TECHNISCHE UNIVERSITÄT WIEN

DISSERTATION

Robustness and Explainable Outlier Detection for Multivariate, Matrix-variate, and Functional Settings

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

und

Horst Lewitschnig
Infineon Technologies Austria AG

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Marcus Mayrhofer
Matrikelnummer: 01607509

Diese Dissertation haben begutachtet:

1. **Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser**
   Institut für Stochastik und Wirtschaftsmathematik, TU Wien
2. **Assoz. Univ.-Prof.in Mag.a Dr.in Helga Wagner**
   Institut für Angewandte Statistik, JKU Linz
3. **Prof. Stefan Van Aelst, PhD**
   Statistics and Data Science Section, KU Leuven

Wien, am 24. Oktober 2024

Marcus Mayrhofer

Peter Filzmoser

Abstract

This work addresses the challenges of robust covariance estimation and interpretable outlier detection for multivariate, matrix-variate, and functional data. The goal is to develop methods that enhance both the robustness and interpretability in these settings.

For outlier interpretability, we propose a novel approach that combines robust Mahalanobis distances with Shapley values to decompose multivariate outlyingness into variable-specific contributions. We present this decomposition in the multivariate setting and demonstrate how our method reduces the exponential computational complexity in the number of variables to linear complexity, while preserving the key properties of the Shapley value. This approach is also extended to the matrix-variate and functional setting, respectively.

For robust location and covariance estimation in the matrix-variate setting, we define the Matrix Minimum Covariance Determinant (MMCD) estimators and prove that they are consistent in the class of matrix-variate elliptical distributions. We show that these estimators are matrix affine equivariant and achieve a higher breakdown point than the maximum attainable for any multivariate affine equivariant covariance estimator applied to vectorized data. We demonstrate that the incorporation of an additional reweighting step improves the efficiency, and finally present and implement a fast algorithm with convergence guarantees.

The MMCD approach naturally extends to the setting of multivariate Functional Data Analysis (FDA), where data are represented using basis functions and coefficient matrices. We establish a connection between stochastic processes with a separable covariance structure and the corresponding matrix-variate distribution of their basis representations. In combination with a multivariate functional Mahalanobis (semi-)distance, the MMCD approach can be used to robustly estimate the mean and covariance functions for multivariate functional data.

The combined use of robust Mahalanobis distances, MMCD estimators, and Shapley value-based outlyingness decomposition offers a comprehensive framework for robust and interpretable data analysis across multivariate, matrix-variate, and functional data structures, with substantial theoretical and practical benefits, verified through simulations and real-world examples.

Kurzfassung

Die vorliegende Dissertation widmet sich den Herausforderungen der robusten Kovarianzschätzung sowie der interpretierbaren Ausreißererkennung für multivariate, matrixwertige und funktionale Daten. Das Ziel der Arbeit besteht darin, Methoden zu entwickeln, die dazu dienen, sowohl die Robustheit als auch die Interpretierbarkeit für diese Datenformate zu verbessern.

Zur Verbesserung der Interpretierbarkeit von Ausreißern präsentieren wir eine neuartige Herangehensweise, bestehend aus einer Kombination von robusten Mahalanobis-Distanzen und Shapley-Werten, mit der die quadrierte Mahalanobis-Distanz in variablen-spezifische Beiträge zerlegt wird. Zuerst wird diese Zerlegung im multivariaten Kontext eingeführt, wobei wir zeigen, dass der exponentielle Rechenaufwand in der Anzahl der Variablen durch unsere Methode auf eine lineare Abhängigkeit von der Variablenanzahl reduziert wird. Des Weiteren demonstrieren wir, dass die postulierte Methode auch auf matrixwertige und funktionale Daten erweitert und angewendet werden kann.

Um eine robuste Schätzung für die Lage und Kovarianz von matrixwertigen Daten zu erhalten, definieren wir die Matrix Minimum Covariance Determinant (MMCD) Schätzer und beweisen ihre Konsistenz in der Klasse der matrixwertigen elliptischen Verteilungen, sowie ihre matrix-affine Äquivarianz. Weiters zeigen wir, dass der Bruchpunkt des MMCD-Schätzers über dem maximal erreichbaren Bruchpunkt von multivariaten affin äquivarianten Schätzern, welche auf vektorisierten Daten berechnet werden, liegt. Durch einen zusätzliche Neugewichtung der Beobachtungen wird die Effizienz der MMCD-Schätzer verbessert. Zur Berechnung der MMCD-Schätzer wird ein schneller Algorithmus mit Konvergenzgarantien implementiert.

Zur Erweiterung des MMCD-Ansatzes auf multivariate funktionale Daten werden die Beobachtung zunächst mithilfe von Basisfunktionen geglättet. Zu diesem Zweck leiten wir den Zusammenhang zwischen stochastischen Prozessen mit einer separierbaren Kovarianzstruktur und der entsprechenden matrixwertigen Verteilung ihrer Koeffizientenmatrizen in der Basisdarstellung her. Wir verwenden die MMCD-Schätzer in Kombination mit einer multivariaten funktionalen Mahalanobis-(Semi-)-Distanz, um Mittel- und Kovarianzfunktion für multivariate funktionale Daten robust zu schätzen.

Die Verknüpfung von robusten Mahalanobis-Distanzen, MMCD-Schätzern und der auf Shapley-Werten basierenden Ausreißererklärung bildet ein umfassendes Framework für robuste und interpretierbare Datenanalyse von multivariaten, matrixwertigen und funktionalen Datenstrukturen. Das vorgestellte Konzept bietet sowohl theoretische als auch praktische Vorteile, welche mithilfe von Simulationen und Anwendungen auf realen Daten illustriert und belegt werden.

Acknowledgement

– Standing on the Shoulders of Giants –

I have received tremendous support from my family, friends, and colleagues while pursuing my PhD, and I want to thank all of you. You are my giants!

First, I want to extend my heartfelt gratitude to my supervisor, Peter. I am truly fortunate to have profited from all the time and effort you dedicate to supervising your PhD students. Thank you for your support, patience, invaluable advice, and for always being there to listen, no matter the topic. Secondly, I want to thank my co-supervisor, Horst. I sincerely appreciate your support during my PhD and especially the numerous occasions you took the time to visit TU Wien to join in discussions and share your valuable insights. Furthermore, I want to express my gratitude to all my colleagues in the computational statistics group. Una, thank you for the countless hours we spent refining the details of our papers together. Your dedication and collaboration were invaluable throughout my PhD. Thank you to my office mate, Pia, I really enjoyed transforming our office into a mini garden. To my fellow PhD colleagues - Barbara, Jeremy, Lukas, Patricia, and Roman - thank you for all the joint lunches that became a cherished tradition in our group and for the shared breaks filled with conversations. Thanks to you, coming to work feels like meeting with a circle of friends.

I am deeply grateful to my family and friends for their support, genuine interest in my endeavors, and for being a source of much-needed comfort and diversion. I am sincerely thankful to my brother Matthias for always being there for me. I have learned so much from looking up to you. I want to express my heartfelt gratefulness to my parents. Your unwavering support has enabled me to pursue my dreams, and I cannot thank you enough for all that you have done for me. Alena, your love and support are invaluable to me, and I cannot find the words to describe how fortunate I feel to share my life with you.

This work is part of the AI4CSM project and has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007326. The JU receives support from the European Union's Horizon 2020 research and innovation programme and national authorities. This work is also funded by the Austrian IKT der Zukunft programme via the Austrian Research Promotion Agency (FFG) and the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK) under project No 884070.

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 24. Oktober 2024

Marcus Mayrhofer

Contents

1 Introduction

Advancements in the development of data collection tools have led to a vast increase in both the volume and complexity of available datasets. While multivariate samples are usually represented as vectors, modern applications often involve more sophisticated data structures, such as matrix-variate or functional data. This inherent data structure enables more detailed data analysis but also presents new challenges, especially if the observations include outliers. Anomalies can distort statistical analyses if left unaddressed, yet they may also provide valuable insights by revealing unusual but important patterns in the data.

Robust statistics provide methods for handling data that contain outliers as well as tools to reliably detect them. However, the vast majority of the approaches are designed for classical vector-valued data. Moreover, most robust procedures focus on the identification of outliers but often fall short of explaining the reason why some observations deviate from the majority of the data. Since the goal of many applications lies not only in detecting outliers but also in investigating which variables contribute most to the outlyingness, providing such explanations is vital for gaining deeper insights into the data and the decisions of robust procedures in a way that is understandable by humans.

This thesis bridges these gaps by presenting theory and methods for robust location and covariance estimation as well as explainable outlier detection for multivariate, matrix-variate, and multivariate functional data. The remainder of the introduction is organized as follows: We provide an overview of robust statistics in Section 1.1, before we briefly describe explainability in AI and outlier detection in Section 1.2. We then introduce matrix-variate data in Section 1.3 and functional data in Section 1.4, respectively. After outlining the introduction, the remainder of the dissertation is structured as follows: In Chapter 2, we introduce outlier explanations based on a combination of Mahalanobis distances and Shapley values. In Chapter 3, the Matrix Minimum Covariance Determinant (MMCD) estimators for robust location and covariance estimation for matrix-variate data are introduced, and we extend explainable outlier detection to the matrix setting. Chapter 4 builds upon the methods developed in the previous chapters and introduces an approach for robust and explainable outlier detection for multivariate functional data. Finally, Chapter 5 summarizes the key findings and outlines potential directions for future research.

1.1 Robust Statistics

Statistical data analysis provides a mathematical framework for analyzing, interpreting, explaining, modeling, and presenting data. Statistical models rely on assumptions about the data, such as randomness, independence, or the underlying distribution, to enable us to effectively summarize the data, quantify uncertainty, and draw valid conclusions. To ensure the reliability of summary statistics and uncertainty measures, it is vital that the model assumptions are met, making it important to thoroughly examine these assumptions.

There are various diagnostic tools and tests available for this purpose. However, many of those tools themselves rely on assumptions and are sensitive to outliers which can create a compounding problem.

The goal of robust statistics is to provide tools in which small deviations from the model assumptions only induce small errors in the final conclusions (Huber and Ronchetti, 2011). In the present context, our focus lies on robust location and covariance estimation. Robust estimators should fit the majority of the data while simultaneously limiting the effects of possible outliers. This type of robustness is called distributional robustness and the most common model can be described as follows: Let us assume that the samples are generated from a clean distribution F with probability $(1-\varepsilon) > 0.5$ and from an unspecified distribution H with probability ε, i.e.,

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon H.$$

In terms of random variables, we can write

$$\boldsymbol{X} = (1-B)\boldsymbol{Y} + B\boldsymbol{Z}, \tag{1.1}$$

with $B \sim \text{Bernoulli}(\varepsilon)$, $\boldsymbol{Y} \sim F$, $\boldsymbol{Z} \sim H$, B independent of $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Hence, $\boldsymbol{X} \sim F_\varepsilon$ and we want to estimate F, but we can only observe F_ε. Other than the independence of B, we do not make any assumptions regarding H. This is known as the Tukey-Huber contamination model after Tukey (1960) and Huber (1964).

In the subsequent paragraphs, we outline the most common multivariate location and covariance estimators, focusing on their properties as detailed in Hampel et al. (1986); Huber and Ronchetti (2011); Maronna et al. (2019). These books provide a comprehensive overview of robust statistics, detailing the various methods developed within the casewise contamination model, which extends beyond location and covariance estimation to include topics such as linear and generalized linear models, Principal Component Analysis (PCA), and time series analysis. We also discuss the more recently proposed cellwise contamination of Alqallaf et al. (2009), which assumes that only individual cells of a data matrix are outlying, rather than entire rows.

1.1.1 Classical Estimators and Their Limitations

We consider a p-variate random variable $\boldsymbol{x} = (x_1, \ldots, x_p)'$ with distribution F taking values in $\mathbb{R}^p$. We analyze the location $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} = \text{Var}[\boldsymbol{x}] = \mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})'] \in \text{PDS}(p)$, where $\text{PDS}(p)$ denotes the class of all $p \times p$ positive definite symmetric matrices. The most common assumption in statistics is that $\boldsymbol{x}$ follows a multivariate normal distribution with density

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \det(\Sigma)}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{1.2}$$

denoted as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Given an i.i.d. sample of multivariate normal observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, for $i = 1, \ldots, n$, the samples can be collected as rows in the data matrix

$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times p}$. The maximum likelihood estimators (MLEs) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the sample mean and covariance given by

$$\hat{\boldsymbol{m}}(\boldsymbol{X}) = \hat{\boldsymbol{m}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \hat{\boldsymbol{S}}(\boldsymbol{X}) = \hat{\boldsymbol{S}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{m}})(\boldsymbol{x}_i - \hat{\boldsymbol{m}})', \tag{1.3}$$

respectively. Equation (1.3) reveals that even if only a few out of the n samples in $\boldsymbol{X}$ deviate from the normality assumptions, they can completely alter and distort the sample estimates.

A more general assumption for the distribution of $\boldsymbol{x}$ is that it follows an elliptical distribution, allowing for moderate departures from normality in some samples. Similar to the normal distribution, it is characterized by its mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In addition to the parametric part, the elliptical distribution is described by its generator function $g : [0, \infty) \to \mathbb{R}$ and its density is given by

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, g) = \frac{1}{\sqrt{\det(\boldsymbol{\Sigma})}} g((\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})). \tag{1.4}$$

A random vector $\boldsymbol{x}$ following an elliptical distribution with mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$, and generator function g is denoted as $\boldsymbol{x} \sim \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$. The normal distribution is the most well-known representative of the class of elliptical distributions class, but it also encompasses heavy-tailed distributions such as the multivariate t-distribution. For any choice of positive g such that the integral in Equation (1.4) equates to one, the level sets

$$L_d(f) = \{\boldsymbol{x} \in R^p : f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, g) = d\},$$

for $d > 0$ are ellipsoids. Figure 1.1 shows the level sets of the centered bivariate normal and centered bivariate t-distribution with 2 degrees of freedom for fixed values of d, respectively. In the elliptical setting, the diagonal entries of $\boldsymbol{\Sigma}$ are ones and off-diagonal entries are 0.7. For the spherical case $\boldsymbol{\Sigma}$ equates to the identity matrix. The comparison between the two distributions highlights that the bivariate t-distribution has a more concentrated center and heavier tails than the bivariate normal distribution.

1.1.2 Outlier detection

Outliers may represent erroneous observations that adversely affect statistical analyses. On the other hand, they often capture rare or extreme events, making them some of the most important pieces of information in a dataset.

A widely used tool for detecting multivariate outliers in statistics is the Mahalanobis distance (Mahalanobis, 1936). For a multivariate observation $\boldsymbol{x} \in \mathbb{R}^p$ from a population with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \mathrm{PDS}(p)$, the squared Mahalanobis distance of an observation $\boldsymbol{x}$ to the mean $\boldsymbol{\mu}$ with respect to the covariance $\boldsymbol{\Sigma}$ is defined as

$$\mathrm{MD}^2(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{MD}^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}). \tag{1.5}$$

If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathrm{MD}^2(\boldsymbol{x}) \sim \chi^2_p$, where χ^2_p denotes the chi-square distribution with p degrees of freedom (Seber, 1984). Usually, the 0.975 or 0.99 quantile of the χ^2_p distribution is used as a detection threshold to flag outliers based on Mahalanobis distance.

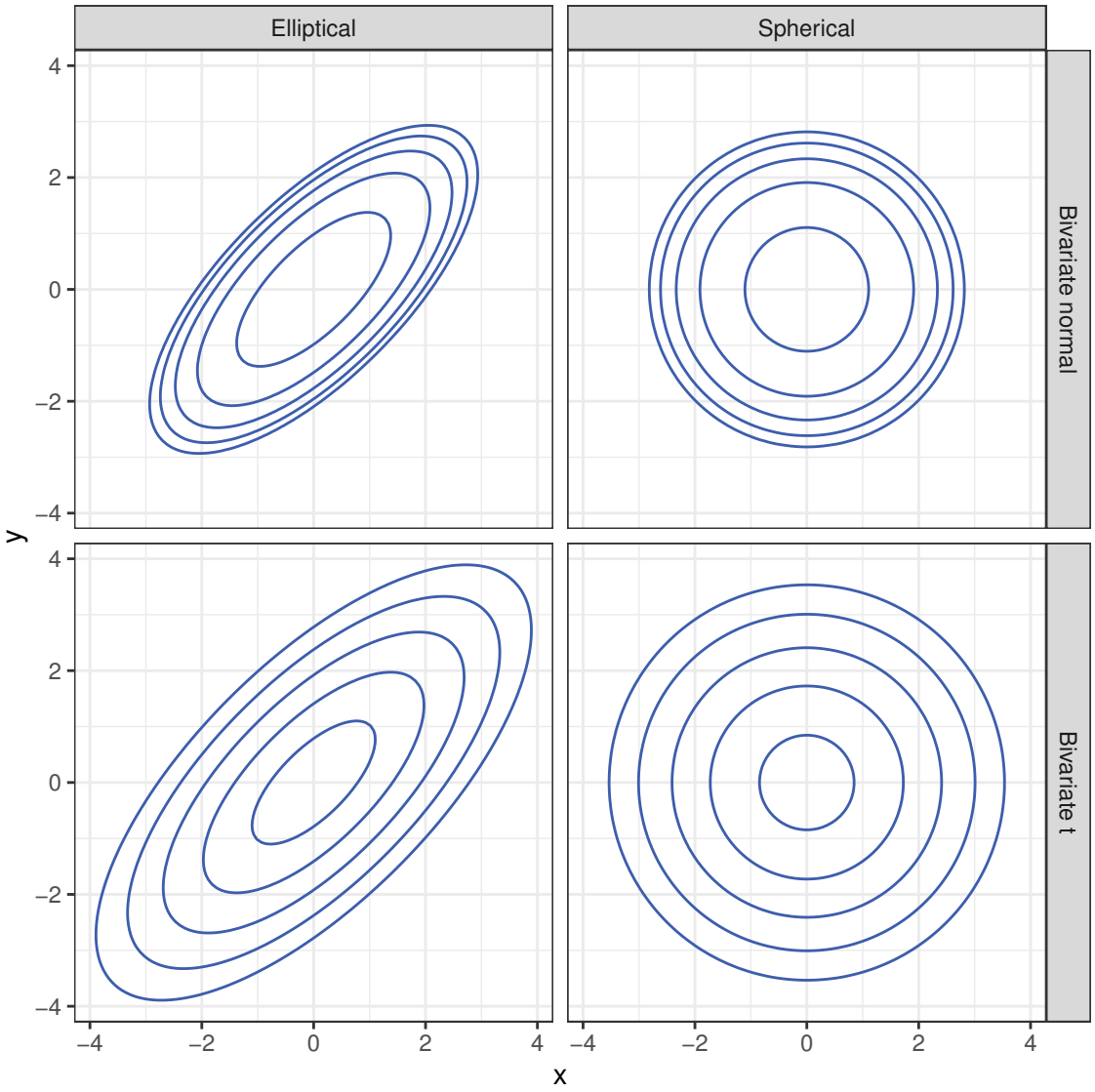
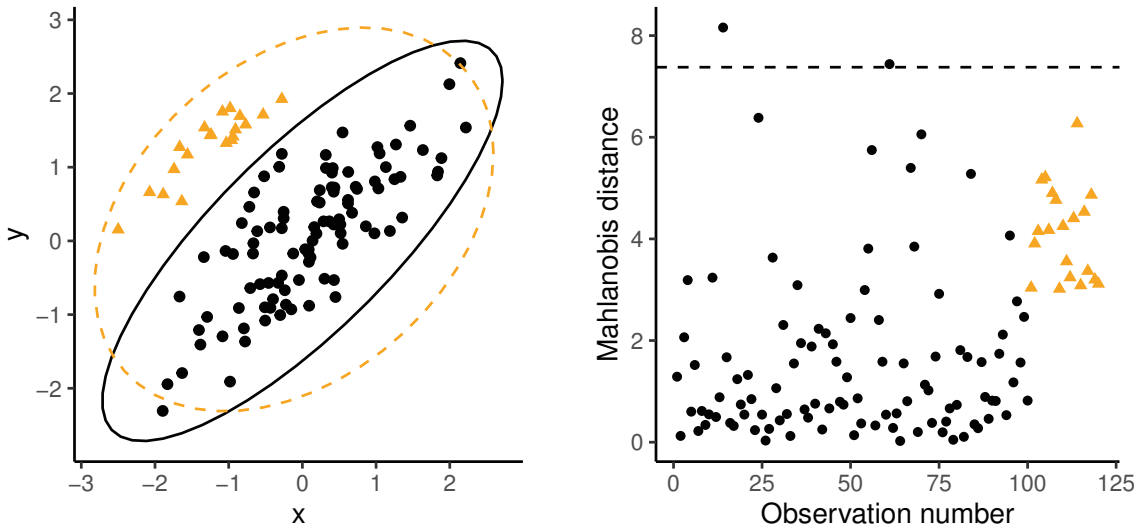


Figure 1.1: Spherical and elliptical level sets $L_d(f)$ of the bivariate normal and bivariate t-distribution with 2 degrees of freedom. Identical values for d are chosen in each plot.

In practice, the true values of the mean and covariance matrix are not known and must instead be estimated from the data. By using sample estimators (1.3), which are themselves highly sensitive to anomalies, many outliers might remain undetected; this phenomenon is known as the *masking effect*, as illustrated in Figure 1.2. The regular data (dots) are sampled from a bivariate normal distribution, and the outliers (triangles) are drawn from a shifted bivariate normal with a smaller spread. The plot also shows the tolerance ellipses based on the true parameters (solid line), i.e., the mean and covariance matrix used to simulate the data, and based on the sample estimates (dashed line). The tolerance ellipses

represent the level sets where $\text{MD}^2(\boldsymbol{x}) = \chi^2_{0.975,2}$, where $\chi^2_{0.975,2}$ denotes the 0.975 quantile of the chi-square distribution with 2 degrees of freedom. Hence, all samples that are within the tolerance ellipse are regular observations while points outside are flagged as outliers. The sample estimates are influenced and biased by the outliers, which distort the location as well as covariance estimation. The right panel of the figure shows the plot of the Mahalanobis distance based on the sample estimates, and the dashed line indicates the chi-square cutoff. This clearly shows that the outliers are masked.



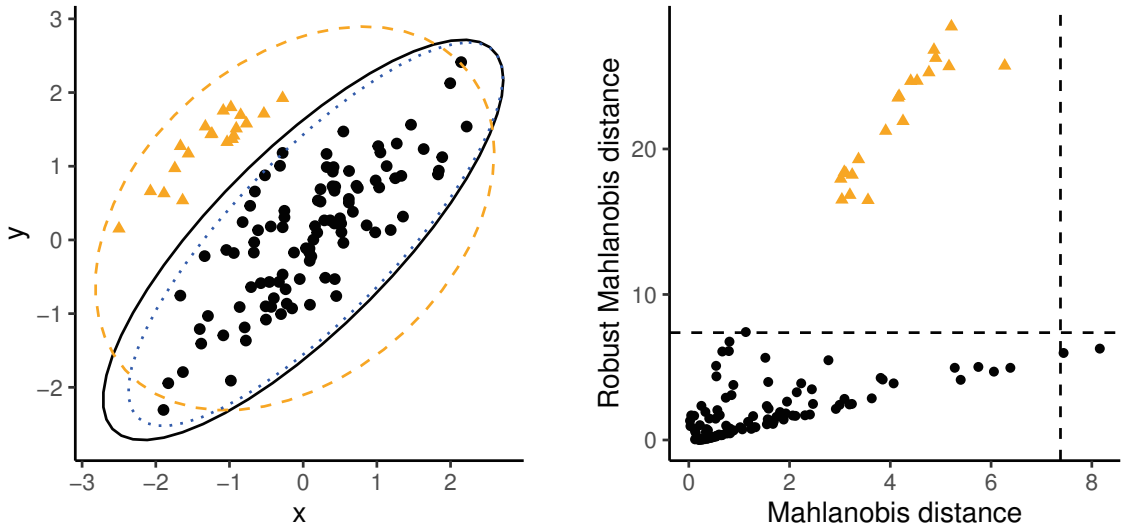
(a) Comparison of 97.5% tolerance ellipses based on the true mean and covariance matrix (solid black line) and sample counterparts (1.3) (dashed orange line).

(b) Mahalanobis distances based on sample mean and covariance matrix.

Figure 1.2: Illustration of the masking effect: Regular data are drawn from a bivariate normal distribution (black dots), and outliers are generated from a shifted bivariate normal distribution (orange triangles).

Robust procedures are designed to capture the overall structure of the data while minimizing the influence of outliers, and they offer diagnostics to help identify and interpret these outliers effectively. By replacing the sample estimates with a robust alternative, we can robustify the Mahalanobis distances and obtain a reliable tool for outlier detection for elliptically distributed data. A popular example of robust location and covariance estimation is the Minimum Covariance Determinant (MCD) approach proposed by Rousseeuw (1985), see Section 1.1.4 for details. Additionally, the distance-distance plot introduced by Rousseeuw and Van Driessen (1999) serves as a diagnostic tool that compares Mahalanobis distances based on sample estimates versus robust counterparts based on the MCD approach. In Figure 1.3, we extend the example illustrating the masking effect from Figure 1.2 by also including the tolerance ellipse for the MCD and the distance-distance plot. The tolerance ellipse of the MCD estimator (dotted line) is slightly narrower than the true ellipse (solid line) but captures its shape and location very accurately. The vertical and horizontal dashed

lines in the distance-distance plot represent the chi-square cutoffs. These lines divide the distance-distance plot into four quadrants: The lower left quadrant shows regular samples according to both classical and robust distances, while the upper right quadrant contains outliers according to both. The lower right quadrant shows samples considered outliers by classical but not by robust distance metrics, while the upper left quadrant highlights samples outlying according to the robust but not the classical distance. As can be seen, the outliers are masked by the classical approach and remain undetected, whereas they are clearly identified using robust Mahalanobis distances based on the MCD estimators.



(a) Comparison of 97.5% tolerance ellipses based on the true mean and covariance matrix (solid black line), sample counterparts (dashed orange line), and robust MCD estimators (dotted blue line).

(b) Mahalanobis distances based on sample mean and covariance matrix, as well as their robust counterparts based on the MCD approach.

Figure 1.3: Comparison of classical and robust Mahalanobis distances.

1.1.3 Properties of Robust Location and Covariance Estimators

Affine equivariance. The class of elliptical distributions is also used to motivate one desirable property of location and covariance estimators, namely affine equivariance. This means that the estimators used for location and covariance should transform in the same way as the parameters of elliptical distributions under affine transformations. Specifically, let $\boldsymbol{a} \in \mathbb{R}^p$ be a fixed vector, $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ an invertible matrix and $\boldsymbol{x} \sim \mathcal{E}(\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}, g)$. Then

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{a} \sim \mathcal{E}(\boldsymbol{\mu}_{\boldsymbol{z}}, \boldsymbol{\Sigma}_{\boldsymbol{z}}, g),$$

where $\boldsymbol{\mu}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{a}$ and $\boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}'$.

Let $\hat{\boldsymbol{\mu}}(\boldsymbol{X})$ and $\hat{\boldsymbol{\Sigma}}(\boldsymbol{X})$ denote location and covariance estimators corresponding to a sample $\boldsymbol{X}$, respectively. If every observation in $\boldsymbol{X}$ is transformed to obtain the dataset $\boldsymbol{Z}$ with

entries $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{a}$, $i = 1, \ldots, n$, then the estimators should transform in the same manner as the parameters of the elliptical distribution, i.e.,

$$\hat{\boldsymbol{\mu}}(\boldsymbol{Z}) = \boldsymbol{A}\hat{\boldsymbol{\mu}}(\boldsymbol{X}) + \boldsymbol{a} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}(\boldsymbol{Z}) = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}(\boldsymbol{X})\boldsymbol{A}'.$$

Influence function. A key principle in robustness is that estimators should remain stable under small deviations from the model assumptions. The influence function quantifies this kind of stability and is defined on the distribution level, rather than on the sample level. To formalize this, let $\hat{\theta}_n(\boldsymbol{X}) = \hat{\theta}_n$ denote an estimator depending on the sample $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times p}$ of size n, consisting of i.i.d. observations with distribution F. The asymptotic value of the estimator at F, denoted $\hat{\theta}_\infty(F) = \hat{\theta}_\infty$, satisfies

$$\hat{\theta}_n \xrightarrow{p} \hat{\theta}_\infty(F),$$

meaning that $\hat{\theta}_n$ converges in probability to $\hat{\theta}_\infty(F)$ as $n \to \infty$. Then, the influence function of the estimator $\hat{\theta}$ is given by

$$\mathrm{IF}_{\hat{\theta}}(\boldsymbol{x}_0, F) = \lim_{\varepsilon \searrow 0} \frac{\hat{\theta}_\infty((1-\varepsilon)F + \varepsilon\delta_{\boldsymbol{x}_0}) - \hat{\theta}_\infty(F)}{\varepsilon}, \tag{1.6}$$

where $\delta_{\boldsymbol{x}_0}$ is the point-mass at $\boldsymbol{x}_0$, and $\lim_{\varepsilon \searrow 0}$ denotes the limit from the right approaching 0. The influence function (1.6) quantifies the sensitivity of the estimator $\hat{\theta}$ to small changes in the data distribution at $\boldsymbol{x}_0$.

Breakdown point. The finite-sample breakdown point of an estimator refers to the smallest proportion of observations that can be replaced by outliers before they may cause the estimator to lose all information about the true parameter. For location and covariance estimators this can be formalized as follows: Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times p}$ denote the collection of clean samples and $\boldsymbol{Z}_m = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)'$ a set of n samples, such that $\boldsymbol{z}_i = \boldsymbol{x}_i$ for $n - m$ samples while the remaining m samples are replaced by arbitrary vectors in $\mathbb{R}^p$. This can be seen as a finite-dimensional version of the casewise contamination model (1.1). Let $(\hat{\boldsymbol{\mu}}(\boldsymbol{Z}_m), \hat{\boldsymbol{\Sigma}}(\boldsymbol{Z}_m))$ and $(\hat{\boldsymbol{\mu}}(\boldsymbol{X}), \hat{\boldsymbol{\Sigma}}(\boldsymbol{X}))$ be location and scatter estimators based on $\boldsymbol{Z}_m$ and $\boldsymbol{X}$, respectively. The finite sample breakdown point of the location estimator $\hat{\boldsymbol{\mu}}$ is defined as

$$\varepsilon^*(\hat{\boldsymbol{\mu}}, \boldsymbol{X}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\boldsymbol{Z}_m} \|\hat{\boldsymbol{\mu}}(\boldsymbol{Z}_m) - \hat{\boldsymbol{\mu}}(\boldsymbol{X})\| = \infty \right\},$$

where the supremum is taken over all possible corrupted collections $\boldsymbol{Z}_m$. The finite sample breakdown point of $\hat{\boldsymbol{\Sigma}}$ is defined as

$$\varepsilon^*(\hat{\boldsymbol{\Sigma}}, \boldsymbol{X}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\boldsymbol{Z}_m} D_\lambda(\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z}_m), \hat{\boldsymbol{\Sigma}}(\boldsymbol{X})) = \infty \right\},$$

with

$$D_\lambda(\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z}_m), \hat{\boldsymbol{\Sigma}}(\boldsymbol{X})) = \max \left\{ \left| \lambda_1(\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z}_m)) - \lambda_1(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X})) \right|, \left| \lambda_p^{-1}(\hat{\boldsymbol{\Sigma}}(\boldsymbol{Z}_m)) - \lambda_p^{-1}(\hat{\boldsymbol{\Sigma}}(\boldsymbol{X})) \right| \right\}.$$

Here, $\lambda_1(\boldsymbol{A}) \ge \cdots \ge \lambda_p(\boldsymbol{A})$ are the ordered eigenvalues of the matrix $\boldsymbol{A}$ and the supremum is taken over all possible corrupted collections $\boldsymbol{Z}_m$.

Efficiency. In statistics, the efficiency of an estimator refers to its ability to estimate a parameter as precisely as possible based on the available data. More formally, the efficiency of an estimator is measured by the ratio between the variance of an optimal estimator and the variance of the estimator in question. Efficiency is often studied in both the finite sample setting and the asymptotic setting. Here, we focus on the finite-sample version; for details regarding the asymptotic setting, see, e.g., Maronna et al. (2019). For location and covariance estimators $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, efficiency is usually measured under the normal model $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ based on the Kullback-Leibler (KL) divergence, denoted as D. Under the normal model, the KL divergence for the mean with known $\boldsymbol{\Sigma}$ is

$$D(\hat{\boldsymbol{\mu}}) = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}),$$

and for the covariance matrix with known $\boldsymbol{\mu}$ it is

$$D(\hat{\boldsymbol{\Sigma}}) = \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\Sigma}}) - \log(\det(\boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\Sigma}})) - p.$$

The normal finite-sample efficiency of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ is then defined as

$$\frac{\mathbb{E}[D(\hat{\boldsymbol{m}})]}{\mathbb{E}[D(\hat{\boldsymbol{\mu}})]} \quad \text{and} \quad \frac{\mathbb{E}[D(\hat{\boldsymbol{S}})]}{\mathbb{E}[D(\hat{\boldsymbol{\Sigma}})]},$$

respectively, where $\hat{\boldsymbol{m}}$ is the sample mean and $\hat{\boldsymbol{S}}$ the sample covariance matrix as defined in Equation (1.3). The finite-sample efficiency of an estimator is usually evaluated based on Monte Carlo simulations, and the expectation $\mathbb{E}[D]$ is replaced by the average simulation score.

Consistency. Another important asymptotic characteristic of an estimator is consistency, which means that an estimator converges to its true parameter value as the sample size increases indefinitely. In robust statistics, consistency of affine equivariant location and covariance estimators $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ is often studied in the context of elliptical distributions. Let $\boldsymbol{X}$ denote a random sample from an elliptical distribution $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ as in Equation (1.4), then $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are consistent estimators if

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \xrightarrow{a.s.} 0, \left\| c\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\| \xrightarrow{a.s.} 0,$$

where c is a constant.

1.1.4 Robust Location and Covariance Estimators

M-estimators of location and covariance can be seen as a generalization of the normal MLEs (1.3). Instead of maximizing the likelihood function of the multivariate normal distribution, they introduce weights based on the squared Mahalanobis distance into the estimating equations. They are defined as solutions of

$$\sum_{i=1}^{n} w_1(d_i)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}) = \boldsymbol{0} \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} w_2(d_i)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})' = \hat{\boldsymbol{\Sigma}},$$

respectively. Here, $w_1(d_i)$ and $w_2(d_i)$ are weight functions based on the squared Mahalanobis distances

$$d_i = \mathrm{MD}^2(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}),$$

for $i = 1, \ldots, n$. One way to obtain those weights is to derive the MLEs based on the elliptical density (1.4). To ensure robustness, it is crucial to limit the weights assigned to observations with large distances, thereby minimizing their influence on the estimators.

The idea is to make the scale of the squared Mahalanobis distances $\mathrm{MD}^2(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ small. This can be achieved by using a robust estimator of scale $\hat{\sigma}$ and minimizing

$$\hat{\sigma}(\mathrm{MD}^2(\boldsymbol{x}_1, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}), \ldots, \mathrm{MD}^2(\boldsymbol{x}_n, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})),$$

where $\hat{\boldsymbol{\mu}} \in \mathbb{R}^p$, $\hat{\boldsymbol{\Sigma}} \in \mathrm{PDS}(p)$, and $\det(\hat{\boldsymbol{\Sigma}}) = 1$, to avoid degenerate solutions. S-estimators are obtained by using an M-estimator of scale for $\hat{\sigma}$.

A drawback of highly robust S-estimators is that they generally have low efficiency. MM-estimators are designed to combine the strengths of S- and M-estimators. They yield estimators that retain the high robustness of S-estimators while achieving the efficiency that M-estimators reach under optimal conditions. For more details regarding M-, S-, and MM-estimators, we refer to the work of Maronna et al. (2019).

MCD estimator. One of the most used location and covariance estimators is the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1985). The objective of the MCD estimator is to find the subset of observations, whose sample covariance matrix has the lowest determinant. This can be formalized as follows: Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times p}$ denote the collection of samples, $H \subseteq \{1, \ldots, n\}$ a h-subset of size $|H| = h$, with $\lfloor (n+p+1)/2 \rfloor \leq h \leq n$, and

$$\hat{\boldsymbol{m}}(H) = \frac{1}{h} \sum_{i \in H} \boldsymbol{x}_i \quad \text{and} \quad \hat{\boldsymbol{S}}(H) = \frac{1}{h-1} \sum_{i \in H} (\boldsymbol{x}_i - \hat{\boldsymbol{m}}(H))(\boldsymbol{x}_i - \hat{\boldsymbol{m}}(H))' \tag{1.7}$$

the sample mean and sample covariance matrix based on the observations in H, respectively. Then the MCD estimators are $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\hat{\boldsymbol{m}}(H^*), \hat{\boldsymbol{S}}(H^*))$, where

$$H^* = \underset{H \subseteq \{1,\ldots,n\}, |H|=h}{\arg\min} \log(\det(\hat{\boldsymbol{S}}(H))). \tag{1.8}$$

In total, there are $\binom{h}{n}$ possible h-subsets. Thus, an efficient strategy should be used to tackle the optimization problem. For this purpose, Rousseeuw and Van Driessen (1999) proposed the Fast-MCD algorithm, with the concentration step (C-step) as a key component. The schematics of the C-step are visualized in Figure 1.4 and can be described as follows: Start with any h-subset $H_{\text{old}} \subset \{1, \ldots, n\}$, with $|H_{\text{old}}| = h$, and compute $(\hat{\boldsymbol{m}}(H_{\text{old}}), \hat{\boldsymbol{S}}(H_{\text{old}}))$ as defined in Equation (1.7). If $\det(\hat{\boldsymbol{S}}(H_{\text{old}})) \neq 0$, compute the squared Mahalanobis distances

$$\mathrm{MD}^2(\boldsymbol{x}_i) = d_i^2(H_{\text{old}}) = (\boldsymbol{x}_i - \hat{\boldsymbol{m}}(H_{\text{old}}))' \hat{\boldsymbol{S}}^{-1}(H_{\text{old}})(\boldsymbol{x} - \hat{\boldsymbol{m}}(H_{\text{old}}))$$

for all $i = 1, \ldots, n$, and sort them in ascending order. This sorting induces a permutation π of $\{1, \ldots, n\}$ such that $d^2_{\pi(1)}(H_{\text{old}}) \leq \ldots \leq d^2_{\pi(n)}(H_{\text{old}})$. Now, define a new h-subset $H_{\text{new}} = \{\pi(1), \ldots, \pi(h)\}$, and compute $\hat{\boldsymbol{m}}(H_{\text{new}})$ and $\hat{\boldsymbol{S}}(H_{new})$. It holds, that $\det(\hat{\boldsymbol{S}}(H_{\text{new}})) \leq$

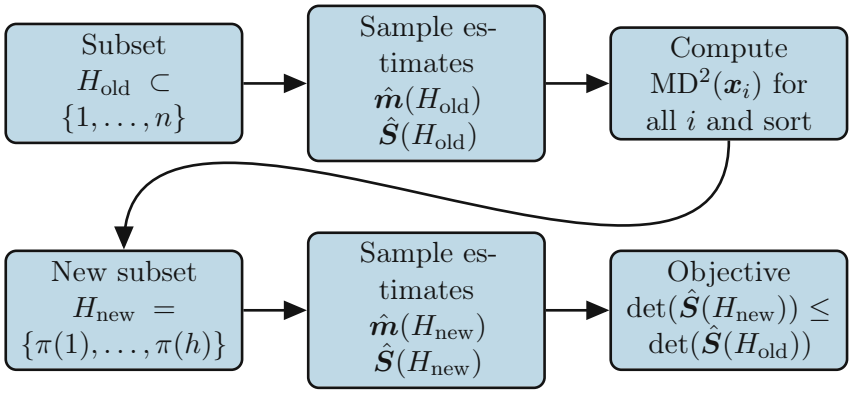


Figure 1.4: Visualization of the main steps of the concentration step (C-step) algorithm.

$\det(\hat{\boldsymbol{S}}(H_{\text{old}}))$, with equality if and only if $\hat{\boldsymbol{S}}(H_{\text{new}}) = \hat{\boldsymbol{S}}(H_{\text{old}})$ (Rousseeuw and Van Driessen, 1999, Theorem 1). Iteratively applying C-steps implies, that the objective function is decreasing in each step and convergence is reached within finitely many iterations.

Since the C-step must not necessarily lead to a global optimum, the Fast-MCD algorithm uses multiple initial h-subsets, iterating C-steps on each one until convergence, and keeps the solution with the lowest determinant. Scaling and reweighting steps are used in the Fast-MCD procedure to increase efficiency and ensure consistency under multivariate normality. Moreover, subsampling strategies are used to deal with large datasets (Rousseeuw and Van Driessen, 1999).

The MCD estimator strikes an effective balance between robustness, efficiency, and computational demands. Moreover, it was shown in Rousseeuw (1985) that the MCD estimator is affine equivariant and attains the highest possible breakdown point for an affine equivariant estimator. Butler et al. (1993) proved consistency, and Croux and Haesbroeck (1999) studied the influence function and proposed a one-step reweighted version of the MCD estimators to improve finite-sample efficiency. Cator and Lopuhaä (2012) extended the theoretical results to more general settings and proved asymptotic normality for both the MCD location and covariance estimators. Computing robust location and covariance estimators is often computationally expensive, and the MCD minimization problem (1.8) is no exception. However, with the computationally efficient Fast-MCD algorithm of Rousseeuw and Van Driessen (1999) this issue was mitigated. Moreover, Hubert et al. (2012) proposed an even faster, deterministic approach to compute the MCD estimator, and Boudt et al. (2020) proposed a regularized version of the MCD. Recently, Raymaekers and Rousseeuw (2023) showed that the MCD optimization problem (1.8) can be rewritten as a restricted maximum likelihood problem: Let $w_i \in \{0, 1\}$, $i = 1, \ldots, n$ denote a set of binary weights such that $\sum_{i=1} n w_i = h$ and $L(\boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ the negative log-likelihood function of the multivariate normal distribution. Then, minimizing

$$\sum_{i=1}^{n} w_i L(\boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

subject to $\sum_{i=1}^{n} w_i = h$, is equivalent to minimizing (1.8). Raymaekers and Rousseeuw (2023) used this connection to define a cellwise robust version of the MCD estimator, and in

Mayrhofer et al. (2024a) (Chapter 3), we proposed a matrix-variate version of the MCD estimator.

1.1.5 Cellwise Contamination

The cellwise contamination model for a p-variate random vector $\boldsymbol{X}$ proposed by Alqallaf et al. (2009) is specified as

$$\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})\boldsymbol{Y} + \boldsymbol{B}\boldsymbol{Z}, \tag{1.9}$$

where $\boldsymbol{B} = \text{diag}(B_1, \ldots, B_p)$ is a diagonal matrix with entries $B_i \sim \text{Bernoulli}(\varepsilon_i)$ for $i = 1, \ldots, p$. By assuming different dependence structures between the random variables B_i, $i = 1, \ldots, p$, different contamination models arise. If they are fully dependent, this model corresponds to the casewise model (1.1). On the other end of the spectrum, if they are fully independent and $\varepsilon_1 = \cdots = \varepsilon_p = \varepsilon_{\text{cell}}$, the probability of observing a realization without contamination is $(1 - \varepsilon_{\text{cell}})^p$. Thus, the casewise contamination probability, i.e., the probability that at least one cell of a p-variate observation is contaminated, is given by $\varepsilon_{\text{case}} = 1 - (1 - \varepsilon_{\text{cell}})^p$. This interplay between cellwise and casewise contamination probability is visualized in Figure 1.5. The blue contours display settings where $\varepsilon_{\text{case}} < 0.5$, which are separated by the black line for which $\varepsilon_{\text{case}} = 0.5$, from the red contours for which $\varepsilon_{\text{case}} > 0.5$.

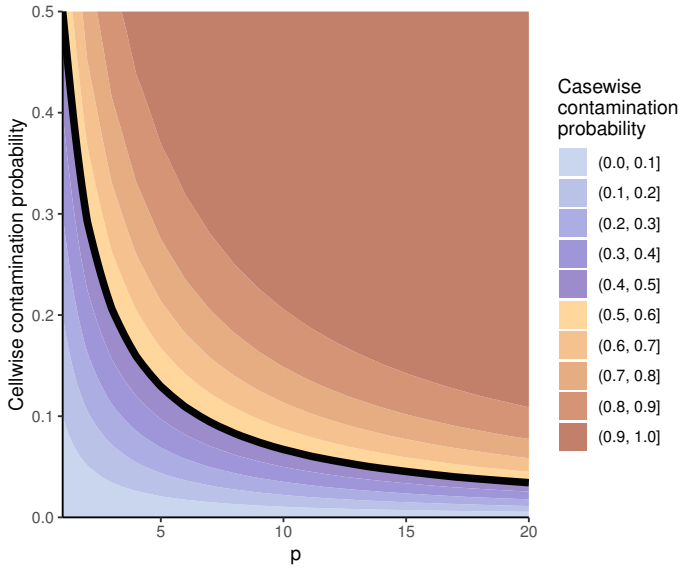


Figure 1.5: Comparison of cellwise and casewise contamination probabilities.

The cellwise and casewise contamination schemes are visualized in Figure 1.6. In both settings, the contamination probability is 10%, and outlying cells are colored black, while regular cells are gray. Under the cellwise contamination model, more than half of the rows are affected by cellwise outliers. Under this model, casewise robust procedures are no longer reliable since there is no *clean* majority of the data left.

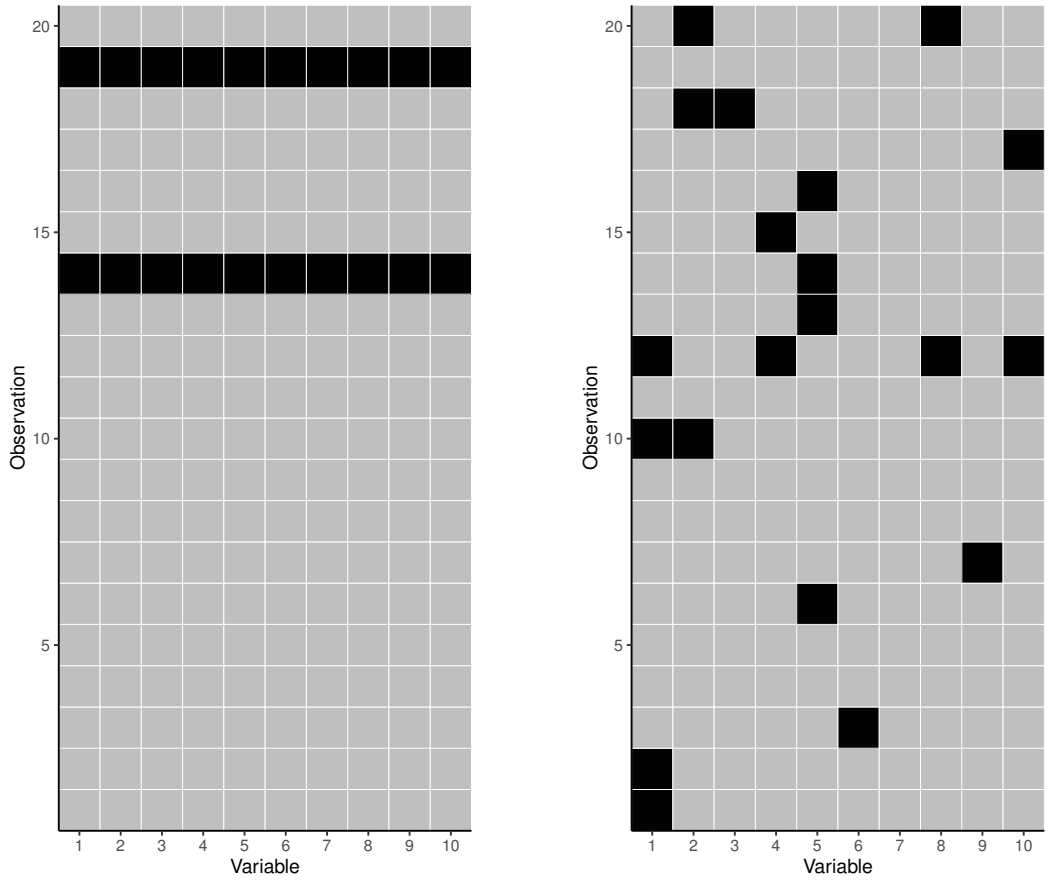


Figure 1.6: Comparison between the casewise (left) and cellwise (right) contamination models.

The cellwise contamination model can be thought of as a setting where individual cells deviate from the values they should have had. This deviation refers to instances where the observed values of certain cells substantially differ from approximations of their expected value based on the values of other cells within the same observation and the general behavior of the population. If an observed value is too different from this prediction, it is flagged as a cellwise outlier. However, this comparison process often involves higher computational demands than many casewise robust approaches, as it requires extensive calculations across all cells to identify and assess potential outliers, and necessitates iterative updates to reflect the population behavior. Therefore, cellwise robust procedures have only risen in popularity rather recently due to rapid advancements in computational power. An overview of cellwise outlier detection is given in Raymaekers and Rousseeuw (2024). Early approaches for cellwise outlier detection focused on the marginal distribution. However, those were only able to detect rather extreme cellwise outliers since dependencies between the variables were not taken into account. On the other hand, approaches that exploit multivariate relations can also detect cellwise outliers that are not marginally outlying. This includes, for instance, the Detecting Deviating Cells (DDC, Rousseeuw and Bossche (2018)) algorithm, which

robustly computes standardized residuals based on linear regressions. Raymaekers and Rousseeuw (2023) proposed a cellwise version of the casewise MCD estimator (Rousseeuw, 1985), to robustly estimate location and covariance under cellwise contamination. All of those approaches are designed to yield reliable results even when more than half of the observations are contaminated by cellwise outliers.

1.2 Explainability

Statistical models are often used to understand and explain relationships between variables and test hypotheses. These intentions lead to the development of many models that are intrinsically interpretable. For example, a classical linear regression model is transparent and understandable. Each coefficient quantifies the effect of a predictor variable on the response variable. Moreover, statistical tests can be employed to validate model assumptions and to determine whether the influence of a predictor variable is significant (Johnson and Wichern, 1998; Hastie et al., 2009).

Due to rapid increases in computational power, optimization-based, data-driven methods that increase prediction accuracy, have gained much popularity in recent times. They include, for example, approaches like gradient-boosting (Friedman, 2001) or deep learning (LeCun et al., 2015), which are focused on achieving better performance on large tabular data as well as on non-tabular data like images or text, rather than making their internal workings easy to explain. This lack of interpretability has given rise to the field of eXplainable Artificial Intelligence (XAI), also known as Interpretable Machine Learning (IML); see, e.g., Molnar (2022); Arrieta et al. (2020) for an overview.

1.2.1 Explainable AI

In XAI, the terms *transparancy*, *interpretablity*, and *explainability* are used to describe desirable properties of machine learning models and XAI approaches. Roscher et al. (2020) provide an overview and summarize the key terminology as follows: *Roughly speaking, transparency considers the machine learning approach, interpretability considers the machine learning model together with data, and explainability considers the model, the data, and human involvement.* For this overview, we will consider XAI as a set of methods to make the decisions and predictions of machine learning models understandable by humans and do not distinguish between interpretability and explainability.

Interpretability can be achieved through two main approaches. The first involves intrinsically interpretable models, which restrict model complexity to obtain interpretable results, such as sparse linear models or short trees. The second approach is based on post-hoc methods to provide explanations of complex models after training. The class of additive feature attribution methods introduced by Lundberg and Lee (2017) is model-agnostic and follows an additive model structure where the explanation for a single prediction is expressed as a sum of effects from each feature. This class includes *Local Interpretable Model-agnostic Explanations* (LIME, Ribeiro et al. (2016)), which explain individual predictions of machine learning models by local approximation with an intrinsically interpretable model. It also encompasses approaches based on Shapley values (Shapley, 1953) from cooperative game

theory to explain individual predictions (Štrumbelj and Kononenko, 2010, 2014). The explanations based on Shapley values can be combined to determine global feature importance measures (Lundberg et al., 2020).

1.2.2 Multivariate Outlier Explanation

In Mayrhofer and Filzmoser (2023) (Chapter 2), we adapted Shapley values to answer the question of why a multivariate observation is flagged as outlying. In robust statistics, the common procedure for outlier detection relies on robust estimation of location and covariance, computing Mahalanobis distances based on the robust estimators, and checking if the distances exceed a certain threshold. While this approach allows for distinguishing between regular observations and outliers, we do not gain any information about which variables might have caused the outlyingness. Using our approach based on Shapley values, we obtain multivariate outlier explanations by decomposing the squared Mahalanobis distance into variable-specific contributions. Due to the properties of Shapley values, we obtain an additive decomposition, i.e., the sum of all variable-specific outlyingness contributions is equal to the squared Mahalanobis distance. The resulting outlyingness scores contain information about all 2^p marginal outlyingness contributions for every observation and can be calculated with linear computational complexity.

Related work includes the method of Debruyne et al. (2019), who answer the same question by estimating the univariate direction of maximum outlyingness using sparse regression. Multivariate outlier explanations are also connected to the cellwise outlier detection described by the model (1.9). The goal of outlier explanation is to provide insights into why an observation is outlying, rather than answering the question of what value a cell should have had, the latter being a common perspective in the cellwise contamination setting.

1.3 Matrix-variate Data

In multivariate statistics, observations are usually given as p-dimensional vectors. In several fields, including image analysis, longitudinal studies, multivariate functional data analysis, and spatio-temporal analysis, data are naturally structured in two dimensions and given as matrices $\boldsymbol{X} \in \mathbb{R}^{p \times q}$. While vector-valued samples are usually stored in an $n \times p$ data matrix, matrix-variate observations are commonly collected in a $p \times q \times n$ data tensor, see Figure 1.7.

Oftentimes, such matrices are vectorized, which means that the cells of the matrix are converted column-by-column to a long vector, resulting in high-dimensional observations. Let $\text{vec}(\cdot)$ denote the vectorization operator, then $\boldsymbol{x} = \text{vec}(\boldsymbol{X}) \in \mathbb{R}^{pq}$ denotes the vectorized version of $\boldsymbol{X}$. With such a treatment, the inherent data structure is lost, and the dimensionality increases quickly. For example, when we stack the pixel information of a 100 by 100 pixel grayscale image, we obtain a vector with $10,000$ entries, and we rarely have enough data points for covariance estimation in such a high dimensional setting. Moreover, computation time is often a limiting factor in those settings.

Alternatively, matrix-variate data can be modeled by assuming that the underlying random matrix follows a matrix-variate distribution. The matrix normal distribution (Dawid, 1981)

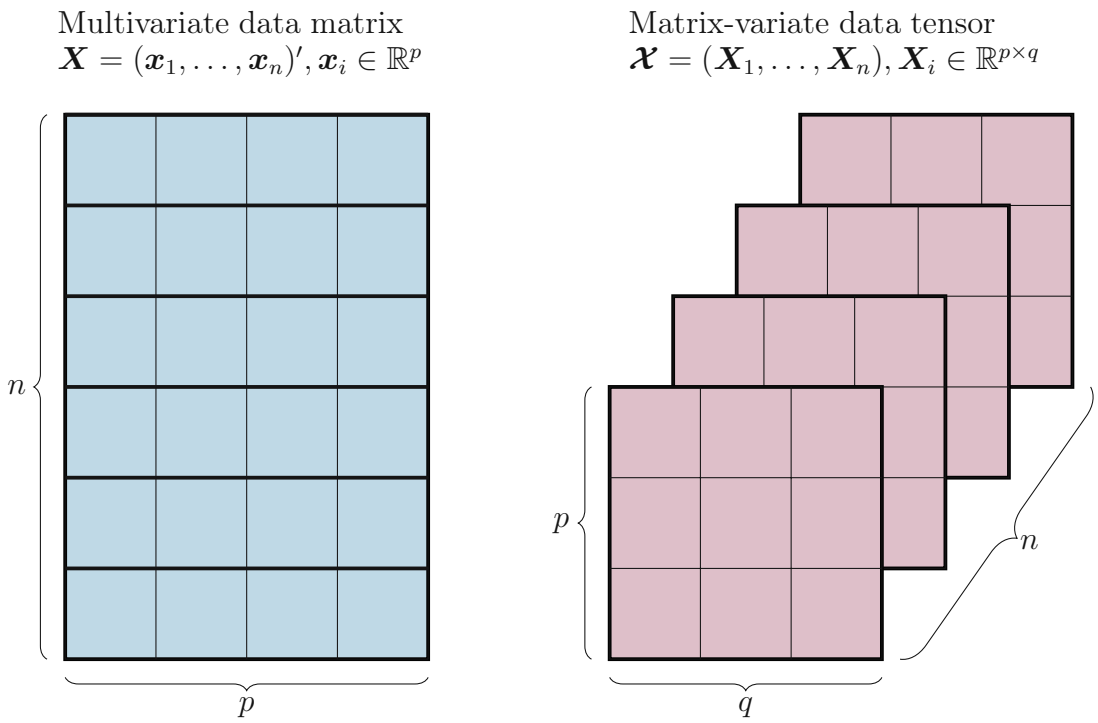


Figure 1.7: Data structure of vector-valued versus matrix-valued samples.

is the matrix-variate counterpart to the multivariate normal distribution (1.2). Formally, a random matrix $\boldsymbol{X} \in \mathbb{R}^{p \times q}$ follows a matrix normal distribution $\mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$ with mean $\mathbf{M} \in \mathbb{R}^{p \times q}$, row covariance $\boldsymbol{\Sigma}^{\text{row}} \in \text{PDS}(p)$, and column covariance $\boldsymbol{\Sigma}^{\text{col}} \in \text{PDS}(q)$, if and only if its vectorized version $\text{vec}(\boldsymbol{X}) \in \mathbb{R}^{pq}$ follows a multivariate normal distribution $\mathcal{N}(\text{vec}(\mathbf{M}), \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}})$, where $\otimes$ denotes the Kronecker product (Gupta and Nagar, 1999). The density function of the matrix normal distribution is given by

$$f(\boldsymbol{X} \,|\, \mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}) = \frac{\exp(-\frac{1}{2} \operatorname{tr}((\boldsymbol{\Sigma}^{\text{col}})^{-1}(\boldsymbol{X} - \mathbf{M})'(\boldsymbol{\Sigma}^{\text{row}})^{-1}(\boldsymbol{X} - \mathbf{M})))}{(2\pi)^{pq/2} \det(\boldsymbol{\Sigma}^{\text{col}})^{p/2} \det(\boldsymbol{\Sigma}^{\text{row}})^{q/2}}. \tag{1.10}$$

For a sample $\boldsymbol{\mathcal{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \in \mathbb{R}^{n \times p \times q}$ drawn from $\mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$, the MLEs for the mean matrix, the row covariance, and the column covariance are given by

$$\begin{aligned} \hat{\mathbf{M}} &= \bar{\boldsymbol{X}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \\ \hat{\boldsymbol{\Sigma}}^{\text{row}} &= \frac{1}{qn} \sum_{i=1}^{n} (\boldsymbol{X}_i - \mathbf{M})(\boldsymbol{\Sigma}^{\text{col}})^{-1}(\boldsymbol{X}_i - \mathbf{M})', \text{ and} \\ \hat{\boldsymbol{\Sigma}}^{\text{col}} &= \frac{1}{pn} \sum_{i=1}^{n} (\boldsymbol{X}_i - \mathbf{M})'(\boldsymbol{\Sigma}^{\text{row}})^{-1}(\boldsymbol{X}_i - \mathbf{M}), \end{aligned} \tag{1.11}$$

respectively. Due to the factored covariance structure, the covariance matrices $\boldsymbol{\Sigma}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}}$ are only identified up to a multiplicative constant $\kappa \neq 0$, since replacing $\boldsymbol{\Sigma}^{\text{row}}$ by $\kappa \boldsymbol{\Sigma}^{\text{row}}$

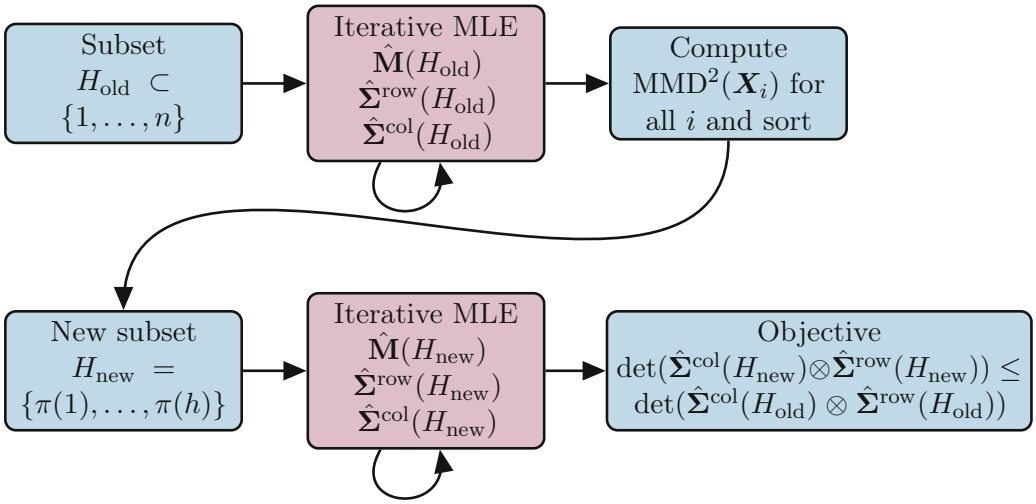


Figure 1.8: Adaptation of the concentration step algorithm for the matrix-variate setting.

and $\mathbf{\Sigma}^{\text{col}}$ by $\mathbf{\Sigma}^{\text{col}}/\kappa$ does not change the matrix normal density (1.10). Since the estimator $\hat{\mathbf{\Sigma}}^{\text{row}}$ for the row covariance depends on the inverse of the column covariance $\mathbf{\Sigma}^{\text{col}}$, and vice versa for $\hat{\mathbf{\Sigma}}^{\text{col}}$, there is no closed-form solution for the covariance maximum likelihood estimators (1.11). However, Dutilleul (1999) proposed an iterative estimation procedure that alternates between an update of $\hat{\mathbf{\Sigma}}^{\text{row}}$ and $\hat{\mathbf{\Sigma}}^{\text{col}}$ based on Equation (1.11). This procedure is often referred to as *flip-flop* algorithm. Similar approaches are also discussed in Mardia and Goodall (1993) and Brown et al. (2001). While for the vectorized samples, the computation of the estimators (1.3) would require $n \geq n + 1$ observations, the MLEs (1.11) yield positive definite estimators with probability one if $n \geq \lfloor p/q + q/p \rfloor + 2$ (Soloveychik and Trushin, 2016).

As in the multivariate case, the matrix normal distribution is a member of the class of the family of matrix elliptical distributions, which offer a more general framework. Those distributions allow us to model the row-wise and column-wise covariance separately, accounting for the inherent data structure. We provide more details about matrix elliptical distributions in Chapter 3 and propose robust estimators for the row and column covariance matrices based on the MCD estimator described in Section 1.1.4. Moreover, we also present an algorithm for an efficient computation based on a C-step for matrix-variate observations. This includes a matrix-variate version of the C-step that replaces the sample mean and covariance estimators, which are the MLEs under the Gaussian model, with the matrix-variate MLEs for the matrix-variate normal distribution. Since the optimization problem of the MCD can be expressed as a trimmed ML problem, the C-step can be seen as an algorithm for computing trimmed ML estimators. Using this connection in combination with the iterative ML estimators for covariance estimation in the matrix-variate setting, we can generalize the C-step to the matrix-variate setting, as illustrated in Figure 1.8. Here, MMD^2 denotes the matrix-variate squared Mahalanobis distance, as detailed in Chapter 3.

Lu and Zimmerman (2005) derived the likelihood ratio test to check whether the separability assumption is valid on the basis of a random sample from a multivariate normal

population. The test requires a large number of samples relative to the dimensions of the data since it relies on an estimate of the full covariance structure when separability is not assumed. Filipiak et al. (2016) proposed Rao's score test for separability, which does not rely on the estimation of the full covariance structure. However, both tests are not robust against anomalies, and few outlying samples could yield a rejection of the separability assumption even if the majority of the data have a separate covariance structure.

1.4 Functional Data Analysis

In Functional Data Analysis (FDA), the observed data are functions instead of real numbers or vectors. FDA offers a set of methods to deal with observations as elements of a function space. In the following, we provide a partial overview and introduction to common methods in FDA based on the books of Ramsay and Silverman (2005); Kokoszka and Reimherr (2017) and reviews of Cuevas (2014); Wang et al. (2016), focusing on the univariate setting. The multivariate setting is described in Chapter 4.

The most used functional space is L^2, the space of square-integrable functions. Let $\mathcal{T} \subset \mathbb{R}$ denote a compact interval and $x, y \in L^2(\mathcal{T})$, then the inner product and norm are given by

$$\langle x, y \rangle = \int_{\mathcal{T}} x(t)y(t)\mathrm{d}t \quad \text{and} \quad \|x\| = \langle x, x \rangle^{1/2},$$

respectively. An L^2 process X has finite second moments, i.e., $\mathbb{E}[\|X(t)\|^2] < \infty$ for all $t \in \mathcal{T}$ and we consider continuous L^2 processes, i.e., $\lim_{h \to 0} \mathbb{E}[\|X(t+h) - X(t)\|^2] = 0$ for every $t \in \mathcal{T}$.

1.4.1 Basis Representation

In FDA, observations can be seen as realizations of a stochastic process. The values are usually recorded on a discrete grid $t_1, \ldots, t_q$, and we observe a possibly high dimensional vector $(X(t_1), \ldots, X(t_q)) \in \mathbb{R}^q$, $t_j \in \mathcal{T}$, $j = 1, \ldots, q$. Preliminary treatment of the discretely observed samples $(X(t_1), \ldots, X(t_q))$ is necessary to transform them into functional data, and basis representation is a very common approach. Let $\{\phi_j(t)\}_{j=1}^{\infty}$, $t \in \mathcal{T}$ be a basis of $L^2(\mathcal{T})$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$, $m \in \mathbb{N}$, $m \leq q$ a finite collection of basis functions that spans an m-dimensional subspace of $L^2(\mathcal{T})$. Then we can obtain a smooth version $\tilde{X}(t) = \sum_{j=1}^{q} a_j \, \phi_j(t)$ of $X(t)$ by minimizing

$$\sum_{k=1}^{q} \left(X(t_k) - \sum_{j=1}^{m} a_j \, \phi_j(t_k) \right)^2 . \tag{1.12}$$

The smoothing step is often motivated either in terms of dimension reduction or noise removal in the data. In its simplest form, the smooth representation is obtained by

$$(X(t_1), \ldots, X(t_q)) \mapsto (a_1, \ldots, a_m) \mapsto (\tilde{X}(t_1), \ldots, \tilde{X}(t_q))$$

and we receive a more compact representation ($m < q$) and a denoised process, since $\tilde{X}(t)$, $t \in \mathcal{T}$, yields a smoothed version of the original process $X(t)$, $t \in \mathcal{T}$ (Cuevas, 2014).

Common choices for basis functions include B-splines or the Fourier basis, the first five basis functions of which are visualized in Figure 1.9, respectively. While the Fourier basis is commonly used for smoothing periodic or seasonal data, B-splines are applied more broadly across various types of data.

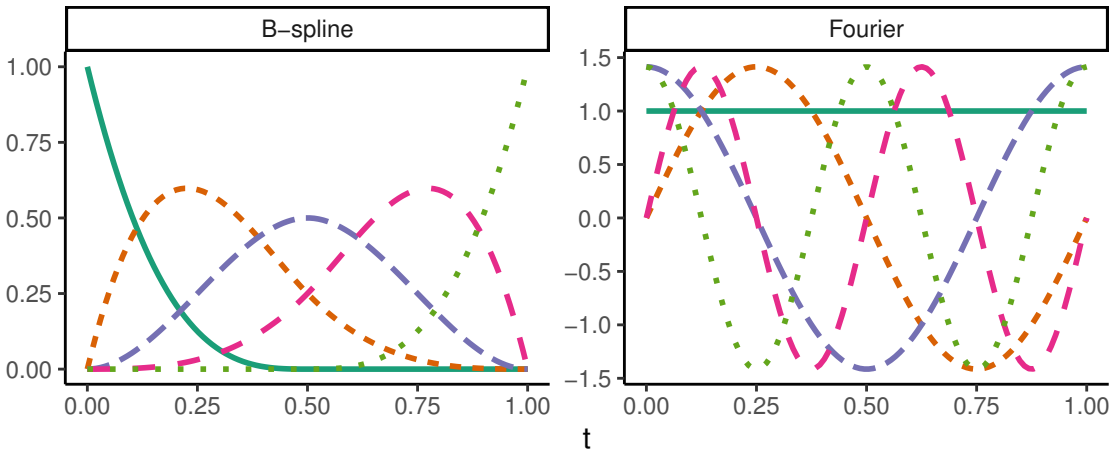


Figure 1.9: First five basis functions of the cubic B-spline basis (left) and the Fourier basis (right).

There exist various other prepossessing procedures such as kernel smoothing, local polynomial fitting, or penalized smoothing. For the latter, a roughness penalty is introduced into the minimization problem (1.12), which allows for finer control over the degree of smoothness of the function and enables us to use more than q basis functions. In Figure 1.10 the raw data are $X(t) = \sin(\pi t) + \varepsilon$, $t \in [0, 1]$, and ε are i.i.d. normal errors sampled at $q = 100$ time points. For illustration purposes, we show smooth approximations using B-splines with either $m = 25$ basis functions and no penalty, or $m = 200$ basis functions and a roughness penalty on the second derivative. In this simple example, the penalized approach yields more accurate results since the underlying function is very smooth.

Based on functional representations of the data, the generalizations of the sample mean and covariance (1.3) are defined as follows: Let $\{x_1(t), \ldots, x_n(t)\}$, $t \in \mathcal{T}$, denote a sample of n functional observations from X, then the sample mean and sample covariance function are given by

$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t) \quad \text{and} \quad \hat{\kappa}(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i(s) - \hat{m}(s))(x_i(t) - \hat{m}(t)),$$

for $s, t \in \mathcal{T}$, respectively. Their population counterparts are given by

$$\mu(t) = \mathbb{E}[X(t)] \quad \text{and} \quad \kappa(s,t) = \operatorname{cov}(X(s), X(t)) = \mathbb{E}\left[(X(s) - \mu(s))(X(t) - \mu(t))\right].$$

1.4.2 Functional Principal Component Analysis (FPCA)

Functional Principal Component Analysis (FPCA) is one of the most widely used tools in FDA. Similar to PCA in multivariate statistics, the goal of FPCA is to decompose functional

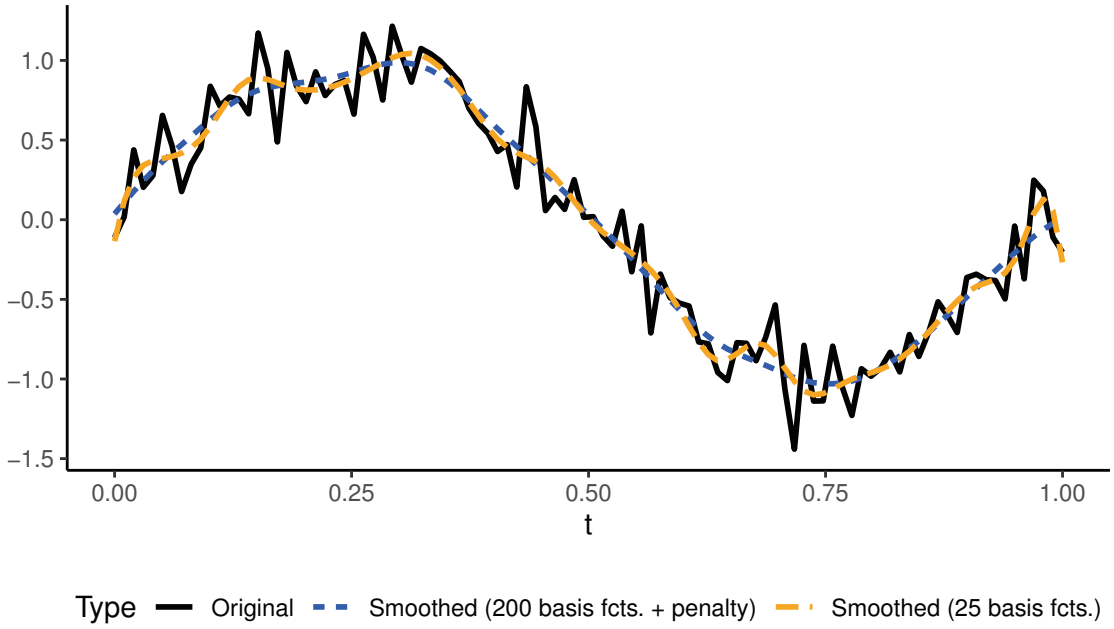


Figure 1.10: Smooth approximations of a random function.

data into a set of uncorrelated functional principal components (FPCs), which explain as much variability within the data as possible. FPCA is often used for exploratory data analysis since it allows us to identify functional features that characterize typical functions by extracting key functional features that describe typical patterns and trends in the data. The FPCs for observed data provide an *optimal* basis because they yield a basis that maximizes the explained variance in each component. Additionally, dimension reduction can be achieved by projecting the observed functions onto those FPCs that explain most of the variability.

Let $X \in L^2(\mathcal{T})$ be a continuous stochastic process with mean function $\mu(t)$ and covariance function $\kappa(s,t)$, $s,t \in \mathcal{T}$, with the associated covariance operator $\mathcal{K} : L^2(\mathcal{T}) \to L^2(\mathcal{T})$ defined as

$$\mathcal{K}f(s) = \int_{\mathcal{T}} c(s,t)f(t)\mathrm{d}t,\ f \in L^2(\mathcal{T}).$$

Then we can characterize the functional principal components by eigenanalysis of the covariance operator $\mathcal{K}$. Mercer's theorem yields the eigendecomposition of $\mathcal{K}$ given by

$$\mathcal{K}\,\xi_k = \lambda_k\,\xi_k \quad \text{and} \quad \kappa(s,t) = \sum_{k=1}^{\infty} \lambda_k\,\xi_k(s)\,\xi_k(t),$$

where $\{\xi_k\}_{k=1}^{\infty} \in L^2(\mathcal{T})$ is the countable sequence of continuous orthonormal eigenfunctions with non-negative decreasing eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$, $\sum_{k=1}^{\infty} \lambda_k < \infty$. The Karhunen-Loève representation theorem implies that X can be written as

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \alpha_k\,\xi_k \quad \text{with} \quad \alpha_k = \int_{\mathcal{T}} \xi_k(t)(X(t) - \mu(t))\mathrm{d}t,$$

where $\alpha_k \sim \mathcal{N}(0, \lambda_k)$, if X is a Gaussian process, see Daw et al. (2022) for more details. The deterministic functions ξ_k are the FPCs, the random variables α_k are the scores, and the total variance of X is equal to the sum of the variances λ_k of the projections of X on the FPCs ξ_k. To compute the empirical FPC $\hat{\xi}_k$, the continuous functional eigenanalysis problem (1.4.2) is typically approximated by either discretizing the functions or using the basis representation to transform it into an approximately equivalent matrix eigenanalysis problem. Figure 1.11 shows the first four empirical FPCs computed from a sample of $n = 300$ smoothed Wiener process trajectories. As for the previous example illustrated in Figure 1.10, the raw data were sampled at $q = 100$ time points and smoothed using either $m = 25$ basis functions and no penalty, or $m = 200$ basis functions and a roughness penalty on the second derivative. If the FPCs are computed based on the coefficient matrix rather than on discretization, the resulting coefficient matrices are of size $n \times m$ and the eigenanalysis is then performed on an $m \times m$ matrix, see, e.g., Ramsay and Silverman (2005) for details. Hence, using only $m = 25$ basis functions instead of $m = 200$ in our example results in a computationally easier task.

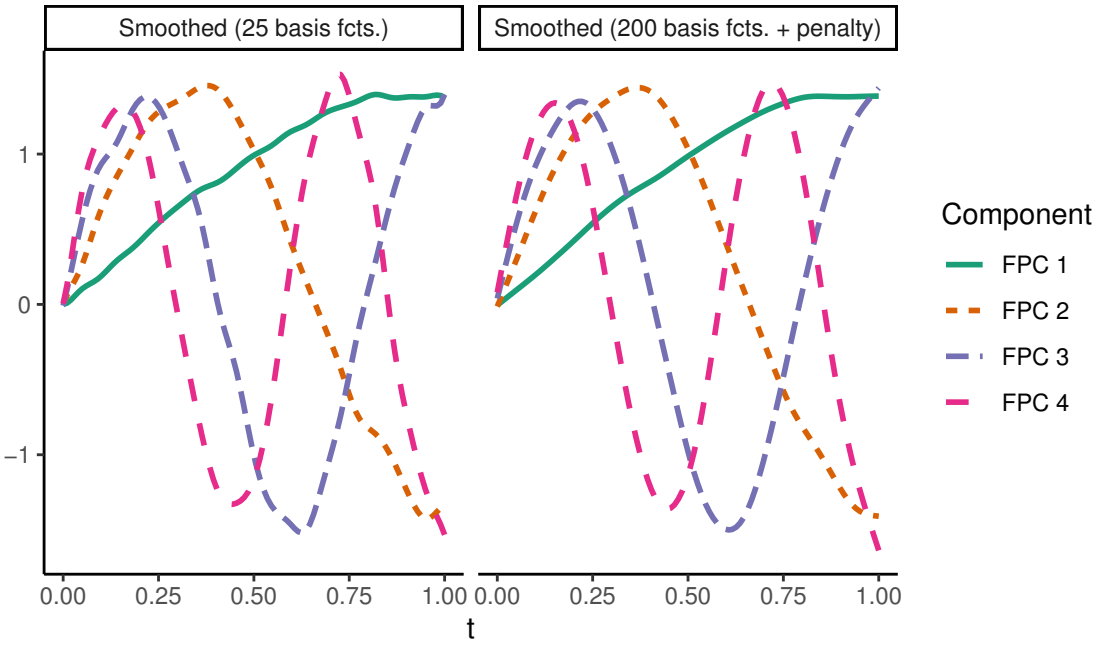


Figure 1.11: Empirical FPCs for a Wiener process.

1.4.3 Outlier Detection

Outlier detection procedures for functional data can generally be categorized as either depth-based or distance-based approaches. Statistical depth functions provide a measure of the centrality of an observation with respect to a dataset or distribution. Functional depth measures provide an ordering within a sample of curves and enable the definition of ranks. The concept of depth was originally introduced in multivariate data analysis as a way to

generalize order statistics, ranks, and medians to higher-dimensional spaces, see, e.g., Zuo and Serfling (2000) for an overview and a discussion of key properties. López-Pintado and Romo (2009) introduced the concept of depth for functional data and proposed the band depth as a measure of centrality for functions. Let $G(x) = \{(t, x(t)) : t \in \mathcal{T}\}$ denote the graph of a real-valued function $x(t), t \in \mathcal{T}$, and let $\{x_1(t), \ldots, x_n(t)\}$ denote a sample of functional observations. The band in $\mathbb{R}^2$ that is delimited by the curves $\{x_i\}_{i\in A}$, $A \subseteq \{1, \ldots, n\}$, is defined as

$$B(\{x_i\}_{i\in A}) = \{(t, y(t)) : t \in \mathcal{I}, x_{\min}(t) \leq y(t) \leq x_{\max}(t)\},$$

where $x_{\min}(t) = \arg\min_{i\in A} x_i(t)$ and $x_{\max}(t) = \arg\max_{i\in A} x_i(t)$. For any function $x \in \{x_i\}_N$, $N = \{1, \ldots, n\}$, $2 \leq j \leq n$, the quantity

$$\mathrm{BD}_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{A\subseteq N, |A|=j} \mathbb{1}\{G(x) \subseteq B(\{x_i\}_{i\in A})\}$$

is the proportion of bands $B(\{x_i\}_{i\in A})$ determined by j different curves $\{x_i\}_{i\in A}$ containing the whole graph of x. Here, $\mathbb{1}\{B\}$ is one if B is true and zero otherwise. The band depth of x for fixed $J \in \mathbb{N}$, $2 \leq J \leq n$ is given by

$$\mathrm{BD}_{n,J}(x) = \sum_{j=2}^{J} \mathrm{BD}_n^{(j)}(x). \tag{1.13}$$

The band depth (1.13) measures how often the function x is enclosed within the bands formed by other subsets of the functional data. López-Pintado and Romo (2009) suggest using $J = 3$ as a default value. The band depth is the basis for the functional boxplot by Sun and Genton (2011), which extends the univariate boxplot by using functional depth instead of traditional univariate ranks. It summarizes a sample of functional observations based on three key features: the functional median, the 50% central region, and the fence. The functional median is the curve with the highest depth and is the most central observation. The band delimited by 50% of the deepest curves defines the 50% central region. The fence is created by expanding this central region by 1.5 times the range of the 50% central region. Using the functional boxplot, a functional observation is identified as an outlier if it is outside of the fence at least at one point. Typically, the band delimiting the regular observations is plotted instead of the fence. For an overview of depth-based outlier detection procedures for univariate and multivariate functional data, we refer to Hubert et al. (2015).

To illustrate the functional boxplot, we consider the monthly sea surface temperature (SST) data related to the El Niño–Southern Oscillation (ENSO) phenomenon. The data cover 74 periods from 1950-1951 to 2023-2024, with SST measurements across four regions in the equatorial Pacific. Further details on the data are provided in Chapter 4. Figure 1.12 shows the functional boxplots for the four regions. The solid line within the shaded area is the functional median, the shaded area is the 50% central region, the area between the dashed lines is the band delimiting the regular observations, the lighter lines are the sample curves, and the labeled lines are the outliers. The period 1997:1998 is flagged as an outlier in the Niño 1+2 and Niño 3 regions. In the other two regions, no observations are flagged.

