} Cov(\boldsymbol{X}_{n,m;\tilde{H}})_{11} &= \frac{1}{h-1}\left(\sum_{j=1}^{k}(x_{i_j1}-\tilde{M}_1)^2+\sum_{j=k+1}^{h}(\tilde{x}_{i_j1}-\tilde{M}_1)^2\right) \\ &= \frac{1}{h-1}\Bigg(\sum_{j=1}^{k}\underbrace{(x_{i_j1}-M_1-L+\frac{k+1}{h}L)^2}_{\mathcal{O}(L^2)} \\ &\quad +\sum_{j=k+1}^{h}\underbrace{(x_{i_j1}+L-M_1-L+\frac{k+1}{h}L)^2}_{\mathcal{O}(L^2)}\Bigg) \\ &= \mathcal{O}(L^2) \end{aligned}$$

Thus, the Frobenius norm of $Cov(\boldsymbol{X}_{n,m;\tilde{H}})$ and also its regularization are $\mathcal{O}(L^2)$ and it follows for some constant $\beta > 0$ that

$$\begin{aligned} \mathcal{O}(L^2) &= \left\|(1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}})+\rho\boldsymbol{T})\right\|_F \\ &\leq \beta\left\|(1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}})+\rho\boldsymbol{T})\right\|_2 \\ &= \beta\lambda_1((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}})+\rho\boldsymbol{T}) \\ &\leq \beta\frac{\det((1-\rho)c_\alpha Cov(\boldsymbol{X}_{n,m;\tilde{H}})}{c^{p-1}}, \end{aligned}$$

due to equivalence of matrix norms in finite dimensional space and c being the constant from above. Choosing L arbitrarily large, we see that the determinant corresponding to a mixed subset is larger than the determinant of the optimal subset H^* of only exchanged observations.

Now suppose $\epsilon_n^*(\hat{\boldsymbol{\mu}}_n;\boldsymbol{X}_n) = m/n < \min(h, n-h+1)/n$ and start from Equation (A.12),

$$\forall C>0\ \exists\ \boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_m^*:\|\hat{\boldsymbol{\mu}}_n(\boldsymbol{X}_{n,m})\|>C.$$

This implies that

$$\forall C>0\ \exists\ \boldsymbol{x}_1^*,\ldots,\boldsymbol{x}_m^*:\sum_{j=1}^{k}\|\boldsymbol{x}_{i_j}\|+\sum_{j=k+1}^{h}\left\|\boldsymbol{x}_{i_j}^*\right\|\geq\left\|\left(\sum_{j=1}^{k}\boldsymbol{x}_{i_j}+\sum_{j=k+1}^{h}\boldsymbol{x}_{i_j}^*\right)\right\|>C,$$

where $i_j \in H^*, j = 1,\ldots,h$. Thus, for all $C > 0$ there exists some $\boldsymbol{x}_{i_j}^*$ whose norm is bigger than C. W.l.o.g. assume it is $\boldsymbol{x}_1^*$ and that the first coordinate is responsible,

$$\forall C>0\ \exists\ \boldsymbol{x}_1^*\in H^*:|x_{11}^*|\geq C.$$

For $m < h < n - h + 1$ there would not be the possibility to only include exchanged points in the subset and it would always be possible to have a subset of h many original observations. This is also the case for $m < n - h + 1 < h$. Thus, there are at least one exchanged point $\boldsymbol{x}_1^*$ and one original point in H^*. But with the same argument as before, the determinant of the mixed subset of original points and arbitrarily large points would eventually contradict optimality, because at one point the determinant would be so large that there would be an h-sized subset of original observations available to get a smaller determinant. □

Proof of Theorem 2.2.2.3. For $i = 1, \ldots, N$, the location estimate is the standard sample mean of h_i many observations from neighborhood a_i selected in a way to minimize the objective function (2.3). By exchanging observations in one neighborhood a_i with arbitrarily large values and keeping the other neighborhoods the same (keeping the matrices $\boldsymbol{K}_j$ bounded), we can apply the results of Theorem 2.2.2.2 for the MRCD structured covariance matrix $\boldsymbol{K}_i$. The location breakdown point for $\boldsymbol{K}_i$ is $\min(n_i - h_i + 1, h_i)/n_i$. Thus, in order to make at least one of the location estimators useless we need to exchange a fraction $\min_{i=1,\ldots,N} \min(n_i - h_i + 1, h_i)/n_i$ of observations of one neighborhood. □

Proof of Theorem 2.2.2.4. Since the spatially smoothed MRCD covariance estimators $\boldsymbol{K}_i$ are regularized on each neighborhood according to the MRCD approach, all eigenvalues are positive and bounded away from zero as long as the target matrix $\boldsymbol{T}$ is regular (see Theorem 2.2.2.2). Hence, none of the covariance estimators will ever be singular and the implosion breakdown point is 1.

For the second part let us fix neighborhood a_i. Note, that the covariance estimator is defined as $\hat{\boldsymbol{\Sigma}}_{SSM,n,i} = (1-\lambda)\boldsymbol{K}_i(\mathcal{H}^*) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^*)$ for the optimal subset $\mathcal{H}^*$. The matrix $\boldsymbol{K}_i = \boldsymbol{K}_i(\mathcal{H}^*)$ is structured in an MRCD manner based on the sample covariance matrix of the subset H_i^* and the target matrix. Since we assume $\boldsymbol{T}$ to be fixed, for $\boldsymbol{K}_i$ we can get arbitrarily large eigenvalues only under the same circumstances as for the MRCD (see Theorem 2.2.2.2). Thus, exchanging a fraction of $(n_i - h_i + 1)/n_i$ by arbitrary values can lead to arbitrarily large eigenvalues of $\boldsymbol{K}_i$. For the explosion breakdown point for one neighborhood covariance estimator $\hat{\boldsymbol{\Sigma}}_{SSM,n,i}$ it is sufficient that at least one $\boldsymbol{K}_j$ has reached its breakdown point (assuming a general setting for $\boldsymbol{W}$ and λ). It follows, that the finite sample explosion breakdown point of $\hat{\boldsymbol{\Sigma}}_{SSM,n,i}$ is

$$\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}; \boldsymbol{X}_n) = \min_{i=1,\ldots,N} \{(n_i - h_i + 1)/n_i\}. \tag{A.13}$$

Since $\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}; \boldsymbol{X}_n)$ is already independent of i, the overall explosion breakdown point for the spatially smoothed MRCD covariance estimators is equal to $\epsilon_n^*(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}; \boldsymbol{X}_n)$. □

For all proofs the target matrix $\boldsymbol{T}$ is assumed to be fixed. In applications it is often the case that $\boldsymbol{T}$ is actually an estimated covariance matrix $\boldsymbol{T}(\boldsymbol{X})$ based on data $\boldsymbol{X}$ and thus, the breakdown point needs the be reevaluated. First, the estimator $\boldsymbol{T}(\boldsymbol{X})$ has to be regular, otherwise the theoretical results are not applicable. Accordingly, the implosion breakdown point of the ssMRCD estimator is still 1. Regarding the explosion

breakdown point, the target matrix enters the estimation for the ssMRCD without any changes. Due to the additive structure of the estimators and the inequalities for the biggest eigenvalue of covariance matrices,

$$\lambda_1(A) + \lambda_1(B) \leq 2\lambda_1(A+B) \leq 2(\lambda_1(A) + \lambda_1(B)),$$

choosing a subset in a way that leads to an unboundedness of the biggest eigenvalue of $\boldsymbol{T}(\boldsymbol{X})$ also leads to unboundedness of the final estimator and unboundedness of the final estimator regarding the biggest eigenvalue implies unboundedness in one of the estimators ($\boldsymbol{T}(\boldsymbol{X})$ or the ssMRCD estimator with a target matrix assumed to be fixed). Thus, the breakdown point of the ssMRCD with estimated target matrix $\hat{\boldsymbol{\Sigma}}^T_{SSM,n,i}$ is exactly the minimum of the two,

$$\epsilon^*_n(\hat{\boldsymbol{\Sigma}}^T_{SSM,n,i}; \boldsymbol{X}_n) = \min\{\epsilon^*_n(\hat{\boldsymbol{\Sigma}}_{SSM,n,i}; \boldsymbol{X}_n), \epsilon^*_n(\boldsymbol{T}_n; \boldsymbol{X}_n))\}.$$

Proofs of Section 2.3

Proof of Theorem 2.3.0.1. This proof is very much along the lines of Boudt et al. (2020). The neighborhood a_i is fixed. Thus, the matrix which determinant should be minimized regarding i is

$$\begin{aligned}
\boldsymbol{A}_1 &:= \left((1-\lambda)\boldsymbol{K}_i(\mathcal{H}^0) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right) \\
&= \left((1-\lambda)[(1-\rho_i)c_{\alpha_i} Cov(\boldsymbol{X}_{H_i^0}) + \rho_i \boldsymbol{T}] + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0) \right) \\
&= \left(\underbrace{(1-\lambda)(1-\rho_i)c_{\alpha_i}}_{:=\tilde{\rho}} Cov(\boldsymbol{X}_{H_i^0}) + \underbrace{(1-\lambda)\rho_i \boldsymbol{T} + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^0)}_{:=\boldsymbol{\Omega}} \right) \\
&= \tilde{\rho}\, Cov(\boldsymbol{X}_{H_i^0}) + \boldsymbol{\Omega}, \\
\boldsymbol{A}_2 &:= \left((1-\lambda)\boldsymbol{K}_i(\tilde{\mathcal{H}}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\tilde{\mathcal{H}}) \right) \\
&= \tilde{\rho}\, Cov(\boldsymbol{X}_{H_i^1}) + \boldsymbol{\Omega},
\end{aligned}$$

where $\boldsymbol{\Omega}$ is a fixed positive definite covariance matrix.

Since the original proof is not restricted to convex linear combinations, we can use the same proof with the matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ and $\boldsymbol{\Omega}$ in place of $\boldsymbol{K}_1$, $\boldsymbol{K}_2$ and $\boldsymbol{\Lambda}$ and adapted factors

$$\begin{aligned}
w_j &= \sqrt{k\tilde{\rho}/h_i}, && j = 1, \ldots, h_i \\
&= \sqrt{k/(p+1)}, && j = h_i + 1, \ldots, k.
\end{aligned}$$

with $k = h_i + p + 1$.[1] □

A.2 Algorithm

Iteration Steps

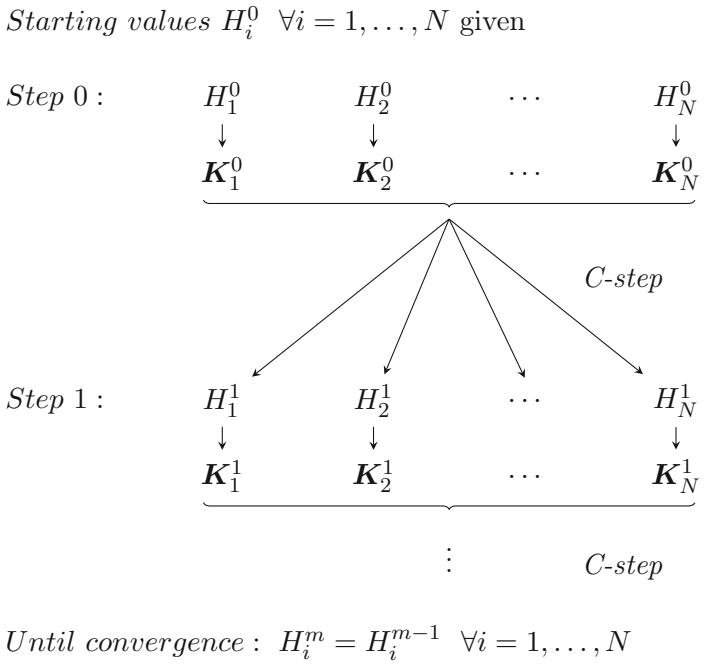


Figure A.8: Illustration of matrices used in the C-step after each iteration step. H_i^j are the selected subsets of neighborhood a_j in step i, and $\boldsymbol{K}_i^j$ the corresponding regular covariance matrices.

Convergence

Here, we want to analyze the algorithm described and motivated in Section 2.3 in more detail. Since Theorem 2.3.0.1 is only valid for one varying covariance matrix and not for multiple varying ones, the convergence properties should be further examined. Figure A.9 shows the objective function values along the iteration procedure for all starting H-set combinations for different parameter settings for both simulation setups. We choose $p = 5$ and a 5% contamination rate achieved by completely random swapping. The weighting matrix is based on inverse distances as mentioned in Section 2.2. Each panel reflects the convergence behavior of the objective function for one simulated data set.

Although the behavior differs in general, it is evident that the algorithm has overall very good convergence properties. A high percentage of monotonically decreasing

[1] Note that this proof can be generalized to any kind of linear combination with fixed matrices and a sample covariance matrix of a subset.

objective functions can often be achieved and the non-monotonically decreasing paths increase only marginally. The reliability of the convergence results in the simulation might possibly be due to the spatially correlated values which lead to rather small changes in the covariance matrices during the algorithm. Thus, for our simulated data sets, the assumption of fixed covariance matrices seems to be sufficiently met. Since in reality local outlier detection mostly makes sense only for spatially correlated data, the theoretical results from Theorem 2.3.0.1 proof to be even more valuable.

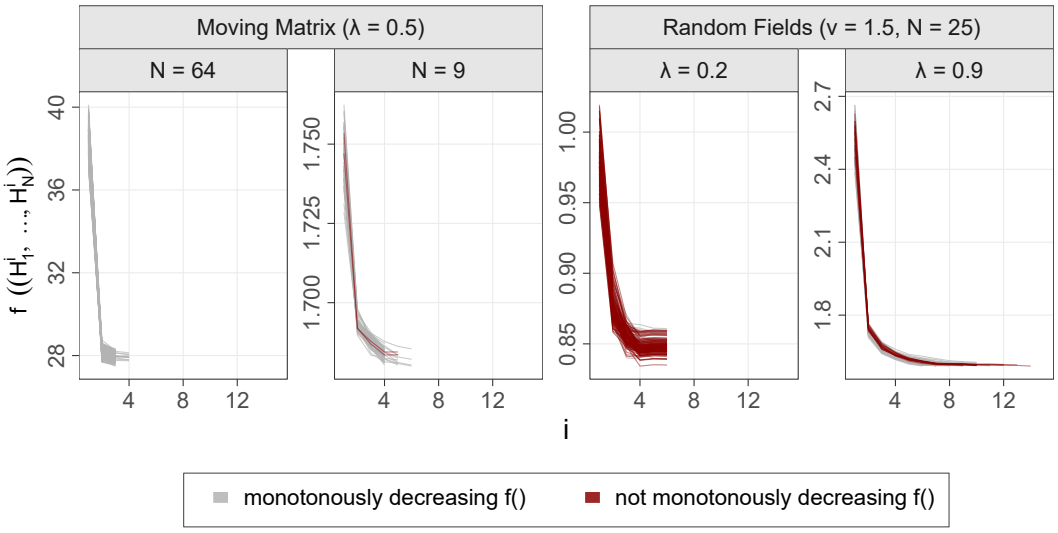


Figure A.9: Different convergence behaviors of the objective function. A $p = 5$ dimensional setting with 5% contamination achieved with completely random swapping. Each line is the path of one initial set of H-sets along the C-step iteration according to the algorithm described in Section 2.3.

Moreover, the convergence takes place fast which might be caused by the choice of the good starting estimates of the detMCD algorithm (Hubert et al., 2012). Another reason might be that the number of observations that can be used is restricted by the neighborhood assignments to a smaller number than in a covariance estimation for the full data set. Very promising is also the rapid improvement at the very beginning, independent of the simulated data sets.

A.3 Analysis of Runtime

Estimation of ssMRCD Estimators

Here, we present some analysis of computational efficiency of the ssMRCD estimator and the algorithm proposed.

The iteration process and the increasing number of starting values with increasing N can have quite an impact on runtime. Nevertheless, as long as the number of neighborhoods N is not too big, the outlier detection method based on the ssMRCD is competitive with other local methods (Ernst and Haesbroeck, 2016), especially

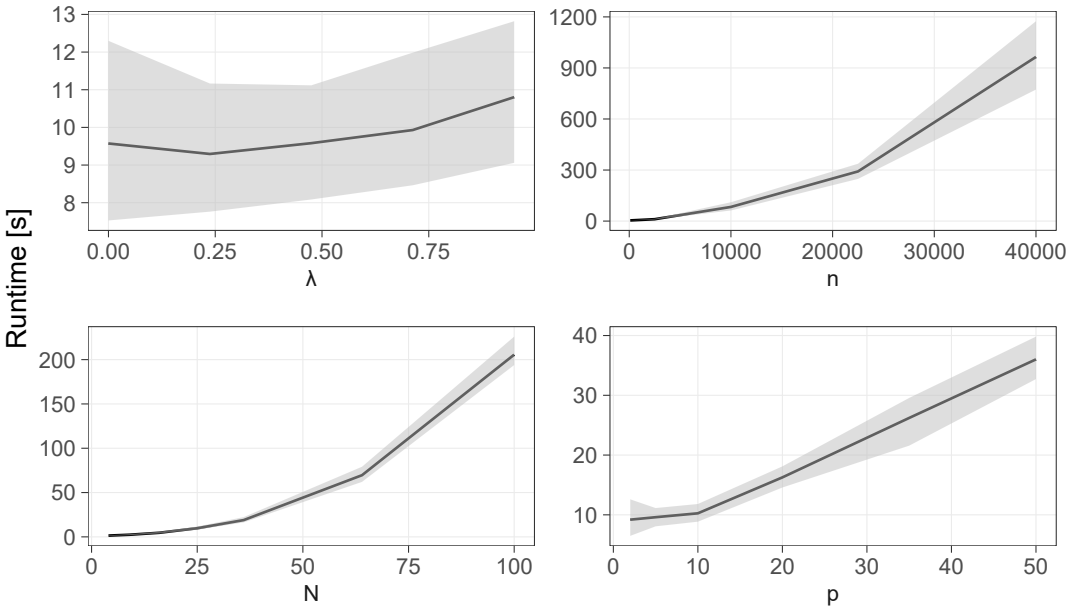


Figure A.10: Analysis of runtime of the ssMRCD for setup 1 with 5% contamination rate, $N_{sim} = 25$ and varying parameter values with default values $\lambda = 0.5$, $p = 5$, $N = 25$ and $n = 1681$ if not varied. The solid line is representing the mean of 100 repetitions, the edges of the band around the mean the 5% and 95% quantile.

if p is large (see also Figure A.11. For simulation setup 1 with $N_{sim} = 25$, a 5% contamination rate through completely random switching and 100 repetitions are used and the parameters p, λ, N and n are each varied univariately. The default values for parameters not being varied are $p = 5$, $\lambda = 0.5$, $N = 25$, and $n = 1681$.

As depicted in Figure A.10, the number of neighborhoods N and the number of observations n have the most influence on runtime with more than a linear increase. The nearly quadratic increase for N is partly due to the number of starting values increasing linearly with N, which could also be reduced to enhance efficiency if necessary. Interestingly, the dimension p has an approximately linear effect on runtime which is overall moderate. The smoothing parameter λ does not significantly change the runtime.

Local Outlier Detection Methods

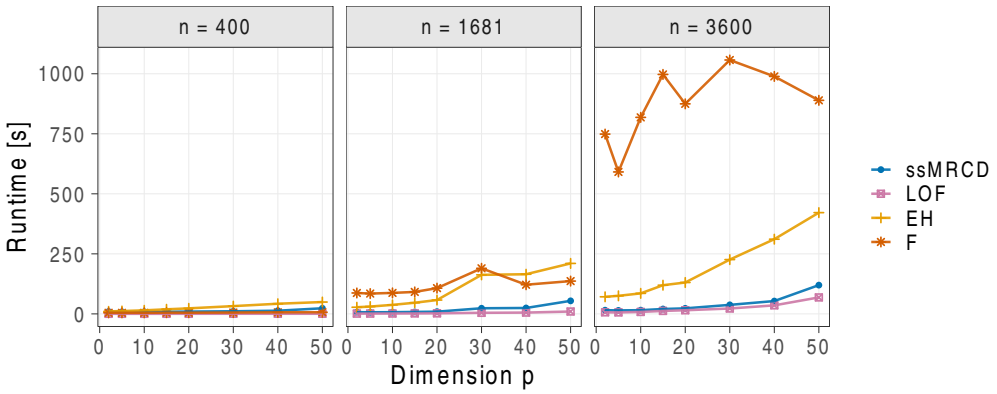


Figure A.11: Comparison of average (mean) runtime for outlier detection algorithms using simulation setup 1 with varying number of observations n and dimension p. The parameter settings from the previous outlier detection performance simulation study with 200 repetitions are used .

With many dimensions and observations efficiency in computing becomes important. The results of a short simulation study regarding the runtime of the four outlier detection techniques with parameter settings of the simulations underlying Figures 2.2 and 2.3 are shown in Figure A.11 for the moving matrix scenario with $N_{sim} = 100$. All methods need more computation time for increasing dimension p, especially the methods based on covariance estimation and inversion (EH, ssMRCD, F). Also, the number of observations seems to have a big effect on runtime, especially for the method of Filzmoser et al. (2013), possibly due to an inefficient implementation of finding the k-nearest neighbors. Although the local outlier factor (LOF) method is reliably fast, the ssMRCD seems to be comparably efficient, even though it involves a complex covariance estimation procedure.

A.4 Parameter Sensitivity

Before comparing the performance in local outlier detection with other methods, the parameter sensitivity of the ssMRCD is analyzed in more detail. This simulation study should simplify the choices of λ and N in particular for real world data and focus on possible issues connected to suboptimal parameter settings.

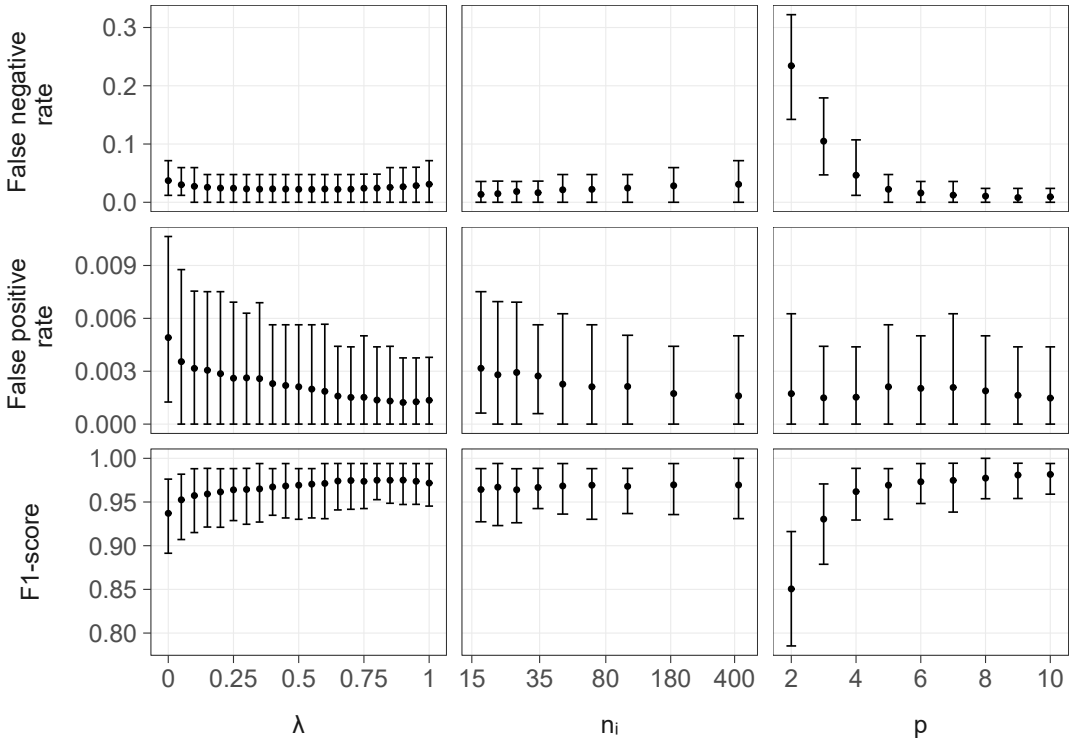


Figure A.12: Outlier detection performance based on the false negative, false positive rate and the F1-score of the ssMRCD outlier detection method for different parameter settings. Each point represents the arithmetic mean and the corresponding bars the 5th and 95th quantile of 100 simulations. For non-varying parameters the default settings are $p = 5$, $N = 25$ (comparable to $n_i \approx 67$) and $\lambda = 0.5$.

For this purpose, setup 2 (random fields) is considered as simulation setting, with parameter $\nu = 3$, and $\beta = 5\%$ completely randomly swapped observations. Special focus is put on the choice of λ and N, but also the effect of dimension p is analyzed. The other parameters are chosen in accordance to possible default settings. The weighting matrix is based on the inverse Euclidean distances of the centers of the neighborhoods a_i. Since all considered methods propose $k = 10$ as a default value, we adhere to this setting for now. Each parameter combination was simulated for 100 different realizations. While Ernst and Haesbroeck (2016) suggest to use Cohen's Kappa as summary statistic of the confusion matrix, we will use the F1-score due to its good interpretability and suitability also for imbalanced classification data.

Figure A.12 shows the false negative rate, the false positive rate and the resulting F1-score plotted against varying values of λ, N and p, with default values of $p = 5$, $N = 25$ (implying $n_i \approx 67$ on average) and $\lambda = 0.5$. These values should reflect a quite general and unspecific parameter setting. For illustration purposes it is more informative to plot the average neighborhood size $n_i \approx 1681/N$ instead of the number of neighborhoods.

The simulation results show that a higher λ decreases the false positive rate, it has marginal reduction effects on the false negative rate until too much smoothing masks real outliers. The overall performance increases moderately in λ, but for λ higher than 0.5 the increase is marginal. Thus, we propose a default value of 0.5 for λ to get the advantage of the decrease in the false positive rate while avoiding the masking effect for higher values. Compared to the influence of λ, the effect of the dimension p is more pronounced. Very small dimensions seem to cover outliers more effectively, probably due to less available information. Interestingly, the size of the neighborhoods seems to be relatively irrelevant, at least in this simulation setting. Only a small masking effect occurs similar to the effects of λ. Too big neighborhoods lead to too much smoothing. Thus, this might imply to choose a strategy of medium sized neighborhoods to increase efficiency in computation and reduce unnecessary regularization. This guidance for the parameter choices might be biased towards this specific simulation setting and not optimal in other settings, but fixing the parameters with at least a sensible value simplifies the overall procedure.

A.5 Local Outlier Detection Performance Analysis

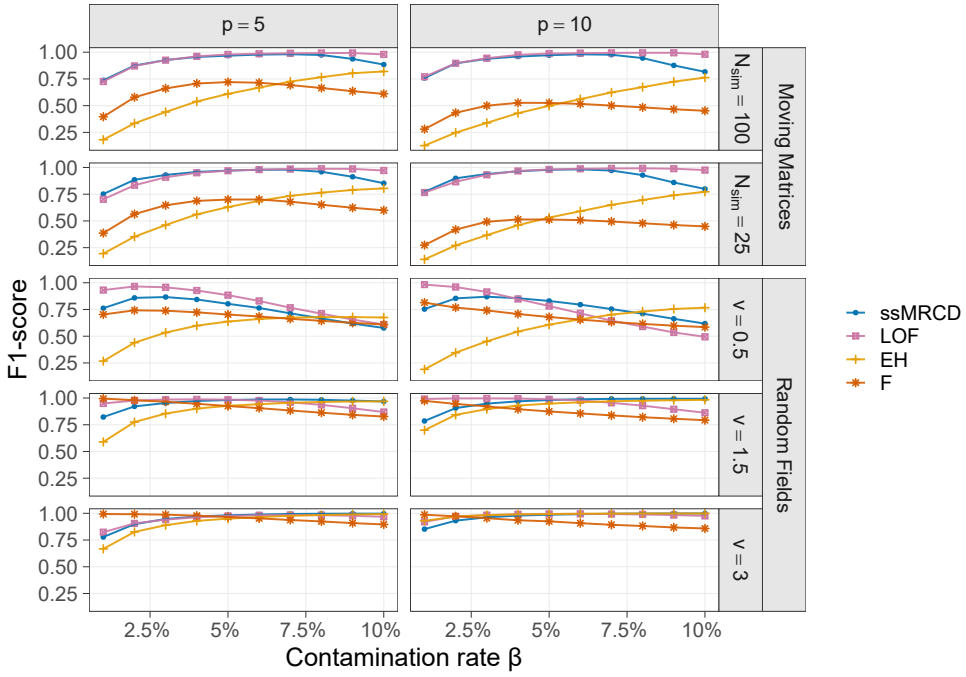


Figure A.13: F1-Score for all four outlier detection methods with varying contamination levels achieved through the switching method of Ernst and Haesbroeck (2016) for different scenarios. Each point represents the mean of 100 repetitions.

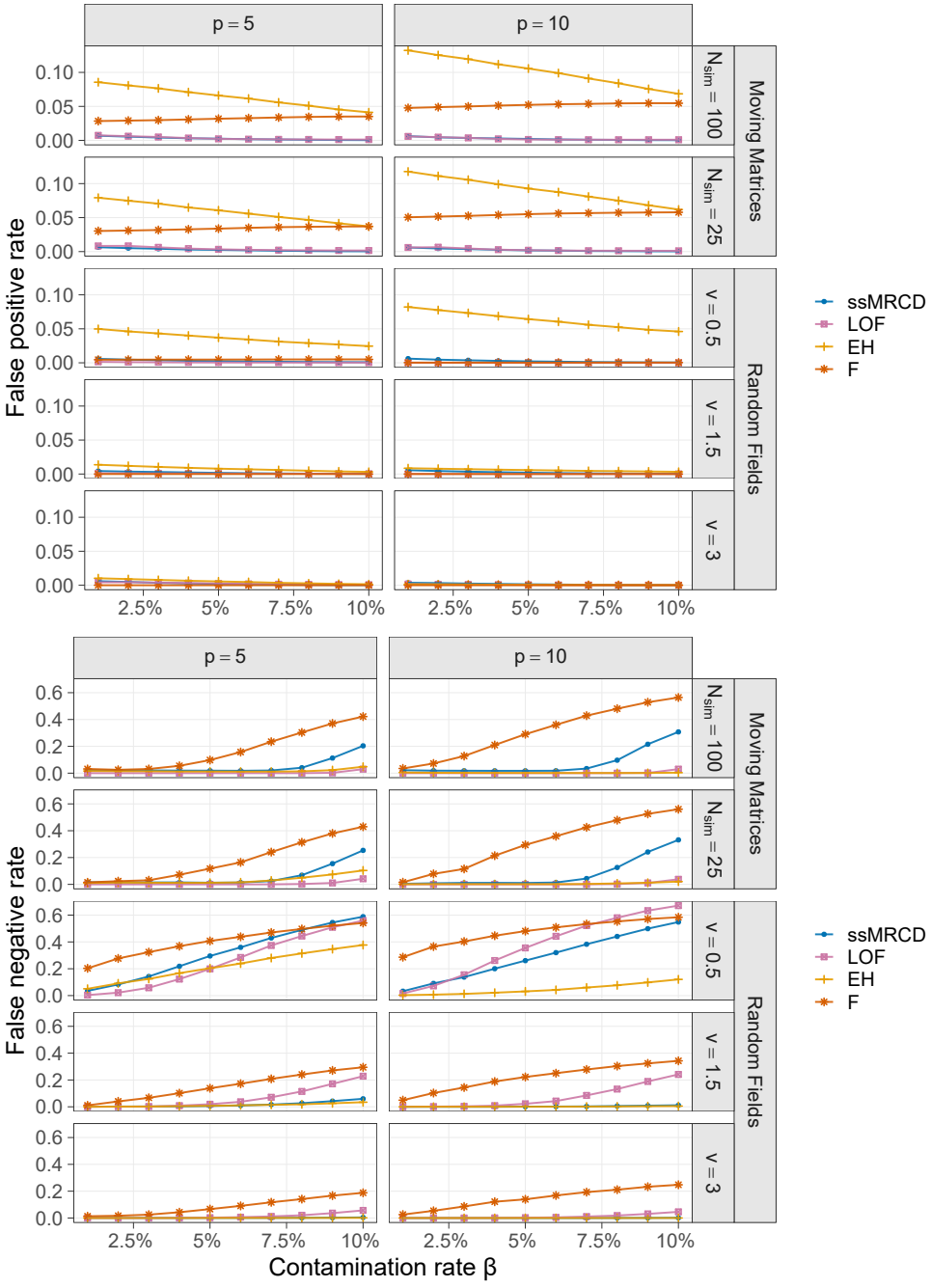


Figure A.14: False positive and false negative rate for all four outlier detection methods with varying contamination levels achieved through the switching method of Ernst and Haesbroeck (2016) for different scenarios. Each point represents the mean of 100 repetitions.

3 A Performance Study of Local Outlier Detection Methods for Mineral Exploration with Geochemical Compositional Data

This chapter was published as Puchhammer, P., Kalubowila, C., Braus, L., Pospiech, S., Sarala, P., and Filzmoser, P. (2024a). A performance study of local outlier detection methods for mineral exploration with geochemical compositional data. *Journal of Geochemical Exploration*, 258:107392. DOI: 10.1016/j.gexplo.2024.107392.

3.1 Introduction

Detecting multivariate outliers is one of the most important steps when analyzing any kind of data. Such outliers could arise from gross errors during data recording, they could be the result of inappropriate data preprocessing, or they could indicate observations which are indeed very different from the rest and thus point at unusual phenomena (Zimek and Filzmoser, 2018). The problem of outlier detection becomes more difficult when analyzing data with additional attributes that need to be considered, such as the locations of observations in a spatial data setting. Here, we are often not interested in the outliers found with standard methods (so-called *global outliers*) but we focus on observations that are anomalous with respect to their spatial surrounding. These observations are called *local outliers*, and they could indicate interesting locations to practitioners, e.g., unknown mineral deposits. On the other hand, methods which use locality (for example geographically weighted methods (e.g. Brunsdon et al., 1998) or geostatistical techniques (see e.g. Cressie, 2015) can also be heavily influenced by local outliers.

While the literature for local outlier detection is not as broad as for global outlier detection, there are still some (multivariate) methods available. We will focus on three methods based on the *pairwise Mahalanobis distance* (see Filzmoser et al., 2013; Puchhammer and Filzmoser, 2024) defined as

$$\mathrm{MD}_{\Sigma}(x,y) = \left[(x-y)^t \Sigma^{-1}(x-y)\right]^{1/2} \qquad \text{for } y \in A(x)$$

for two (multivariate) observations x, y with a robust covariance estimate Σ which can depend on the spatial attributes of x and y. The set-valued function $A(x)$ returns observations that are spatially close to an observation x. The three methods differ in their covariance estimation, specifically in the degree of its locality. The fourth

method comes from the area of machine learning and is solely distance-based. All of the four methods compare each observation with its k-nearest neighbors ($A(x)$ returns the k-nearest spatial neighbors) and calculate a degree of outlyingness that together with a method-specific cutoff value flag observations as outliers.

The first method introduced by Filzmoser et al. (2013), in the following called *robust local outlier detection method* (ROB), is available in the R-package `mvoutlier` (Filzmoser and Gschwandtner, 2012) and uses the pairwise Mahalanobis distances together with a global and robustly estimated covariance matrix, ignoring the spatial context of the data. The measure of outlyingness for each observation is based on theoretical properties connected to χ^2-quantiles. For more details we refer to the respective paper by Filzmoser et al. (2013). In contrast, the method of Ernst and Haesbroeck (2016), here called *regularized spatial detection technique* (REG), estimates local covariance matrices based on the k-nearest spatial neighbors for each observation separately. Thus, for a fixed observation x, the covariance estimation is only based on observations in $A(x)$. The measure of outlyingness (also called next-distance) is just the minimum of all MD, $\min_{y \in A(x)}$ MD (x, y) of each observation x, and the final cutoff value to determine outlyingness is the upper fence of an adjusted boxplot based on all next-distances. Next-distances above the cutoff value indicate local outliers. As a compromise between using only one covariance estimation and using a covariance estimation for the local neighborhood of each observation individually, the third method of Puchhammer and Filzmoser (2024) is bridging the gap by partitioning the space into groups (e.g. by country boundaries for socioeconomic data, or via grids or clustering for data without known clear grouping) and estimating a covariance matrix for each group using the so-called *ssMRCD estimator* implemented in the R package `ssMRCD` (Puchhammer and Filzmoser, 2023). The concept of next-distances from REG is also applied here to identify outliers. Simulation studies in Puchhammer and Filzmoser (2024) show that the method ROB tends to have an increased false negative rate since the global covariance matrix seems to not being strict enough in its estimation of the local covariance. The method REG leads to an increased false positive rate, because using only the k-nearest neighbors for the covariance estimation seems to be too strict by not putting the local estimation into the global perspective. Outlier detection based on ssMRCD includes some spatial smoothing among spatially close groups, and thus the broader structure is also taken into account which balances the false positive and false negative rate.

The last considered method for local outlier detection is the *local outlier factor* (LOF) introduced by Breunig et al. (2000) and adapted to the spatial setting according to Schubert et al. (2012). Since the LOF is purely (Euclidean) distance-based and does not use the pairwise Mahalanobis distance, there is no need to estimate a covariance matrix. Instead, a local density based on the Euclidean distance in the feature space is calculated for each observation and compared with the density of its k-nearest spatial neighbors. Formally, the base of the LOF is the so-called reachability distance g_k between two objects x and y which is defined by

$$g_k(x, y) = \max\{d_k(x), d(x, y)\}$$

where d is the Euclidean distance and $d_k(x)$ the (Euclidean) distance of x to its k-

nearest neighbor. The density used, also called the local reachability density, is defined by

$$lrd_k(x) = \left(\frac{\sum_{y \in A_k(x)} g_k(x, y)}{|A_k(x)|} \right)^{-1}$$

with $A_k(x)$ being the spatial k-nearest neighbors. If the density of an observation is considerably lower than the density of its neighbors,measured by a local outlier factor

$$LOF_k(x) = \frac{\sum_{y \in A_k(x)} \frac{lrd_k(y)}{lrd_k(x)}}{|A_k(x)|}$$

bigger than 1, the observation is considered a local outlier. The original LOF method of Breunig et al. (2000) is implemented in the R package `DescTools` (see Signorell et mult. al., 2017).

Finding these local outliers is quite important for mineral exploration especially in the context of geochemical data. Though there are a number of methods such as geological mapping, geochemistry, geophysical surveys and remote sensed imagery that are used in mineral exploration to find potential areas for mineral deposits (Marjoribanks, 2010), in this paper, we are focusing on a geochemical approach in connection with local outlier detection. In the areas having transported cover, such as glaciated terrains, mineral deposits are typically found as sub-outcropping under till-cover. In addition, many ore deposits locate buried under the bedrock surface or even hundreds of meters depth in the bedrock without outcrop on the surface. That type of buried deposits are challenging for the mineral exploration due to poor recognition with surface techniques. However, geochemical data of till and bedrock may provide good targeting criteria for identifying both sub-outcropping and buried mineral deposits. Local outliers reveal anomalous data points which highly deviate from the surrounding data variability in geochemical data sets and may be indicators for mineral deposits in geochemical explorations (Filzmoser, 2004). Thus, geochemical anomaly detection in general is crucial for exploring unknown mineral deposits, and applying local outlier detection techniques in particular can be beneficial in achieving this goal. The type of geochemical data (i.e. elements) that should be used to identify outliers and then predict possible deposits may depend on the type of targeted mineral deposit. When detecting Ni - Cu deposits, as an example, outliers can be associate with high Ni, Cu, PGE, Ti, V, S, Cr and Co (Maier, 2015).

In this context, also certain relations of element concentrations are often very insightful. This is connected to the compositional nature of element concentrations which is an essential aspect and needs to be addressed by any method when applied to geochemical data. While the assumption of a normal distribution seems valid for many measurements, the underlying distribution of geochemical data has an inert structure that must not be ignored. Since geochemical measurements (also called analytical results) constitute a composition of elements, the sum of the concentrations or *parts* of each sample is fixed to the same number. Thus, the underlying geometry of the data is not the Euclidean but the Aitchison geometry (Pawlowsky-Glahn et al., 2015) and

the relevant information is not in the absolute values but in the pairwise (logarithmic) ratios of the parts. Although this geometry seems complicated, many methods can be applied after appropriately transforming the compositional data to the Euclidean geometry while additionally taking the original structure (i.e. the simplex) or the pairwise (logarithmic) ratios of the parts into account for interpretation.

There are various transformations suited for this task (see, e.g., Filzmoser et al., 2018). We will focus on two of them that are easy to apply and have good theoretical properties. The first transformation leads to the so-called *centered-logratio (clr)* coefficients. For a composition $\boldsymbol{x} = (x_1, \ldots, x_D)$, the clr transformation is defined as

$$clr(\boldsymbol{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{k=1}^{D} x_k}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{k=1}^{D} x_k}} \right), \tag{3.1.1}$$

which is essentially the logarithm (per variable) of the observed composition standardized by its geometric mean. The clr-transformation is isometric, meaning that it preserves the distance of the Aitchison geometry when using the Euclidean distance in the transformed space. Also, the interpretation is desirably straightforward and based on relative information with respect to the (geometric) mean. However, a drawback present in many applications is that the transformed data matrix does not have full rank since clr coefficients are based on a generating system and not on a basis of the Aitchison geometry. This can be overcome by using one of infinitely many orthonormal coordinates. The transformation of choice in this paper is based on *isometric logratio* (ilr) coordinates and known under the name *pivot coordinates* (e.g. Filzmoser et al., 2018). The j-th entry of the pivot coordinates of $\boldsymbol{x}$ is defined as

$$ilr(\boldsymbol{x})_j = \sqrt{\frac{D-j}{D-j+1}} \ln \left(\frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}} \right) \tag{3.1.2}$$

for $j = 1, \ldots, D-1$. Since an orthonormal basis is used, we reduce the dimension of the transformed composition by one and resolve the problem of singularity in general present for the clr-transformation. Up to a constant we have equality in the first entry of the ilr and clr transformation, $ilr(\boldsymbol{x})_1 \propto clr(\boldsymbol{x})_1$, and thus, the first entry of ilr coordinates can be interpreted just as easily as the clr-transformation. Note that although there is a close connection between the first coordinates, it should be kept in mind that ilr coordinates represent dominance while clr indicates the average of a composition. However, this close relationship does not apply to the other coordinates of the ilr-transformed composition which is essentially the one major disadvantage of the orthonormal basis. We will use both transformations according to their properties, and choose the transformation based on the questions and requirements arising in our data analysis.

In this paper we analyze geochemical data by applying local outlier detection techniques to three data sets differing in scale and data quality. We show the importance of data preprocessing steps and the usage of compositional data analysis methods, describe the problems encountered with data having insufficient quality and debate

possible solutions that adequately account for the compositional nature. Moreover, we show how different local outlier detection techniques perform on different scales and analyze in which cases some methods might be less appropriate to find mineral deposits. Some ideas on outlier diagnostics, method evaluation, and filtering of outliers based on common compositional data transformations are also discussed to complete a thorough local outlier analysis in the compositional data setting.

The paper is organized as follows. In Section 3.2 we describe the three data sets and corresponding preprocessing steps before applying the four local outlier detection methods in Section 3.3. The final two sections summarize and discuss the findings and provide overall conclusions.

3.2 Data Description and Preprocessing

For illustration purposes we choose three data sets differing in spatial scale, sampling scale and data quality to showcase the differences and specifics of the four selected outlier detection methods. The locations of the samples of the different data sets are depicted in Figure 3.2.1.

The first data set is the so-called *GEMAS* data set described in Reimann et al. (2014a,b). The data consists of agricultural soil samples that cover most of Europe in a density of 1 sample per 2500 km^2, see Figure 3.2.1 left. The 2108 samples were analyzed by X-ray fluorescence, following tight quality control procedures, resulting in concentration values for 41 chemical elements. Here we use the data set published in the R-package `robCompositions` (see Templ et al., 2011), named as data set `gemas`. It contains only elements with less than 3% of the analytical results below the detection limit, resulting in 18 main elements with good data quality.

The other two data sets are used for till geochemical analysis (regional till geochemistry, targeting till geochemistry and mineral deposits) in Finland. They are provided by the Geological Survey of Finland (1995, 2013, 2016) (GTK) and modified as described below. The regional till data set covers whole Finland and it has been collected during the period of 1983 to 1991. This data set contains the concentrations of 22 - 26 elements (depending on the map sheet – in the selected area we have 22 elements available), see Table 3.2.1. The samples have been collected from the C horizon, which contains unaltered till. The sampling depth is approximately 1.5 - 2 m. The sampling density is 1 sample per 4 km^2 and the full data set comprises of 82 062 samples. Furthermore, concentrations of 22 - 26 elements that can be extracted by aqua regia have been analyzed for fine fraction of the till material less than 0.06 mm and the data has been published by 1:400 000 map sheets (Salminen and Tarvainen, 1995).

The final data set, the targeting till geochemical data set, contains around 385 000 soil samples collected by GTK along sample lines in certain areas between the years from 1971 to 1983. Most of them are till samples, however samples from sorted mineral soils, weathered bedrock and mixed intermediate forms also exist in the data set. In this paper, only till samples collected using percussion drilling and test pitting methods from the C horizon which contain fine (less than 0.06 mm) fractions are considered. The samples have been collected by 1:100 000 map sheets. The point density of the

samples lies between 6 - 12 samples per 1 km^2 where the line interval is 500 – 2000 m and the distance between two points is 100 - 400 m. The average depth of the samples is 2 m, and an emission quantometer method has been used to measure the concentration of 17 elements listed in Table 3.2.1 (Gustavsson et al., 1979).

For the data analysis in Section 3.3 we do not use the complete regional and targeting till data sets, but choose data subsets covering only the area from Central Lapland depicted in Figure 3.2.1 (right), which is partitioned into four smaller areas or *map sheets* by GTK. This area contains many known mineral deposits and provides sufficient data quality in terms of enough reliable measurements, which is not provided in all areas for the targeting till data set. By taking the same sampling area for these two data sets, we are also able to compare their usefulness for mineral exploration with local outlier detection methods.

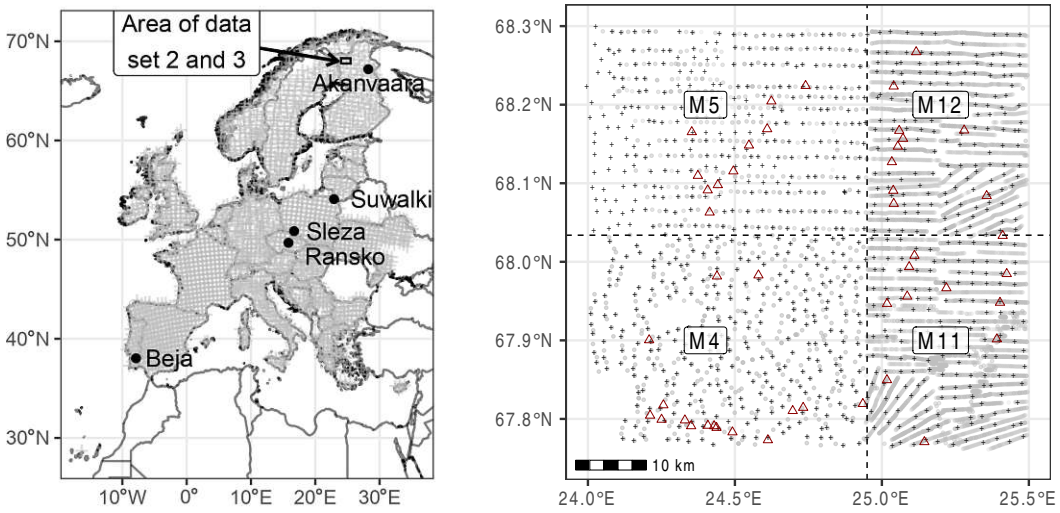


Figure 3.2.1: Map of research areas. *Left:* Grey crosses indicate sample locations of the GEMAS project while the black dots represent the reference sites of the SEMACRET Project (2023). The rectangle in the Northern part of Finland represents the four selected map sheets of the regional and targeting till data set shown on the right. *Right:* Sample locations of regional till (black crosses) and targeting till (gray dots) data, partitioned into four map sheets. Each triangle indicates a known mineral deposits.

3.2.1 Data Preprocessing

The element selection based on the detection limit threshold of 3% for the GEMAS data set constitutes a compromise between rejecting too many elements, and keeping too many elements with low data quality. Removing said elements ensures that most of the reliable information of this data set is extracted. Due to the high data quality in general, no further preprocessing is necessary.

The selected subset of the regional till data set generally has good data quality. However, for some elements it contains values below the lower detection limit, and other data quality issues. Therefore, additional data cleaning is required. The right part of Table 3.2.1 shows the percentages of values with the data problems mentioned before for the selected 4 map sheets. We decided to exclude Zr and Th for all further analyses. The remaining data quality issues are not connected to detection limit problems, since only Zr and Th where erroneous in this way. A small amount of analytical results have additional markers in the data with unclear encoding. The benefit of the additional information saved trough this procedure outweigh possible negative effects on data analysis and the differently marked analytical results are kept in the data. We refer to Mert et al. (2016) for an analysis of contamination on compositional data transformations. The resulting regional till data set has thus 870 samples and 20 variables.

Compared to the previous data sets, the targeting till geochemical data set has serious data quality issues, typically connected to values related to detection limits, zero and even negative analytical results, and values marked with special symbols. Therefore, extensive data cleaning is required in order to perform further statistical analyses and modeling procedures. The percentages of insufficient quality of samples per element and map sheet areas are calculated and shown in Table 3.2.1 left. Eventually, elements which contain more than 30 percent of problematic samples over all map sheets (e.g., Ag, Pb and Zn) are removed from all further analyses.

Furthermore, the geochemical analysis of the targeting till data set has been carried out at different times and map sheets. Therefore, it is necessary to analyze a possible mismatch and if the measurements are comparable. Figure 3.2.2 illustrates the spatial concentration of Fe in both till data sets separately. It is evident that there are discontinuities at the map sheet boundaries in the targeting till data set due to inconsistencies during the geochemical analysis done by quantometer method. These discontinuities are not present in the regional till data, where there is a change of geological units from Archean and Proterozoic to only Proterozoic origin visible. Thus, after displaying clearly visible map sheet boundaries and discrepancies between map sheets at least for Fe that are not due to the underlying geology it was decided to analyze the map sheets (1:100 000 scale) separately for the targeting till data set, as the smaller areas also contain enough sample points to carry out the analysis.

To improve data cleaning further, also Q-Q plots are used to examine the distribution of concentrations between different map sheets in the regional as well as in the targeting till data set where only elements with less than 30% of quality issues are included. For the Q-Q plots, we focus on the elements Co, Cr, Cu, Fe, Ni, V, and Ti, which are important ore metals in ultramafic rocks, and thus of special interest for mineral exploration. As example, the concentration values of Fe for all four map sheets separately are shown by Q-Q plots for the targeting till data set in Figure 3.2.3(a) and for the regional till data set in Figure 3.2.3(c) as well as the corresponding clr transformed values in Figure 3.2.3(b) and in Figure 3.2.3(d), respectively. The Q-Q plots for Fe vary between map sheets (M4, M5, M11, M12) but especially between the two data sets. With respect to the average concentration level per map sheet we even see adverse ordering in original as well as clr transformed values for regional

	Targeting till (%)				*Regional till (%)*			
Element	M4	M5	M11	M12	M4	M5	M11	M12
Ag	100	100	100	100	-	-	-	-
Al	91.17	93.84	4.64	16.63	0	0	0	0
Ba	-	-	-	-	0	0	0	0
Ca	1.13	2.02	27.09	53.87	0	0	0	0
Co	16.27	18.52	3.04	8.24	0	0	0	0
Cr	86.65	6.82	6.33	4.16	0	0	0	0
Cu	1.27	4.39	0.96	1.93	0	0.36	0	0
Fe	0.03	0.87	0.73	1.80	0	0	0	0
K	1.96	3.44	35.66	9.60	0	8.02	0	0
La	-	-	-	-	0	0	0	0
Li	-	-	-	-	0	0	0	0
Mg	0.03	0.06	0.01	0.03	0	0	0	0
Mn	2.17	1.14	2.71	2.74	0	0	0	0
Na	0.37	0.40	15.90	4.38	-	-	-	-
Ni	0.24	0.33	0.18	0.46	0	1.45	0	0
P	-	-	-	-	0	0	0	0
Pb	99.58	91.95	97.35	97.71	-	-	-	-
Sc	-	-	-	-	0	0	0	0
Si	0	0.47	0.01	0.07	-	-	-	-
Sr	-	-	-	-	0	0	0	0
Th	-	-	-	-	5.88	15.32	9.52	1.92
Ti	0	0.27	0.01	0.01	0	0	0	0
V	0.27	0.94	1.03	3.72	0	0	0	0
Y	-	-	-	-	0	0	0	0
Zn	98	22.04	91.48	60.58	0	0	0	0
Zr	-	-	-	-	11.02	7.66	36.90	44.87

Table 3.2.1: Percentages of problematic data quality of the targeting till and regional till data set for different elements with respect to corresponding map sheets. The values of the elements per map sheet used after the final data cleaning are underlined.

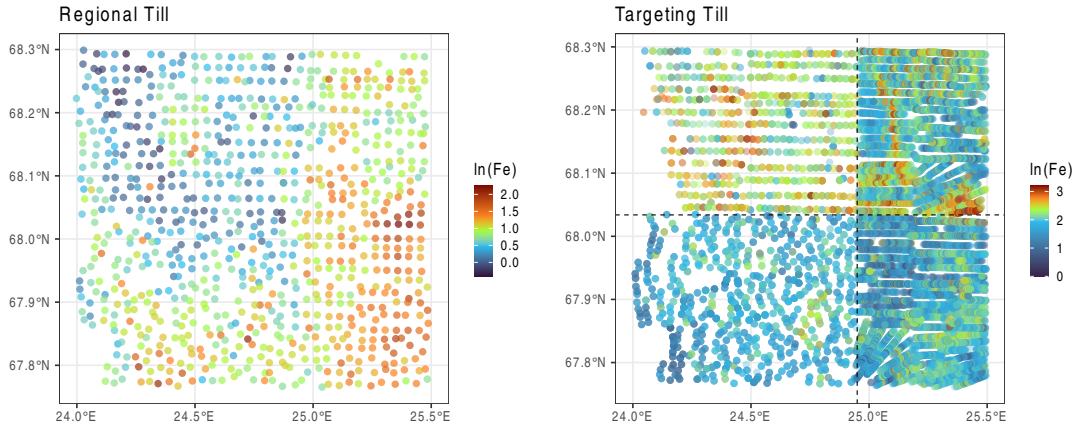


Figure 3.2.2: Illustration of discontinuities of Fe (%) between map sheets in the targeting till (right) compared to regional till data set (left). Clear boundaries are visible between the map sheets in the targeting till data set.

and targeting till, which is congruent with Figure 3.2.2. Note that Q-Q plots alone are not sufficient to diagnose map sheet leveling problems but the adverse ordering could still be a strong indicator of them. However, other quantitative differences between the two data sets might be mainly due to different analytical techniques. Interestingly, the clr transformation based in the regional till data set reorders the average relative level of Fe between map sheet M4 and M12 indicating that using the appropriate compositional data structure adds important information which would be ignored otherwise. Regarding differences between the map sheets per data set for other elements (Co, Cr, Cu, Ni, V, and Ti) shown in B.1 it is less clear whether they originate from map sheet problems or from spatially changing lithology. Finally, the distributions of elements for all map sheets, elements and data sets seem to be plausible. Apart from some lower detection limit problems in the targeting till data set, which will be taken care of in the next step, we do not need to account for any extensive rounding, grouping or other distributional issues that might occur.

