Statistical Papers (2021) 62:1583–1609
https://doi.org/10.1007/s00362-019-01148-1

REGULAR ARTICLE

Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators

Elisa Cabana[1] · **Rosa E. Lillo**[2] · **Henry Laniado**[3]

Received: 10 December 2018 / Revised: 14 May 2019 / Published online: 20 November 2019

Abstract
A collection of robust Mahalanobis distances for multivariate outlier detection is proposed, based on the notion of shrinkage. Robust intensity and scaling factors are optimally estimated to define the shrinkage. Some properties are investigated, such as affine equivariance and breakdown value. The performance of the proposal is illustrated through the comparison to other techniques from the literature, in a simulation study and with a real dataset. The behavior when the underlying distribution is heavy-tailed or skewed, shows the appropriateness of the method when we deviate from the common assumption of normality. The resulting high true positive rates and low false positive rates in the vast majority of cases, as well as the significantly smaller computation time show the advantages of our proposal.

Keywords Multivariate distance · Robust location and covariance matrix estimation · Comedian matrix · Multivariate L_1-median

1 Introduction

The detection of outliers in multivariate data is an important task in Statistics, since that kind of data can distort any statistical procedure (Tarr et al. 2016). The task of

This research was partially supported by MINISTERIO DE ECONOMIA, INDUSTRIA Y COMPETITIVIDAD, award number: ECO2015-66593-P.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00362-019-01148-1) contains supplementary material, which is available to authorized users.

✉ Elisa Cabana
ecab.estadistica@gmail.com

1 Department of Statistics, University Carlos III of Madrid, Madrid, Spain

2 Department of Statistics, and UC3M-Santander Big Data Institute, University Carlos III of Madrid, Madrid, Spain

3 Department of Mathematical Sciences, University EAFIT, Medellín, Colombia

detecting multivariate outliers can be useful in various fields such as quality control, medicine, finance, image analysis, and chemistry (Vargas 2003; Brettschneider et al. 2008; Hubert et al. 2008; Hubert and Debruyne 2010; Perrotta and Torti 2010; Choi et al. 2016). The concept of outlier is not well-defined in the literature since different authors tend to have varying definitions. Although they are generally defined as observations resulting from a secondary process, which differ from the underlying distribution. Thus, the outliers will differ from the main bulk of the data. They do not need to be especially high or low for all the variables in the data set (Alqallaf et al. 2009), that is why the task of identifying multivariate outliers with the classical univariate methods commonly fail. In the multivariate sense, there must be considered both the distance of an observation from the centroid of the data, and the shape of the data. The Mahalanobis distance (Mahalanobis 1936) is a well-known measure which takes it both into account. For multivariate Gaussian data, the distribution of the squared Mahalanobis distance, MD^2, is known (Gnanadesikan and Kettenring 1972) to be chi-squared with p (the dimension of the data, the number of variables) degrees of freedom, i.e. χ^2_p. Then, the adopted rule for identifying the outliers is selecting the threshold as the 0.975 quantile of the χ^2_p (Maronna and Zamar 2002).

However, outliers need not necessarily have large MD values (Masking problem) and not all observations with large MD values are necessarily outliers (Swamping problem). The problems of masking and swamping arise due to the influence of outliers on classical location and scatter estimates (sample mean and sample covariance matrix), which implies that the estimated distance will not be robust to outliers. The solution is to consider robust estimators of centrality and covariance matrix to obtain a robust Mahalanobis distance (*RMD*). Many robust estimators for location and covariance have been introduced in the literature (Maronna and Yohai 1976). Rousseeuw (1985) proposed the minimum covariance determinant (*MCD*) estimator based on the computation of the ellipsoid with the smallest volume or with the smallest covariance determinant that would encompass at least half of the data points. The procedure required naive subsampling for minimizing the objective function of the *MCD*, but an improvement much more effective, the *Fast-MCD*, was introduced by Rousseeuw and Driessen (1999) and a code is available in MATLAB (Verboven and Hubert 2005). Unfortunately *Fast-MCD* still requires substantial running times for large p, because the number of candidate solutions grows exponentially with the dimension p of the sample and, as a consequence, the procedure becomes computationally expensive for even moderately sized problems.