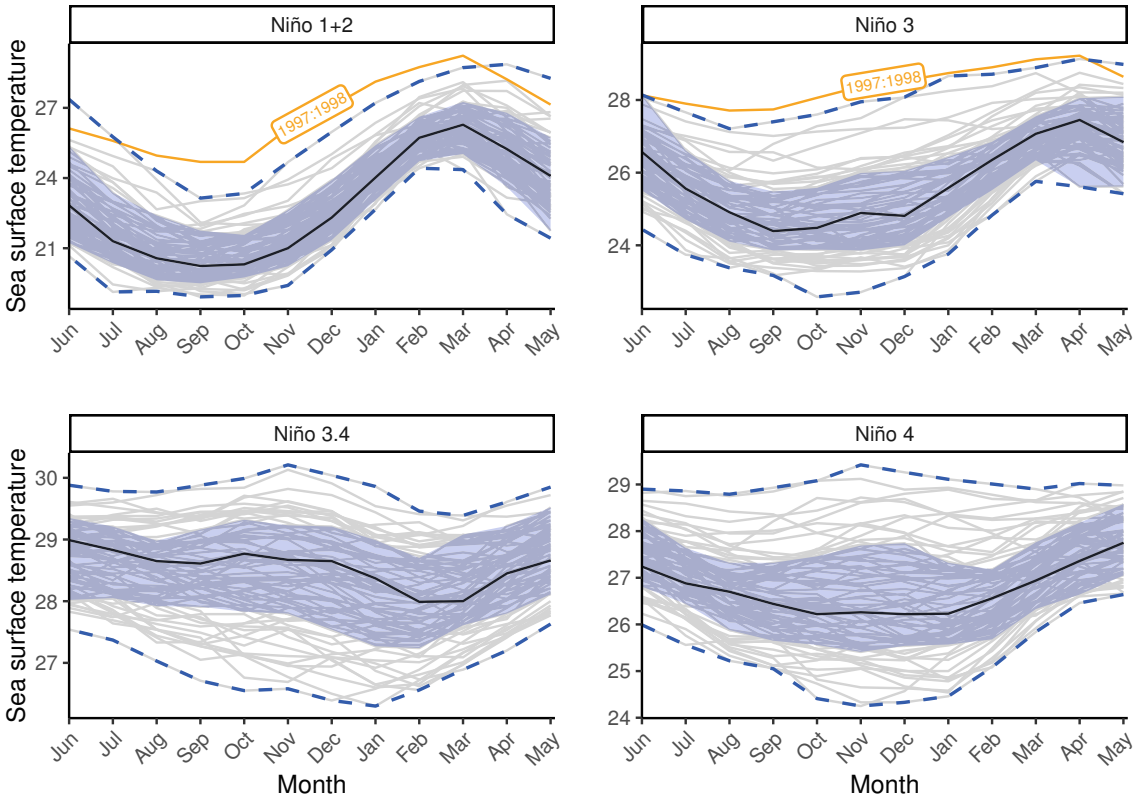


Figure 1.12: Functional boxplots of the monthly SST measurements across four regions in the equatorial Pacific. The solid black line in the shaded blue area represents the functional median. The shaded area indicates the 50% central region. The region between the dashed blue lines delineates the regular observations. The light gray lines depict the sample curves, and the labeled orange lines mark outliers.

In Chapter 4, we use a distance-based approach that accounts for the dependencies between the four SST measurement functions, and provide a comprehensive analysis of the data.

Many distance-based approaches for functional outlier detection focus on a generalization of the Mahalanobis distance to the functional setting. The standard Mahalanobis distance (1.5) is based on the inverse covariance matrix. The covariance operator $\mathcal{K}$ takes the same role in the functional context as the covariance matrix does in the multivariate setting, but it is not invertible in general. Hence, it is not straightforward how a functional Mahalanobis distance should be defined. Galeano et al. (2015) proposed a method that relies on spectral cutoff regularization to define a regularized covariance operator. Ghiglietti et al. (2017) and Berrendero et al. (2020) proposed further approaches to regularize the covariance operator. Most research regarding functional Mahalanobis distance considers univariate functional data. In Chapter 4, we introduce a method to robustly estimate location and covariance for multivariate functional data and compute truncated Mahalanobis (semi-)distances based on the approach of Galeano et al. (2015).

2 Multivariate Outlier Explanations Using Shapley Values and Mahalanobis Distances

This chapter was published as Mayrhofer, M. and Filzmoser, P. (2023). Multivariate outlier explanations using Shapley values and Mahalanobis distances. *Econometrics and Statistics*. DOI: 10.1016/j.ecosta.2023.04.003.

Contributions: M. Mayrhofer developed the methodology, established the proofs, and implemented the proposed procedures in the `R` package `ShapleyOutlier` (Mayrhofer and Filzmoser, 2022). He wrote the first draft of the paper, participated in discussions, and finalized the paper together with the co-author.

2.1 Introduction

Multivariate outlier detection is a topic of unabated popularity in statistics and computer science. Not only does there exist a wide variety of approaches but also the terminology varies; anomaly detection, novelty detection, or fraud detection all refer to the problem of identifying unusual behavior (Zimek and Filzmoser, 2018). In a dataset with n observations measured at p variables, one is interested in identifying observations that do not conform to their expected behavior according to the remaining (neighboring) observations (Chandola et al., 2009; Grubbs, 1969).

A widespread tool for the detection of multivariate outliers in statistics is based on the Mahalanobis distance (Mahalanobis, 1936). Generally, for an observation vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$ from a population with expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance matrix $\boldsymbol{\Sigma}$, the squared Mahalanobis distance of $\boldsymbol{x}$ to $\boldsymbol{\mu}$ with respect to $\boldsymbol{\Sigma}$ is given as

$$\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{2.1}$$

and will be denoted as $\mathrm{MD}^2(\boldsymbol{x})$. To specify the outlyingness of an observation from a given sample, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ need to be estimated, with their estimators being denoted as $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. If the underlying distribution is a multivariate normal distribution, it is common to use the 0.975 quantile of a chi-square distribution with p degrees of freedom $\chi^2_{p;0.975}$ as a cutoff value (Rousseeuw and Zomeren, 1990). Observations with a squared Mahalanobis distance exceeding this cutoff are identified as multivariate outliers. It is evident that for outlier detection, the estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ themselves must be robust against such outliers. Many different proposals for robust estimation of multivariate location and covariance can be found throughout the literature, with one of the most popular being the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985).

The squared Mahalanobis distance from Equation (2.1) can also be written as

$$\mathrm{MD}^2(\boldsymbol{x}) = \sum_{j=1}^{p}\sum_{k=1}^{p}(x_j - \mu_j)(x_k - \mu_k)\omega_{jk}, \tag{2.2}$$

where ω_{jk} denotes the element (j,k) of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. This outlyingness measure collects distance contributions of all pairwise variable combinations, weighted by ω_{jk}, resulting in a single number. However, this value cannot be interpreted in the sense of contributions from individual variables, which would be vital for determining the effect of the single variables on the overall outlyingness.

Analyzing the contributions of individual variables is also of major interest in Explainable Artificial Intelligence, which is often referred to as Interpretable Machine Learning. For example, suppose a "black-box" classifier has been trained on a dataset; it is often essential to know how and why the individual variables of an observation contribute to the model's decision to assign an observation to a particular class (Ribeiro et al., 2016). Various tools have been established for this purpose, and Shapley values are among the most popular ones. Although the Shapley value (Shapley, 1953) was initially proposed in the context of game theory in 1953, it was applied much later in the context of machine learning by Štrumbelj and Kononenko (2010, 2014) and its popularity increased greatly after the publications of Lundberg and Lee (2017); Lundberg et al. (2018, 2020). We refer to Molnar (2022) and Biecek and Burzykowski (2021) for a more exhaustive discussion of these methods.

In this paper, we propose using the Shapley value for multivariate outlier explanation, which will be directly based on the squared Mahalanobis distance. Our method allows us to determine the individual variable contributions to the outlyingness and to answer the question *why* an observation is flagged as a multivariate outlier. The arguably most critical disadvantage of the Shapley values in a general setting is their high computational complexity, which exponentially increases with the number of variables. However, we will show that the Shapley values resulting from our approach can be expressed as a simplified problem, substantially facilitating their computation, even in a higher dimension. In addition, we present an extension of this concept that enables the assignment of outlyingness scores to pairs of variables, allowing the evaluation of interaction effects.

It should be mentioned that an alternative approach to answer which variables contribute the most to the multivariate outlyingness of an observation has been presented by Debruyne et al. (2019), who estimate the univariate direction of maximum outlyingness using sparse regression. Nevertheless, this method does not result in an additive decomposition of the squared Mahalanobis distance.

Another approach closely related to outlier explanation is called cellwise outlier detection. For an overview of this relatively recent research field, we refer to Raymaekers and Rousseeuw (2021). Its main idea is to investigate the outlyingness of each cell of a data matrix instead of focusing on entire observations. In general terms, cellwise outlyingness is based on the difference of the actual value of a cell compared to the value we would have expected.

Computing the amount by which a cell is anomalous is also related to multivariate outlier explanation. However, the approach is somehow reversed: To obtain explanations for the outlyingness of a single row, we decompose its squared Mahalanobis distance into outlyingness contributions for every cell using the Shapley value. In comparison, cellwise

outlier detection methods must first identify a subset of clean cells for every row, which is used to derive the value a cell should have had. This is commonly done using the conditional expectation, and the resulting outlyingness scores are then based on the difference between a cell's actual value and its conditional expectation.

The remainder of this paper is structured as follows: In Section 2.2 we introduce Shapley values before we derive in detail how to apply them for multivariate outlier explanation using squared Mahalanobis distances. Moreover, we outline how to combine those results with the concept of cellwise outlier detection, leading to the cellwise robust outlier explanation algorithms described in Section 2.3. The performance of those outlier explanation tools for cellwise outlier detection is demonstrated via the numerical experiments presented in Section 2.4. In Section 2.5 we analyze the performance of our method on real-world examples. The final Section 2.6 summarizes the key points of our findings.

2.2 Shapley Values for Outlier Explanation

In the following, we propose a method for the interpretation of multivariate outliers that combines squared Mahalanobis distances with Shapley values (Shapley, 1953). The concept of Shapley values is briefly introduced based on its nascent field of research, namely cooperative game theory (Peters, 2008).

2.2.1 Shapley Values and Cooperative Game Theory

In cooperative game theory, players can form coalitions that produce a payoff and decide how their coalitions' proceeds are distributed among them.

Definition 2.2.1.1. *A coalitional (cooperative) game with transferable utility (TU-game) (T, v) is given by a set of players $T = \{1, 2, \ldots, t\}$ and the characteristic function v, which assigns the worth $v(S) \in \mathbb{R}$ to each coalition $S \subseteq T$, such that $v(\emptyset) = 0$.*

In other words, the function v tells us how much collective payoff a coalition S of players can gain by cooperating. A payoff distribution for the grand coalition T is given by $\boldsymbol{\varphi}(v) = (\varphi_1(v), \ldots, \varphi_t(v))'$, where $\varphi_j(v) \in \mathbb{R}$ is the payoff to player j. There are several proposals on how the payoff should be assigned to the players $j \in T$ to obtain a *fair* distribution. While there are different concepts and notions of fairness in the literature, we will focus on the one introduced by Shapley (1953). The *Shapley value* $\boldsymbol{\phi}(v)$, with coordinates

$$\phi_j(v) = \sum_{S \subseteq T \setminus \{j\}} \frac{|S|!(t - |S| - 1)!}{t!} \left(v(S \cup \{j\}) - v(S)\right), \tag{2.3}$$

is the *unique* payoff distribution that fulfills the following conditions (Young, 1985):

- *Efficiency:* The payoff to individual players $\varphi_j(v)$ must add up to the worth of the grand coalition $v(T)$, hence $\sum_{j=1}^{p} \varphi_j(v) = v(T)$.
- *Symmetry:* If $v(S \cup \{j\}) = v(S \cup \{k\})$ holds for all $S \subseteq T \setminus \{j, k\}$ for two players j and k, then $\varphi_j(v) = \varphi_k(v)$.

- *Monotonicity:* If for any two games (T, v_1) and (T, v_2) and all $S \subseteq T$ the condition

$$v_1(S \cup \{j\}) - v_1(S) \geq v_2(S \cup \{j\}) - v_2(S)$$

is satisfied, then $\phi_j(v_1) \geq \phi_j(v_2)$.

Therefore, the Shapley value permits the definition of a fair payoff distribution for the grand coalition T. The term $v(S \cup \{j\}) - v(S)$ describes the marginal contribution of player j to a coalition S. The corresponding Shapley value $\phi_j(v)$ is then given as the weighted mean of the marginal contributions formed over all possible coalitions.

2.2.2 Linking Shapley Value and Mahalanobis Distance

Let us consider an observation vector $\boldsymbol{x} = (x_1, \ldots, x_p)'$ from a population with expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance matrix $\boldsymbol{\Sigma}$. We would like to investigate the contribution of the j-th coordinate x_j to the outlyingness of $\boldsymbol{x}$. The set of players is denoted as $P = \{1, \ldots, p\}$, and it contains the indices of all variables. A coalition S is formed by a subset of P. We define the characteristic function v mentioned above as the squared Mahalanobis distance

$$\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^S) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \tag{2.4}$$

with $\hat{\boldsymbol{x}}^S = (\hat{x}_1^S, \ldots, \hat{x}_p^S)'$ and

$$\hat{x}_j^S := \begin{cases} x_j & \text{if } j \in S \\ \mu_j & \text{if } j \notin S \end{cases}, \tag{2.5}$$

which fulfills $\mathrm{MD}^2(\hat{\boldsymbol{x}}^S) = 0$, if $S = \emptyset$ is the empty set.

In this setting, the k-th coordinate of the Shapley value from Equation (2.3) is given as the weighted average of the marginal contributions

$$\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) := \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^S)$$

over all 2^{p-1} subsets $S \subseteq P \setminus \{k\}$. This suggests an exponential computational complexity, which becomes costly, especially if p is large. However, the following theorem shows that this highly demanding problem can be reduced to linear complexity.

Theorem 2.2.2.1. *Given two vectors $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^p$ and a non-singular matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, the contribution of the k-th variable to the squared Mahalanobis distance $\mathrm{MD}^2(\boldsymbol{x})$ based on the Shapley value is given by*

$$\phi_k(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \sum_{S\subseteq P\setminus\{k\}} \frac{|S|!(p - |S| - 1)!}{p!} \Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \tag{2.6}$$

$$= (x_k - \mu_k) \sum_{j=1}^{p} (x_j - \mu_j)\omega_{jk}, \tag{2.7}$$

with $\boldsymbol{\Sigma}^{-1} =: \boldsymbol{\Omega} = (\omega_{jk})_{j,k=1,\ldots,p}$ and $\hat{\boldsymbol{x}}^S$ as in Equation (2.5).

Proof. The proof of this theorem is given in A.1. □

Indeed, we can compute the expression of Equation (2.7) as an intermediate result when we compute the squared Mahalanobis distance, see Equation (2.2).

The Shapley value of an observation $\boldsymbol{x}$ resulting from Theorem 2.2.2.1 is given by the vector

$$\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\phi_1(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \ldots, \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}))' \tag{2.8}$$

and we will simply denote it as $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), ..., \phi_p(\boldsymbol{x}))'$, whenever the (robustly estimated) mean and covariance matrix are employed for its computation. Considering Theorem 2.2.2.1, it is straightforward to see that

$$\boldsymbol{\phi}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu}) \circ \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{2.9}$$

where $\circ$ denotes the element-wise product.

Since $\boldsymbol{\phi}(\boldsymbol{x})$ is based on the Shapley value, it is the only decomposition of the squared Mahalanobis distance with the characteristic function defined in Equation (2.4) that fulfills the following properties:

- *Efficiency:* The contributions $\phi_j(\boldsymbol{x})$, for $j = 1, \ldots, p$, sum up to the squared Mahalanobis distance of $\boldsymbol{x}$, hence

$$\sum_{j=1}^{p} \phi_j(\boldsymbol{x}) = \mathrm{MD}^2(\boldsymbol{x}). \tag{2.10}$$

- *Symmetry:* If $\mathrm{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{j\}}) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{k\}})$ holds for all subsets $S \subseteq P \setminus \{j, k\}$ for two coordinates j and k, then $\phi_j(\boldsymbol{x}) = \phi_k(\boldsymbol{x})$.

- *Monotonicity:* Let $\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}} \in \mathbb{R}^p$ be two vectors and $\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{p\times p}$ be two non-singular matrices. If

$$\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^{S\cup\{j\}}) - \mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^{S}) \geq \mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{x}}^{S\cup\{j\}}) - \mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{x}}^{S})$$

  holds for all subsets $S \subseteq P$, then $\phi_j(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \geq \phi_j(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$.

A single coordinate $\phi_j(\boldsymbol{x})$ of the Shapley value defined in Theorem 2.2.2.1 can be interpreted as the average marginal contribution of the j-th variable to the squared Mahalanobis distance of an individual observation $\boldsymbol{x}$. While the squared Mahalanobis distance aggregates all distance contributions and results in an outlyingness measure for an entire observation, Equation (2.7) reveals that a coordinate $\phi_j(\boldsymbol{x})$ of the Shapley value only accounts for those distance contributions that are related to the j-th variable. This also implies that the outlyingness contribution of the j-th variable is connected to all other variables since a large distance of another variable to its mean influences the contribution of the j-th variable. However, the weighting based on the precision matrix alleviates this issue, since only variables that are not conditionally independent influence the score, at least in the case of elliptically distributed data (Baba et al., 2004). The efficiency property stated in Equation (2.10) indicates that we obtain an additive decomposition of the squared Mahalanobis distance into variable contributions, where a large value of $\phi_j(\boldsymbol{x})$ indicates a large contribution of the

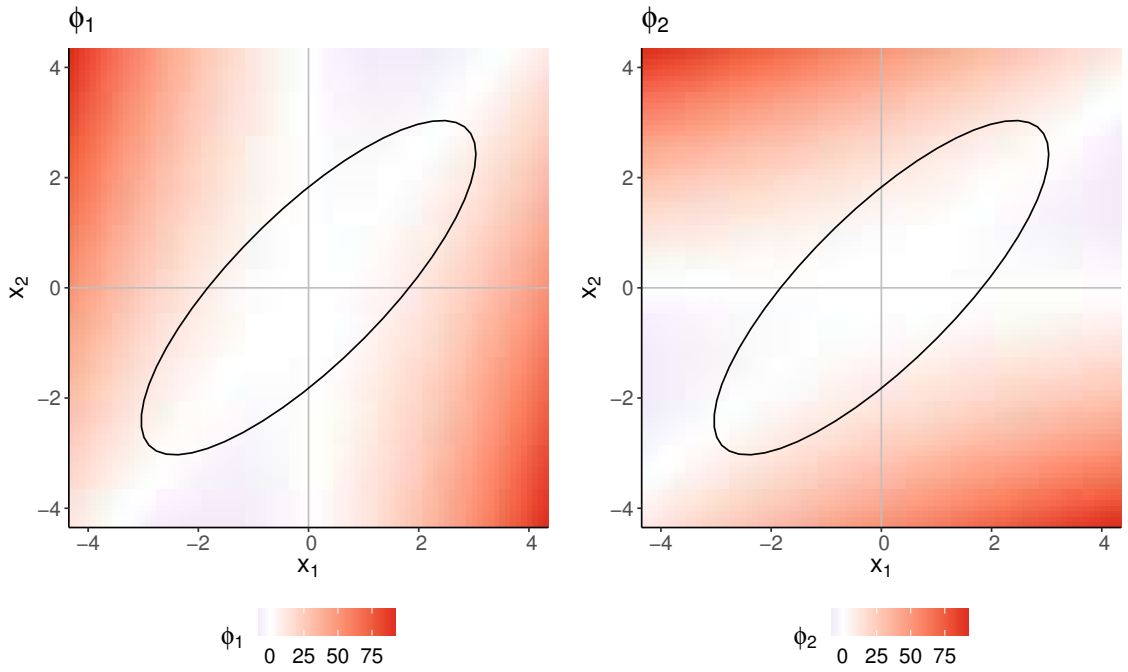


Figure 2.1: Plots illustrating a two-dimensional visualization of the Shapley values $\boldsymbol{\phi}(\boldsymbol{x})$ for $\boldsymbol{x} \in [-4, 4] \times [-4, 4]$ with mean $\boldsymbol{\mu} = (0, 0)'$ and covariance matrix $\boldsymbol{\Sigma}$, with elements $\sigma_{12} = \sigma_{21} = 0.8$, and $\sigma_{11} = \sigma_{22} = 1$. The graphs are colored according to the components $\phi_1(\boldsymbol{x})$ and $\phi_2(\boldsymbol{x})$ of the Shapley value, respectively, and both panels show the 99-percentile confidence ellipse.

j-th coordinate to $\text{MD}^2(\boldsymbol{x})$. It should be noted that the contributions can also be negative, as illustrated in Figure 2.1.

Remark: The definition given in Equation (2.5), where $\hat{x}_j^S = \mu_j$ if $j \notin S$, could also be modified. In the literature, it is often suggested to use the conditional expectation of x_j, given all other variables with index contained in S, instead of the expected value (Lundberg and Lee, 2017). However, evaluating the conditional expectation explicitly requires us to impose distributional assumptions or to apply approximation techniques, and this may also lead to high computational complexity, as for every coordinate $j \in P$ there are 2^{p-1} possible subsets S. Our definition of $\hat{\boldsymbol{x}}^S$ results in two major advantages:

1. The computational complexity of computing the Shapley value reduces from an exponential to a linear one; see Theorem 2.2.2.1.

2. For any S and the resulting $\hat{\boldsymbol{x}}^S$, the definition of Equation (2.5) results in the fact that $\text{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^S)$ is identical to $(\boldsymbol{x}_S - \boldsymbol{\mu}_S)'\boldsymbol{\Omega}_S(\boldsymbol{x}_S - \boldsymbol{\mu}_S)$, the squared Mahalanobis distance of $\boldsymbol{x}_S = (x_j)_{j \in S}$, where $\boldsymbol{x}_S$ and $\boldsymbol{\mu}_S$ only consist of the coordinates of $\boldsymbol{x}$ and $\boldsymbol{\mu}$ contained in the set S, respectively, and $\boldsymbol{\Omega}_S$ is the submatrix of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ with rows and columns included in S. Therefore, analyzing the outlyingness of $\hat{\boldsymbol{x}}^S$ using the squared Mahalanobis distance is equivalent to an analysis of the outlyingness

of the lower dimensional version $\boldsymbol{x}_S$, see also Equation (2.2).

2.2.3 Shapley Interaction Index

The analysis of interactions between players is often of interest in cooperative game theory, and one of the first proposals for such an analysis is due to Owen (1972). In more recent developments Grabisch and Roubens (1999); Fujimoto et al. (2006) introduced a framework that allows for an axiomatic generalization of the Shapley value to the so-called Shapley interaction index, which is also used in the field of Explainable AI (Lundberg et al., 2018). Applying this method within our framework, we can investigate pairwise outlyingness contributions of variables.

Using the notation of cooperative game theory, as in Section 2.2.1, the Shapley interaction index for S, with fixed $|S| = s$, is given by

$$I_{Sh}(v, S) = \sum_{T \subseteq P \setminus S} \frac{t!(p-t-s)!}{(p-s+1)!} \Delta_S v(T), \tag{2.11}$$

with $t = |T|$, and $\Delta_S v(T) = \sum_{L \subseteq S} (-1)^{s-l} v(T \cup L), l = |L|$, also known as the *discrete* or *set function derivative* (Grabisch, 2016). We refer to the previously mentioned articles of Grabisch and Roubens (1999); Fujimoto et al. (2006) for more details regarding the theory and properties connected to this concept.

As before, we decompose the squared Mahalanobis distance using the characteristic function defined in Equation (2.4). Moreover, we only focus on the *pairwise* Shapley interaction index ($|S| = 2$) because higher order Shapley interaction indices ($|S| \geq 3$) turn out to be zero in this setting (see A.2 for a proof).

Theorem 2.2.3.1. *Given two vectors $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^p$ and a non-singular matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, the pairwise contributions of the variable pair (j, k) of an observation $\boldsymbol{x}$ to the squared Mahalanobis distance $\mathrm{MD}^2(\boldsymbol{x})$, based on the Shapley interaction index as defined in Equation (2.11), are collected in the matrix $\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the off-diagonal elements are given by*

$$\Phi_{jk}(\boldsymbol{x}) := \sum_{T \subseteq P \setminus \{j,k\}} \frac{t!(p-t-2)!}{(p-1)!} \Delta_{\{j,k\}} \mathrm{MD}^2(\hat{\boldsymbol{x}}^T) \tag{2.12}$$

$$= 2(x_j - \mu_j)(x_k - \mu_k)\omega_{jk}, \tag{2.13}$$

with

$$\Delta_{\{j,k\}} \mathrm{MD}^2(\hat{\boldsymbol{x}}^T) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T \cup \{j,k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T \cup \{j\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T \cup \{k\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^T). \tag{2.14}$$

The diagonal elements are defined as

$$\Phi_{jj}(\boldsymbol{x}) := \phi_j(\boldsymbol{x}) - \sum_{k \neq j} \Phi_{jk}(\boldsymbol{x}) \tag{2.15}$$

$$= (x_j - \mu_j)^2 \omega_{jj} - (x_j - \mu_j) \sum_{k \neq j} (x_k - \mu_k)\omega_{jk}, \tag{2.16}$$

where $\phi_j(\boldsymbol{x})$ is the j-th coordinate of the Shapley value as in Theorem 2.2.2.1.

Proof. The proof of this theorem is given in A.2. □

To gain a better understanding of what the Shapley interaction index measures, we start by rewriting Equation (2.14) as

$$\begin{aligned}\Delta_{\{j,k\}} \mathrm{MD}^2(\hat{\boldsymbol{x}}^T) = (&\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^T)) \\ &- (\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^T)) \\ &- (\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^T)).\end{aligned}$$

This reveals that $\Delta_{\{j,k\}} \mathrm{MD}^2(\hat{\boldsymbol{x}}^T)$ measures the difference in squared Mahalanobis distance between simultaneous, pairwise and individual, marginal replacement of the variables x_j and x_k with their means μ_j and μ_k, respectively. The Shapley interaction index $\Phi_{jk}(\boldsymbol{x})$ then aggregates the pairwise differences $\Delta_{\{j,k\}} \mathrm{MD}^2(\hat{\boldsymbol{x}}^T)$ across all 2^{p-2} subsets $T \subseteq P \setminus \{j,k\}$ and measures the average effect of a pairwise versus a marginal replacement. Theorem 2.2.3.1 shows that the Shapley interaction index can be simplified, such that $\Phi_{jk}(\boldsymbol{x})$ only depends on the deviation of the j-th and the k-th coordinate from their mean, weighted by the corresponding entry of the precision matrix. This discloses that the Shapley interaction index $\Phi_{jk}(\boldsymbol{x})$ isolates the outlyingness contribution of the variable pair (j,k), while the Shapley value $\phi_j(\boldsymbol{x})$ accounts for all marginal contributions in which the j-th variable is involved. Figure 2.2 provides an illustration of the Shapley interaction index between the first and the second variable in a simple two-dimensional example.

The definition of the diagonal elements $\Phi_{jj}(\boldsymbol{x})$ is chosen such that a generalization of the *Efficiency* property given in Equation (2.10) is possible:

$$\phi_j(\boldsymbol{x}) = \sum_{k=1}^{p} \Phi_{jk}(\boldsymbol{x}) \quad \text{and} \quad \mathrm{MD}^2(\boldsymbol{x}) = \sum_{j=1}^{p}\sum_{k=1}^{p} \Phi_{jk}(\boldsymbol{x}).$$

Thus, the Shapley values for every variable can be decomposed into pairwise interactions with the remaining variables. Since the covariance matrix only contains information about the pairwise, linear relationship between variables, it is quite intuitive that no further decomposition is possible.

It is worth mentioning that there are other suggestions on how to generalize the Shapley value such that an explicit definition of $\Phi_{jj}(\boldsymbol{x}), j = 1,\ldots,p$, is not necessary (e.g. Sundararajan et al., 2020, *Shapley-Taylor interaction index*).

2.3 Cellwise Robust Outlier Explanation

Cellwise outlier detection focuses on identifying unusual *cells* rather than rows in a data matrix. Such a procedure is particularly justified when dealing with datasets containing many variables: If only individual cells of an observation are contaminated, then the majority of non-contaminated cells still contains valuable information that should not be discarded. Moreover, already a small proportion of outlying cells spread out over the whole data matrix could, in a rowwise treatment, soon lead to a setting where the majority of observations would have to be considered as traditional rowwise outliers. However, rowwise robust methods can only deal with settings where at least half of the observations are not corrupted. To

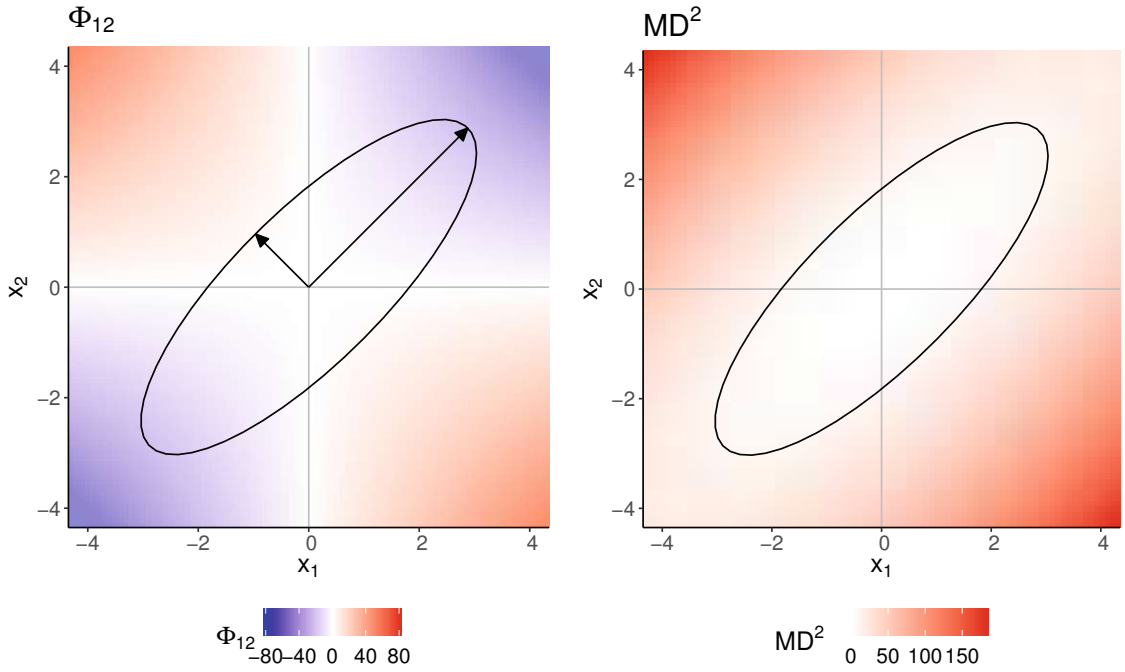


Figure 2.2: Using the same setup as for the example described in Figure 2.1, the pairwise contributions Φ_{12} of x_1 and x_2 to the squared Mahalanobis distance are visualized in the left panel. In this simple two-dimensional example, we can see that the pairwise contributions are highest in the direction of the eigenvector of $\boldsymbol{\Sigma}$ with the smallest eigenvalue. Hence, observations with a large multivariate outlyingness and a small univariate outlyingness are assigned high pairwise outlyingness scores Φ_{12}. In the right graph, we display the squared Mahalanobis distance, and both panels include the 99-percentile confidence ellipse.

deal with such settings, Alqallaf et al. (2009) formalized the cellwise contamination model. Several papers that build on this concept are referred to in Raymaekers and Rousseeuw (2021), and they also introduce a novel procedure for cellwise outlier identification.

As already outlined in Section 2.1, the key objective of this work concerns the explanation of multivariate outliers based on the Shapley value for given or appropriately estimated parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. To obtain cellwise robust covariance estimates, the 2SGS approach of Agostinelli et al. (2015a), the DDC method of Rousseeuw and Bossche (2018), or the cellMCD estimator of Raymaekers and Rousseeuw (2023) can be used. Since the Shapley value enables an additive decomposition of the squared Mahalanobis distance, it can be used to identify outlying cells. However, these contributions do not inform about the supposed cell values under the assumption that they were not contaminated. Apart from detecting outlying cells, estimating the values the cells were supposed to have is of major importance when handling cellwise outliers. In this section, we outline how to combine the ideas of cellwise outlier detection and multivariate outlier explanation to obtain *cellwise robust outlier explanations.*

2.3.1 SCD (Shapley Cell Detector) Algorithm

As a starting point, we take another look at the decomposition derived in Theorem 2.2.2.1, where we obtain the average marginal contributions of each component to the squared Mahalanobis distance. Equations (2.5) and (2.6) allow us to interpret said contributions in more detail: The value of $\phi_j(\boldsymbol{x})$ represents the average change in $\mathrm{MD}^2(\boldsymbol{x})$ across all 2^{p-1} possible variations of other variables, when the j-th component of $\boldsymbol{x}$ is replaced by its mean. Hence, positive values of $\phi_j(\boldsymbol{x})$ indicate that replacing x_j with μ_j would lead to an average reduction in $\mathrm{MD}^2(\boldsymbol{x})$, whereas negative values indicate that such a replacement would have the opposite effect.

The information provided by the Shapley value can now be used to design an algorithm for identifying outlying cells and replacing their values. We propose a stepwise procedure, which is described in detail in Algorithm 1. We call this method *Shapley Cell Detector*, abbreviated as SCD. The set R is updated in each step and will finally contain the indices of all cells of an observation $\boldsymbol{x}$ which are marked as outlying. The set of outlying coordinates R can be related to Equation (2.5), where $S = \bar{R} = P \setminus R$ denotes the cells which are not replaced. In the course of each iteration, we replace the coordinates of $\boldsymbol{x}$ that have the highest scores according to the Shapley value $\boldsymbol{\phi}(\boldsymbol{x})$ until the modified observation $\tilde{\boldsymbol{x}}$ is no longer a multivariate outlier. As is common in multivariate outlier detection, a cutoff based on the chi-square distribution is used, which entails the same distributional assumptions discussed in the introduction. The replaced value does not directly correspond to the mean but rather to a value towards the direction of the mean whereby the magnitude of the correction is controlled by a step size parameter $\delta \in (0, 1]$. This is done for each set R until a score resulting from the complement $\bar{R} := P \setminus R$ of R is larger than one obtained from the set R. Here, the r-dimensional subvector $\tilde{\boldsymbol{x}}_R = (\tilde{x}_j)_{j \in R}$ of the modified observation $\tilde{\boldsymbol{x}}$ consists of the replaced values, which are dependent on $\boldsymbol{\mu}_R = (\mu_j)_{j \in R}$. Note that the maximum in line 6 of Algorithm 1 is usually unique, implying that $k = 1$ and only one index is added to R per iteration.

Algorithm 1 Shapley Cell Detector (SCD)

1: **procedure** SCD($\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta$)
2: $\quad\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x}$
3: $\quad R \leftarrow \emptyset$
4: $\quad\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)' \leftarrow \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\phi_1(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \ldots, \phi_p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}))'$
5: $\quad$**while** $\mathrm{MD}^2(\tilde{\boldsymbol{x}}) > \chi^2_{p,0.99}$ **do**
6: $\quad\quad R \leftarrow R \cup \{j_1, \ldots, j_k\}$, where $(\phi_{j_l})_{l=1,\ldots,k} = \max_{i=1,\ldots,p} \phi_i$
7: $\quad\quad$**while** $\max_{j \in R} \phi_j > \max_{j \in \bar{R}} \phi_j$ **do**
8: $\quad\quad\quad\tilde{\boldsymbol{x}}_R \leftarrow \tilde{\boldsymbol{x}}_R - (\tilde{\boldsymbol{x}}_R - \boldsymbol{\mu}_R)\delta$
9: $\quad\quad\quad\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$
10: $\quad\quad$**end while**
11: $\quad$**end while**
12: $\quad$**return** $\tilde{\boldsymbol{x}}$
13: **end procedure**

Example 2.3.1.1. *We illustrate the working principle of Algorithm 1 by considering a 5-dimensional observation* $\boldsymbol{x} = (0, 1, 2, 2.2, 2.5)'$ *from a population with mean* $\boldsymbol{\mu} = (0, 0, 0, 0, 0)'$ *and covariance matrix* $\boldsymbol{\Sigma}$*, with elements* $\sigma_{jk} = 0.9, j \neq k$*, and* $\sigma_{jj} = 1$*. Here,* $\boldsymbol{x}$ *would be marked as a multivariate outlier since* $\mathrm{MD}^2(\boldsymbol{x}) = 44.90 > 15.09 = \chi^2_{5,0.99}$ *and we can employ the Shapely value of Theorem 2.2.2.1 to explain this multivariate outlier, resulting in* $\boldsymbol{\phi}(\boldsymbol{x}) = (0, -5.07, 9.87, 15.26, 24.84)'$*. Those outlyingness scores are then used in Algorithm 1 to flag outlying cells, and, for simplicity, we analyze the case where* $\delta = 1$*. In this scenario, the coordinate* x_5 *is identified first, followed by* x_4*, and then* x_3*. Each variable in turn is replaced by* μ_5, μ_4*, and* μ_3*, respectively. This results in an altered version* $\tilde{\boldsymbol{x}}$ *of the original observation* $\boldsymbol{x}$*, which is no longer outlying, and therefore the algorithm stops.*

It should be noted that in this example, we have no information about which cells are truly outlying or have been manipulated. However, in general, it seems desirable to keep the number of modified coordinates as small as possible.

Algorithm 1 is easy to implement and fast to compute. The discrepancy between the original and replaced cells indicates the outlyingness in the particular variables. However, this simplicity results from our definition of the Shapley value in Theorem 2.2.2.1, which leads to a replacement by a value towards the mean in Algorithm 1.

Figure 2.3 provides a further illustration of the SCD procedure for a two-dimensional example. It schematically displays five specific observations, denoted by `A` to `E`, to which Algorithm 1 is applied. The left plot shows the result when setting $\delta = 1$ in the algorithm, while the right plot corresponds to $\delta = 0.1$. The points in the plots highlight the individual computation steps of the algorithm, and the ellipse indicates the stopping criterion $\chi^2_{2,0.99}$. While for $\delta = 1$ the algorithm uses at most two steps, this behavior changes for the case of $\delta = 0.1$. Using a smaller step size leads to different replacement values for the points `B`, `D`, and `E`. Comparing the computation steps for points `B` and `D`, the results in the right plot seem more meaningful since they avoid increasing the Mahalanobis distance during the computation, and the final replacement is more similar to the original points.

Until now, we only considered a replacement of outlying cells by the mean or by a value towards the direction of the mean. However, the algorithm is only stopped by a sufficient reduction of the squared Mahalanobis distance. Therefore, the task at hand can thus be redefined further: Find the optimal replacements for outlying cells to achieve the highest possible reduction in squared Mahalanobis distance. As before, the Shapley value should determine the outlyingness of the cells.

2.3.2 MOE (Multivariate Outlier Explainer) Algorithm

Based on the definition of the Shapley value in Equation (2.7), a coordinate has a low outlyingness contribution if it is close to its mean. Consequently, it is unlikely that this cell is flagged as outlying. This center-outward ordering is induced by the squared Mahalanobis distance computed with respect to the mean, and thus it explains the *global outlyingness* of an observation. However, the described procedure might not be optimal for detecting cellwise outliers, where *local outlyingness* is emphasized, because the information contained in the regular cells of an observation could be incorporated to define an optimal replacement. For this purpose, an alternative approach to using the mean as the center for computing

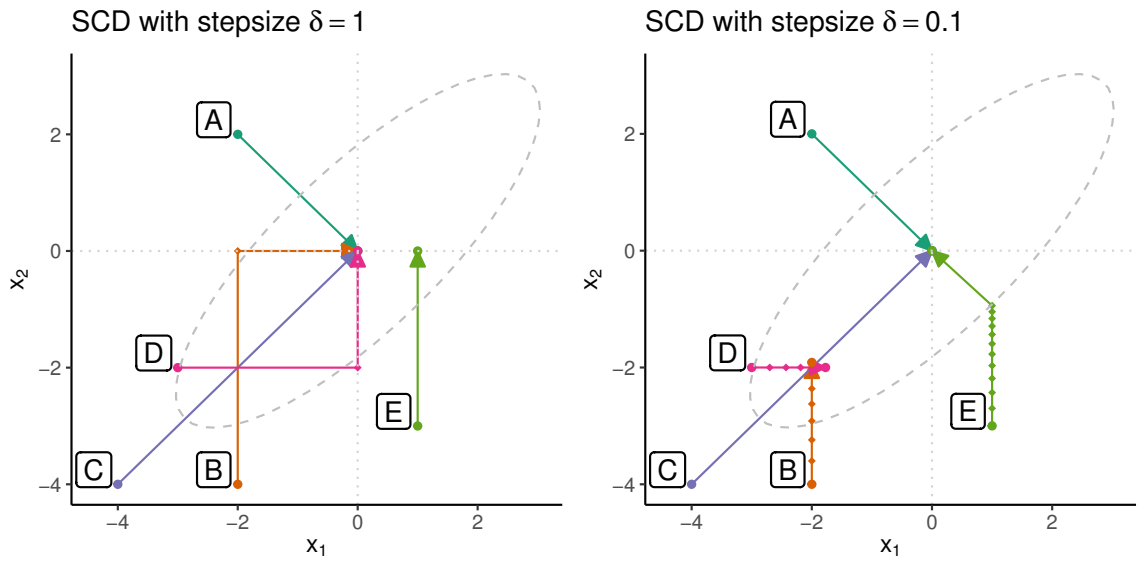


Figure 2.3: In this figure, two graphs are displayed to illustrate the operating principle of Algorithm 1 in a two-dimensional setting. Both plots show the position of the five outlying points `A` to `E` and their replacements. The plot on the left side shows the results when cells are directly replaced by their corresponding mean, while the right side illustrates the stepwise approach.

Mahalanobis distances and Shapley values is outlined in the following paragraphs. We call the newly proposed center parameter *reference point*. This new Shapley value will also be used later for an outlier replacement strategy.

The question of how to best replace cells of an observation to minimize the squared Mahalanobis distance has been addressed in Raymaekers and Rousseeuw (2021). Here we assume that the set R of outlying cells is fixed (and $R \neq \emptyset$) for an observation $\boldsymbol{x} = (x_1, \ldots, x_p)'$, and the cells x_j should be shifted to the values $\tilde{x}_j$, for $j \in R$. Explicitly we can write this as $\boldsymbol{x} - \boldsymbol{E}_R\boldsymbol{\beta}$ where $\boldsymbol{E}_R$ denotes the $p \times r$ matrix with the standard basis vectors $\boldsymbol{e}_j, j \in R$ as columns. The squared Mahalanobis distance of this expression can now be rewritten as follows,

$$\begin{aligned}
\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x} - \boldsymbol{E}_R\boldsymbol{\beta}) &= (\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{E}_R\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{E}_R\boldsymbol{\beta}) \\
&= \left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{E}_R\boldsymbol{\beta})\right\|_2^2 \\
&= \left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}) - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{E}_R\boldsymbol{\beta}\right\|_2^2.
\end{aligned}$$

Minimizing this expression corresponds to a least-squares problem, which leads to the least-squares estimator

$$\hat{\boldsymbol{\beta}}(R) = \underset{\boldsymbol{\beta}\in\mathbb{R}^s}{\arg\min}\, \mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x} - \boldsymbol{E}_R\boldsymbol{\beta}) = (\boldsymbol{E}_R'\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_R)^{-1}\boldsymbol{E}_R'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{2.17}$$

and the replaced values are given by $\tilde{\boldsymbol{x}}_R = \boldsymbol{x}_R - \hat{\boldsymbol{\beta}}(R)$, which are equal to the conditional means under multivariate normality, *i.e.* $\tilde{\boldsymbol{x}}_R = \mathbb{E}[\boldsymbol{x}_R|\boldsymbol{x}_{\bar{R}}]$ (Raymaekers and Rousseeuw, 2021).

If R consists of only one element, say $R = \{j\}$, for $j \in P$, then the solution of Equation (2.17) simplifies to

$$\hat{\beta}(j) = \frac{1}{\omega_{jj}}(\omega_{j1}, \ldots, \omega_{jp})(\boldsymbol{x} - \boldsymbol{\mu}), \tag{2.18}$$

where ω_{ij} denotes the element (i, j) of $\boldsymbol{\Sigma}^{-1}$, and the modification for observation $\boldsymbol{x}$ is given by $\tilde{x}_j = x_j - \hat{\beta}(j)$.

Building on those findings, we can now define the new reference point $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$ for a fixed set of outlying cells R, by setting each coordinate to

$$\tilde{\mu}_j(\boldsymbol{x}, R) = x_j - \hat{\beta}_{(j)}(R \cup \{j\}), \tag{2.19}$$

where $\hat{\beta}_{(j)}(R \cup \{j\})$ is the component of $\hat{\boldsymbol{\beta}}(R \cup \{j\})$ corresponding to the index j. To determine the set R, we adapt the SCD procedure by incorporating $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$ as a reference point for the Mahalanobis distance and updating it in each iteration. We refer to this procedure as Multivariate Outlier Explainer (MOE) and outline its general workflow in Algorithm 2.

The MOE procedure is initialized by computing the reference point $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$, with $R = \emptyset$. For the initial computation of $\hat{\boldsymbol{\beta}}$ we can simply apply Equation (2.18) to each coordinate of $\boldsymbol{x}$, which can be done in one step by matrix multiplication. Using this initial reference point, we obtain the squared Mahalanobis distance $\mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\boldsymbol{\Sigma}}(\tilde{\boldsymbol{x}})$, which is in turn used to define the corresponding Shapley value $\boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ according to Equation (2.9). We want to emphasize that the properties of the Shapley value listed in Section 2.2 remain unchanged, particularly the *Efficiency* property: The sum of the coordinates of the Shapley value $\boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ equals the squared Mahalanobis distance with respect to the new reference point $\tilde{\boldsymbol{\mu}}$. Outlying cells are then identified based on the Shapley value and corrected in the direction of their corresponding entries of $\tilde{\boldsymbol{\mu}}$, resulting in the modified observation $\tilde{\boldsymbol{x}}$. The process of updating the reference point $\tilde{\boldsymbol{\mu}}$, identifying outlying cells based on their Shapley values, and correcting them in the direction of $\tilde{\boldsymbol{\mu}}$, is then repeated until the vector $\tilde{\boldsymbol{x}}$ is no longer marked as outlying. Aside from using the reference point $\tilde{\boldsymbol{\mu}}$ in the MOE procedure instead of $\boldsymbol{\mu}$, the concept of the algorithm is similar to the SCD procedure, but there are two other important distinctions:

- The outlier cutoff value used in line 7 is adapted to the new reference point. Filzmoser et al. (2014) have shown that for a sample $\boldsymbol{x}$ drawn from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the conditional distribution of the squared Mahalanobis distance $\mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\boldsymbol{\Sigma}}(\boldsymbol{x})$ given $\tilde{\boldsymbol{\mu}}$ is a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda = \mathrm{MD}^2(\tilde{\boldsymbol{\mu}})$, denoted as $\chi^2_p(\lambda)$. Therefore, the 0.99 quantile of this distribution is taken as the cutoff value to exit the loop.

- Since the goal of this procedure is cellwise outlier detection, we want to avoid flagging coordinates that were only shifted by a negligible amount. Therefore, we monitor

the distance $\boldsymbol{d}$ by which each cell of $\boldsymbol{x}$ is shifted in the direction of $\tilde{\boldsymbol{\mu}}$. Initially, this distance is set to $d_j = 0, j = 1, \ldots, p$, followed by an iterative update of the distance variable in line 11. Moreover, we adjust $\boldsymbol{d}$ such that the distances are independent of the scale of the single coordinates. We then update the set of outlying coordinates R by only choosing coordinates for which $d_j > \eta \max_{l=1,\ldots,p} d_l$, with $\eta \in [0, 1]$. In simulations not included in this work, $\eta = 0.2$ resulted in a good trade-off between the recall, meaning the fraction of correctly identified cells among all contaminated cells, and the precision, meaning the fraction of correctly identified cells among all detected cells, of the procedure and is therefore chosen as a default value. Finally, we amend $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$, $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, and $\tilde{\boldsymbol{x}}$ according to the updated set R.

Algorithm 2 Multivariate Outlier Explainer (MOE)

1: **procedure** MOE($\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta, \eta$)
2: $\quad\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x}$
3: $\quad R \leftarrow \emptyset$
4: $\quad\boldsymbol{d} = (d_1, \ldots, d_p)' \leftarrow (0, \ldots, 0)'$
5: $\quad\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_p)' \leftarrow \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R) = (x_1 - \hat{\beta}(1), \ldots, x_p - \hat{\beta}(p))'$
6: $\quad\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)' \leftarrow \boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = (\phi_1(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}), \ldots, \phi_p(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}))'$
7: $\quad$**while** $\mathrm{MD}^2_{\tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}}(\tilde{\boldsymbol{x}}) > \chi^2_{p,0.99}(\mathrm{MD}^2(\tilde{\boldsymbol{\mu}}))$ **do**
8: $\quad\quad R \leftarrow R \cup \{j_1, \ldots, j_k\}$, where $(\phi_{j_l})_{l=1,\ldots,k} = \max_{i=1,\ldots,p} \phi_i$
9: $\quad\quad$**while** $\max_{j \in R} \phi_j > \max_{j \in \bar{R}} \phi_j$ **do**
10: $\quad\quad\quad\boldsymbol{c} \leftarrow (\tilde{\boldsymbol{x}}_R - \tilde{\boldsymbol{\mu}}_R)\delta$
11: $\quad\quad\quad\boldsymbol{d}_R \leftarrow \boldsymbol{d}_R + \boldsymbol{c}$
12: $\quad\quad\quad\tilde{\boldsymbol{x}}_R \leftarrow \tilde{\boldsymbol{x}}_R - \boldsymbol{c}$
13: $\quad\quad\quad\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$
14: $\quad\quad$**end while**
15: $\quad\quad\tilde{\boldsymbol{\mu}} \leftarrow \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$
16: $\quad$**end while**
17: $\quad\boldsymbol{d} = (d_1, \ldots, d_p)' \leftarrow (d_1/\sqrt{\sigma_{11}}, \ldots, d_p/\sqrt{\sigma_{pp}})'$
18: $\quad R \leftarrow \{j_1, \ldots, j_m\}$, for which $(d_{j_l})_{l=1,\ldots,m} > \eta \max_{i=1,\ldots,p} d_i$
19: $\quad\tilde{\boldsymbol{\mu}} \leftarrow \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$
20: $\quad\boldsymbol{\phi} \leftarrow \boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$
21: $\quad\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x}$
22: $\quad\tilde{\boldsymbol{x}}_R \leftarrow \tilde{\boldsymbol{\mu}}_R$
23: $\quad$**return** $\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\phi}$
24: **end procedure**

Algorithm 2 allows us to detect and impute cellwise outliers, and it also yields a local explanation of the outlyingness. Furthermore, the Shapley values computed with respect to the reference point $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$ can be used to explain the results of other cellwise outlier detection procedures. To this end, we merely need to compute $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, R)$ for a given set of outlying cells R of an observation $\boldsymbol{x}$. By subsequently determining the Shapley value $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, we can therefore explain *why* the observation is outlying.

Example 2.3.2.1. *We reiterate Example 2.3.1.1 with the MOE procedure, using a step size of* δ *of* 0.1. *The first two coordinates* x_1 *and* x_2 *are marked as outlying, resulting in* $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, \{1, 2\}) = (2.19, 2.19, 2.27, 2.13, 2.04)$ *and* $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = (34.89, 7.07, -0.86, 1.28, 4.88)$.

Comparing the results of Algorithms 1 and 2, it can be seen that the sets of outlying cells for the two algorithms are disjoint. Therefore, the interpretations of the results are different. The reason for this discrepancy is mainly that we no longer decompose $\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x})$, *but instead the squared Mahalanobis distance of the amended reference point* $\tilde{\boldsymbol{\mu}}$, $\mathrm{MD}^2_{\tilde{\boldsymbol{\mu}},\boldsymbol{\Sigma}}(\boldsymbol{x})$. *While the Shapley value* $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *used in Algorithm 1 explains the global outlyingness, the Shapley value* $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ *used in Algorithm 2 provides us with a local understanding of the outlyingness, which is better suited to the setting of cellwise outlyingness.*

In Figure 2.4, we compare the final Shapley values yielded by the SCD and MOE algorithms. In Figure 2.5, we show the Shapley values computed during each iteration for both algorithms, using a step size $\delta = 0.1$. *Both figures indicate the squared Mahalanobis distance (black bar) and the corresponding (non-)central chi-square quantile (dotted line).*

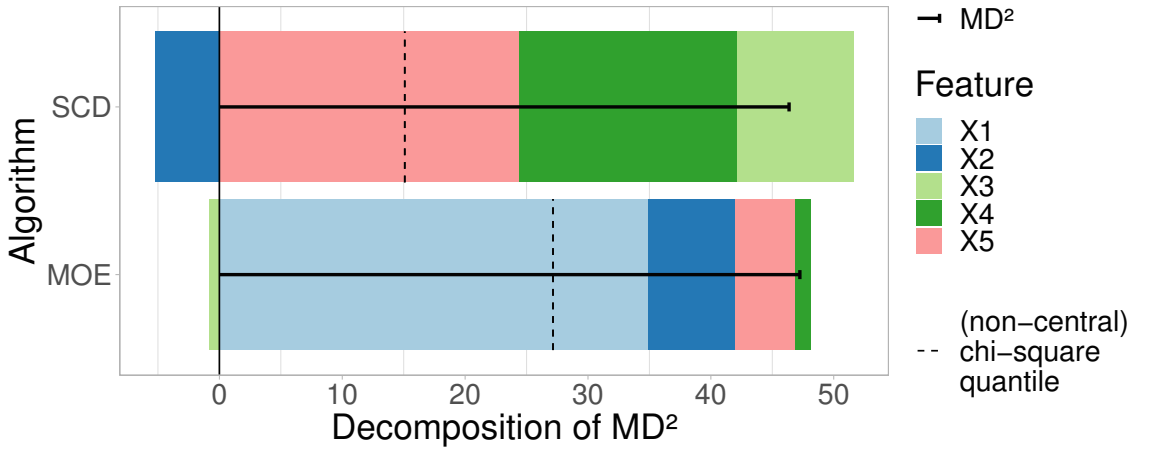


Figure 2.4: Comparison of the Shapley values $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ used in Algorithm 1 to explain the *global* outlyingness, and $\boldsymbol{\phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, used in Algorithm 2 to gain *local* insights on the outlyingness, with the input values defined in Example 2.3.1.1. The SCD procedure identifies the three coordinates x_3, x_4, and x_5, which are furthest from the mean $\boldsymbol{\mu}$. On the other hand, the MOE algorithm uses the alternative reference point $\tilde{\boldsymbol{\mu}}$ to identify variables x_1 and x_2.

2.4 Simulations

The simple numerical example from the previous section has illustrated that the SCD and MOE algorithms can lead to quite different outcomes. However, it needs to be emphasized that their purposes also differ: While the SCD procedure aims at global outlier explanation, i.e. with respect to the distribution of the entire dataset, the MOE procedure is locally applicable and builds on the local information contained in the regular cells of an individual observation. Nevertheless, it can be interesting to compare both procedures in terms of

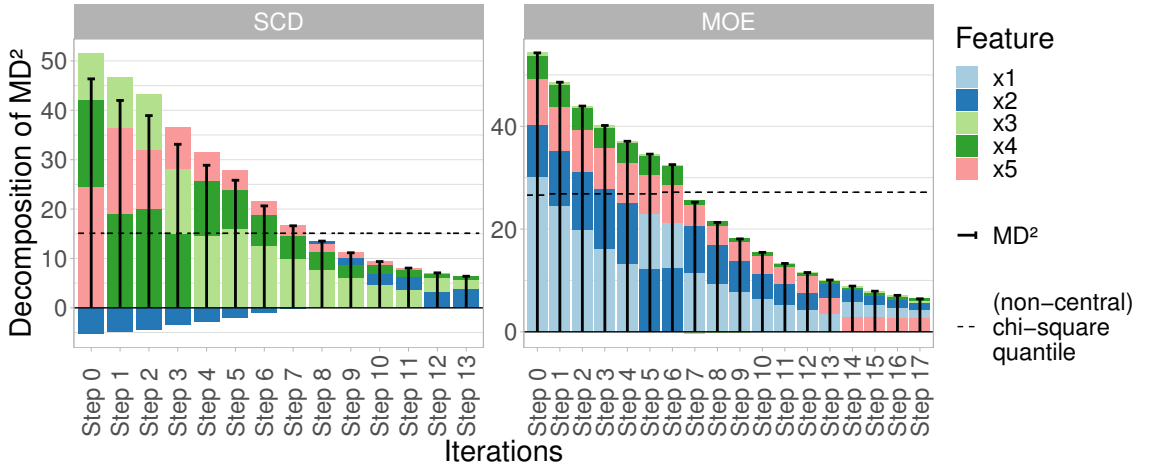


Figure 2.5: Comparison between the Shapley values calculated for each iteration of Algorithm 1 (left) and Algorithm 2 (right), respectively, for Example 2.3.1.1. While the outlyingness is monotonically decreasing in both cases, the sets of identified variables are disjoint. Both the SCD and MOE procedures reduce the outlyingness by iteratively shifting the identified variables toward the corresponding coordinates of $\boldsymbol{\mu}$ or $\tilde{\boldsymbol{\mu}}$, respectively.

their ability to identify cellwise outliers and, in particular, to examine their performance in comparison to a reference method, namely the cellHandler procedure introduced by Raymaekers and Rousseeuw (2021). We choose standard parameters for all three procedures, meaning that both the SCD and MOE algorithms are set up with a step size of $\delta = 0.1$, and the MOE procedure additionally uses a detection threshold $\eta = 0.2$.

In our analysis, we compare two different mechanisms for generating outliers and analyze the effects of various parameter configurations, which are summarized in Table 2.1 and described in more detail in the following paragraphs. For each specific parameter combination, we repeat the simulations 50 times and compute averages of the resulting measures Recall, Precision, and F-Score.

For both outlier generation procedures, we generate data matrices with p columns and $n = 20p$ rows from multivariate normal distributions with mean $\boldsymbol{\mu} = \mathbf{0}$ and three different types of covariance matrices $\boldsymbol{\Sigma}$, namely $\boldsymbol{C}_{\text{mod}}$, $\boldsymbol{C}_{\text{mix}}$, and $\boldsymbol{C}_{\text{low}}$. In all three cases, the diagonal elements are set to 1. For $\boldsymbol{C}_{\text{mod}}$, the off-diagonal elements are chosen as 0.5, resulting in moderate correlations. The off-diagonal elements of $\boldsymbol{C}_{\text{mix}}$ correspond to $(-0.9)^{|j-k|}, j \neq k$, yielding both high and low correlations. For $\boldsymbol{C}_{\text{low}}$, the off-diagonal elements are randomly generated as described in Agostinelli et al. (2015a), generally resulting in low correlations.

To analyze the effect of highly correlated shift-outliers, we randomly select $\lceil n\epsilon_2 \rceil$ rows, and for each of those rows, we replace $r = \lceil p\epsilon_1 \rceil$ randomly selected cells by r-variate outliers. Those follow a Gaussian distribution with mean $\boldsymbol{\mu} = (\gamma, \ldots, \gamma)'$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}$, with elements $\tilde{\sigma}_{jk} = 0.7, j \neq k$, and $\tilde{\sigma}_{jj} = 1$. The magnitude of the outliers is determined by the value γ, which is selected according to Table 2.1. Following this approach, the fraction of outlying cells ranges between 0.01 and 0.16.

Table 2.1: Summary of the parameters used for the two simulation scenarios on cellwise outlier detection discussed in Section 2.4.

Parameters	Shift outliers	Structured outliers
Dimension, p	$5, 10, 20, 30, 40$	$5, 10, 20, 30, 40$
Covariance, $\boldsymbol{\Sigma}$	$\boldsymbol{C}_{\text{mix}}$, $\boldsymbol{C}_{\text{low}}$, $\boldsymbol{C}_{\text{mod}}$	$\boldsymbol{C}_{\text{mix}}$, $\boldsymbol{C}_{\text{low}}$, $\boldsymbol{C}_{\text{mod}}$
Fraction of outlying columns, ϵ_1	$0.1, 0.2, 0.3, 0.4$	-
Fraction of outlying rows, ϵ_2	$0.1, 0.2, 0.3, 0.4$	-
Fraction of outlying cells, ϵ_3	-	$0.1, 0.2, 0.3, 0.4$
Magnitude of outlyingness, γ	$1, 2, 3$	$2, 3, 4, 5, 6$
Total combinations	720	300

For the second scenario, outliers are generated such that they are structurally outlying but have low univariate outlyingness, as proposed by Raymaekers and Rousseeuw (2021). For this purpose, $n\epsilon_3$ cells are selected randomly in each column. Like this, each row contains a subset $K \subseteq P$ of cells $\boldsymbol{x}_K$ which are subsequently replaced by the vector $\gamma\sqrt{k}\boldsymbol{u}'/\operatorname{MD}_{\boldsymbol{\mu}_K,\boldsymbol{\Sigma}_K}(\boldsymbol{u})$, where $k = |K|$, and $\boldsymbol{u}$ is the eigenvector of $\boldsymbol{\Sigma}_K$ that corresponds to the smallest eigenvalue.

We summarize the overall results in Table 2.2, comparing Precision, Recall, and F-Score. The performance metrics are averaged over all parameters not listed in the table (p, ϵ_1, ϵ_2, ϵ_3, and γ) and all replications. Regarding Precision, the MOE algorithm exhibits the best results in 4 out of 6 settings. Concerning Recall, the SCD procedure performs best when the correlations are low to moderate, while the cellHandler procedure performs best when the correlations are moderate or mixed. Finally, when comparing the F-Score, we see that each algorithm outperforms the remaining two at least once. However, the results listed in the table are averaged over a wide range of parameter settings. Therefore, we study the individual effects of the different parameters in more detail in the following.

In Figure 2.6, we analyze the effect of the dimension p on the cellwise outlier detection performance. We focus on the case of highly correlated shift outliers, with fixed $\epsilon_1 = \epsilon_2 = 0.4$ and $\gamma = 3$. This results in a situation with many moderately contaminated cells. We observe an increase in Precision for all three algorithms and covariance structures as p increases. The SCD procedure shows the most substantial increase and the highest overall Precision in case of low and moderate correlations. For the mixed correlations, the MOE procedure exhibits the highest Precision. Moving on to Recall, we see an initial increase followed by a very slight decline for all three methods for mixed and moderate correlations. For low correlations, the MOE procedure shows a severe drop in Recall, while the other two procedures only show a slight decline in performance. This is related to the default choice of the tuning parameter $\eta = 0.2$ for the MOE procedure, and the Recall could be improved by choosing a smaller value of η. While the Recall is similar for all methods in case of moderate and mixed correlations, we observe that the SCD procedure has the highest Recall in case of low correlations, followed by the cellHandler algorithm.

For the structured outliers, we illustrate the influence of γ for fixed $\epsilon_3 = 0.4$ and $p = 30$ in Figure 2.7. As expected, Precision and Recall are increasing as the magnitude of outlyingness,

Table 2.2: Summary of the results of the simulations described in Section 2.4. The performance metrics Precision, Recall, and F-Score listed in this table are averaged over all replications and parameter combinations.

		Shift outliers			Structured outliers		
$\boldsymbol{\Sigma}$	Algorithm	Precision	Recall	F-Score	Precision	Recall	F-Score
$\boldsymbol{C}_{mix}$	SCD	0.690	0.737	0.708	0.546	0.551	0.540
$\boldsymbol{C}_{mix}$	MOE	0.894	0.707	**0.782**	0.916	0.545	**0.668**
$\boldsymbol{C}_{mix}$	cellHandler	0.760	0.743	0.741	0.854	0.564	0.667
$\boldsymbol{C}_{low}$	SCD	0.713	0.510	**0.574**	0.767	0.715	**0.729**
$\boldsymbol{C}_{low}$	MOE	0.678	0.396	0.478	0.880	0.597	0.695
$\boldsymbol{C}_{low}$	cellHandler	0.599	0.473	0.508	0.900	0.630	0.722
$\boldsymbol{C}_{mod}$	SCD	0.767	0.405	0.507	0.859	0.530	0.627
$\boldsymbol{C}_{mod}$	MOE	0.808	0.421	**0.528**	0.954	0.476	0.599
$\boldsymbol{C}_{mod}$	cellHandler	0.649	0.471	0.522	0.917	0.513	**0.634**

controlled by γ, increases. The MOE procedure shows the highest overall Precision. However, regarding Recall, the SCD procedure performs better for mixed and high correlations. For low correlations, the cellHandler procedure exhibits the steepest increase in Recall as γ increases.

In conclusion, these simulations show that our approaches based on the Shapley value, particularly the MOE procedure, yield comparable results to one of the current state-of-the-art methods, namely the cellHandler procedure. While cellwise outlier detection presents the focus of the latter method, our approach is instead based on utilizing cellwise outlier detection specifically to enhance and robustify the outlyingness scores based on Theorem 2.2.2.1, with respect to an observation's "expected" position, as outlined in Equations (2.17) and (2.19).

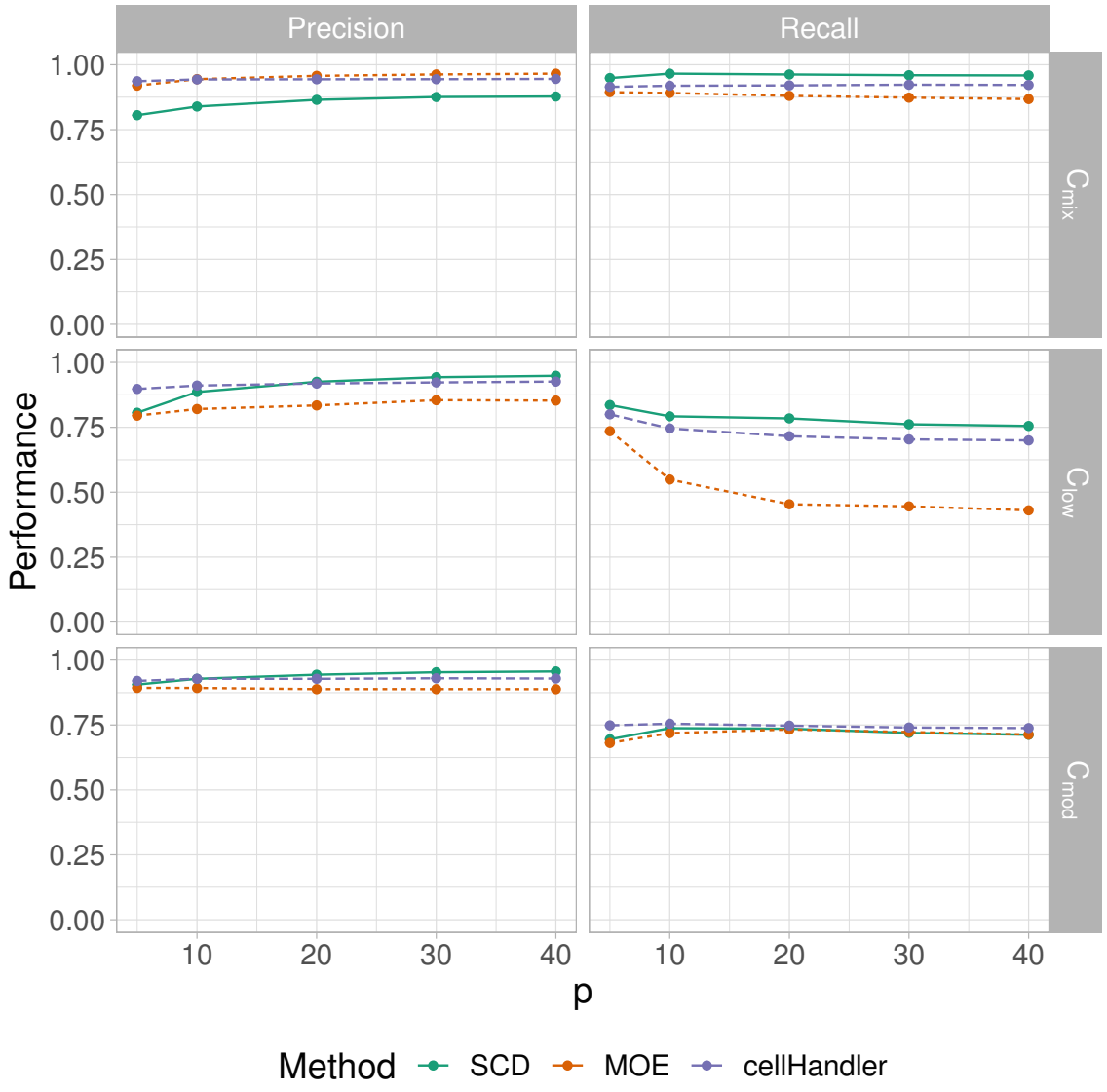


Figure 2.6: Comparison between the SCD, MOE, cellHandler procedures in the simulation setting of cellwise shift outliers outlined in Section 2.4, with simulation parameters $\epsilon_1 = \epsilon_2 = 0.4$ and $\gamma = 3$. The performance scores Precision (left) and Recall (right) of the individual algorithms are listed separately for each type of covariance structure.

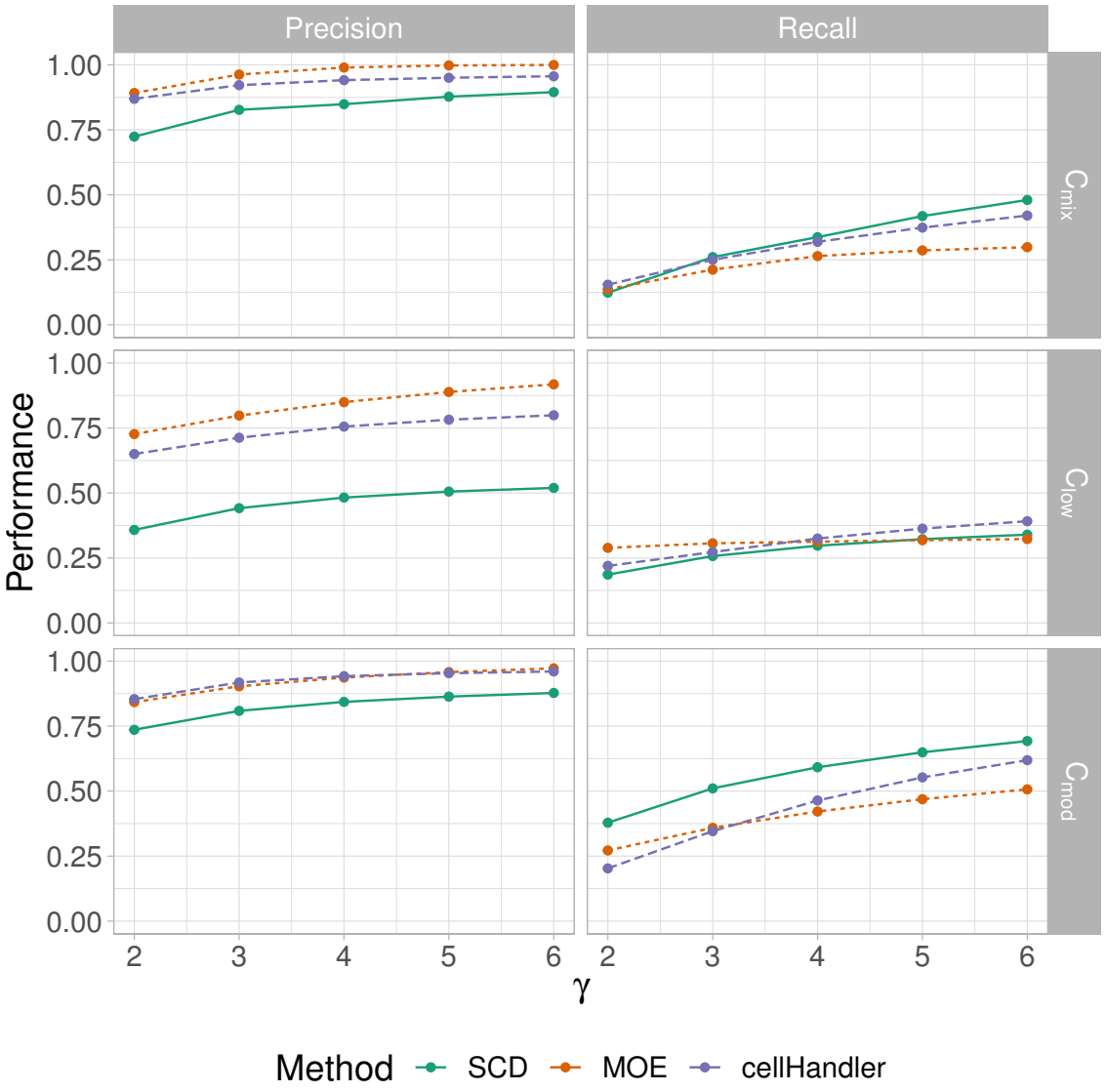


Figure 2.7: Comparison between the SCD, MOE, cellHandler procedures in the simulation setting of structured cellwise outliers outlined in Section 2.4, with simulation parameters $\epsilon_3 = 0.4$ and $p = 30$. The performance scores Precision (left) and Recall (right) of the individual algorithms are listed separately for each type of covariance structure.

2.5 Applications

While the simulations shown in the previous section have demonstrated the performance of the methods and algorithms introduced in Section 2.2 and 2.3 on simulated datasets, we now apply them to two real-world data. To this end, we analyze the *Top Gear* dataset from Alfons (2021) and the *Weather in Vienna* dataset from Stadt Wien (2022).

2.5.1 Top Gear

The *Top Gear* dataset comprises measurements of 11 numerical attributes (see Figure 2.8) of 245 complete data instances of cars featured on the website of the BBC television series. We apply a logarithmic transformation to five variables for data preprocessing to obtain more symmetrical marginal distributions. Additionally, each column is robustly centered and scaled based on the median and the MAD. Furthermore, we estimate the covariance using the MCD estimator before applying the SCD, MOE, and cellHandler procedures.

In the following, we use three different types of plots to analyze the results of all three tested algorithms on this dataset: Figure 2.8 summarizes the Shapley values, Figure 2.9 shows the outlying cells, and Figure 2.10 displays the Shapley interaction indices, respectively.

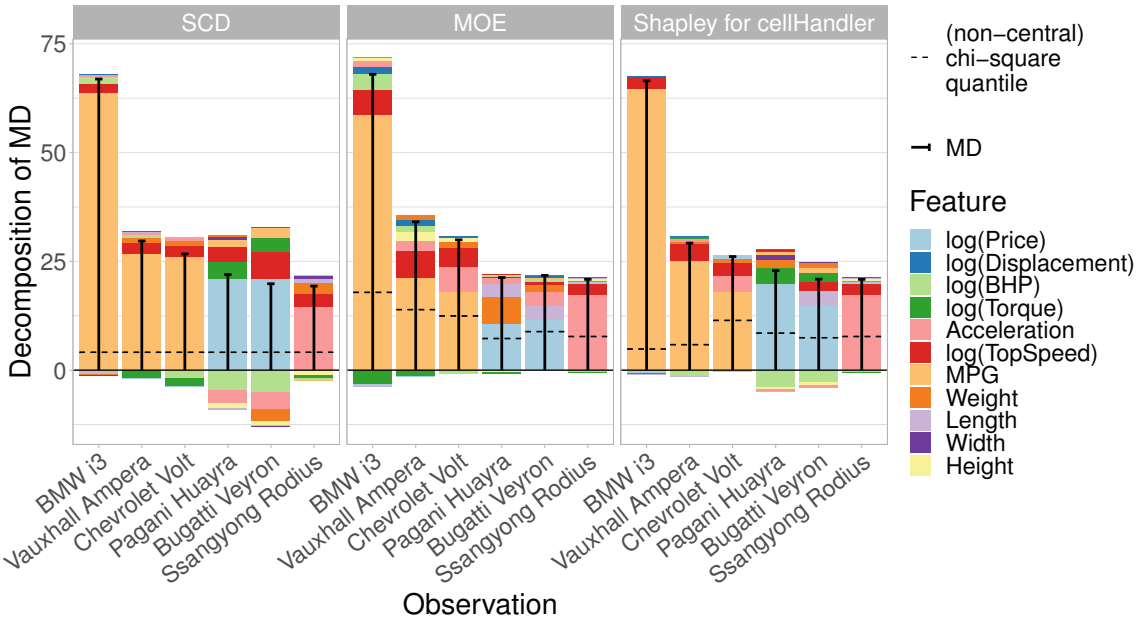


Figure 2.8: Comparison of the outlyingness scores resulting from the SCD (left), MOE (center), and cellHandler (right) procedures. Each graph shows a visualization of the Shapley values for the six most outlying observations.

In detail, Figure 2.8 consists of three graphs, each displaying the outlyingness decompositions according to the applied algorithm of the six cars with the highest Mahalanobis distance. In the left panel, we see the results generated using the SCD procedure, where we use the center of the data as a reference point. In the center panel, we show the results of

using the MOE algorithm with the non-central chi-square cutoff. For both procedures, we use a step size $\delta = 0.1$, and the MOE algorithm's detection threshold is $\eta = 0.2$. In the right panel, we show the results of using the cellHandler procedure to flag outlying cells, and then employing the Shapley value, with reference point $\tilde{\boldsymbol{\mu}}(\boldsymbol{x}, S)$ according to Equation (2.17), to enhance the interpretability of the results, as outlined in Section 2.3. Since we are analyzing multiple observations with large differences in squared Mahalanobis distance, plotting the squared distance is ineligible, and we display the square root instead. However, since we are decomposing the squared distance in Theorem 2.2.2.1, we must scale the outlyingness scores. For this reason, we derive each variable's proportional contribution to the squared distance and multiply it by the (not-squared) Mahalanobis distance. While this results in a somewhat distorted graph, this workflow enables us to analyze and compare multiple observations using a stacked bar chart.

Analyzing Figure 2.8, we first want to focus on the three cars with the highest outlyingness. For those cars, the main contribution to the squared Mahalanobis distance in the three graphs is caused by the variable `MPG`. Considering that these specific models are hybrid vehicles, it seems reasonable that their fuel consumption differs strongly from that of gasoline and diesel cars. All three methods lead to similar results in this case. For the two sports cars `Bugatti Veyron` and `Pagani Huayra`, we see that the `Price` variable is contributing the most to the outlyingness, which is again visible in the results of all three methods. For these two cars, most characteristics are similar to a certain extent, except for their weight: The Bugatti weighs 1990 kg while the Pagani has only a weight of 1350 kg. This fact becomes clearly visible when applying the MOE algorithm, where the `Weight` variable has a high contribution to the squared Mahalanobis distance of the Pagani but not for the Bugatti. Again, the three procedures agree for the `Ssangyong Rodius`, where `Acceleration` contributes the most. In fact, the listed value for `Acceleration` is 0, which is clearly an error in the published dataset itself.

In Figure 2.9, we show the results of applying the MOE procedure (top) to the TopGear dataset, as well as the Shapley values based on the cellHandler procedure (bottom). In these plots, the original values of the variables are displayed in each cell. White rectangles represent regular cells, while outlying cells are colored red or blue, depending on whether the cell's original value is higher (red) or lower (blue) than the replacement. The color intensity is given according to the Shapley values of the cells. The biggest differences between the MOE and the cellHandler algorithm can be seen between the two sports cars `Bugatti Veyron` and `Pagani Huayra`, where the cellHandler procedure results in many more outlying cells. However, it is surprising that the `Acceleration` parameter is not flagged since both cars have an exceptionally fast acceleration.

Finally, Figure 2.10 consists of heatmaps displaying the Shapley interaction indices and barplots showing the corresponding Shapley values for the `Chevrolet Volt` (left) and `Pagani Huayra` (right). The Shapley values and interaction indices are based on the reference point obtained from Algorithm 2. For the Chevrolet, we see a single outstanding index for `MPG`. On the other hand, the Pagani not only shows a high index for `Price` but also for the pairwise outlyingness score between `Weight` and `Price`, which indicates that for an expensive sports car, it is unexpectedly lightweight.

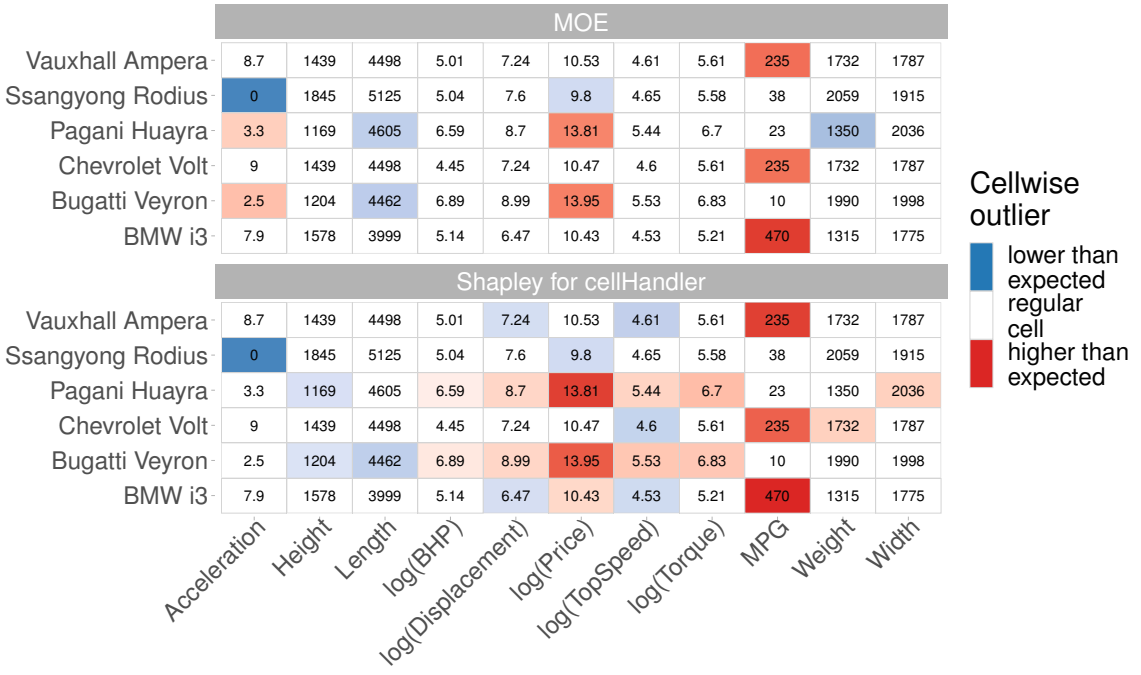


Figure 2.9: Outlying cells according to Algorithm 2 (top) and the cellHandler procedure (bottom). Each cell shows the original value from the dataset, color coding indicates whether those values were higher (red) or lower (blue) than the imputed values, and the color intensity is based on the magnitude of the Shapley value.

2.5.2 Weather in Vienna

As a second real-world example, we analyze monthly weather data from the weather station "Hohe Warte" in Vienna (Stadt Wien, 2022). Therefore we consider 16 numerical attributes, which are described in Table A.3 in A.3, over a time period spanning from 1955 to 2022. Furthermore, we restrict our investigation to the three summer months, June, July, and August, and compute average values for the considered variables, which yields 68 annual observations for each variable. As for the previous example, we center and scale the data using median and MAD and estimate the covariance using the MCD estimator before applying the SCD and MOE algorithms using the same setup as before.

Figure 2.11 displays the outlying cells of the entire 68 years of measurements: The top panel shows the results from the SCD algorithm, and the bottom panel displays those from the MOE algorithm. Both panels reveal that the number of detected anomalies has increased over the years. The SCD procedure further yields results that we would expect to find given that we are currently experiencing an anthropogenic climate change, such as an increasing number of hot days over the years or an increased minimum and maximum mean daily temperature. We emphasize that the SCD procedure results in a global outlyingness measure with respect to the overall mean. On the other hand, the MOE algorithm acts as a local measure: With given values of the regular cells in a particular year, the outlyingness in the remaining variables is determined.