After the extensive map sheet analysis of the targeting till, the final data cleaning is necessarily done per map sheet. In order to use as many elements as possible, we start by removing samples that have at least one zero value of element concentration. Also observations with more than 30% of problematic values over all elements provide a restricted amount of reliable information and are removed. Due to the high sample density, we still keep enough observations to make sensible analysis when applying the rather strict row cleaning (M4: 2417 samples, M5: 1399 samples, M11: 5821 samples, M12: 4557 samples). Note, that for data sets of lower sample density, the decision between having less samples or less elements available after data cleaning is less clear than in this case. Finally, only the elements that have less than 5% problematic values per map sheet are used, which are underlined in Table 3.2.1. This is again rather strict, but we hope to reduce the number of local outliers connected to poor data quality

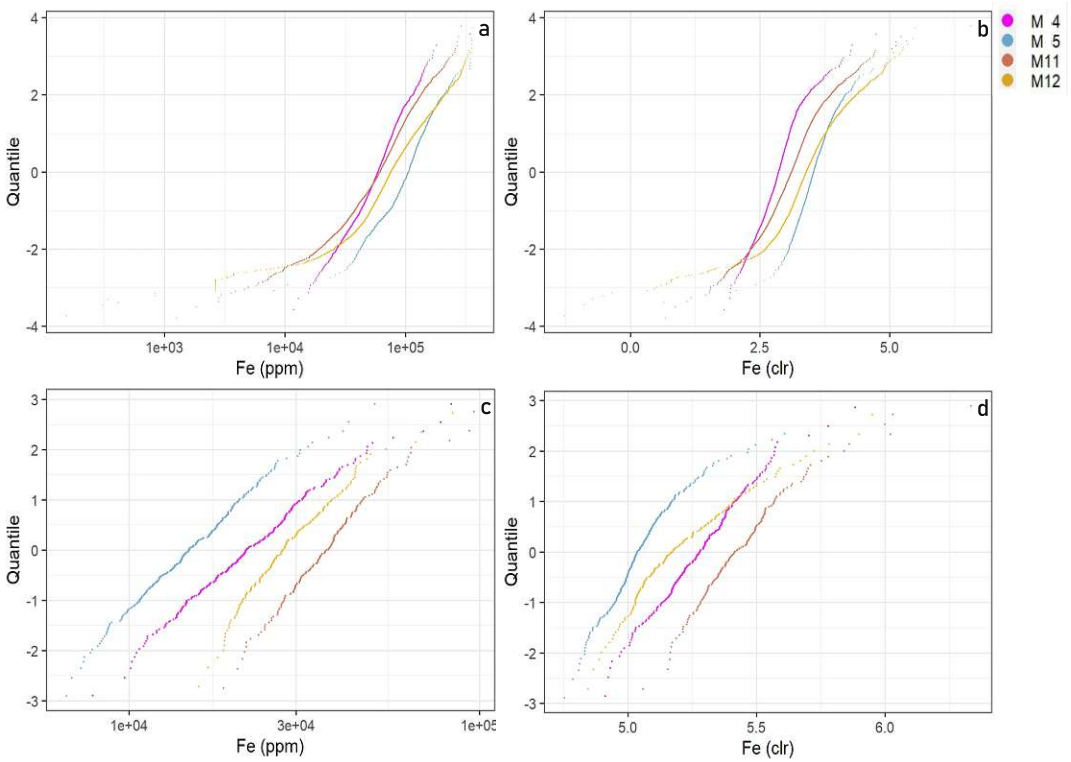


Figure 3.2.3: Q-Q plots of Fe: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till, n (targeting till) = 16460, n (regional till) = 870.

or detection limit problems. Imposing the even stricter limit of 3% similar to the GEMAS data set is not applicable for this data set, since many more elements would be lost for the analysis. However, note that the effects of some data quality issues on compositional transformations will be present but limited (Mert et al., 2016).

After the final data cleaning of both data sets, targeting and regional till, the last preprocessing step is to address the compositional nature of the geochemical data (for targeting till again per map sheet). Although the clr-transformation is isometric regarding the Aitchison geometry and easy to interpret, the linear dependency introduced is problematic. In the case of covariance estimation we would get a singular matrix which is not invertible. However, this is a necessity for the pairwise Mahalanobis distance and the three methods based on it. Thus, it is sensible to choose ilr coordinates for the regularized spatial outlier detection technique, the robust local outlier detection method and the ssMRCD-based outlier detection technique to avoid this problem. Since LOF does not need a covariance estimation and is strictly (Euclidean) distance-based, any transformation for compositional data which is isometric can be applied. Thus, both ilr and clr can be used, and due to isometry both lead to the same local outlier factor and thus, to the same local outliers.

3.3 Data Analysis

After finishing the preprocessing and data cleaning steps and the compositional data transformations, the four outlier detection methods can be applied to the transformed data. Regarding the parameters, we generally adhere to default settings wherever sensible. For all four methods we compare each observation with the same amount of k nearest neighbors. Setting k influences the locality of the local outlier detection in all methods since we are looking for an anomaly compared to samples from a larger area. For a more detailed analysis of the effects of different k values we refer to Braus (2023). Since the considered data sets differ in sample size and density, k is adjusted to the data sets. The parameters for the ssMRCD-based method are a smoothing parameter $\lambda = 0.5$, representing a compromise between local and global covariance estimation. Neighborhoods are defined data specific and the weighting matrix for smoothing between neighborhoods is defined as the pairwise inverse distance of the neighborhood spatial centers, which is the most natural choice for data without inherent spatial structural breaks. For LOF a value above 1.5 is flagged as outlying, and for the regularized spatial detection technique we want to include all observations ($\beta_{REG} = 1$) independent of local heterogeneity. As regularized covariance estimator, the Minimum Regularized Covariance Determinant estimator (Boudt et al., 2020) with a trimming percentage $\alpha = 75\%$ is chosen, meaning that 75% of the k-nearest neighbors are used for the local covariance estimation. Regarding ROB, all other parameters including the amount of neighbors allowed to be similar as well as the cutoff value are adjusted to the spatial scale of the data.

3.3.1 GEMAS Data

For the GEMAS data set we use the following parameter setting for the different local outlier detection methods: We choose $k = 10$ for all methods, a standard settings, and $N = 50$ neighborhoods for the ssMRCD estimation which reflects an appropriate level of locality given the sample density and ensures sufficient observations per neighborhood. The neighborhoods are selected by k-means clustering based only on the spatial attributes of the data and checked for reasonably sized spatial clusters. For the robust local outlier detection method (ROB) we allow 10% of the neighbors to be similar ($\beta_{ROB} = 0.1$) due to the large area covered and the sparse sampling of the GEMAS data set, and choose an isolation degree bigger than 0.2 as cutoff value. These parameter choices are also supported by the work of Braus (2023). As some indicator for performance we also include the reference sites of the SEMACRET Project (2023) in Figure 3.2.1 (left). Finding these reference sites can be interpreted as analysis goals. However, on the one hand this approach is unbalanced since a high number of found outliers already leads to an improved performance, and on the other hand unknown mineral deposits are not taken into account. Nevertheless, we get more insight into possible drawbacks of different methods.

The flagged outliers per method are shown in Figure 3.3.1. Starting with the most global method, the robust local outlier detection method (ROB), problems connected to the global covariance estimation are evident. For the robust covariance estimation, the MCD estimator (Rousseeuw, 1985) is used which selects a subsample of the data with the lowest determinant of the sample covariance based on this subsample. On the GEMAS data set, the subsample contains mainly observations from Middle to Northern European countries, thus leading to a covariance not representing Southern Europe and to an unbalanced and somewhat biased spatial distribution of the outliers. For the regularized local outlier detection technique (REG), the outliers are more or less evenly distributed. However, the problem of a high number (almost 15% of the observations) of outliers arises which is likely connected to an increased false positive rate. The high amount of outliers makes it difficult to get more valuable insight into the data rendering the method essentially useless without further processing. Both, the ssMRCD-based and the LOF-based outlier detection method seem promising, however it seems that the ssMRCD finds most reference sites, including a strong signal very close to the ultra-mafic intrusion body in Beja (see Figure 3.3.2b). Note that the mineral deposit in Suwalki in North Poland is assumed to be multiple hundreds of meters deep under the surface, so it is unlikely that it affects the soil sufficiently. Moreover, LOF does not flag the soil sample from the Canary Islands as outlying which would be sensible given that the ten nearest neighbors are located far away somewhere in South Spain.

Although the element selection in the GEMAS data set might not be oriented towards mineral exploration, the data quality is very good and it is well suited to discuss further processing steps. After applying the four methods we end up with many potential locations for mineralisation or other anomalous observations. Thus, a closer look at the identified outliers is quite important since finding mineralised areas can be very expensive, and additional analysis can improve the identification of important locations. We are interested in utilizing potential mineralisations by mining, hence high values of

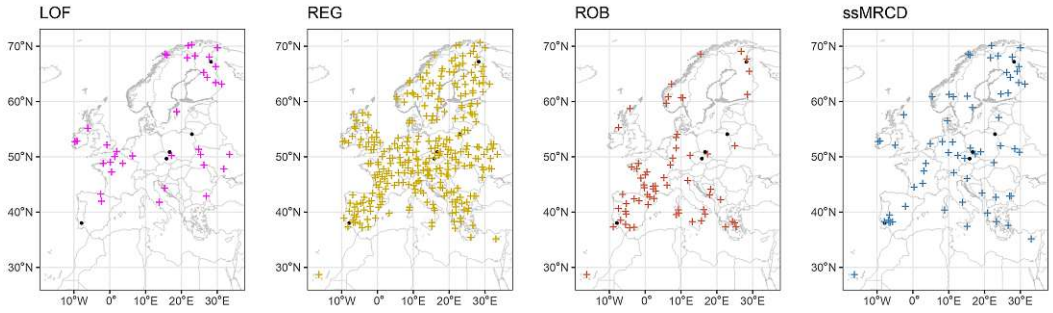
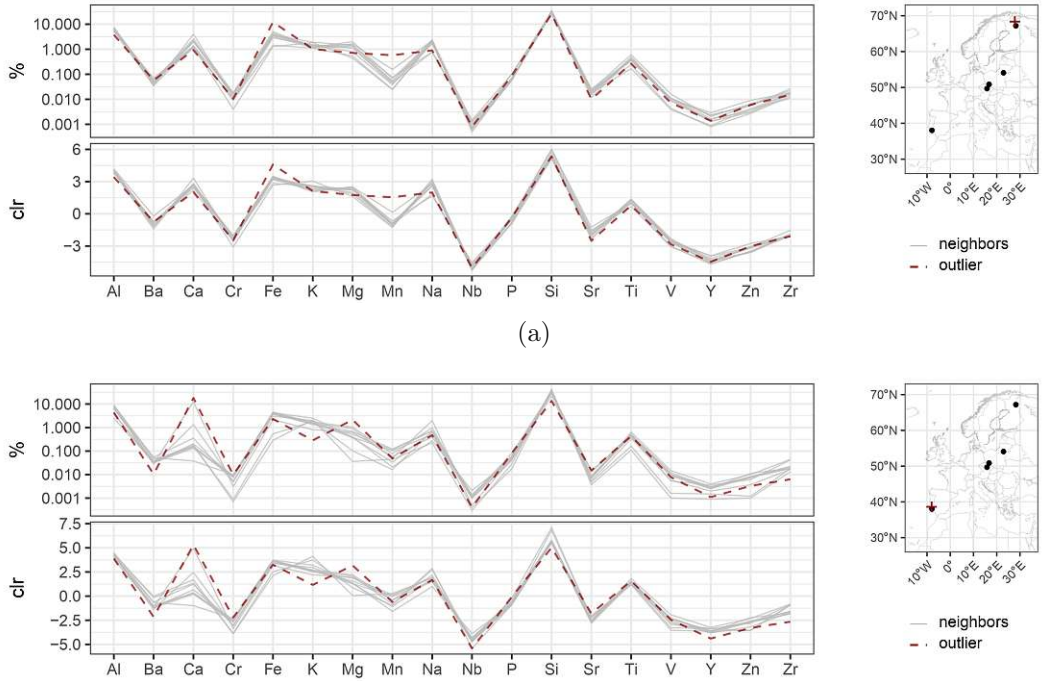


Figure 3.3.1: Spatial locations of flagged outliers (marked as cross) by each method separately on the GEMAS data set. The black dots represent areas of interest in the SEMACRET Project (2023) where mineral deposits are anticipated.



(a)

(b)

Figure 3.3.2: Outlier diagnostics for (a) observation 530 which is closest to the Akanvaara area, and for (b) observation 189, closest to the ultra-mafic intrusion body in Beja, each colored in red. The two parallel coordinate plots show the multivariate structure of the observations and corresponding 10 nearest neighbors in gray, once in percent (upper part) and once in clr-transformed values (lower part).

elements in clr coordinates (meaning high concentrations relative to other elements) but also in total concentrations are desirable. Thus, we employ a filtering procedure on all flagged outliers keeping only those observations which have clr and measured concentration levels simultaneously above the global 95% quantile for at least one element. In Table 3.3.1 the total numbers of outliers, filtered and unfiltered, are shown for all three data sets. Depending on geological knowledge and the type of mineral deposits that should be found, the selection of elements for the filtering procedure can be further specified in concrete applications. For finding potential Ni-Cu mineralisation, the elements in the filtering procedure can be tailored specifically to high Ni and Cu and other connected elements (see also subsection 3.3.3).

The number of outliers is reduced by the filtering procedure, but for single observations we can still improve on the analysis to increase the chance of finding valuable mineral deposits. A possible diagnostic tool is based on parallel coordinate plots which can give insight into the multivariate structure. Each observation is represented by a line, and the values of each variable on the horizontal axis are connected. We focus on the comparison with the k-nearest neighbors. Together with insight into the underlying bedrock, the corresponding observation can be interpreted as interesting new target for further exploration or discarded as uninteresting anomaly. In Figures 3.3.2a and 3.3.2b, two of the flagged outliers which are closest to the Akanvaara area and Beja are analyzed in comparison to their 10 nearest neighbors (colored in gray) using the parallel coordinate plot.

Regarding the outlier next to Beja in Portugal, which was flagged only by the ssMRCD-based method, we see high values in Ca and Mg and particularly low values in K. This fits well to the known geology in this region. While the neighbors are mostly located on sand (3 samples) and on the South-Portuguese Flysch zones (4-5 samples) which are composed of higher Al, Si, Fe, K as well as hardly any Ca and Mg (Jorge et al., 2013), respectively, the flagged outlier lies on the layered Gabbroic Sequence at Beja which is consistent with the elemental composition of the outlier as it contains olivine bearing gabbroic rocks which are bordered by heterogeneous diorites (Jesus et al., 2014). Gabbro usually contains minerals which associate with Ca and Mg such as pyroxene, plagioclase, and olivine of which weathering release Ca and Mg. The depicted high values in Ca and Mg are thus indicators for the Caliche type of weathering, which is typical in that type of climate for (ultra-)mafic lithologies. Also, low Si and slightly higher Cr with respect to neighbors indicate weathering of gabbroic rocks.

For the outlier indicated by the methods LOF, REG and ssMRCD near the Akanvaara deposit and the so-called Koitelainen deposit north-western of Akanvaara, higher values can be observed for Fe and Mn with respect to the nearest neighbors (Figure 3.3.2a). The Akanvaara deposit is located in Northern Finland (eastern part of the Central Lapland greenstone belt) and it is considered as a layered mafic intrusion which hosts vanadium mineralisation in layers of magnetite gabbro and also in chromitite layers within gabbro. These two layers have been mineralised by massive, semi-massive and disseminated magnetite, pyrite, chalcopyrite and chromite (Lutynski, 2019). Koitelainen also an ultramafic deposit which is enriched by commodities such as Cr_2O_3, V, Fe and PGE. The flagged outlier is closer to the Koitelainen deposit than the Akanvaara

deposit where the distances from the outlier to the deposits are approximately 17 km and 72 km respectively. Thus, when considering the flagged outlier for these deposits, elevated amounts of elements such as Cr, V, Cu are also expected other than Fe in order to identify it as an indicator for Akanvaara and Koitelainen. However, the GEMAS data are for grassland areas and Akanvaara locates inside largely forested area without close vicinity to grasslands. Furthermore, since this flagged outlier associates with only high Fe and Mn, it cannot be 100 percent certain that it indicates the Akanvaara or Koitelainen deposits, but it is certain that it indicates a mafic environment where there is a possibility for a mineral deposit.

3.3.2 Regional Till Data

For the regional till data set some parameter settings are adjusted. We again compare single observations with their $k = 10$ nearest neighbors. For the ssMRCD-based method, each of the 4 map sheets is chosen as an own neighborhood. This choice is due to the very dense sampling grid, but simulations in Puchhammer and Filzmoser (2024) also suggest that the method is rather insensitive to the number of neighbors, as long as some smoothing by the parameter λ is performed. For the robust local outlier detection method (ROB), we increase the percentage of neighbors allowed to be similar to 30% ($\beta_{ROB} = 0.3$) due to the smaller scale of the sampling area and choose an appropriate cutoff value for the isolation degree of 0.4. We refer to Braus (2023) for sensitivity analyses with respect to the choice of these parameters.

Interestingly, due to the smaller scale of the data we have the advantage of known mineral deposits (Geological Survey of Finland, 2016). There are 48 known mineral deposits of various types in the research area depicted as red rectangles in the right part of Figure 3.2.1. Ideally, our methods find these locations. However, since generally there are no samples directly on the deposits, we define a deposit to be found if an outlier is located 4 km or closer to the deposit. This might seem like quite far, but for an average density of one sample per 4 km^2 and historical glacial movement this distance is quite reasonable. Note, that this is not a guarantee that the outlying sample detecting the deposit has a typical element composition connected to the specific deposit type. Hence, it might be possible, that the sample is outlying due to other processes. Moreover, it would be preferable if the methods find the deposits as the most extreme outliers. Thus, we rank the outliers according to their outlyingness value, and analyze how many deposits are found until which outlier rank.

The left part of Figure 3.3.3 shows how many deposits are found by outliers up to the rank depicted on the horizontal axis for the regional till data set, with and without the filtering procedure described in the prior subsection. We see that filtering outliers reduces the number of outliers overall. However, for the regularized spatial outlier detection technique, the ssMRCD-based method and also the LOF there is an improvement in accuracy, meaning more deposits are found earlier. The degree of the improvement differs among methods, from strong for REG to negligible for LOF. Nevertheless, the filtering tool proves to be valuable if a subselection of outliers is necessary.

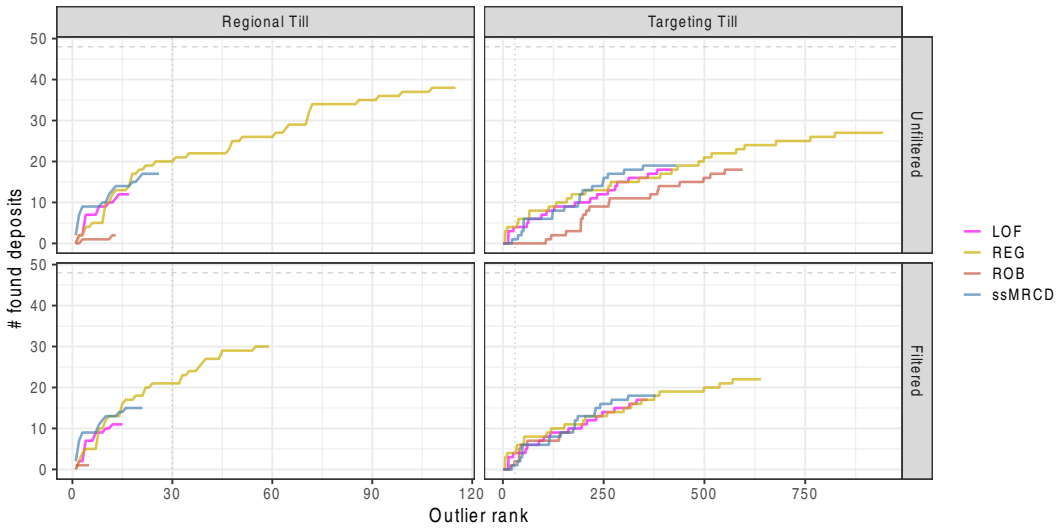


Figure 3.3.3: Performance of local outlier detection methods on regional (left) and targeting (right) till data for filtered and unfiltered flagged outliers. The dashed line represents the number of known mineral deposits. The methods applied to the regional till data set have better performance than for targeting till, as can be seen for the first 30 outliers (dotted line).

3.3.3 Targeting Till Data

Finally, the targeting till data set is used for the analysis. As discussed in Section 3.2 it is most sensible to analyze the four map sheets separately. The data also provides a structure of smaller sub-mapsheets, 12 for M4 and M5 and 6 for M11 and M12, respectively, that are used as neighborhood structure for the ssMRCD-based method. The only other parameter setting that is changed compared to the regional till data analysis is k, the number of neighbors to be compared with each observation. Due to the high sampling density, we increase k to 30 to find appropriate local outliers. Since we have 16 times more observations than in the regional till data set for the same area, the necessary distance to a known mineral deposit for it to be defined as found is reduced to 1 km in order to compare the performance of the data sets fairly.

Due to the separation of the map sheets in the analysis and the fact that we have a different set of elements per map sheet, we cannot compare the degree of outlyingness without adjustments. Thus, for each map sheet the outlyingness is standardized with its cutoff value to reduce the effects of separate analysis, and then the observations is ranked jointly by the standardized outlyingness.

The results can be seen in the right panel of Figure 3.3.3, again with unfiltered and filtered outliers. The methods flag many samples as outlying and for ROB and REG filtering significantly reduces the number of outliers while increasing accuracy. Ideally, the curves would jump at the very beginning up towards the number of known deposits. We can see that the ssMRCD-based method is closest to the ideal, both with and without filtering of outliers.

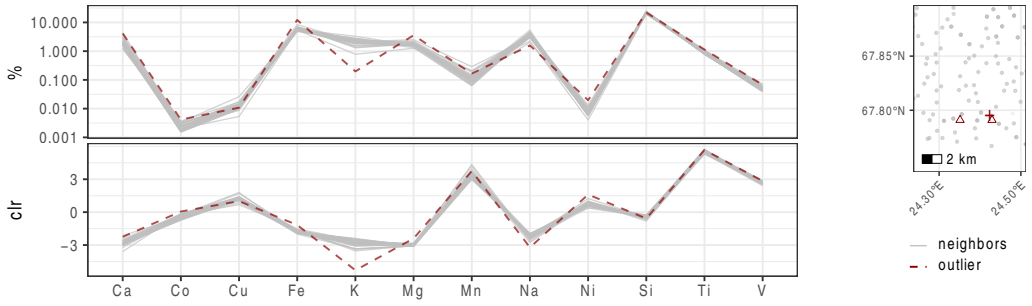


Figure 3.3.4: Outlier diagnostics for an outlier (red cross) close to the Saattopora-Cu deposit (right triangle). The two parallel coordinate plots show the multivariate structure of the observations and its corresponding 30 nearest neighbors in gray, once in percent (upper part) and once in clr-transformed values (lower part).

As mentioned before it is also possible to use a specific set of elements for filtering that match a deposit type of interest. As illustration we now try to find a (known) Ni-Cu deposit by filtering according to Ni, Cu, Ti, V, Co and Cr (see section 3.1). Three of the four methods (LOF, REG, ssMRCD) flag the sample analyzed in Figure 3.3.4 as outlier, which is less than 1 km away from the Saattopora-Cu deposit hosting Cu together with Au, Ni, Co and Ag. High values in Ni, Ti and Co of the flagged sample imply that the Ni-Cu deposit is connected to its outlyingness.

Comparing the results of the two data sets shown in Figure 3.3.3, we can clearly see that significantly more mineral deposits are potentially found by less flagged outliers using the regional till data set. In the case of mineral exploration this is definitely desirable since each outlier would need to be analyzed more closely. By providing that valuable outliers have high ranks in outlyingness, the effort and time spent on additional analysis is reduced. Note that outlier detection with the regional till data set might be more accurate than with the targeting till data set just because of the availability of more elements. This seems to be an important factor in finding certain types of ore deposits compared to a higher sampling density.

Another interesting approach is to analyze if the (potentially) found mineral deposits are the same or if the data sets lead to different results. In Figure 3.3.5 the outlier rank of the found deposits for both data sets are shown, for filtered and unfiltered outliers, and summarized in Table 3.3.2. In most cases, the number of found deposits is hardly affected by the filtering procedure. This indicates that the filtering process designed for subselecting outliers really leads to more accuracy in finding mineral deposits. Again, we can see that analyzing outliers from the regional till data set is effectively detecting ore deposits since many of them are found with a much lower outlier rank. Interestingly, the ore deposits found differ between the data sets used. This reflects also the size and type of ore deposits which would mean that with sparse sampling grids bigger outcropping or sub-outcropping mineralisations are possibly found but with increased sampling density the detection of smaller sub-outcropping and buried

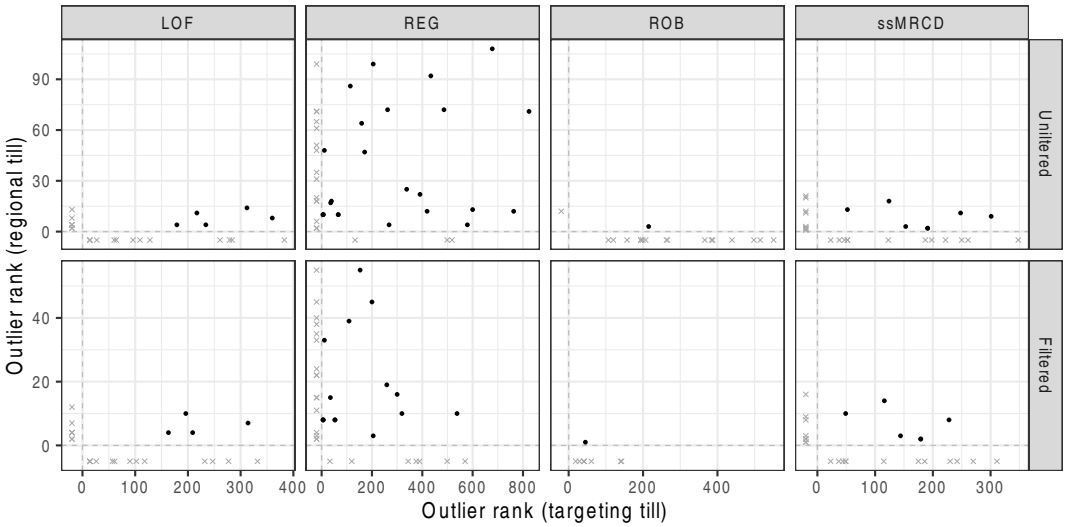


Figure 3.3.5: Found mineral deposits per outlier rank for unfiltered outliers (a) and filtered outliers (b) in either the regional and/or the targeting till data set. Jointly found deposits are marked by dots, deposits found only by one method are marked as gray crosses.

deposits is improved.

	Unfiltered			*Filtered*		
Method	GEMAS	Regional	Targeting	GEMAS	Regional	Targeting
LOF	36	17	420	24	15	359
REG	311	115	943	182	59	640
ROB	66	13	595	28	5	167
ssMRCD	64	26	431	48	21	379
# samples	2108	870	14194	-	-	-

Table 3.3.1: Number of flagged outliers for each method and data set, unfiltered and filtered by high values in clr values and non-transformed measurements in at least one element.

3.4 Summary and Discussion

In this paper we demonstrated the suitability of local outlier detection methods for the purpose of mineral exploration in geochemistry. Generally, local outlier detection incorporates the spatial neighborhood of the samples in order to identify local anomalies in the multivariate element composition. The analyzed data sets are of different scale, sample density and data quality, and they also vary in the number of available element concentrations. However, the geochemical data sets have in common that they are

	Unfiltered			*Filtered*		
Method	Regional	Targeting	Both	Regional	Targeting	Both
LOF	12	18	5	11	17	4
REG	39	27	24	31	22	15
ROB	2	18	1	1	9	1
ssMRCD	17	19	7	15	18	6

Table 3.3.2: Number of found deposits for each method and data set, unfiltered and filtered by high values in clr values and non-transformed measurements in at least one element. Maximum number of deposits possible to find is 48.

of compositional nature, which made it necessary to process them with tools from compositional data analysis.

The different methods for multivariate local outlier detection mainly vary in the way how they estimate the covariance matrix to compute pairwise Mahalanobis distances. The simplest approach is to use a joint global covariance matrix. The other extreme is to use separate covariance estimates for each local neighborhood. A third, recently proposed methods tries to find a compromise between those two extremes, with the idea that the robust covariance estimation should change smoothly across the neighborhoods. These methods are compared to a procedure called LOF (Local Outlier Factor), which incorporates Euclidean distances between the multivariate observations, and thus is based on a very different concept.

While all methods find mineralisations, we have shown that they also have their limitations, ranging from biased covariance estimation to an extensive flagging of outliers and not finding reasonable spatial outliers. With known mineral deposits it is possible to evaluate the methodologies on real data and analyze their performance in more detail. However, the considered mineral deposits are of very different type, and one might have to go into much more detail to see if the compositions of the identified outliers really reflect the type of mineralisation, or if the elements used in the analysis are even appropriate for this purpose. Moreover, it can also happen that some of the identified outliers point at new yet unknown mineralisations, which makes the evaluation used in this paper biased.

Thus, next to appropriate outlier detection methods, it is also important to use diagnostic tools to verify if the indicated outliers indeed point at mineralisations. We introduced exploratory procedures that combine relative and absolute information, as outliers are supposed to be atypical in the multivariate compositional data space, but at the same time they are supposed to have high concentration values for particular elements.

Next to a data subset from the GEMAS project we evaluated the procedures for two data sets from the same area in Finland, measured in different years, with a very different sampling density, and yielding different sets of elements with different data quality. The main question was if higher sampling density would also lead to higher accuracy for mineral identification. However, the crucial point for mineral identification seems that not only the commodity elements need to be available, but

also complementing elements that allow to understand and characterize the geological situation.

3.5 Conclusions

A general but possibly obvious conclusion is that also for local outlier detection, data quality is more important than quantity. However, it is not just quality which matters, it is also the set of elements which needs to be big enough in order to cover the complexity of the geochemistry that experts would expect to find at mineralised zones. Here, rare elements such as gold could be very valuable, provided that they are measured with sufficient quality. Elements measured with low quality, as for example with a high proportion of values below the detection limit, will negatively affect the log-ratio transformations used in compositional data analysis. In more detail, an observation where just one element has a value below the detection limit could end up in a multivariate observation of the compositionally transformed data set with all entries being distorted. This could lead to a very high proportion of outliers, where local outlier detection methods could fail to work correctly.

For the tested local outlier detection methods it is known that some are very sensitive and may lead to a too strict rule for indicating outliers. Also the way how the methods work internally is very different, and therefore these methods are flagging different sets of outliers. From a theoretical point of view, the ssMRCD method will be preferable over the methods REG and ROB in case where the investigated area shows geochemical differences, e.g. as a result of different underlying processes (pollution sources, soil formation, environmental conditions, etc.). The LOF method tends to identify data points that are isolated in the multivariate space. Thus, if the sampling is dense and the observations continuously change towards the mineralisation, this method may fail to see samples on top of mineralised zones as outliers. Nevertheless, a strategy could be to use multiple local outlier detection methods to balance their advantages and limitations.

For sampling strategies it follows that a lower density with more analyzed elements is desirable to high density sampling with low data quality. When interesting locations are found with sparse data, the density can then still be increased in further studies adjusted to the specific ore type and deposit size to also find smaller targets (for example, vein type or small sub-outcropping deposits). Nevertheless, statistical analysis alone is limited and always needs cooperation with experts providing interpretation of outliers and classifying them as potential mineral deposits worth to be analyzed further.

Appendix B

B.1 Q-Q Plots

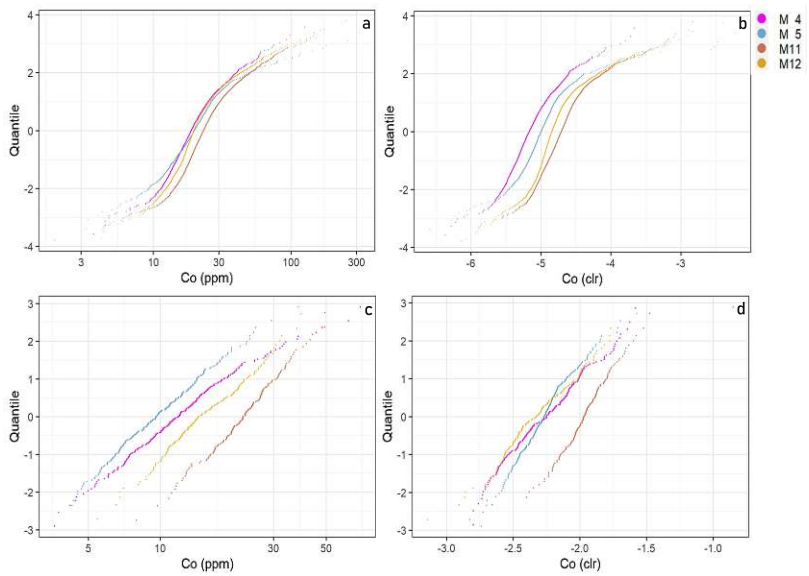


Figure B.1: Q-Q plots of Co: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

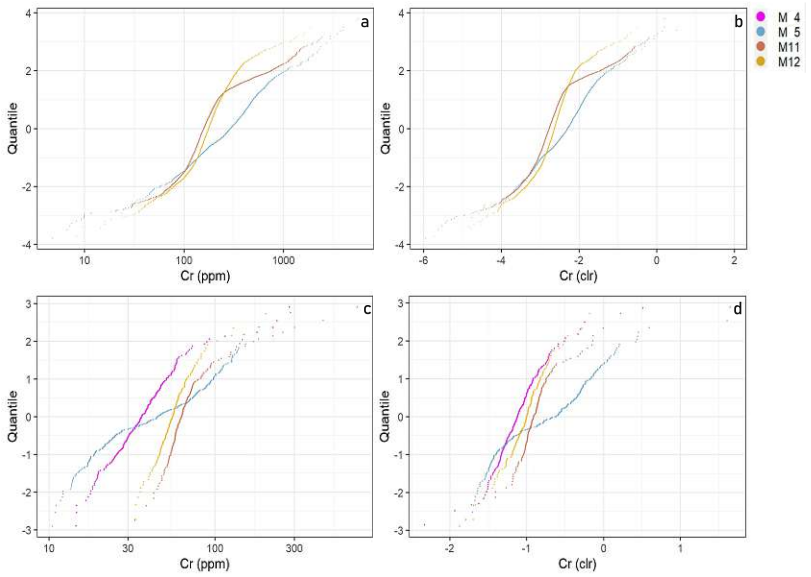


Figure B.2: Q-Q plots of Cr: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

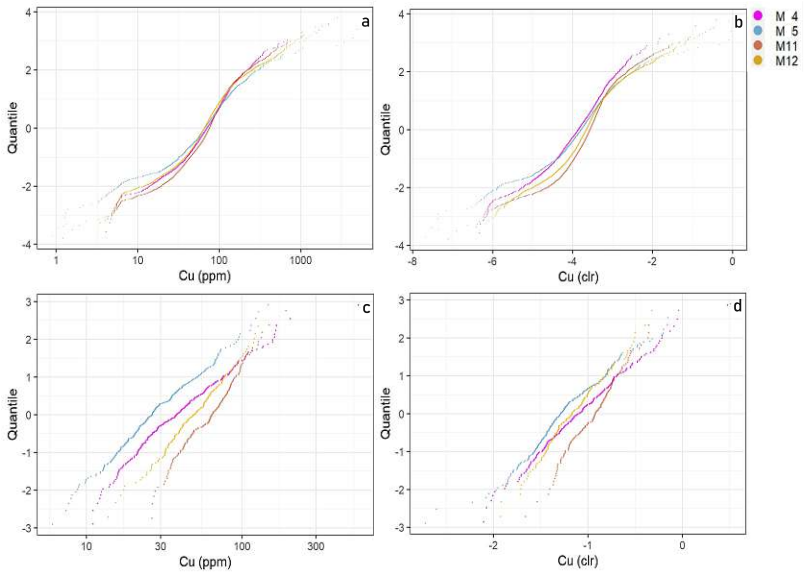


Figure B.3: Q-Q plots of Cu: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

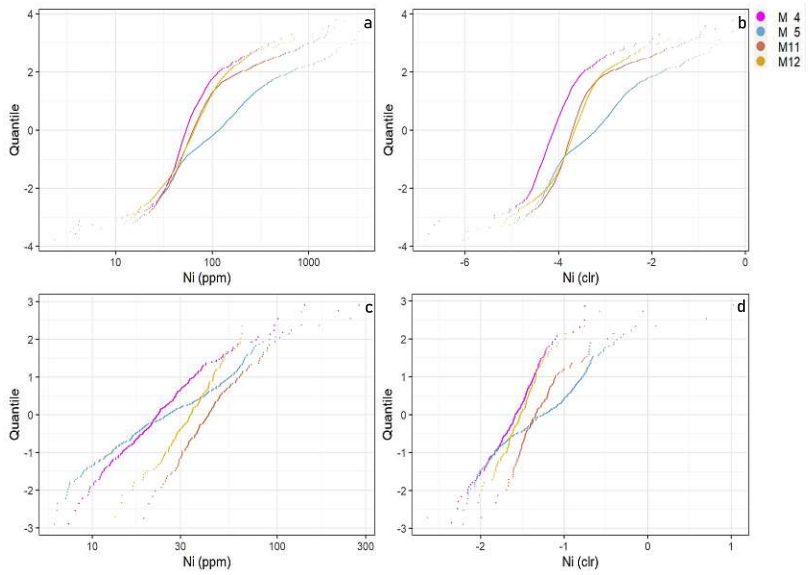


Figure B.4: Q-Q plots of Ni: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

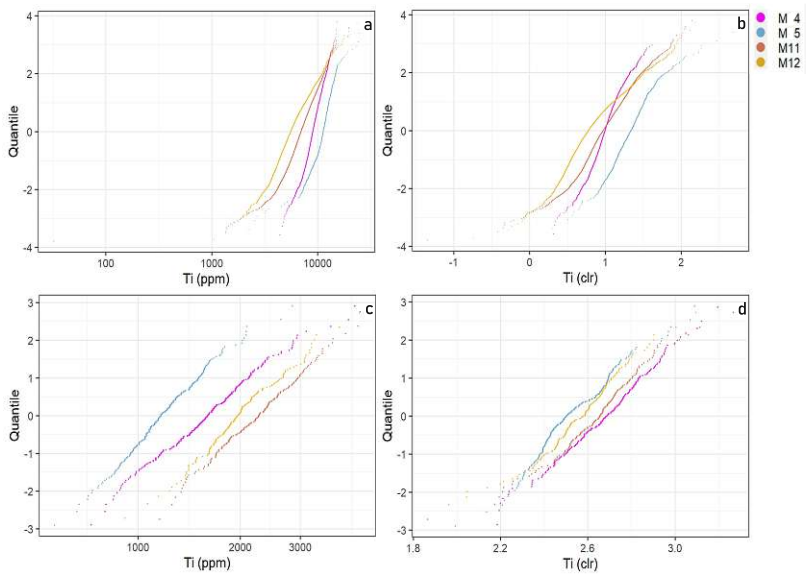


Figure B.5: Q-Q plots of Ti: (a) original concentration in targeting till, (b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

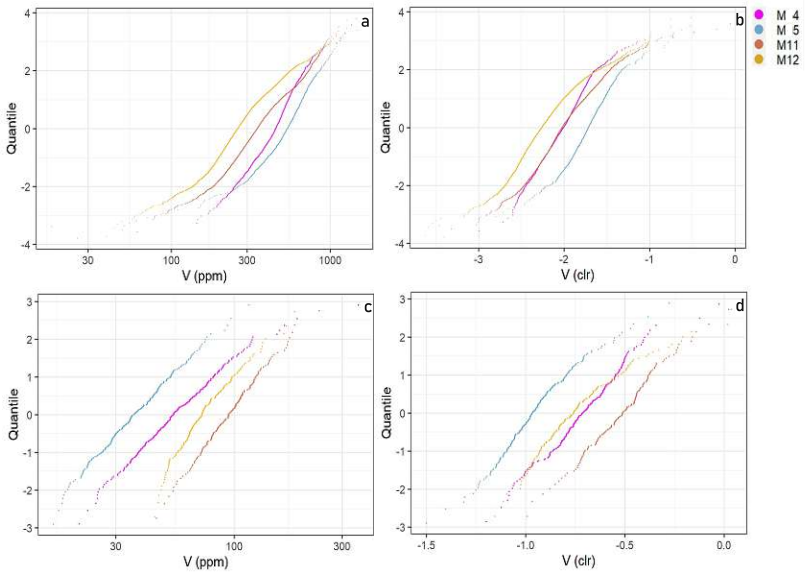


Figure B.6: Q-Q plots of V: (a) original concentration in targeting till, b) clr transformed concentration in targeting till, (c) original concentration in regional till, (d) clr transformed concentration in regional till.

4 Sparse Outlier-Robust PCA for Multi-Source Data

This chapter was published as Puchhammer, P., Wilms, I., and Filzmoser, P. (2024b). Sparse outlier-robust PCA for multi-source data. *arXiv preprint arXiv:2407.16299*.

4.1 Introduction

Principal component analysis (PCA) is undoubtedly one of the most important unsupervised statistical methods available. The basic idea is to project the observations in a given data set onto a new vector space with orthonormal basis where each basis vector is a linear combination of the original variables constructed to capture the highest variability for the first basis vector, the second highest variability for the second basis vector and so on. The new variables are called *principal components* (PC), the coordinates of the PCs in the original variable space are called *loadings* and the coordinates of the observations with respect to the PCs are called *scores*. Often, only the first few PCs that catch a majority of the variance and thus of the available information are analyzed. As such, PCA finds widespread application across numerous areas, such as dimensionality reduction, visualization, clustering, feature engineering and many more.

Standard PCA—PCA based on the sample covariance—has, however, three important shortcomings when it comes to analyzing modern data sets. First, modern data sets often consist of many variables. Then sensible, efficient and correct interpretation of scores and loadings can get difficult since the loadings obtained via standard PCA are often a combination of all variables involved. Moreover, by focusing the interpretation on large absolute loading entries and ignoring small ones, whether intentionally or not, misleading interpretation results can be produced as discussed in Cadima and Jolliffe (1995). Therefore, inducing *sparsity* in the loading entries is necessary to ensure correct interpretation of PCA results. Sparse PCA (see Section 4.1.1) has become fundamentally important in a variety of applications.

Secondly, standard PCA is often applied to a single data set, yet many modern applications entail multiple related data sets from different sources for which PCA needs to be performed jointly. Classical examples of such multi-source data are time series data that can be grouped based on time increments like months or years, spatial data with groups based on spatial proximity or nationality, or more general subgroups based on e.g., demographics, socioeconomic status or other external variables. Even though PCA can still be applied globally on the whole data and structural changes might still be identified in the scores, the question of which variables drive the variance

in different groups or data sets remains unanswered as such. On the other hand, fully localizing PCA by applying it on each subgroup individually ignores the overall inert link between the subgroups, rendering the individualistic approach inappropriate. Thus, the local and global aspect of the data needs to be addressed simultaneously. Moreover, in the multi-source PCA setting with N sources, sparsity and especially structured sparsity patterns are well suited. By analyzing multiple related data sets, we end up with N-times more loading entries than for a global PCA approach. Thus, sparsity in each entry is important. However, due to the interconnection of the data sets and, thus of the loadings, structured sparsity, here meaning sparsity in entries of the same variable for all sources simultaneously, can be present in the data sets as well. Including a structured sparsity inducing combination of groupwise and elementwise penalties can then increase accuracy in PCA or also regression results as demonstrated by Jenatton et al. (2010) and Simon et al. (2013), respectively, for groupings of variables in a global context. Although other disciplines have already explored multi-source data successfully, multi-source PCA analysis remains under-explored (see Section 4.1.1).

Third, standard PCA is not robust to outliers (aka anomalies). Yet, variability analysis in modern multi-source data sets also requires that outliers are taken care of reliably. By definition, outliers do not behave like the majority of the data and lie outside of the multivariate point cloud of regular observations. Thus, outliers inherently increase variability measured by classical estimators and distort the direction of high variability towards them. Since we are interested in the direction of variability of the data majority, robustness in estimators for variability, i.e. covariance matrices, must be used. Well-known robust covariance estimators are for example the *minimum covariance determinant estimator* (MCD, Rousseeuw, 1985; Rousseeuw and Driessen, 1999) or its regularized variant, the *minimum regularized covariance determinant estimator* (MRCD, Boudt et al., 2020) that can be used to robustify PCA by a so-called *plug-in* approach (Croux and Haesbroeck, 2000) that we also adopt in this paper. see Section 4.1.1 for other approaches.

In this paper, we offer the first multi-source PCA approach that delivers sparse loadings and is robust to outliers. A key ingredient of our method is the *spatially smoothed MRCD* (ssMRCD) estimator (Puchhammer and Filzmoser, 2024, 2023), an outlier-robust covariance estimator that jointly estimates covariance matrices for multiple, related data sets by inducing smoothing. We tailor the ssMRCD estimator towards our multi-source PCA set-up and then adopt the popular plug-in approach in robustness, using the tailored ssMRCD as plug-in, to perform sparse, outlier-robust PCA for multiple, related data sets. We employ standard sparsity as well as structured sparsity penalties to mirror the relations between the multiple sources (also called *neighborhoods* in the ssMRCD context given the focus on spatial settings, yet the ssMRCD estimator is generally suited for analyzing multi-source data). By jointly analyzing the covariance matrices—via the ssMRCD—and sparsity in the loadings—via the structured sparsity penalties—we can better differentiate between global structures indicated by similarities between sources, and local structures indicated by differences in our variability analysis. Apart from this main methodological contribution, we also offer an important computational contribution by designing an alternating direction method of multipliers algorithm and by carefully fine tuning it to solve the multi-source

PCA problem in a computationally efficient manner. We offer computing code in the form of a publicly available R-package `ssMRCD` (Puchhammer and Filzmoser, 2023) to facilitate adoptability and practical use.

4.1.1 Related Work

Many proposals in the literature exist that focus either on sparse PCA, multi-source PCA or outlier-robust PCA; we review these strands below. Yet, few studies address two out of these three aspects simultaneously, and no PCA method exists, to the best of our knowledge, that delivers all three features jointly; a gap that this paper fills.

Sparse PCA

The literature on sparse PCA is rich. We provide a compact, yet incomplete, overview here but refer the interested reader to Bertsimas et al. (2022) for a recent review on the different literature strands. Starting with the work of Jolliffe et al. (2003), who introduce the *least absolute shrinkage and selection operator* (LASSO) into PCA with the algorithm known as SCoTLASS, Zou et al. (2006) include an elastic-net penalty to PCA reformulated as regression problem. Further developments include the work of Shen and Huang (2008) approaching the problem from a regularized singular value decomposition, Ma (2013) focusing on a thresholding approach for high-dimensional data, d'Aspremont et al. (2008) deriving a greedy algorithm based on a semi-definite relaxation variation, and Journée et al. (2010) with a convex reformulation of sparse PCA. Recently, Bertsimas and Kitane (2023) proposed GeoSPCA, a sparse PCA approach that builds on a geometrical interpretation of the problem.

Multi-Source PCA

While the need for multi-source data analysis has already been addressed in a variety of other disciplines (e.g., Price and Sherwood, 2018; Wang et al., 2013 for regression settings, Puchhammer and Filzmoser, 2024 for covariance estimation, Danaher et al., 2014; Price et al., 2021 for inverse covariance estimation, or Barbaglia et al., 2016; Wilms et al., 2018 for time series data), multi-source PCA analysis is still under-explored. A recent exception is Shi and Kontar (2024). They propose *personalized PCA* (PerPCA), a systematic approach to analyze data collected from different sources with heterogeneous trend thereby decoupling shared (global) and unique (local) features. A similar but different approach to analyzing connected data sets is multi-block PCA, where features are grouped together instead of observations. Recent advances of this field is *integrated principal components analysis* (iPCA) by Tang and Allen (2021), who propose a model-based framework of the classical PCA problem suited for analyzing multiple data sources with features of different types that are measured on the same set of samples, and offer sparse as well as non-sparse iPCA estimators. Other PCA-related methods for multi-block data are *joint and individual variance explained* (JIVE; Lock et al., 2013) and *common and individual feature extraction* (CIFE; Zhou et al., 2016), both focusing on low-rank approximations.

Outlier-Robust PCA

Early work by Croux and Haesbroeck (2000) considers robust PCA using robust covariance estimators as plug-ins instead of the regular sample covariance matrix that is non-robust. Leyder et al. (2024) recently proposed *Generalized spherical PCA*, a robust PCA method based on the generalized spatial sign covariance matrix. Other robust PCA approaches are based on projection pursuit (e.g., Li and Chen, 1985; Hubert et al., 2002; Croux and Ruiz-Gazen, 2005) or a combination of both called ROBPCA (Hubert et al., 2005). Hubert et al. (2016) further extend ROBPCA to sparse PCA (ROSPCA) while Croux et al. (2013) develop a robust PCA method with standard sparsity based on projection pursuit. A well-known PCA approach by Candès et al. (2011) delivers robust PCA results under additional weak assumptions and is of special interest among others in the fields of video processing and face recognition. More recently, Yi et al. (2017) offer *joint sparse principal component analysis* (JSPCA) that simultaneously selects useful features and provides protection against outliers whereas Wang and Van Aelst (2020) develop a sparse PCA based on least trimmed squares (LTS-SPCA) which is then also compared to ROBPCA and ROSPCA in simulations. For robustness in the case where entries of data set columns have been corrupted by permutations Yao et al. (2024) propose data analysis via *unlabeled principal component analysis* (UPCA). Some robust PCA methods tailored towards high-dimensional data include Schmitt and Vakili (2016), whose *FastHCS* algorithm selects a subset of observations, as well as the work of Fayomi et al. (2024), where PCA is based on a Cauchy likelihood reformulation.

4.1.2 Outline

The remainder of the paper is structured as follows. In Section 4.2, we derive the objective function to perform sparse, outlier-robust PCA for multiple related data sets. In Section 4.3, we present and carefully discuss the computationally-efficient algorithm to perform multi-source PCA based on the alternating direction method of multipliers (ADMM). In Section 4.4 our proposal is tested on simulated data and compared to state-of-the-art PCA alternatives. Two real data examples are analyzed in Section 4.5 and finally, conclusions are given in Section 4.6.

4.2 Multi-Source PCA Based on the ssMRCD

In Section 4.2.1 we introduce the optimization problem to perform sparse PCA for multiple related data sets thereby focusing on the first PC. We then expand the problem to multiple PCs in Section 4.2.2. Finally, in Section 4.2.3 we detail the ssMRCD estimator, a key ingredient in our plug-in approach to achieve robustness on multi-source data.

4.2.1 First Principal Component

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$ be data sets from N sources consisting of $\boldsymbol{X}_i = (\boldsymbol{x}'_{i,1}, \ldots, \boldsymbol{x}'_{i,n_i})' \in \mathbb{R}^{n_i \times p}$ observations per source $i = 1, \ldots, N$ of the same p variables. For each source i,

corresponding locally estimated covariance matrices are denoted by $\hat{\boldsymbol{\Sigma}}_i \in \mathbb{R}^{p\times p}$ and estimated means by $\hat{\boldsymbol{\mu}}_i \in \mathbb{R}^p$.

When considering the PC loadings for multi-source data each loading can be written as a matrix, where each column represents a source, and each row one variable. The loadings matrix of the k-th principal component is denoted as

$$\boldsymbol{V}^k = \begin{pmatrix} v_{11}^k & \dots & v_{1N}^k \\ \vdots & & \vdots \\ v_{p1}^k & \dots & v_{pN}^k \end{pmatrix} = (\boldsymbol{v}_{\cdot 1}^k, \dots, \boldsymbol{v}_{\cdot N}^k) = (\boldsymbol{v}_{1\cdot}^{k\prime}, \dots, \boldsymbol{v}_{p\cdot}^{k\prime})' \,. \tag{4.2.1}$$

The loadings matrix of the first PC is obtained by solving the following optimization problem

$$\boldsymbol{V}^1 = \underset{\substack{\boldsymbol{V}\in\mathbb{R}^{p\times N} \\ ||\boldsymbol{v}_{\cdot i}||_2=1, i=1,\dots,N}}{\operatorname{argmin}} - \sum_{i=1}^{N} \boldsymbol{v}_{\cdot i}' \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v}_{\cdot i} + \eta\gamma \sum_{j=1}^{p} \underbrace{\sum_{i=1}^{N} |v_{ji}|}_{=||\boldsymbol{v}_{j\cdot}||_1} + \eta(1-\gamma)\sqrt{N} \sum_{j=1}^{p} ||\boldsymbol{v}_{j\cdot}||_2, \tag{4.2.2}$$

where $\eta \geq 0$ regularizes the overall degree of sparsity, and $\gamma \in [0,1]$ distributes the sparsity between local ($\gamma = 1$) and global ($\gamma = 0$) sparsity patterns. The groupwise penalty induces global sparsity structures—that is sparsity for all loadings of a given variable across all sources—and is equivalent to the groupwise penalty used in Simon et al. (2013). As in Simon et al. (2013), the term $\sqrt{N}$ balances the size of the two penalties since the minimal penalty for the L_1-norm under the given constraints is N, whereas the minimal groupwise penalty is $\sqrt{N}$. Thus, we can compare the effect of increasing η among different levels of γ more easily.

4.2.2 Multiple Principal Components

The loadings for the k-th principal component $\boldsymbol{V}^k$ are the solutions to the optimization problem Equation (4.2.2) with an additional constraint to account for orthogonality per source,

$$\boldsymbol{v}_{\cdot i}^k \perp \boldsymbol{v}_{\cdot i}^l \quad l = 1, \dots, k-1, i = 1, \dots, N.$$

The orthogonality constraints between the loadings per source constitute non-standard optimization constraints, especially in the context of (sparse) PCA. Since the groupwise sparsity induces a non-separable objective function and existing solutions rely on standard orthogonality constraints, they cannot be applied, and new solutions are needed.

To facilitate notation and optimization, we rewrite the problem into stacked-column notation. The matrix $\boldsymbol{V}^k$ can be stacked into one collective vector $\boldsymbol{v}^k$ and the covariance matrices into a block-diagonal (positive semi-definite) matrix,

$$\boldsymbol{v}^k = \begin{pmatrix} \boldsymbol{v}_{\cdot 1}^k \\ \vdots \\ \boldsymbol{v}_{\cdot N}^k \end{pmatrix}, \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_1 & & \\ & \ddots & \\ & & \hat{\boldsymbol{\Sigma}}_N \end{pmatrix}.$$

Then, the objective function for the loadings of the k-th set of PCs can be rewritten in a form known from standard PCA with adapted penalty terms and linear and quadratic equality and inequality constraints,

$$\begin{aligned} \boldsymbol{v}^k = \underset{\boldsymbol{v}\in\mathbb{R}^{pN}}{\operatorname{argmin}} \quad & -\boldsymbol{v}'\hat{\boldsymbol{\Sigma}}\boldsymbol{v} + \eta\gamma||\boldsymbol{v}||_1 + \eta(1-\gamma)\sqrt{N}\sum_{j=1}^{p}\sqrt{\boldsymbol{v}'\boldsymbol{C}_j\boldsymbol{v}} \\ \text{s.t.} \quad & \boldsymbol{v}'\boldsymbol{B}_i\boldsymbol{v} = 1 \quad \forall i = 1,\ldots,N \\ & \boldsymbol{v}'\boldsymbol{B}_i\boldsymbol{v}^l = 0 \quad \forall l = 1,\ldots,k-1,\ i = 1,\ldots,N. \end{aligned} \tag{4.2.3}$$

The $pN \times pN$ matrices $\boldsymbol{C}_j$ and $\boldsymbol{B}_i$, extract the j-th row (variable) and i-th column (source) of $\boldsymbol{V}^k$ from the stacked column vector $\boldsymbol{v}^k$, respectively, and are defined as

$$(\boldsymbol{C}_j)_{ik} = \begin{cases} 1, & \text{if } i = k = pl + j, \text{ where } l = 0,\ldots,N-1, \\ 0, & \text{otherwise.} \end{cases}$$

$$(\boldsymbol{B}_i)_{jk} = \begin{cases} 1, & \text{if } j = k = p(i-1) + l, \text{ where } l = 1,\ldots,p, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, $\boldsymbol{C}_j = \boldsymbol{C}_j' = \boldsymbol{C}_j'\boldsymbol{C}_j$ and $\boldsymbol{C}_1 + \ldots + \boldsymbol{C}_p = \boldsymbol{I}_{pN}$ for $j = 1,\ldots,p$ and $\boldsymbol{B}_i = \boldsymbol{B}_i' = \boldsymbol{B}_i'\boldsymbol{B}_i$ and $\boldsymbol{B}_1 + \ldots + \boldsymbol{B}_p = \boldsymbol{I}_{pN}$ for $i = 1,\ldots,N$.

Once the loadings $\boldsymbol{v}^1,\ldots,\boldsymbol{v}^k$ are obtained from the data, the scores of each locally centered observation $\boldsymbol{x}_{i,\iota} - \hat{\boldsymbol{\mu}}_i$ for $\iota = 1,\ldots,n_i$ of source i are calculated by

$$\boldsymbol{t}_{i,\iota} = (\boldsymbol{x}_{i,\iota} - \hat{\boldsymbol{\mu}}_i)\left(\boldsymbol{v}_{\cdot i}^1,\ldots,\boldsymbol{v}_{\cdot i}^k\right) \tag{4.2.4}$$

and collected in $\boldsymbol{T}_{\boldsymbol{i}} = \left(\boldsymbol{t}_{i,1}',\ldots,\boldsymbol{t}_{i,n_i}'\right)' \in \mathbb{R}^{n_i\times k}$.

4.2.3 Outlier-Robustness via ssMRCD Plug-In

Optimization of problem (4.2.3) for the loadings requires plug-in estimators for the covariance matrix of each source, whereas computation of the scores in Equation (4.2.4) additionally requires mean estimators. Standard choices to this end would be the sample covariance matrices and sample means computed for each source separately. Such estimators face, however, two problems. First, they are not robust to outliers. One may resort to traditional robust estimators such as the MCD or median for each source separately to circumvent this problem. Since only robustly estimated covariances and means are used further, no additional robustification steps are necessary. Second, these classical (non-robust or robust) estimators still treat each source in isolation thereby ignoring potential connections and interactions between them. Therefore, it is crucial to incorporate both local and global information to leverage available information across multiple sources more extensively, enhancing the accuracy and reliability of the resulting covariance and mean estimators.

