Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-021-00471-6

REGULAR ARTICLE

The minimum weighted covariance determinant estimator for high-dimensional data

Jan Kalina[1,2] · Jan Tichavský[1]

Received: 14 January 2021 / Revised: 25 August 2021 / Accepted: 27 September 2021
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
In a variety of diverse applications, it is very desirable to perform a robust analysis of high-dimensional measurements without being harmed by the presence of a possibly larger percentage of outlying measurements. The minimum weighted covariance determinant (MWCD) estimator, based on implicit weights assigned to individual observations, represents a promising and flexible extension of the popular minimum covariance determinant (MCD) estimator of the expectation and scatter matrix of mlutivariate data. In this work, a regularized version of the MWCD denoted as the minimum regularized weighted covariance determinant (MRWCD) estimator is proposed. At the same time, it is accompanied by an outlier detection procedure. The novel MRWCD estimator is able to outperform other available robust estimators in several simulation scenarios, especially in estimating the scatter matrix of contaminated high-dimensional data.

Keywords High-dimensional data · Regularization · Robust estimation · Implicit weighting · Scatter matrix

Mathematics Subject Classification 62H12 · 62F35

1 Introduction

This paper is interested in highly robust methods proposed for the fundamental task to estimate the expectation and scatter matrix of elliptically symmetric unimodal distributions. Starting with data with the number of observations n exceeding the number

✉ Jan Kalina
kalina@cs.cas.cz

[1] The Czech Institute of Sciences, Institute of Computer Science, Prague 8, Czech Republic

[2] The Czech Institute of Sciences, Institute of Information Theory and Automation, Prague 8, Czech Republic

Published online: 07 October 2021

of variables p, the minimum covariance determinant (MCD) estimator of Rousseeuw (1984) represents a popular choice, the properties of which were overviewed by Hubert and Debruyne (2010) or Hubert et al. (2018). It can be computed by means of the FAST-MCD algorithm of Rousseeuw and Van Driessen (1999), which is based on concentration steps (C-steps). The minimum weighted covariance determinant (MWCD) estimator represents an extension of the MCD to any non-negative weights (not only equal to 1 or 0) proposed by Roelant et al. (2009), who proved the breakdown point of the MWCD to be high (if suitable weights are considered) and derived its influence function and asymptotic covariance matrix.

The MCD and MWCD estimators use implicit weighting and represent multivariate counterparts (for data with $n > p$) of the least weighted squares (LWS) estimator, where the latter was proposed for linear regression in Víšek (2011). The MWCD estimator was successful e.g. within a classification analysis of contaminated EEG signals of Kalina et al. (2016). In simulations, the MWCD estimator yields more accurate estimates (with smaller quadratic errors) of both the expectation and scatter matrix than other estimators (including MCD and MM-estimators) for data with moderate or larger sample sizes with $n > p$ and with intermediate outliers, i.e. outliers that are relatively close to the bulk of the data (Roelant et al. 2009). Such results were confirmed in Kalina (2021), who considered also alternative definitions of the MWCD (including a two-stage procedure) for $n > p$. The MCD and MWCD estimators (as solutions of approximate algorithms) are precisely equal to an iteratively defined affine equivariant L-estimator of Section 8.5.1 of Jurečková et al. (2013). Other robust estimators of parameters of multivariate data with $n > p$ were overviewed in Chapter 5 of Jurečková et al. (2019) or in a more theoretical context in Chapter 8 of Jurečková et al. (2013).

Robust methods are recommendable also for high-dimensional data (with $n < p$ or even $n \ll p$), which may be encountered in a variety of disciplines including biomedicine, engineering or econometrics with an increasing intensity. Our particular task is to estimate expectation and scatter of high-dimensional data contaminated by outliers. As the empirical covariance matrix is singular for data with the number of observations n smaller than the number of variables p, various alternative estimates have been proposed to estimate the scatter matrix by a regular matrix also for $n < p$. Various such regularized (shrinkage) estimates of the scatter matrix (Pourahmadi 2013), including the estimator of Ledoit and Wolf (2004) with an explicit form for the optimal regularization parameter, are however not robust with respect to the presence of outliers in the data (Chen et al. 2011).

From the practical point of view, robust estimates of parameters suitable also for high-dimensional data are of a great importance. So far, robust statistical methods for high-dimensional data have been more elaborated for dimensionality reduction (Hubert et al. 2005), regression (Hastie et al. 2015), and outlier detection (Filzmoser et al. 2008). On the other hand, robust estimates of expectation and scatter of high-dimensional data have started to appear only in the last decade (Filzmoser and Todorov 2011). Some inspiring approaches are overviewed in Sect. 1.1. Section 2 recalls selected robust regularized estimates for high-dimensional data and proposes a novel regularized version of the MWCD estimator, i.e. a weighted extension of the MRCD estimator of Boudt et al. (2020). A numerical illustration is presented in Sect. 3 and simulations follow in Sect. 4. Section 5 brings conclusions.

1.1 Available robust estimators for high-dimensional data

This section recalls several inspiring estimators of location and scatter for high-dimensional data. Chen et al. (2011) proposed a regularized version of the multivariate M-estimator, where the latter is often denoted as Tyler's estimator. The optimal value for the regularization parameter for Chen's estimator was derived by Ashurbekova et al. (2019). Asymptotics for the estimators of Chen et al. (2011) and Ledoit and Wolf (2004) was derived by Couillet and McKay (2014). A (non-robust) shrinkage estimator of the covariance matrix for Hotelling's T^2 test was proposed by Karjanto et al. (2015), who considered replacing standard means by trimmed means.

Gschwandtner and Filzmoser (2013) defined the RegMCD estimator of expectation and inverse scatter by minimizing the L_1-regularized likelihood evaluated over h observations with the smallest values of Mahalanobis distances, where h is chosen by the user. For the computations, they used a direct extension of the FAST-MCD algorithm. They performed outlier detection based on comparing robust regularized Mahalanobis distances of individual observations with the quantile $\chi^2_p(0.975)$, i.e. the 97.5 % quantile of χ^2_p. This quantile is also the default value for the MCD (i.e. for $n > p$) and comes from the idea that population Mahalanobis distances have a χ^2 distribution (Rousseeuw and Van Zomeren 1990). Simulations revealed such outlier detection to perform reliably. A different regularization of the likelihood was used by Fritsch et al. (2011) to define the R-MCD estimator as an extension of the MCD. This approach was used mainly to perform outlier detection as by Gschwandtner and Filzmoser (2013).

Ro et al. (2015) considered only diagonal elements of the covariance matrix to define their minimum diagonal product (MDP) estimator. Their diagonal estimator of Σ does not need any additional regularization. They derived the asymptotics of the estimator, proved its breakdown point to be high, and also proposed a sophisticated outlier detection rule based on the MDP estimator.

Boudt et al. (2020) proposed the MRCD estimator based on minimizing the determinant of a Tikhonov-regularized empirical covariance matrix evaluated over all h-subsets of the data. A sophisticated algorithm was proposed for the computation. It does not need numerous random initial choices of subsets of observations, which are typical for the FAST-MCD algorithm (Rousseeuw and Van Driessen 1999) and its extensions. Simulations revealed the MRCD estimator to be the best for a variety of contaminated situations with $n < p$.

While properties of the available regularized versions of the MCD estimator (including their local sensitivity) remain unknown, let us recall that robust methods based on a complete rejection of individual observations may achieve a high local sensitivity. This is the case of the least trimmed squares (LTS) estimator (Rousseeuw and Leroy 1987) in linear regression as discussed in Víšek (2006) or in Chapter 4.9 of Jurečková et al. (2019). On the other hand, the LWS estimator allowing to assign weights (not only zeros or ones) to individual observations has a smaller local sensitivity (Kalina and Tichavský 2020). This motivates us to consider an extension of the MRCD to a more flexible estimator allowing to assign continuous weights to individual observations.

2 Robust estimation for multivariate (possibly high-dimensional) data

2.1 Assumptions and notation

We use the notation $PSD(p)$ and $PD(p)$ for the set of all positive semidefinite symmetric matrices and positive definite symmetric matrices of size $p \times p$, respectively. Throughout the paper, the following assumptions are assumed to be fulfilled; if not explicitly stated, it is possible that $n < p$.

We consider the total number n of p-dimensional i.i.d. observations $X_i = (X_{i1}, \dots, X_{ip})^T$ for $i = 1, \dots, n$, where $\mu \in \mathbb{R}^p$. The matrix with elements X_{ij} with $i = 1, \dots, n$ and $j = 1, \dots, p$ will be denoted as X; its rows correspond to observations. The notation $\lfloor x \rfloor$ will be used for the integer part of $x \in \mathbb{R}$.