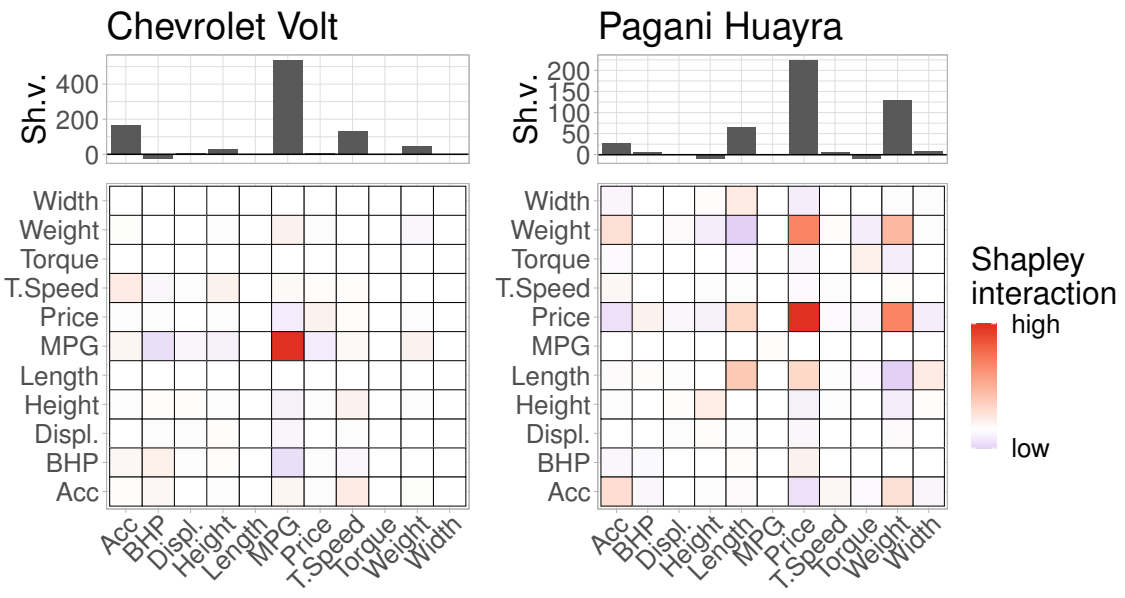


Figure 2.10: The two graphs in the lower portion of this figure show the Shapley interaction indices $\boldsymbol{\Phi}(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}(\boldsymbol{x}, S), \Sigma)$ for the `Chevrolet Volt` and `Pagani Huayra`, which are computed with respect to the reference point provided by Algorithm 2. The corresponding Shapley values are displayed above the heatmaps.

A more detailed analysis of the results can be made by comparing the Shapley values and pairwise outlyingness scores obtained from each procedure. Such an analysis is representatively carried out for the year 2021. The results are displayed in Figure 2.12, where we can observe a clear distinction between the results of the SCD and MOE procedures, respectively. Both algorithms detect anomalies in the average temperature minimum (`avg_t_min`) and total precipitation (`precp_sum`). However, using the local reference point enables the MOE procedure to detect outliers in the number of sun hours (`sun_h`) and the number of clear days (`num_clear`). According to the results of the local MOE procedure given in Figures 2.11 and 2.12, the weather of Vienna in 2021 was unusually hot, with more rain than we would expect. When considering the trend of increasing temperature over the years at this specific weather station, we would generally expect fewer sun hours and more clear days than observed in 2021.

2.6 Discussion and Conclusions

This paper introduced Shapley values in connection with Mahalanobis distances for multivariate outlier explanation. The Mahalanobis distance is commonly employed for multivariate outlier detection in statistics. Then again, the Shapley value is a concept that originated in cooperative game theory and recently gained popularity in the field of Explainable AI. There it is used to explain the predictions of complex machine learning models by providing information about the contributions of the individual features to a model's prediction. Combining

Figure 2.11: Comparison of outlying cells according to Algorithm 1 (top) and Algorithm 2 (bottom) for the weather data of Vienna. It is visible in the results of both procedures that the number of anomalies is increasing over the years.

the Shapley value with the squared Mahalanobis distance enables us to derive outlyingness scores for each coordinate of an observation. Those scores consider all 2^p possible combinations of p variables of a single instance and allow us to additively decompose the squared Mahalanobis distance into contributions originating from the individual variables. Without further simplification, the computation would entail evaluating the squared Mahalanobis distance for those 2^p combinations, which would pose a substantial computational challenge. However, we showed that our approach leads to a much simpler and computationally efficient form of the Shapley value. Moreover, the Shapley interaction indices generalize Shapley values and can be used to derive outlyingness scores for pairs of variables.

Outlier explanation, and thus identifying the contributions of a variable to the outlyingness of a particular observation, is closely related to cellwise outlyingness, where one aims to identify unusual cells instead of entire observations. We have adopted cellwise outlyingness into the framework of Shapley values and have proposed two procedures for simultaneous outlier detection and explanation. First, we introduced the SCD procedure as a straightforward implementation of Shapley values for cellwise outlier detection. This algorithm is iteratively replacing anomalous cells with a value towards their mean until the observation is no longer outlying. The more sophisticated MOE procedure takes the information of the non-outlying cells into account and determines a local reference point

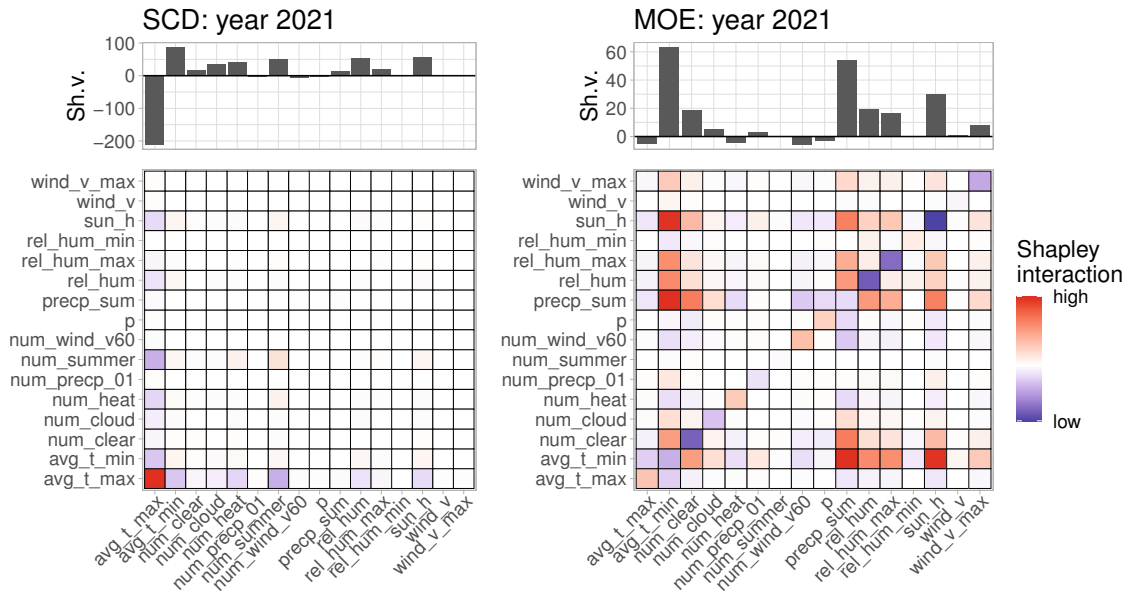


Figure 2.12: The two graphs in the lower panel show the Shapley interaction indices of the year 2021 for the SCD procedure (left) and the MOE procedure (right). The corresponding Shapley values are displayed above the heatmaps.

based on this added input. As a result, one again obtains an additive decomposition of the squared Mahalanobis distance, but with contributions that explain the *local* outlyingness of an observation.

The performance of the two cellwise outlier detection and explanation procedures has been evaluated in simulations and on real-world datasets. It has further been compared to the recently published cellHandler procedure. However, we want to emphasize that the goal of our work is clearly defined as outlier explanation rather than cellwise outlier detection. In particular, Mahalanobis distances rely on a robustly estimated covariance matrix, which has not been in focus in this paper.

We believe that Shapley values are a powerful tool for providing humanly interpretable explanations that allow us to gain further insights into the results of models and methods used in statistics and computer science. They show great potential for further use in this area, especially when a simplification of the computation is possible, as is the case when combining them with Mahalanobis distances. Possible extensions of Shapley values for outlier detection in functional data analysis will be the subject of our future research.

Software and data availability: The methods introduced in this work are available in the `R` package `ShapleyOutlier` on CRAN, including the weather dataset and a vignette to reproduce the examples presented in Section 2.5.

Appendix A

A.1 Proof of Theorem 2.2.2.1

Lemma A.1.1. *The contributions* $\Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S) = \operatorname{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{k\}}) - \operatorname{MD}^2(\hat{\boldsymbol{x}}^S)$ *can be expressed as*

$$\Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S) = 2(x_k - \mu_k)\left(\sum_{j\in S\cup\{k\}} (x_j - \mu_j)\omega_{jk}\right) - (x_k - \mu_k)^2\omega_{kk}, \tag{A.20}$$

for any subset $S \subseteq P \setminus \{k\}$.

Proof.

$$\begin{aligned}
\Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S) &= \operatorname{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{k\}}) - \operatorname{MD}^2(\hat{\boldsymbol{x}}^S) \\
&= (\hat{\boldsymbol{x}}^{S\cup\{k\}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{x}}^{S\cup\{k\}} - \boldsymbol{\mu}) - (\hat{\boldsymbol{x}}^S - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{x}}^S - \boldsymbol{\mu}) \\
&= \sum_{j=1}^{p}\sum_{l=1}^{p}(\hat{x}_j^{S\cup\{k\}} - \mu_j)(\hat{x}_l^{S\cup\{k\}} - \mu_l)\omega_{jl} - \sum_{j=1}^{p}\sum_{l=1}^{p}(\hat{x}_j^{S} - \mu_j)(\hat{x}_l^{S} - \mu_l)\omega_{jl} \\
&= \sum_{j\in S\cup\{k\}}\sum_{l\in S\cup\{k\}}(x_j - \mu_j)(x_l - \mu_l)\omega_{jl} - \sum_{j\in S}\sum_{l\in S}(x_j - \mu_j)(x_l - \mu_l)\omega_{jl} \\
&= \sum_{j\in S\cup\{k\}}(x_k - \mu_k)(x_j - \mu_j)\underbrace{\omega_{kj}}_{=\omega_{jk}} + \sum_{j\in S}(x_k - \mu_k)(x_j - \mu_j)\omega_{jk} \\
&= (x_k - \mu_k)^2\omega_{kk} + 2(x_k - \mu_k)\sum_{j\in S}(x_j - \mu_j)\omega_{jk} = \text{(A.20)}
\end{aligned}$$

□

Now that we have derived a simpler form for the contributions $\Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S)$, we can use this result to rewrite Equation (2.6) for the k-th component of the Shapley value $\phi_k(\boldsymbol{x})$. We apply Lemma A.1.1 in the first step of the proof below, and for a simpler notation, we write

$$w(|S|) := \frac{|S|!(p - |S| - 1)!}{p!},$$

for which $\sum_{S\subseteq P\setminus\{k\}} w(|S|) = 1$ holds.

Proof of Theorem 2.2.2.1.

$$
\begin{aligned}
\phi_k(\boldsymbol{x}) &= \sum_{S\subseteq P\setminus\{k\}} w(|S|)\Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S) \\
&= \sum_{S\subseteq P\setminus\{k\}} w(|S|)\Big((x_k-\mu_k)^2\omega_{kk} + 2(x_k-\mu_k)\sum_{j\in S}(x_j-\mu_j)\omega_{jk}\Big) \\
&= (x_k-\mu_k)^2\omega_{kk}\underbrace{\left(\sum_{S\subseteq P\setminus\{k\}} w(|S|)\right)}_{=1} + 2(x_k-\mu_k)\sum_{S\subseteq P\setminus\{k\}}\left(w(|S|)\sum_{j\in S}(x_j-\mu_j)\omega_{jk}\right) \\
&= (x_k-\mu_k)^2\omega_{kk} + 2(x_k-\mu_k)\sum_{s=1}^{p-1}\left(w(s)\sum_{\substack{S\subseteq P\setminus\{k\}\\|S|=s}}\sum_{j\in S}(x_j-\mu_j)\omega_{jk}\right) \\
&= (x_k-\mu_k)^2\omega_{kk} + 2(x_k-\mu_k)\sum_{s=1}^{p-1}\left(w(s)\binom{p-2}{s-1}\sum_{j\in P\setminus\{k\}}(x_j-\mu_j)\omega_{jk}\right) \\
&= (x_k-\mu_k)^2\omega_{kk} + 2(x_k-\mu_k)\sum_{s=1}^{p-1}\left(\frac{s}{p(p-1)}\sum_{j\in P\setminus\{k\}}(x_j-\mu_j)\omega_{jk}\right) \\
&= (x_k-\mu_k)^2\omega_{kk} + (x_k-\mu_k)\sum_{j\in P\setminus\{k\}}(x_j-\mu_j)\omega_{jk} \\
&= (x_k-\mu_k)\sum_{j\in P}(x_j-\mu_j)\omega_{jk} = (x_k-\mu_k)\left(\sum_{j=1}^{p}(x_j-\mu_j)\omega_{jk}\right)
\end{aligned}
$$

□

A.2 Proof of Theorem 2.2.3.1

Proof of Theorem 2.2.3.1. To derive the off-diagonal elements defined in Equation (2.12), we start with rewriting $\Delta_{\{j,k\}}\operatorname{MD}^2(\hat{\boldsymbol{x}}^T), T\subseteq P\setminus\{j,k\}$, by applying Lemma A.1.1:

$$
\begin{aligned}
&\Delta_{\{j,k\}}\operatorname{MD}^2(\hat{\boldsymbol{x}}^T) \\
=&\big[\operatorname{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) - \operatorname{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}})\big] - \big[\operatorname{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) - \operatorname{MD}^2(\hat{\boldsymbol{x}}^{T})\big] \\
=&2(x_k-\mu_k)\left(\sum_{l\in T\cup\{j\}}(x_l-\mu_l)\omega_{jk} - \sum_{l\in T\cup\{k\}}(x_l-\mu_l)\omega_{jk}\right) + 2(x_k-\mu_k)^2\omega_{kk} \\
=&2(x_k-\mu_k)\left((x_j-\mu_j)\omega_{jk} - (x_k-\mu_k)\omega_{kk}\right) + 2(x_k-\mu_k)^2\omega_{kk} \\
=&2(x_k-\mu_k)(x_j-\mu_j)\omega_{jk}
\end{aligned}
$$

Moving on, we plug the result into the formula for Φ_{jk} for $j \neq k$, given in Equation (2.12), and we obtain

$$\begin{aligned}\Phi_{jk} &= \sum_{T \subseteq P \setminus \{j,k\}} \frac{t!(p-t-2)!}{(p-1)!} \Delta_{\{j,k\}} \operatorname{MD}^2(\hat{\boldsymbol{x}}^T) \\ &= \sum_{T \subseteq P \setminus \{j,k\}} \frac{t!(p-t-2)!}{(p-1)!} 2(x_k - \mu_k)(x_j - \mu_j)\omega_{jk} \\ &= 2(x_k - \mu_k)(x_j - \mu_j)\omega_{jk},\end{aligned}$$

where the last equality is obtained by following the same structure as in the proof of Theorem 2.2.2.1. Finally, we have to derive the diagonal elements Φ_{jj} given by

$$\begin{aligned}\Phi_{jj} &= \phi_j - \sum_{k \neq j} \Phi_{jk} \\ &= (x_j - \mu_j) \sum_{k=1}^{p} (x_k - \mu_k)\omega_{jk} - 2(x_j - \mu_j) \sum_{k \neq j} (x_k - \mu_k)\omega_{jk} \\ &= (x_j - \mu_j)^2 \omega_{jj} - (x_j - \mu_j) \sum_{k \neq j} (x_k - \mu_k)\omega_{jk}.\end{aligned}$$

□

Higher Order Interactions

Proof. To show that all interactions of order three or higher are zero, it is sufficient to show that for the three-way interactions the *set function derivative* $\Delta_{\{j,k,l\}} \operatorname{MD}^2(\hat{\boldsymbol{x}}^T)$ is zero for all $T \subseteq P \setminus \{j,k,l\}$. This follows from the iterative definition of the set function derivative for $S \cap \{j\} = \emptyset$ (Grabisch, 2016), which is given by

$$\Delta_{S \cup \{j\}} \operatorname{MD}^2(\hat{\boldsymbol{x}}^T) = \Delta_S(\Delta_j \operatorname{MD}^2(\hat{\boldsymbol{x}}^T)).$$

Hence, to show that all Shapley interaction indices

$$I_{Sh}(v, S) = \sum_{T \subseteq P \setminus S} \frac{t!(p-t-s)!}{(p-s+1)!} \Delta_S v(T),$$

with $|S| \geq 3$ are zero, we only have to prove that $\Delta_{\{j,k,l\}} \mathrm{MD}^2(\hat{\boldsymbol{x}}^T) = 0, \forall T \subseteq P \setminus \{j,k,l\}$. For this purpose, we first rewrite the above expression and then apply Lemma A.1.1:

$$\begin{aligned}
\Delta_{\{j,k,l\}} \mathrm{MD}^2(\hat{\boldsymbol{x}}^T) = &- \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k,l\}}) \\
&+ \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,l\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k,l\}}) \\
&- \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{l\}}) \\
&+ \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T}) \\
= &- \left[\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k,l\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k,l\}})\right] \\
&+ \left[\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,l\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{l\}})\right] \\
&+ \left[\mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j,k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{j\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T\cup\{k\}}) + \mathrm{MD}^2(\hat{\boldsymbol{x}}^{T})\right] \\
= &- \left[(x_j - \mu_j)^2\omega_{jj} + 2(x_j - \mu_j) \sum_{m\in T\cup\{k,l\}} (x_m - \mu_m)\omega_{jm}\right] \\
&+ \left[(x_j - \mu_j)^2\omega_{jj} + 2(x_j - \mu_j) \sum_{m\in T\cup\{l\}} (x_j - \mu_j)\omega_{jk}\right] \\
&+ \left[2(x_k - \mu_k)(x_j - \mu_j)\right] \\
= &- 2(x_k - \mu_k)(x_j - \mu_j) + 2(x_k - \mu_k)(x_j - \mu_j) = 0
\end{aligned}$$

□

A.3 Weather in Vienna - Parameters

Table A.3 shows the parameter descriptions for the *Weather in Vienna* dataset, which have been adapted and translated from Stadt Wien (2022).

Table A.3: Description of the parameters of the *Weather in Vienna* dataset

Parameter	Desciption
`avg_t_max`	Mean daily maximum air temperature in °C
`avg_t_min`	Mean daily minimum air temperature in °C
`num_summer`	Number of summer days (days with a temperature maximum $t_{\max} \geq 25.0$ °C)
`num_heat`	Number of hot days (days with a temperature maximum $t_{\max} \geq 30.0$ °C)
`p`	Daily mean air pressure in hPa (mean of all measurements at 7 a.m., 2 p.m., 7 p.m. CET; before 1971 9 p.m. instead of 7 p.m.)
`sun_h`	Monthly total sunshine duration in hours
`num_clear`	Number of clear days (daily mean cloudiness $< 20/100$)
`num_cloud`	Number of cloudy days (daily mean cloudiness $> 80/100$)
`rel_hum`	Daily mean relative humidity in percent (2 x RH7 mean + RH14 mean + RH19 mean)/4; before 1971 9 p.m. instead of 7 p.m.)
`rel_hum_max`	Relative humidity maximum in percent
`rel_hum_min`	Relative humidity minimum in percent
`wind_v`	Monthly average wind speed in km/h
`num_wind_v60`	Number of days with wind peaks ≥ 60 km/h
`wind_v_max`	Maximum wind speed in km/h
`precp_sum`	Monthly total precipitation in mm
`num_precp_01`	Number of days with precipitation ≥ 0.1 mm

3 Robust Covariance Estimation and Explainable Outlier Detection for Matrix-valued Data

This chapter is based on the work Mayrhofer, M., Radojičić, U., and Filzmoser, P. (2024a). Robust covariance estimation and explainable outlier detection for matrix-valued data. *arXiv preprint arXiv:2403.03975*.

Contributions: M. Mayrhofer developed the methodological framework, implemented the procedures in `C++` and `R` in the package `robustmatrix` (Mayrhofer et al., 2024b), and wrote the first draft of the paper. He established the proofs together with co-author Radojičić U. All co-authors were involved in the discussions and collaborated on writing the final paper.

3.1 Introduction

Thanks to modern data collection tools, the amount and complexity of available information are increasing rapidly, and matrix-valued data are often observed. Compared to classical multivariate observations, where values for p variables are recorded for one subject, matrix-valued observations are recorded on a grid of $p \times q$ variables. These are then naturally represented as a matrix with p rows and q columns. Some examples include image data, where p and q are given by the resolution of the image, or multivariate data measured on p variables, where the measurements for a subject are available for q replications (e.g., different time points, different spatial locations, different experimental conditions, etc.). Frequently, matrix-valued data are analyzed as classical multivariate data by stacking the matrix columns (or rows) to a vector of length $p \cdot q$. Thus, if n observations are available, the data are arranged in a matrix of dimension $n \times pq$. Depending on the dimensions, this can create high-dimensional data, possibly with a sample size lower than the resulting dimensionality, which constitutes a limitation for multivariate statistical methods.

As an alternative to vectorizing matrix-valued observations, we model them under the assumption that they originate from a certain matrix-variate distribution. As in the multivariate setting, the class of matrix-elliptical distributions (Gupta et al., 2013), serves as a natural ground for studying covariance estimation. The matrix-elliptical family is a semi-parametric class of distributions parametrized by the mean $\mathbf{M} \in \mathbb{R}^{p \times q}$, row covariance $\mathbf{\Sigma}^{\text{row}} \in \text{PDS}(p)$, column covariance $\mathbf{\Sigma}^{\text{col}} \in \text{PDS}(q)$, and the so-called density generator function $g : [0, \infty) \to \mathbb{R}$. Here, $\text{PDS}(a)$, with $a \in \mathbb{N}$, denotes the class of all positive definite symmetric $a \times a$ matrices. More specifically, a random matrix $\boldsymbol{X}$ with an absolutely continuous distribution has an elliptical distribution, denoted $\mathcal{ME}(\mathbf{M}, \mathbf{\Sigma}^{\text{row}}, \mathbf{\Sigma}^{\text{col}}, g)$, if its

density can be written as

$$f(\boldsymbol{X}) = \det(\boldsymbol{\Sigma}^{\text{row}})^{-q/2} \det(\boldsymbol{\Sigma}^{\text{col}})^{-p/2} g(\operatorname{tr}(\boldsymbol{\Omega}^{\text{col}}(\boldsymbol{X}-\mathbf{M})'\boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X}-\mathbf{M}))), \tag{3.1.1}$$

with $\boldsymbol{\Omega}^{\text{row}} = (\boldsymbol{\Sigma}^{\text{row}})^{-1}$ and $\boldsymbol{\Omega}^{\text{col}} = (\boldsymbol{\Sigma}^{\text{col}})^{-1}$ denoting the precision matrices among the rows and columns, respectively. Matrix elliptical distributions can also be related to their multivariate counterparts. Formally, a random matrix $\boldsymbol{X}$ follows a matrix elliptical distribution $\mathcal{ME}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}, g)$ if and only if its vectorized version $\operatorname{vec}\boldsymbol{X}$ follows a multivariate elliptical distribution $\mathcal{E}(\operatorname{vec}(\mathbf{M}), \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}, g)$ (Gupta et al., 2013). Here, $\operatorname{vec}(\cdot)$ is the vectorization operator, stacking the columns of a matrix on top of each other, $\otimes$ is the Kronecker product. Probably the most studied matrix elliptical distribution is the matrix normal distribution (Dawid, 1981), denoted $\mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$, with density

$$f(\boldsymbol{X} \,|\, \mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}) = \frac{\exp(-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Omega}^{\text{col}}(\boldsymbol{X}-\mathbf{M})'\boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X}-\mathbf{M})))}{(2\pi)^{pq/2} \det(\boldsymbol{\Sigma}^{\text{col}})^{p/2} \det(\boldsymbol{\Sigma}^{\text{row}})^{q/2}}. \tag{3.1.2}$$

Regarding the estimation of location and covariance for an i.i.d. sample $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ in $\mathbb{R}^{n\times p\times q}$, with $\boldsymbol{X}_i \sim \mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$, we can either work with the vectorized observations or directly with the matrices. In the former setting, the existence and uniqueness of the maximum likelihood estimator (MLE) for the covariance is guaranteed almost surely if $n \geq pq + 1$. However, this approach does not take advantage of the Kronecker structure of the covariance matrix and instead directly estimates the entire pq-dimensional matrix $\boldsymbol{\Sigma}$. In contrast, if we utilize the knowledge of the inherent data structure, we only need to estimate the p-dimensional rowwise covariance matrix $\boldsymbol{\Sigma}^{\text{row}}$ and the q-dimensional columnwise covariance matrix $\boldsymbol{\Sigma}^{\text{col}}$. For the matrix-variate sample $\mathfrak{X}$, the MLEs for the mean, as well as for the rowwise and columnwise covariance, are given by (Dutilleul, 1999):

$$\hat{\mathbf{M}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \tag{3.1.3}$$

$$\hat{\boldsymbol{\Sigma}}^{\text{row}} = \frac{1}{qn}\sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\mathbf{M}})\hat{\boldsymbol{\Omega}}^{\text{col}}(\boldsymbol{X}_i - \hat{\mathbf{M}})' \tag{3.1.4}$$

$$\hat{\boldsymbol{\Sigma}}^{\text{col}} = \frac{1}{pn}\sum_{i=1}^{n} (\boldsymbol{X}_i - \hat{\mathbf{M}})'\hat{\boldsymbol{\Omega}}^{\text{row}}(\boldsymbol{X}_i - \hat{\mathbf{M}}) \tag{3.1.5}$$

Soloveychik and Trushin (2016) showed that for n i.i.d. samples from a continuous $p\times q$ matrix-variate distribution, there exists no unique maximum of the matrix normal likelihood function if $n < \max(p/q, q/p) + 1$, and that a unique maximum exists almost surely if $n \geq \lfloor p/q + q/p \rfloor + 2$. Although there are no closed-form solutions for the maximum likelihood estimates (MLEs) of $\boldsymbol{\Sigma}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}}$, Dutilleul (1999) proposed an iterative estimation procedure. The idea of the so-called *flip-flop* algorithm is to alternate between the computation of $\hat{\boldsymbol{\Sigma}}^{\text{row}}$ and $\hat{\boldsymbol{\Sigma}}^{\text{col}}$ based on Equations (3.1.4) and (3.1.5), respectively, until a convergence criterion is met. The algorithm is constructed such that positive definite estimates of subsequent iterations are nondecreasing in likelihood (Lu and Zimmerman, 2005), and it converges almost surely to the unique maximum from any symmetric positive definite initialization of either $\hat{\boldsymbol{\Sigma}}^{\text{row}}$ or $\hat{\boldsymbol{\Sigma}}^{\text{col}}$, if $n \geq \lfloor p/q + q/p \rfloor + 2$ (Soloveychik and Trushin, 2016).

Existing proposals for robust covariance estimation include a generalization of Tyler's M-estimator (Tyler, 1987) introduced by Soloveychik and Trushin (2016), a robust estimator for structured covariance matrices with Kronecker structure as a particular case (Sun et al., 2016), distribution-free robust covariance estimation (Zhang et al., 2022), and ML estimation for the matrix t-distribution (Thompson et al., 2020).

We propose novel robust estimators for the parameters $\mathbf{M}$, $\mathbf{\Sigma}^{\text{row}}$, and $\mathbf{\Sigma}^{\text{col}}$, termed the matrix minimum covariance determinant (MMCD) estimators. These estimators generalize the minimum covariance determinant (MCD) approach (Rousseeuw, 1985), one of the most widely used approaches for robustly estimating the mean and covariance of multivariate (vector-valued) data. We show that the MMCD estimators are equivariant under matrix affine transformations and surpass the maximal attainable breakdown point of any multivariate, affine equivariant location/covariance estimator when applied to the vectorized data, such as the MCD estimator. Additionally, we show that the MMCD estimators are consistent for the finite-dimensional parameters $(\mathbf{M}, \mathbf{\Sigma}^{\text{row}}, \mathbf{\Sigma}^{\text{col}})$ of the matrix elliptical distribution, thus bridging a gap between the individual, distribution-specific, estimators in the elliptical family. Furthermore, a concentration step (C-step) algorithm is developed to efficiently compute the MMCD estimators; see Rousseeuw and Van Driessen (1999) for more details on C-step for MCD. Additionally, we introduce a reweighting step that preserves the properties of the MMCD estimators and greatly increases finite-sample efficiency.

The robust MMCD estimators can then be employed for outlier detection using the Mahalanobis distances (Mahalanobis, 1936) for matrix-valued observations. Because it is essential to understand the reasons for the outlyingness, we extend the concept of Shapley values introduced in Mayrhofer and Filzmoser (2023) for outlier explanation in the multivariate case to the matrix-variate setting. Shapley values (Shapley, 1953) are well-known from explainable AI (Lundberg and Lee, 2017), but their computation is usually time-consuming. Our proposal is computationally efficient, and the resulting Shapley values preserve their attractive properties (Shapley, 1953).

The paper is organized as follows. In Section 3.2, we introduce the MMCD estimators, then proceed to derive their theoretical properties in Section 3.3. Section 3.4 is devoted to computational details for the MMCD estimators. In Section 3.5, we propose Shapley values for outlier explanation and present their properties. In Sections 3.6 and 3.7, we illustrate the performance of the proposed methods on numerical simulations and real-world examples. Section 3.8 concludes our findings. The supplementary materials contain more information on the theoretical background in this context, proofs, technical derivations, code, and additional numerical results.

3.2 The MMCD Estimators

The MLEs given in Equations (3.1.3)-(3.1.5), much like the multivariate normal MLEs, i.e. sample mean and covariance, also serve as valid (consistent) parameter estimators in the class of elliptical distributions; see Remark 3.3.0.1. However, just like their multivariate counterparts, these are not robust against outlying observations. In order to obtain robust estimators for the finite-dimensional parameters $(\mathbf{M}, \mathbf{\Sigma}^{\text{row}}, \mathbf{\Sigma}^{\text{col}})$ in Equation (3.1.1), we optimize the weighted version of the matrix-normal (log-)likelihood function. This principle

has been similarly used in the context of other robust estimators (e.g., Neykov et al., 2007; García-Escudero et al., 2010; Kurnaz et al., 2018), and in particular, Raymaekers and Rousseeuw (2023) show that the MCD estimator can be reformulated in terms of likelihood; the objective of the MCD estimator is to identify the subset of h out of n samples ($n/2 \leq h \leq n$) with the smallest determinant of the sample covariance matrix. This is equivalent to determining a subset of size h that maximizes the multivariate normal (log-)likelihood function.

Extending the concept of the multivariate MCD approach, we introduce weights $\boldsymbol{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ for a given sample $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ that is independently drawn from $\mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$ to formulate the weighted log-likelihood function $l(\boldsymbol{w}, \mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}} | \mathfrak{X})$ as

$$-\frac{1}{2}\sum_{i=1}^{n} w_i \Big(p \ln(\det(\boldsymbol{\Sigma}^{\text{col}})) + q \ln(\det(\boldsymbol{\Sigma}^{\text{row}})) + \text{MMD}^2(\boldsymbol{X}) + pq \ln(2\pi) \Big), \tag{3.2.1}$$

where $\text{MMD}^2(\boldsymbol{X})$ denotes the squared matrix Mahlanobis distance defined as

$$\text{MMD}^2(\boldsymbol{X}) := \text{MMD}^2(\boldsymbol{X}; \mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}) = \text{tr}(\boldsymbol{\Omega}^{\text{col}}(\boldsymbol{X} - \mathbf{M})' \boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X} - \mathbf{M})). \tag{3.2.2}$$

Setting $w_i = 1$ for all $i = 1, \ldots, n$, yields the traditional log-likelihood function, and its maximization yields the MLEs of Equation (3.1.3)-(3.1.5). However, by taking binary weights, $w_i \in \{0, 1\}$, with the constraint that $\sum_{i=1}^{n} w_i = h$, we see that $n - h$ contributions are trimmed. Since contributions from outliers should be trimmed, the task is to identify the subset of regular observations $H \subset \{1, \ldots, n\}$ with $|H| = h$, where $w_i = 1$ for $i \in H$ and 0 otherwise. The resulting constrained optimization problem of finding the weighted MLE can be written as

$$\begin{aligned} &\max_{\boldsymbol{w}, \mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}} l(\boldsymbol{w}, \mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}} | \mathfrak{X}) \\ &\text{subject to} \quad w_i \in \{0, 1\} \text{ for all } i = 1, \ldots, n \quad \text{and} \quad \sum_{i=1}^{n} w_i = h. \end{aligned} \tag{3.2.3}$$

To improve clarity, we will use the following notation for subsamples of $\mathfrak{X}$ and estimators based on it: Let $H \subseteq \{1, \ldots, n\}$ be a subset of size $h = |H|$, then $\mathfrak{X}_H := (\boldsymbol{X}_i)_{i \in H}$ denotes an h-subset of $\mathfrak{X}$. An estimator for a parameter θ based on the sample $\mathfrak{X}$ is denoted as $\hat{\theta}_{\mathfrak{X}}$ or simply as $\hat{\theta}$ if it is clear on which sample the estimator is computed. Similarly, if an estimator is based on an h-subset, it is denoted as $\hat{\theta}_H$ or as $\hat{\theta}_{\mathfrak{X}_H}$.

Proposition 3.2.0.1. *Let $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, $n/2 \leq h \leq n$ and $h \geq \lfloor p/q + q/p \rfloor + 2$, be an i.i.d. sample from $\mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$. Maximizing the weighted log-likelihood function* (3.2.3) *is equivalent to minimizing*

$$\ln(\det(\hat{\boldsymbol{\Sigma}}_H^{\text{col}} \otimes \hat{\boldsymbol{\Sigma}}_H^{\text{row}})) = p \ln(\det(\hat{\boldsymbol{\Sigma}}_H^{\text{col}})) + q \ln(\det(\hat{\boldsymbol{\Sigma}}_H^{\text{row}})) \tag{3.2.4}$$

across all subsets $H \subset \{1, \ldots, n\}$ *with* $|H| = h$. *In Equation* (3.2.4),

$$\hat{\mathbf{M}}_H = \frac{1}{h} \sum_{i \in H} \boldsymbol{X}_i, \tag{3.2.5}$$

$$\hat{\boldsymbol{\Sigma}}^{\text{row}}_H = \frac{1}{qh} \sum_{i \in H} (\boldsymbol{X}_i - \hat{\mathbf{M}}_H) \hat{\boldsymbol{\Omega}}^{\text{col}}_H (\boldsymbol{X}_i - \hat{\mathbf{M}}_H)', \text{ and} \tag{3.2.6}$$

$$\hat{\boldsymbol{\Sigma}}^{\text{col}}_H = \frac{1}{ph} \sum_{i \in H} (\boldsymbol{X}_i - \hat{\mathbf{M}}_H)' \hat{\boldsymbol{\Omega}}^{\text{row}}_H (\boldsymbol{X}_i - \hat{\mathbf{M}}_H) \tag{3.2.7}$$

denote the MLEs based on the observations in H, *and* $\hat{\boldsymbol{\Omega}}^{\text{row}}_H = (\hat{\boldsymbol{\Sigma}}^{\text{row}}_H)^{-1}$ *and* $\hat{\boldsymbol{\Omega}}^{\text{col}}_H = (\hat{\boldsymbol{\Sigma}}^{\text{col}}_H)^{-1}$ *denote the corresponding precision matrices.*

A proof is given in Supplement B.2. Based on this proposition, we obtain a matrix-variate counterpart to the multivariate MCD estimator's objective, resulting in robust estimators of the parameters $\mathbf{M}$, $\boldsymbol{\Sigma}^{\text{row}}$, and $\boldsymbol{\Sigma}^{\text{col}}$.

Definition 3.2.0.1. *Let* $\boldsymbol{\mathcal{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, $n/2 \leq h \leq n$ *and* $h \geq \lfloor p/q + q/p \rfloor + 2$, *be an i.i.d sample of a continuous* $p \times q$ *matrix-variate distribution. The* raw *matrix minimum covariance determinant (MMCD) estimators are defined as*

$$(\hat{\mathbf{M}}_{H^*}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H^*}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H^*}) := \underset{\substack{\hat{\mathbf{M}}_H, \hat{\boldsymbol{\Sigma}}^{\text{row}}_H, \hat{\boldsymbol{\Sigma}}^{\text{col}}_H \\ H \subset \{1,\ldots,n\}, |H| = h}}{\arg\min} \; p \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{col}}_H)) + q \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{row}}_H)), \tag{3.2.8}$$

with $\hat{\mathbf{M}}_H$, $\hat{\boldsymbol{\Sigma}}^{\text{row}}_H$, *and* $\hat{\boldsymbol{\Sigma}}^{\text{col}}_H$ *as in Equations* (3.2.5), (3.2.6), *and* (3.2.7), *respectively.*

The estimators in Definition (3.2.0.1) almost surely exist and are positive definite if $h \geq \lfloor p/q + q/p \rfloor + 2$ (Soloveychik and Trushin, 2016). If $p = 1$ and/or $q = 1$, optimization problem (3.2.8) coincides with the optimization problem of the MCD estimator, and one obtains the univariate or multivariate MCD estimator, respectively.

3.3 Properties of the MMCD Estimators

Matrix affine equivariance. The concept of affine equivariance in multivariate analysis is rooted in the idea that the estimators used for location and covariance should transform in the same way as the parameters of elliptically symmetrical unimodal distributions (referred to as elliptical distributions hereafter), see Maronna et al. (2019). We can define the matrix-variate analog of affine equivariance based on the properties of matrix-variate elliptical distributions, which are frequently employed to study the robustness properties of normal theory under nonnormal situations (Gupta and Nagar, 1999).

Linear functions of a random matrix $\boldsymbol{X} \sim \mathcal{ME}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}, g)$ also have an elliptical distribution (Gupta et al., 2013). This means that for constant matrices $\boldsymbol{A} \in \mathbb{R}^{r \times p}$, $\text{rank}(\boldsymbol{A}) = r \leq p$, $\boldsymbol{B} \in \mathbb{R}^{q \times s}$, $\text{rank}(\boldsymbol{B}) = s \leq q$, and $\boldsymbol{C} \in \mathbb{R}^{r \times s}$, the transformed random matrix $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{B} + \boldsymbol{C}$ has density

$$\boldsymbol{Z} \sim \mathcal{ME}(\boldsymbol{A}\,\mathbf{M}\,\boldsymbol{B} + \boldsymbol{C}, \boldsymbol{A}\boldsymbol{\Sigma}^{\text{row}}\boldsymbol{A}', \boldsymbol{B}'\boldsymbol{\Sigma}^{\text{col}}\boldsymbol{B}, g). \tag{3.3.1}$$

Let $\hat{\mathbf{M}}_{\mathfrak{X}}$, $\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}}$, and $\hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{X}}$ denote the estimators based on a sample $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ generated by $f(\mathbf{M}, \mathbf{\Sigma}^{\text{row}}, \mathbf{\Sigma}^{\text{col}})$. Then the estimators of the sample $\mathfrak{Z} = (\boldsymbol{A}\boldsymbol{X}_1\boldsymbol{B}+\boldsymbol{C}, \ldots, \boldsymbol{A}\boldsymbol{X}_n\boldsymbol{B}+\boldsymbol{C})$ should transform in the same way as the parameters in Equation (3.3.1), i.e.,

$$\hat{\mathbf{M}}_{\mathfrak{Z}} = \boldsymbol{A}\hat{\mathbf{M}}_{\mathfrak{X}}\boldsymbol{B} + \boldsymbol{C}, \quad \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Z}} = \boldsymbol{A}\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}}\boldsymbol{A}', \quad \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Z}} = \boldsymbol{B}'\hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{X}}\boldsymbol{B}. \tag{3.3.2}$$

Properties (3.3.2) provide a suitable generalization of affine equivariance to the matrix-variate setting, and it is easy to verify that they hold for the estimators given in (3.1.3)-(3.1.5). However, they do not imply affine equivariance of the location and covariance estimators for the vectorized observations. This would only hold for transformations with the Kronecker structure $\text{vec}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B} + \boldsymbol{C}) = (\boldsymbol{B}' \otimes \boldsymbol{A})\,\text{vec}(\boldsymbol{X}) + \text{vec}(\boldsymbol{C})$. We refer to Properties (3.3.2) as *matrix affine equivariance* to avoid confounding definitions.

Lemma 3.3.0.1. *Let $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ be a sample of $p \times q$ matrices, where $\boldsymbol{X}_i \sim \mathcal{ME}(\mathbf{M}_{\mathfrak{X}}, \mathbf{\Sigma}^{\text{row}}_{\mathfrak{X}}, \mathbf{\Sigma}^{\text{col}}_{\mathfrak{X}}, g)$, and let $\mathfrak{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ be the affine transformation of $\mathfrak{X}$, i.e., $\boldsymbol{Z}_i = \boldsymbol{A}\boldsymbol{X}_i\boldsymbol{B} + \boldsymbol{C}$, $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, $\boldsymbol{B} \in \mathbb{R}^{q\times q}$, $\boldsymbol{A}, \boldsymbol{B}$ invertible, and $\boldsymbol{C} \in \mathbb{R}^{p\times q}$. The following then holds:*

(a) The MMCD estimators as in Definition 3.2.0.1 are matrix affine equivariant.

(b) $\text{MMD}^2(\boldsymbol{Z}_i; \hat{\mathbf{M}}_{\mathfrak{Z}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Z}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Z}}) = \text{MMD}^2(\boldsymbol{X}_i; \hat{\mathbf{M}}_{\mathfrak{X}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{X}})$, *where* $(\hat{\mathbf{M}}_{\mathfrak{Z}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Z}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Z}})$ *are matrix affine equivariant location and covariance estimators of the transformed sample* $\mathfrak{Z}$.

Lemma 3.3.0.1 shows that the MMCD estimators are equivariant under matrix affine transformations, and a proof is given in Supplement B.2.

Breakdown point. The finite sample breakdown point of an estimator evaluates its resilience to contamination. It refers to the largest proportion of observations that may be arbitrarily replaced by outliers such that the estimator still contains some information about the true parameter (Maronna et al., 2019). Let $\mathfrak{X}$ be a sample of n matrix-variate observations in $\mathbb{R}^{p\times q}$ and suppose $\mathfrak{Y}$ is a corrupted version, obtained by replacing m samples of $\mathfrak{X}$ by arbitrary matrices. The finite sample breakdown point of a location estimator $\hat{\mathbf{M}}$ is given by

$$\varepsilon^*(\hat{\mathbf{M}}, \mathfrak{X}) = \min_{1\le m\le n} \left\{ \frac{m}{n} : \sup_m \left\| \hat{\mathbf{M}}_{\mathfrak{X}} - \hat{\mathbf{M}}_{\mathfrak{Y}} \right\| = \infty \right\} \tag{3.3.3}$$

and the (joint) finite sample breakdown point of row and columnwise covariance estimators $\hat{\mathbf{\Sigma}}^{\text{row}}$ and $\hat{\mathbf{\Sigma}}^{\text{col}}$ is given by

$$\varepsilon^*(\hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}, \mathfrak{X}) = \min_{1\le m\le n} \left\{ \frac{m}{n} : \sup_m \max_{i,j} \left| D_\lambda(\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{X}}) \right| = \infty \right\}, \tag{3.3.4}$$

where

$$D_\lambda(\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{X}}) = \log(\lambda_i(\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}})\lambda_j(\hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}})) - \log(\lambda_i(\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}})\lambda_j(\hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{X}})).$$

While the MCD and the MMCD estimators coincide for the case that $p = 1$ and/or $q = 1$, the following theorem shows that the MMCD estimators achieve a higher breakdown point than the MCD estimators applied to the vectorized samples if $p \ge 2$ and $q \ge 2$.

Theorem 3.3.0.1. *Let $\mathfrak{X}$ be a collection of n i.i.d. samples from a continuous $p \times q$ matrix-variate distribution, where $d = \lfloor p/q + q/p \rfloor$, $p, q \in \mathbb{N}, p \geq 2, q \geq 2$, and let $\hat{\mathbf{M}}$, $\hat{\mathbf{\Sigma}}^{\text{row}}$, and $\hat{\mathbf{\Sigma}}^{\text{col}}$ denote the MMCD estimators, then*

$$\varepsilon^*(\hat{\mathbf{M}}, \mathfrak{X}) = \varepsilon^*(\hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}, \mathfrak{X}) = \frac{1}{n}\lfloor \min(n - h + 1, h - (d + 1)) \rfloor =: \frac{m}{n},$$

with $n/2 \leq h \leq n$ and $h \geq d + 2$.

The proof extends established methodologies from Rousseeuw (1985) and Lopuhaa and Rousseeuw (1991) to address the matrix variate setting, leveraging additional insights and techniques outlined in Supplement B.2. Theorem 3.3.0.1 implies that the maximum breakdown point of the MMCD estimators is $1/n\lfloor (n-d)/2 \rfloor$ and is attained if $h = \lfloor (n+d+2)/2 \rfloor$. This means that the maximum breakdown point of the MMCD covariance estimators for $p \geq 2, q \geq 2$ is higher than the upper bound for the breakdown point of affine equivariant covariance estimators applied to vectorized samples, which is given by $1/n\lfloor (n-pq+1)/2 \rfloor$ (Davies, 1987; Lopuhaa and Rousseeuw, 1991). However, as mentioned earlier, affine equivariance in the matrix-variate setting does not imply affine equivariance in the multivariate setting. Thus, the mentioned upper bound for the vectorized observations does not apply. In other words, since affine equivariance (in the vectorized case) is not a requirement for matrix-variate affine equivariance, it is possible to achieve a higher breakdown point for the MMCD estimators than for any affine equivariant multivariate estimator applied to the vectorized data.

To illustrate the advantage of respecting the inherent data structure of matrix-variate data for the breakdown properties, we compare the maximum breakdown points of the MCD and MMCD estimators in Figure 3.3.1 for different combinations of p and q, and for different sample sizes n. Here, the MCD estimator is applied to the vectorized data, and the dimensionality of the samples is pq, which can get large. This affects the computability of the MCD estimator since it requires a subset size larger than the dimension.

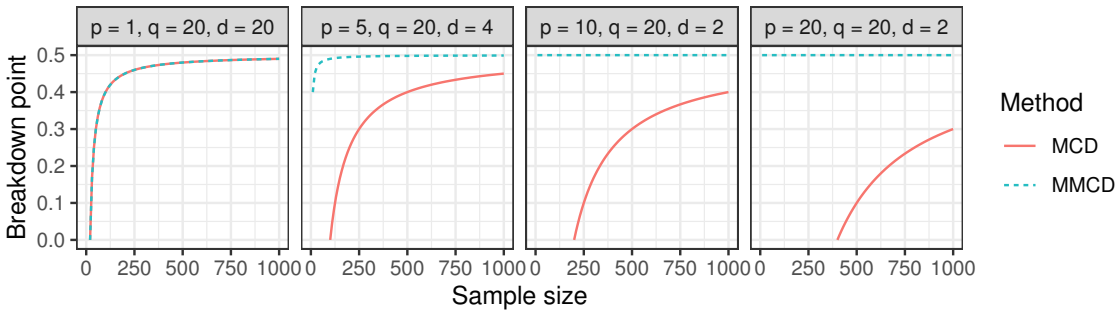


Figure 3.3.1: Comparison of the maximum breakdown point of the MMCD estimators for matrix-variate data with $p = 1, 5, 10, 20$ rows and $q = 20$ columns, and the MCD estimator applied to the vectorized data. When $p = 1$, both estimators and their breakdown points coincide. However, increasing the number of rows yields better breakdown properties for the MMCD estimators, as the proportion between the number of rows and columns $d = \lfloor p/q + q/p \rfloor$ is approaching 2.

Consistency for elliptical distributions. Let us now consider the asymptotic behavior of the MMCD estimators. By scaling the rowwise or columnwise MMCD covariance estimator by a distribution-specific consistency factor, we can achieve consistency for elliptical distributions.

Theorem 3.3.0.2. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample from an elliptical matrix-variate distribution $\mathcal{ME}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{row}}, g)$ with positive definite covariances $\boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}$, and let $(\hat{\mathbf{M}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}, \hat{\boldsymbol{\Sigma}}^{\text{col}})$ be the corresponding MMCD estimators. Then, it holds that*

$$\left\|\hat{\mathbf{M}} - \mathbf{M}\right\| \xrightarrow{a.s.} 0, \quad \left\|c(\alpha)\hat{\boldsymbol{\Sigma}}^{\text{col}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}} - \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}\right\| \xrightarrow{a.s.} 0,$$

where $c(\alpha), \alpha = h/n \in [0.5, 1]$, is a distribution-specific consistency factor as in Croux and Haesbroeck (1999).

The proof of the consistency of the *raw* MMCD estimators relies on the strong consistency of the MCD estimator given in Butler et al. (1993); Cator and Lopuhaä (2012) and is provided in Supplement B.2. It shows that the consistency factor of the MCD estimator and the MMCD estimator must coincide, and therefore, we use the consistency factor

$$c(\alpha) = \frac{\alpha}{F_{\chi^2_{pq+2}}(\chi^2_{\alpha;pq})} \tag{3.3.5}$$

proposed by Croux and Haesbroeck (1999) to obtain consistency at the normal model, where $F_{\chi^2_{pq+2}}$ denotes the CDF of the chi-square distribution with $pq+2$ degrees of freedom, and $\chi^2_{pq;\alpha}$ denotes the α quantile of the chi-square distribution with pq degrees of freedom.

Remark 3.3.0.1. *Note first that for $h = n$, the corresponding MMCD estimators coincide with the ones defined in* (3.1.3)-(3.1.5)*. Therefore, a simple, yet not discussed in the literature, consequence of Theorem 3.3.0.2 is that the estimators obtained maximizing the likelihood under the matrix-normal model, are consistent estimators of the corresponding finite-dimensional parameters $(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{row}})$ in the semi-parametric, matrix elliptical family.*

Reweighted MMCD - improving efficiency. The *raw* MMCD estimators are most robust when about half of the observations are trimmed, i.e., $h = \lfloor (n+d+2)/2 \rfloor$. However, this leads to a low efficiency at the normal model. While efficiency could be increased by trimming fewer samples, this would lead to lower robustness. To enhance a robust estimator's efficiency without compromising robustness, Lopuhaa and Rousseeuw (1991); Maronna et al. (2019) proposed a one-step reweighing procedure. We can apply this technique for the MMCD estimators by defining weighted ML estimators with weights depending on the Mahalanobis distances given the raw MMCD estimators.

Definition 3.3.0.1. *Let $\mathfrak{X}$ be a collection of n i.i.d. samples from a continuous $p \times q$ matrix-variate distribution, where $d = \lfloor p/q + q/p \rfloor$, $p, q \in \mathbb{N}, p \geq 2, q \geq 2$, and let $\hat{\mathbf{M}}$, $\hat{\boldsymbol{\Sigma}}^{\text{row}}$, and $\hat{\boldsymbol{\Sigma}}^{\text{col}}$ denote the* raw *MMCD estimators as in Definition 3.2.0.1. The* reweighted *MMCD*

estimators are given by

$$\tilde{\mathbf{M}} = \frac{1}{\sum_{i=1}^{n} w(\mathrm{MMD}(\boldsymbol{X}_i))} \sum_{i=1}^{n} w(\mathrm{MMD}(\boldsymbol{X}_i))\boldsymbol{X}_i, \tag{3.3.6}$$

$$\tilde{\boldsymbol{\Sigma}}^{\mathrm{row}} = \frac{1}{q\sum_{i=1}^{n} w(\mathrm{MMD}(\boldsymbol{X}_i))} \sum_{i=1}^{n} w(\mathrm{MMD}(\boldsymbol{X}_i))(\boldsymbol{X}_i - \tilde{\mathbf{M}})\tilde{\boldsymbol{\Omega}}^{\mathrm{col}}(\boldsymbol{X}_i - \tilde{\mathbf{M}})', \text{ and} \tag{3.3.7}$$

$$\tilde{\boldsymbol{\Sigma}}^{\mathrm{col}} = \frac{1}{p\sum_{i=1}^{n} w(\mathrm{MMD}(\boldsymbol{X}_i))} \sum_{i=1}^{n} w(\mathrm{MMD}(\boldsymbol{X}_i))(\boldsymbol{X}_i - \tilde{\mathbf{M}})'\tilde{\boldsymbol{\Omega}}^{\mathrm{row}}(\boldsymbol{X}_i - \tilde{\mathbf{M}}), \tag{3.3.8}$$

where $w : [0,\infty) \to [0,\infty)$ *is a non-increasing and bounded weight function such that* $w(\mathrm{MMD}(\boldsymbol{X}_i)) > 0$ *for at least* $\lfloor (n+d+2)/2 \rfloor$ *observations that vanishes for large distances, i.e.,* $w(\mathrm{MMD}(\boldsymbol{X}_i)) = 0$ *if* $\mathrm{MMD}(\boldsymbol{X}_i) > c_1 > 0$.

The following theorem shows that the *reweighted* MMCD estimator preserves the breakdown point of the original estimator. The simulations presented in Section 3.6 illustrate substantial improvements in the efficiency of the reweighed MMCD estimators. With increasing sample size, the finite sample efficiency exceeds 90% across various selections of p and q.

Theorem 3.3.0.3. *Let* $\mathfrak{X}$ *be a collection of* n *i.i.d. samples from a continuous* $p \times q$ *matrix-variate distribution, where* $d = \lfloor p/q + q/p \rfloor$, $p, q \in \mathbb{N}, p \geq 2, q \geq 2$, *and let* $\hat{\mathbf{M}}_{\mathfrak{X}}$, $\hat{\boldsymbol{\Sigma}}^{\mathrm{row}}_{\mathfrak{X}}$, *and* $\hat{\boldsymbol{\Sigma}}^{\mathrm{col}}_{\mathfrak{X}}$ *denote the* raw *MMCD estimators as in Definition 3.2.0.1 with breakdown points*

$$\varepsilon^*(\hat{\mathbf{M}}_{\mathfrak{X}}, \mathfrak{X}) = \varepsilon^*(\hat{\boldsymbol{\Sigma}}^{\mathrm{row}}_{\mathfrak{X}}, \hat{\boldsymbol{\Sigma}}^{\mathrm{col}}_{\mathfrak{X}}, \mathfrak{X}) = \frac{1}{n}\lfloor \min(n-h+1, h-(d+1)) \rfloor =: \frac{m}{n},$$

and let $\tilde{\mathbf{M}}_{\mathfrak{X}}$, $\tilde{\boldsymbol{\Sigma}}^{\mathrm{row}}_{\mathfrak{X}}$, *and* $\tilde{\boldsymbol{\Sigma}}^{\mathrm{col}}_{\mathfrak{X}}$ *denote the* reweighted *estimators as in Definition 3.3.0.1. Then,*

$$\varepsilon^*(\tilde{\mathbf{M}}_{\mathfrak{X}}, \mathfrak{X}) \geq \frac{m}{n} \quad \text{and} \quad \varepsilon^*(\tilde{\boldsymbol{\Sigma}}^{\mathrm{row}}_{\mathfrak{X}}, \tilde{\boldsymbol{\Sigma}}^{\mathrm{col}}_{\mathfrak{X}}, \mathfrak{X}) \geq \frac{m}{n}.$$

A proof is given in Supplement B.2. For the algorithm used to compute the reweighted MMCD estimators introduced in the following section, we use the weight function $w : [0,\infty) \mapsto \{0,1\}$ with

$$w(\mathrm{MMD}^2(\boldsymbol{X}_i)) := \begin{cases} 1 & \text{if} \quad i \in H \vee \mathrm{MMD}^2(\boldsymbol{X}_i) < \chi^2_{pq;0.975} \\ 0 & \text{otherwise} \end{cases}. \tag{3.3.9}$$

Note that the h observations in the h-subset of the raw MMCD estimator have the lowest MMDs, and the condition that all observations $i \in H$ get a positive weight ensures that the reweighting step does not lead to an estimator that uses fewer than h samples.

3.4 Algorithm

Rousseeuw and Van Driessen (1999) proposed the Fast-MCD algorithm to efficiently compute the MCD estimator. The key idea to find the h-subset with the lowest covariance determinant is based on the concentration step (C-step): after each C-step, the objective function is smaller or equal as before, and by repeatedly applying C-steps convergence is reached within finitely many iterations.

3.4.1 Adapting the C-step

Adapting the structure of the Fast-MCD algorithm to the matrix-variate setting leads to the development of the MMCD algorithm. This adaptation necessitates a modification in the covariance estimation during the C-step to derive suitable counterparts for computing the MMCD estimators. However, this process encounters a challenge due to the involvement of two covariance matrices, as depicted in Equations (3.2.6) and (3.2.7), both lacking closed-form solutions for their estimation. To address this issue, we incorporate the flip-flop algorithm introduced by Dutilleul (1999) within the C-step. Consider a matrix-variate random sample $\boldsymbol{\mathcal{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, with $\boldsymbol{X}_i \in \mathbb{R}^{p\times q}$, and any h-subset $H_{\text{old}} \subset \{1, \ldots, n\}$, with $|H_{\text{old}}| = h > \lfloor p/q + q/p \rfloor + 2$. First, the MLEs $(\hat{\mathbf{M}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}})$ are computed based on the observations in the subset H_{old} using the flip-flop algorithm, which is non-decreasing in likelihood. Next, compute the squared Mahalanobis distances $d_i^2(H_{\text{old}}) := \text{MMD}^2(\boldsymbol{X}_i, \hat{\mathbf{M}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}})$ for all $i = 1, \ldots, n$. In Proposition 3.2.0.1, we showed that $\sum_{i\in H_{\text{old}}} d_i^2(H_{\text{old}}) = hpq$, hence only the terms of the log-likelihood function involving the determinants change in this step. To construct the new subset H_{new}, sort the squared MMDs in ascending order, resulting in a permutation π of $\{1, \ldots, n\}$ such that $d^2_{\pi(1)}(H_{\text{old}}) \leq \ldots \leq d^2_{\pi(n)}(H_{\text{old}})$, and define a new h-subset $H_{\text{new}} = \{\pi(1), \ldots, \pi(h)\}$. Since the estimators do not change in this step, the terms involving the determinant are constant, and by construction, the sum of the Mahalanobis distances either decreases or stays constant. Hence, the reordering is non-decreasing in likelihood. Finally, the estimators are updated using the flip-flop algorithm based on the observations in the subset H_{new}, resulting in estimators $(\hat{\mathbf{M}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}})$, increasing the likelihood once more, and it follows that

$$p \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}})) + q \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}})) \leq p \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}})) + q \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}})). \tag{3.4.1}$$

By repeatedly applying such C-steps, we can decrease the covariance determinant in subsequent iterations as in Equation (3.4.1). This results in a decreasing and non-negative sequence of determinants that must converge after exploring finitely many h-subsets. Similar to the multivariate case, we obtain equality of the determinants from one h-subset to the next if and only if the estimators do not change from one to the next iteration. However, this does not necessarily imply that we have found a global optimum. A pseudo-code for this matrix-variate version of the C-step is given in Algorithm 3 in Supplement B.3.

3.4.2 The MMCD Algorithm

The MMCD algorithm is a matrix-variate extension of the Fast-MCD procedure of Rousseeuw and Van Driessen (1999), aiming to alleviate the C-steps dependence on the initial subset by using multiple initial subsets, iteratively conducting C-steps on each until convergence, and ultimately selecting the solution with the lowest determinant. While this explains the idea of the algorithm, there are more computational considerations and adjustments in the full MMCD algorithm. A pseudo-code of the MMCD Algorithm 4 is given in Supplement B.3.

As in its multivariate counterpart, the MMCD procedure uses so-called *elemental* subsets to initialize the procedure. This means that we use m subsets of size $d + 2, d = \lfloor p/q + q/p \rfloor$, instead of size h, to increase the probability of obtaining at least one clean initial subset. Using $m = 500$ elemental subsets by default allows for a reasonable tradeoff between a

wide variety of settings where we likely obtain at least one clean elemental subset and the computational demands of computing initial estimators. If either $p \ll q$ or $q \gg p$, d will be large, and using more initial subsets is recommended. Using elemental subsets increases not only the robustness of the initial estimators but also the computational efficiency.

Moreover, the MMCD procedure only uses 2 C-step and MLE iterations for the initial elemental subsets to ensure even faster computation of the initial estimators. In the MLE procedure, Werner et al. (2008) demonstrated that the same asymptotic efficiency can be attained using only two iterations instead of iterating until convergence. As for the C-step, Rousseeuw and Van Driessen (1999) outlined that after two iterations, subsets with the lowest covariance determinant during the procedure can already be identified, even before reaching convergence. Moreover, simulations show that we can identify those initial subsets that yield robust solutions after 2 C-steps, whether we use 2 MLE iterations or iterate the flip-flop algorithm until convergence. This is described in detail in Supplement B.3, where we also show that elemental subsets indeed yield more robust solutions than their larger counterparts in case of high contamination.

The initialization step of the MMCD procedure yields m initial estimators, and we keep the 10 estimators with the lowest covariance determinant. Using those as initial estimators, we iterate C-steps until convergence on the complete dataset $\mathfrak{X}$. The solution with the lowest covariance determinant then yields the raw MMCD estimators.

The raw MMCD estimators are scaled using the consistency factor $c(\alpha)$ given in Equation (3.3.5) to achieve consistency at the normal model as outlined in Theorem 3.3.0.2. Based on those rescaled raw MMCD estimators, the reweighted estimators described in Definition 3.3.0.1 are computed using the weights given in Equation (3.3.9). The reweighted MMCD estimators are then scaled using $c(\tilde{\alpha}) = c(\tilde{h}/n)$, where $\tilde{h}$ denotes the number of observations with weights one.

The MMCD algorithm repeatedly computes Mahalanbois distances for all n samples, which is computationally expensive when n gets large. To improve the computational efficiency for settings where n is large, we implemented the subsampling approach proposed by Rousseeuw and Van Driessen (1999). The idea is to split the sample of n observations into several smaller subsamples and compute initial estimators on those subsamples before working on the large set with n observations.

3.5 Outlier Detection and Explainability

Given a sample $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ of matrix-variate observations, the task for outlier detection is to identify those observations which are "far away" from the center of the data cloud with respect to its shape. In robust statistics, it is common to consider the Mahalanobis distance for this purpose, assume an underlying normal distribution of the observations, and use a quantile of the chi-square distribution as an outlier cutoff value (Maronna et al., 2019). Here, we follow the same idea: an observation $\boldsymbol{X}_i$ is flagged as an outlier if

$$\mathrm{MMD}^2(\boldsymbol{X}_i; \hat{\mathbf{M}}, \hat{\boldsymbol{\Sigma}}^{\mathrm{row}}, \hat{\boldsymbol{\Sigma}}^{\mathrm{col}}) > \chi^2_{pq;0.975} \ ,$$

for $i \in \{1, \ldots, n\}$ and the MMCD estimators $\hat{\mathbf{M}}$, $\hat{\boldsymbol{\Sigma}}^{\mathrm{row}}$, and $\hat{\boldsymbol{\Sigma}}^{\mathrm{col}}$.

Even though this information is valuable in practice, it is not very useful for understanding the reasons for outlyingness. This is the goal of outlier explainability, where the contributions of the cells/rows/columns of the matrix-valued observations are investigated in more detail. We will use the concept of Shapley values for this purpose and first briefly review how this is applied to multivariate data before extending it to the matrix-variate case. For details, we refer to Mayrhofer and Filzmoser (2023).

3.5.1 Shapley Values for Multivariate Data

Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$ denote an observation vector from a population with expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance matrix $\boldsymbol{\Sigma}$, and $P = \{1, \ldots, p\}$ the index set of the variables. Then the outlyingness contributions $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_p(\boldsymbol{x}))$ based on the Shapley value assign each variable its average marginal contribution to the squared Mahalanobis distance, i.e.,

$$\phi_k(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{S \subseteq P \setminus \{k\}} \frac{|S|!(p - |S| - 1)!}{p!} \Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S) = (x_k - \mu_k) \sum_{j=1}^{p} (x_j - \mu_j)\omega_{jk}, \quad (3.5.1)$$

with marginal contributions

$$\Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S) := \operatorname{MD}^2(\hat{\boldsymbol{x}}^{S \cup \{k\}}) - \operatorname{MD}^2(\hat{\boldsymbol{x}}^S) \quad \text{and} \quad \hat{x}_j^S := \begin{cases} x_j & \text{if } j \in S \\ \mu_j & \text{if } j \notin S \end{cases} \quad (3.5.2)$$

as the components of $\hat{\boldsymbol{x}}^S$. Here, ω_{jk} is the element (j, k) of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Since $\boldsymbol{\phi}(\boldsymbol{x})$ is based on the Shapley value, it is the only decomposition of the squared Mahalanobis distance based on Equation (3.5.2) that fulfills the following properties:

- *Efficiency:* The contributions $\phi_j(\boldsymbol{x})$, for $j = 1, \ldots, p$, sum up to the squared Mahalanobis distance of $\boldsymbol{x}$, hence $\sum_{j=1}^{p} \phi_j(\boldsymbol{x}) = \operatorname{MD}^2(\boldsymbol{x})$.
- *Symmetry:* If $\operatorname{MD}^2(\hat{\boldsymbol{x}}^{S \cup \{j\}}) = \operatorname{MD}^2(\hat{\boldsymbol{x}}^{S \cup \{k\}})$ holds for all subsets $S \subseteq P \setminus \{j, k\}$ for two coordinates j and k, then $\phi_j(\boldsymbol{x}) = \phi_k(\boldsymbol{x})$.
- *Monotonicity:* Let $\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}} \in \mathbb{R}^p$ be two vectors and $\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}} \in \operatorname{PDS}(p)$ be two matrices. If
$$\operatorname{MD}^2_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^{S \cup \{j\}}) - \operatorname{MD}^2_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^S) \geq \operatorname{MD}^2_{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{x}}^{S \cup \{j\}}) - \operatorname{MD}^2_{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{x}}^S)$$
holds for all subsets $S \subseteq P$, then $\phi_j(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \geq \phi_j(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$.

In words, the coordinate $\phi_k(\boldsymbol{x})$ of the Shapley value is the average marginal contribution of the k-th variable to the squared Mahalanobis distance and is obtained by averaging over all marginal outlyingness contributions $\Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S)$ across all possible subsets $S \subseteq P \setminus \{k\}$. Although this suggests an exponential computational complexity, which becomes costly, especially if p is large, the second equality in Equation (3.5.1) reveals just linear complexity; for a proof we refer to Mayrhofer and Filzmoser (2023). Equation (3.5.1) allows for another insight into the Shapley value by comparing it to the squared Mahalanobis distance, which can be written as $\sum_{j,k=1}^{p} (x_j - \mu_j)(x_k - \mu_k)\omega_{jk}$. While the latter calculates an outlyingness measure by aggregating the contributions $(x_j - \mu_j)(x_k - \mu_k)\omega_{jk}$ of all variables for the entire observation, Equation (3.5.1) shows that a coordinate $\phi_k(\boldsymbol{x})$ of the Shapley value only considers the contributions that are associated with the k-th variable.

3.5.2 Shapley Balue for Matrix-valued Data

To define Shapley values for matrix-variate data, we can use the connection between the matrix and multivariate Mahalanobis distance; see Equation (3.2.2). Let $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ be a matrix-variate sample with mean $\mathbf{M} \in \mathbb{R}^{p\times q}$ and covariance matrices $\boldsymbol{\Sigma}^{\text{row}} \in \text{PDS}(p)$ and $\boldsymbol{\Sigma}^{\text{col}} \in \text{PDS}(q)$. The pq-dimensional vectorized observation is denoted as $\boldsymbol{x} = \text{vec}(\boldsymbol{X})$, with mean $\boldsymbol{\mu} = \text{vec}(\mathbf{M})$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}$. Based on Equation (3.5.1), we can obtain outlyingness contributions for every coordinate of $\boldsymbol{x}$ and hence for every cell of the matrix $\boldsymbol{X}$ by

$$\phi_a(\boldsymbol{x}) = (x_a - \mu_a)\sum_{b=1}^{pq}(x_b - \mu_b)\omega_{ab} = (x_{jk} - m_{jk})\sum_{i=1}^{p}\sum_{l=1}^{q}(x_{il} - m_{il})\omega_{ij}^{\text{row}}\omega_{kl}^{\text{col}} = \phi_{jk}(\boldsymbol{X}),$$

with $a = i+(l-1)p$ and $b = j+(k-1)p$, and for $j = 1,\dots,p$ and $k = 1,\dots,q$. Using matrix operations, we can efficiently compute the $p \times q$ matrix containing the cellwise Shapley values $\phi_{jk}(\boldsymbol{X})$ as

$$\boldsymbol{\Phi}(\boldsymbol{X}) = (\boldsymbol{X} - \mathbf{M}) \circ \boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X} - \mathbf{M})\boldsymbol{\Omega}^{\text{col}} \in \mathbb{R}^{p\times q}, \tag{3.5.3}$$

where $\circ$ refers to element-wise multiplication.

Next, we discuss how matrix affine transformations as in Equation (3.3.2) affect the cellwise Shapley values for matrix-variate data.

Proposition 3.5.2.1. *Let $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ be a sample from $\mathcal{ME}(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}, g)$, $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, $\boldsymbol{B} \in \mathbb{R}^{q\times q}$, $\boldsymbol{A}, \boldsymbol{B}$ invertible, and $\boldsymbol{C} \in \mathbb{R}^{p\times q}$. Then, the cellwise Shapley values are* not *matrix affine equivariant, i.e., $\boldsymbol{\Phi}(\boldsymbol{AXB}) \neq \boldsymbol{A\Phi X})\boldsymbol{B}$ for general positive definite $\boldsymbol{A}$ and $\boldsymbol{B}$. However, they are*

(a) shift invariant, i.e., $\boldsymbol{\Phi}(\boldsymbol{X} + \boldsymbol{C}) = \boldsymbol{\Phi}(\boldsymbol{X})$,

(b) scale invariant, i.e., if $\boldsymbol{A}$ and $\boldsymbol{B}$ are scaling matrices, thus diagonal matrices with non-zero entries, then $\boldsymbol{\Phi}(\boldsymbol{AXB}) = \boldsymbol{\Phi}(\boldsymbol{X})$,

(c) permutation equivariant, i.e., if $\boldsymbol{A}$ and $\boldsymbol{B}$ are permutation matrices, then $\boldsymbol{\Phi}(\boldsymbol{AXB}) = \boldsymbol{A\Phi}(\boldsymbol{X})\boldsymbol{B}$, and

The proofs are given in Supplement B.4. When considering gray-scale image data, shifting or rescaling the gray-scale information would not change the cellwise Shapley values. Further, exchanging rows and columns of the image; in particular mirroring or rotating the image by 90°, would equivalently transform the Shapley values. Similarly to the setting of cellwise outliers (Alqallaf et al., 2009), cellwise Shapley values are tied to the original coordinate system and are not matrix affine equivariant.

It can be preferable in some applications to obtain outlyingness explanations for a complete row or column of the matrix-valued observations, especially when we want to compare multiple observations. In the following, we show how Shapley values for rows can be obtained; Shapley values for columns can be computed based on the transposed matrix or by adapting the following notation accordingly for columns.

Consider again the set $P = \{1, \dots, p\}$, and $S \subseteq P \setminus \{j\}$. The rowwise marginal contributions to the matrix Mahalanobis distance are defined as as

$$\Delta_j \operatorname{MMD}(\hat{\boldsymbol{X}}^S) := \operatorname{MMD}(\hat{\boldsymbol{X}}^{S\cup\{j\}}) - \operatorname{MMD}(\hat{\boldsymbol{X}}^S),$$

where the i-th row of $\hat{\boldsymbol{X}}^S$ is given as $(x_{i1}, \dots, x_{iq})$ if $i \in S$ and $(m_{i1}, \dots, m_{iq})$ if $i \notin S$.

Proposition 3.5.2.2. *The j-th coordinate of the rowwise Shapley value is given by*

$$\phi_{j\cdot}(\boldsymbol{X}) := \sum_{S\subseteq P\setminus\{j\}} \frac{|S|!(p-|S|-1)!}{p!} \Delta_j \operatorname{MMD}(\hat{\boldsymbol{X}}^S) \tag{3.5.4}$$

$$= \sum_{i=1}^{p}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{jl}-m_{jl})(x_{ik}-m_{ik})\omega_{ij}^{\text{row}}\omega_{kl}^{\text{col}} = \sum_{k=1}^{q}\phi_{jk}(\boldsymbol{X}), \tag{3.5.5}$$

A proof for Equation (3.5.5) can be found in Supplement B.4. Thus, a rowwise Shapley value is obtained by summing up the cellwise Shapley values for the corresponding row, which is equivalent to adapting the marginal contributions to a rowwise replacement. The vectors containing the rowwise or columnwise Shapley values can also be computed by

$$\boldsymbol{\phi}_{\text{row}}(\boldsymbol{X}) = \operatorname{diag}(\boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X}-\mathbf{M})\boldsymbol{\Omega}^{\text{col}}(\boldsymbol{X}-\mathbf{M})') \in \mathbb{R}^p \quad \text{and} \tag{3.5.6}$$

$$\boldsymbol{\phi}_{\text{col}}(\boldsymbol{X}) = \operatorname{diag}((\boldsymbol{X}-\mathbf{M})'\boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X}-\mathbf{M})\boldsymbol{\Omega}^{\text{col}}) \in \mathbb{R}^q, \tag{3.5.7}$$

respectively. The properties listed in Proposition 3.5.2.1 also apply in this setting.

3.6 Simulations

In the simulation studies outlined in this section, our primary focus is to rigorously assess the performance of the MMCD estimators. We aim to validate their demonstrated theoretical properties and compare efficiency against ML estimators. Despite our initial intention to include various robust estimators as mentioned in Section 3.1, practical constraints arose as the relevant routines were exclusively accessible in `Matlab`, while our current framework operates within `R`. For the sake of consistency and practical implementation, we concentrate on comparing the efficiency of the raw and reweighted MMCD alongside an in-depth analysis of the MLEs, (reweighted) MMCD estimators, and MCD estimator based on the vectorized samples on contaminated data. To ensure the highest possible breakdown point across all simulations and examples discussed in this paper, we set $h = \lfloor (n+d+2)/2 \rfloor$ for the MMCD estimators and $h = \lfloor (n+pq+1)/2 \rfloor$ for the MCD estimator. We conduct 100 repetitions for each simulation setting and visualize the results through either line plots or boxplots. In the line plots, the solid lines represent average scores, while the shaded areas depict the one standard error regions.

Finite-sample efficiency. To analyze the finite-sample efficiency, we generate samples from a centered matrix normal distribution with dimensions $(p, q) \in \{(5, 20), (50, 20), (100, 50)\}$ for various sample sizes $n \in \{20, 50, 100, 300, 1000\}$. For the rowwise covariance matrix we adopt the covariance matrices proposed by Agostinelli et al. (2015b), denoted $\boldsymbol{\Sigma}^{\text{row}} =$

$\mathbf{\Sigma}^{\text{rnd}} \in \text{PDS}(p)$, which have random entries and generally yield low correlations. For the columnwise covariance, we use $\mathbf{\Sigma}^{\text{col}} = \mathbf{\Sigma}^{\text{mix}}(0.7) \in \text{PDS}(q)$, with entries $\sigma_{jk}^{\text{mix}}(0.7) = 0.7^{|j-k|}$. We assess the normal finite-sample efficiency by comparing the ratio

$$\frac{D(\hat{\mathbf{\Sigma}}^{\text{row}}_{\text{MLE}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\text{MLE}})}{D(\hat{\mathbf{\Sigma}}^{\text{row}}_{\text{MMCD}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\text{MMCD}})},$$

where $D(\hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}})$ denotes the Kullback-Leiber (KL) divergence of the estimators $\hat{\mathbf{\Sigma}}^{\text{row}}$ and $\hat{\mathbf{\Sigma}}^{\text{col}}$ in the matrix normal setting $\mathcal{MN}(\mathbf{M}, \mathbf{\Sigma}^{\text{row}}, \mathbf{\Sigma}^{\text{col}})$, which is given by

$$\begin{aligned} D(\hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}) = &\operatorname{tr}(\mathbf{\Omega}^{\text{row}}\hat{\mathbf{\Sigma}}^{\text{row}}) \operatorname{tr}(\mathbf{\Omega}^{\text{col}}\hat{\mathbf{\Sigma}}^{\text{col}}) \\ &- q \log(\det(\mathbf{\Omega}^{\text{row}}\hat{\mathbf{\Sigma}}^{\text{row}})) - p \log(\det(\mathbf{\Omega}^{\text{col}}\hat{\mathbf{\Sigma}}^{\text{col}})) - pq, \end{aligned} \tag{3.6.1}$$

with $\mathbf{\Omega}^{\text{row}} = (\mathbf{\Sigma}^{\text{row}})^{-1}$ and $\mathbf{\Omega}^{\text{col}} = (\mathbf{\Sigma}^{\text{col}})^{-1}$. As shown in Figure 3.6.1, the efficiency of the raw MMCD estimators is below 0.5 on average. In contrast, the reweighted estimators' efficiency is above 0.5 for $n = 100$ and it rises to over 0.9 as the sample size increases.

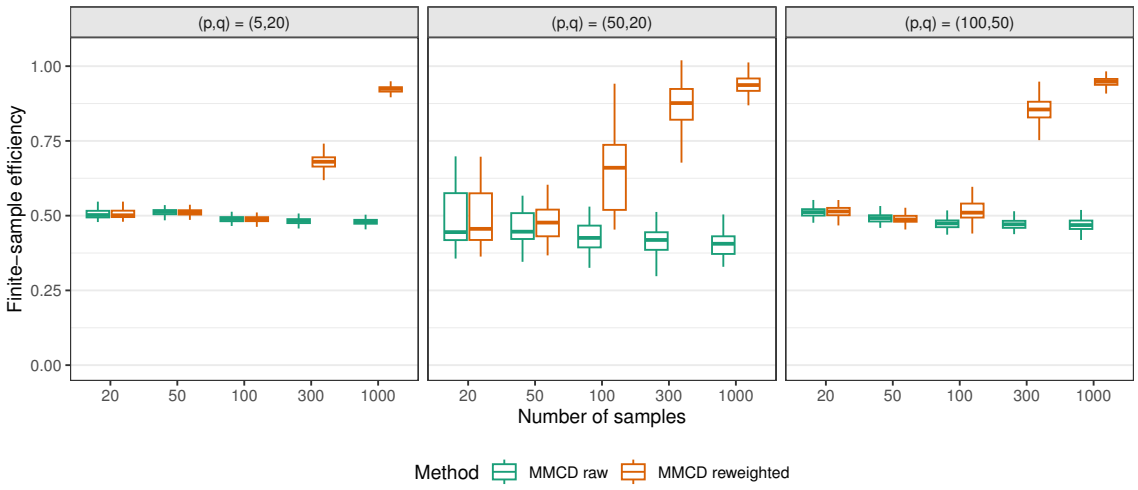


Figure 3.6.1: Comparison of the finite-sample efficiency of raw and reweighted MMCD.

Robustness and matrix size. For the setting with contamination, we consider matrix-variate samples with $p \in \{2, \ldots, 30\}$ rows and $q = \{10, 20, 30\}$ columns for sample sizes $n \in \{100, 1000\}$. The clean data are generated from a centered matrix normal distribution with $\mathbf{\Sigma}^{\text{row}} = \mathbf{\Sigma}^{\text{rnd}}$ and $\mathbf{\Sigma}^{\text{col}} = \mathbf{\Sigma}^{\text{mix}}(0.7)$. A fraction, $\varepsilon = 0.1$, of the clean data is replaced by outliers, sampled from a matrix normal distribution with a mean matrix where all entries are equal to $\gamma = 1$. The covariance matrices of the outliers are the same as for the regular observations.

We use KL divergence (3.6.1) to analyze the quality of the covariance estimation. Additionally, we analyze outlier detection capabilities of the squared Mahalanobis distance based on the estimators, with the $\chi^2_{pq,0.99}$ quantile as a detection threshold. We also include the

Mahalanobis distances based on true parameters used to generate the data as a benchmark and measure performance by precision and recall. Due to the excessively long computation times of the Fast-MCD procedure in higher-dimensional scenarios, we used the deterministic MCD (Hubert et al., 2012) when $pq > 300$. Since the MCD estimator requires $n > pq$, it is only computed for those settings.

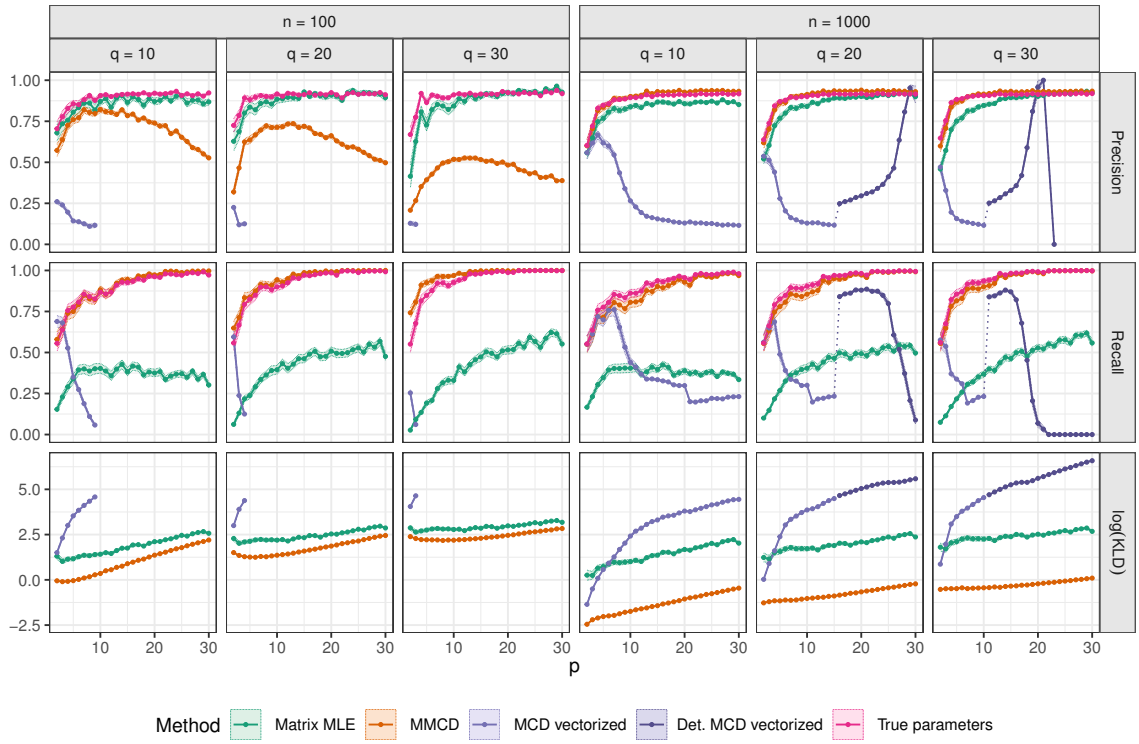


Figure 3.6.2: Comparison of precision, recall, and KL divergence for ML and MMCD estimators, (deterministic) MCD estimators with vectorized data, and true parameters as a benchmark for outlier detection for simulated data from a matrix normal distribution with 10% contamination.