An outlier-robust covariance and mean estimator tailored towards this global-local scenario is the ssMRCD estimator (Puchhammer and Filzmoser, 2024) that is implemented in the `R`-package `ssMRCD` (Puchhammer and Filzmoser, 2023). The ssMRCD

estimator has originally been developed for local outlier detection in spatial data. However, due to its general requirements, it can be extended to general data sets within a multi-source setting, which we do in this paper. To make the paper self-contained, we fully detail the ssMRCD estimator tailored towards the multi-source PCA problem set-up, including the selection of its hyperparameters in Section (4.3.2).

Starting with a partition of the data into multiple sources, the ssMRCD estimator selects a subset H_i of size $|H_i| = h_i$ consisting of an $\alpha \in [0.5, 1]$ percentage of observations of $\boldsymbol{X}_i$ by minimizing the objective function over all H-subset combinations $\mathcal{H} = (H_i)_{i=1,\dots,N}$

$$\mathcal{H}^* = \underset{\mathcal{H}=(H_i)_{i=1,\dots,N}}{\operatorname{argmin}} \sum_{i=1}^{N} \det \left((1-\lambda)\boldsymbol{K}_i(\mathcal{H}) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}) \right),$$

similar to the MCD (Rousseeuw, 1985; Rousseeuw and Driessen, 1999), or the MRCD (Boudt et al., 2020) estimator, thus choosing subsets with least-outlying observations. The weight matrix $\boldsymbol{W}$, with entries $\omega_{ij}, i, j = 1, \dots, N$, provides a measure of similarity between data sources which is used to leverage global information more targeted. For example, for spatial or time series data, the weights could be based on inverse distances, or for groupings based on known properties, the similarity between these properties' levels might be an appropriate choice for $\boldsymbol{W}$ (see also Section 4.5). The matrices $\boldsymbol{K}_i(\mathcal{H})$ are constructed in an MRCD manner, regularizing the sample covariance matrix of the H-subsets of source i $\operatorname{Cov}(\boldsymbol{X}_{H_i})$ with a global target matrix $\boldsymbol{R}$ and a factor ζ_i,

$$\boldsymbol{K}_i(\mathcal{H}) = \zeta_i \boldsymbol{R} + (1-\zeta_i) c_\alpha \operatorname{Cov}(\boldsymbol{X}_{H_i}),$$

making the estimator suitable also for high-dimensional data. The target matrix $\boldsymbol{R}$ can be any robustly estimated regular covariance matrix, and ζ_i is set to ensure a low condition number for starting values (see Boudt et al., 2020). The factor c_α is required for consistency and described in more detail in Croux and Haesbroeck (1999). Finally, the ssMRCD covariance estimators are defined as

$$\hat{\boldsymbol{\Sigma}}_i = (1-\lambda)\boldsymbol{K}_i(\mathcal{H}^*) + \lambda \sum_{j=1, j\neq i}^{N} \omega_{ij} \boldsymbol{K}_j(\mathcal{H}^*),$$

and the mean estimators $\hat{\boldsymbol{\mu}}_i$ as the sample mean of the selected observations $\boldsymbol{X}_{H_i^*}$ per source. The most prominent parameter for the ssMRCD estimator is $\lambda \in [0, 1]$ which defines the amount of smoothing between the covariances of sources weighted with $\boldsymbol{W}$. The parameter λ thus describes how much of the global data is exploited compared to the local source-specific data; the closer its value to one, the more the local data sources are exploited.

4.3 Algorithm

We propose a computationally efficient algorithm tailored towards solving optimization problem (4.2.3). Since the optimization of (4.2.3) is difficult due to its non-differentiable

norm penalties and overall non-convexity, we develop an alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011) specifically fine-tuned to solving it. The proposed ADMM algorithm is based on solving the following equivalent representation of problem (4.2.3), namely

$$\min_{\substack{\boldsymbol{v}_{(1)},\boldsymbol{v}_{(2)},\\ \boldsymbol{v}_{(3)},\boldsymbol{v}_{(0)}}} \quad \underbrace{-\boldsymbol{v}_{(1)}'\hat{\boldsymbol{\Sigma}}\boldsymbol{v}_{(1)} + I_\infty\{\boldsymbol{v}_{(1)}'\boldsymbol{B}_i\boldsymbol{v}_{(1)} = 1,\ \boldsymbol{v}_{(1)}'\boldsymbol{B}_i\boldsymbol{v}^l = 0\ \forall 1 \le i \le N, 1 \le l < k\}}_{f_1(\boldsymbol{v}_{(1)})}$$

$$+ \underbrace{\eta\gamma||\boldsymbol{v}_{(2)}||_1}_{f_2(\boldsymbol{v}_{(2)})} + \underbrace{\eta(1-\gamma)\sqrt{N}\sum_{j=1}^{p}\sqrt{\boldsymbol{v}_{(3)}'\boldsymbol{C}_j\boldsymbol{v}_{(3)}}}_{f_3(\boldsymbol{v}_{(3)})} \tag{4.3.1}$$

$$\text{s.t.} \quad \boldsymbol{v}_{(i)} - \boldsymbol{v}_{(0)} = 0, \quad i = 1,2,3, \tag{4.3.2}$$

where $I_\infty\{\cdot\}$ denotes the indicator function with an infinite amount of weight if the condition inside the brackets is not fulfilled.The introduction in problem (4.3.1) of the helper variables $\boldsymbol{v}_{(1)}, \boldsymbol{v}_{(2)}$ and $\boldsymbol{v}_{(3)}$, coupled together to $\boldsymbol{v}_{(0)}$ via the constraints in Equation (4.3.2), allows us to efficiently decouple the optimization problem into corresponding subproblems. The ADMM then solves these subproblems iteratively until convergence; the updates for the m-th iteration step are

$$\begin{aligned}
\boldsymbol{v}_{(i)}^{m+1} &= \arg\min_{\boldsymbol{v}_{(i)}}(f_i(\boldsymbol{v}_{(i)}) + \frac{\rho}{2}||\frac{1}{\rho}\boldsymbol{u}_{(i)}^m + \boldsymbol{v}_{(i)} - \boldsymbol{v}_{(0)}^m||_2^2 \\
\boldsymbol{v}_{(0)}^{m+1} &= \frac{1}{3}\sum_{i=1}^{3}\left(\boldsymbol{v}_{(i)}^{m+1} + \frac{1}{\rho}\boldsymbol{u}_{(i)}^m\right), \\
\boldsymbol{u}_{(i)}^{m+1} &= \boldsymbol{u}_{(i)}^m + \rho\left(\boldsymbol{v}_{(i)}^{m+1} - \boldsymbol{v}_{(0)}^{m+1}\right),
\end{aligned} \tag{4.3.3}$$

with penalty parameter $\rho > 0$ enforcing consensus between the helper variables. The solutions of the three new optimization problems (4.3.3) are detailed in Appendix C.1. In the following, we discuss convergence in Section 4.3.1, and provide guidance to select the hyperparameters in Section 4.3.2.

4.3.1 Convergence

A globally optimal solution to problem (4.2.3) exists since we have a compact variable space and a continuous objective function. Convergence of the iterative ADMM is therefore based on monitoring the primal and dual residuals at each iteration m,

$$\begin{aligned}
r^m &= ||\boldsymbol{v}_{(1)}^m - \boldsymbol{v}_{(0)}^m||_2^2 + ||\boldsymbol{v}_{(2)}^m - \boldsymbol{v}_{(0)}^m||_2^2 + ||\boldsymbol{v}_{(3)}^m - \boldsymbol{v}_{(0)}^m||_2^2 \\
s^m &= 3\rho^2||\boldsymbol{v}_{(0)}^m - \boldsymbol{v}_{(0)}^{m-1}||_2^2.
\end{aligned}$$

The primal residual r^m measures the coherence between the optimizers of the three subproblems in the sense of constraint (4.3.2), whereas the dual residual s^m gives the overall change compared to the previous iteration. The tolerances for the two

convergence criteria are then

$$\epsilon^m_{prime} = \sqrt{Np}\,\epsilon_{ADMM} + \epsilon_{ADMM}\max\{||\boldsymbol{v}^m_{(1)}||_2, ||\boldsymbol{v}^m_{(2)}||_2, ||\boldsymbol{v}^m_{(3)}||_2, ||\boldsymbol{v}^m_{(0)}||_2\} \quad (4.3.4)$$
$$\epsilon^m_{dual} = \sqrt{Np}\,\epsilon_{ADMM} + \epsilon_{ADMM}\max\{||\boldsymbol{u}^m_{(1)}||_2, ||\boldsymbol{u}^m_{(2)}||_2, ||\boldsymbol{u}^m_{(3)}||_2\},$$

for a given tolerance ϵ_{ADMM}, implying the same absolute and relative tolerance. The algorithm stops when $r^m < \epsilon^m_{prime}$ and $s^m < \epsilon^m_{dual}$. We set to $\epsilon_{ADMM} = 10^{-4}$ in all simulations and real data examples.

To avoid convergence problems due to the sign ambiguity of the PCA vectors (since two vectors that are directly opposed to each other result in the same explained variance and sparsity pattern), we impose an additional constraint to the optimality problem in Equations (4.2.3). Taking a fixed vector $\boldsymbol{z} \in \mathbb{R}^{pN}$, we enforce the solution for the k-th vectorized loadings matrix to have a non-negative scalar product with $\boldsymbol{z}$ for each source,

$$\boldsymbol{z}'\boldsymbol{B}_i\boldsymbol{v}^k \geq 0, \quad \forall i = 1, \ldots, N, \quad (4.3.5)$$

and add the term $I_\infty\{\boldsymbol{z}'\boldsymbol{B}_i\boldsymbol{v}_{(1)} \geq 0,\ \forall 1 \leq i \leq N\}$ to the objective function $f_1(\boldsymbol{v}_{(1)})$. Unless $\boldsymbol{z}$ is exactly orthogonal to the real solution in at least one source, we consistently choose the solution for $\boldsymbol{v}_{(1)}$ in the same direction. The best choice for $\boldsymbol{z}$ would be the true solution. Since the true solution is, however, unknown, the next best possibility is the well-chosen starting value that we derive in Section 4.3.1. Section 4.3.1 further zooms into the choice of the penalty parameter ρ and Section 4.3.1 ends with further algorithmic enhancements.

Starting Value

The starting value plays a crucial role in convergence, especially due to the non-convexity of problem (4.2.3). Also, the projection approach with $\boldsymbol{z}$ being the chosen starting value is more stable for a good choice of the starting value since the orthogonality issues described just above are less likely to occur.

To find a good starting value, a compromise between the two extreme cases of sparsity is needed. When considering problem (4.2.3) with $\eta = 0$ (no sparsity), it boils down to N separate standard PCA problems, one for each covariance $\hat{\boldsymbol{\Sigma}}_i$ $(i = 1, \ldots, N)$. The solutions for the k-th PC are the k-th eigenvectors $\boldsymbol{y}_k(\hat{\boldsymbol{\Sigma}}_i)$ of the covariances calculated per source i,

$$\boldsymbol{y}^0_k = (\boldsymbol{y}_k(\hat{\boldsymbol{\Sigma}}_1)', \ldots, \boldsymbol{y}_k(\hat{\boldsymbol{\Sigma}}_N)')'.$$

The extreme solution for $\eta \to \infty$ depends on γ, as denoted by $\boldsymbol{y}^\infty_k(\gamma)$, and can be calculated based on the results of Proposition 4.3.1.1.[1] The proof of the proposition is given in Appendix C.2.

Proposition 4.3.1.1. *Using the notation of Equation* (4.2.2)*, define*

$$G_1(\boldsymbol{v}) = \sum_{j=1}^{p}\sum_{i=1}^{N}|v_{ji}| = \sum_{i=1}^{N}||\boldsymbol{v}_{\cdot i}||_1, \qquad G_2(\boldsymbol{v}) = \sum_{j=1}^{p}||\boldsymbol{v}_{j\cdot}||_2.$$

[1] Note, that the indices in Proposition 4.3.1.1 are not necessarily unique, such as for correlation matrices, which are addressed in Appendix C.2.

a. *For each source* $i = 1, \ldots, N$ *and given the normality constraint* $||\boldsymbol{v}_{\cdot i}||_2 = 1$, *the minimal value of* $||\boldsymbol{v}_{\cdot i}||_1$ *is attained, if there exists a variable* $j'(i)$ *such that* $|v_{j'(i)i}| = 1$ *and* $v_{ji} = 0$ *for all* $j \neq j'(i)$.

b. *Given the normality constraints* $||\boldsymbol{v}_{\cdot i}||_2 = 1, i = 1, \ldots, N$, *the minimal value of* $G_2(\boldsymbol{v})$ *is attained if there exists a variable* j' *for all sources* $i = 1, \ldots, N$ *such that* $|v_{j'i}| = 1$ *and* $v_{ji} = 0$ *for all* $j \neq j'$.

c. *The minimizers of* $G_1(\boldsymbol{v})$ *with the highest explained variance* $\sum_{i=1}^{N} \boldsymbol{v}_{\cdot i}' \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v}_{\cdot i}$ *have corresponding indices for non-zero entries* $j'(i) \in \arg\max_{j=1,\ldots,p} \left(\hat{\boldsymbol{\Sigma}}_i\right)_{jj}$ *for each source* i. *The minimizers of* $G_2(\boldsymbol{v})$ *with highest explained variance have non-zero entries only for the variable indexed by* $j' \in \arg\max_{j=1,\ldots,p} \sum_{i=1}^{N} \left(\hat{\boldsymbol{\Sigma}}_i\right)_{jj}$.

Proposition 4.3.1.1 forms the basis for constructing the extreme solution set. Starting with the extreme solution of the first component, $\boldsymbol{y}_1^{\infty}(\gamma)$, when $\gamma = 1$, only G_1 is included in the penalty term, with penalty weight η increasing without bounds. Thus, we focus solely on the minimizers of G_1 with the highest variance as an extreme solution. For $\gamma \neq 1$, we also need to minimize G_2. However, since the minimizers of G_2 are also minimizers of G_1, we ultimately seek minimizers of G_2 that explain most of the variance.

Secondly, the set of extreme solutions for a subsequent PC, $\boldsymbol{y}_k^{\infty}(\gamma)$ for $k > 1$, is constructed iteratively. For the k-th component the orthogonality constraints need to be maintained. Under the assumption that the prior PCs are extreme solutions with non-zero variables indexed by $j_1(i), \ldots, j_{k-1}(i)$, $i = 1, \ldots, N$, every minimizer of the penalty terms with non-zero entries indexed by $j_k(i) \notin \{j_1(i), \ldots, j_{k-1}(i)\}, i = 1, \ldots, N$, satisfies the orthogonality constraints.[2] Thus, we can focus on minimizers with the k-th highest explained variance without further adjustments to the orthogonality constraints.

For a given η and γ, we then average the two extreme solutions, $\boldsymbol{y}_k^0$ and $\boldsymbol{y}_k^{\infty}(\gamma)$, to obtain an appropriate starting value for the k-th component, implicitly assuming some continuity of the solution in η. Finally, we project the starting value onto the feasible subspace defined by the optimization constraints using the projection defined in Equation (4.3.6). Thus, the starting value for the k-th PC is

$$\boldsymbol{y}_k = P_{\left(\boldsymbol{v}^{1:(k-1)}\right)^{\perp}} \left(\frac{1}{2}(\boldsymbol{y}_k^0 + \boldsymbol{y}_k^{\infty}(\gamma))\right).$$

The good performance of the starting value is investigated in an additional simulation study in Appendix C.2.

Choice of Penalty Parameter ρ

An open question connected to convergence is how to select the value of the penalty parameter ρ as it vastly influences convergence and also convergence speed. On the one hand, small values for ρ keep the primal residuals larger, implying larger changes

[2] In the case of minimizers of G_2 it holds that $j_l(i) = j_l', i = 1, \ldots, N, l = 1, \ldots, k-1$.

in updates and possibly fewer iterations, hence faster convergence. On the other hand, large values of ρ improve the stability of solving the first minimization problem of the ADMM stated in Equation (4.3.3) for $i = 1$. Hence, a good balance for ρ needs to be found.

Regarding stability of the first subproblem in (4.3.3), note that it can be rephrased to a root finding problem by applying the Karush-Kuhn-Tucker-Theorem (KKT) (see Appendix C.1 for more details) where we use either the proposal from Section 4.3.1 as starting value (for the first iteration; hence $m = 1$), or the outcome of the previous ADMM iteration $\boldsymbol{v}_{(0)}^{m}$ (for iterations $m > 1$). Choosing ρ on the larger side moves the solution for $\boldsymbol{v}_{(1)}^{m+1}$ closer to the starting value $\boldsymbol{v}_{(0)}^{m}$ for the root problem, thus the iteration to a root that also fulfills the inequality condition in the KKT-Theorem becomes more stable.

Finally, based on initial experiments, we found that using a principal component specific penalty parameter, denoted by ρ_k for the k-th component, works well regarding both convergence speed of the ADMM and the root finding problem. In particular, we use

$$\rho_k = \eta + \frac{1}{2N} \sum_{i=1}^{N} \left(\sum_{j=1}^{p} \left(\hat{\boldsymbol{\Sigma}}_i \right)_{jj} - \sum_{l=1}^{k-1} (\boldsymbol{v}^l)' \hat{\boldsymbol{\Sigma}}_i \boldsymbol{v}^l \right),$$

and then increase ρ_k sequentially by 1 if there is either no convergence in the residuals or if the root found is not feasible. [3] This approach is supported by the findings of Ghadimi et al. (2014) regarding the optimal ρ for quadratic problems, which depends on η and the eigenvalues of the matrix in the quadratic term.

Algorithmic Enhancements

To further improve performance and enhance computational speed, we implement several additional algorithmic enhancements.

First, since we are interested in sparse loadings that often cannot be exactly zero due to the iterative nature of the algorithm, in our calculations we round loading entries whose absolute values are below a tolerance of $\epsilon_{thr} = 5 \cdot 10^{-3}$. While a fixed tolerance seems somewhat arbitrary, it enables fair comparisons between different parameter settings and possibly differing algorithm accuracies. Due to rounding small values to zero, the orthogonality constraints might be slightly violated.

Second, we apply *inexact minimization / early termination* for the subproblem regarding the function $f_1(\boldsymbol{v}_{(1)})$, meaning that iterations of the root finder are stopped before full convergence during each step of the ADMM. Thus, fewer iterations per ADMM iteration are needed, leading to speed gains without significant loss of accuracy overall. The tolerance ϵ_{root} indicates an error for finding the root of the function f

[3] Alternatively, one could resort to the often-used approach to dynamically adapt ρ based on the size of the residuals (see Boyd et al., 2011). However, it does not perform well in our case. This likely occurs because the penalty for violating the constraints is reduced too heavily in some steps, leading to non-appropriate root finding starting values and consequently no convergence to a feasible root.

Algorithm 2 ADMM $(\hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_N, \eta, \gamma, \boldsymbol{v}^{1:(k-1)}, \boldsymbol{y}_k, m_{max}, \epsilon_{ADMM}, \epsilon_{root}, \rho)$

Initialize $\boldsymbol{v}_{(1)}, \boldsymbol{v}_{(2)}, \boldsymbol{v}_{(3)}, \boldsymbol{u}_{(1)}, \boldsymbol{u}_{(2)}, \boldsymbol{u}_{(3)} \leftarrow \mathbf{0} \in \mathbb{R}^{Np}$, $\boldsymbol{v}_{(0)}^{new} \leftarrow \boldsymbol{y}_k$, $m \leftarrow 1$
while $m \leq m_{max}$ **do**
  $\boldsymbol{v}_{(0)}^{old} \leftarrow \boldsymbol{v}_{(0)}^{new}$
  Solve subproblems: ▷ See also Appendix C.1
  $\boldsymbol{v}_{(1)} \leftarrow$ solution of Equation (4.3.3) for $i = 1$ with $\boldsymbol{u}_{(1)}$ and $\boldsymbol{v}_{(0)}^{new}$
  $\boldsymbol{v}_{(2)} \leftarrow S(\boldsymbol{v}_{(0)}^{new} - \frac{1}{\rho}\boldsymbol{u}_{(2)}, \frac{\eta\gamma}{\rho})$
  $\boldsymbol{v}_{(3)} \leftarrow S_G(\boldsymbol{v}_{(0)}^{new} - \frac{1}{\rho}\boldsymbol{u}_{(3)}, \frac{\eta(1-\gamma)\sqrt{N}}{\rho})$
  $\boldsymbol{v}_{(0)}^{new} \leftarrow \frac{1}{3}\sum_{i=1}^{3}(\boldsymbol{v}_{(i)} + \frac{1}{\rho}\boldsymbol{u}_{(i)})$
  $\boldsymbol{v}_{(0)}^{new} \leftarrow P_{(\boldsymbol{v}^{1:(k-1)})^{\perp}}(\boldsymbol{v}_{(0)}^{new})$
  $\boldsymbol{u}_{(i)} = \boldsymbol{u}_{(i)} + \rho(\boldsymbol{v}_{(i)} - \boldsymbol{v}_{(0)}^{new})$
  $r \leftarrow \sum_{i=1}^{3} ||\boldsymbol{v}_{(i)} - \boldsymbol{v}_{(0)}^{new}||_2^2$ (primal residual)
  $s \leftarrow 3\rho^2 ||\boldsymbol{v}_{(0)}^{new} - \boldsymbol{v}_{(0)}^{old}||_2^2$ (dual residual)
  Set $\epsilon_{prime}, \epsilon_{dual}$ according to Equation (4.3.4)
  if $r < \epsilon_{prime}$ and $s < \epsilon_{dual}$ **then**
    break
  end if
end while
$\boldsymbol{v}_{(0)}^{new} \leftarrow P_{(\boldsymbol{v}^{1:(k-1)})^{\perp}}(\boldsymbol{v}_{(0)}^{new})$
Set entries of $\boldsymbol{v}_{(0)}^{new}$ with absolute value lower than $\epsilon_{thr} = 0.005$ to 0
Normalize $\boldsymbol{v}_{(0)}^{new} \leftarrow P_{(\mathbf{0})^{\perp}}(\boldsymbol{v}_{(0)}^{new})$
Return $\boldsymbol{v}_{(0)}^{new}$

of $10^{-1}\epsilon_{root}|f| + 10^{-1}\epsilon_{root}$ (see function `multiroot` in package `rootSolve`, Soetaert, 2009) and we allow an increased error of $10\epsilon_{root}$ in the constraints. In our calculations, ϵ_{root} is set to 10^{-2} or 10^{-1} for increased speed.

Third, the algorithm is considerably faster if we project $\boldsymbol{v}_{(0)}^{m}$ after each ADMM iteration step onto the feasible subspace given by the optimization constraints in Equation (4.2.3). Denote the matrix containing all calculated loadings of source i as $\boldsymbol{v}_{\cdot i}^{1:(k-1)} = (\boldsymbol{v}_{\cdot i}^{1}, \ldots, \boldsymbol{v}_{\cdot i}^{k-1})$. Then, the function projecting a vector $\boldsymbol{v} = (\boldsymbol{v}'_{\cdot 1}, \ldots, \boldsymbol{v}'_{\cdot N})'$ to the feasible space for the k-th PC is defined as

$$P_{(\boldsymbol{v}^{1:(k-1)})^{\perp}}(\boldsymbol{v}) = \begin{pmatrix} \frac{\boldsymbol{v}_{\cdot 1} - \sum_{l=1}^{k-1}\langle \boldsymbol{v}_{\cdot 1}, \boldsymbol{v}_{\cdot 1}^{l}\rangle \boldsymbol{v}_{\cdot 1}^{l}}{||\boldsymbol{v}_{\cdot 1} - \sum_{l=1}^{k-1}\langle \boldsymbol{v}_{\cdot 1}, \boldsymbol{v}_{\cdot 1}^{l}\rangle \boldsymbol{v}_{\cdot 1}^{l}||_2} \\ \vdots \\ \frac{\boldsymbol{v}_{\cdot N} - \sum_{l=1}^{k-1}\langle \boldsymbol{v}_{\cdot N}, \boldsymbol{v}_{\cdot N}^{l}\rangle \boldsymbol{v}_{\cdot N}^{l}}{||\boldsymbol{v}_{\cdot N} - \sum_{l=1}^{k-1}\langle \boldsymbol{v}_{\cdot N}, \boldsymbol{v}_{\cdot N}^{l}\rangle \boldsymbol{v}_{\cdot N}^{l}||_2} \end{pmatrix}. \tag{4.3.6}$$

If $k = 1$, we project onto to orthogonal space of the null vector $\mathbf{0}$, thus we are normalizing only.

Finally, an overview of the ADMM algorithm for the k-th PC is summarized in Algorithm 2. We achieve convergence in all simulations and real data examples using the algorithmic fine tuning and choices described throughout Section 4.3.1.

4.3.2 Hyperparameter Selection

In Section 4.3.2 we provide criteria to select the sparsity hyperparameters, in Section 4.3.2 hyperparameter of the ssMRCD estimator are discussed and Section 4.3.2 provides guidance on how to determine the number of PCs.

Tuning Parameters for (Joint) Sparsity

The tuning parameters of the objective function (4.2.3), i.e. the parameter for the overall amount of sparsity η and the parameter for the trade-off between global and local sparsity γ, need to be selected.

To select γ, we propose an optimality criterion for the first principal component which balances explained variance and sparsity in the loadings; a balance that is commonly desired for sparse PCA. In particular, we maximize the explained variance $\mathcal{V}(\boldsymbol{v}) = \boldsymbol{v}'\hat{\boldsymbol{\Sigma}}\boldsymbol{v}$ and the mean $\mathcal{S}(\boldsymbol{v})$ of the standardized entry- and groupwise sparsity

$$\mathcal{S}(\boldsymbol{v}) = \frac{1}{2}\left(\frac{\#\{v_{ji} = 0, i = 1, \ldots, N, j = 1, \ldots, p\}}{N(p-1)} + \frac{\#\{||\boldsymbol{v}_{j\cdot}||_2 = 0, j = 1, \ldots, p\}}{p-1}\right).$$

Since $\mathcal{V}(\boldsymbol{v})$ and $\mathcal{S}(\boldsymbol{v})$ vary also over η, the optimal γ is then chosen to be the maximizer of the area under the curve (AUC) of sparsity $\mathcal{S}(\boldsymbol{v})$ and the explained variance, standardized to the two extreme solutions,

$$\frac{\mathcal{V}(\boldsymbol{v}) - \mathcal{V}(\boldsymbol{y}_1^{\infty})}{\mathcal{V}(\boldsymbol{y}_1^{0}) - \mathcal{V}(\boldsymbol{y}_1^{\infty})}, \tag{4.3.7}$$

along the trajectory path for varying η, stopping at full sparsity. For computational efficiency, we keep the same selected γ for the higher-order PCs.

To select η, we propose a simple approach of optimizing the trade-off product (TPO) of the standardized number of zero entries and the standardized explained variance (4.3.7) for the first principal component,[4]

$$TPO = \left(\frac{\#\{v_{ji} = 0, i = 1, \ldots, N, j = 1, \ldots, p\}}{N(p-1)}\right)\left(\frac{\mathcal{V}(\boldsymbol{v}) - \mathcal{V}(\boldsymbol{y}_1^{\infty})}{\mathcal{V}(\boldsymbol{y}_1^{0}) - \mathcal{V}(\boldsymbol{y}_1^{\infty})}\right).$$

For further PCs, the optimal η is adjusted according to the residual variance to distribute sparsity more equally across the loadings. The degree of sparsity for the l-th PC, given the optimally selected η of the first PC, is $\eta_l = g_l\eta$, $l > 1$. To calculate g_l, we use the projected covariance matrix $\hat{\boldsymbol{\Sigma}}_i$ of the orthogonal space of the corresponding first $l-1$ PCs, per source. We propose to use the sum of the first eigenvalues, $\tilde{\lambda}_1$, of the projected covariance matrices as scaling factor,

$$\tilde{g}_l = \sum_{i=1}^{N} \tilde{\lambda}_1 \left(\left(\boldsymbol{I}_p - \boldsymbol{v}_{\cdot i}^{1:(l-1)}\left(\boldsymbol{v}_{\cdot i}^{1:(l-1)}\right)'\right)\hat{\boldsymbol{\Sigma}}_i\left(\boldsymbol{I}_p - \boldsymbol{v}_{\cdot i}^{1:(l-1)}\left(\boldsymbol{v}_{\cdot i}^{1:(l-1)}\right)'\right)\right),$$

with $\boldsymbol{I}_p$ being the p-dimensional identity matrix. Finally, the standardized value $g_l = \tilde{g}_l/\tilde{g}_1$ is used for scaling.

[4] Alternatively, one can resort to BIC-based approaches as in Hubert et al. (2016) and Croux et al. (2013). We prefer the approach based on the TPO instead since a BIC-based approach becomes more complicated with multiple covariances and the additional groupwise sparsity penalty regarding degrees of freedom.

Hyperparameters for the ssMRCD

The ssMRCD plug-in estimator requires additional hyperparameters that need to be set (see Puchhammer and Filzmoser, 2024, 2023). Since the partition into multiple sources and the weights $\boldsymbol{W}$ between them are data dependent, there is not a general rule how to set them, except the notions that are elaborated on in Section 4.2.3.

For the smoothing parameter λ, the selection criterion described in Puchhammer and Filzmoser (2024) is not applicable in our case since it is based on local outlier detection. We therefore derive a new approach to set λ in a more general setting based on the idea that data should be described as well as possible by the means and covariances that are produced by the ssMRCD model. The model residuals per source $\boldsymbol{r}_{i,\iota}$, $\iota = 1, \ldots, n_i$, are

$$\boldsymbol{r}_{i,\iota} = \hat{\boldsymbol{\Sigma}}_i^{-1/2}(\boldsymbol{x}_{i,\iota} - \hat{\boldsymbol{\mu}}_i),$$

and if we have a good estimation of the data, the mean of the smallest α-fraction of residual norms over all sources

$$R = \frac{1}{h_1 + \ldots + h_N} \sum_{\iota=1}^{h_1+\ldots+h_N} ||\boldsymbol{r}_{(\iota)}||_2, \tag{4.3.8}$$

where $\boldsymbol{r}_{(1)}, \ldots, \boldsymbol{r}_{(h_1+\ldots+h_N)}$ are the $h_1 + \ldots + h_N$ smallest residuals, should be small. Hence, minimizing R will be the criterion for the optimal λ. If the partition cannot be derived from the data context, the same approach can be used to find a good grouping or even good weights, although computationally very expensive.

Number of Principal Components

Finally, the number of principal components necessary to describe the data appropriately needs to be selected. We propose to use the cumulative percent variation (CPV), as suggested in Hubert et al. (2016). The number of components should at least cover a certain threshold, for instance 80% of the overall (global) variation,

$$\frac{\sum_{l=1}^{k} (\boldsymbol{v}^l)' \hat{\boldsymbol{\Sigma}} \boldsymbol{v}^l}{\text{trace}(\hat{\boldsymbol{\Sigma}})} \geq 80\%. \tag{4.3.9}$$

Depending on the research question, other summary statistics connected to CPV on a source level are applicable as well, e.g., the minimal CPV over all sources should be at least 80%, or an adapted scree plot consisting of boxplots can be used (see also Section 4.5).

4.4 Simulations

We introduce two simulation setups. In Section 4.4.1, we investigate the induced sparsity patterns of the multi-source PCA method for varying degrees of global and local sparsity. In Section 4.4.2 the performance of the method first in absence and

then in presence of outliers is investigated. Finally, note that in Appendix C.2, we present results of additional simulation experiments on the suitability and efficiency of the proposed starting values.

Across simulations, the different methods are compared using multiple evaluation criteria that are averaged over repetitions and sources. First, we compute the angle between the real subspace and the estimated subspace spanned by the first (for $k = 1$) or first and second (for $k = 2$) real and estimated loading, respectively, according to Hubert et al. (2016, 2005), standardized to $[0, 1]$. Second, we obtain the orthogonal distance of a non-contaminated observation $\boldsymbol{x}_{i,\iota}$,

$$OD_{i,\iota} = ||\boldsymbol{x}_{i,\iota} - \hat{\boldsymbol{\mu}}_i - \left(\boldsymbol{v}_{\cdot i}^1, \ldots, \boldsymbol{v}_{\cdot i}^k\right) \boldsymbol{t}_{i,\iota}||_2, \tag{4.4.1}$$

as measure for good data projection and reconstruction abilities of the components. The OD means of clean observations over all sources per simulation scenario are linearly scaled to $[0, 1]$ for illustration purposes.

Next, concerning sparsity, we first present the level of sparsity selected by each method as a reference

$$\frac{\#\{v_{ji} = 0, i = 1, \ldots, N, j = 1, \ldots, p\}}{N(p-1)}.$$

From a sparsity recognition point of view, we then analyze standard evaluation techniques to check if methods correctly specify sparse and non-sparse variables. The *true negative rate* (TNR) specifies the percentage of correctly identified non-zero entries of the real loadings $\tilde{\boldsymbol{v}}$ and the *true positive rate* (TPR) the correctly found zero entries of $\tilde{\boldsymbol{v}}$,

$$TNR = \frac{1}{N} \sum_{i=1}^{N} \frac{\#\{j \in 1, \ldots, p : v_{ji} \neq 0, \tilde{v}_{ji} \neq 0\}}{\#\{j \in 1, \ldots, p : \tilde{v}_{ji} \neq 0\}},$$

$$TPR = \frac{1}{N} \sum_{i=1}^{N} \frac{\#\{j \in 1, \ldots, p : v_{ji} = 0, \tilde{v}_{ji} = 0\}}{\#\{j \in 1, \ldots, p : \tilde{v}_{ji} = 0\}}.$$

We also include three evaluation measures to combine these two measures into one. The well known F1-Score and the zero-measure (Z-measure) introduced in Hubert et al. (2016), which calculates the percentage of overall correctly identified entries without partitioning into groups first, are applicable in a balanced setting. However, in an unbalanced setting with high sparsity, F1 and the Z-measure essentially lead to ignorance of correctly identified non-sparse entries and the amount of correctly identified sparse entries drives a “good" performance and gives incentives to overestimate the sparsity pattern. In such settings, we prefer to use the geometric mean (G-Mean) of TNR and TPR instead.

4.4.1 Detecting Sparsity Patterns

We start by investigating whether the proposed sparse multi-source PCA method delivers entry- and groupwise sparse loadings as desired when no outliers are present.

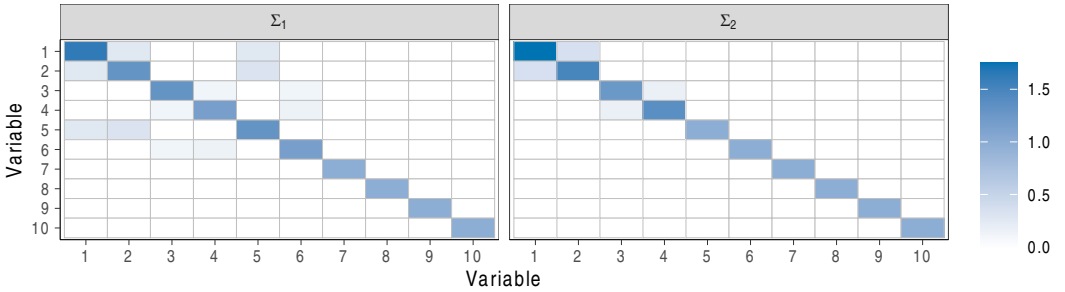


Figure 4.4.1: Heat map of the two covariance matrices used as basis for all simulation settings with $p = 10$ variables. Each covariance entry is colored according to its value.

To this end, we construct two covariance matrices for $N = 2$ groups using different sparse loading matrices $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, similar to the simulation setting of Croux et al. (2013),

$$\boldsymbol{P}_1 = \begin{bmatrix} \sqrt{\frac{1}{2}} & 0 & -\sqrt{\frac{1}{2}} & 0 & 0 & 0 \\ \sqrt{\frac{1}{4}} & 0 & \sqrt{\frac{1}{4}} & 0 & -\sqrt{\frac{1}{2}} & 0 \\ 0 & \sqrt{\frac{1}{2}} & 0 & -\sqrt{\frac{1}{2}} & 0 & 0 \\ 0 & \sqrt{\frac{1}{4}} & 0 & \sqrt{\frac{1}{4}} & 0 & -\sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{4}} & 0 & \sqrt{\frac{1}{4}} & 0 & \sqrt{\frac{1}{2}} & 0 \\ 0 & \sqrt{\frac{1}{4}} & 0 & \sqrt{\frac{1}{4}} & 0 & \sqrt{\frac{1}{2}} \\ & & \mathbf{0}_{(p-6)\times 6} & & & \boldsymbol{I}_{p-6} \end{bmatrix}, \boldsymbol{P}_2 = \begin{bmatrix} \sqrt{\frac{2}{3}} & 0 & -\sqrt{\frac{1}{3}} & 0 \\ \sqrt{\frac{1}{3}} & 0 & \sqrt{\frac{2}{3}} & 0 \\ 0 & \sqrt{\frac{1}{3}} & 0 & -\sqrt{\frac{2}{3}} \\ 0 & \sqrt{\frac{2}{3}} & 0 & \sqrt{\frac{1}{3}} \\ \mathbf{0}_{(p-4)\times 4} & & & \boldsymbol{I}_{p-4} \end{bmatrix},$$

and a common eigenvalue diagonal matrix, $\boldsymbol{D} = \text{Diag}(2, 1.5, 1.25, 1.125, 1, \ldots, 1)$. The covariance matrices are then constructed based on the eigen-decomposition for each source as $\boldsymbol{\Sigma}_i = \boldsymbol{P}_i \boldsymbol{D} \boldsymbol{P}_i'$, for $i = 1, 2$ and visualized in Figure 4.4.1. Random noise $\epsilon \sim \mathcal{N}(0, 0.1)$ is (symmetrically) added for each entry of the covariance matrices for each simulation run individually to address possible uncertainty in the covariance estimation of $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$. While the first loadings are directionally similar, the second loadings imply opposing directions of highest variance between different sources to cover also the scenario of non-compliant dominating groups. We consider 100 simulation repetitions for $p = 10$ variables, we set $\epsilon_{root} = 0.1$, and take different values of the sparsity parameters $\eta = 0, 0.05, \ldots, 1.25, \gamma = 0, 0.5, 1$. Then, we apply our sparse multi-source PCA procedure using the real covariance matrices $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ as plug-in and obtain the first two principal components. For the first PC, variables 1,2 and 5 have non-zero loadings in group 1 (see non-zero entries in the first column of $\boldsymbol{P}_1$), whereas only variables 1 and 2 have non-zero loadings in group 2 (see non-zero entries of the first column in $\boldsymbol{P}_2$). For the second PC, similarly, variables 3, 4 and 6 have non-zero loadings in the first group whereas only the former two have non-zero loadings in the second group.

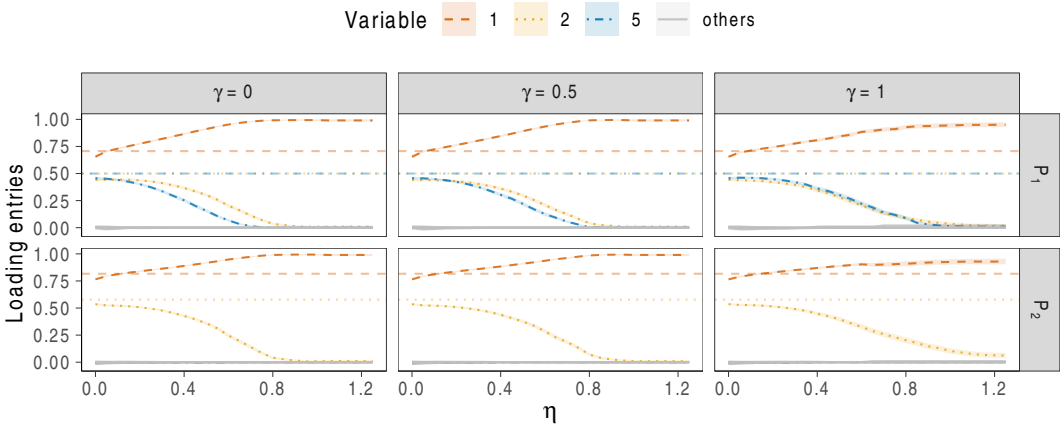


Figure 4.4.2: The mean of loading entries of the first PC with a band for the standard error. The seven solid gray lines depict the entries that are sparse by construction, the dashed and dotted lines depict the variables that are not sparse in at least one source loading. The horizontal lines indicate the true value of the corresponding loading entries.

The resulting loading entries for varying parameters η and γ are visualized in Figure 4.4.2 for the first PC and in Figure 4.4.3 for the second PC. Colored and non-solid lines indicate the variables whose loading entries are non-zero, and the corresponding horizontal lines the respective values, according to the true loadings $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$. The gray lines correspond to variables with zero loading entries, and the shaded area around each loading entry indicates the standard error. We can clearly see that the true structured sparsity patterns are recovered successfully, and the sparse multi-source PCA method thus succeeds in separating the important variables with non-zero loadings from the unimportant variables with zero loadings. The estimated loadings for the first PC are very similar across different values of the hyperparameter γ (in the different panels of Figure 4.4.2). In contrast, for the second PC, we see large differences in the variables 3 and 4, that have non-zero loading entries in both sources. By increasing γ they are kept at more accurately high levels for a larger range of η until the rise of the gray solid lines around $\eta \geq 0.8$ indicates a trickling down of the variability of the real first PC, that is not accounted for in the estimated fully sparse first PC for high η.

Altogether, the proposed multi-source PCA method succeeds in recovering the true sparsity patterns present in the PCA loadings in an idealized setting where the covariance matrices are known upfront. In the next section, we evaluate the performance of the method across different evaluation metrics; and this in the realistic scenario where the covariance matrices of the N sources need to be estimated, not only in case of clean data but we also discuss the impact of different outlier scenarios on its performance.

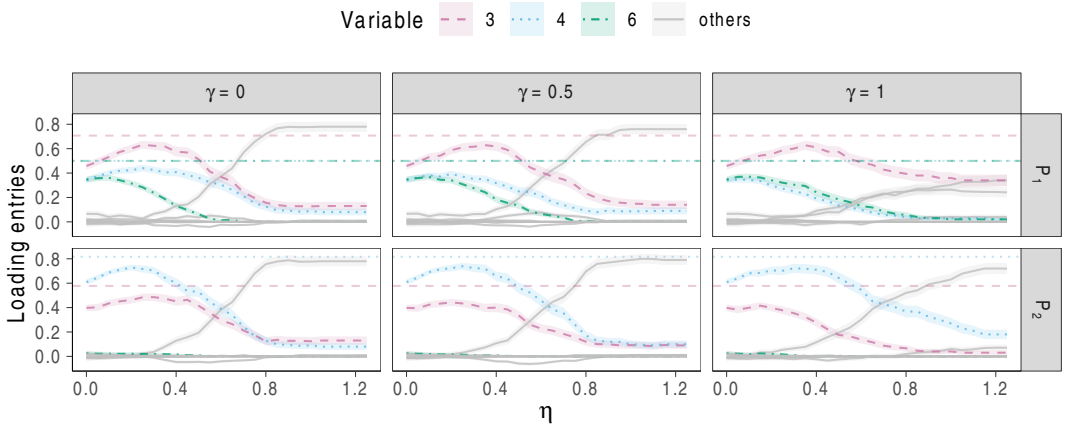


Figure 4.4.3: The mean of loading entries of the second PC with a band for the standard error. The seven solid gray lines depict the entries that are sparse by construction, the dashed and dotted lines the variables that are not sparse in at least one source loading. The horizontal lines indicate the true value of the corresponding loading entries.

4.4.2 Outlier Robustness

We now evaluate the performance of the proposed sparse outlier-robust PCA method for multi-source data in more detail, first when no outliers are present and we subsequently discuss the impact on its performance in presence of outliers. We hereby estimate the covariances matrices of the N sources using the ssMRCD estimator.

We construct linearly shifting covariance matrices by using a convex combination of the two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ of the simulation setting described in Section 4.4.1. With $N \geq 2$ being the number of sources, the covariance $\tilde{\mathbf{\Sigma}}_i$ for source i for $i = 1, \ldots, N$ is constructed according to

$$\tilde{\mathbf{\Sigma}}_i = \left(1 - \frac{i-1}{N-1}\right)\mathbf{\Sigma}_1 + \frac{i-1}{N-1}\mathbf{\Sigma}_2.$$

The corresponding real loadings are then just the eigenvectors of $\tilde{\mathbf{\Sigma}}_i$.

Similar to the simulation setting of Croux et al. (2013), clean data points for each source i are drawn from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \tilde{\mathbf{\Sigma}}_i)$ and a certain ϵ_{out} fraction of shift outliers are drawn from $\mathcal{N}(\boldsymbol{\mu}_{out}, \boldsymbol{I}_p)$ with

$$\boldsymbol{\mu}_{out} = \sqrt{2}(2, 4, 2, 4, 0, -1, 1, 0, 1, -1, \ldots, 0, 1, -1)'$$

per source.

We compute several versions of the multi-source PCA method, to appropriately evaluate the contributions of its three main features, namely in (1) delivering structured sparse loadings, (2) exploiting the multi-source aspect and in (3) providing protection against outliers. Our proposal that delivers all three aspects is labeled *ssMRCD-PCA* in the remainder. It computes the ssMRCD estimates with $\alpha = 0.5$, a band matrix,

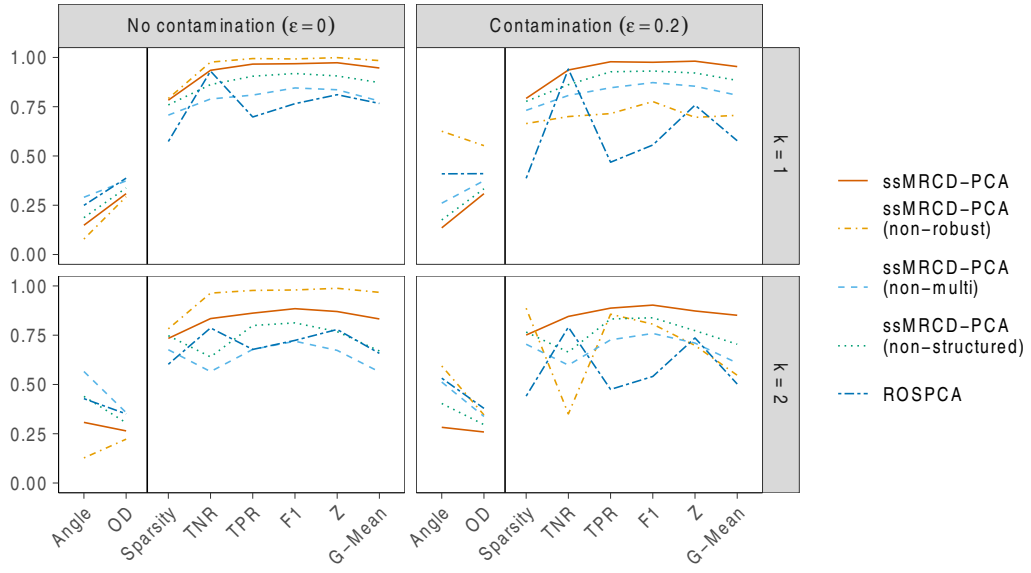


Figure 4.4.4: Comparison of performance of the first two components for equally contaminated data sources ($N = 10$).

constructed with ones on the off-diagonals, zeros in the diagonal and appropriately scaled, as weight matrix $\boldsymbol{W}$, and the optimal smoothing criteria from Equation (4.3.8) for the selection of λ. Regarding the PCA sparsity parameters, they are optimized using the optimization approaches described in Section 4.3.2 on a grid of $\gamma = 0, 0.1, \ldots, 1$ and $\eta = 0, 0.1, \ldots, 5$, or stopped prior if full sparsity is achieved. Next, *ssMRCD-PCA (non-robust)* uses the ssMRCD estimator as described above but with $\alpha = 1$, thereby using all observations for its computation and, hence, providing no protection against outliers. *ssMRCD-PCA (non-multi)* neglects the multi-source aspect in total and computes PCA without exploiting neither the joint estimation across the multiple sources of the ssMRCD estimator ($\lambda = 0$) nor the shared sparsity patterns ($\gamma = 1$). Finally, *ssMRCD-PCA (non-structured)* uses the proposed ssMRCD plug-in estimator but no structured sparsity in the PCA step, hence $\gamma = 1$. Moreover, we also compare these four versions of the multi-source PCA method to the ROSPCA method introduced by Hubert et al. (2016) which is a state-of-the-art benchmark for sparse outlier-robust PCA for which open-source code is easily available. To this end, we use the R-package `rospca` (Reynkens, 2018) with their implemented optimal sparsity approach. Note that this method is not tailored towards multi-source data, hence we apply it for each source individually.

In Figure 4.4.4 our proposed method ssMRCD-PCA is compared to its three variants that either neglect robustness (non-robust), multi-sourceness (non-multi) or structured sparsity (non-structured), as well as to the benchmark method ROSPCA. For each method we calculate the first two PCs for 100 simulation repetitions and $p = 10$ with varying contamination level $\epsilon = 0\%$ (hence no contamination) and $\epsilon = 20\%$ and number of data observations per source $n = 500$, both constant over all $N = 10$ sources. First,

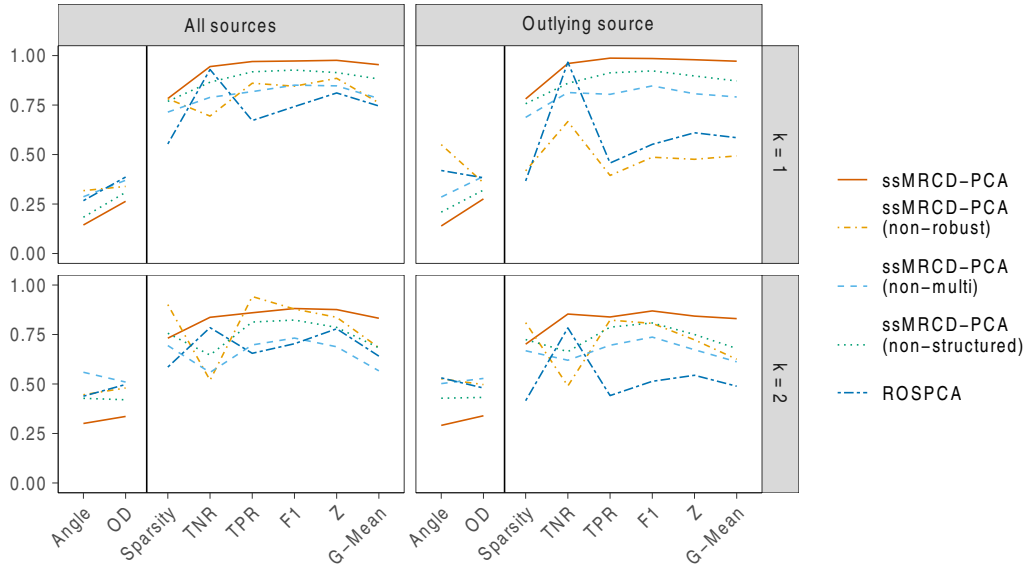


Figure 4.4.5: Comparison of performance of the first two components with locally contaminated data (with contamination level of $\epsilon = 20\%$) in source 5 of $N = 10$ sources for $n = 500$. The results for data source 5 are shown in the right panel, the mean over all sources is shown in the left panel.

when no outliers are present ("No contamination" panel in Figure 4.4.4), the proposed ssMRCD-PCA with a robust plug-in estimator is, as expected, slightly less effective than its version that uses a non-robust plug-in, ssMRCD-PCA (non-robust). Still, it is, generally, more effective than its benchmark ROSPCA.

Furthermore, the price for neglecting the structured sparsity patterns (see ssMRCD-PCA (non-structured) and also ssMRCD-PCA (non-multi)) concerns all evaluation criteria, however, the increased TNR for the second component is especially evident. Comparing these two versions with the proposed ssMRCD-PCA, clear benefits are noticeable for both, considering structured sparsity as well as considering the multi-source aspect also for variance computation.

When outliers are present (panel "Contamination" in Figure 4.4.4), the importance of using a robust method becomes directly apparent, since the variant ssMRCD-PCA (non-robust) is heavily affected by the outliers: the criteria connected to data reconstruction, i.e. the angle and OD, show inferior performance compared to the proposed ssMRCD-PCA method with robust plug-in, while the detection of sparsity patterns is comparable to ROSPCA. We can also see that for the first component the proposed method ssMRCD-PCA provides better results than ROSPCA in all measurements and settings. Especially interesting compared to ROSPCA is the combination of higher sparsity with a lower angle and low OD. This implies a good fit of the highly sparse loadings to the data and the subspace of highest variation. Moreover, the sparsity recognition metrics confirm that the correct sparsity structure is found. The cost of neglecting either the multi-source aspect or the structured sparsity

remains similar to the uncontaminated case.

Finally, another interesting outlier configuration to analyze in the context of multi-source PCA is how *locally contaminated data*, in this context meaning contamination in only one source, affects the PCA results. Therefore, we stick to the data contamination setting with $N = 10$ data sources but instead of contaminating all sources equally, we only contaminate the fifth of $N = 10$ sources with $\epsilon = 20\%$ outlying observations. We use 100 simulation repetitions, $n = 500$ observations, and $p = 10$ variables. The results are shown in Figure 4.4.5.

We can see that the proposed multi-source PCA method provides consistently better results in both PCs than its benchmark ROSPCA. When we analyze the results for the single contaminated data source (right panel in Figure 4.4.5), we see very stable results for the proposed ssMRCD-PCA method. This is in contrast to the results of ROSPCA with contamination, where the performance on the contaminated source is clearly worse than the average over all sources (left panel in Figure 4.4.5) in almost all performance measures and both components. The non-robust version of the ssMRCD-PCA based method also shows reasonable performance when averaged over all sources. Yet, even with local contamination we can see a stark performance decline in the contaminated source, especially in the first PC. This indicates that provided a multi-source scenario, inherent similarities in the covariances between groups should be leveraged. Applying additional smoothing, further stabilizes the covariance estimation, even in a non-robust setting, and in combination with groupwise sparsity we achieve reliable sparse loadings.

4.5 Applications

We demonstrate the usefulness of the proposed sparse multi-source PCA method on two diverse applications, namely one on multivariate time series data from an Austrian weather stations (Section 4.5.1) and the second on measurements of plant geochemistry (Section 4.5.2).

4.5.1 Weather Analysis at Hohe Warte

We analyze daily weather measurements of the weather station *Hohe Warte* in Vienna, Austria over the years 1960-2023 as provided by GeoSphere Austria (2024). The data set consists of $p = 13$ variables covering the amounts of sunshine, wind, cloud coverage as well as temperature, humidity, air pressure and visibility (see Appendix C.3 for the full list) for $N = 64$ sources corresponding to the different years and overall $\sum_{i=1}^{64} n_i = 23,372$ observations. For preprocessing, we standardize the variables to the corresponding medians and the mean absolute deviations from the years 1960 to 1980 that are used as a baseline for proceeding climatic developments.

For the computation of the ssMRCD plug-in estimator, we assume that each year is similar to five prior and five subsequent years with a linear decrease in similarity

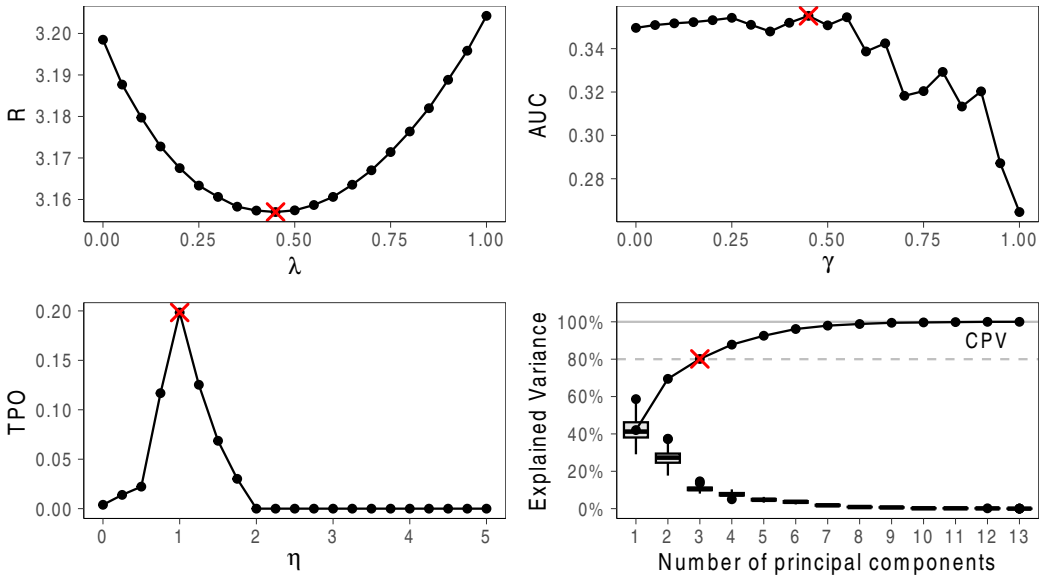


Figure 4.5.1: Hohe Warte weather station: Optimal smoothing and sparsity parameters.

