The Minimum Regularized Covariance Determinant estimator

Kris Boudt
Solvay Business School, Vrije Universiteit Brussel
Faculty of Economics and Business, Vrije Universiteit Amsterdam

Peter J. Rousseeuw
Department of Mathematics, KU Leuven

Steven Vanduffel
Solvay Business School, Vrije Universiteit Brussel

Tim Verdonck
Department of Mathematics, KU Leuven

March 22, 2018

Abstract

The Minimum Covariance Determinant (MCD) approach estimates the location and scatter matrix using the subset of given size with lowest sample covariance determinant. Its main drawback is that it cannot be applied when the dimension exceeds the subset size. We propose the Minimum Regularized Covariance Determinant (MRCD) approach, which differs from the MCD in that the subset-based covariance matrix is a convex combination of a target matrix and the sample covariance matrix. A data-driven procedure sets the weight of the target matrix, so that the regularization is only used when needed. The MRCD estimator is defined in any dimension, is well-conditioned by construction and preserves the good robustness properties of the MCD. We prove that so-called concentration steps can be performed to reduce the MRCD objective function, and we exploit this fact to construct a fast algorithm. We verify the accuracy and robustness of the MRCD estimator in a simulation study and illustrate its practical use for outlier detection and regression analysis on real-life high-dimensional data sets in chemistry and criminology.

Keywords: Breakdown value; High-dimensional data; Regularization; Robust covariance estimation.

1 Introduction

The Minimum Covariance Determinant (MCD) method (Rousseeuw, 1984, 1985) is a highly robust estimator of multivariate location and scatter. Given an $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, its objective is to find h observations whose sample covariance matrix has the lowest possible determinant. Here $h < n$ is fixed. The MCD estimate of location is then the average of these h points, whereas the scatter estimate is a multiple of their covariance matrix. Consistency and asymptotic normality of the MCD estimator have been shown by Butler et al. (1993) and Cator and Lopuhaä (2012). The MCD has a bounded influence function (Croux and Haesbroeck, 1999) and has the highest possible breakdown value (i.e. 50%) when $h = \lfloor (n + p + 1)/2 \rfloor$ (Lopuhaä and Rousseeuw, 1991). The MCD approach has been applied to various fields such as chemistry, finance, image analysis, medicine, and quality control, see e.g. the review paper of Hubert et al. (2008).

A major restriction of the MCD approach is that the dimension p must satisfy $p < h$ for the covariance matrix of any h-subset to be non-singular. In fact, for accuracy of the estimator it is often recommended to take $n > 5p$, e.g. in Rousseeuw et al. (2012). This limitation creates a gap in the availability of high breakdown methods for so-called "fat data", in which the number of rows (observations) is small compared to the number of columns (variables). To fill this gap we propose a modification of the MCD to make it applicable to high dimensions. The basic idea is to replace the subset-based covariance by a regularized covariance estimate, defined as a weighted average of the sample covariance of the h-subset and a predetermined positive definite target matrix. The proposed Minimum Regularized Covariance Determinant (MRCD) estimator is then the regularized covariance based on the h-subset which makes the overall determinant the smallest.

In addition to its availability for high dimensions, the main features of the MRCD estimator are that it preserves the good breakdown properties of the MCD estimator and is well-conditioned by construction. Since the estimated covariance matrix is guaranteed to be invertible it is suitable for computing robust distances, and for linear discriminant analysis and graphical modeling (Öllerer and Croux, 2015). Furthermore, we will generalize the C-step theorem of Rousseeuw and Van Driessen (1999) by showing that the objective function is reduced when concentrating the h-subset to the h observations with the smallest robust distance computed from the regularized covariance. This C-step theorem forms the theoretical basis for the proposed fast MRCD estimation algorithm.

The remainder of the paper is organized as follows. In Section 2 we introduce the MRCD covariance estimator and discuss its properties. Section 3 proposes a practical and fast algorithm for the MRCD. The extensive simulation study in Section 4 confirms the good properties of the method. Section 5 uses the MRCD estimator for outlier detection and regression analysis on real data sets from chemistry and criminology. The main findings and suggestions for further research are summarized in the conclusion.

2 From MCD to MRCD

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a dataset in which $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ denotes the i-th observation ($i = 1, \ldots, n$). The observations are stored in the $n \times p$ matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$. We assume that most of them come from an elliptical distribution with location $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$. The remaining observations can be arbitrary outliers, and we do not know beforehand which ones they are. The problem is to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ despite the outliers.

2.1 The MCD estimator

The MCD approach searches for an h-subset of the data (where $n/2 \leqslant h < n$) whose sample covariance matrix has the lowest possible determinant. Clearly, the subset size h affects the efficiency of the estimator as well as its robustness to outliers. For robustness, $n - h$ should be at least the number of outliers. When many outliers could occur one may set $h = \lceil 0.5n \rceil$. Typically one sets $h = \lceil 0.75n \rceil$ to get a better efficiency. Throughout the paper, H denotes a set of h indices reflecting the observations included in the subset, and $\mathcal{H}_h$ is the collection of all such sets. For a given H in $\mathcal{H}_h$ we denote the corresponding $h \times p$ submatrix of $\mathbf{X}$ by $\mathbf{X}_H$. Throughout the paper, we use the term h-subset to denote both H and $\mathbf{X}_H$ interchangeably. The mean and sample covariance matrix of $\mathbf{X}_H$ are then

$$\begin{aligned} \mathbf{m}_{\boldsymbol{X}}(H) &= h^{-1}\mathbf{X}_H'\mathbf{1}_h & (1) \\ \mathbf{S}_{\boldsymbol{X}}(H) &= h^{-1}(\mathbf{X}_H - \mathbf{m}_{\boldsymbol{X}}(H))'(\mathbf{X}_H - \mathbf{m}_{\boldsymbol{X}}(H)) \ . & (2) \end{aligned}$$

The MCD approach then aims to minimize the determinant of $\mathbf{S}_{\boldsymbol{X}}(H)$ among all $H \in \mathcal{H}_h$:

$$H_{MCD} = \underset{H \in \mathcal{H}_h}{\operatorname{argmin}} \left(\det(\mathbf{S}_{\boldsymbol{X}}(H))^{1/p} \right) \tag{3}$$

where we take the p-th root of the determinant for numerical reasons. Note that the p-th root of the determinant of the covariance matrix is the geometric mean of its eigenvalues; SenGupta (1987) calls it the standardized generalized variance. The MCD can also be seen as a multivariate least trimmed squares estimator in which the trimmed observations have the largest Mahalanobis distance with respect to the sample mean and covariance of the h-subset (Agulló et al., 2008).

The MCD estimate of location $\mathbf{m}_{MCD}$ is defined as the average of the h-subset, whereas the MCD scatter estimate is given as a multiple of its sample covariance matrix:

$$\begin{aligned} \mathbf{m}_{MCD} &= \mathbf{m}_{\boldsymbol{X}}(H_{MCD}) && (4) \\ \mathbf{S}_{MCD} &= c_\alpha \mathbf{S}_{\boldsymbol{X}}(H_{MCD}) && (5) \end{aligned}$$

where c_α is a consistency factor such as the one given by Croux and Haesbroeck (1999), and depends on the trimming percentage $\alpha = (n-h)/n$. Butler et al. (1993) and Cator and Lopuhaä (2012) prove consistency and asymptotic normality of the MCD estimator, and Lopuhaä and Rousseeuw (1991) show that it has the highest possible breakdown value (i.e., 50%) when $h = \lfloor (n+p+1)/2 \rfloor$. Accurately estimating a covariance matrix requires a sufficiently high number of observations. A rule of thumb is to require $n > 5p$ (Rousseeuw and Van Zomeren, 1990; Rousseeuw et al., 2012). When $p > h$ the MCD is ill-defined since all $\mathbf{S}_{\boldsymbol{X}}(H)$ have zero determinant.

2.2 The MRCD estimator

We will generalize the MCD estimator to high dimensions. As is common in the literature, we first standardize the p variables. For this we compute the median of each variable and stack them in a location vector $\nu_{\boldsymbol{X}}$. We also estimate the scale of each variable by the Qn estimator of Rousseeuw and Croux (1993), and put these scales in a diagonal matrix $\mathbf{D}_{\boldsymbol{X}}$.