Figure 3.6.2 shows that the MMCD estimators have lower KL divergence than the competing methods and attain a recall similar to the benchmark approach based on the true parameters used to generate the data across all settings. The precision of the MMCD estimators depends on the dimensionality of the matrix-variate samples as well as on the sample size. For $n = 100$, the precision decreases with increasing dimensionality pq, but the effect is mitigated by an improving performance when $\max\{p/q, q/p\}$ is small. For $n = 1000$, the precision is close to the precision based on the true parameters. This suggests that for small sample sizes, a correction similar to the one proposed by Pison et al. (2002) for the MCD could lead to a better performance. In the matrix-variate setting, such a correction would not only be dependent on pq and n but also on p/q and q/p.

For small p and q, the comparison between the MMCD estimators and the MCD for the vectorized observations is of special interest. For $n = 1000$ and $q = 10$ they have a similar

recall when $p \leq 6$, and for $q \in \{20, 30\}$ the MCD estimators show substantial improvements when the deterministic MCD approach is used instead of the Fast-MCD. This can be explained by the dependence of the Fast-MCD on the robustness of the initial solutions, and with an increasing pq, the probability of obtaining a clean subset becomes very small. The MCD estimator shows a steep drop in precision as p increases when Fast-MCD is used. For the deterministic MCD, we see a trade-off between precision and recall with increasing dimensionality, but the KL divergence remains high. With increasing dimensionality, even the nonrobust matrix MLEs outperform the MCD estimator which highlights the importance of respecting the inherent data structure of matrix-variate observations.

Robustness and contamination type. In addition to the shift outliers we also consider block and cell contamination for matrix normal samples of size $(p, q) = (5, 20)$. In all three settings, we consider a fraction of $\varepsilon = 0.1$ contaminated samples. Let $\boldsymbol{X} = (x_{jk}), j = 1, \ldots, 5, k = 1, \ldots, 20$, denote a sample from a centered matrix normal distribution with rowwise covariance $\boldsymbol{\Sigma}^{\text{row}} = \boldsymbol{\Sigma}^{\text{rnd}}$ and columnwise covariance $\boldsymbol{\Sigma}^{\text{col}} = \boldsymbol{\Sigma}^{\text{mix}}(0.7)$. For block contamination, we replaced the top left 2×5 block, corresponding to the entries $x_{jk}, j = 1, 2, k = 1, \ldots, 5$, with entries from a shifted matrix normal distribution with a mean matrix where all entries are equal to $\gamma = 1$ and covariance matrices corresponding to the top left block of $\boldsymbol{\Sigma}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}}$. For cell contamination, a fraction of 0.1 of the cells of the outlying observations are randomly permuted. The shift outliers are generated with a mean shift $\gamma = 1$ as before.

Figure 3.6.3 shows that the MMCD estimators are better suited for outlier detection and yield more robust covariance estimates than the matrix MLEs as well as the MCD estimator on the vectorized observations. Overall, the results are similar across all three simulation scenarios, only for mean shift contamination we see higher variation than in the other two settings. This is likely because the block and cell contamination interfere with covariance estimation more profoundly, i.e., the KL divergence of the matrix MLEs is highest for block contamination followed by cell and shift contamination.

The supplementary materials B.5 provide in-depth simulation studies that expand upon the scenarios discussed in this section. These simulations analyze the effects of the level of contamination and mean shifts for multiple types of covariance matrices. Additionally, we extend our analysis beyond the normal model to include samples generated from a matrix t-distribution, examining performance across a range of degrees of freedom. For this scenario, we also compute the ML estimators for the matrix t-distribution (Thompson et al., 2020). We include a summary of computation time and consider additional performance metrics, such as the F-score (harmonic mean of precision and recall), Frobenius error, and the angle between eigenvalues of covariance matrices.

3.7 Examples

3.7.1 Glacier Weather Data – Sonnblick Observatory

We analyze the publicly available weather data from Austria's highest weather station, located in the Austrian Central Alps at an elevation of 3106 m above sea level on top of the glaciated mountain "Hoher Sonnblick" (datasource: GeoSphere Austria - `https://data.hub.geosphere.at`). The observed parameters are monthly averages of temperature

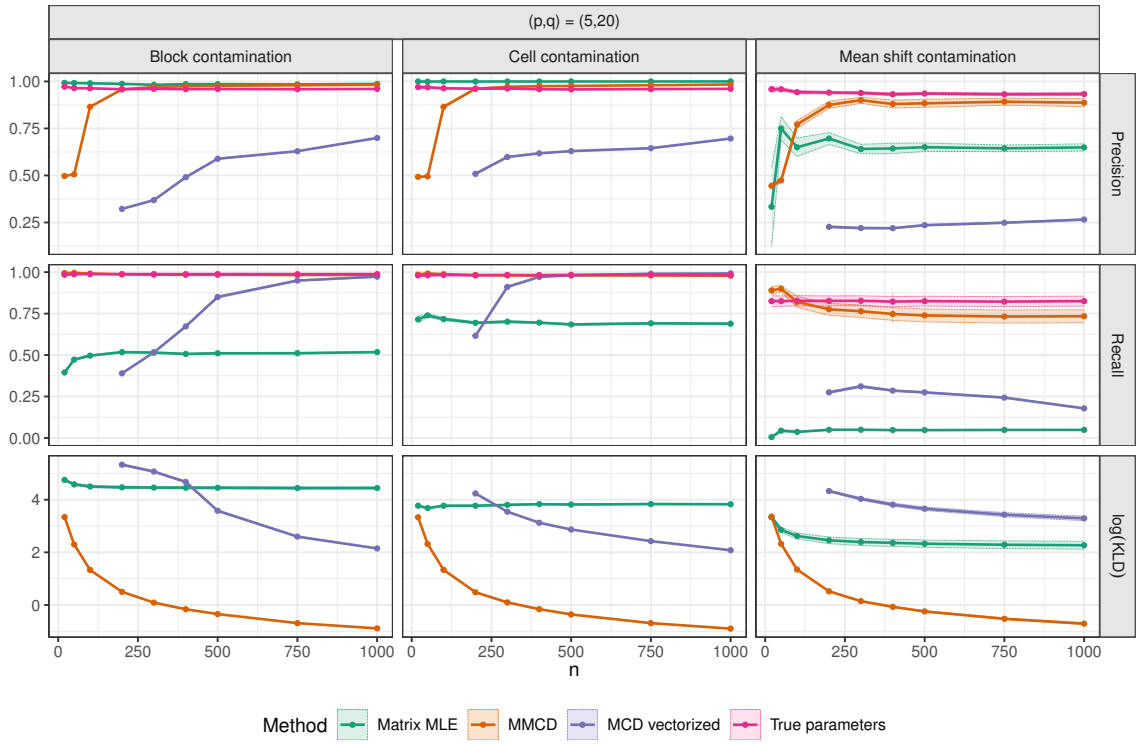


Figure 3.6.3: Precicion, recall, and logarithm of KL divergence comparing block, cell, and sample contamination.

(T), precipitation (P), proportion of solid precipitation (SP), air pressure (AP), and sunshine hours (SH). We consider the monthly values between 1891 and 2022 and exclude five years with missing values, yielding $n = 127$ observations of $p = 5$ times $q = 12$ dimensional matrices. Our goal is to identify observations that show a different weather pattern than the majority of the data and explain why the corresponding years deviate from the majority. We did not adjust for a possible yearly trend in this exploratory analysis as we wish to understand long-term patterns and shifts in climate without the influence of adjustments.

In total, outlier detection based on the MMCD estimators flags 23 outlying matrices, which are indicated in Figure 3.7.1 as colored years: If the aggregated monthly measurements are above their average, the cells are colored red; otherwise, they are colored blue. The rowwise Shapley value is then used to determine color brightness, i.e., the larger the outlyingness contribution, the darker the color. Years with missing observations are grey; years with only white cells refer to regular observations. It is visible that the outlier frequency increases in the last period. Moreover, more recent outliers are characterized by increased temperature, precipitation, air pressure, and a lack of solid precipitation (e.g. snow) – a clear signal of a climate change.

In Figure 3.7.2, we use cellwise Shapley values to understand which parameters in which months contributed most to the outlyingness of 1895 and 2022, corresponding to the first and last outlying observation in the dataset, respectively, where the color scheme is inherited

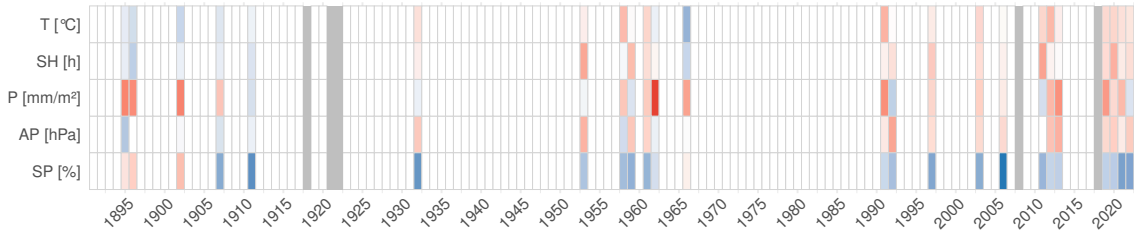


Figure 3.7.1: Yearly outlyingness contributions for the glacier weather data. Regular years are white, and years that contain missing data are gray. Outliers are colored as follows: blue for "above average", red for "below average", and color intensity proportional to rowwise Shapley value.

Figure 3.7.2: Outlyingess contributions based on cellwise Shapley values for the years 1895 and 2022 of the glacier weather data using the same color scheme as in Figure 3.7.1.

from Figure 3.7.1. The largest outlyingness contribution is due to an unusually large amount of precipitation in March 1895. Overall, high amounts of precipitation were observed that year, with a high percentage of snow even in the summer months. In contrast, the largest outlyingness contributions in 2022 are due to a very sunny March and low percentages of snowfall in May, June, and August.

3.7.2 Darwin Data

We consider the DARWIN (Diagnosis AlzheimeR WIth haNdwriting) (Cilia et al., 2022) dataset containing handwriting samples of 174 subjects, 89 diagnosed with Alzheimer's disease (AD), and 85 healthy subjects (H). Each individual completed 25 handwriting tasks on paper, and the pen movements were recorded using a graphic tablet. The tasks are ordered in difficulty. From the raw handwriting data, 18 features were extracted: Total Time, Air Time, Paper Time, Mean Speed on paper, Mean Speed in air, Mean Acceleration on paper, Mean Acceleration in air, Mean Jerk on paper, Mean Jerk in air, Pressure Mean, Pressure Variance, Generalization of the Mean Relative Tremor (GMRT) on paper, GMTR in air, Mean GMRT, Pendowns Number, Max X Extension, Max Y Extension, and Dispersion

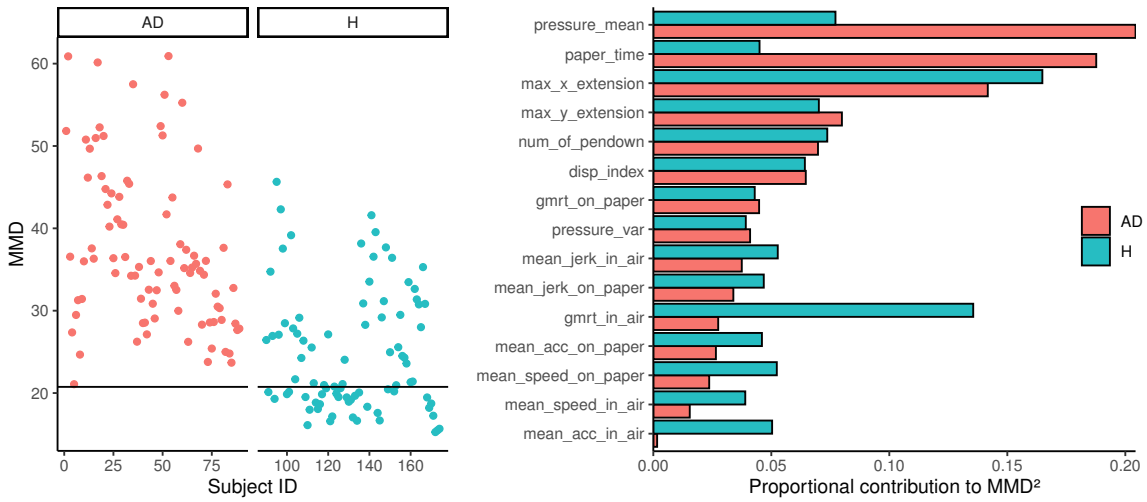


Figure 3.7.3: Plot of robust MMD based on MMCD estimators for the Darwin data on the left, and average proportional rowwise Shapley values for the H and AD subjects on the right.

Index. For a more detailed description of the data, we refer to Cilia et al. (2018). In Cilia et al. (2022), each task was considered separately to train a classifier, and the combination of the classifiers led to an improvement in the classification of subjects. Our focus here lies not in the classification task but rather in explaining the differences between AD and H groups. We treat the observations as matrices, with the rows representing the extracted features and the columns representing the tasks. Because of linear dependencies, the variables Total Time and Mean GMRT were excluded. Further, the variable Air Time had several extreme and unreliable measurements and was thus also excluded. This yields observation matrices with $p = 15$ features and $q = 25$ tasks.

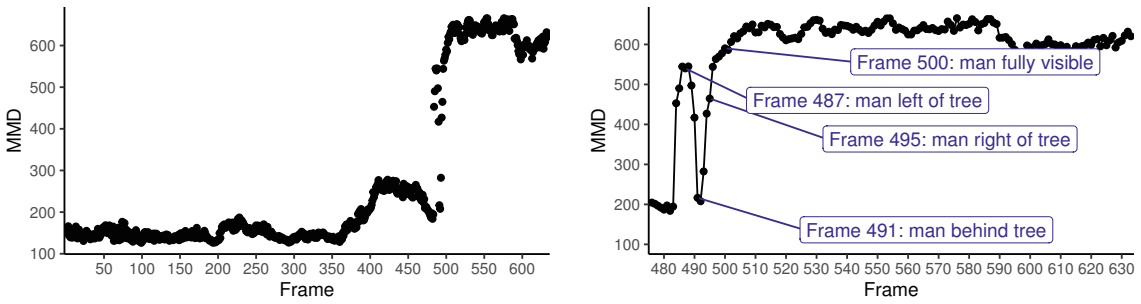
We applied the MMCD procedure only on the healthy subjects and used the robust estimators to compute MMDs for all observations. Thus, the MMDs presented in Figure 3.7.3 left are generally smaller for the H group, whereas all observations from the AD group exceed the outlier cutoff value. The fact that healthy subjects also exceed the cutoff value shows the heterogeneity in this group. In the right panel of Figure 3.7.3, we consider the *average proportional contributions* of the variables to the MMDs for the H and AD groups. The outlyingness contributions are based on the rowwise Shapley values, resulting in 15 scores for each individual. Since those scores sum up to the squared MMD, we can divide them by the squared MMD to get proportional contributions, and by averaging over all individuals in the H and AD groups, respectively, we obtain the values shown in this plot. Large differences between the AD and H groups indicate variables that are important to distinguish between healthy individuals and those who have Alzheimer's disease. For example, Pressure Mean and Paper Time are evidently higher in the AD group.

3.7.3 Video Data

In this example, we examine a surveillance video of a beach sourced from Li et al. (2004). The video comprises 633 frames, each sized at 128×160 pixels; five selected frames are shown in Figure 3.7.4. The majority of the frames depict the beach scene. Around frame 500, a man walks into the scene from the left and partly disappears behind the tree. As he continues walking, he reappears on the right side of the tree and remains in the video until the end.

Figure 3.7.4: Selected frames of the video data.

For our analyis, we converted the original RGB video to a grayscale video, applied the MMCD procedure, and obtained MMDs for all 633 frames, which are visualized in Figure 3.7.5. The plot on the left shows the robust MMDs for all 633 frames, and the one on the right for frames 471 to 633 to better highlight the increase in MMD when the man enters the scenery, with a short drop in MMD when he disappears behind the tree. We indicate frames 487, 491, and 495, also presented in Figure 3.7.6 in terms of their cellwise Shapley values. We see that the pixels that form the contours of the man and most of the pixels of the man's head contribute most to the outlyingness. When the man disappears behind the tree, there are fewer pixels with high outlyingness contributions. Since the sum of the contributions amounts to the squared MMD of an observation, this explains the behavior of the MMDs of the frames shown in Figure 3.7.5b. It is interesting to see a certain increase in



(a) Frames 1 to 633.

(b) Frames 471 to 633.

Figure 3.7.5: Plot of robust MMD based on MMCD estimators for the video data.

the MMD in Figure 3.7.5a between frames 400 and 450. Here, the Shapley values on the contour of the palm tree contribute the most to the outlyingness. This could be caused by a slight shifting of the camera or a small movement of the palm tree due to wind.

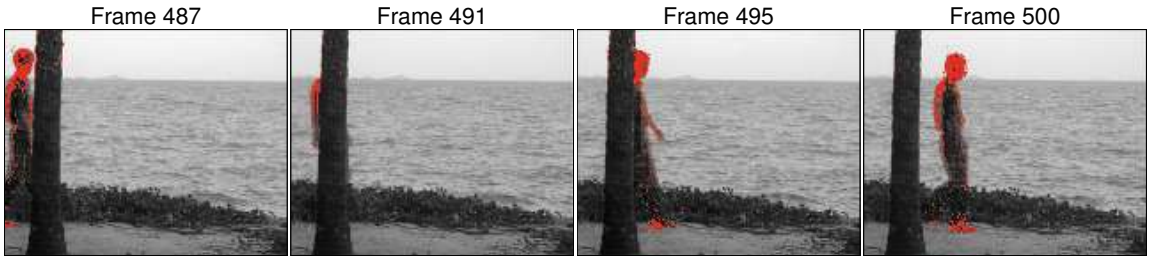


Figure 3.7.6: Outlyingness scores based on cellwise Shapley values are shown in red, where darker colors indicate higher outlyingness contributions, and the grayscale video frames are displayed in the background.

3.8 Summary and Conclusions

Matrix-valued observations, like images or dual-factor data tables, are common in various fields. To apply multivariate methods on matrix-valued data, the matrices are typically converted to vectors by stacking either the rows or columns. This disrupts the inherent data structure and increases dimensionality, thereby complicating parameter estimation. Thus, it is often preferable to model matrix-valued data directly with matrix-variate distributions. In this setting, Maximum Likelihood (ML) estimation methods exist for estimating the mean, as well as the row and column covariances, respectively. However, these estimators are sensitive to deviations caused by outliers among matrix-valued observations.

This work introduced the MMCD (matrix minimum covariance determinant) estimators as a robust counterpart to the ML estimators in the matrix-variate normal model. Several desirable properties are achieved: equivariance under matrix affine transformations, high breakdown point, and consistency under elliptical matrix-variate distributions. The proposed reweighted versions lead to higher efficiency but not to any loss in terms of breakdown point. An algorithm along the lines of the Fast-MCD procedure (Rousseeuw and Van Driessen, 1999) allows for efficient computation of the estimators. Simulation experiments validate the theoretical properties and advantages. Depending on the ratio of the number of rows and columns of the matrix-valued observations, the MMCD estimators show a big advantage over robust estimation for vectorized observations regarding breakdown and computational efficiency.

We further extended the outlier explanation concept based on Shapley values (Mayrhofer and Filzmoser, 2023) to the matrix-variate setting. This allows for an additive decomposition of the matrix-variate Mahalanobis distance of an observation into Shapley contributions of either the rows, the columns, or the matrix cells. The resulting Shapley values greatly aid with diagnostics, particularly in revealing those cells (rows, columns) of the matrix with the most substantial contributions to the outlyingness of the observation.

The efficiency of MMCD estimators in outlier detection for large sample sizes is evident from the simulations. However, our future research aims to improve and extend these estimators. For instance, smaller sample sizes might benefit from integrating finite sample corrections proposed by Pison et al. (2002) to enhance the results. Furthermore, the iterative computation of MMCD covariance estimators, which involves inverse covariance matrices,

requires data that ensures full-rank estimates at each iteration. This requirement may be impeded for example in image data, in case certain rows or columns maintain constant pixel values across all observations. To solve this, regularization involving a linear combination of the covariance matrix with a full-rank target matrix can be used (Ledoit and Wolf, 2004), similarly to the multivariate setting (Boudt et al., 2020).

The MMCD objective can be expressed as a trimmed maximum likelihood problem, and thus, can be extended to tensor-valued data using ML estimation for the tensor normal distribution Manceur and Dutilleul (2013). The framework of Raymaekers and Rousseeuw (2023) can be used to develop a cellwise robust version of the MMCD. Our ongoing research focuses on extending the MMCD estimators and outlier explanations based on Shapley values to the field of functional data analysis. Our goal is to introduce robust estimators and enhance interpretability for multivariate functional data. In the future, we also plan to incorporate these robust estimators as plug-in estimators to robustify established multivariate methodologies in the matrix-variate domain, like principal component analysis and discriminant analysis.

Software and data availability

The `R` package `robustmatrix` includes a parallelized `C++` implementation of the MMCD algorithm and a vignette to reproduce the examples presented in this paper.

Appendix B

B.1 Preliminaries

Consider an i.i.d. sample $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \in \mathbb{R}^{n \times p \times q}$, with $\boldsymbol{X}_i \sim \mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$. Due to the factored covariance structure of matrix normal data, the rowwise and columnwise covariance matrices $\boldsymbol{\Sigma}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}}$ are only identified up to a multiplicative constant $\kappa \neq 0$, since replacing $\boldsymbol{\Sigma}^{\text{row}}$ by $\kappa\boldsymbol{\Sigma}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}}$ by $1/\kappa\boldsymbol{\Sigma}^{\text{col}}$ does not change the pdf of $\boldsymbol{X}$. While the Kronecker product $\boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}$ can be uniquely identified, the issue of trivial non-uniqueness of $\boldsymbol{\Sigma}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}}$ is commonly solved by either fixing a diagonal entry, the determinant, or the norm of either matrix (Roś et al., 2016; Soloveychik and Trushin, 2016). For simplicity, we assume that the first diagonal entry of $\boldsymbol{\Sigma}^{\text{col}}$ is set to one. This implies that the uniqueness of $\boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}$ is equivalent to the uniqueness of $\boldsymbol{\Sigma}^{\text{col}}$ and $\boldsymbol{\Sigma}^{\text{row}}$ with the identifiability constraint $\sigma_{11}^{\text{col}} = 1$. The multiplicative constant for their estimators is also chosen such that $\hat{\sigma}_{11}^{\text{col}} = 1$.

Instead of using Equations (3.1.3)-(3.1.5) for mean and covariance estimation, it is also possible to consider the vectorized samples $\boldsymbol{x}_i = \text{vec}(\boldsymbol{X}_i) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, \ldots, n$, where $\boldsymbol{\mu} = \text{vec}(\boldsymbol{M})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}$ denote the mean and covariance matrix, respectively. Then the maximum likelihood estimators for mean and covariance are given by

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})', \tag{B.1}$$

respectively. The computation of the MLEs for matrix-variate samples based on Equations (3.1.3)-(3.1.5) involves estimating $p(p+1)/2 + q(q+1)/2 + pq$ parameters instead of $pq(pq+1)/2 + pq$ parameters for the vectorized observations according to Equation (B.1). This raises the question of whether fewer than $pq + 1$ observations are sufficient for guaranteeing the existence and uniqueness of MLEs for i.i.d. samples from a matrix normal distribution. This question was investigated in several papers, such as Dutilleul (1999); Lu and Zimmerman (2005); Srivastava et al. (2008); Roś et al. (2016); Soloveychik and Trushin (2016). We rely on the latter for the most recent proof of those conditions. Note that it is not necessary to assume that the sample consists of i.i.d. observations. In fact, the i.i.d. assumption can be relaxed to allow for statistically dependent samples and it is not even necessary to require identical distribution (Soloveychik and Trushin, 2016, Remarks 2 and 6). The critical condition for existence and uniqueness is that the sample contains at least $n \geq \lfloor p/q + q/p \rfloor + 2$ observations that are not collinear. The same holds for the existence and uniqueness of the MMCD estimators, where n is replaced by h, and for properties like the breakdown point the assumptions could be relaxed only requiring that the sample is in general position, i.e., no subset of $r, 2 \leq r \leq \lfloor p/q + q/p \rfloor + 2$ samples lies on an $r-2$ dimensional subspace. However, the i.i.d. assumption is still necessary when we consider properties like consistency.

The idea of the multivariate MCD estimator is as follows: Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$ denote the i-th observation of a dataset in the multivariate setting, where $i = 1, \ldots, n$. The objective of the MCD estimator is to find the subset of h out of n observations whose sample covariance matrix has the lowest determinant, with $n/2 \leq h \leq n$ and $h > p$. In total, there are $\binom{h}{n}$ possible h-subsets, and thus, a strategy needs to be used to tackle the optimization problem efficiently. This has been done with the so-called Fast-MCD algorithm (Rousseeuw and Van Driessen, 1999), which internally sorts the observations based on their Mahalanobis distances. For an observation $\boldsymbol{x}_i$ from a population with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \mathrm{PDS}(p)$ it is given by

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

Since the Mahalanobis distance is vital for the computation of the MCD estimator, it will also be crucial in a matrix-variate extension, where it can be directly derived from the Mahalanobis distance of a vectorized matrix-variate observation $\boldsymbol{X}$ as

$$\begin{aligned}
\mathrm{MMD}^2(\boldsymbol{X}) = \mathrm{MMD}^2(\boldsymbol{X}; \mathbf{M}, \boldsymbol{\Sigma}^{\mathrm{row}}, \boldsymbol{\Sigma}^{\mathrm{col}}) &= \mathrm{MD}^2(\mathrm{vec}(\boldsymbol{X})) \\
&= \mathrm{vec}(\boldsymbol{X} - \mathbf{M})' (\boldsymbol{\Omega}^{\mathrm{col}} \otimes \boldsymbol{\Omega}^{\mathrm{row}}) \, \mathrm{vec}(\boldsymbol{X} - \mathbf{M}) \\
&= \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{q} \sum_{l=1}^{q} (x_{ik} - m_{ik})(x_{jl} - m_{jl}) \omega_{ij}^{\mathrm{row}} \omega_{kl}^{\mathrm{col}} \\
&= \mathrm{tr}(\boldsymbol{\Omega}^{\mathrm{col}} (\boldsymbol{X} - \mathbf{M})' \boldsymbol{\Omega}^{\mathrm{row}} (\boldsymbol{X} - \mathbf{M})),
\end{aligned}$$

where m_{ij}, $\omega_{ij}^{\mathrm{row}}$ and $\omega_{ij}^{\mathrm{col}}$ denote the elements (i, j) of the matrices $\mathbf{M}$, $\boldsymbol{\Omega}^{\mathrm{row}}$ and $\boldsymbol{\Omega}^{\mathrm{col}}$, respectively. If $\boldsymbol{X}$ has a matrix normal distribution, then the squared matrix Mahalanbois distance has a χ^2 distribution with pq degrees of freedom, $\mathrm{MMD}^2(\boldsymbol{X}) \sim \chi^2_{pq}$ (Gupta and Nagar, 1999).

B.2 Proofs of Section 3.2

Proof of Proposition 3.2.0.1. In optimization problem (3.2.3) we want to maximize

$$\begin{aligned}
l(\boldsymbol{w}, \mathbf{M}, \boldsymbol{\Sigma}^{\mathrm{row}}, \boldsymbol{\Sigma}^{\mathrm{col}} | \mathfrak{X}) = &- \frac{1}{2} \sum_{i=1}^{n} w_i \Big(p \ln(\det(\boldsymbol{\Sigma}^{\mathrm{col}})) + q \ln(\det(\boldsymbol{\Sigma}^{\mathrm{row}})) \Big) \\
&- \frac{1}{2} \sum_{i=1}^{n} w_i \, \mathrm{MMD}^2(\boldsymbol{X}_i) - hpq \ln(2\pi)
\end{aligned} \tag{B.2}$$

subject to $w_i \in \{0, 1\}$ for all $i = 1, \ldots, n$ and $\sum_{i=1}^{n} w_i = h$. In Equation (B.2), $\mathrm{MMD}^2(\boldsymbol{X}_i)$ is defined as in Equation (3.2.2).

For any random h-subset H (or equivalently the corresponding set of weights $\boldsymbol{w}$) the

constrained MLEs for $\mathbf{M}$, $\mathbf{\Sigma}^{\text{row}}$, and $\mathbf{\Sigma}^{\text{col}}$ of Equation (B.2) can be written as:

$$\hat{\mathbf{M}}_H = \frac{1}{h}\sum_{i=i}^{n} w_i \boldsymbol{X}_i = \frac{1}{h}\sum_{i\in H} \boldsymbol{X}_i$$

$$\hat{\mathbf{\Sigma}}_H^{\text{row}} = \frac{1}{qh}\sum_{i=i}^{n} w_i (\boldsymbol{X}_i - \hat{\mathbf{M}}_H)\hat{\mathbf{\Omega}}_H^{\text{col}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_H)' = \frac{1}{qh}\sum_{i\in H} (\boldsymbol{X}_i - \hat{\mathbf{M}}_H)\hat{\mathbf{\Omega}}_H^{\text{col}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_H)'$$

$$\hat{\mathbf{\Sigma}}_H^{\text{col}} = \frac{1}{ph}\sum_{i=i}^{n} w_i (\boldsymbol{X}_i - \hat{\mathbf{M}}_H)'\hat{\mathbf{\Omega}}_H^{\text{row}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_H) = \frac{1}{ph}\sum_{i\in H} (\boldsymbol{X}_i - \hat{\mathbf{M}}_H)'\hat{\mathbf{\Omega}}_H^{\text{row}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_H)$$

Using those estimators to compute the sum of the Mahalanobis distances $\text{MMD}^2(\boldsymbol{X}_i)$ in Equation (B.2) we obtain

$$\begin{aligned}
\sum_{i=1}^{n} w_i \,\text{MMD}^2(\boldsymbol{X}_i) &= \sum_{i\in H} \text{tr}\left(\hat{\mathbf{\Omega}}_H^{\text{col}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_H)'\hat{\mathbf{\Omega}}_H^{\text{row}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_H)\right) \\
&= \sum_{i\in H} \text{tr}\left((\boldsymbol{X}_i - \hat{\mathbf{M}}_H)\hat{\mathbf{\Omega}}_H^{\text{col}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_H)'\hat{\mathbf{\Omega}}_H^{\text{row}}\right) \\
&= \text{tr}\left(\sum_{i\in H}((\boldsymbol{X}_i - \hat{\mathbf{M}}_H)\hat{\mathbf{\Omega}}_H^{\text{col}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_H)')\hat{\mathbf{\Omega}}_H^{\text{row}}\right) \\
&= \text{tr}\left(qh\hat{\mathbf{\Sigma}}_H^{\text{row}}\hat{\mathbf{\Omega}}_H^{\text{row}}\right) = hpq.
\end{aligned}$$

Thus, the terms in the second row of Equation (B.2) are all constant, and it is sufficient to maximize only the term in the first row, which contains the (negative) determinant of Equation (3.2.4). □

Properties of MMCD estimators

Proof of Lemma 3.3.0.1. Ad (a): We show that the MMCD estimators are matrix affine equivariant. Let us consider the objective of the MMCD for the transformed samples, which is to minimize

$$\begin{aligned}
\det(\hat{\mathbf{\Sigma}}_{\mathfrak{Z}_H}^{\text{col}} \otimes \hat{\mathbf{\Sigma}}_{\mathfrak{Z}_H}^{\text{row}}) &= \det\left((\boldsymbol{B}'\hat{\mathbf{\Sigma}}_{\mathfrak{X}_H}^{\text{col}}\boldsymbol{B}) \otimes (\boldsymbol{A}\hat{\mathbf{\Sigma}}_{\mathfrak{X}_H}^{\text{row}}\boldsymbol{A}')\right) \\
&= \left[\det(\boldsymbol{B}'\hat{\mathbf{\Sigma}}_{\mathfrak{X}_H}^{\text{col}}\boldsymbol{B})\right]^p \left[\det(\boldsymbol{A}\hat{\mathbf{\Sigma}}_{\mathfrak{X}_H}^{\text{row}}\boldsymbol{A}')\right]^q \\
&= \left[\det(\boldsymbol{B}')\det(\hat{\mathbf{\Sigma}}_{\mathfrak{X}_H}^{\text{col}})\det(\boldsymbol{B})\right]^p \left[\det(\boldsymbol{A})\det(\hat{\mathbf{\Sigma}}_{\mathfrak{X}_H}^{\text{row}})\det(\boldsymbol{A}')\right]^q \\
&= 4\det(\boldsymbol{B})^p \det(\boldsymbol{A})^q \det(\hat{\mathbf{\Sigma}}_{\mathfrak{X}_H}^{\text{col}})^p \det(\hat{\mathbf{\Sigma}}_{\mathfrak{X}_H}^{\text{row}})^q.
\end{aligned}$$

Since $4\det(\boldsymbol{B})^p \det(\boldsymbol{A})^q$ is constant, the objective does not change, and we obtain the same h-subset. Since the MMCD estimators correspond to the trimmed MLEs and the objective is not affected by the transformation, the matrix affine equivariance of the MMCD estimators follows from the matrix affine equivariance of the MLEs.

Ad (b): Suppose that $(\hat{\mathbf{M}}_{\mathfrak{Z}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Z}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Z}})$ are matrix affine equivariant estimators of location and covariance of the transformed sample $\mathfrak{Z}$, then

$$\begin{aligned}
&\text{MMD}^2(\boldsymbol{Z}_i; \hat{\mathbf{M}}_{\mathfrak{Z}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Z}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Z}}) \\
&= \text{tr}(\hat{\mathbf{\Omega}}^{\text{col}}_{\mathfrak{Z}}(\boldsymbol{Z}_i - \hat{\mathbf{M}}_{\mathfrak{Z}})' \hat{\mathbf{\Omega}}^{\text{row}}_{\mathfrak{Z}}(\boldsymbol{Z}_i - \hat{\mathbf{M}}_{\mathfrak{Z}})) \\
&= \text{tr}\Big(\big(\boldsymbol{B}^{-1}\hat{\mathbf{\Omega}}^{\text{col}}_{\mathfrak{X}}(\boldsymbol{B}')^{-1}(\boldsymbol{A}\boldsymbol{X}_i\boldsymbol{B} + \boldsymbol{C} - (\boldsymbol{A}\hat{\mathbf{M}}_{\mathfrak{X}}\boldsymbol{B} + \boldsymbol{C}))'\big) \\
&\qquad \big((\boldsymbol{A}')^{-1}\hat{\mathbf{\Omega}}^{\text{row}}_{\mathfrak{X}}\boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{X}_i\boldsymbol{B} + \boldsymbol{C} - (\boldsymbol{A}\hat{\mathbf{M}}_{\mathfrak{X}}\boldsymbol{B} + \boldsymbol{C}))\big)\Big) \\
&= \text{tr}(\boldsymbol{B}^{-1}\hat{\mathbf{\Omega}}^{\text{col}}_{\mathfrak{X}}(\boldsymbol{B}')^{-1}\boldsymbol{B}'(\boldsymbol{X}_i - \hat{\mathbf{M}}_{\mathfrak{X}})'\boldsymbol{A}'(\boldsymbol{A}')^{-1}\hat{\mathbf{\Omega}}^{\text{row}}_{\mathfrak{X}}\boldsymbol{A}^{-1}\boldsymbol{A}(\boldsymbol{X}_i - \hat{\mathbf{M}}_{\mathfrak{X}})\boldsymbol{B}) \\
&= \text{tr}(\hat{\mathbf{\Omega}}^{\text{col}}_{\mathfrak{X}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_{\mathfrak{X}})'\hat{\mathbf{\Omega}}^{\text{row}}_{\mathfrak{X}}(\boldsymbol{X}_i - \hat{\mathbf{M}}_{\mathfrak{X}})) = \text{MMD}^2(\boldsymbol{X}_i; \hat{\mathbf{M}}_{\mathfrak{X}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{X}}).
\end{aligned}$$

□

The proofs of Theorems 3.3.0.1 and 3.3.0.3 require some definitions and properties related to the vector space of matrices, which are introduced before the proofs of the theorems. Since all matrices of a fixed size form a vector space, objects such as ellipsoids or a simplex that are defined on the more common vector spaces are also defined here. Let

$$E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V}) = \{\boldsymbol{X} : \text{tr}(\boldsymbol{V}^{-1}(\boldsymbol{X} - \boldsymbol{T})'\boldsymbol{U}^{-1}(\boldsymbol{X} - \boldsymbol{T})) \leq 1\} \tag{B.3}$$

be the ellipsoid containing the matrices $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ with $\text{MMD}^2(\boldsymbol{X}; \boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V}) \leq 1$, where $\boldsymbol{T} \in \mathbb{R}^{p\times q}$, $\boldsymbol{U} \in \text{PDS}(p)$ and $\boldsymbol{V} \in \text{PDS}(q)$. The volume of this ellipsoid is given by

$$\text{vol}(E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})) = \underbrace{\frac{\pi^{pq/2}}{\Gamma(pq/2 + 1)}}_{=:\beta_{pq}} \prod_{i=1}^{p}\prod_{j=1}^{q} \sqrt{\lambda_i(\boldsymbol{U})\lambda_j(\boldsymbol{V})} = \beta_{pq} \underbrace{\det(\boldsymbol{U})^{q/2}\det(\boldsymbol{V})^{p/2}}_{=:\det(E(\boldsymbol{T},\boldsymbol{U},\boldsymbol{V}))}, \tag{B.4}$$

where Γ is the gamma function, $0 < \lambda_p(\boldsymbol{U}) \leq \ldots \leq \lambda_1(\boldsymbol{U})$ and $0 < \lambda_q(\boldsymbol{V}) \leq \ldots \leq \lambda_1(\boldsymbol{V})$ are the eigenvalues of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Moreover, the axes have lengths $\sqrt{\lambda_i(\boldsymbol{U})\lambda_j(\boldsymbol{V})}$.

Let $\boldsymbol{A}$ be a symmetric nonnegative definite $p \times p$ matrix, then

$$\lambda_1(\boldsymbol{A}) = \sup_{\boldsymbol{z}\in\mathbb{R}^p} \frac{\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}} \quad \text{and} \quad \lambda_n(\boldsymbol{A}) = \inf_{\boldsymbol{z}\in\mathbb{R}^p} \frac{\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}}. \tag{B.5}$$

Consider another symmetric nonnegative definite $p \times p$ matrix $\boldsymbol{B}$, then using Equation (B.5) we get that

$$\lambda_1(\boldsymbol{A} + \boldsymbol{B}) \leq \lambda_1(\boldsymbol{A}) + \lambda_1(\boldsymbol{B}) \quad \text{and} \quad \lambda_p(\boldsymbol{A} + \boldsymbol{B}) \geq \lambda_p(\boldsymbol{A}) + \lambda_p(\boldsymbol{B}). \tag{B.6}$$

If $\boldsymbol{A} \in \text{PDS}(p)$ with eigenvalues $0 < \lambda_p(\boldsymbol{A}) \leq \ldots \leq \lambda_1(\boldsymbol{A})$ then the eigenvalues of $\boldsymbol{A}^{-1}$ are the reciprocals of the eigenvalues of $\boldsymbol{A}$, i.e. $\lambda_i(\boldsymbol{A}^{-1}) = \lambda_i^{-1}(\boldsymbol{A})$. Hence, we have that

$$\frac{1}{\lambda_1(\boldsymbol{A})} = \inf_{\boldsymbol{z}\in\mathbb{R}^p} \frac{\boldsymbol{z}'\boldsymbol{A}^{-1}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}},$$

which implies that for any $\boldsymbol{x} \in \mathbb{R}^p$

$$\frac{1}{\lambda_1(\boldsymbol{A})} \leq \frac{\boldsymbol{x}'\boldsymbol{A}^{-1}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} \Leftrightarrow \boldsymbol{x}'\boldsymbol{x} \leq \boldsymbol{x}'\boldsymbol{A}^{-1}\boldsymbol{x}\lambda_1(\boldsymbol{A}). \tag{B.7}$$

Suppose $\boldsymbol{A} \in \mathrm{PDS}(p)$, $\boldsymbol{B} \in \mathrm{PDS}(q)$ and let $\lambda(\boldsymbol{A})$ be an eigenvalue of $\boldsymbol{A}$ with corresponding eigenvector $\boldsymbol{v}(\boldsymbol{A})$, and $\lambda(\boldsymbol{B})$ an eigenvalue of $\boldsymbol{B}$ with corresponding eigenvector $\boldsymbol{v}(\boldsymbol{B})$. Then $\lambda(\boldsymbol{A})\lambda(\boldsymbol{B})$ is an eigenvalue of $\boldsymbol{B} \otimes \boldsymbol{A}$ with corresponding eigenvector $\boldsymbol{v}(\boldsymbol{B}) \otimes \boldsymbol{v}(\boldsymbol{A})$. We denote the sequence of eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$ as $0 < \lambda_p(\boldsymbol{A}) \leq \ldots \leq \lambda_1(\boldsymbol{A})$ and $0 < \lambda_q(\boldsymbol{b}) \leq \ldots \leq \lambda_1(\boldsymbol{b})$, respectively. It follows that the smallest eigenvalue $\lambda_{pq}(\boldsymbol{A}, \boldsymbol{B}) = \lambda_p(\boldsymbol{A})\lambda_q(\boldsymbol{B})$ and the largest eigenvalue $\lambda_1(\boldsymbol{A}, \boldsymbol{B}) = \lambda_1(\boldsymbol{A})\lambda_1(\boldsymbol{B})$. Moreover note that for $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$

$$\mathrm{vec}(\boldsymbol{Z})'(\boldsymbol{B} \otimes \boldsymbol{A})\,\mathrm{vec}(\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z}),$$

as in Equation (3.2.2), which implies that

$$\lambda_{pq}(\boldsymbol{A}, \boldsymbol{B}) = \inf_{\boldsymbol{Z}\in\mathbb{R}^{p\times q}} \frac{\mathrm{tr}(\boldsymbol{B}\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z})}{\mathrm{tr}(\boldsymbol{Z}'\boldsymbol{Z})} \quad \text{and} \quad \lambda_1(\boldsymbol{A}, \boldsymbol{B}) = \sup_{\boldsymbol{Z}\in\mathbb{R}^{p\times q}} \frac{\mathrm{tr}(\boldsymbol{B}\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z})}{\mathrm{tr}(\boldsymbol{Z}'\boldsymbol{Z})}.$$

This leads us to the matrix-variate version of Equation (B.7), where for any matrix $\boldsymbol{X} \in \mathbb{R}^{p\times q}$

$$\|\boldsymbol{X}\|_F^2 = \mathrm{tr}(\boldsymbol{X}'\boldsymbol{X}) \leq \mathrm{tr}(\boldsymbol{B}^{-1}\boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X})\lambda_1(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}^{-1}\boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X})\lambda_1(\boldsymbol{A})\lambda_1(\boldsymbol{B}) \tag{B.8}$$

Lemma B.2.1. *Take $p, q \in \mathbb{N}$, $d = \lfloor p/q + q/p \rfloor$, $d + 2 \leq s \leq pq$, and matrices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_s \in \mathbb{R}^{p\times q}$ that are in general position, i.e., no subset of $r, 2 \leq r \leq s$ samples lies on an $r-2$ dimensional subspace. For an ellipsoid $E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})$ as in Equation* (B.3)*, containing the matrices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_s$, it holds that for every $C > 0$ there exists a constant $\alpha := \alpha(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_s) > 0$ only depending on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_s$ such that $\|\boldsymbol{T}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{T}'\boldsymbol{T})} > \alpha$ implies $\det(E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})) > C$, i.e.,*

$$\forall C > 0\ \exists \alpha > 0 : \|\boldsymbol{T}\|_F > \alpha \implies \det(E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})) > C.$$

Proof. The samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_s$ are in general position, which implies that they span a nonempty $s-1$ simplex. Since $E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})$ contains those samples, it also contains the simplex spanned by those matrices. This implies that there exists a constant $a > 0$ only depending on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_s$, such that the length of k, $s - 1 \leq k \leq pq$, of the pq axes of the ellipsoid $E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})$ is at least a, i.e., there are k out of pq indices $(i, j), 1 \leq i \leq p, 1 \leq j \leq q$ such that

$$\sqrt{\lambda_i(\boldsymbol{U})\lambda_j(\boldsymbol{V})} > a. \tag{B.9}$$

In Equation (B.9), $\lambda_i(\boldsymbol{U}), i \in \{1, \ldots, p\}$, are the eigenvalues of $\boldsymbol{U}$, and $\lambda_j(\boldsymbol{V}), j \in \{1, \ldots, q\}$, are the eigenvalues of $\boldsymbol{V}$. For any matrix $\boldsymbol{X}$ contained in $E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})$, Equations (B.3) and (B.8) imply that

$$\begin{aligned} \|\boldsymbol{X} - \boldsymbol{T}\|_F^2 &= \mathrm{tr}((\boldsymbol{X} - \boldsymbol{T})'(\boldsymbol{X} - \boldsymbol{T})) \\ &\leq \mathrm{tr}(\boldsymbol{V}^{-1}(\boldsymbol{X} - \boldsymbol{T})'\boldsymbol{U}^{-1}(\boldsymbol{X} - \boldsymbol{T}))\lambda_1(\boldsymbol{U})\lambda_1(\boldsymbol{V}) \\ &\leq \lambda_1(\boldsymbol{U})\lambda_1(\boldsymbol{V}). \end{aligned} \tag{B.10}$$

Without loss of generality, we assume that the matrix of all zeros $\mathbf{0} \in \mathbb{R}^{p\times q}$ is contained in the ellipsoid $E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})$, then Equation (B.10) implies that $\|\boldsymbol{T}\|_F^2 \leq \lambda_1(\boldsymbol{U})\lambda_1(\boldsymbol{V})$. Take $\alpha = C/(a^{pq-1})$, then we have that

$$\frac{C}{a^{pq-1}} < \|\boldsymbol{T}\|_F \leq \sqrt{\lambda_1(\boldsymbol{U})\lambda_1(\boldsymbol{V})} \Leftrightarrow C < \sqrt{\lambda_1(\boldsymbol{U})\lambda_1(\boldsymbol{V})}a^{pq-1}$$

and from Equaiton (B.9) it follows that

$$\begin{aligned}\det(E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})) :&= \det(\boldsymbol{U})^{q/2}\det(\boldsymbol{V})^{p/2} \\ &= \prod_{i=1}^{p}\prod_{j=1}^{q}\sqrt{\lambda_i(\boldsymbol{U})\lambda_j(\boldsymbol{V})} \\ &> \sqrt{\lambda_1(\boldsymbol{U})\lambda_1(\boldsymbol{V})}a^{pq-1} > C.\end{aligned}$$

□

Proof of Theorem 3.3.0.1. We show that the breakdown points of the MMCD estimators of location and covariance defined in Equations (3.3.3) and (3.3.4), respectively, are both m/n, with $m = \lfloor \min(n-h+1, h-(d+1)) \rfloor$, $d = \lfloor p/q + q/p \rfloor$. First, we prove that $\varepsilon^*(\hat{\mathbf{M}}, \mathfrak{X}) = \varepsilon^*(\hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}, \mathfrak{X}) \geq m/n$. Let $\mathfrak{Y}$ be the sample obtained by replacing at most $m-1$ matrices of $\mathfrak{X}$ by arbitrary $p \times q$ matrices. Since $n-(m-1) \geq h$, $\mathfrak{Y}$ contains at least h matrices of the orginial sample $\mathfrak{X}$ and because $m-1 \leq h-(d+1)-1$, every subset of size h of $\mathfrak{Y}$ includes at least $d+2$ matrices of the original sample $\mathfrak{X}$. Hence, the MMCD estimators can almost surely be computed for any h-subset of $\mathfrak{Y}$. Let us consider three ellipsoids:

- Let $E_{\max} = E(\mathbf{0}, c_{\max}\boldsymbol{I}, \boldsymbol{I})$ denote the smallest sphere that contains all samples in $\mathfrak{X}$, where $c_{\max}$ is chosen accordingly.
- Let $E_h = E(\mathbf{0}, c_h\boldsymbol{I}, \boldsymbol{I})$ denote the smallest sphere that contains the h samples of $\mathfrak{X}$ that are also in $\mathfrak{Y}$, where c_h is chosen accordingly.
- Let $E_{\text{MMCD}} = E(\hat{\mathbf{M}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}})$ denote the MMCD ellipsoid.

It follows that $\det(E_{\text{MMCD}}) \leq \det(E_h) \leq \det(E_{\max}) =: \alpha$, where for an ellipsoid $E = E(\boldsymbol{T}, \boldsymbol{U}, \boldsymbol{V})$, $\det(E)$ is defined in (B.4). Note that $\mathfrak{X}$ is a collection of random samples from a continuous distribution and therefore it is in general position almost surely. Further, E_{MMCD} covers at least h samples, and those include at least $d+2$ samples of $\mathfrak{X}$, which span a nonempty $d+1$ simplex. Lemma B.2.1 shows that there exists a constant $\alpha > 0$ that only depends on those $d+2$ samples such that, if $\left\|\hat{\mathbf{M}}_{\mathfrak{Y}}\right\|_F > C$ it would imply $\det(E_{\text{MMCD}}) > \alpha$. As shown above, this is not possible, hence $\left\|\hat{\mathbf{M}}_{\mathfrak{Y}}\right\|_F \leq C$.

Similarly, since $\mathfrak{Y}$ contains at least $d+2$ matrices of the original sample $\mathfrak{X}$, the MMCD estimators almost surely yield positive definite covariance estimates $\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}}$ and $\hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}}$. More specifically, let $\mathfrak{X}_T$, $T \subseteq H$ be the subset of the at least $d+2$ matrices of the original sample that are in $\mathfrak{Y}$. Since $|T| \geq d+2 = \lfloor p/q + q/p \rfloor + 2$ the MLE estimators $(\hat{\boldsymbol{M}}_{\mathfrak{X}_T}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}_T}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}_T})$ of this subsample are almost surely positive definite. Let $E_T = E(\hat{\boldsymbol{M}}_{\mathfrak{X}_T}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}_T}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{X}_T})$ denote

the corresponding ellipsoid which is the smallest ellipsoid, of the type $E = E(\mathbf{T}, \mathbf{U}, \mathbf{V})$ as in Equation (B.3), containing the samples $\mathfrak{X}_T$ as one can think of it as the MMCD ellipsoid for those $|T| \geq d+2$ samples with $H = T$. This further implies that the volume of the corresponding ellipsoid E_T is bounded from below by a constant only depending on $\mathfrak{X}$, i.e. $\det(E_T) \geq v > 0$. As E_{MMCD} is also an ellipsoid containing the samples $\mathfrak{X}_T$, $\det(E_{\text{MMCD}}) \geq \det(E_T) \geq v > 0$. Moreover, it also means that there exists a constant k depending only on $\mathfrak{X}$, such that $E_T \subseteq kE_{\text{MMCD}}$, implying that there exists a constant $\gamma > 0$ depending only on $\mathfrak{X}$, such that $\lambda_i(\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}})\lambda_j(\hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}}) > \gamma, 1 \leq i \leq p, 1 \leq j \leq q$. Especially, $\lambda_p(\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}})\lambda_q(\hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}}) > \gamma$. Since also $\det(E_{\text{MMCD}}) \leq \alpha$ there exists a constant $\delta > 0$, depending only on $\mathfrak{X}$ such that $\lambda_i(\hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}})\lambda_j(\hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}}) < \delta, 1 \leq i \leq p, 1 \leq j \leq q$.

Next we show that $\varepsilon^*(\hat{\mathbf{M}}, \mathfrak{X}) = \varepsilon^*(\hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}, \mathfrak{X}) \leq m/n$. If $m = n - h + 1$, we replace $m = n - h + 1$ matrices of $\mathfrak{X}$ to obtain $\mathfrak{Y}$, then $n - m = h - 1$, implying that every subset of h samples of $\mathfrak{Y}$ contains at least one conaminated sample. Hence, $E_{\text{MMCD}} = E(\hat{\mathbf{M}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}})$ also includes at least one contaminated sample. Let $\|\boldsymbol{X}\|_F \to \infty$ for all contaminated samples $\boldsymbol{X}$, then at least one eigenvalue of E_{MMCD} explodes and the MMCD location and covariance estimators break down. Finally, consider the case where $m = h-(d+1)$. To construct $\mathfrak{Y}$, take any $d+1$ samples of $\mathfrak{X}$ and consider the d dimensional hyperplane L they determine. Replace $h - (d+1)$ samples that are not in L and replace them with matrices on L. Then L contains h points of $\mathfrak{Y}$ and the ellipsoid covering those points has volume zero and hence determinant zero. Since $\mathfrak{X}$ is in general position, we can construct $\mathfrak{Y}$ such that no other lower dimensional hyperplane contains h points of $\mathfrak{Y}$. Hence, $\hat{\mathbf{M}}_{\mathfrak{Y}}$ lies on L and $E_{\text{MMCD}} = E(\hat{\mathbf{M}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{row}}_{\mathfrak{Y}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\mathfrak{Y}})$ has zero determinant. This implies that at least one eigenvalue is zero, hence the MMCD location and covariance estimators break down. □

Proof of Theorem 3.3.0.2. Let $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ be a sample of matrix-variate observations and $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{x}_i = \text{vec}(\boldsymbol{X}_i)$, $i = 1, \ldots n$ its vectorized form. The MCD estimator can also be found as a solution to the following maximization problem:

$$\max_{\boldsymbol{w}, \hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}} l(\boldsymbol{w}, \hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}} | (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) = \max_{\boldsymbol{w}, \hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}} -\frac{1}{2} \sum_{i=1}^{n} w_i (\ln(\det(\hat{\mathbf{\Sigma}})) + pq \ln(2\pi) + \text{MD}^2(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}))$$

subject to $w_1, \ldots, w_n \in \{0, 1\}$, $\sum_{i=1}^{n} w_i = h$, $\hat{\boldsymbol{\mu}} \in \mathbb{R}^{pq}$, $\hat{\mathbf{\Sigma}} \in \text{PDS}(pq)$; see Raymaekers and Rousseeuw (2023) for more insight. Similarly, the MMCD estimator is a solution to the following maximization problem:

$$\begin{aligned} &\max_{\boldsymbol{w}, \hat{\mathbf{M}}, \hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}} l(\boldsymbol{w}, \hat{\mathbf{M}}, \hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}} | (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)) \\ = &\max_{\boldsymbol{w}, \hat{\mathbf{M}}, \hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}} -\frac{1}{2} \sum_{i=1}^{n} w_i \Big(p \ln(\det(\hat{\mathbf{\Sigma}}^{\text{col}})) + q \ln(\det(\hat{\mathbf{\Sigma}}^{\text{row}})) + \text{MMD}^2(\boldsymbol{X}_i) + pq \ln(2\pi) \Big) \end{aligned}$$

subject to $w_1, \ldots, w_n \in \{0, 1\}$, $\sum_{i=1}^{n} w_i = h$, $\hat{\mathbf{M}} \in \mathbb{R}^{p \times q}$, $\hat{\mathbf{\Sigma}}^{\text{row}} \in \text{PDS}(p)$, $\hat{\mathbf{\Sigma}}^{\text{col}} \in \text{PDS}(q)$; see Proposition 3.2.0.1.

Denote further $(\boldsymbol{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\mathbf{\Sigma}}_{\text{MCD}})$ and $(\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\mathbf{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\mathbf{\Sigma}}^{\text{row}}_{\text{MMCD}})$ weights, mean and covariance estimators for the vectorized sample $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, based on MCD and

MMCD, respectively. As $\boldsymbol{X}_i \sim \mathcal{ME}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}, g)$, then $\boldsymbol{x}_i \sim \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}, g)$, with $\mathbb{E}(\boldsymbol{x}_i) = \boldsymbol{\mu} = \text{vec}(\mathbf{M})$, $\text{cov}\,(\boldsymbol{x}_i) = c_g \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}$, where c_g is a distribution-specific scaling parameter; for more details see Theorem 2.11 in Gupta et al. (2013). Moreover, the mean estimator $\hat{\boldsymbol{\mu}}_{\text{MCD}}$ and properly scaled covariance estimator $\hat{\boldsymbol{\Sigma}}_{\text{MCD}}$ are strongly consistent for the population counterparts $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}$; see e.g. Croux and Haesbroeck (1999) and Cator and Lopuhaä (2012). Especially, this implies that for every $\delta > 0$ there exists $n \in \mathbb{N}$ such that

$$\|\hat{\boldsymbol{\mu}}_{\text{MCD}} - \boldsymbol{\mu}\| \overset{a.s.}{<} \delta, \quad \left\|\hat{\boldsymbol{\Sigma}}_{\text{MCD}} - \boldsymbol{A} \otimes \boldsymbol{B}\right\| \overset{a.s.}{<} \delta,$$

for some $\boldsymbol{A} \otimes \boldsymbol{B} \in \text{PDS}(p) \otimes \text{PDS}(q)$. In the following, we will drop *a.s.* superscript from (in)equality signs when it is clear from the context. For fixed weights $\boldsymbol{w}$, the log-likelihood function $l_{\cdot,\boldsymbol{w}|(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)} : (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{w}, (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))$ is continuous in both $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$, and its continuity implies

$$\left|l(\boldsymbol{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) - l(\boldsymbol{w}_{\text{MCD}}, \boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{B}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))\right| < \varepsilon,$$

for $\varepsilon = \varepsilon(\delta) > 0$. The solution $(\boldsymbol{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MCD}})$ is optimal for $l(\cdot|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))$, implying that

$$0 < l(\boldsymbol{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) - l(\boldsymbol{w}_{\text{MCD}}, \boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{B}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) < \varepsilon. \quad \text{(B.11)}$$

Similarly, $(\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}})$ is a maximizer of $l(\cdot|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))$ in the set of all feasible weights, means, and covariances with Kronecker product structure. As $(\boldsymbol{w}_{\text{MCD}}, \boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{B})$ belongs to the same set,

$$l(\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) > l(\boldsymbol{w}_{\text{MCD}}, \boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{B}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)).$$

Denote further $\hat{\boldsymbol{S}}_{\text{MMCD}} = \frac{1}{h} \sum_{i=1}^{n} w_{\text{MMCD},i} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\text{MMCD}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\text{MMCD}})'$ to be the estimate of $\boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}$, based on the weights (subset) produced by the MMCD algorithm. As $\hat{\boldsymbol{S}}_{\text{MMCD}}$ is optimal for l given fixed weights $\boldsymbol{w}_{\text{MMCD}}$,

$$\begin{aligned}
&l(\boldsymbol{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) \\
>&l(\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\boldsymbol{S}}_{\text{MMCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) \\
>&l(\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) \\
>&l(\boldsymbol{w}_{\text{MCD}}, \boldsymbol{\mu}, \boldsymbol{A} \otimes \boldsymbol{B}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)). \qquad \text{(B.12)}
\end{aligned}$$

(B.11) and (B.12) now give that

$$0 < l(\boldsymbol{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) - l(\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\boldsymbol{S}}_{\text{MMCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) < \varepsilon,$$

i.e., due to Proposition 3.2.0.1,

$$0 < \det(\hat{\boldsymbol{S}}_{\text{MMCD}}) - \det(\hat{\boldsymbol{\Sigma}}_{\text{MCD}}) < \varepsilon, \quad \text{(B.13)}$$

for $\varepsilon = \varepsilon(n) > 0$, arbitrarily small ($\varepsilon(n) \to 0$, $n \to \infty$). As both $\hat{\boldsymbol{\Sigma}}_{\text{MCD}}$ and $\hat{\boldsymbol{S}}_{\text{MMCD}}$ are weighted sample covariances for the random sample of vectorized observations calculated

using the weights satisfying the same constraints, Corollary 4.1. in Cator and Lopuhaä (2012) (taking P_t to be the empirical measure based on the sample $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$) implies that

$$\hat{\boldsymbol{\mu}}_{\text{MMCD}} \xrightarrow{a.s.} \boldsymbol{\mu}, \quad \hat{\boldsymbol{S}}_{\text{MMCD}} \xrightarrow{a.s.} c(\alpha)^{-1}\boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}} \quad , \tag{B.14}$$

where $c(\alpha) > 0$ is a distribution-specific consistency factor of the MCD given in Croux and Haesbroeck (1999).

To complete the proof consider reparametrization of $l(\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{A}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))$ for fixed weights $\boldsymbol{w} \in \mathbb{R}^n$, mean $\boldsymbol{a} \in \mathbb{R}^{pq}$, and covariance $\boldsymbol{A} \in \text{PDS}(pq)$ in terms of the precicion matrix $\boldsymbol{B} = \boldsymbol{A}^{-1}$. Denote this new parametrization as $g(\boldsymbol{B}|\boldsymbol{w}, \boldsymbol{a}, (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) = l(\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{B}^{-1}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, which is now concave in $\boldsymbol{B}$. Especially, for $\boldsymbol{w} = \boldsymbol{w}_{\text{MMCD}}$ and $\boldsymbol{a} = \hat{\boldsymbol{\mu}}_{\text{MMCD}}$, the function $g(\boldsymbol{B}|\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))$ is concave in $\boldsymbol{B}$ and achieves a unique global maximum at $\boldsymbol{B} = \hat{\boldsymbol{S}}_{\text{MMCD}}$. Equations (B.11) and (B.12) then give

$$\begin{aligned} 0 <& l(\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\boldsymbol{S}}_{\text{MMCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) \\ & - l(\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}}|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) < \varepsilon, \end{aligned}$$

further implying that

$$\begin{aligned} 0 <& g(\hat{\boldsymbol{S}}^{-1}_{\text{MMCD}}|\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) \\ & - g((\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}})^{-1}|\boldsymbol{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)) < \varepsilon, \end{aligned}$$

as both $\hat{\boldsymbol{S}}_{\text{MMCD}}$ and $\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}}$ are a.s. positive definite for n large enough. Concavity of g and the fact that $\hat{\boldsymbol{S}}^{-1}_{\text{MMCD}}$ is its global maximum further imply that

$$\|\hat{\boldsymbol{S}}^{-1}_{\text{MMCD}} - (\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}})^{-1}\| < \delta_1,$$

for $\delta_1 = \delta_1(\varepsilon) \to 0$ as $n \to \infty$. Almost sure positive definiteness of $\hat{\boldsymbol{S}}_{\text{MMCD}}$ and $\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}}$, and continuity of matrix inverse imply that

$$\|\hat{\boldsymbol{S}}_{\text{MMCD}} - \hat{\boldsymbol{\Sigma}}^{\text{col}}_{\text{MMCD}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\text{MMCD}}\| < \delta, \tag{B.15}$$

for $\delta = \delta(\varepsilon) \to 0$ as $n \to \infty$. Equations (B.14) and (B.15) now complete the proof. Observe that the proof indicates that the distribution-specific consistency factor is inherited from the MCD covariance estimator; see Croux and Haesbroeck (1999). □

Proof of Theorem 3.3.0.3. We show that the breakdown points of the reweighted MMCD estimators are at least as high as the breakdown points of the raw MMCD estimators. Let $\mathfrak{Y}$ be the sample obtained by replacing at most $m - 1$ matrices of $\mathfrak{X}$ by arbitrary $p \times q$ matrices. Let $\hat{\mathbf{M}}_{\mathfrak{Y}}$, $\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}}$, and $\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}}$ denote the *raw* MMCD estimators and $\tilde{\mathbf{M}}_{\mathfrak{Y}}$, $\tilde{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}}$, and $\tilde{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}}$ denote the *reweighted* MMCD estimators based on the corrupted sample $\mathfrak{Y}$. Further, $d(\boldsymbol{Y}_i) = \text{MMD}(\boldsymbol{Y}_i; \hat{\mathbf{M}}_{\mathfrak{Y}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}})$, $i \in N = \{1, \ldots, n\}$, denote the matrix Mahalanbois distances of the corrupted sample based on the *raw* MMCD estimators. Since $m \leq \varepsilon^*(\hat{\mathbf{M}}_{\mathfrak{X}}, \mathfrak{X}) - 1 = \varepsilon^*(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{X}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{X}}, \mathfrak{X}) - 1$ it follows that there exist constants k_0, k_1, and k_2 that only depend on $\mathfrak{X}$, such that

$$\begin{aligned} & \left\|\hat{\mathbf{M}}_{\mathfrak{Y}}\right\| \leq k_0 < \infty \quad \text{and} \\ & 0 < k_1 < \lambda_p(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}})\lambda_q(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}}) \leq \lambda_1(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}})\lambda_1(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}}) \leq k_2 < \infty. \end{aligned} \tag{B.16}$$

Since at least $\lfloor (n+d+2)/2 \rfloor$ have a positive weight and at most $\lfloor (n-d)/2 \rfloor - 1$ observations are replaced, there are at least $d+2$ observations of the orginal sample $\mathfrak{X}$ contained in $\mathfrak{Y}$ that have a positive weight. Let $T \subseteq N$ denote the indices of those samples, then we have that

$$\sum_{i=1}^{n} w(d(\boldsymbol{Y}_i)) = \sum_{i \in N \setminus T} w(d(\boldsymbol{Y}_i)) + \sum_{i \in T} w(d(\boldsymbol{X}_i)) \geq \sum_{i \in T} w(d(\boldsymbol{X}_i)) \geq (d+2)c_0 > 0, \quad \text{(B.17)}$$

with $c_0 := \min_{i \in T} w(d(\boldsymbol{X}_i)) > 0$. This implies that the denominators of $\tilde{\mathbf{M}}_{\mathfrak{Y}}$, $\tilde{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}}$, and $\tilde{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}}$ are always positive.

Let us now show that there exists a constant $\alpha_0 < \infty$ only dependent on $\mathfrak{X}$ such that $\left\| \tilde{\mathbf{M}}_{\mathfrak{Y}} \right\|_F < \alpha_0$. From Equation (B.8) we have that

$$\begin{aligned} \left\| \boldsymbol{Y}_i - \hat{\mathbf{M}}_{\mathfrak{Y}} \right\|_F^2 &\leq \operatorname{tr}(\hat{\boldsymbol{\Omega}}^{\text{col}}_{\mathfrak{Y}} (\boldsymbol{Y}_i - \hat{\mathbf{M}}_{\mathfrak{Y}})' \hat{\boldsymbol{\Omega}}^{\text{row}}_{\mathfrak{Y}} (\boldsymbol{Y}_i - \hat{\mathbf{M}}_{\mathfrak{Y}})) \lambda_1(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}}) \lambda_1(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}}) \\ &= d(\boldsymbol{Y}_i) \lambda_1(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}}) \lambda_1(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}}). \end{aligned}$$

When computing $\tilde{\mathbf{M}}_{\mathfrak{Y}}$ we have that $w(d(\boldsymbol{Y}_i)) = 0$ if $d(\boldsymbol{Y}_i) > c_1$ and for all $\boldsymbol{Y}_i \in \mathfrak{Y}$ that are assigned positive weights, Equation (B.16) yields

$$\|\boldsymbol{Y}_i\|_F^2 \leq \left\| \boldsymbol{Y}_i - \hat{\mathbf{M}}_{\mathfrak{Y}} \right\|_F^2 + \left\| \hat{\mathbf{M}}_{\mathfrak{Y}} \right\|_F^2 \leq c_1 k_2 + k_0^2. \quad \text{(B.18)}$$

Since the denominator of is $\tilde{\mathbf{M}}_{\mathfrak{Y}}$ bounded according to Equation (B.17), w is non-increasing and bounded, and k_0 and k_2 are only dependent on $\mathfrak{X}$, there exsits a constant α_0 only dependent on $\mathfrak{X}$ such that

$$\left\| \tilde{\mathbf{M}}_{\mathfrak{Y}} \right\|_F \leq \alpha_0 < \infty. \quad \text{(B.19)}$$

To show that the covariance does not break down, we first consider the case of the weight function $w(d_i) = \mathbb{1}(d_i \leq c_1)$, for $c_1 > 0$. Let $S \subseteq \{1, \ldots, N\}$ denote the subset of indices of the $s = |S|$ samples of $\mathfrak{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ for which $d_i \leq c_1, i \in S$. Observe that the h samples $\boldsymbol{Y}_i$, $i \in H$ are those with the smallest MD, hence $T \subseteq H \subseteq S$. Let $\hat{\boldsymbol{M}}_{\mathfrak{Y}_T}, \hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}_T}, \hat{\boldsymbol{\Omega}}_{\mathfrak{Y}_T}$ and $\hat{\boldsymbol{M}}_{\mathfrak{Y}_S}, \hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}_S}, \hat{\boldsymbol{\Omega}}_{\mathfrak{Y}_S}$ denote the MLE estimators of $\mathfrak{Y}_T = (\boldsymbol{Y}_i)_{i \in T}$ and $\mathfrak{Y}_S = (\boldsymbol{Y}_i)_{i \in S}$, respectively. Consider the following three ellipsoids:

- Let $E_T = E(\hat{\boldsymbol{M}}_{\mathfrak{Y}_T}, \hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}_T}, \hat{\boldsymbol{\Omega}}_{\mathfrak{Y}_T})$ denote the ellipsoid corresponding to the MLEs of $\mathfrak{Y}_T$, i.e., the smallest ellipsoid containing those at least $d+2$ samples.
- Let $E_S = E(\hat{\boldsymbol{M}}_{\mathfrak{Y}_S}, \hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}_S}, \hat{\boldsymbol{\Omega}}_{\mathfrak{Y}_S})$ denote the ellipsoid corresponding to the MLEs of $\mathfrak{Y}_S$.
- Let $E_0 = E(\mathbf{0}, k\boldsymbol{I}_p, \boldsymbol{I}_q)$ denote the smallest sphere containing the samples $\mathfrak{Y}_S$, where $k = c_1 k_2 + k_0^2$ is as in (B.18).

Observe first that as E_T is the smallest ellipsoid containing the samples $\mathfrak{Y}_T = \mathfrak{X}_T$ that are also in E_S, there exists a constant a_1 depending only on $\mathfrak{X}_T$, such that $E_T \subseteq a_1 E_S := E(\hat{\boldsymbol{M}}_{\mathfrak{Y}_S}, a_1 \hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}_S}, \hat{\boldsymbol{\Omega}}_{\mathfrak{Y}_S})$. On the other hand, E_S is the smallest ellipsoid containing $\mathfrak{Y}_S$. As these points are also in E_0, then there exist $\alpha = \alpha(c_1)$ such that $\det(E_S) \leq \det(E_0) \leq \alpha$.

Equivalent argumentation as in the proof of Theorem 3.3.0.1 completes the first part of the proof.

Let now $w = w(d_i)$ be an arbitrarily, nondecreasing, bounded weight function, such that $w(d_i) = 0$ if $d_i > c_1$, $i = 1, \ldots, n$. The weighted log-likelihood function for the sample $\mathfrak{Y}$, with the weights satisfying $\sum_{i=1}^{n} w_i = s$ is given by

$$\begin{aligned} l(\boldsymbol{w}, \mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}} | \mathfrak{Y}) &= -\frac{1}{2} \sum_{i=1}^{m} w_i \Big(p \ln(\det(\boldsymbol{\Sigma}^{\text{col}})) + q \ln(\det(\boldsymbol{\Sigma}^{\text{row}})) \\ &\quad + \operatorname{tr}(\boldsymbol{\Omega}^{\text{col}} (\boldsymbol{Y}_i - \mathbf{M})' \boldsymbol{\Omega}^{\text{row}} (\boldsymbol{Y}_i - \mathbf{M})) + pq \ln(2\pi) \Big) \\ &= -\frac{1}{2} \Big(s \big(p \ln(\det(\boldsymbol{\Sigma}^{\text{col}})) + q \ln(\det(\boldsymbol{\Sigma}^{\text{row}})) \big) \\ &\quad + \sum_{i=1}^{s} \operatorname{tr}(\boldsymbol{\Omega}^{\text{col}} (\boldsymbol{Z}_i - \mathbf{M})' \boldsymbol{\Omega}^{\text{row}} (\boldsymbol{Z}_i - \mathbf{M})) + pq \ln(2\pi) \Big) \\ &= l(\tilde{\boldsymbol{w}}, \boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}} | \mathfrak{Z}), \end{aligned}$$

where $\tilde{\boldsymbol{w}} = (\tilde{w}(d_1), \ldots, \tilde{w}(d_n))$, the new weight function satisfies $\tilde{w}(d_i) = \mathbb{1}(d_1 \leq c_1)$, $\mathfrak{Z} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$, and $\boldsymbol{Z}_i = \sqrt{w_i} \boldsymbol{Y}_i$, $i = 1, \ldots, n$. To complete the proof it is sufficient to observe the following: $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_h$ contains at least $d + 2$ points of the form $\sqrt{w_i} \boldsymbol{X}_i$ and are in a general position, as $w_i \geq a_2 > 0$, for some constant depending only on $\mathfrak{X}$. Moreover, $\|\boldsymbol{Z}_i\|_F^2 = w_i \|\boldsymbol{Y}_i\|_F^2 \leq w_i (c_1 k_2 + k_0^2) \leq w(0)(c_1 k_2 + k_0^2)$, $i = 1, \ldots, s$. The statement now follows from the first part of the proof, observing that assumption $\sum_{i=1}^{m} w_i = s$ without loss of generality, since $0 < w(0) \leq \sum_{i=1}^{n} w_i \leq s w(0) < \infty$. □

B.3 MMCD Algorithm

Algorithm 3 Iterative C-step procedure for the MMCD estimators

procedure CSTEP($(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n), H_{\text{old}}, \varepsilon > 0$)
  $(\hat{\mathbf{M}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}}) = \text{MLE}((\boldsymbol{X}_i)_{i \in H_{\text{old}}})$
  $h = |H_{\text{old}}|$
  repeat
    $(\hat{\mathbf{M}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}}) = (\hat{\mathbf{M}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}})$
    $\boldsymbol{d} = (\text{MMD}^2(\boldsymbol{X}_1; \hat{\mathbf{M}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}}), \ldots, \text{MMD}^2(\boldsymbol{X}_n; \hat{\mathbf{M}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}}))$
    $\pi_1(i) = \{\{1, \ldots, n\} \to \{1, \ldots, n\} : i \mapsto j : d_{\pi(1)} \leq \ldots \leq d_{\pi(n)}\}$
    $H_{\text{new}} = \{\pi(1), \pi(2), \ldots, \pi(h)\}$
    $(\hat{\mathbf{M}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}}) = \text{MLE}((\boldsymbol{X}_i)_{i \in H_{\text{new}}})$
  until $\left| p(\ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}})) - \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}}))) + q(\ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}})) - \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}}))) \right| < \varepsilon$
  return $\hat{\mathbf{M}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}}, \boldsymbol{d}, H_{\text{new}}$
end procedure

Algorithm 4 Fast *reweighted* MMCD procedure

procedure MMCD($\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$)
  $h = \lfloor {}^{(n+d+2)}/_{2} \rfloor$
  $\alpha = {}^{h}/_{n}$
  $N = \{1, \ldots, n\}$
  for $k = 1$ to 500 **do**
    $H_k = \text{sample}(N, \text{size} = d + 2)$
    $(\hat{\mathbf{M}}_k, \hat{\mathbf{\Sigma}}_k^{\text{row}}, \hat{\mathbf{\Sigma}}_k^{\text{col}}, \boldsymbol{d}_k, H_k) = \text{CSTEP}_2(\mathfrak{X}, H_k)$ ▷ 2 MLE and C-step iterations
    $\delta_k = p \ln(\det(\hat{\mathbf{\Sigma}}_k^{\text{col}})) + q \ln(\det(\hat{\mathbf{\Sigma}}_k^{\text{row}}))$
  end for
  $\pi_\delta(i) = \{\{1, \ldots, 500\} \to \{1, \ldots, 500\} : i \mapsto j : \delta_{\pi_\delta(1)} \leq \ldots \leq \delta_{\pi_\delta(500)}\}$
  for $l \in \{\pi_\delta(1), \pi_\delta(2), \ldots, \pi_\delta(10)\}$ **do**
    $(\hat{\mathbf{M}}_l, \hat{\mathbf{\Sigma}}_l^{\text{row}}, \hat{\mathbf{\Sigma}}_l^{\text{col}}, \boldsymbol{d}_l, H_l) = \text{CSTEP}(\mathfrak{X}, H_l)$ ▷ Iterating C-steps until convergence
    $\delta_l = p \ln(\det(\hat{\mathbf{\Sigma}}_l^{\text{col}})) + q \ln(\det(\hat{\mathbf{\Sigma}}_l^{\text{row}}))$
  end for
  $j = \arg\min_{k \in N}(\delta_k)$
  $(\hat{\mathbf{M}}, \hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}) = (\hat{\mathbf{M}}_j, c(\alpha)\hat{\mathbf{\Sigma}}_j^{\text{row}}, \hat{\mathbf{\Sigma}}_j^{\text{col}})$ ▷ Consistency scaling for raw MMCD
  $\boldsymbol{d} = (\text{MMD}^2(\boldsymbol{X}_1; \hat{\mathbf{M}}, \hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}), \ldots, \text{MMD}^2(\boldsymbol{X}_n; \hat{\mathbf{M}}, \hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}))$
  $H = H_j \cup \{i \in N | d_i < \chi^2_{0.975;pq}\}$
  $(\hat{\mathbf{M}}, \hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}) = \text{MLE}(\boldsymbol{X}_{i \in H})$ ▷ Computation of reweighted MMCD
  $\tilde{\alpha} = {}^{|H|}/_{n}$
  $(\hat{\mathbf{M}}_*, \hat{\mathbf{\Sigma}}_*^{\text{row}}, \hat{\mathbf{\Sigma}}_*^{\text{col}}) = (\hat{\mathbf{M}}, c(\tilde{\alpha})\hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}})$ ▷ Consistency scaling for reweighted MMCD
  return $\hat{\mathbf{M}}_*, \hat{\mathbf{\Sigma}}_*^{\text{row}}, \hat{\mathbf{\Sigma}}_*^{\text{col}}$
end procedure

Elemental Subsets

For large n, the probability of obtaining at least one clean subset with $d+2$ observations among m random subsets tends to

$$1 - (1 - (1 - \varepsilon)^{d+2})^m,$$

with ε denoting the percentage of outliers, see also Rousseeuw and Van Driessen (1999). Hence, the number of subsets we must investigate to obtain at least one clean subset with a probability of β is

$$\lceil \log(1-\beta)/log(1-(1-\varepsilon)^{d+2}) \rceil. \tag{B.20}$$

In Figure B.1, we plot the number of necessary subsets according to Equation (B.20) for $\beta = 0.99$ for d between 1 and 50 and ε between 0 and 0.5. The different green-shaded areas starting from the bottom right indicate settings where up to $m = 500$ initial subsets of size $d+2$ are sufficient to obtain at least one clean subset with a probability of $\beta = 0.99$ and the various shades of orange indicate settings where we need more elemental subsets.

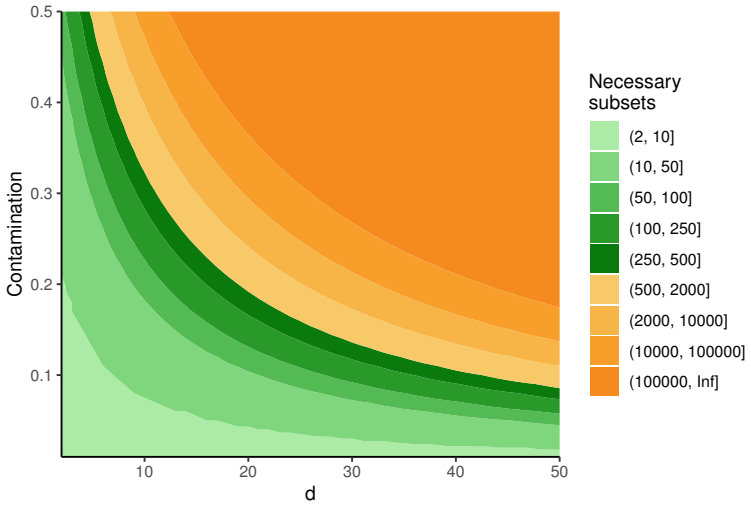


Figure B.1: Number of subsets of size $d+2$ we have to investigate for various levels of contamination, to obtain at least one clean subset with a probability of 99%

We assess the influence of using only 2 C-step and MLE iterations on the MMCD estimators' objective, the determinant of $\hat{\mathbf{\Sigma}}^{\text{col}} \otimes \hat{\mathbf{\Sigma}}^{\text{row}}$. We consider a setting with $n = 200$ observations with $p = 2$ rows and $q = 8$ columns. The clean observations are generated by a centered matrix normal distribution with $\mathbf{\Sigma}^{\text{row}} = \mathbf{\Sigma}^{\text{fix}}(0.7)$ and $\mathbf{\Sigma}^{\text{col}} = \mathbf{\Sigma}^{\text{mix}}(0.7)$, with diagonal entries $\sigma^{\text{fix}}_{jj} = \sigma^{\text{mix}}_{jj} = 1$ and off-diagonal entries $\sigma^{\text{fix}}_{jk}(0.7) = 0.7$ and $\sigma^{\text{mix}}_{jk}(0.7) = 0.7^{|j-k|}$, respectively. The outliers have a mean of 5 and the same covariance as the regular observations. We use 100 random subsets and plot $\det(\hat{\mathbf{\Sigma}}^{\text{col}} \otimes \hat{\mathbf{\Sigma}}^{\text{row}})$ for subsequent C-step iterations with 40% of contamination. We compare the setting when we limit the number of MLE iterations to 2 or iterate until convergence and/or use elemental subsets with $d+2 = 6$ instead of h-subsets of size $n/2 = 100$. Comparing the top and bottom row of Figure B.2, we see that there is virtually no difference in the objective function while limiting the ML iterations increases

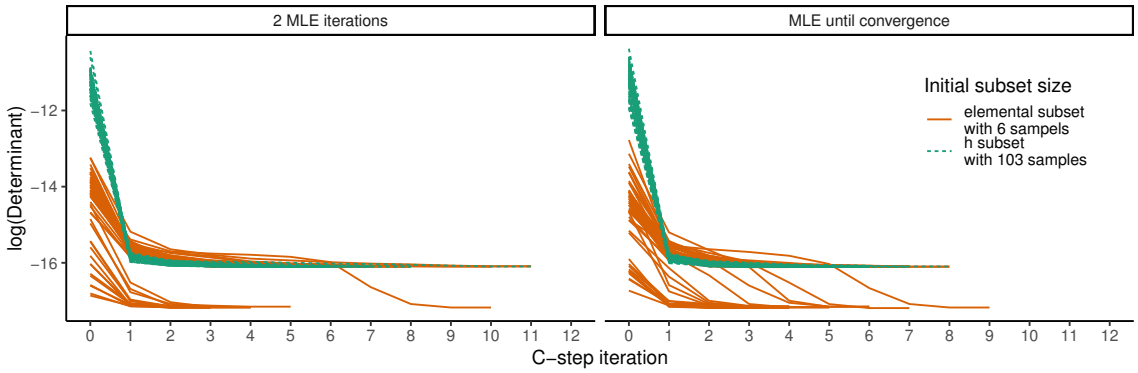


Figure B.2: Logarithm of determinant for successive C-step iterations to analyze the effects of initial subset size and the number of ML iterations.

the computation speed. For the subset size, we see that several of the elemental subsets yield robust solutions with a lower covariance determinant than the larger h-subsets and that most of them are identified after 1 or 2 iterations. While 40% contamination is not often encountered in practice, it shows that the algorithm can deal with settings with such a high level of contamination. We also analyzed settings with lower contamination, and using elemental subsets and fewer ML iterations had no negative effects in those settings, however, the larger h-subsets also led to robust solutions more frequently.