Assumption 1 We consider p-dimensional i.i.d. random vectors $X_1, \dots, X_n$ following an elliptically symmetric unimodal (ESU) distribution (Hubert et al. 2018) with the location parameter $\mu \in \mathbb{R}^p$ and scatter matrix $\Sigma \in PD(p)$. Further, it is assumed that any $\lfloor n/2 \rfloor + 1$ observations among the total number of n observations give a non-singular estimate of Σ.

Definition 1 (*Weight function*) Let $\psi : [0, 1] \to [0, 1]$ be a non-increasing and continuous function on $[0, 1]$, let $\psi(0) = 1$ and $\psi(1) = 0$. Then, the function ψ is called a weight function.

For a fixed n, we consider weights generated by ψ in the form

$$w_i = \psi\left(\frac{i - 1/2}{n}\right) \quad \text{for} \quad i = 1, \dots, n. \tag{1}$$

We use the notation $\hat{j} = (1, 1, \dots, 1)^T \in \mathbb{R}^n, \mathcal{I}_n$ for the unit matrix of size $n \times n$, Φ^{-1} for the quantile function of normal $N(0, 1)$ distribution, det for the determinant and *med* for the median. Let us now use standard basis vectors $e_i = (e_{i1}, \dots, e_{in})^T$ for $i = 1, \dots, n$, where $e_{ii} = 1$ and $e_{ij} = 0$ for $j \neq i$. If we consider fixed non-negative weights $w = (w_1, \dots, w_n)^T$, the diagonal matrix $W \in \mathbb{R}^{n \times n}$ with diagonal elements $w_1, \dots, w_n$ can be expressed as

$$W = diag(w) = \sum_{i=1}^{n} e_i^T w e_i e_i^T. \tag{2}$$

The weighted mean of the data with weights w, transformed to the natural requirement $\sum_{i=1}^n w_i = 1$, equals $\bar{X}_w = Xw = \sum_{i=1}^n w_i X_i$ and the (empirical) weighted covariance matrix with weights w can be expressed as

$$S(w) = \sum_{i=1}^{n} w_i (X_i - \bar{X}_w)(X_i - \bar{X}_w)^T = X^T(\mathcal{I}_n - w\hat{j}_n^T)W(\mathcal{I}_n - \hat{j}_n w^T)X. \tag{3}$$

2.2 The MWCD estimator

Let us assume data under Assumptions 1 with $n > p$. In this paper, we define the MWCD estimator of μ and Σ simply as the solution of the algorithm of Roelant et al. (2009), i.e. of a natural extension of the FAST-MCD algorithm of Rousseeuw and Van Driessen (1999). The algorithm denoted here as FAST-MWCD is based on concentration steps (C-steps), while a proof that a C-step improves the loss function was presented also in Roelant et al. (2009).

The definition of the MWCD estimator will be now formally expressed by means of several alternative (and novel) forms. The MWCD estimator searches for the optimal permutation of weights. Let the set of all permutation matrices of size $n \times n$ be denoted as $\mathbb{P}_n$. Instead of the magnitudes (1) of the weights, which have to fulfil $\sum_{i=1}^{n} w_i = 1$, we search for their (unknown) permutation, uniquely described by a given permutation matrix $P \in \mathbb{P}_n$. Thus, instead of considering $S(w)$, the MWCD estimator is defined as

$$\operatorname*{argmin}_{P \in \mathbb{P}_n} \det S(Pw). \tag{4}$$

Exploiting properties of permutation matrices, we may write

$$\begin{aligned} S(Pw) &= X^T(\mathcal{I}_n - Pw\hat{j}_n^T)PWP^T(\mathcal{I}_n - \hat{j}_n(Pw)^T)X \\ &= X^T(\mathcal{I}_nP - Pw\hat{j}_n^TP)W(P^T\mathcal{I}_n - P^T\hat{j}_nw^TP^T)X \\ &= X^T(P - Pw\hat{j}_n^T)W(P^T - \hat{j}_nw^TP^T)X \\ &= X^TP(\mathcal{I}_n - w\hat{j}_n^T)W(\mathcal{I}_n - \hat{j}_nw^T)P^TX \\ &= X^TP\Xi\Xi^TP^TX \\ &= (X^TP\Xi)(X^TP\Xi)^T \end{aligned} \tag{5}$$

with $\Xi = (\mathcal{I}_n - w\hat{j}_n^T)W^{1/2}$. Thus, $S(Pw)$ can be expressed as being composed of the data matrix X, a factor Ξ depending only on magnitudes of the weights, and the permutation matrix P.

If the permutation matrix P corresponding to the optimal permutation (solution of (4)) is denoted as P^*, the MWCD-estimates are defined as

$$\bar{X}_{MWCD} = X^TP^*w \quad \text{and} \quad S_{MWCD} = c_\psi(X^TP^*\Xi)(X^TP^*\Xi)^T, \tag{6}$$

where c_ψ is a consistency factor.

2.3 Regularized MWCD estimator

In this section, we propose a novel minimum regularized weighted covariance determinant (**MRWCD**) estimator, where the estimate of μ will be denoted by $\bar{X}_{MRWCD}$ and the corresponding estimate of Σ by S_{MRWCD}. It is based on finding the optimal permutation of given magnitudes of weights, naturally extending the MRCD estimator of Boudt et al. (2020) to a weighted version, and extending (6) to a regularized version. The user specifies a given target matrix T, which must be symmetric positive definite of size $p \times p$. The simplest choice is the identity matrix $T = \mathcal{I}_p$.

Definition 2 We assume Assumptions 1, the magnitudes of the weights defined by (1) transformed to $\sum_{i=1}^{n} w_i = 1$, a given regular matrix $T \in PD(p)$, and a given $\rho > 0$. The MRWCD is defined as

$$\underset{P \in \mathbb{P}_n}{\operatorname{argmin}} \det \left(\rho T + (1-\rho) S(Pw) \right). \tag{7}$$

If the permutation matrix P corresponding to the optimal permutation (solution of (7)) is denoted as P^*, the MRWCD-estimates of μ and Σ are defined by

$$\bar{X}_{MRWCD} = X^T P^* w \quad \text{and} \quad S_{RMWCD} = \rho T + (1-\rho) c_\psi (X^T P^* \Xi)(X^T P^* \Xi)^T, \tag{8}$$

respectively, where c_ψ is a consistency factor.

Algorithm 1 for approximating the MRWCD estimator performs generalized C-steps (i.e. modified versions compared to basic C-steps of the MWCD; cf. Kalina 2021) extending the algorithm of Boudt et al. (2020). Matrix operations allow to evaluate robust regularized Mahalanobis distances in an efficient way. Algorithm 1 starts with a data transform replacing the raw observations by transformed data $Z_1, \ldots, Z_n$ to ensure location and scale equivariance. In our notation, $Q_n(X_{1j}, \ldots, X_{nj})$ is the Q_n statistic of Rousseeuw and Croux (1993) evaluated for the j-th variable for $j = 1, \ldots, p$.

Further, 6 initial estimates of scatter are applied on the transformed data. These estimates previously used within the DetMCD algorithm (see Section 3.1 of Hubert et al. 2012) are chosen as a small diverse set of sufficiently robust estimates. Our numerical evidence confirms these 6 initial estimates to be much more suitable compared to naïve choices including e.g. diagonal estimates of Σ or based on random permutations of weights. These estimates are overviewed in Table 1, where *corr* denotes the correlation matrix. We remark that these 6 estimates are actually not necessarily estimates of Σ, but are related to scatter; we can say that a matrix Δ related to scatter of the transformed data is considered and its estimate is denoted by $\hat{\Delta}$ here. In fact, correlation or shape matrices serve as Δ here, as we need them within (robust and regularized) Mahalanobis distances only to obtain ranking of observations. For the computation of the raw OGK estimator, we use the implementation in the package rrcov (Todorov and Filzmoser 2009). For all the other estimates, we spent only little effort to perform our own straightforward implementations. In case of ties, average ranking method is used for rank-based estimates.

The method of Hubert et al. (2012) for obtaining regularized versions of the scatter estimates is formulated here as a separate Algorithm 2, where *cmed* denotes the coordinate-wise median. This method further denoted as hubertization is applied on the obtained 6 estimates of Δ. Within Algorithm 1, estimates of μ corresponding to hubertized estimates of Δ are computed.