The standardized observations are then

$$\boldsymbol{u}_i = \mathbf{D}_{\boldsymbol{X}}^{-1}(\boldsymbol{x}_i - \nu_{\boldsymbol{X}}) \ \ . \tag{6}$$

This disentangles the location-scale and correlation problems, as in Boudt et al. (2012).

In a second step, we use a predetermined and well-conditioned symmetric and positive definite target matrix $\mathbf{T}$. Following Won et al. (2013), we call such a matrix well-conditioned if the condition number (i.e., the ratio between the largest and smallest eigenvalues) is at most 1000. We also use a scalar weight coefficient ρ, henceforth called the regularization parameter. We then define the regularized covariance matrix of an h-subset H of the standardized data $\boldsymbol{U}$ as

$$\mathbf{K}(H) = \rho\ \mathbf{T} + (1-\rho)c_\alpha \mathbf{S}_{\boldsymbol{U}}(H) \tag{7}$$

where $\mathbf{S}_U(H)$ is as defined in (2) but for $\boldsymbol{U}$, and c_α is the same consistency factor as in (5).

It will be convenient to use the singular value decomposition $\mathbf{T} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'$ where $\boldsymbol{\Lambda}$ is the diagonal matrix holding the eigenvalues of $\mathbf{T}$ and $\mathbf{Q}$ is the orthogonal matrix holding the corresponding eigenvectors. We can then rewrite the regularized covariance matrix $\mathbf{K}(H)$ as

$$\mathbf{K}(H) = \mathbf{Q}\boldsymbol{\Lambda}^{1/2}[\rho\ \mathbf{I} + (1-\rho)c_\alpha \mathbf{S}_{\boldsymbol{W}}(H)]\boldsymbol{\Lambda}^{1/2}\mathbf{Q}' \tag{8}$$

where the $n \times p$ matrix $\boldsymbol{W}$ consists of the transformed standardized observations $\boldsymbol{w}_i = \boldsymbol{\Lambda}^{-1/2}\mathbf{Q}'\boldsymbol{u}_i$. It follows that $\mathbf{S}_{\boldsymbol{W}}(H) = \boldsymbol{\Lambda}^{-1/2}\mathbf{Q}'\mathbf{S}_{\boldsymbol{U}}(H)\mathbf{Q}\boldsymbol{\Lambda}^{-1/2}$.

The MRCD subset H_{MRCD} is defined by minimizing the determinant of the regularized covariance matrix $\mathbf{K}(H)$ in (8):

$$H_{MRCD} = \underset{H \in \mathcal{H}_h}{\operatorname{argmin}} \left(\det(\mathbf{K}(H))^{1/p}\right) \ \ . \tag{9}$$

Since $\mathbf{T}$, $\mathbf{Q}$ and $\boldsymbol{\Lambda}$ are fixed, H_{MRCD} can also be written as

$$H_{MRCD} = \underset{H \in \mathcal{H}_h}{\operatorname{argmin}} \left(\det(\rho\ \mathbf{I} + (1-\rho)c_\alpha \mathbf{S}_{\boldsymbol{W}}(H))^{1/p}\right) \ \ . \tag{10}$$

Once H_{MRCD} is determined, the MRCD location and scatter estimates of the original data matrix $\mathbf{X}$ are defined as

$$\mathbf{m}_{MRCD} = \nu_{\boldsymbol{X}} + \mathbf{D}_{\boldsymbol{X}}\mathbf{m}_{\boldsymbol{U}}(H_{MRCD}) \quad (11)$$

$$\mathbf{K}_{MRCD} = \mathbf{D}_{\boldsymbol{X}}\mathbf{Q}\boldsymbol{\Lambda}^{1/2}[\rho\,\mathbf{I} + (1-\rho)\mathbf{S}_{\boldsymbol{W}}(H_{MRCD})]\boldsymbol{\Lambda}^{1/2}\mathbf{Q}'\mathbf{D}_{\boldsymbol{X}}. \quad (12)$$

Because of the initial standardization step, the MRCD scatter estimate is location invariant and scale equivariant. This means that for any diagonal $p \times p$ matrix $\mathbf{A}$ and any $p \times 1$ vector $\mathbf{b}$ the MRCD scatter estimate $S(\mathbf{AX}+\mathbf{b})$ equals $\mathbf{AS}(\mathbf{X})\mathbf{A}'$.

2.3 The MRCD precision matrix

The precision matrix is the inverse of the scatter matrix, and is needed for the calculation of robust MRCD-based Mahalanobis distances, for linear discriminant analysis, for graphical modeling (Öllerer and Croux, 2015), and for many other computations. The MRCD scatter matrix (12) is computationally convenient to invert by its construction, yielding the expression

$$\mathbf{K}_{MRCD}^{-1} = \mathbf{D}_{\boldsymbol{X}}^{-1}\mathbf{Q}'\boldsymbol{\Lambda}^{-1/2}[\rho\,\mathbf{I}_p + (1-\rho)\mathbf{S}_{\boldsymbol{W}}(H_{MRCD})]^{-1}\boldsymbol{\Lambda}^{-1/2}\mathbf{Q}\mathbf{D}_{\boldsymbol{X}}^{-1} \ . \quad (13)$$

When $p > h$, a computationally more convenient form can be easily obtained by the Sherman-Morrison-Woodbury identity (Sherman and Morrison, 1950; Woodbury, 1950; Bartlett, 1951).

Note that the MRCD should not be confused with the Regularized Minimum Covariance Determinant (RMCD) estimator of Croux et al. (2012). The latter assumes sparsity of the precision matrix, and maximizes the penalized log-likelihood function of each $h-$subset by the GLASSO algorithm of Friedman et al. (2008). The repeated application of GLASSO is time-consuming.

2.4 Choice of target matrix and calibration of ρ

The MRCD estimate depends on two quantities: the target matrix $\mathbf{T}$ and the regularization parameter ρ. For the target matrix $\mathbf{T}$ on $\boldsymbol{U}$ we can take the identity matrix; relative to the original data $\boldsymbol{X}$ this is the diagonal matrix with the robustly estimated univariate scales on the diagonal. Depending on the application, we can also take a non-diagonal target matrix

$\mathbf{T}$. When this matrix is estimated in a first step, it should be robust to outliers in the data. A reasonable choice is to compute a rank correlation matrix of $\boldsymbol{U}$, which incorporates some of the relation between the variables. When we have reasons to suspect an equicorrelation structure, we can set $\mathbf{T}$ equal to

$$\mathbf{R}_c = c\mathbf{J}_p + (1-c)\mathbf{I}_p \tag{14}$$

with $\mathbf{J}_p$ the $p \times p$ matrix of ones, $\mathbf{I}_p$ the identity matrix, and $-1/(p-1) < c < 1$ to ensure positive definiteness. The parameter c in the equicorrelation matrix (14) can be estimated by averaging robust correlation estimates over all pairs of variables, under the constraint that the determinant of $\mathbf{R}_c$ is above a minimum threshold value.

When the regularization parameter ρ equals zero $\mathbf{K}(H)$ becomes the sample covariance $\mathbf{S}_{\boldsymbol{U}}(H)$, and when ρ equals one $\mathbf{K}(H)$ becomes the target. We require $0 \leqslant \rho \leqslant 1$ to ensure that $\mathbf{K}(H)$ is positive definite (as it is a convex combination of positive definite matrices), hence invertible.

In practice, we recommend a data-driven approach which sets ρ at the lowest nonnegative value for which $\rho\, \mathbf{I} + (1-\rho)c_\alpha \mathbf{S}_{\boldsymbol{W}}(H)$ is well-conditioned. This is easy to implement, as we only need to compute the eigenvalues λ of $c_\alpha \mathbf{S}_{\boldsymbol{W}}(H)$ once, since the eigenvalues of $\rho\, \mathbf{I} + (1-\rho)c_\alpha \mathbf{S}_{\boldsymbol{W}}(H)$ equal

$$\rho + (1-\rho)\lambda. \tag{15}$$

Note that by this heuristic we only use regularization when needed. Indeed, if $\mathbf{S}_{\boldsymbol{W}}(H)$ is well-conditioned, the heuristic sets ρ equal to zero. Also note that the eigenvalues in (15) are at least ρ, so the smallest eigenvalue of the MRCD scatter estimate is bounded away from zero. Therefore the MRCD scatter estimator has a 100% implosion breakdown value, compared to the 50% implosion breakdown value of the MCD.

