

Supplementary Material for Robust covariance estimation and explainable outlier detection for matrix-valued data

Marcus Mayrhofer¹, Una Radojičić², and Peter Filzmoser³

^{1,2,3}Institute of Statistics and Mathematical Methods in Economics, TU Wien

A Preliminaries

Consider an i.i.d. sample $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n) \in \mathbb{R}^{n \times p \times q}$, with $\mathbf{X}_i \sim \mathcal{MN}(\mathbf{M}, \Sigma^{\text{row}}, \Sigma^{\text{col}})$. Due to the factored covariance structure of matrix normal data, the rowwise and columnwise covariance matrices Σ^{row} and Σ^{col} are only identified up to a multiplicative constant $\kappa \neq 0$, since replacing Σ^{row} by $\kappa \Sigma^{\text{row}}$ and Σ^{col} by $1/\kappa \Sigma^{\text{col}}$ does not change the pdf of \mathbf{X} . While the Kronecker product $\Sigma^{\text{col}} \otimes \Sigma^{\text{row}}$ can be uniquely identified, the issue of trivial non-uniqueness of Σ^{row} and Σ^{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 Σ^{col} is set to one. This implies that the uniqueness of $\Sigma^{\text{col}} \otimes \Sigma^{\text{row}}$ is equivalent to the uniqueness of Σ^{col} and Σ^{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)-(5) for mean and covariance estimation, it is also possible to consider the vectorized samples $\mathbf{x}_i = \text{vec}(\mathbf{X}_i) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $i = 1, \dots, n$, where $\boldsymbol{\mu} = \text{vec}(\mathbf{M})$ and $\Sigma = \Sigma^{\text{col}} \otimes \Sigma^{\text{row}}$ denote the mean and covariance matrix, respectively. Then the maximum likelihood estimators for mean and covariance are given by

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

respectively. The computation of the MLEs for matrix-variate samples based on Equations (3)-(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 (A.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 $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})' \in \mathbb{R}^p$ denote the i -th observation of a data set in the multivariate setting, where $i = 1, \dots, n$. 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 Driessen, 1999), which internally sorts the observations based on their Mahalanobis distances. For an observation \mathbf{x}_i from a population with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \text{PDS}(p)$ it is given by

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

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 \mathbf{X} as

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

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

B Proofs of Section 2

Proof of Proposition 2.0.1. In optimization problem (8) we want to maximize

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

subject to $w_i \in \{0, 1\}$ for all $i = 1, \dots, n$ and $\sum_{i=1}^n w_i = h$. In Equation (B.1), $\text{MMD}^2(\mathbf{X}_i)$ is defined as in Equation (7).

For any random h -subset H (or equivalently the corresponding set of weights \mathbf{w}) the constrained MLEs for \mathbf{M} , Σ^{row} , and Σ^{col} of Equation (B.1) can be written as:

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

Using those estimators to compute the sum of the Mahalanobis distances $\text{MMD}^2(\mathbf{X}_i)$ in Equation (B.1) we obtain

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

Thus, the terms in the second row of Equation (B.1) are all constant, and it is sufficient to maximize only the term in the first row, which contains the (negative) determinant of Equation (9). \square

B.1 Properties of MMCD estimators

Proof of Lemma 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

$$\begin{aligned}\det(\hat{\Sigma}_{\mathbf{3}_H}^{\text{col}} \otimes \hat{\Sigma}_{\mathbf{3}_H}^{\text{row}}) &= \det \left((\mathbf{B}' \hat{\Sigma}_{\mathbf{x}_H}^{\text{col}} \mathbf{B}) \otimes (\mathbf{A} \hat{\Sigma}_{\mathbf{x}_H}^{\text{row}} \mathbf{A}') \right) \\ &= \left[\det(\mathbf{B}' \hat{\Sigma}_{\mathbf{x}_H}^{\text{col}} \mathbf{B}) \right]^p \left[\det(\mathbf{A} \hat{\Sigma}_{\mathbf{x}_H}^{\text{row}} \mathbf{A}') \right]^q \\ &= \left[\det(\mathbf{B}') \det(\hat{\Sigma}_{\mathbf{x}_H}^{\text{col}}) \det(\mathbf{B}) \right]^p \left[\det(\mathbf{A}) \det(\hat{\Sigma}_{\mathbf{x}_H}^{\text{row}}) \det(\mathbf{A}') \right]^q \\ &= 4 \det(\mathbf{B})^p \det(\mathbf{A})^q \det(\hat{\Sigma}_{\mathbf{x}_H}^{\text{col}})^p \det(\hat{\Sigma}_{\mathbf{x}_H}^{\text{row}})^q.\end{aligned}$$

Since $4 \det(\mathbf{B})^p \det(\mathbf{A})^q$ 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 $(\hat{\mathbf{M}}_{\mathbf{Z}}, \hat{\Sigma}_{\mathbf{Z}}^{\text{row}}, \hat{\Sigma}_{\mathbf{Z}}^{\text{col}})$ are matrix affine equivariant estimators of location and covariance of the transformed sample \mathbf{Z} , then

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

□

The proofs of Theorems 3.0.1 and 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

$$E(\mathbf{T}, \mathbf{U}, \mathbf{V}) = \{\mathbf{X} : \text{tr}(\mathbf{V}^{-1}(\mathbf{X} - \mathbf{T})'\mathbf{U}^{-1}(\mathbf{X} - \mathbf{T})) \leq 1\} \quad (\text{B.2})$$

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

$$\text{vol}(E(\mathbf{T}, \mathbf{U}, \mathbf{V})) = \underbrace{\frac{\pi^{pq/2}}{\Gamma(pq/2 + 1)}}_{=: \beta_{pq}} \prod_{i=1}^p \prod_{j=1}^q \sqrt{\lambda_i(\mathbf{U})\lambda_j(\mathbf{V})} = \beta_{pq} \underbrace{\det(\mathbf{U})^{q/2} \det(\mathbf{V})^{p/2}}_{=: \det(E(\mathbf{T}, \mathbf{U}, \mathbf{V}))}, \quad (\text{B.3})$$

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

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

$$\lambda_1(\mathbf{A}) = \sup_{\mathbf{z} \in \mathbb{R}^p} \frac{\mathbf{z}'\mathbf{A}\mathbf{z}}{\mathbf{z}'\mathbf{z}} \quad \text{and} \quad \lambda_n(\mathbf{A}) = \inf_{\mathbf{z} \in \mathbb{R}^p} \frac{\mathbf{z}'\mathbf{A}\mathbf{z}}{\mathbf{z}'\mathbf{z}}. \quad (\text{B.4})$$

Consider another symmetric nonnegative definite $p \times p$ matrix \mathbf{B} , then using Equation (B.4) we get that

$$\lambda_1(\mathbf{A} + \mathbf{B}) \leq \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B}) \quad \text{and} \quad \lambda_p(\mathbf{A} + \mathbf{B}) \geq \lambda_p(\mathbf{A}) + \lambda_p(\mathbf{B}). \quad (\text{B.5})$$

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

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

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

$$\frac{1}{\lambda_1(\mathbf{A})} \leq \frac{\mathbf{x}'\mathbf{A}^{-1}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \Leftrightarrow \mathbf{x}'\mathbf{x} \leq \mathbf{x}'\mathbf{A}^{-1}\mathbf{x}\lambda_1(\mathbf{A}). \quad (\text{B.6})$$

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

$$\text{vec}(\mathbf{Z})'(\mathbf{B} \otimes \mathbf{A})\text{vec}(\mathbf{Z}) = \text{tr}(\mathbf{B}\mathbf{Z}'\mathbf{A}\mathbf{Z}),$$

as in Equation (7), which implies that

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

This leads us to the matrix-variate version of Equation (B.6), where for any matrix $\mathbf{X} \in \mathbb{R}^{p \times q}$

$$\|\mathbf{X}\|_F^2 = \text{tr}(\mathbf{X}'\mathbf{X}) \leq \text{tr}(\mathbf{B}^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})\lambda_1(\mathbf{A}, \mathbf{B}) = \text{tr}(\mathbf{B}^{-1}\mathbf{X}'\mathbf{A}^{-1}\mathbf{X})\lambda_1(\mathbf{A})\lambda_1(\mathbf{B}) \quad (\text{B.7})$$

Lemma B.1. Take $p, q \in \mathbb{N}$, $d = \lfloor p/q + q/p \rfloor$, $d+2 \leq s \leq pq$, and matrices $\mathbf{X}_1, \dots, \mathbf{X}_s \in \mathbb{R}^{p \times q}$ 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 $E(\mathbf{T}, \mathbf{U}, \mathbf{V})$ as in Equation (B.2), containing the matrices $\mathbf{X}_1, \dots, \mathbf{X}_s$, it holds that for every $C > 0$ there exists a constant $\alpha := \alpha(\mathbf{X}_1, \dots, \mathbf{X}_s) > 0$ only depending on $\mathbf{X}_1, \dots, \mathbf{X}_s$ such that $\|\mathbf{T}\|_F = \sqrt{\text{tr}(\mathbf{T}'\mathbf{T})} > \alpha$ implies $\det(E(\mathbf{T}, \mathbf{U}, \mathbf{V})) > C$, i.e.,

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

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

$$\sqrt{\lambda_i(\mathbf{U})\lambda_j(\mathbf{V})} > a. \quad (\text{B.8})$$

In Equation (B.8), $\lambda_i(\mathbf{U})$, $i \in \{1, \dots, p\}$, are the eigenvalues of \mathbf{U} , and $\lambda_j(\mathbf{V})$, $j \in \{1, \dots, q\}$, are the eigenvalues of \mathbf{V} . For any matrix \mathbf{X} contained in $E(\mathbf{T}, \mathbf{U}, \mathbf{V})$, Equations (B.2) and (B.7) imply that

$$\begin{aligned} \|\mathbf{X} - \mathbf{T}\|_F^2 &= \text{tr}((\mathbf{X} - \mathbf{T})'(\mathbf{X} - \mathbf{T})) \\ &\leq \text{tr}(\mathbf{V}^{-1}(\mathbf{X} - \mathbf{T})'\mathbf{U}^{-1}(\mathbf{X} - \mathbf{T}))\lambda_1(\mathbf{U})\lambda_1(\mathbf{V}) \\ &\leq \lambda_1(\mathbf{U})\lambda_1(\mathbf{V}). \end{aligned} \quad (\text{B.9})$$

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 $E(\mathbf{T}, \mathbf{U}, \mathbf{V})$, then Equation (B.9) implies that $\|\mathbf{T}\|_F^2 \leq \lambda_1(\mathbf{U})\lambda_1(\mathbf{V})$. Take $\alpha = C/(a^{pq-1})$, then we have that

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

and from Equation (B.8) it follows that

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

□

Proof of Theorem 3.0.1. We show that the breakdown points of the MMCD estimators of location and covariance defined in Equations (16) and (17), 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 $\varepsilon^*(\hat{\mathbf{M}}, \mathfrak{X}) = \varepsilon^*(\hat{\Sigma}^{\text{row}}, \hat{\Sigma}^{\text{col}}, \mathfrak{X}) \geq m/n$. 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 original 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 $E_{\max} = E(\mathbf{0}, c_{\max} \mathbf{I}, \mathbf{I})$ denote the smallest sphere that contains all samples in \mathfrak{X} , where c_{\max} is chosen accordingly.
- Let $E_h = E(\mathbf{0}, c_h \mathbf{I}, \mathbf{I})$ 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 $E_{\text{MMCD}} = E(\hat{\mathbf{M}}_{\mathfrak{Y}}, \hat{\Sigma}_{\mathfrak{Y}}^{\text{row}}, \hat{\Sigma}_{\mathfrak{Y}}^{\text{col}})$ denote the MMCD ellipsoid.

It follows that $\det(E_{\text{MMCD}}) \leq \det(E_h) \leq \det(E_{\max}) =: \alpha$, where for an ellipsoid $E = E(\mathbf{T}, \mathbf{U}, \mathbf{V})$, $\det(E)$ is defined in (B.3). 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.1 shows that there exists a constant $\alpha > 0$ that only depends on those $d + 2$ samples such that, if $\|\hat{\mathbf{M}}_{\mathfrak{Y}}\|_F > C$ it would imply $\det(E_{\text{MMCD}}) > \alpha$. As shown above, this is not possible, hence $\|\hat{\mathbf{M}}_{\mathfrak{Y}}\|_F \leq C$.

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 $\hat{\Sigma}_{\mathfrak{Y}}^{\text{row}}$ and $\hat{\Sigma}_{\mathfrak{Y}}^{\text{col}}$. 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 $(\hat{\mathbf{M}}_{\mathfrak{X}_T}, \hat{\Sigma}_{\mathfrak{X}_T}^{\text{row}}, \hat{\Sigma}_{\mathfrak{X}_T}^{\text{col}})$ of this subsample are almost surely positive definite. Let $E_T = E(\hat{\mathbf{M}}_{\mathfrak{X}_T}, \hat{\Sigma}_{\mathfrak{X}_T}^{\text{row}}, \hat{\Sigma}_{\mathfrak{X}_T}^{\text{col}})$ denote the corresponding ellipsoid which is the smallest ellipsoid, of the type $E = E(\mathbf{T}, \mathbf{U}, \mathbf{V})$ as in Equation (B.2), 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 $E_T \subseteq k E_{\text{MMCD}}$, implying that there exists a constant $\gamma > 0$

depending only on \mathfrak{X} , such that $\lambda_i(\hat{\Sigma}_{\mathfrak{Y}}^{\text{row}})\lambda_j(\hat{\Sigma}_{\mathfrak{Y}}^{\text{col}}) > \gamma, 1 \leq i \leq p, 1 \leq j \leq q$. Especially, $\lambda_p(\hat{\Sigma}_{\mathfrak{Y}}^{\text{row}})\lambda_q(\hat{\Sigma}_{\mathfrak{Y}}^{\text{col}}) > \gamma$. Since also $\det(E_{\text{MMCD}}) \leq \alpha$ there exists a constant $\delta > 0$, depending only on \mathfrak{X} such that $\lambda_i(\hat{\Sigma}_{\mathfrak{Y}}^{\text{row}})\lambda_j(\hat{\Sigma}_{\mathfrak{Y}}^{\text{col}}) < \delta, 1 \leq i \leq p, 1 \leq j \leq q$.

Next we show that $\varepsilon^*(\hat{\mathbf{M}}, \mathfrak{X}) = \varepsilon^*(\hat{\Sigma}^{\text{row}}, \hat{\Sigma}^{\text{col}}, \mathfrak{X}) \leq m/n$. 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 contaminated sample. Hence, $E_{\text{MMCD}} = E(\hat{\mathbf{M}}_{\mathfrak{Y}}, \hat{\Sigma}_{\mathfrak{Y}}^{\text{row}}, \hat{\Sigma}_{\mathfrak{Y}}^{\text{col}})$ also includes at least one contaminated sample. Let $\|\mathbf{X}\|_F \rightarrow \infty$ for all contaminated samples \mathbf{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 $E_{\text{MMCD}} = E(\hat{\mathbf{M}}_{\mathfrak{Y}}, \hat{\Sigma}_{\mathfrak{Y}}^{\text{row}}, \hat{\Sigma}_{\mathfrak{Y}}^{\text{col}})$ has zero determinant. This implies that at least one eigenvalue is zero, hence the MMCD location and covariance estimators break down. \square

Proof of Theorem 3.0.2. Let $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ be a sample of matrix-variate observations and $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\mathbf{x}_i = \text{vec}(\mathbf{X}_i)$, $i = 1, \dots, n$ its vectorized form. The MCD estimator can also be found as a solution to the following maximization problem:

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

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

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

subject to $w_1, \dots, w_n \in \{0, 1\}$, $\sum_{i=1}^n w_i = h$, $\hat{\mathbf{M}} \in \mathbb{R}^{p \times q}$, $\hat{\Sigma}^{\text{row}} \in \text{PDS}(p)$, $\hat{\Sigma}^{\text{col}} \in \text{PDS}(q)$; see Proposition 2.0.1.