Algorithm 1 MRWCD for high-dimensional data (under Assumptions 1) using $T = \mathcal{I}_p$.

```
Input: $X_1, \ldots, X_n$, where $X_i \in \mathbb{R}^p$ for each $i = 1, \ldots, n$
Input: $\varepsilon > 0$
Input: Weight function $\psi$ transformed so that $\sum_{i=1}^n \psi((i - 1/2)/n) = 1$
Input: Consistency factor $c_\psi$ (depends on $p$, $n$, and $\psi$)
Input: $\kappa = 50$
Input: $\Omega = 30$
Output: $\bar{X}_{MRWCD}$ and $S_{MRWCD}$
  $v := \left(med\{X_{11}, \ldots, X_{n1}\}, \ldots, med\{X_{1p}, \ldots, X_{np}\}\right)^T \in \mathbb{R}^p$
  $D := \text{diag}\{Q_n(X_{11}, \ldots, X_{n1}), \ldots, Q_n(X_{1p}, \ldots, X_{np})\}$
  Eigendecomposition $T = E \Lambda E^T$
  $Z_i := \Lambda^{-1/2} E^T D^{-1}(X_i - v), \quad i = 1, \ldots, n$
  for $r = 1$ to $6$ do
    Use Algorithm 2 to obtain a positive definite estimate of $\Delta$ (say $S_{0r}$) and a corresponding estimate of $\mu$ (say $m_{0r}$) from $Z_1, \ldots, Z_n$ using the $r$-th method (estimator)
    Eigendecomposition $S_{0r} = E_r \Lambda_r E_r^T$
    $d_{ri}^2(m_{0r}, S_{0r}) := ||\Lambda_r^{-1/2} E_r^T (Z_i - m_{0r})||^2, \quad i = 1, \ldots, n$
    $R_i^r$ := rank of $d_{ri}^2(m_{0r}, S_{0r})$ among $d_{0r}^2(m_{0r}, S_{0r}), \ldots, d_{rn}^2(m_{0r}, S_{0r})$, where $i = 1, \ldots, n$
    $m_{1r} := \sum_{i=1}^n \psi\left((R_i^r - 1/2)/n\right) Z_i$
    $S_{1r} := c_\psi \sum_{i=1}^n \psi\left((R_i^r - 1/2)/n\right) (Z_i - m_{1r})(Z_i - m_{1r})^T$
    $\ell_{1r} := \det S_{1r}$
    $\lambda^* :=$ max. eigenvalue of $S_{1r}$
    $\lambda_* :=$ min. eigenvalue of $S_{1r}$
    $\rho_r := \max\left\{0, \mathbb{1}[\lambda^*/\lambda_* < \kappa] \cdot (\lambda^* - \kappa\lambda_*)/(\kappa + \lambda^* - \kappa\lambda_* - 1)\right\}$
    $S_{1r} := \rho_r T + (1 - \rho_r) S_{1r}$
  end for
  $\rho := \max_r\{\rho_r\} \mathbb{1}\left[\max_r\{\rho_r\} \leq 0.1\right] + \max\{0.1, med_r\{\rho_r\}\} \cdot \mathbb{1}\left[\max_r\{\rho_r\} > 0.1\right]$
  for $r = 1$ to $6$ do
    $\omega := 0$
    if $\rho_r \leq \rho$ then
      repeat // Generalized C-step
        $\omega := \omega + 1$
        $m_{0r} := m_{1r}; \quad S_{0r} := S_{1r}; \quad \ell_{0r} := \ell_{1r};$
        Eigendecomposition $S_{0r} = E_r \Lambda_r E_r^T$
        $d_{ri}^2(m_{0r}, S_{0r}) := ||\Lambda_r^{-1/2} E_r^T (Z_i - m_{1r})||^2, \quad i = 1, \ldots, n$
        $R_i^r$ := rank of $d_{ri}^2(m_{0r}, S_{0r})$ among $d_{1r}^2(m_{0r}, S_{0r}), \ldots, d_{rn}^2(m_{0r}, S_{0r})$, where $i = 1, \ldots, n$
        $m_{1r} := \sum_{i=1}^n \psi\left((R_i^r - 1/2)/n\right) Z_i$
        $S_{1r} := \rho T + (1 - \rho) c_\psi \sum_{i=1}^n \psi\left((R_i^r - 1/2)/n\right) (Z_i - m_{1r})(Z_i - m_{1r})^T$
        $\ell_{1r} := \det S_{1r}$
      until $((\ell_{1r} + \varepsilon > \ell_{0r})$ or $(\omega \geq \Omega))$
    end if
  end for
  $\check{r} := \underset{r=1,\ldots,6}{\text{argmin}}\ \ell_{0r}$
  $\bar{X}_{MRWCD} := v + D m_{0\check{r}}$
  $S_{MRWCD} := D E \Lambda^{1/2} S_{0\check{r}} \Lambda^{1/2} E^T D$
```

Table 1 The 6 initial estimates of scatter, applied on transformed observations $Z_1, \ldots, Z_n$ in Algorithm 2

Index	Estimate of Δ (or estimate of Σ)
1	$corr(\tanh(Z))$
2	Spearman correlation matrix of Z computed as $corr(R)$, where $R_{1j}, \ldots, R_{nj}$ are ranks of $Z_{1j}, \ldots, Z_{nj}$ for $j = 1, \ldots, p$
3	$corr(\Phi^{-1}(R^*))$, where $R^*_{ij} = (R_{ij} - 1/3)/(n + 1/3)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$
4	$S = (1/n) \sum_{i=1}^{N} k_i k_i^T$, where $k_i = Z_i/\vert\vert Z_i\vert\vert$
5	Empirical covariance matrix of $\tilde{H}$, where $\tilde{H}$ is the subset with $\lceil n/2 \rceil$ observations with the smallest norm
6	Raw orthogonalized Gnanadesikan–Kettenring (OGK) estimator

Algorithm 2 Hubertization: Computing a positive definite estimate $\hat{\Delta}_H$ of Δ and a corresponding estimate $\hat{\mu}_H$ of μ for a given scatter estimate $\hat{\Delta}$, i.e. for one of the 6 estimates of Table 1.

```
Input: Data $Z_1, \ldots, Z_n$, where $Z_i \in \mathbb{R}^p$
Input: Matrix $\hat{\Delta} \in PSD(p)$
Output: $\hat{\Delta}_H \in PD(p)$
Output: $\hat{\mu}_H$
  Eigendecomposition $\hat{\Delta} = E \Lambda E^T$
  $B := ZE$
  $R := E \operatorname{diag}\{Q_n(B_{11}, \ldots, B_{n1}), \ldots, Q_n(B_{1p}, \ldots, B_{np})\}$
  $\hat{\Delta}_H := RR^T$
  $\hat{\mu} := \hat{\Delta}_H^{1/2} cmed(Z \hat{\Delta}_H^{-1/2})$
```

Computing the MRWCD estimator requires to find an approximation to the consistency factor c_ψ depending on p, n and ψ. We propose the consistency factor can be evaluated on normal data with $N(0, \mathcal{I}_p)$ without the knowledge of the proper Σ. However, estimating c_ψ requires regularization, while finding ρ within Algorithm 1 already depends on the known c_ψ. Thus, we used Algorithm 1 (including the regularization) with $c_\psi = 1$ and averaged the results over 1000 simulations to obtain a scatter matrix estimate (say S). Further, c_ψ was found in a grid search to solve

$$\min_{c>0} ||cS - \mathcal{I}_p||_F^2, \tag{9}$$

where the Frobenius norm of a matrix $A = \left(a_{ij}\right)_{i,j=1}^p \in \mathbb{R}^{p \times p}$ is defined as

$$||A||_F = \left(\sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij}^2 \right)^{1/2}. \tag{10}$$

The computational complexity of the MRWCD estimator is comparable to that of the MRCD. Thanks to starting with the 6 initial estimates, the computation of both these estimators is much more efficient compared to RegMCD, where the latter estimator computed by a modification of FAST-MCD requires a large number of initial subsets of data. The computations are feasible also for larger p; for each of the 6 initial estimates, there is typically up to 5 C-steps performed, and we set a maximum of 30 C-steps as in Roelant et al. (2009).

2.4 Regularization parameter and choice of target matrix

Naturally, one can always use cross validation to find a suitable value of the regularization parameter ρ. Such approach is computationally very demanding, especially for larger p, so we use here the approach of Boudt et al. (2020) to get a simple estimate of ρ. We consider an upper bound for the condition number of the matrix $\rho T + (1-\rho)c_\psi \hat{\Delta}_r$ with a known matrix denoted here as $\hat{\Delta}_r$ ($r = 1, \ldots, 6$), where the latter has the maximal and the minimal eigenvalue denoted as λ^* and λ_*, respectively. The equation requiring the condition number to be equal to a selected threshold κ has the form

$$\frac{\text{maximal eigenvalue of } \left(\rho T + (1-\rho)c_\psi \hat{\Delta}_r\right)}{\text{minimal eigenvalue of } \left(\rho T + (1-\rho)c_\psi \hat{\Delta}_r\right)} = \kappa. \tag{11}$$

The resulting value of ρ_r for the default choice $T = \mathcal{I}_p$ is explicitly expressed within Algorithm 1. Combining individual values of $\rho_1, \ldots, \rho_6$ to a single value of ρ ensures the obtained matrix S_{MRWCD} to be regular and positive definite even for $n \ll p$, while the estimator does not possess affine equivariance.