On the other hand, the squared *RMD* distributional fit usually breaks down, i.e. it does not necessarily have to follow a chi-squared distribution when you deviate from the Gaussian distribution. Thus, determining exact cutoff values for outlying distances continues to be a difficult problem and it has found much attention because no universally applicable method has been proposed. Despite this fact, the $\chi^2_{p;0.975}$ quantile is often considered as threshold for recognizing outliers in the robust distance case, but this approach may have some drawbacks. Evidence of this behavior is now well documented even in moderately large samples, especially when the number of variables increases (Becker and Gather 1999; Hardin and Rocke 2005; Cerioli et al. 2009; Cerioli et al. 2008). It is crucial to determine the threshold for the distances in

order to decide whether an observation is an outlier. Filzmoser et al. (2005) proposed to use an adjusted quantile, instead of the classical choice of the $\chi^2_{p;0.975}$ quantile. The adjusted threshold is estimated adaptively from the data, but their proposal is defined for a specific robust Mahalanobis distance, the one based on the *MCD* estimator. Let us call this method *Adj MCD*. Peña and Prieto (2001) and Peña and Prieto (2007) proposed an algorithm called *Kurtosis*, based on the analysis of the projections of the sample points onto a certain set of directions obtained by maximizing and minimizing the kurtosis coefficient of the projections, and some random directions generated by a stratified sampling scheme. With the combination of random and specific directions, the authors proposed a powerful procedure for robust estimation and outlier detection. However, this procedure has some drawbacks when the dimension p of the sample space grows, and in presence of correlation between the variables, the method looses power (Marcano and Fermín 2013). Maronna and Zamar (2002) proposed the Orthogonalized Gnanadesikan–Kettenring (*OGK*) estimator. It was the result of applying a general method to the pairwise robust scatter matrix from Gnanadesikan and Kettenring (1972), in order to obtain a positive-definite scatter matrix. On the other hand, a reweighing step can be used to identify outliers, where atypical observations get weight 0 and normal observations get weight 1. Sajesh and Srinivasan (2012) proposed the Comedian method (*COM*) to detect outliers from multivariate data based on the comedian matrix estimator from Falk (1997). The method is found to be efficient under various simulation scenarios and suitable in high-dimensional data. Furthermore, there are several real scenarios where the number of variables is high in which outlier detection is very important. For example, medical imaging datasets often contain deviant observations due to pre-processing artifacts or large intrinsic inter-subject variability (Lazar 2008; Lindquist 2008; Monti 2011; Poline and Brett 2012), in biological and financial studies (Chen et al. 2010; Zeng et al. 2015), and also in geochemical data, because of their complex nature (Reimann and Filzmoser 2000; Templ et al. 2008).

In this article, a collection of *RMD's* are proposed for outlier detection especially in high dimension. They are based on considering different combinations of robust estimators of location and covariance matrix. Two basic options are considered for the location parameter: a component-wise median and the L_1 multivariate median (Gower 1974; Brown 1983; Dodge 1987; Small 1990). A notion called *shrinkage estimator* (Ledoit and Wolf 2003a, b, 2004; DeMiguel et al. 2013; Gao 2016; Sun et al. 2018; Steland 2018) is considered, which is aimed to reduce estimation error. The shrinkage is applied to both of the previous mentioned location estimators. As for the covariance matrix, the options basically consists on a shrinkage estimator over special cases of *comedian matrices* (Hall and Welsh 1985; Falk 1997), which are based on a location parameter that will be estimated using a robust estimator of centrality in a way that a *RMD* can be obtained with meaningful combinations of both location and covariance matrix estimators. Simulation results demonstrates the satisfactory practical performance of our proposal, especially when the number of variables grows. The computational cost is studied by both simulations and a real dataset example.

The paper is organized as follows. Section 2 describes the shrinkage estimators both for the location and the covariance matrix, and the proposed combinations of these estimators in order to define a *RMD*. Section 3 shows a simulation study with

contaminated multivariate Gaussian data and when we deviate from the Gaussian assumption, e.g. with skewed or heavy-tailed data, to compare the proposal with the other robust approaches: *MCD*, *Adj MCD*, *Kurtosis*, *OGK* and *COM*. In Sect. 4 other simulation scenarios are proposed with correlated data, transformed data and large contaminated data to investigate the properties of affine equivariance, breakdown value and contamination under correlation. The computational times are also introduced in this section. Section 5 shows the behavior with a real dataset example. Finally, Sect. 6 provides some conclusions.