Denote further $(\mathbf{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\Sigma}_{\text{MCD}})$ and $(\mathbf{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\Sigma}_{\text{MMCD}}^{\text{col}} \otimes \hat{\Sigma}_{\text{MMCD}}^{\text{row}})$ weights, mean and covariance estimators for the vectorized sample $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, based on MCD and MMCD, respectively, noting also that these estimators are implicit functions of the sample size n too. However, for the simplicity of the notation, we omit adding the additional subscript n everywhere, unless explicitly needed. As $\mathbf{X}_i \sim \mathcal{ME}(\mathbf{M}, \Sigma^{\text{row}}, \Sigma^{\text{col}}, g)$, then $\mathbf{x}_i \sim \mathcal{E}(\boldsymbol{\mu}, \Sigma^{\text{col}} \otimes \Sigma^{\text{row}}, g)$, with $\mathbb{E}(\mathbf{x}_i) = \boldsymbol{\mu} = \text{vec}(\mathbf{M})$, $\text{cov}(\mathbf{x}_i) = c_g \Sigma^{\text{col}} \otimes \Sigma^{\text{row}}$, where c_g is a distribution-specific scaling parameter; for more details see Theorem 2.11 in [Gupta and Varga \(2012\)](#). Moreover, the mean estimator $\hat{\boldsymbol{\mu}}_{\text{MCD}}$ and properly scaled covariance estimator $\hat{\Sigma}_{\text{MCD}}$ are strongly consistent for the population counterparts $\boldsymbol{\mu}$ and $\Sigma^{\text{col}} \otimes \Sigma^{\text{row}}$; see e.g. [Croux and Haesbroeck \(1999\)](#) and [Cator and Lopuhaä \(2012\)](#). Especially, this implies that for every $\delta > 0$ there exists $n_0 \in \mathbb{N}$ such that

$$\|\hat{\boldsymbol{\mu}}_{\text{MCD},n} - \boldsymbol{\mu}\| \stackrel{a.s.}{<} \delta, \quad \|\hat{\Sigma}_{\text{MCD},n} - \mathbf{A} \otimes \mathbf{B}\| \stackrel{a.s.}{<} \delta,$$

for some $\mathbf{A} \otimes \mathbf{B} \in \text{PDS}(p) \otimes \text{PDS}(q)$ and all $n \geq n_0$. In the following, we will drop *a.s.* superscript from (in)equality signs when it is clear from the context. For fixed weights \mathbf{w} , the log-likelihood function $l_{\cdot, \mathbf{w}}(\mathbf{x}_1, \dots, \mathbf{x}_n) : (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto l(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{w}, (\mathbf{x}_1, \dots, \mathbf{x}_n))$ is continuous in both $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Take now $\varepsilon > 0$. The continuity then implies that there exists $\delta > 0$ such that,

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

for $\|\hat{\boldsymbol{\mu}}_{\text{MCD}} - \boldsymbol{\mu}\| \stackrel{a.s.}{<} \delta$ and $\|\hat{\boldsymbol{\Sigma}}_{\text{MCD}} - \mathbf{A} \otimes \mathbf{B}\| \stackrel{a.s.}{<} \delta$. Moreover, the solution $(\mathbf{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MCD}})$ is optimal for $l(\cdot | (\mathbf{x}_1, \dots, \mathbf{x}_n))$, implying that

$$0 < l(\mathbf{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MCD}} | (\mathbf{x}_1, \dots, \mathbf{x}_n)) - l(\mathbf{w}_{\text{MCD}}, \boldsymbol{\mu}, \mathbf{A} \otimes \mathbf{B} | (\mathbf{x}_1, \dots, \mathbf{x}_n)) < \varepsilon. \quad (\text{B.10})$$

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

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

Denote further $\hat{\mathbf{S}}_{\text{MMCD}} = \frac{1}{h} \sum_{i=1}^n w_{\text{MMCD},i} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MMCD}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\text{MMCD}})'$ 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 $\hat{\mathbf{S}}_{\text{MMCD}}$ is optimal for l given fixed weights \mathbf{w}_{MMCD} ,

$$\begin{aligned} l(\mathbf{w}_{\text{MCD}}, \hat{\boldsymbol{\mu}}_{\text{MCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MCD}} | (\mathbf{x}_1, \dots, \mathbf{x}_n)) &> l(\mathbf{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\mathbf{S}}_{\text{MMCD}} | (\mathbf{x}_1, \dots, \mathbf{x}_n)) \\ &> l(\mathbf{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, \hat{\boldsymbol{\Sigma}}_{\text{MMCD}}^{\text{col}} \otimes \hat{\boldsymbol{\Sigma}}_{\text{MMCD}}^{\text{row}} | (\mathbf{x}_1, \dots, \mathbf{x}_n)) \\ &> l(\mathbf{w}_{\text{MCD}}, \boldsymbol{\mu}, \mathbf{A} \otimes \mathbf{B} | (\mathbf{x}_1, \dots, \mathbf{x}_n)). \end{aligned} \quad (\text{B.11})$$

(B.10) and (B.11) now give that

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

i.e., due to Proposition 2.0.1,

$$0 < \det(\hat{\mathbf{S}}_{\text{MMCD}}) - \det(\hat{\boldsymbol{\Sigma}}_{\text{MCD}}) < \varepsilon, \quad (\text{B.12})$$

for $\varepsilon = \varepsilon(n) > 0$, arbitrarily small ($\varepsilon(n) \rightarrow 0, n \rightarrow \infty$). As both $\hat{\boldsymbol{\Sigma}}_{\text{MCD}}$ and $\hat{\mathbf{S}}_{\text{MMCD}}$ 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 $(\mathbf{x}_1, \dots, \mathbf{x}_n)$) implies that

$$\hat{\boldsymbol{\mu}}_{\text{MMCD}} \xrightarrow{a.s.} \boldsymbol{\mu}, \quad \hat{\mathbf{S}}_{\text{MMCD}} \xrightarrow{a.s.} c(\alpha)^{-1} \boldsymbol{\Sigma}^{\text{col}} \otimes \boldsymbol{\Sigma}^{\text{row}}, \quad (\text{B.13})$$

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 $l(\mathbf{w}, \mathbf{a}, \mathbf{A} | (\mathbf{x}_1, \dots, \mathbf{x}_n))$ for fixed weights $\mathbf{w} \in \mathbb{R}^n$, mean $\mathbf{a} \in \mathbb{R}^{pq}$, and covariance $\mathbf{A} \in \text{PDS}(pq)$ in terms of the precision matrix $\mathbf{B} = \mathbf{A}^{-1}$. Denote this new parametrization as $g(\mathbf{B} | \mathbf{w}, \mathbf{a}, (\mathbf{x}_1, \dots, \mathbf{x}_n)) = l(\mathbf{w}, \mathbf{a}, \mathbf{B}^{-1} | \mathbf{x}_1, \dots, \mathbf{x}_n)$, which is now concave in \mathbf{B} . Especially, for $\mathbf{w} = \mathbf{w}_{\text{MMCD}}$ and

$\mathbf{a} = \hat{\boldsymbol{\mu}}_{\text{MMCD}}$, the function $g(\mathbf{B}|\mathbf{w}_{\text{MMCD}}, \hat{\boldsymbol{\mu}}_{\text{MMCD}}, (\mathbf{x}_1, \dots, \mathbf{x}_n))$ is concave in \mathbf{B} and achieves a unique global maximum at $\mathbf{B} = \hat{\mathbf{S}}_{\text{MMCD}}$. Equations (B.10) and (B.11) then give

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

further implying that

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

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

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

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

$$\|\hat{\mathbf{S}}_{\text{MMCD}} - \hat{\boldsymbol{\Sigma}}_{\text{MMCD}}^{\text{col}} \otimes \hat{\boldsymbol{\Sigma}}_{\text{MMCD}}^{\text{row}}\| < \delta, \quad (\text{B.14})$$

for $\delta = \delta(\varepsilon) \rightarrow 0$ as $n \rightarrow \infty$. Equations (B.13) and (B.14) 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 \(1999\)](#). \square

Proof of Theorem 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 $\hat{\mathbf{M}}_{\mathfrak{Y}}$, $\hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{row}}$, and $\hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{col}}$ denote the *raw* MMCD estimators and $\tilde{\mathbf{M}}_{\mathfrak{Y}}$, $\tilde{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{row}}$, and $\tilde{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{col}}$ denote the *reweighted* MMCD estimators based on the corrupted sample \mathfrak{Y} . Further, $d(\mathbf{Y}_i) = \text{MMD}(\mathbf{Y}_i; \hat{\mathbf{M}}_{\mathfrak{Y}}, \hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{row}}, \hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{col}})$, $i \in N = \{1, \dots, n\}$, denote the matrix Mahalanobis distances of the corrupted sample based on the *raw* MMCD estimators. Since $m \leq \varepsilon^*(\hat{\mathbf{M}}_{\mathfrak{X}}, \mathfrak{X}) - 1 = \varepsilon^*(\hat{\boldsymbol{\Sigma}}_{\mathfrak{X}}^{\text{row}}, \hat{\boldsymbol{\Sigma}}_{\mathfrak{X}}^{\text{col}}, \mathfrak{X}) - 1$ it follows that there exist constants k_0 , k_1 , and k_2 that only depend on \mathfrak{X} , such that

$$\begin{aligned} \|\hat{\mathbf{M}}_{\mathfrak{Y}}\| &\leq k_0 < \infty \quad \text{and} \\ 0 &< k_1 < \lambda_p(\hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{row}}) \lambda_q(\hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{col}}) \leq \lambda_1(\hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{row}}) \lambda_1(\hat{\boldsymbol{\Sigma}}_{\mathfrak{Y}}^{\text{col}}) \leq k_2 < \infty. \end{aligned} \quad (\text{B.15})$$

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 original 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

$$\sum_{i=1}^n w(d(\mathbf{Y}_i)) = \sum_{i \in N \setminus T} w(d(\mathbf{Y}_i)) + \sum_{i \in T} w(d(\mathbf{X}_i)) \geq \sum_{i \in T} w(d(\mathbf{X}_i)) \geq (d+2)c_0 > 0, \quad (\text{B.16})$$

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

Let us now show that there exists a constant $\alpha_0 < \infty$ only dependent on \mathfrak{X} such that $\|\tilde{\mathbf{M}}_{\mathfrak{Y}}\|_F < \alpha_0$. From Equation (B.7) we have that

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

When computing $\tilde{\mathbf{M}}_{\mathfrak{Y}}$ we have that $w(d(\mathbf{Y}_i)) = 0$ if $d(\mathbf{Y}_i) > c_1$ and for all $\mathbf{Y}_i \in \mathfrak{Y}$ that are assigned positive weights, Equation (B.15) yields

$$\|\mathbf{Y}_i\|_F^2 \leq \|\mathbf{Y}_i - \hat{\mathbf{M}}_{\mathfrak{Y}}\|_F^2 + \|\hat{\mathbf{M}}_{\mathfrak{Y}}\|_F^2 \leq c_1 k_2 + k_0^2. \quad (\text{B.17})$$

Since the denominator of $\tilde{\mathbf{M}}_{\mathfrak{Y}}$ is bounded according to Equation (B.16), w is non-increasing and bounded, and k_0 and k_2 are only dependent on \mathfrak{X} , there exists a constant α_0 only dependent on \mathfrak{X} such that

$$\|\tilde{\mathbf{M}}_{\mathfrak{Y}}\|_F \leq \alpha_0 < \infty. \quad (\text{B.18})$$

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 $S \subseteq \{1, \dots, N\}$ denote the subset of indices of the $s = |S|$ samples of $\mathfrak{Y} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$ for which $d_i \leq c_1, i \in S$. Observe that the h samples $\mathbf{Y}_i, i \in H$ are those with the smallest MD, hence $T \subseteq H \subseteq S$. Let $\hat{\mathbf{M}}_{\mathfrak{Y}_T}, \hat{\Sigma}_{\mathfrak{Y}_T}, \hat{\Omega}_{\mathfrak{Y}_T}$ and $\hat{\mathbf{M}}_{\mathfrak{Y}_S}, \hat{\Sigma}_{\mathfrak{Y}_S}, \hat{\Omega}_{\mathfrak{Y}_S}$ denote the MLE estimators of $\mathfrak{Y}_T = (\mathbf{Y}_i)_{i \in T}$ and $\mathfrak{Y}_S = (\mathbf{Y}_i)_{i \in S}$, respectively. Consider the following three ellipsoids:

- Let $E_T = E(\hat{\mathbf{M}}_{\mathfrak{Y}_T}, \hat{\Sigma}_{\mathfrak{Y}_T}, \hat{\Omega}_{\mathfrak{Y}_T})$ denote the ellipsoid corresponding to the MLEs of \mathfrak{Y}_T , i.e., the smallest ellipsoid containing those at least $d + 2$ samples.
- Let $E_S = E(\hat{\mathbf{M}}_{\mathfrak{Y}_S}, \hat{\Sigma}_{\mathfrak{Y}_S}, \hat{\Omega}_{\mathfrak{Y}_S})$ denote the ellipsoid corresponding to the MLEs of \mathfrak{Y}_S .
- Let $E_0 = E(\mathbf{0}, k\mathbf{I}_p, \mathbf{I}_q)$ denote the smallest sphere containing the samples \mathfrak{Y}_S , where $k = c_1 k_2 + k_0^2$ is as in (B.17).