3 An algorithm for the MRCD estimator

A naive algorithm for the optimization problem (9) would be to compute $\det(\mathbf{K}(H))$ for every possible h-subset H. However, for realistic sample sizes this type of brute force evaluation is infeasible.

The original MCD estimator (3) has the same issue. The current solution for the MCD consists of either selecting a large number of randomly chosen initial subsets (Rousseeuw and Van Driessen, 1999) or starting from a smaller number of deterministic subsets (Hubert et al., 2012). In either case one iteratively applies so-called C-steps. The C-step of MCD improves an h-subset H_1 by computing its mean and covariance matrix, and then puts the h observations with smallest Mahalanobis distance in a new subset H_2. The C-step theorem of Rousseeuw and Van Driessen (1999) proves that the covariance determinant of H_2 is lower than or equal to that of H_1 , so C-steps lower the MCD objective function.

We will now generalize this theorem to regularized covariance matrices.

Theorem 1. *Starting from an h-subset H_1 , one can compute $\mathbf{m}_1 = \frac{1}{h}\sum_{i\in H_1}\mathbf{x}_i$ and $\mathbf{S}_1 = \frac{1}{h}\sum_{i\in H_1}(\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)'$. The matrix*

$$\mathbf{K}_1 = \rho\mathbf{T} + (1-\rho)\mathbf{S}_1$$

is positive definite hence invertible, so we can compute

$$d_1(i) = (\mathbf{x}_i - \mathbf{m}_1)'\mathbf{K}_1^{-1}(\mathbf{x}_i - \mathbf{m}_1)$$

for $i = 1, \ldots, n$. Let H_2 be an h-subset for which

$$\sum_{i\in H_2} d_1(i) \leq \sum_{i\in H_1} d_1(i) \tag{16}$$

and compute $\mathbf{m}_2 = \frac{1}{h}\sum_{i\in H_2}\mathbf{x}_i$, $\mathbf{S}_2 = \frac{1}{h}\sum_{i\in H_2}(\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)'$ and $\mathbf{K}_2 = \rho\mathbf{T} + (1-\rho)\mathbf{S}_2$. Then

$$\det(\mathbf{K}_2) \leq \det(\mathbf{K}_1) \tag{17}$$

with equality if and only if $\mathbf{m}_2 = \mathbf{m}_1$ and $\mathbf{K}_2 = \mathbf{K}_1$.

Note that for $\rho = 0$ our Theorem 1 specializes to Theorem 1 of Rousseeuw and Van Driessen (1999), but with the weaker condition (16), thereby strengthening the result. The proof of Theorem 1 is given in Appendix A.

Making use of the generalized C-step we can now construct the actual algorithm to find the MRCD subset in step 3 of the pseudocode.

MRCD algorithm

1. Compute the standardized observations $\boldsymbol{u}_i$ as defined in (6) using the median and the Qn estimator for univariate location and scale.

2. Perform the singular value decomposition of $\mathbf{T}$ into $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'$ where $\boldsymbol{\Lambda}$ is the diagonal matrix holding the eigenvalues of $\mathbf{T}$ and $\mathbf{Q}$ is the orthogonal matrix whose columns are the corresponding eigenvectors. Compute $\boldsymbol{w}_i = \boldsymbol{\Lambda}^{-1/2}\mathbf{Q}'\boldsymbol{u}_i$.

3. Find the MRCD subset:

   3.1. Follow Subsection 3.1 in Hubert et al. (2012) to obtain six robust, well-conditioned initial location estimates $\boldsymbol{m}^i$ and scatter estimates $\mathbf{S}^i$ $(i = 1, \ldots, 6)$.

   3.2. Determine the subsets H_0^i of $\boldsymbol{W}$ containing the h observations with lowest Mahalanobis distance in terms of $\boldsymbol{m}^i$ and $\mathbf{S}^i$.

   3.3. For each subset H_0^i, determine the smallest value of $0 \leq \rho^i < 1$ for which $\rho^i\,\mathbf{I} + (1-\rho^i)c_\alpha \mathbf{S}_{\boldsymbol{W}}(H_0^i)$ is well-conditioned. Denote this value as ρ_0^i .

   3.4. If $\max_i \rho_0^i \leq 0.1$, set $\rho = \max_i \rho_0^i$, else set $\rho = \max\{0.1; \text{median}_i \rho_0^i\}$.

   3.5. For the initial subset H_0^i for which $\rho_0^i \leq \rho$, repeat the generalized C-steps from Theorem 1 using $\rho\,\mathbf{I} + (1-\rho)c_\alpha \mathbf{S}_{\boldsymbol{W}}(H_0^i)$ until convergence. Denote the resulting subsets as H^i .

   3.6. Let H_{MRCD} be the subset for which $\rho\,\mathbf{I} + (1-\rho)c_\alpha \mathbf{S}_{\boldsymbol{W}}(H^i)$ has the lowest determinant among the candidate subsets.

4. From H_{MRCD} compute the final MRCD location and scatter estimates as in (12).

In Step 3.1, we first determine the initial scatter estimates $\boldsymbol{S}^i$ of $\boldsymbol{W}$ in the same way as in the DetMCD algorithm of Hubert et al. (2012). This includes the use of steps 4a and 4b of the OGK algorithm of Maronna and Zamar (2002) to correct for inaccurate eigenvalues and guarantee positive definiteness of the initial estimates. For completeness, the OGK algorithm is provided in Appendix B. Given the six initial location and scatter estimates, we then determine in step 3.2 the corresponding six initial subsets of h observations with the lowest Mahalanobis distance. In step 3.3, we compute, for each subset, a regularized covariance, where we use line search and formula (15) to calibrate the regularization parameter in such a way that the corresponding condition number is at most 1000. This leads to potentially six different regularization parameters ρ_i.

To ensure comparability of the MRCD covariance estimates on different subsets, we need a unique regularization parameter. In step 3.4, we set by default the final value of the regularization parameter ρ as the largest value of the initial regularization parameters. This is a conservative choice ensuring that the MRCD covariance computed on each subset is well-conditioned. In case of outliers in one of the initial subsets, this may however lead to a too large value of the regularization parameter. To safeguard the estimation against this outlier inflation of ρ, we change the default choice, when the largest value of all initial ρ_i's exceeds 0.1. We then set the regularization parameter at the median value of the initial regularization parameters, when this median value exceeds 0.1. Otherwise we take 0.1. In the simulation study, we find that in practice ρ tends to be well below 0.1, as long as the MRCD is implemented with a subset size h that is small enough to resist the outlier contamination. A robust implementation of the MRCD thus ensures that regularization is only used when needed.

In step 3.6, we recalculate the regularized covariance using ρ instead of ρ_i for each subset with $\rho_i \leq \rho$. We then apply C-steps until the subset no longer changes, which typically requires only a few steps. Finally, out of the resulting subsets we select the one with the lowest objective value, and use it to compute our final location and scatter estimates according to (12).

4 Simulation study

We now investigate the empirical performance of the MRCD. We compare the MRCD estimator to the OGK estimator of Maronna and Zamar (2002), which can also robustly estimate location and scatter in high dimensions but by itself does not guarantee that the scatter matrix is well-conditioned. The OGK estimator, as described in Appendix B, does not result from optimizing a explicit objective function like the M(R)CD approach. Nevertheless it often works well in practice.

Data generation setup. In the simulation experiment we generated $M = 500$ contaminated samples of size n from a p-variate normal distribution, with $n \times p$ taken as either 800×100, 200×100, 200×200 and 200×400. Since both the MRCD and OGK estimators are location and scale equivariant, we follow Agostinelli et al. (2015), henceforth ALYZ, by assuming without loss of generality that the mean $\boldsymbol{\mu}$ is $\mathbf{0}$, and that the diagonal elements of $\boldsymbol{\Sigma}$ are all equal to unity. As in ALYZ, we account for the lack of affine equivariance of the proposed MRCD estimator by generating in each replication the correlation matrix randomly such that the performance of the estimator is not tied to a particular choice of correlation matrix. We use the procedure of Section 4 in ALYZ, including the iterative correction to ensure that the condition number of the generated correlation matrix is within a tolerance interval around 100. To contaminate the data sets, we follow Maronna and Zamar (2002) and randomly replace $\lfloor \varepsilon n \rfloor$ observations by outliers along the eigenvector direction of $\boldsymbol{\Sigma}$ with smallest eigenvalue, since this is the direction where the contamination is hardest to detect. The distance between the outliers and the mean of the good data is denoted by k, which is set to 50 for medium-sized outlier contamination and to 100 for far outliers. We let the fraction of contamination ε be either 0% (clean data), 20% or 40%.