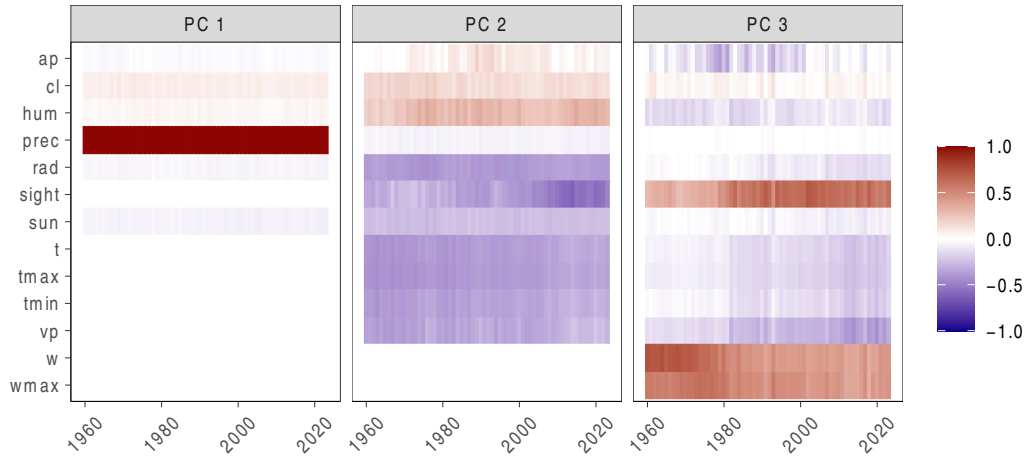


Figure 4.5.2: Hohe Warte weather station: Heat map of the loadings for the $p = 13$ variables (rows) over time (columns) of the first three components (panels).

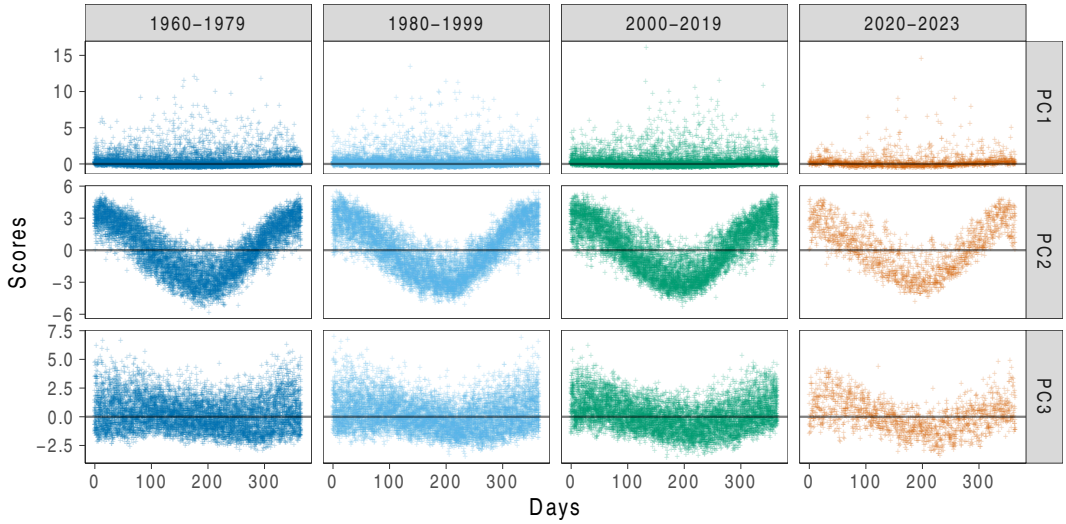


Figure 4.5.3: Hohe Warte weather station: Scores for each observation of the first three components (rows) partitioned into four consecutive time subsets (columns).

leading to the weighting matrix $\boldsymbol{W}$ structured as a band matrix given by

$$\begin{pmatrix} 0 & 5 & \dots & 1 & & \\ 5 & \ddots & \ddots & & \ddots & \\ \vdots & \ddots & \ddots & \ddots & & 1 \\ 1 & & \ddots & \ddots & \ddots & \vdots \\ & \ddots & & \ddots & \ddots & 5 \\ & & 1 & \dots & 5 & 0 \end{pmatrix},$$

where each row is then scaled to have an overall sum of 1 (see also Puchhammer and Filzmoser, 2024). The amount of smoothness, λ, of the ssMRCD estimator is optimized over the interval $[0, 1]$ with step size 0.05 to minimize the residual norm in Equation (4.3.8). The upper left part of Figure 4.5.1 shows the residual norm varying over the amount of smoothing and the optimal value of $\lambda = 0.45$. Based on the optimally smoothed set of covariances, the sparsity parameters of the proposed sparse multi-source PCA method are selected using a step size of 0.05 for γ and 0.25 for η. The optimal values are then $\gamma = 0.45, \eta = 1$ (see upper right and lower left part of Figure 4.5.1, respectively). The corresponding boxplot-based scree-plot is shown in the lower left part of Figure 4.5.1. Each boxplot is constructed per PC and is based on the individually explained variance per year for all $N = 64$ years. According to the CPV-criterion Equation (4.3.9), we analyze only the first three components, since they explain 80% of the overall variance.

Figure 4.5.2 presents the sparsity patterns of the loadings for the first three PCs obtained by our sparse multi-source PCA method. We see three different causes of

variation. The first PC (left panel) is mainly composed of precipitation and displays a clear global pattern across all years, meaning that precipitation drives most of the weather variation over the year. Figure 4.5.3 displays the corresponding scores (top row), partitioned into time subsets, and shows rather constant variation over time. The second component (middle panel in Figure 4.5.2) consists mainly of temperature, vapor pressure, sight, radiation and sun as well as humidity and cloud cover in the opposed direction. By inspecting the scores in the middle row of Figure 4.5.3, we can see that the second component captures the seasonality and the corresponding variability over each year. Again, the pattern seems to be rather stable over the 64 years.

Finally, the loadings for the third component are visualized in the right panel of Figure 4.5.2 and consist again of temperature variables and sight. Yet, in contrast to the first two PCs, a trend becomes visible in their loadings on the third PC as can be seen from darker colors in the heat map for the more recent years. Also maximal wind speed and wind velocity display important loading entries, which are stable or rather decreasing in importance for variability over the years, respectively. This changing pattern over the years is also visible in the shape of the scores displayed in Figure 4.5.3, bottom row. While the years 1960-1979 do not seem to exhibit seasonality in the scores, such a pattern becomes more apparent for the more recent years. While pinpointing the source of variability for the third component is more difficult than for the first two components, one possible source could be related to climate change, apparent from the evolving trend over the long time frame of 64 years. The smooth transitioning over time as mainly visible in the third component is directly detectable from our multi-source PCA method whereas it would remain unnoticed from a standard (global) PCA analysis.

Finally, via two score-related measures we also demonstrate the need for robust methods, namely by showing the presence of outliers in the data. First, the OD defined in Equation (4.4.1) measures the distance of each observation from the estimated principal component subspace and thus, how strongly the observation disagrees with the direction of highest variance of the data majority. A single-source upper cut-off value for outlier detection is proposed by Hubert et al. (2005) as $(\hat{\mu}_{MCD} + \hat{\sigma}_{MCD} z_{0.975})^{3/2}$, where $\hat{\mu}_{MCD}$ and $\hat{\sigma}_{MCD}$ are univariate MCD estimates of OD and $z_{0.975}$ is the 97.5% quantile of the standard normal distribution.

Secondly, the so-called *score distance* (SD) for observation $\boldsymbol{x}_{i,\iota}$ is defined as

$$SD_{i,\iota} = \sqrt{\boldsymbol{t}_{i,\iota}' \boldsymbol{L}_i^{-1} \boldsymbol{t}_{i,\iota}},$$

where $\boldsymbol{L}_i$ denotes the diagonal matrix of the eigenvalues of the first k PCs of source i. Observations with high SD are not outlying with respect to the direction of variance as for OD, but they are outlying from the main data cloud within the projected subspace. For robust PCA, a typical cut-off value of $\sqrt{\chi^2_{k,0.975}}$ is often proposed for SD (see also Hubert et al., 2016). Note that both cut-off values are based on theoretical results that are not directly applicable in a multi-source context and thus, the cut-off values should be used as orientation rather than fixed cut-offs for outlier detection.

In Figure 4.5.4 the densities of OD and SD are shown separately for the four time spans from Figure 4.5.3 together with the respective cut-off values. While there are

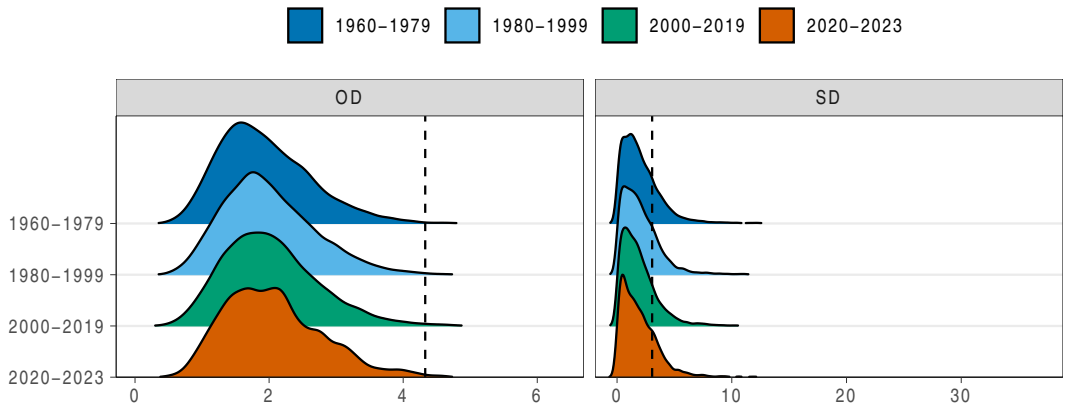


Figure 4.5.4: Hohe Warte weather station: Orthogonal distance (OD, left panel) and score distance (SD, right panel) densities for the considered time subsets based on the first three PCs as well as standard cut-off values for OD and SD as vertical dashed lines. The horizontal axis spans to the maximum value of SD and OD respectively.

observations with high SD and OD—also exceeding the cut-off values—present in all time subsets showcasing the importance of robust estimation procedures in general, we also see additional bumps around the OD range of $2.5 - 4$ for the years 2020-2023 specifically. The frequency of high SD seems to rise as well, together indicating an increase in extreme weather observations in the most recent years which thus justifies the need for an outlier-robust joint PCA method.

4.5.2 Geochemical Plant Analysis

Our second application demonstrates the usefulness of sparse PCA for multi-source data with a more general grouping structure. The data consists of $n = 547$ observations of $p = 19$ element concentrations originating of $N = 6$ different plant species (Norway Spruce, Common Juniper and Scots Pine) and organs (bark, needle, twig) and was collected during the NEXT project funded by the EU (NEXT, 2021) in order to draw conclusions for mineral exploration. The aim is to explore differences and similarities of variance among the different plant groups and the suitability of the scores to discriminate between mineralizations and non-mineralizations. Since mineralizations are supposed to have a geochemical composition that is different from non-mineralized areas, these observations form outliers, which calls for robust procedures.

For applying the ssMRCD estimator, we assume equal amounts of similarity between observations of the same plant species or of the same organs when constructing the weight matrix $\boldsymbol{W}$. Moreover, due to the compositional nature of element concentrations, we apply the standard *isometric log-ratio*(ilr) transformation to the data (see e.g., Filzmoser et al., 2018) known from compositional data analysis. Based on the optimal smoothing criteria R, the optimal value for group smoothing is $\lambda = 0.35$.

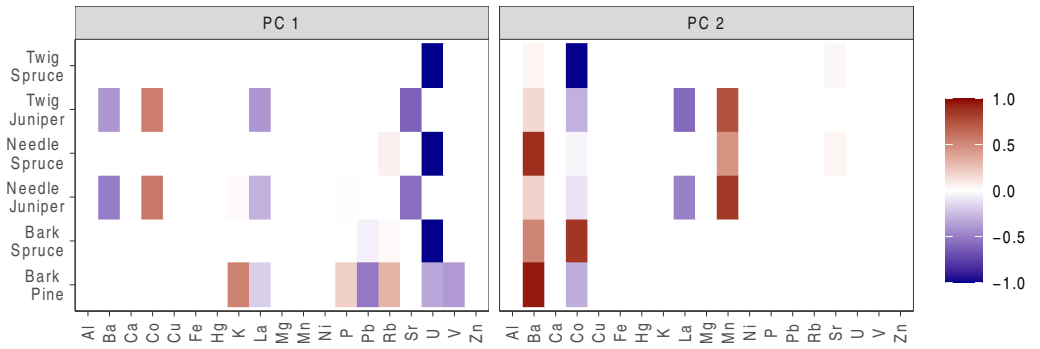


Figure 4.5.5: Plant geochemistry: Heat map of loadings for the $p = 19$ variables (columns) over plant-organ species combination (rows) on the first and second principal component (panels)

.

After the calculation of the ssMRCD covariance matrices, they are (linearly) transformed from ilr to *centered log-ratios* (clr) coordinates to increase the interpretability of the principal components' loadings and scores. The clr-transformation essentially standardizes each variable with the geometric mean per observation, followed by a log-transformation. While clr leads to linear dependent variables opposed to ilr, leading to numerical issues for covariance estimation (especially determinant based estimators like the ssMRCD estimator), it is possible to intuitively interpret clr as relative importance of elements which is not possible for ilr. Thus, we apply the sparse multi-source PCA algorithm to the transformed covariance matrix of clr variables. Optimal parameters are then given by $\gamma = 0.6, \eta = 0.15$ (see also Figure C.3 in Appendix C.4).

In Figure 4.5.5 the loadings of the first and second principal components explaining around 33% of overall variance are shown per source,[5] being a combination of a plant species and organ. We see clear similarities across all organs of the juniper species in both components and of the spruce species for the first component. Only in the second component the organs of the same species (spruce) start to show differences. Moreover, pine bark has the most complexity in the loading structure of the first component. This combination of heavy metals like uranium (U), vanadium (V) and lead (Pb) against phosphorus (P), potassium (K) and rubidium (Rb) is to be expected from physiological characteristics of pine bark.

When it comes to mineral exploration, a goal of the NEXT project, it would be interesting if we can find a plant organ-species combination and a direction of variation along which the discrimination between mineralizations and non-mineralizations is visible. To investigate this, we use the geological classification between calcsilicate rocks and mafic rocks. Mafic rocks are often associated with volcanic and intrusive activities and they can indicate the presence of specific mineral deposits like nickel, copper, and platinum group elements.

Taking a look at Figure 4.5.6 we can see the distribution of the scores connected to

[5] Since 10 PCs are needed to explain 80% of the data (see also Figure C.3) we will focus on the first two components for interpretation.

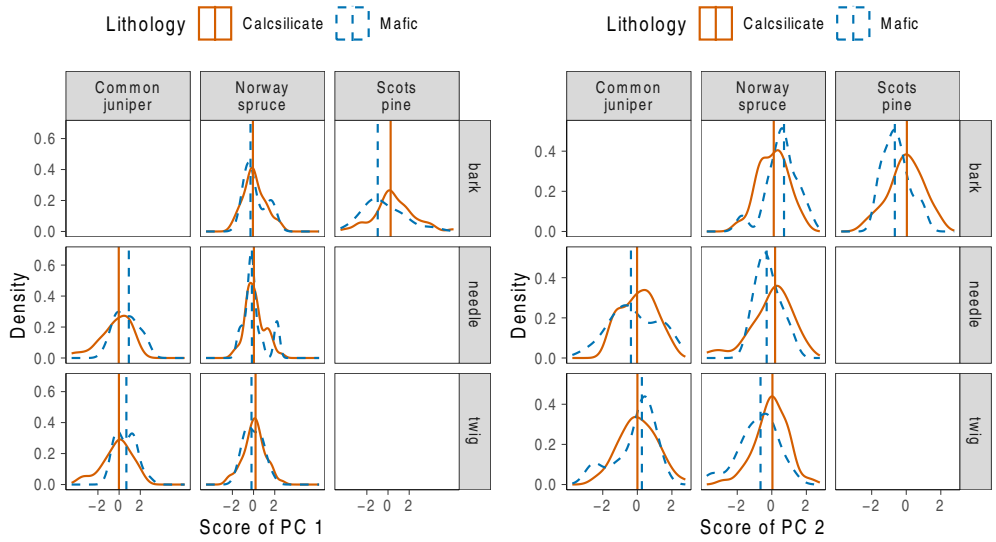


Figure 4.5.6: Plant geochemistry: Density and median (vertical line) for calcsilicate rock (solid) and mafic rock (dashed) measurements for the first PC for all plant species (columns) - organ (rows) combinations. Note that empty panels correspond to combinations of plant species and organs not present in the considered data set.

the first (left) and second (right) PC as density and the median as vertical lines for different groups split into observations connected to calcsilicate and mafic rocks. Other geological units are not shown. The bark of Scots pine has the most differentiable peaks and medians, indicating the possibility to use this plant organ-species combination with the elements of the first loading for mineral exploration. Similar conclusions can be made for the second PC. Here, Norway spruce tends to differentiate the most between the two geologies across all organs, indicating potential leverage for geological and mineral exploration. However, the geology and other external variables of the respective data set can vary heavily and other sources of variation like soil moisture, amount of till, fine fraction of the sample or physiological effects of the plants mentioned before can also be part of variation described by the PCs.

4.6 Conclusion

We introduce sparse PCA analysis for multiple related data sources to permit the detection of global as well as local, source-specific sparsity patterns in the PCA loadings. To this end, we propose an optimization problem that maximizes explained variances across the multiple data sources while inducing structured sparsity patterns. The ssMRCD estimator is used as plug-in into the optimization problem and perfectly fits the spirit of combined global-local sparsity patterns by delivering local covariances that are smoothed over groups. Moreover, it provides protection against the presence of outliers in the data.

We provide a computationally efficient algorithm based on the ADMM to obtain sparse outlier-robust PCA loadings. Algorithmic parameters are fine-tuned and convergence is achieved in all applications and simulations. Care is given to optimally select the hyperparameters controlling the degree of sparsity and smoothing properties of the ssMRCD estimator tailored to the PCA context. The proposed ssMRCD-PCA method is publicly available in the R-package `ssMRCD` (Puchhammer and Filzmoser, 2023).

The proposed sparse multi-source PCA method performs well in simulation settings mimicking structured sparsity and it outperforms non-robust counterparts as well as the state-of-the-art sparse, robust PCA method ROSPCA when outliers are present. The versatility of the multi-source method is illustrated on two different applications.

Possible further application scenarios entail also a wide variety of data, where the grouping structure is not fixed upfront. The flexibility of the method regarding the source-definition can also be leveraged for, e.g., the large field of spatial data. Finally, our multi-source perspective to sparse, outlier-robust PCA holds also promise for other popular multivariate analyses such as discriminant analysis, graphical modeling or canonical correlation analysis.

Acknowledgements

We sincerely thank our esteemed colleague, Solveig Pospiech, for her invaluable geological and biological expertise on plants, geology and geochemistry. Her generous contributions have greatly enhanced the precision and depth of our interpretations.

Data Availability

The data for Section 5.1 (weather data) is part of the supplementary files as well as publicly available under `weatherHoheWarte` in the R-package `ssMRCD` hosted on `CRAN`.

Appendix C

C.1 ADMM Minimization Problems

For ease of notation, we introduce the matrix notation of the vectorized ADMM components $\boldsymbol{u}_{(1)}, \boldsymbol{u}_{(2)}, \boldsymbol{u}_{(3)}$ and $\boldsymbol{v}_{(0)}, \boldsymbol{v}_{(1)}, \boldsymbol{v}_{(2)}$ and $\boldsymbol{v}_{(3)}$ similar to (4.2.1) as

$$\boldsymbol{U}^m_{(i)} = \begin{pmatrix} \boldsymbol{u}^m_{(i),1} & \dots & \boldsymbol{u}^m_{(i),(N-1)p+1} \\ \vdots & & \vdots \\ \boldsymbol{u}^m_{(i),p} & \dots & \boldsymbol{u}^m_{(i),Np} \end{pmatrix} = (\boldsymbol{u}^m_{(i),\cdot 1}, \dots, \boldsymbol{u}^m_{(i),\cdot N}) = (\boldsymbol{u}^{m'}_{(i),1\cdot}, \dots, \boldsymbol{u}^{m'}_{(i),p\cdot})',$$

$$\boldsymbol{V}^m_{(i)} = \begin{pmatrix} \boldsymbol{v}^m_{(i),1} & \dots & \boldsymbol{v}^m_{(i),(N-1)p+1} \\ \vdots & & \vdots \\ \boldsymbol{v}^m_{(i),p} & \dots & \boldsymbol{v}^m_{(i),Np} \end{pmatrix} = (\boldsymbol{v}^m_{(i),\cdot 1}, \dots, \boldsymbol{v}^m_{(i),\cdot N}) = (\boldsymbol{v}^{m'}_{(i),1\cdot}, \dots, \boldsymbol{v}^{m'}_{(i),p\cdot})'.$$

The notation for the variables without superscript is likewise as well as for the vector $\boldsymbol{z}$ used in Equation (4.3.5).

Minimization Problem 1

Due to the block-diagonal structure of $\hat{\boldsymbol{\Sigma}}$, the additivity of the quadratic Frobenius norm and the separable constraints, the minimization problem can be separated among sources. Thus, per source i, we have the following minimization problem in $\boldsymbol{v} \in \mathbb{R}^p$ for each iteration step m

$$\begin{aligned} \min_{v} \quad & -\boldsymbol{v}'\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} + \frac{\rho}{2}||\boldsymbol{v} + \underbrace{\frac{1}{\rho}\boldsymbol{u}^m_{(1),\cdot i} - \boldsymbol{v}^m_{(0),\cdot i}}_{=:\boldsymbol{c}}||_2^2 \\ s.t. \quad & \boldsymbol{v}'\boldsymbol{v} = 1, \\ & \boldsymbol{v}'\boldsymbol{v}^l_{\cdot i} = 0, \quad 1 \le l < k \\ & \boldsymbol{z}'_{\cdot i}\boldsymbol{v} \ge 0. \end{aligned}$$

The problem is non-convex but differentiable and, thus, can be solved by calculating the Lagrangian

$$\mathcal{L}(\boldsymbol{v}) = -\boldsymbol{v}'\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} + \frac{\rho}{2}||\boldsymbol{c} + \boldsymbol{v}||_2^2 - \mu\boldsymbol{z}'_{\cdot i}\boldsymbol{v} + \lambda_0(\boldsymbol{v}'\boldsymbol{v} - 1) + \sum_{l=1}^{k-1}\lambda_l(\boldsymbol{v}'\boldsymbol{v}^l_{\cdot i})$$

and applying the Karush-Kuhn-Tucker (KKT) conditions,

$$\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{v}) = -2\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} + \rho(\boldsymbol{c}+\boldsymbol{v}) - \mu\boldsymbol{z}_{\cdot i} + 2\lambda_0\boldsymbol{v} + \sum_{l=1}^{k-1}\lambda_l\boldsymbol{v}_{\cdot i}^l = \boldsymbol{0}, \tag{C.1}$$

$$\begin{aligned} g(\boldsymbol{v}) &= -\boldsymbol{z}_{\cdot i}'\boldsymbol{v} \leq 0, \\ h_0(v) &= \boldsymbol{v}'\boldsymbol{v} - 1 = 0, \\ h_l(v) &= \boldsymbol{v}'\boldsymbol{v}_{\cdot i}^l = 0, \quad \forall 1 \leq l < k, \\ \mu &\geq 0, \end{aligned} \tag{C.2}$$

$$\mu\boldsymbol{z}_{\cdot i}'\boldsymbol{v} = 0. \tag{C.3}$$

For speed we can derive a term for λ_0 by multiplying equation (C.1) from left with $\boldsymbol{v}'$, which cancels λ_l and μ due to the optimality conditions (C.3) and (C.2),

$$0 = -2\boldsymbol{v}'\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} + \rho\boldsymbol{v}'(\boldsymbol{c}+\boldsymbol{v}) - \underbrace{\mu\boldsymbol{z}_{\cdot i}'\boldsymbol{v}}_{=0} + 2\lambda_0\underbrace{\boldsymbol{v}'\boldsymbol{v}}_{=1} + \sum_{l=1}^{k-1}\lambda_l\underbrace{\boldsymbol{v}'\boldsymbol{v}_{\cdot i}^l}_{=0}$$

$$\lambda_0 = \boldsymbol{v}'\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - \frac{\rho}{2}\boldsymbol{v}'(\boldsymbol{c}+\boldsymbol{v}).$$

It is also possible to calculate λ_l as a function of μ and $\boldsymbol{v}$ by multiplying with $(\boldsymbol{v}_{\cdot i}^l)'$,

$$0 = -2(\boldsymbol{v}_{\cdot i}^l)'\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} + \rho\underbrace{(\boldsymbol{v}_{\cdot i}^l)'(\boldsymbol{c}+\boldsymbol{v})}_{=\bar{v_j}'\boldsymbol{c}} - \mu(\boldsymbol{v}_{\cdot i}^l)'\boldsymbol{z}_{\cdot i} + \underbrace{\sum_{l=1}^{k-1}\lambda_l(\boldsymbol{v}_{\cdot i}^l)'\boldsymbol{v}_{\cdot i}^l}_{=\lambda_l}$$

$$\lambda_l = 2(\boldsymbol{v}_{\cdot i}^l)'\hat{\boldsymbol{\Sigma}}_i\boldsymbol{v} - \rho(\boldsymbol{v}_{\cdot i}^l)'\boldsymbol{c} + \mu(\boldsymbol{v}_{\cdot i}^l)'\boldsymbol{z}_{\cdot i}.$$

However, substituting λ_l with the exact expression derived above has proven to deteriorate precision in the orthogonality constraints without a significant gain in speed.

It is not possible to derive an analytical solution due to third and higher-order terms after substituting the multiplier λ_0 into Equation (C.1). We have to resort to solving the root constraints (C.1), (C.2) and (C.3) numerically using the function `multiroot` from the R-package `rootSolve` (Soetaert, 2009). Additionally, we need to ensure that all other constraints are also fulfilled after finding a root. We apply the concept of *warm starts* and use $\boldsymbol{v}_{(0),\cdot i}^m$ as the starting value for the root finder. If no feasible root is found, we increase ρ until a feasible root is found.

Regarding regularity conditions, we can check the *linear independence constraint qualification* (LICQ) condition, where we need linear independence of all $\nabla h_l(\boldsymbol{v}) = \boldsymbol{v}_{\cdot i}^l$, $\nabla h_0(\boldsymbol{v}) = \boldsymbol{v}$ and $\nabla g(\boldsymbol{v}) = \boldsymbol{z}_{\cdot i}$ if $g(\boldsymbol{v}) = \boldsymbol{z}_{\cdot i}'\boldsymbol{v} = 0$. By design, $\nabla h_l(\boldsymbol{v})$ and $\nabla h_0(\boldsymbol{v})$ are independent since the components are all orthogonal. If $g(\boldsymbol{v}) = \boldsymbol{z}_{\cdot i}'\boldsymbol{v} = 0$, we are orthogonal to $\nabla h_0(\boldsymbol{v})$. Additionally choosing $\boldsymbol{z}_{\cdot i}$ orthogonal to all prior loadings, the regularity condition is fulfilled for all $\boldsymbol{v}$, implying that it is sufficient to look at points fulfilling the KKT conditions to find the optimum. Since for each source i $\boldsymbol{z}_{\cdot i}$ is chosen as the starting value $\boldsymbol{y}_{k,\cdot i}$ which is in the given feasible space and thus part of the orthogonality space of $\boldsymbol{v}_{\cdot i}^l$, the LICQ condition is fulfilled.

Minimization Problem 2

The objective function to minimize,

$$\eta\gamma||\boldsymbol{v}_{(2)}||_1 + \frac{\rho}{2}||\frac{1}{\rho}\boldsymbol{u}^m_{(2)} + \boldsymbol{v}_{(2)} - \boldsymbol{v}^m_{(0)}||^2_2,$$

is separable across sources due to the squared Frobenius norm and the L_1-norm. The analytical solution is thus given by the proximal operator of the L_1-norm, i.e. element-wise soft-thresholding (Boyd et al., 2011) for each entry of $\boldsymbol{v}_{(2)}$,

$$\boldsymbol{v}^{m+1}_{(2),i} = S\left(\boldsymbol{v}^m_{(0),i} - \boldsymbol{u}^m_{(2),i}/\rho, \eta\gamma/\rho\right) \qquad \forall i = 1, \ldots, Np,$$

with $S(x, \lambda) = \text{sign}(x)\max(|x| - \lambda, 0)$.

Minimization Problem 3

The part of the minimization function connected to the groupwise sparsity, $f_3(\boldsymbol{v}_{(3)})$, can be rewritten as

$$f_3(\boldsymbol{v}_{(3)}) = \eta(1-\gamma)\sqrt{N}\sum_{j=1}^{p}\sqrt{\boldsymbol{v}'_{(3)}\boldsymbol{C}_j\boldsymbol{v}_{(3)}} = \eta(1-\gamma)\sqrt{N}\sum_{j=1}^{p}||\boldsymbol{v}_{(3),j\cdot}||_2.$$

Thus, we can use the groupwise/block soft-thresholding operator, which is the proximal operator of the L_1-norm of subgroups (Boyd et al., 2011)

$$\boldsymbol{v}^{m+1}_{(3),j\cdot} = S_G(\boldsymbol{v}^m_{(0),j\cdot} - \boldsymbol{u}^m_{(3),j\cdot}/\rho, \eta(1-\gamma)\sqrt{N}/\rho),$$

with $S_G(\boldsymbol{x}, \lambda) = \max(1 - \lambda/||\boldsymbol{x}||_2, 0)\boldsymbol{x}$.

C.2 Starting Values

The extreme solution for $\eta \to \infty$ depends on γ, $\boldsymbol{y}^\infty_k(\gamma)$.

Proof of Corollary 4.3.1.1. a. First, we know for any $\boldsymbol{v}$

$$1 = ||\boldsymbol{v}_{\cdot i}||^2_2 = \sum_{j=1}^{p} v^2_{ji} \leq \sum_{j=1}^{p} v^2_{ji} + 2\sum_{j'<j}|v_{j'i}||v_{ji}| = ||\boldsymbol{v}_{\cdot i}||^2_1,$$

and that the proposed minimizer of $||\boldsymbol{v}_{\cdot i}||_1$ has the minimal objective function value of 1. Moreover, all other minimizer have to fulfill that $|v_{j'i}||v_{ji}| = 0$ for all $j' < j$ to reach the minimal objective function value of 1. Thus, all minimizers have exactly one entry unequal to zero per source.

b. Define $x_j = \sqrt{\sum_{i=1}^{N} v^2_{ji}} = ||\boldsymbol{v}_{j\cdot}||_2$. Then, based on the inequality of part a, it holds that

$$N = \sum_{i=1}^{N}||\boldsymbol{v}_{\cdot i}||^2_2 = \sum_{j=1}^{p}||\boldsymbol{v}_{j\cdot}||^2_2 = ||x||^2_2 \leq ||x||^2_1 = \left(\sum_{j=1}^{p}||\boldsymbol{v}_{j\cdot}||_2\right)^2 = G_2(\boldsymbol{v})^2.$$

The extreme solution proposed above has an objective function value of N, which is thus minimal. Again, the same argument applies as before implies, that all mixed terms $|x_{j'}||x_j|$ must be equal to zero for all $j' \neq j$ to reach equality of the two norms.

c. Trivial.

□

For the special but important case of correlation matrices, we need an adaptation since the variable chosen by the explained variance is not unique. In order to ensure consistent behavior, we calculate the k-th eigenvectors of each correlation matrix, scale it with the root of the respective eigenvalue, and take the mean. The variable with the absolute highest value will be taken as the groupwise solution for $\eta \to \infty$. Although all variables are valid solutions for $\eta \to \infty$, choosing an extreme solution close to $\boldsymbol{y}_k^0$, we get more consistency over varying λ and thus better convergence.

Simulation Results

In Figure C.1 the performance of the proposed starting value for simulation scenario 1 (Section 4.4.1) and the first four principal components is illustrated. We apply $\rho = p$, $\epsilon_{root} = 10^{-1}, \epsilon_{ADMM} = 10^{-4}, \epsilon_{thr} = 0.005$. For PCs 2 to 4 we iteratively use the best solution of all starting values for the orthogonality constraints. We simulate 100 random starting values, where each entry of the starting values is drawn from a standard normal distribution, and the vector is then projected onto the feasible space using the projection defined in Equation (4.3.6). The values of the objective function of the random starting values are shown as boxplots for varying $\eta \in [0, 2]$ (horizontal axis) and $\gamma = 0, 0.5, 1$ (panels). The crosses indicate the objective function value for the proposed starting value. It is clearly visible, that the proposed starting value reliably produces optimal solutions and is thus a valid alternative to using many random starting values.

In Figure C.2 the results for correlation matrices are shown. Regarding correlation matrices, there are multiple optimal extreme solutions, since all variables have the same amount of variance univariately. If there are multiple optimal solutions obtained in the simulations the one with the highest absolute scalar product with the extreme solution $\boldsymbol{y}_k^0$ is taken for the orthogonality constraints for the simulation for the next PC. This accounts for the fact that the extreme solutions $\boldsymbol{y}_k^\infty(\gamma)$ (as well as $\boldsymbol{y}_k^0$ for higher components) are based on the prior extreme solution components.

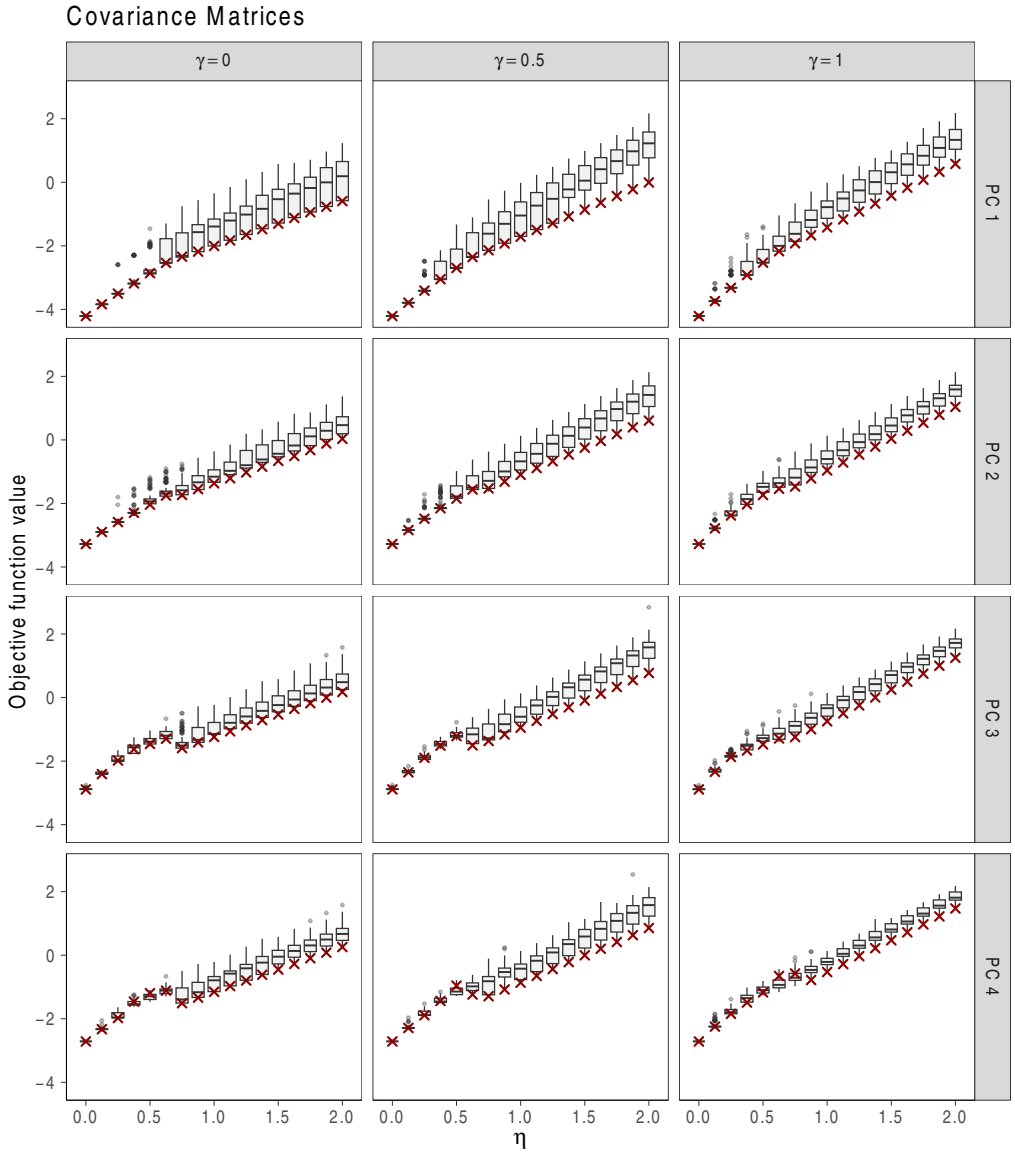


Figure C.1: Objective function values for the proposed starting value (cross) and for random starting values (boxplot) for covariance matrices.

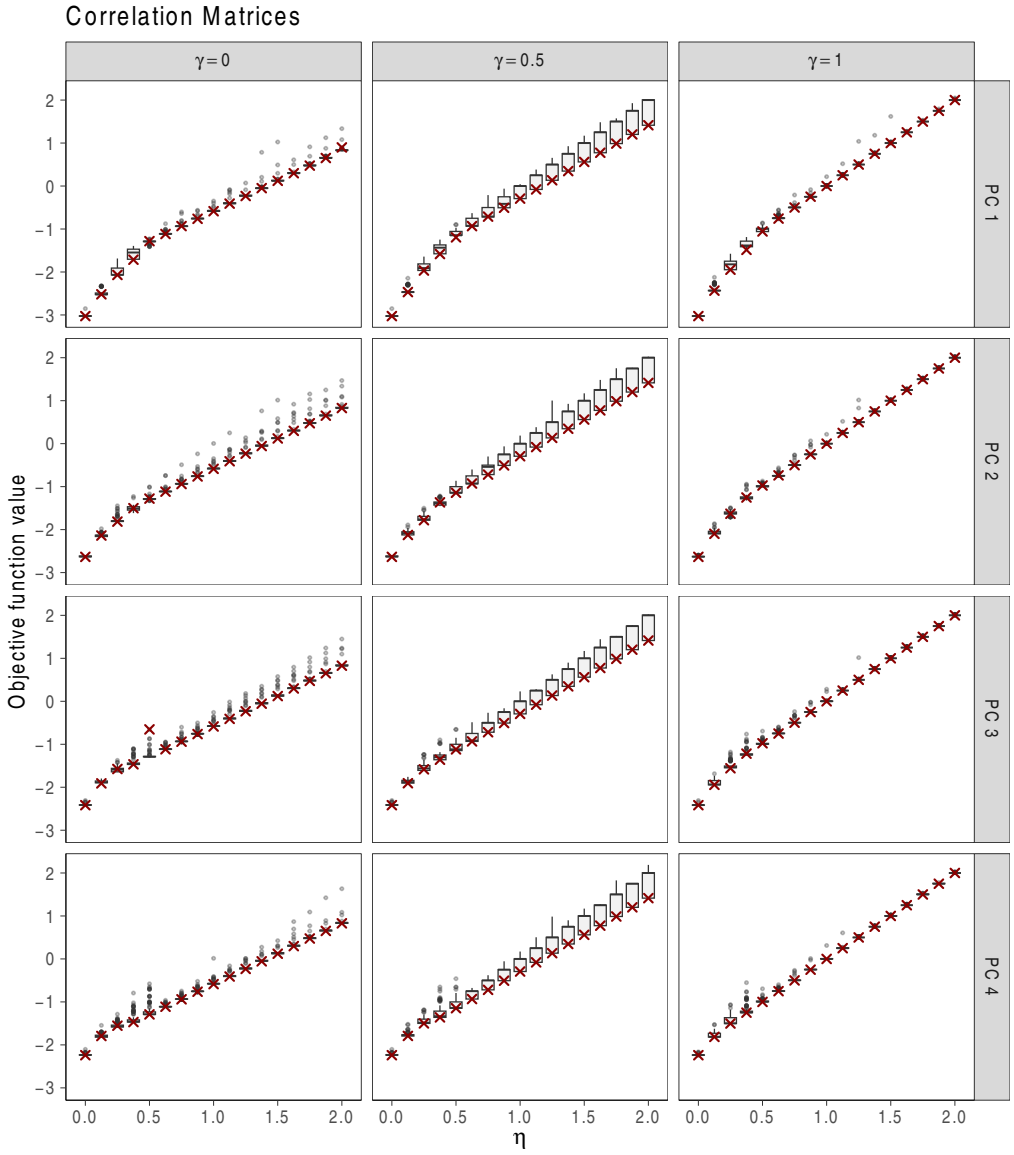


Figure C.2: Objective function values for the proposed starting value (cross) and for random starting values (boxplot) for correlation matrices.

C.3 Weather Analysis at Hohe Warte

The variables used in the real data example collected from the weather station Hohe Warte are listed and described in Table C.1.

Name	Description	Unit
cl	Cloud coverage, daily mean	$1, \ldots, 100$
rad	Global radiation, daily sum	J/cm^2
vp	Vapour pressure, daily mean	hPa
wmax	Maximal wind speed, daily maximum	m/s
ap	Air pressure, daily mean	hPa
hum	Relative air humidity, daily mean	%
prec	Precipitation, daily sum	mm
sight	Sight distance, sight at 1pm	m
sun	Sunshine duration, daily sum	h
tmax	Maximal air temperature at 2m, daily maximum	°C
tmin	Minimal air temperature at 2m, daily minimum	°C
t	Air temperature at 2m, daily mean	°C
w	Wind speed, daily mean	m/s

Table C.1: Hohe Warte weather station: List of variables.

C.4 Geochemical Plant Analysis

Figure C.3 shows the optimal parameter selection for the geochemical plant data set.

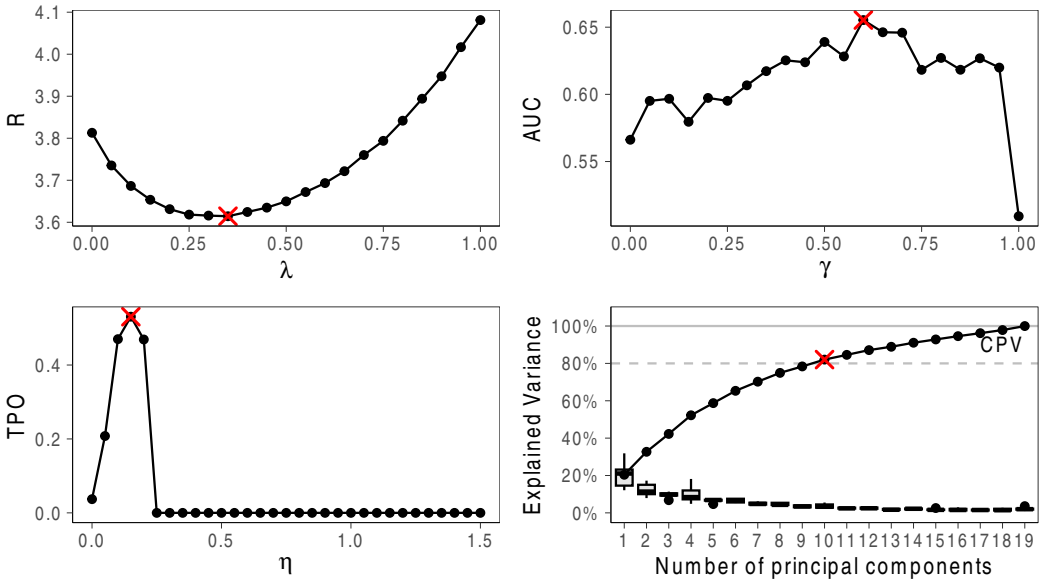


Figure C.3: Plant geochemistry: Optimal smoothing and sparsity parameters.

5 A Smooth Multi-Group Gaussian Mixture Model for Cellwise Robust Covariance Estimation

This chapter was published as Puchhammer, P., Wilms, I., and Filzmoser, P. (2025). A smooth multi-group Gaussian Mixture Model for cellwise robust covariance estimation. *arXiv preprint arXiv:2504.02547*. DOI: 10.48550/arXiv.2504.02547.

5.1 Introduction

The continuous increase in data volumes confronts statisticians with increasingly complex data structures. External information in addition to the measured features is often available and can be leveraged in the analysis. An example of external information are data with a partitioning of the observations into groups. This can be either a partition such as healthy persons and patients, but it could also be related to an expert grouping or to groups based on some hypothesis. However, in contrast to traditional classification tasks, the group information is considered uncertain to some extent, and thus the intended groups need more flexible modeling. Examples common in the medical context are progressive diseases, where patients are in transition from a healthy status towards more and more sever stages of a disease. Overall, groups cannot be dissociated from each other leading to a multi-group setting for the analysis.

Analyzing the groups separately might offer some insight, but overall trends or connections between groups would be lost or at least difficult to extract. On the other extreme, removing the grouping structure also poses analytical obstacles. Methodologies that assume identically distributed observations might fail because of the lack of coherency between the groups. Other approaches based on multiple distributions, such as mixture models or clustering methods, can deliver groups of data, however, they are not necessarily connected to the provided grouping and thus model something we might not be interested in. Therefore, more flexible models that can account for an underlying, possibly smooth connection among data groups defined by external information on a prior partition are needed to draw proper insights from data sets often present in real life.

There are many practical problem settings of this kind: When analyzing spatial data, as in the geosciences, underlying structures such as terrain type or country borders can dictate the grouping structure. Although the underlying basis are (continuous) spatial coordinates, the focus for the analysis still lies on the specifics of provided groups, but also on their common characteristics. The same applies to time-series data

structured by some fixed time interval, such as months or years, or by specific events. An important area where separation based on smooth external variables is common is medicine, where many diagnoses are based on continuous measurements with specific thresholds. An example is diabetes, where the diagnosis is based on measured blood sugar. Moreover, even if the diagnosis is not based on continuous external variables, most diseases are progressive, so measured features vary in a smooth way between people with different health conditions. Thus, taking the diagnosis classification as granted will not only lead to mistakes, but also misses information of persons being at a transition, as well as the reasons for this transition. The idea extends to many other fields, such as groups based on socio-economic status, or failure of components due to abrasion in industrial technology.

When it comes to real-life data, outliers are often present. Their effect on data analysis should be minimized to obtain robust and reliable results. Especially in settings with complex data structures, they can be masked more easily and can have a greater effect on the results if not detected. With multivariate data, outlying observations can be entirely different from the data majority, or they can just differ in single variables. The latter are called cellwise outliers, and methods were developed for their identification in one coherent data set, such as the *detecting deviating data cells* algorithm (DDC, Rousseeuw and Bossche, 2018), or the *cellMCD* estimator (Raymaekers and Rousseeuw, 2023) for cellwise robust covariance estimation. A cellwise robust version of a Gaussian mixture model was recently proposed by Zaccaria et al. (2024, cellGMM) – however, the method is limited to delivering the best clusters independent of prior information from the grouping structure.

We extend the setting of Gaussian mixture models (GMMs) to multi-group data sets to address the additional focus given by the pre-defined groups. Assuming that a smooth process underlies the partition into groups, we model each group having a main distribution and being mixed with distributions of other groups. This allows us to match the resulting distributions to the pre-defined groups and to put unusual observations into a bigger context. An observation can either be unusual in the original group and might fit better to another group, indicating a possible mismatch, or an observation is generally unusual because of possibly outlying cells. For a mismatch, it is worth checking the group assignment for errors. In case of outlying cells, these may refer to unreliable or extreme measurements that should either be corrected or removed for further analysis. By specifying the probabilities of group membership for each observation, we can also shed light on the transition mechanisms of observations moving from their predefined group to another one, and thus identify potentially influential variables during this transition.

The remainder of the paper is structured as follows. Section 5.2 provides more detailed information on the relevant literature, as well as an introduction to the model setup and the objective function. Section 5.3 details the algorithm and hyperparameter settings. Theoretical results on robustness properties are reported in Section 5.4, and experimental simulation results on robustness are described in Section 5.5. Three real-life data examples from meteorology, medicine and oenology, the science of wine and wine making, are illustrated in Section 5.6, and Section 5.7 concludes.

5.2 Methodology

We introduce the multi-group Gaussian mixture model in Section 5.2.1. The objective function based on the log-likelihood is proposed in Section 5.2.2, and finally connections and differences to related methods are discussed in Section 5.2.3.

5.2.1 Model and Notation

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$ be data sets from N groups consisting of independent observations $\boldsymbol{X}_g = ((\boldsymbol{x}_{g,1})', \ldots, (\boldsymbol{x}_{g,n_g})')' \in \mathbb{R}^{n_g \times p}$ per group $g = 1, \ldots, N$ of the same p variables. Let $n = \sum_{g=1}^{N} n_g$, and assume that observations $\boldsymbol{x}_{g,i}$ from group g, $i = 1, \ldots, n_g$, originate from a Gaussian mixture

$$\boldsymbol{x}_{g,i} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \text{ with probability } \pi_{g,k} \geq 0 \tag{5.2.1}$$

for $k = 1, \ldots, N$. Note that observations of a particular group can originate not only from a single distribution but from a Gaussian mixture of all group distributions. In the multi-group setting we assume that a pre-specified group is more coherent than the combined data, and thus it consists of a main distribution assigned to it. Therefore, we enforce $\pi_{g,g} \geq \alpha \geq 0.5$, where the constant α specifies how coherent each group should be.

Based on Equation (5.2.1) it follows that the expected value and the covariance of any $\boldsymbol{x}_g$ from group g are

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}_g] &= \sum_{k=1}^{N} \pi_{g,k} \boldsymbol{\mu}_k, \\ \operatorname{Cov}[\boldsymbol{x}_g] &= \sum_{k=1}^{N} \pi_{g,k} \boldsymbol{\Sigma}_k + \sum_{k=1}^{N} \pi_{g,k} (\boldsymbol{\mu}_k - \mathbb{E}[\boldsymbol{x}_g])(\boldsymbol{\mu}_k - \mathbb{E}[\boldsymbol{x}_g])', \end{aligned} \tag{5.2.2}$$

see Appendix D.1 for the derivation. The covariance corresponding to group g is then a smoothed covariance consisting of the covariance from the major distribution, $\boldsymbol{\Sigma}_g$, with a minimum weight of α, and of the other covariance matrices $\boldsymbol{\Sigma}_k$, with weights $\pi_{g,k}$ specifying the amount of overlap to other distributions as well as the variability of the means around the expected value.

In the following we define our notation used throughout the paper. The multivariate normal density with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$ of an observation $\boldsymbol{x}_{g,i}$ is denoted by

$$\varphi(\boldsymbol{x}_{g,i}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}_{g,i} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_{g,i} - \boldsymbol{\mu}_k)\right)}{\sqrt{(2\pi)^p \det \boldsymbol{\Sigma}_k}}.$$

Since outlying cells will be considered missing in the likelihood, observed and missing cells of $\boldsymbol{x}_{g,i}$ are denoted by a binary vector $\boldsymbol{w}_{g,i} = (w_{g,i1}, \ldots, w_{g,ip})$, where a value of 1 indicates observed variables, and 0 indicates missing or outlying values. We will put $(\boldsymbol{w}_{g,i})$ as superscript, as in $\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}$, $\boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})}$ and $\boldsymbol{\Sigma}_k^{(\boldsymbol{w}_{g,i})}$, if we only consider the subset of variables that are observed, i.e. $\{j : w_{g,ij} = 1, j = 1, \ldots, p\}$. Moreover, for any binary

vectors $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$, the notation $\boldsymbol{\Sigma}_k^{(\boldsymbol{w}|\tilde{\boldsymbol{w}})}$ denotes the submatrix of $\boldsymbol{\Sigma}_k$ that includes rows and columns indicated by $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$, respectively. Also, $(1-\boldsymbol{w})$ indicates missing cells instead of observed ones, $\{j : w_{g,ij} = 0, j = 1, \ldots, p\}$.

When considering the multivariate normal density $\varphi(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}; \boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{w}_{g,i})})$ of a partially observed observation, conventions regarding fully non-observed observations ($\boldsymbol{w}_{g,i} = \boldsymbol{0}$) are as follows. The density $\varphi(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}; \boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})}, \boldsymbol{\Sigma}_k^{(\boldsymbol{w}_{g,i})})$ and the covariance determinant $\det(\boldsymbol{\Sigma}_k^{(\boldsymbol{w}_{g,i})})$ are equal to 1, the squared Mahalanobis distance $(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})} - \boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})})'(\boldsymbol{\Sigma}_k^{(\boldsymbol{w}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})} - \boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})})$ is equal to zero.

5.2.2 Objective Function

For our cellwise robust estimation of the statistical model described above we denote the model parameters that need to be estimated as $\boldsymbol{\pi} = (\pi_{g,k})_{g,k=1}^N$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_k)_{k=1}^N$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_k)_{k=1}^N$, and their estimates as $\hat{\boldsymbol{\pi}} = (\hat{\pi}_{g,k})_{g,k=1}^N$, $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_k)_{k=1}^N$ and $\hat{\boldsymbol{\Sigma}} = (\hat{\boldsymbol{\Sigma}}_k)_{k=1}^N$.

Based on the proposed model in Equation (5.2.1) we use a likelihood approach to estimate the parameters. Robustness against cellwise outliers is achieved by considering outlying cells to be missing values indicated by a set of matrices $\boldsymbol{W} = (\boldsymbol{W}_g)_{g=1}^N$ consisting of binary vectors $\boldsymbol{w}_{g,i}, i = 1, \ldots, n_g$, which also need to be estimated, $\hat{\boldsymbol{W}} = (\hat{\boldsymbol{W}}_g)_{g=1}^N$. These missing values are removed from the likelihood estimation by using the observed likelihood.

For defining the objective function, the approach of the cellMCD (Raymaekers and Rousseeuw, 2023) is extended. We combine the observed log-likelihood for the model described in Equation (5.2.1) with a penalty term for the number of missing cells. The estimators are then the minimizers of the *observed penalized log-likelihood* $\mathrm{Obj}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{W})$, defined as

$$\sum_{g=1}^{N}\sum_{i=1}^{n_g}\left[-2\ln\left(\sum_{k=1}^{N}\pi_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}; \boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})}, \boldsymbol{\Sigma}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)\right) + \sum_{j=1}^{p} q_{g,ij}(1-w_{g,ij})\right] \tag{5.2.3}$$

subject to the constraints

$$\boldsymbol{\Sigma}_{reg,k} = (1-\rho_k)\boldsymbol{\Sigma}_k + \rho_k \boldsymbol{T}_k \tag{5.2.4}$$

$$\sum_{i=1}^{n_g} w_{g,ij} \geq h_g \qquad \forall j = 1, \ldots, p, \forall g = 1, \ldots, N \tag{5.2.5}$$

$$\sum_{k=1}^{N} \pi_{g,k} = 1 \qquad \forall g = 1, \ldots, N \tag{5.2.6}$$

$$\pi_{g,g} \geq \alpha \geq 0.5. \tag{5.2.7}$$

The first part of Equation (5.2.3) is the observed likelihood of each observation $\boldsymbol{x}_{g,i}$ given a missingness pattern $\boldsymbol{w}_{g,i}$. The second part introduces the penalty term to reduce the number of flagged cells and increases accuracy as also shown in Raymaekers and Rousseeuw (2023). Flagging a cell of an observation $x_{g,ij}$ costs a value of $q_{g,ij}$ in the

objective function. The penalty constant $q_{g,ij}$ is derived by the notion of a standardized residual. If the (absolute) residual is atypically large (measured by a χ^2-quantile), the minimizing effects on the likelihood exceed the additional cost flagging the cell. If the residual is too small, it will not be flagged and included in the estimation. In that way, only clearly outlying cells are flagged and overflagging is reduced. For more details on choosing $q_{g,ij}$, we refer to Section 5.3.4.

Regarding the constraints, Equation (5.2.4) provides regularization of the covariance matrices by a convex combination with a regular diagonal matrix $\boldsymbol{T}_k$ of univariate robust scale for group k, and a regularization factor $\rho_k > 0$, similar to the MRCD (Boudt et al., 2020). Regularity provides stability for grouped data settings, where groups can also consist of just a few observations, as well as for high-dimensional settings. The proposed values for ρ_k and $\boldsymbol{T}_k$ are described in more detail in Section 5.3.4.

The number of cells flagged per group and variable is constrained by Equation (5.2.5), where at least half of the cells per group need to be included in the parameter estimation of the mixture model, $h_g \geq \lceil 0.5 n_g \rceil$. However, due to the possible instability of the covariance estimation between two variables, we set the default value to $h_g = \lceil 0.75 n_g \rceil$ and thus allow for a maximum of 25% of flagged cells per variable and group.

Lastly, the two constraints in Equations (5.2.6) and (5.2.7) originate from the proposed multi-group GMM. The parameter α specifies how strict the model is regarding the pre-defined groups. A value of $\alpha = 1$ allows no group change of observations from their given groups. When α decreases, more and more flexibility among the groups is allowed. Therefore, a gradual increase in flexibility can illuminate observations located in the transition between groups.

5.2.3 Connections to Related Work

Our method combines elements of clustering via mixture models, robustness, missing data, and multi-group data analysis.

