Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: www.tandfonline.com/journals/utch20

Robust Covariance Estimation and Explainable Outlier Detection for Matrix-Valued Data

Marcus Mayrhofer, Una Radojičić & Peter Filzmoser

To cite this article: Marcus Mayrhofer, Una Radojičić & Peter Filzmoser (2025) Robust Covariance Estimation and Explainable Outlier Detection for Matrix-Valued Data, Technometrics, 67:3, 516-530, DOI: 10.1080/00401706.2025.2475781

To link to this article: https://doi.org/10.1080/00401706.2025.2475781

TECHNOMETRICS
2025, VOL. 67, NO. 3, 516–530
https://doi.org/10.1080/00401706.2025.2475781

Taylor & Francis
Taylor & Francis Group

Robust Covariance Estimation and Explainable Outlier Detection for Matrix-Valued Data

Marcus Mayrhofer, Una Radojičić, and Peter Filzmoser

Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wien, Austria

ABSTRACT

This work introduces the Matrix Minimum Covariance Determinant (MMCD) method, a novel robust location and covariance estimation procedure designed for data that are naturally represented in the form of a matrix. Unlike standard robust multivariate estimators, which would only be applicable after a vectorization of the matrix-variate samples leading to high-dimensional datasets, the MMCD estimators account for the matrix-variate data structure and consistently estimate the mean matrix, as well as the rowwise and columnwise covariance matrices in the class of matrix-variate elliptical distributions. Additionally, we show that the MMCD estimators are matrix affine equivariant and achieve a higher breakdown point than the maximal achievable one by any multivariate, affine equivariant location/covariance estimator when applied to the vectorized data. An efficient algorithm with convergence guarantees is proposed and implemented. As a result, robust Mahalanobis distances based on MMCD estimators offer a reliable tool for outlier detection. Additionally, we extend the concept of Shapley values for outlier explanation to the matrix-variate setting, enabling the decomposition of the squared Mahalanobis distances into contributions of the rows, columns, or individual cells of matrix-valued observations. Notably, both the theoretical guarantees and simulations show that the MMCD estimators outperform robust estimators based on vectorized observations, offering better computational efficiency and improved robustness. Moreover, real-world data examples demonstrate the practical relevance of the MMCD estimators and the resulting robust Shapley values.

ARTICLE HISTORY
Received May 2024
Accepted February 2025

KEYWORDS
Covariance with Kronecker structure; Explainable artificial intelligence; Image data; Matrix-variate distributions; Minimum covariance determinant; Shapley values

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 and Varga 2012), serves as a natural ground for studying covariance estimation. The matrix-elliptical family is a semi-parametric class of distributions parameterized by the mean $\boldsymbol{M} \in \mathbb{R}^{p\times q}$, row covariance $\boldsymbol{\Sigma}^{\text{row}} \in \text{PDS}(p)$, column covariance $\boldsymbol{\Sigma}^{\text{col}} \in \text{PDS}(q)$, 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 $\mathcal{ME}(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}, g)$, if its density can be written as

$$f(\boldsymbol{X}) = \det(\boldsymbol{\Sigma}^{\text{row}})^{-q/2} \det(\boldsymbol{\Sigma}^{\text{col}})^{-p/2} \\ g(\text{tr}(\boldsymbol{\Omega}^{\text{col}}(\boldsymbol{X} - \boldsymbol{M})'\boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X} - \boldsymbol{M}))), \quad (1)$$

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 $\mathcal{ME}(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}, g)$ if and only if its vectorized version $\text{vec}\,\boldsymbol{X}$ follows a multivariate elliptical distribution $\mathcal{E}(\text{vec}(\boldsymbol{M}), \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}, g)$ (Gupta and Varga 2012). Here, $\text{vec}(\cdot)$ 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 $\mathcal{MN}(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$, with density

CONTACT Marcus Mayrhofer marcus.mayrhofer@tuwien.ac.at Institute of Statistics and Mathematical Methods in Economics, TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria.

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

Regarding the estimation of location and covariance for an iid sample $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \in \mathbb{R}^{n\times p\times q}$, with $\boldsymbol{X}_i \sim \mathcal{MN}(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$, 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 use 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):

$$\hat{\boldsymbol{M}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i \quad (3)$$

$$\hat{\boldsymbol{\Sigma}}^{\text{row}} = \frac{1}{qn}\sum_{i=1}^{n}(\boldsymbol{X}_i - \hat{\boldsymbol{M}})\hat{\boldsymbol{\Omega}}^{\text{col}}(\boldsymbol{X}_i - \hat{\boldsymbol{M}})' \quad (4)$$

$$\hat{\boldsymbol{\Sigma}}^{\text{col}} = \frac{1}{pn}\sum_{i=1}^{n}(\boldsymbol{X}_i - \hat{\boldsymbol{M}})'\hat{\boldsymbol{\Omega}}^{\text{row}}(\boldsymbol{X}_i - \hat{\boldsymbol{M}}). \quad (5)$$

Soloveychik and Trushin (2016) showed that for n iid 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 (4) and (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, Babu, and Palomar 2016), distribution-free robust covariance estimation (Zhang, Shen, and Kong 2022), and ML estimation for the matrix t-distribution (Thompson et al. 2020).

We propose novel robust estimators for the parameters $\boldsymbol{M}$, $\boldsymbol{\Sigma}^{\text{row}}$, and $\boldsymbol{\Sigma}^{\text{col}}$, 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 $(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$ 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 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 article is organized as follows. In Section 2, we introduce the MMCD estimators, then proceed to derive their theoretical properties in Section 3. Section 4 is devoted to computational details for the MMCD estimators. In Section 5, we propose Shapley values for outlier explanation and present their properties. In Sections 6 and 7, we illustrate the performance of the proposed methods on numerical simulations and real-world examples. Section 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.

2. The MMCD Estimators

The MLEs given in (3)–(5), much like the multivariate normal MLEs, that is sample mean and covariance, also serve as valid (consistent) parameter estimators in the class of elliptical distributions; see Remark 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 $(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$ in (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, Hoffmann, and Filzmoser 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 $\mathcal{MN}(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$ to formulate the weighted log-likelihood function $l(\boldsymbol{w}, \boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}} | \mathfrak{X})$ as

$$-\frac{1}{2}\sum_{i=1}^{n} w_i \Big(p \ln(\det(\boldsymbol{\Sigma}^{\text{col}})) + q \ln(\det(\boldsymbol{\Sigma}^{\text{row}})) + \text{MMD}^2(\boldsymbol{X}_i) + pq\ln(2\pi) \Big), \tag{6}$$

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

$$\begin{aligned}\text{MMD}^2(\boldsymbol{X}) &:= \text{MMD}^2(\boldsymbol{X}; \boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}) \\ &= \text{tr}(\boldsymbol{\Omega}^{\text{col}}(\boldsymbol{X} - \boldsymbol{M})' \boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X} - \boldsymbol{M})).\end{aligned} \tag{7}$$