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 $E_T \subseteq a_1 E_S := E(\hat{\mathbf{M}}_{\mathfrak{Y}_S}, a_1 \hat{\Sigma}_{\mathfrak{Y}_S}, \hat{\Omega}_{\mathfrak{Y}_S})$. 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 3.0.1 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 $d_i > c_1, i = 1, \dots, n$. The weighted log-likelihood function for the sample \mathfrak{Y} ,

with the weights satisfying $\sum_{i=1}^n w_i = s$ is given by

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

where $\tilde{\mathbf{w}} = (\tilde{w}(d_1), \dots, \tilde{w}(d_n))$, the new weight function satisfies $\tilde{w}(d_i) = \mathbf{1}(d_1 \leq c_1)$, $\mathfrak{Z} = \{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}$, and $\mathbf{Z}_i = \sqrt{w_i} \mathbf{Y}_i$, $i = 1, \dots, n$. To complete the proof it is sufficient to observe the following: $\mathbf{Z}_1, \dots, \mathbf{Z}_h$ contains at least $d + 2$ points of the form $\sqrt{w_i} \mathbf{X}_i$ and are in a general position, as $w_i \geq a_2 > 0$, for some constant depending only on \mathfrak{X} . Moreover, $\|\mathbf{Z}_i\|_F^2 = w_i \|\mathbf{Y}_i\|_F^2 \leq w_i(c_1 k_2 + k_0^2) \leq w(0)(c_1 k_2 + k_0^2)$, $i = 1, \dots, s$. 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 $0 < w(0) \leq \sum_{i=1}^n w_i \leq sw(0) < \infty$. \square

C MMCD algorithm

Algorithm 1 Iterative C-step procedure for the MMCD estimators

```

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

```

Algorithm 2 Fast *reweighted* MMCD procedure

```

1: procedure MMCD( $\mathfrak{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ )
2:    $h = \lfloor (n+d+2)/2 \rfloor$ 
3:    $\alpha = h/n$ 
4:    $N = \{1, \dots, n\}$ 
5:   for  $k = 1$  to 500 do
6:      $H_k = \text{sample}(N, \text{size} = d + 2)$ 
7:      $(\hat{\mathbf{M}}_k, \hat{\Sigma}_k^{\text{row}}, \hat{\Sigma}_k^{\text{col}}, \mathbf{d}_k, H_k) = \text{CSTEP}_2(\mathfrak{X}, H_k)$   $\triangleright$  2 MLE and C-step iterations
8:      $\delta_k = p \ln(\det(\hat{\Sigma}_k^{\text{col}})) + q \ln(\det(\hat{\Sigma}_k^{\text{row}}))$ 
9:   end for
10:   $\pi_\delta(i) = \{ \{1, \dots, 500\} \rightarrow \{1, \dots, 500\} : i \mapsto j : \delta_{\pi_\delta(1)} \leq \dots \leq \delta_{\pi_\delta(500)} \}$ 
11:  for  $l \in \{\pi_\delta(1), \pi_\delta(2), \dots, \pi_\delta(10)\}$  do
12:     $(\hat{\mathbf{M}}_l, \hat{\Sigma}_l^{\text{row}}, \hat{\Sigma}_l^{\text{col}}, \mathbf{d}_l, H_l) = \text{CSTEP}(\mathfrak{X}, H_l)$   $\triangleright$  Iterating C-steps until convergence
13:     $\delta_l = p \ln(\det(\hat{\Sigma}_l^{\text{col}})) + q \ln(\det(\hat{\Sigma}_l^{\text{row}}))$ 
14:  end for
15:   $j = \arg \min_{k \in N} (\delta_k)$ 
16:   $(\hat{\mathbf{M}}, \hat{\Sigma}^{\text{row}}, \hat{\Sigma}^{\text{col}}) = (\hat{\mathbf{M}}_j, c(\alpha) \hat{\Sigma}_j^{\text{row}}, \hat{\Sigma}_j^{\text{col}})$   $\triangleright$  Consistency scaling for raw MMCD
17:   $\mathbf{d} = (\text{MMD}^2(\mathbf{X}_1; \hat{\mathbf{M}}, \hat{\Sigma}^{\text{row}}, \hat{\Sigma}^{\text{col}}), \dots, \text{MMD}^2(\mathbf{X}_n; \hat{\mathbf{M}}, \hat{\Sigma}^{\text{row}}, \hat{\Sigma}^{\text{col}}))$ 
18:   $H = H_j \cup \{i \in N | d_i < \chi_{0.975; pq}^2\}$ 
19:   $(\hat{\mathbf{M}}, \hat{\Sigma}^{\text{row}}, \hat{\Sigma}^{\text{col}}) = \text{MLE}(\mathbf{X}_{i \in H})$   $\triangleright$  Computation of reweighted MMCD
20:   $\tilde{\alpha} = |H|/n$ 
21:   $(\hat{\mathbf{M}}_*, \hat{\Sigma}_*^{\text{row}}, \hat{\Sigma}_*^{\text{col}}) = (\hat{\mathbf{M}}, c(\tilde{\alpha}) \hat{\Sigma}^{\text{row}}, \hat{\Sigma}^{\text{col}})$   $\triangleright$  Consistency scaling for reweighted MMCD
22:  return  $\hat{\mathbf{M}}_*, \hat{\Sigma}_*^{\text{row}}, \hat{\Sigma}_*^{\text{col}}$ 
23: end procedure

```

C.1 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 Driessen \(1999\)](#). Hence, the number of subsets we must investigate to obtain at least one clean subset with a probability of β is

$$\lceil \log(1 - \beta) / \log(1 - (1 - \varepsilon)^{d+2}) \rceil. \quad (\text{C.1})$$

In Figure C.1, we plot the number of necessary subsets according to Equation (C.1) 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.

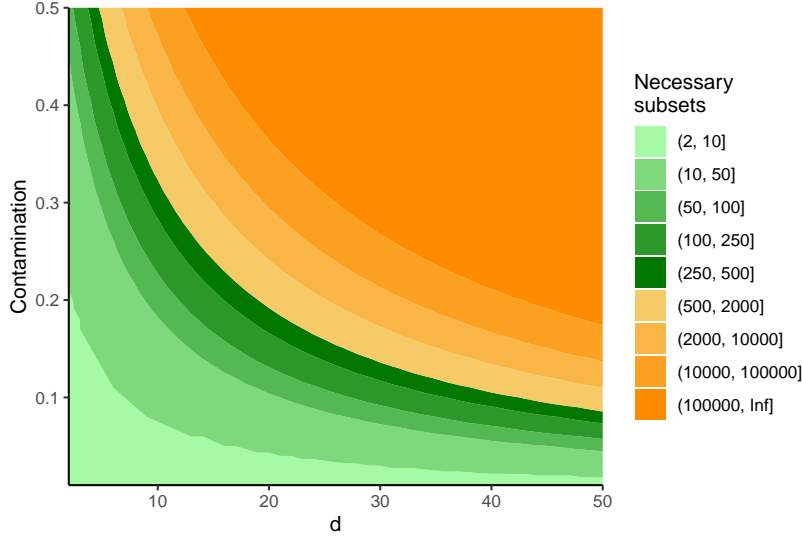


Figure C.1: Number of subsets of size $d + 2$ we have to investigate for various levels of contamination, to obtain at least one clean subset with a probability of 99%

We assess the influence of using only 2 C-step and MLE iterations on the MMCD estimators' objective, the determinant of $\hat{\Sigma}^{\text{col}} \otimes \hat{\Sigma}^{\text{row}}$. 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 $\Sigma^{\text{row}} = \Sigma^{\text{fix}}(0.7)$ and $\Sigma^{\text{col}} = \Sigma^{\text{mix}}(0.7)$, with diagonal entries $\sigma_{jj}^{\text{fix}} = \sigma_{jj}^{\text{mix}} = 1$ and off-diagonal entries $\sigma_{jk}^{\text{fix}}(0.7) = 0.7$ and $\sigma_{jk}^{\text{mix}}(0.7) = 0.7^{|j-k|}$, respectively. The outliers have a mean of 5 and the same covariance as the regular observations. We use 100 random subsets and plot $\det(\hat{\Sigma}^{\text{col}} \otimes \hat{\Sigma}^{\text{row}})$ 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 C.2, we see that there is virtually no difference in the objective function while limiting the ML iterations increases the computation speed. For the subset size, we see that several of

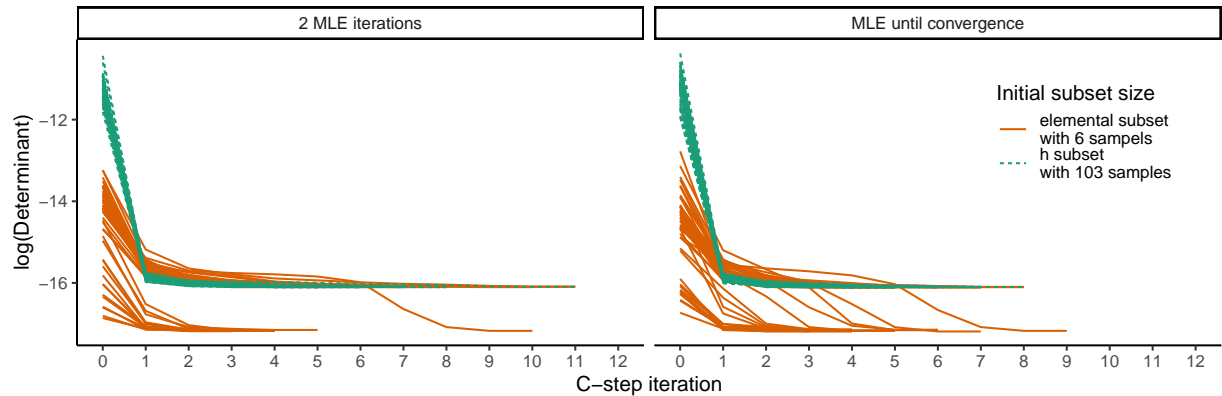


Figure C.2: Logarithm of determinant for successive C-step iterations to analyze the effects of initial subset size and the number of ML iterations.

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 C.1.1. *Instead of using the consistency factor $c(\alpha)$ given in Equation (18), we could also scale the estimators to align the MMDs with a quantile of the chi-square distribution as in [Rousseeuw and 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.*

D Shapley proofs

Proof of Proposition 5.2.1. To show that cellwise Shapley values are not matrix affine equivariant, we consider a rowwise addition matrix \mathbf{A} that adds the w -th row to the v -th row. For simplicity, let \mathbf{B} be the identity matrix. Then Equation (D.1) yields

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

while

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

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

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

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

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

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

If \mathbf{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 \mathbf{B} . This implies that

$$\phi_{jk}(\mathbf{A}\mathbf{X}\mathbf{B}) = x_{jk}y_{jk} = (\mathbf{X} \circ \mathbf{Y})_{jk} = \phi_{jk}(\mathbf{X}),$$

showing the scale invariance.

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

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

and similarly for \mathbf{B} . Hence Equation (D.1) becomes

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

verifying the permutation equivariance. □

Proof of Theorem 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 \text{MMD}(\hat{\mathbf{X}}^S) &:= \text{MMD}(\hat{\mathbf{X}}^{S \cup \{a\}}) - \text{MMD}(\hat{\mathbf{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 $\phi_a(\mathbf{X})$ of the Shapley value $\phi(\mathbf{X})$ are given by ($w(|S|) = \frac{|S|!(p-|S|-1)!}{p!}$)

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

and we can simplify the first term of the sum as

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

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

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

which completes the proof. \square

E 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

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

and angle error between eigenvalues given as

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

where $\hat{\mathbf{a}}$ and \mathbf{a} are the vectors of sorted eigenvalues of $\hat{\Sigma}^{\text{col}} \otimes \hat{\Sigma}^{\text{row}}$ and $\Sigma^{\text{col}} \otimes \Sigma^{\text{row}}$, 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.

E.1 Effects of dimensionality and computation time

We start by considering additional metrics for the simulations discussed in Section 6. Figure E.1 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 E.2 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 E.3 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.

E.2 Cellwise and block contamination

We also consider the additional metrics for the simulations comparing the three different contamination types in Figures E.4 and E.5. 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 E.2.1. *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.

E.3 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 Σ^{rnd} , as proposed by Agostinelli et al. (2015), is randomly generated with low correlations. The

covariance matrix $\Sigma^{\text{fix}}(0.7)$ 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 $\Sigma^{\text{mix}}(0.7)$ 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 $\Sigma^{\text{row}} = \Sigma^{\text{rnd}}$, $\Sigma^{\text{col}} = \Sigma^{\text{mix}}(0.7)$, and $s = 1$ as detailed in Section 6. An overview of all parameters for the simulations is provided in Table E.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 Σ^{row}	$\Sigma^{\text{fix}}(0.7)$, Σ^{rnd}
Columnwise covariance Σ^{col}	$\Sigma^{\text{mix}}(0.7)$, Σ^{rnd}
Mean shift γ	1, 2, 3, 4, 5
Covariance multiplier s	1, 2, 3, 4

Table E.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 E.6, 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 E.7 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 E.8 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 E.9 and E.10 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 E.10 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 E.11 and E.12 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 E.8), 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 E.13 we use $\gamma = 1$ and in Figure E.14 we increase the mean shift to $\gamma = 5$. The F-score based on the true parameters is included as a reference. When $\Sigma^{\text{row}} = \Sigma^{\text{fix}}(0.7)$, $\Sigma^{\text{col}} = \Sigma^{\text{mix}}(0.7)$, 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.

E.3.1 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 E.15, 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 E.16 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.

E.4 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 $\Sigma^{\text{row}} = \Sigma^{\text{rnd}} \in \text{PDS}(p)$ and $\Sigma^{\text{col}} = \Sigma^{\text{mix}}(0.7) \in \text{PDS}(q)$ 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 E.17, 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 3.0.2 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 (18) 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 E.3.

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.

E.5 Banded covariance

Zhang et al. (2022) proposed distribution-free regularized covariance estimation methods for matrix-valued data assuming separability and a banded or tapering covariance structure on both Σ^{row} and Σ^{col} . They also proposed robust versions of their banded and tapering estimators for heavy-tailed distributions. In their simulations for the robust setting, Zhang et al. (2022) assume a banded structure with entries of Σ^{row} and Σ^{col} equal to $\sigma_{jk} = 0.5^{|j-k|} \mathbb{1}_{[|j-k| < 2]}$, resulting in tridiagonal covariance matrices. Here, $\mathbb{1}_{[a]}$ denotes the indicator function, and $j, k \in \{1, \dots, p\}$ or $j, k \in \{1, \dots, q\}$, for Σ^{row} or Σ^{col} , respectively. They compared their estimators to the sample covariance matrix of the vectorized data for samples generated from a heavy-tailed t-distribution with 3 degrees of freedom.

We compare the standard and robust banded and tapering covariance estimators of Zhang et al. (2022) in a setting where a fraction of $\varepsilon = 0.1$ of the samples are replaced by outliers. In this setting, we use the same tridiagonal covariance structure, $(p, q) = (20, 10)$, $n \in \{100, 500, 1000\}$, and consider different shifts $\gamma \in \{0, 0.2, \dots, 4.8, 5\}$ of the outlying observations. The clean data are sampled from a matrix normal distribution with a mean of zero, and the outliers are sampled from a matrix normal distribution with a mean matrix where all entries are equal to δ . As performance metrics we compute the Frobenius error as well as the KL divergence of the covariance estimators. Procedures to compute the banded and tapering covariance estimators are only available in `Matlab` and we used the code provided in the supplement to Zhang et al. (2022) to perform the simulations. Comparing the sample estimates based on the vectorized clean and contaminated data both in `R` and `Matlab` confirms the comparability of the two environments.

Figure E.18 shows the average score for 100 replications as well as the standard error, which is very small and barely visible in the plot, confirming the stability of all procedures. The standard and robust banded and tapering estimators yield similar results on average, only for $n = 100$ and increasing δ , the robust banded and tapering procedures show better performance than the standard procedures. The simulations highlight the difficulties of the banded and tapering estimators for higher values of δ , and they prove to be less reliable than the MLEs for the matrix normal distribution. In terms of Frobenius error, the banded and tapering estimators show favorable properties for low levels of contamination since they force many entries of the covariance to be exactly zero. This is not reflected in the KL divergence, which accounts for the underlying distribution while the banded and tapering estimators are distribution-free. Overall, only the MMCD estimators remain robust for increasing levels of contamination and they clearly outperform the sample covariance estimator computed on the clean data.