Regarding robustness, many methods exist for the rowwise setting, where an entire observation is considered an outlier (Maronna et al., 2019). A recent rise in methodologies is visible for the cellwise paradigm, introduced by Alqallaf et al. (2009), where single cells of an observation are considered outlying. Standard rowwise robust estimators of covariance and location are the Minimum Covariance Determinant (MCD; Rousseeuw, 1984, 1985) estimator, typically proposed for $n \geq 5p$ (with n the number of observations and p the number of variables), and its regularized version, the Minimum Regularized Covariance Determinant (MRCD; Boudt et al., 2020) estimator. Both search for a subset of observations that minimize the resulting sample covariance.

In the cellwise paradigm, the cellwise robust MCD (cellMCD; Raymaekers and Rousseeuw, 2023) is a recent proposal to extend the likelihood formulation of the MCD to the cellwise outlier setting, leveraging the idea that outlying cells can be considered to be missing values in the estimation procedure. The objective function of the cellMCD consists of the observed likelihood (Little and Rubin, 2019), where outlying cells are declared as missing, plus a penalty term reducing the number of flagged cells and thus increasing estimation accuracy. The objective function is then optimized in an iterative manner, switching between covariance and location estimation via an Expected

Maximization (EM) algorithm and updating flagged outlying cells. Again, $n \geq 5p$ is suggested. An alternative in high-dimensional settings is the covariance estimator of Öllerer and Croux (2015) based on pairwise correlations.

Regarding finite mixture models, rowwise robust proposals for standard GMMs (Neykov et al., 2007) were recently extended to cellwise robustness (cellGMM, Zaccaria et al., 2024). Similar to the cellMCD, the objective function consists of an observed likelihood incorporating the mixture model and a penalty term. However, due to the model structure, the penalty weights need to be estimated for each observation separately in the first step before the outliers can be flagged more accurately in the second step. While cellGMM is cellwise robust and allows for multiple distributions, it does not account for the pre-defined grouping structure and estimated clusters are not directly matched to the given groups.

One rowwise robust method that is applicable in the scenario described above is the spatially smoothed MRCD (ssMRCD) estimation proposed by Puchhammer and Filzmoser (2024). Originally developed for spatial data, it relies on predefined groups that are connected by a bigger picture, and in contrast to a standard GMM also provides a match between pre-defined groups and covariance and location estimates. However, the ssMRCD is not formulated as a mixture model, as it yields a covariance estimate for a group by incorporating overall and group-wise information, where the group contributions are pre-specified by weights. For achieving robustness, the ssMRCD estimator targets the determinant of specific covariance matrices, similar to MCD and MRCD.

Compared to the ssMRCD, there are certain advantages of the proposed probabilistic model-based approach when it comes to selecting hyperparameters. While the amount of smoothing and the smoothing weights need to be prespecified for the ssMRCD estimator, which correspond to the mixture weights in the specified mixture model, here these parameters can be estimated within the probabilistic model. Also the amount of flexibility (referred to as smoothing for the ssMRCD) is not a fixed parameter given to the model, but it can vary between groups and is only restricted by the hyperparameter α.

5.3 Algorithm

The algorithm for the multi-group GMM consists of two steps, iteratively minimizing the objective function over two sets of parameters, similar to Raymaekers and Rousseeuw (2023). The W-step minimizes over $\boldsymbol{W}$ and the Expectation Minimization (Maximization) (EM, Dempster et al., 1977; McLachlan and Krishnan, 2008) step minimizes over $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Especially the EM-step is adapted to the multi-group setting by accounting for constraint (5.2.7) and by regularizing the covariance, see Equation (5.2.4). Given initial starting values for the parameters described in Appendix D.3,we iteratively repeat the W-step and the EM-step until the estimated covariance matrices have converged. A pseudo code of the main algorithmic structure is given in Algorithm 3.

5.3.1 W-Step

The calculation of the $(\tau+1)$-th step is based on the estimated parameters in the τ-th step, $\hat{\boldsymbol{\pi}}^{\tau} = (\hat{\pi}^{\tau}_{g,k})^{N}_{g,k=1}$, $\hat{\boldsymbol{\mu}}^{\tau} = (\hat{\boldsymbol{\mu}}^{\tau}_{k})^{N}_{k=1}$, $\hat{\boldsymbol{\Sigma}}^{\tau} = (\hat{\boldsymbol{\Sigma}}^{\tau}_{k})^{N}_{k=1}$, $\hat{\boldsymbol{W}}^{\tau} = (\hat{\boldsymbol{W}}^{\tau}_{g})^{N}_{g=1}$. Here, we minimize the objective function Equation (5.2.3) corresponding to the parameter $\boldsymbol{W}$. For an estimate $\hat{\boldsymbol{W}}^{\tau}$, a copy $\tilde{\boldsymbol{W}}$ is defined and modified for each variable step by step to reduce the objective function value, starting with $j = 1$. Although the exact results depend on the order of the variables, Raymaekers and Rousseeuw (2023) have shown by simulations that this effect is small or even negligible.

Based on the fixed variable j, for each group g and observation i we calculate the difference in the objective function for including the cell in the estimation, $\tilde{w}_{g,ij} = 1$ $({}_1\tilde{\boldsymbol{w}}_{g,i})$ and flagging the cell, $\tilde{w}_{g,ij} = 0$ $({}_0\tilde{\boldsymbol{w}}_{g,i})$ while all other entries stay unmodified. Note that the results are order independent regarding groups or observations. Thus, the difference $\Delta_{g,ij}$ is

$$\begin{aligned}\Delta_{g,ij} = &-2\ln\left(\sum_{k=1}^{N}\hat{\pi}^{\tau}_{g,k}\varphi\left(\boldsymbol{x}^{({}_1\tilde{\boldsymbol{w}}_{g,i})}_{g,i};\hat{\boldsymbol{\mu}}^{\tau({}_1\tilde{\boldsymbol{w}}_{g,i})}_{k},\hat{\boldsymbol{\Sigma}}^{\tau}_{reg,k}{}^{({}_1\tilde{\boldsymbol{w}}_{g,i})}\right)\right)\\ &+2\ln\left(\sum_{k=1}^{N}\hat{\pi}^{\tau}_{g,k}\varphi\left(\boldsymbol{x}^{({}_0\tilde{\boldsymbol{w}}_{g,i})}_{g,i};\hat{\boldsymbol{\mu}}^{\tau({}_0\tilde{\boldsymbol{w}}_{g,i})}_{k},\hat{\boldsymbol{\Sigma}}^{\tau}_{reg,k}{}^{({}_0\tilde{\boldsymbol{w}}_{g,i})}\right)\right) - q_{g,ij}.\end{aligned}$$

For all observations with $\Delta_{g,ij} < 0$, we set $\tilde{w}_{g,ij}$ equal to 1 for further calculations. If there are less than h_g observations per group g with $\Delta_{g,ij} < 0$, we set those $\tilde{w}_{g,ij}$ equal to 1 for which $\Delta_{g,ij}$ is among the lowest h_g values of $\{\Delta_{g,ij} : i = 1,\ldots,n_g\}$. Then, the same procedure is applied to the next variable with the updated $\tilde{\boldsymbol{W}}$, until the flagging is updated for all variables. Overall, the updated $\tilde{\boldsymbol{W}}$ after all variables is the next estimate $\hat{\boldsymbol{W}}^{\tau+1}$. We always modify $\tilde{\boldsymbol{W}}$ such that the objective function is at least not increasing given the constraints, and thus the whole W-step does not increase the objective function value.

5.3.2 EM-Step

Given $\hat{\boldsymbol{W}}^{\tau+1}$, the parameters of the mixture model can be estimated to minimize the unpenalized observed likelihood of the GMM with missing values thus minimizing the overall objective function. Eirola et al. (2014) provide an EM-based algorithm for GMMs with missing data that will be adapted to the multi-group setting incorporating the additional constraints given by Equations (5.2.4) and (5.2.7). More details and derivations are provided in Appendix D.3.

The expected probability that observation $\boldsymbol{x}_{g,i}$ is from distribution k conditional on the observed values indicated by $\hat{\boldsymbol{w}}^{\tau+1}_{g,i}$ and on the previous estimates $\hat{\boldsymbol{\pi}}^{\tau} = (\hat{\pi}^{\tau}_{g,k})^{N}_{g,k=1}$, $\hat{\boldsymbol{\mu}}^{\tau} = (\hat{\boldsymbol{\mu}}^{\tau}_{k})^{N}_{k=1}$, $\hat{\boldsymbol{\Sigma}}^{\tau} = (\hat{\boldsymbol{\Sigma}}^{\tau}_{k})^{N}_{k=1}$, is

$$\hat{t}^{\tau+1}_{g,i,k} = \frac{\hat{\pi}^{\tau}_{g,k}\varphi\left(\boldsymbol{x}^{(\hat{\boldsymbol{w}}^{\tau+1}_{g,i})}_{g,i};\hat{\boldsymbol{\mu}}^{\tau(\hat{\boldsymbol{w}}^{\tau+1}_{g,i})}_{k},\hat{\boldsymbol{\Sigma}}^{\tau}_{reg,k}{}^{(\hat{\boldsymbol{w}}^{\tau+1}_{g,i})}\right)}{\sum_{l=1}^{N}\hat{\pi}^{\tau}_{g,l}\varphi\left(\boldsymbol{x}^{(\hat{\boldsymbol{w}}^{\tau+1}_{g,i})}_{g,i};\hat{\boldsymbol{\mu}}^{\tau(\hat{\boldsymbol{w}}^{\tau+1}_{g,i})}_{l},\hat{\boldsymbol{\Sigma}}^{\tau}_{reg,l}{}^{(\hat{\boldsymbol{w}}^{\tau+1}_{g,i})}\right)}. \tag{5.3.1}$$

Due the constraints in Equation (5.2.7) and (5.2.6), the mixture probability updates are adapted according to

$$\hat{\pi}_{g,g}^{\tau+1} = \max\left\{\alpha, \frac{1}{n_g}\sum_{i=1}^{n_g}\hat{t}_{g,i,g}^{\tau+1}\right\}, \quad \hat{\pi}_{g,k}^{\tau+1} = (1-\hat{\pi}_{g,g}^{\tau+1})\frac{\frac{1}{n_g}\sum_{i=1}^{n_g}\hat{t}_{g,i,k}^{\tau+1}}{1-\frac{1}{n_g}\sum_{i=1}^{n_g}\hat{t}_{g,i,g}^{\tau+1}}.$$

Further, for an observation $\boldsymbol{x}_{g,i}$ with current missingness pattern $\hat{\boldsymbol{w}}_{g,i}^{\tau+1}$, the conditional expectation $\hat{\boldsymbol{x}}_{g,i}^{\tau+1}$ assuming that $\boldsymbol{x}_{g,i}$ comes from distribution k is calculated by

$$\begin{aligned}\hat{\boldsymbol{x}}_{g,i}^{\tau+1\ (1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1})} &= \hat{\boldsymbol{\mu}}_k^{\tau(1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1})} + \hat{\boldsymbol{\Sigma}}_{reg,k}^{\tau\ (1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1}|\hat{\boldsymbol{w}}_{g,i}^{\tau+1})} \\ &\quad \times \left(\hat{\boldsymbol{\Sigma}}_{reg,k}^{\tau\ (\hat{\boldsymbol{w}}_{g,i}^{\tau+1}|\hat{\boldsymbol{w}}_{g,i}^{\tau+1})}\right)^{-1}\left(\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i}^{\tau+1})} - \hat{\boldsymbol{\mu}}_k^{\tau(\hat{\boldsymbol{w}}_{g,i}^{\tau+1})}\right) \qquad (5.3.2)\\ \hat{\boldsymbol{x}}_{g,i}^{\tau+1\ (\hat{\boldsymbol{w}}_{g,i}^{\tau+1})} &= \boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i}^{\tau+1})}. \qquad (5.3.3)\end{aligned}$$

The new estimate for $\hat{\boldsymbol{\mu}}_k^{\tau+1}$ is then

$$\hat{\boldsymbol{\mu}}_k^{\tau+1} = \frac{1}{\bar{t}_k}\sum_{g=1}^{N}\sum_{i=1}^{n_g}\hat{t}_{g,i,k}^{\tau+1}\hat{\boldsymbol{x}}_{g,i}^{\tau+1}$$

with $\bar{t}_k = \sum_{g=1}^{N}\sum_{i=1}^{n_g}\hat{t}_{g,i,k}^{\tau+1}$.

For estimating the covariance based on $\hat{\boldsymbol{x}}_{g,i}^{\tau+1}$, an additional term needs to be added. Assuming that observation $\boldsymbol{x}_{g,i}$ originates from distribution k, the correction term is calculated according to

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}_{reg,k}^{\tau\ (1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1}|1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1})} &= \hat{\boldsymbol{\Sigma}}_{reg,k}^{\tau\ (1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1}|1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1})} - \hat{\boldsymbol{\Sigma}}_{reg,k}^{\tau\ (1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1}|\hat{\boldsymbol{w}}_{g,i}^{\tau+1})} \\ &\quad \times \left(\hat{\boldsymbol{\Sigma}}_{reg,k}^{\tau\ (\hat{\boldsymbol{w}}_{g,i}^{\tau+1}|\hat{\boldsymbol{w}}_{g,i}^{\tau+1})}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{reg,k}^{\tau\ (\hat{\boldsymbol{w}}_{g,i}^{\tau+1}|1-\hat{\boldsymbol{w}}_{g,i}^{\tau+1})}\end{aligned}$$

for unobserved variables, $\hat{\boldsymbol{w}}_{g,i}^{\tau+1}$ equal to 0, and 1 otherwise. The new estimate $\hat{\boldsymbol{\Sigma}}_{reg,k}^{\tau+1}$ is then calculated as

$$\hat{\boldsymbol{\Sigma}}_{reg,k}^{\tau+1} = \rho_k\boldsymbol{T}_k + (1-\rho_k)\frac{1}{\bar{t}_k}\sum_{g=1}^{N}\sum_{i=1}^{n_g}\hat{t}_{g,i,k}^{\tau+1}\left[(\hat{\boldsymbol{x}}_{g,i}^{\tau+1}-\hat{\boldsymbol{\mu}}_k^{\tau+1})(\hat{\boldsymbol{x}}_{g,i}^{\tau+1}-\hat{\boldsymbol{\mu}}_k^{\tau+1})' + \tilde{\boldsymbol{\Sigma}}_{reg,k}^{\tau}\right].$$

5.3.3 Convergence of the Algorithm

The algorithm iterates between the W-step and the EM-step until the maximal absolute change in any entry of all covariance matrices, $\max_{k,j,j'}|\hat{\Sigma}_{reg,k,jj'}^{\tau} - \hat{\Sigma}_{reg,k,jj'}^{\tau+1}|$, is smaller than $\epsilon_{conv} = 10^{-4}$.

Since the regularization of the covariance matrices acts on the maximization step of the EM-algorithm, the same argumentation as in Proposition 6 from Raymaekers and Rousseeuw (2023) can be applied to show that each W-step and EM-step reduce the objective function or leaves it unchanged while all constraints are fulfilled. Thus, the algorithm converges to a local minimum.

5.3.4 Choice of Hyperparameters

In the objective function (5.2.3), the parameters ρ_k, $\boldsymbol{T}_k$ and $q_{g,ij}$ are used but not yet specified.

First, regarding the regularization, we choose a diagonal matrix $\boldsymbol{T}_k$ consisting of robust univariate scale estimates for observations from group k, $\boldsymbol{T}_k = \text{diag}(\hat{\sigma}_{k,1}, \ldots, \hat{\sigma}_{k,p})$. Here, we choose the univariate MCD estimator applied to each variable separately. For the amount of regularization we opt for a condition number of 100 for each covariance. However, due to multiple groups, this is not always possible since $\boldsymbol{T}_k$ could vary heavily and possibly already have a higher condition number for one specific k. Thus, the condition number to achieve for distribution k is $\kappa_k = \max(1.1 \operatorname{cond} \boldsymbol{T}_k, 100)$, where the factor 1.1 allows for multivariate data input if the condition number of $\boldsymbol{T}_k$ is high. Given the initial estimates $\hat{\boldsymbol{\Sigma}}_k^0$, the regularization factor ρ_k is chosen as small as possible and such that the condition number fulfills $\rho_k \boldsymbol{T}_k + (1-\rho_k)\hat{\boldsymbol{\Sigma}}_k^0 \leq \kappa_k$.

Second, the penalty weights $q_{g,ij}$ are chosen per observation and variable. In the cellMCD algorithm (Raymaekers and Rousseeuw, 2023), the weights only depend on the initial estimate of the conditional variance per variable j, and a cell is flagged if

$$\ln(C_{ij}) + \ln(2\pi) + (x_{ij} - \hat{x}_{ij})^2 / C_{ij} > q_j,$$

where $\hat{x}_{ij}$ and C_{ij} are conditional mean and variance of x_{ij} given the current estimates and observed cells for observation i. The penalty weight q_j is chosen as $q_j = \chi^2_{1,0.99} + \ln(2\pi) + \ln(C_{ij})$ such that cells are flagged if the standardized residuals exceed a χ^2-quantile,

$$\frac{(x_{ij} - \hat{x}_{ij})^2}{C_{ij}} > \chi^2_{1,0.99},$$

the 99-th quantile of the chi-square distribution with one degree of freedom.

In the multi-group GMM, the original distributions of the observations are not clear, and we first need an initial estimate to which distribution each observation belongs to. Given initial estimates $\hat{\boldsymbol{\pi}}^0$, $\hat{\boldsymbol{\mu}}^0$ and $\hat{\boldsymbol{\Sigma}}^0$, we can calculate the probabilities $\hat{t}^0_{g,i,k}$ according to Equation (5.3.1) and use a weighted penalty parameter for each observation,

$$q_{g,ij} = \chi^2_{1,0.99} + \ln(2\pi) + \sum_{k=1}^{N} \hat{t}^0_{g,i,k} \ln(C^0_{k,j}),$$

where $C^0_{k,j} = \frac{1}{(\hat{\boldsymbol{\Sigma}}^0_{reg,k})^{-1}_{jj}}$.

5.4 Robustness Properties

In this section, we introduce an extension of the additive breakdown point for cluster and finite mixture model settings to the cellwise paradigm. As common in these settings, the breakdown point is data dependent and in unfavorable constellations, a robust estimator can break down if even one point is added. Thus, often an idealized setting of well clustered data points is considered, introduced by Hennig (2004) for univariate

Algorithm 3 Multi-group GMM

Input: $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$; initial estimates $\hat{\boldsymbol{\Sigma}}^0_{reg}$, $\hat{\boldsymbol{\mu}}^0$, $\hat{\boldsymbol{\pi}}^0$, $\hat{\boldsymbol{W}}^0$; hyperparameters $q_{g,ij}$, $\boldsymbol{T}_k$, ρ_k, ϵ_{conv}, h_g, α

$\boldsymbol{W} \leftarrow \hat{\boldsymbol{W}}^0$
$(\boldsymbol{\Sigma}_{reg}, \boldsymbol{\mu}, \boldsymbol{\pi}) \leftarrow (\hat{\boldsymbol{\Sigma}}^0_{reg}, \hat{\boldsymbol{\mu}}^0, \hat{\boldsymbol{\pi}}^0)$
`crit` $\leftarrow \infty$
while `crit` $> \epsilon_{conv}$ **do**
    $\boldsymbol{\Sigma}^{prev}_{reg} \leftarrow \boldsymbol{\Sigma}_{reg}$
    $\boldsymbol{W} \leftarrow$ `wstep`$(\boldsymbol{X}, \boldsymbol{\Sigma}_{reg}, \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{W}, q_{g,ij}, h_g)$
    $(\boldsymbol{\Sigma}_{reg}, \boldsymbol{\mu}, \boldsymbol{\pi}) \leftarrow$ `emstep`$(\boldsymbol{X}, \boldsymbol{\Sigma}_{reg}, \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{W}, \boldsymbol{T}, \rho, \alpha)$
    `crit` $\leftarrow \max_{k,j,j'} |\Sigma^{prev}_{reg,k,jj'} - \Sigma_{reg,k,jj'}|$
end while
return $\boldsymbol{\Sigma}_{reg}, \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{W}$

and extended by Cuesta-Albertos et al. (2008) to multivariate data in the rowwise paradigm (described in Appendix D.2). In this section we transfer the idealized setting from the rowwise outlier paradigm to the notion of cellwise outliers (see Section 5.4.1) as well as to the complex grouped structure of the targeted data sets (see Section 5.4.2) and prove the corresponding breakdown point of the proposed estimator.

5.4.1 Cellwise Breakdown in an Idealized Scenario

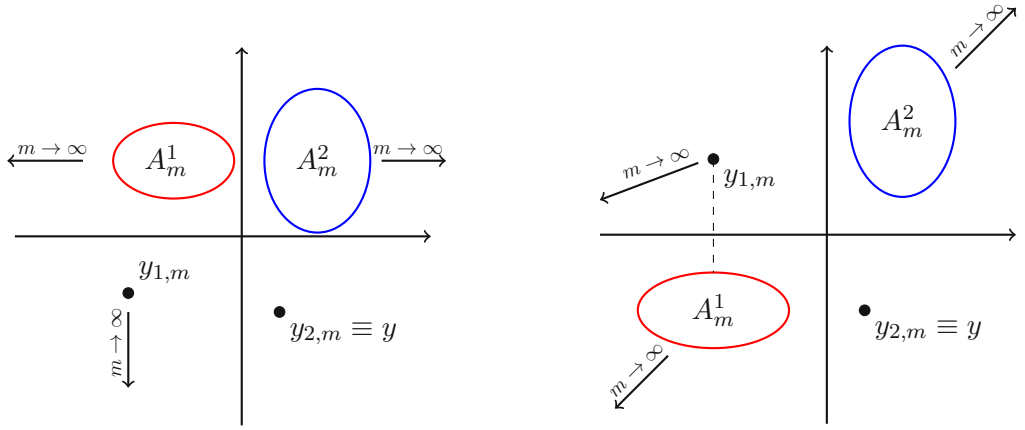
Compared to the well-known rowwise outliers, where an outlier is considered to be a whole observation, in the cellwise outlier paradigm introduced by Alqallaf et al. (2009), outliers are considered to be only single cells of observations. For the corresponding cellwise replacement breakdown point, only single cells are replaced by arbitrary values. The maximal fraction of contaminated cells per variable without breakdown of the estimator is then its breakdown point (Raymaekers and Rousseeuw, 2023).

When considering cellwise outlyingness in a mixture model setting, the scenario of well-clustered data used for the assessment of the breakdown behavior in the rowwise paradigm is not sufficiently separating the clusters when it comes to cellwise outlyingness. In the cellwise contamination scheme, the removal of a subset of variables could still lead to cluster overlap (see Figure 5.4.1a) and thus, the ideal scenario should be adapted to cluster separation in all subsets (see Figure 5.4.1b). Note that a separation in all variable subsets is equivalent to a separation in each variable.

To formalize well-separated clusters in the cellwise paradigm, a sequence of clusters $(\mathcal{X}_m)_{m\in\mathbb{N}}$ is considered ideal when the distances of observations within clusters are bounded by a constant $b < \infty$ and observations from different clusters are increasingly far away. Formally, let $s \geq 2$ be the number of clusters, and $\tilde{n}_1 < \tilde{n}_2 < \ldots < \tilde{n}_s = \tilde{n} \in \mathbb{N}$. For each m-th part of the sequence, the data $\mathcal{X}_m$ are clustered into s clusters $A^1_m, \ldots, A^s_m$ such that

$$A^1_m = \{\boldsymbol{x}_{1,m}, \ldots, \boldsymbol{x}_{\tilde{n}_1,m}\}, \ldots, A^s_m = \{\boldsymbol{x}_{\tilde{n}_{s-1}+1,m}, \ldots, \boldsymbol{x}_{\tilde{n}_s,m}\}$$

and $\mathcal{X}_m = \bigcup_{l=1}^s A^l_m$ and $\boldsymbol{x}_{i,m} = (x_{i1,m}, \ldots, x_{ip,m})$ for $i = 1, \ldots, \tilde{n}, m \in \mathbb{N}$.



(a) Not ideal in cellwise paradigm. Clusters A_m^1, A_m^2 and $y_{2,m}$ not separated vertically, $y_{1,m}$ and $y_{2,m}$ not separated horizontally.

(b) Ideal in cellwise paradigm ($w_{y_{1,m}} = (1,0)$, $w_{y_{2,m}} = \mathbf{0}$, $y_{1,m} \in B_m^1$, $y_{2,m}$ in any B_m^l). The dashed line for $y_{2,m}$ indicates bounded horizontal but increasing vertical distance.

Figure 5.4.1: Horizontally overlapping clusters in Figure a) and ideally separated clusters in the cellwise outlier paradigm in Figure b).

Thus, to ensure that clusters are well separated in each variable, we enforce

$$\lim_{m\to\infty} \min\{|x_{i'j,m} - x_{ij,m}| : \boldsymbol{x}_{i',m} \in A_m^l, \boldsymbol{x}_{i,m} \in A_m^h, h \neq l, j = 1,\dots,p\} = \infty. \quad (5.4.1)$$

Additionally, well-clustered also means that data points of each cluster are close to each other. Thus, a bounded distance within clusters in all variables separately is assumed,

$$\max_{1\le l\le s} \max\{|x_{i'j,m} - x_{ij,m}| : \boldsymbol{x}_{i',m}, \boldsymbol{x}_{i,m} \in A_m^l, j = 1,\dots,p\} < b \quad \forall m \in \mathbb{N}. \quad (5.4.2)$$

Note, that Equation (5.4.2) is equivalent to the corresponding assumption in the rowwise setting stated in Equation (D.3).

We now consider added cellwise outliers, $\mathcal{Y}_m = \{\boldsymbol{y}_{1,m},\dots,\boldsymbol{y}_{\tilde{r},m}\}$, such that $0 \le \tilde{r}_1 \le \dots \le \tilde{r}_s = \tilde{r}$ and

$$B_m^1 = \{\boldsymbol{y}_{1,m},\dots,\boldsymbol{y}_{\tilde{r}_1,m}\},\dots,B_m^s = \{\boldsymbol{y}_{\tilde{r}_{s-1}+1,m},\dots,\boldsymbol{y}_{\tilde{r}_s,m}\}.$$

For each added observation $\boldsymbol{y}_{i,m}$, there exists a $\boldsymbol{w}(\boldsymbol{y}_{i,m}) \in \{0,1\}^p$ indicating the outlying cells by $w(\boldsymbol{y}_{i,m})_j = 0$ and non-outlying cells by $w(\boldsymbol{y}_{i,m})_j = 1$. The non-outlying part of cellwise outliers should originate from one of the constructed clusters,

$$\max_{1\le l\le s} \max\{|y_{i'j,m} - x_{ij,m}| : \boldsymbol{x}_{i,m} \in A_m^l, \boldsymbol{y}_{i',m} \in B_m^l, \\ j = 1,\dots,p \text{ with } w(\boldsymbol{y}_{i',m})_j = 1\} < b \quad \forall m \in \mathbb{N},$$

and outlying cells should be infinitely far away from all other outlying cells and clusters,

$$\lim_{m\to\infty} \min\{|y_{i'j,m} - x_{ij,m}| : \boldsymbol{x}_{i,m} \in \mathcal{X}_m, \boldsymbol{y}_{i',m} \in \mathcal{Y}_m, w(\boldsymbol{y}_{i',m})_j = 0\} = \infty, \quad (5.4.3)$$

$$\lim_{m\to\infty} \min\{|y_{i'j,m} - y_{ij,m}| : \boldsymbol{y}_{i',m}, \boldsymbol{y}_{i,m} \in \mathcal{Y}_m, i \neq i', w(\boldsymbol{y}_{i',m})_j = 0\} = \infty. \quad (5.4.4)$$

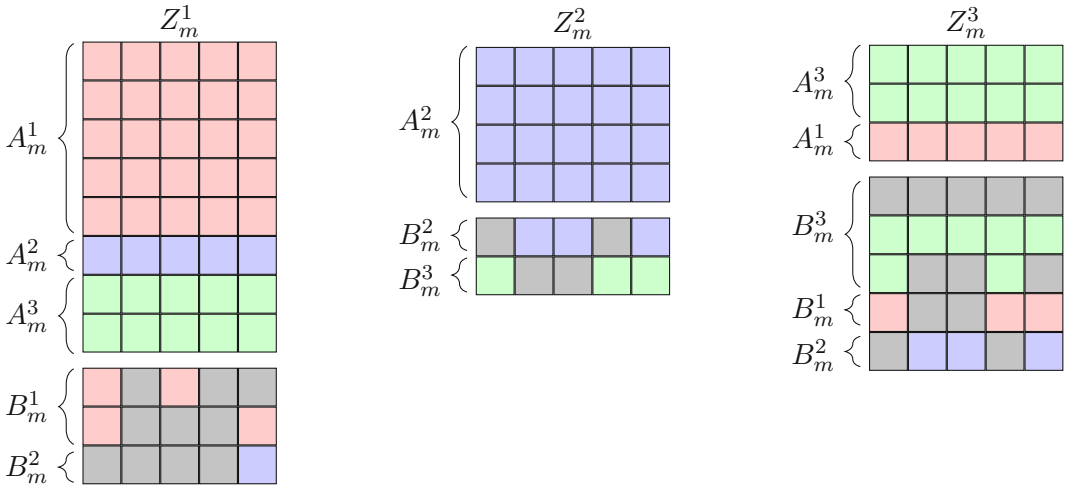


Figure 5.4.2: Possible group structure of groups for $N = 3$. Each column block corresponds to a group and each row within a column block to an observation. Red, violet and green rows are indicating from which cluster the observation originates from, gray indicates outlying cells. The gray line in the third block is assigned to B_m^3, but could stem from any other cluster too.

The breakdown of an estimator $\hat{E}$ of location, covariance or cluster weight is defined equivalently to the rowwise setting. Thus, the breakdown of an estimator is relatively defined by estimates based on $\mathcal{X}_m$ and on $\mathcal{X}_m \cup \mathcal{Y}_m$ and the location breakdown for a cluster l occurs, if for all $h = 1, \ldots, N$

$$||\hat{\boldsymbol{\mu}}_l(\mathcal{X}_m) - \hat{\boldsymbol{\mu}}_h(\mathcal{X}_m \cup \mathcal{Y}_m)||_2 \to \infty, \tag{5.4.5}$$

where $||\cdot||_2$ denotes the Euclidean norm. Denoting the smallest and largest eigenvalue of a covariance matrix with λ_p and λ_1, respectively, a covariance estimator of a cluster l would implode (explode) if $\lambda_p(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m)) \to 0$ ($\lambda_1(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m)) \to \infty$) and $\lambda_p(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m \cup \mathcal{Y}_m)) \nrightarrow 0$ ($\lambda_1(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m \cup \mathcal{Y}_m)) \nrightarrow \infty$) or vice versa. The weight estimator $\hat{\pi}_l$ of a cluster l breaks down if $\hat{\pi}_l \in \{0, 1\}$.

The cellwise additive breakdown point is then defined as

$$\epsilon^*(\hat{E}) = \min\left\{ \frac{\max_{j=1,\ldots,p} \sum_{i=1}^{\tilde{r}} (1 - w(\boldsymbol{y}_{i,m})_j)}{\tilde{n} + \tilde{r}} : \hat{E} \text{ breaks down} \right\},$$

where $\sum_{i=1}^{\tilde{r}} (1 - w(\boldsymbol{y}_{i,m})_j)$ denotes the number of contaminated cells per column j.

5.4.2 Cellwise Breakdown for Multi-Group Data

For analyzing the breakdown point in an ideal setting for a multi-group mixture model as described in Section 5.2.1, we assume N many underlying clusters and outliers constructed to be cellwise, separated as described in Section 5.4.1. All observations $\mathcal{X}_m \cup \mathcal{Y}_m$, contaminated or not, are partitioned into groups $\boldsymbol{Z}_m^1, \ldots, \boldsymbol{Z}_m^N$ of size $n_1 +$

$r_1, \ldots, n_N + r_N$ (where n_g is the number of clean and r_g is the number of added observations of group g) by a function $\tilde{g} : \mathcal{X}_m \bigcup \mathcal{Y}_m \to \{1, \ldots, N\}$, thus $\mathcal{Z}_m = \bigcup_{g=1}^{N} \boldsymbol{Z}_m^g = \mathcal{X}_m \bigcup \mathcal{Y}_m$. Moreover, we assume that for each group g a certain fraction $\tilde{\alpha}_g$ of its n_g observations and r_g added outliers are from cluster g,

$$\frac{|\{\boldsymbol{x} : \boldsymbol{x} \in A_m^g, \tilde{g}(\boldsymbol{x}) = g\}|}{n_g} \geq \tilde{\alpha}_g, \quad \frac{|\{\boldsymbol{y} : \boldsymbol{y} \in B_m^g, \tilde{g}(\boldsymbol{y}) = g\}|}{r_g} \geq \tilde{\alpha}_g, \tag{5.4.6}$$

thus, reflecting the major distribution per group. An illustration of the groups and the cluster origins per observation for a fictitious ideal data set is shown in Figure 5.4.2, where each row corresponds to an observation, each column block corresponds to a group and each column per group to a variable. The first (row) block per group includes the clean data, and the second block the added, possibly contaminated data. The color indicates the ideal cluster each observation is originating from (red, green, violet) for clean cells or whether a cell is outlying (grey). For each group the majority of observations comes from the main cluster for clean and for contaminated observations, respectively. Cellwise contamination can affect single cells (group 2), all cells of single variables (group 1, variable 2 and 4) and/or whole observations (group 3, first contaminated row).

For the ideal scenario we assume that at least $\left\lceil \frac{n_g + r_g + 1}{2} \right\rceil$ observations from group g are from cluster g and thus, $\tilde{\alpha}_g$ is restricted to fulfill $(n_g + r_g)\tilde{\alpha}_g \geq \left\lceil \frac{n_g + r_g + 1}{2} \right\rceil$ for all $g = 1, \ldots, N$. Note, for the proposed estimation this implies that for any variable j and group g there always exists at least one observation in $\boldsymbol{Z}_m^g$ originating from cluster g which is observed for variable j.

Cellwise breakdown is defined equivalently to the ungrouped setting and the breakdown point is defined as the minimal fraction of outlying cells for at least one variable in at least one group necessary to break down one estimator $\hat{E}$,

$$\epsilon^*_{group}(\hat{E}) = \min_{g=1,\ldots,N} \min \left\{ \frac{\max_{j=1,\ldots,p} \sum_{\boldsymbol{y} \in \boldsymbol{Z}_m^g \cap \mathcal{Y}_m} (1 - w(\boldsymbol{y})_j)}{n_g + r_g} : \hat{E} \text{ breaks down} \right\}.$$

Corollary 5.4.2.1. *Given the ideal setting and fixed $\rho_k > 0, \boldsymbol{T}_k > 0$ (positive definite), the following statements hold.*

a. For all m and no contamination, $\mathcal{Z}_m = \mathcal{X}_m$, there exist feasible estimates $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ such that the objective function is finite for any feasible set of $\boldsymbol{W}$ in Equation (5.2.5)*. Thus, the value of the objective function for a minimizer of Equation* (5.2.3) *under the constraints* (5.2.4) *to* (5.2.7) *is bounded.*

b. Given the contaminated data $\mathcal{Z}_m$ and sets of estimates $\hat{\boldsymbol{\pi}}(\mathcal{Z}_m)$, $\hat{\boldsymbol{\mu}}(\mathcal{Z}_m)$, $\hat{\boldsymbol{\Sigma}}(\mathcal{Z}_m)$, $\hat{\boldsymbol{W}}(\mathcal{Z}_m)$ for $m \in \mathbb{N}$. If there exists an l such that $\lambda_1(\hat{\boldsymbol{\Sigma}}_{reg,l}(\mathcal{Z}_m)) \to \infty$ for $m \to \infty$, then the value of the objective function of the estimates goes to infinity.

c. Given the contaminated data $\mathcal{Z}_m$ and sets of estimates $\hat{\boldsymbol{\pi}}(\mathcal{Z}_m)$, $\hat{\boldsymbol{\mu}}(\mathcal{Z}_m)$, $\hat{\boldsymbol{\Sigma}}(\mathcal{Z}_m)$, $\hat{\boldsymbol{W}}(\mathcal{Z}_m)$ for $m \in \mathbb{N}$. If there exists a variable j^*, l, k and a constant $\tilde{b}$ such that $|\hat{\mu}_{k,j^*}(\mathcal{Z}_m) - \hat{\mu}_{l,j^*}(\mathcal{Z}_m)| < \tilde{b}$ for $l \neq k$, then the objective function of these estimates goes to infinity.

The proof leverages the ideal scenario and subsequent intuition about reasonable estimates to bound the objective function in the uncontaminated case and to further show that an observation cannot “escape" from one cluster to another if it is originating from an exploding cluster since clusters move apart from each other. It is given in Appendix D.2.

Theorem 5.4.2.1 (Breakdown point)**.** *For the ideal scenario and fixed $\rho_k, \boldsymbol{T}_k > 0$ the following breakdown results in the cellwise paradigm hold.*

a. The implosion breakdown point is 1*.*

b. The weight breakdown point is 1*.*

c. The explosion breakdown point is at least $\min_g\{(n_g - h_g + 1)/n_g\}$*.*

d. The location breakdown point is 0.

e. The explosion breakdown point is exactly $\min_g\{(n_g - h_g + 1)/n_g\}$*, when assuming that the location estimator is not broken down.*

The proof leverages the strong cellwise separation between the clusters and the results of Corollary 5.4.2.1 and is given in Appendix D.2.

5.5 Simulations

In order to test the proposed method, we focus on five main scenarios: 1) a basic setting with $N = 2$ balanced groups, 2) a balanced setting with $N = 5$ groups, 3) an unbalanced two-group setting, 4) a balanced two-group setting with increasing singularity issues, and 5) a high-dimensional balanced two-group setting. Setting 1) and 2) are described in detail in the main text; for the remaining settings and further detailed evaluations we refer to Appendix D.4.

In Section 5.5.1 the generation of clean and contaminated data for two covariance structures is described in detail. Competing methods and evaluation criteria are summarized in Section 5.5.2 and 5.5.3, respectively, and corresponding results are shown in Section 5.5.4.

5.5.1 Data Generation

Clean data are generated according to the underlying multi-group Gaussian mixture model, formulated in Equation (5.2.1), for given dimensions $p \in \{10, 20, 60\}$. For $N \in \{2, 5\}$ groups we vary the mixture between the groups indicated by the parameter $\pi_{diag} \in \{0.75, 0.9\}$. The mixture probabilities are then given by $\pi_{gg} = \pi_{diag}$ and $\pi_{g,k} = \frac{1-\pi_{diag}}{N-1}$ for $g, k = 1, \ldots, N, g \neq k$.

We differentiate between two different covariance structures applied to all covariances in the mixture distributions. The first type is of Toeplitz structure (similar to Raymaekers and Rousseeuw, 2023) and each covariance $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ is constructed by $\boldsymbol{\Sigma}_{k,ij} = \zeta_k^{|i-j|}$ where ζ_k is randomly drawn from a uniform distribution in $[0.5, 1]$.

Toeplitz covariances share the relationships between variables but to a different extent. The second type is based on the approach of Agostinelli et al. (2015) (ALYZ) to construct well-conditioned correlation matrices. We allow for more variation of the variances and stop the iterative procedure early, specifically when the trace of a covariance is bounded by $[p/2, 2p]$. Compared to the Toeplitz structure, here the correlation between the variables can vary more strongly between the groups, making it more difficult for local methods to account for outliers.

Two types of scenarios are discussed for the mean of the distributions. On the one hand, we consider a scenario where there are just differences in the covariance, thus setting all means to zero, $\boldsymbol{\mu}_k = \mathbf{0}$. On the other hand, the more realistic scenario with different means is considered, by applying the concept of c-separation (Dasgupta, 1999) that gives a notion of how strongly the distributions overlap. We assume significant overlap (0.5-separated clusters) due to an underlying smooth variable and construct the means inductively, starting with $\boldsymbol{\mu}_1 = \mathbf{0}_p$. Given $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k-1}$ a new vector $\boldsymbol{\mu}_{tmp}$ is drawn from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p)$. To ensure a certain level of separation and overlap we set the next distributional mean to $\boldsymbol{\mu}_k = t^*(\boldsymbol{\mu}_{tmp} - \frac{1}{k-1}\sum_{l=1}^{k-1}\boldsymbol{\mu}_l) + \frac{1}{k-1}\sum_{l=1}^{k-1}\boldsymbol{\mu}_l$, where t^* fulfills

$$||\boldsymbol{\mu}_l - \boldsymbol{\mu}_k||_2 \geq 0.5\sqrt{p\max(\lambda_1(\boldsymbol{\Sigma}_l), \lambda_1(\boldsymbol{\Sigma}_k))}$$

for all $l = 1, \ldots, k-1$, with equality for at least one l. Each group g consists of $n_g \in \{30, 40, 50, 100\}$ many clean observations drawn with probability $\pi_{g,k}$ from $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

For each group a percentage $\epsilon_{cell} = 10\%$ of random cells per variable is contaminated as in Raymaekers and Rousseeuw (2023). Given an observation from group g which is drawn from distribution k and where a subset of variables indexed with $\mathcal{J}$ should be contaminated, cells indexed by $\mathcal{J}$ are replaced with

$$\boldsymbol{\mu}_{k,\mathcal{J}} + \boldsymbol{v}_{k,\mathcal{J}}\frac{\gamma_{cell}\sqrt{|\mathcal{J}|}}{\sqrt{\boldsymbol{v}'_{k,\mathcal{J}}\boldsymbol{\Sigma}^{-1}_{k,\mathcal{J}}\boldsymbol{v}_{k,\mathcal{J}}}}.$$

Here, $\mathcal{J}$ as subscript denotes the part of the vectors/matrices corresponding to the indexed variables, and $\boldsymbol{v}_{k,\mathcal{J}}$ denotes the eigenvector with the smallest eigenvalue of $\boldsymbol{\Sigma}_{k,\mathcal{J}}$. The parameter $\gamma_{cell} \in \{2, 6, 10\}$ controls the strength of the outlyingness of contaminated cells with respect to $\boldsymbol{\mu}_k$. For $\gamma_{cell} = 2$ the cellwise outliers are hard to distinguish from regular cells, while $\gamma_{cell} = 10$ produces clear outliers which are easier to detect for robust methods, and very influential to non-robust procedures.

5.5.2 Competing Methods

Regarding the performance comparison of our proposed method, we include the following seven methods in our simulation study, starting with their acronyms.

cellgGMM: The proposed cellwise robust multi-group GMM.

sample: The sample covariance applied to each group separately as a non-robust alternative.

mclust: A non-robust basic finite GMM implemented via an EM-algorithm in the R-Package `mclust` (Fraley et al., 2024) applied globally, with the correct number of groups provided. Since there is no clear attribution of an estimated cluster to a group, mclust will only be calculated for two-group settings and clusters will be assigned to groups in the most favorable way[1].

MRCD (Boudt et al., 2020): Rowwise robust covariance estimator applicable to high dimensions and applied separately to each group. It is available in the R-package `rrcov` (Todorov, 2024).

ssMRCD (Puchhammer and Filzmoser, 2024): An estimator targeted towards a multi-group setting robust against rowwise contamination available in the R-package `ssMRCD` (Puchhammer and Filzmoser, 2023). It is calculated with the default values for smoothing and equal weights for all groups, and the unsmoothed covariance estimates are assumed to correspond to the covariance matrices of the mixture distribution.

cellMCD (Raymaekers and Rousseeuw, 2023): A cellwise robust method for covariance and location available in the R-package `cellWise`(Raymaekers et al., 2023).

OC (Öllerer and Croux, 2015): The cellwise robust covariance estimator is applied separately to each group. The OC-estimator does not provide a location estimate but it can calculate a covariance matrix in high-dimensional settings. A fast implementation is available in the R-package Filzmoser et al. (2009).

5.5.3 Evaluation Criteria

The performance of covariance estimation is compared across all methods, where possible. Given an estimated covariance $\hat{\boldsymbol{\Sigma}}_k$, the Kullback-Leibler divergence to the real covariance $\boldsymbol{\Sigma}_k$ is used as evaluation criterion,

$$KL(\hat{\boldsymbol{\Sigma}}_k, \boldsymbol{\Sigma}_k) = \operatorname{tr}(\hat{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_k^{-1}) - p - \log\det(\hat{\boldsymbol{\Sigma}}_k \boldsymbol{\Sigma}_k^{-1}).$$

For $N \geq 2$, the final performance metric is the average over all distributions, $KL = \frac{1}{N}\sum_{k=1}^{N} KL(\hat{\boldsymbol{\Sigma}}_k, \boldsymbol{\Sigma}_k)$.

The mean estimates $\hat{\boldsymbol{\mu}}_k$ and the mixture probabilities $\hat{\boldsymbol{\pi}}$ are evaluated by the Mean Squared Error (MSE)

$$MSE(\hat{\boldsymbol{\mu}}_k, \boldsymbol{\mu}_k) = \frac{1}{p}\sum_{j=1}^{p}(\mu_{kj} - \hat{\mu}_{kj})^2,$$

$$MSE(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}) = \frac{1}{N^2}\sum_{g=1}^{N}\sum_{k=1}^{N}(\pi_{g,k} - \hat{\pi}_{g,k})^2$$

[1]The assignment of groups and clusters is such that it minimizes the evaluation measure of the KL-divergence. Thus, it is possible that the performance of estimating locations might suffer for the considered performance criteria.

and averaged over the groups for the mean, $MSE(\mu) = \frac{1}{N}\sum_{k=1}^{N} MSE(\hat{\boldsymbol{\mu}}_k, \boldsymbol{\mu}_k)$.

Additionally, the correctness of flagged cellwise outliers is measured by the standard recall, precision and F1-score and compared only to the cellMCD, since this is the only other method providing flagged cells.

5.5.4 Results

As introduced at the beginning of Section 5.5, we focus on two out of the five different settings in the main text, and additional figures regarding the MSE for location and mixture probabilities as well as outlier detection performance are included in Appendix D.4. Each combination is repeated 100 times. Note that cellMCD cannot be calculated if too many marginal outliers are present, in which case the failed runs are removed for all methods reducing the number of repetitions shown in the plots (see Appendix D.4 for corresponding tables stating the number of effective runs).

We start with the basic balanced setting where we consider $p = 10$ variables, $N = 2$ groups and $n_1 = n_2 = 100$ observations per group. Figure 5.5.1 and 5.5.2 show the KL-divergence for covariance estimation across all seven competing methods and a varying strength of outlyingness γ_{cell} for the Toeplitz and ALYZ covariance structure, respectively. The four subpanels differ regarding the coherency in the predefined groups. For example, observations of one group are very coherent for $\pi_{diag} = 0.9$ and $\mu = 0$ (top right panel) or less coherent for $\pi_{diag} = 0.75$ and varying μ. For both covariance structures and among all four coherency types it is visible that only the cellwise robust methods can manage outlying cells as γ_{cell} increases. Our proposed method cellgGMM and cellMCD are the most reliable while OC local is somehow robust against an increase in the degree of outlyingness of cells. However, OC local starts already with suboptimal estimates for $\gamma_{cell} = 2$. At the bottom panels it is evident that differences in location, even for strong overlapping distributions like here, is sufficient to drastically decrease performance for all competitor methods regarding covariance estimation. Especially for the cellMCD, non-coherency in the mean and covariance structures (ALYZ structure) confuse the algorithm in detecting cells and precision deteriorates (see also Figure D.2 and D.4 in the appendix) while for the proposed cellgGMM it facilitates the correct clustering (see also Figure D.1 and D.3).

In the setting with an extended number of $N = 5$ groups, $p = 10$ variables and $n_1 = \ldots = n_5 = 100$ observations per group, we see similar and even more prominent patterns. In Figure 5.5.3, the KL-divergence for the ALYZ covariance structure[2] is shown. Again, methods that are not robust against cellwise outliers suffer increasingly with the degree of outlyingness when it comes to covariance estimation. While for varying μ, the findings are the same as in the basic setting, we see that here cellgGMM performs better than cellMCD even in the most coherent setting (top right panel). Thus, the more groups are available to our proposed method, the better it can leverage the given context.

[2]Due to the difficulties of the cellMCD based on the amount of marginal outliers, some parameter combinations for the Toeplitz-structured covariances lead to a very low number of repetitions (down to 16). Thus, corresponding results are stated in the appendix and should be treated with caution.

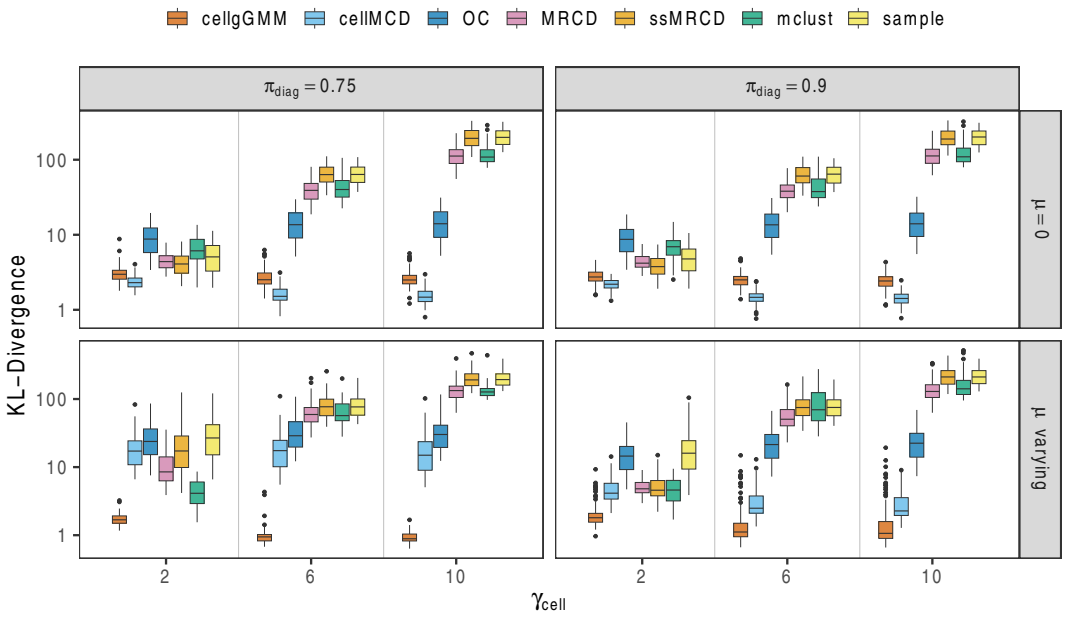


Figure 5.5.1: KL-divergence for the basic balanced setting and Toeplitz covariance structure for varying strength γ of outlyingness.

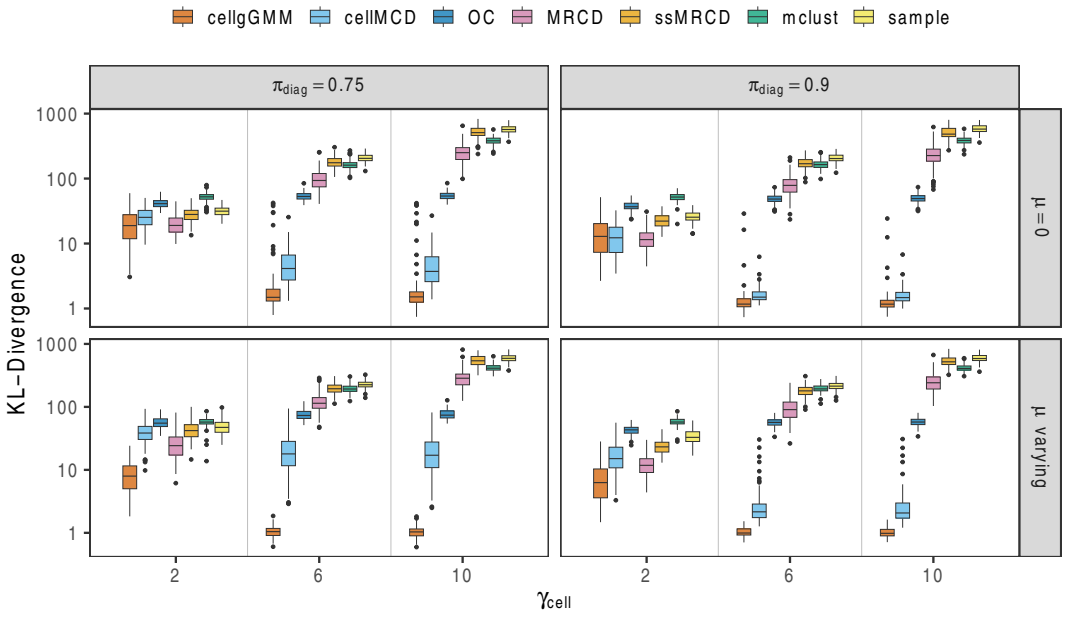


Figure 5.5.2: KL-divergence for the basic balanced setting and ALYZ covariance structure for varying strength γ of outlyingness.

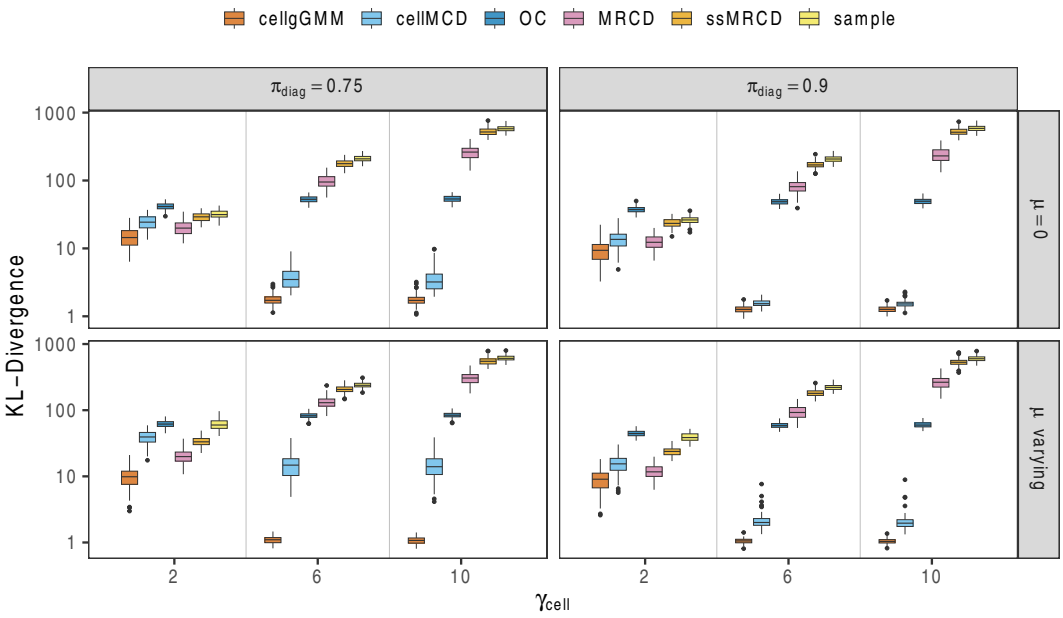


Figure 5.5.3: KL-divergence for the balanced setting with five groups and ALYZ covariance structure for varying strength γ of outlyingness.

With respect to the other three settings, the findings are similarly good for the proposed cellgGMM. The results of the competing methods in the unbalanced setting with $N = 2, p = 10, n_1 = 100$ and $n_2 = 50$ are comparable to the balanced settings described above. When increasing the p-to-n-ratio ($N = 2$, $p = 20$, $n_1 = n_2 = 30$), we see that cellMCD struggles a lot with flagging cellwise outliers due to low precision and subsequently with covariance estimation, often delivering worse covariance estimates than the OC local method. In the high dimensional scenario ($N = 2$, $p = 60$, $n_1 = n_2 = 40$) the results depend on the covariance structure. For the Toeplitz structure, OC local performs comparably well, while for ALYZ-structured covariances, cellgGMM generally outperforms OC local more clearly.

In general, cellgGMM consistently performs well in all five settings considered and in multiple coherency constellations. While it is often comparable to cellMCD when $\mu = 0$, in real multi-group settings this is a rare exception and one has to consider real life data to be closer to settings where locations vary over groups. In these simulation scenarios, cellgGMM outperforms all other considered methods.

5.6 Applications

We illustrate possible application scenarios of the proposed method by data from the fields meteorology, medicine and oenology. Weather measurements of Austrian weather stations are analyzed in Section 5.6.1, and in Section 5.6.2 we investigate handwriting data of healthy and Alzheimer patients. In the third application in Section 5.6.3 we

analyze patterns of high to low rated wine samples.

5.6.1 Austrian Weather Data

We illustrate our method on data provided by GeoSphere Austria (2024), with $p = 6$ monthly measured weather variables at 183 Austrian weather stations, including air pressure (p) and temperature (t), amount of rain (rsum), relative humidity (rel), hours of sunshine (s) and wind velocity (vv), which are averaged over the year 2021. The data set is publicly available in the R-Package `ssMRCD` (Puchhammer and Filzmoser, 2023) under the name `weatherAUT2021` on CRAN. Figure 5.6.1 shows the spatial locations and the underlying diverse geographical and thus also meteorological structure caused by the Alps. We proposed a separation of the stations into $N = 5$ more coherent groups, visible by the dashed lines in the figure. The most western area (group 1, $n_1 = 31$) is characterized by very mountainous terrain, which extends to the east into the next area (group 2, $n_2 = 80$), where high and low mountains are present. The most northern part (group 3, $n_3 = 35$) consists of low mountains and hills along the Danube river which flows through Vienna and the Vienna Basin (group 5, $n_5 = 21$). The last area to the East (group 4, $n_4 = 16$) hosts some hills but is mainly flat.