Setting $w_i = 1$ for all $i = 1, \ldots, n$, yields the traditional log-likelihood function, and its maximization yields the MLEs of (3)–(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

$$\begin{aligned}&\max_{\boldsymbol{w}, \boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}} l(\boldsymbol{w}, \boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}} | \mathfrak{X}) \\ &\text{s.t.} \quad w_i \in \{0, 1\} \text{ for all } i = 1, \ldots, n \quad \text{and} \quad \sum_{i=1}^{n} w_i = h.\end{aligned} \tag{8}$$

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 2.0.1. Let $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, $n/2 \le h \le n$ and $h \ge \lfloor p/q + q/p \rfloor + 2$, be an iid sample from $\mathcal{MN}(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$. Maximizing the weighted log-likelihood function (8) is equivalent to minimizing

$$\ln(\det(\hat{\boldsymbol{\Sigma}}_H^{\text{col}} \otimes \hat{\boldsymbol{\Sigma}}_H^{\text{row}})) = p \ln(\det(\hat{\boldsymbol{\Sigma}}_H^{\text{col}})) + q \ln(\det(\hat{\boldsymbol{\Sigma}}_H^{\text{row}})) \tag{9}$$

across all subsets $H \subset \{1, \ldots, n\}$ with $|H| = h$. In (9),

$$\hat{\boldsymbol{M}}_H = \frac{1}{h}\sum_{i \in H} \boldsymbol{X}_i, \tag{10}$$

$$\hat{\boldsymbol{\Sigma}}_H^{\text{row}} = \frac{1}{qh}\sum_{i \in H} (\boldsymbol{X}_i - \hat{\boldsymbol{M}}_H)\hat{\boldsymbol{\Omega}}_H^{\text{col}}(\boldsymbol{X}_i - \hat{\boldsymbol{M}}_H)', \text{ and} \tag{11}$$

$$\hat{\boldsymbol{\Sigma}}_H^{\text{col}} = \frac{1}{ph}\sum_{i \in H} (\boldsymbol{X}_i - \hat{\boldsymbol{M}}_H)'\hat{\boldsymbol{\Omega}}_H^{\text{row}}(\boldsymbol{X}_i - \hat{\boldsymbol{M}}_H) \tag{12}$$

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

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

Definition 2.0.1. Let $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, $n/2 \le h \le n$ and $h \ge \lfloor p/q + q/p \rfloor + 2$, be an iid sample of a continuous $p \times q$ matrix-variate distribution. The *raw* matrix minimum covariance determinant (MMCD) estimators $(\hat{\boldsymbol{M}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}, \hat{\boldsymbol{\Sigma}}^{\text{col}})$ are defined as

$$\underset{\substack{\hat{\boldsymbol{M}}_H, \hat{\boldsymbol{\Sigma}}_H^{\text{row}}, \hat{\boldsymbol{\Sigma}}_H^{\text{col}} \\ H \subset \{1, \ldots, n\}, |H| = h}}{\arg\min} \quad p \ln(\det(\hat{\boldsymbol{\Sigma}}_H^{\text{col}})) + q \ln(\det(\hat{\boldsymbol{\Sigma}}_H^{\text{row}})), \tag{13}$$

with $\hat{\boldsymbol{M}}_H$, $\hat{\boldsymbol{\Sigma}}_H^{\text{row}}$, and $\hat{\boldsymbol{\Sigma}}_H^{\text{col}}$ as in (10), (11), and (12), respectively.

The estimators in Definition 2.0.1 almost surely exist and are positive definite if $h \ge \lfloor p/q + q/p \rfloor + 2$ (Soloveychik and Trushin 2016). If $p = 1$ and/or $q = 1$, optimization problem (13) coincides with the optimization problem of the MCD estimator, and one obtains the univariate or multivariate MCD estimator, respectively.

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 $\boldsymbol{X} \sim \mathcal{ME}(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}}, g)$ also have an elliptical distribution (Gupta and Varga 2012). This means that for constant matrices $\boldsymbol{A} \in \mathbb{R}^{r \times p}$, $\text{rank}(\boldsymbol{A}) = r \le p$, $\boldsymbol{B} \in \mathbb{R}^{q \times s}$, $\text{rank}(\boldsymbol{B}) = s \le q$, and $\boldsymbol{C} \in \mathbb{R}^{r \times s}$, the transformed random matrix $\boldsymbol{Z} = \boldsymbol{AXB} + \boldsymbol{C}$ has density

$$\boldsymbol{Z} \sim \mathcal{ME}(\boldsymbol{AMB} + \boldsymbol{C}, \boldsymbol{A}\boldsymbol{\Sigma}^{\text{row}}\boldsymbol{A}', \boldsymbol{B}'\boldsymbol{\Sigma}^{\text{col}}\boldsymbol{B}, g). \tag{14}$$

Let $\hat{\boldsymbol{M}}_{\mathfrak{X}}$, $\hat{\boldsymbol{\Sigma}}_{\mathfrak{X}}^{\text{row}}$, and $\hat{\boldsymbol{\Sigma}}_{\mathfrak{X}}^{\text{col}}$ denote the estimators based on a sample $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ generated by $f(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{col}})$. Then the estimators of the sample $\mathfrak{Z} = (\boldsymbol{AX}_1\boldsymbol{B} + \boldsymbol{C}, \ldots, \boldsymbol{AX}_n\boldsymbol{B} + \boldsymbol{C})$ should transform in the same way as the parameters in (14), that is,

$$\begin{aligned}&\hat{\boldsymbol{M}}_{\mathfrak{Z}} = \boldsymbol{A}\hat{\boldsymbol{M}}_{\mathfrak{X}}\boldsymbol{B} + \boldsymbol{C}, \quad \hat{\boldsymbol{\Sigma}}_{\mathfrak{Z}}^{\text{row}} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\mathfrak{X}}^{\text{row}}\boldsymbol{A}', \\ &\hat{\boldsymbol{\Sigma}}_{\mathfrak{Z}}^{\text{col}} = \boldsymbol{B}'\hat{\boldsymbol{\Sigma}}_{\mathfrak{X}}^{\text{col}}\boldsymbol{B}.\end{aligned} \tag{15}$$

Properties (15) 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)–(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 $\text{vec}(\boldsymbol{AXB} + \boldsymbol{C}) = (\boldsymbol{B}' \otimes \boldsymbol{A})\,\text{vec}(\boldsymbol{X}) + \text{vec}(\boldsymbol{C})$. We refer to Properties (15) as *matrix affine equivariance* to avoid confounding definitions.

Lemma 3.0.1. Let $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ be a sample of $p \times q$ matrices, where $\boldsymbol{X}_i \sim \mathcal{ME}(\boldsymbol{M}_{\mathfrak{X}}, \boldsymbol{\Sigma}^{\text{row}}_{\mathfrak{X}}, \boldsymbol{\Sigma}^{\text{col}}_{\mathfrak{X}}, g)$, and let $\mathfrak{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ be the affine transformation of $\mathfrak{X}$, that is, $\boldsymbol{Z}_i = \boldsymbol{A}\boldsymbol{X}_i\boldsymbol{B} + \boldsymbol{C}$, $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, $\boldsymbol{B} \in \mathbb{R}^{q\times q}$, $\boldsymbol{A}, \boldsymbol{B}$ invertible, and $\boldsymbol{C} \in \mathbb{R}^{p\times q}$. The following then holds:

(a) The MMCD estimators as in Definition 2.0.1 are matrix affine equivariant.