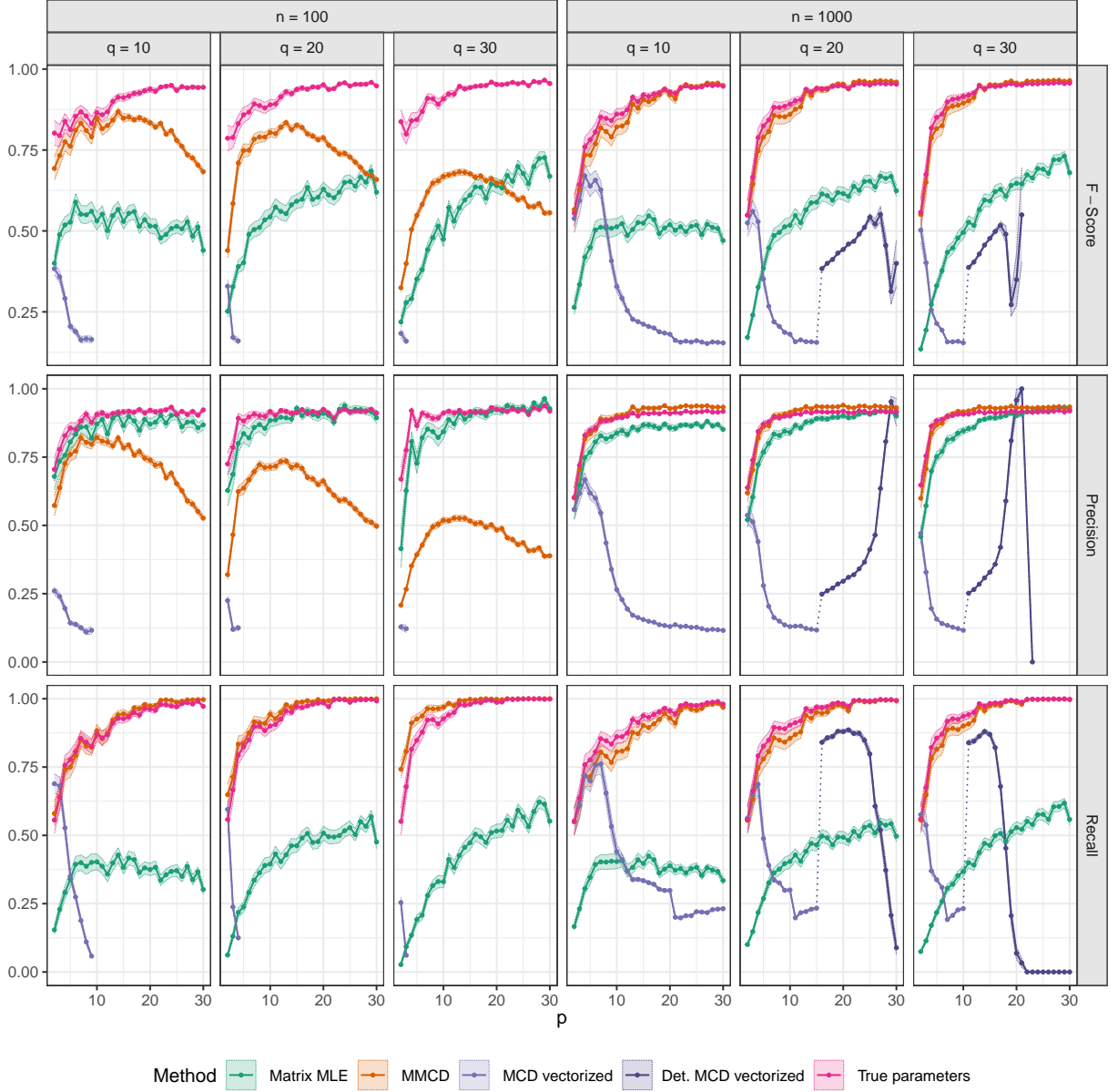


Figure E.1: Outlier detection capabilities comparing multiple matrix sizes $p \in \{2, \dots, 30\}$ and $q \in \{10, 20, 30\}$ for $n \in \{100, 1000\}$, $\gamma = 1$, $\varepsilon = 0.1$.

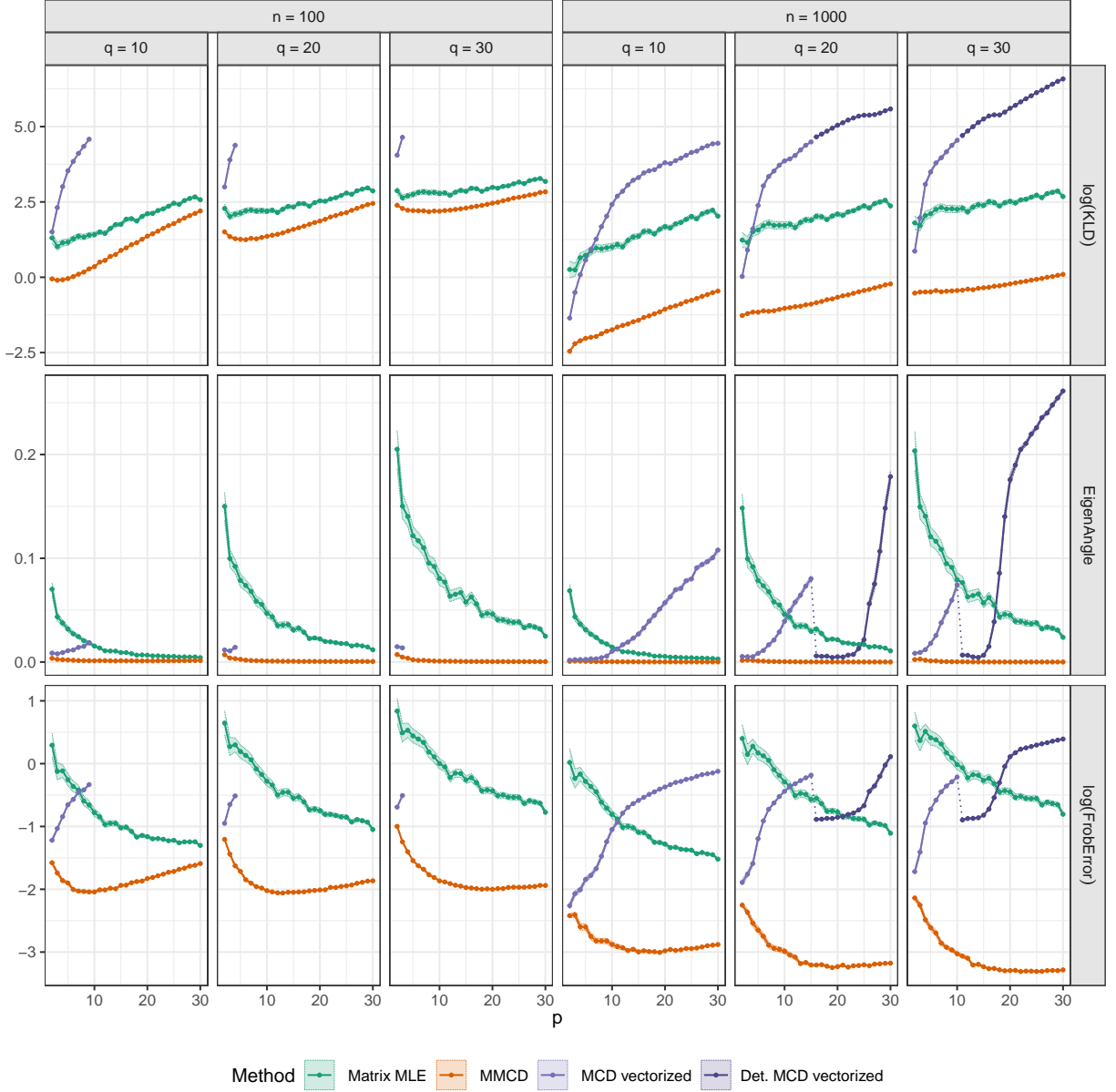


Figure E.2: Quality of covariance estimation comparing multiple matrix sizes $p \in \{2, \dots, 30\}$ and $q \in \{10, 20, 30\}$ for $n \in \{100, 1000\}$, $\gamma = 1$, $\varepsilon = 0.1$.

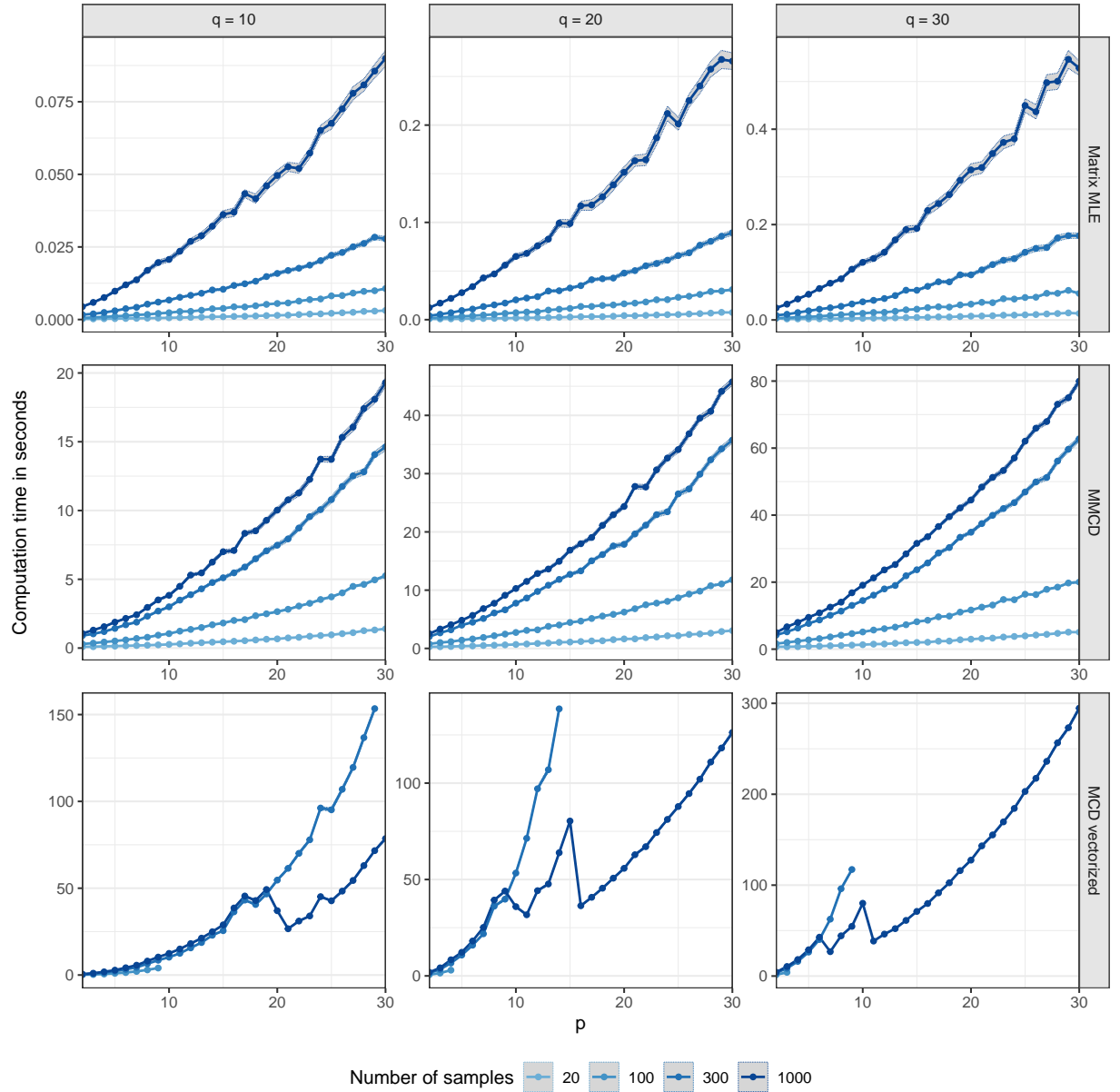


Figure E.3: Comparison of computation time in seconds for multiple matrix sizes $p \in \{2, \dots, 30\}$ and $q \in \{10, 20, 30\}$ for $n \in \{20, 100, 300, 1000\}$.

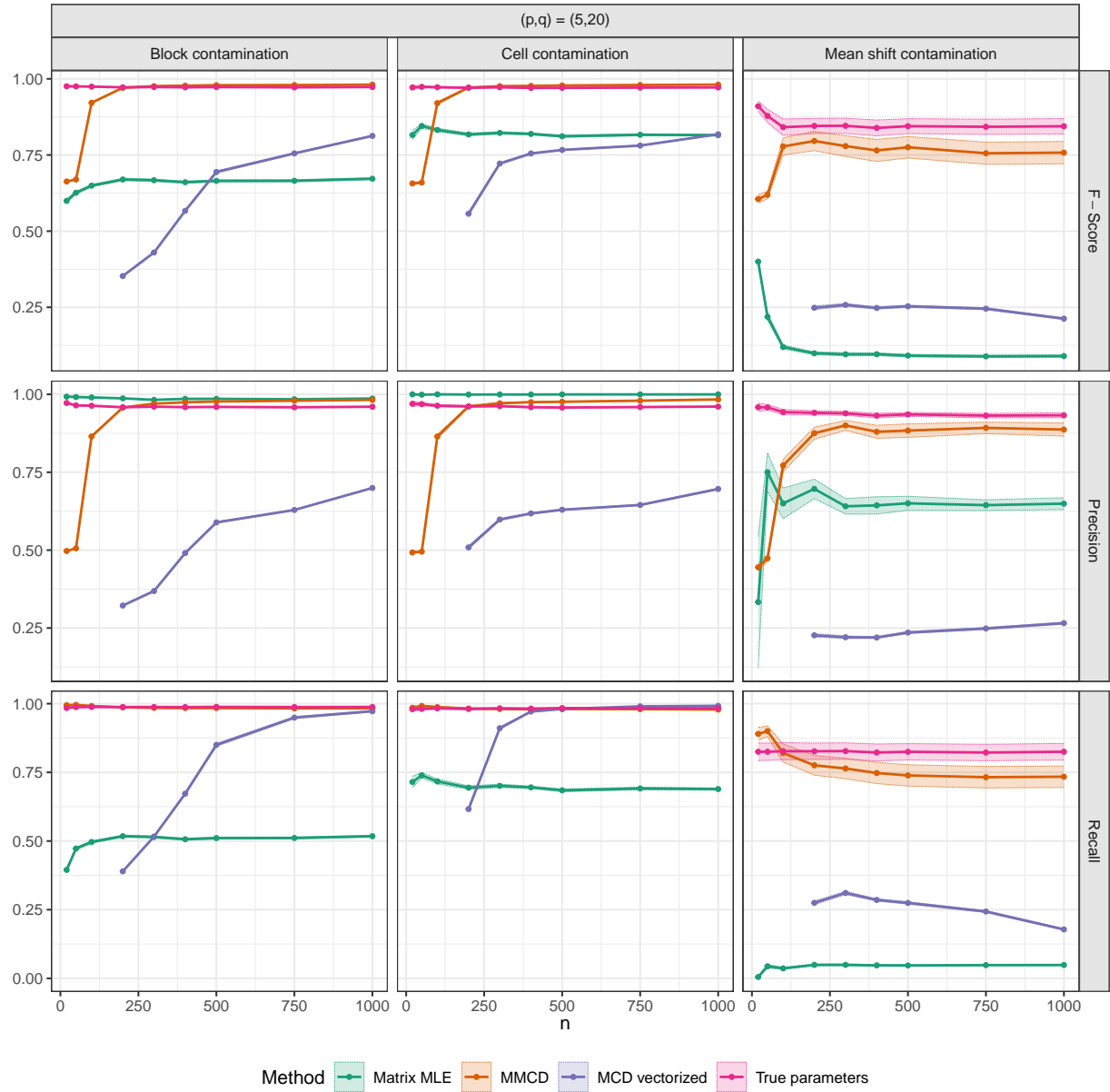


Figure E.4: Quality of covariance estimation comparing block, cell, and sample contamination.