Our goal is to identify weather stations with cellwise outliers given the spatial context and to further analyze why these stations are atypical. Moreover, we are also interested in the coherency of the pre-defined groups. To this end, we apply our method with default values $h_g = 0.75n_g$, allowing for up to 25% of flagged cells per variable, and $\alpha = 0.5$, indicating a strong flexibility of observations to switch between the five groups. The highest class probabilities $\max_k \hat{t}_{g,i,k}$ per observations are shown in Figure 5.6.1 with different plot symbols.

Observations with at least one flagged cell are shown in Figure 5.6.2. The top panel shows the estimated class probabilities $\hat{t}_{g,i,k}$ by the color of the tiles, while the membership to one of the original groups is marked by a dot. In the bottom panel, outlying cells are colored according to their standardized residuals $r_{g,ij}$ (Raymaekers and Rousseeuw, 2023),

$$r_{g,ij} = \sum_{k=1}^{N} \hat{t}_{g,i,k} \frac{x_{g,ij} - \hat{x}^k_{g,ij}}{\sqrt{\hat{\boldsymbol{\Sigma}}^{(j|j)}_{reg,k} - \hat{\boldsymbol{\Sigma}}^{(j|\hat{\boldsymbol{w}}_{g,i})}_{reg,k} \left(\hat{\boldsymbol{\Sigma}}^{(\hat{\boldsymbol{w}}_{g,i}|\hat{\boldsymbol{w}}_{g,i})}_{reg,k}\right)^{-1} \hat{\boldsymbol{\Sigma}}^{(\hat{\boldsymbol{w}}_{g,i}|j)}_{reg,k}}},$$

where $\hat{x}^k_{g,ij}$ denotes the expected value of $x_{g,ij}$ given that it is from distribution k and using only unflagged cells $\hat{\boldsymbol{w}}_{g,i}$, see also Equation (5.3.2). The proposed method can identify if observations are outlying in all groups, indicated by a high number of cellwise outliers (e.g. half of the cells are outlying), or whether they are outlying specifically in their pre-defined group, indicated by a high probability for another group. In the upper panel of Figure 5.6.2 showing only observations with outlying cells, this is expressed by non-overlapping dots and dark blue tiles.

Positive values of the residual indicate that the observed value is higher than what would be expected, and negative values refer to observed values which are lower than expected, given the other non-flagged cells. Many outliers are connected to cell outliers

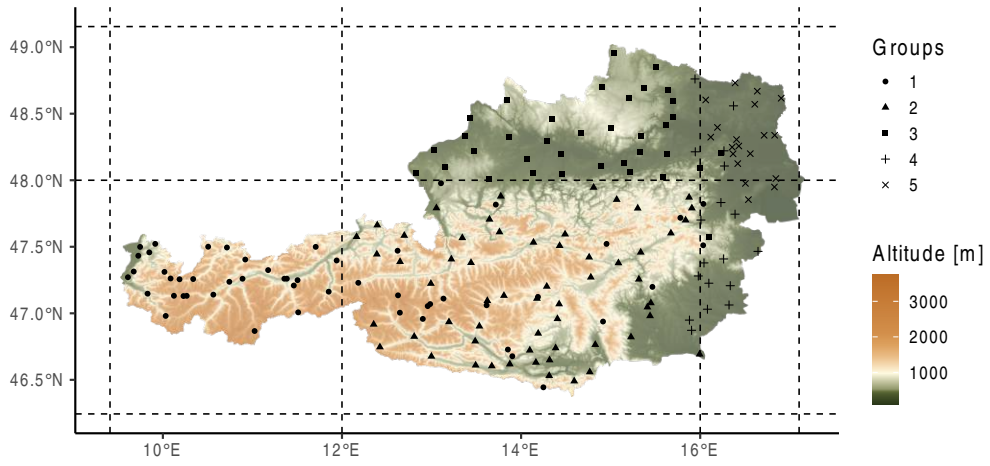


Figure 5.6.1: Altitude map of Austria with spatial locations of weather stations indicated by black symbols and separation into groups indicated by dashed grid lines. Shapes are based on the maximal class probability of the corresponding observation.

in the variable wind velocity, likely due to the diverse exposure of weather stations even in the same area. Moreover, a pattern of unexpected high values in wind velocity and low values in air pressure and temperature is visible for the five weather stations with half of their cells outlying (Villacher Alpe, Sonnblick, Rudolfshütte, Patscherkofel, Galzig) - exactly the five highest weather stations with an altitude of more than 2000 meters.

Figure 5.6.3 presents a more detailed analysis of the variables wind velocity and air temperature. The tolerance ellipses, based on the estimated locations and covariance matrices per group, show a smooth transition from groups connected to mountainous landscapes (group 1 and 2) with higher variation in temperature to flatter landscapes (group 3 to 5) with increased variation in wind velocity and generally higher temperature. The only cellwise outlier with unexpectedly high temperature is the weather station Wien-IS, which is located in the city center of the capital Vienna.

5.6.2 Darwin - Alzheimer Disease

Alzheimer disease is a non-curable neuro-degenerative disease which progresses over time, leading to cognitive impairment. To mitigate the negative effects of Alzheimer disease on affected patients and their loved ones, early diagnoses and treatment is essential. In contrast to Cilia et al. (2022) who train a classifier to discriminate between the two groups, we propose to use the developed multi-group GMM methodology as a tool to analyze the gray area between diagnosed Alzheimer patients and subjects considered healthy. While the groups are given by an official diagnosis, some persons can be on the verge to Alzheimer and not yet being diagnosed or at very early stages.

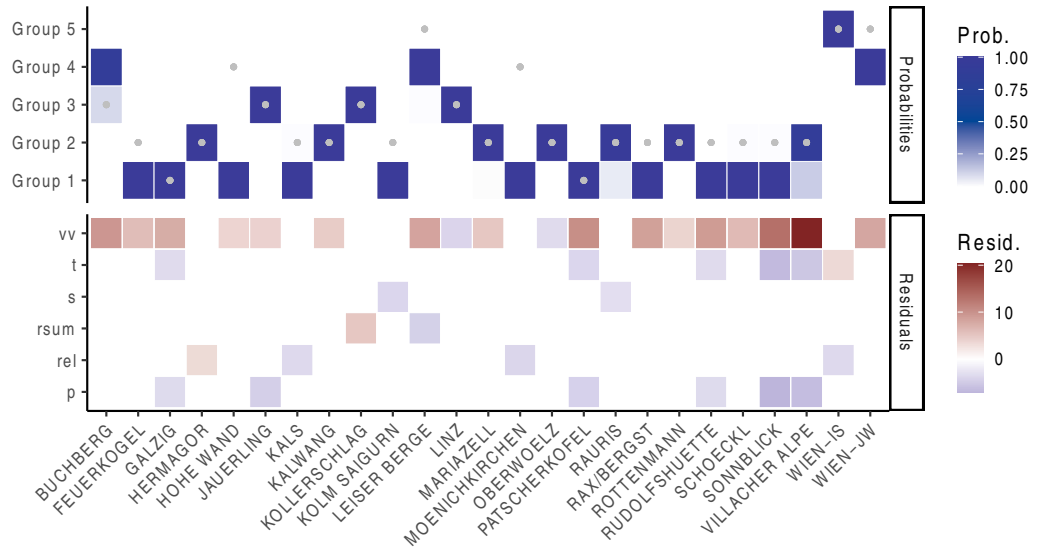


Figure 5.6.2: Outlying weather stations with group probabilities $\hat{t}_{g,i,k}$ on the top panel with dots at the original groups and the residuals of each cell on the bottom panel.

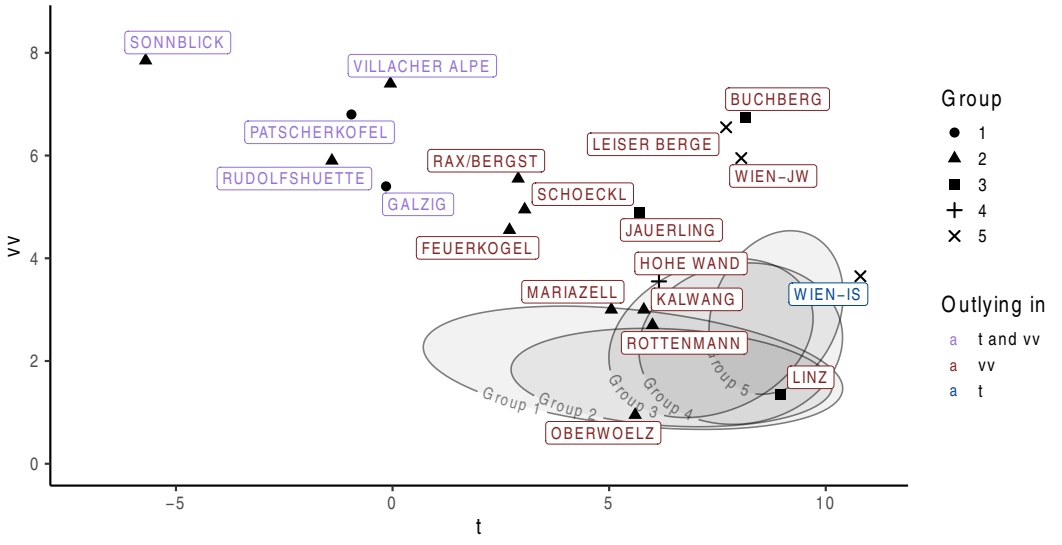


Figure 5.6.3: Bivariate feature space of wind velocity (vv) and air temperature (t). The 95% tolerance ellipses are based on the estimated smoothed covariance matrices and locations per group. Stations outlying in at least one of the two variables are shown. Shapes correspond to the original group of each observation, the color of the label indicates which cells are outlying.

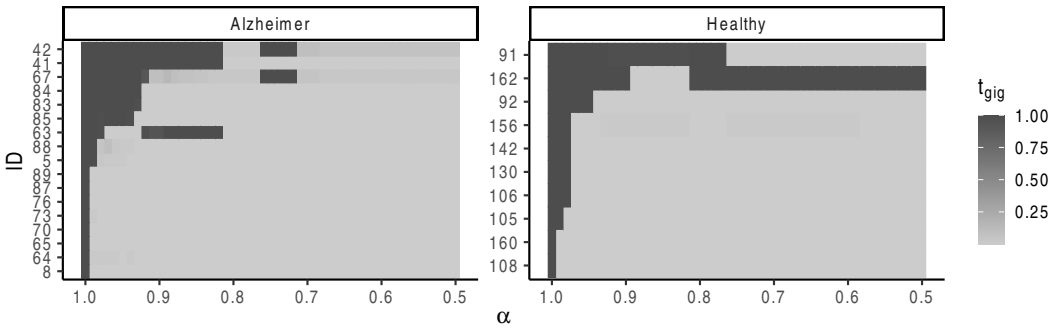


Figure 5.6.4: Class probabilities $t_{g,i,g}$ for subjects whose probabilities change based on α sorted by time of switching.

Thus, the strict separation into groups might not be beneficial, and a more smoothed approach can help to better analyze the intertwinings between the two groups and identify corresponding influential variables.

The DARWIN (Diagnosis AlzheimeR WIth haNdwriting) data set (Cilia et al., 2022), available in the R-package `robustmatrix` (Mayrhofer et al., 2024), contains handwriting samples from $n_1 = 85$ healthy persons and $n_2 = 89$ patients with diagnosed Alzheimer disease (AD). Each subject was asked to execute 25 different handwriting tasks on a tablet from which 18 summary features where extracted: total time, air time, paper time, mean speed on paper, mean speed in air, mean acceleration on paper, mean acceleration on air, mean jerk on paper, mean jerk in air, mean of pressure, variance of pressure, generalization of the mean relative tremor (GMRT) on paper, GMRT in air, mean GMRT, number of pendowns, maximal x-extension, maximal y-extension and dispersion index. For a detailed explanation of the tasks and measured variables we refer to Cilia et al. (2018). Similar to Mayrhofer et al. (2025) we also exclude the variables total time, mean GMRT and air time due to linear dependencies and unreliable measurements. The remaining variables are summarized over the 25 tasks by the median and the median absolute deviation (mad). Thus, we include $p = 30$ variables and the groups are given by the Alzheimer disease status ($N = 2$).

One way to focus on the overlap of the two groups is to vary the parameter $\alpha \in \{1, 0.99, \ldots, 0.51, 0.5\}$ in the calculations. While $\alpha = 1$ forces the observations to belong to the predefined group, decreasing values are less and less strict and enable switching to the other group if the multivariate distribution of that group is more appropriate. Figure 5.6.4 presents the class probabilities $\hat{t}_{g,i,g}$ for varying α for subjects whose probability of being in their predefined class $\hat{t}_{g,i,g}$ is lower than 50% for at least one value of α (*switchers*). We can see that a subset of 8 AD diagnosed patients and 2 healthy subjects move to the opposite group as soon as the procedure starts to allow for a switch, i.e. when $\alpha < 1$, indicating strong multivariate similarities to the opposite group.

Figure 5.6.5 shows all cells of the data matrix, with the observations split into Alzheimer patients and healthy people. Additionally, within these groups we show the switchers, which are sorted according to increasing values of α, thus in the same order

as shown in Figure 5.6.4. The cells of the matrix present information about outlyingness of the cells when varying α (no symbol, crosses or dots), and color according to the standard deviation of the residuals over varying α. If a cell is white, it is not outlying for all α. Cells marked by dots are outlying for several or even all values of α. Higher variability of the residuals can occur for different reasons: (a) the person switches to the other group, (b) the cell is identified as an outlier for particular values of α, or both (a) and (b) occur. Case (a) mainly appears for the switching persons. For example, the variable `pressure_mean` (both median and mad) which shows many cells with increased residual variability. Several of those cells are outliers as soon as the given diagnosis is not enforced to the statistical model, revealing the inhomogeneity of the subjects with respect to this variable. However, there is also a block of cells which are not outliers, and this block appears for persons switching from the healthy to the AD group, as this group provides a better model fit. It might be worth looking closer at the data collection of this variable, since either possible unfavorable measurement conditions or other undiagnosed or progressive diseases affecting the variable could cause the detected unusual behavior. The variable `pressure_mean` (as well as some other features) also leads to cellwise outliers for many observations, while other variables such as `mean_speed_in_air` are inconspicuous.

This plot also provides insights into multivariate cluster overlaps given by the distribution estimates for values of a specific subject. For example, Alzheimer patient 8 switches immediately to the healthy group without any change in residuals, indicating that patient 8 is at the overlap of the clusters in all variables but relatively closer to the center of the healthy cluster. It is likely that such persons have an early diagnosis of Alzheimer and low cognitive impairment.

5.6.3 Wine Quality

Lastly, we leverage the model flexibility to investigate how qualitative expert evaluations of wine are consistent with their quantitative chemical features. To this end, we use a data set of Cortez et al. (2009b), available at the UCI Machine Learning Repository (Cortez et al., 2009a). The data were collected over the years 2004 to 2007 and consist of $p = 11$ physicochemical measurements, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH-level, sulphates, and alcohol percentage, for $n = 4898$ samples of white *vinho verde*, a known Portuguese wine. Additionally, each wine was qualitatively graded from 0 (very bad) to 10 (excellent) by three different sensory assessors by blind tasting. The median of the three grades is reported as the variable quality.

Originally, Cortez et al. (2009b) trained a Support Vector Machine classifier given the quality variable. However, we are more interested in the coherency of each group and whether expert evaluations are consistent regarding the chemical features reported. We partition the data into three groups based on the quality assessment: the first group with low wine quality includes $n_1 = 1640$ wine samples with quality assessments 3 to 5 (20 wine samples with quality level 3, 163 with 4, and 1457 with 5), the second group with medium quality contains $n_2 = 2198$ samples with quality level 6, and the third group includes $n_3 = 1060$ good quality wine samples (880 samples with level 7,

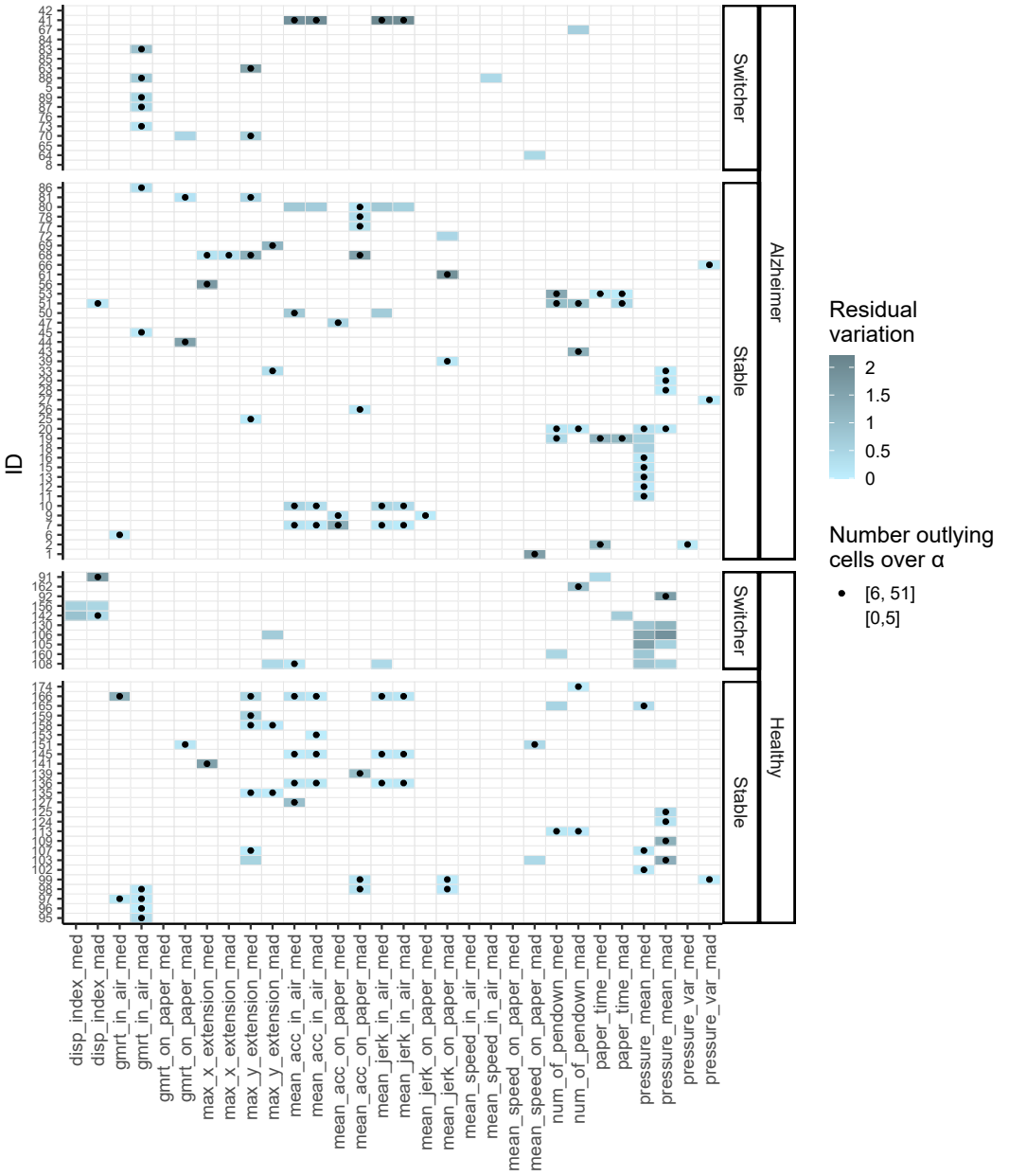


Figure 5.6.5: Standard deviation of residuals over α shown in color per individual and variable. Dots indicate that cells are outlying frequently for 51 different α values. White tiles represent non-outlying cells (no dot) and no calculable variation of residuals.

175 with 8, and 5 samples with quality 9). Due to prominent skewness in multiple variables we apply a robust transformation to each variable to achieve central normality (see Raymaekers and Rousseeuw, 2024b). We then apply the cellgGMM estimator with $\alpha = 0.75$. The increase in α compared to the minimal 0.5 should stabilize the estimation due to the low number of unbalanced groups as well as some incoherency within the groups.

The parallel coordinate plot in Figure 5.6.6 displays the resulting parameter estimates alongside the feature values of the wine samples. Each panel represents wine samples that are of low, medium or high quality according to the experts (column) and of low, medium or high quality according to the predicted group assignment of our model (rows). Consequently, the diagonal panels highlight wine samples where both expert evaluations and statistical methodology agree on their quality.

Panels below the main diagonal show wine samples that experts rate lower than their physicochemical measurements would suggest, while panels above the diagonal show samples rated higher than expected based on their quantitative features. Additionally, each panel includes the estimated location (solid black line) and standard deviation (black error bars) provided by the cellgGMM for the expert-proposed group (thus they are identical in each column). We see a strong heterogeneity within each expert group. While the wine samples where experts and cellgGMM agree are quite coherent, clear structural differences are visible in case of deviations. The two bottom left panels show quantitatively good wines that are rated low by experts. They differ clearly from less qualitative wines, most prominently by low density and residual sugar while containing a relatively high amount of alcohol. On the opposite, wines rated too high by experts (middle right panel) show adverse results for residual sugar, density and alcohol.

Moreover, there are many cellwise outliers detected by the algorithm that are also visible in the parallel coordinate plot. Especially the high amount of outlying chloride values is noticeable, as well as low citric acid values. Here, robustness against cellwise outliers is key to get reliable estimates and to avoid clusters basically modeling one variable with a high number of extreme values.

Overall, we get a good insight into the physicochemical features connected to the quality of wines as given by experts. While we achieve a nice pattern for high quality wines by our proposed multi-group GMM, the heterogeneity of the expert ratings is high. Possible factors might be chemical or physical properties that are not measured but are decisive for assessors when rating wine highly, a somewhat subjective notion of quality, or both. The strong heterogeneity together with multiple prominent cellwise outliers might also explain why previous classification attempts for this specific data set only achieve an accuracy of up to 64.6% in Cortez et al. (2009b).

5.7 Summary and Conclusions

We establish a flexible GMM that accounts for external group information and can deliver moment estimates matched to given groups. Underlying progressive structures of the multi-group setting are present in many multi-group data sets and can be leveraged. To this end, we introduce a probabilistic multi-group GMM allowing observations to

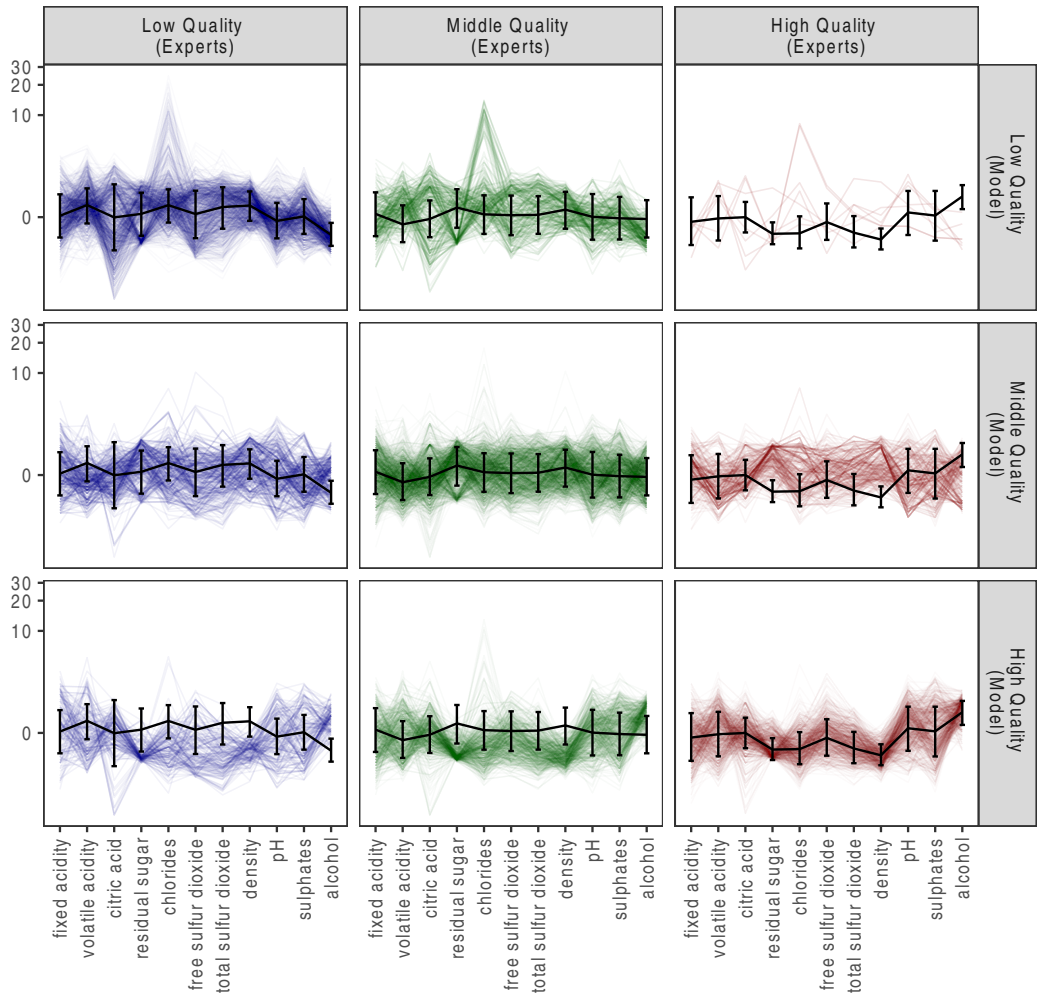


Figure 5.6.6: Differences in wine quality assessment of expert rating (columns) and physicochemical features. Black lines show estimated location and standard deviation for expert groups, colored lines show wine measurements, divided in expert group (column) and the statistically most likely group (rows).

originate from other than their pre-defined group. An objective function is formulated based on its likelihood together with a penalty term.

A further contribution of this paper is the introduction of an appropriate notion of breakdown of the estimator in the cellwise multi-group setting. A novel setting of ideally cellwise well-clustered data is described for which the cellwise robustness properties can theoretically be evaluated and compared between different methods. This optimal setting is further extended to multi-group data for which we prove the breakdown point for the proposed cellwise robust multi-group GMM.

An iterative algorithm based on the EM algorithm guarantees convergence to a local optimum and due to the additional regularization the resulting estimator is applicable in high-dimensions. The robustness of the estimator is confirmed also in extensive simulation covering multiple relevant scenarios, and its usefulness is further demonstrated on three versatile real life examples where possible interpretation angles of the rich output of the method are illustrated in detail.

Compared to other methods, the cellgGMM provides a one-to-one match of estimated covariance and location parameters with pre-defined groups while allowing observations to be assigned flexibly to other groups if they are better fitting – a combination not offered by other methods. In contrast, classical GMMs deliver estimates that are not clearly matching known groups, and separate analysis forces observation to be always of the original group. The approach is in a way also more refined than robust discriminant analysis (for an overview see Hubert et al., 2024) which would discard observations in the covariance estimation that might not fit to the pre-defined groups due to misgrouping or being in the gray area between groups, e.g. when groups are related to progressive medical diseases or diagnoses. In applications, especially the parameter α that specifies the strictness of the membership to the given groups is a particularly well-suited tool to shed light on transition dynamics when varied. In a broader sense, the parameter α continuously bridges the gap between a separate parameter estimation via the cellMCD for each group when $\alpha = 1$ and a classical (cellwise robust) GMM with a given number of clusters in the extreme (and excluded) case of $\alpha = 0$.

The proposed method is applicable in many fields of research where assignments to pre-defined groups can be viewed more flexibly. Future research might leverage the resulting moment estimates for other prominent multivariate methods like principal component analysis, discriminant analysis or graphical modeling, and possibly further extend classical methods towards a joint approach for group dependent and group independent features.

Appendix D

D.1 Derivation of Group Moments

Given the multi-group Gaussian mixture model in Equation (5.2.1) we can derive the group moments.

Expected value: Due to the law of total expectation it follows that

$$\mathbb{E}[\boldsymbol{x}_g] = \sum_{k=1}^{N} \mathbb{P}[\boldsymbol{x}_g \in k]\mathbb{E}[\boldsymbol{x}_g | \boldsymbol{x}_g \in k] = \sum_{k=1}^{N} \pi_{g,k}\boldsymbol{\mu}_k.$$

Covariance: We want to show Equation (5.2.2),

$$\mathrm{Cov}[\boldsymbol{x}_g] = \sum_{k=1}^{N} \pi_{g,k}\boldsymbol{\Sigma}_k + \sum_{k=1}^{N} \pi_{g,k}(\boldsymbol{\mu}_k - \mathbb{E}[\boldsymbol{x}_g])(\boldsymbol{\mu}_k - \mathbb{E}[\boldsymbol{x}_g])'.$$

For fixed variables j, j' (that can also be equal), the corresponding covariance based on Equation (5.2.2) can be reformulated as

$$
\begin{aligned}
\mathrm{Cov}[\boldsymbol{x}_g]_{j,j'} &= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,j,j'} + \sum_{k=1}^{N} \pi_{g,k}((\boldsymbol{\mu}_k - \mathbb{E}[\boldsymbol{x}_g])(\boldsymbol{\mu}_k - \mathbb{E}[\boldsymbol{x}_g])')_{j,j'} \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,j,j'} + \sum_{k=1}^{N} \pi_{g,k}(\boldsymbol{\mu}_k - \mathbb{E}[\boldsymbol{x}_g])_j(\boldsymbol{\mu}_k - \mathbb{E}[\boldsymbol{x}_g])_{j'} \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,j,j'} \\
&\quad + \sum_{k=1}^{N} \pi_{g,k}(\mu_{k,j}\mu_{k,j'} - \mu_{k,j}\mathbb{E}[\boldsymbol{x}_g]_{j'} - \mu_{k,j'}\mathbb{E}[\boldsymbol{x}_g]_j + \mathbb{E}[\boldsymbol{x}_g]_{j'}\mathbb{E}[\boldsymbol{x}_g]_j) \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,j,j'} + \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j}\mu_{k,j'} - \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j}\mathbb{E}[\boldsymbol{x}_g]_{j'} \\
&\quad - \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j'}\mathbb{E}[\boldsymbol{x}_g]_j + \sum_{k=1}^{N} \pi_{g,k}\mathbb{E}[\boldsymbol{x}_g]_{j'}\mathbb{E}[\boldsymbol{x}_g]_j \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,j,j'} + \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j}\mu_{k,j'} - \mathbb{E}[\boldsymbol{x}_g]_{j'} \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j} \\
&\quad - \mathbb{E}[\boldsymbol{x}_g]_j \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j'} + \mathbb{E}[\boldsymbol{x}_g]_{j'}\mathbb{E}[\boldsymbol{x}_g]_j \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,j,j'} + \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j}\mu_{k,j'} - \mathbb{E}[\boldsymbol{x}_g]_{j'}\mathbb{E}[\boldsymbol{x}_g]_j. \qquad \text{(D.1)}
\end{aligned}
$$

We can introduce the random variable $Z_{g,i}$ indicating from which distribution observation $\boldsymbol{x}_g$ comes. From the law of total covariance we get that

$$
\begin{aligned}
\mathrm{Cov}(\boldsymbol{x}_{g,j}, \boldsymbol{x}_{g,j'}) &= \mathbb{E}[\mathrm{Cov}(\boldsymbol{x}_{g,j}, \boldsymbol{x}_{g,j'}|Z)] + \mathrm{Cov}(\mathbb{E}[\boldsymbol{x}_{g,j'}|Z], \mathbb{E}[\boldsymbol{x}_{g,j}|Z]) \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,jj'} + \mathrm{Cov}(\mu_{Z,j'}, \mu_{Z,j}) \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,jj'} + \mathbb{E}(\mu_{Z,j'}\mu_{Z,j}) - \mathbb{E}(\mu_{Z,j'})\mathbb{E}(\mu_{Z,j}) \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,jj'} + \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j'}\mu_{k,j} - (\sum_{k=1}^{N} \pi_{g,k}\mu_{k,j'})(\sum_{k=1}^{N} \pi_{g,k}\mu_{k,j}) \\
&= \sum_{k=1}^{N} \pi_{g,k}\Sigma_{k,jj'} + \sum_{k=1}^{N} \pi_{g,k}\mu_{k,j'}\mu_{k,j} - \mathbb{E}[\boldsymbol{x}_g]_{j'}\mathbb{E}[\boldsymbol{x}_g]_j. \qquad \text{(D.2)}
\end{aligned}
$$

We can see that the right hand sides of Equation (D.1) and (D.2) are the same.

D.2 Derivation of the Breakdown Point

Idealized Scenario in a Rowwise Outlier Paradigm

The classical finite sample addition (replacement) breakdown point describes the maximal fraction of observations that need to be added to (replaced with arbitrary values in) a given sample to make the estimator useless. An estimator of location $\hat{\boldsymbol{\mu}}$ breaks down if it becomes unbounded, $||\hat{\boldsymbol{\mu}}||_2 \to \infty$. An estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$ becomes either unbounded and explodes (*explosion breakdown point*), $\lambda_1(\hat{\boldsymbol{\Sigma}}) \to \infty$, or singular (*implosion breakdown point*), $\lambda_p(\hat{\boldsymbol{\Sigma}}) = 0$, where λ_1 and λ_p describe the largest and smallest eigenvalue, respectively.

However, in the setting of mixture models, pathological settings where (robust) estimators break down by just changing one observation can occur. Thus, we focus on the additive breakdown point for parameter estimation in ideal settings of well-clustered data points for mixture models, as described in Hennig (2004) for univariate and extended by Cuesta-Albertos et al. (2008) to multivariate data. A sequence of clusters $(\mathcal{X}_m)_{m\in\mathbb{N}}$ is considered to be ideal when the distances of observations within clusters are bounded by a constant $b < \infty$ and observations from different clusters are increasingly far away. Formally, let $s \geq 2$ be the number of clusters and $\tilde{n}_1 < \tilde{n}_2 < \ldots < \tilde{n}_s = \tilde{n} \in \mathbb{N}$. For each m-th part of the sequence, the data $\mathcal{X}_m$ is clustered into s clusters $A_m^1, \ldots, A_m^s$ such that

$$A_m^1 = \{\boldsymbol{x}_{1,m}, \ldots, \boldsymbol{x}_{\tilde{n}_1,m}\}, \ldots, A_m^s = \{\boldsymbol{x}_{\tilde{n}_{s-1}+1,m}, \ldots, \boldsymbol{x}_{\tilde{n}_s,m}\}$$

and $\mathcal{X}_m = \bigcup_{l=1}^s A_m^l$. The formal conditions for ideal clusters above are

$$\max_{1\leq l\leq s} \max\{||\boldsymbol{x}_{i',m} - \boldsymbol{x}_{i,m}||_2 : \boldsymbol{x}_{i',m}, \boldsymbol{x}_{i,m} \in A_m^l\} < b \quad \forall m \in \mathbb{N}, \tag{D.3}$$

$$\lim_{m\to\infty} \min\{||\boldsymbol{x}_{i',m} - \boldsymbol{x}_{i,m}||_2 : \boldsymbol{x}_{i',m} \in A_m^l, \boldsymbol{x}_{i,m} \in A_m^h, h \neq l\} = \infty, \tag{D.4}$$

where $||.||_2$ denotes the Euclidean norm. The added outliers denoted as $\mathcal{Y}_m = \{\boldsymbol{y}_{1,m}, \ldots, \boldsymbol{y}_{\tilde{r},m}\}$ should be clearly distinguished from all clusters and not build a cluster on their own,

$$\lim_{m\to\infty} \min\{||\boldsymbol{y}_{i',m} - \boldsymbol{x}_{i,m}||_2 : \boldsymbol{x}_{i,m} \in \mathcal{X}_m, \boldsymbol{y}_{i',m} \in \mathcal{Y}_m\} = \infty,$$
$$\lim_{m\to\infty} \min\{||\boldsymbol{y}_{i',m} - \boldsymbol{y}_{i,m}||_2 : \boldsymbol{y}_{i',m}, \boldsymbol{y}_{i,m} \in \mathcal{Y}_m, i \neq i'\} = \infty.$$

The breakdown of an estimator is then relatively defined by estimates based on $\mathcal{X}_m$ and on $\mathcal{X}_m \cup \mathcal{Y}_m$. Location breakdown for a cluster l occurs, if for all $h = 1, \ldots, N$

$$||\hat{\boldsymbol{\mu}}_l(\mathcal{X}_m) - \hat{\boldsymbol{\mu}}_h(\mathcal{X}_m \cup \mathcal{Y}_m)||_2 \to \infty. \tag{D.5}$$

A covariance estimator of a cluster l would implode if $\lambda_p(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m)) \to 0$ and $\lambda_p(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m \cup \mathcal{Y}_m)) \nrightarrow 0$ or if $\lambda_p(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m)) \nrightarrow 0$ and $\lambda_p(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m \cup \mathcal{Y}_m)) \to 0$. Analogously, the explosion breakdown occurs when $\lambda_1(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m)) \to \infty$ and $\lambda_1(\hat{\boldsymbol{\Sigma}}_l(\mathcal{X}_m \cup \mathcal{Y}_m)) \nrightarrow \infty$ or vice versa. The weight estimator $\hat{\pi}_l$ of a cluster l breaks down if $\hat{\pi}_l \in \{0, 1\}$. The addition breakdown point is then defined as $\frac{\tilde{r}}{\tilde{n}+\tilde{r}}$ where $\tilde{r}$ is the minimal number of added outliers necessary to break down the parameter estimate. Both illustrations in Figure 5.4.1 depict ideal settings in the rowwise outlier paradigm.

Proof of Corollary 5.4.2.1

Proof of Corollary 5.4.2.1. For ease of notation we drop the superscript m for observations and the explicit dependence of the estimators on $\mathcal{Z}_m$ or $\mathcal{X}_m$. All limits are corresponding to $m \to \infty$. The notation $\boldsymbol{w}(\boldsymbol{y})$ marks the real outlying cells of $\boldsymbol{y}$ while the notation $\boldsymbol{w}_{\boldsymbol{y}}$ indicates the missingness pattern of $\boldsymbol{y}$ for a given $\boldsymbol{W}$ from the objective function if the indexation of $\boldsymbol{y}$ is irrelevant. Then penalty term can generally be left out since it is always bounded,

$$|\sum_{g=1}^{N}\sum_{i=1}^{n_g}\sum_{j=1}^{p} q_{g,ij}(1-w_{g,ij})| \leq pN \max_g n_g \max_{g,i,j} q_{g,ij} < \infty.$$

a. Given a data matrix $\mathcal{X}$ we construct a set of estimators with finite value of the objective function. For all $k = 1, \ldots, N$ set $\hat{\Sigma}_{k,jj} = 1$ and zero otherwise and $\hat{\boldsymbol{\mu}}_k = \frac{1}{|A_m^k|}\sum_{\boldsymbol{x}\in A_m^k} \boldsymbol{x}$, where $|A_m^k|$ denotes the number of elements in A_m^k. Then, also regularized covariance matrices $\hat{\boldsymbol{\Sigma}}_{reg,k}$ have finite positive eigenvalues.

   1. Assume $\alpha \neq 1$. Set $\hat{\pi}_{k,k} = \alpha \geq 0.5$, $\hat{\pi}_{k,l} = \frac{1-\alpha}{N-1} > 0$ for $k \neq l$. For each observation $\boldsymbol{x}_{g,i}$ with $\boldsymbol{w}_{g,i}$ originating from any cluster l it holds that

$$\begin{aligned}
&\ln\left(\sum_{k=1}^{N}\hat{\pi}_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}; \hat{\boldsymbol{\mu}}_k^{(\boldsymbol{w}_{g,i})}, \hat{\boldsymbol{\Sigma}}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)\right) \\
&\quad \geq \ln\left(\frac{1-\alpha}{N-1}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}; \hat{\boldsymbol{\mu}}_l^{(\boldsymbol{w}_{g,i})}, \hat{\boldsymbol{\Sigma}}_{reg,l}^{(\boldsymbol{w}_{g,i})}\right)\right) \\
&\quad = \ln\frac{1-\alpha}{N-1} + \ln\frac{e^{-\frac{1}{2}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}-\hat{\boldsymbol{\mu}}_l^{(\boldsymbol{w}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,l}^{(\boldsymbol{w}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}-\hat{\boldsymbol{\mu}}_l^{(\boldsymbol{w}_{g,i})})}}{\sqrt{(2\pi)^{\sum_j w_{g,ij}}\det\hat{\boldsymbol{\Sigma}}_{reg,l}^{(\boldsymbol{w}_{g,i})}}} \\
&\quad = \ln\frac{1-\alpha}{N-1} - \frac{1}{2}\Bigg((\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})} - \hat{\boldsymbol{\mu}}_l^{(\boldsymbol{w}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,l}^{(\boldsymbol{w}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})} - \hat{\boldsymbol{\mu}}_l^{(\boldsymbol{w}_{g,i})}) \\
&\qquad + \sum_j w_{g,ij}\ln(2\pi) + \ln\det\hat{\boldsymbol{\Sigma}}_{reg,l}^{(\boldsymbol{w}_{g,i})}\Bigg) \\
&\quad \geq \ln\frac{1-\alpha}{N-1} - \frac{1}{2}(\boldsymbol{b}^{(\boldsymbol{w}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,l}^{(\boldsymbol{w}_{g,i})})^{-1}(\boldsymbol{b}^{(\boldsymbol{w}_{g,i})}) \\
&\qquad - \frac{1}{2}p\ln(2\pi) - \frac{1}{2}\ln\det\hat{\boldsymbol{\Sigma}}_{reg,l}^{(\boldsymbol{w}_{g,i})},
\end{aligned}$$

   where $\boldsymbol{b}$ denotes the vector $\boldsymbol{b} = (b, \ldots, b) \in \mathbb{R}^p$ with b corresponding to Equation (5.4.2) and the last inequality follows from Equation (D.3) with the Euclidean norm. Since all terms on the right hand side are bounded, the objective function is bounded from above. For the lower bound, it follows

that

$$\begin{aligned}
&\ln\left(\sum_{k=1}^{N}\hat{\pi}_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})};\hat{\boldsymbol{\mu}}_k^{(\boldsymbol{w}_{g,i})},\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)\right)\\
&\quad\leq \ln N+\max_k \ln\left(\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})};\hat{\boldsymbol{\mu}}_k^{(\boldsymbol{w}_{g,i})},\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)\right)\\
&\quad\leq \ln N+\max_k(\underbrace{-\frac{1}{2}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}-\hat{\boldsymbol{\mu}}_k^{(\boldsymbol{w}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\boldsymbol{w}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}-\hat{\boldsymbol{\mu}}_k^{(\boldsymbol{w}_{g,i})})}_{\leq 0})\\
&\qquad\underbrace{-\frac{1}{2}\sum_j w_{g,ij}\ln(2\pi)}_{\leq 0}+\max_k(-\frac{1}{2}\ln\det\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\boldsymbol{w}_{g,i})})\\
&\quad\leq \ln N-\frac{1}{2}\ln\min_k\det\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\boldsymbol{w}_{g,i})}.
\end{aligned}$$

Since the covariance estimates are finite, the objective function is bounded for any feasible $\boldsymbol{W}$.

2. Assume $\alpha=1$. Set $\hat{\pi}_{k,k}=1$, $\hat{\pi}_{k,l}=0$ for all $k\neq l$. All observations from a group g originate from cluster g, $\boldsymbol{Z}^g=A^g$, see Equation (5.4.6). Thus, for any $\boldsymbol{x}_{g,i}$ it holds that

$$\begin{aligned}
&\ln\left(\sum_{k=1}^{N}\hat{\pi}_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})};\hat{\boldsymbol{\mu}}_k^{(\boldsymbol{w}_{g,i})},\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)\right)\\
&\qquad=-\frac{1}{2}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}-\hat{\boldsymbol{\mu}}_g^{(\boldsymbol{w}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,g}^{(\boldsymbol{w}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}-\hat{\boldsymbol{\mu}}_g^{(\boldsymbol{w}_{g,i})})\\
&\qquad\quad-\frac{1}{2}\sum_j w_{g,ij}\ln(2\pi)-\frac{1}{2}\ln\det\hat{\boldsymbol{\Sigma}}_{reg,g}^{(\boldsymbol{w}_{g,i})}\\
&\qquad\geq-\frac{1}{2}\left((\boldsymbol{b}^{(\boldsymbol{w}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,g}^{(\boldsymbol{w}_{g,i})})^{-1}(\boldsymbol{b}^{(\boldsymbol{w}_{g,i})})+p\ln(2\pi)+\ln\det\hat{\boldsymbol{\Sigma}}_{reg,g}^{(\boldsymbol{w}_{g,i})}\right)
\end{aligned}$$

and the objective function is bounded from above. For the lower bound, it

follows

$$\begin{aligned}
&\ln\left(\sum_{k=1}^{N}\hat{\pi}_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})};\hat{\boldsymbol{\mu}}_{k}^{(\boldsymbol{w}_{g,i})},\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)\right)\\
&\qquad=\underbrace{-\frac{1}{2}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}-\hat{\boldsymbol{\mu}}_{g}^{(\boldsymbol{w}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,g}^{(\boldsymbol{w}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})}-\hat{\boldsymbol{\mu}}_{g}^{(\boldsymbol{w}_{g,i})})}_{\leq 0}\\
&\qquad\quad\underbrace{-\frac{1}{2}\sum_{j}w_{g,ij}\ln(2\pi)}_{\leq 0}-\frac{1}{2}\ln\det\hat{\boldsymbol{\Sigma}}_{reg,g}^{(\boldsymbol{w}_{g,i})}\\
&\qquad\leq-\frac{1}{2}\ln\det\hat{\boldsymbol{\Sigma}}_{reg,g}^{(\boldsymbol{w}_{g,i})}.
\end{aligned}$$

Thus, the objective function is bounded for any feasible $\boldsymbol{W}$.

b. Assume that under the given estimates the objective function is bounded. By construction, the estimated covariances $\hat{\boldsymbol{\Sigma}}_{reg,k}$ are regular and thus, the lowest eigenvalues $\lambda_p(\hat{\boldsymbol{\Sigma}}_{reg,k}) \geq \tilde{b}_k(\rho_k, \boldsymbol{T}_k) > 0$ are bounded away from zero. According to the proof of Proposition 2b) from Raymaekers and Rousseeuw (2023) it holds for all k and any feasible $\hat{\boldsymbol{w}}$ that

$$\ln\det\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}})}\geq\ln\max_{j=1,\ldots,p}\hat{\Sigma}_{reg,k,jj}^{(\hat{\boldsymbol{w}})}+(p-1)\ln\tilde{b}_k(\rho_k,\boldsymbol{T}_k).$$

where $\tilde{b}_k(\rho_k, \boldsymbol{T}_k)$ is a constant depending only on ρ_k and $\boldsymbol{T}_k$.

From part a. we know that for all $\boldsymbol{x}_{g,i}$ from group g it holds that

$$\begin{aligned}
&\ln\left(\sum_{k=1}^{N}\hat{\pi}_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i})};\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i})},\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i})}\right)\right)\\
&\quad\leq\ln N-\frac{1}{2}\min_{k}\Bigg((\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i})}-\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i})}-\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i})})\\
&\qquad+\ln\det\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i})}\Bigg)\\
&\quad\leq\ln N-\frac{1}{2}\min_{k}\Bigg((\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i})}-\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i})})'(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i})}-\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i})})\\
&\qquad+\ln\max_{j=1,\ldots,p}\hat{\Sigma}_{reg,k,jj}^{(\hat{\boldsymbol{w}}_{g,i})}+(p-1)\ln\tilde{b}_k(\rho_k,\boldsymbol{T}_k)\Bigg)\\
&\quad\leq\ln N-\frac{1}{2}\min_{k}(p-1)\ln\tilde{b}_k(\rho_k,\boldsymbol{T}_k)-\frac{1}{2}\min_{k}\Bigg((\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i})}-\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i})})'\\
&\qquad\times(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i})})^{-1}(\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i})}-\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i})})+\ln\max_{j=1,\ldots,p}\hat{\Sigma}_{reg,k,jj}^{(\hat{\boldsymbol{w}}_{g,i})}\Bigg).
\end{aligned}\tag{D.6}$$

Let $j^*(l) = \max_{j=1,\dots,p} \hat{\Sigma}_{reg,l,jj}$ for the distribution where $\lambda_1(\hat{\boldsymbol{\Sigma}}_{reg,l}) \to \infty$. For each group g there exists at least one observation $\boldsymbol{x}_{g,i^*(g)}$ from cluster g for which variable $j^*(l)$ is observed, $w_{g,i^*(g)j^*(l)} = 1$. For these observations, we have

$$\begin{aligned}(\boldsymbol{x}_{g,i^*(g)}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})} - \hat{\boldsymbol{\mu}}_l^{(\hat{\boldsymbol{w}}_{g,i^*(g)})})'(\hat{\boldsymbol{\Sigma}}_{reg,l}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})})^{-1} \\ \times(\boldsymbol{x}_{g,i^*(g)}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})} - \hat{\boldsymbol{\mu}}_l^{(\hat{\boldsymbol{w}}_{g,i^*(g)})}) + \ln \max_{j=1,\dots,p} \hat{\Sigma}_{reg,l,jj}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})} &\geq \ln \max_{j=1,\dots,p} \hat{\Sigma}_{reg,l,jj} \\ &= \ln \max_{j,j'=1,\dots,p} |\hat{\Sigma}_{reg,l,jj'}| \\ &\geq \ln \frac{\lambda_1(\hat{\boldsymbol{\Sigma}}_{reg,l})}{p} \to \infty.\end{aligned}$$

Thus, for all $\boldsymbol{x}_{g,i^*(g)}, g = 1, \dots, N$ the argument l cannot be the minimizer.

Without loss of generality, assume that all other covariance matrices are bounded, $\lambda_1(\hat{\boldsymbol{\Sigma}}_{reg,k}) < \infty$ if $k \neq l$. Due to Equation (5.4.1), (5.4.3) and (5.4.4) it holds that $|x_{g,i^*(g)j^*(l)} - x_{h,i^*(h)j^*(l)}| \to \infty$ if $g \neq h$. Also,

$$\begin{aligned}(\boldsymbol{x}_{g,i^*(g)}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})} - \hat{\boldsymbol{\mu}}_k^{(\hat{\boldsymbol{w}}_{g,i^*(g)})})'(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})})^{-1}(\boldsymbol{x}_{g,i^*(g)}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})} - \hat{\boldsymbol{\mu}}_k^{(\hat{\boldsymbol{w}}_{g,i^*(g)})}) \\ \geq (x_{g,i^*(g)j^*(l)} - \hat{\mu}_{k,j^*(l)})^2 (\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})})^{-1}_{j^*(l)j^*(l)}.\end{aligned}$$

If $(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})})^{-1}_{j^*(l)j^*(l)} \to 0$, then the smallest eigenvalue goes to zero,

$$\lambda_p((\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})})^{-1}) \to 0$$

implying $\lambda_1(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})}) \to \infty$ as well as $\lambda_1(\hat{\boldsymbol{\Sigma}}_{reg,k}) \to \infty$, which contradicts that the other covariances are bounded in the first eigenvalue. Thus, $(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i^*(g)})})^{-1}_{j^*(l)j^*(l)}$ is bounded away from zero.

Since all observations are increasingly far away, there exists at least one $\boldsymbol{x}_{g',i^*(g')}$ such that $(x_{g',i^*(g')j^*(l)} - \hat{\mu}_{k,j^*(l)})^2 \to \infty$ for all $k = 1, \dots, N, k \neq l$ and for which the minimum from Equation (D.6) goes to infinity. Moreover, all parts are bounded from above,

$$\ln\left(\sum_{k=1}^{N} \hat{\pi}_{g,k} \varphi\left(\boldsymbol{x}_{g,i}^{(\hat{\boldsymbol{w}}_{g,i})}; \hat{\boldsymbol{\mu}}_k^{(\hat{\boldsymbol{w}}_{g,i})}, \hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i})}\right)\right) \leq \ln N - \frac{p}{2} \min_k \ln \tilde{b}_k(\rho_k, \boldsymbol{T}_k).$$

Thus, the objective function has to explode.

c. Assume that the objective function of the estimators $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{W}}$ is finite. Then $\hat{\boldsymbol{\Sigma}}_{reg,k}$ are regular and not exploding due to part b. For all groups g there exists at least one observation $\boldsymbol{x}_{g,i*(g)} \in (A^g \cup B^g) \cap \boldsymbol{Z}^g$ such that $\hat{w}_{g,i*(g)j^*} = 1$. Let $C_1 = \min_{k,\hat{\boldsymbol{w}},j} \hat{\boldsymbol{\Sigma}}_{reg,k,jj}^{(\hat{\boldsymbol{w}})} > 0$ and $C_2 = \min_{k,\hat{\boldsymbol{w}},j} (\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}})})^{-1}_{jj} > 0$ (see part b.), then

it holds

$$\begin{aligned}
&\ln\left(\sum_{k=1}^{N}\hat{\pi}_{g,k}\varphi\left(\boldsymbol{x}_{g,i*(g)}^{(\hat{\boldsymbol{w}}_{g,i*(g)})};\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i*(g)})},\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i*(g)})}\right)\right)\\
&\quad\leq \ln N-\frac{1}{2}\min_{k}(p-1)\ln\tilde{b}_{k}(\rho_{k},\boldsymbol{T}_{k})-\frac{1}{2}\min_{k}\ln\max_{j=1,\ldots,p}\hat{\boldsymbol{\Sigma}}_{reg,k,jj}^{(\hat{\boldsymbol{w}}_{g,i*(g)})}\\
&\qquad-\frac{1}{2}\min_{k}\Bigg((\boldsymbol{x}_{g,i*(g)}^{(\hat{\boldsymbol{w}}_{g,i*(g)})}-\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i*(g)})})'(\hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{g,i*(g)})})^{-1}\\
&\qquad\times(\boldsymbol{x}_{g,i*(g)}^{(\hat{\boldsymbol{w}}_{g,i*(g)})}-\hat{\boldsymbol{\mu}}_{k}^{(\hat{\boldsymbol{w}}_{g,i*(g)})})\Bigg)\\
&\quad\leq \ln N-\frac{1}{2}\min_{k}(p-1)\ln\tilde{b}_{k}(\rho_{k},\boldsymbol{T}_{k})-\frac{1}{2}\ln C_{1}\\
&\qquad-\frac{1}{2}C_{2}\min_{k}\left((x_{g,i^*(g)j^*}-\hat{\mu}_{k,j^*})^{2}\right).
\end{aligned}$$

There are N many observations observed in j^* that move increasingly far away from each other in variable j^*. Since there exists l', l such that to $|\hat{\mu}_{l',j^*}-\hat{\mu}_{l,j^*}|<\tilde{b}$ there are only $N-1$ location estimates that move infinitely far away from each other. It follows that $\max_g \min_k (x_{g,i^*(g)j^*}-\hat{\mu}_{k,j^*})^2\to\infty$ and thus, there is one term in the objective function that explodes, while the others are bounded (see part b.).

□

Proof of Breakdown Points in Theorem 5.4.2.1

Proof of Theorem 5.4.2.1. a. Clear, since the lowest eigenvalues are always bound away from zero (see also proof of Theorem 2c in Puchhammer and Filzmoser, 2024).

b. Since constraint (5.2.7) restricts the estimates $\hat{\boldsymbol{\pi}}(\mathcal{Z}_m)$ such that $\hat{\pi}(\mathcal{Z}_m)_{g,g}\geq\alpha>0$ for all g, the weight of each cluster k is $\frac{1}{N}\sum_{g=1}^{N}\hat{\pi}(\mathcal{Z}_m)_{g,k}\geq\frac{\alpha}{N}>0$. Thus, all clusters have non-zero weight.

c. From Corollary 5.4.2.1a., we know for uncontaminated data $\mathcal{X}_m$ that the objective function is finite for the minimizers, and from Corollary 5.4.2.1b. we know that the covariance matrix estimates are not exploding. Thus, a breakdown occurs only when there exists an l such that $\lambda_1(\hat{\boldsymbol{\Sigma}}_{reg,l}(\mathcal{Z}_m))\to\infty$.

Assume that for each group g only up to n_g-h_g cells per column are contaminated and outlying in the idealized scenario. It is possible to set $\hat{\boldsymbol{W}}$ such that $w_{\boldsymbol{y},j}=0$ for all cells of added outliers $\boldsymbol{y}$ exactly when $w(\boldsymbol{y})_j=0$. Thus, there exists a copy of an uncontaminated ideal scenario $\tilde{\mathcal{X}}_m$, that has the same values if cells are observed as indicated by $\hat{\boldsymbol{W}}$ and non-outlying values if $w_{\boldsymbol{y},j}=0$. From Corollary 5.4.2.1a. for the given $\hat{\boldsymbol{W}}$ it follows that there exist candidate estimates

with finite objective function for $\tilde{\mathcal{X}}_m$ and the value of the objective function on $\mathcal{X}_m \cup \mathcal{Y}_m$ is the same (and finite). From Corollary 5.4.2.1b. it follows that if a covariance matrix explodes, the objective function explodes as well and the estimates cannot be minimizers of the objective function because there exist candidate estimates with a lower objective function. Thus, the breakdown point is at least $\min_g\{(n_g - h_g + 1)/n_g\}$.

d. We produce a special setting that is ideal and uncontaminated and in which there are two possible estimates for location that have increasing distance from each other for $m \to \infty$.