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

Lemma 3.0.1 shows that the MMCD estimators are equivariant under matrix affine transformations, and a proof is given in Supplement B.1.

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 $\hat{\boldsymbol{M}}$ is given by

$$\varepsilon^*(\hat{\boldsymbol{M}}, \mathfrak{X}) = \min_{1\le m\le n} \left\{ \frac{m}{n} : \sup_{\mathfrak{Y}} \left\| \hat{\boldsymbol{M}}_{\mathfrak{X}} - \hat{\boldsymbol{M}}_{\mathfrak{Y}} \right\| = \infty \right\} \tag{16}$$

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

$$\varepsilon^*(\hat{\boldsymbol{\Sigma}}^{\text{row}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}, \mathfrak{X}) = \min_{1\le m\le n} \left\{ \frac{m}{n} : \sup_{\mathfrak{Y}} \max_{i,j} \left| \log(\lambda_i(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{Y}})\lambda_j(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{Y}})) - \log(\lambda_i(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{X}})\lambda_j(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{X}})) \right| = \infty \right\}, \tag{17}$$

where the supremum $\sup_{\mathfrak{Y}}$ in (16) and (17) is taken over all possible samples of m $(p \times q)$ - matrices used to contaminate sample $\mathfrak{X}$, and $\lambda_j(\mathbf{A})$ denotes the jth largest eigenvalue of the symmetric matrix $\mathbf{A}$.

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

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

with $n/2 \le h \le n$ and $h \ge d + 2$.

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.1. Theorem 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 $p \ge 2, q \ge 2$ 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 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.

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

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

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, Davies, and Jhun (1993) and Cator and Lopuhaä (2012) and is provided in Supplement B.1. It shows that the consistency factor of the MCD estimator and the MMCD estimator must coincide, and therefore, we use the consistency factor

$$c(\alpha) = \frac{\alpha}{F_{\chi^2_{pq+2}}(\chi^2_{\alpha;pq})} \tag{18}$$

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-

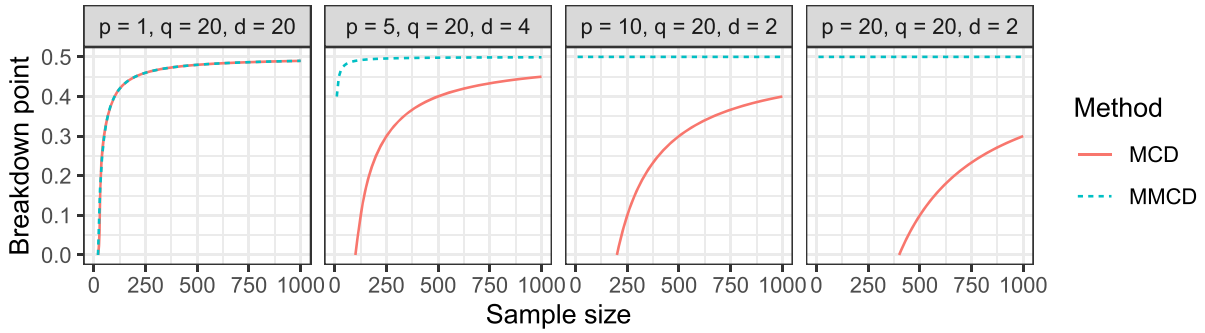


Figure 1. Comparison of the maximum breakdown point of the MMCD estimators for matrix-variate data with $p = 1, 5, 10, 20$ rows and $q = 20$ columns, and the MCD estimator applied to the vectorized data. When $p = 1$, both estimators and their breakdown points coincide. However, increasing the number 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.

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 3.0.1. Note first that for $h = n$, the corresponding MMCD estimators coincide with the ones defined in (3)-(5). Therefore, a simple, yet not discussed in the literature, consequence of Theorem 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 $(\boldsymbol{M}, \boldsymbol{\Sigma}^{\text{row}}, \boldsymbol{\Sigma}^{\text{row}})$ in the semi-parametric, matrix elliptical family. One should also bear in mind that the need for scaling arises from the trimmed nature of the covariance estimator.

Reweighted MMCD - improving efficiency. The *raw* MMCD estimators are most robust when about half of the observations are trimmed, that is, $h = \lfloor (n+d+2)/2 \rfloor$. 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) and 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 3.0.1. Let $\mathfrak{X}$ be a collection of n iid 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 $\hat{\boldsymbol{M}}$, $\hat{\boldsymbol{\Sigma}}^{\text{row}}$, and $\hat{\boldsymbol{\Sigma}}^{\text{col}}$ denote the *raw* MMCD estimators as in Definition 2.0.1. The *reweighted* MMCD estimators are given by

$$\tilde{\boldsymbol{M}} = \frac{1}{\sum_{i=1}^n w(\text{MMD}(\boldsymbol{X}_i))} \sum_{i=1}^n w(\text{MMD}(\boldsymbol{X}_i))\boldsymbol{X}_i, \tag{19}$$

$$\tilde{\boldsymbol{\Sigma}}^{\text{row}} = \frac{1}{q \sum_{i=1}^n w(\text{MMD}(\boldsymbol{X}_i))} \sum_{i=1}^n w(\text{MMD}(\boldsymbol{X}_i))(\boldsymbol{X}_i - \tilde{\boldsymbol{M}})\tilde{\boldsymbol{\Omega}}^{\text{col}}(\boldsymbol{X}_i - \tilde{\boldsymbol{M}})', \quad \text{and} \tag{20}$$

$$\tilde{\boldsymbol{\Sigma}}^{\text{col}} = \frac{1}{p \sum_{i=1}^n w(\text{MMD}(\boldsymbol{X}_i))} \sum_{i=1}^n w(\text{MMD}(\boldsymbol{X}_i))(\boldsymbol{X}_i - \tilde{\boldsymbol{M}})'\tilde{\boldsymbol{\Omega}}^{\text{row}}(\boldsymbol{X}_i - \tilde{\boldsymbol{M}}), \tag{21}$$

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

The following theorem shows that the *reweighted* MMCD estimator preserves the breakdown point of the original estimator. The simulations presented in Section 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 3.0.3. Let $\mathfrak{X}$ be a collection of n iid 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 $\hat{\boldsymbol{M}}_{\mathfrak{X}}$, $\hat{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{X}}$, and $\hat{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{X}}$ denote the *raw* MMCD estimators as in Definition 2.0.1 with breakdown points

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

and let $\tilde{\boldsymbol{M}}_{\mathfrak{X}}$, $\tilde{\boldsymbol{\Sigma}}^{\text{row}}_{\mathfrak{X}}$, and $\tilde{\boldsymbol{\Sigma}}^{\text{col}}_{\mathfrak{X}}$ denote the *reweighted* estimators as in Definition 3.0.1. Then,

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

A proof is given in Supplement B.1. 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

$$w(\text{MMD}^2(\boldsymbol{X}_i)) := \begin{cases} 1 & \text{if} \quad i \in H \vee \text{MMD}^2(\boldsymbol{X}_i) < \chi^2_{pq;0.975} \\ 0 & \text{otherwise} \end{cases}. \tag{22}$$

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.