Apart from the simplest and most natural choice of the target matrix $T = \mathcal{I}_p$, it is possible to consider e.g. two target matrices exploited already by Schäfer and Strimmer 2005. Let the largest diagonal element of the matrix $\hat{\Delta}_r$ of above be denoted as S^* and its smallest diagonal element as S_*. The first target matrix

$$T = \left(T_{ij}\right)_{i,j=1}^{p}, \quad T_{ij} = \begin{cases} \nu := \frac{1}{p}\sum_{i=1}^{p} S_{ii}, & \text{if } i = j, \\ c := \frac{1}{p(p-1)/2}\sum_{i=1}^{p-1}\sum_{j>i} S_{ij}, & \text{if } i \neq j, \end{cases} \tag{12}$$

considers a common covariance and a common variance, while the second target matrix

$$T = \left(T_{ij}\right)_{i,j=1}^{p}, \quad T_{ij} = \begin{cases} S_{ii}, & \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases} \tag{13}$$

is diagonal and considers unequal variances. For target matrices (12) or (13), we modify now the approach for the upper bound of the condition number. Particularly for the target matrix (12), the Eq. (11) has the solution

$$\rho_r = \frac{\lambda^* - \kappa\lambda_*}{\kappa\nu + \lambda^* - \kappa\lambda_* - \nu}, \tag{14}$$

which can be used in Algorithm 1 replacing there ρ_r derived for the default choice $T = \mathcal{I}_p$. If the target matrix (13) is considered, (11) has the solution

$$\rho_r = \frac{\lambda^* - \kappa\lambda_*}{\kappa S_* + \lambda^* - \kappa\lambda_* - S^*}. \tag{15}$$

2.5 Weighting schemes for implicitly weighted estimators

We now present several possible choices for the weight functions, which can be used for the MRWCD and MWCD estimators. The simplest choice are linearly decreasing weights generated by

$$\psi_1(t) = 1 - t, \quad t \in [0, 1]. \tag{16}$$

Weights generated by the logistic curve (logistic function)

$$\psi_2(t) = \frac{1 + \exp\{-s/2\}}{1 + \exp\{s(t - \frac{1}{2})\}}, \quad t \in [0, 1], \tag{17}$$

with a fixed choice $s = 10$ remain larger for linear ones for the good data, while they are very small for the most outlying observations. As shown in Fig. 1, weights generated by the standardized density of normal distribution

$$\psi_3(t) = \exp\left\{-\frac{t^2}{2\sigma^2}\right\}, \quad t \in [0, 1], \tag{18}$$

with a fixed choice $\sigma = 0.8$ decrease more slowly and can be thus expected to achieve a better performance for non-contaminated data.

For each of the three choices, it is possible to consider more robust alternatives defined for a fixed $\tau \in [1/2, 1)$, so that the estimators consider only $h = \lfloor \tau n \rfloor$ observations and trim away the remaining ones. The versions based on hard trimming (HT)

$$\psi_i^{HT}(t) = \psi(t) \cdot \mathbb{1}[t < \tau], \quad t \in [0, 1], \quad i \in \{1, 2, 3\}, \tag{19}$$

are depicted in Fig. 1. Alternatively, one may consider such their versions

$$\psi_i^{D}(t) = \psi\left(\frac{t}{\tau}\right) \cdot \mathbb{1}[t < \tau], \quad t \in [0, 1], \quad i \in \{1, 2, 3\}, \tag{20}$$

for which the weights decrease more quickly to 0 (so they obtain the superscript D) and thus may be more robust but less efficient. The weight function

$$\psi_4(t) = \mathbb{1}[t > \tau], \quad t \in [0, 1], \tag{21}$$

performs a hard trimming (cf. the discussion in Cerioli et al. (2018)) so that the MRWCD equals to the MRCD estimator, and the MWCD to the MCD.

Algorithm 3 Outlier detection based on a reweighted MRWCD estimator in two versions (FDR-χ^2 or FDR-F)

```
.
Input: Data $X_1, \dots, X_n$, where $X_i \in \mathbb{R}^p$ for each $i = 1, \dots, n$
Input: Weight function $\psi$ transformed so that $\sum_{i=1}^n \psi((i - 1/2)/n) = 1$
Input: Consistency factor $c_\psi$ (depends on $p$, $n$, and $\psi$)
Input: $\delta = 0.025$
Input: $\alpha = 0.05$
Output: List of outliers in the set $\{X_1, \dots, X_n\}$
  Compute $\bar{X}_{MRWCD}$ and $S_{MRWCD}$ using $\psi$ and $c_\psi$ according to Algorithm 1
  Compute $d_i^2(\bar{X}_{MRWCD}, S_{MRWCD})$ using (30)
  FDR-$\chi^2$:
      $$\omega_i := \mathbb{1}\left[d_i^2(\bar{X}_{MRWCD}, S_{MRWCD}) \leq \chi^2_{p,1-\delta}\right], \quad i = 1, \dots, n \tag{22}$$

  FDR-$F$:
      $$\omega_i := \mathbb{1}\left[d_i^2(\bar{X}_{MRWCD}, S_{MRWCD}) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(1-\delta)\right], \quad i = 1, \dots, n \tag{23}$$

  $m := \sum_{i=1}^n w_i$
  $\bar{X}_{RMRWCD} := \sum_{i=1}^n w_i X_i / m$
  $S_{RMRWCD} :=$
      $$:= \frac{1-\delta}{P(\chi^2_{p+2} < \chi^2_{p,1-\delta})} \sum_{i=1}^n \frac{w_i (X_i - \bar{X}_{RMRWCD})(X_i - \bar{X}_{RMRWCD})^T}{m-1} \tag{24}$$

  Compute $d_i^2(\bar{X}_{RMRWCD}, S_{RMRWCD})$ using (30)
  Compute
      $$p_i := \begin{cases} P\left[\zeta \geq \frac{m}{(m-1)^2} d_i^2(\bar{X}_{RMRWCD}, S_{RMRWCD})\right], & \omega_i = 1, \\ P\left[\xi \geq \frac{m(m-p)}{(m+1)(m-1)p} d_i^2(\bar{X}_{RMRWCD}, S_{RMRWCD})\right], & \omega_i = 0, \end{cases} \tag{25}$$
  for $i = 1, \dots, n$, where $\zeta \sim Beta\left(\frac{p}{2}, \frac{m-p-1}{2}\right)$ and $\xi \sim F(p, m-p)$
  Arrange $p_1, \dots, p_n$ in ascending order as $p_{(1)} \leq \cdots p_{(n)}$
      $$H := \underset{\eta=1,\dots,n}{\text{argmax}} \left[p_{(\eta)} \leq \frac{\eta}{n}\alpha\right] \tag{26}$$

  for $i = 1$ to $n$ do
      Assign $X_i$ to be an outlier if and only if $p_{(i)} \leq p_{(H)}$
  end for
```

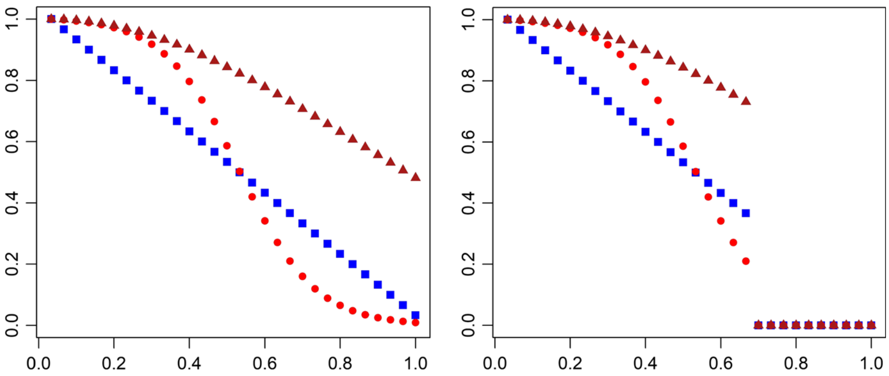

Fig. 1 Various weight functions suitable for implicitly weighted estimators (including the novel MRWCD). Left: weight functions ψ_1 (squares), ψ_2 (circles), and ψ_3 (triangles). Right: weight functions ψ_1^{HT} (squares), ψ_2^{HT} (circles), and ψ_3^{HT} (triangles), always with $\tau = 2/3$. Horizontal axis: parameter $t \in [0, 1]$. Vertical axis: values of the weight functions