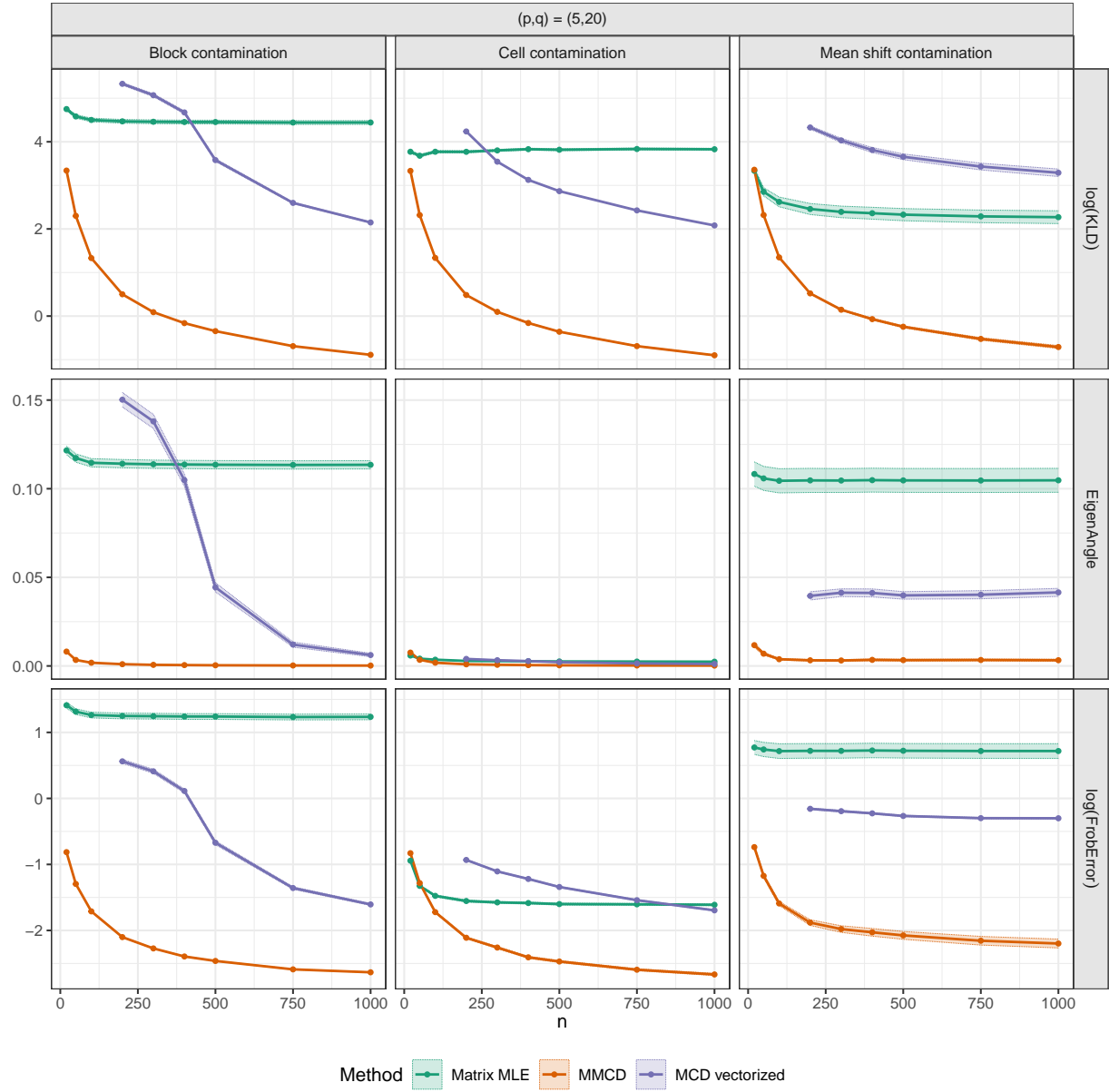


Figure E.5: Outlier detection capabilities comparing block, cell, and sample contamination.

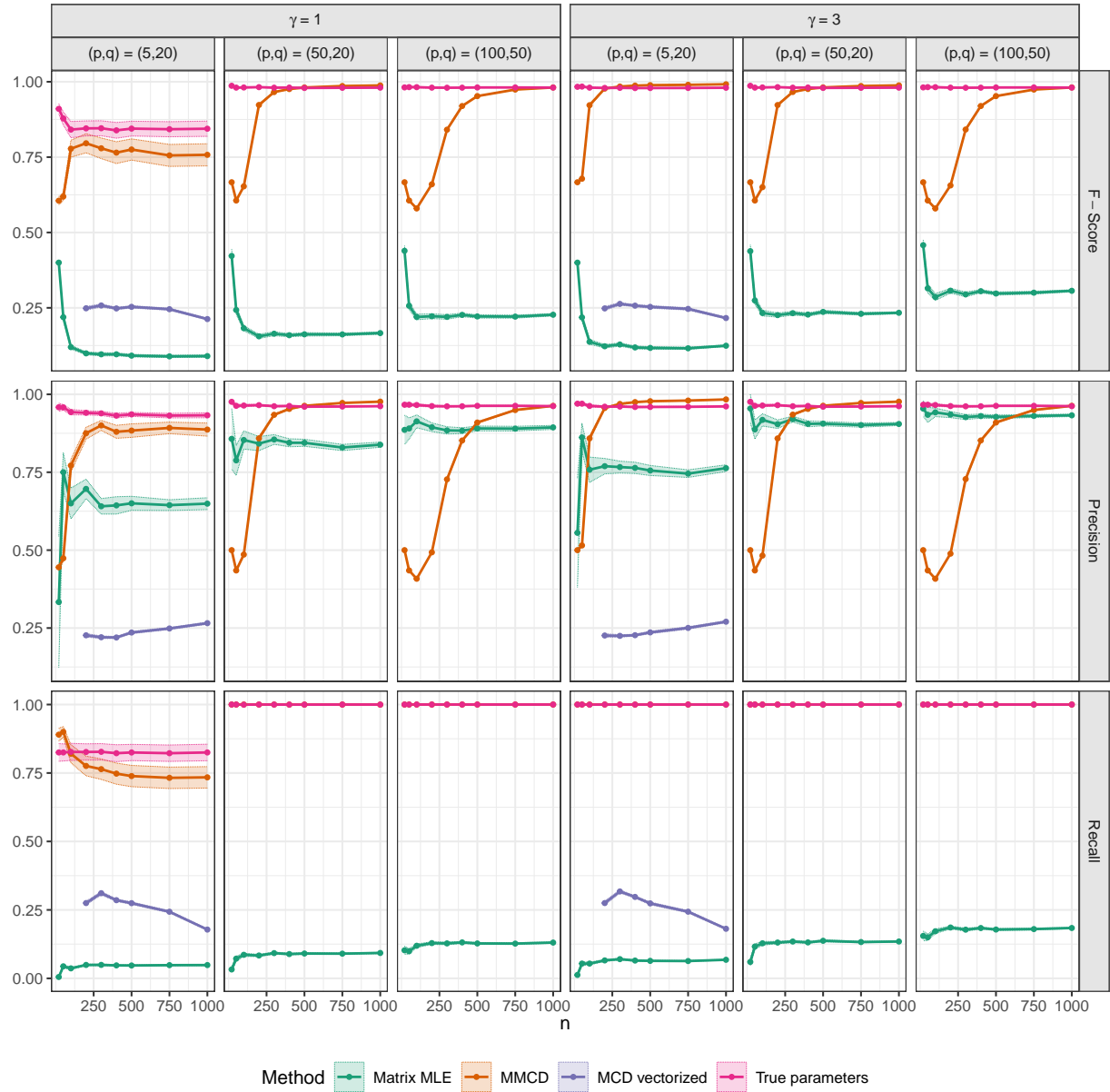


Figure E.6: Overview of simulation results with a fraction $\varepsilon = 0.2$ of contaminated samples. The outlier detection capabilities are measured by F-score, precision, and recall.

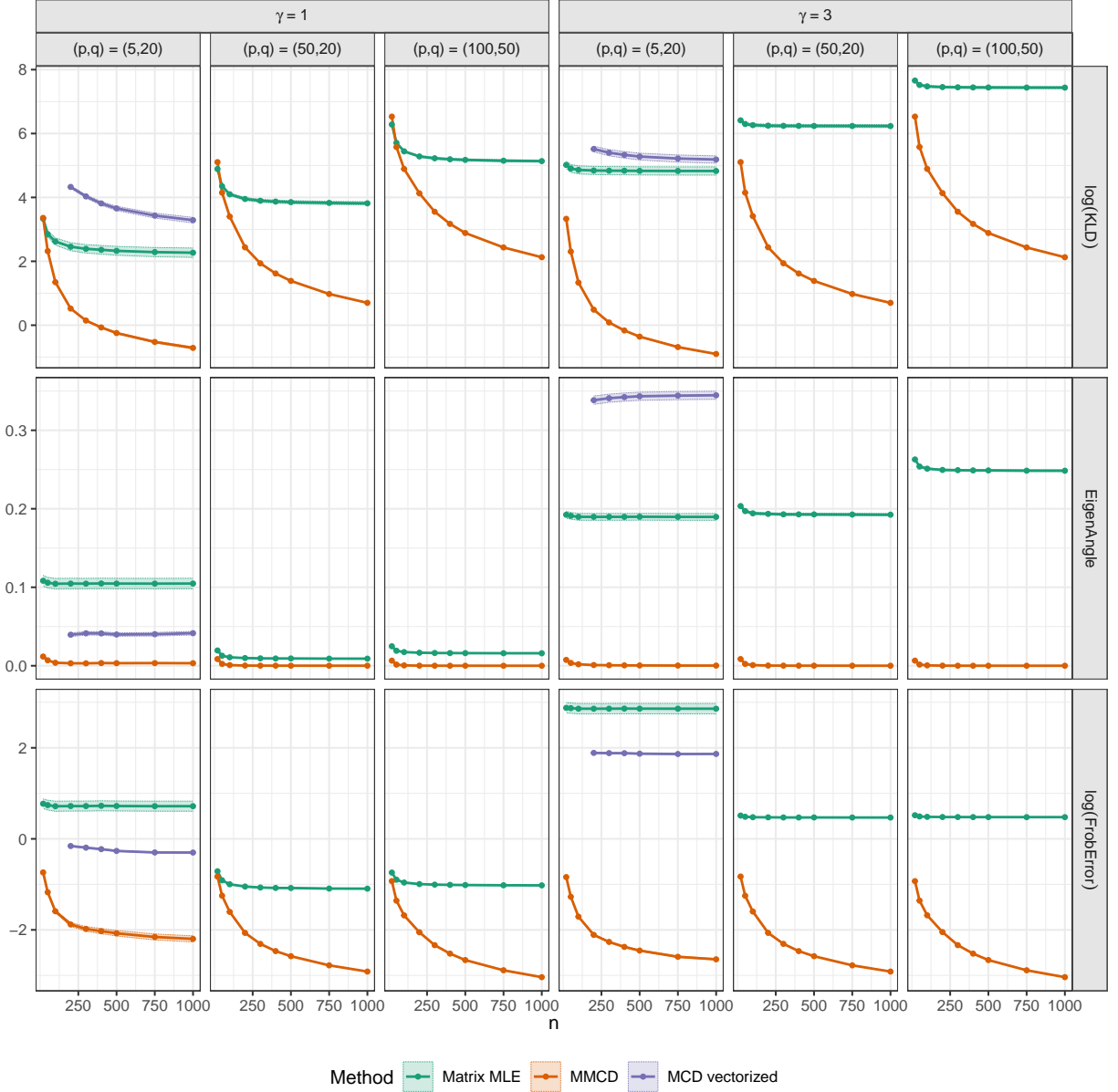


Figure E.7: Overview of simulation results with a fraction $\varepsilon = 0.2$ of contaminated samples. The quality of covariance estimation is evaluated based on the logarithm of KL divergence, angle error between eigenvalues, and the logarithm of relative Frobenius error.

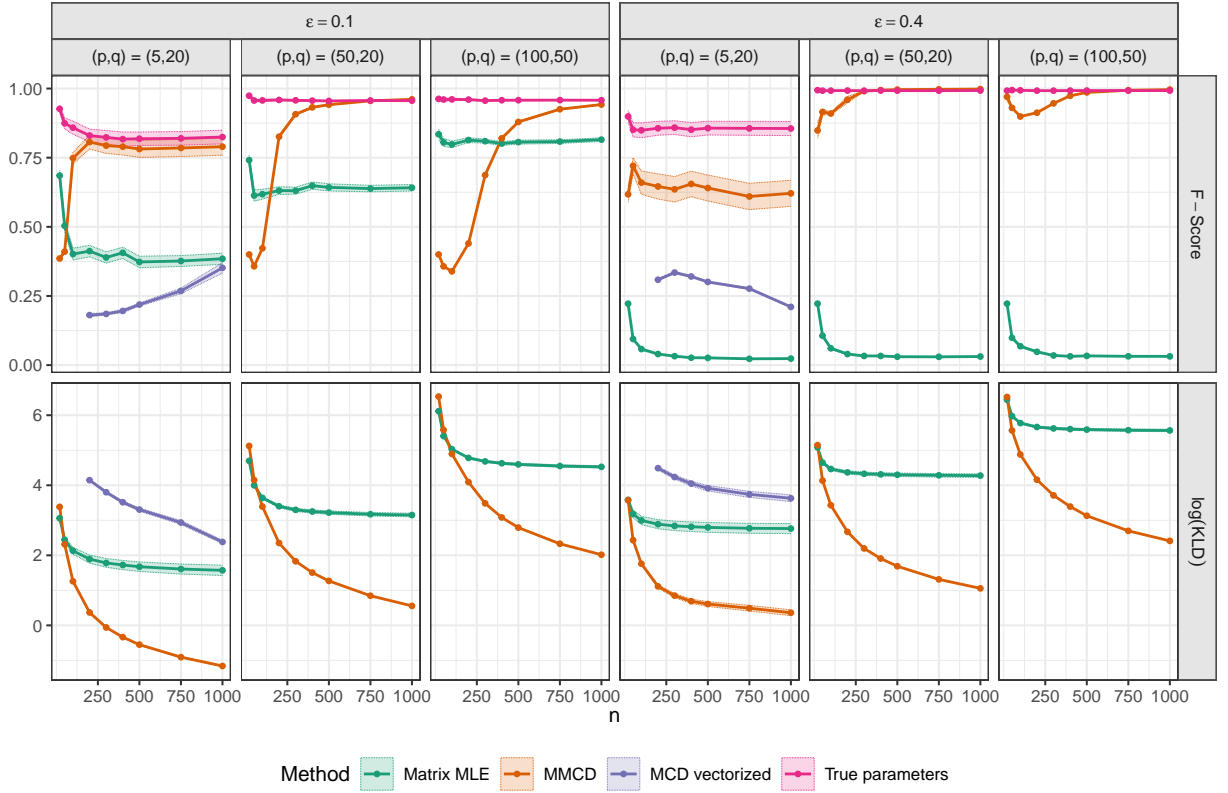


Figure E.8: F-score and logarithm of KL divergence for simulations with mean shift $\gamma = 1$.