4. Algorithm

Rousseeuw and 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.

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 (11) and (12), 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 $\mathfrak{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, 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 $(\hat{\boldsymbol{M}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}})$ are computed based on the observations in the subset H_{old} using the flip-flop algorithm, which is nondecreasing in likelihood. Next, compute the squared Mahalanobis distances $d_i^2(H_{\text{old}}) := \text{MMD}^2(\boldsymbol{X}_i, \hat{\boldsymbol{M}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}})$ for all $i = 1, \ldots, n$. In Proposition 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 $d^2_{\pi(1)}(H_{\text{old}}) \leq \cdots \leq d^2_{\pi(n)}(H_{\text{old}})$, 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 nondecreasing in likelihood. Finally, the estimators are updated using the flip-flop algorithm based on the observations in the subset H_{new}, resulting in estimators $(\hat{\boldsymbol{M}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}}, \hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}})$, increasing the likelihood once more, and it follows that

$$p \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{new}}})) + q \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{new}}})) \leq p \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{col}}_{H_{\text{old}}})) + q \ln(\det(\hat{\boldsymbol{\Sigma}}^{\text{row}}_{H_{\text{old}}})). \quad (23)$$

By repeatedly applying such C-steps, we can decrease the covariance determinant in subsequent iterations as in (23). This results in a decreasing and nonnegative 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 1 in Supplement C.

4.2. The MMCD Algorithm

The MMCD algorithm is a matrix-variate extension of the Fast-MCD procedure of Rousseeuw and 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 2 is given in Supplement C.

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, Jansson, and Stoica (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 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 C.1, 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 (18) to achieve consistency at the normal model as outlined in Theorem 3.0.2. Based on those rescaled raw MMCD estimators, the reweighted estimators described in Definition 3.0.1 are computed using the weights given in (22). 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 Mahalanobis 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 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.

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

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

for $i \in \{1, \ldots, n\}$ and the MMCD estimators $\hat{\boldsymbol{M}}$, $\hat{\boldsymbol{\Sigma}}^{\mathrm{row}}$, and $\hat{\boldsymbol{\Sigma}}^{\mathrm{col}}$. The consistency of MMCD estimators for matrix-normal distributed data implies that the robust estimate of the squared matrix Mahalanobis distance, $\mathrm{MMD}^2(\boldsymbol{X}_i; \hat{\boldsymbol{M}}, \hat{\boldsymbol{\Sigma}}^{\mathrm{row}}, \hat{\boldsymbol{\Sigma}}^{\mathrm{col}})$, is asymptotically distributed as $\chi^2(pq)$. This asymptotic behavior justifies the use of this cutoff for large samples, assuming Gaussianity. 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).

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, that is,

$$\begin{aligned}\phi_k(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{S \subseteq P \setminus \{k\}} \frac{|S|!(p - |S| - 1)!}{p!} \Delta_k \,\mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \\ &= (x_k - \mu_k) \sum_{j=1}^{p} (x_j - \mu_j)\omega_{jk}, \end{aligned} \tag{24}$$

with marginal contributions

$$\Delta_k \,\mathrm{MD}^2(\hat{\boldsymbol{x}}^S) := \mathrm{MD}^2(\hat{\boldsymbol{x}}^{S \cup \{k\}}) - \mathrm{MD}^2(\hat{\boldsymbol{x}}^S) \quad \text{and}$$
$$\hat{x}_j^S := \begin{cases} x_j & \text{if } j \in S \\ \mu_j & \text{if } j \notin S \end{cases} \tag{25}$$

as the components of $\hat{\boldsymbol{x}}^S$. Here, $\mathrm{MD}^2(\mathbf{x}) = \mathrm{MD}^2(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Mahalanobis distance of $\mathbf{x}$ from the mean $\boldsymbol{\mu}$ with respect to the covariance $\boldsymbol{\Sigma}$, and ω_{jk} is the element (j, k) of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. For $k \in \{1, \ldots, p\}$ and $S \subset P \setminus \{k\}$, a positive value of the marginal contribution $\Delta_k \,\mathrm{MD}^2(\hat{\boldsymbol{x}}^S)$ indicates that replacing the kth variable with its mean, in addition to replacing the variables in $P \setminus S$, reduces the corresponding squared Mahalanobis distance. Conversely, $\Delta_k \,\mathrm{MD}^2(\hat{\boldsymbol{x}}^S) < 0$ implies that this replacement increases the squared Mahalanobis distance. This outcome reflects the use of the overall mean as a replacement, rather than the conditional mean, which impacts the contribution of the replaced variables, see Mayrhofer and Filzmoser (2023) for more details.

Since $\boldsymbol{\phi}(\boldsymbol{x})$ is based on the Shapley value, it is the only decomposition of the squared Mahalanobis distance based on (25) 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, $\sum_{j=1}^{p} \phi_j(\boldsymbol{x}) = \mathrm{MD}^2(\boldsymbol{x})$.
- *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 $\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}} \in \mathrm{PDS}(p)$ be two matrices. If

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

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

The coordinate $\phi_k(\boldsymbol{x})$ of the Shapley value is the average marginal contribution of the kth variable to the squared Mahalanobis distance and is obtained by averaging over all marginal outlyingness contributions $\Delta_k \,\mathrm{MD}^2(\hat{\boldsymbol{x}}^S)$ across all possible subsets $S \subseteq P \setminus \{k\}$. This reflects the average effect of replacing the kth variable with its mean on the squared Mahalanobis distance. Although this suggests an exponential computational complexity, which becomes costly, especially if p is large, the second equality in (24) reveals just linear complexity; for a proof we refer to Mayrhofer and Filzmoser (2023). Equation (24) 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 (24) shows that a coordinate $\phi_k(\boldsymbol{x})$ of the Shapley value only considers the contributions that are associated with the kth variable.

5.2. Shapley Value 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 (7). Let $\boldsymbol{X} \in \mathbb{R}^{p \times q}$ be a matrix-variate sample with mean $\boldsymbol{M} \in \mathbb{R}^{p \times q}$ and covariance matrices $\boldsymbol{\Sigma}^{\mathrm{row}} \in \mathrm{PDS}(p)$ and $\boldsymbol{\Sigma}^{\mathrm{col}} \in \mathrm{PDS}(q)$. The pq-dimensional vectorized observation is denoted as $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{X})$, with mean $\boldsymbol{\mu} = \mathrm{vec}(\boldsymbol{M})$ and covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\mathrm{col}} \otimes \boldsymbol{\Sigma}^{\mathrm{row}}$. Based on (24), 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 $j = 1, \ldots, p$ and $k = 1, \ldots, q$. Using matrix operations, we can efficiently compute the $p \times q$ matrix containing the cellwise Shapley values $\phi_{jk}(\boldsymbol{X})$ as

$$\boldsymbol{\Phi}(\boldsymbol{X}) = (\boldsymbol{X} - \boldsymbol{M}) \circ \boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X} - \boldsymbol{M})\boldsymbol{\Omega}^{\text{col}} \in \mathbb{R}^{p\times q}, \tag{26}$$

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

Next, we discuss how matrix affine transformations as in (15) affect the cellwise Shapley values for matrix-variate data.