2.6 Outlier detection

Robust estimates of parameters of multivariate data have also been successfully applied to outlier detection. The simplest approach (denoted here as the plain approach) for a given estimator $\hat{\mu}$ and $\hat{\Sigma}$ of μ and Σ, respectively, is to assign such observations X_i to be outlying, for which the Mahalanobis distances $d_i^2(\hat{\mu}, \hat{\Sigma})$ fulfil $d_i^2(\hat{\mu}, \hat{\Sigma}) > \chi^2_{p,1-\delta}$, where $\chi^2_{p,1-\delta}$ is the $(1-\delta)$-quantile of χ^2_p distribution. The choice $\delta = 0.025$ is standard in this context (Hubert et al. 2005). Useful discussions about critical values have been presented already for non-regularized estimators. Let us denote

$$\omega_i = \mathbb{1}\left[d_i^2(\mu, \Sigma) > \chi^2_{p,1-\delta}\right], \quad i = 1, \ldots, n, \tag{27}$$

and $m = \sum_{i=1}^n \omega_i$. Inspired by Hardin and Rocke (2005), approximations conditioning on ω_i in the form

$$\frac{m}{(m-1)^2} d_i^2(\mu, \Sigma) \dot{\sim} Beta\left(\frac{p}{2}, \frac{m-p-1}{2}\right), \quad \text{if } \omega_i = 1, \tag{28}$$

and

$$\frac{m(m-p)}{(m+1)(m-1)p} d_i^2(\mu, \Sigma) \dot{\sim} F_{p,m-p}, \quad \text{if } \omega_i = 0, \tag{29}$$

were considered for $i = 1, \ldots, n$ by Cerioli (2010), who also considered (28) and (29) to be approximately valid if replacing μ and Σ by MCD-estimates. Cerioli and Farcomeni (2011) extended the method to consider a multiple testing procedure based on (28) and (29) controling the false discovery rate (FDR) by the B–H approach (Benjamini and Hochberg 1995).

While the plain approach has been criticized for regularized estimators (Boudt et al. 2020), the method of Cerioli and Farcomeni (2011) may be applied as an approximation also to regularized estimators. For the MRWCD estimator, let us now formally define the Mahalanobis distances of the observation X_i from $\bar{X}_{MRWCD}$ as

$$d_i^2(\bar{X}_{MRWCD}, S_{MRWCD}) = (X_i - \bar{X}_{MRWCD})^T S_{MRWCD}^{-1}(X_i - \bar{X}_{MRWCD}) \tag{30}$$

for $i = 1, \ldots, n$. However, for the sake of improving numerical stability, we recommend to avoid the computation of (30), which is expensive of order p^3. Instead, eigendecomposition of S_{MRWCD} in the form $S_{MRWCD} = Q\Lambda Q^{-1}$ with an orthogonal matrix Q and a diagonal matrix Λ is more efficient, which allows to express (30) as

$$\begin{aligned} d_i^2(\bar{X}_{MRWCD}, S_{MRWCD}) &= (X_i - \bar{X}_{MRWCD})^T Q\Lambda Q^{-1}(X_i - \bar{X}_{MRWCD}) \\ &= ||\Lambda^{-1/2}Q^{-1}(X_i - m)||^2, \quad i = 1, \ldots, n. \end{aligned} \tag{31}$$

Let us finally formulate a multivariate outlier detection procedure based on the MRWCD estimator which tests the null hypothesis that individual observations are not outlying. The procedure inspired by Cerioli and Farcomeni (2011) is formulated in Algorithm 3 with two different approximations (22) or (23) to obtain a reweighted version of the MRWCD estimator (RMRWCD). The corresponding versions of the FDR-based procedure will be denoted as FDR-χ^2 and FDR-F, respectively. Approximate p-values obtained using (25) are corrected for multiple testing. The same approximate FDR-χ^2 and FDR-F procedures may be also applied to other regularized estimators.

3 A real data example

The cardiovascular genetic case-control study dataset of Kalina et al. (2016) will be now considered in order to illustrate how the MRWCD estimator may be used for outlier detection. Gene expressions of $p = 38\,590$ gene transcripts were measured by means of HumanWG-6 Illumina BeadChip microarrays (version 3) on 48 individuals in the years 2006–2011, namely on 24 patients immediately after cerebrovascular stroke (CVS) and 24 control persons, where the latter individuals were without a manifested cardiovascular disease. The original aim of the research was to identify genes associated with excess genetic risk for the incidence of cerebrovascular stroke. As in Marozzi et al. (2020), we randomly select a subset of the whole genetic dataset with $p = 1000$, however using only the CVS patients (i.e. $n = 24$ here). As the microarray data were pre-processed in a sophisticated way, they contain 7 valid digits and we can assume that no ties occur. The computations are performed in R software (R Core Team 2018).

We compute the MRWCD estimator with the robust weight function ψ_3^D. The loss function values for various h are shown in Fig. 2(left). Based on the figure, we come

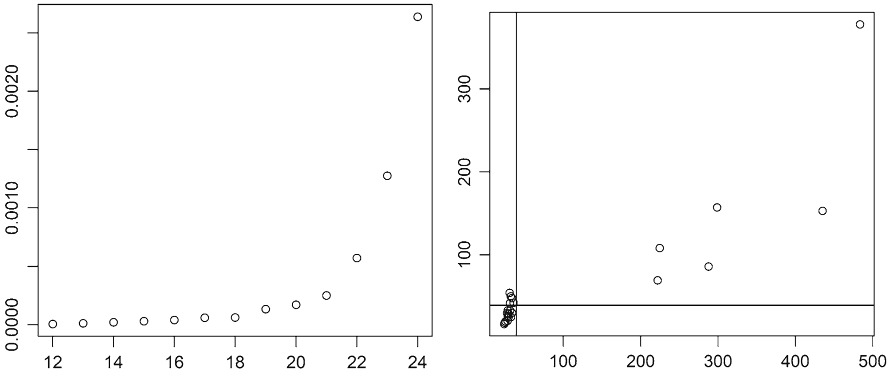

Fig. 2 Results for the gene expression dataset of Sect. 3 with $p = 1000$ and $n = 24$. Left: loss function (vertical axis) of MRWCD with ψ_3^{HT} for different values of $h = \lfloor \tau n \rfloor \geq n/2$ (horizontal axis). Right: robust regularized Mahalanobis distance for individual observations, based on MRCD with $\tau = 0.75$ (horizontal axis) and on MRWCD with ψ_3^{HT} with $\tau = 0.75$ (vertical axis)

to detecting 6 outliers, i.e. choosing $h = 18$ as the suitable value. Using $\tau = 0.75$, Fig. 2(right) shows robust Mahalanobis distances of individual observations by means of MRWCD with ψ_3^D against those by means of MRCD. The lines in the figure correspond to the quantile $\chi_p^2(0.975)$. Although the Mahalanobis distances of some additional points exceed $\chi_p^2(0.975)$, they turn out not to be significant in the FDR-F outlier detection procedure, which finds 6 outliers here. Simulation comparisons of the outlier detection performance of various methods presented in Sect. 2.6 will be presented later in Sect. 4.2.

4 Simulations

Several simulation studies aiming at investigating the performance of the newly proposed MRWCD estimator, particularly over high-dimensional data, will be now presented. In each of the simulations, we randomly generate the data 1000 times in the following way. First, we randomly generate n observations from $N(\mu, \Sigma)$, where we always use $\mu = (0, \ldots, 0)^T$ and $\Sigma = \mathcal{I}_p$. Then, we randomly select $\lfloor \alpha n \rfloor$ of the observations (for a given α), which are replaced in one of three ways. In each case, the outliers are generated as i.i.d. random vectors from p-variate normal distribution. The choice $\alpha = 0$ corresponds to non-contaminated normal data.

(A) Outliers generated from $N((3, \ldots, 3)^T, \Sigma^A)$, where Σ^A denotes the matrix $\mathcal{I}_p$ with additional covariance $\Sigma_{12}^A = \Sigma_{21}^A = 0.5$;

(B) Outliers generated from $N(\mu^B, c\Sigma)$ with a given $\mu^B \in \mathbb{R}^p$ and $c > 0$;

(C) Multivariate outliers of Fritsch et al. (2011). We take a random vector $a \in \mathbb{R}^p$ with coordinates generated from the alternative (Bernoulli) distribution $Alt(1/2)$ and generate the outliers from $N(\mu^C, \Sigma^C)$, where $\Sigma^C = \Sigma + 5aa^T/||a||_2^2$.

The following estimators (and their implementations) are used in the simulations.