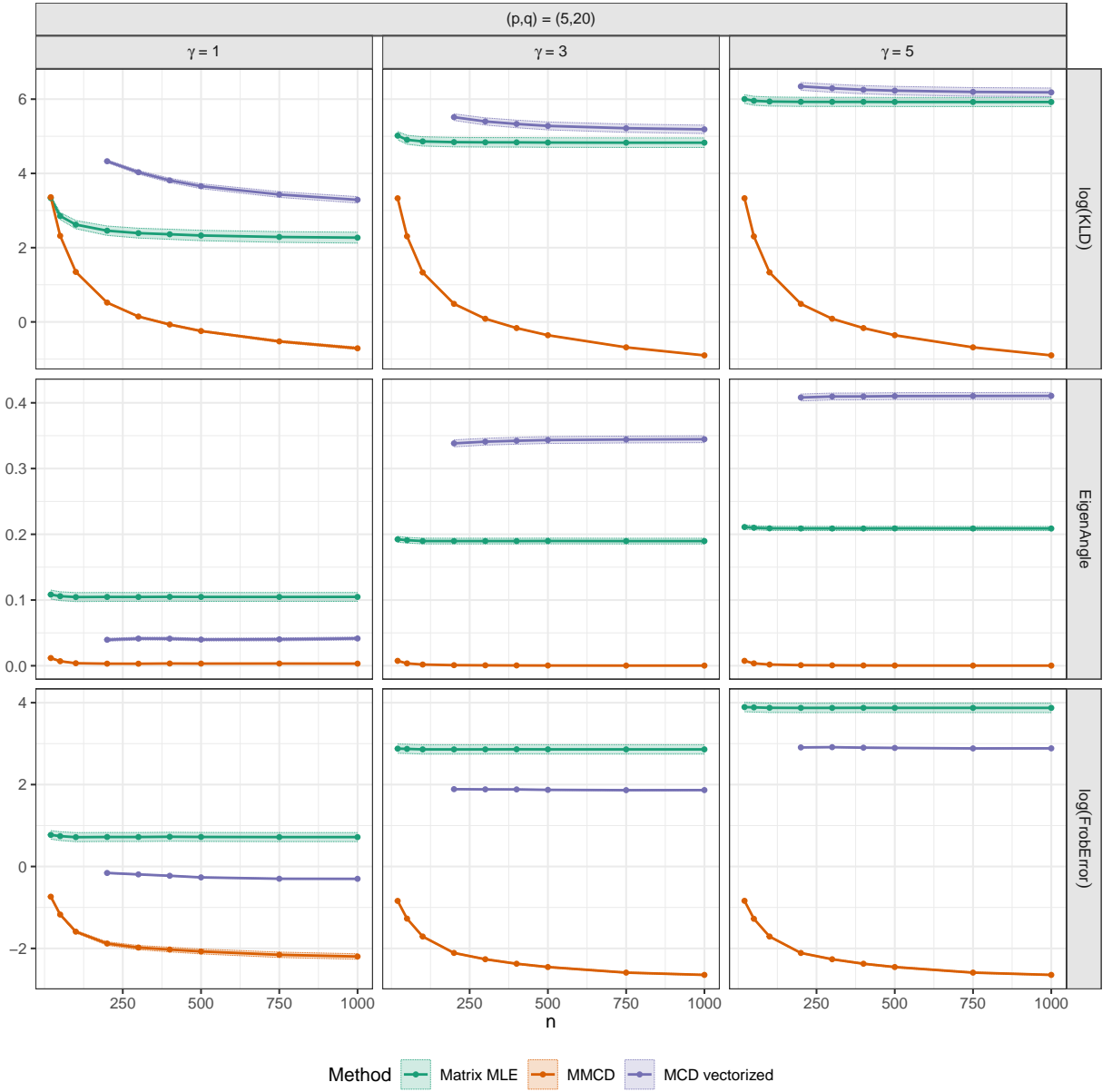


Figure E.9: Quality of covariance estimation for simulations with $\varepsilon = 0.2$ and $(p, q) = (5, 20)$.

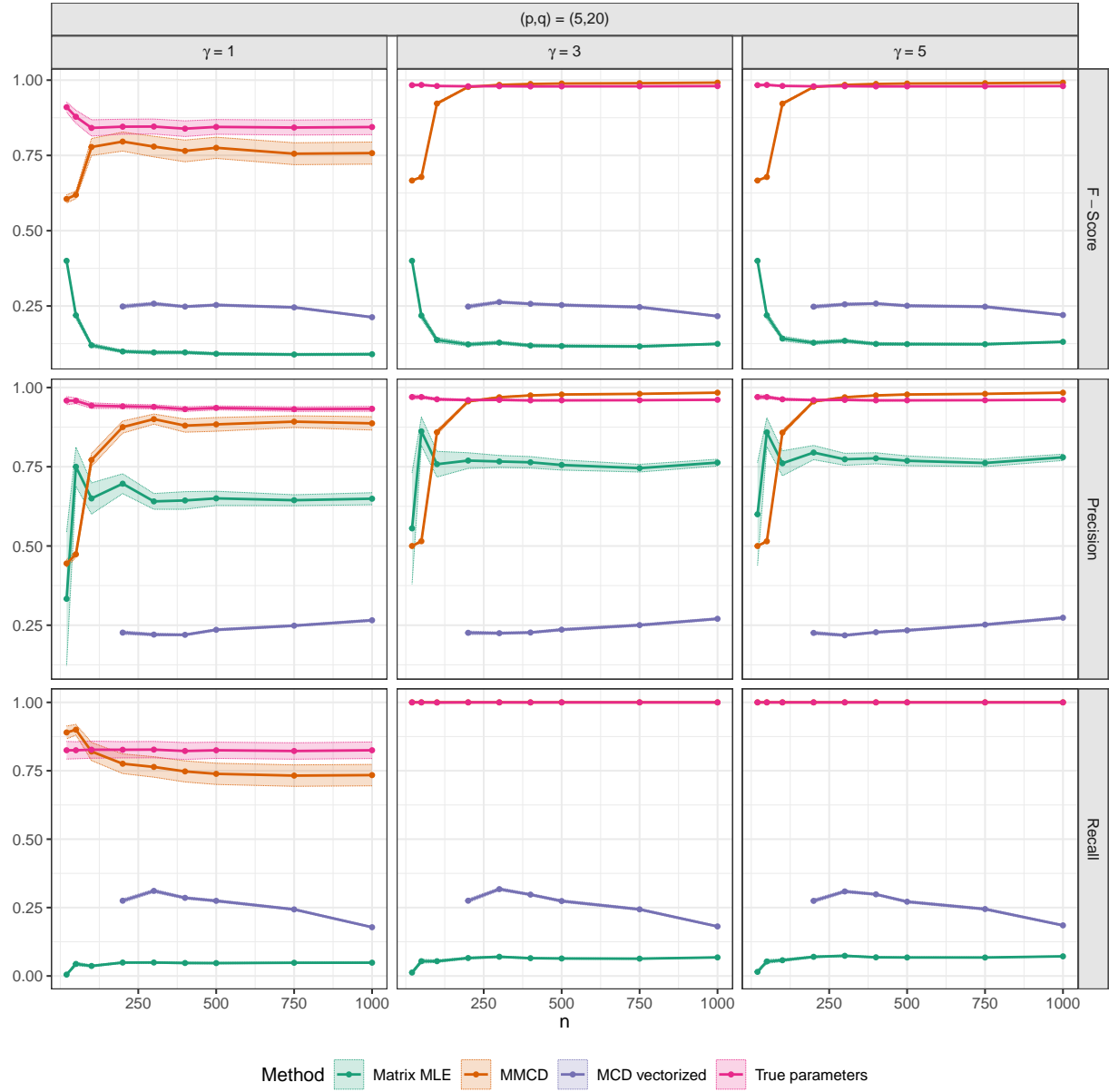


Figure E.10: Outlier detection capabilities for simulations with $\varepsilon = 0.2$ and $(p, q) = (5, 20)$.

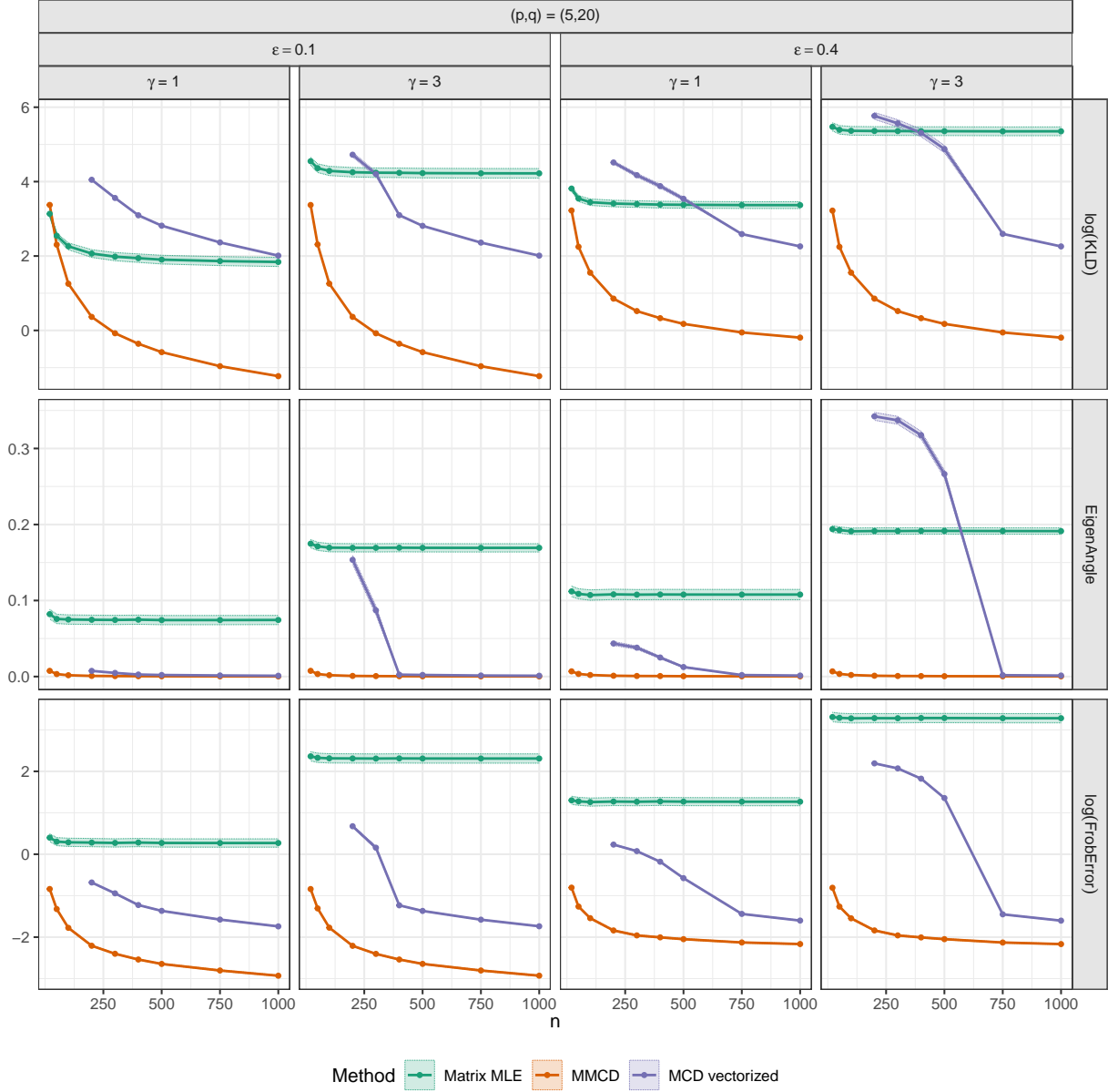


Figure E.11: Quality of covariance estimation for simulations where the covariance of the outliers is scaled by $s = 2$.

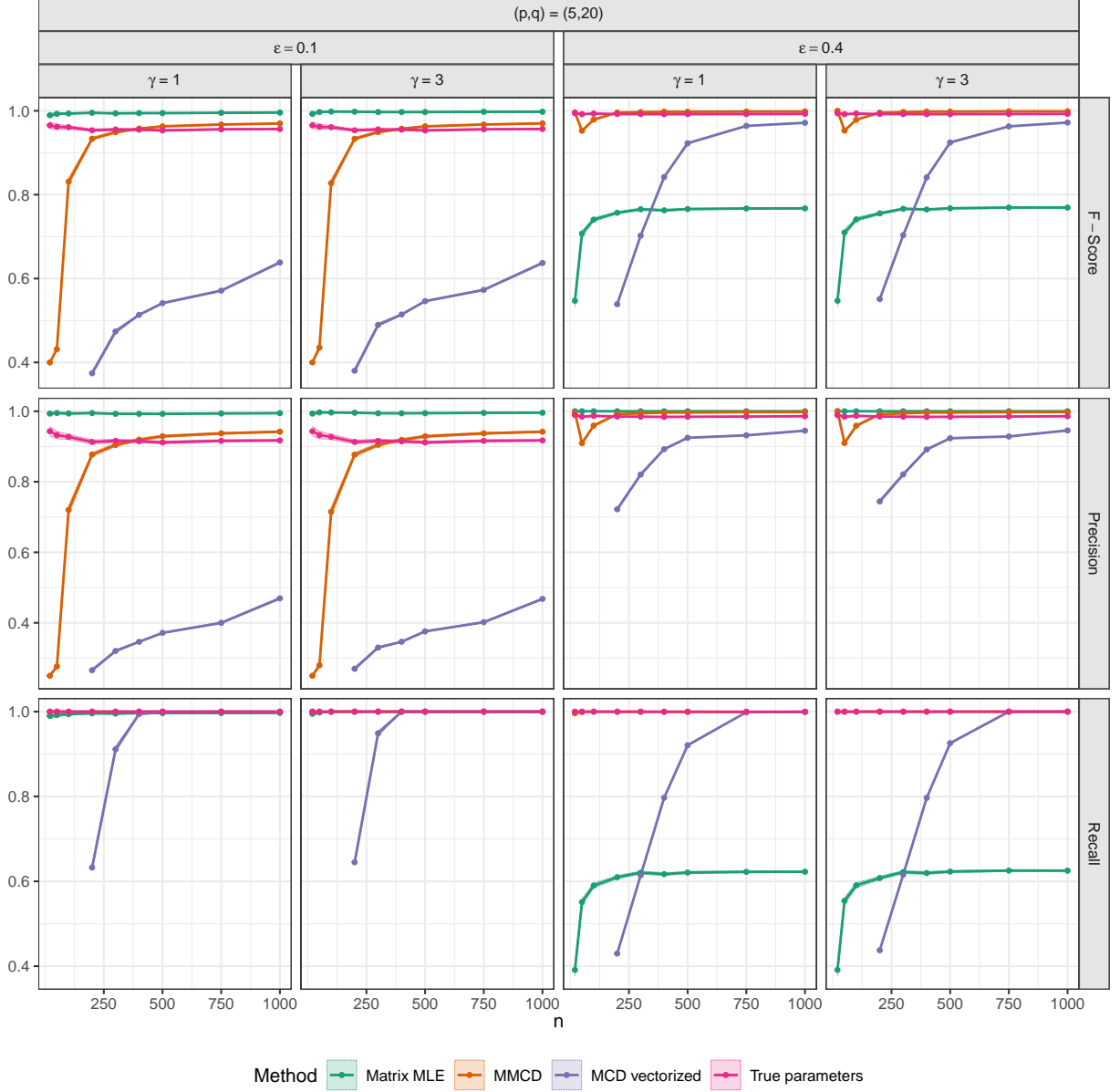


Figure E.12: Outlier detection capabilities for simulations where the covariance of the outliers is scaled by $s = 2$.