Assume $N = 2$ many groups[3], $\alpha = 0.5$, n_g is even for all $g = 1, \ldots, N$, and that there are no added outliers, $\mathcal{Y}_m = \emptyset$. Further assume that we have minimizing estimates of the objective function, $\hat{\boldsymbol{\pi}}$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{W}}$. Assume $\mathcal{X}_m$ such that the minimizing $\hat{\boldsymbol{W}}$ has zeros in the first column and in the first $n_g/2$ cells and for all other columns there are zeros in the last half of the cells, for both groups. Assume $\mathcal{X}_m$ such that $\hat{\boldsymbol{\Sigma}}_{reg,1} = \hat{\boldsymbol{\Sigma}}_{reg,2}$ as well as $\hat{\pi}_{1,1} = \hat{\pi}_{2,2} = 0.5$. Construct $\tilde{\boldsymbol{\mu}}_1 = (\hat{\mu}_{2,1}, \hat{\mu}_{1,2}, \ldots, \hat{\mu}_{1,p})$ and $\tilde{\boldsymbol{\mu}}_2 = (\hat{\mu}_{1,1}, \hat{\mu}_{2,2}, \ldots, \hat{\mu}_{2,p})$ by exchanging the first coordinate of $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$.

Then it holds for the constructed $\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2$ that

$$
\begin{aligned}
\varphi\left(\boldsymbol{x}_{1,i}^{(\hat{\boldsymbol{w}}_{1,i})}; \hat{\boldsymbol{\mu}}_1^{(\hat{\boldsymbol{w}}_{1,i})}, \hat{\boldsymbol{\Sigma}}_{reg,1}^{(\hat{\boldsymbol{w}}_{1,i})}\right) &= \varphi\left(\boldsymbol{x}_{1,i}^{(\hat{\boldsymbol{w}}_{1,i})}; \tilde{\boldsymbol{\mu}}_1^{(\hat{\boldsymbol{w}}_{1,i})}, \hat{\boldsymbol{\Sigma}}_{reg,1}^{(\hat{\boldsymbol{w}}_{1,i})}\right) \quad \forall i \leq n_1/2 \\
\varphi\left(\boldsymbol{x}_{1,i}^{(\hat{\boldsymbol{w}}_{1,i})}; \hat{\boldsymbol{\mu}}_1^{(\hat{\boldsymbol{w}}_{1,i})}, \hat{\boldsymbol{\Sigma}}_{reg,1}^{(\hat{\boldsymbol{w}}_{1,i})}\right) &= \varphi\left(\boldsymbol{x}_{1,i}^{(\hat{\boldsymbol{w}}_{1,i})}; \tilde{\boldsymbol{\mu}}_2^{(\hat{\boldsymbol{w}}_{1,i})}, \hat{\boldsymbol{\Sigma}}_{reg,2}^{(\hat{\boldsymbol{w}}_{1,i})}\right) \quad \forall i \geq 1 + n_2/2 \\
\varphi\left(\boldsymbol{x}_{2,i}^{(\hat{\boldsymbol{w}}_{2,i})}; \hat{\boldsymbol{\mu}}_2^{(\hat{\boldsymbol{w}}_{2,i})}, \hat{\boldsymbol{\Sigma}}_{reg,2}^{(\hat{\boldsymbol{w}}_{2,i})}\right) &= \varphi\left(\boldsymbol{x}_{2,i}^{(\hat{\boldsymbol{w}}_{2,i})}; \tilde{\boldsymbol{\mu}}_2^{(\hat{\boldsymbol{w}}_{2,i})}, \hat{\boldsymbol{\Sigma}}_{reg,2}^{(\hat{\boldsymbol{w}}_{2,i})}\right) \quad \forall i \leq n_1/2 \\
\varphi\left(\boldsymbol{x}_{2,i}^{(\hat{\boldsymbol{w}}_{2,i})}; \hat{\boldsymbol{\mu}}_2^{(\hat{\boldsymbol{w}}_{2,i})}, \hat{\boldsymbol{\Sigma}}_{reg,2}^{(\hat{\boldsymbol{w}}_{2,i})}\right) &= \varphi\left(\boldsymbol{x}_{2,i}^{(\hat{\boldsymbol{w}}_{2,i})}; \tilde{\boldsymbol{\mu}}_1^{(\hat{\boldsymbol{w}}_{2,i})}, \hat{\boldsymbol{\Sigma}}_{reg,1}^{(\hat{\boldsymbol{w}}_{2,i})}\right) \quad \forall i \geq 1 + n_2/2.
\end{aligned}
$$

Thus, the value of the objective function is the same and finite and the constructed estimates $\hat{\boldsymbol{\pi}}$, $\tilde{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{W}}$ are also optimizers. However, $||\hat{\boldsymbol{\mu}}_l(\mathcal{X}_m) - \tilde{\boldsymbol{\mu}}_h(\mathcal{X}_m \cup \mathcal{Y}_m)||_2 \to \infty$ for all $l, h \in \{1, 2\}$ due to Corollary 5.4.2.1c.

e. For ease of notation we drop the superscript m for observations and the explicit dependence of the estimators of $\mathcal{Z}_m$ or $\mathcal{X}_m$. All limits are corresponding to $m \to \infty$. We construct a counter example that shows that the covariance needs to explode if the location estimator is not breaking down within the idealized scenario.

Given an uncontaminated sample $\mathcal{X}$ and one variable j^*, we assume that all cells from variable j^* of the uncontaminated data are positive. The uncontaminated data $\mathcal{X}$ is partitioned into groups $\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^N$ and only one group g' is contaminated with $n_{g'} - h_{g'} + 1$ many cellwise outliers $\mathcal{Y}$, outlying only in variable j^* with negative values. Thus, for any $\boldsymbol{W}_{g'}$ there is always at least one outlying cell in variable j^*, that is observed. The data used in the contaminated case is then

[3]This setting can be generalized to $N > 2$, e.g. by adding groups which consist entirely of one cluster each.

$\mathcal{Z} = \bigcup_{g=1}^{N} \boldsymbol{Z}^g$. For an estimator $\hat{\boldsymbol{W}}(\mathcal{Z})$ let $\tilde{\boldsymbol{y}}$ be an outlier for which variable j^* is observed, $w(\tilde{\boldsymbol{y}})_{j^*} = 0$ and $\hat{w}_{\tilde{\boldsymbol{y}},j^*} = 1$.

Let $\hat{t}_k(\boldsymbol{z})$ denote the probability of an observation $\boldsymbol{z} \in \boldsymbol{Z}_g$ that it belongs to distribution k given the estimates $\hat{\boldsymbol{\pi}}(\mathcal{Z})$, $\hat{\boldsymbol{\mu}}(\mathcal{Z})$, $\hat{\boldsymbol{\Sigma}}(\mathcal{Z})$ and $\hat{\boldsymbol{W}}(\mathcal{Z})$,

$$\hat{t}_k(\boldsymbol{z}) = \frac{\hat{\pi}_{g,k}\varphi\left(\boldsymbol{z}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}; \hat{\boldsymbol{\mu}}_k^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}, \hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}\right)}{\sum_{l=1}^{N} \hat{\pi}_{g,l}\varphi\left(\boldsymbol{z}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}; \hat{\boldsymbol{\mu}}_l^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}, \hat{\boldsymbol{\Sigma}}_{reg,l}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}\right)}.$$

Note that due to the regularity of the covariance estimates the density goes to zero, $\varphi\left(\boldsymbol{z}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}; \hat{\boldsymbol{\mu}}_k^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}, \hat{\boldsymbol{\Sigma}}_{reg,k}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}\right) \to 0$, if $||\boldsymbol{z}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})} - \hat{\boldsymbol{\mu}}_k^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}||_2 \to \infty$ and thus $\hat{t}_k(\boldsymbol{z}) \to 0$. Since there are N many possible distributions, for $\tilde{\boldsymbol{y}}$ there exists a distribution k^* with $\hat{t}_{k^*}(\tilde{\boldsymbol{y}}) \geq \frac{1}{N} > 0$.

Upon convergence of the EM-algorithm the location estimate of the j^*-th variable of distribution k^* is

$$\hat{\mu}_{k^*j^*}(\mathcal{Z}) = \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{z} \in \boldsymbol{Z}_g} \hat{t}_{k^*}(\boldsymbol{z}) \hat{z}_{j^*},$$

with $\bar{t}_{k^*} = \sum_{g=1}^{N} \sum_{\boldsymbol{z} \in \boldsymbol{Z}_g} \hat{t}_{k^*}(\boldsymbol{z})$ and $\hat{z}_{j^*}$ being the imputed value of $\boldsymbol{z}$ for variable j^*. For $\hat{w}_{\boldsymbol{z},j^*} = 1$ it is equal to z_{j^*} and for $\hat{w}_{\boldsymbol{z},j^*} = 0$ it is equal to

$$\hat{\mu}_{k^*j^*} + \hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(j^*|\hat{\boldsymbol{w}}_{\boldsymbol{z}})} \left(\hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}}|\hat{\boldsymbol{w}}_{\boldsymbol{z}})}\right)^{-1} \left(\boldsymbol{z}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})} - \hat{\boldsymbol{\mu}}_{k^*}^{(\hat{\boldsymbol{w}}_{\boldsymbol{z}})}\right),$$

where $\hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(j^*|\hat{\boldsymbol{w}}_{\boldsymbol{z}})}$ indicates the submatrix $\hat{\boldsymbol{\Sigma}}_{reg,k^*}$ consisting of the j^*-th row and the observed variables of $\boldsymbol{z}$ as columns, see also Equations (5.3.2) and (5.3.3).

Denoting the set of observations of $\mathcal{Z}$ where variable j^* is observed as $\mathcal{O}_{j^*} =$

$\{\boldsymbol{z} \in \mathcal{Z} : \hat{w}_{\boldsymbol{z},j^*} = 1\}$, we can separate the sum term into

$$
\begin{aligned}
\hat{\mu}_{k^*j^*}(\mathcal{Z}) &= \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{z} \in \boldsymbol{Z}_g} \hat{t}_{k^*}(\boldsymbol{z}) \hat{z}_{j^*} \\
&= \frac{1}{\bar{t}_{k^*}} \sum_{g \neq g'} \sum_{\boldsymbol{x} \in \boldsymbol{Z}_g} \hat{t}_{k^*}(\boldsymbol{x}) \hat{x}_{j^*} + \frac{1}{\bar{t}_{k^*}} \sum_{\boldsymbol{z} \in \boldsymbol{Z}_{g'}} \hat{t}_{k^*}(\boldsymbol{z}) \hat{z}_{j^*} \\
&= \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{x} \in \boldsymbol{Z}_g \cap \mathcal{X}} \hat{t}_{k^*}(\boldsymbol{x}) \hat{x}_{j^*} + \frac{1}{\bar{t}_{k^*}} \sum_{\boldsymbol{y} \in \boldsymbol{Z}_{g'} \cap \mathcal{Y}} \hat{t}_{k^*}(\boldsymbol{y}) \hat{y}_{j^*} \\
&= \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{x} \in \boldsymbol{Z}_g \cap \mathcal{X} \cap \mathcal{O}_{j^*}} \hat{t}_{k^*}(\boldsymbol{x}) \hat{x}_{j^*} + \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{x} \in \boldsymbol{Z}_g \cap \mathcal{X} \cap \mathcal{O}_{j^*}^c} \hat{t}_{k^*}(\boldsymbol{x}) \hat{x}_{j^*} \\
&\quad + \frac{1}{\bar{t}_{k^*}} \sum_{\boldsymbol{y} \in \boldsymbol{Z}_{g'} \cap \mathcal{Y} \cap \mathcal{O}_{j^*}} \hat{t}_{k^*}(\boldsymbol{y}) \hat{y}_{j^*} + \frac{1}{\bar{t}_{k^*}} \sum_{\boldsymbol{y} \in \boldsymbol{Z}_{g'} \cap \mathcal{Y} \cap \mathcal{O}_{j^*}^c} \hat{t}_{k^*}(\boldsymbol{y}) \hat{y}_{j^*} \\
&= \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{x} \in \boldsymbol{Z}_g \cap \mathcal{X} \cap \mathcal{O}_{j^*}} \hat{t}_{k^*}(\boldsymbol{x}) x_{j^*} + \frac{1}{\bar{t}_{k^*}} \sum_{\boldsymbol{y} \in \boldsymbol{Z}_{g'} \cap \mathcal{Y} \cap \mathcal{O}_{j^*}} \hat{t}_{k^*}(\boldsymbol{y}) y_{j^*} \\
&\quad + \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{x} \in \boldsymbol{Z}_g \cap \mathcal{X} \cap \mathcal{O}_{j^*}^c} \hat{t}_{k^*}(\boldsymbol{x}) \Big[\hat{\mu}_{k^*j^*}(\mathcal{Z}) + \hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(j^*|\hat{\boldsymbol{w}}_{\boldsymbol{x}})} \left(\hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(\hat{\boldsymbol{w}}_{\boldsymbol{x}}|\hat{\boldsymbol{w}}_{\boldsymbol{x}})} \right)^{-1} \\
&\quad \times \left(\boldsymbol{x}^{(\hat{\boldsymbol{w}}_{\boldsymbol{x}})} - \hat{\boldsymbol{\mu}}_{k^*}(\mathcal{Z})^{(\hat{\boldsymbol{w}}_{\boldsymbol{x}})} \right) \Big] + \frac{1}{\bar{t}_{k^*}} \sum_{\boldsymbol{y} \in \boldsymbol{Z}_{g'} \cap \mathcal{Y} \cap \mathcal{O}_{j^*}^c} \hat{t}_{k^*}(\boldsymbol{y}) \Big[\hat{\mu}_{k^*j^*}(\mathcal{Z}) \\
&\quad + \hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(j^*|\hat{\boldsymbol{w}}_{\boldsymbol{y}})} \left(\hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(\hat{\boldsymbol{w}}_{\boldsymbol{y}}|\hat{\boldsymbol{w}}_{\boldsymbol{y}})} \right)^{-1} \left(\boldsymbol{y}^{(\hat{\boldsymbol{w}}_{\boldsymbol{y}})} - \hat{\boldsymbol{\mu}}_{k^*}(\mathcal{Z})^{(\hat{\boldsymbol{w}}_{\boldsymbol{y}})} \right) \Big].
\end{aligned}
$$

Subtracting the estimated location on the uncontaminated sample $\hat{\mu}_{k^*j^*}(\mathcal{X})$ and

using that the location estimator is not breaking down, we further get

$$
\begin{aligned}
&\underbrace{\hat{\mu}_{k^*j^*}(\mathcal{Z}) - \hat{\mu}_{k^*j^*}(\mathcal{X})}_{\text{bounded}} = \\
&= \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{x} \in \boldsymbol{Z}_g \cap \mathcal{X} \cap \mathcal{O}_{j^*}} \hat{t}_{k^*}(\boldsymbol{x}) \underbrace{(x_{j^*} - \hat{\mu}_{k^*j^*}(\mathcal{X}))}_{*} \\
&+ \frac{1}{\bar{t}_{k^*}} \sum_{\boldsymbol{y} \in \boldsymbol{Z}_{g'} \cap \mathcal{Y} \cap \mathcal{O}_{j^*}} \hat{t}_{k^*}(\boldsymbol{y}) \underbrace{(y_{j^*} - \hat{\mu}_{k^*j^*}(\mathcal{X}))}_{\to -\infty} \\
&+ \frac{1}{\bar{t}_{k^*}} \sum_{g=1}^{N} \sum_{\boldsymbol{x} \in \boldsymbol{Z}_g \cap \mathcal{X} \cap \mathcal{O}_{j^*}^c} \hat{t}_{k^*}(\boldsymbol{x}) \\
&\qquad \left[\underbrace{\hat{\mu}_{k^*j^*}(\mathcal{Z}) - \hat{\mu}_{k^*j^*}(\mathcal{X})}_{\text{bounded}} + \hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(j^*|\hat{\boldsymbol{w}}_{\boldsymbol{x}})} \left(\hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(\hat{\boldsymbol{w}}_{\boldsymbol{x}}|\hat{\boldsymbol{w}}_{\boldsymbol{x}})} \right)^{-1} \underbrace{\left(\boldsymbol{x}^{(\hat{\boldsymbol{w}}_{\boldsymbol{x}})} - \hat{\boldsymbol{\mu}}_{k^*}(\mathcal{Z})^{(\hat{\boldsymbol{w}}_{\boldsymbol{x}})} \right)}_{*} \right] \\
&+ \frac{1}{\bar{t}_{k^*}} \sum_{\boldsymbol{y} \in \boldsymbol{Z}_{g'} \cap \mathcal{Y} \cap \mathcal{O}_{j^*}^c} \hat{t}_{k^*}(\boldsymbol{y}) \\
&\qquad \left[\underbrace{\hat{\mu}_{k^*j^*}(\mathcal{Z}) - \hat{\mu}_{k^*j^*}(\mathcal{X})}_{\text{bounded}} + \hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(j^*|\hat{\boldsymbol{w}}_{\boldsymbol{y}})} \left(\hat{\boldsymbol{\Sigma}}_{reg,k^*}^{(\hat{\boldsymbol{w}}_{\boldsymbol{y}}|\hat{\boldsymbol{w}}_{\boldsymbol{y}})} \right)^{-1} \left(\boldsymbol{y}^{(\hat{\boldsymbol{w}}_{\boldsymbol{y}})} - \hat{\boldsymbol{\mu}}_{k^*}(\mathcal{Z})^{(\hat{\boldsymbol{w}}_{\boldsymbol{y}})} \right) \right].
\end{aligned}
$$

Due to Corollary 5.4.2.1a. the objective function of the uncontaminated sample is finite and due to Theorem 5.4.2.1, part a. and c., the estimated covariances on the uncontaminated sample are bounded and regular. Since we assume that the location estimator is not breaking down, variables cannot be separated (otherwise a similar counter example to part d. can be constructed). Thus, for all $\boldsymbol{x} \in \mathcal{X}$ there exists k such that $|\boldsymbol{x}^{(\boldsymbol{w})} - \hat{\boldsymbol{\mu}}_k^{(\boldsymbol{w})}(\mathcal{X})|$ bounded for all feasible $\boldsymbol{w}$ – otherwise the objective function would explode – and thus, if $|\boldsymbol{x}^{(\boldsymbol{w})} - \hat{\boldsymbol{\mu}}_l^{(\boldsymbol{w})}(\mathcal{X})| \to \infty$ for $l \neq k$ it follows that $\hat{t}_l(\boldsymbol{x}) \to 0$ and $t_l(\boldsymbol{x})(\boldsymbol{x}^{(\boldsymbol{w})} - \hat{\boldsymbol{\mu}}_l^{(\boldsymbol{w})}(\mathcal{X})) \to 0$. Thus, all subtraction parts marked with $*$ are bounded. The last term $\hat{t}_{k^*}(\boldsymbol{y}) \left(\boldsymbol{y}^{(\hat{\boldsymbol{w}}_{\boldsymbol{y}})} - \hat{\boldsymbol{\mu}}_{k^*}(\mathcal{Z})^{(\hat{\boldsymbol{w}}_{\boldsymbol{y}})} \right)$ is also bounded, since outliers are only outlying in variable j^* and otherwise they are part of one cluster. Thus, with the same argument as for uncontaminated data, the term is bounded.

Since $\hat{t}_{k^*}(\tilde{\boldsymbol{y}}) \geq 1/N$ and $\tilde{\boldsymbol{y}} \in \boldsymbol{Z}_{g'} \cap \mathcal{Y} \cap \mathcal{O}_{j^*}$ the whole sum of $\in \boldsymbol{Z}_{g'} \cap \mathcal{Y} \cap \mathcal{O}_{j^*}$ goes to minus infinity. To enable the equality of both sides, at least one of the covariances needs to explode (in variable j^*) to counteract the exploding sum.

□

D.3 Algorithm

In this section details on initialization and additional derivations for the EM-step are provided.

Initialization

First, all data sets are standardized robustly on a global scale (meaning as if the group structure is not known) using the wrapped location (see also default options in function `estLocScale` from the R-package `cellWise` Raymaekers et al., 2023). This leads to global scale and shift invariance and is helpful to stabilize the regularization approach based on the condition number of the estimated covariance matrices. For a given α the initial estimate for $\hat{\boldsymbol{\pi}}^0$ is

$$\hat{\boldsymbol{\pi}}^0 = \begin{bmatrix} \alpha & \frac{1-\alpha}{N-1} & \cdots & \frac{1-\alpha}{N-1} \\ \frac{1-\alpha}{N-1} & \alpha & \cdots & \frac{1-\alpha}{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-\alpha}{N-1} & \frac{1-\alpha}{N-1} & \cdots & \alpha \end{bmatrix}.$$

Then the other initial values are estimated for each group separately according to the following steps:

1. Based on the scaled and centered data sets, local robust scales $\hat{\sigma}_{k,j}$ for group k and variable j are calculated using the univariate MCD. The regularization matrices are then defined as $\boldsymbol{T}_k = \text{diag}(\hat{\sigma}_{k,1}, \ldots, \hat{\sigma}_{k,p})$.

2. Define the condition number to achieve for distribution k as

$$\kappa_k = \max\left\{100, 1.1\frac{\lambda_1(\boldsymbol{T}_k)}{\lambda_p(\boldsymbol{T}_k)}\right\}.$$

3. We use the DDCW as in Raymaekers and Rousseeuw (2023), applied separately for each group, to get initial estimates $\hat{\boldsymbol{\Sigma}}^0_{reg,k}$ and $\hat{\boldsymbol{\mu}}^0_k$. While this approach is not feasible in normal clustering, here we assume that each group has a main distribution enforced by Equation (5.2.7). Thus, taking a robust estimate of the covariance and mean of the main bulk of the observations for each group separately is reasonable and a good initial estimate of the corresponding main distribution. To ensure regularity also in cases with low number of observation in a group k, each time a covariance is calculated during the DDCW-algorithm, it is regularized with regularization matrix $\boldsymbol{T}_k$ and an adaptive regularization factor ρ_k ensuring a maximal condition number of κ_k.

4. Similar to the initialization in Raymaekers and Rousseeuw (2023) the entries of the matrices $\boldsymbol{W}^0$ are all set to one.

After the convergence of the algorithm all data are rescaled to the original scale.

EM-Step

The Expectation-Maximization Algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) is often used to find maximum likelihood estimates in setting where data is incomplete - meaning that some random variables are not observed. Here, this includes the values of missing cells indicated by the given $\boldsymbol{W}$ and the class of an observation which is an often used approach in the context of mixture models.

For each observation $\boldsymbol{x}_{g,i}$ a binary random variable $z_{g,i,k}$ indicates whether it was drawn from distribution k. The likelihood resulting from including the additional random variables $z_{g,i,k}$ is called the *complete log-likelihood* and the resulting objective function the *complete objective function* $\text{CObj}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{W}, \boldsymbol{Z})$ is -2 times

$$\sum_{g=1}^{N}\sum_{i=1}^{n_g}\left[\sum_{\substack{k=1\\ \pi_{g,k}\neq 0}}^{N} z_{g,i,k}\ln\left(\pi_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})};\boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})},\boldsymbol{\Sigma}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)\right)+\sum_{j=1}^{p} q_{g,i,j}(1-w_{g,ij})\right],$$

where $\boldsymbol{Z}$ includes all random variables $z_{g,i,k}$. When taking the conditional expectation of $z_{g,i,k}$,

$$t_{g,i,k}=\mathbb{E}[z_{g,i,k}|\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{W}]=\frac{\pi_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})};\boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})},\boldsymbol{\Sigma}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)}{\sum_{l=1}^{N}\pi_{g,l}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})};\boldsymbol{\mu}_l^{(\boldsymbol{w}_{g,i})},\boldsymbol{\Sigma}_{reg,l}^{(\boldsymbol{w}_{g,i})}\right)},$$

we can formulate the *expected objective function* $\text{EObj}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{W})$, which is -2 times

$$\sum_{g=1}^{N}\sum_{i=1}^{n_g}\left[\sum_{\substack{k=1\\ \pi_{g,k}\neq 0}}^{N} t_{g,i,k}\ln\left(\pi_{g,k}\varphi\left(\boldsymbol{x}_{g,i}^{(\boldsymbol{w}_{g,i})};\boldsymbol{\mu}_k^{(\boldsymbol{w}_{g,i})},\boldsymbol{\Sigma}_{reg,k}^{(\boldsymbol{w}_{g,i})}\right)\right)+\sum_{j=1}^{p} q_{g,i,j}(1-w_{g,ij})\right]. \tag{D.7}$$

The Expectation-Maximization algorithm then leverages that we can iteratively take the expectation and then maximize the expected objective function in Equation (D.7). Overall this approach gives us at least the same or more optimal next estimates after each iteration.

The extension of the maximization step regarding the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the Gaussian Mixture Model with missing values (Eirola et al., 2014) to the multi-group GMM with missing values is straight forward since the group structure can be ignored once the conditional expectation of $z_{g,i,k}$ is calculated.

The only difference is the estimation of the mixture probabilities $\boldsymbol{\pi}$ due to the constraint $\pi_{g,g} \geq \alpha$ and $\sum_{k=1}^{N}\pi_{g,k}=1$ for all $g=1,\ldots,N$. To find the optimal mixture probability the Karush-Kuhn-Tucker theorem can be applied. We set the derivative of the expected objective function in Equation (D.7) with respect to $\pi_{g,l}$ to

zero, then the following conditions have to hold

$$\begin{aligned}\frac{\partial[EObj + \lambda(1 - \sum_{k=1}^{N} \pi_{g,k}) + \mu(\alpha - \pi_{g,g})]}{\partial \pi_{g,l}} &= 0\\ \mu(\alpha - \pi_{g,g}) &= 0\\ \mu &\geq 0\\ 1 - \sum_{k=1}^{N} \pi_{g,k} &= 0.\end{aligned}$$

Plugging in the concrete formula from Equation (D.7) leads to ($\mathbb{I}$ denoting the indicator function)

$$\begin{aligned}0 =& \frac{-2\sum_{i=1}^{n_g} t_{g,i,l}}{\pi_{g,l}} - \lambda - \mu \mathbb{I}_{l=g}\\ -2\sum_{i=1}^{n_g} t_{g,i,l} =& \lambda \pi_{g,l} + \mu \mathbb{I}_{l=g} \pi_{g,l}\\ -2\sum_{i=1}^{n_g}\sum_{l=1,l\neq g}^{N} t_{g,i,l} =& \lambda \sum_{l=1,l\neq g}^{N} \pi_{g,l} + \mu \sum_{l=1,l\neq g}^{N} \mathbb{I}_{l=g} \pi_{g,l}\\ \lambda =& \frac{-2\sum_{i=1}^{n_g}\sum_{l=1,l\neq g}^{N} t_{g,i,l}}{(1-\pi_{g,g})} = \frac{-2\sum_{i=1}^{n_g}(1 - t_{g,i,g})}{(1-\pi_{g,g})},\end{aligned}$$

where we sum over all $l \neq g$ from the third row on. Plugging λ in leads to

$$\pi_{g,l} = \frac{(1-\pi_{g,g})\sum_{i=1}^{n_g} t_{g,i,l}}{\sum_{i=1}^{n_g}(1 - t_{g,i,g})} = (1-\pi_{g,g})\frac{\frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,l}}{1 - \frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g}}.$$

For the Lagrange parameter μ we finally have

$$\begin{aligned}\frac{-\frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g}}{\pi_{g,g}} + \frac{(1 - \frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g})}{(1-\pi_{g,g})} =& \frac{\mu}{2n_g} \geq 0\\ \frac{\pi_{g,g}}{(1-\pi_{g,g})} \geq& \frac{\frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g}}{(1 - \frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g})}\end{aligned}$$

Since $f(x) = x/(1-x)$ is monotonously increasing, this is fulfilled if $\pi_{g,g} \geq \frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g}$. Thus, if the inequality is strict, $\mu > 0$ and $\pi_{g,g} = \alpha$. Otherwise, $\pi_{g,g} = \frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g}$ is a feasible solution which is equal to the unconstrained minimization problem. Overall, we have

$$\pi_{g,g} = \max\left\{\alpha, \frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g}\right\}, \quad \pi_{g,l} = (1-\pi_{g,g})\frac{\frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,l}}{1 - \frac{1}{n_g}\sum_{i=1}^{n_g} t_{g,i,g}}.$$

Also the regularity condition *linear independence constraint qualification* (LICQ) is fulfilled for all feasible $\boldsymbol{\pi}$.

D.4 Additional Simulation Results

In the following subsections additional results for the five settings from Section 5.5 are presented. The settings analyzed are the balanced basic setting ($N = 2, p = 10, n_1 = n_2 = 100$), an unbalanced setting ($N = 2, p = 10, n_1 = 100, n_2 = 50$) as well as a balanced setting with nearly as many variables as observations per group ($N = 2, p = 20, n_1 = 30, n_2 = 30$), a setting with more groups ($N = 5, p = 10, n_1 = \ldots = n_5 = 100$) and a high-dimensional setting ($N = 2, p = 60, n_1 = n_2 = 40$).

For each setting, the performance of parameter estimation compared to competing methods is visualized as well as the correctness of flagging outlying cells. Moreover, for each setting a table with the number of repetitions considered in the figures is given. They can deviate from the default number of 100 due to the restriction of the cellMCD regarding the number of marginal outliers.

Basic Balanced Setting

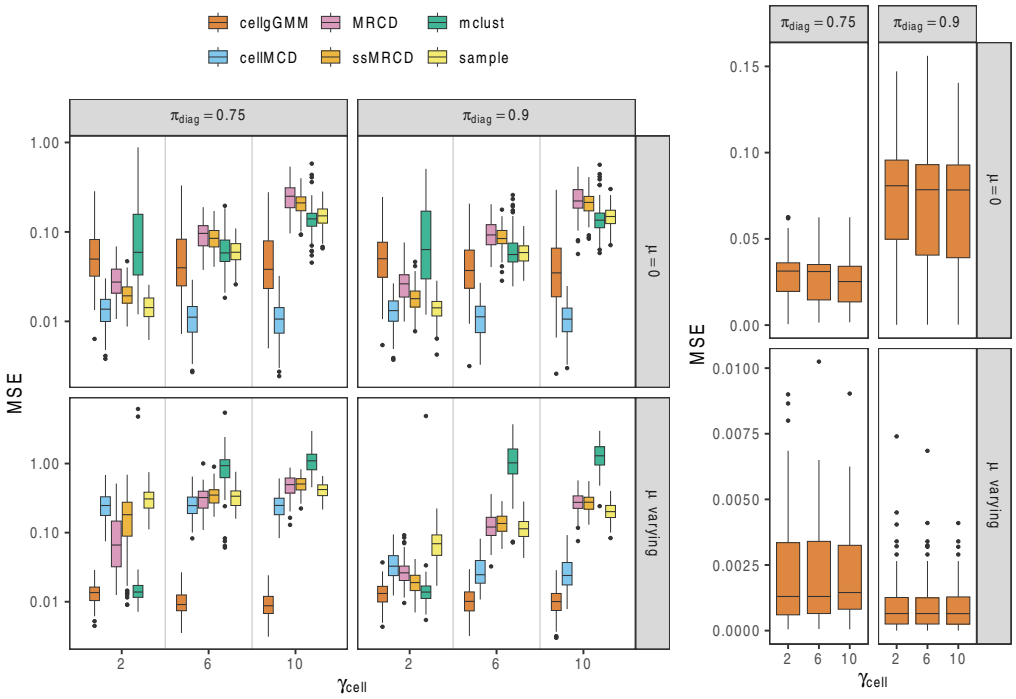


Figure D.1: Parameter estimates for the basic balanced setting ($N = 2, p = 10, n_1 = n_2 = 100$) with Toeplitz structured covariances. In the left panel MSE of the mean estimation and in the right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

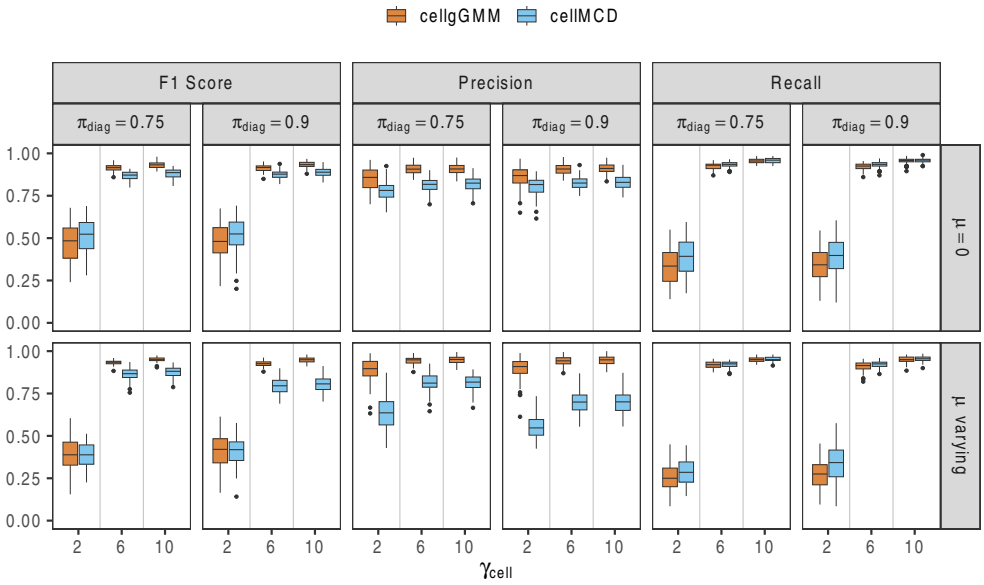


Figure D.2: Performance of cellwise outlier detection in the basic balanced setting ($N = 2, p = 10, n_1 = n_2 = 100$) with Toeplitz structured covariances evaluated by precision, recall and F1-score.

γ_{cell}	π_{diag}	μ	#
10	0.75	0	100
10	0.75	varying	58
10	0.90	0	100
10	0.90	varying	98
6	0.75	0	100
6	0.75	varying	61
6	0.90	0	100
6	0.90	varying	99
2	0.75	0	100
2	0.75	varying	84
2	0.90	0	100
2	0.90	varying	100

(a) Toeplitz structure.

γ_{cell}	π_{diag}	μ	#
10	0.75	0	100
10	0.75	varying	100
10	0.90	0	100
10	0.90	varying	100
6	0.75	0	100
6	0.75	varying	100
6	0.90	0	100
6	0.90	varying	100
2	0.75	0	100
2	0.75	varying	100
2	0.90	0	100
2	0.90	varying	100

(b) Agostinelli et al. (2015) structure.

Table D.1: Number of successful replications for the two covariance structures in the basic balanced setting ($N = 2, p = 10, n_1 = n_2 = 100$), depending on simulation parameters.

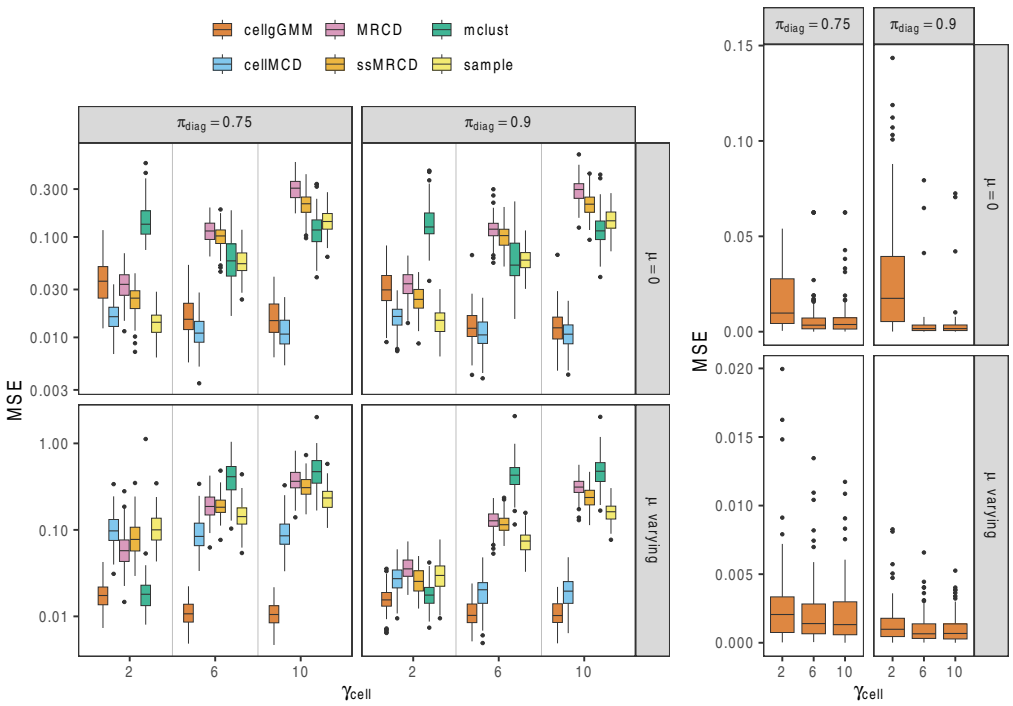


Figure D.3: Parameter estimates for the basic balanced setting ($N = 2, p = 10, n_1 = n_2 = 100$) with covariances according to Agostinelli et al. (2015). In the left panel MSE of the mean estimation and in the right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

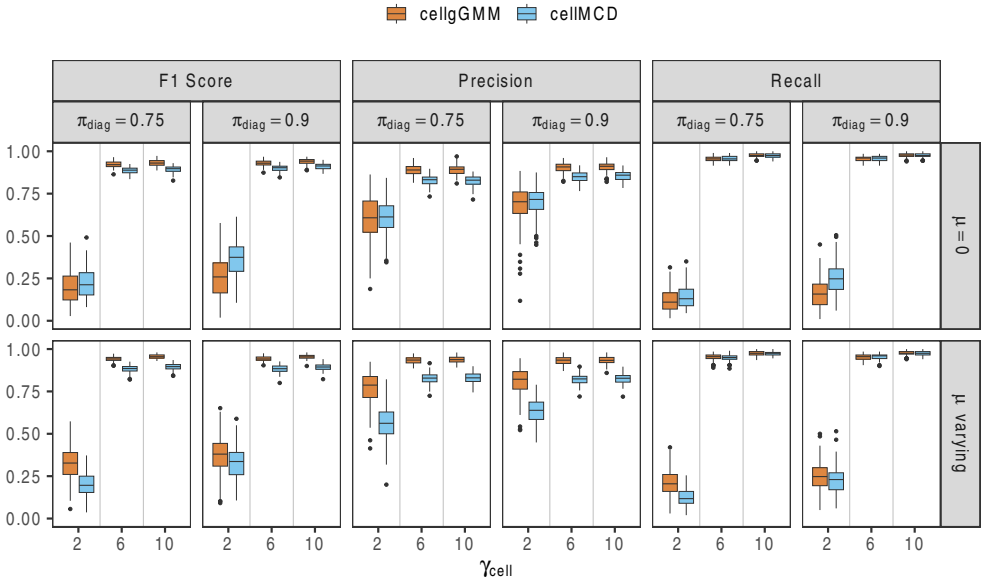


Figure D.4: Performance of cellwise outlier detection in the basic balanced setting ($N = 2, p = 10, n_1 = n_2 = 100$) with covariances according to Agostinelli et al. (2015) evaluated by on precision, recall and F1-score.

Balanced Setting with Increased Group Number

γ_{cell}	π_{diag}	μ	#
10	0.75	0	100
10	0.75	varying	16
10	0.90	0	100
10	0.90	varying	99
6	0.75	0	100
6	0.75	varying	21
6	0.90	0	100
6	0.90	varying	100
2	0.75	0	100
2	0.75	varying	61
2	0.90	0	100
2	0.90	varying	100

(a) Toeplitz structure.

γ_{cell}	π_{diag}	μ	#
10	0.75	0	100
10	0.75	varying	96
10	0.90	0	100
10	0.90	varying	100
6	0.75	0	100
6	0.75	varying	96
6	0.90	0	100
6	0.90	varying	100
2	0.75	0	100
2	0.75	varying	100
2	0.90	0	100
2	0.90	varying	100

(b) Agostinelli et al. (2015) structure.

Table D.2: Number of successful replications for the two covariance structures in the balanced setting with increased number of groups ($N = 5, p = 10, n_1 = \ldots = n_5 = 100$), depending on simulation parameters.

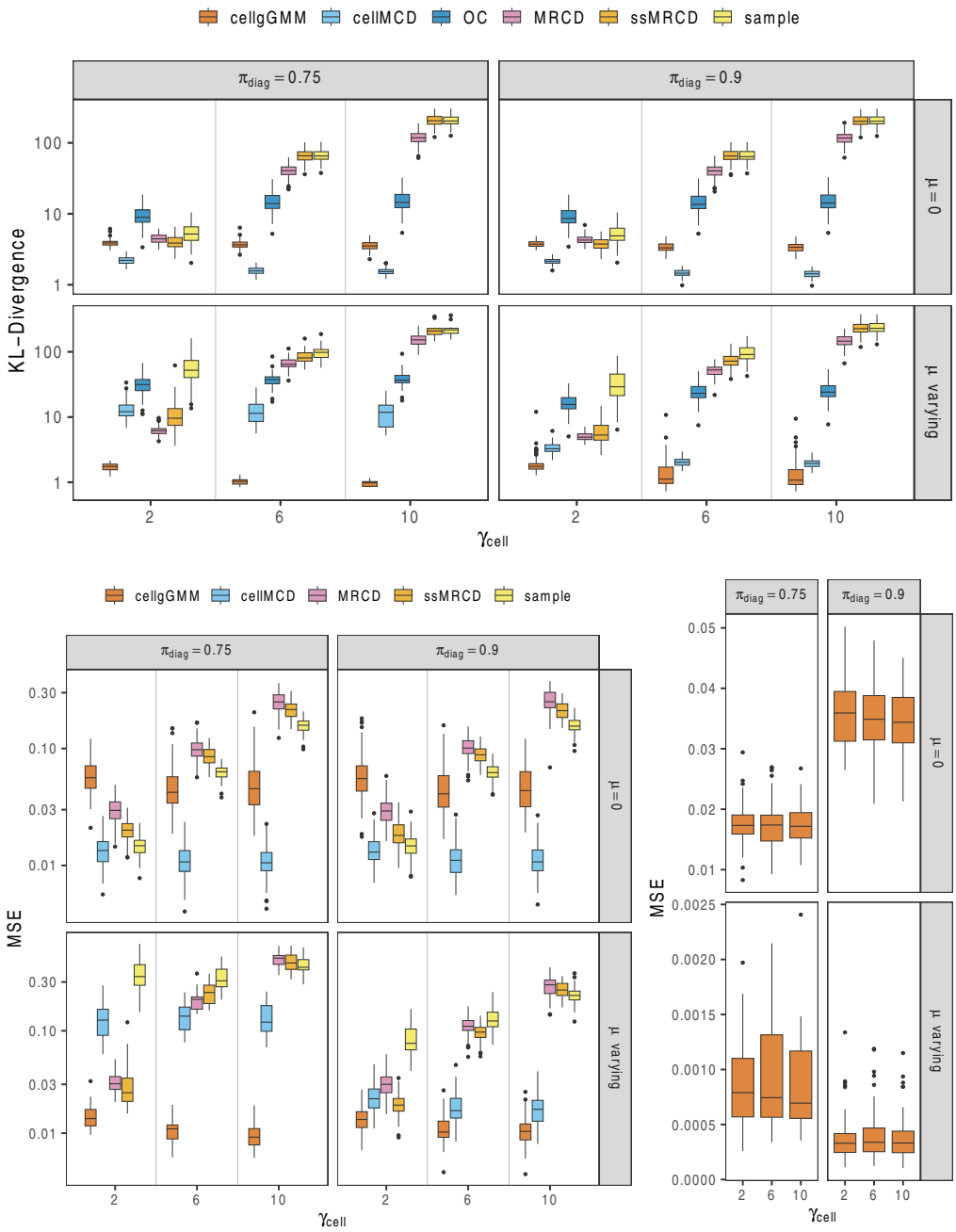


Figure D.5: Parameter estimates for the balanced setting with increased number of groups ($N = 5, p = 10, n_1 = \ldots = n_5 = 100$) and Toeplitz structured covariances. On top the KL-divergence of the covariance estimates. On the bottom left panel MSE of the mean estimation and on the bottom right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

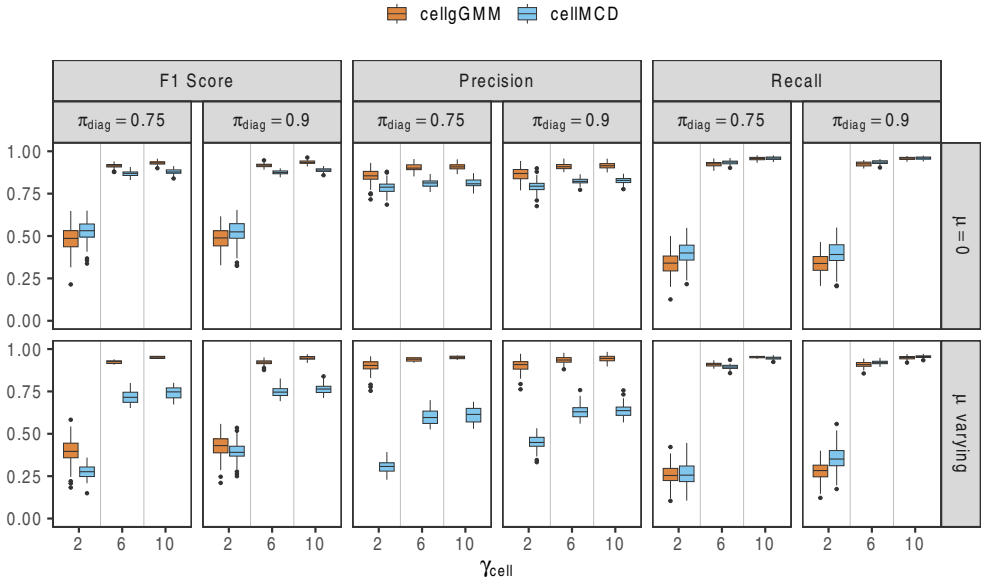


Figure D.6: Performance of cellwise outlier detection in the balanced setting with increased number of groups ($N = 5, p = 10, n_1 = \ldots = n_5 = 100$) and Toeplitz structured covariances evaluated by precision, recall and F1-score.

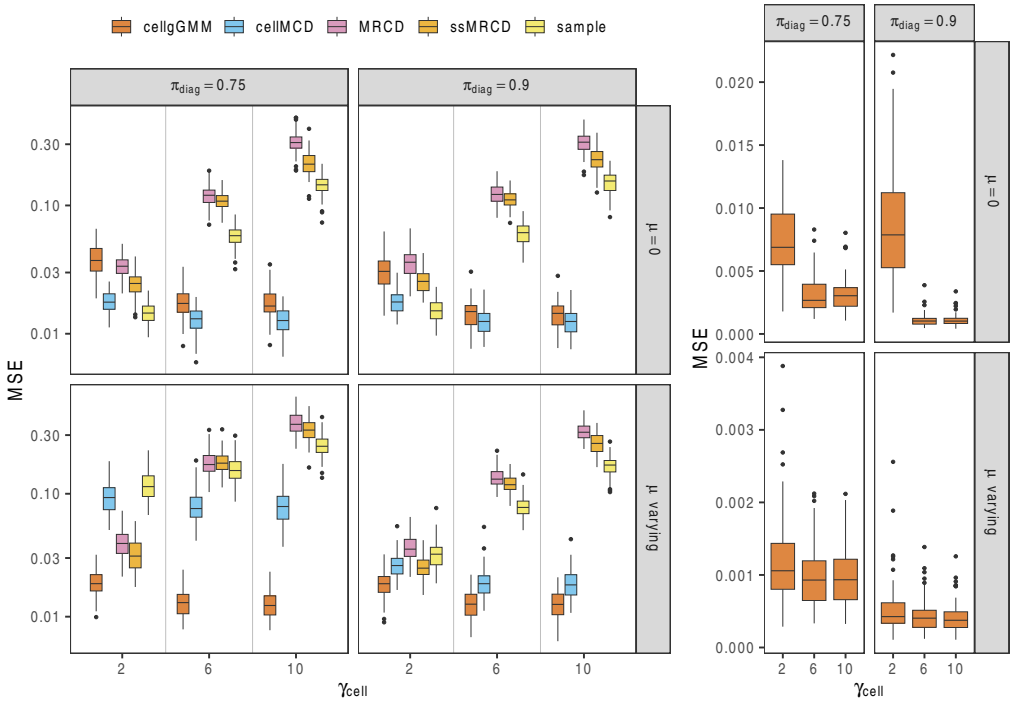


Figure D.7: Parameter estimates for the balanced setting with increased number of groups ($N = 5, p = 10, n_1 = \ldots = n_5 = 100$) and covariances according to Agostinelli et al. (2015). in the left panel MSE of the mean estimation and in the right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

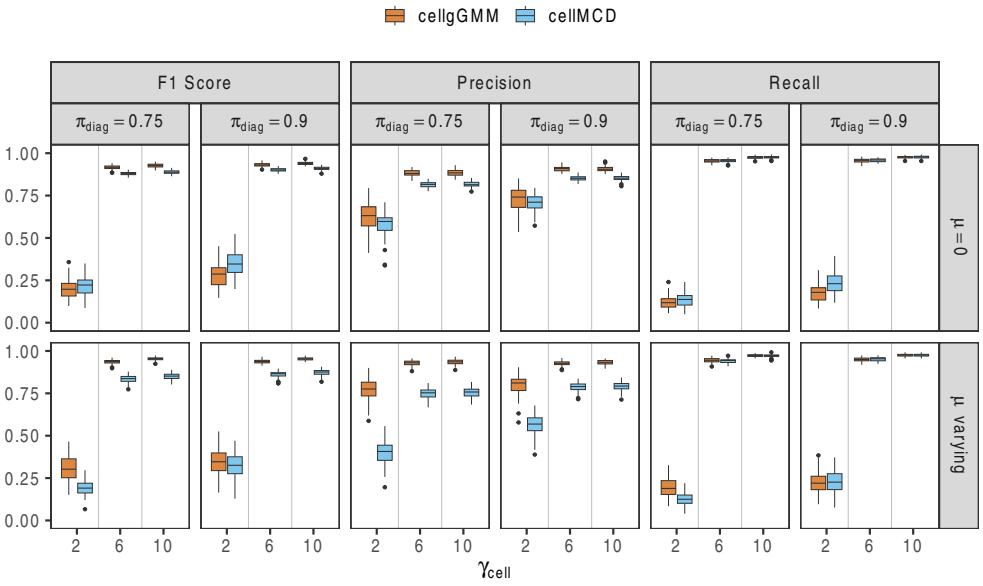


Figure D.8: Performance of cellwise outlier detection in the balanced setting with increased number of groups ($N = 5, p = 10, n_1 = \ldots = n_5 = 100$) and covariances according to Agostinelli et al. (2015) evaluated by on precision, recall and F1-score.

Unbalanced Groups

γ_{cell}	π_{diag}	μ	#
10	0.75	0	100
10	0.75	varying	58
10	0.90	0	100
10	0.90	varying	93
6	0.75	0	100
6	0.75	varying	68
6	0.90	0	100
6	0.90	varying	96
2	0.75	0	100
2	0.75	varying	84
2	0.90	0	100
2	0.90	varying	100

(a) Toeplitz structure.

γ_{cell}	π_{diag}	μ	#
10	0.75	0	100
10	0.75	varying	99
10	0.90	0	100
10	0.90	varying	100
6	0.75	0	100
6	0.75	varying	99
6	0.90	0	100
6	0.90	varying	100
2	0.75	0	100
2	0.75	varying	100
2	0.90	0	100
2	0.90	varying	100

(b) Agostinelli et al. (2015) structure.

Table D.3: Number of successful replications for the two covariance structures in the unbalanced setting ($N = 2, p = 10, n_1 = 100, n_2 = 50$), depending on simulation parameters.

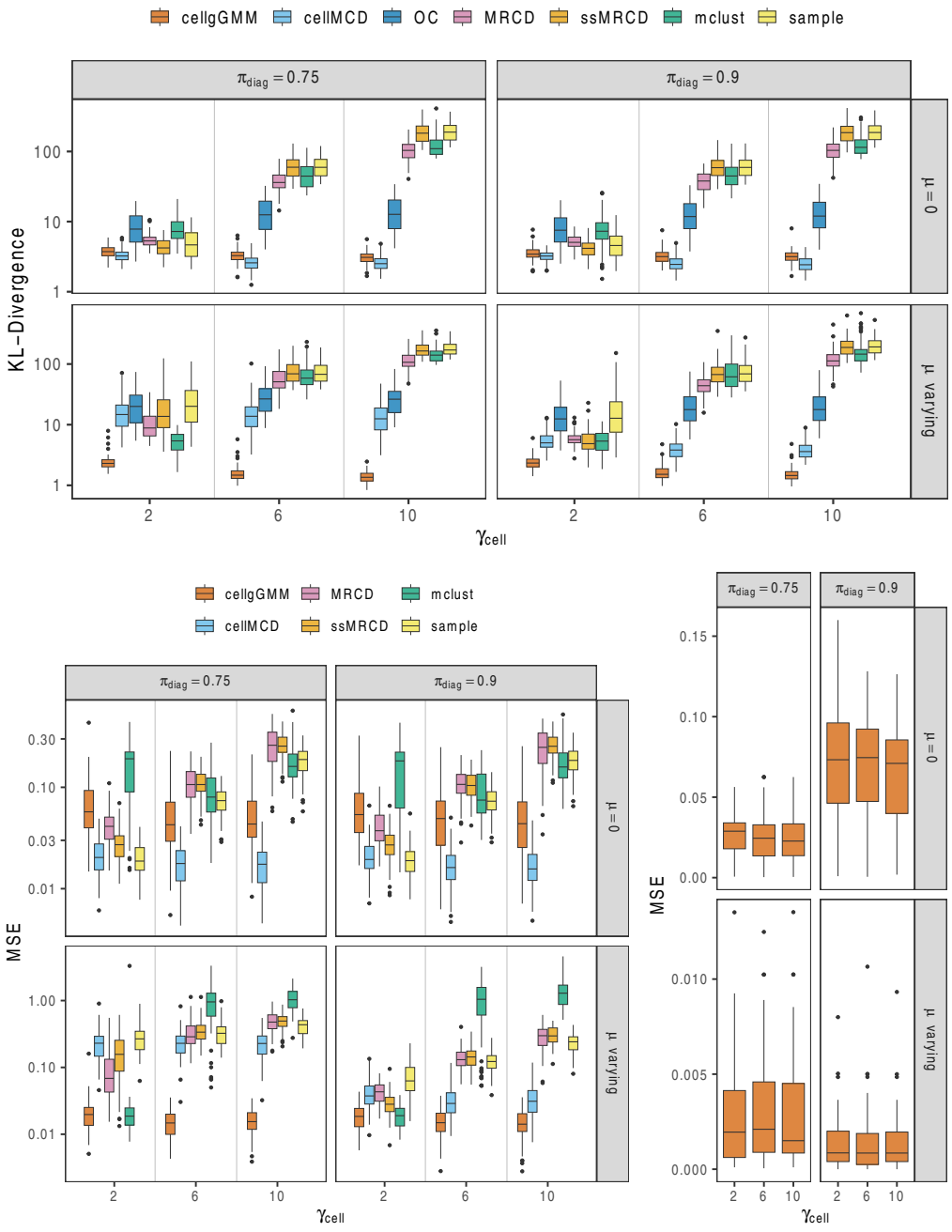


Figure D.9: Parameter estimates for the unbalanced setting ($N = 2, p = 10, n_1 = 100, n_2 = 50$) with Toeplitz structured covariances. On top the KL-divergence of the covariance estimates. On the bottom left panel MSE of the mean estimation and on the bottom right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

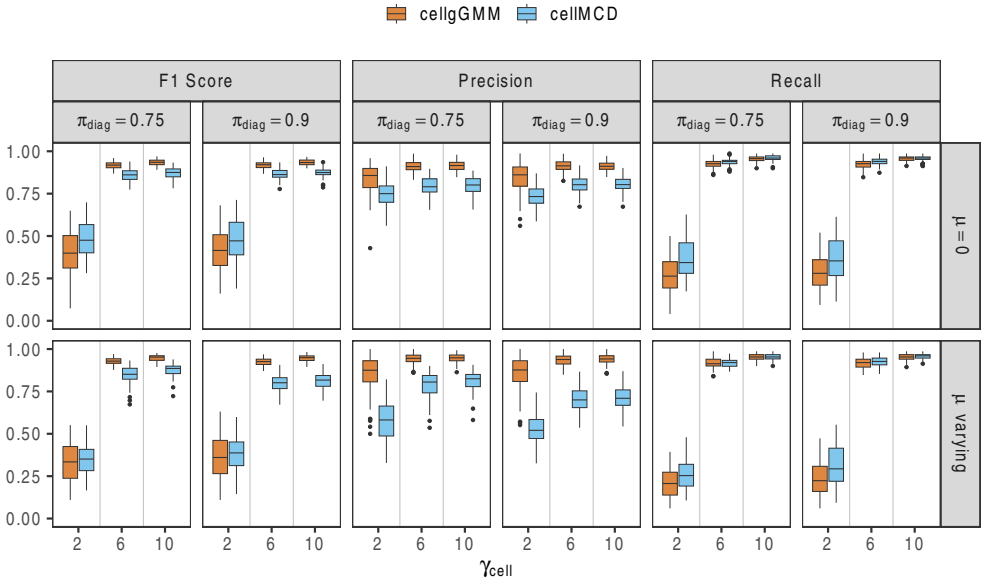


Figure D.10: Performance of cellwise outlier detection in the unbalanced setting ($N = 2, p = 10, n_1 = 100, n_2 = 50$) with Toeplitz structured covariances evaluated by precision, recall and F1-score.