Proposition 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, that is, $\boldsymbol{\Phi}(\boldsymbol{AXB}) \neq \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{X})\boldsymbol{B}$ for general positive definite $\boldsymbol{A}$ and $\boldsymbol{B}$. However, they are

(a) shift invariant, that is, $\boldsymbol{\Phi}(\boldsymbol{X} + \boldsymbol{C}) = \boldsymbol{\Phi}(\boldsymbol{X})$,
(b) scale invariant, that is, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are scaling matrices, thus, diagonal matrices with nonzero entries, then $\boldsymbol{\Phi}(\boldsymbol{AXB}) = \boldsymbol{\Phi}(\boldsymbol{X})$,
(c) permutation equivariant, that is, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are permutation matrices, then $\boldsymbol{\Phi}(\boldsymbol{AXB}) = \boldsymbol{A}\boldsymbol{\Phi}(\boldsymbol{X})\boldsymbol{B}$, and

The proofs are given in Supplement D. 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 $P = \{1, \ldots, p\}$, and $S \subseteq P \setminus \{j\}$. The rowwise marginal contributions to the matrix Mahalanobis distance are defined as

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

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

Proposition 5.2.2. The jth coordinate of the rowwise Shapley value is given by

$$\phi_{j\cdot}(\boldsymbol{X}) := \sum_{S\subseteq P\setminus\{j\}} \frac{|S|!(p - |S| - 1)!}{p!}\Delta_j \text{MMD}(\hat{\boldsymbol{X}}^S) \tag{27}$$
$$= \sum_{i=1}^{p}\sum_{k=1}^{q}\sum_{l=1}^{q}(x_{jl} - m_{jl})(x_{ik} - m_{ik})\omega_{ij}^{\text{row}}\omega_{kl}^{\text{col}}$$
$$= \sum_{k=1}^{q}\phi_{jk}(\boldsymbol{X}). \tag{28}$$

A proof for (28) can be found in supplement D. 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

$$\boldsymbol{\phi}_{\text{row}}(\boldsymbol{X}) = \text{diag}(\boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X} - \boldsymbol{M})\boldsymbol{\Omega}^{\text{col}}(\boldsymbol{X} - \boldsymbol{M})') \in \mathbb{R}^p \quad \text{and} \tag{29}$$

$$\boldsymbol{\phi}_{\text{col}}(\boldsymbol{X}) = \text{diag}((\boldsymbol{X} - \boldsymbol{M})'\boldsymbol{\Omega}^{\text{row}}(\boldsymbol{X} - \boldsymbol{M})\boldsymbol{\Omega}^{\text{col}}) \in \mathbb{R}^q, \tag{30}$$

respectively. The properties listed in Proposition 5.2.1 also apply in this setting.

6. Simulations

In this section, we present simulation studies designed to rigorously evaluate the performance of the MMCD estimators, validate their theoretical properties, and compare their efficiency against ML estimators.

Among the four robust covariance estimators for matrix-valued data mentioned in Section 1, we could only find implementations for the methods proposed by Thompson et al. (2020) and Zhang, Shen, and Kong (2022). While our focus is on robustness under the Tukey-Huber contamination model (Tukey 1960; Huber 1964), which assumes a mixture of clean and contaminated observations, their methods target robustness in a heavy-tailed data setting. Consequently, we expect that these estimators cannot compete with the MMCD estimators under the Tukey-Huber contamination model, which is confirmed by simulations included in the supplementary materials E.

In the following, we compare the efficiency of the raw and reweighted MMCD estimators under the normal model without contamination to confirm that the reweighting step improves efficiency. Moreover, we provide 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 regions depict one standard error interval.

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. (2015), denoted $\boldsymbol{\Sigma}^{\text{row}} =$

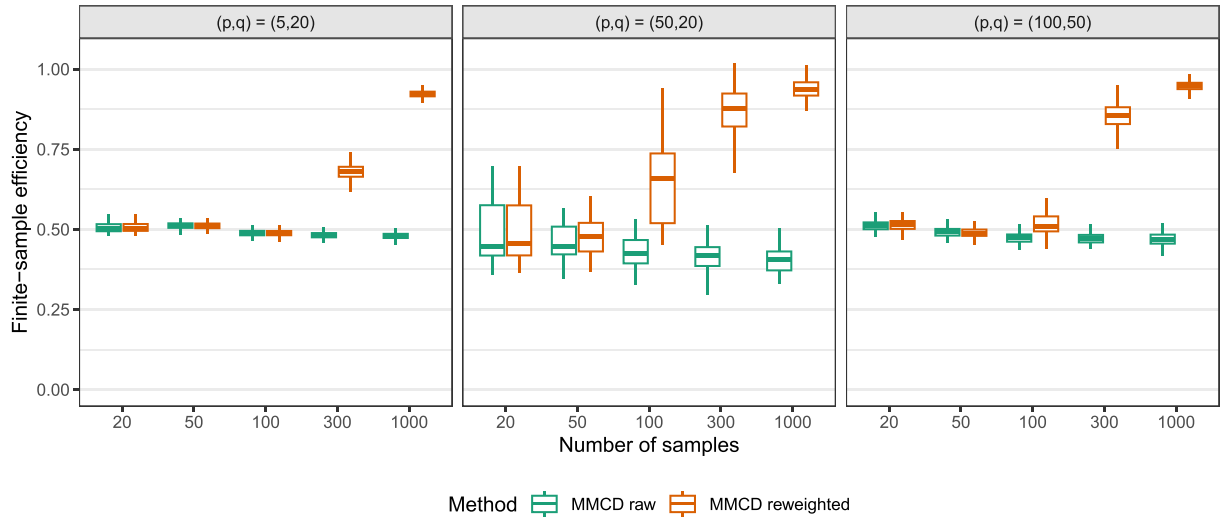


Figure 2. Comparison of the finite-sample efficiency of raw and reweighted MMCD.

$\mathbf{\Sigma}^{\text{rnd}} \in \text{PDS}(p)$, 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 $\mathcal{MN}(\boldsymbol{M}, \mathbf{\Sigma}^{\text{row}}, \mathbf{\Sigma}^{\text{col}})$, which is given by

$$\begin{aligned} D(\hat{\mathbf{\Sigma}}^{\text{row}}, \hat{\mathbf{\Sigma}}^{\text{col}}) = &\operatorname{tr}(\mathbf{\Omega}^{\text{row}}\hat{\mathbf{\Sigma}}^{\text{row}}) \operatorname{tr}(\mathbf{\Omega}^{\text{col}}\hat{\mathbf{\Sigma}}^{\text{col}}) \\ &- q \log(\det(\mathbf{\Omega}^{\text{row}}\hat{\mathbf{\Sigma}}^{\text{row}})) \\ &- p \log(\det(\mathbf{\Omega}^{\text{col}}\hat{\mathbf{\Sigma}}^{\text{col}})) - pq, \end{aligned} \tag{31}$$

with $\mathbf{\Omega}^{\text{row}} = (\mathbf{\Sigma}^{\text{row}})^{-1}$ and $\mathbf{\Omega}^{\text{col}} = (\mathbf{\Sigma}^{\text{col}})^{-1}$. As shown in Figure 2, 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.