- Classical estimates, i.e. the mean and empirical covariance matrix, where the latter can be obtained as (3) with weights $w_i = 1/n$ for $i = 1, \ldots, n$;
- The regularized (shrinkage) empirical covariance matrix of Σ of Ledoit and Wolf (2004), using package CovTools (Lee and You 2019) of R software;
- RegMCD proposed in Gschwandtner and Filzmoser (2013), using package rrlda (Gschwandtner et al. 2012) of R software;
- MRCD of Boudt et al. (2020) with the trimming constant $h = \lfloor 3n/4 \rfloor$, using package rrcov (Todorov and Filzmoser 2009);
- The novel MRWCD with a specified weight function, using our implementation of Algorithm 1; if ψ_1^{HT} or ψ_3^{HT} is used, then $\tau = 3/4$ is chosen.

We consider the target matrix $T = \mathcal{I}_p$ for all regularized estimators. Except for the Ledoit-Wolf estimator, all methods estimate both expectation and scatter. We verified that Algorithm 1 with $\psi(t) = \mathbb{1}[t < \tau]$ for $t \in [0, 1]$ yields results very close to those obtained with the MRCD estimator implemented in package rrcov. We also realized that results of MRWCD with ψ_i^D are rather similar to those obtained with ψ_i^{HT} for each fixed $i \in \{1, 2, 3\}$. Importantly, we verified that regularization is used only when necessary (for small values of n/p).

Denoting the parameters of non-contaminated data again by μ and Σ, we consider three measures to evaluate the performance of an estimate $\hat{\mu}$ of μ and the performance of a corresponding estimate $\hat{\Sigma}$ of Σ. These are defined as

$$E_1 = ||\hat{\mu} - \mu||_2^2, \quad E_2 = ||\hat{\Sigma} - \Sigma||_F^2, \quad E_3 = \operatorname{tr}(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1}) - p, \tag{32}$$

where $\operatorname{tr}(A)$ denotes the trace of a matrix. Naturally, the error measures are averaged over the simulated datasets.

Results of simulation A are presented in Table 2 for various values of p, n, and α. The empirical covariance matrix turns out to be outperformed by regularized counterparts even for normal data, as it becomes ill-conditioned in high dimensions. There are not big differences between results of MRCD (Boudt et al. 2020) and RegMCD (Gschwandtner and Filzmoser 2013). MRWCD outperforms them in some contaminated situations in terms of E_2. E_3 remains undefined for classical estimates for $n < p$; for other estimates, the results obtained with E_3 do not correspond much to those obtained with E_2. We do not consider E_3 to be a suitable measure as such, because it may be derived as the Kullback–Leibler divergence between two multivariate normal distributions with regular covariance matrices under the assumption of their common expectation. This also explains why the only application of E_3 to high-dimensional data appears to be presented in Boudt et al. (2020).

Results of simulations B and C are presented in Tables 3 and 4, respectively. Most often, the best estimates of Σ in terms of E_2 are obtained using MRWCD. Its performance seems especially good if the data are more heavily contaminated by severe outliers, which are shifted from the bulk of the data as well as having a large variability. For estimating μ, MRWCD stays only slightly behind the best results (if not the best), but the main contribution of MRWCD (also compared to MRCD) is mainly in estimating Σ.

Table 2 Measures of performance of estimating μ (E_1) and Σ (E_2 and E_3) in simulation A

Estimator	E_1	E_2	E_3	E_1	E_2	E_3	E_1	E_2	E_3
	Simulation A with $p = 50, n = 60$								
	$\alpha = 0$			$\alpha = 0.1$			$\alpha = 0.2$		
Classical	**0.91**	6.64	84.3	2.32	42.35	120.7	4.32	75.03	152.9
Ledoit-Wolf	–	**1.35**	**50.9**	–	40.53	87.4	–	72.42	118.6
RegMCD	1.10	7.22	62.0	**1.06**	7.27	**62.2**	**1.04**	7.57	**63.1**
MRCD	1.09	7.33	82.9	1.07	7.60	83.1	**1.04**	7.96	83.1
MRWCD ψ_1	1.07	6.10	73.7	1.12	6.32	69.1	1.48	13.23	68.2
MRWCD ψ_1^{HT}	1.28	6.77	81.1	1.29	6.56	80.9	1.31	6.97	81.5
MRWCD ψ_2	1.17	6.47	76.7	1.20	**6.17**	75.4	1.22	**6.89**	71.1
MRWCD ψ_3	0.94	6.12	71.4	1.65	13.00	65.1	3.21	22.50	75.4
MRWCD ψ_3^{HT}	1.26	6.77	81.1	1.27	6.57	81.0	1.29	6.98	81.7
	Simulation A with $p = 50, n = 200$								
	$\alpha = 0$			$\alpha = 0.05$			$\alpha = 0.20$		
Classical	**0.50**	3.59	57.0	1.17	22.01	75.3	4.28	72.82	125.3
Ledoit-Wolf	–	**0.71**	**50.3**	–	21.61	68.6	–	72.07	117.8
RegMCD	0.61	5.03	57.8	0.60	5.10	**57.9**	**0.58**	5.34	**58.6**
MRCD	0.59	4.37	60.8	**0.58**	4.43	60.7	**0.58**	4.51	60.3
MRWCD ψ_1	0.58	3.95	60.5	**0.58**	**3.89**	60.6	1.30	17.47	69.4
MRWCD ψ_1^{HT}	0.69	4.59	67.4	0.69	4.42	67.6	0.76	**4.46**	67.5
MRWCD ψ_2	0.64	4.30	63.6	0.64	4.14	63.6	0.74	5.29	64.9
MRWCD ψ_3	0.51	3.61	57.7	0.83	12.86	66.6	3.34	28.10	78.3
MRWCD ψ_3^{HT}	0.68	4.55	66.7	0.68	4.40	67.0	0.74	4.44	66.9

For each scenario, the choices of p, n and α are specified. The smallest value in each simulation is shown in boldface

MRWCD estimates with non-robust weights perform particularly well. The reason why MRWCD estimates with robust weights stay (only slightly) behind is that MRWCD with hard trimming is always computed with trimming away 25 % of observations, although the true contamination level is lower in all simulations. An adaptive search for data-dependent weights, extending the ideas of Čížek (2011) or Cerioli et al. (2018), may have a potential to lead to further improvements. The effect of replacing $T = \mathcal{I}_p$ by an alternative target matrix is revealed in Table 4 to be rather small.

4.1 Effect of regularization

To study how the estimated values of ρ are affected by different choices of p, n and α, where the last corresponds to the true contamination level, we present averaged values of ρ evaluated within a particular study over 1000 randomly generated datasets in Table 5. Simulation B with $\mu_B = (2, \ldots, 2)^T$, $c = 5$, and ψ_1^T is considered there. It turns out that no regularization is used when it is not needed. Further, ρ increases with an increasing ratio p/n, and decreases with an increasing contamination, because the

Table 3 Measures of performance of estimating μ (E_1) and Σ (E_2 and E_3) in simulation B

Estimator	E_1	E_2	E_3	E_1	E_2	E_3
	Simulation B with $p = 200$ and $n = 50$					
	$\alpha = 0.1, c = 5$			$\alpha = 0.2, c = 3$		
	$\mu^B = (0, \ldots, 0)^T$			$\mu^B = (1, \ldots, 1)^T$		
Classical	2.40	53.6	–	3.66	62.5	–
Ledoit-Wolf	—	**7.7**	**219**	–	32.0	**245**
RegMCD	2.33	30.9	307	2.37	31.2	309
MRCD	2.30	28.5	434	**2.31**	29.3	433
MRWCD ψ_1	2.30	21.6	335	2.44	22.8	329
MRWCD ψ_1^{HT}	2.45	22.2	337	2.46	23.0	337
MRWCD ψ_2	2.52	22.6	331	2.55	23.3	331
MRWCD ψ_3	**2.18**	22.7	307	2.99	**21.1**	273
MRWCD ψ_3^T	2.35	22.3	347	2.36	22.9	347
	Simulation B with $p = 400$ and $n = 100$					
	$\alpha = 0.1, c = 5$			$\alpha = 0.1, c = 3$		
	$\mu^B = (2, \ldots, 2)^T$			$\mu^B = (3, \ldots, 3)^T$		
Classical	4.63	177.5	–	6.41	343.9	–
Ledoit-Wolf	–	139.6	569	–	321.1	729
RegMCD	2.31	43.5	616	2.34	43.5	615
MRCD	**2.29**	40.5	858	2.33	40.9	861
MRWCD ψ_1	2.33	26.6	562	2.38	22.0	486
MRWCD ψ_1^{HT}	2.39	31.0	656	2.38	31.1	659
MRWCD ψ_2	2.49	31.8	647	2.47	32.1	649
MRWCD ψ_3	3.29	**16.4**	**439**	4.26	**18.3**	**430**
MRWCD ψ_3^{HT}	2.30	31.1	678	**2.27**	31.0	681

For each scenario, the choices of p, n and parameters characterizing the data contamination are specified. The smallest value in each simulation is shown in boldface

outliers contribute to improving the numerical stability (and decreasing the condition number) of the estimated scatter matrix.