Evaluation setup. On each generated data set we run the MRCD with different subset sizes h, taken as 50%, 75%, and 100% of the sample size n, and compare the results obtained when using $\rho = 0$ (MCD estimator) versus using the data-driven choice of ρ with the condition number at most 1000. As the target matrix, we take either the identity matrix ($\mathbf{T} = \mathbf{I}_p$) or the equicorrelation matrix ($\mathbf{T} = \mathbf{R}_c$), with equicorrelation parameter robustly estimated as the average Kendall rank correlation. We measure the inaccuracy of our scatter estimates $\boldsymbol{S}_m$ compared to the true covariance $\boldsymbol{\Sigma}_m$ by their mean squared error (MSE) given by

$$MSE = \frac{1}{M}\frac{1}{p^2}\sum_{m=1}^{M}\sum_{k=1}^{p}\sum_{l=1}^{p}\left(\mathbf{S}_m - \boldsymbol{\Sigma}_m\right)_{k,l}^2 \quad .$$

Note that the true $\boldsymbol{\Sigma}_m$ differs across values of m when generating data according to ALYZ.

Table 1: Mean squared error of the MRCD and OGK scatter matrices, together with the average value of ρ, across 500 replications of the ALYZ data generating process.

	MSE				*Average value of ρ*			
	800×100	200×100	200×200	200×400	800×100	200×100	200×200	200×400
Panel A: Clean data								
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{I}_p$	0.0024	0.0087	0.0085	0.0105	0	0.0047	0.0067	0.0108
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{R}_c$	0.0024	0.0087	0.0085	0.0106	0	0.0050	0.0067	0.0110
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{I}_p$	0.0017	0.0064	0.0062	0.0066	0	0.0001	0.0054	0.0080
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{R}_c$	0.0017	0.0064	0.0062	0.0065	0	0.0001	0.0054	0.0081
$h = n$, $\mathbf{T} = \mathbf{I}_p$	0.0013	0.0050	0.0050	0.0049	0	0	0.0047	0.0064
$h = n$, $\mathbf{T} = \mathbf{R}_c$	0.0013	0.0050	0.0050	0.0049	0	0	0.0046	0.0065
OGK	0.0015	0.0060	0.0058	0.0058				
Panel B: 20% *contamination,* $k = 50$								
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{I}_p$	0.0024	0.0088	0.0089	0.0102	0	0.0023	0.0032	0.0053
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{R}_c$	0.0024	0.0088	0.0089	0.0102	0	0.0024	0.0033	0.0052
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{I}_p$	0.0017	0.0066	0.0065	0.0066	0	0	0.0027	0.0039
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{R}_c$	0.0017	0.0066	0.0065	0.0066	0	0	0.0027	0.0039
$h = n$, $\mathbf{T} = \mathbf{I}_p$	17.4482	15.6942	7.9168	4.3830	0.0220	0.1234	0.1828	0.2251
$h = n$, $\mathbf{T} = \mathbf{R}_c$	17.5846	15.6247	7.9131	4.3830	0.0182	0.1253	0.1830	0.2251
OGK	0.0079	0.0187	0.0159	0.0146				
Panel C: 20% *contamination,* $k = 100$								
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{I}_p$	0.0024	0.0088	0.0089	0.0102	0	0.0023	0.0030	0.0049
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{R}_c$	0.0024	0.0089	0.0089	0.0102	0	0.0022	0.0031	0.0049
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{I}_p$	0.0017	0.0066	0.0065	0.0066	0	0	0.0024	0.0037
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{R}_c$	0.0017	0.0065	0.0065	0.0066	0	0	0.0026	0.0036
$h = n$, $\mathbf{T} = \mathbf{I}_p$	59.1521	47.9720	23.4138	12.2040	0.0989	0.2311	0.2957	0.3528
$h = n$, $\mathbf{T} = \mathbf{R}_c$	58.9257	48.3764	23.4233	12.2020	0.1006	0.2279	0.2956	0.3529
OGK	0.0087	0.0217	0.0173	0.0160				
Panel D: 40% *contamination,* $k = 50$								
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{I}_p$	0.0025	0.0094	0.0095	0.0099	0	0.0011	0.0015	0.0022
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{R}_c$	0.0025	0.0094	0.0095	0.0100	0	0.0011	0.0015	0.0022
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{I}_p$	2.8783	3.462	3.5194	3.1857	0	0.0227	0.0506	0.0842
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{R}_c$	2.9372	3.4802	3.5407	3.2100	0	0.0237	0.0509	0.0841
$h = n$, $\mathbf{T} = \mathbf{I}_p$	66.8405	60.5137	30.6172	16.4693	0.0367	0.1055	0.1468	0.1736
$h = n$, $\mathbf{T} = \mathbf{R}_c$	66.0929	60.4384	30.6125	16.4699	0.0421	0.1059	0.1469	0.1736
OGK	0.0398	0.0744	0.0557	0.0477				
Panel E: 40% *contamination,* $k = 100$								
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{I}_p$	0.0025	0.0094	0.0095	0.0100	0	0.0009	0.0012	0.0018
$h = \lceil 0.5n \rceil$, $\mathbf{T} = \mathbf{R}_c$	0.0025	0.0094	0.0095	0.0100	0	0.0009	0.0012	0.0019
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{I}_p$	10.8897	13.2065	13.1731	11.1849	0	0.0423	0.0817	0.1212
$h = \lceil 0.75n \rceil$, $\mathbf{T} = \mathbf{R}_c$	11.0477	13.0273	13.1299	11.2233	0	0.0423	0.0821	0.1215
$h = n$, $\mathbf{T} = \mathbf{I}_p$	236.0263	205.1611	101.3645	52.8403	0.0953	0.1757	0.2235	0.2599
$h = n$, $\mathbf{T} = \mathbf{R}_c$	235.1515	205.3006	101.3751	52.8334	0.0969	0.1755	0.2234	0.2599
OGK	0.0589	0.1074	0.0789	0.0658				

Discussion of results. The results are reported in Table 1, where the left panel shows the MSE and the right panel lists the average value of the data-driven ρ for the MRCD covariance estimator. The top panel shows the results in the absence of outlier contamination, i.e. $\varepsilon = 0\%$. In the lower panels we see the effects of 20% and 40% contamination in the data, for different values of k.

In terms of the MSE of the covariance estimates, we find that, in the case of no outlier contamination, the MRCD covariance estimate with $h = n$ has the lowest MSE, and that the OGK covariance estimator is second best. The MSE of the MRCD estimator with $h = \lceil 0.5n \rceil$ is almost double the MSE of the MRCD covariance obtained when $h = n$. This lower efficiency is compensated by the high breakdown robustness. In fact, for both 20% and 40% outlier contamination, the MSE of the MRCD with $h = \lceil 0.5n \rceil$ is similar to the one in absence of outliers, and it is substantially lower than the MSE of the OGK covariance estimator.

The simulation study also sheds light on how the structure of the data and the presence of outlier contamination affect the calibration of the regularization parameter ρ. Recall that the MRCD uses the smallest value of $0 \leqslant \rho < 1$ for which the scatter matrix is well-conditioned, so when the MCD is well-conditioned the MRCD also obtains $\rho = 0$ and thus coincides with the MCD in that case. We indeed find that $\rho = 0$ in the scenarios where $h > p$ and $h < n(1 - \epsilon)$, and that ρ remains close to zero when the subset size h is small enough to resist the outlier contamination. It follows that the choice between the identity matrix or the robustly calibrated equicorrelation matrix as target matrix has only a negligible impact on the MSE, provided the MRCD is implemented with a subset size h that is small enough to resist the outlier contamination. When the number of outliers exceeds the subset size, we see that outliers induce higher ρ values.