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 (31) 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, Rousseeuw, and Verdonck 2012) when $pq > 300$. Since the MCD estimator requires $n > pq$, it is only computed for those settings.

Figure 3 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, Van Aelst, and Willems (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.

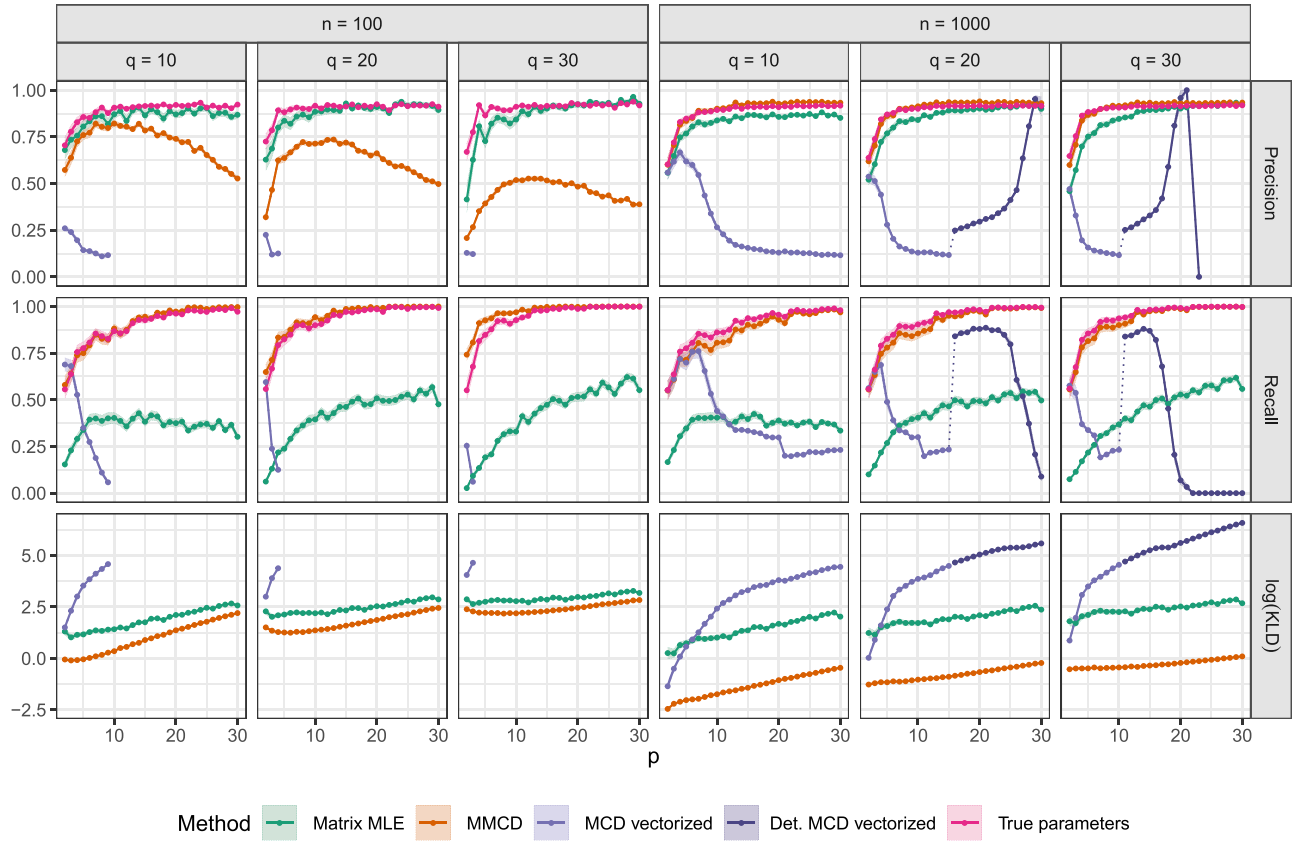


Figure 3. Comparison of precision, recall, and KL divergence for ML and MMCD estimators, (deterministic) MCD estimators with vectorized data, and true parameters as a benchmark for outlier detection for simulated data from a matrix normal distribution with 10% contamination.

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 4 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, that is, the KL divergence of the matrix MLEs is highest for block contamination followed by cell and shift contamination.

The supplementary materials E 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). Finally, we include a comparison to the distribution-free estimators Zhang, Shen, and Kong (2022) for banded row and column covariance matrices, 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.

7. Examples

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 (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