Remark B.3.1. *Instead of using the consistency factor $c(\alpha)$ given in Equation* (3.3.5)*, we could also scale the estimators to align the MMDs with a quantile of the chi-square distribution as in Rousseeuw and Van Driessen (1999). Across the simulations and the examples considered in this paper, we have only seen very slight changes in the resulting estimators for both the raw and reweighted MMCD.*

B.4 Shapley Proofs

Proof of Proposition 3.5.2.1. To show that cellwise Shapley values are not matrix affine equivariant, we consider a rowwise addition matrix $\boldsymbol{A}$ that adds the w-th row to the v-th row. For simplicity, let $\boldsymbol{B}$ be the identity matrix. Then Equation (B.21) yields

$$((\boldsymbol{AX}) \circ (\boldsymbol{CY}))_{jk} = \begin{cases} (x_{jk} + x_{wk})(y_{jk} - y_{wk}) & j = v \\ x_{jk}y_{jk} & j \neq v \end{cases}$$

while

$$(\boldsymbol{A}(\boldsymbol{X} \circ \boldsymbol{Y}))_{jk} = \begin{cases} x_{jk}y_{jk} + x_{wk}y_{wk} & j = v \\ x_{jk}y_{jk} & j \neq v \end{cases}.$$

Hence, we do not get invariance nor equivariance for rowwise or columnwise addition matrices. This also implies that the cellwise Shapley values are not, in general, matrix affine equivariant.

Shift invariance follows from

$$\boldsymbol{\Phi}(\boldsymbol{X}+\boldsymbol{C}) = ((\boldsymbol{X}+\boldsymbol{C}) - (\mathbf{M}+\boldsymbol{C})) \circ \boldsymbol{\Omega}^{\text{row}}((\boldsymbol{X}+\boldsymbol{C}) - (\mathbf{M}+\boldsymbol{C}))\boldsymbol{\Omega}^{\text{col}} = \boldsymbol{\Phi}(\boldsymbol{X}),$$

which means that we can assume that $\boldsymbol{X}$ has zero mean without loss of generality.

Let $\boldsymbol{Y} := \boldsymbol{\Omega}^{\text{row}}\boldsymbol{X}\boldsymbol{\Omega}^{\text{col}}$, $\boldsymbol{C} := (\boldsymbol{A}')^{-1}$ and $\boldsymbol{D} := (\boldsymbol{B}')^{-1}$, then we can write the cellwise Shapley values as $\boldsymbol{\Phi}(\boldsymbol{AXB}) = (\boldsymbol{AXB}) \circ (\boldsymbol{CYD})$. The jk-th entry of this matrix can be written as

$$\begin{aligned}
\phi_{jk}(\boldsymbol{AXB}) =& ((\boldsymbol{AXB}) \circ (\boldsymbol{CYD}))_{jk} = (\boldsymbol{AXB})_{jk}(\boldsymbol{CYD})_{jk} \\
=& \sum_{i=1}^{p}\sum_{l=1}^{q} a_{ji}x_{il}b_{lk} \sum_{m=1}^{p}\sum_{n=1}^{q} c_{jm}y_{mn}d_{nk} \\
=& \sum_{i=1}^{p}\sum_{l=1}^{q}\sum_{m=1,m\neq i}^{p}\sum_{n=1,n\neq l}^{q} a_{ji}c_{jm}x_{il}y_{mn}b_{lk}d_{nk} \\
&+ \sum_{i=1}^{p}\sum_{l=1}^{q}\sum_{n=1,n\neq l}^{q} a_{ji}c_{ji}x_{il}y_{in}b_{lk}d_{nk} \\
&+ \sum_{i=1}^{p}\sum_{l=1}^{q}\sum_{m=1,m\neq i}^{p} a_{ji}c_{jm}x_{il}y_{ml}b_{lk}d_{lk} \\
&+ \sum_{i=1}^{p}\sum_{l=1}^{q} a_{ji}c_{ji}x_{il}y_{il}b_{lk}d_{lk}.
\end{aligned} \tag{B.21}$$

If $\boldsymbol{A}$ is a scaling matrix, i.e., a diagonal matrix with non-zero entries, we have that

$$a_{ji}c_{jm} = \begin{cases} 1 & j = i = m \\ 0 & \text{otherwise} \end{cases},$$

and similarly for $\boldsymbol{B}$. This implies that

$$\phi_{jk}(\boldsymbol{AXB}) = x_{jk}y_{jk} = (\boldsymbol{X} \circ \boldsymbol{Y})_{jk} = \phi_{jk}(\boldsymbol{X}),$$

showing the scale invariance.

If $\boldsymbol{A}$ is a permutation matrix, i.e., a matrix consisting of any permutation of the canonical basis vectors, we have that $(\boldsymbol{A}')^{-1} = \boldsymbol{A}$ and

$$a_{ji}c_{jm} = a_{ji}a_{jm} = \begin{cases} a_{ji} & i = m \\ 0 & i \neq m \end{cases},$$

and similarly for $\boldsymbol{B}$. Hence Equation (B.21) becomes

$$((\boldsymbol{AXB}) \circ (\boldsymbol{CYD}))_{jk} = \sum_{i=1}^{p}\sum_{l=1}^{q} a_{ji}c_{ji}x_{il}y_{il}b_{lk}d_{lk} = (\boldsymbol{A}(\boldsymbol{X} \circ \boldsymbol{Y})\boldsymbol{B})_{jk},$$

verifying the permutation equivariance. □

Proof of Theorem 3.5.2.2. To show that the computation of the rowwise Shapley value can be simplified, we start by rewriting the rowwise marginal contributions to the matrix Mahalanobis distance.

$$
\begin{aligned}
\Delta_a \operatorname{MMD}(\hat{\boldsymbol{X}}^S) &:= \operatorname{MMD}(\hat{\boldsymbol{X}}^{S\cup\{a\}}) - \operatorname{MMD}(\hat{\boldsymbol{X}}^S) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{q}\sum_{l=1}^{q}(\hat{x}_{ik}^{S\cup\{a\}} - m_{ik})(\hat{x}_{jl}^{S\cup\{a\}} - m_{jl})\omega_{lk}^{\text{col}}\omega_{ij}^{\text{row}} \\
&\quad - \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{q}\sum_{l=1}^{q}(\hat{x}_{ik}^{S} - m_{ik})(\hat{x}_{jl}^{S} - m_{jl})\omega_{lk}^{\text{col}}\omega_{ij}^{\text{row}} \\
&= \sum_{i\in S\cup\{a\}}\sum_{j\in S\cup\{a\}}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ik} - m_{ik})(x_{jl} - m_{jl})\omega_{lk}^{\text{col}}\omega_{ij}^{\text{row}} \\
&\quad - \sum_{i\in S}\sum_{j\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ik} - m_{ik})(x_{jl} - m_{jl})\omega_{lk}^{\text{col}}\omega_{ij}^{\text{row}} \\
&= \sum_{i\in S\cup\{a\}}\sum_{j\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ik} - m_{ik})(x_{jl} - m_{jl})\omega_{lk}^{\text{col}}\omega_{ij}^{\text{row}} \\
&\quad + \sum_{i\in S\cup\{a\}}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ik} - m_{ik})(x_{al} - m_{al})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}} \\
&\quad - \sum_{i\in S}\sum_{j\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ik} - m_{ik})(x_{jl} - m_{jl})\omega_{lk}^{\text{col}}\omega_{ij}^{\text{row}} \\
&= \sum_{i\in S}\sum_{j\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ik} - m_{ik})(x_{jl} - m_{jl})\omega_{lk}^{\text{col}}\omega_{ij}^{\text{row}} \\
&\quad - \sum_{i\in S}\sum_{j\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ik} - m_{ik})(x_{jl} - m_{jl})\omega_{lk}^{\text{col}}\omega_{ij}^{\text{row}} \\
&\quad + \sum_{j\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ak} - m_{ak})(x_{jl} - m_{jl})\omega_{lk}^{\text{col}}\omega_{aj}^{\text{row}} \\
&\quad + \sum_{i\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ik} - m_{ik})(x_{al} - m_{al})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}} \\
&\quad + \sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ak} - m_{ak})(x_{al} - m_{al})\omega_{lk}^{\text{col}}\omega_{aa}^{\text{row}} \\
&= 2\sum_{i\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{al} - m_{al})(x_{ik} - m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}} \\
&\quad + \sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ak} - m_{ak})(x_{al} - m_{al})\omega_{lk}^{\text{col}}\omega_{aa}^{\text{row}}.
\end{aligned}
$$

Now the coordinates $\phi_a(\boldsymbol{X})$ of the Shapley value $\boldsymbol{\phi}(\boldsymbol{X})$ are given by ($w(|S|) = \frac{|S|!(p-|S|-1)!}{p!}$)

$$\begin{aligned}\phi_a(\boldsymbol{X}) = &\sum_{S\subseteq P\setminus\{a\}} w(|S|)\Delta_a \operatorname{MMD}(\hat{\boldsymbol{X}}^S)\\ =&2\sum_{S\subseteq P\setminus\{a\}} w(|S|)\sum_{i\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}\\ &+\sum_{S\subseteq P\setminus\{a\}} w(|S|)\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ak}-m_{ak})(x_{al}-m_{al})\omega_{lk}^{\text{col}}\omega_{aa}^{\text{row}}\end{aligned}$$

and we can simplify the first term of the sum as

$$\begin{aligned}&2\sum_{S\subseteq P\setminus\{a\}} w(|S|)\sum_{i\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}\\ &=2\sum_{s=1}^{p-1} w(|S|)\sum_{S\subseteq P\setminus\{a\},|S|=s}\sum_{i\in S}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}\\ &=2\sum_{s=1}^{p-1}\sum_{k=1}^{q}\sum_{l=1}^{t} w(|S|)\sum_{S\subseteq P\setminus\{a\},|S|=s}\sum_{i\in S}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}\\ &=2\sum_{s=1}^{p-1}\sum_{k=1}^{q}\sum_{l=1}^{q}\frac{|S|!(p-|S|-1)!}{p!}\binom{p-2}{s-1}\sum_{i\in P\setminus\{a\}}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}\\ &=2\frac{1}{p(p-1)}\sum_{s=1}^{p-1}s\sum_{k=1}^{q}\sum_{l=1}^{q}\sum_{i\in P\setminus\{a\}}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}\\ &=2\frac{1}{p(p-1)}\frac{p(p-1)}{2}\sum_{k=1}^{q}\sum_{l=1}^{q}\sum_{i\in P\setminus\{a\}}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}\\ &=\sum_{k=1}^{q}\sum_{l=1}^{q}\sum_{i\in P\setminus\{a\}}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}.\end{aligned}$$

Since the second term is independent of the subset S and $\sum_{S\subseteq P\setminus\{a\}} w(|S|) = 1$, we obtain

$$\begin{aligned}\phi_a(\boldsymbol{X}) =&\sum_{k=1}^{q}\sum_{l=1}^{q}\sum_{i\in P\setminus\{a\}}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}}\\ &+\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{ak}-m_{ak})(x_{al}-m_{al})\omega_{lk}^{\text{col}}\omega_{aa}^{\text{row}}\\ =&\sum_{i=1}^{p}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{al}-m_{al})(x_{ik}-m_{ik})\omega_{lk}^{\text{col}}\omega_{ia}^{\text{row}},\end{aligned}$$

which completes the proof. □

B.5 Further Simulation Results

In order to select a simulation setting, one has to consider that the ML estimators for the parameters of the matrix-variate normal distribution employ an iterative algorithm, which is commonly initialized by setting either the rowwise or columnwise covariance matrix equal to the identity matrix (Dutilleul, 1999). Therefore, identity covariance matrices will not be used for data generation as this could lead to an undesirable advantage for the estimation.

To assess the quality of covariance estimation, we consider two additional measures to the KL divergence: the relative Frobenius error given as

$$\frac{\left\| \hat{\boldsymbol{\Sigma}}^{\text{col}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}} - \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}} \right\|_F}{\left\| \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}} \right\|_F},$$

and angle error between eigenvalues given as

$$1 - \frac{\hat{\boldsymbol{a}}^\top \boldsymbol{a}}{\sqrt{\hat{\boldsymbol{a}}^\top \hat{\boldsymbol{a}}} \sqrt{\boldsymbol{a}^\top \boldsymbol{a}}},$$

where $\hat{\boldsymbol{a}}$ and $\boldsymbol{a}$ are the vectors of sorted eigenvalues of $\hat{\boldsymbol{\Sigma}}^{\text{col}} \otimes \hat{\boldsymbol{\Sigma}}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}$, respectively. Large values of the KL divergence and the relative Frobenius error indicate difficulties in the estimation of the covariances. The angle error between the eigenvalues is in the interval $[0, 1]$, and a large value means that the shape of the covariance matrix is not appropriately estimated. To assess the efficacy of outlier detection, we include the F-score in addition to precision and recall. The F-score is defined as the harmonic mean of precision and recall, where precision denotes the proportion of correctly identified outliers among all detected samples, while recall represents the proportion of correctly identified outliers among all contaminated samples. The `R` code of the simulations and all simulation results are available in the online supplement.

Effects of Dimensionality and Computation Time

We start by considering additional metrics for the simulations discussed in Section 3.6. Figure B.3 shows the F-score in addition to precision and recall. The F-score shows that for $n = 100$ and increasing dimensionality the robust MMCD estimators and the MLEs yield similar results. This is due to an increasing recall of the MLEs and a decreasing precision of the MMCD estimators. For $n = 1000$, the F-score of the MMCD estimators is close to the benchmark and for the MCD we see the advantage of using the deterministic MCD approach over the Fast-MCD method with increasing sample size. In Figure B.4 we see that the MCD performs best in all settings across all evaluation measures. For the MCD we do not see a difference in the KL divergence when swapping to the deterministic procedure. However, the angle error between eigenvalues shows clear improvements, indicating that the estimation of the shape of the covariance matrix improves. Both in terms of the angle and Frobenius error, the MCD estimator attains better scores than the MLEs even for higher pq, while the MLEs have better KL divergence.

We also analyze the computation times of the estimators in this setting. Figure B.5 clearly shows that computation time depends on the dimensionality of the matrix-variate samples

and the number of samples for all approaches. The relative increases of computation time of the matrix MLEs and the MMCD estimators are similar for $n \in \{20, 100, 300\}$ but for $n = 1000$ the relative increase in computation time for the matrix MLEs is larger than for the MMCD estimators, highlighting the effectiveness of the subsampling approach with increasing sample size. For the MCD, we observe a decrease in computation time when $pq > 300$ since the deterministic MCD is used instead of the Fast-MCD procedure. However, computing the MCD still takes longer than the MMCD approach. Hence, the matrix-variate approach does yield higher robustness and more accurate covariance estimation with shorter computation times. Although parallel processing is available for the MMCD procedure, it was not utilized in the simulations to ensure better comparability for the algorithms. Depending on the number of available threads, parallel processing yields substantial improvements in computation time.

Cellwise and Block Contamination

We also consider the additional metrics for the simulations comparing the three different contamination types in Figures B.6 and B.7. The robustness of the MMCD estimators is again confirmed using all three metrics assessing the quality of the covariance estimation. The angle error reveals that the cell contamination has less effect on the shape of the covariance matrix than the other two scenarios and that all three estimators seemingly do a good job of estimating the covariance shape. For block contamination, the MCD yields better results than the MLEs with increasing sample size and even gets close to the MMCD in terms of angle error.

Remark B.5.1. *Our cell contamination setting does not correspond to the setting of cellwise outliers (Alqallaf et al., 2009). We first select a subset of outlying observations and permute the cells for this selection while Alqallaf et al. (2009) select a fraction of all cells from all samples. In our setting, we can guarantee that only 10 percent of the samples are contaminated while the cellwise contamination scheme of Alqallaf et al. (2009) would likely lead to more than half of the samples being contaminated.*

In further simulations, we considered different fractions of contaminated samples as well as multiple rowwise and columnwise covariance matrices for cellwise and block contamination. Additionally, we analyzed the effect of the fraction of permuted cells per observation for cell contamination, and for block contamination, we considered different mean matrices. Those simulation results are not discussed here but are available in the online supplement.

Shift Outliers

For shift outliers, we include an in-depth analysis of the effect of the various simulation parameters. The simulations involve generating regular and outlying samples from a matrix normal distribution. A fraction, ε, of the clean data is replaced by outliers. The clean observations are drawn from a centered distribution, while the mean of the outliers shifts based on the parameter γ, i.e., the mean of the outliers is set to a matrix with all entries equal to γ. Three types of covariance matrices are considered: The covariance matrix $\mathbf{\Sigma}^{\text{rnd}}$, as proposed by Agostinelli et al. (2015b), is randomly generated with low correlations. The

covariance matrix $\mathbf{\Sigma}^{\text{fix}}(0.7)$ induces a relatively collinear setting, with entries defined as:

$$\sigma_{jk}^{\text{fix}}(0.7) = \begin{cases} 1 & \text{if } j = k \\ 0.7 & \text{if } j \neq k \end{cases}.$$

The covariance matrix $\mathbf{\Sigma}^{\text{mix}}(0.7)$ exhibits both large and small correlations, featuring entries as follows:

$$\sigma_{jk}^{\text{mix}}(0.7) = \begin{cases} 1 & \text{if } j = k \\ 0.7^{|j-k|} & \text{if } j \neq k \end{cases}.$$

While maintaining the same covariance structure for both outliers and clean samples, we explore the impact of increasing the outlier covariance by scaling the covariance of clean observations by the parameter s. Each simulation setting is replicated 100 times. Unless specified otherwise, we set $\mathbf{\Sigma}^{\text{row}} = \mathbf{\Sigma}^{\text{rnd}}$, $\mathbf{\Sigma}^{\text{col}} = \mathbf{\Sigma}^{\text{mix}}(0.7)$, and $s = 1$ as detailed in Section 3.6. An overview of all parameters for the simulations is provided in Table B.1. For $(p, q) = (5, 20)$, all listed parameter combinations are considered, while for $(p, q) \in (50, 20), (100, 50)$, we only consider $s = 1$.

Parameter	Parameter values
Sample size n	$20, 50, 100, 200, 300, 400, 500, 750, 1000$
Contamination ε	$0.1, 0.2, 0.3, 0.4$
Rowwise covariance $\mathbf{\Sigma}^{\text{row}}$	$\mathbf{\Sigma}^{\text{fix}}(0.7), \mathbf{\Sigma}^{\text{rnd}}$
Columnwise covariance $\mathbf{\Sigma}^{\text{col}}$	$\mathbf{\Sigma}^{\text{mix}}(0.7), \mathbf{\Sigma}^{\text{rnd}}$
Mean shift γ	$1, 2, 3, 4, 5$
Covariance multiplier s	$1, 2, 3, 4$

Table B.1: Parameters considered for the simulations with $p, q = (5, 20)$.

We analyze the effect of the mean shift in a setting with contamination of $\varepsilon = 0.2$ and compare $\gamma = 1$ and $\gamma = 3$. In the upper row of Figure B.8, the boxplots depict F-scores across various parameter configurations. Notably, the MMCD estimators exhibit improved performance as sample sizes increase across all settings, consistently outperforming ML estimators. However, for $(p, q) = (5, 20)$, in a scenario involving a minor mean shift, the F-scores derived from MMCD exhibit some volatility with larger sample sizes. This situation arises due to the proximity of outliers to regular observations, posing challenges in their identification. Notably, a more pronounced mean shift significantly simplifies outlier detection. Moreover, we see that the recall of the MMCD estimators is close to one across all settings, except for $(p, q) = (5, 20)$ and a small mean shift. The MLE estimators only detect the most severe outliers due to the masking effect, leading to a median recall below 0.25 across all settings. With an increasing sample size, the precision of the MMCD is improving and has very low variability. On the other hand, the MLE shows very unstable results.

Figure B.9 presents the scores depicting covariance estimation. For the MMCD estimators the covariance estimation performance is improving with the sample size across all settings. On the other hand, the sample size has a negligible effect on the quality of the MLE

estimators in the presence of outliers and a larger mean shift decreases performance. For small sample sizes, MLE and MMCD estimators are close in terms of KL divergence, but the angle error and Frobenius error indicate worse performance of the MLE estimators also for small sample sizes. The relative Frobenius error of MMCD estimators is smaller than one and thus only plotted on $[0, 1]$. For the MLE estimators, it is often above one and those settings are not visible in plots.

Figure B.10 shows the difference between a contamination of $\varepsilon = 0.1$ and $\varepsilon = 0.4$ with mean shift $\gamma = 1$. The KL divergence reveals that the MMCD estimator yields more accurate results across all settings. However, for $\varepsilon = 0.1$, the F-scores of the MLE are increasing with the dimensionality and perform better than the MMCD for small sample sizes. For $\varepsilon = 0.4$, only the MMCD yields reliable results.

For the setting with $(p, q) = (5, 20)$ and $\varepsilon = 0.2$, we also computed the MCD on the vectorized samples in addition to the matrix MLE and MMCD and considered the true mean and covariance used to generate the data as a benchmark. Figures B.11 and B.12 summarize the results and reveal that the MCD on the vectorized observations does not lead to robust estimators. This issue arises because the robustness of the MCD and MMCD depends on the dimensionality of the data. For the MCD it depends on $p \cdot q$ and for the MMCD it depends on $p/q + q/p$. To achieve a 99% probability of obtaining at least one clean initial subset with $(p, q) = (5, 20)$ and a contamination $\varepsilon = 0.2$, MCD requires approximately $2.8 \cdot 10^{10}$ initial subsets, while MMCD only needs 16. For the setting with the smallest mean shift, the comparison between MMCD and the actual parameters in Figure B.12 highlights the difficulty of this setting since even using the actual parameters; the recall shows a lot of variability.

In addition to shifting the mean of the outliers by $\gamma \in \{1, \dots, 5\}$, we now consider the effect of scaling the covariance by $s \in \{1, \dots, 4\}$. The difference between $s \in \{2, 3, 4\}$ was negligible and Figures B.13 and B.14 summarize the results for $s = 2$. While the MLE performs quite well for outlier detection, especially compared to the setting with $s = 1$ (see Figure B.10), the estimated covariance matrices are not accurate. The overall performance of the MCD computed on the vectorized samples improves with increasing sample size n but even more samples would be necessary to obtain similar results to the MMCD.

Finally, we compare the 4 different combinations of row- and columnwise covariance matrices with $\varepsilon = 0.2$. In Figure B.15 we use $\gamma = 1$ and in Figure B.16 we increase the mean shift to $\gamma = 5$. The F-score based on the true parameters is included as a reference. When $\mathbf{\Sigma}^{\text{row}} = \mathbf{\Sigma}^{\text{fix}}(0.7)$, $\mathbf{\Sigma}^{\text{col}} = \mathbf{\Sigma}^{\text{mix}}(0.7)$, and $\gamma = 1$, the mean shift is too small and the outliers cannot be separated from the regular observations. Increasing the mean shift to $\gamma = 5$, the separation becomes clearer and the MMCD yields robust results. If $\gamma = 1$, we still see a lot of variability in the F-score if only the rowwise or columnwise covariance matrix is generated randomly. However, if both are generated randomly the distinction between outliers and regular observations is easier.

Effects of Fine-grained Mean Shifts

To get a more in-depth view of the effect of the mean shift we consider a finer grid for the parameter $\gamma \in \{0.1, 0.2, \dots, 2\}$ for $n \in \{20, 100, 1000\}$, $(p, q) = (5, 20)$, $\varepsilon = 0.1$. In Figure B.17, we see that for $n = 20$, the MMCD has a low precision but an even higher

recall than we can achieve using the actual parameters used to generate the data to compute the Mahalanobis distances for outlier detection. For larger sample sizes, the precision of the MMCD increases while the recall remains high, resulting in an F-score close to the one achieved by the actual parameters. For $n = 1000$, we also computed the MCD on the vectorized observations, it attains a higher recall than the matrix MLEs but lower precision and performs worse than the MMCD in all settings. Likewise, to the results for outlier detection, Figure B.18 shows similar results for covariance estimation. While the MMCD performs best in most settings it shows potential for improvement for small n and γ. The simulations also show that at a level of 10 percent contamination, even a small shift γ of the outliers negatively impacts covariance estimation and, consequently, outlier detection due to the masking effect.

Beyond Normality, Contaminated t-distribution

To analyze the effect of deviations from the matrix normal distribution we consider samples from a matrix t-distribution. Similar to the matrix normal distribution, the matrix t-distribution is parameterized by a mean matrix, rowwise and columnwise covariance matrices, and degrees of freedom as an additional parameter, see Gupta and Nagar (1999) for more details. We also consider the ML estimators for the matrix t-distribution proposed by Thompson et al. (2020), which are implemented in the `R` package `MixMatrix`. We consider samples from a $p \times q = 5 \times 20$ centered matrix t-distribution with $\nu \in \{1, \dots, 30\}$ degrees of freedom with $\mathbf{\Sigma}^{\text{row}} = \mathbf{\Sigma}^{\text{rnd}} \in \text{PDS}(p)$ and $\mathbf{\Sigma}^{\text{col}} = \mathbf{\Sigma}^{\text{mix}}(0.7) \in \text{PDS}(q)$ for $n \in \{20, 100, 1000\}$, $(p, q) = (5, 20)$, $\varepsilon \in \{0.1, 0.2\}$. The outliers are generated from a shifted distribution with a mean matrix of all ones with the same covariance structure and the same degrees of freedom. In Figure B.19, we analyze the influence of the degrees of freedom on precision, recall, angle error between eigenvalues, and the logarithm of the relative Frobenius error for various estimators, number of samples, and levels of contamination. The angle and Frobenius error clearly show the advantage of the MMCD estimators for covariance estimation. If the distribution of the samples is known, the consistency correction outlined in Theorem 3.3.0.2 allows us to obtain consistency for any matrix elliptical distribution. Since we do not know the underlying distribution in practice, we use the consistency factor for the normal model given in Equation (3.3.5) which does affect the scale of the covariance but not the shape. This is also reflected in the difference between the angle and Frobenius error of the MMCD estimators and MLEs for the matrix t-distribution since the scale of the covariance has a more profound impact on the Frobenius error. In terms of angle error, the MMCD estimators perform better than the MLEs for the matrix t-distribution for all degrees of freedom ν while the MMCD shows high Frobenius errors for $\nu \leq 4$.

While the MMCD estimators and MLEs for the matrix t-distribution have a recall close to one in all settings, we see a difference in precision depending on the fraction of contaminated samples and the number of samples. For $\varepsilon = 0.1$, the MLEs for the matrix t-distribution show a steep increase in precision with rising degrees of freedom for all the sample sizes. On the other hand, for $\varepsilon = 0.2$, the precision is constant and low for all n. Similarly to the simulations based on the normal model, the precision of the MMCD estimators is low for $n = 20$ and remains low for increasing degrees of freedom. While the precision increases alongside the degrees of freedom for larger sample sizes it is still low. However, this is what

we would expect since the matrix t-distribution has heavier tails than the matrix normal distribution and the mean shift is rather small, such that we do not see the full potential of the MMCD estimators even under the normal model, see Section B.5.

Both the normal MLEs and the MCD estimators computed on the vectorized samples perform poorly for covariance estimation and outlier detection when the samples are generated from a matrix t-distribution.

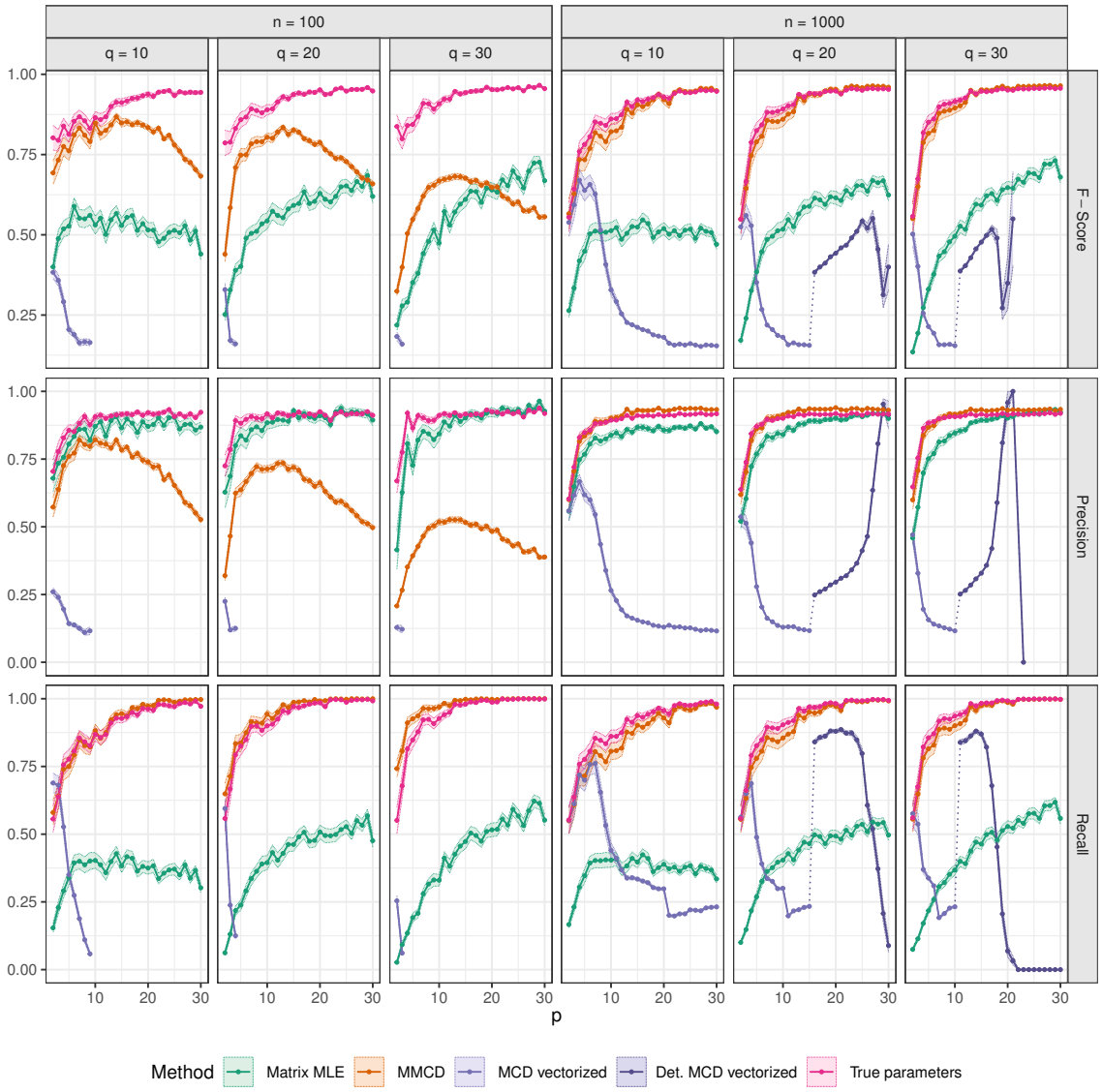


Figure B.3: Outlier detection capabilities comparing multiple matrix sizes $p \in \{2, \ldots, 30\}$ and $q \in \{10, 20, 30\}$ for $n \in \{100, 1000\}$, $\gamma = 1$, $\varepsilon = 0.1$.

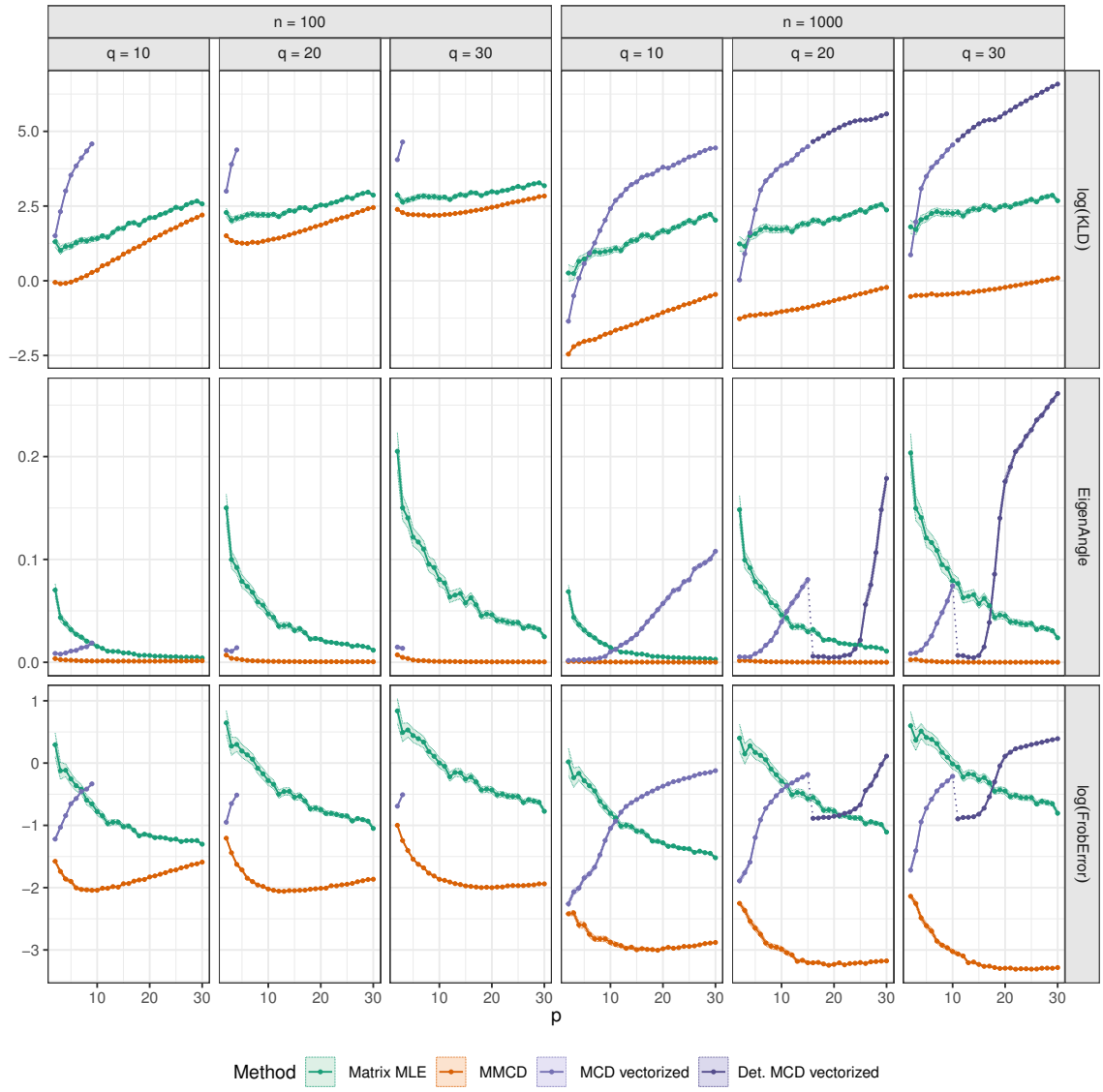


Figure B.4: Quality of covariance estimation comparing multiple matrix sizes $p \in \{2, \ldots, 30\}$ and $q \in \{10, 20, 30\}$ for $n \in \{100, 1000\}$, $\gamma = 1$, $\varepsilon = 0.1$.

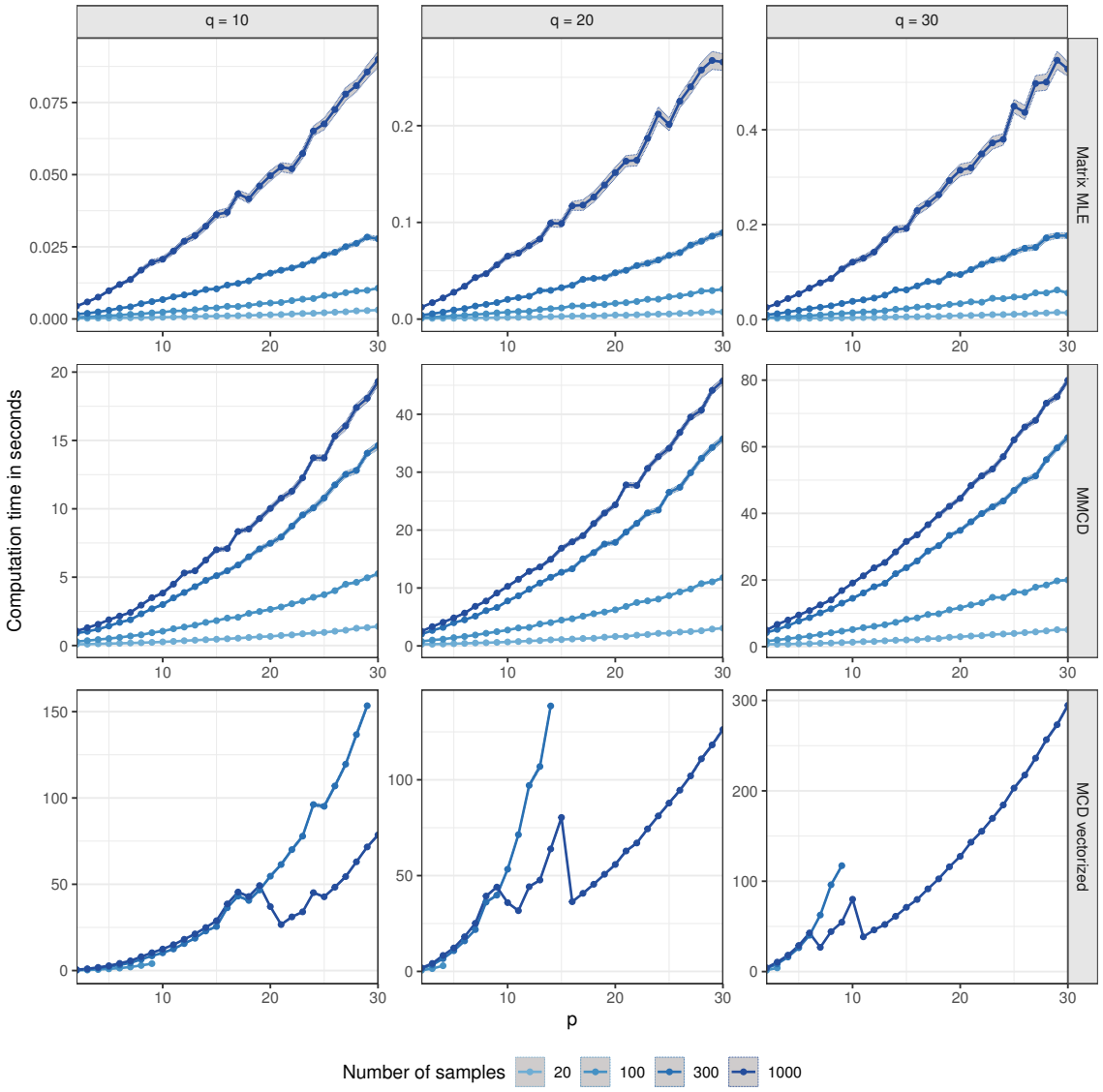


Figure B.5: Comparison of computation time in seconds for multiple matrix sizes $p \in \{2, \ldots, 30\}$ and $q \in \{10, 20, 30\}$ for $n \in \{20, 100, 300, 1000\}$.

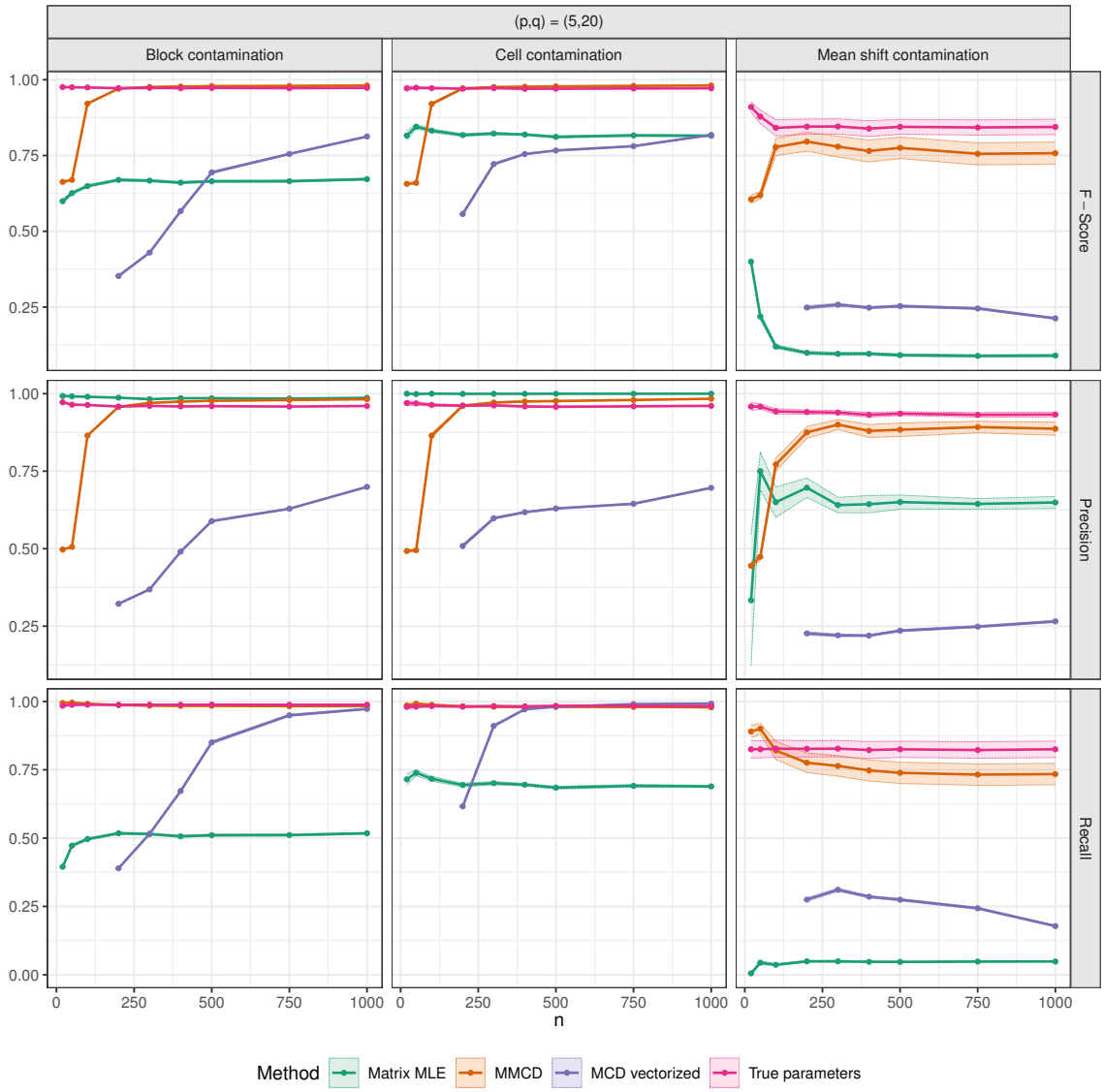


Figure B.6: Quality of covariance estimation comparing block, cell, and sample contamination.

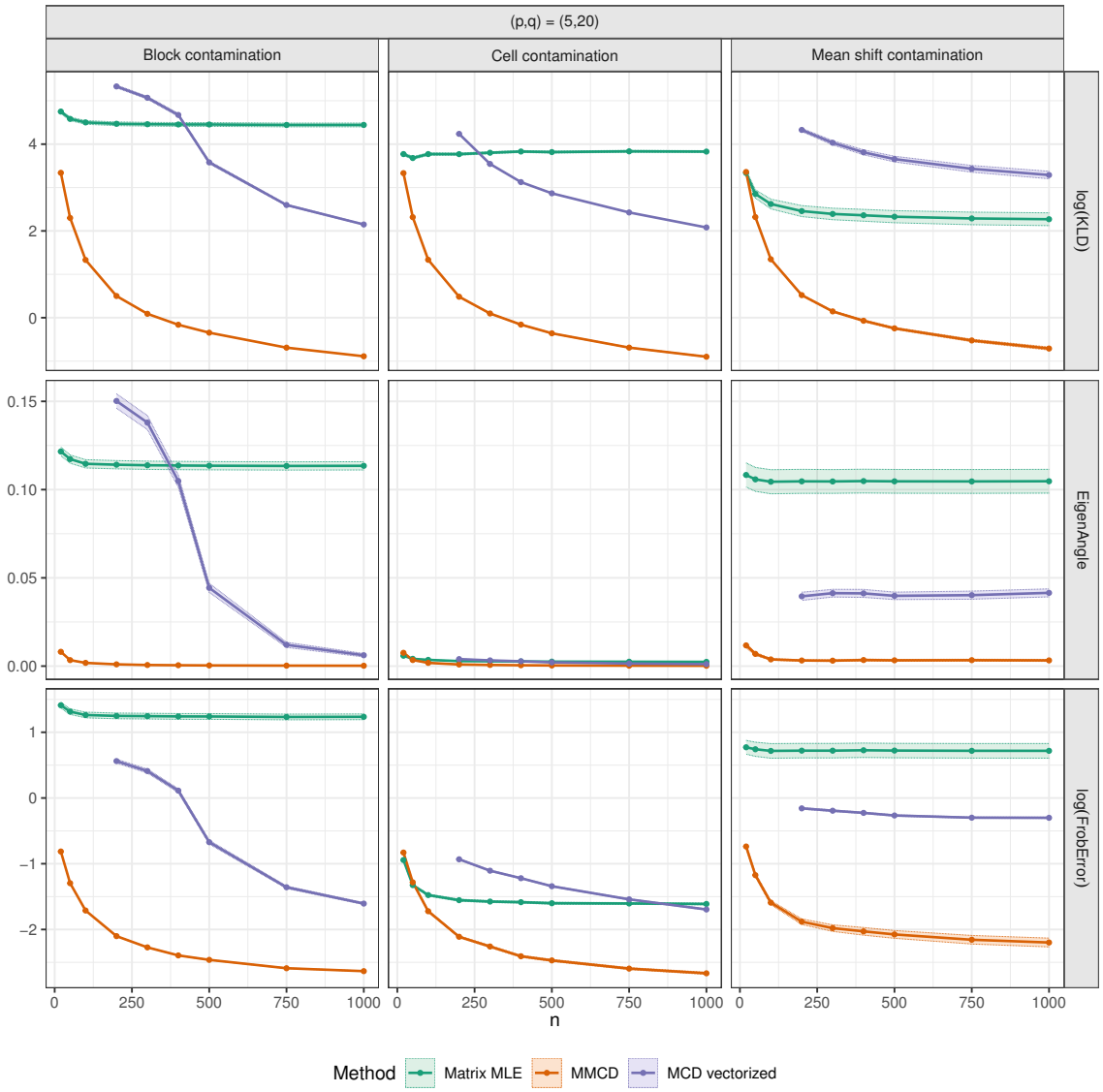


Figure B.7: Outlier detection capabilities comparing block, cell, and sample contamination.

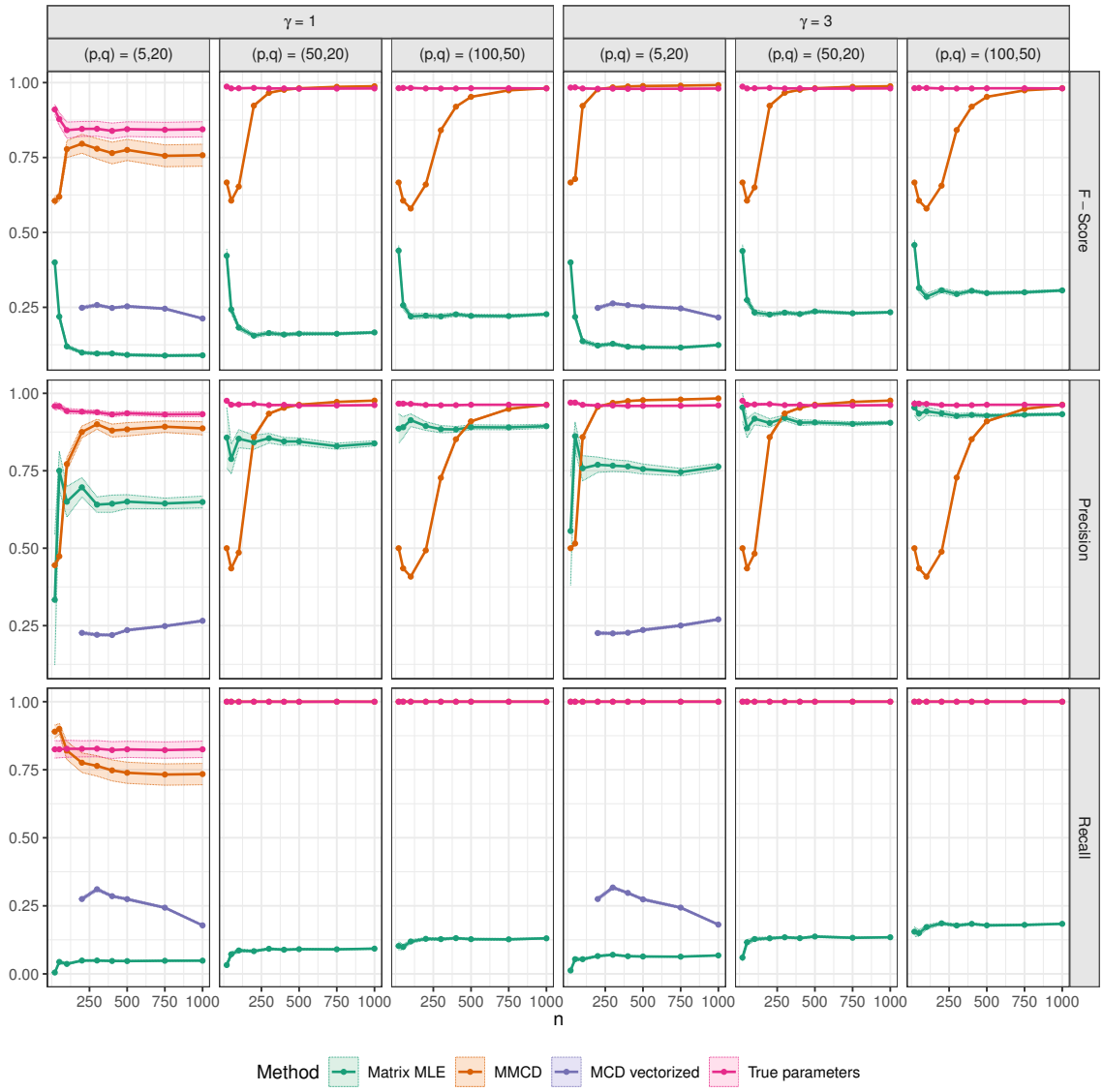


Figure B.8: Overview of simulation results with a fraction $\varepsilon = 0.2$ of contaminated samples. The outlier detection capabilities are measured by F-score, precision, and recall.

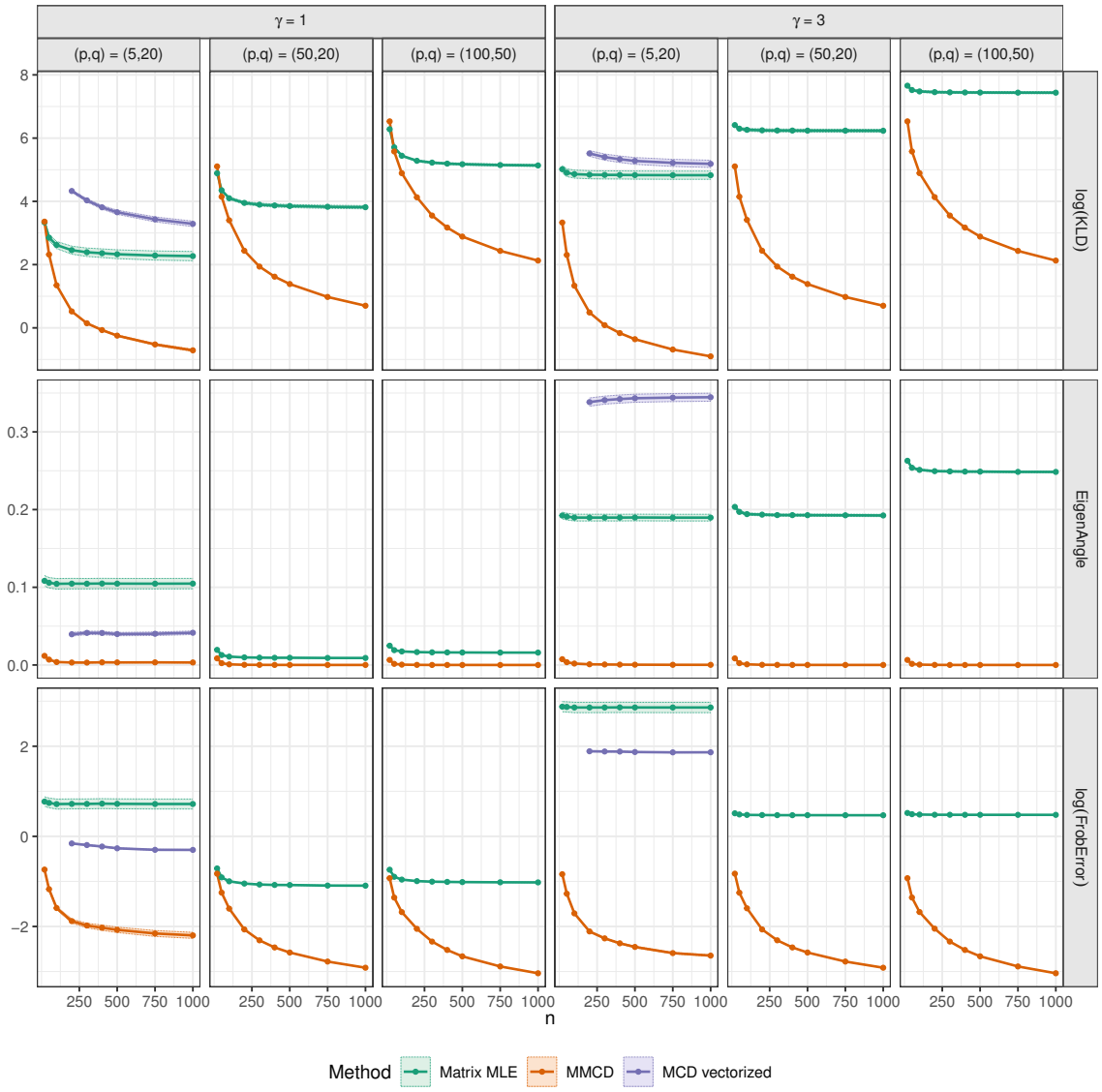


Figure B.9: Overview of simulation results with a fraction $\varepsilon = 0.2$ of contaminated samples. The quality of covariance estimation is evaluated based on the logarithm of KL divergence, angle error between eigenvalues, and the logarithm of relative Frobenius error.

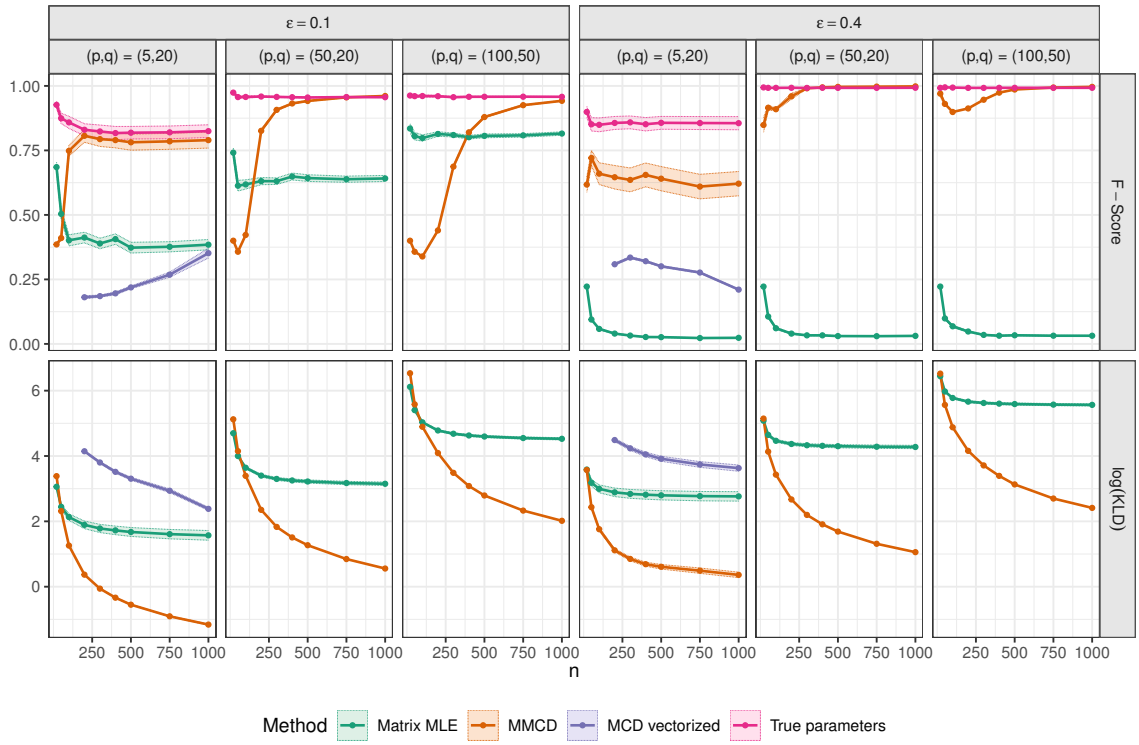


Figure B.10: F-score and logarithm of KL divergence for simulations with mean shift $\gamma = 1$.

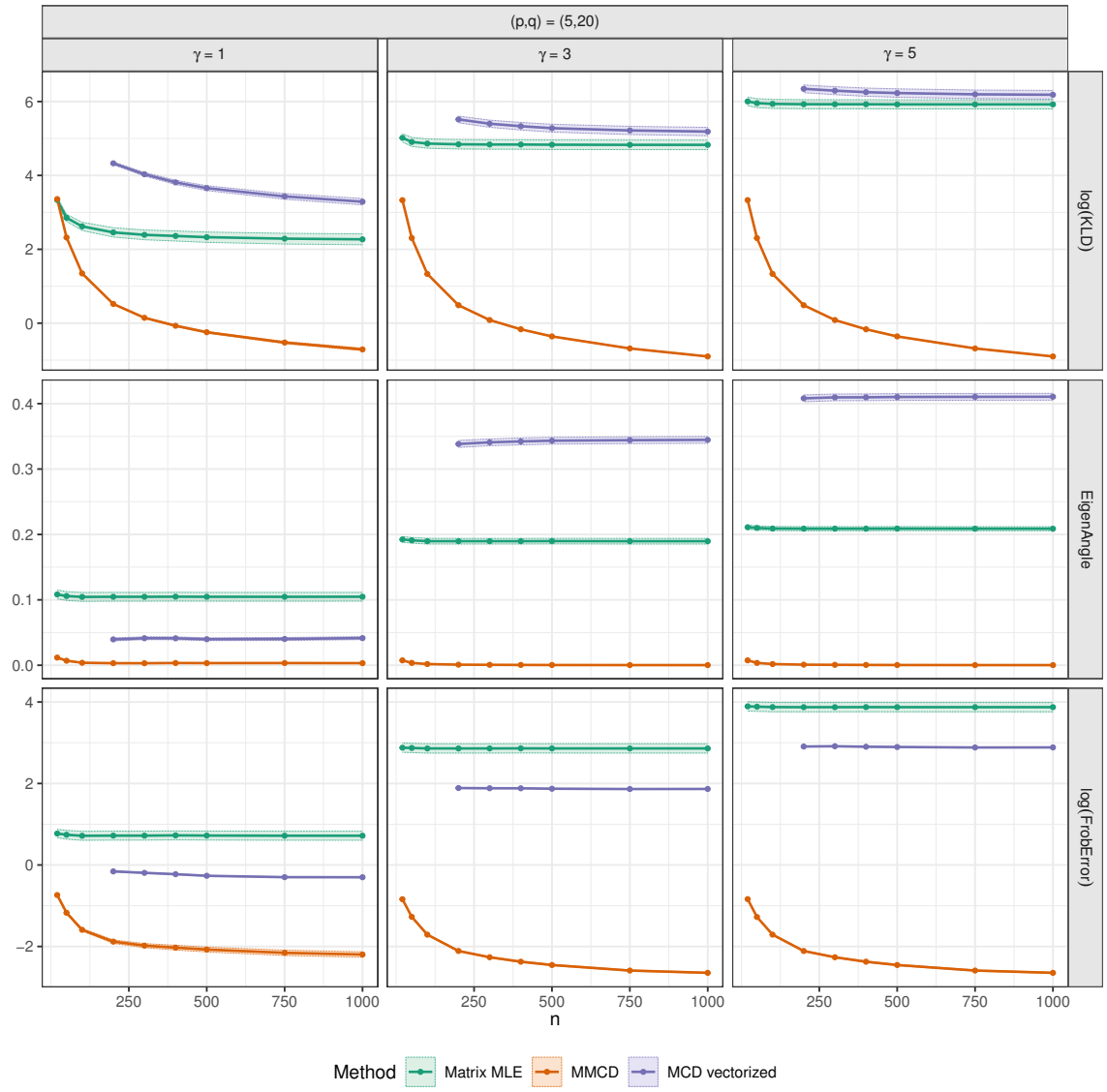


Figure B.11: Quality of covariance estimation for simulations with $\varepsilon = 0.2$ and $(p, q) = (5, 20)$.

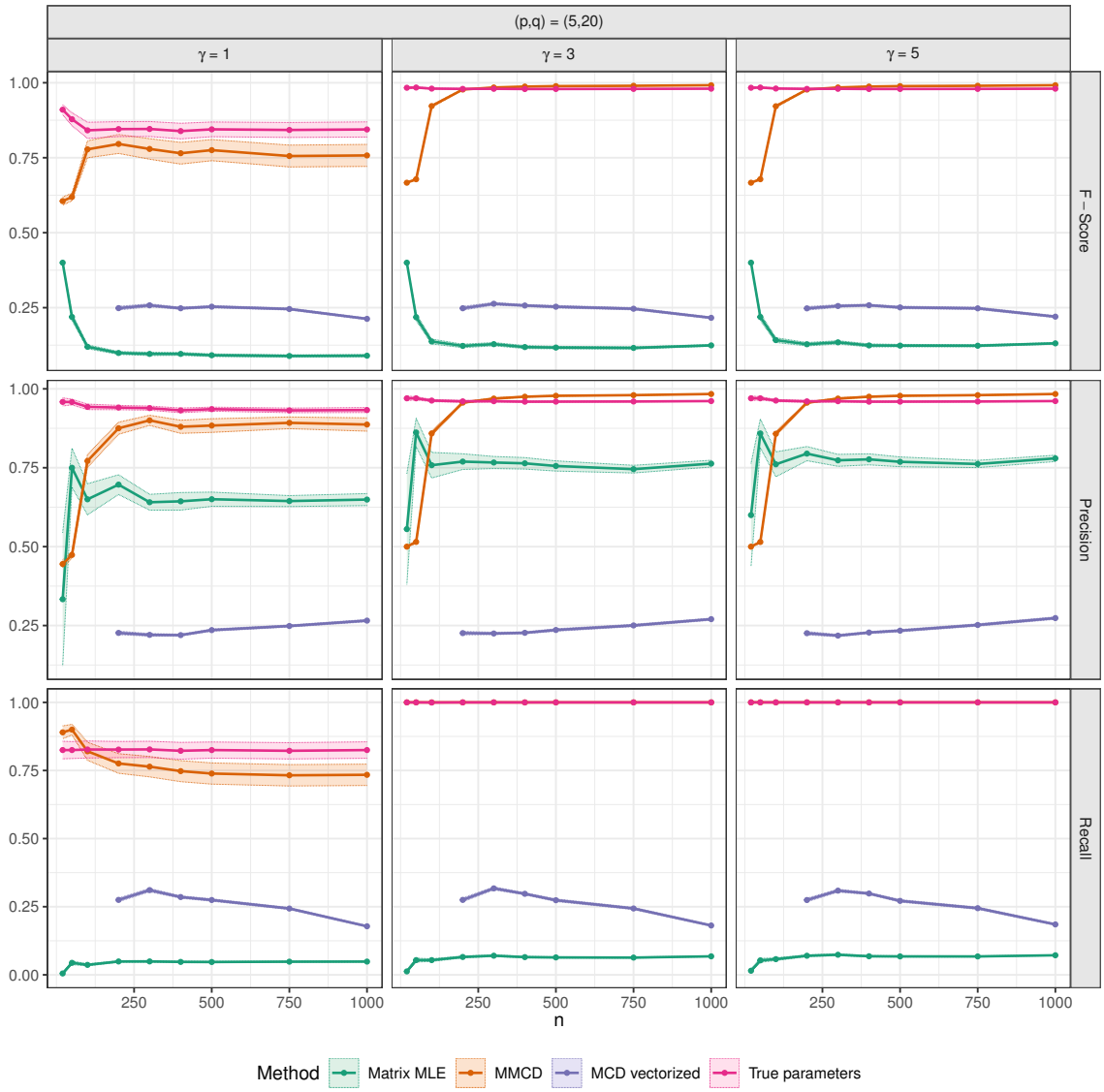


Figure B.12: Outlier detection capabilities for simulations with $\varepsilon = 0.2$ and $(p, q) = (5, 20)$.

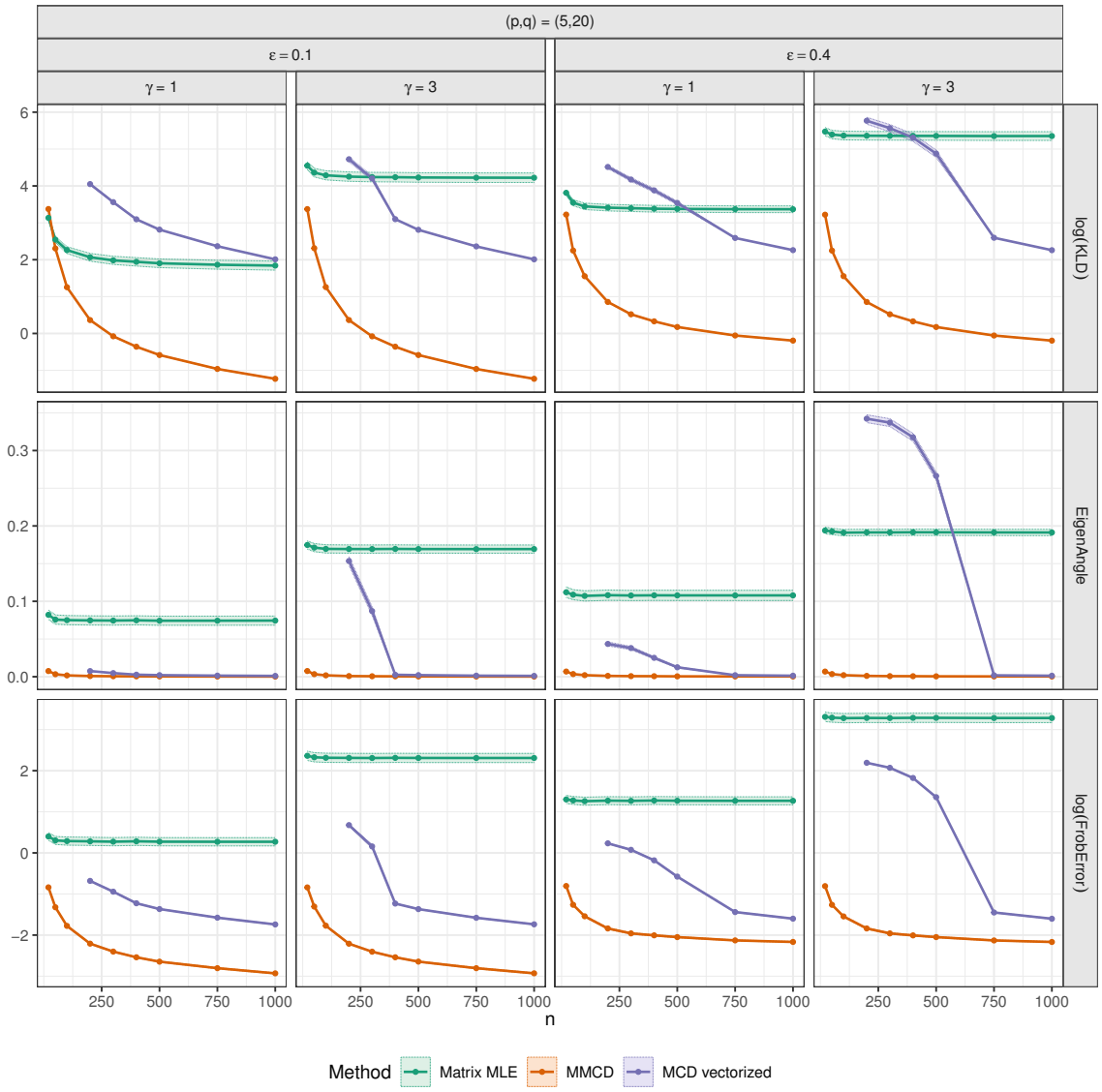


Figure B.13: Quality of covariance estimation for simulations where the covariance of the outliers is scaled by $s = 2$.

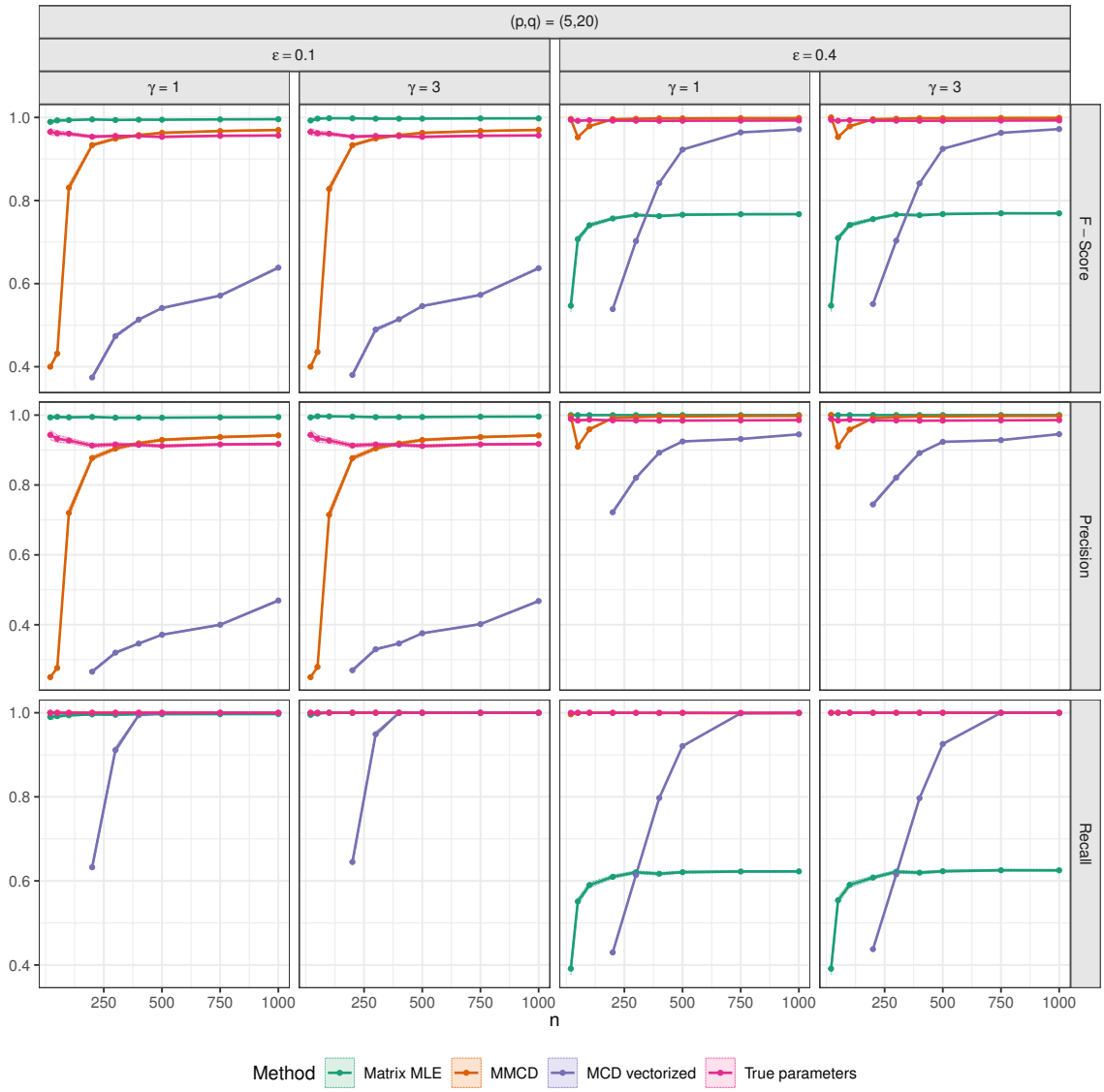


Figure B.14: Outlier detection capabilities for simulations where the covariance of the outliers is scaled by $s = 2$.

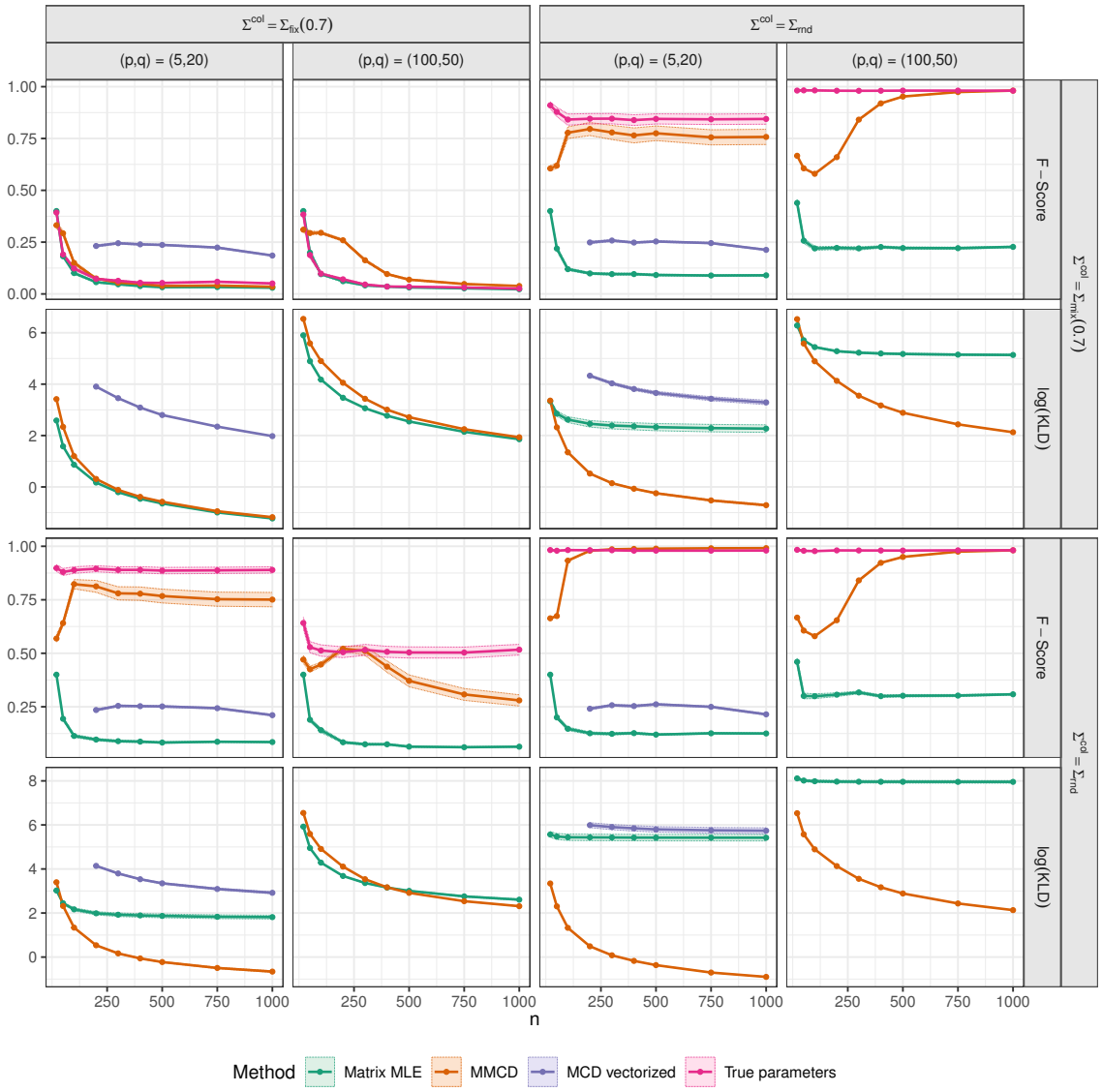


Figure B.15: F-score and logarithm of KL divergence comparing 4 different combinations of row- and columnwise covariance matrices, $\gamma = 1$, and $\varepsilon = 0.2$.

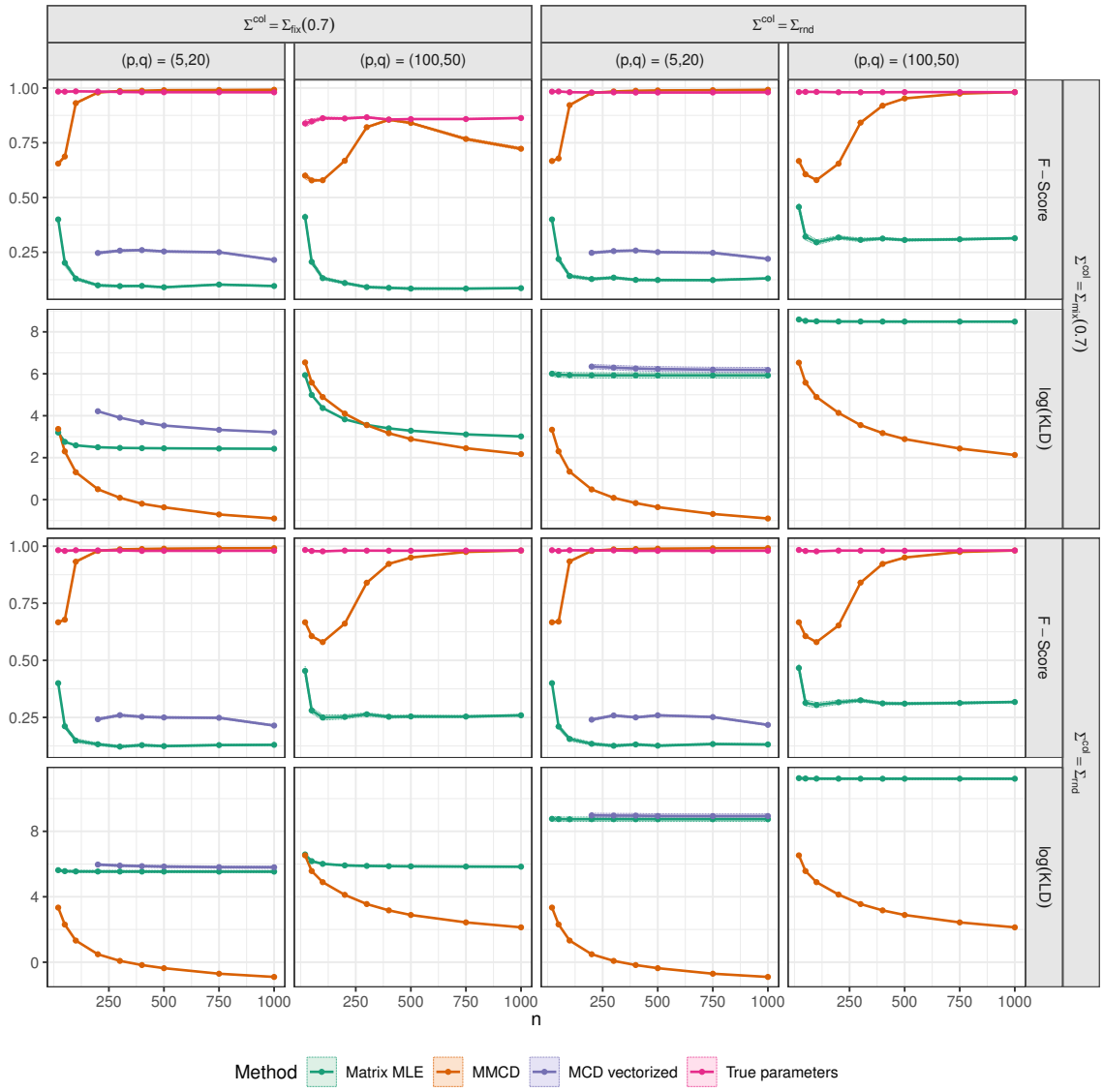


Figure B.16: F-score and logarithm of KL divergence comparing 4 different combinations of row- and columnwise covariance matrices, $\gamma = 5$, and $\varepsilon = 0.2$.