In conclusion, the simulation study confirms that the MRCD is a good method for estimating location and scatter in high dimensions. It only regularizes when needed. When h is less than p and the number of clean observations, the resulting ρ is typically less than 0.05, implying that the MRCD strikes a balance between being similar to the MCD for tall data and achieving a well-conditioned estimate in the case of fat data.

5 Real data examples

We illustrate the MRCD on two datasets with low n/p, so using the original MCD is not indicated. The MRCD is implemented using the identity matrix as target matrix.

5.1 Octane data

The octane data set described in Esbensen et al. (1996) consists of near-infrared absorbance spectra with $p = 226$ wavelengths collected on $n = 39$ gasoline samples. It is known that the samples 25, 26, 36, 37, 38 and 39 are outliers which contain added ethanol (Hubert et al., 2005). Of course, in most applications the number of outliers is not known in advance hence it is not obvious to set the subset size h. The choice of h matters because increasing h improves the efficiency at uncontaminated data but hurts the robustness to outliers. Our recommended default choice is $h = \lceil 0.75n \rceil$, safeguarding the MRCD covariance estimate against up to 25% of outliers.

Alternatively, one could employ a data-driven approach to select h. It consists of computing the MRCD for a range of h values, and looking for an important change in the objective function or the estimates at some value of h. This is not too hard, since we only need to obtain the initial estimates $\boldsymbol{S}^i$ once. Figure 1 plots the MRCD objective function (10) for each value of h, while Figure 2 shows the Frobenius distance between the MRCD scatter matrices of the standardized data ($i.e.$, $\rho\, \mathbf{I} + (1-\rho)\mathbf{S}_{\boldsymbol{W}}(H_{MRCD})$), as defined in (12)) obtained for $h-1$ and h. Both figures clearly indicate that there is an important change at $h = 34$, so we choose $h = 33$. The total computation time to produce these plots was only 12 seconds on an Intel(R) Core(TM) i7-5600U CPU with 2.60 GHz.

We then calculate the MRCD estimator with $h = 33$, yielding $\rho = 0.1149$. The condition number of the scatter matrix equals 720. Figure 3 shows the corresponding robust distances

$$RD(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \mathbf{m}_{MRCD})'\mathbf{K}_{MRCD}^{-1}(\boldsymbol{x}_i - \mathbf{m}_{MRCD})} \tag{18}$$

where $\mathbf{m}_{MRCD}$ and $\mathbf{K}_{MRCD}$ are the MRCD location and scatter estimates of (12). The flagged outliers (red triangles) stand out, showing the MRCD has correctly identified the 6 samples with added ethanol.

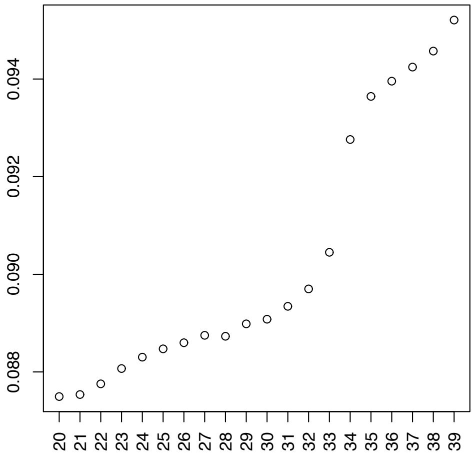

Figure 1: Octane data: MRCD objective value (10) for different values of h.

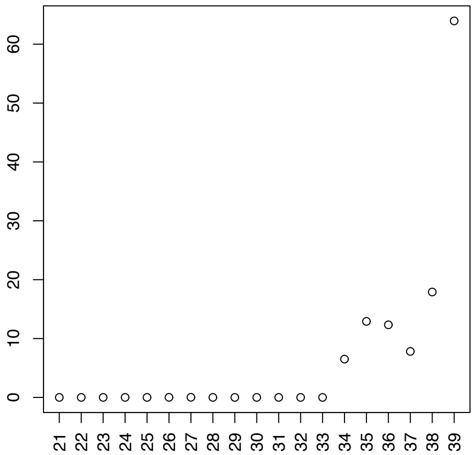

Figure 2: Octane data: Frobenius distance between MRCD scatter matrices on standardized data for $h-1$ and h.

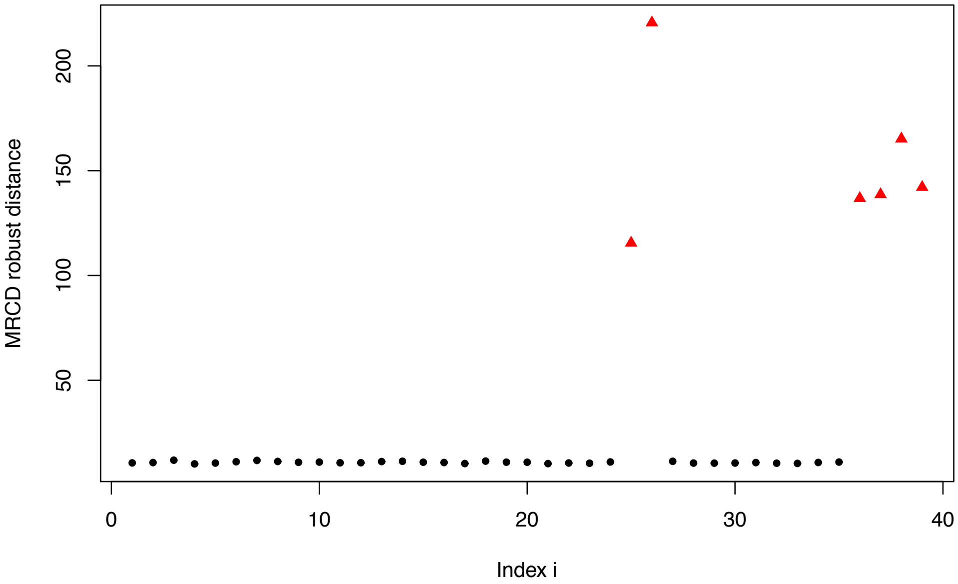

Figure 3: Robust distances of the octane data, based on the MRCD with $h = 33$.

5.2 Murder rate data

Khan et al. (2007) regress the murder rate per 100,000 residents in the $n = 50$ states of the US in 1980 on 25 demographic predictors, and mention that graphical tools reveal one clear outlier.

For lower-dimensional data, Rousseeuw et al. (2004) applied the MCD estimator to the response(s) and predictors together to robustly estimate a multivariate regression. Here we investigate whether for high-dimensional data the same type of analysis can be carried out based on the MRCD. In the murder rate data this yields a total of 26 variables.

As for the octane data, we compute the MRCD estimates for the candidate range of h. In Figure 4 we see a big jump in the objective function when going from $h = 49$ to $h = 50$. But in the plot of the Frobenius distance between successive MRCD scatter matrices (Figure 5) we see evidence of four outliers, which lead to a substantial change in the MRCD when included in the subset.

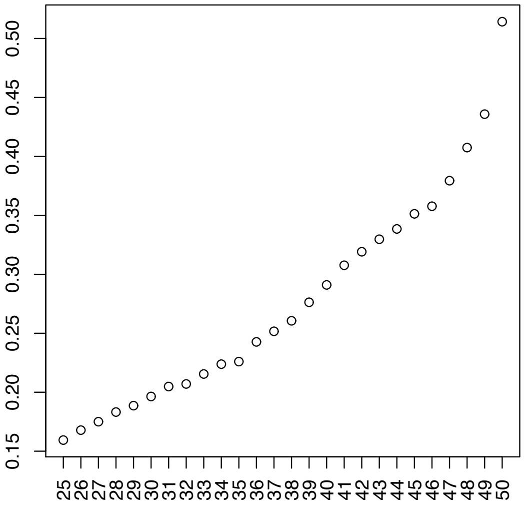

Figure 4: Murder rate data: MRCD objective value (10) for different values of h.