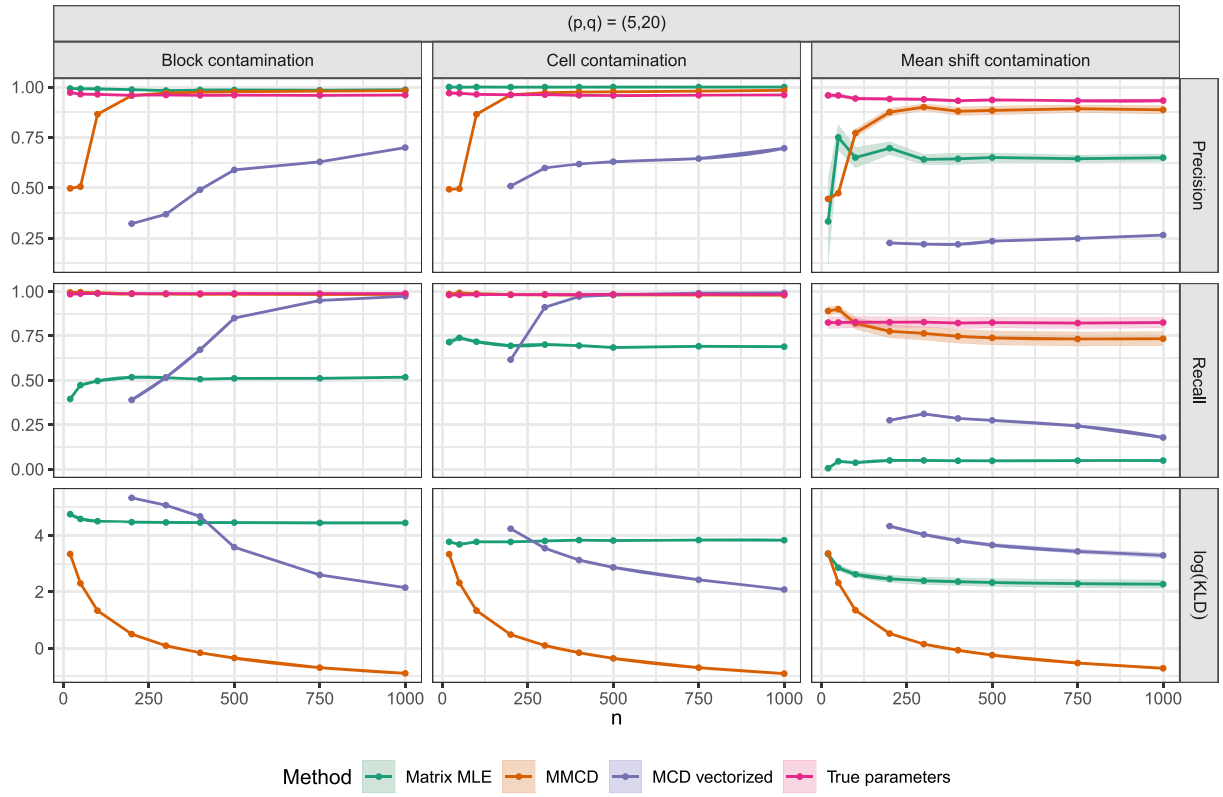


Figure 4. Precision, recall, and logarithm of KL divergence comparing block, cell, and sample contamination.

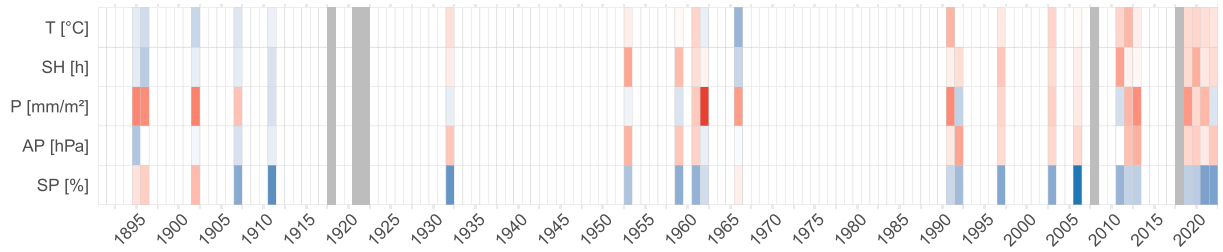


Figure 5. Yearly outlyingness contributions for the glacier weather data. Regular years are white, and years that contain missing data are gray. Outliers are colored as follows: blue for "below average", red for "above average", and color intensity proportional to the rowwise Shapley value.

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 5 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, that is, 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 6, 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 from Figure 5. 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.

7.2. Darwin Data

We consider the DARWIN (Diagnosis AlzheimeR WIth haNdwriting) (Cilia et al. 2022) data set 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:

Figure 6. 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 5.

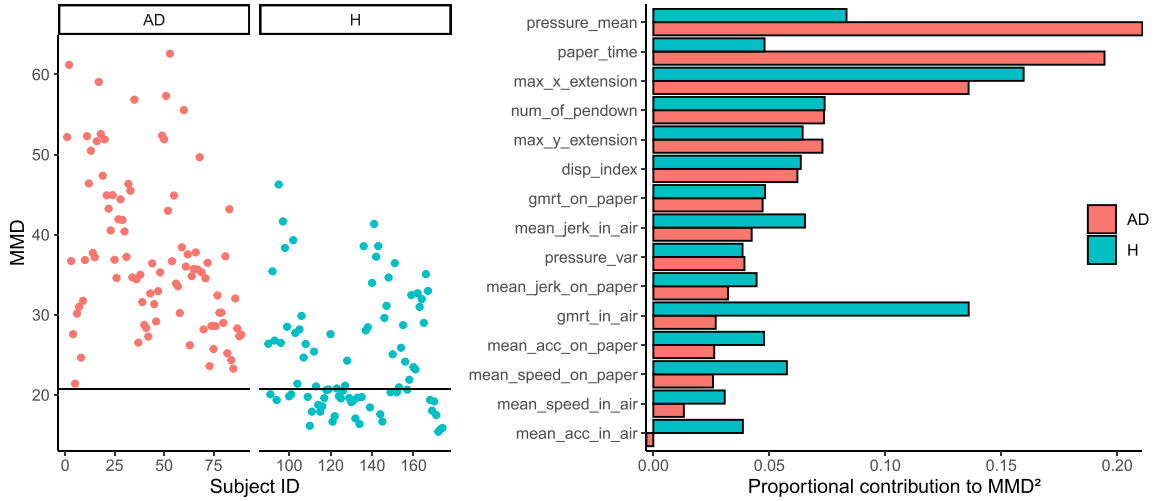


Figure 7. Plot of robust MMD based on MMCD estimators for the Darwin data on the left, and average proportional rowwise Shapley values for the H and AD subjects on the right.

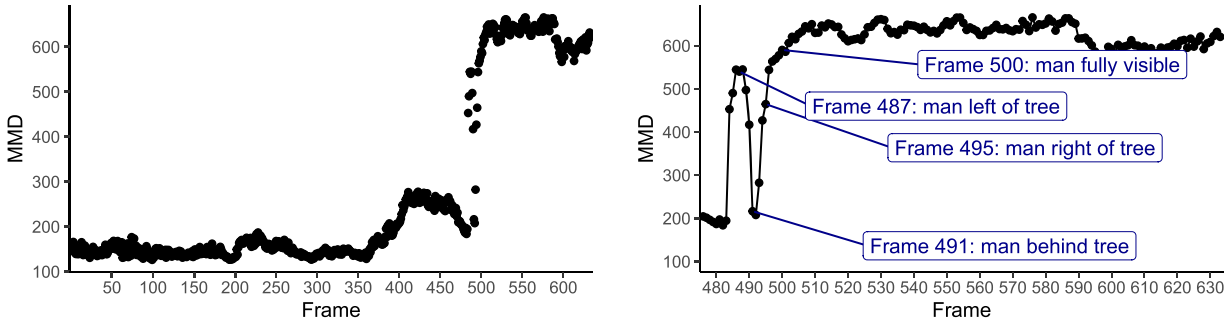
Total Time, Air Time, Paper Time, Mean Speed on paper, Mean Speed in air, Mean Acceleration on paper, Mean Acceleration in air, Mean Jerk on paper, Mean Jerk in air, Pressure Mean, Pressure Variance, Generalization of the Mean Relative Tremor (GMRT) on paper, GMTR in air, Mean GMRT, Pendowns Number, Max X Extension, Max Y Extension, and Dispersion 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 7 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 7, 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.

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 8. 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 8. Selected frames of the video data.



(a) Frames 1 to 633. (b) Frames 471 to 633.

Figure 9. Plot of robust MMD based on MMCD estimators for the video data.