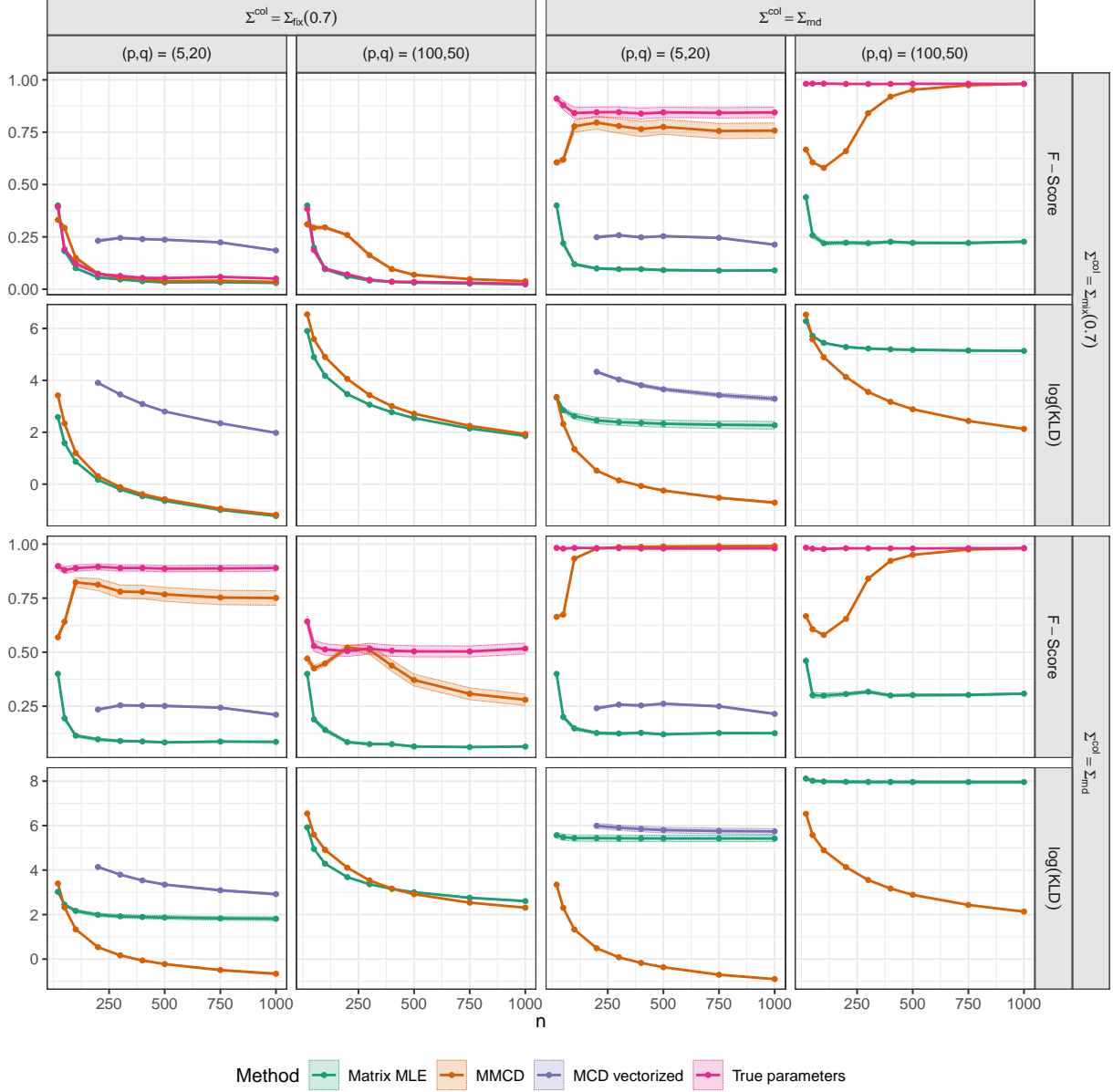


Figure E.13: F-score and logarithm of KL divergence comparing 4 different combinations of row- and columnwise covariance matrices, $\gamma = 1$, and $\varepsilon = 0.2$.

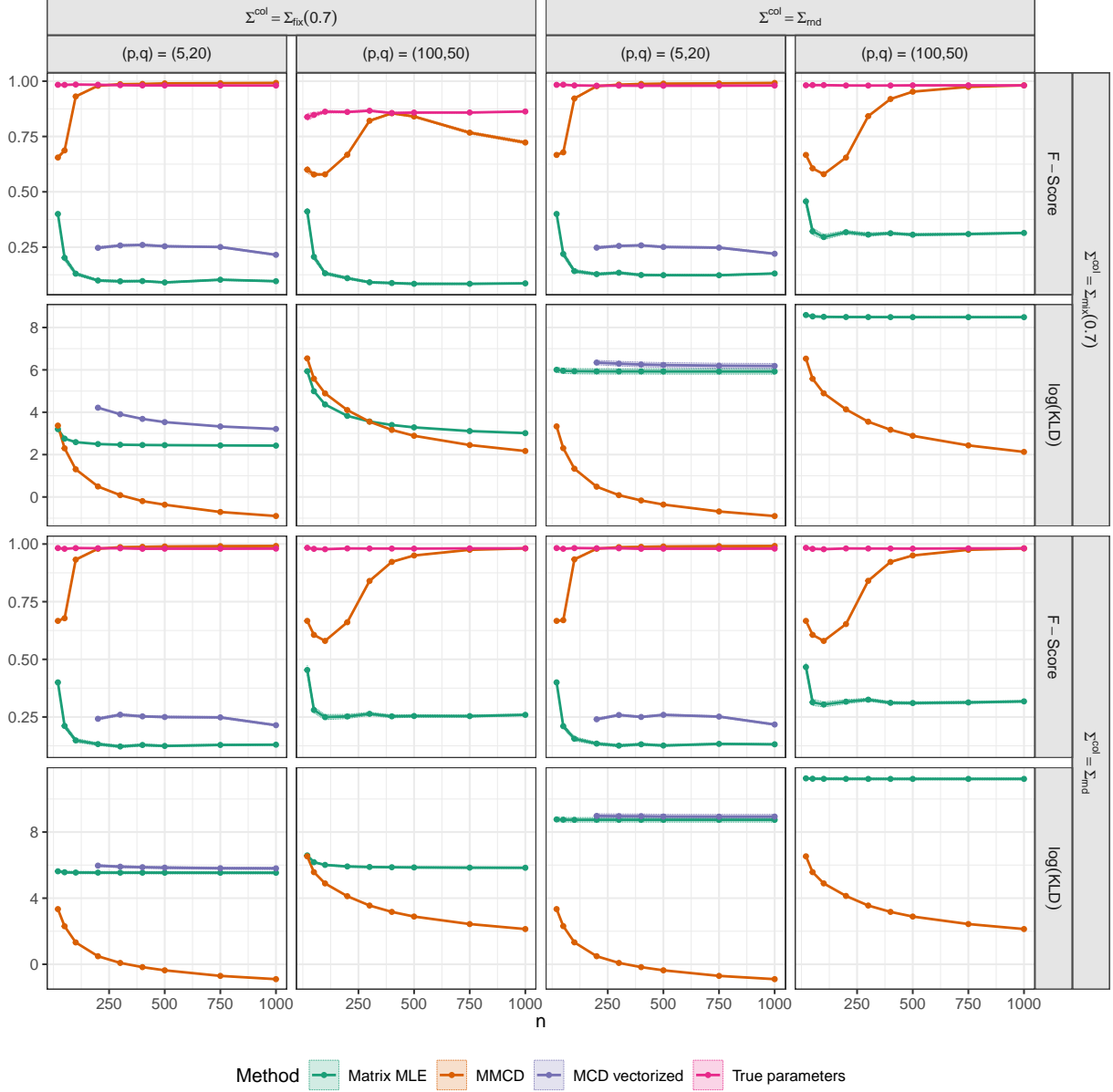


Figure E.14: F-score and logarithm of KL divergence comparing 4 different combinations of row- and columnwise covariance matrices, $\gamma = 5$, and $\varepsilon = 0.2$.

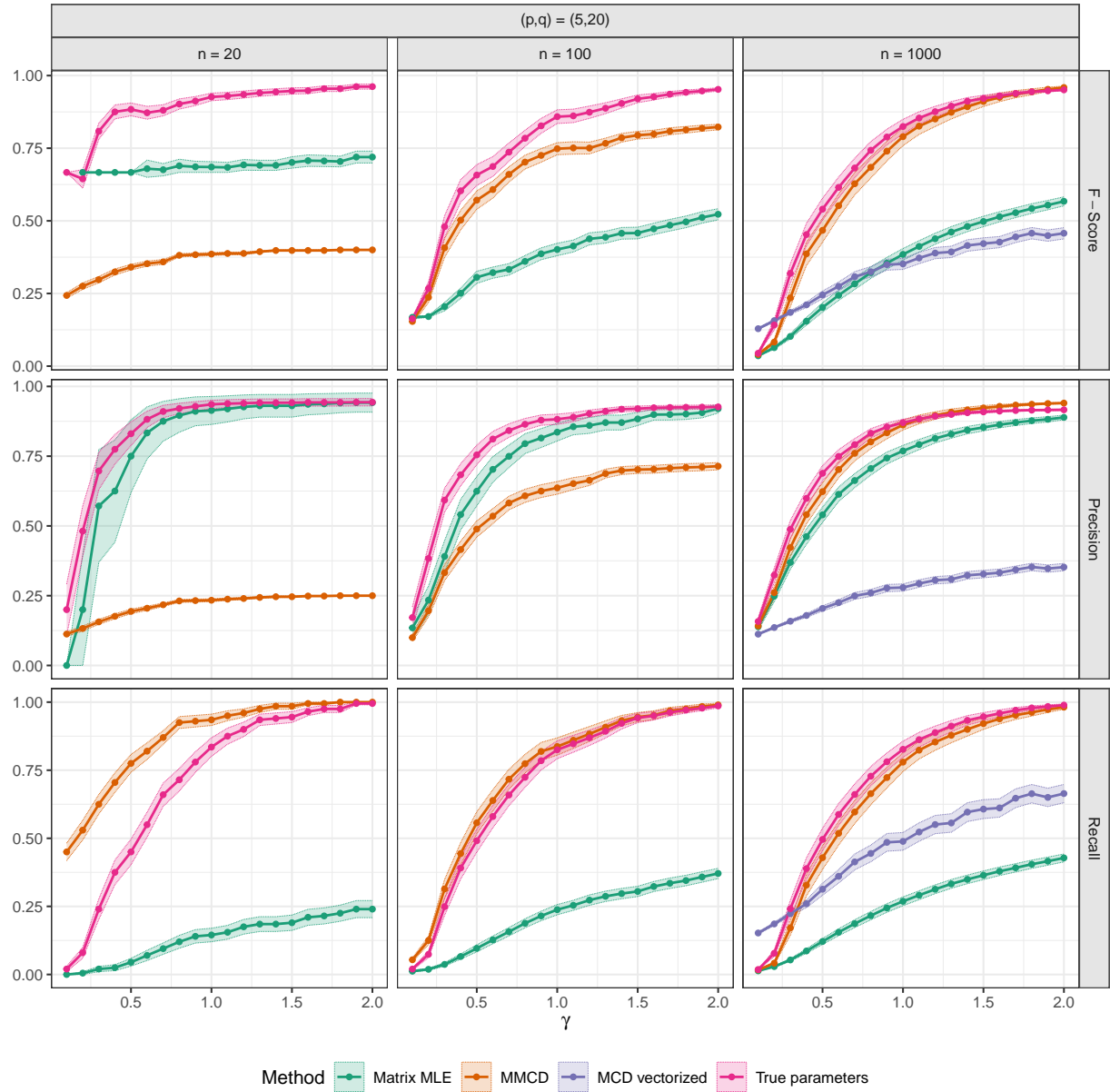


Figure E.15: Outlier detection capabilities for simulations with mean shift $\gamma \in \{0.1, 0.2, \dots, 2\}$ for $n \in \{20, 100, 1000\}$, $\varepsilon = 0.1$.

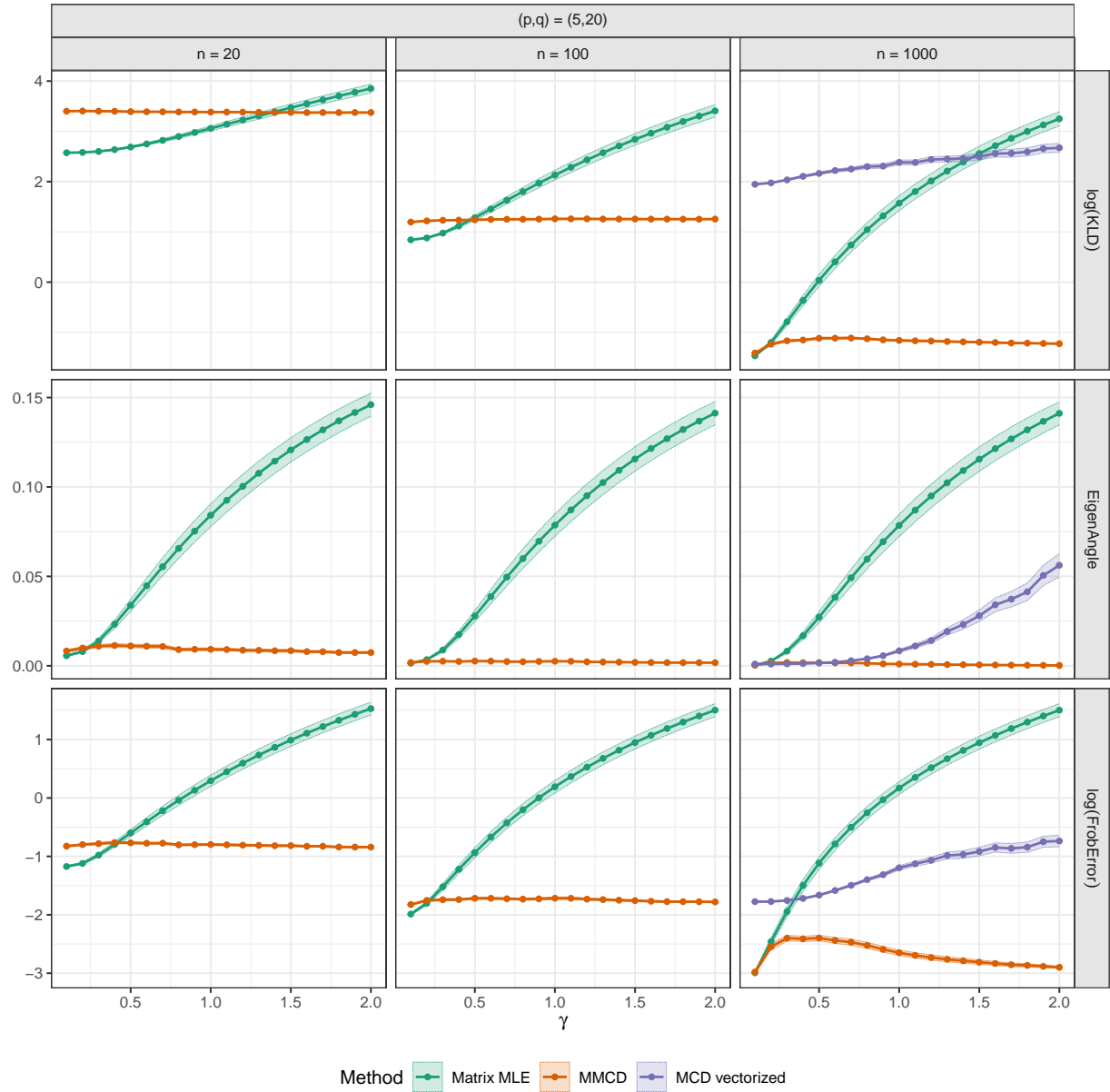


Figure E.16: Quality of covariance estimation for simulations with mean shift $\gamma \in \{0.1, 0.2, \dots, 2\}$ for $n \in \{20, 100, 1000\}$, $\varepsilon = 0.1$.