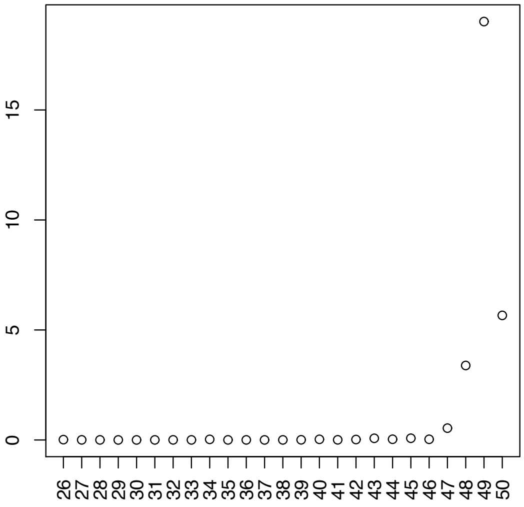

Figure 5: Murder rate data: Frobenius distance between MRCD scatter matrices on standardized data for $h - 1$ and h.

As a conservative choice we set $h = 44$, which allows for up to 6 outliers. We then partition the MRCD scatter matrix on all 26 variables as follows:

$$\mathbf{K}_{MRCD} = \begin{pmatrix} \mathbf{K}_{xx} & \mathbf{K}_{xy} \\ \mathbf{K}_{xy} & \mathbf{K}_{yy} \end{pmatrix},$$

where x stands for the vector of predictors and y is the response variable. The resulting estimate of the slope vector is then

$$\hat{\beta}_{MRCD} = \mathbf{K}_{xx}^{-1}\mathbf{K}_{xy} \, .$$

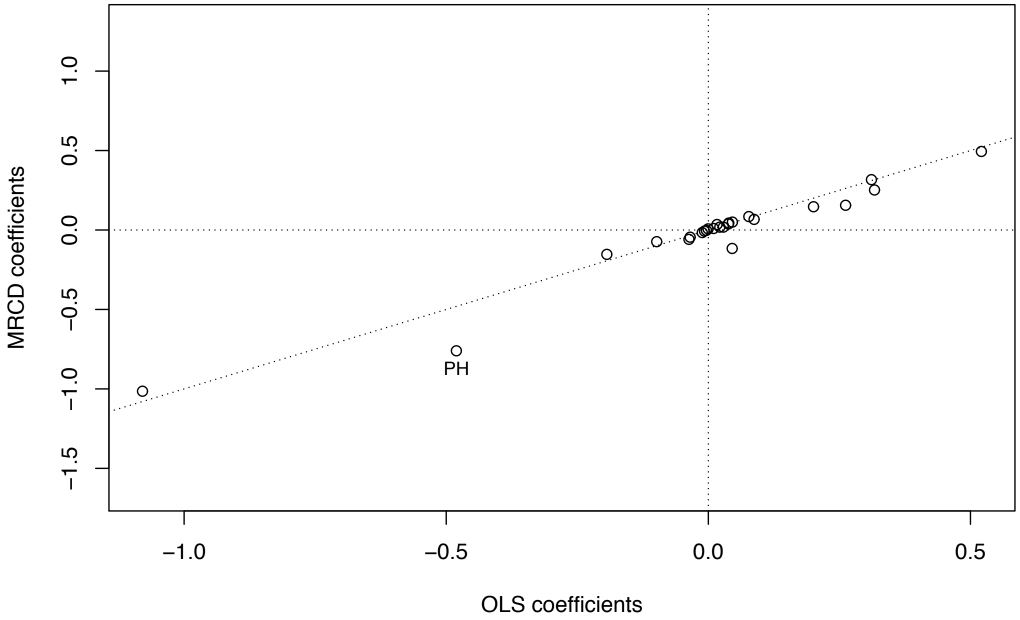

Figure 6: MRCD slopes versus OLS slopes of regressing the murder rate on demographic variables.

The resulting MRCD and OLS slope coefficients are shown in Figure 6. For most variables the coefficients are similar, except for the variable "PH" corresponding to the number of telephones per 100 residents, for which the MRCD coefficient (-0.76) is about 1.5 times the OLS coefficient (-0.48). The telephone density might serve as a proxy for the technological level of the state. A negative coefficient then indicates that on average the more technologically advanced the state, the lower the murder rate, other things being equal (which they rarely are).

Figure 7 plots the murder rate against the phone density. In this scatter plot, Arkansas and Nevada appear as outliers. Arkansas is a bad leverage point: it has the lowest phone density (in 1980) but an average murder rate. Nevada is a vertical outlier, as it lies above the downward sloping regression line fitting the bulk of the data, meaning that its murder rate is higher than one would expect on the basis of the phone density alone. The red triangles are the observations not included in the MRCD subset. We see that MRCD regression on all predictors has effectively flagged Arkansas and Nevada. Omitting them, as implicitly done in the MRCD regression, has led to a more negative value of this slope.

Finally, we note that MRCD regression can be plugged into existing robust algorithms

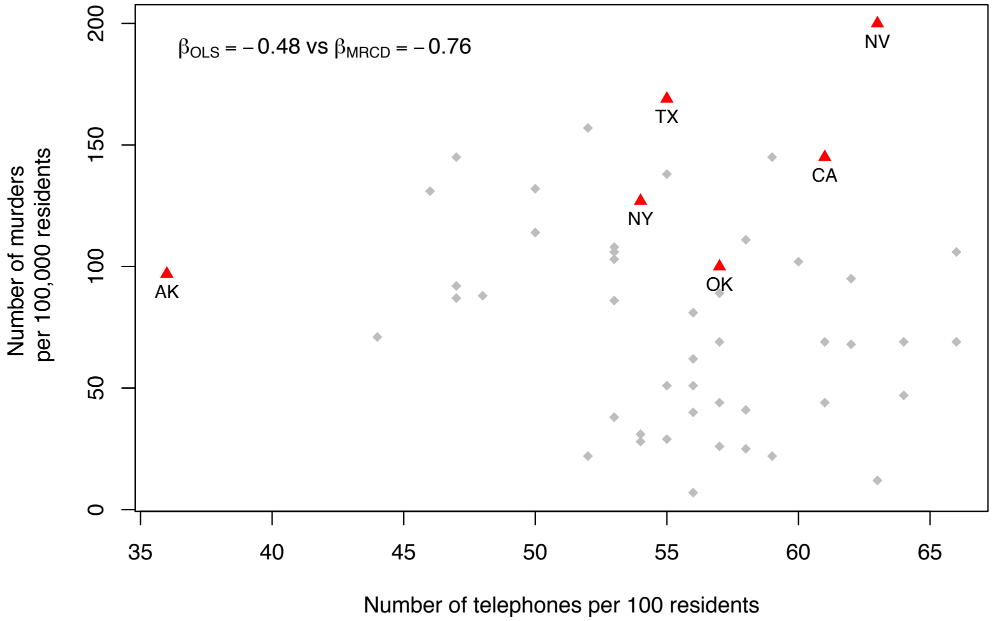

Figure 7: Scatter plot of murder rate per state against phone density. The red triangles indicate the observations that are not included in the final MRCD subset with $h = 44$.

for variable selection, which avoids the limitation mentioned in Khan et al. (2007) that "a robust fit of the *full* model may not be feasible due to the numerical complexity of robust estimation when [the dimension] d is large (e.g., $d \geq 200$) or simply because d exceeds the number of cases, n." The MRCD could be used in such situations because its computation remains feasible in higher dimensions.

6 Concluding remarks

In this paper we generalize the Minimum Covariance Determinant estimation approach of Rousseeuw (1985) to higher dimensions, by regularizing the sample covariance matrices of subsets before minimizing their determinant. The resulting Minimum Regularized Covariance Determinant (MRCD) estimator is well-conditioned by construction, even when $p > n$, and preserves the good robustness of the MCD. We were able to construct a fast algorithm for the MRCD by generalizing the C-step used by the MCD, and proving that this generalized C-step is guaranteed to reduce the covariance determinant. We verified the performance of the MRCD estimator in an extensive simulation study including both clean and contaminated data. The simulation study also confirms that the MRCD can be interpreted as a generalization of the MCD, because when n is sufficiently large compared to p and the MCD is well-conditioned the regularization parameter in MRCD becomes zero so the MRCD estimate coincides with the MCD. Finally, we illustrated the use of the MRCD for outlier detection and robust regression on two fat data applications from chemistry and criminology, for which $p > n/2$.