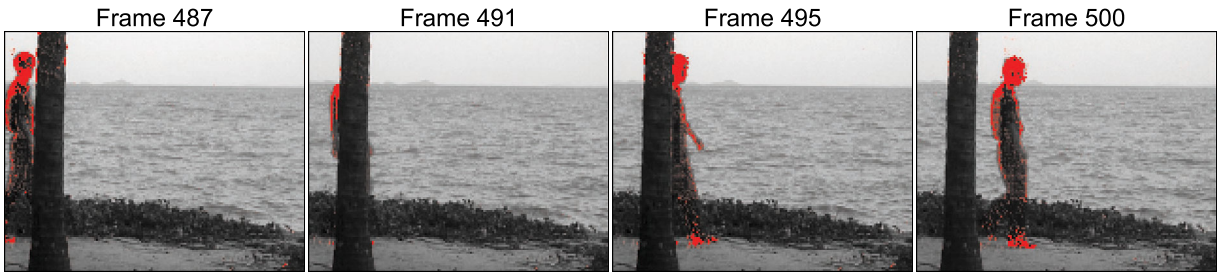


Figure 10. Outlyingness scores based on cellwise Shapley values are shown in red, where darker colors indicate higher outlyingness contributions, and the grayscale video frames are displayed in the background.

For our analysis, 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 9. The plot on the left shows the robust MMDs for all 633 frames, and the one on the right for frames 471–633 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 10 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 9(b). It is interesting to see a certain increase in the MMD in Figure 9(a) 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.

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 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, Van Aelst, and Willems (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.

Supplementary Materials

The supplementary material consists of five sections. Section A covers preliminaries for matrix-variate data. Section B contains the proofs for all results concerning the MMCD estimators. Section C includes pseudo-code and further details regarding the C-step and MMCD procedures. Section D contains the proofs related to outlier explanations. Section E provides additional simulation results.

Software and data availability: The `R` package `robustmatrix` includes a parallelized `C++` implementation of the MMCD algorithm (Mayrhofer, Radojičić, and Filzmoser 2024). The vignette `MMCD_examples` and the `R` code in the supplementary material replicate the examples presented in this paper.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This work was supported by the AI4CSM project and has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 101007326; and 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. This work was supported by the Austrian Science Fund (FWF), project number I 5799-N. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

References

Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015), "Robust Estimation of Multivariate Location and Scatter in the Presence of Cellwise and Casewise Contamination," *Test*, 24, 441–461. [523]

Alqallaf, F., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009), "Propagation of Outliers in Multivariate Data," *The Annals of Statistics*, 37, 311–331. [523]

Boudt, K., Rousseeuw, P. J., Vanduffel, S., and Verdonck, T. (2020), "The Minimum Regularized Covariance Determinant Estimator," *Statistics and Computing*, 30, 113–128. [529]

Butler, R., Davies, P., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics*, 21, 1385–1400. [519]

Cator, E. A., and Lopuhaä, H. P. (2012), "Central Limit Theorem and Influence Function for the MCD Estimators at General Multivariate Distributions," *Bernoulli*, 18, 520–551. [519]

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. [526,527]

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. [527]

Croux, C., and Haesbroeck, G. (1999), "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis*, 71, 161–190. [519]

Davies, P. L. (1987), "Asymptotic Behaviour of S-estimates of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics*, 15, 1269–1292. [519]

Dawid, A. P. (1981), "Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application." *Biometrika*, 68, 265–274. [516]

Dutilleul, P. (1999), "The MLE Algorithm for the Matrix Normal Distribution," *Journal of Statistical Computation and Simulation*, 64, 105–123. [517,521]

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. [517]

Gupta, A., and Nagar, D. (1999), *Matrix Variate Distributions*. Monographs and Surveys in Pure and Applied Mathematics, New York: Taylor & Francis. [518]

Gupta, A. K., and Varga, T. (2012), *Elliptically Contoured Models in Statistics* (Vol. 240), Dordrecht: Springer. [516,518]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101. [523]

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012), "A Deterministic Algorithm for Robust Location and Scatter," *Journal of Computational and Graphical Statistics*, 21, 618–637. [524]

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. [517]

Ledoit, O., and Wolf, M. (2004), "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices," *Journal of Multivariate Analysis*, 88, 365–411. [529]

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, 1459–1472. [527]

Lopuhaa, H. P., and Rousseeuw, P. J. (1991), "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices," *The Annals of Statistics*, 19, 229–248. [519,520]

Lu, N., and Zimmerman, D. L. (2005), "The Likelihood Ratio Test for a Separable Covariance Matrix," *Statistics & Probability Letters*, 73, 449–457. [517]

Lundberg, S. M., and Lee, S.-I. (2017), "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, pp. 4765–4774, Curran Associates, Inc. [517]

Mahalanobis, P. C. (1936), "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, 2, 49–55. [517]

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. [529]

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2019), *Robust Statistics: Theory and Methods (with R)*, Hoboken, NJ: Wiley. [518,519,520,522]

Mayrhofer, M., and Filzmoser, P. (2023), "Multivariate Outlier Explanations Using Shapley Values and Mahalanobis Distances," *Econometrics and Statistics*. [517,522,529]

Mayrhofer, M., Radojičić, U., Filzmoser, P. (2024), robustmatrix: Robust Matrix-Variate Parameter Estimation R Package Version 0.1.3. Available at https://CRAN.R-project.org/package=robustmatrix.

Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2007), "Robust Fitting of Mixtures Using the Trimmed Likelihood Estimator," *Computational Statistics & Data Analysis*, 52, 299–308. [517]

Pison, G., Van Aelst, S., and Willems, G. (2002), "Small Sample Corrections for LTS and MCD," *Metrika*, 55, 111–123. [524,529]

Raymaekers, J., and Rousseeuw, P. J. (2023), "The Cellwise Minimum Covariance Determinant Estimator," *Journal of the American Statistical Association*, 119, 2610–2621. [517,529]

Rousseeuw, P. (1985), "Multivariate Estimation with High Breakdown Point," *Mathematical Statistics and Applications Vol. B*, eds. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht: Reidel Publishing Company, pp. 283–297. [517,519]

Rousseeuw, P. J., and Driessen, K. V. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223. [517,521,529]

Shapley, L. S. (1953), "A Value for n-person Games," *Contributions to the Theory of Games*, 2, 307–317. [517]

Soloveychik, I., and Trushin, D. (2016), "Gaussian and Robust Kronecker Product Covariance Estimation: Existence and Uniqueness," *Journal of Multivariate Analysis*, 149, 92–113. [517,518]

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, 3576–3590. [517]

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, 668–674. [517,523,525]

Tukey, J. W. (1960), "A Survey of Sampling From Contaminated Distributions," in *Contributions to Probability and Statistics*, ed. I. Oklin, pp. 448–485, Redwood City, CA: Stanford University Press. [523]

Tyler, D. E. (1987), "A Distribution-Free m-estimator of Multivariate Scatter," *The Annals of Statistics*, 15, 234–251. [517]

Werner, K., Jansson, M., and Stoica, P. (2008), "On Estimation of Covariance Matrices with Kronecker Product Structure," *IEEE Transactions on Signal Processing*, 56, 478–491. [521]

Zhang, Y., Shen, W., and Kong, D. (2022), "Covariance Estimation for Matrix-Valued Data," *Journal of the American Statistical Association*, 118, 2620–2631. [517,523,525]