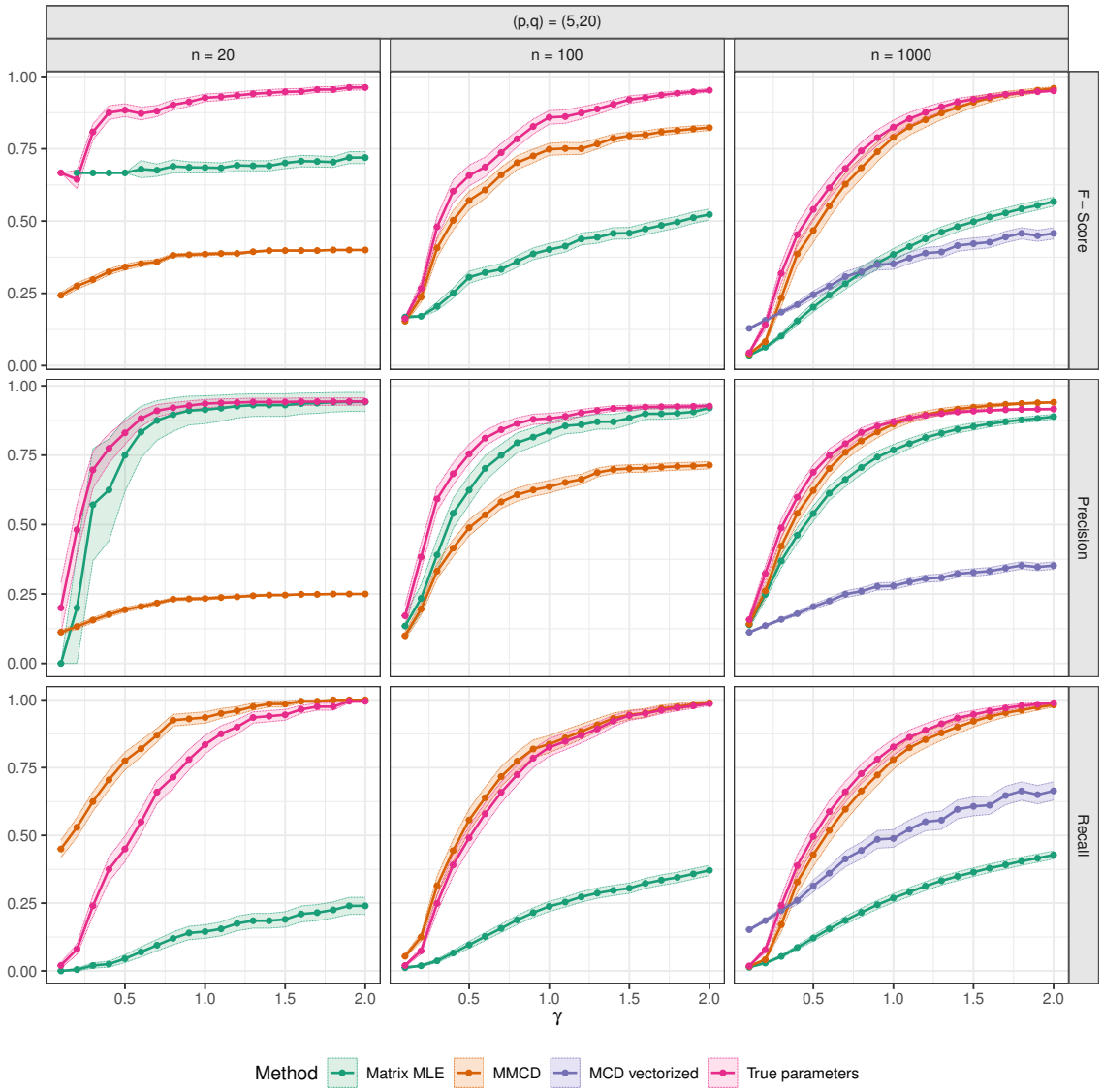


Figure B.17: Outlier detection capabilities for simulations with mean shift $\gamma \in \{0.1, 0.2, \ldots, 2\}$ for $n \in \{20, 100, 1000\}$, $\varepsilon = 0.1$.

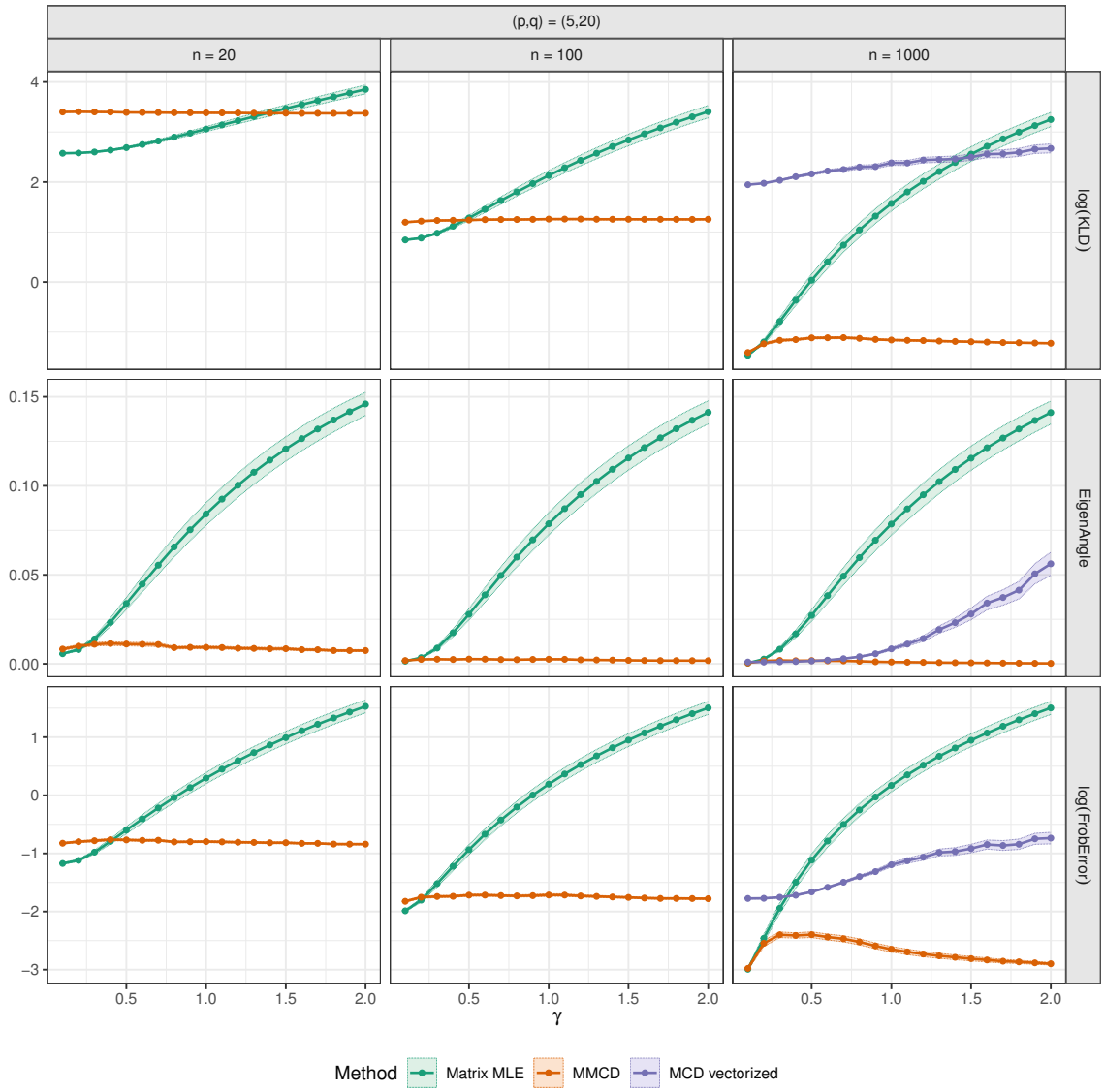


Figure B.18: Quality of covariance estimation for simulations with mean shift $\gamma \in \{0.1, 0.2, \ldots, 2\}$ for $n \in \{20, 100, 1000\}$, $\varepsilon = 0.1$.

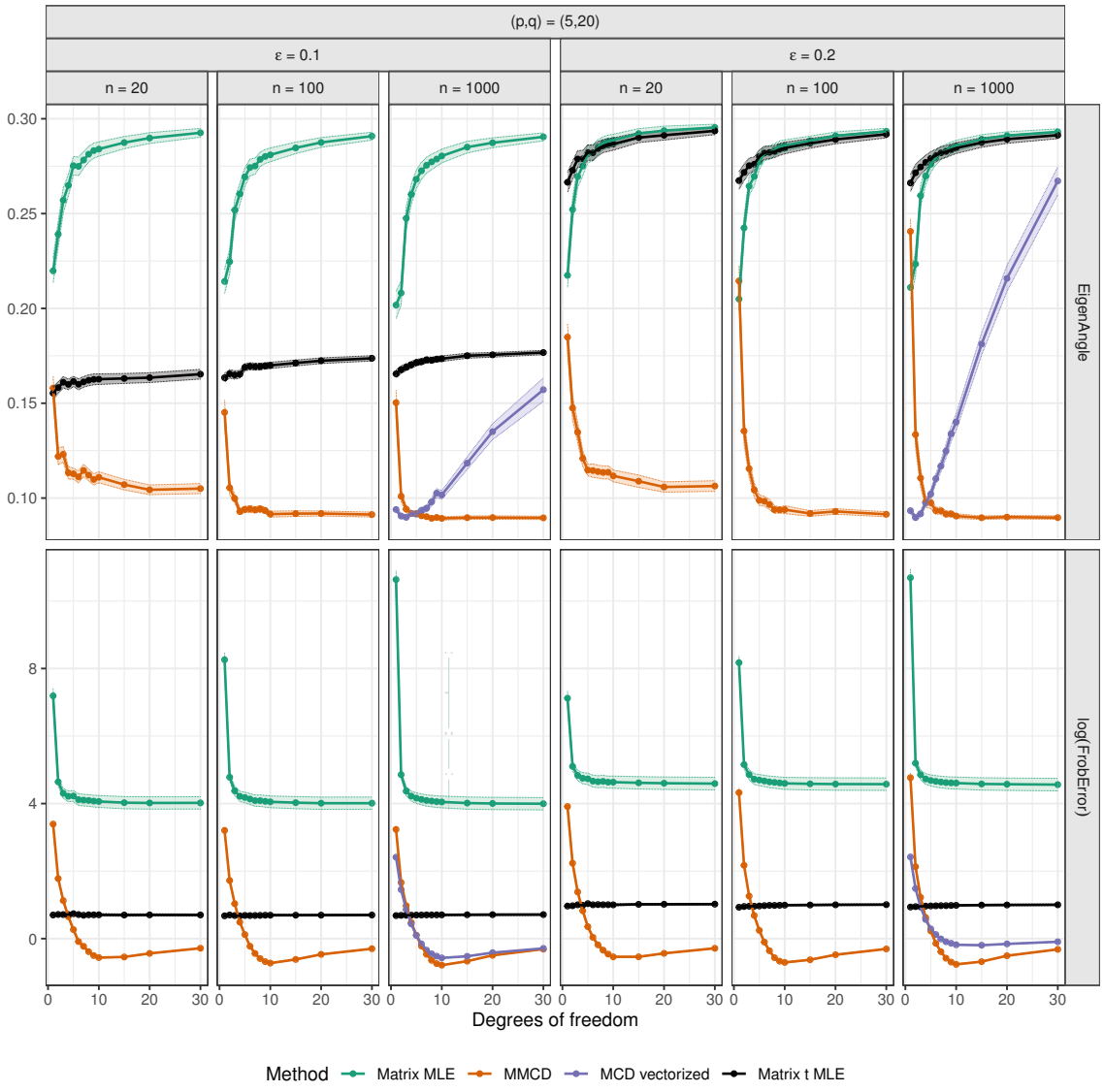


Figure B.19: Precicion, recall, eigen angle and logarithm of relative Frobenius error of samples from a contaminated t-distribution with $\nu \in \{1, \ldots, 30\}$ degrees of freedom for $n \in \{20, 100, 1000\}$, $\gamma = 1$, $\varepsilon \in \{0.1, 0.2\}$.

4 Explainable Outlier Detection for Multivariate Functional Data Based on a Functional Mahalanobis Distance

This chapter is based on the joint work of M. Mayrhofer, U. Radojičić, H. Lewitschnig, and P. Filzmoser.

Contributions: M. Mayrhofer developed the methodological framework and implemented the procedures in `R`. He and co-author Radojičić U. collaborated in establishing the proofs and writing the first draft. All co-authors were involved in the discussions and collaborated on writing the final paper.

4.1 Introduction

Functional Data Analysis (FDA) encompasses statistical models and methods to analyze data that are naturally represented as functions. With the advancement of modern data collection tools, multivariate functional observations are increasingly common since data are now often recorded repeatedly across multiple time points. These functional observations can be seen as finite-dimensional realizations of continuous stochastic processes, providing a framework for modeling and analyzing such data effectively (Cuevas, 2014; Wang et al., 2016).

Unlike traditional approaches that treat data as vectors or matrices, models with a functional structure inherently account for key characteristics of the underlying random process generating the observations, such as smoothness (Ramsay and Silverman, 2005; Ferraty, 2006). In this context, the estimation of mean and covariance plays a central role in understanding the underlying structure and variability. However, rather than direct covariance estimation, Functional Principal Component Analysis (FPCA) is often the dominant focus in the literature. Shang (2014) provides a comprehensive review of methods for univariate FPCA, Chiou et al. (2014) provide extensions to multivariate functional data, and Happ and Greven (2018) establish and discuss the connection between univariate and multivariate FPCA.

In the context of applications like climate and weather monitoring, medical data analysis, signal processing, or financial modeling, where multivariate functional data often arise, robust and explainable methods are essential to ensure accurate conclusions. Robust procedures enable reliable statistical data analyses by focusing on common patterns while providing clarity on anomalies, which is a key step for drawing informed, data-driven conclusions.

The presence of outliers can severely distort estimates of mean and covariance, as well as the performance of non-robust FPCA. Thus, robust approaches and tools for outlier detection are

needed. Several methods for robust FPCA have been developed in the univariate functional setting; see, e.g., Boente and Salibián-Barrera (2015, 2021) for an overview. While there are some methods for distance-based outlier detection in the univariate functional setting, see, e.g., Galeano et al. (2015); Ghiglietti et al. (2017); Berrendero et al. (2020); Oguamalam et al. (2024), non-parametric, depth based approaches are more commonly employed in the multivariate case. Hubert et al. (2015) provide an overview of outlier detection in multivariate functional data.

Given that much of the focus in multivariate FDA is on FPCA and depth-based outlier detection, there is a clear need for robust mean and covariance estimation methods. Our contribution is a detailed framework for robust covariance estimation and distance-based, explainable outlier detection for smooth multivariate functional data with a focus on processes with a separable covariance structure. This approach leverages the Matrix Minimum Covariance Determinant (MMCD) estimators of Mayrhofer et al. (2024a) for robust covariance estimation and a generalization of the trimmed functional Mahalanobis distance of Galeano et al. (2015) to the multivariate functional setting to identify outliers. Further, a framework for outlier explanations based on Shapley values (Shapley, 1953; Mayrhofer and Filzmoser, 2023) is proposed which enables an additive decomposition of the trimmed multivariate functional Mahalanobis distance into time-coordinate-specific contributions.

The structure of the paper is as follows: Section 4.2 outlines the theoretical framework and the already existing notion of univariate functional Mahalanobis distance. In Section 4.3 we propose a trimmed multivariate functional Mahalanobis distance and provide an in-depth analysis of its properties with a focus on processes with a separable covariance structure. Section 4.4 focuses on processes that are expressed on a finite basis and contains the main theoretical contribution of this work, which shows how the separability of the multivariate covariance operator of a random function translates onto the distribution of the random coefficient matrix of its smoothed counterpart. This leads to the connection between the trimmed multivariate functional Mahalanobis distance and the matrix-variate Mahalanobis distance. The connection is paramount to enable efficient computation with real-world data, and a step-by-step algorithm detailing the method is provided. Section 4.5 outlines the framework and computational details for outlier explanations based on Shapley values in the functional datasetting. Section 4.6 demonstrates the performance of our method for outlier detection and covariance estimation in an extensive simulation study, including a comparison with state-of-the-art methods. Section 3.7 shows the usefulness of the robust procedure and outlier explanations for real-world examples, and in Section 4.8 we discuss, summarize, and conclude our results.

4.2 Preliminaries

4.2.1 Multivariate Stochastic Processes

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\mathcal{T} \subset \mathbb{R}$ a compact interval, commonly thought of as time, then

$$\boldsymbol{X} = \{\boldsymbol{X}(t,\omega), t \in \mathcal{T}\} = \{(X_1(t,\omega), \ldots, X_p(t,\omega))', t \in \mathcal{T}\} : \mathcal{T} \times \Omega \to \mathbb{R}^p$$

is a time-continuous vector-valued stochastic process. Hence, $\boldsymbol{X}$ is a collection of random variables defined on a common probability space index by the continuous set $\mathcal{T}$; for every fixed $t \in \mathcal{T}$ the process defines a random variable $\boldsymbol{X}(t)$, and for every fixed $\omega \in \Omega$ a sample path or trajectory $\boldsymbol{X}(.,\omega)$, i.e., a function of $t \in \mathcal{T}$.

In functional data analysis (FDA) it is commonly assumed that those realizations are elements of a Hilbert space such as $\mathcal{H} := L_p^2(\mathcal{T}) = L^2(\mathcal{T}) \times \cdots \times L^2(\mathcal{T})$, the space of p-dimensional square-integrable functions (Jacques and Preda, 2014; Wang et al., 2016). The inner product and norm of $\boldsymbol{x} = (x_1, \ldots, x_p)'$ and $\boldsymbol{y} = (y_1, \ldots, y_p)'$ in $\mathcal{H}$ are given by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{H}} = \sum_{j=1}^{p} \langle x_j, y_j \rangle = \sum_{j=1}^{p} \int_{\mathcal{T}} x_j(t) y_j(t) \mathrm{d}t \quad \text{and} \quad \|\boldsymbol{x}\|_{\mathcal{H}} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{H}}^{1/2}, \tag{4.2.1}$$

respectively. A stochastic process is called an L^2 process if and only if it has finite second moments; $\mathbb{E}[\|\boldsymbol{X}(t)\|_{\mathcal{H}}^2] < \infty$ for all $t \in \mathcal{T}$. In the following, we consider L^2-continuous processes, i.e., L^2 processes for which $\lim_{h \to 0} \mathbb{E}[\|\boldsymbol{X}(t+h) - \boldsymbol{X}(t)\|_{\mathcal{H}}^2] = 0$ for every $t \in \mathcal{T}$, see, e.g., Ash and Gardner (2014) for more details. To simplify the notation, we will omit the subscript $\mathcal{H}$ from the inner product and norm when it is clear from the context which inner product or norm is being referenced.

In this setting, each component X_j of $\boldsymbol{X}$ is an L^2-continuous stochastic process for all $j = 1, \ldots, p$ with continuous mean and covariance function given by

$$\boldsymbol{\mu}(t) = \mathbb{E}[\boldsymbol{X}(t)] = \begin{pmatrix} \mathbb{E}[X_1(t)] \\ \vdots \\ \mathbb{E}[X_p(t)] \end{pmatrix}, \quad \mathbf{K}(s,t) = \begin{pmatrix} \kappa_{11}(s,t) & \cdots & \kappa_{1p}(s,t) \\ \vdots & \ddots & \vdots \\ \kappa_{p1}(s,t) & \cdots & \kappa_{pp}(s,t) \end{pmatrix}, \tag{4.2.2}$$

respectively. Here, κ_{ij}, $i, j = 1 \ldots, p$, are (cross) covariance functions (kernels) given by

$$\kappa_{ij}(s,t) = \mathrm{cov}(X_i(s), X_j(t)) = \mathbb{E}\left[(X_i(s) - \mu_i(s))(X_j(t) - \mu_j(t))\right]. \tag{4.2.3}$$

The covariance operator $\boldsymbol{\mathcal{K}} : \mathcal{H} \to \mathcal{H}$ of $\boldsymbol{X}$ associated with kernel $\mathbf{K}(s,t)$ is defined as

$$\boldsymbol{\mathcal{K}}\, \boldsymbol{x}(s) = \int_{\mathcal{T}} \mathbf{K}(s,t) \boldsymbol{x}(t) \mathrm{d}t, \quad \boldsymbol{x} \in \mathcal{H}.$$

Since $\boldsymbol{X}$ is L^2-continuous, the covariance operator $\boldsymbol{\mathcal{K}}$ is a Hilbert-Schmidt operator, and the multivariate Mercer's theorem (Withers, 1974; Daw et al., 2022) implies that there exist

countable sequences of continuous orthonormal eigenfunctions $\{\boldsymbol{\psi}_k\}_{k\geq 1}$ and non-negative decreasing eigenvalues $\{\pi_k\}_{k\geq 1}$ with $\sum_{k=1}^{\infty}\pi_k < \infty$ such that

$$\boldsymbol{\mathcal{K}}\,\boldsymbol{\psi}_k = \pi_k\,\boldsymbol{\psi}_k \quad \text{and} \quad \mathbf{K}(s,t) = \sum_{k=1}^{\infty}\pi_k\,\boldsymbol{\psi}_k(s)\,\boldsymbol{\psi}_k'(t). \tag{4.2.4}$$

The multivariate Karhunen-Loève representation theorem then implies that there exists a unique sequence of uncorrelated random variables $\{\beta_k\}_{k\geq 1}$ such that

$$\boldsymbol{X}(t) = \boldsymbol{\mu}(t) + \sum_{k=1}^{\infty}\beta_k\,\boldsymbol{\psi}_k \quad \text{with} \quad \beta_k = \langle \boldsymbol{X}-\boldsymbol{\mu}, \boldsymbol{\psi}_k\rangle_{\mathcal{H}} = \int_{\mathcal{T}} \boldsymbol{\psi}_k'(t)(\boldsymbol{X}(t)-\boldsymbol{\mu}(t))\mathrm{d}t, \tag{4.2.5}$$

where $\beta_k \sim \mathcal{N}(0,\pi_k)$ if $\boldsymbol{X}$ is a multivariate Gaussian process (Daw et al., 2022).

In continuation, an L^2-continuous multivariate stochastic process $\boldsymbol{X}$ with mean function $\boldsymbol{\mu}$ and covariance function $\mathbf{K}$ is denoted as $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$.

4.2.2 Notion of Mahalanobis Distance

Before we start discussing the functional setting, let us first review the concept of the Mahalanobis distance for random vectors. For a p-variate random vector $\boldsymbol{x}$ from a population with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} \in \mathrm{PDS}(p)$, the squared Mahalanobis distance of a random vector $\boldsymbol{x}$ (from mean $\boldsymbol{\mu}$, with respect to covariance $\boldsymbol{\Sigma}$) is given by

$$\mathrm{MD}^2(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathrm{MD}^2(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}). \tag{4.2.6}$$

Here $\mathrm{PDS}(p)$ denotes the set of all $(p\times p)$ positive definite symmetric matrices. Let $\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}' = \boldsymbol{\Sigma}$ denote the spectral decomposition of $\boldsymbol{\Sigma}$, where $\boldsymbol{D} = \mathrm{diag}(\lambda_1,\ldots,\lambda_p)$ is a diagonal matrix containing the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of $\boldsymbol{\Sigma}$, and the matrix $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ contains the corresponding eigenvectors. Based on the spectral decomposition, we can rewrite $\mathrm{MD}^2(\boldsymbol{x})$ in terms of the principal components $\boldsymbol{z} = \boldsymbol{V}'(\boldsymbol{x}-\boldsymbol{\mu})$ as follows:

$$\mathrm{MD}^2(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{V}'(\boldsymbol{x}-\boldsymbol{\mu}) = \sum_{j=1}^{p}\frac{z_j^2}{\lambda_j}. \tag{4.2.7}$$

In the context of functional data, the covariance operator $\boldsymbol{\mathcal{K}}$ serves as the functional analog to the covariance matrix in multivariate statistics. Therefore, when defining the Mahalanobis distance for functional data in a manner analogous to the multivariate case, the inverse of the covariance operator plays a crucial role in the formulation. However, as a Hilbert-Schmidt operator, see Preliminaries 4.2.1, the covariance operator is, in general, not invertible, and a regularized covariance operator should be used instead. For univariate functional data, there have been several proposals on how to define a notion of Mahalanobis distance in infinite-dimensional L^2 space: Galeano et al. (2015) introduced a method based on a spectral cutoff regularization, where the covariance operator is truncated to a finite number of components, making it invertible. Specifically, let $X \in L^2(\mathcal{T})$ be a univariate stochastic process with mean μ and covariance function κ, denoted as $X \sim \mathcal{SP}(\mu,\kappa)$, then

its squared truncated functional Mahalanobis distance with truncation level $m \in \mathbb{N}$ is given by

$$\mathrm{fMD}^2(X, \mu; \kappa, m) = \mathrm{fMD}^2(X; m) = \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X - \mu, \xi_i \rangle^2, \tag{4.2.8}$$

where (λ_i, ξ_i), $i = 1, \ldots, p$, $\lambda_1 \geq \cdots \geq \lambda_m > 0$, denote the first m eigenpairs of the covariance operator $\mathcal{K}$ with kernel κ. This approach is computationally efficient and well-suited for smoothed functions represented by a finite basis. Ghiglietti et al. (2017) proposed an alternative method that introduces regularization through an additional parameter, offering greater flexibility and addressing convergence issues encountered in Galeano et al. (2015). Berrendero et al. (2020) further extended the Mahalanobis distance definition by incorporating smoothing through reproducing kernel Hilbert spaces (RKHS), embedding regularization directly into the distance computation.

4.3 Multivariate Functional Mahalanobis Distance

In this section, we introduce the notion of multivariate truncated Mahalanobis semi-distance, which extends the univariate functional version defined by Galeano et al. (2015). This particular choice of univariate functional Mahalanobis distance was made for computational simplicity and the favorable properties when applied to functions represented by a finite basis. While the concept has been briefly discussed in Martino et al. (2019), primarily in the context of simulations and examples as a competitive method, to the best of our knowledge, this is the first rigorous formalization and study of the Mahalanobis distance for multivariate functional data and its properties.

Definition 4.3.0.1. *Let $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$ and $\boldsymbol{Y} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$. For $M \in \mathbb{N}$ such that $\pi_1 \geq \cdots \geq \pi_M > 0$, the squared truncated functional multivariate Mahalanobis semi-distance (fMMD) between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (w.r.t. $\mathbf{K}$) is given by*

$$\mathrm{fMMD}^2(\boldsymbol{X}, \boldsymbol{Y}; \mathbf{K}, M) = \sum_{k=1}^{M} \frac{1}{\pi_k} \langle \boldsymbol{X} - \boldsymbol{Y}, \boldsymbol{\psi}_k \rangle^2,$$

where, $(\pi_k, \boldsymbol{\psi}_k)$ denotes the kth eigenpair of the covariance operator $\boldsymbol{\mathcal{K}}$ with kernel $\mathbf{K}$, $k = 1, \ldots, m$, and $M \in \mathbb{N}$ determines the spectral cutoff.

Using Definition 4.3.0.1, we define the squared truncated functional Mahalanobis semi-distance of $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$ (w.r.t. $\boldsymbol{\mu}$ and $\mathbf{K}$) as

$$\mathrm{fMMD}^2(\boldsymbol{X}; M) := \mathrm{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \sum_{k=1}^{M} \frac{1}{\pi_k} \langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{\psi}_k \rangle^2. \tag{4.3.1}$$

As shown in Galeano et al. (2015) for univariate functions and discussed in Martino et al. (2019) for the multivariate setting, fMMD in Definition 4.3.0.1 is a semi-distance, since it lacks the identifiability condition due to truncation. I.e., if the projections of $\boldsymbol{X}$ and $\boldsymbol{Y}$ coincide on $\mathrm{span}(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M)$, for a fixed $M \in \mathbb{N}$, then $\mathrm{fMMD}(\boldsymbol{X}, \boldsymbol{Y}) = 0$, even if $\boldsymbol{X} \neq \boldsymbol{Y}$. For the sake of conciseness, we will simply write Mahalanobis distance.

4.3.1 L^2-Multivariate Stochastic Processes

The following lemma shows that fMMD given in Definition 4.3.0.1 is affine invariant.

Lemma 4.3.1.1. *Let* $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu_X}, \mathbf{K_X})$ *be such that* $\text{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu_X}; \mathbf{K_X}, M)$ *is well defined for* $M \in \mathbb{N}$*. Then, for any regular matrix* $\boldsymbol{A} \in \mathbb{R}^{p\times p}$*, fixed vector-variate function* $\boldsymbol{\nu} : \mathcal{T} \to \mathbb{R}^p$ *and* $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{\nu}$*, the following holds,*

i) $\boldsymbol{Y} \sim \mathcal{MSP}(\boldsymbol{\mu_Y}, \mathbf{K_Y})$*, with* $\boldsymbol{\mu_Y} = \boldsymbol{A\mu_X} + \boldsymbol{\nu}, \mathbf{K_Y} = \boldsymbol{A}\,\mathbf{K_X}\,\boldsymbol{A}'$*,*

ii) $\text{fMMD}^2(\boldsymbol{Y}, \boldsymbol{\mu_Y}; \mathbf{K_Y}, M) = \text{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu_X}; \mathbf{K_X}, M)$.

For a detailed proof, see Appendix C.2. Lemma 4.3.1.2 gives another desirable property of fMMD (Definition 4.3.0.1); in the case where the components of $\boldsymbol{X}$ are uncorrelated, its squared Mahalanobis distance reduces to the weighted sum of univariate functional Mahalanobis distances (4.2.8), where the amount of truncation depends on the magnitude of the eigenvalues of the individual processes.

Lemma 4.3.1.2. *Let* $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$ *be a multivariate process with uncorrelated components* $X_j \in L^2(\mathcal{T})$*,* $j = 1, \ldots, p$*. Let further* $\pi_1 \geq \cdots \geq \pi_M > \pi_{M+1}$ *be the largest* M *eigenvalues of the covariance operator* $\boldsymbol{\mathcal{K}} = \text{diag}(\mathcal{K}_1, \ldots, \mathcal{K}_p)$ *associated with the covariance function* $\mathbf{K} = \text{diag}(\kappa_1, \ldots, \kappa_p)$*. Then, for* $M = m_1 + \cdots + m_p$*,*

$$\text{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \sum_{j=1}^{p} \text{fMD}^2(X_j, \mu_j; \kappa_j, m_j),$$

with $(\lambda_i^{(j)}, \xi_i^{(j)}), i = 1, \ldots, m_j$ *denoting the* m_j *largest eigenpairs of the covariance operator* $\mathcal{K}_j$ *with kernel* κ_j*, and* $m_j = |\{\lambda_1^{(j)}, \ldots \lambda_M^{(j)}\} \cap \{\pi_1, \ldots \pi_M\}|$ *is the number of eigenvalues of the covariance operator* $\mathcal{K}_j$ *that belongs to* $\{\pi_1, \ldots, \pi_M\}$*,* $i = 1, \ldots, p$*, where we count also the multiplicities in* $\{\lambda_1^{(j)}, \ldots \lambda_M^{(j)}\}$.

A proof is given in Appendix C.2. Lemma 4.3.1.2 implicitly implies that the (uncorrelated) components in the multivariate process should be transformed to similar scales. We also note that the additive property of fMMD discussed in Lemma 4.3.1.2 extends to processes with uncorrelated blocks of components. However, we refrain from formally presenting this more general result to keep the paper accessible and avoid unnecessary technical complexity.

A key challenge in computing fMMD lies in accurately estimating the eigenfunctions and the covariance operator. The common approach is to vectorize the p-variate process $\boldsymbol{X} = (X_1, \ldots, X_p)' : \mathcal{T} \to \mathbb{R}^p$ by concatenating the individual processes and then applying the univariate method (Ramsay and Silverman, 2005). However, this strategy neglects the potential structure in $\boldsymbol{X}$ and can substantially increase the dimensionality, making the problem computationally prohibitive when p is large. Hence, in the following, we focus on the family of multivariate processes with separable covariance structure; for an overview, see, e.g., Chen et al. (2021).

4.3.2 Separable Covariance Processes

We say that the multivariate stochastic process $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$ has separable covariance structure if for every $s, t \in \mathcal{T}$, the covariance $\mathbf{K}$ can be decomposed into

$$\mathbf{K}(s,t) = \boldsymbol{\Sigma}^{\text{row}}\, \kappa(s,t), \tag{4.3.2}$$

for $\boldsymbol{\Sigma}^{\text{row}} \in \text{PDS}(p)$ representing the cross-covariance structure between the p components and a positive definite kernel κ capturing the common temporal covariance structure. We write $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$. The separability property significantly simplifies the covariance estimation since it allows for the within-function and the between-component second-order dependence to be studied (and interpreted) separately; see, e.g., Chen et al. (2021, 2023); Genton (2007); Cressie and Huang (1999); Rodríguez-Iturbe and Mejía (1974).

Remark 4.3.2.1. *It should be noted that $\boldsymbol{\Sigma}^{\text{row}}$ and $\kappa(s,t)$ in decomposition (4.3.2) are only identifiable up to a multiplicative constant; for any $c > 0$, $\mathbf{K}(s,t) = (c\boldsymbol{\Sigma}^{\text{row}})\,(c^{-1}\kappa(s,t))$. This ambiguity has little practical relevance. However, for reproducibility, we fix the scale of κ using the strategy presented in Mayrhofer et al. (2024a).*

The equivalent of Lemma 4.3.1.2 for the processes with separable covariance structure is given in Corollary 4.3.2.1.

Corollary 4.3.2.1. *Let $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$ be a multivariate process with separable covariance and components $X_j \in L^2(\mathcal{T})$, $j = 1, \ldots, p$. Then, for M being a multiple of p, i.e., $M = mp > 0$, with $m \in \mathbb{N}$, the following holds:*

(i) $$\text{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \boldsymbol{\Sigma}^{\text{row}}, \kappa, M) = \sum_{j=1}^{p} \text{fMD}^2((\boldsymbol{\Sigma}^{\text{row}})^{-1/2} X_j, \boldsymbol{e}_j'(\boldsymbol{\Sigma}^{\text{row}})^{-1/2}\boldsymbol{\mu}; \kappa, m).$$

(ii) If $\boldsymbol{X}$ has uncorrelated components, i.e., $\boldsymbol{\Sigma}^{\text{row}} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with $\sigma_1, \ldots, \sigma_p > 0$, then

$$\text{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \boldsymbol{\Sigma}^{\text{row}}, \kappa, M) = \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \text{fMD}^2(X_j, \mu_j; \kappa, m).$$

Here, $(\lambda_i, \xi_i), i = 1, \ldots, m$, correspond to the first m eigenpairs of the covariance operator $\mathcal{K}$ with kernel κ, and $\boldsymbol{e}_j$ is the jth vector of the canonical basis of $\mathbb{R}^p$.

See Appendix C.2 for the proof. Separability of the covariance structure transfers further to certain separability of the eigendecomposition of $\boldsymbol{\mathcal{K}}$, thus reducing the calculation of eigenpairs $(\pi_i, \boldsymbol{\psi}_i)$, $i \geq 1$, to a separate univariate functional eigendecomposition of the covariance associated with κ, and a multivariate eigendecomposition of $\boldsymbol{\Sigma}^{\text{row}}$. For more insight, see Appendix C.1.

Remark 4.3.2.2. *It is important to note that requiring M to be a multiple of p is not arbitrary, but rather a natural choice given the structure of the problem. To illustrate, if $\boldsymbol{\Sigma}^{\text{row}} = \boldsymbol{I}_p$, the eigenvalues of the covariance operator $\boldsymbol{\mathcal{K}}$ appear with multiplicity p, reflecting the inherent symmetries in the data. When projecting onto an M-dimensional space, M is typically chosen to capture a desired amount of explained variance, or based on a threshold*

related to the significance of the eigenfunctions. Given that each eigenvalue corresponds to p linearly independent components, it is reasonable to select all p components associated with any given eigenvalue when deciding on the projection dimension. This ensures that the projection retains the intrinsic structure of the data, avoiding arbitrary truncation of the eigenspaces and preserving the full contribution of the variance associated with each eigenvalue.

One of the most prominent members of the separable-covariance class of processes is a multivariate Gaussian process: $\boldsymbol{X}$ is a multivariate Gaussian process if every finite collection of realizations has a matrix-variate normal distribution. We provide a brief overview of matrix normal distribution in Appendix C.1, and for more details on matrix-variate distributions, see, e.g., Gupta and Nagar (1999); Gupta et al. (2013). Multivariate Gaussian processes are fully characterized by their first and second moments, and in continuation, we write $\boldsymbol{X} \sim \mathcal{MGP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$. For a formal definition, see Appendix C.1, and for an overview of properties of multivariate Gaussian processes, see, e.g., Chen et al. (2017, 2023). Lemma 4.3.2.1 gives the distribution of fMMD under the assumption of Gaussianity.

Lemma 4.3.2.1. *Let $\boldsymbol{X} \sim \mathcal{MGP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$ be a multivariate Gaussian process. Then, for $M > 0$,*

$$\text{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \boldsymbol{\Sigma}^{\text{row}}, \kappa, M) \sim \chi^2(M),$$

where $\chi^2(M)$ is the chi-square distribution with M degrees of freedom.

The proof is presented in Appendix C.2.

4.4 Robust Parameter Estimation for Separable Processes

In practice we do not observe continuous functions, but rather a discrete set of functional values. As discussed in Basna et al. (2022), a fundamental step in FDA is often to transform these discretely recorded data into a functional form, allowing each observed function to be evaluated at any point within its continuous domain $t \in \mathcal{T}$. Typically, the functional object is approximated by linear combinations of a finite number of basis functions, where this representation is exact only for functions of finite rank.

4.4.1 Finite Basis Representation

Let $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$,

$$\boldsymbol{X}_i = \begin{pmatrix} X_{i,1}(t_1) & \cdots & X_{i,1}(t_q) \\ \vdots & \ddots & \vdots \\ X_{i,p}(t_1) & \cdots & X_{i,p}(t_q) \end{pmatrix} \in \mathbb{R}^{p \times q},\ i = 1, \ldots, n,$$

be an i.i.d. sample of the multivariate random processes $\mathcal{MSP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$ with separable covariance $\mathbf{K} = \boldsymbol{\Sigma}^{\text{row}} \kappa$, observed at time points $t_1, \ldots, t_q \in \mathcal{T}$, for $q \in \mathbb{N}$.

Given $m \in \mathbb{N}$, let $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots \phi_m(t))'$, $t \in \mathcal{T}$ be a fixed basis that spans an m-dimensional subspace of $L^2(\mathcal{T})$. We transform each discretely observed $\boldsymbol{X}_i$, $i = 1, \ldots, n$, to

a *functional form* by representing its components as a linear combination of basis functions in $\boldsymbol{\phi}$:

$$X_{i,j}^{(m)}(t) = \sum_{k=1}^{m} a_{j,k}^{(i)} \phi_k(t) = \boldsymbol{a}_j^{(i)\prime} \boldsymbol{\phi}(t), \quad j = 1, \ldots, p,\ i = 1, \ldots, n, \tag{4.4.1}$$

where superscript (m) emphasizes that $X_{i,j}^{(m)}$, $j = 1, \ldots, p$, $i = 1, \ldots, n$, is at most a *rank-m* process, meaning that its covariance operator has at most m non-zero eigenvalues.

Collecting all coefficients $\boldsymbol{a}_j^{(i)} = (a_{j,1}^{(i)}, \ldots, a_{j,m}^{(i)})'$, $j = 1, \ldots, p$, corresponding to the ith observation, in a matrix $\boldsymbol{A}_i = (\boldsymbol{a}_1^{(i)}, \ldots, \boldsymbol{a}_p^{(i)}) \in \mathbb{R}^{m \times p}$, we can represent each observation as

$$\boldsymbol{X}_i^{(m)}(t) = \boldsymbol{A}_i' \boldsymbol{\phi}(t), \quad i = 1, \ldots, n. \tag{4.4.2}$$

For simplicity of the notation, we drop the superscript (m) in the following and write $\boldsymbol{X}_i(t) = \boldsymbol{A}_i' \boldsymbol{\phi}(t)$, $i = 1, \ldots, n$.

The coefficients $\boldsymbol{a}_j^{(i)}$, $j = 1, \ldots, p$, $i = 1, \ldots, n$, are usually determined through least squares estimation, where the goal is to minimize the difference between the observed data and their approximation while ensuring smoothness; see Ramsay and Silverman (2005) for more details. Common choices of basis functions include splines, wavelets, and Fourier bases, among others; see, e.g., Ramsay and Silverman (2005) for an overview. The choice of the basis $\boldsymbol{\phi}$ is beyond the scope of this paper, and for simplicity we use B-splines; see, e.g., Eilers and Marx (1996). For more details on a connection between a finite-basis representation (4.4.2) and noise smoothening in additive noise models, see Appendix C.1.

The following theorem shows how the separability of the covariance $\mathbf{K}$ translates onto the distribution of a random matrix $\boldsymbol{A}$, and gives a direct connection between fMMD^2 of $\boldsymbol{X}$ and the squared matrix Mahalanobis distance of $\boldsymbol{A}$, defined as

$$\mathrm{MMD}^2(\boldsymbol{A}, \mathbf{M}, \boldsymbol{\Sigma}^{\mathrm{row}}, \boldsymbol{\Sigma}^{\mathrm{col}}) = \mathrm{MMD}^2(\boldsymbol{A}) = \mathrm{tr}((\boldsymbol{\Sigma}^{\mathrm{col}})^{-1}(\boldsymbol{A} - \mathbf{M})'(\boldsymbol{\Sigma}^{\mathrm{row}})^{-1}(\boldsymbol{A} - \mathbf{M})) \tag{4.4.3}$$

for an $m \times p$ random matrix $\boldsymbol{A}$, with mean matrix $\mathbf{M} \in \mathbb{R}^{m \times p}$, and row and column covariance $\boldsymbol{\Sigma}^{\mathrm{row}} \in \mathrm{PDS}(m)$ and $\boldsymbol{\Sigma}^{\mathrm{col}} \in \mathrm{PDS}(p)$, respectively (Mayrhofer et al., 2024a).

Theorem 4.4.1.1. *Let $\boldsymbol{X}(t) = \boldsymbol{A}' \boldsymbol{\phi}(t)$ be a rank m separable covariance process with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{K} = \boldsymbol{\Sigma}^{\mathrm{row}} \kappa$, with a regular matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p) \in \mathbb{R}^{m \times p}$, and a vector of basis functions $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$. Then the following holds:*

(i) $\boldsymbol{A}$ has a matrix-variate distribution with mean $\boldsymbol{M_A}$ and covariance $\mathrm{Cov}(\mathrm{vec}(\boldsymbol{A})) = \boldsymbol{\Sigma}^{\mathrm{row}} \otimes \boldsymbol{\Sigma}^{\mathrm{col}}$, for $\boldsymbol{\Sigma}^{\mathrm{col}} \in \mathrm{PDS}(m)$, satisfying

$$\boldsymbol{M_A}' \boldsymbol{\phi}(t) = \boldsymbol{\mu}(t) \quad \text{and} \quad \boldsymbol{\phi}'(s) \boldsymbol{\Sigma}^{\mathrm{col}} \boldsymbol{\phi}(t) = \kappa(s, t),$$

for every $s, t \in \mathcal{T}$.

(ii) $\mathrm{fMMD}^2(\boldsymbol{X}; mp) = \mathrm{tr}\left((\boldsymbol{\Sigma}^{\mathrm{col}})^{-1}(\boldsymbol{A} - \boldsymbol{M_A})'(\boldsymbol{\Sigma}^{\mathrm{row}})^{-1}(\boldsymbol{A} - \boldsymbol{M_A})\right) = \mathrm{MMD}^2(\boldsymbol{A})$.

(iii) If additionally $\boldsymbol{X} \sim \mathcal{MGP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\mathrm{row}}, \kappa)$ is a multivariate Gaussian process, then $\boldsymbol{A} \sim \mathcal{MN}(\boldsymbol{M_A}, \boldsymbol{\Sigma}^{\mathrm{col}}, \boldsymbol{\Sigma}^{\mathrm{row}})$ follows a matrix normal distribution with mean $\boldsymbol{M_A}$ and positive definite covariances $\boldsymbol{\Sigma}^{\mathrm{row}}$ and $\boldsymbol{\Sigma}^{\mathrm{col}}$ as in (i).

For a detailed proof, see Appendix C.2. By leveraging the relationships between the moments of $\boldsymbol{A}$ and those of $\boldsymbol{X}$, the parameter estimation for $\boldsymbol{X}$ becomes straightforward once the mean and covariance of $\boldsymbol{A}$ are estimated. Additionally, Theorem 4.4.1.1 (ii) implies that the robust estimator of $\mathrm{MMD}(\boldsymbol{A})$ serves also as a robust estimator of $\mathrm{fMMD}(\boldsymbol{X})$. Details on covariance estimation for the random matrices can be found in Dutilleul (1999); Soloveychik and Trushin (2016).

The equivalent of Theorem 4.4.1.1 for univariate processes of finite rank is given in Corollary 4.4.1.1.

Corollary 4.4.1.1. *Let $X(t) = \boldsymbol{a}'\boldsymbol{\phi}(t)$ be a rank $m \in \mathbb{N}$ stochastic process with mean μ and covariance κ, with coefficients $\boldsymbol{a} \in \mathbb{R}^m$ and basis $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$. Then the following holds:*

(i) $\boldsymbol{a}$ has a multivariate distribution with mean $\boldsymbol{m_a}$ and covariance $\mathrm{cov}(\boldsymbol{a}) = \boldsymbol{\Sigma} \in \mathrm{PDS}(m)$ such that $\boldsymbol{m_a}'\boldsymbol{\phi}(t) = \boldsymbol{\mu}(t)$ and $\boldsymbol{\phi}'(s)\boldsymbol{\Sigma}\boldsymbol{\phi}(t) = \kappa(s,t)$ for all $s,t \in \mathcal{T}$.

(ii) $\mathrm{fMD}^2(X;m) = (\boldsymbol{a}-\boldsymbol{m_a})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{a}-\boldsymbol{m_a}) = \mathrm{MD}^2(\boldsymbol{a})$.

(iii) If X is a Gaussian process, then $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{m_a}, \boldsymbol{\Sigma})$ has a multivariate normal distribution with mean $\boldsymbol{m_a}$ and covariance matrix $\boldsymbol{\Sigma}$.

The corollary follows directly from Theorem 4.4.1.1. As the primary objective is robust covariance estimation, we employ robust estimators of the moments of the random matrix $\boldsymbol{A}$, specifically using the MMCD estimators for mean and covariance, as introduced in Mayrhofer et al. (2024a). For completeness, we briefly review the MMCD method in the following.

4.4.2 Matrix Minimum Covariance Determinant Estimator (MMCD)

For a random sample $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n \in \mathbb{R}^{m\times p}$ from a matrix-elliptical semi-parametric distribution (Gupta et al., 2013), with the parametric part parameterized by the mean $\mathbf{M}_{\boldsymbol{A}}$ and covariance matrices $\boldsymbol{\Sigma}^{\mathrm{row}}$ and $\boldsymbol{\Sigma}^{\mathrm{col}}$, Mayrhofer et al. (2024a) proposed robust mean and covariance estimators. The robust MMCD estimators $(\hat{\mathbf{M}}_{\boldsymbol{A},H^*}, \hat{\boldsymbol{\Sigma}}^{\mathrm{row}}_{H^*}, \hat{\boldsymbol{\Sigma}}^{\mathrm{col}}_{H^*})$ solve

$$\underset{\substack{\hat{\mathbf{M}}_{\boldsymbol{A},H}, \hat{\boldsymbol{\Sigma}}^{\mathrm{row}}_H, \hat{\boldsymbol{\Sigma}}^{\mathrm{col}}_H \\ H\subset\{1,\ldots,n\}, |H|=h}}{\arg\min} \quad p\ln(\det(\hat{\boldsymbol{\Sigma}}^{\mathrm{col}}_H)) + q\ln(\det(\hat{\boldsymbol{\Sigma}}^{\mathrm{row}}_H)), \tag{4.4.4}$$

where

$$\hat{\mathbf{M}}_{\boldsymbol{A},H} = \frac{1}{h}\sum_{i\in H}\boldsymbol{X}_i,$$

$$\hat{\boldsymbol{\Sigma}}^{\mathrm{row}}_H = \frac{1}{qh}\sum_{i\in H}(\boldsymbol{X}_i - \hat{\mathbf{M}}_{\boldsymbol{A},H})(\hat{\boldsymbol{\Sigma}}^{\mathrm{col}}_H)^{-1}(\boldsymbol{X}_i - \hat{\mathbf{M}}_{\boldsymbol{A},H})', \tag{4.4.5}$$

$$\hat{\boldsymbol{\Sigma}}^{\mathrm{col}}_H = \frac{1}{ph}\sum_{i\in H}(\boldsymbol{X}_i - \hat{\mathbf{M}}_{\boldsymbol{A},H})'(\hat{\boldsymbol{\Sigma}}^{\mathrm{row}}_H)^{-1}(\boldsymbol{X}_i - \hat{\mathbf{M}}_{\boldsymbol{A},H}). \tag{4.4.6}$$

Here, $h = \alpha n, \alpha \in [0.5, 1]$ represents the size of the *clean* subset of the sample used for moment estimation. For $h = 0.5n$, the estimators achieve a maximal breakdown point of

$n/2 - \lfloor p/m + m/p \rfloor - 1$. With the proper scaling, the method yields consistent estimators in this context, while the finite-sample efficiency can be further improved with an additional reweighting step.

As there are no closed-form solutions for the robust MMCD-estimators, Mayrhofer et al. (2024a) proposed a nested iterative estimation procedure based on a *concentration step* algorithm (Rousseeuw and Van Driessen, 1999) for solving (4.4.4), and an iterative *flip-flop* algorithm (Dutilleul, 1999) for computing the maximum likelihood estimates (4.4.5)-(4.4.6). Starting from any positive definite initialization, the proposed procedure is shown to converge almost surely to the positive definite covariance estimates, provided $h \geq \lfloor p/m + m/p \rfloor + 2$. The convergence also holds if the ellipticity assumption is violated. For technical and implementation details, see Mayrhofer et al. (2024a).

A detailed pseudocode for the robust parameter estimation of the separable covariance processes based on a finite basis representation (4.4.2) and the MMCD estimators (4.4.4) is given in Algorithm 5.

Algorithm 5 Robust estimation of mean and covariance function

Input: $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, $\boldsymbol{X}_i \in \mathbb{R}^{p \times q}$, $i = 1, \ldots, n$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$, $\mathcal{T}$

1: **Create functional data object**
   Estimate coefficient matrices $\mathfrak{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)$ by smoothing $\mathfrak{X}$;
   Obtain finite basis representation $\boldsymbol{X}_i(t) = \boldsymbol{A}_i' \boldsymbol{\phi}(t)$, $i = 1, \ldots, n$, $t \in \mathcal{T}$;

2: **MMCD** (Mayrhofer et al., 2024a, Algorithm 2)
   Run MMCD procedure on $\mathfrak{A}$ and get $(\hat{\mathbf{M}}_{\boldsymbol{A},H^*}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H^*}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H^*}, \text{MMD}(\mathfrak{A}))$;

3: **Obtain functional data objects for mean and covariance**
   $\hat{\boldsymbol{\mu}}(t) = \hat{\mathbf{M}}_{\boldsymbol{A},H^*}\, \boldsymbol{\phi}(t)$;
   $\hat{\boldsymbol{\Sigma}}^{\text{row}} = \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H^*}$;
   $\hat{\kappa}(s,t) = \boldsymbol{\phi}'(s) \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H^*}\, \boldsymbol{\phi}(t)$;
   $\text{fMMD}(\boldsymbol{X}_i) = \text{fMMD}(\boldsymbol{X}_i, \hat{\boldsymbol{\mu}}; \hat{\boldsymbol{\Sigma}}^{\text{row}}, \hat{\kappa}, mp) = \text{MMD}(\boldsymbol{A}_i, \hat{\mathbf{M}}_{\boldsymbol{A},H^*}; \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H^*}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H^*})$

Output: $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}, \hat{\kappa}, (\text{fMMD}(\boldsymbol{X}_1), \ldots, \text{fMMD}(\boldsymbol{X}_n))$

Algorithm 5 yields robust estimators. Similarly, non-robust counterparts can be obtained by replacing MMCD by the iterative matrix maximum likelihood estimation (MMLE) procedure of Dutilleul (1999) in `step 2`. Since covariance estimation and FPCA are closely related, we outline how to compute the robust functional principal components in the separable covariance setting in Algorithm 6 in Appendix C.3.

Algorithm 5 can be easily adapted for the analysis raw data: `step 1` in the algorithm is omitted, and $p \times q$ matrices of raw data observations are supplied to the MMCD in `step 2`. The pointwise estimates of mean and covariance evaluated at observed time points $t_1, \ldots, t_q$ are the output of MMCD `step 2`. Usually post-smoothing is applied to those estimates to extend them to the functional setting, see Ramsay and Silverman (2005). However, this is beyond the scope of this paper. The rationale for using the algorithm on raw data lies in the fact that, for certain classes of separable covariance processes (e.g., Gaussian and Student's-t processes, see Chen et al. (2021, 2023)), all finite-dimensional projections belong to the same family of matrix-variate distributions with separable covariance structure. Moreover, in those cases, parameters estimated on finite-dimensional projections correspond to pointwise

evaluations of the process parameters, as described above.

4.5 Explainable Outlier Detection

The combination of the MMCD estimators (Mayrhofer et al., 2024a) with the truncated multivariate functional Mahalanobis distances provides a reliable framework for outlier detection. To understand why an observation is outlying, we propose a method for decomposing the truncated (multivariate) functional Mahalanobis distance into time-coordinate-specific outlyingness contributions. As in Mayrhofer and Filzmoser (2023), we use Shapley values (Shapley, 1953) to obtain those decompositions; they were originally introduced in cooperative game theory (Peters, 2008) and gained popularity in the field of Explainable AI (Lundberg and Lee, 2017).

4.5.1 Outlier Explanations for Multivariate and Matrix-variate Data

For a p-variate observation $\boldsymbol{x} = (x_1, \ldots, x_p)'$ from a population with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$, covariance matrix $\boldsymbol{\Sigma} \in \mathrm{PDS}(p)$, and $P = \{1, \ldots, p\}$ the index set of the variables, the outlyingness contributions $\boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\theta}(\boldsymbol{x}) = (\theta_1(\boldsymbol{x}), \ldots, \theta_p(\boldsymbol{x}))'$ assign each variable its average marginal contribution to the squared Mahalanobis distance (4.2.6), i.e.,

$$\theta_k(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{S \subseteq P \setminus \{k\}} \frac{|S|!(p - |S| - 1)!}{p!} \Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) = (x_k - \mu_k) \sum_{j=1}^{p} (x_j - \mu_j) \omega_{jk}, \quad (4.5.1)$$

with marginal outlyingness contributions

$$\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) := \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S \cup \{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \quad \text{and} \quad \hat{x}_j^S := \begin{cases} x_j & \text{if } j \in S \\ \mu_j & \text{if } j \notin S \end{cases} \quad (4.5.2)$$

as the components of $\hat{\boldsymbol{x}}^S$. Here, ω_{jk} is the element (j, k) of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Since $\boldsymbol{\theta}(\boldsymbol{x})$ is based on the Shapley value, it is the only decomposition of the squared Mahalanobis distance based on Equation (4.5.2) that fulfills the following properties:

- *Efficiency:* The contributions $\theta_j(\boldsymbol{x})$, for $j = 1, \ldots, p$, sum up to the squared Mahalanobis distance of $\boldsymbol{x}$, hence $\sum_{j=1}^{p} \theta_j(\boldsymbol{x}) = \mathrm{MD}^2(\boldsymbol{x})$.

- *Symmetry:* If $\mathrm{MD}^2(\hat{\boldsymbol{x}}^{S \cup \{j\}}) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S \cup \{k\}})$ holds for all subsets $S \subseteq P \setminus \{j, k\}$ for two coordinates j and k, then $\theta_j(\boldsymbol{x}) = \theta_k(\boldsymbol{x})$.

- *Monotonicity:* Let $\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}} \in \mathbb{R}^p$ be two vectors and $\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}} \in \mathrm{PDS}(p)$ be two matrices. If

  $$\mathrm{MD}^2_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^{S \cup \{j\}}) - \mathrm{MD}^2_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^S) \geq \mathrm{MD}^2_{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{x}}^{S \cup \{j\}}) - \mathrm{MD}^2_{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}}(\hat{\boldsymbol{x}}^S)$$

  holds for all subsets $S \subseteq P$, then $\theta_j(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \geq \theta_j(\boldsymbol{x}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$.

This means that the kth coordinate of the Shapley value $\theta_k(\boldsymbol{x})$ represents the average marginal contribution of the kth coordinate to the squared Mahalanobis distance. This is calculated by averaging over all marginal outlyingness contributions $\Delta_k \operatorname{MD}^2(\hat{\boldsymbol{x}}^S)$ across all possible subsets $S \subseteq P \setminus \{k\}$. The second equality in Equation (4.5.1) demonstrates that the exponential computational complexity of the Shapley value can be reduced to a linear complexity in this setting. For a proof, we refer to Mayrhofer and Filzmoser (2023). Equation (4.5.1) provides further insight into the Shapley value by comparing it to the squared Mahalanobis distance $\operatorname{MD}^2(\boldsymbol{x}) = \sum_{j,k=1}^{p}(x_j - \mu_j)(x_k - \mu_k)\omega_{jk}$. While $\operatorname{MD}^2(\boldsymbol{x})$ yields an outlyingness measure that aggregates the contributions $(x_j - \mu_j)(x_k - \mu_k)\omega_{jk}$ of all p variables, the outlyingness scores $\theta_k(\boldsymbol{x})$ only consider the contributions associated with the kth coordinate. In Mayrhofer et al. (2024a), the concept was extended to the matrix-variate setting. For a random matrix $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ with mean $\mathbf{M} \in \mathbb{R}^{p\times q}$, row covariance $\boldsymbol{\Sigma}^{\text{row}} \in \operatorname{PDS}(p)$ and column covariance $\boldsymbol{\Sigma}^{\text{col}} \in \operatorname{PDS}(q)$, cellwise, rowwise, and columnwise outlyingness contributions to the squared matrix Mahalanobis distance (4.4.3) are given by

$$\boldsymbol{\Theta}(\boldsymbol{X}) = (\boldsymbol{X} - \mathbf{M}) \circ (\boldsymbol{\Sigma}^{\text{row}})^{-1}(\boldsymbol{X} - \mathbf{M})(\boldsymbol{\Sigma}^{\text{col}})^{-1} \in \mathbb{R}^{p\times q}, \tag{4.5.3}$$

$$\boldsymbol{\theta}_{\text{row}}(\boldsymbol{X}) = \operatorname{diag}((\boldsymbol{\Sigma}^{\text{row}})^{-1}(\boldsymbol{X} - \mathbf{M})(\boldsymbol{\Sigma}^{\text{col}})^{-1}(\boldsymbol{X} - \mathbf{M})') \in \mathbb{R}^{p}, \tag{4.5.4}$$

$$\boldsymbol{\theta}_{\text{col}}(\boldsymbol{X}) = \operatorname{diag}((\boldsymbol{X} - \mathbf{M})'(\boldsymbol{\Sigma}^{\text{row}})^{-1}(\boldsymbol{X} - \mathbf{M})(\boldsymbol{\Sigma}^{\text{col}})^{-1}) \in \mathbb{R}^{q}, \tag{4.5.5}$$

respectively, where $\circ$ is an element-wise product. The cellwise Shapley values (4.5.3) are based on the multivariate Shapley values (4.5.1) of vectorized observations. The row- and columnwise Shapley values can be obtained by adding up the cellwise Shapley values for the respective row or column, or by adjusting the individual contributions for row-wise replacements.

4.5.2 Outlier Explanations for Functional Data

Here we extend outlier explanations based on Shapley values to the setting of univariate and multivariate functional data. We consider square integrable stochastic processes defined on the domain $\mathcal{T} \subseteq \mathbb{R}$ which we can decompose into disjoint subintervals $\mathcal{T} = \mathcal{T}_1 \cup \cdots \cup \mathcal{T}_d$ where $\mathcal{T}_a \cap \mathcal{T}_b = \emptyset$ for all $a \neq b$.

For univariate stochastic processes $X \sim \mathcal{SP}(\mu, \kappa)$, we show that we can decompose $\operatorname{fMD}^2(X, \mu; \kappa, m)$ into time-specific outlyingness contributions $\theta_{\mathcal{T}_a}(X, \mu; \kappa, m)$ for the disjoint subintervals $\mathcal{T}_a, a = 1, \ldots, d$, of $\mathcal{T}$. For a multivariate process $\boldsymbol{X} = (X_1, \ldots, X_p)' \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$, we show that we can decompose $\operatorname{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ into coordinate-specific contributions $\theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$, time-specific contributions $\theta_{\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$, as well as time-coordinate-specific contributions $\Theta_{k,\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$, for the disjoint subintervals $\mathcal{T}_a, a = 1, \ldots, d$, of $\mathcal{T}$ and $k = 1, \ldots, p$. Since these decompositions are based on the Shapley value, they inherit desirable properties (see Section 4.5.1) such as efficiency, implying that the outlyingness contributions sum up to the squared (multivariate) truncated Mahalanobis distance.

Univariate Time-specific Outlyingness Contributions

Consider two univariate stochastic processes $X, Y \in L^2(\mathcal{T})$ with $\mathcal{T} = \mathcal{T}_1 \cup \cdots \cup \mathcal{T}_d$, $\mathcal{T}_a \cap \mathcal{T}_b = \emptyset$

for all $a \neq b$, then the inner product $\langle X, Y \rangle$ can be written as

$$\langle X, Y \rangle = \sum_{a=1}^{d} \langle X, Y \rangle_{\mathcal{T}_a} \quad \text{with} \quad \langle X, Y \rangle_{\mathcal{T}_a} = \int_{\mathcal{T}_a} X(t)Y(t)\mathrm{d}t. \tag{4.5.6}$$

Let us now generalize Equation (4.5.2) to the functional setting. Consider $X \sim \mathcal{SP}(\mu, \kappa)$, then we define the marginal outlyingness contribution on the subinterval $\mathcal{T}_a$ to $\mathrm{fMD}^2(X, \mu; \kappa, m)$ as

$$\Delta_{\mathcal{T}_a} \mathrm{fMD}^2(\hat{X}^R, \mu; \kappa, m) := \mathrm{fMD}^2(\hat{X}^{R \cup \{a\}}, \mu; \kappa, m) - \mathrm{fMD}^2(\hat{X}^R, \mu; \kappa, m) \tag{4.5.7}$$

with $R \subseteq D \setminus \{a\}, D = \{1, \ldots, d\}$, and

$$\hat{X}^R(t) := \begin{cases} X(t) & \text{if } t \in \bigcup_{b \in R} \mathcal{T}_b \\ \mu(t) & \text{if } t \notin \bigcup_{b \in R} \mathcal{T}_b \end{cases}. \tag{4.5.8}$$

Proposition 4.5.2.1. *For $X \sim \mathcal{SP}(\mu, \kappa)$ and $\Delta_{\mathcal{T}_a} \mathrm{fMD}^2(\hat{X}^R, \mu; \kappa, m)$ as in Equation (4.5.7), the time-specific outlyingness contribution within the subinterval $\mathcal{T}_a$ based on the Shapley value is given by*

$$\begin{aligned} \theta_{\mathcal{T}_a}(X, \mu; \kappa, m) &:= \sum_{R \subseteq D \setminus \{a\}} \frac{|R|!(d - |R| - 1)!}{(d)!} \Delta_{\mathcal{T}_a} \mathrm{fMD}^2(\hat{X}^R, \mu; \kappa, m) \\ &= \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X - \mu, \xi_i \rangle_{\mathcal{T}_a} \langle X - \mu, \xi_i \rangle, \end{aligned} \tag{4.5.9}$$

where $(\lambda_i, \xi_i), i = 1, \ldots, m$, denote the eigenpairs of the covariance operator $\mathcal{K}$ with kernel κ.

A proof is given in Appendix C.4. The following lemma outlines how to efficiently compute Equation (4.5.9) for smooth functions represented in a basis. Let $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_m(t))'$ for $t \in \mathcal{T}$ and $m \in \mathbb{N}$ be a family of basis functions in $L^2(\mathcal{T})$. The rank $m \times m$ matrix of inner products of $\boldsymbol{\phi}$ is denoted as $\boldsymbol{W} = \int_{\mathcal{T}} \boldsymbol{\phi}(t)\, \boldsymbol{\phi}'(t)\mathrm{d}t$ and $\boldsymbol{W}_{\mathcal{T}_a} = \int_{\mathcal{T}_a} \boldsymbol{\phi}(t)\, \boldsymbol{\phi}'(t)\mathrm{d}t$ represents the matrix of inner products restricted to $\mathcal{T}_a \subseteq \mathcal{T}$.

Lemma 4.5.2.1. *Let $X \sim \mathcal{SP}(\mu, \kappa)$ be a rank $m \in \mathbb{N}$ stochastic process as in Corollary 4.4.1.1, with $X(t) = \boldsymbol{a}'\, \boldsymbol{\phi}(t)$, $\mu(t) = \boldsymbol{m}_{\boldsymbol{a}}'\, \boldsymbol{\phi}(t)$, and $\kappa(s, t) = \boldsymbol{\phi}'(s) \boldsymbol{\Sigma}\, \boldsymbol{\phi}(t)$, for $s, t \in \mathcal{T}$, then*

$$\theta_{\mathcal{T}_a}(X, \mu; \kappa, m) = (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}})' \boldsymbol{W}_{\mathcal{T}_a} \boldsymbol{W}^{-1} \boldsymbol{\Sigma}^{-1} (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}}).$$

See Appendix C.4 for a proof.

Multivariate Coordinate-specific Outlyingness Contributions

Let us consider the p-variate stochastic process $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$, with $\boldsymbol{X} = (X_1, \ldots, X_p)'$, $P = \{1, \ldots, p\}$ the index set of variables, and $(\pi_k, \boldsymbol{\psi}_k), k = 1, \ldots, M$ the eigenpairs of the covariance operator $\boldsymbol{\mathcal{K}}$ with kernel $\mathbf{K}$.

We would like to investigate the contribution of the kth coordinate function X_k to $\mathrm{fMMD}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ using Shapley values. Similar as in Equation (4.5.2), the marginal outlyingness contributions to fMMD^2 are given by

$$\Delta_k \,\mathrm{fMMD}^2(\hat{\boldsymbol{X}}^S, \boldsymbol{\mu}; \mathbf{K}, M) := \mathrm{fMMD}^2(\hat{\boldsymbol{X}}^{S\cup\{k\}}, \boldsymbol{\mu}; \mathbf{K}, M) - \mathrm{fMMD}^2(\hat{\boldsymbol{X}}^S, \boldsymbol{\mu}; \mathbf{K}, M) \quad (4.5.10)$$

with $\hat{\boldsymbol{X}}^S = (\hat{X}_1^S, \ldots, \hat{X}_p^S)'$, $S \subseteq P$, and

$$\hat{X}_j^S := \begin{cases} X_j & \text{if } j \in S \\ \mu_j & \text{if } j \notin S \end{cases}. \quad (4.5.11)$$

Proposition 4.5.2.2. *For $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$ and $\Delta_k \,\mathrm{fMMD}^2(\hat{\boldsymbol{X}}^S, \boldsymbol{\mu}; \mathbf{K}, M)$ as in Equation (4.5.7), the coordinate-specific outlyingness contribution to $\mathrm{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ of the kth coordinate function based on the Shapley value is given by*

$$\begin{aligned}\theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) &:= \sum_{S\subseteq P\setminus\{k\}} \frac{|S|!(p-|S|-1)!}{p!} \Delta_k \,\mathrm{fMMD}^2(\hat{\boldsymbol{X}}^S, \boldsymbol{\mu}; \mathbf{K}, M) \\ &= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k - \mu_k, \psi_{i,k}\rangle \langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{\psi}_i\rangle, \end{aligned} \quad (4.5.12)$$

with $\psi_{i,k} = \boldsymbol{\psi}_i' \boldsymbol{e}_k$ denoting the kth component of the ith eigenfunction $\boldsymbol{\psi}_i$ of covariance operator $\boldsymbol{\mathcal{K}}$ with kernel $\mathbf{K}$.

The proof is given in Appendix C.4. The following corollary outlines how to compute the coordinate-specific outlyingness contributions for a multivariate stochastic process with a separable covariance structure.

Corollary 4.5.2.1. *Let $\boldsymbol{X} \sim \mathcal{MSP}(\mathbf{0}, \boldsymbol{\Sigma}^{\mathrm{row}}, \kappa)$ with covariance operator $\boldsymbol{\mathcal{K}} = \boldsymbol{\Sigma}^{\mathrm{row}}\mathcal{K}$ and covariance kernel $\mathbf{K}(s,t) = \boldsymbol{\Sigma}^{\mathrm{row}}\kappa(s,t)$, then*

$$\theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \boldsymbol{\Sigma}^{\mathrm{row}}\kappa, M) = \sum_{i=1}^{m}\sum_{j=1}^{p} \frac{1}{\lambda_i^{\mathrm{ker}} \lambda_j^{\mathrm{row}}} \left(\langle X_k - \mu_k, \xi_i\rangle v_{j,k}^{\mathrm{row}} \sum_{l=1}^{p} \langle X_l - \mu_l, \xi_i\rangle v_{j,l}^{\mathrm{row}} \right).$$

Here, $(\lambda_i^{\mathrm{ker}}, \xi_i)$, $i = 1, \ldots, m$, denote the m largest eigenpairs of $\mathcal{K}$, $(\lambda_j^{\mathrm{row}}, \boldsymbol{v}_j^{\mathrm{row}})$, $j = 1, \ldots, p$, the eigenpairs of $\boldsymbol{\Sigma}^{\mathrm{row}}$, and $v_{j,k}^{\mathrm{row}} = \boldsymbol{e}_k' \boldsymbol{v}_j^{\mathrm{row}}$.

A proof is given in C.4. We can efficiently compute the outlyingness contributions for all p variables using matrix operations. Let $\tilde{\boldsymbol{A}} \in \mathbb{R}^{p\times m}$ with entries $(\alpha_{ji}) = \langle X_j - \mu_j, \xi_i\rangle$ denote the matrix of inner products of the coordinate functions $X_j, j = 1, \ldots, p$, with the functional principal components $\xi_i, i = 1, \ldots, m$, and $\boldsymbol{D}^{\mathrm{ker}} = \mathrm{diag}(\lambda_1^{\mathrm{ker}}, \ldots, \lambda_m^{\mathrm{ker}})$ the diagonal matrix of the corresponding ordered eigenvalues $\lambda_1^{\mathrm{ker}} \geq \cdots \geq \lambda_m^{\mathrm{ker}}$ of the kernel function κ. Then the vector $\boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ with entries $\theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \boldsymbol{\Sigma}^{\mathrm{row}}\kappa, M)$, for $k = 1, \ldots, p$, can be computed as

$$\boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \mathrm{diag}(\tilde{\boldsymbol{A}}(\boldsymbol{D}^{\mathrm{ker}})^{-1}\tilde{\boldsymbol{A}}'(\boldsymbol{\Sigma}^{\mathrm{row}})^{-1}).$$

For smooth multivariate functional data represented in a finite basis $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$, the outlyingness scores can be computed using the coefficients.

Lemma 4.5.2.2. *Let* $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$ *be a rank* $M \in \mathbb{N}$ *multivariate stochastic process as in Theorem 4.4.1.1, with* $\boldsymbol{X}(t) = \boldsymbol{A}'\boldsymbol{\phi}(t)$, $\boldsymbol{\mu}(t) = \boldsymbol{M}_{\boldsymbol{A}}'\boldsymbol{\phi}(t)$, *and* $\kappa(s,t) = \boldsymbol{\phi}'(s)\boldsymbol{\Sigma}^{\text{col}}\boldsymbol{\phi}(t)$, *for* $s,t \in \mathcal{T}$. *Then it holds that*

$$\theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \sum_{j=1}^{p} \frac{1}{\lambda_j^{\text{row}}} v_{j,k}^{\text{row}} (\boldsymbol{a}_k - \boldsymbol{m}_{\boldsymbol{A},k})' (\boldsymbol{\Sigma}^{\text{col}})^{-1} (\boldsymbol{A} - \boldsymbol{M}_{\boldsymbol{A}})' \boldsymbol{v}_j^{\text{row}},$$

with $(\lambda_j^{\text{row}}, \boldsymbol{v}_j^{\text{row}})$, $j = 1, \ldots, p$, *the eigenpairs of* $\boldsymbol{\Sigma}^{\text{row}}$, *and* $v_{j,k}^{\text{row}} = \boldsymbol{e}_k' \boldsymbol{v}_j^{\text{row}}$.

See Appendix C.4 for a proof. Using matrix operations we obtain the vector of coordinate-specific outlyingness contributions $\boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ with entries $\theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$, for $k = 1, \ldots, p$, by

$$\boldsymbol{\theta}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \text{diag}((\boldsymbol{\Sigma}^{\text{row}})^{-1} (\boldsymbol{A} - \boldsymbol{M}_{\boldsymbol{A}}) (\boldsymbol{\Sigma}^{\text{col}})^{-1} (\boldsymbol{A} - \boldsymbol{M}_{\boldsymbol{A}})').$$

Here, diag(.) of a square matrix denotes the vector of diagonal entries.

Multivariate Time and Time-coordinate Outlyingness Contributions

To obtain the marginal outlyingness contribution of the kth coordinate function in the time interval $\mathcal{T}_a$ to fMMD^2 based on the Shapley value, we modify Equation (4.5.2) and define them as

$$\Delta_{k,\mathcal{T}_c} \text{fMMD}^2(\hat{\boldsymbol{X}}^{S,R}, \boldsymbol{\mu}; \mathbf{K}, M) := \text{fMMD}^2(\hat{\boldsymbol{X}}^{S \cup \{k\}, R \cup c}, \boldsymbol{\mu}; \mathbf{K}, M) - \text{fMMD}^2(\hat{\boldsymbol{X}}^{S,R}, \boldsymbol{\mu}; \mathbf{K}, M) \tag{4.5.13}$$

with $\hat{\boldsymbol{X}}^{S,R} = (\hat{X}_1^{S,R}, \ldots, \hat{X}_p^{S,R})'$,

$$\hat{X}_j^{S,R}(t) := \begin{cases} X_j(t) & \text{if } j \in S \wedge t \in \bigcup_{a \in R} \mathcal{T}_a \\ \mu_j(t) & \text{if } j \notin S \vee t \notin \bigcup_{a \in R} \mathcal{T}_a \end{cases}. \tag{4.5.14}$$

Proposition 4.5.2.3. *For* $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$ *and* $\Delta_{k,\mathcal{T}_c} \text{fMMD}^2(\hat{\boldsymbol{X}}^{S,R}, \boldsymbol{\mu}; \mathbf{K}, M)$ *as in Equation* (4.5.7), *the outlyingness contributions of the* k*th coordinate in the time-interval* $\mathcal{T}_a$ *to* $\text{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ *based on the Shapley value are given by*

$$\Theta_{k,\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k - \mu_k, \psi_{i,k} \rangle_{\mathcal{T}_a} \langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{\psi}_i \rangle_{\mathcal{T}}, \tag{4.5.15}$$

with $\psi_{i,k} = \boldsymbol{\psi}_i' \boldsymbol{e}_k$ *denoting the* k*th component of the* i*th eigenfunction* $\boldsymbol{\psi}_i$ *of covariance operator* $\boldsymbol{\mathcal{K}}$ *with kernel* $\mathbf{K}$.

The proof is given in Appendix C.4; it relies on concatenating the coordinate functions and Proposition 4.5.2.1. We can modify Equations (4.5.13) and (4.5.14) to obtain time-specific outlyingness contributions by replacing all coordinate functions $X_j(t)$ by their mean $\mu_j(t)$ for a given interval $t \in \mathcal{T}_a$ for all $j = 1, \ldots, p$ coordinates instead of only one. This yields the time-specific outlyingness contributions

$$\theta_{\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \sum_{i=1}^{M} \frac{1}{\pi_i} \langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{\psi}_i \rangle_{\mathcal{T}_a} \langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{\psi}_i \rangle_{\mathcal{T}}. \tag{4.5.16}$$

Corollary 4.5.2.2. *For a multivariate stochastic process* $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$ *with separable covariance operator* $\boldsymbol{\mathcal{K}} = \boldsymbol{\Sigma}^{\text{row}}\mathcal{K}$ *and covariance kernel* $\mathbf{K}(s,t) = \boldsymbol{\Sigma}^{\text{row}}\kappa(s,t)$ *Equation* (4.5.15) *becomes*

$$\Theta_{k,\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \boldsymbol{\Sigma}^{\text{row}}\kappa, M) = \sum_{i=1}^{m}\sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}}\lambda_j^{\text{row}}} \left(\langle X_k - \mu_k, \xi_i\rangle_{\mathcal{T}_a} v_{j,k}^{\text{row}} \sum_{l=1}^{p} \langle X_l - \mu_l, \xi_i\rangle_{\mathcal{T}} v_{j,l}^{\text{row}} \right).$$

Here $(\lambda_i^{\text{ker}}, \xi_i)$, $i = 1, \ldots, m$, *denote the* m *largest eigenpairs of* $\mathcal{K}$, $(\lambda_j^{\text{row}}, \boldsymbol{v}_j^{\text{row}})$, $j = 1, \ldots, p$, *the eigenpairs of* $\boldsymbol{\Sigma}^{\text{row}}$, *and* $v_{j,k}^{\text{row}} = \boldsymbol{e}_k'\boldsymbol{v}_j^{\text{row}}$.

The proof follows from the same arguments as the proof of Corollary 4.5.2.1. We can compute Shapley values for the time interval $\mathcal{T}_a$ for all coordinate functions simultaneously. Let $\tilde{\boldsymbol{A}}^{\mathcal{T}} \in \mathbb{R}^{p\times m}$ with entries $(\alpha_{ji}^{\mathcal{T}}) = \langle X_j - \mu_j, \xi_i\rangle_{\mathcal{T}}$ and $\tilde{\boldsymbol{A}}^{\mathcal{T}_a} \in \mathbb{R}^{p\times m}$ with entries $(\alpha_{ji}^{\mathcal{T}_a}) = \langle X_j - \mu_j, \xi_i\rangle_{\mathcal{T}_a}$, $j = 1, \ldots, p, i = 1, \ldots, m$. Then the vector $\boldsymbol{\Theta}_{\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ with entries $\Theta_{k,\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \boldsymbol{\Sigma}^{\text{row}}\kappa, M)$, for $k = 1, \ldots, p$, can be computed as

$$\boldsymbol{\Theta}_{\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \text{diag}((\boldsymbol{\Sigma}^{\text{row}})^{-1}\tilde{\boldsymbol{A}}^{\mathcal{T}_a}(\boldsymbol{D}^{\text{ker}})^{-1}(\tilde{\boldsymbol{A}}^{\mathcal{T}})'),$$

for each interval subinterval $\mathcal{T}_a, a = 1, \ldots, d$, of $\mathcal{T}$.

Lemma 4.5.2.3. *Let* $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$ *be a rank* $M \in \mathbb{N}$ *multivariate stochastic process as in Theorem 4.4.1.1, with* $\boldsymbol{X}(t) = \boldsymbol{A}'\boldsymbol{\phi}(t)$, $\boldsymbol{\mu}(t) = \boldsymbol{M}_{\boldsymbol{A}}'\boldsymbol{\phi}(t)$, *and* $\kappa(s,t) = \boldsymbol{\phi}'(s)\boldsymbol{\Sigma}^{\text{col}}\boldsymbol{\phi}(t)$, *for* $s, t \in \mathcal{T}$. *Then the following holds,*

$$\Theta_{\mathcal{T}_a,k}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \sum_{j=1}^{p} \frac{1}{\lambda_j^{\text{row}}} v_{j,k}^{\text{row}} (\boldsymbol{a}_k - \boldsymbol{m}_{\boldsymbol{A},k})'\boldsymbol{W}_{\mathcal{T}_a}\boldsymbol{W}^{-1}(\boldsymbol{\Sigma}^{\text{col}})^{-1}(\boldsymbol{A} - \boldsymbol{M}_{\boldsymbol{A}})'\boldsymbol{v}_j^{\text{row}},$$

with $(\lambda_j^{\text{row}}, \boldsymbol{v}_j^{\text{row}})$, $j = 1, \ldots, p$, *the eigenpairs of* $\boldsymbol{\Sigma}^{\text{row}}$, *and* $v_{j,k}^{\text{row}} = \boldsymbol{e}_k'\boldsymbol{v}_j^{\text{row}}$.

A proof is given in Appendix C.4. Using matrix operations we get

$$\boldsymbol{\Theta}_{\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) = \text{diag}((\boldsymbol{\Sigma}^{\text{row}})^{-1}(\boldsymbol{A} - \boldsymbol{M}_{\boldsymbol{A}})\boldsymbol{W}_{\mathcal{T}_a}\boldsymbol{W}^{-1}(\boldsymbol{\Sigma}^{\text{col}})^{-1}(\boldsymbol{A} - \boldsymbol{M}_{\boldsymbol{A}})').$$

4.6 Simulations

Outlier detection in multivariate FDA is particularly challenging as various types of outliers can emerge, such as shifts, shape outliers, isolated spikes, or changes in the dependence structure. We performed a simulation study to assess the performance of different outlier detection methods by simulating multivariate functional data and introducing several types of outliers. In our comparison of outlier detection methods, we examine two categories: distance-based approaches and depth-based methods.

The *distance-based* approach uses Mahalanobis distance, which is either based on the trimmed functional Mahalanobis distance (4.3.1) applied to B-spline coefficient matrices, or the matrix Mahalanobis distance (4.4.3) applied to the raw data. The parameters for the Mahalanobis distance are estimated using classical maximum likelihood estimation (Dutilleul, 1999) or the robust matrix minimum covariance determinant approach (Mayrhofer et al.,

2024a). Both approaches are implemented in the `robustmatrix` R package (Mayrhofer et al., 2024b).As a benchmark, we also include ML estimators of the clean data that are used to compute relative scores to asses mean and covariance estimation.

On the other hand, *depth-based* methods determine outliers by measuring the centrality of a function within a data cloud and define outliers as observations with low depth values, which indicates that they lie far from the central bulk of the data, see, e.g., Zuo and Serfling (2000) for more details. There are various depth-based outlyingness measures for multivariate functional data, such as Stahel-Donoho outlyingness/projection depth (fSDO; Stahel (1981); Donoho (1982); Zuo (2003)), skewness-adjusted outlyingness/skewness-adjusted projection depth (fAO; Hubert and Van der Veeken (2008); Hubert et al. (2015)), or directional outlyingness/directional projection depth (fDO; Rousseeuw et al. (2018)). Those methods are implemented in the `R` package `mrfDepth` (Segaert et al., 2024). Additionally, we consider the Magnitude-Shape Plot (MS; Dai and Genton (2018)) from the `R` package `fdaoutlier` (Ojo et al., 2023), which is based on directional outlyingness as defined by Dai and Genton (2019). These methods are applied to the raw and smoothed data as well as to the coefficient matrices. The depth-based methods are not specifically designed to be used on the coefficient matrices. However, Theorem 4.4.1.1 implies that the distribution of the finite-dimensional projections of a Gaussian process can be transferred to the coefficients. This provides a lower-dimensional representation of the data that is easier to handle computationally and is thus included in the comparison.

4.6.1 Setup

The random functions are drawn from a multivariate Gaussian process with a separable covariance structure, by generating finite-dimensional realizations at $q = 100$ time points.