We believe that the MRCD is a valuable addition to the tool set for robust multivariate analysis, especially in high dimensions. We look forward to further research on its use in principal component analysis where the original MCD has proved useful (Croux and Haesbroeck, 2000; Hubert et al., 2005), and analogously in factor analysis (Pison et al., 2003), classification (Hubert and Van Driessen, 2004), clustering (Hardin and Rocke, 2004), multivariate regression (Rousseeuw et al., 2004), penalized maximum likelihood estimation (Croux et al., 2012) and other multivariate techniques. A further research topic is to study the finite sample distribution of the robust distances computed from the MRCD. Our experiments have shown that the usual chi-square and F-distribution results for the MCD distances (Hardin and Rocke, 2005) are no longer good approximations when p is large relatively to n. A better approximation would be useful for improving the accuracy of the MRCD by reweighting.

Acknowledgments. This research has benefited from the financial support of the Flemish Science Foundation (FWO) and project C16/15/068 of Internal Funds KU Leuven. We are grateful to Yukai Yang for his initial assistance to this work. We also thank Dries Cornilly, Christophe Croux, Sebastiaan Höppner, Stefan Van Aelst and Marjan Wauters for their constructive comments.

References

Agostinelli, C., A. Leung, V. Yohai, and R. Zamar (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test 24*(3), 441–461.

Agulló, J., C. Croux, and S. Van Aelst (2008). The multivariate least trimmed squares estimator. *Journal of Multivariate Analysis 99*, 311–338.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics 22*(1), 107–111.

Boudt, K., J. Cornelissen, and C. Croux (2012). Jump robust daily covariance estimation by disentangling variance and correlation components. *Computational Statistics & Data Analysis 56*(11), 2993–3005.

Butler, R., P. Davies, and M. Jhun (1993). Asymptotics for the Minimum Covariance Determinant estimator. *The Annals of Statistics 21*(3), 1385–1400.

Cator, E. and H. Lopuhaä (2012). Central limit theorem and influence function for the MCD estimator at general multivariate distributions. *Bernoulli 18*(2), 520–551.

Croux, C. and C. Dehon (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications 19*(4), 497–515.

Croux, C., S. Gelper, and G. Haesbroeck (2012). Regularized Minimum Covariance Determinant estimator. *Mimeo.*

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.

Croux, C. and G. Haesbroeck (2000). Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika 87*, 603–618.

Esbensen, K., T. Midtgaard, and S. Schönkopf (1996). *Multivariate Analysis in Practice: A Training Package.* Camo As.

Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*, 186–197.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(2), 432–441.

Gnanadesikan, R. and J. Kettenring (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics 28*, 81–124.

Grübel, R. (1988). A minimal characterization of the covariance matrix. *Metrika 35*(1), 49–52.

Hardin, J. and D. Rocke (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis 44*, 625–638.

Hardin, J. and D. Rocke (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics 14*(4), 928–946.

Hubert, M., P. Rousseeuw, and S. Van Aelst (2008). High breakdown robust multivariate methods. *Statistical Science 23*, 92–119.

Hubert, M., P. Rousseeuw, and K. Vanden Branden (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics 47*, 64–79.

Hubert, M., P. Rousseeuw, and T. Verdonck (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics 21*(3), 618–637.

Hubert, M. and K. Van Driessen (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis 45*, 301–320.

Khan, J., S. Van Aelst, and R. H. Zamar (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association 102*(480), 1289–1299.

Lopuhaä, H. and P. Rousseeuw (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics 19*, 229–248.

Maronna, R. and R. H. Zamar (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics 44*(4), 307–317.

Öllerer, V. and C. Croux (2015). Robust high-dimensional precision matrix estimation. In *Modern Nonparametric, Robust and Multivariate Methods*, pp. 325–350. Springer.

Pison, G., P. Rousseeuw, P. Filzmoser, and C. Croux (2003). Robust factor analysis. *Journal of Multivariate Analysis 84*, 145–172.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*(388), 871–880.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (Eds.), *Mathematical Statistics and Applications, Vol. B*, pp. 283–297. Reidel Publishing Company, Dordrecht.

Rousseeuw, P. and C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association 88*(424), 1273–1283.

Rousseeuw, P., C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler (2012). Robustbase: Basic Robust Statistics. R package version 0.92-3.

Rousseeuw, P., S. Van Aelst, K. Van Driessen, and J. Agulló (2004). Robust multivariate regression. *Technometrics 46*, 293–305.

Rousseeuw, P. and K. Van Driessen (1999). A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics 41*, 212–223.

Rousseeuw, P. and B. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association 85*(411), 633–639.

SenGupta, A. (1987). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *Journal of Multivariate Analysis 23*(2), 209–219.

Sherman, J. and W. J. Morrison (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics 21*(1), 124–127.

Won, J.-H., J. Lim, S.-J. Kim, and B. Rajaratnam (2013). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 427–450.

Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum report 42*, 106.

Appendix A: Proof of Theorem 1

Generate a p-variate sample $\mathbf{Z}$ with $p+1$ points for which $\boldsymbol{\Lambda} = \frac{1}{p+1}\sum_{j=1}^{p+1}(\boldsymbol{z}_i - \overline{z})(\boldsymbol{z}_i - \overline{z})'$ is non-singular and $\overline{z} = \frac{1}{p+1}\sum_{j=1}^{p+1}\boldsymbol{z}_i$. Then $\tilde{\boldsymbol{z}}_i = \boldsymbol{\Lambda}^{-1/2}(\boldsymbol{z}_i - \overline{z})$ has mean zero and covariance matrix $\mathbf{I}_p$. Now compute $\boldsymbol{y}_i = \mathbf{T}^{1/2}\tilde{\boldsymbol{z}}_i$, hence $\mathbf{Y}$ has mean zero and covariance matrix $\mathbf{T}$.

Next, create the artificial dataset

$$\tilde{\mathbf{X}}^1 = \big(w_1(\boldsymbol{x}_1^1 - \mathbf{m}_1), \ldots, w_h(\boldsymbol{x}_h^1 - \mathbf{m}_1), w_{h+1}\boldsymbol{y}_1, \ldots, w_k\boldsymbol{y}_{p+1}\big)$$

with $k = h + p + 1$ points, where $\boldsymbol{x}_1^1, \ldots, \boldsymbol{x}_h^1$ are the members of H_1. The factors w_i are given by

$$w_i = \begin{cases} \sqrt{k(1-\rho)/h} & \text{for } i = 1, \ldots, h \\ \sqrt{k\rho/(p+1)} & \text{for } i = h+1, \ldots, k \ . \end{cases}$$

The mean and covariance matrix of $\tilde{\mathbf{X}}^1$ are then

$$\frac{1}{k}\sum_{i=1}^{k}\tilde{\boldsymbol{x}}_i^1 = \sqrt{\frac{1-\rho}{kh}}\sum_{i=1}^{h}(\boldsymbol{x}_i^1 - \mathbf{m}_1) + \sqrt{\frac{\rho}{k(p+1)}}\sum_{j=1}^{p+1}\boldsymbol{y}_j = 0$$

and

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^{k}\tilde{\boldsymbol{x}}_i^1(\tilde{\boldsymbol{x}}_i^1)' &= \frac{1-\rho}{h}\sum_{i=1}^{h}(\boldsymbol{x}_i^1 - \mathbf{m}_1)(\boldsymbol{x}_i^1 - \mathbf{m}_1)' + \frac{\rho}{p+1}\sum_{j=1}^{p+1}\boldsymbol{y}_j\boldsymbol{y}_j' \\ &= (1-\rho)\mathbf{S}_1 + \rho\mathbf{T} = \mathbf{K}_1 \ .\end{aligned}$$

The regularized covariance matrix $\mathbf{K}_1$ is thus the actual covariance matrix of the combined data set $\tilde{\mathbf{X}}^1$. Analogously we construct

$$\tilde{\mathbf{X}}^2 = \big(w_1(\boldsymbol{x}_1^2 - \mathbf{m}_2), \ldots, w_h(\boldsymbol{x}_h^2 - \mathbf{m}_2), w_{h+1}\boldsymbol{y}_1, \ldots, w_k\boldsymbol{y}_{p+1}\big)$$

where $\boldsymbol{x}_1^2, \ldots, \boldsymbol{x}_h^2$ are the members of H_2 . $\tilde{\mathbf{X}}_2$ has zero mean and covariance matrix $\mathbf{K}_2 =$

$(1-\rho)\mathbf{S}_2 + \rho\mathbf{T}$.