4.2 Outlier detection

The outlier detection performance of methods of Sect. 2.6 will now be compared in a simulation study inspired by Gschwandtner and Filzmoser (2013). Each dataset contains the "good" observations randomly generated as independent values from $N_p(0, \sigma^*)$, where Σ^* corresponds to $\mathcal{I}_p$ except for the covariances $\Sigma^*_{12} = \Sigma^*_{21} = 0.7$. In addition, each dataset contains $\lfloor n\alpha \rfloor$ contaminated observations, which are obtained as independent $N(\mu_{out}, \sigma_{out}\mathcal{I}_p)$ values. The plain approach assigns a given (i-th) observation as an outlier if and only if its corresponding robust Mahalanobis distance exceeds $\chi^2_{p,1-\delta}$. The MRCD and RegMCD estimators triming away 25 % of the

Table 4 Measures of performance of estimating μ (E_1) and Σ (E_2 and E_3) in Simulation C

Estimator	E_1	E_2	E_3	E_1	E_2	E_3
	Simulation C					
	$p = 200, n = 50$			$p = 300, n = 200$		
	$\alpha = 0.2$			$\alpha = 0.1$		
	$\mu^C = (2, \dots, 2)^T$			$\mu^C = (1, \dots, 1)^T$		
Classical	3.96	86.5	–	2.49	173.1	–
Ledoit-Wolf	–	80.8	188	–	170.0	153
RegMCD	3.49	18.4	**162**	**2.24**	15.1	**136**
MRCD	**3.46**	19.3	507	**2.24**	16.7	259
	MRWCD with $T = \mathcal{I}_p$					
MRWCD ψ_1	3.62	19.6	416	2.33	18.4	312
MRWCD ψ_1^{HT}	3.58	18.2	327	2.31	15.9	227
MRWCD ψ_2	3.74	21.5	458	2.47	19.0	385
MRWCD ψ_3	3.78	20.1	330	2.45	19.1	323
MRWCD ψ_3^{HT}	3.61	**18.0**	283	2.26	**14.7**	256
	MRWCD with target matrix (12)					
MRWCD ψ_1	3.61	19.9	441	2.33	18.8	330
MRWCD ψ_1^{HT}	3.58	19.1	389	2.30	15.5	248
MRWCD ψ_2	3.75	21.2	492	2.48	20.6	361
MRWCD ψ_3	3.77	20.8	373	2.45	19.7	347
MRWCD ψ_3^{HT}	3.61	18.4	325	2.27	16.3	266
	MRWCD with target matrix (13)					
MRWCD ψ_1	3.62	20.6	424	2.34	18.8	322
MRWCD ψ_1^{HT}	3.58	19.5	396	2.31	15.7	232
MRWCD ψ_2	3.74	21.1	524	2.47	20.4	369
MRWCD ψ_3	3.76	20.9	351	2.44	19.2	341
MRWCD ψ_3^{HT}	3.62	19.1	290	2.26	16.6	252

For each scenario, the choices of p, n, α and μ_C are specified. The smallest value in each simulation is shown in boldface

Table 5 Simulation comparison of the effect of regularization of Sect. 4.1: average values of ρ in Simulation B with $\mu_B = (2, \dots, 2)^T$, $c = 5$, and ψ_1^T

p	n	α	Average ρ	p	n	α	Average ρ
50	100	0.0	0.000	200	200	0.0	0.088
50	100	0.1	0.000	200	200	0.1	0.069
50	100	0.2	0.000	200	200	0.2	0.056
100	100	0.0	0.085	400	100	0.0	0.195
100	100	0.1	0.073	400	100	0.1	0.156
100	100	0.2	0.056	400	100	0.2	0.134
200	100	0.0	0.121	400	200	0.0	0.131
200	100	0.1	0.106	400	200	0.1	0.109
200	100	0.2	0.094	400	200	0.2	0.100

observations are used here, and the MRWCD estimator is used with the loss function ψ_1^{HT}.

The results obtained for different values of p, n, α and μ_{out} are presented in Table 6 for $\sigma_{out} = 1$. In the table, averaged ratios of false negative (FN) and false positive (FP) results computed across 100 simulations are reported. For example, the FDR-F method used with the MRWCD estimator yields the average FN equal to 0.04 for $p = 50, n = 100, \alpha = 0.1$ (i.e. 10 outliers), and $\mu_{out} = 1$.

It follows from Table 6 that the outlier detection performance increases with an increasing ratio p/n (when all other parameters are retained). The performance is also improved with an increasing μ_{out}, although false positivity does not decrease to 0 in some situations, especially for MRCD; this is a consequence of considering 75 % of observations in the estimate, without reflecting the true contamination level. FDR-χ^2 and especially FDR-F approaches yield better results compared to the plain ones; the improvement of FDR-based hypothesis testing is remarkable for such observations, for which the robust Mahalanobis distances are only moderately above $\chi^2_{p,0.975}$. The choice of σ_{out} turns out to have a negligible effect here. For $p > n$, RegMCD turns out to represent the best estimator for outlier detection for larger μ_{out}, while MRWCD with the FDR-F procedure outperforms all other tools for smaller μ_{out}.

5 Conclusions

This work fills the gap of robust implicitly weighted estimation of expectation and scatter of high-dimensional data. A real data example with gene expressions illustrates a data-driven outlier detection (and finding a suitable number of observations detected as outliers). In the task to estimate the expectation and scatter matrix of high-dimensional (possibly contaminated) data, simulations reveal the MRWCD estimator to outperform other tools in some situations. This is especially true in the task to estimate the scatter matrix when using E_2 as the criterion for the estimation performance. The MRWCD estimator, which depends on the choice of a particular weight function, performs the best in several simulation scenarios, especially in estimating the scatter matrix of contaminated high-dimensional data. Simulations of Sect. 4.2 reveal the MRWCD estimator to perform well also in the outlier detection task, especially if using the FDR-F approach.

The MRWCD estimator is computationally feasible for data with the number of variables in the order of thousands. This is true in spite of the fact that the MRWCD estimator extends the MRCD to allow for any continuous weights. Indeed, Algorithm 1 is much less demanding compared to a direct extension of the FAST-MCD, as it requires only a relatively small number of generalized C-steps; in fact, the computation of the 6 initial estimates is the most demanding part of the computation.

None of the weight functions seems to be the best across a wide range of scenarios. It seems as a possibility for future research to use metalearning (automatic method selection) to predict the most suitable weights for a given dataset based on suitable data features, as a multivariate analogy of the regression approach of Kalina and Tichavský (2019). We agree with Tong et al. (2018) that the choice of T in regularized covariance matrices should not be performed (estimated) based on the data. Instead, it

Table 6 Outlier detection performance in simulations of Sect. 4.2: the average percentages of false negatives (FN) and false positives (FP) for three different outlier detection approaches of Sect. 2.6