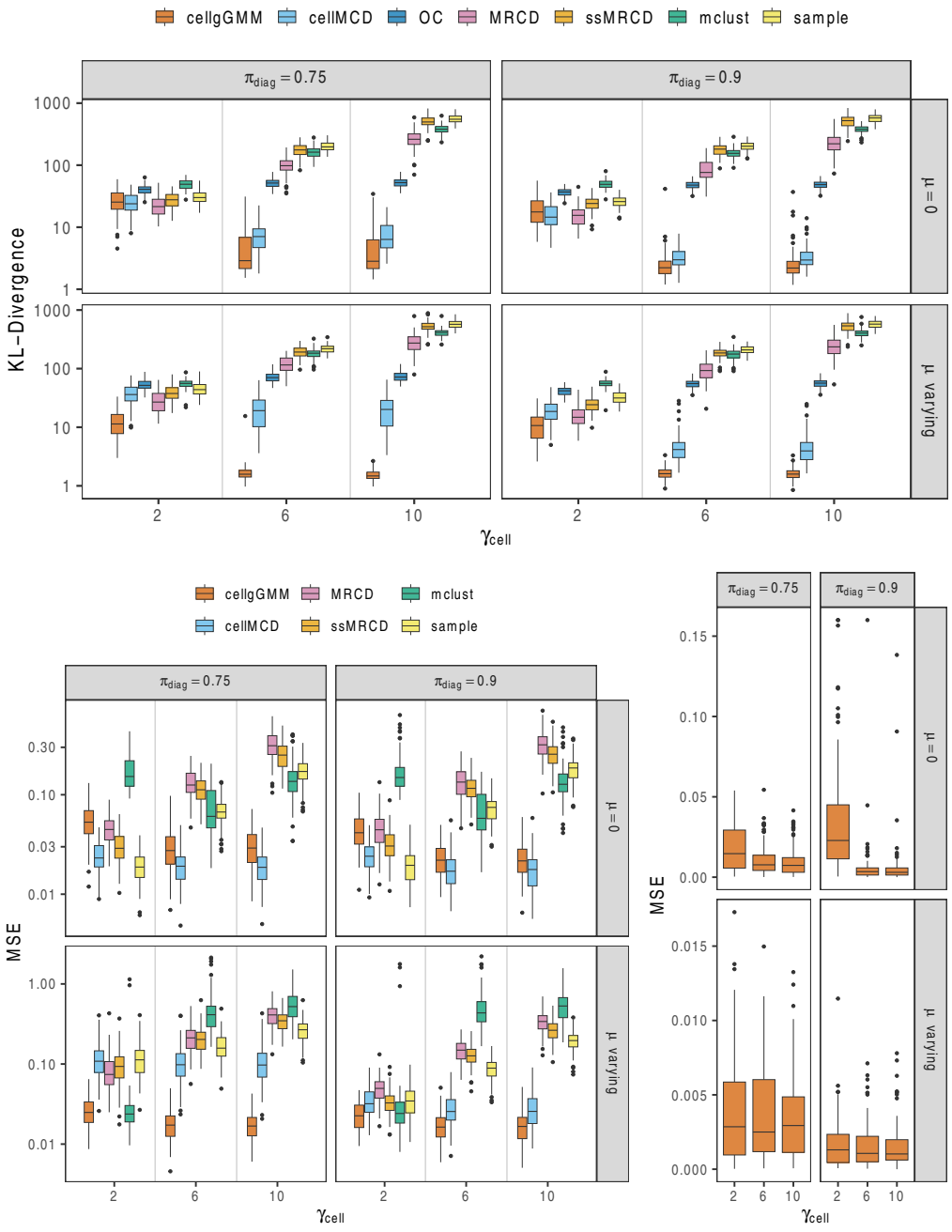


Figure D.11: Parameter estimates for the unbalanced setting ($N = 2, p = 10, n_1 = 100, n_2 = 50$) with covariances according to Agostinelli et al. (2015) On top the KL-divergence of the covariance estimates. On the bottom left panel MSE of the mean estimation and on the bottom right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

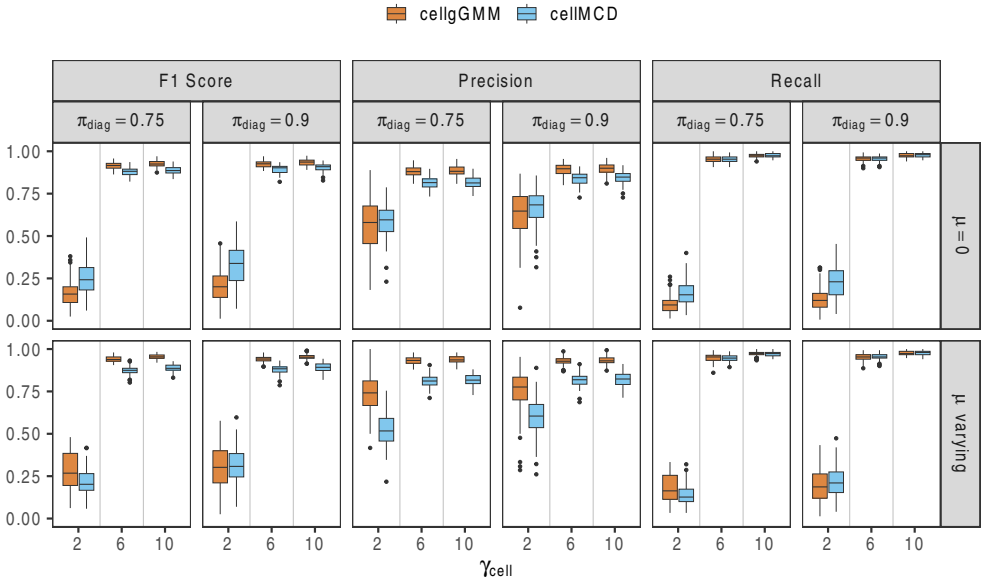


Figure D.12: Performance of cellwise outlier detection in the unbalanced setting ($N = 2, p = 10, n_1 = 100, n_2 = 50$) with covariances according to Agostinelli et al. (2015) evaluated by on precision, recall and F1-score.

Balanced Setting with Similar n and p

γ_{cell}	π_{diag}	μ	#
10	0.75	0	84
10	0.75	varying	12
10	0.90	0	84
10	0.90	varying	53
6	0.75	0	84
6	0.75	varying	14
6	0.90	0	85
6	0.90	varying	55
2	0.75	0	89
2	0.75	varying	38
2	0.90	0	88
2	0.90	varying	85

(a) Toeplitz structure.

γ_{cell}	π_{diag}	μ	#
10	0.75	0	79
10	0.75	varying	81
10	0.90	0	82
10	0.90	varying	82
6	0.75	0	81
6	0.75	varying	82
6	0.90	0	83
6	0.90	varying	82
2	0.75	0	92
2	0.75	varying	92
2	0.90	0	88
2	0.90	varying	85

(b) Agostinelli et al. (2015) structure.

Table D.4: Number of successful replications for the two covariance structures in the balanced setting with similar sized p and n ($N = 2, p = 20, n_1 = 30, n_2 = 30$), depending on simulation parameters.

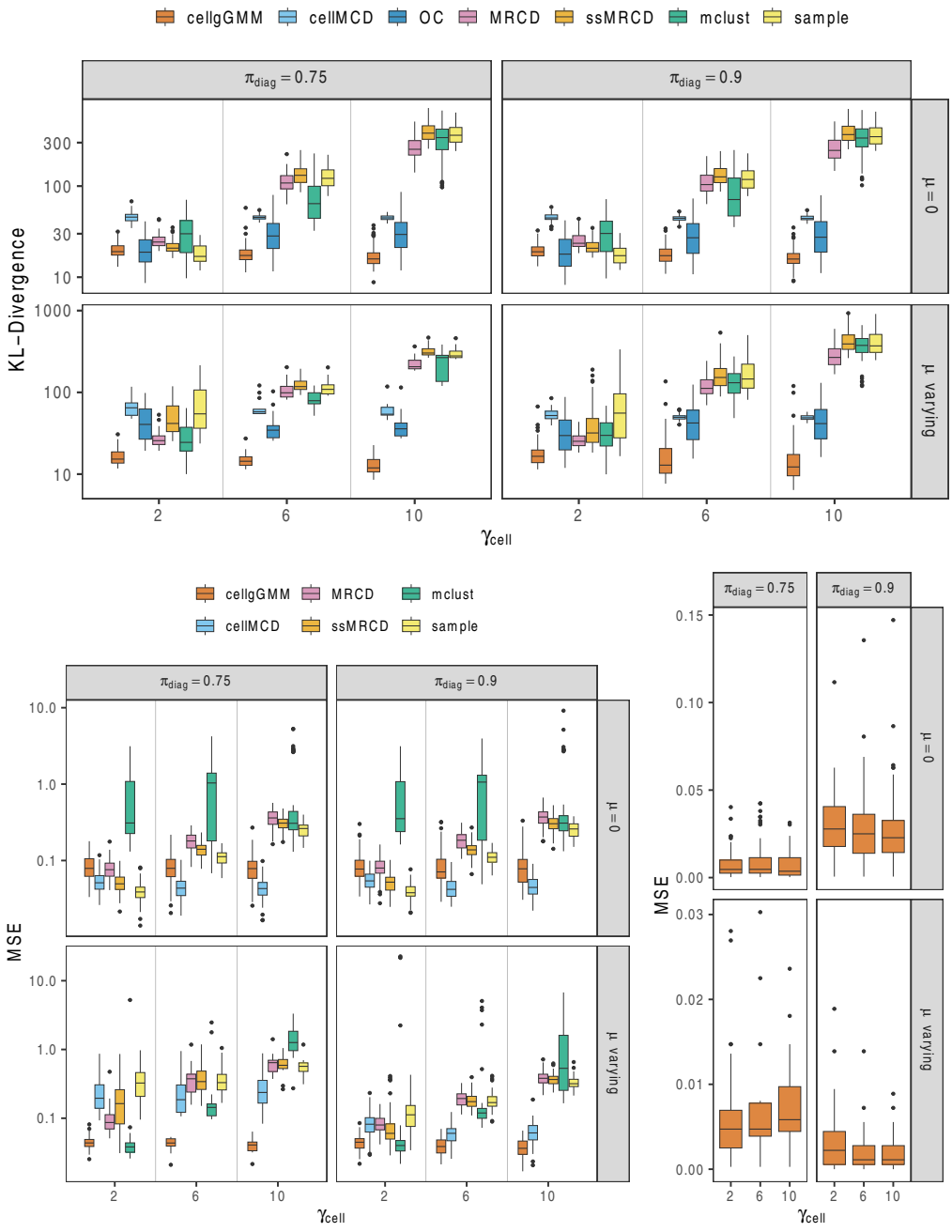


Figure D.13: Parameter estimates for the balanced setting with n close to p ($N = 2, p = 20, n_1 = 30, n_2 = 30$) and Toeplitz structured covariances. On top the KL-divergence of the covariance estimates. On the bottom left panel MSE of the mean estimation and on the bottom right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

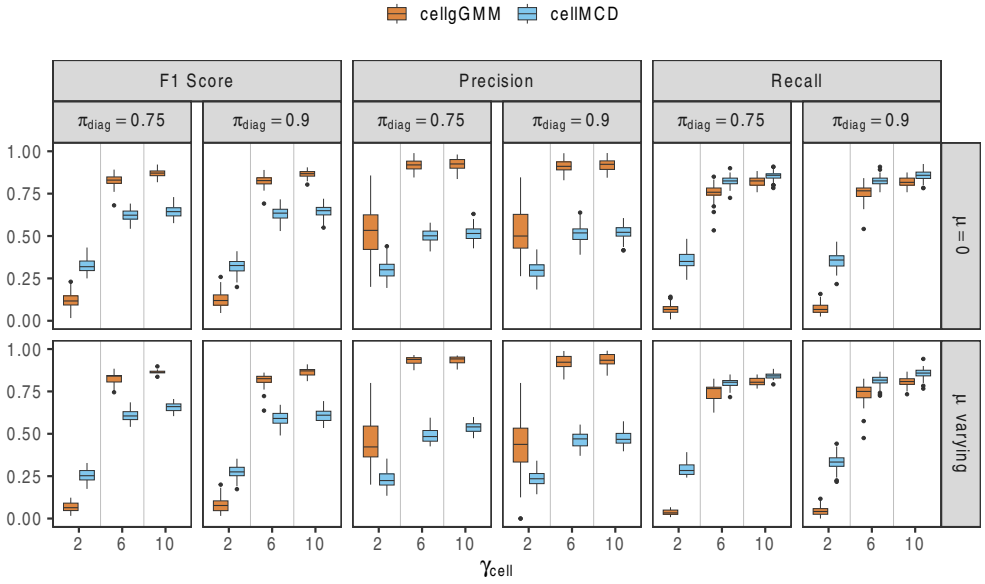


Figure D.14: Performance of cellwise outlier detection in the balanced setting with n close to p ($N = 2, p = 20, n_1 = 30, n_2 = 30$) and Toeplitz structured covariances evaluated by precision, recall and F1-score.

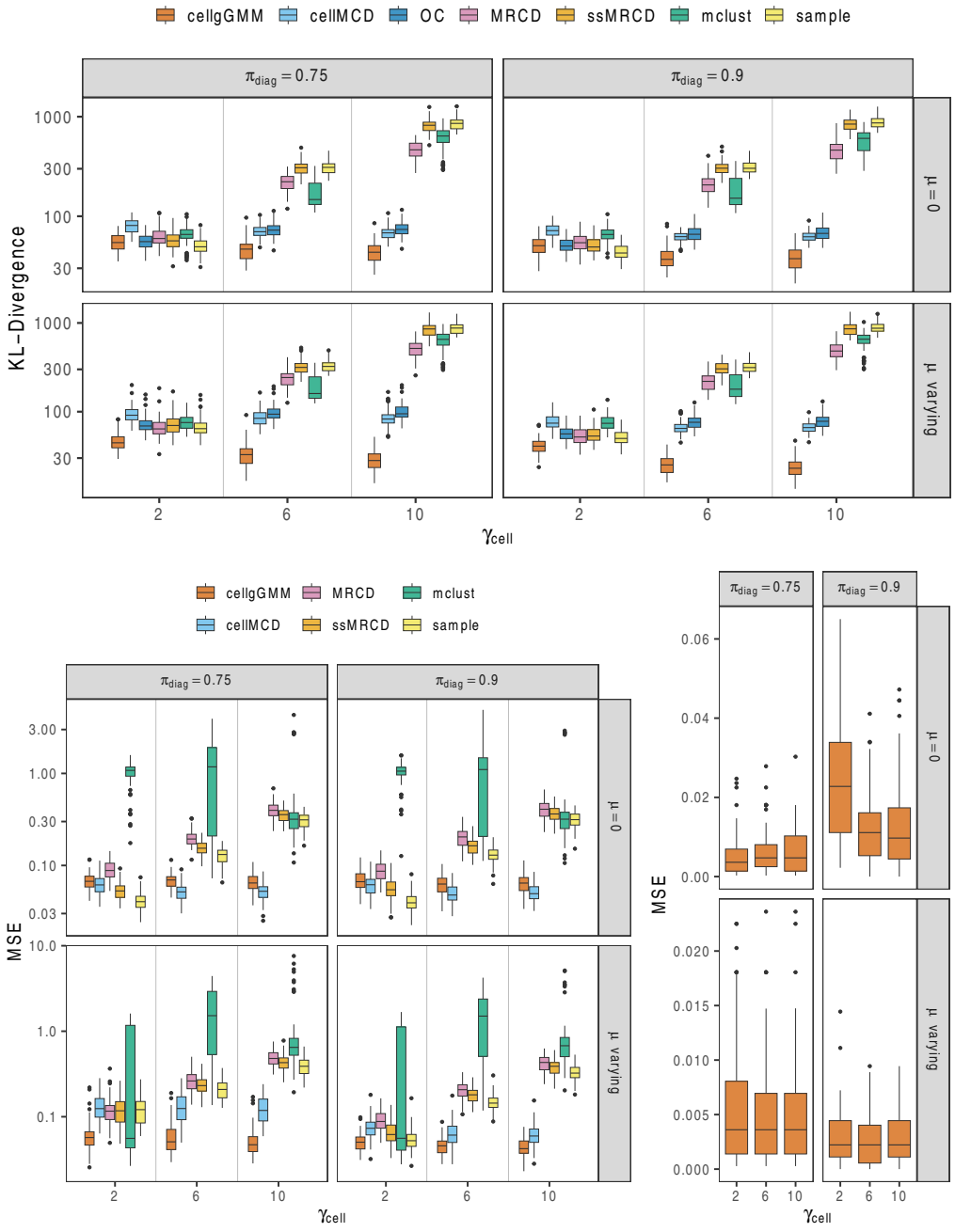


Figure D.15: Parameter estimates for the balanced setting with n close to p ($N = 2, p = 20, n_1 = 30, n_2 = 30$) with covariances according to Agostinelli et al. (2015). On top the KL-divergence of the covariance estimates. On the bottom left panel MSE of the mean estimation and on the bottom right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

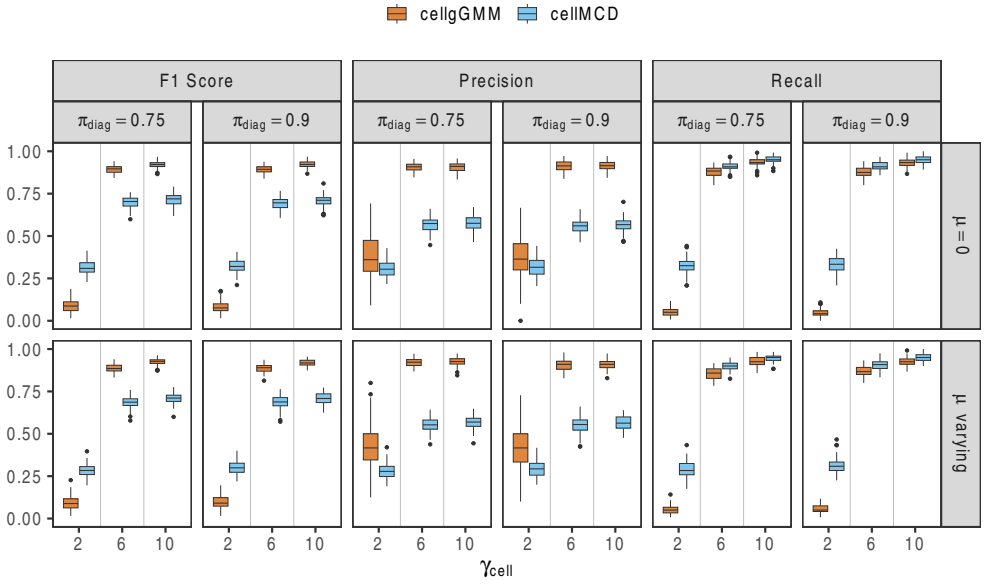


Figure D.16: Performance of cellwise outlier detection in the balanced setting with n close to p ($N = 2, p = 20, n_1 = 30, n_2 = 30$) and covariances according to Agostinelli et al. (2015) evaluated by on precision, recall and F1-score.

High-Dimensional Setting

γ_{cell}	π_{diag}	μ	#
10	0.75	0	100
10	0.75	varying	100
10	0.90	0	100
10	0.90	varying	100
6	0.75	0	100
6	0.75	varying	100
6	0.90	0	100
6	0.90	varying	100
2	0.75	0	100
2	0.75	varying	100
2	0.90	0	100
2	0.90	varying	100

(a) Toeplitz structure.

γ_{cell}	π_{diag}	μ	#
10	0.75	0	100
10	0.75	varying	100
10	0.90	0	100
10	0.90	varying	100
6	0.75	0	100
6	0.75	varying	100
6	0.90	0	100
6	0.90	varying	100
2	0.75	0	100
2	0.75	varying	100
2	0.90	0	100
2	0.90	varying	100

(b) Agostinelli et al. (2015) structure.

Table D.5: Number of successful replications for the two covariance structures in the balanced high-dimensional setting ($N = 2, p = 60, n_1 = n_2 = 40$), depending on simulation parameters.

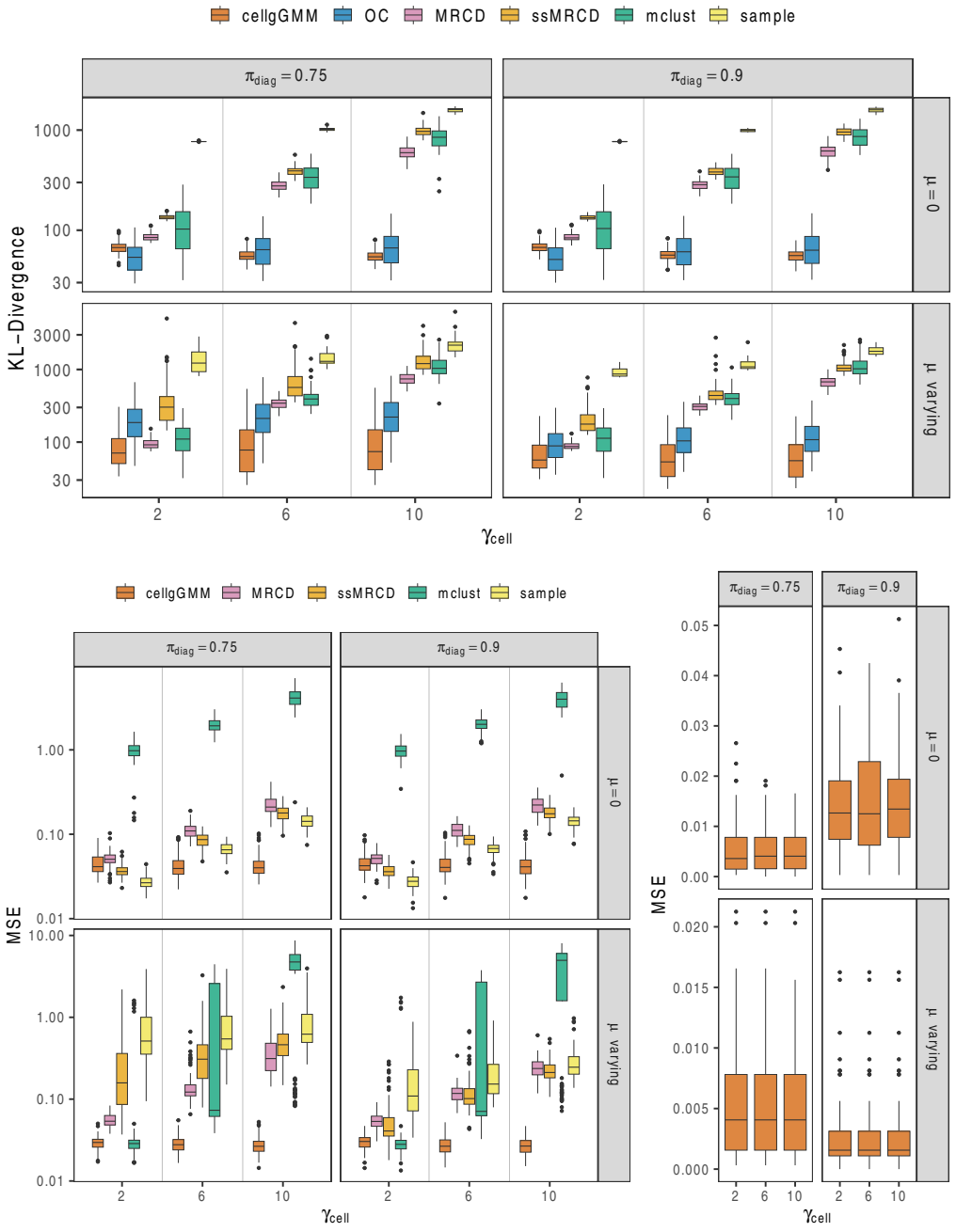


Figure D.17: Parameter estimates for the balanced high-dimensional setting ($N = 2, p = 60, n_1 = n_2 = 40$) with Toeplitz structured covariances. On top the KL-divergence of the covariance estimates. On the bottom left panel MSE of the mean estimation and on the bottom right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

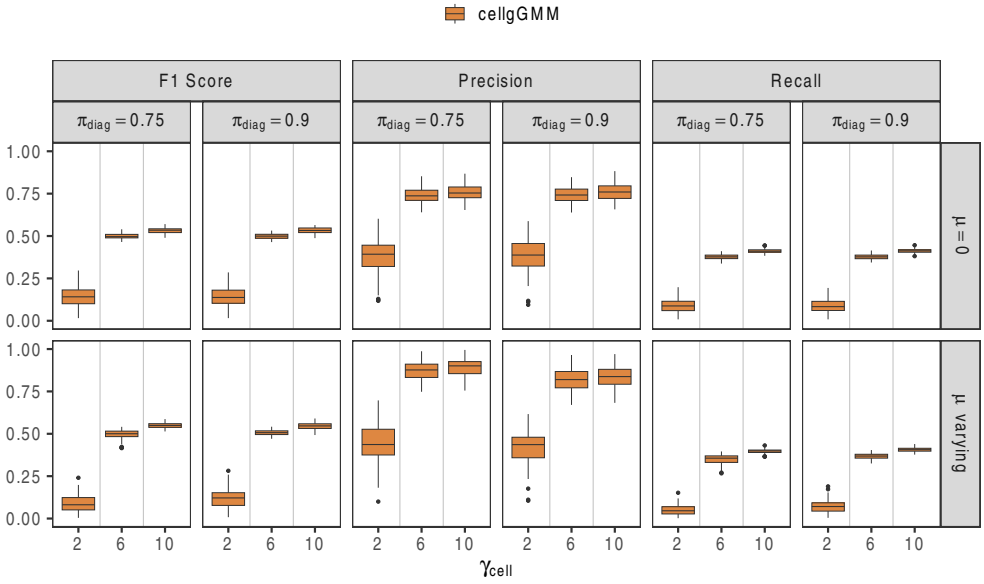


Figure D.18: Performance of cellwise outlier detection in the balanced high-dimensional setting ($N = 2, p = 60, n_1 = n_2 = 40$) with Toeplitz structured covariances evaluated by precision, recall and F1-score.

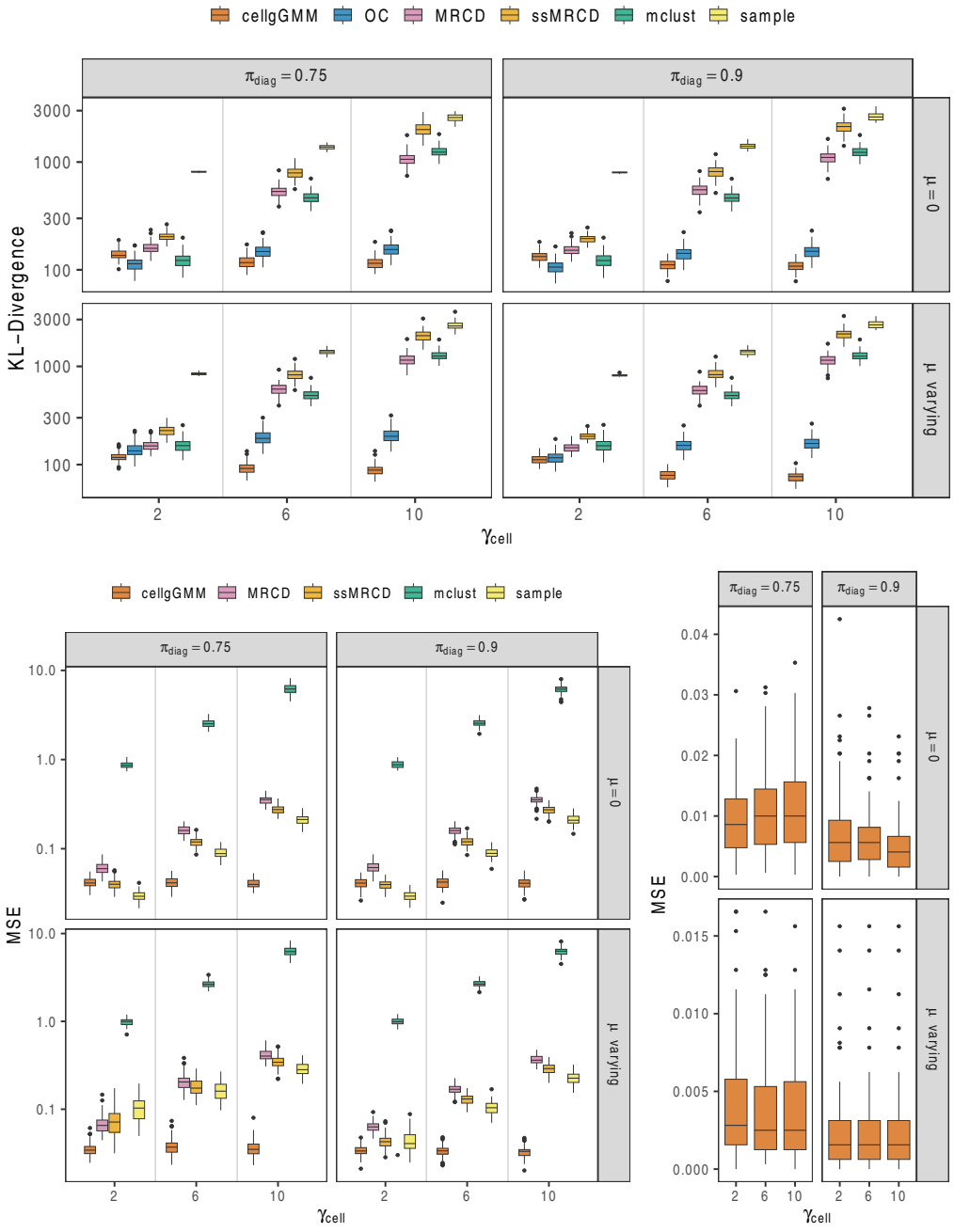


Figure D.19: Parameter estimates for the balanced high-dimensional setting ($N = 2, p = 60, n_1 = n_2 = 40$) with covariances according to Agostinelli et al. (2015). On top the KL-divergence of the covariance estimates. On the bottom left panel MSE of the mean estimation and on the bottom right the MSE of the mixture probabilities $\boldsymbol{\pi}$.

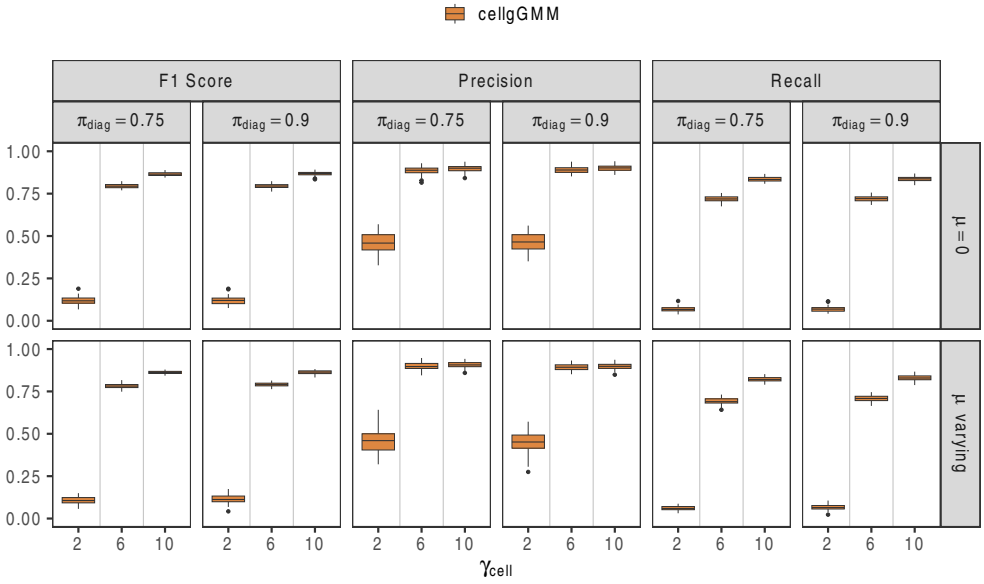


Figure D.20: Performance of cellwise outlier detection in the balanced high-dimensional setting ($N = 2, p = 60, n_1 = n_2 = 40$) with covariances according to Agostinelli et al. (2015) evaluated by on precision, recall and F1-score.

6 Conclusions

This thesis contributes with three novel methods to the robust statistical toolbox, designed for the analysis of multi-group and spatial data and for effective outlier detection. In addition, it provides an empirical evaluation of spatial outlier detection in the context of geochemistry. The following sections summarize the key findings of each chapter, outline the limitations of the proposed methods, and discuss potential directions for future research.

Chapter 2 introduces the *spatially smoothed Minimum Regularized Covariance Determinant (ssMRCD)* estimator, a rowwise robust covariance estimator tailored for grouped spatial data. This method combines robustness via minimization of determinants with spatial smoothing through weighted combinations of covariances across groups. It facilitates local outlier detection via pairwise Mahalanobis distances. Theoretical results include derivations of the breakdown point for the MRCD and ssMRCD, alongside an efficient algorithm. Simulation results and a real-world data example highlight the method's robustness and interpretability

A limitation lies in the subjectivity involved in defining groups based on spatial proximity, currently a task left to the statistician. Automating this step through a nested clustering algorithm that prioritizes spatial over feature proximity would enhance objectivity. Similarly, the optimal number of neighbors used in comparisons remains an open problem. While the method is developed for spatial data, it may also be applicable to time series or spatio-temporal data. Extensions could include hierarchical smoothing or varying smoothing strength across groups. The general principle of weighted covariances may also be adapted for other robust estimators, such as M-estimators.

Chapter 3 applies the ssMRCD and competing methods to geochemical data of varying scale, sampling density, and quality. The multi-group structure helps account for systematic measurement biases, and the method's utility for mineral exploration is demonstrated. Issues related to data preprocessing for compositional data are also addressed.

A challenge in this context is the partial validation of detected outliers: they are compared to known mineral deposits, yet not all existing deposits are likely to be documented. Ideally, field investigations would validate the flagged regions, however such efforts are extremely costly. Integrating geophysical or geological data could improve group definition and enhance outlier detection. Additionally, cellwise robust methods may help to better characterize why certain observations are outlying.

Chapter 4 presents a sparse, robust PCA method for multi-group data. It facilitates the interpretation by inducing shared sparsity patterns among all groups. A non-convex optimization problem is solved using a tailored and fine-tuned ADMM algorithm, which is extensively tested. Real life data applications show the increased interpretation

potential of loadings.

Future extensions could include the construction of compositionally coherent sparse clr-loadings which are constrained to sum up to zero. Then, each clr-loading has a clear compositional meaning. Also more elaborate visualizations could incorporate geographical and PCA information at the same time similar to dynamic PCA. This proved to be difficult, thus, research in that area could contribute to a better understanding. Moreover, a multi-group biplot framework would be a valuable but non-trivial extension. Hierarchical group structures could also be incorporated to reflect more complex group relationships.

Finally, **Chapter 5** develops a novel Gaussian mixture model that incorporates the multi-group aspect by allowing observations to be mislabeled and to originate from another group. This provides more flexibility and overall smoothness between estimated covariances. Cellwise robustness ensures reliable results and detected outliers can shed light into group-overlaps. Especially when varying the degree of flexibility of the groups, observations can be identified as being very representative of a certain group, being at the verge between groups or being totally mislabeled and outlying in their original group. An efficient algorithm is developed and results prove to be reliably superior as illustrated in simulations and real data examples. Moreover, a theoretical framework was presented to analyze the breakdown point of cellwise robust clustering models for well-clustered data and further used to prove the breakdown point for the multi-group mixture model.

While extensive simulations studies prove efficient estimation of mixture parameters as well as detection of cellwise outliers, a simulation study in the presence of rowwise outliers is not considered yet. Moreover, cellwise robustness is especially useful in the context of high-dimensional data. Although a simulation setting is considered with more variables than observations per group, it would be valuable if the algorithm is still feasible for more calculation-heavy settings with a very high absolute number of variables–a setting well suited for sparse precision matrices and the joint graphical lasso (Danaher et al., 2014). Based on the formulation via a likelihood, further theoretical results like consistency or the influence function could be derived. Moreover, it is also possible to use the proposed covariance estimates for cellwise spatial outlier detection when applied to spatial data similar to Chapter 2. New insights and interpretation possibilities could be gained when applying the cellwise paradigm to spatial or also high-dimensional time series data.

Overall, this thesis presents three robust statistical methods tailored to multi-group data, enhancing parameter estimation and outlier interpretation by leveraging contextual information. These contributions offer valuable insights into group transitions and misclassifications. Each chapter opens avenues for further research, both methodologically and theoretically, and the findings lay a solid foundation for advancing robust analysis in complex data settings.

Bibliography

Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24:441–461.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Alqallaf, F., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, pages 311–331.

Barbaglia, L., Wilms, I., and Croux, C. (2016). Commodity dynamics: A sparse multi-class approach. *Energy Economics*, 60:62–72.

Bellino, A., Alfani, A., Riso, L. D., and Baldantoni, D. (2019). Multivariate spatial analysis for the identification of criticalities and of the subtended causes in river ecosystems. *Environmental Science and Pollution Research*, 27(25):30969–30976.

Bertsimas, D., Cory-Wright, R., and Pauphilet, J. (2022). Solving large-scale sparse PCA to certifiable (near) optimality. *Journal of Machine Learning Research*, 23(13):1–35.

Bertsimas, D. and Kitane, D. L. (2023). Sparse PCA: A geometric approach. *Journal of Machine Learning Research*, 24(32):1–33.

Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30(1):113–128.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Braus, L. (2023). *Local Outlier Detection for Compositional Data*. Diploma thesis, Technische Universität Wien.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD '00*. ACM Press.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443.

Cadima, J. and Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2):203–214.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37.

Cilia, N. D., De Gregorio, G., De Stefano, C., Fontanella, F., Marcelli, A., and Parziale, A. (2022). Diagnosing alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*, 111:104822.

Cilia, N. D., De Stefano, C., Fontanella, F., and Di Freca, A. S. (2018). An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis. *Procedia Computer Science*, 141:466–471.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009a). Wine Quality. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56S3T.

Cortez, P., Cerdeira, A. L., Almeida, F., Matos, T., and Reis, J. (2009b). Modeling wine preferences by data mining from physicochemical properties. *Decis. Support Syst.*, 47:547–553.

Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons.

Croux, C., Filzmoser, P., and Fritz, H. (2013). Robust sparse principal component analysis. *Technometrics*, 55(2):202–214.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2):161–190.

Croux, C. and Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87(3):603–618.

Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: The projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226.

Cuesta-Albertos, J., Matrán, C., and Mayo-Iscar, A. (2008). Robust estimation in the normal mixture model based on robust clustering. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(4):779–802.

Cuesta-Albertos, J. A., Gordaliza, A., and Matrán, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(2):373–397.

Dasgupta, S. (1999). Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE.

d'Aspremont, A., Bach, F., and El Ghaoui, L. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9(7).

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, pages 1269–1292.

Dempster, A. P., Laird, N. M., and 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.

Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators*. Ph.D. qualifying paper, Harvard University.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, 157184.

Eirola, E., Lendasse, A., Vandewalle, V., and Biernacki, C. (2014). Mixture of Gaussians for distance estimation with missing data. *Neurocomputing*, 131:32–42.

Ernst, M. and Haesbroeck, G. (2016). Comparison of local outlier detection techniques in spatial multivariate data. *Data Mining and Knowledge Discovery*, 31(2):371–399.

Farcomeni, A. (2014a). Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, 56(1):102–111.

Farcomeni, A. (2014b). Snipping for robust k-means clustering under component-wise contamination. *Statistics and Computing*, 24:907–919.

Fayomi, A., Pantazis, Y., Tsagris, M., and Wood, A. T. (2024). Cauchy robust principal component analysis with applications to high-dimensional data sets. *Statistics and Computing*, 34(1):26.

Filzmoser, P. (2004). A multivariate outlier detection method. In *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, volume 1, pages 18–22, Minsk, Belarus. Belarusian State University.

Filzmoser, P., Fritz, H., and Kalcher, K. (2009). *pcaPP: Robust PCA by projection pursuit*.

Filzmoser, P. and Gschwandtner, M. (2012). *mvoutlier: Multivariate outlier detection based on robust methods*. R package version 2.1.1.

Filzmoser, P., Hron, K., and Templ, M. (2018). Applied compositional data analysis. *Cham: Springer*.

Filzmoser, P., Ruiz-Gazen, A., and Thomas-Agnan, C. (2013). Identification of local multivariate outliers. *Statistical Papers*, 55(1):29–47.

Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B., and Fop, M. (2024). *mclust: Gaussian mixture modelling for model-based clustering, classification, and density estimation*. R package version 6.6.1.

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., and Thirion, B. (2012). Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. *Medical Image Analysis*, 16(7):1359–1370.

Gallegos, M. T. (2002). Maximum likelihood clustering with outliers. In *Classification, Clustering, and Data Analysis: Recent Advances and Applications*, pages 247–255. Springer.

Gallegos, M. T. (2003). Clustering in the presence of outliers. In *Exploratory Data Analysis in Empirical Research: Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation eV, University of Munich, March 14–16, 2001*, pages 58–66. Springer.

Gallegos, M. T. and Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, 33(1):347–380.

Garcia-Escudero, L. A. and Gordaliza, A. (1999). Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447):956–969.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3):1324–1345.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4:89–109.

García-Escudero, L. A., Rivera-García, D., Mayo-Íscar, A., and Ortega, J. (2021). Cluster analysis with cellwise trimming and applications for the robust clustering of curves. *Information Sciences*, 573:100–124.

Geological Survey of Finland (1995). Regional till geochemistry. `https://hakku.gtk.fi/en/locations/search`.

Geological Survey of Finland (2013). Targeting till geochemistry. `https://hakku.gtk.fi/en/locations/search`.

Geological Survey of Finland (2016). Mineral deposits. `https://hakku.gtk.fi/en/locations/search`.

GeoSphere Austria (2024). `https://data.hub.geosphere.at`.

Ghadimi, E., Teixeira, A., Shames, I., and Johansson, M. (2014). Optimal parameter selection for the alternating direction method of multipliers (ADMM): Quadratic problems. *IEEE Transactions on Automatic Control*, 60(3):644–658.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177.

Gustavsson, N., Noras, P., and Tanskanen, H. (1979). Summary: Report on geochemical mapping methods. Technical report, Geological Survey of Finland.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley & Sons, New York.

Harris, P., Brunsdon, C., Charlton, M., Juggins, S., and Clarke, A. (2013). Multivariate spatial outlier detection using robust geographically weighted methods. *Mathematical Geosciences*, 46(1):1–31.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location–scale mixtures. *Ann. Statist.*, 32(1):1313–1340.

Hennig, C. (2008). Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, 99(6):1154–1176.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Hubert, M., Raymaekers, J., and Rousseeuw, P. J. (2024). Robust discriminant analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(5):e70003.

Hubert, M., Reynkens, T., Schmitt, E., and Verdonck, T. (2016). Sparse PCA for high-dimensional data with outliers. *Technometrics*, 58(4):424–434.

Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79.

Hubert, M., Rousseeuw, P. J., and Verboven, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60(1-2):101–111.

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3):618–637.

Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 52(12):5186–5201.

Jenatton, R., Obozinski, G., and Bach, F. (2010). Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 366–373. JMLR Workshop and Conference Proceedings.

Jesus, A. P., Mateus, A., Munhá, J. M., and Tassinari, C. (2014). Internal architecture and Fe–Ti–V oxide ore genesis in a Variscan synorogenic layered mafic intrusion, the Beja Layered Gabbroic Sequence (Portugal). *Lithos*, 190:111–136.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547.

Jorge, R., Fernandes, P., Rodrigues, B., Pereira, Z., and Oliveira, J. (2013). Geochemistry and provenance of the Carboniferous Baixo Alentejo Flysch Group, South Portuguese Zone. *Sedimentary Geology*, 284-285:133–148.

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2).

Leung, A., Yohai, V., and Zamar, R. (2017). Multivariate location and scatter matrix estimation under cellwise and casewise contamination. *Computational Statistics & Data Analysis*, 111:59–76.

Leyder, S., Raymaekers, J., and Verdonck, T. (2024). Generalized spherical principal component analysis. *Statistics and Computing*, 34(3):104.

Li, G. and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and monte carlo. *Journal of the American Statistical Association*, 80(391):759–766.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523.

Lutynski, P. (2019). Akanvaara Project, Finland. Technical report, Strategic Resources Inc, Canada.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801.

Maier, W. (2015). Geology and petrogenesis of magmatic Ni-Cu-PGE-Cr-V deposits: An introduction and overview. In *Mineral Deposits of Finland*, pages 73–92. Elsevier.

Marjoribanks, R. (2010). *Geological Methods in Mineral Exploration and Mining*. Springer Science & Business Media.

Markatou, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, 56(2):483–486.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons, Chichester.

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

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons.

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317.

Mayrhofer, M., Radojičić, U., and Filzmoser, P. (2025). Robust covariance estimation and explainable outlier detection for matrix-valued data. *Technometrics*, (just-accepted):1–23.

Mayrhofer, M., Radojičić, U., and Filzmoser, P. (2024). *robustmatrix: Robust matrix-variate parameter estimation*. R package version 0.1.3.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons.

Mert, M. C., Filzmoser, P., and Hron, K. (2016). Error propagation in isometric log-ratio coordinates for compositional data: theoretical and practical considerations. *Mathematical Geosciences*, 48:941–961.

Mouret, F., Hippert-Ferrer, A., Pascal, F., and Tourneret, J.-Y. (2023). A robust and flexible EM algorithm for mixtures of elliptical distributions with missing data. *IEEE Transactions on Signal Processing*, 71:1669–1682.

NEXT (2021). Grant agreement Nr: 776804. `https://doi.org/10.3030/776804`.

Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299–308.

Öllerer, V. and Croux, C. (2015). Robust high-dimensional precision matrix estimation. *Modern Nonparametric, Robust and Multivariate Methods*, pages 325–350.

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Wiley, Chichester.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348.

Price, B. S., Molstad, A. J., and Sherwood, B. (2021). Estimating multiple precision matrices with cluster fusion regularization. *Journal of Computational and Graphical Statistics*, 30(4):823–834.

Price, B. S. and Sherwood, B. (2018). A cluster elastic net for multivariate regression. *Journal of Machine Learning Research*, 18(232):1–39.

Puchhammer, P. and Filzmoser, P. (2023). *ssMRCD: Spatially smoothed MRCD estimator*. R package version 1.1.0.

Puchhammer, P. and Filzmoser, P. (2024). Spatially smoothed robust covariance estimation for local outlier detection. *Journal of Computational and Graphical Statistics*, 33(3):928–940.

Puchhammer, P., Kalubowila, C., Braus, L., Pospiech, S., Sarala, P., and Filzmoser, P. (2024a). A performance study of local outlier detection methods for mineral exploration with geochemical compositional data. *Journal of Geochemical Exploration*, 258:107392.

Puchhammer, P., Wilms, I., and Filzmoser, P. (2024b). Sparse outlier-robust PCA for multi-source data. *arXiv preprint arXiv:2407.16299*.

Puchhammer, P., Wilms, I., and Filzmoser, P. (2025). A smooth multi-group Gaussian Mixture Model for cellwise robust covariance estimation. *arXiv preprint arXiv:2504.02547*.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raymaekers, J. and Rousseeuw, P. (2021). Handling cellwise outliers by sparse regression and robust covariance. *Journal of Data Science, Statistics, and Visualisation*, 1(3):1–19.

Raymaekers, J., Rousseeuw, P., den Bossche, W. V., and Hubert, M. (2023). *cellWise: Analyzing data with cellwise outliers*. R package version 2.5.3.

Raymaekers, J. and Rousseeuw, P. J. (2023). The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association*, pages 1–12.

Raymaekers, J. and Rousseeuw, P. J. (2024a). Challenges of cellwise outliers. *Econometrics and Statistics*.

Raymaekers, J. and Rousseeuw, P. J. (2024b). Transforming variables to central normality. *Machine Learning*, 113(8):4953–4975.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., and O'Connor, P., editors (2014a). *Chemistry of Europe's Agricultural Soils - Part A: Methodology and Interpretation of the GEMAS Data Set*. Schweizerbart Science Publishers, Stuttgart, Germany.

Reimann, C., Birke, M., Demetriades, A., Filzmoser, P., and O'Connor, P., editors (2014b). *Chemistry of Europe's Agricultural Soils - Part B: General Background Information and Further Analysis of the GEMAS Data Set*. Schweizerbart Science Publishers, Stuttgart, Germany.

Reynkens, T. (2018). *rospca: Robust sparse PCA using the ROSPCA Algorithm*. R package version 1.0.4.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 8(283-297):37.

Rousseeuw, P. J. and Bossche, W. V. D. (2018). Detecting deviating data cells. *Technometrics*, 60(2):135–145.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Salminen, R. and Tarvainen, T. (1995). Geochemical mapping and databases in Finland. *Journal of Geochemical Exploration*, 55(1-3):321–327.

Schmitt, E. and Vakili, K. (2016). The fasthcs algorithm for robust pca. *Statistics and Computing*, 26(6):1229–1242.

Schubert, E., Zimek, A., and Kriegel, H.-P. (2012). Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1):190–237.

SEMACRET Project (2023). `https://semacret.eu`.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034.

Shi, N. and Kontar, R. A. (2024). Personalized PCA: Decoupling shared and unique features. *Journal of Machine Learning Research*, 25:1–82.

Signorell et mult. al., A. (2017). *DescTools: Tools for descriptive statistics*. R package version 0.99.23.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Soetaert, K. (2009). *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*. R package version 1.6.

Stahel, W. A. (1981). Breakdown of covariance estimators. Research Report 31, Fachgruppe für Statistik, ETH Zürich.

Tang, T. M. and Allen, G. I. (2021). Integrated principal components analysis. *Journal of Machine Learning Research*, 22(198):1–71.

Templ, M., Hron, K., and Filzmoser, P. (2011). *robCompositions: an R-package for Robust Statistical Analysis of Compositional Data*. John Wiley and Sons.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(sup1):234–240.

Todorov, V. (2024). *rrcov: Scalable robust estimators with high breakdown point*. R package version 1.7-6.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.

Tyler, D. E. (2010). A note on multivariate location and scatter statistics for sparse data sets. *Statistics & Probability Letters*, 80(17):1409–1413.

Van Aelst, S., Vandervieren, E., and Willems, G. (2011). Stahel-donoho estimators with cellwise weights. *Journal of Statistical Computation and Simulation*, 81(1):1–27.

Wang, W., Liang, Y., and Xing, E. (2013). Block regularized lasso for multivariate multi-response linear regression. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 608–617, Scottsdale, Arizona, USA. PMLR.

Wang, Y. and Van Aelst, S. (2020). Sparse principal component analysis based on least trimmed squares. *Technometrics*, 62(4):473–485.

Wilms, I., Barbaglia, L., and Croux, C. (2018). Multiclass vector auto-regressive models for multistore sales data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67(2):435–452.

Yao, Y., Peng, L., and Tsakiris, M. C. (2024). Unlabeled principal component analysis and matrix completion. *Journal of Machine Learning Research*, 25(77):1–38.

Yi, S., Lai, Z., He, Z., Cheung, Y.-m., and Liu, Y. (2017). Joint sparse principal component analysis. *Pattern Recognition*, 61:524–536.

Zaccaria, G., García-Escudero, L. A., Greselin, F., and Mayo-Íscar, A. (2024). Cellwise outlier detection in heterogeneous populations. *arXiv preprint arXiv:2409.07881*.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. P. (2016). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2426–2439.

Zimek, A. and Filzmoser, P. (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6):e1280.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

Zuo, Y. (2001). Some quantitative relationships between two types of finite sample breakdown point. *Statistics & Probability Letters*, 51(4):369–375.

Curriculum Vitae

Personal Data

Name	**Patricia Puchhammer**
Date of Birth	September 21, 1995
Nationality	Austrian
E-mail	patricia.puchhammer@tuwien.ac.at

Professional Experience

06/2022 – present	**Project assistant (Prae Doc)** in the CSTAT group, Institute of Statistics and Mathematical Methods in Economics, *TU Wien*, Austria
11/2021 – 05/2022	**Data Scientist** at *Ludwig Boltzmann Institute for Lung Health*, Vienna, Austria
06/2018 – 11/2018	**Student Assistant** at the group Companies, Industries and Regions in the *Institute for Advanced Studies*, Vienna, Austria
03/2017 – 06/2019	**Tutor** in the ECON and ASTAT group, Institute of Statistics and Mathematical Methods in Economics, *TU Wien*, Austria

---

Education

06/2022 – present	**Doctoral program** in Engineering Sciences - Technical Mathematics, TU Wien, Austria
07/2018 – 06/2021	**Master's program** in Statistics and Mathematical Methods in Economics, TU Wien, Austria *Graduated with honors*
10/2014 – 07/2018	**Bachelor's program** in Statistics and Mathematical Methods in Economics, TU Wien, Austria
09/2010 – 06/2014	**Matura** (A levels, higher education entrance qualification), ORG St. Ursula, Vienna, Austria *Graduated with honors*

List of Publications

Puchhammer, P., Wilms, I., and Filzmoser, P. (2025). A smooth multi-group Gaussian Mixture Model for cellwise robust covariance estimation. *arXiv preprint arXiv:2504.02547.*

Puchhammer, P., Wilms, I., and Filzmoser, P. (2024). Sparse outlier-robust PCA for multi-source data. *arXiv preprint arXiv:2407.16299.*

Puchhammer, P., Kalubowila, C., Braus, L., Pospiech, S., Sarala, P., and Filzmoser, P. (2024). A performance study of local outlier detection methods for mineral exploration with geochemical compositional data. *Journal of Geochemical Exploration*, 258, 107392. DOI: 10.1016/j.gexplo.2024.107392.

Puchhammer, P., and Filzmoser, P. (2024). Spatially smoothed robust covariance estimation for local outlier detection. *Journal of Computational and Graphical Statistics*, 33(3), 928-940. DOI: 10.1080/10618600.2023.2277875.

Ofenheimer, A., Breyer, M. K., Wouters, E. F., Schiffers, C., Hartl, S., Burghuber, O. C., Krach, F., Maninno, D., Franssen, F., Mraz, T., Puchhammer, P., and Breyer-Kohansal, R. (2024). The effect of body compartments on lung function in childhood and adolescence. *Clinical Nutrition*, 43(2), 476-481.

Danninger, K., Burghuber, O., Breyer, M. K., Puchhammer, P., Ofenheimer, A., Breyer-Kohansal, R., Kaufmann, C., Hartl,S., and Weber, T. (2023). Prevalence and determinants of vascular aging: the LEAD study. *Artery Research*, 29(1).

Schiffers, C., Faner, R., Ofenheimer, A., Sunanta, O., Puchhammer, P., Mraz, T., Breyer, M.K., Burghuber, O.C., Hartl, S., Agusti, A. and Breyer-Kohansal, R. (2023). Supranormal lung function: prevalence, associated factors and clinical manifestations across the lifespan. *Respirology*, 28(10), 942-953.

Puchhammer, P. (2021). The effects of quantitative easing in a new Keynesian small open economy model. Master's thesis, Technische Universität Wien.

Schnabl, A., Lappöhn, S., Plank, K., and Puchhammer, P. (2019). Umwegrentabilität des österreichischen EU-Ratsvorsitzes 2018. Bundeskanzleramt Österreich.

Schnabl, A., Lappöhn, S., Plank, K., and Puchhammer, P. (2019). Ökonomische Effekte der Wien Holding für Österreich und seine Bundesländer. Wien Holding GmbH.

List of Published Software

Puchhammer, P. and Filzmoser, P. (2023). *ssMRCD: Spatially smoothed MRCD estimator*. R package version 1.1.0

List of Presentations

Puchhammer, P., and Filzmoser, P. (2025). Leveraging spatial anomaly detection for mineral exploration. *EGU General Assembly 2025*, Vienna, Austria.

Puchhammer, P., Wilms, I., and Filzmoser, P. (2024). Robust sparse PCA for spatial data. *ICORS meets DSSV 2024*, Fairfax, United States of America.

Puchhammer, P., Filzmoser, P., and Wilms, I. (2024). Groupwise sparse PCA for spatial data. *Austrian Statistical Days 2024*, Vienna, Austria.

Puchhammer, P., and Filzmoser, P. (2023). Spatial outlier detection using the spatially smoothed MRCD. *22nd Annual Conference of the International Association for Mathematical Geosciences 2023 (IAMG 2023)*, Trondheim, Norway.

Puchhammer, P., and Filzmoser, P. (2023). Detecting Local Outliers Using the Spatially Smoothed MRCD Estimator. *Olomoucian Days of Applied Mathematics (ODAM 2023)*, Olomouc, Czech Republic.

Puchhammer, P., and Filzmoser, P. (2023). A spatially smoothed MRCD estimator for local outlier detection. *International Conference on Robust Statistics (ICORS 2023)*, Toulouse, France.

April 30, 2025

Patricia Puchhammer