Denote $d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{x}}'(\mathbf{K}_1)^{-1}\tilde{\boldsymbol{x}}$. We can then prove that:

$$\frac{1}{k}\sum_{i=1}^{h} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^2) = \frac{1-\rho}{h}\sum_{i=1}^{h} d_{\mathbf{K}_1}(\boldsymbol{x}_i^2 - \mathbf{m}_2) \tag{19}$$

$$\leq \frac{1-\rho}{h}\sum_{i=1}^{h} d_{\mathbf{K}_1}(\boldsymbol{x}_i^2 - \mathbf{m}_1) \tag{20}$$

$$\leq \frac{1-\rho}{h}\sum_{i=1}^{h} d_{\mathbf{K}_1}(\boldsymbol{x}_i^1 - \mathbf{m}_1) \tag{21}$$

$$= \frac{1}{k}\sum_{i=1}^{h} d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^1) \tag{22}$$

in which the second inequality (21) is the condition (16).

The first inequality (20) can be shown as follows. Put $\boldsymbol{z}_i = (\mathbf{K}_1)^{-1/2}\boldsymbol{x}_i^2$ and $\tilde{\boldsymbol{z}} = (\mathbf{K}_1)^{-1/2}\mathbf{m}_1$ and note that $\overline{\boldsymbol{z}} = (\mathbf{K}_1)^{-1/2}\mathbf{m}_2$ is the average of the $\boldsymbol{z}_i$. Then (20) becomes

$$\sum_{i=1}^{h} \|\boldsymbol{z}_i - \overline{\boldsymbol{z}}\|^2 \leq \sum_{i=1}^{h} \|\boldsymbol{z}_i - \tilde{\boldsymbol{z}}\|^2,$$

which follows from the fact that $\tilde{\boldsymbol{z}}$ is the unique minimizer of the least squares objective $\sum_{i=1}^{k} \|\boldsymbol{z}_i - c\|^2$, so (20) becomes an equality if and only if $\tilde{\boldsymbol{z}} = \overline{\boldsymbol{z}}$ which is equivalent to $\mathbf{m}_2 = \mathbf{m}_1$.

It follows that

$$\begin{aligned}\sum_{i=1}^{k} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^2) &= \sum_{i=1}^{h} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^2) + \frac{\rho}{p+1}\sum_{j=1}^{p+1} d_{\mathbf{K}_1}(\boldsymbol{y}_j)\\ &\leq \sum_{i=1}^{h} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^1) + \frac{\rho}{p+1}\sum_{j=1}^{p+1} d_{\mathbf{K}_1}(\boldsymbol{y}_j)\\ &= \sum_{i=1}^{k} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^1) \quad .\end{aligned}$$

Now put

$$b = \frac{\sum_{i=1}^{k} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^2)}{\sum_{i=1}^{k} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^1)} \leq 1 \quad .$$

If we now compute distances relative to $b\mathbf{K}_1$, we find

$$\frac{1}{k}\sum_{i=1}^{k} d_{b\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^2) = \frac{1}{b}\frac{1}{k}\sum_{i=1}^{k} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^2) = \frac{1}{k}\sum_{i=1}^{k} d_{\mathbf{K}_1}(\tilde{\boldsymbol{x}}_i^1) = \frac{1}{k}\sum_{i=1}^{k}(\tilde{\boldsymbol{x}}_i^1)'(\mathbf{K}_1)^{-1}\tilde{\boldsymbol{x}}_i^1$$
$$= \frac{1}{k}\sum_{i=1}^{k}(\mathbf{K}_1^{-1/2}\tilde{\boldsymbol{x}}_i^1)'(\mathbf{K}_1^{-1/2}\tilde{\boldsymbol{x}}_i^1) = \text{Trace}\left(\frac{1}{k}\sum_{i=1}^{k}(\mathbf{K}_1^{-1/2}\tilde{\boldsymbol{x}}_i^1)'(\mathbf{K}_1^{-1/2}\tilde{\boldsymbol{x}}_i^1)\right)$$
$$= \text{Trace}\left((\mathbf{K}_1)^{-1/2}\left(\frac{1}{k}\sum_{i=1}^{k}(\tilde{\boldsymbol{x}}_i^1)(\tilde{\boldsymbol{x}}_i^1)'\right)(\mathbf{K}_1)^{-1/2}\right) = \text{Trace}(\mathbf{I}_p) = p \ .$$

From the theorem in Grübel (1988), it follows that $\mathbf{K}_2$ is the unique minimizer of $\det(\mathbf{S})$ among all $\mathbf{S}$ for which $\frac{1}{k}\sum_{i=1}^{k} d_{\mathbf{S}}(\tilde{\boldsymbol{x}}_i^2) = p$ (note that the mean of $\tilde{\boldsymbol{x}}_i^2$ is zero). Therefore

$$\det(\mathbf{K}_2) \leq \det(b\mathbf{K}_1) \leq \det(\mathbf{K}_1) \ .$$

We can only have $\det(\mathbf{K}_2) = \det(\mathbf{K}_1)$ if both of these inequalities are equalities. For the first, by uniqueness we can only have equality if $\mathbf{K}_2 = b\mathbf{K}_1$. For the second inequality, equality holds if and only if $b = 1$. Combining both yields $\mathbf{K}_2 = \mathbf{K}_1$. Moreover, $b = 1$ implies that (20) becomes an equality, hence $\mathbf{m}_2 = \mathbf{m}_1$. This concludes the proof of Theorem 1.

Appendix B: The OGK estimator

Maronna and Zamar (2002) presented a general method to obtain positive definite and approximately affine equivariant robust scatter matrices starting from a robust bivariate scatter measure. This method was applied to the bivariate covariance estimate of Gnanadesikan and Kettenring (1972). The resulting multivariate location and scatter estimates are called orthogonalized Gnanadesikan-Kettenring (OGK) estimates and are calculated as follows:

1. Let $m(.)$ and $s(.)$ be robust univariate estimators of location and scale.

2. Construct $\boldsymbol{y}_i = \boldsymbol{D}^{-1}\boldsymbol{x}_i$ for $i = 1, \ldots, n$ with $\boldsymbol{D} = \text{diag}(s(X_1), \ldots, s(X_p))$.

3. Compute the 'pairwise correlation matrix' $\boldsymbol{U}$ of the variables of $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$, given by $u_{jk} = 1/4(s(Y_j + Y_k)^2 - s(Y_j - Y_k)^2)$. This $\boldsymbol{U}$ is symmetric but not necessarily positive definite.

4. Compute the matrix $\boldsymbol{E}$ of eigenvectors of $\boldsymbol{U}$ and

   (a) project the data on these eigenvectors, i.e. $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{E}$;

   (b) compute 'robust variances' of $\boldsymbol{V} = (V_1, \ldots, V_p)$, i.e. $\boldsymbol{\Lambda} = \mathrm{diag}(s^2(V_1), \ldots, s^2(V_p))$;

   (c) set the $p \times 1$ vector $\hat{\boldsymbol{\mu}}(\boldsymbol{Y}) = \boldsymbol{E}\boldsymbol{m}$ where $\boldsymbol{m} = (m(V_1), \ldots, m(V_p))^T$, and compute the positive definite matrix $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}) = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^T$.

5. Transform back to $\boldsymbol{X}$, i.e. $\hat{\boldsymbol{\mu}}_{\mathrm{OGK}} = \boldsymbol{D}\hat{\boldsymbol{\mu}}(\boldsymbol{Y})$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{OGK}} = \boldsymbol{D}\hat{\boldsymbol{\Sigma}}(\boldsymbol{Y})\boldsymbol{D}^T$.

Step 2 makes the estimate location invariant and scale equivariant, whereas the next steps replace the eigenvalues of $\boldsymbol{U}$ (some of which may be negative) by positive numbers. In the simulation study and empirical analysis, we set $m(.)$ to the median and $s(.)$ to the median absolute deviation.