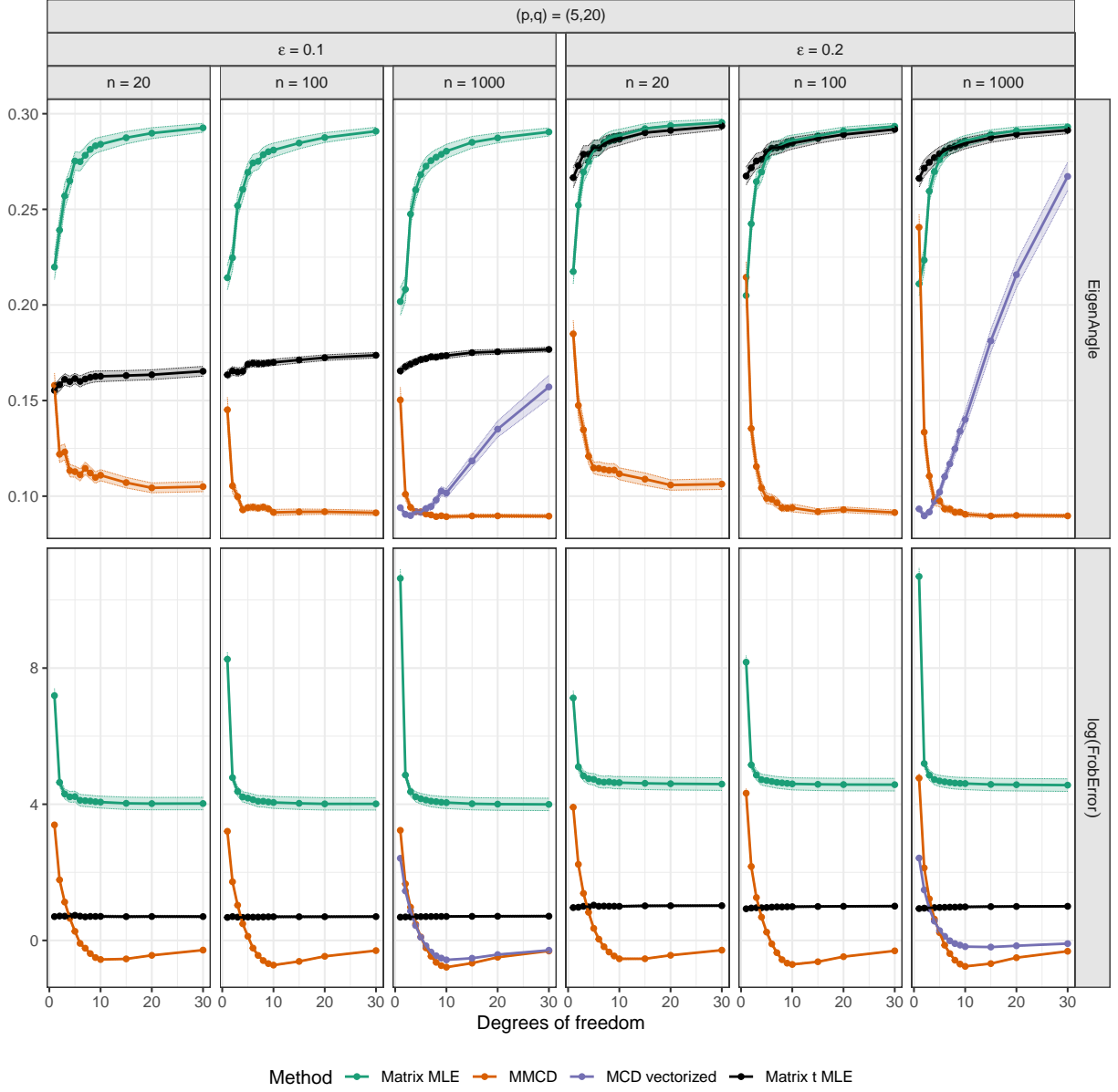


Figure E.17: The eigen angle and the logarithm of relative Frobenius error of samples from a contaminated t-distribution with $\nu \in \{1, \dots, 30\}$ degrees of freedom for $n \in \{20, 100, 1000\}$, $\gamma = 1$, $\varepsilon \in \{0.1, 0.2\}$.

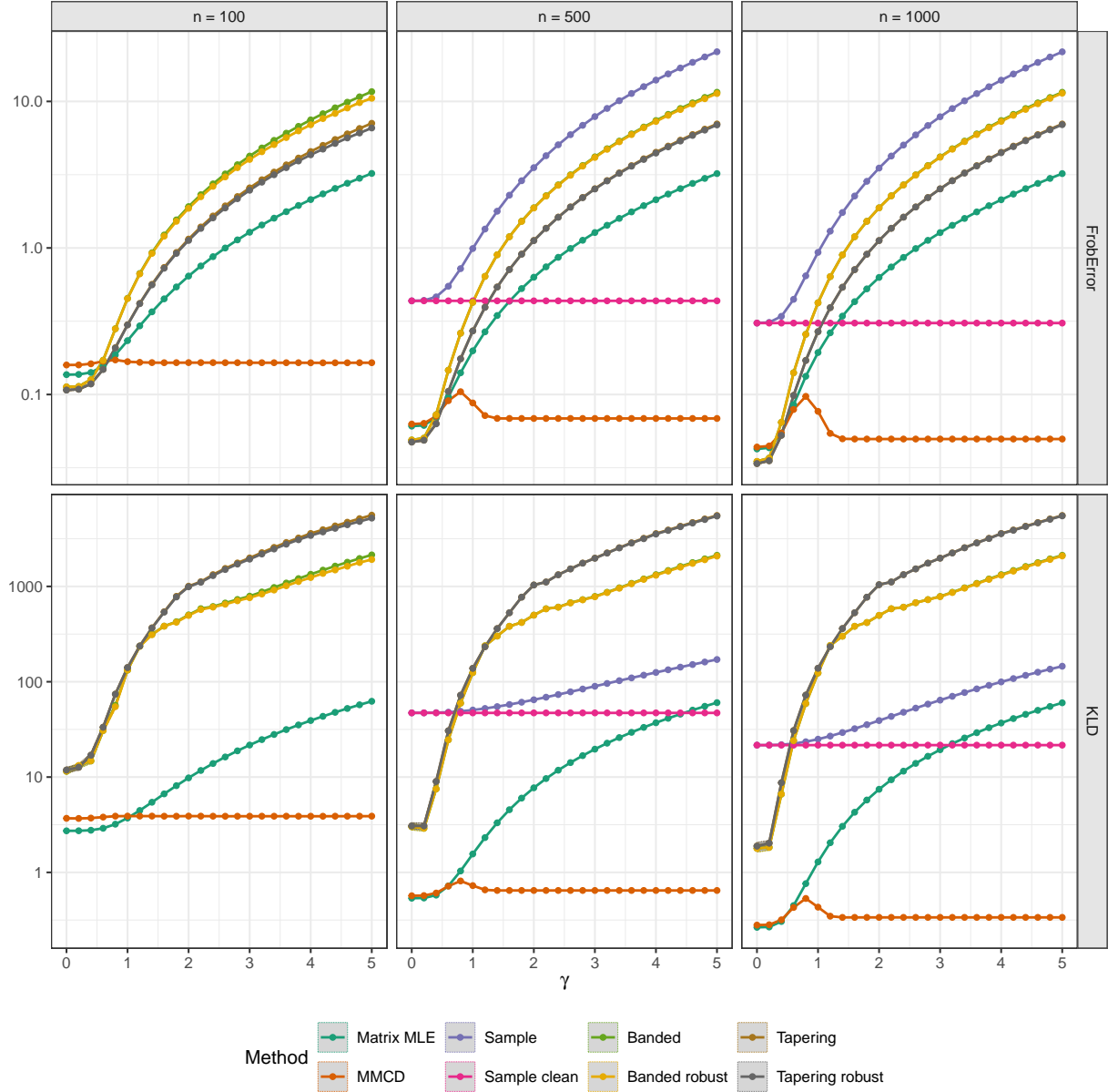


Figure E.18: Covariance estimation performance, where both the rowwise and columnwise covariance matrices are banded, evaluated with $\gamma \in \{0, 0.2, \dots, 5\}$, $n \in \{100, 500, 1000\}$, $\varepsilon = 0.1$.

Theorem E.1. *Let $n \geq d + 1$, and the sample is in a general position. Then the breakdown point of any matrix affine equivariant (in the sense of REF) location and scatter estimator is at most $\frac{1}{n} \lfloor (n - d + 1)/2 \rfloor$.*

Proof of Theorem E.1. Let $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ be a sample of $p \times q$ -matrices in the general position such that $n \geq d + 1$ for $d = \lfloor \frac{p}{q} + \frac{q}{p} \rfloor$, and denote further $\chi = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i = \text{vec}(\mathbf{X}_i) \in \mathbb{R}^{pq}$, $i = 1, \dots, n$. Consider first the class of location and scatter estimators $(\hat{\mathbf{T}}, \hat{\Sigma})$ in \mathbb{R}^{pq} that are affine equivariant under the affine transformation with Kronecker-structure matrices, i.e. for any regular $\mathbf{A} \in \mathbb{R}^{p \times p}$, $\mathbf{B} \in \mathbb{R}^{q \times q}$, $\hat{\Sigma}((\mathbf{A} \otimes \mathbf{B})\mathbf{x}) = (\mathbf{A} \otimes \mathbf{B})\hat{\Sigma}(\mathbf{x})(\mathbf{A}' \otimes \mathbf{B}')$. We first show that the breakdown point of such an estimator, when applied to the vectorized matrices in a general position is at most $\frac{1}{n} \lfloor (n - d + 1)/2 \rfloor$.

Let \mathcal{S}' be the contaminated sample with at most $\lfloor (n - d + 1)/2 \rfloor$ contaminated points. Then, there are at least $n - \lfloor (n - d + 1)/2 \rfloor \geq d$ uncontaminated points. Let those points be precisely $\mathbf{x}_1, \dots, \mathbf{x}_{n - \lfloor (n - d + 1)/2 \rfloor}$.

Choose now any d out of those $n - \lfloor (n - d + 1)/2 \rfloor \geq d$ uncontaminated points that lay on the plane $\mathbf{a}'\mathbf{X}\mathbf{b} = 0$, for $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{b} \in \mathbb{R}^q$, i.e. $(\mathbf{a} \otimes \mathbf{b})'\mathbf{x} = 0$. W.l.o.g., assume that these points are $\mathbf{x}_1, \dots, \mathbf{x}_d$. Due to the affine equivariance of the estimator, we can choose \mathbf{a} and \mathbf{b} to be the first vectors of the corresponding canonical bases.

There are then $n' = n - \lfloor (n - d + 1)/2 \rfloor \geq d - d$ uncontaminated points in \mathcal{S}' . W.l.o.g., let those points be $\mathbf{x}_{d+1}, \dots, \mathbf{x}_{d+n'}$. Since $n' \leq \lfloor (n - d + 1)/2 \rfloor$, choose n' points from $\lfloor (n - d + 1)/2 \rfloor$ contaminated ones and replace them with the points of the form:

$$\mathbf{y}_i = \begin{pmatrix} u & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-1} \end{pmatrix} \otimes \begin{pmatrix} u & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q-1} \end{pmatrix} \mathbf{x}_i,$$

for $u > 0$, $i = d + 1, \dots, d + n'$. The contaminated sample $\mathcal{S}'(u)$ now consists of the points: $\mathcal{S}'(u) = \{\mathbf{x}_1, \dots, \mathbf{x}_d, \mathbf{x}_{d+1}, \dots, \mathbf{x}_{d+n'}, \dots, \mathbf{y}_{d+1}, \dots, \mathbf{y}_{d+n'}, \dots, \mathbf{x}_n\}$. Denoting $\Lambda_p(u) = \text{diag}(u, \mathbf{I}_{p-1})$, $\Lambda_q(u) = \text{diag}(u, \mathbf{I}_{q-1})$, construct now the second contaminated sample $\mathcal{S}'' = \Lambda_p(u)^{-1} \otimes \Lambda_q(u)^{-1} \mathcal{S}'$.

□

References

- Agostinelli, C., A. Leung, V. J. Yohai, and R. H. Zamar (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test* 24, 441–461.
- Alqallaf, F., S. Van Aelst, V. J. Yohai, and R. H. Zamar (2009). Propagation of outliers in multivariate data. *The Annals of Statistics* 37(1), 311–331.
- Cator, E. A. and H. P. Lopuhaä (2012). Central limit theorem and influence function for the MCD estimators at general multivariate distributions. *Bernoulli* 18(2), 520 – 551.
- Croux, C. and G. Haesbroeck (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis* 71(2), 161–190.

- Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* 64(2), 105–123.
- Gupta, A. and D. Nagar (1999). *Matrix Variate Distributions*. Monographs and Surveys in Pure and Applied Mathematics. Taylor & Francis.
- Gupta, A. K. and T. Varga (2012). *Elliptically contoured models in statistics*, Volume 240. Springer Science & Business Media.
- Lu, N. and D. L. Zimmerman (2005). The likelihood ratio test for a separable covariance matrix. *Statistics & Probability Letters* 73(4), 449–457.
- Raymaekers, J. and P. J. Rousseeuw (2023). The cellwise minimum covariance determinant estimator. *Journal of the American Statistical Association* 0(0), 1–12.
- Roś, B., F. Bijma, J. C. de Munck, and M. C. de Gunst (2016). Existence and uniqueness of the maximum likelihood estimator for models with a kronecker product covariance structure. *Journal of Multivariate Analysis* 143, 345–361.
- Rousseeuw, P. J. and K. V. Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41(3), 212–223.
- Soloveychik, I. and D. Trushin (2016). Gaussian and robust kronecker product covariance estimation: Existence and uniqueness. *Journal of Multivariate Analysis* 149, 92–113.
- Srivastava, M. S., T. von Rosen, and D. Von Rosen (2008). Models with a kronecker product covariance structure: estimation and testing. *Mathematical methods of statistics* 17, 357–370.
- Thompson, G. Z., R. Maitra, W. Q. Meeker, and A. F. Bastawros (2020). Classification with the matrix-variate-t distribution. *Journal of Computational and Graphical Statistics* 29(3), 668–674.
- Zhang, Y., W. Shen, and D. Kong (2022). Covariance estimation for matrix-valued data. *Journal of the American Statistical Association*, 1–12.