The data are smoothed using $d = 30$ cubic B-Spine basis functions without a penalty. This choice of basis functions captures the essential functional features. Since the goal of this simulation study is to compare the methods and not determine the best smoothing strategy, we kept this fixed across all settings. We consider sample sizes of $n \in \{300, 1000\}$ observations with $p \in \{3, 10, 50\}$ coordinate functions. For the covariance structure between the coordinate functions we adopt the covariance matrices proposed by Agostinelli et al. (2015b), denoted by $\mathbf{\Sigma}^{row}$, which have random entries and generally yield low correlations. For the covariance function κ we consider both Ornstein-Uhlenbeck κ_{OU} as well as Matérn-type $\kappa_{\text{Matérn}}$ covariance structures, which are defined as

$$\kappa_{\mathrm{OU}}(s,t) = \sigma_1^2 \exp\left(\frac{-|s-t|}{\sigma_2}\right) \quad \text{and} \quad \kappa_{\text{Matérn}}(s,t) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\tau\,|s-t|)^{\nu} K_{\nu}(\tau\,|s-t|),$$

respectively. For the clean data, we use parameters $\sigma_1 = \sqrt{0.3}$ and $\sigma_2 = 0.3$ for the Ornstein-Uhlenbeck covariance function, and $\sigma = 1$, $\tau = 5$, and $\nu = 0.5$ for the Matérn-type covariance function. Except for the isolated outliers, the mean function of every coordinate is given by $\mu_j(t) = 30t(1-t)^{1.5}$, for $j = 1, \ldots, p$. Similar choices for the mean and covariance function were considered by Arribas-Gil and Romo (2014), Dai and Genton (2018), and Oguamalam et al. (2024) for outlier detection in univariate functional data. Outliers are added to the datasets by randomly replacing a fraction $\varepsilon \in \{0.1, 0.3\}$ of the clean observations. Each

simulation setting is replicated 100 times. Four considered outlier settings are visualized in Figure 4.6.1.

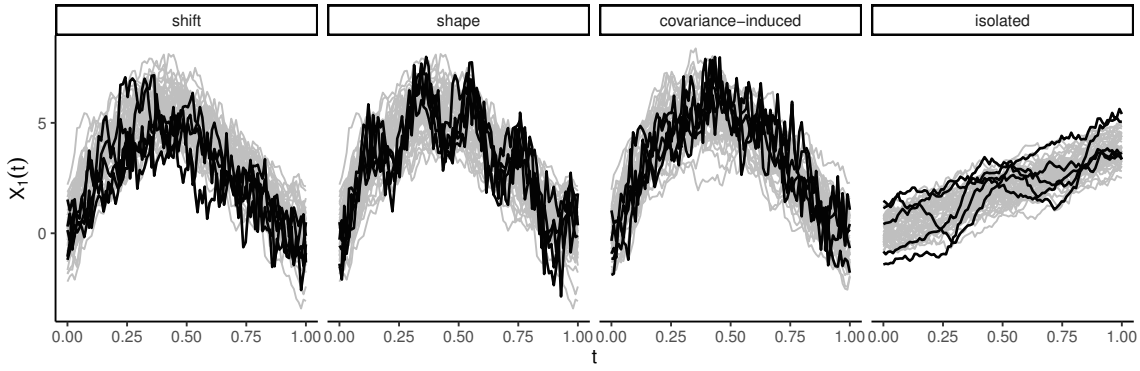


Figure 4.6.1: Visualization of the first coordinate function, $X_1(t)$, of the multivariate process $\boldsymbol{X}(t) = (X_1(t), X_2(t), X_3(t))$, for the 4 outlier types considered in this simulation study. Each plot shows 45 clean sample curves (gray) and 5 outliers (black). For the shift and shape outliers, $\lambda = 9$ and $\kappa = \kappa_{\mathrm{OU}}$. For the covariance-induced outliers, $\nu = 0.2$ and $\tau = 10$, and for the isolated outliers, $\lambda = 0.5$.

Shift outliers are created in all p coordinates by introducing perturbations along the first eigenfunction ξ_1, capturing the largest mode of variation. On the other hand, *shape outliers* are created by perturbing along the tenth eigenfunction ξ_{10}, affecting local features without a global shift of the functions. The coordinate functions are given by

$$X_j^{\mathrm{shift}}(t) = X_j + \lambda\,\xi_1 \quad \text{and} \quad X_j^{\mathrm{shape}}(t) = X_j + \lambda\,\xi_{10},$$

respectively. Here, $j = 1, \ldots, p$, and $\lambda \in \{6, 9, 12, 15\}$.

To introduce *covariance-induced outliers*, we modify the covariance structure using the Matérn-type covariance function. By altering the smoothness parameter $\nu \in \{0.1, 0.2, 0.5\}$ and range parameter $\tau \in \{7, 10, 15\}$, we generate functions with unusually high variability or erratic behavior compared to the regular observations.

The final setting considers *isolated outliers*, which deviate from the regular observations at specific time points, while the remaining function remains unchanged. The mean of the clean data is $\mu_j = 4t$ while the outliers have random mean functions $\mu_j = 4t + \lambda(-1)^u \left(1.8 - \frac{1}{\sqrt{0.02\pi}} \exp\left(\frac{-(t-\alpha)^2}{0.02}\right)\right)$, for $j = 1, \ldots, p$. Here $u \sim \mathit{Bernoulli}(0.5)$ is a Bernoulli random variable, and $\alpha \sim U(0.25, 0.75)$ follows a uniform distribution. In comparison to the first three settings, the mean function is random, and hence, the outliers neither follow the same distribution nor form a cluster.

4.6.2 Results

The outlier detection performance of the methods is compared based on precision, recall, and their harmonic mean, i.e., the F-score. In cases where none of the flagged observations are

identified correctly, precision and F-score may be undefined, as this would lead to division by zero. To handle these cases, we assign the highest rank to them in rank-based comparisons, ensuring they are considered appropriately. For the comparison of the actual performance measures, such cases are excluded from the plots to avoid skewing the results.

To evaluate mean and covariance estimation we consider

$$\frac{1}{p\,|\mathcal{T}|}\int \|\boldsymbol{\mu}(t)-\hat{\boldsymbol{\mu}}(t))\|_2^2\,dt \quad \text{and} \quad \frac{1}{(p\,|\mathcal{T}|)^2}\iint \left\|\boldsymbol{\Sigma}^{\text{row}}k(s,t)-\hat{\boldsymbol{\Sigma}}^{\text{row}}\hat{k}(s,t))\right\|_2^2 dsdt,$$

respectively. Here $|\mathcal{T}| = \max(\mathcal{T}) - \min(\mathcal{T})$ denotes the length of the interval $\mathcal{T}$. For the distance-based methods, these errors can be computed based on the estimated parameters. Since the depth-based approaches are non-parametric, we first apply the outlier detection methods to identify and remove the outliers before robustly computing the maximum likelihood estimates on the cleaned subset. Additionally, relative errors are computed by dividing the estimation errors of each method by the benchmark estimation error attained by the ML estimates computed on the clean data.

Among the depth-based methods, the average results across all settings are only slightly influenced by smoothing. Therefore, only the results based on the raw data are reported. For a comparison of the depth-based methods, see Appendix C.5. For the distance-based methods, smoothing has a more apparent influence, and results for both raw and smoothed data are reported.

To get an overview of the overall performance for all described simulation settings, we compare the methods by ranking them according to precision, recall, F-Score, and mean as well as covariance estimation errors in Figure 4.6.2. All ranks are ordered such that the method with the lowest rank is best. The dots show the average ranks and the intervals are based on non-parametric multiple comparisons using the Friedman and the post-hoc Nemenyi tests; see Hollander (2013) for details. Whenever two methods' intervals do not overlap, they are statistically significantly different at a 99 percent confidence level. All methods that have overlapping intervals with the best method are colored black while the others are gray.

For shift outliers, the robust distances computed on the smoothed data and the MS plot perform best. The distance-based approach performs significantly better than the MS plot for mean estimation and in terms of recall, while the MS plot is better in terms of F-Score and covariance estimation error. For the shape outliers, the robust distances computed on the smoothed data work best. The MS plot is close in terms of recall while the robust distances computed on the raw data are similar in terms of covariance estimation error. For the covariance-induced outliers the distance-based methods work significantly better than the depth-based approaches. The non-robust distances perform very well for outlier detection but have higher mean and covariance estimation errors. Finally, for the isolated outliers, the robust distances computed on the smooth data work best for outlier detection, while the MS plot and the robust distances based on the raw data have the lowest covariance estimation errors. While there is no overall best method for all settings, the robust distances computed on the smoothed and raw data yield the most reliable results for mean and covariance estimation. For outlier detection, the robust distances computed on the smoothed data and the MS plot perform best.

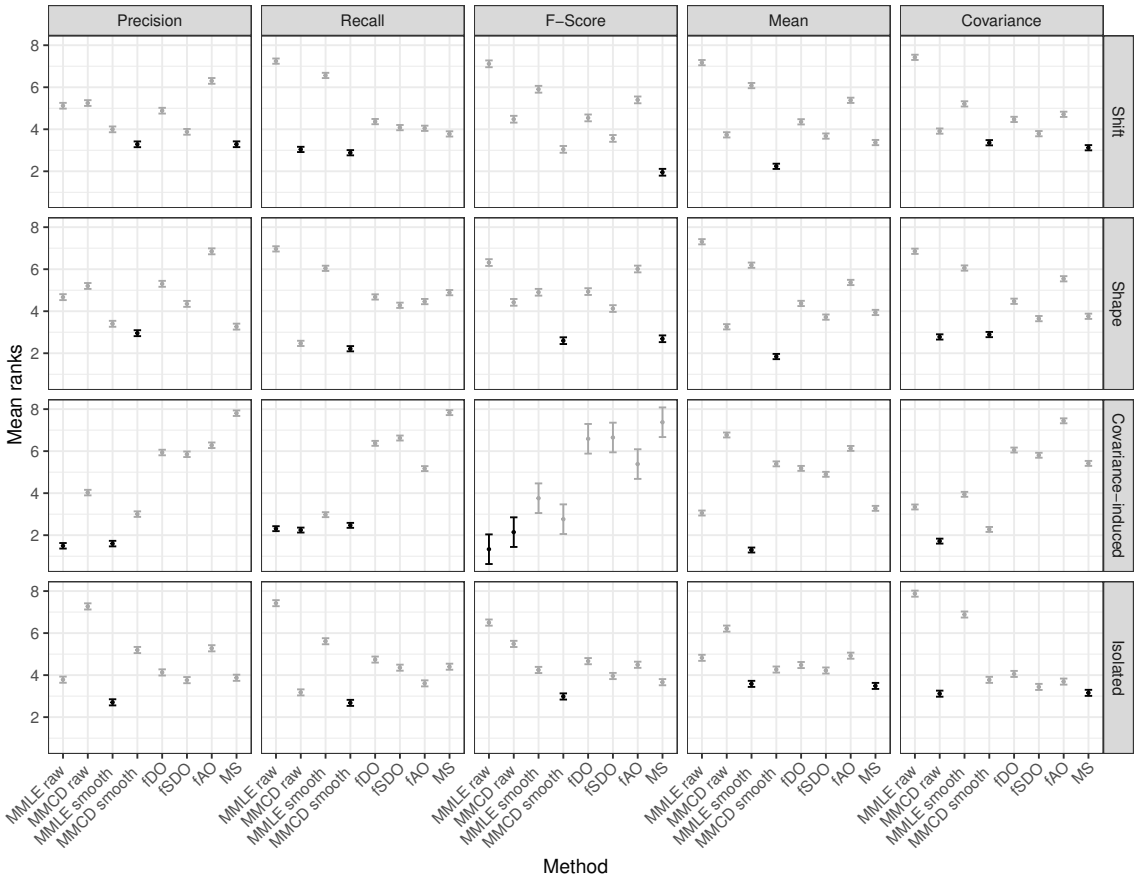


Figure 4.6.2: Rank-based comparison between the methods for all described simulation settings with intervals based on non-parametric multiple comparisons using the Friedman and the post-hoc Nemenyi tests. Methods with overlapping intervals with the best method (lowest average rank) are not statistically significantly different at a 99% confidence level and are colored black, while the others are gray. The horizontal facets describe the different performance metrics, and the vertical facets distinguish between the four outlier types.

To get a more in-depth understanding of the simulation results, we analyze the results for $p \in \{3, 50\}$, $\varepsilon \in \{0.1, 0.3\}$, $n = 1000$, and $\kappa = \kappa_{\mathrm{M}}$ in more detail. We consider two parameter settings for each scenario that either result in slightly outlying observations that are rather difficult to detect (setting A) or more clearly outlying observations that are easier to detect but have a stronger influence on parameter estimation when they remain undetected (setting B). Specifically, for shift and shape outliers, $\lambda = 6$ in setting A and $\lambda = 15$ in setting B. For covariance-induced outliers, we fix $\nu = 0.2$ in both settings and use $\tau = 7$ in setting A and $\tau = 15$ in setting B. Finally, for the isolated outliers, $\lambda = 0.2$ and $\lambda = 1$ in settings A and B, respectively.

Figures 4.6.3 and 4.6.4 compare F-Score and relative covariance estimation errors, respectively. Relative covariance estimation errors close to 1 (dashed-line) indicate that the

method's error is similar to the error of the benchmark approach. Plots of recall, precision, and relative mean estimation errors are given in Appendix C.5 for more details.

For the shift outliers, the MS plot performs best when the fraction of contaminated samples is low. However, the MS plot fails for a higher fraction of contaminated samples and higher dimension ($p = 50$ and $\varepsilon = 0.3$). Robust distances (raw and smoothed) perform well in setting B. The other depth-based methods show mixed results, with strong sensitivity to contamination at $p = 3$ and more stable results at $p = 50$. Covariance estimation errors follow a similar pattern, with no clear best method, but robust distances on smoothed data provide the most reliable results.

For shape outliers, smaller alterations in setting A are best detected by robust distance-based methods across all dimensions. In setting B, depth-based methods improve but are less stable compared to the robust distance-based methods.

For covariance-induced outliers, distance-based methods (robust and non-robust) excel in outlier detection. Covariance estimation is most accurate with robust distance-based methods.

Considering the isolated outliers, robust distances on smooth data offer the most stable results for both outlier detection and covariance estimation. At $p = 3$, the MS plot performs comparably. The other depth-based approaches work well at $\varepsilon = 0.1$ but decline in performance as contamination increases. Robust distances on the raw data perform well but fall short compared to the smoothed approach. At $p = 50$, the MS plot shows extreme differences between settings A and B, with F-scores close to 0 and 1, respectively. The remaining depth-based approaches work well but show decreases in F-Score in setting B at $\varepsilon = 0.1$. Distance-based methods still work well, distances based on raw data occasionally outperform their smoothed counterparts.

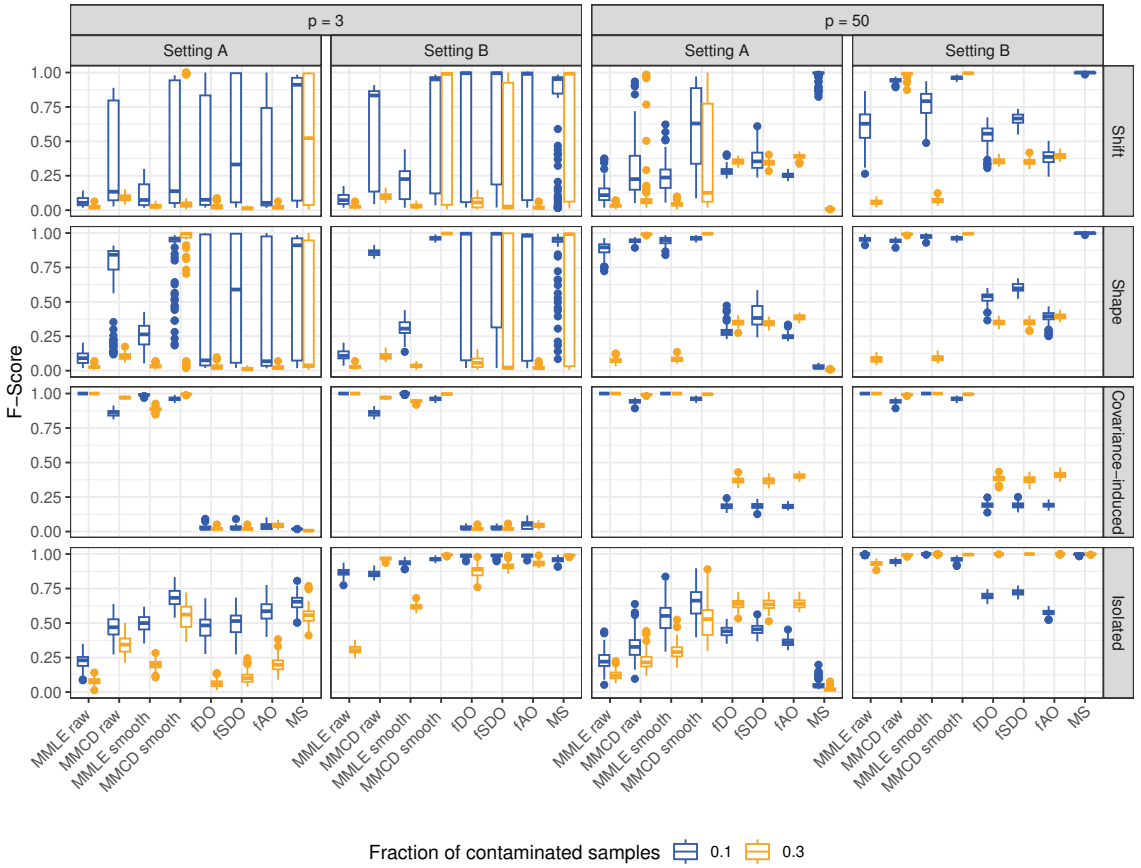


Figure 4.6.3: Comparison of simulation results based on F-Score for $n = 1000$ and $\kappa = \kappa_{\mathrm{M}}$: The top-level horizontal facets represent different dimensionalities, with $p \in \{3, 50\}$. The nested horizontal facets correspond to moderate (Setting A) and severe (Setting B) outliers. The vertical facets distinguish between the four outlier types. Boxplot colors (blue and orange) correspond to contamination levels of $\varepsilon = 0.1$ and $\varepsilon = 0.3$, respectively.

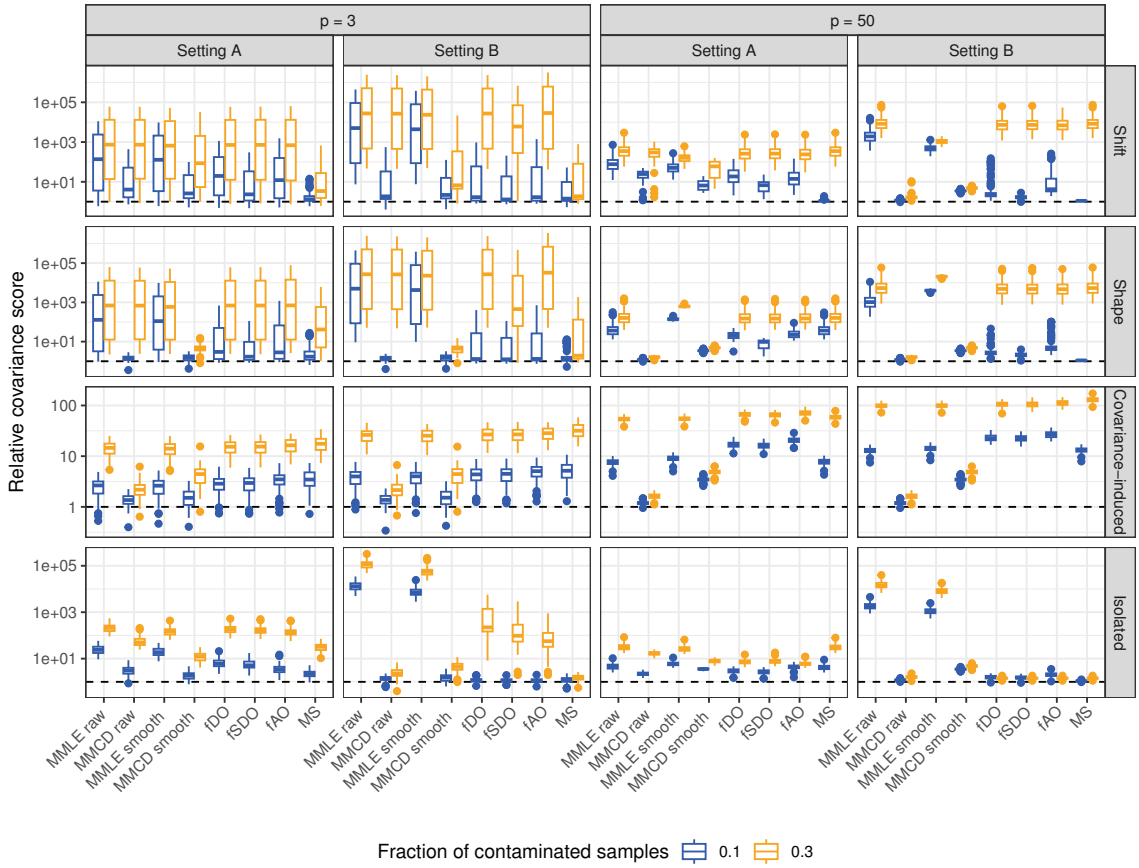


Figure 4.6.4: Comparison of simulation results based on relative covariance estimation errors given on a log-scale for $n = 1000$ and $\kappa = \kappa_{\mathrm{M}}$: The top-level horizontal facets represent different dimensionalities, with $p \in \{3, 50\}$. The nested horizontal facets correspond to moderate (Setting A) and severe (Setting B) outliers. The vertical facets distinguish between the four outlier types. Boxplot colors (blue and orange) correspond to contamination levels of $\varepsilon = 0.1$ and $\varepsilon = 0.3$, respectively.

4.7 Examples

4.7.1 Fertility rates

We analyze the annual age-specific female fertility curves for several countries/regions from the Human Fertility Database (2024). Specifically, we consider the age-specific fertility rate (ASFR), defined as

$$\text{ASFR}(s,t) = \frac{\text{number of live births to women aged } s \text{ in year } t}{\text{population of women aged } s \text{ in year } t},$$

for women aged 15 to 45. We selected the subset of $n = 22$ countries/regions that contain no missing values for the years between 1960 and 2019. To facilitate the interpretability of the results, we aggregate the annual ASFRs into five-year intervals (1960:1964, 1965:1969, ..., 2015:2019), which results in observations that that are naturally arranged in 12×44 matrices for each country. This matrix structure reflects the average ASFRs for each of the 12 five-year periods across the 44 age groups.

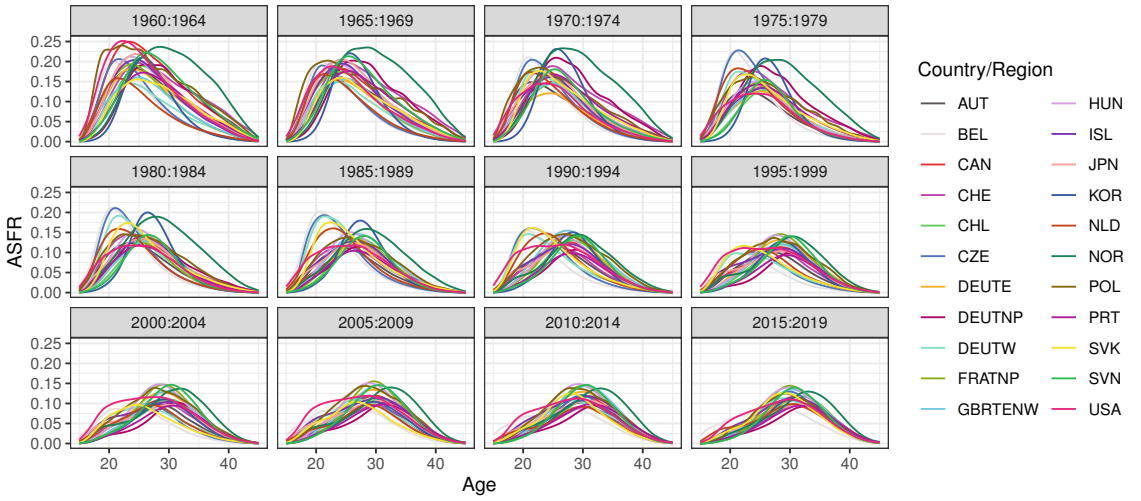
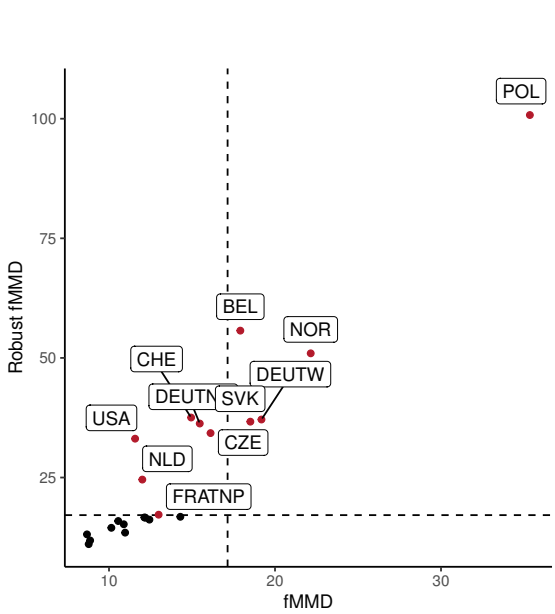


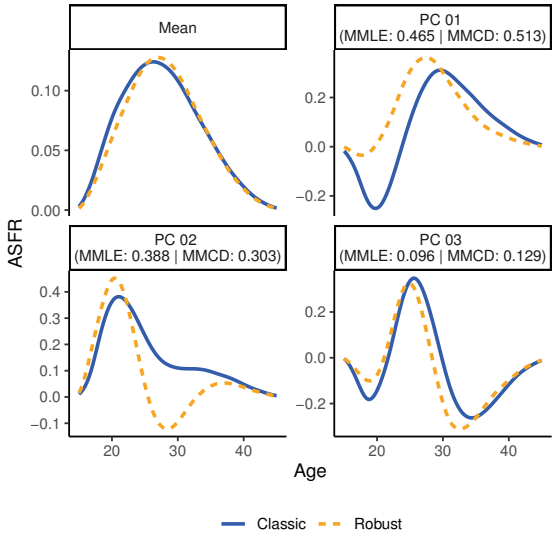
Figure 4.7.1: Smoothed age-specific fertility curves for all 22 countries/regions.

Our goal is to treat these ASFRs as functional data with continuous age s for each of the 12 five-year intervals. To achieve this, the matrix-variate data are transformed into multivariate functional data through the following process. First, a log transformation is applied to the ASFR data. Next, the log-transformed data are smoothed into multivariate functional data using a cubic B-spline basis consisting of 20 basis functions. Finally, the smoothed coefficient matrices are exponentiated to return the data to their original scale. This method ensures the positivity of the smoothed ASFRs, thereby maintaining the inherent characteristics of the fertility rates. The smoothed curves are shown in Figure 4.7.1. Here, every plot shows the fertility curves for one of the five-year intervals. Overall, fertility is declining and women give birth at older ages as time progresses. Moreover, the curves are rather similar in the last period while there is more difference between the countries in the earlier years. We see that some countries/regions form a cluster with left-skewed fertility,

i.e., women give birth at younger ages. On the other hand we see that, for example, Norway (NOR) stands out because of very high fertility in the first years and right-skewed fertility in the later years.



(a) Comparison of classic and robust Mahalanobis distances of the ASFRs.



(b) Comparison of classic (MMLE) and robust (MMCD) mean as well as the first three eigenfunctions. The explained variance of the eigenfunctions is given in parentheses.

Figure 4.7.2: Robust analysis of the smoothed ASFRs.

We compare the classic and robust fMMD of the smoothed samples in the distance-distance plot shown in Figure 4.7.2a. The distances are based on parameter estimates from the MMLE and MMCD procedures, respectively, applied to the coefficient matrices. Outliers are detected using $(\chi^2_{0.99,20\cdot12})^{1/2}$ as the cutoff, where $\chi^2_{0.99,d}$ denotes the 0.99 quantile of a χ^2-distribution with d degrees of freedom. When comparing the classic and robust fMMD, we observe that several countries, such as the US (USA) and Germany (DEUTNP), are masked when using the classic distance.

The influence of outliers on the principal component functions is illustrated in Figure 4.7.2b. The classic and robust mean functions are similar, but their eigenfunctions are quite different. The first classic eigenfunction captures the fertility differences between younger and older women across countries/regions and explains around 47% of the variance, while the first robust eigenfunction, which mirrors the form of the robust mean, indicates overall high or low fertility levels and explains around 51% of the variance. This difference arises because the robust eigenfunctions are estimated from a subset of the data that excludes countries with exceptionally high fertility rates among young women. The second robust eigenfunction closely resembles the first classic eigenfunction and explains 31% of the variance, while the second classic eigenfunction highlights generally higher fertility rates, particularly for younger women, and explains 39% of the variance. The first two robust eigenfunctions provide a clear and coherent interpretation, i.e., overall high or low fertility from the first eigenfunction and

higher and lower fertility for younger or older women based on the second eigenfunction, whereas the classic eigenfunctions exhibit contrasting effects on the fertility of young women, complicating their interpretation. The third robust and classical eigenfunctions explain about 10% and 13% of the variance, respectively, and are quite similar, showing a concentrated increase in fertility around age 25, with lower fertility rates for both younger and older women.

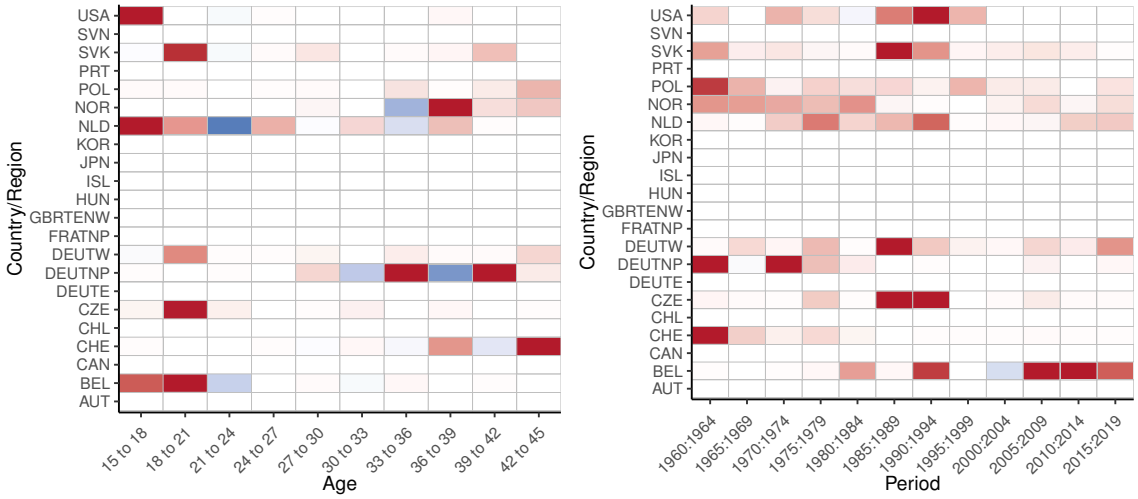


Figure 4.7.3: Age-specific (left) or year-specific (right) outlyingness contributions based on Shapley values for the smoothed ASFRs.

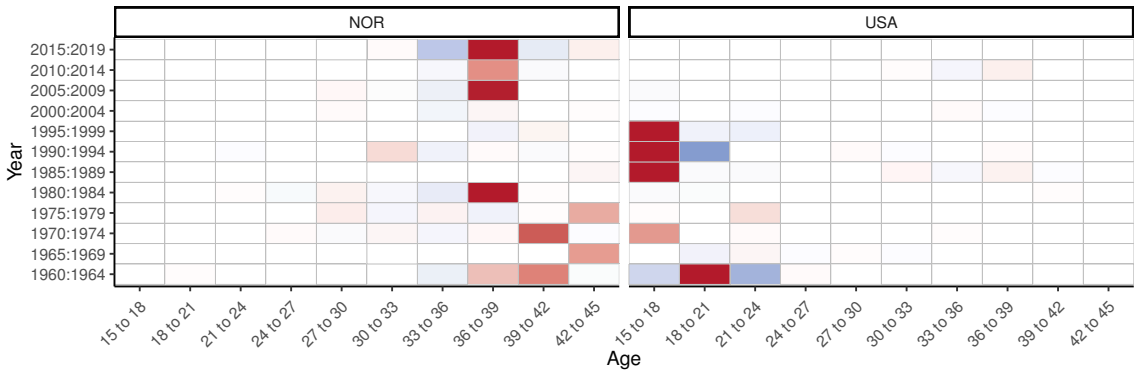


Figure 4.7.4: Age-specific and year-specific outlyingness contributions based on Shapley values for the smoothed ASFRs of Norway (left) and the US (right).

We also analyze the year-specific and/or age-specific outlyingness contributions based on Shapley values for the outlying countries, visualized in Figures 4.7.3 and 4.7.4. Here, positive contributions are colored red, while negative contributions are colored blue. The color intensity depends on the absolute value of the contributions. For Belgium (BEL), Czech Republic (CZE), Netherlands (NLD), Slovakia (SVK), and the US we see high outlyingness

contributions for young women. On the other hand, countries like Switzerland (CHE) and Norway stand out due to the high outlyingness contributions of older women. When we compare the outlyingness over the years, we see that Poland (POL) shows moderately high contributions over almost all years while countries like Belgium stand out due to changes in fertility dynamics more recently. For Norway and the US, we consider a more detailed analysis in Figure 4.7.4. Norway shows the highest outlyingness contributions for women aged 36 to 39 in recent periods and for women aged 39 to 45 in the earlier time periods. Analyzing Figure 4.7.1, we see that Norway shows an almost linear decay in ASFR for older women between 1960 and 1984 while the other countries show more of an exponential decay in ASFR for the same age group, explaining the higher outlyingness contributions for those years and ages. Additionally, from 2005 onwards, Norway is the only country where the ASFR peaks for women older than 30 and is clearly higher up to an age of 40. In contrast, the US stands out because of the high fertility of young women from 1960 to 1964 as well as 1985 to 1999. This is reflected by an upward shift of the smooth ASFR curves for young women as well as by the difference in the curvature of the functions.

4.7.2 El Niño la Niña data

El Niño–Southern Oscillation (ENSO) is a periodic climate phenomenon describing recurring changes in sea surface temperature (SST) and atmospheric pressure in the equatorial Pacific Ocean (Trenberth, 1997). The ENSO phenomenon thereby impacts temperature and precipitation patterns across the planet. ENSO is divided into three states: *El Niño* refers to the warm phase, *La Niña* to the cold phase, and in the middle of the continuum there is the *Neutral* state.

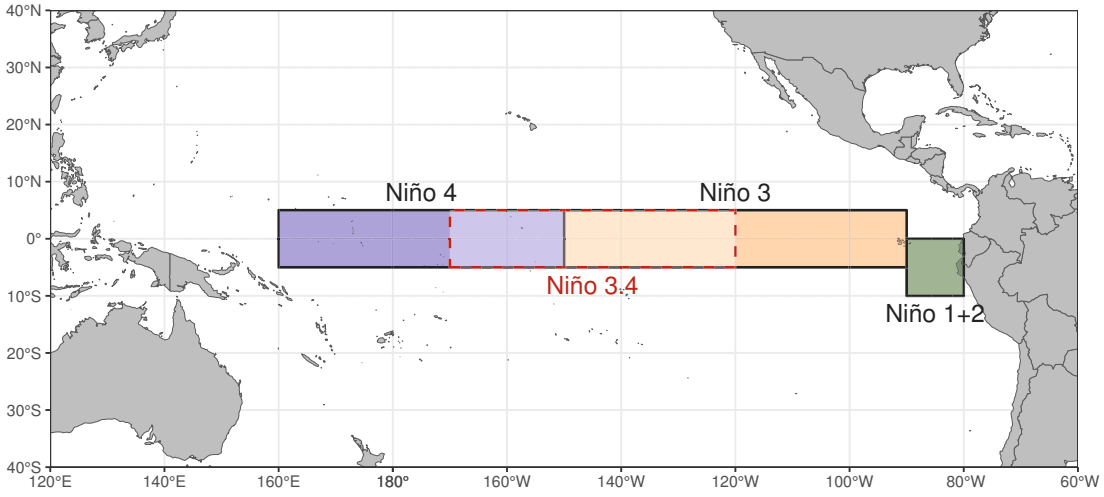


Figure 4.7.5: Map of the 4 regions in the equatorial Pacific Ocean where SST related to ENSO are measured: Niño 1+2 (0° to 10°S, 90°W to 80°W), Niño 3 (5°N to 5°S, 150°W to 90°W), Niño 4 (5°N to 5°S, 160°E to 150°W), Niño 3.4 (5°N to 5°S, 170°W to 120°W).

We analyze the SST data provided by the US Climate Prediction Center (CPC). Specifically,

we consider the monthly SST measurements (datasource: `https://www.cpc.ncep.noaa.gov/data/indices/`) based on the *Extended Reconstructed Sea Surface Temperature, Version 5* (ERSSTv5), see Huang et al. (2017) for more details. The CPC defines the onset of an El Niño/La Niña episode when the 3-month average SST anomaly exceeds +/-0.5°C in the Niño 3.4 region shown in Figure 4.7.5. Here, SST anomalies are computed as the differences between the 3-month average SST and a 30-year baseline average. This method of classification of the phases is known as the *Oceanic Niño Index*, and the classification across different institutions is similar but not identical; see, e.g., Trenberth (1997), for a discussion.

The threshold for El Niño/La Niña events is further divided into four categories: weak (0.5 to 0.9°C absolute SST anomaly), moderate (1.0 to 1.4°C), strong (1.5 to 1.9°C), and very strong ($\geq$ 2.0°C) episodes. Instead of grouping the months into calendar years, they are grouped into 12-year periods, running from June to May. A period is classified according to the most intense episode (El Niño or La Niña) that occurs within it. If no such episodes are present, the period is classified as neutral.

We consider the 74 periods from 1950-1951 to 2023-2024. In each period we have 12 monthly SST measurements for each of the four regions depicted in Figure 4.7.5. In total, there are 573 neutral months, 223 La Niña months, and 92 El Niño months in this dataset. However, there are only 18 neutral periods while there are 36 La Niña periods and 18 El Niño periods. Our goal is to compare a univariate analysis of the SST data based on the region *Niño 3.4* with a multivariate approach that incorporates the SST measurements of all 4 regions.

As a first step, the raw data are smoothed using 6 B-spline basis functions without any penalty. We then compare the robust fMD based on the SST data of region *Niño 3.4* and the fMMD based on SST data of all 4 regions in the distance-distance plot in Figure 4.7.6. The robust distances are based on parameter estimates from the MCD (Rousseeuw and Van Driessen, 1999) and MMCD (Mayrhofer et al., 2024a) procedures applied to the coefficient vectors or coefficient matrices, respectively. The vertical line represents the $(\chi^2_{0.99,6})^{1/2}$ cutoff for fMD, and the horizontal line is its multivariate counterpart $(\chi^2_{0.99,24})^{1/2}$ for fMMD. The dots represent the 74 periods from 1950:1951 to 2023:2024 and their size indicate the number of months during which an El Niño or La Niña event occurred. The color of the dots corresponds to the strongest sea surface temperature (SST) anomaly recorded in each period. The multivariate treatment of the data allows us to identify more unusual observations, many of which are (very) strong El Niño periods that are not outlying based on the univariate distance. This suggests that considering the SST measurements of all four Niño regions is better suited to identify extreme periods.

Figure 4.7.7 displays the smoothed SST measurements of all $n = 74$ periods. Regular observations are depicted in gray while outliers are colored in either red or blue, indicating whether they correspond to El Niño or La Niña periods, respectively. The thicker black line is the smoothed robust mean function. To avoid label overlap, each period is labeled with the last two digits of its starting year (e.g., label 50 represents the period 1950:1951). Most outlying La Niña periods show the same curvature as the regular observations but with a downward shift, while the detected El Niño periods often show both a different curvature and an upward shift.

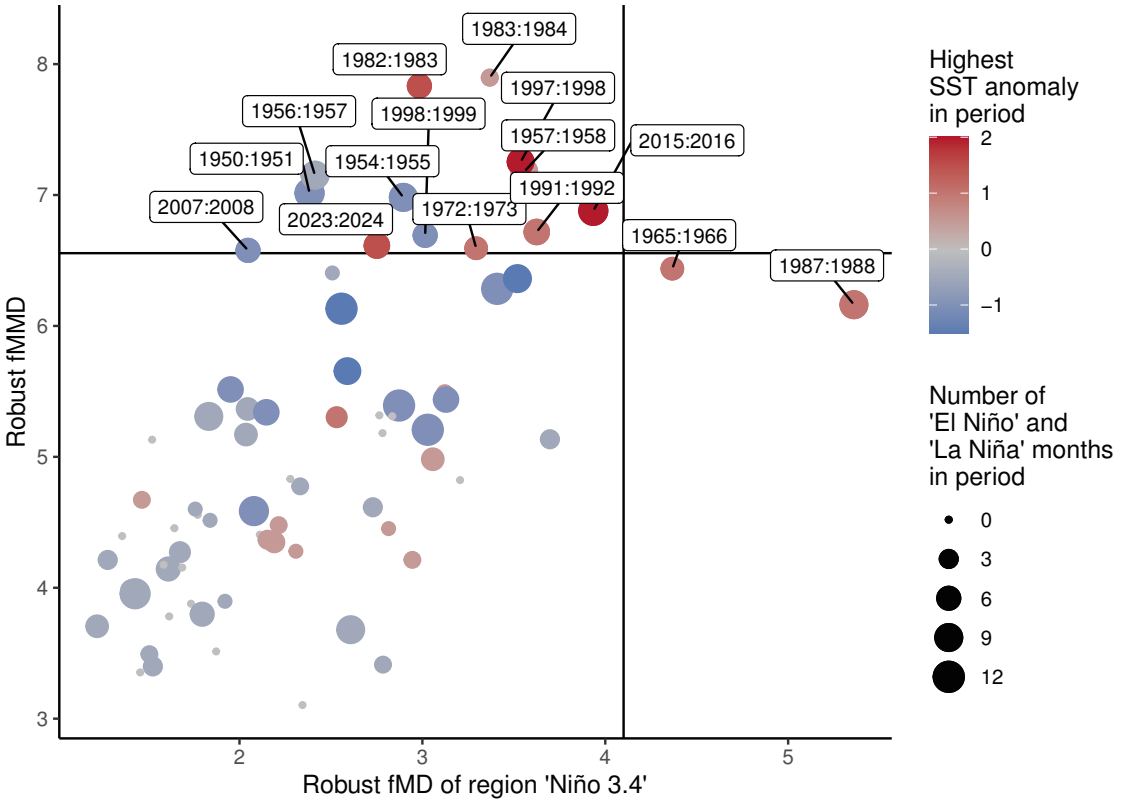


Figure 4.7.6: Comparison of robust fMD based on the region *Niño 3.4* and fMMD based on all 4 regions.

In combination with Figure 4.7.8 we gain a deeper insight into the data. This depicts the month- and station-specific outlyingness contributions based on Shapley values for the eight periods with the highest outlyingness. Here positive contributions are shown in red, while negative contributions are in blue. For instance, during the period 1950:1951, the SST remains declining in Niño 4 in September and October while the mean rises. Furthermore, the SST measurements in Niño 1+2 and 3.4 are one of the lowest observed. In the period 1957:1958, there is a steep decline in SST in Niño 1+2 contrasted by almost constant SST in Niño 4 early in the period; this joint behavior is contrary to the main trend. Additionally, the curvature of the SST function in Niño 3.4 deviates from the mean in both July and January-February. For 2015:2016, we observe an abrupt decline in SST in Niño 3 near the end of the period, accompanied by unusually high, stable SST in Niño 4. The periods 1950:1951 and 1957:1958 do not stand out as clearly in Niño 3.4 and would remain undetected by the univariate approach based on SST of only Niño 3.4. On the other hand, the period 2015:2016 stands out visually in all regions. While it still falls short of the univariate detection threshold, the joint analysis clearly reveals unusual behavior.

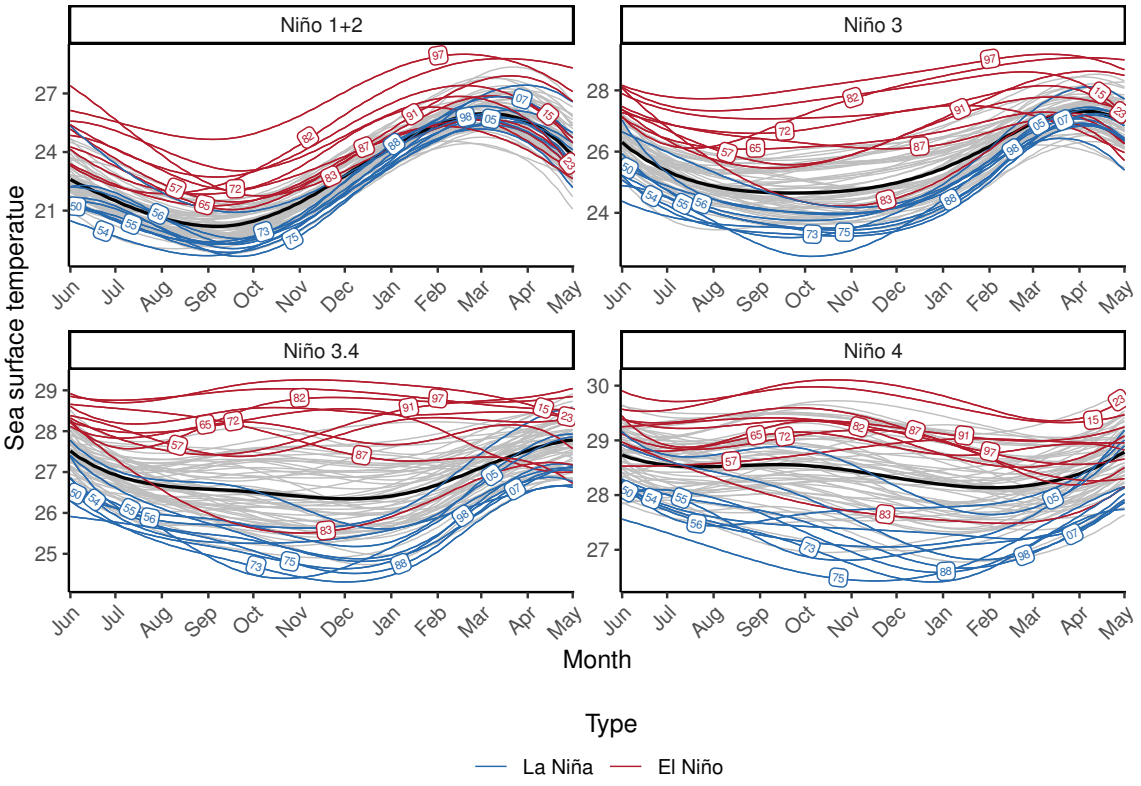


Figure 4.7.7: Smoothed SST measurements of all four Niño regions with outlying observations colored either red or blue, depending on whether the outliers are El Niño or La Niña periods.

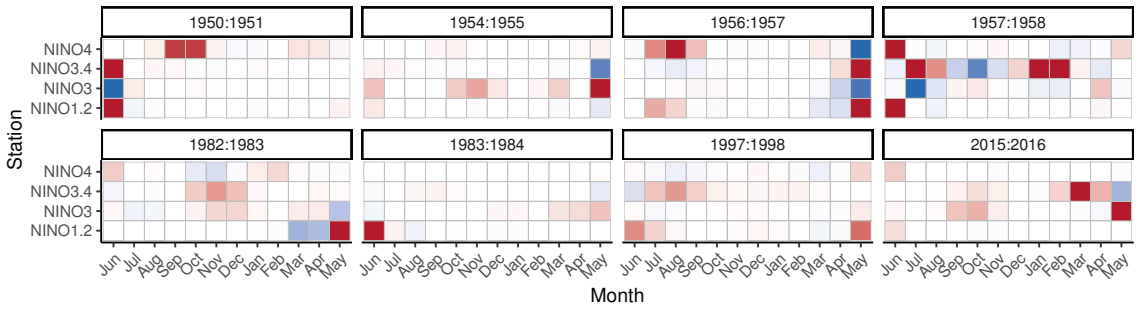


Figure 4.7.8: Month- and station-specific outlyingness contributions based on Shapley values.

4.8 Discussion and conclusions

In this paper, we introduce an approach to multivariate functional outlier detection based on the Mahalanobis distance, a method commonly used for traditional multivariate data. While most existing methods rely on various depth measures (Hubert et al., 2015), our approach

offers an alternative distance-based approach from classical multivariate analysis. While it is not obvious how to define a Mahalanobis distance for multivariate functional data, there is the additional difficulty of robustly estimating a corresponding covariance. We proposed the functional Mahalanobis (semi-)distance and derived several properties such as affine invariance. If the components of the multivariate process are uncorrelated, the multivariate functional Mahalanobis distance reduces to the sum of the functional Mahalanobis distance of the individual components introduced by Galeano et al. (2015). This can be further exploited with the assumption of separability of the multivariate covariance operator. In that case we have shown how a multivariate random function is connected to the distribution of the random coefficient matrix resulting from a basis representation of the smoothed functions. As a consequence it is possible to link the trimmed multivariate functional Mahalanobis distance with the matrix-variate Mahalanobis distance, for which robust parameter estimates have been proposed in Mayrhofer et al. (2024a) to reliably identify outliers.

A further contribution is outlier explainability by means of Shapley values, which allow for an additive decomposition of the univariate and multivariate squared functional Mahalanobis distance. The outlyingness contributions can be evaluated for the individual components of the multivariate functions, for non-overlapping time domains, or simultaneously as time-coordinate-specific contributions. The strength of these diagnostics was demonstrated in the example section.

The proposed robust distances can either be computed with raw data or with a smoothed version, and performance results were reported in the simulation section, where both versions show good performance. However, there are computational arguments for using smoothed data: First, the parameter estimation and distance computation for the smoothed data is performed on the coefficient matrix, which has much lower dimension than the raw data matrix, resulting in a 5- to 10-fold increase in computation speed, see Appendix C.5 for a detailed comparison. Second, the required sample size and breakdown point of the MMCD estimators depend on the ratio between the number of rows and columns. As this ratio approaches one, fewer samples are needed, and the breakdown point increases.

This paper provides a framework for the generalization of other Mahalanobis distance-based outlier detection approaches, like the α-Mahalanobis distances proposed by Berrendero et al. (2020). We leave the exploration of these generalizations for future research. As demonstrated in the examples, robust covariance estimates can be used for FPCA. Furthermore, robust Mahalanobis distance can subsequently be used for clustering and classification purposes.

Appendix C

C.1 Further Preliminaries

Matrix Normal Distribution

A random matrix $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ follows a matrix normal distribution, denoted as $\boldsymbol{X} \sim \mathcal{MN}(\mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$, with mean $\mathbf{M} \in \mathbb{R}^{p\times q}$, row covariance $\boldsymbol{\Sigma}^{\text{row}} \in \text{PDS}(p)$, and column covariance $\boldsymbol{\Sigma}^{\text{col}} \in \text{PDS}(q)$, if and only if its vectorized form $\text{vec}(\boldsymbol{X}) \in \mathbb{R}^{pq}$ has a multivariate normal distribution $\mathcal{N}(\text{vec}(\mathbf{M}), \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}})$ (Gupta and Nagar, 1999). Here, the class of all positive definite symmetric $a \times a$ matrices is denoted by $\text{PDS}(a)$. The vectorization operator $\text{vec}(\cdot)$ stacks the columns of a matrix on top of each other, and $\otimes$ represents the Kronecker product. The probability density function (pdf) of a matrix normal random variable $\boldsymbol{X}$ is given by

$$f(\boldsymbol{X} \,|\, \mathbf{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}) = \frac{\exp(-\frac{1}{2}\,\text{tr}((\boldsymbol{\Sigma}^{\text{col}})^{-1}(\boldsymbol{X}-\mathbf{M})'(\boldsymbol{\Sigma}^{\text{row}})^{-1}(\boldsymbol{X}-\mathbf{M})))}{(2\pi)^{pq/2}\det(\boldsymbol{\Sigma}^{\text{col}})^{p/2}\det(\boldsymbol{\Sigma}^{\text{row}})^{q/2}}. \tag{C.1}$$

Importantly, $\boldsymbol{\Sigma}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}}$ are only identified up to a multiplicative constant $\kappa \neq 0$. Specifically, replacing $\boldsymbol{\Sigma}^{\text{row}}$ by $\kappa\boldsymbol{\Sigma}^{\text{row}}$ and $\boldsymbol{\Sigma}^{\text{col}}$ by $1/\kappa\boldsymbol{\Sigma}^{\text{col}}$ leafs the pdf (C.1) unchanged. To resolve the non-identifiability, one can fix a diagonal entry, the determinant, or norm of either matrix (Roś et al., 2016; Soloveychik and Trushin, 2016).

Multivariate Gaussian Process

A stochastic process $\boldsymbol{X}$ is a multivariate Gaussian process if every finite collection of realizations has a joint normal distribution. A finite collection of realizations from a stochastic process $\boldsymbol{X}$ at time points $\boldsymbol{t} = (t_1, \ldots, t_q), t_1 < t_2 < \cdots < t_q, t_k \in \mathcal{T}, k = 1, \ldots, q$, is denoted by $\boldsymbol{X}_{\boldsymbol{t}} = (\boldsymbol{X}(t_1), \ldots, \boldsymbol{X}(t_q))' \in \mathbb{R}^{p\times q}$, yielding a matrix-variate sample. For the mean function, we have $\mathbf{M}_{\boldsymbol{t}} = (\boldsymbol{\mu}(t_1), \ldots, \boldsymbol{\mu}(t_q))' \in \mathbb{R}^{p\times q}$, and the covariance function yields a block-partitioned matrix $\mathbf{K}_{\boldsymbol{t}} \in \mathbb{R}^{pq\times pq}$ as in Equation (4.2.2), with entries $\kappa_{ij}(t_k, t_l)$, for $i, j = 1, \ldots, p$ and $k, l = 1, \ldots, q$.

The joint normality of $\boldsymbol{X}_{\boldsymbol{t}}$ is described using a matrix-variate approach, which directly models the matrix-valued realizations as in Chen et al. (2017, 2023). In this case, $\boldsymbol{X}_{\boldsymbol{t}} \sim \mathcal{MN}(\boldsymbol{M}_{\boldsymbol{t}}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}_{\boldsymbol{t}})$, where $\boldsymbol{\Sigma}^{\text{row}}$ represents the row-wise covariance matrix, capturing the dependencies between individual coordinate functions, and $\boldsymbol{\Sigma}^{\text{col}}_{\boldsymbol{t}} = (\kappa(t_k, t_l))_{k,l=1}^{q} \in \mathbb{R}^{q\times q}$ is the column-wise covariance matrix, accounting for the temporal correlations between different time points. This formulation leverages a single kernel function $\kappa(s, t)$ to model time dependencies, allowing the covariance structure to be factorized as $\mathbf{K}(s, t) = \boldsymbol{\Sigma}^{\text{row}}\kappa(s, t)$, significantly reducing complexity compared to using $p(p + 1)$ separate kernels, as mentioned in Equation (4.2.2).

Remark C.1.1. *Alternatively, the joint normality can also be expressed by vectorizing the matrix-valued realizations, as in Alvarez et al. (2012). In this case, the columns of* $\boldsymbol{X_t}$ *are stacked into a vector* $\boldsymbol{x_t} = \text{vec}(\boldsymbol{X_t})$*, and the process is modeled as* $\boldsymbol{x_t} \sim \mathcal{N}(\boldsymbol{m_t}, \mathbf{K_t})$*, where* $\boldsymbol{m_t} = \text{vec}(\boldsymbol{M_t})$.

Finite-basis Representation and Additive Noise Model

One of the key principles of FDA is to work with *smooth* functions. This means that adjacent values are linked together to some degree and are unlikely to be too different from each other. If the functions were not smooth, there would be no significant advantage to treating it as functional data instead of just multivariate (Ramsay and Silverman, 2005).

In practice, functional data are observed at discrete time points and the raw observed data may contain noise or fluctuations that obscure the underlying patterns or trends present in the true functional form. We can formalize this using an additive noise model, see, e.g., Ramsay and Silverman (2005) and Zhu et al. (2016),

$$\boldsymbol{Y}(t) = \begin{pmatrix} Y_1(t) \\ \vdots \\ Y_p(t) \end{pmatrix} = \begin{pmatrix} X_1(t) \\ \vdots \\ X_p(t) \end{pmatrix} + \begin{pmatrix} \varepsilon_1(t) \\ \vdots \\ \varepsilon_p(t) \end{pmatrix} = \boldsymbol{X}(t) + \boldsymbol{\varepsilon}(t),$$

where $\boldsymbol{Y}$ is the observed process, and $\boldsymbol{X}$ is an underlying *signal* process we are interested in, and $\boldsymbol{\varepsilon}$ are additive errors that are independent of $\boldsymbol{X}$.

To reduce the noise and reveal the underlying structure or behavior of the process we employ smoothing techniques. The most common approach is to represent the observed process by basis functions. Let $\{\phi_k\}_{k\geq 1}$ denote a family of orthonormal basis functions of $L^2(\mathcal{T})$, then each component X_j of $\boldsymbol{X} \in \mathcal{H}$ can be expressed in terms of this basis as

$$X_j = \sum_{k=1}^{\infty} a_{jk}\, \phi_k,\ j = 1, \ldots, p.$$

Because the basis is fixed, the randomness of the stochastic process $\boldsymbol{X}$ is captured by the coefficients a_{jk}, $j = 1, \ldots, p$, $k \geq 1$. Using only a sufficiently large number m of basis functions, we can approximate X_j arbitrarily well and rewrite the coordinates of the observed process as

$$Y_j(t) = \sum_{k=1}^{m} a_{jk}\, \phi_k(t) + \tilde{\varepsilon}_j(t),$$

where the error term now consists of the approximation error and the measurement errors, i.e.,

$$\tilde{\varepsilon}_j(t) = \varepsilon_j(t) + \sum_{k=m+1}^{\infty} a_{jk}\, \phi_k(t).$$

Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$ denote the vector consisting of the first m basis functions of $\{\phi_k\}_{k\geq 1}$ and $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jm})' \in \mathbb{R}^m$ the vector of coefficients, we can write

$$Y_j(t) = \boldsymbol{a}_j' \, \boldsymbol{\phi}(t) + \tilde{\varepsilon}_j(t), \quad j = 1 \ldots, p.$$

By collecting the coefficients in a matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p)' \in \mathbb{R}^{p \times m}$ we can write the multivariate process as

$$\boldsymbol{Y}(t) = \boldsymbol{A}\,\boldsymbol{\phi}(t) + \tilde{\boldsymbol{\varepsilon}}(t).$$

The coefficients a_{jk}, $j = 1, \ldots, p$, $k = 1, \ldots, m$, are usually determined based on a least squares approach, and often a roughness penalty is involved; see Ramsay and Silverman (2005) for more details.

FPCA for Multivariate Stochastic Process with Separable Covariance

Let $\mathcal{K}$ denote the covariance operator corresponding to the covariance kernel $\kappa(s,t)$ of the multivariate stochastic process $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$. Then we have that

$$\mathcal{K}\,\xi(s) = \int_{\mathcal{T}} \kappa(s,t)\,\xi_i(t)\mathrm{d}t = \lambda_i^{\text{ker}}\,\xi_i(s), \quad i = 1, \ldots, m,$$

and

$$\boldsymbol{\Sigma}^{\text{row}}\boldsymbol{v}_j^{\text{row}} = \lambda_j^{\text{row}}\boldsymbol{v}_j^{\text{row}}, \quad j = 1, \ldots, p.$$

For the multivariate covariance operator $\boldsymbol{\mathcal{K}}$ with corresponding kernel $\mathbf{K}(s,t)$ we have

$$\boldsymbol{\mathcal{K}}\,\boldsymbol{\psi}_k(s) = \int_{\mathcal{T}} \mathbf{K}(s,t)\,\boldsymbol{\psi}_k(t)\mathrm{d}t = \pi_k\,\boldsymbol{\psi}_k(s), \quad k = 1, \ldots, M, \tag{C.2}$$

where $M = pm$. In the separable setting, $\boldsymbol{\mathcal{K}} = \boldsymbol{\Sigma}^{\text{row}}\mathcal{K}$ with kernel $\mathbf{K}(s,t) = \boldsymbol{\Sigma}^{\text{row}}\kappa(s,t)$. Consider the eigendecomposition $\boldsymbol{\Sigma}^{\text{row}} = \boldsymbol{V}^{\text{row}}\boldsymbol{D}^{\text{row}}(\boldsymbol{V}^{\text{row}})'$, where $\boldsymbol{V}^{\text{row}} = ((\boldsymbol{v}_1^{\text{row}})', \ldots, (\boldsymbol{v}_p^{\text{row}})')$ is the matrix of eigenvectors and $\boldsymbol{D}^{\text{row}} = \operatorname{diag}(\lambda_1^{\text{row}}, \ldots, \lambda_p^{\text{row}})$ the diagonal matrix of ordered eigenvalues $\lambda_1^{\text{row}} \geq \cdots \geq \lambda_p^{\text{row}}$, where we assume the uniqueness of the eigenvalues for simplicity; a proper generalization of the results holds also in the case of non-simple eigenvalues. Using the indexation $k = k(i,j) = 1, \ldots, M = pm$,

$$\pi_k = \lambda_i^{\text{ker}}\lambda_j^{\text{row}}, \quad \boldsymbol{\psi}_k(t) = \xi_i(t)\boldsymbol{v}_j^{\text{row}}, \quad t \in \mathcal{T}, \quad i = 1, \ldots, m, \quad j = 1 \ldots, p. \tag{C.3}$$

To see that the relations in (C.3) indeed hold, observe first that orthogonality of $\boldsymbol{V}^{\text{row}}$ and orthonormality of ξ_i, $i = 1, \ldots, m$ give the orthonormality of the corresponding products. Furthermore,

$$\begin{aligned} \lambda_i^{\text{ker}}\lambda_j^{\text{row}}(\xi_i(s)\boldsymbol{v}_j^{\text{row}}) &= \lambda_j^{\text{row}} \int_{\mathcal{T}} \kappa(s,t)\xi_i(t)\boldsymbol{v}_j^{\text{row}}\,\mathrm{d}t \\ &= \int_{\mathcal{T}} \boldsymbol{\Sigma}^{\text{row}}\kappa(s,t)\xi_i(t)\boldsymbol{v}_j^{\text{row}}\,\mathrm{d}t = \int_{\mathcal{T}} \mathbf{K}(s,t)(\xi_i(t)\boldsymbol{v}_j^{\text{row}})\,\mathrm{d}t. \end{aligned} \tag{C.4}$$

The uniqueness of the eigendecomposition (C.2) yields the desired claim. For a more general connection between multivariate FPCA of $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{\mu}, \mathbf{K})$ and univariate FPCA of its components $X_1, \ldots, X_p$, see Happ and Greven (2018).

C.2 Mahalanobis Distance Proofs

Proof of Lemma 4.3.1.1. The affine equivariance of the mean and the covariance follow directly from the affine equivariance of their multivariate counterparts:

$$\boldsymbol{\mu}_{\boldsymbol{Y}} = \mathbb{E}[\boldsymbol{Y}(t)] = \mathbb{E}[\boldsymbol{A}\boldsymbol{X}(t) + \boldsymbol{\nu}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}(t)] + \boldsymbol{\nu} = \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{\nu},$$

$$\begin{aligned}
\mathbf{K}_{\boldsymbol{Y}}(s,t) = \operatorname{cov}(\boldsymbol{Y}(s), \boldsymbol{Y}(t)) &= \mathbb{E}[(\boldsymbol{Y}(s) - \mathbb{E}[\boldsymbol{Y}(s)])(\boldsymbol{Y}(s) - \mathbb{E}[\boldsymbol{Y}(t)])'] \\
&= \mathbb{E}[(\boldsymbol{A}\boldsymbol{X}(s) - \mathbb{E}[\boldsymbol{A}\boldsymbol{X}(s)])(\boldsymbol{A}\boldsymbol{X}(s) - \mathbb{E}[\boldsymbol{A}\boldsymbol{X}(t)])'] \\
&= \mathbb{E}[(\boldsymbol{A}\boldsymbol{X}(s) - \boldsymbol{A}\mathbb{E}[\boldsymbol{X}(s)])(\boldsymbol{A}\boldsymbol{X}(s) - \boldsymbol{A}\mathbb{E}[\boldsymbol{X}(t)])'] \\
&= \boldsymbol{A}\mathbb{E}[(\boldsymbol{X}(s) - \mathbb{E}[\boldsymbol{X}(s)])(\boldsymbol{X}(s) - \mathbb{E}[\boldsymbol{X}(t)])']\boldsymbol{A}' \\
&= \boldsymbol{A}\,\mathbf{K}_{\boldsymbol{X}}(s,t)\boldsymbol{A}'.
\end{aligned}$$

For simplicity of the notation assume $\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{\nu} = \mathbf{0}$, and denote $\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}$ and $\boldsymbol{\mathcal{K}}_{\boldsymbol{Y}}$ to be the (matrices of) covariance operators associated with $\mathbf{K}_{\boldsymbol{X}}$ and $\mathbf{K}_{\boldsymbol{Y}}$, respectively. Then, for any $\boldsymbol{f} \in \mathcal{H}$, $u \in \mathcal{T}$,

$$\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}\,\boldsymbol{f}(u) = \int_{\mathcal{T}} \mathbf{K}_{\boldsymbol{X}}(u,v)\boldsymbol{f}(v)\mathrm{d}v = \boldsymbol{A}\int_{\mathcal{T}} \mathbf{K}_{Y}(u,v)(\boldsymbol{A}'\boldsymbol{f}(v))\mathrm{d}v = \boldsymbol{A}\,\boldsymbol{\mathcal{K}}_{\boldsymbol{Y}}(\boldsymbol{A}'\boldsymbol{f})(u).$$

Denoting $\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)}$ and $\boldsymbol{\mathcal{K}}_{\boldsymbol{Y}}^{(M)}$ to be the truncation of $\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}$ and $\boldsymbol{\mathcal{K}}_{\boldsymbol{Y}}$ onto the corresponding first M components, i.e., for $\boldsymbol{f} \in \mathcal{H}$,

$$\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)}\,\boldsymbol{f}(u) = \sum_{i=1}^{M} \pi_{\boldsymbol{X},i}\langle \boldsymbol{f}, \boldsymbol{\psi}_{\boldsymbol{X},i}\rangle\,\boldsymbol{\psi}_{\boldsymbol{X},i}(u), \quad \boldsymbol{\mathcal{K}}_{\boldsymbol{Y}}^{(M)}\,\boldsymbol{f}(u) = \sum_{i=1}^{M} \pi_{\boldsymbol{Y},i}\langle \boldsymbol{f}, \boldsymbol{\psi}_{\boldsymbol{Y},i}\rangle\,\boldsymbol{\psi}_{\boldsymbol{Y},i}(u),$$

it is straightforward to verify that

$$(\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)})^{-1}\boldsymbol{f}(u) = (\boldsymbol{A}')^{-1}(\boldsymbol{\mathcal{K}}_{\boldsymbol{Y}}^{(M)})^{-1}(\boldsymbol{A}^{-1}\boldsymbol{f})(u),$$

provided M is such that $\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)}$ is invertible. Additionally, we can write

$$\mathrm{fMMD}^2(\boldsymbol{X}; k) = \langle (\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)})^{-1}\boldsymbol{X}, \boldsymbol{X}\rangle, \quad \mathrm{fMMD}^2(\boldsymbol{Y}; k) = \langle (\boldsymbol{\mathcal{K}}_{\boldsymbol{Y}}^{(M)})^{-1}\boldsymbol{Y}, \boldsymbol{Y}\rangle.$$

Finally,

$$\begin{aligned}
\mathrm{fMMD}^2(\boldsymbol{Y}, \boldsymbol{\mu}_{\boldsymbol{Y}}; \mathbf{K}_{\boldsymbol{Y}}, M) &= \langle (\boldsymbol{\mathcal{K}}_{\boldsymbol{Y}}^{(M)})^{-1}\boldsymbol{Y}, \boldsymbol{Y}\rangle \\
&= \langle (\boldsymbol{A}')^{-1}(\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)})^{-1}(\boldsymbol{A}^{-1}\boldsymbol{Y}), \boldsymbol{Y}\rangle = \langle (\boldsymbol{A}')^{-1}(\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)})^{-1}\boldsymbol{X}, \boldsymbol{A}\boldsymbol{X}\rangle \\
&= \int_{\mathcal{T}} ((\boldsymbol{A}')^{-1}(\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)})^{-1}\boldsymbol{X}(u))'\boldsymbol{A}\boldsymbol{X}(u)\mathrm{d}u \\
&= \int_{\mathcal{T}} ((\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)})^{-1}\boldsymbol{X})'\boldsymbol{X}(u)\mathrm{d}u \\
&= \langle (\boldsymbol{\mathcal{K}}_{\boldsymbol{X}}^{(M)})^{-1}\boldsymbol{X}, \boldsymbol{X}\rangle = \mathrm{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}_{\boldsymbol{X}}; \mathbf{K}_{\boldsymbol{X}}, M).
\end{aligned}$$

□

Proof of Lemma 4.3.1.2. Let $(\lambda_i^{(j)}, \xi_i^j)$, be the ith eigenpair of the covariance $\mathcal{K}_j$ of the j-component of $\boldsymbol{X}$, $i \geq 1$, $j = 1, \ldots, p$. Construct now the following set of multivariate functions: $\xi_i^{(j)}\, \boldsymbol{e}_j$, $i \geq 1$, $j = 1, \ldots, p$, where $\boldsymbol{e}_j$ is the jth vector of the canonical basis of $\mathbb{R}^p$. It is then straightforward to verify that this set is indeed orthonormal. Additionally, as components in $\boldsymbol{X}$ are uncorrelated, $\boldsymbol{\mathcal{K}} = \operatorname{diag}(\mathcal{K}_1, \ldots, \mathcal{K}_p)$. Simple algebra gives further $\boldsymbol{\mathcal{K}}\, \xi_i^{(j)}\, \boldsymbol{e}_j = \lambda_i^{(j)}\, \xi_i^{(j)}\, \boldsymbol{e}_j$, for any $i \geq 1$ and $j = 1 \ldots, p$. Thus, functions in $\{\xi_i^{(j)}\, \boldsymbol{e}_j : i \geq 1,\, j = 1 \ldots, p\}$, are the eigenfunctions of $\boldsymbol{\mathcal{K}}$, while $\lambda_i^{(j)}$, $i \geq 1$, $j = 1 \ldots, p$ are the corresponding eigenvalues. In other words, the spectrum of $\boldsymbol{\mathcal{K}}$ corresponds to the union of the spectra of individual covariance operators $\mathcal{K}_j$, $j = 1, \ldots, p$. Then, for $m_1, \ldots, m_p$ as described in the statement of the result, the M largest eigenpairs of $\boldsymbol{\mathcal{K}}$ are $(\lambda_i^{(j)}, \xi_i^{(j)}\, \boldsymbol{e}_j)$, $j = 1, \ldots, p$, $i = 1, \ldots, m_j$. Observe that since $\pi_{M+1} < \pi_M$, these eigenvalues are chosen from the spectrum of $\boldsymbol{\mathcal{K}}$ in a unique way. The following now holds:

$$\begin{aligned}\mathrm{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) &= \sum_{j=1}^{p} \sum_{j=1}^{m_j} \frac{1}{\lambda_i^{(j)}} \langle \boldsymbol{X} - \boldsymbol{\mu}, \xi_i^{(j)}\, \boldsymbol{e}_j \rangle^2 = \sum_{j=1}^{p} \left(\sum_{j=1}^{m_j} \frac{1}{\lambda_i^{(j)}} \langle X_j - \mu_j, \xi_i^{(j)} \rangle^2 \right) \\ &= \sum_{j=1}^{p} \mathrm{fMD}^2(X_j, \mu_j; \kappa_j, m_j).\end{aligned}$$

□

Proof of Corollary 4.3.2.1. The proof of the statement (i) follows from the following claims: Lemma 4.3.1.2, the fact that the components of the separable covariance processes, which have uncorrelated components, share, up to scale, common covariance kernel, and Lemma 4.3.1.1 by taking $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1/2}$ and observing that the process transformed that way has uncorrelated components. Claim (ii) follows directly from (i). □

Proof of Lemma 4.3.2.1. Let $\boldsymbol{X}$ be the multivariate Gaussian process with covariance $\mathbf{K} = \kappa \boldsymbol{\Sigma}^{\mathrm{row}}$ and mean $\boldsymbol{\mu}$, where for the simplicity of the notation we assume that $\boldsymbol{\mu} = \mathbf{0}$. Let further $\pi_i \geq \pi_2 \geq \cdots \geq \pi_M > 0$, and $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M$ the leading M eigenvalues and eigenfunctions of covariance operator $\boldsymbol{\mathcal{K}}$ associated with covariance function $\mathbf{K}$, respectively, i.e., $\boldsymbol{\mathcal{K}}\, \boldsymbol{\psi}_i = \pi_i\, \boldsymbol{\psi}_i$, for $i = 1, \ldots, M$. Then $\beta_i = \langle \boldsymbol{X}, \boldsymbol{\psi}_i \rangle \sim \mathcal{N}(0, \pi_i)$, $i = 1, \ldots, M$ are uncorrelated, i.e., independent random variables; see, e.g., Wang (2008) for more details. Denoting $\eta_i = \beta_i / \sqrt{\pi_i} \sim \mathcal{N}(0, 1)$ to be i.i.d. random variables from standard normal distribution, we can write

$$\mathrm{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \boldsymbol{\Sigma}^{\mathrm{row}}, \kappa, M) = \sum_{i=1}^{M} \eta_i^2 \sim \chi^2(M),$$

as a sum of M squared independent standard normal random variables. □

Proof of Theorem 4.4.1.1. (i) For $t \in \mathcal{T}$ let $\boldsymbol{X}(t) = \boldsymbol{A}'\, \boldsymbol{\phi}(t)$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$, for $\boldsymbol{A} = (\boldsymbol{a}_1 \ldots \boldsymbol{a}_p)$. First observe that for every $t \in \mathcal{T}$, $\boldsymbol{\mu}(t) = \mathbb{E}(\boldsymbol{X}(t)) = \mathbb{E}(\boldsymbol{A}'\, \boldsymbol{\phi}(t)) = \mathbb{E}(\boldsymbol{A}')\, \boldsymbol{\phi}(t) = \mathbb{E}(\boldsymbol{A})'\, \boldsymbol{\phi}(t) = \boldsymbol{M}_{\boldsymbol{A}}'\, \boldsymbol{\phi}(t)$. For the simplicity of the notation, since all

the quantities involved are centered, we further take $\boldsymbol{M_A} = \mathbf{0}$. Proceed now first by assuming $\boldsymbol{\Sigma}^{\text{row}} = \boldsymbol{I}_p$. Then, for $t \in \mathcal{T}$

$$\kappa(t,t)\delta_{i,j} = \mathbb{E}(X_i(t)X_j(t)) = \boldsymbol{e}_i'\mathbb{E}(\boldsymbol{X}(t)\boldsymbol{X}'(t))\boldsymbol{e}_j = \mathbb{E}(X_i(t)X_j(t)) \tag{C.5}$$

$$= \boldsymbol{e}_i'\mathbb{E}\left(\boldsymbol{A}'\,\boldsymbol{\phi}(t)(\boldsymbol{A}'\,\boldsymbol{\phi}(t))'\right)\boldsymbol{e}_j = \boldsymbol{\phi}'(t)\mathbb{E}\left(\boldsymbol{a}_i\boldsymbol{a}_j'\right)\boldsymbol{\phi}(t), \quad i,j = 1,\ldots,p, \tag{C.6}$$

where Kronecker delta $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. Since (C.5) holds for every $t \in \mathcal{T}$,

$$\mathbb{E}\left(\boldsymbol{a}_i\boldsymbol{a}_j'\right) = 0 \text{ for } i \neq j, \quad \text{and} \quad \mathbb{E}\left(\boldsymbol{a}_i\boldsymbol{a}_i'\right) = \mathbb{E}\left(\boldsymbol{a}_j\boldsymbol{a}_j'\right), \text{ for } i = 1,\ldots,p.$$

Denoting $\boldsymbol{\Sigma}^{\text{col}} := \text{Cov}(\boldsymbol{a}_1)$

$$\text{Cov}(\text{vec}(\boldsymbol{A})) = \text{Cov}\begin{pmatrix}\boldsymbol{a}_1\\ \vdots \\ \boldsymbol{a}_p\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}^{\text{col}} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}^{\text{col}} & \cdots & \mathbf{0}\\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & & \boldsymbol{\Sigma}^{\text{col}}\end{pmatrix} = \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{I}_p,$$

where for every $s,t \in \mathcal{T}$ the matrix $\boldsymbol{\Sigma}^{\text{col}}$ satisfies

$$\kappa(s,t) = \boldsymbol{\phi}'(s)\boldsymbol{\Sigma}^{\text{col}}\,\boldsymbol{\phi}(t). \tag{C.7}$$

To see that $\boldsymbol{\Sigma}^{\text{col}}$ is indeed positive, observe the following: As $\boldsymbol{W} := \int_{\mathcal{T}} \boldsymbol{\phi}(t)\,\boldsymbol{\phi}'(t)\text{d}t$ is a positive definite matrix, then there exists $m_0 \in \mathbb{N}$ and $t_1,\ldots,t_{m_0}$, such that the Riemann sum $1/m_0 \sum_{i=1}^{m_0} \boldsymbol{\phi}(t_i)\,\boldsymbol{\phi}'(t_i)$ approximating the integral in $\boldsymbol{W}$ is the positive definite matrix. Therefore, there exist m linearly independent vectors in $\{\boldsymbol{\phi}(t_i) : i = 1,\ldots,m_0\}$. Without loss of generality assume that these are $\boldsymbol{\phi}(t_1),\ldots,\boldsymbol{\phi}(t_m)$. Take now $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{y} \neq 0$. Due to the independence of $\boldsymbol{\phi}(t_1),\ldots,\boldsymbol{\phi}(t_m)$, $\boldsymbol{y}$ can represented as $\boldsymbol{y} = \sum_{i=1}^m c_i\,\boldsymbol{\phi}(t_i)$, for some $(c_1,\ldots,c_m) \neq \mathbf{0}$. Then

$$\boldsymbol{y}'\boldsymbol{\Sigma}^{\text{col}}\boldsymbol{y} = \sum_{i,j=1}^m c_i c_j\,\boldsymbol{\phi}'(t_i)\boldsymbol{\Sigma}^{\text{col}}\,\boldsymbol{\phi}(t_j) = \sum_{i,j=1}^m c_i c_j \kappa(t_i,t_j) > 0,$$

where the last inequality holds due to the positive definiteness of κ.