Estimator	Outlier detection	FN $\mu_{out} = 0.5$	FP	FN $\mu_{out} = 1$	FP	FN $\mu_{out} = 5$	FP
$p = 50, n = 100, \alpha = 0.1$							
RegMCD	Plain	1.00	0.00	0.59	0.00	0.00	0.00
RegMCD	FDR-χ^2	1.00	0.00	0.23	0.12	0.00	0.08
RegMCD	FDR-F	1.00	0.00	0.21	0.12	0.00	0.08
MRCD	Plain	0.52	0.23	0.08	0.18	0.00	0.18
MRCD	FDR-χ^2	0.26	0.19	0.05	0.11	0.00	0.09
MRCD	FDR-F	0.23	0.11	0.04	0.07	0.00	0.05
MRWCD	Plain	0.51	0.18	0.08	0.17	0.00	0.18
MRWCD	FDR-χ^2	0.31	0.17	0.07	0.07	0.00	0.02
MRWCD	FDR-F	0.28	0.10	0.04	0.05	0.00	0.02
$p = 100, n = 100, \alpha = 0.1$							
RegMCD	Plain	1.00	0.00	0.22	0.00	0.00	0.00
RegMCD	FDR-χ^2	0.85	0.00	0.15	0.07	0.00	0.06
RegMCD	FDR-F	0.33	0.00	0.13	0.05	0.00	0.05
MRCD	Plain	0.44	0.22	0.00	0.18	0.00	0.17
MRCD	FDR-χ^2	0.19	0.15	0.05	0.09	0.00	0.08
MRCD	FDR-F	0.17	0.13	0.04	0.06	0.00	0.05
MRWCD	Plain	0.23	0.17	0.00	0.14	0.00	0.13
MRWCD	FDR-χ^2	0.17	0.11	0.00	0.02	0.00	0.02
MRWCD	FDR-F	0.14	0.09	0.00	0.02	0.00	0.02
$p = 200, n = 100, \alpha = 0.1$							
RegMCD	Plain	0.98	0.00	0.01	0.00	0.00	0.00
RegMCD	FDR-χ^2	0.62	0.10	0.00	0.03	0.00	0.03
RegMCD	FDR-F	0.29	0.09	0.00	0.01	0.00	0.01
MRCD	Plain	0.27	0.21	0.00	0.17	0.00	0.14
MRCD	FDR-χ^2	0.15	0.12	0.04	0.08	0.00	0.06
MRCD	FDR-F	0.13	0.10	0.03	0.05	0.00	0.04
MRWCD	Plain	0.33	0.15	0.00	0.11	0.00	0.11
MRWCD	FDR-χ^2	0.16	0.08	0.00	0.02	0.00	0.01
MRWCD	FDR-F	0.13	0.07	0.00	0.02	0.00	0.01
$p = 400, n = 100, \alpha = 0.1$							
RegMCD	Plain	0.85	0.00	0.00	0.00	0.00	0.00
RegMCD	FDR-χ^2	0.28	0.08	0.00	0.00	0.00	0.00
RegMCD	FDR-F	0.21	0.05	0.00	0.00	0.00	0.00
MRCD	Plain	0.22	0.13	0.00	0.09	0.00	0.06

Table 6 continued

Estimator	Outlier detection	FN $\mu_{out} = 0.5$	FP	FN $\mu_{out} = 1$	FP	FN $\mu_{out} = 5$	FP
MRCD	FDR-χ^2	0.13	0.10	0.03	0.03	0.00	0.03
MRCD	FDR-F	0.09	0.07	0.01	0.03	0.00	0.02
MRWCD	Plain	0.24	0.11	0.00	0.05	0.00	0.05
MRWCD	FDR-χ^2	0.10	0.06	0.00	0.01	0.00	0.01
MRWCD	FDR-F	0.07	0.04	0.00	0.01	0.00	0.01

The MRWCD estimator is used with ψ_1^{HT}

has been recommended to choose T not to contain information learned from the given data (Schäfer and Strimmer 2005). On the other hand, the choice of T may reflect a prior knowledge related to the structure of data in a Bayesian setup, as discussed in DeMiguel et al. (2013).

Possible applications or extensions of the MRWCD estimator can be found within robust regularized linear discriminant analysis or within machine learning (Rusiecki 2008), or for cell-wise down-weighting of cell-wise contaminated high-dimensional observations within methods of Agostinelli et al. (2015) or Van Aelst (2016).

Acknowledgements The research was supported by the projects GA21-05325S and GA19-05704S of the Czech Science Foundation. The authors would like to thank Jurjen Duintjer Tebbens for discussion, and would like to thank the anonymous referees, an associate editor, and the editor-in-chief for their time and constructive advice.

References

Agostinelli C, Leung A, Yohai VJ, Zamar RH (2015) Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST 24:441–461

Ashurbekova K, Usseglio-Carleve A, Forbes F, Achard S (2019) Optimal shrinkage for robust covariance matrix estimators in a small sample size setting. https://hal.archives-ouvertes.fr/hal-02378034

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Boudt K, Rousseeuw PJ, Vanduffel S, Verdonck T (2020) The minimum regularized covariance determinant estimator. Stat Comput 30:113–128

Cerioli A (2010) Multivariate outlier detection with high-breakdown estimators. J Am Stat Assoc 105:147–156

Cerioli A, Farcomeni A (2011) Error rates for multivariate outlier detection. Comput Stat Data Anal 55:544–553

Cerioli A, Riani M, Atkinson AC, Corbellini A (2018) The power of monitoring: how to make the most of a contaminated multivariate sample. Stat Methods Appl 27:559–587

Chen Y, Wiesel A, Hero AO (2011) Robust shrinkage estimation of high dimensional covariance matrices. IEEE Trans Signal Process 59:4097–4107

Čížek P (2011) Semiparametrically weighted robust estimation of regression models. Comput Stat Data Anal 55:774–788

Couillet R, McKay M (2014) Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. J Multivar Anal 131:99–120

DeMiguel V, Martin-Utrera A, Nogales FJ (2013) Size matters: optimal calibration of shrinkage estimators for portfolio selection. J Bank Finance 37:3018–3034

Filzmoser P, Todorov V (2011) Review of robust multivariate statistical methods in high dimension. Anal Chinica Acta 705:2–14

Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Comput Stat Data Anal 52:1694–1711

Fritsch V, Varoquaux G, Thyreau B, Poline JB, Thirion B (2011) Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant. Lect Notes Comput Sci 6893:264–271

Gschwandtner M, Filzmoser P (2013) Outlier detection in high dimension using regularization. In: Kruse R et al (eds) Synergies of soft computing and statistics. Springer, Berlin, pp 37–244

Gschwandtner M, Filzmoser P, Croux C, Haesbroeck G (2012) rrlda: robust regularized linear discriminant analysis. R package version 1.1. https://CRAN.R-project.org/package=rrlda

Hardin J, Rocke DM (2005) The distribution of robust distances. J Comput Graph Stat 14:928–946

Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press, Boca Raton

Hubert M, Debruyne M (2010) Minimal covariance determinant. Wiley Interdiscip Rev Comput Stat 2:36–43

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

Hubert M, Rousseeuw PJ, Verdonck T (2012) A deterministic algorithm for robust location and scatter. J Comput Graph Stat 21:618–637

Hubert M, Debruyne M, Rousseeuw PJ (2018) Minimum covariance determinant and extensions. WIREs Comput Stat 10:e1421

Jurečková J, Sen PK, Picek J (2013) Methodology in robust and nonparametric statistics. CRC Press, Boca Raton

Jurečková J, Picek J, Schindler M (2019) Robust statistical methods with R, 2nd edn. CRC Press, Boca Raton

Kalina J (2021) The minimum weighted covariance determinant estimator revisited. Commun Stat Simul Comput. https://doi.org/10.1080/03610918.2020.1725818

Kalina J, Tichavský J (2019) Statistical learning for recommending (robust) nonlinear regression methods. J Appl Math Stat Inform 15(2):47–59

Kalina J, Tichavský J (2020) On robust estimation of error variance in (highly) robust regression. Meas Sci Rev 20:6–14

Kalina J, Hlinka J, (2017) Implicitly weighted robust classification applied to brain activity research. In: Fred A, Gamboa H (eds) Biomedical engineering systems and technologies BIOSTEC, (2016) Communications in Computer and Information Science 690. Springer, Cham, pp 87–107

Karjanto S, Ramli NM, Ghani NAM, Aripin R, Yusop NM (2015) Shrinkage covariance matrix approach based on robust trimmed mean in gene sets detection. AIP Conf Proc 1643:225–231

Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 88:365–411

Lee K, You K (2019) CovTools: statistical tools for covariance analysis. R package version 0.5.3. https://CRAN.R-project.org/package=CovTools

Marozzi M, Mukherjee A, Kalina J (2020) Interpoint distance tests for high-dimensional comparison studies. J Appl Stat 47:653–665

Pourahmadi M (2013) High-dimensional covariance estimation. Wiley, Hoboken

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org

Ro K, Zou C, Wang Z (2015) Outlier detection for high-dimensional data. Biometrika 102:589–599

Roelant E, Van Aelst S, Willems G (2009) The minimum weighted covariance determinant estimator. Metrika 70:177–204

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880

Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. J Am Stat Assoc 88:1273–1283

Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223

Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York

Rousseeuw PJ, Van Zomeren BC (1990) Unmasking multivariate outliers and leverage points. J Am Stat Assoc 85:633–639

Rusiecki A (2008) Robust MCD-based backpropagation learning algorithm. Lect Notes Artif Intell 5097:154–163

Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4:32

Todorov V, Filzmoser P (2009) An object-oriented framework for robust multivariate analysis. J Stat Softw 32(3):1–47

Tong J, Hu R, Xi J, Xiao Z, Guo Q, Yu Y (2018) Linear shrinkage estimation of covariance matrices using low-complexity cross-validation. Signal Process 148:223–233

Van Aelst S (2016) Stahel–Donoho estimation for high-dimensional data. Int J Comput Math 93:628–639

Víšek JÁ (2006) The least trimmed squares. Part I: consistency. Kybernetika 42:1–36

Víšek JÁ (2011) Consistency of the least weighted squares under heteroscedasticity. Kybernetika 47:179–206

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.