Let now $\boldsymbol{X} = \boldsymbol{A}'\,\boldsymbol{\phi} \sim \mathcal{MSP}(\mathbf{0},\boldsymbol{\Sigma}^{\text{row}},\kappa)$. Lemma 4.3.1.1 then implies that $\boldsymbol{Y} := \boldsymbol{\Sigma}^{\text{row}\,-1/2}\boldsymbol{X} \sim \mathcal{MSP}(\mathbf{0},\boldsymbol{I}_p,\kappa)$. Additionally, for $\boldsymbol{A_Y} = \boldsymbol{A}\boldsymbol{\Sigma}^{\text{row}\,-1/2}$ is $\boldsymbol{Y} = \boldsymbol{A}_{\boldsymbol{Y}}'\,\boldsymbol{\phi}$. The first part of the proof now shows that

$$\text{Cov}(\text{vec}(\boldsymbol{A_Y})) = \boldsymbol{I}_p \otimes \boldsymbol{\Sigma}^{\text{col}},$$

where $\boldsymbol{\Sigma}^{\text{col}}$ depends only on the kernel κ as given in (C.7). Note further that

$$\text{vec}(\boldsymbol{A}) = \text{vec}(\boldsymbol{I}_m\boldsymbol{A_Y}(\boldsymbol{\Sigma}^{\text{row}})^{1/2}) = \left((\boldsymbol{\Sigma}^{\text{row}})^{1/2} \otimes \boldsymbol{I}_m\right)\text{vec}(\boldsymbol{A_Y}),$$

finally giving

$$\text{Cov}(\text{vec}(\boldsymbol{A})) = \left((\boldsymbol{\Sigma}^{\text{row}})^{1/2} \otimes \boldsymbol{I}_m\right)\left(\boldsymbol{I}_p \otimes \boldsymbol{\Sigma}^{\text{col}}\right)\left((\boldsymbol{\Sigma}^{\text{row}})^{1/2} \otimes \boldsymbol{I}_m\right) = \boldsymbol{\Sigma}^{\text{row}} \otimes \boldsymbol{\Sigma}^{\text{col}}.$$

(ii) To prove the statement (ii) we begin by assuming $\boldsymbol{\Sigma}^{\text{row}} = \boldsymbol{I}_p$. Additionally, without loss of generality and for the simplicity of the notation take $\boldsymbol{\mu} = \mathbf{0}$. Let further $(\lambda_1, \xi_1), \ldots, (\lambda_m, \xi_m)$, $\lambda_1 \geq \cdots \lambda_m > 0$ be the eigenpairs of κ, e.i. for every $s \in \mathcal{T}$

$$\int_{\mathcal{T}} \kappa(s,t)\, \xi_i(t) \mathrm{d}t = \lambda_i\, \xi_i(s). \tag{C.8}$$

Expressing the eigenfunctions ξ_i, $i = 1, \ldots, m$ in $\boldsymbol{\phi}$ basis gives

$$\xi_i = \boldsymbol{b}_i' \boldsymbol{\phi}, \quad i = 1, \ldots, m. \tag{C.9}$$

For any $k \in \{1, \ldots, p\}$, $i = 1, \ldots, m$, and $s \in \mathcal{T}$ (C.8) and (C.9) imply the following relations are equivalent:

$$\begin{aligned}
\int_{\mathcal{T}} \kappa(s,t)\, \xi_i(t) \mathrm{d}t = \lambda_i\, \xi_i(s) \iff & \int_{\mathcal{T}} \mathbb{E}(X_k(s) X_k(t))\, \xi_i(t) \mathrm{d}t = \lambda_i\, \xi_i(s), \\
\iff & \boldsymbol{\phi}'(s) \mathbb{E}(\boldsymbol{a}_k \boldsymbol{a}_k') \left(\int_{\mathcal{T}} \boldsymbol{\phi}(t)\, \boldsymbol{\phi}'(t) \mathrm{d}t \right) \boldsymbol{b}_i = \lambda_i\, \boldsymbol{\phi}'(s) \boldsymbol{b}_i, \\
\iff & \boldsymbol{\phi}'(s) \mathbb{E}(\boldsymbol{a}_k \boldsymbol{a}_k') \boldsymbol{W} \boldsymbol{b}_i = \lambda_i\, \boldsymbol{\phi}'(s) \boldsymbol{b}_i,
\end{aligned} \tag{C.10}$$

where $\boldsymbol{W} := \int_{\mathcal{T}} \boldsymbol{\phi}(t)\, \boldsymbol{\phi}'(t) \mathrm{d}t$ is a positive definite matrix; see proof of part (i) for more details. Since (C.10) holds for every $s \in \mathcal{T}$, and since $\mathbb{E}(\mathbf{a}_k \mathbf{a}_k') = \boldsymbol{\Sigma}^{\text{col}}$; see proof of (i) for details, we obtain

$$\boldsymbol{\Sigma}^{\text{col}} \boldsymbol{W} \boldsymbol{b}_i = \lambda_i \boldsymbol{b}_i \iff (\boldsymbol{\Sigma}^{\text{col}})^{-1} = \boldsymbol{W}^{1/2} \boldsymbol{\Sigma}^{\text{col}} \boldsymbol{W}^{1/2} \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \tag{C.11}$$

where $\boldsymbol{u}_i = \boldsymbol{W}^{1/2} \boldsymbol{b}_i$, $i = 1, \ldots, m$. Equation (C.11) also implies that $(\lambda_i, \boldsymbol{u}_i)$, $i = 1, \ldots, m$ are eigenpairs of symmetric matrix $\boldsymbol{W}^{1/2} \boldsymbol{\Sigma}^{\text{col}} \boldsymbol{W}^{1/2}$. Finally, (C.11) further implies

$$\boldsymbol{W}^{1/2} \boldsymbol{\Sigma}^{\text{col}} \boldsymbol{W}^{1/2} = \sum_{i=1}^{m} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i' = \boldsymbol{W}^{1/2} \left(\sum_{i=1}^{m} \lambda_i \boldsymbol{b}_i \boldsymbol{b}_i' \right) \boldsymbol{W}^{1/2} \iff \boldsymbol{\Sigma}^{\text{col}} = \sum_{i=1}^{m} \lambda_i \boldsymbol{b}_i \boldsymbol{b}_i'.$$

Moreover,

$$\begin{aligned}
\boldsymbol{W}^{-1/2} (\boldsymbol{\Sigma}^{\text{col}})^{-1} \boldsymbol{W}^{-1/2} &= (\boldsymbol{W}^{1/2} \boldsymbol{\Sigma}^{\text{col}} \boldsymbol{W}^{1/2})^{-1} \\
&= \sum_{i=1}^{m} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i' = \boldsymbol{W}^{1/2} \left(\sum_{i=1}^{m} \frac{1}{\lambda_i} \boldsymbol{b}_i \boldsymbol{b}_i' \right) \boldsymbol{W}^{1/2} \\
\iff (\boldsymbol{\Sigma}^{\text{col}})^{-1} &= \boldsymbol{W} \left(\sum_{i=1}^{m} \frac{1}{\lambda_i} \boldsymbol{b}_i \boldsymbol{b}_i' \right) \boldsymbol{W}.
\end{aligned} \tag{C.12}$$

Corollary 4.3.2.1, together with relation $\langle X_j, \xi_i \rangle = \langle \boldsymbol{a}_j' \boldsymbol{\phi}, \boldsymbol{b}_i' \boldsymbol{\phi} \rangle = \boldsymbol{a}_j' \boldsymbol{W} \boldsymbol{b}_i$ now implies

$$\begin{aligned}
\text{fMMD}^2(\boldsymbol{X}; mp) &= \sum_{i=1}^{m} \sum_{j=1}^{p} \lambda_i^{-1} \langle X_j, \xi_i \rangle^2 = \sum_{i=1}^{m} \sum_{j=1}^{p} \lambda_i^{-1} (\boldsymbol{a}_j' \boldsymbol{W} \boldsymbol{b}_i)^2 \\
&= \sum_{j=1}^{p} \boldsymbol{a}_j' \boldsymbol{W} \sum_{i=1}^{m} \left(\lambda_i^{-1} \boldsymbol{b}_i \boldsymbol{b}_i' \right) \boldsymbol{W} \boldsymbol{a}_j = \sum_{j=1}^{p} \boldsymbol{a}_j' (\boldsymbol{\Sigma}^{\text{col}})^{-1} \boldsymbol{a}_j \\
&= \text{tr}(\boldsymbol{A}' (\boldsymbol{\Sigma}^{\text{col}})^{-1} \boldsymbol{A}) = \text{MMD}^2(\boldsymbol{A}).
\end{aligned}$$

Affine invariance of fMMD; Lemma 4.3.1.1 and MMD (Mayrhofer et al., 2024a, Lemma 3.0.1) complete the proof of (ii).

(iii) As in (i), for $t \in \mathcal{T}$, $\boldsymbol{\mu}(t) = \mathbb{E}(\boldsymbol{X}(t)) = \mathbb{E}(\boldsymbol{A}'\,\boldsymbol{\phi}(t)) = \mathbb{E}(\boldsymbol{A}')\,\boldsymbol{\phi}(t) = \mathbb{E}(\boldsymbol{A})'\,\boldsymbol{\phi}(t) = \boldsymbol{M}_{\boldsymbol{A}}'\,\boldsymbol{\phi}(t)$. For the simplicity of the notation, we again take $\boldsymbol{M}_{\boldsymbol{A}} = \boldsymbol{0}$.

Proceed first by assuming first that $\boldsymbol{\Sigma}^{\text{row}} = \boldsymbol{I}_p$. Take further $t_1, \ldots, t_m$, such that $\boldsymbol{\phi}_{t_1,\ldots,t_m} := (\boldsymbol{\phi}(t_1) \ldots \boldsymbol{\phi}(t_m))$ is a full column rank matrix; for the proof of the existence see proof of (i). Gaussianity of $\boldsymbol{X}$ then implies that

$$(\boldsymbol{X}(t_1), \ldots, \boldsymbol{X}(t_{\text{m}})) = \boldsymbol{A}'(\boldsymbol{\phi}(t_1), \ldots, \boldsymbol{\phi}(t_{\text{m}})) \sim \mathcal{MN}(\boldsymbol{0}, \boldsymbol{I}_p, \boldsymbol{\Sigma}^{\text{col}}_{t_1,\ldots,t_{\text{m}}}),$$

where $\boldsymbol{\Sigma}^{\text{col}}_{t_1,\ldots,t_{\text{m}}} = [\kappa(t_i, t_j)]_{i,j}$ is a positive definite matrix, due to the positive definiteness of kernel κ. Therefore,

$$\text{vec}(\boldsymbol{I}_p\boldsymbol{A}'\,\boldsymbol{\phi}_{t_1,\ldots,t_m}) = \left(\boldsymbol{\phi}_{t_1,\ldots,t_m} \otimes \boldsymbol{I}_p\right) \text{vec}(\boldsymbol{A}') \sim \mathcal{N}_{mp}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\text{col}}_{t_1,\ldots,t_{\text{m}}} \otimes \boldsymbol{I}_p),$$

i.e.,

$$\text{vec}(\boldsymbol{\phi}_{t_1,\ldots,t_m}\,\boldsymbol{A}\boldsymbol{I}_p) = \left(\boldsymbol{I}_p \otimes \boldsymbol{\phi}'_{t_1,\ldots,t_m}\right) \text{vec}(\boldsymbol{A}) \sim \mathcal{N}_{mp}(\boldsymbol{0}, \boldsymbol{I}_p \otimes \boldsymbol{\Sigma}^{\text{col}}_{t_1,\ldots,t_{\text{m}}}).$$

Regularity of $\boldsymbol{\phi}_{t_1,\ldots,t_m}$ and the fact that inversion of Kronecker product of two matrices, as well as the product of two Kronecker products, retains the Kronecker structure completes the first part of the proof, where the particular form of the covariance $\boldsymbol{\Sigma}^{\text{col}}$ is given by (i). A short note to the reader: Given the general result in (i), it was enough to show that $\boldsymbol{A}$ has normally distributed entries.

Finally, for $\boldsymbol{X} \sim \mathcal{MSP}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\text{row}}, \kappa)$, let $\boldsymbol{Y} := (\boldsymbol{\Sigma}^{\text{row}})^{-1/2}\boldsymbol{X} = \left(\boldsymbol{A}(\boldsymbol{\Sigma}^{\text{row}})^{-1/2}\right)'\boldsymbol{\phi} \sim \mathcal{MSP}(\boldsymbol{0}, \boldsymbol{I}_p, \kappa)$. The first part of the proof shows that

$$\boldsymbol{A}(\boldsymbol{\Sigma}^{\text{row}})^{-1/2} \sim \mathcal{MN}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\text{col}}, \boldsymbol{I}_p).$$

Matrix affine equivariance of matrix normal distribution (Gupta and Nagar, 1999) finally gives that

$$\boldsymbol{A} \sim \mathcal{MN}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\text{col}}, \boldsymbol{\Sigma}^{\text{row}}),$$

thus completing the proof.

□

C.3 PCA Algorithm

Algorithm 6 Robust FPCA for separable processes

Input: $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, $\boldsymbol{X}_i \in \mathbb{R}^{p \times q}$, $i = 1, \ldots, n$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$, $\mathcal{T}$

1: **Create functional data object**
   Estimate coefficient matrices $\mathfrak{A} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)$ by smoothing $\mathfrak{X}$;
   Obtain finite basis representation $\boldsymbol{X}_i(t) = \boldsymbol{A}_i' \, \boldsymbol{\phi}(t)$, $i = 1, \ldots, n$, $t \in \mathcal{T}$;
2: **MMCD** (Algorithm 2 in Mayrhofer et al. (2024a))
   Run MMCD procedure on $\mathfrak{A}$ and get $(\hat{\mathbf{M}}_{\boldsymbol{A},H^*}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H^*}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H^*}, \text{MMD}(\mathfrak{A}))$;
3: **Compute coefficient representations of FPCs**
   Compute matrix of inner products of basis functions $\boldsymbol{W} = \int_{\mathcal{T}} \boldsymbol{\phi}(t) \, \boldsymbol{\phi}'(t) \mathrm{d}t$;
   Eigendecomposition of $\boldsymbol{W}^{1/2} \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H^*} \boldsymbol{W}^{1/2} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}'$;
      Matrix of eigenvalues $\boldsymbol{\Lambda} = \text{diag}(\lambda_1^{\text{ker}}, \ldots, \lambda_m^{\text{ker}})$;
      Matrix of eigenvectors $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$;
   Compute coefficients $\boldsymbol{b}_j = \boldsymbol{W}^{-1/2} \boldsymbol{u}_j$, $j = 1, \ldots, m$, of FPCs;
4: **Obtain univariate FPCs**
   Univariate eigenpairs $(\xi_j, \lambda_j^{\text{ker}})$ with $\xi_j(t) = \boldsymbol{b}_j' \, \boldsymbol{\phi}(t)$, $t \in \mathcal{T}$, $j = 1, \ldots, m$;
5: **Obtain multivariate FPCs**
   Eigendecomposition of $\hat{\boldsymbol{\Sigma}}^{\text{row}} = \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H^*} = \boldsymbol{V} \boldsymbol{\Gamma} \boldsymbol{V}'$;
      Matrix of eigenvalues $\boldsymbol{\Gamma} = \text{diag}(\lambda_1^{\text{row}}, \ldots, \lambda_p^{\text{row}})$;
      Matrix of eigenvectors $\boldsymbol{V} = (\boldsymbol{v}_1^{\text{row}}, \ldots, \boldsymbol{v}_1^{\text{row}})$;
   Define indexation $k = k(j,l) = 1, \ldots, M = pm$ for $j = 1, \ldots, m$, $l = 1, \ldots, p$;
   Multivariate eigenpairs $(\pi_k, \boldsymbol{\psi}_k) = (\lambda_j^{\text{ker}} \lambda_k^{\text{row}}, \boldsymbol{v}_l^{\text{row}} \, \xi_j)$;

Output: $(\pi_1, \ldots, \pi_M), (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M)$

C.4 Shapley Proofs

Proof of Proposition 4.5.2.1. The outlyingness contribution $\theta_{\mathcal{T}_a}(X;m) = \theta_{\mathcal{T}_a}(X,\mu;\kappa,m)$ of the kth coordinate to $\mathrm{fMD}^2(X,m) = \mathrm{fMD}^2(X,\mu;\kappa,m)$ is given as the weighted average of the marginal outlyingness contributions $\Delta_{\mathcal{T}_a}\mathrm{fMD}^2(\hat{X}^R;m) = \Delta_{\mathcal{T}_a}\mathrm{fMD}^2(\hat{X}^R,\mu;\kappa,m)$. Then the $\Delta_{\mathcal{T}_a}\mathrm{fMD}^2(\hat{X}^R;m)$ can be simplified as follows:

$$\begin{aligned}
\Delta_{\mathcal{T}_a}\mathrm{fMD}^2(\hat{X}^R;m) &= \mathrm{fMD}^2(\hat{X}^{R\cup\{a\}};m) - \mathrm{fMD}^2(\hat{X}^R;m) \\
&= \sum_{i=1}^{m}\frac{1}{\lambda_i}\left(\left(\sum_{b=1}^{d}\langle\hat{X}^{R\cup\{a\}},\xi_i\rangle_{\mathcal{T}_b}\right)^2 - \left(\sum_{b=1}^{d}\langle\hat{X}^{R},\xi_i\rangle_{\mathcal{T}_b}\right)^2\right) \\
&= \sum_{i=1}^{m}\frac{1}{\lambda_i}\left(\sum_{b\in R\cup\{a\}}\sum_{c\in R\cup\{a\}}\langle X,\xi_i\rangle_{\mathcal{T}_b}\langle X,\xi_i\rangle_{\mathcal{T}_c}\right. \\
&\qquad\left. - \sum_{b\in R}\sum_{c\in R}\langle X,\xi_i\rangle_{\mathcal{T}_b}\langle X,\xi_i\rangle_{\mathcal{T}_c}\right) \\
&= \sum_{i=1}^{m}\frac{1}{\lambda_i}\left(\sum_{b\in R}\sum_{c\in R}\langle X,\xi_i\rangle_{\mathcal{T}_b}\langle X,\xi_i\rangle_{\mathcal{T}_c} + 2\langle X,\xi_i\rangle_{\mathcal{T}_a}\sum_{b\in R}\langle X,\xi_i\rangle_{\mathcal{T}_b}\right. \\
&\qquad\left. + \langle X,\xi_i\rangle_{\mathcal{T}_a}^2 - \sum_{b\in R}\sum_{c\in R}\langle X,\xi_i\rangle_{\mathcal{T}_b}\langle X,\xi_i\rangle_{\mathcal{T}_c}\right) \\
&= \sum_{i=1}^{m}\frac{1}{\lambda_i}\left(2\langle X,\xi_i\rangle_{\mathcal{T}_a}\sum_{b\in R}\langle X,\xi_i\rangle_{\mathcal{T}_b} + \langle X,\xi_i\rangle_{\mathcal{T}_a}^2\right).
\end{aligned}$$

Further, we write

$$w(|R|) := \frac{r!(d-r-1)!}{d!},$$

with $r = |R|$ for which $\sum_{R\subseteq D\setminus\{a\}} w(|S|) = 1$ holds. With this, the time-specific outlyingness contribution within the subinterval $\mathcal{T}_a$ based on the Shapley value simplifies to

$$\begin{aligned}
\theta_{\mathcal{T}_a}(X,\mu;\kappa,m) &= \sum_{R\subseteq D\setminus\{a\}}\frac{|R|!(d-|R|-1)!}{(d)!}\Delta_{\mathcal{T}_a}\mathrm{fMD}^2(\hat{X}^R;m) \\
&= \sum_{R\subseteq D\setminus\{a\}} w(|R|)\left(\sum_{i=1}^{m}\frac{1}{\lambda_i}\left(2\langle X,\xi_i\rangle_{\mathcal{T}_a}\sum_{b\in R}\langle X,\xi_i\rangle_{\mathcal{T}_b} + \langle X,\xi_i\rangle_{\mathcal{T}_a}^2\right)\right) \\
&= \sum_{i=1}^{m}\frac{1}{\lambda_i}\langle X,\xi_i\rangle_{\mathcal{T}_a}^2 + 2\sum_{i=1}^{m}\frac{1}{\lambda_i}\langle X,\xi_i\rangle_{\mathcal{T}_a}\sum_{R\subseteq D\setminus\{a\}} w(|R|)\sum_{b\in R}\langle X,\xi_i\rangle_{\mathcal{T}_b} \\
&= \sum_{i=1}^{m}\frac{1}{\lambda_i}\langle X,\xi_i\rangle_{\mathcal{T}_a}^2 + 2\sum_{i=1}^{m}\frac{1}{\lambda_i}\langle X,\xi_i\rangle_{\mathcal{T}_a}\sum_{r=1}^{d-1} w(r)\sum_{\substack{R\subseteq D\setminus\{a\}\\|R|=r}}\sum_{b\in R}\langle X,\xi_i\rangle_{\mathcal{T}_b}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X, \xi_i \rangle_{\mathcal{T}_a}^2 + 2 \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X, \xi_i \rangle_{\mathcal{T}_a} \sum_{r=1}^{d-1} w(r) \binom{d-2}{r-1} \sum_{b \in D \setminus \{a\}} \langle X, \xi_i \rangle_{\mathcal{T}_b} \\
&= \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X, \xi_i \rangle_{\mathcal{T}_a}^2 + 2 \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X, \xi_i \rangle_{\mathcal{T}_a} \sum_{r=1}^{d-1} \frac{r}{d(d-1)} \sum_{b \in D \setminus \{a\}} \langle X, \xi_i \rangle_{\mathcal{T}_b} \\
&= \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X, \xi_i \rangle_{\mathcal{T}_a}^2 + \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X, \xi_i \rangle_{\mathcal{T}_a} \sum_{b \in D \setminus \{a\}} \langle X, \xi_i \rangle_{\mathcal{T}_b} \\
&= \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X, \xi_i \rangle_{\mathcal{T}_a} \sum_{b \in D} \langle X, \xi_i \rangle_{\mathcal{T}_b} \\
&= \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X, \xi_i \rangle_{\mathcal{T}_a} \langle X, \xi_i \rangle.
\end{aligned}
$$

□

Proof of Lemma 4.5.2.1. We have that $X(t) = \boldsymbol{a}' \boldsymbol{\phi}(t)$, $\mu(t) = \boldsymbol{m}_{\boldsymbol{a}}' \boldsymbol{\phi}(t)$, and $\kappa(s,t) = \boldsymbol{\phi}'(s) \boldsymbol{\Sigma} \boldsymbol{\phi}(t)$, for $s, t \in \mathcal{T}$. Based on Proposition 4.5.2.1 we obtain

$$
\begin{aligned}
\phi_{\mathcal{T}_a}(X, \mu, \kappa; m) &= \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle X - \mu, \xi_i \rangle_{\mathcal{T}_a} \langle X - \mu, \xi_i \rangle \\
&= \sum_{i=1}^{m} \frac{1}{\lambda_i} \langle (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}})' \boldsymbol{\phi}, \boldsymbol{b}_i \rangle_{\mathcal{T}_a} \langle (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}})' \boldsymbol{\phi}, \xi_i \rangle \\
&= \sum_{i=1}^{m} \frac{1}{\lambda_i} (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}})' \langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle_{\mathcal{T}_a} \boldsymbol{b}_i (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}}') \langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle \boldsymbol{b}_i \\
&= (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}})' \boldsymbol{W}_{\mathcal{T}_a} \underbrace{\left(\sum_{i=1}^{m} \frac{1}{\lambda_i} \boldsymbol{b}_i \boldsymbol{b}_i' \right) \boldsymbol{W}}_{= \boldsymbol{W}^{-1} (\boldsymbol{\Sigma}^{\mathrm{col}})^{-1}, \text{ see Eq. (C.12)}} (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}}) \\
&= (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}})' \boldsymbol{W}_{\mathcal{T}_a} \boldsymbol{W}^{-1} \boldsymbol{\Sigma}^{-1} (\boldsymbol{a} - \boldsymbol{m}_{\boldsymbol{a}}),
\end{aligned}
$$

where $\boldsymbol{W} = \langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle_{\mathcal{T}} := \int_{\mathcal{T}} \boldsymbol{\phi}(t) \, \boldsymbol{\phi}'(t) \mathrm{d}t$ and $\boldsymbol{W}_{\mathcal{T}_a} = \langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle_{\mathcal{T}_a} := \int_{\mathcal{T}_a} \boldsymbol{\phi}(t) \, \boldsymbol{\phi}'(t) \mathrm{d}t$. □

Proof of Proposition 4.5.2.2. The outlyingness contribution $\theta_k(\boldsymbol{X}; M) = \theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ of the kth coordinate to $\mathrm{fMMD}^2(\boldsymbol{X}; M) = \mathrm{fMMD}^2(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M)$ is the weighted average of the marginal outlyingness contributions $\Delta_k \, \mathrm{fMMD}^2(\hat{\boldsymbol{X}}^S; M) = \Delta_k \, \mathrm{fMMD}^2(\hat{\boldsymbol{X}}^S, \boldsymbol{\mu}; \mathbf{K}, M)$. Without loss of generality, we assume that the data are centered to simplify the notation, i.e., $\boldsymbol{\mu}(t) = \mathbf{0} = (0, \ldots, 0)' \in \mathbb{R}^p, t \in \mathcal{T}$, simply denoted as $\boldsymbol{\mu} = \mathbf{0}$. Then the marginal

outlyingness contributions can be written as

$$
\begin{aligned}
\Delta_k \operatorname{fMMD}^2(\hat{\boldsymbol{X}}^S; M) &= \operatorname{fMMD}^2(\hat{\boldsymbol{X}}^{S\cup\{k\}}; M) - \operatorname{fMMD}^2(\hat{\boldsymbol{X}}^S; M) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle \hat{\boldsymbol{X}}^{S\cup\{k\}}, \boldsymbol{\psi}_i \rangle^2 - \sum_{i=i}^{M} \frac{1}{\pi_i} \langle \hat{\boldsymbol{X}}^{S}, \boldsymbol{\psi}_i \rangle^2 \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\left(\sum_{j=1}^{p} \langle \hat{X}_j^{S\cup\{k\}}, \psi_{i,j} \rangle \right)^2 - \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\sum_{j=1}^{p} \langle \hat{X}_j^{S}, \psi_{i,j} \rangle \right)^2 \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\left(\sum_{j=1}^{p} \langle \hat{X}_j^{S\cup\{k\}}, \psi_{i,j} \rangle \right)^2 - \left(\sum_{j=1}^{p} \langle \hat{X}_j^{S}, \psi_{i,j} \rangle \right)^2 \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\left(\sum_{j=1}^{p} \langle \hat{X}_j^{S\cup\{k\}}, \psi_{i,j} \rangle \right)^2 - \left(\sum_{j=1}^{p} \langle \hat{X}_j^{S}, \psi_{i,j} \rangle \right)^2 \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\sum_{j\in S\cup k} \sum_{l\in S\cup k} \langle X_l, \psi_{i,l} \rangle \langle X_j, \psi_{i,j} \rangle - \sum_{j\in S} \sum_{l\in S} \langle X_l, \psi_{i,l} \rangle \langle X_j, \psi_{i,j} \rangle \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\sum_{j\in S\cup k} \sum_{l\in S} \langle X_l, \psi_{i,l} \rangle \langle X_j, \psi_{i,j} \rangle + \langle X_k, \psi_{i,k} \rangle \sum_{j\in S\cup k} \langle X_j, \psi_{i,j} \rangle \right. \\
&\qquad \left. - \sum_{j\in S} \sum_{l\in S} \langle X_l, \psi_{i,l} \rangle \langle X_j, \psi_{i,j} \rangle \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\sum_{j\in S} \sum_{l\in S} \langle X_l, \psi_{i,l} \rangle \langle X_j, \psi_{i,j} \rangle + \langle X_k, \psi_{i,k} \rangle \sum_{j\in S\cup k} \langle X_j, \psi_{i,j} \rangle \right. \\
&\qquad \left. + \langle X_k, \psi_{i,k} \rangle \sum_{l\in S} \langle X_l, \psi_{i,l} \rangle - \sum_{j\in S} \sum_{l\in S} \langle X_l, \psi_{i,l} \rangle \langle X_j, \psi_{i,j} \rangle \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\langle X_k, \psi_{i,k} \rangle^2 + 2 \langle X_k, \psi_{i,k} \rangle \sum_{j\in S} \langle X_j, \psi_{i,j} \rangle \right)
\end{aligned}
$$

Further, we write

$$
w(|S|) := \frac{|S|!(p - |S| - 1)!}{p!},
$$

for which $\sum_{S\subseteq P\setminus\{k\}} w(|S|) = 1$ holds. Then the contribution of the kth coordinate to the squared truncated functional Mahalanobis distance $\operatorname{fMMD}^2(\boldsymbol{X}; M)$ based on the Shapley

value is given by

$$
\begin{aligned}
\theta_k(\boldsymbol{X}; M) &= \sum_{S \subseteq P \setminus \{k\}} w(|S|) \Delta_k \, \mathrm{fMMD}^2(\hat{\boldsymbol{X}}^S; M) \\
&= \sum_{S \subseteq P \setminus \{k\}} w(|S|) \left(\sum_{i=1}^{M} \frac{1}{\pi_i} \left(\langle X_k, \psi_{i,k} \rangle^2 + 2 \langle X_k, \psi_{i,k} \rangle \sum_{j \in S} \langle X_j, \psi_{i,j} \rangle \right) \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle^2 + 2 \sum_{S \subseteq P \setminus \{k\}} w(|S|) \left(\sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle \sum_{j \in S} \langle X_j, \psi_{i,j} \rangle \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle^2 + 2 \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle \sum_{s=1}^{p-1} w(s) \sum_{\substack{S \subseteq P \setminus \{k\} \\ |S| = s}} \sum_{j \in S} \langle X_j, \psi_{i,j} \rangle \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle^2 + 2 \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle \sum_{s=1}^{p-1} w(s) \binom{p-2}{s-1} \sum_{j \in P \setminus \{k\}} \langle X_j, \psi_{i,j} \rangle \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle^2 + 2 \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle \sum_{s=1}^{p-1} \frac{s}{p(p-1)} \sum_{j \in P \setminus \{k\}} \langle X_j, \psi_{i,j} \rangle \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle^2 + \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k, \psi_{i,k} \rangle \sum_{j \in P \setminus \{k\}} \langle X_j, \psi_{i,j} \rangle \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\langle X_k, \psi_{i,k} \rangle^2 + \langle X_k, \psi_{i,k} \rangle \sum_{j \in P \setminus \{k\}} \langle X_j, \psi_{i,j} \rangle \right) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\langle X_k, \psi_{i,k} \rangle \sum_{j=1}^{p} \langle X_j, \psi_{i,j} \rangle \right) = \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\langle X_k, \psi_{i,k} \rangle \langle \boldsymbol{X}, \boldsymbol{\psi}_i \rangle \right)
\end{aligned}
$$

□

Proof of Corollary 4.5.2.1. From Equation (C.4) it follows that

$$\psi_{kl}(t) = \boldsymbol{\psi}_k'(t)\boldsymbol{e}_l = \xi_i(t)\boldsymbol{v}_j^{\text{row}}\boldsymbol{e}_l = \xi_i(t)v_{j,l}^{\text{row}},$$

for some i, j, k, hence

$$\begin{aligned}
\theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) &= \sum_{i=1}^{M} \frac{1}{\pi_i} \left(\langle X_k, \psi_{i,k} \rangle \sum_{l=1}^{p} \langle X_l, \psi_{i,l} \rangle \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}} \lambda_j^{\text{row}}} \left(\langle X_k, \xi_i \, v_{j,k}^{\text{row}} \rangle \sum_{l=1}^{p} \langle X_l, \xi_i \, v_{j,l}^{\text{row}} \rangle \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}} \lambda_j^{\text{row}}} \left(\langle X_k, \xi_i \rangle v_{j,k}^{\text{row}} \sum_{l=1}^{p} \langle X_l, \xi_i \rangle v_{j,l}^{\text{row}} \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}} \lambda_j^{\text{row}}} \left(\alpha_{ki} v_{j,k}^{\text{row}} \sum_{l=1}^{p} \alpha_{li} v_{j,l}^{\text{row}} \right)
\end{aligned}$$

□

Proof of Lemma 4.5.2.2. We have that $\boldsymbol{X}(t) = \boldsymbol{A}' \boldsymbol{\phi}(t)$, $\boldsymbol{\mu}(t) = \boldsymbol{M}_{\boldsymbol{A}}' \boldsymbol{\phi}(t)$, and $\kappa(s,t) = \boldsymbol{\phi}'(s)\boldsymbol{\Sigma}^{\text{col}} \boldsymbol{\phi}(t)$, for $s, t \in \mathcal{T}$, see Theorem 4.4.1.1. Combined with Corollary 4.5.2.1 we obtain

$$\begin{aligned}
\theta_k(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) &= \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}} \lambda_j^{\text{row}}} \left(\langle X_k, \xi_i \rangle v_{j,k}^{\text{row}} \sum_{l=1}^{p} \langle X_l, \xi_i \rangle v_{j,l}^{\text{row}} \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}} \lambda_j^{\text{row}}} \left(\boldsymbol{a}_k' \langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle \boldsymbol{b}_i v_{j,k}^{\text{row}} \sum_{l=1}^{p} \boldsymbol{a}_l' \langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle \boldsymbol{b}_i v_{j,l}^{\text{row}} \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}} \lambda_j^{\text{row}}} \left(\boldsymbol{a}_k' \boldsymbol{W} \boldsymbol{b}_i v_{j,k}^{\text{row}} \sum_{l=1}^{p} \boldsymbol{a}_l' \boldsymbol{W} \boldsymbol{b}_i v_{j,l}^{\text{row}} \right) \\
&= \sum_{j=1}^{p} \frac{1}{\lambda_j^{\text{row}}} v_{j,k}^{\text{row}} \sum_{l=1}^{p} \boldsymbol{a}_k' \underbrace{\boldsymbol{W} \left(\sum_{i=1}^{m} \frac{1}{\lambda_i^{\text{ker}}} \boldsymbol{b}_i \boldsymbol{b}_i' \right) \boldsymbol{W}}_{=(\boldsymbol{\Sigma}^{\text{col}})^{-1}, \text{ see Eq. (C.12)}} \boldsymbol{a}_l v_{j,l}^{\text{row}} \\
&= \sum_{j=1}^{p} \frac{1}{\lambda_j^{\text{row}}} v_{j,k}^{\text{row}} \sum_{l=1}^{p} \boldsymbol{a}_k' (\boldsymbol{\Sigma}^{\text{col}})^{-1} \boldsymbol{a}_l v_{j,l}^{\text{row}} \\
&= \sum_{j=1}^{p} \frac{1}{\lambda_j^{\text{row}}} v_{j,k}^{\text{row}} \boldsymbol{a}_k' (\boldsymbol{\Sigma}^{\text{col}})^{-1} \boldsymbol{A}' \boldsymbol{v}_j^{\text{row}},
\end{aligned}$$

where $\boldsymbol{W} = \langle \boldsymbol{\phi}, \boldsymbol{\phi}' \rangle := \int_{\mathcal{T}} \boldsymbol{\phi}(t) \, \boldsymbol{\phi}'(t) \mathrm{d}t$. □

Proof of Proposition 4.5.2.3. Let Y denote the univariate process concatenating each coordinate function X_j for $j = 1, \ldots, p$ of $\boldsymbol{X}$ which is then defined on the interval $\tilde{\mathcal{T}} = \underbrace{\mathcal{T} \cup \cdots \cup \mathcal{T}}_{p \text{ times}}$.
With this concatenating approach, the mean function for $\tilde{t} \in [\mathcal{T} \cdot (j-1), \mathcal{T} \cdot j] \subset \tilde{\mathcal{T}}, j = 1, \ldots, p$, is given as

$$\mu(\tilde{t}) = \mu_j(t),$$

with $t \in \mathcal{T}$, and the covariance function for $\tilde{s} \in [\mathcal{T} \cdot (j-1), \mathcal{T} \cdot j] \subset \tilde{\mathcal{T}}, j = 1, \ldots, p$, and $\tilde{t} \in [\mathcal{T} \cdot (k-1), \mathcal{T} \cdot k] \subset \tilde{\mathcal{T}}, k = 1, \ldots, p$, is given as

$$\kappa(\tilde{s}, \tilde{t}) = \kappa_{jk}(s, t),$$

with $s, t \in \mathcal{T}$. The eigendecomposition of Y is then given as

$$\mathcal{K}\, \xi_i = \pi_i\, \xi_i \quad \text{and} \quad \kappa(\tilde{s}, \tilde{t}) = \sum_{i=1}^{\infty} \pi_i\, \xi_i(\tilde{s})\, \xi_i'(\tilde{t}),$$

with $\xi_i(\tilde{t}) = \boldsymbol{\psi}_i(t)\boldsymbol{e}_j = \psi_{i,j}(t)$ for $\tilde{t} \in [\mathcal{T} \cdot (j-1), \mathcal{T} \cdot j] \subset \tilde{\mathcal{T}}, j = 1, \ldots, p$. Let $\mathcal{T}_{a_k} \subseteq [\mathcal{T} \cdot (k-1), \mathcal{T} \cdot k] \subset \tilde{\mathcal{T}}, a \in \{1, \ldots, d\}, k \in \{1, \ldots, p\}$, $\tilde{D} = \{1, \ldots, pd\}$, and $\tilde{\mathcal{T}} = \bigcup_{j=1}^{p} \bigcup_{b=1}^{d} \mathcal{T}_{b_j}$, where $\mathcal{T}_{b_j} = \mathcal{T}_b$ and $\mathcal{T} = \bigcup_{b=1}^{d} \mathcal{T}_b$. Then we can define time-coordinate-specific outlyingness contributions based on Shapley values using Proposition 4.5.2.1, which yields

$$\begin{aligned}
\Theta_{k,\mathcal{T}_a}(\boldsymbol{X}, \boldsymbol{\mu}; \mathbf{K}, M) &= \theta_{\mathcal{T}_{a_k}}(Y, \mu; \kappa, M) \\
&= \sum_{R \subseteq \tilde{D} \setminus \{a_k\}} \frac{|R|!(d - |R| - 1)!}{(d)!} \Delta_{\mathcal{T}_{a_k}} \mathrm{fMD}^2(\hat{Y}^R, \mu; \kappa, M) \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle Y - \mu, \xi_i \rangle_{\mathcal{T}_{a_k}} \langle Y - \mu, \xi_i \rangle_{\tilde{\mathcal{T}}} \\
&= \sum_{i=1}^{M} \frac{1}{\pi_i} \langle X_k - \mu_k, \psi_{i,k} \rangle_{\mathcal{T}_a} \langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{\psi}_i \rangle_{\mathcal{T}}
\end{aligned}$$

□

Proof of Lemma 4.5.2.3. We have that $\boldsymbol{X}(t) = \boldsymbol{A}'\,\boldsymbol{\phi}(t)$, $\boldsymbol{\mu}(t) = \boldsymbol{M}_{\boldsymbol{A}}'\,\boldsymbol{\phi}(t)$, and $\kappa(s,t) = \boldsymbol{\phi}'(s)\boldsymbol{\Sigma}^{\text{col}}\,\boldsymbol{\phi}(t)$, for $s,t \in \mathcal{T}$, see Theorem 4.4.1.1. In combination with Corollary 4.5.2.2 we obtain

$$
\begin{aligned}
&\Theta_{\mathcal{T}_a,k}(\boldsymbol{X},\boldsymbol{\mu};\mathbf{K},M) = \\
&= \sum_{i=1}^{m}\sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}}\lambda_j^{\text{row}}} \left(\langle X_k - \mu_k, \xi_i\rangle_{\mathcal{T}_a} v_{j,k}^{\text{row}} \sum_{l=1}^{p} \langle X_l, \xi_i\rangle_{\mathcal{T}} v_{j,l}^{\text{row}} \right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}}\lambda_j^{\text{row}}} \left(\langle (\boldsymbol{a}_k - \boldsymbol{m}_{\boldsymbol{A},k})'\,\boldsymbol{\phi}, \boldsymbol{\phi}'\,\boldsymbol{b}_i\rangle_{\mathcal{T}_a} v_{j,k}^{\text{row}} \sum_{l=1}^{p} \langle (\boldsymbol{a}_l - \boldsymbol{m}_{\boldsymbol{A},l})'\,\boldsymbol{\phi}, \boldsymbol{\phi}'\,\boldsymbol{b}_i\rangle_{\mathcal{T}} v_{j,l}^{\text{row}} \right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}}\lambda_j^{\text{row}}} \left((\boldsymbol{a}_k - \boldsymbol{m}_{\boldsymbol{A},k})' \langle \boldsymbol{\phi}, \boldsymbol{\phi}'\rangle_{\mathcal{T}_a} \boldsymbol{b}_i v_{j,k}^{\text{row}} \sum_{l=1}^{p} \boldsymbol{b}_i' \langle \boldsymbol{\phi}, \boldsymbol{\phi}'\rangle_{\mathcal{T}} (\boldsymbol{a}_l - \boldsymbol{m}_{\boldsymbol{A},l}) v_{j,l}^{\text{row}} \right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{p} \frac{1}{\lambda_i^{\text{ker}}\lambda_j^{\text{row}}} \left((\boldsymbol{a}_k - \boldsymbol{m}_{\boldsymbol{A},k})' \boldsymbol{W}_{\mathcal{T}_a} \boldsymbol{b}_i v_{j,k}^{\text{row}} \sum_{l=1}^{p} \boldsymbol{b}_i' \boldsymbol{W} (\boldsymbol{a}_l - \boldsymbol{m}_{\boldsymbol{A},l}) v_{j,l}^{\text{row}} \right) \\
&= \sum_{j=1}^{p} \frac{1}{\lambda_j^{\text{row}}} v_{j,k}^{\text{row}} (\boldsymbol{a}_k - \boldsymbol{m}_{\boldsymbol{A},k})' \boldsymbol{W}_{\mathcal{T}_a} \underbrace{\left(\sum_{i=1}^{m} \frac{1}{\lambda_i^{\text{ker}}} \boldsymbol{b}_i \boldsymbol{b}_i' \right) \boldsymbol{W}}_{= \boldsymbol{W}^{-1}(\boldsymbol{\Sigma}^{\text{col}})^{-1}, \text{ see Eq. (C.12)}} \sum_{l=1}^{p} (\boldsymbol{a}_l - \boldsymbol{m}_{\boldsymbol{A},l}) v_{j,l}^{\text{row}} \\
&= \sum_{j=1}^{p} \frac{1}{\lambda_j^{\text{row}}} v_{j,k}^{\text{row}} (\boldsymbol{a}_k - \boldsymbol{m}_{\boldsymbol{A},k})' \boldsymbol{W}_{\mathcal{T}_a} \boldsymbol{W}^{-1} (\boldsymbol{\Sigma}^{\text{col}})^{-1} \sum_{l=1}^{p} (\boldsymbol{a}_l - \boldsymbol{m}_{\boldsymbol{A},l}) v_{j,l}^{\text{row}} \\
&= \sum_{j=1}^{p} \frac{1}{\lambda_j^{\text{row}}} v_{j,k}^{\text{row}} (\boldsymbol{a}_k - \boldsymbol{m}_{\boldsymbol{A},k})' \boldsymbol{W}_{\mathcal{T}_a} \boldsymbol{W}^{-1} (\boldsymbol{\Sigma}^{\text{col}})^{-1} (\boldsymbol{A} - \boldsymbol{M}_{\boldsymbol{A}})' \boldsymbol{v}_j^{\text{row}},
\end{aligned}
$$

where $\boldsymbol{W} = \langle \boldsymbol{\phi}, \boldsymbol{\phi}'\rangle_{\mathcal{T}} := \int_{\mathcal{T}} \boldsymbol{\phi}(t)\,\boldsymbol{\phi}'(t)\mathrm{d}t$ and $\boldsymbol{W}_{\mathcal{T}_a} = \langle \boldsymbol{\phi}, \boldsymbol{\phi}'\rangle_{\mathcal{T}_a} := \int_{\mathcal{T}_a} \boldsymbol{\phi}(t)\,\boldsymbol{\phi}'(t)\mathrm{d}t$. □

C.5 Further simulation results

Computation Time

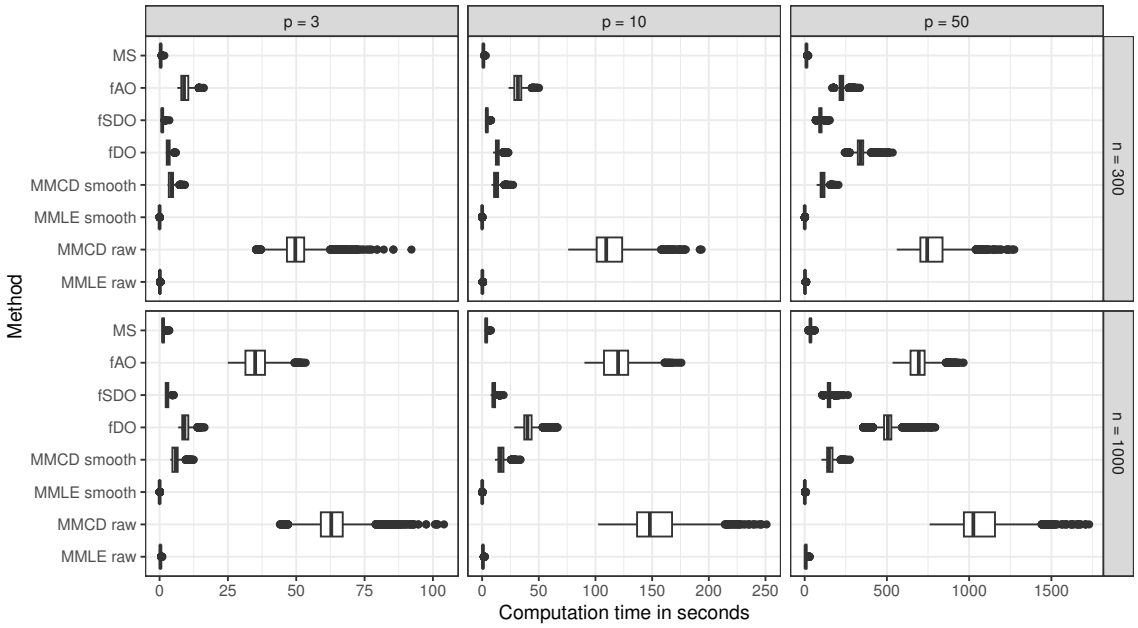


Figure C.1: Comparison of the computation time for all described simulation settings in Section 4.6 divided into facets according to dimensionality $p \in \{3, 10, 50\}$ and number of observations $n \in \{300, 1000\}$.

Comparison of the Depth-based Approaches

Figures C.2 and C.3 show that there are only slight differences in F-Score and the quality of covariance estimation whether depth-based outlier detection is performed on the raw data, smoothed data (evaluated on the raw data's time grid), or the coefficient matrices of the smoothed data.

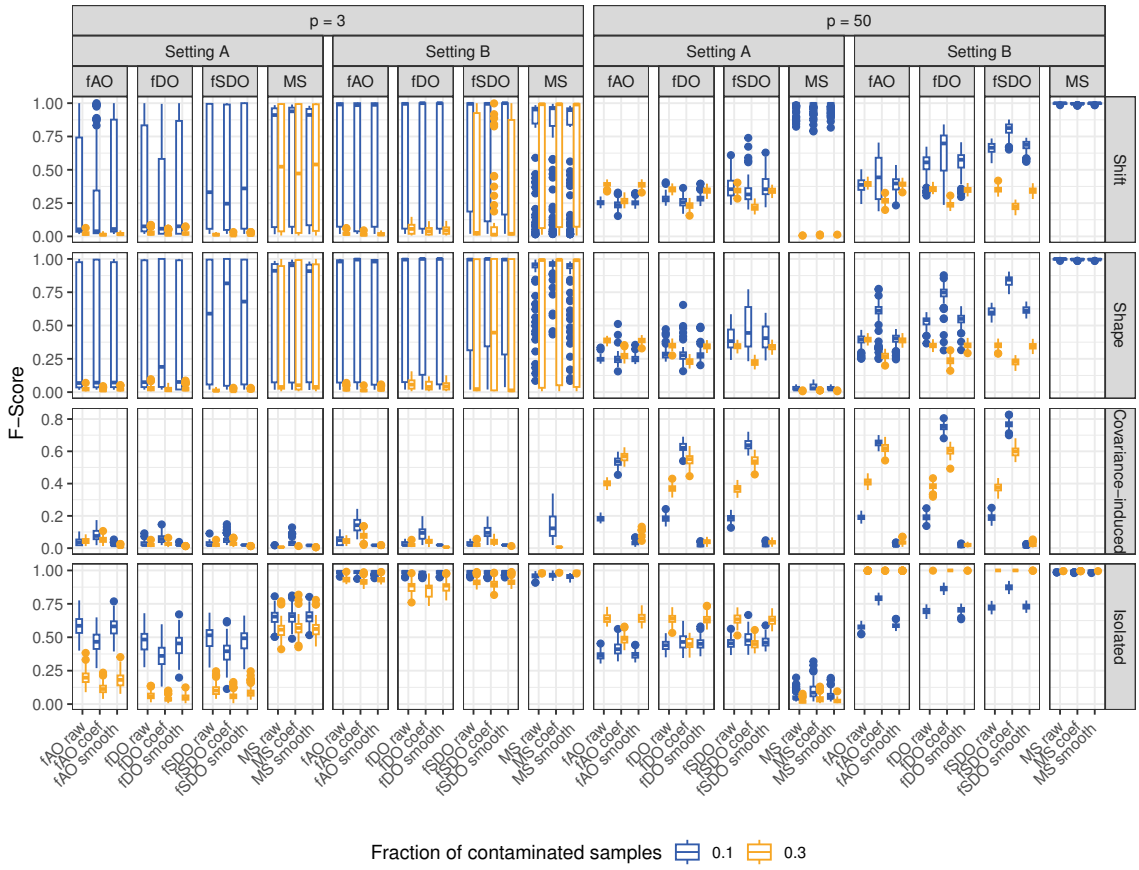


Figure C.2: Comparison of simulation results for depth-based methods in terms of F-Score for $n = 1000$ and $\kappa = \kappa_{\mathrm{M}}$: The top-level horizontal facets represent different dimensionalities, with $p \in \{3, 50\}$. The nested horizontal facets correspond to moderate (Setting A) and severe (Setting B) outliers. The lowest horizontal facets represent the four depth-based methods, each applied to raw data, smoothed data (evaluated on the raw data's time grid), or coefficient matrices. The vertical facets distinguish between the four outlier types. Boxplot colors (blue and orange) correspond to contamination levels of $\varepsilon = 0.1$ and $\varepsilon = 0.3$, respectively.

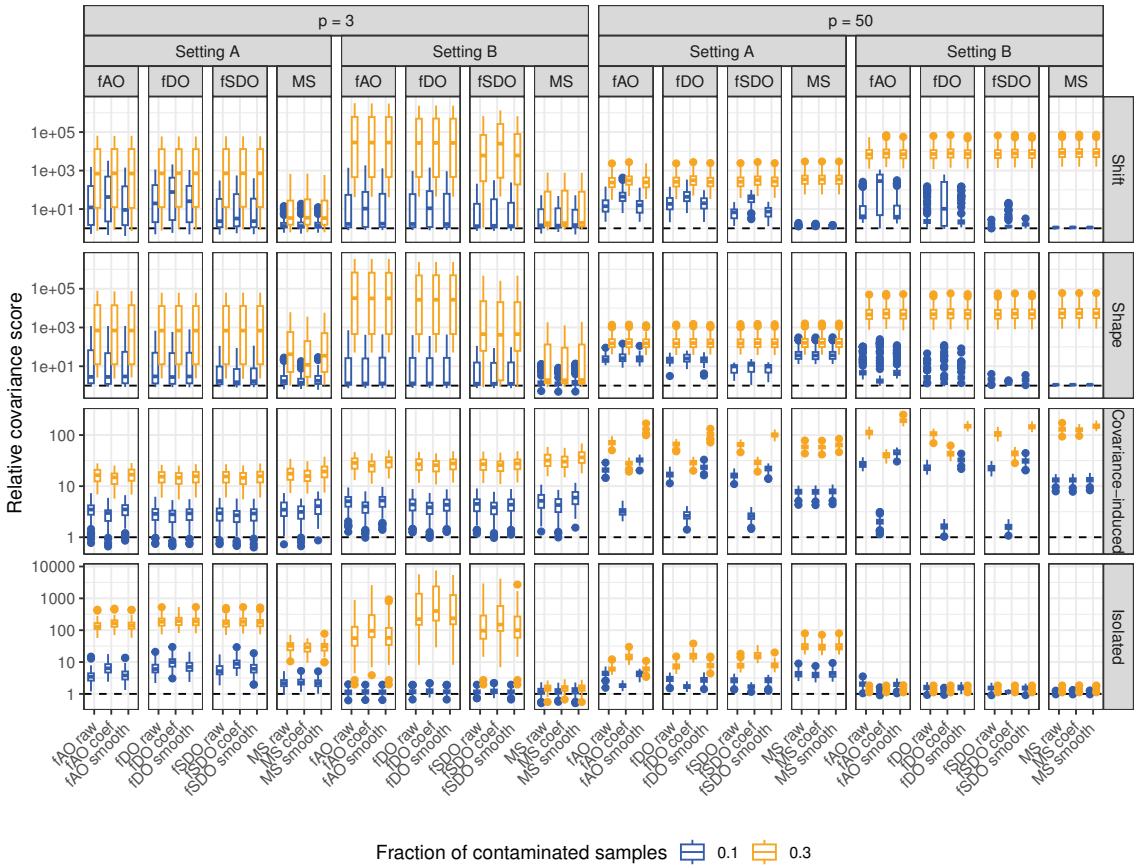


Figure C.3: Comparison of simulation results for depth-based methods in terms of relative covariance estimation error for $n = 1000$ and $\kappa = \kappa_{\mathrm{M}}$: The top-level horizontal facets represent different dimensionalities, with $p \in \{3, 50\}$. The nested horizontal facets correspond to moderate (Setting A) and severe (Setting B) outliers. The lowest horizontal facets represent the four depth-based methods, each applied to raw data, smoothed data (evaluated on the raw data's time grid), or coefficient matrices. The vertical facets distinguish between the four outlier types. Boxplot colors (blue and orange) correspond to contamination levels of $\varepsilon = 0.1$ and $\varepsilon = 0.3$, respectively.

Additional Performance Metrics

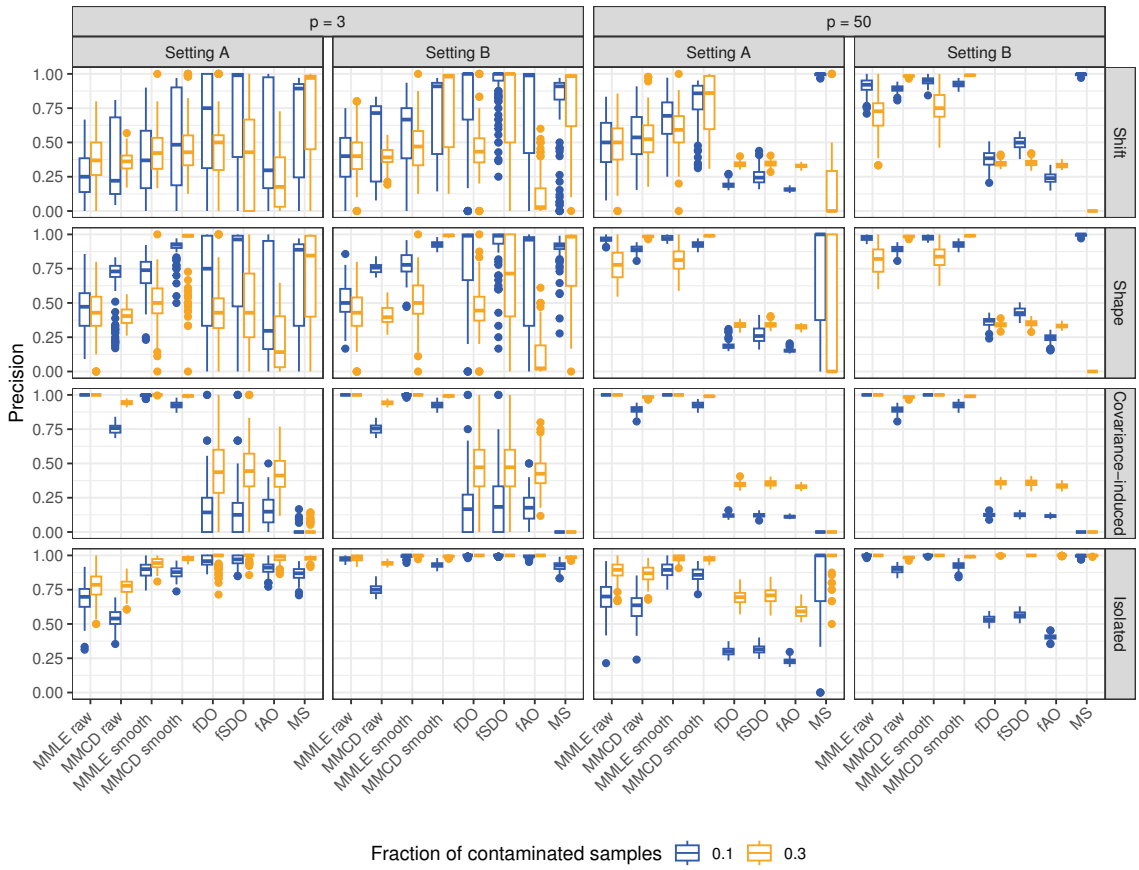


Figure C.4: Comparison of simulation results based on precision for $n = 1000$ and $\kappa = \kappa_{\mathrm{M}}$: The top-level horizontal facets represent different dimensionalities, with $p \in \{3, 50\}$. The nested horizontal facets correspond to moderate (Setting A) and severe (Setting B) outliers. The vertical facets distinguish between the four outlier types. Boxplot colors (blue and orange) correspond to contamination levels of $\varepsilon = 0.1$ and $\varepsilon = 0.3$, respectively.

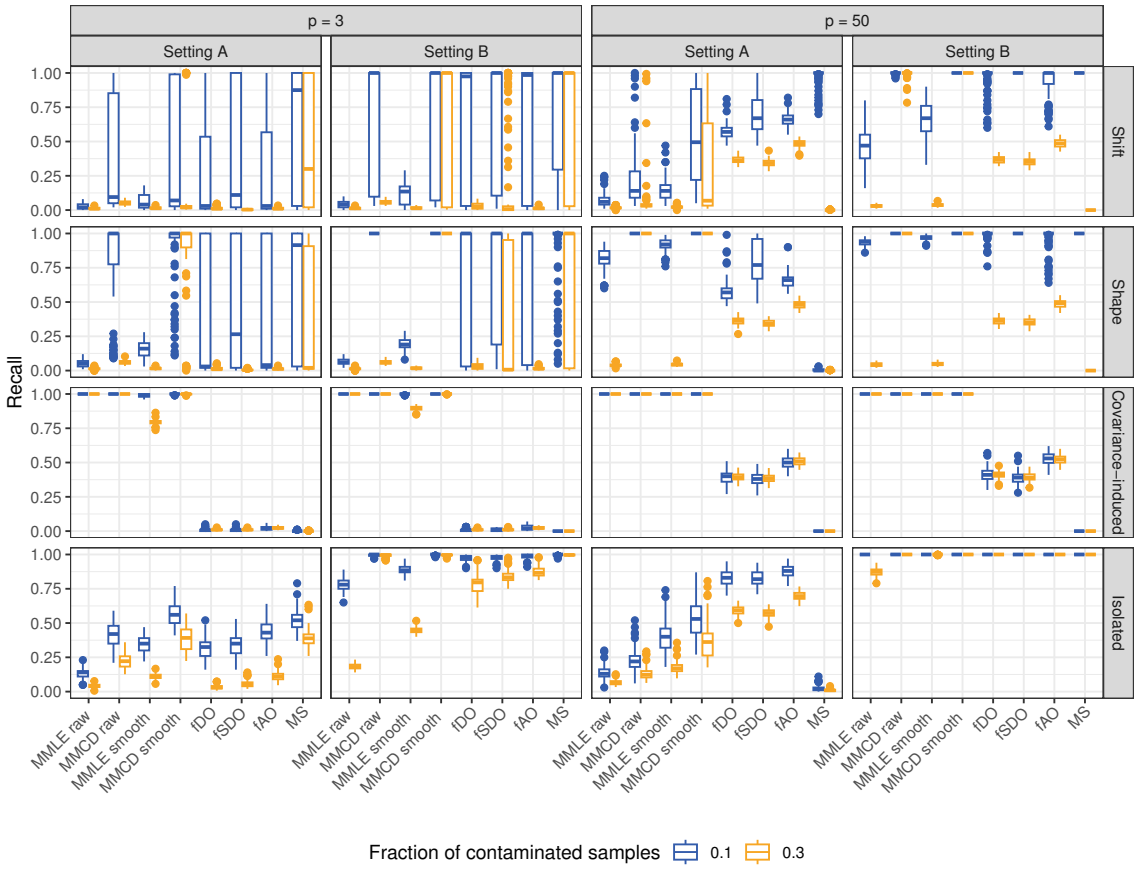


Figure C.5: Comparison of simulation results based on recall for $n = 1000$ and $\kappa = \kappa_{\mathrm{M}}$: The top-level horizontal facets represent different dimensionalities, with $p \in \{3, 50\}$. The nested horizontal facets correspond to moderate (Setting A) and severe (Setting B) outliers. The vertical facets distinguish between the four outlier types. Boxplot colors (blue and orange) correspond to contamination levels of $\varepsilon = 0.1$ and $\varepsilon = 0.3$, respectively.

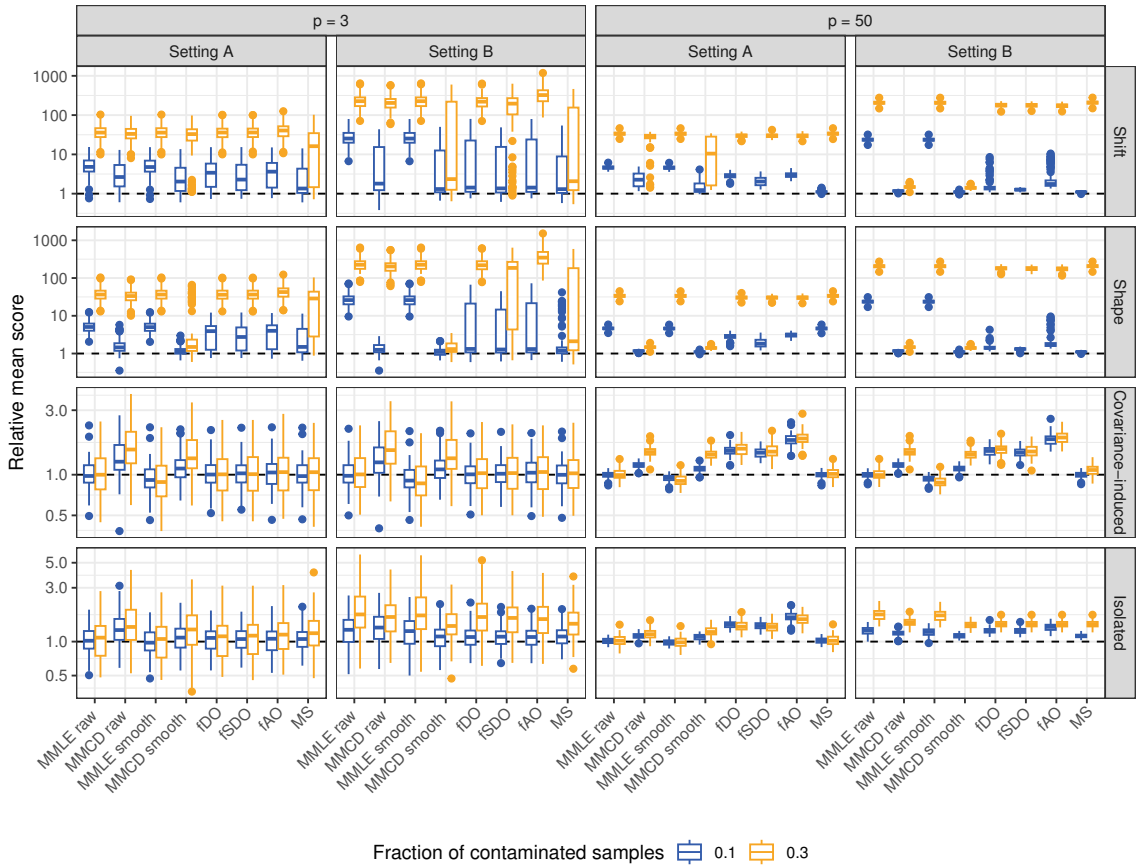


Figure C.6: Comparison of simulation results based on relative mean estimation errors given on a log-scale for $n = 1000$ and $\kappa = \kappa_{\mathrm{M}}$: The top-level horizontal facets represent different dimensionalities, with $p \in \{3, 50\}$. The nested horizontal facets correspond to moderate (Setting A) and severe (Setting B) outliers. The vertical facets distinguish between the four outlier types. Boxplot colors (blue and orange) correspond to contamination levels of $\varepsilon = 0.1$ and $\varepsilon = 0.3$, respectively.

5 Conclusions

We presented a framework for robust location and covariance estimation as well as explainable outlier detection for multivariate, matrix-variate, and functional data. In the following, we summarize the key contributions and findings of each chapter, providing a brief recap of the methodologies and results. Following each summary, we elaborate on the rationale behind our chosen approaches and highlight potential avenues for future research.

In Chapter 2, we proposed multivariate outlier explanations based on the Shapley value, a method rooted in cooperative game theory that has gained popularity in the field of explainable AI. We showed that we can use the Shapley value to decompose the squared Mahalanobis distance of a multivariate observation into variable-specific contributions to multivariate outlyingness. This decomposition is additive, meaning that the sum of the contributions is equal to the squared Mahalanobis distance. Further, the decomposition based on the Shapley value contains information about all 2^p marginal outlyingness contributions of a p-variate observation. We showed that the contributions can be computed with linear computational complexity in p. They rely on the robustness of the associated Mahalanobis distance and, hence, the robustness of the location and covariance estimates, respectively. Therefore, it is imperative to use robust estimates to obtain meaningful outlyingness decompositions. The performance of our proposed cellwise outlier detection procedures based on the variable-specific outlyingness contributions was evaluated using both simulations and real-world datasets. The methods are implemented in `R` (R Core Team, 2024) in the package `ShapleyOutlier`, which is publicly available on CRAN (Mayrhofer and Filzmoser, 2022). We further extended the multivariate outlier explanations based on Shapley values to the matrix-variate and functional setting in Chapters 3 and 4. The visualizations shown in these chapters allow us to gain deeper insights and understandings of why an observation is outlying, confirming the benefits of the Shapley values in real-world applications.

The Shapley value and, thus, the variable-specific outlyingness contributions depend on the characteristic function of the game assigning a value to every possible coalition of players. We proposed to use the squared Mahalanobis distance $\mathrm{MD}^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\hat{\boldsymbol{x}}^S)$ as the characteristic function, where $\hat{\boldsymbol{x}}^S$ is a modified version of $\boldsymbol{x} = (x_1, \ldots, x_p)$ in which all coordinates $j \in \{1, \ldots, n\} \setminus S$, $S \subseteq \{1, \ldots, n\}$, are replaced by their mean μ_j. This replacement strategy allows us to reduce the exponential computational complexity in p to a linear one. However, the resulting marginal contributions $\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) = \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S\cup\{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^S)$ are not necessarily positive. This indicates that the characteristic function is not monotonic, as replacing an additional coordinate with its mean does not always reduce the squared Mahalanobis distance. While the properties of the Shapley value guarantee that the decomposition of the squared Mahalanobis distance is additive, some of the variable-specific outlyingness scores can thus become negative. A negative outlyingness contribution of a coordinate has a clear interpretation in our setting, i.e., replacing this coordinate with its mean would lead to an average increase in the squared Mahalanobis distance. We can

obtain solely positive marginal contributions $\Delta_k \mathrm{MD}^2(\hat{\boldsymbol{x}}^S)$ by replacing the coordinates $j \in \{1, \ldots, n\} \setminus S$, $S \subseteq \{1, \ldots, n\}$ by their conditional expectations $\mathbb{E}[x_{\{1,\ldots,n\}\setminus S}|x_S]$. However, we have not yet found a way to significantly facilitate the computation of the Shapley values, if we use the replacement strategy based on conditional expectations. Therefore, the conditional expectations and the squared Mahalanobis distances would need to be computed for all 2^p combinations to obtain the variable-specific outlyingness contributions based on Shapley values in this setting, which is computationally infeasible even for rather small values of p. Since the outlyingness contributions based on replacement with the mean proposed in Chapter 2 have a clear interpretation, whether they are positive or negative, we did not investigate this approach further, but it may provide an interesting topic for further research.

In Chapter 3, we introduced the Matrix Minimum Covariance Determinant (MMCD) estimators to robustly estimate the location and covariance for matrix-variate data. The MMCD estimators account for the matrix-variate data structure and do not require vectorization of the matrix-variate samples, which would lead to impractically high-dimensional datasets. We showed that the estimators are consistent under matrix-variate elliptical distributions, are matrix-affine equivariant, and achieve a higher breakdown point than the maximum attainable for any multivariate affine equivariant covariance estimator applied to vectorized data. We implemented an efficient algorithm with convergence guarantees in `C++` and integrated it into `R` to ensure efficient execution (Eddelbuettel and François, 2011; R Core Team, 2024). The implementation is publicly available in the `robustmatrix` package on CRAN (Mayrhofer et al., 2024b). The simulations and examples we presented confirm the robustness of the MMCD estimators and demonstrate that the robust Mahalanobis distances based on MMCD estimators provide a reliable method for detecting outliers.

While we studied many robustness properties of the MMCD estimators, deriving and analyzing their influence function remains a topic for future research. Moreover, based on our framework of robust covariance estimation for matrix elliptical distributions, further robust estimators can be generalized to the matrix-variate setting. Based on these, the deterministic Minimum Covariance Determinant (MCD) procedure (Hubert et al., 2012) could be used to compute the MMCD estimators. Another interesting direction for further research is the generalization of cellwise robust procedures such as the cellwise MCD (Raymaekers and Rousseeuw, 2023). Such an approach is particularly relevant in the matrix-variate setting, where sample size is often limited. Discarding entire observations due to a few outlying cells within a matrix can be overly restrictive in many cases. Beyond location and covariance estimation, robust regression approaches like the least trimmed squares estimator of Rousseeuw (1984) could be considered in the matrix-variate setting. We are currently developing robust classification and clustering methods based on the MMCD estimators and extending the MMCD approach to the tensor-variate setting based on maximum likelihood estimators for the tensor normal distribution (Manceur and Dutilleul, 2013). Another topic we would like to pursue in the future is developing a robust test for separability of the covariance matrix in matrix- and tensor-variate settings, based on the works of Lu and Zimmerman (2005) and Filipiak et al. (2016).

In Chapter 4, we proposed a framework for explainable outlier detection and robust estimation of the mean and covariance functions of multivariate functional data. We established a connection between stochastic processes with separable covariance structures and the corresponding matrix-variate distribution of their basis representation. Leveraging

this connection, we employed the MMCD estimators in conjunction with a truncated multivariate functional Mahalanobis semi-distance for robust parameter estimation. Simulations demonstrate that our approach works well for the estimation of the mean and covariance functions, as well as outlier detection in the presence of shift, shape, covariance-induced, and isolated outliers.

In the functional data setting, the covariance operator is generally not invertible due to its infinite-dimensional nature. Specifically, the operator usually has an infinite number of eigenvalues, many of which are close to zero, making inversion infeasible. In our approach, we used basis functions to smooth the raw data and obtain finite-dimensional coefficient matrices. This finite-dimensional representation provides a regularized and invertible version of the covariance operator. In situations where this prior smoothing step is not desired, one can either use our approach based on the raw data or incorporate the smoothing step into the distance computation as proposed by Berrendero et al. (2020) for univariate functional data. Another interesting extension concerns function-valued stochastic processes where each realization is a function rather than a vector (Chen et al., 2017). An example of this approach would be the age-specific fertility rates analyzed in Chapter 4, which we averaged over five-year periods to obtain multivariate functional data. In this context, both the years and age could be treated as continuous variables.

Bibliography

Agostinelli, C., Leung, A., Yohai, V., and Zamar, R. (2015a). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24:441–461.

Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015b). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24:441–461.

Alfons, A. (2021). robustHD: An R package for robust regression with high-dimensional data. *Journal of Open Source Software*, 6(67):3786.

Alqallaf, F., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, pages 311–331.

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266.

Arribas-Gil, A. and Romo, J. (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Ash, R. B. and Gardner, M. F. (2014). *Topics in Stochastic Processes: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*, volume 27. Academic press.

Baba, K., Shibata, R., and Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664.

Basna, R., Nassar, H., and Podgórski, K. (2022). Data driven orthogonal basis selection for functional data analysis. *Journal of Multivariate Analysis*, 189:104868.

Berrendero, J. R., Bueno-Larraz, B., and Cuevas, A. (2020). On Mahalanobis distance in functional settings. *Journal of Machine Learning Research*, 21(9):1–33.

Biecek, P. and Burzykowski, T. (2021). *Explanatory Model Analysis*. Chapman and Hall/CRC, New York.

Boente, G. and Salibián-Barrera, M. (2015). S-estimators for functional principal component analysis. *Journal of the American Statistical Association*, 110(511):1100–1111.

Boente, G. and Salibián-Barrera, M. (2021). Robust functional principal components for sparse longitudinal data. *Metron*, 79(2):159–188.

Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020). The minimum regularized covariance determinant estimator. *Statistics and Computing*, 30(1):113–128.

Brown, P. J., Kenward, M. G., and Bassett, E. E. (2001). Bayesian discrimination with longitudinal data. *Biostatistics*, 2(4):417–432.

Butler, R., Davies, P., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, pages 1385–1400.

Cator, E. A. and Lopuhaä, H. P. (2012). Central limit theorem and influence function for the MCD estimators at general multivariate distributions. *Bernoulli*, 18(2):520 – 551.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3).

Chen, K., Delicado, P., and Müller, H.-G. (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):177–196.

Chen, W., Genton, M. G., and Sun, Y. (2021). Space-time covariance structures and models. *Annual Review of Statistics and Its Application*, 8(1):191–215.

Chen, Z., Fan, J., and Wang, K. (2023). Multivariate Gaussian processes: definitions, examples and applications. *METRON*, pages 1–11.

Chiou, J.-M., Chen, Y.-T., and Yang, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, pages 1571–1596.

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.

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1339.

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.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23.

Dai, W. and Genton, M. G. (2018). Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, 27(4):923–934.

Dai, W. and Genton, M. G. (2019). Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, 131:50–65.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, pages 1269–1292.

Daw, R., Simpson, M., Wikle, C. K., Holan, S. H., and Bradley, J. R. (2022). An overview of univariate and multivariate Karhunen Loève expansions in statistics. *Journal of the Indian Society for Probability and Statistics*, 23(2):285–326.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274.

Debruyne, M., Höppner, S., Serneels, S., and Verdonck, T. (2019). Outlyingness: Which variables contribute most? *Statistics and Computing*, 29:707–723.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Technical report, Harvard University, Boston.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89 – 121.

Ferraty, F. (2006). *Nonparametric Functional Data Analysis*. Springer.

Filipiak, K., Klein, D., and Roy, A. (2016). Score test for a separable covariance structure with the first component as compound symmetric correlation matrix. *Journal of Multivariate Analysis*, 150:105–124.

Filzmoser, P., Ruiz-Gazen, A., and Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, 55.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

Fujimoto, K., Kojadinovic, I., and Marichal, J.-L. (2006). Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55:72–99.

Galeano, P., Joseph, E., and Lillo, R. E. (2015). The Mahalanobis distance for functional data with applications to classification. *Technometrics*, 57(2):281–291.

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.

Genton, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics: The Official Journal of the International Environmetrics Society*, 18(7):681–695.

Ghiglietti, A., Ieva, F., and Paganoni, A. M. (2017). Statistical inference for stochastic processes: two-sample hypothesis tests. *Journal of Statistical Planning and Inference*, 180:49–68.

Grabisch, M. (2016). *Set Functions, Games and Capacities in Decision Making*. Springer Publishing Company, Incorporated, 1st edition.

Grabisch, M. and Roubens, M. (1999). An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28:547–565.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Gupta, A. and Nagar, D. (1999). *Matrix Variate Distributions*. Monographs and Surveys in Pure and Applied Mathematics. Taylor & Francis.

Gupta, A. K., Varga, T., and Bodnar, T. (2013). *Elliptically Contoured Models in Statistics and Portfolio Theory*. Springer, 2 edition.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley-Interscience, New York.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning*, volume 2 of *Springer Series in Statistics*. Springer New York.

Hollander, M. (2013). *Nonparametric Statistical Methods*. John Wiley & Sons Inc.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M. (2017). Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *Journal of Climate*, 30(20):8179–8205.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

Huber, P. J. and Ronchetti, E. M. (2011). *Robust Statistics*. John Wiley & Sons.

Hubert, M., Rousseeuw, P. J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–202.

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 Van der Veeken, S. (2008). Outlier detection for skewed data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 22(3-4):235–246.

Human Fertility Database (2024). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at `www.humanfertility.org`.

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis*. Prentice Hall, 4th edition edition.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC.

Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 172:211–222.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Li, L., Huang, W., Gu, I. Y.-H., and Tian, Q. (2004). Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American statistical Association*, 104(486):718–734.

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

Lu, N. and Zimmerman, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statistics & Probability Letters*, 73(4):449–457.

Lundberg, S. M., Erion, G., Chen, H., Degrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., and et al. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

Manceur, A. M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, 239:37–49.

Mardia, K. V. and Goodall, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate Environmental Statistics*, 6(76):347–385.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons.

Martino, A., Ghiglietti, A., Ieva, F., and Paganoni, A. M. (2019). A k-means procedure based on a Mahalanobis type distance for clustering multivariate functional data. *Statistical Methods & Applications*, 28:301–322.

Mayrhofer, M. and Filzmoser, P. (2022). *ShapleyOutlier: Multivariate Outlier Explanations using Shapley Values and Mahalanobis Distances*. R package version 0.1.2.

Mayrhofer, M. and Filzmoser, P. (2023). Multivariate outlier explanations using Shapley values and Mahalanobis distances. *Econometrics and Statistics*.

Mayrhofer, M., Radojičić, U., and Filzmoser, P. (2024a). Robust covariance estimation and explainable outlier detection for matrix-valued data. *arXiv preprint arXiv:2403.03975*.

Mayrhofer, M., Radojičić, U., and Filzmoser, P. (2024b). *robustmatrix: Robust Matrix-Variate Parameter Estimation*. R package version 0.1.3.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

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.

Oguamalam, J., Radojičić, U., and Filzmoser, P. (2024). Minimum regularized covariance trace estimator and outlier detection for functional data. *Technometrics*, pages 1–12.

Ojo, O. T., Lillo, R. E., and Fernandez Anta, A. (2023). *fdaoutlier: Outlier Detection Tools for Functional Data Analysis*. R package version 0.2.1.

Owen, G. (1972). Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79.

Peters, H. (2008). *Game Theory*. Springer, Berlin Heidelberg.

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

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer.

Raymaekers, J. and Rousseeuw, P. (2021). Handling cellwise outliers by sparse regression and robust covariance. *Journal of Data Science, Statistics, and Visualisation*, 1.

Raymaekers, J. and Rousseeuw, P. J. (2023). The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association*, 0(0):1–12.

Raymaekers, J. and Rousseeuw, P. J. (2024). Challenges of cellwise outliers. *Econometrics and Statistics*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*.

Rodríguez-Iturbe, I. and Mejía, J. M. (1974). The design of rainfall networks in time and space. *Water Resources Research*, 10(4):713–728.

Roś, B., Bijma, F., de Munck, J. C., and de Gunst, M. C. (2016). Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure. *Journal of Multivariate Analysis*, 143:345–361.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications Vol. B*, pages 283–297.

Rousseeuw, P. and Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of The American Statistical Association*, 85:633–639.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

Rousseeuw, P. J. and Bossche, W. V. D. (2018). Detecting deviating data cells. *Technometrics*, 60(2):135–145.

Rousseeuw, P. J., Raymaekers, J., and Hubert, M. (2018). A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, 27(2):345–359.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Seber, G. A. F. (1984). *Multivariate Observations*. Wiley series in probability and statistics. John Wiley & Sons.

Segaert, P., Hubert, M., Rousseeuw, P., and Raymaekers, J. (2024). *mrfDepth: Depth Measures in Multivariate, Regression and Functional Settings*. R package version 1.0.17.

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98:121–142.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

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

Srivastava, M. S., von Rosen, T., and Von Rosen, D. (2008). Models with a Kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics*, 17:357–370.

Stadt Wien (2022). Monthly data from the weather station Hohe Warte since April 1872 - Vienna.

Stahel, W. A. (1981). *Breakdown of covariance estimators*. Fachgruppe für Statistik, Eidgenössische Techn. Hochsch.

Sun, Y., Babu, P., and Palomar, D. P. (2016). Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Transactions on Signal Processing*, 64(14):3576–3590.

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. (2020). The Shapley Taylor interaction index. In *International Conference on Machine Learning*, pages 9259–9268. PMLR.

Thompson, G. Z., Maitra, R., Meeker, W. Q., and Bastawros, A. F. (2020). Classification with the matrix-variate-t distribution. *Journal of Computational and Graphical Statistics*, 29(3):668–674.

Trenberth, K. E. (1997). The definition of el nino. *Bulletin of the American Meteorological Society*, 78(12):2771–2778.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pages 448–485.

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

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Review of functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295.

Wang, L. (2008). *Karhunen-Loeve expansions and their applications*. London School of Economics and Political Science (United Kingdom).

Werner, K., Jansson, M., and Stoica, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2):478–491.

Withers, C. (1974). Mercer's Theorem and Fredholm resolvents. *Bulletin of the Australian Mathematical Society*, 11(3):373–380.

Young, H. (1985). Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14:65–72.

Zhang, Y., Shen, W., and Kong, D. (2022). Covariance estimation for matrix-valued data. *Journal of the American Statistical Association*, pages 1–12.

Zhu, H., Strawn, N., and Dunson, D. B. (2016). Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17.

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:e1280.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *The Annals of Statistics*, 31(5):1460–1490.

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

Štrumbelj, E. and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665.

Curriculum Vitae

Personal Data

Name	**Marcus Mayrhofer**
Date of Birth	29.11.1995
Nationality	Austrian
E-mail	`marcus.mayrhofer@tuwien.ac.at`

Professional Experience

11/2021 – present	**Project assistant** in the CSTAT group, Institute of Statistics and Mathematical Methods in Economics, *TU Wien*, Austria
07/2021	**Scientific internship (Master's thesis)** at Technik-steuerung/Angewandte Statistik, *voestalpine Stahl GmbH*, Linz, Austria
10/2020 – 07/2021	**Tutor** in the CSTAT group, Institute of Statistics and Mathematical Methods in Economics, *TU Wien*, Austria
07/2020 – 08/2020	**Scientific internship** at CE Numerical Simulation, Research & Development, *Fronius International GmbH*, Thalheim, Austria
08/2020 – 09/2020	**Scientific internship (Bachelor's thesis)** in the CE Numerical Simulation group, Research & Development, *Fronius International GmbH*, Thalheim bei Wels, Austria
07/2019	**Scientific internship** at Technik-steuerung/Angewandte Statistik, *voestalpine Stahl GmbH*, Linz, Austria
07/2017 – 09/2017	**IT internship**, *voestalpine Automotive components GmbH*, Aguascalientes, Mexico
07/2015 – 09/2015	**IT internship**, *voestalpine Stamptec GmbH*, Shenyang, China

---

Education

11/2021 – present	**Doctoral program** in Engineering Sciences - Technical Mathematics, TU Wien, Austria
06/2020 – 10/2021	**Master's program** in Statistics and Mathematical Methods in Economics, TU Wien, Austria *Graduated with honors*
07/2016 – 06/2020	**Bachelor's program** in Statistics and Mathematical Methods in Economics, TU Wien, Austria
09/2010 – 06/2015	**Matura** (A levels, higher education entrance qualification), BBS Rohrbach, Austria *Graduated with honors*

List of Publications

Mayrhofer, M., Radojičić, U., and Filzmoser, P. (2024). Robust covariance estimation and explainable outlier detection for matrix-valued data. *arXiv preprint arXiv:2403.03975*.

Mayrhofer, M. and Filzmoser, P. (2023). Multivariate outlier explanations using Shapley values and Mahalanobis distances. *Econometrics and Statistics*. DOI: 10.1016/j.ecosta.2023.04.003.

Mayrhofer, M. (2021). Explainable artificial intelligence methods for modeling categorical responses. Master's thesis, Technische Universität Wien

List of Published Software

Mayrhofer, M., Radojičić, U., and Filzmoser, P. (2024b). *robustmatrix: Robust Matrix-Variate Parameter Estimation*. R package version 0.1.3

Mayrhofer, M. and Filzmoser, P. (2022). *ShapleyOutlier: Multivariate Outlier Explanations using Shapley Values and Mahalanobis Distances*. R package version 0.1.2

List of Presentations

Mayrhofer, M., Radojicic, U., and Filzmoser, P. (2024). A minimum covariance determinant approach for matrix-variate data. *Statistische Woche 2024*, Regensburg, Germany.

Mayrhofer, M., Radojicic, U., and Filzmoser, P. (2024). Robust covariance estimation for matrix-valued data. *Bernoulli-IMS 11th World Congress in Probability and Statistics*, Bochum, Germany.

Mayrhofer, M., Radojicic, U., and Filzmoser, P. (2024). Explainable anomaly detection using Shapley values. *Austrian Statistical Days 2024*, Vienna, Austria.

Mayrhofer, M. and Filzmoser, P. (2023). Explainable Multivariate Outlier Detection based on Shapley Values. *Olomoucian Days of Applied Mathematics (ODAM 2023)*, Olomouc, Czech Republic

Mayrhofer, M. and Filzmoser, P. (2023). Explainable outlier detection based on Shapley values. *25th International Conference on Computational Statistics (COMPSTAT 2023)*, London, United Kingdom

Mayrhofer, M., Rieser, C., and Filzmoser, P. (2023). L0 Regularized Cellwise Outlier Detection and Covariance Estimation. *Joint conference of Data Science, Statistics & Visualisation and the European Conference on Data Analysis*, Antwerp, Belgium

Mayrhofer, M., Lewitschnig, H., and Filzmoser, P. (2023). New Mission Profile Model Using Functional Data Analysis. *Infineon meets University 2023*, Munich, Germany.

Mayrhofer, M., Radojicic, U., Lewitschnig, H., and Filzmoser, P. (2023). Outlier detection and explanation for matrix-valued data. *International Conference on Robust Statistics (ICORS 2023)*, Toulouse, France.

Mayrhofer, M., and Filzmoser, P. (2022). Outlier explanation using Shapley values and Mahalanobis distances. *International Conference on Robust Statistics (ICORS 2022)*, Waterloo, Canada.

Vienna, October 24, 2024

Marcus Mayrhofer