2 A robust Mahalanobis distance based on shrinkage estimator

The classical Mahalanobis distance is defined for every p-dimensional observation $\mathbf{x}_i$ of the multivariate sample $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, as:

$$MD_i = ((\mathbf{x}_i - \hat{\boldsymbol{\mu}})\hat{\Sigma}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T)^{1/2}, \tag{1}$$

where $\hat{\boldsymbol{\mu}}$ is the estimated multivariate location (sample mean) and $\hat{\Sigma}$ is the estimated covariance matrix (sample covariance matrix).

Since the problem with this definition is that the classical estimates of location and covariance matrix are often highly influenced by the presence of outliers (Rousseeuw and Van Zomeren 1990), the solution is to consider robust estimates of centrality and covariance matrix, i.e. resistant against the influence of outlying observations, giving rise to a robust Mahalanobis distance, defined as:

$$RMD_i := ((\mathbf{x}_i - \hat{\boldsymbol{\mu}}_R)\hat{\Sigma}_R^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_R)^T)^{1/2}, \tag{2}$$

where $\hat{\boldsymbol{\mu}}_R$ and $\hat{\Sigma}_R$ are robust estimators of centrality and covariance matrix, respectively.

We propose to use a notion which is frequently used in finance and portfolio optimization, known as *shrinkage* (Eq. 3). It is widely used in those fields because its good performance for "large p small n" problems (see Couillet and McKay 2014; Chen et al. 2011; Steland 2018), although we focus on data with $n > p$. This estimator $\hat{E}_{Sh}$ relies on the fact that "shrinking" an estimator $\hat{E}$ of a parameter θ towards a target estimator $\hat{T}$, would help to reduce the estimation error, because although the shrinkage target is usually biased, it also contains less variance than the estimator $\hat{E}$. Therefore, under general conditions, there exists a *shrinkage intensity* η, so the resulting shrinkage estimator would contain less estimation error than $\hat{E}$ (James and Stein 1961).

$$\hat{E}_{Sh} = (1 - \eta)\hat{E} + \eta\hat{T}. \tag{3}$$

The main advantage of using a shrinkage estimator is to obtain a trade-off between bias and variance. This approach can be applied to estimate both the location and dispersion parameters obtaining different meaningful combinations to define robust

Mahalanobis distances. In the case of covariance matrices shrinkage has the additional advantage that it is always positive definite and well conditioned.

2.1 Location parameter

Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ be the $n \times p$ data matrix with n being the sample size and p the number of variables. Based on the fact that the *median* is a better choice in terms of robustness, we start by considering as a location estimator the *component-wise median:*

$$\hat{\boldsymbol{\mu}}_{CCM} = (median(\mathbf{x}_1), ..., median(\mathbf{x}_p)), \tag{4}$$

where *median* denotes the univariate median and $(\mathbf{x}_j) = (x_{1j}, ..., x_{nj})^T$ for all $j = 1, ..., p$ is the jth column of $\mathbf{x}$.

Another option is to consider a multivariate median $\hat{\boldsymbol{\mu}}_{MM}$ called L_1*-median* which is a robust and highly efficient estimator of central tendency (Lopuhaa and Rousseeuw 1991; Vardi and Zhang 2000; Oja 2010). It is defined as:

$$\hat{\boldsymbol{\mu}}_{MM} = argmin_{\mathbf{x}_m,\ m\in\{1,...,n\}} \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_m - \mathbf{x}_i||_1 \,. \tag{5}$$

DeMiguel et al. (2013) proposed a shrinkage estimator over the sample mean, towards a scaled vector of ones as the target. In the same way we propose to study shrinkage estimators for both (4) and (5). Consider $\nu_{\boldsymbol{\mu}}\mathbf{e}$ as the target estimator $\hat{T}$ in (3), where $\mathbf{e}$ is the p-dimensional vector of ones, and consider $\hat{\boldsymbol{\mu}}_{CCM}$ as the sample estimator $\hat{E}$. Then, the shrinkage estimator over the component-wise median is:

$$\hat{\boldsymbol{\mu}}_{Sh(CCM)} = (1 - \eta)\hat{\boldsymbol{\mu}}_{CCM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e}\,. \tag{6}$$

The scaling factor $\nu_{\boldsymbol{\mu}}$ and the intensity η should minimize the expected quadratic loss, that is:

$$\begin{array}{ll} \min_{\nu_{\boldsymbol{\mu}},\eta} & E\left[\left\|\hat{\boldsymbol{\mu}}_{Sh(CCM)} - \boldsymbol{\mu}\right\|_2^2\right] \\ \text{s.t.} & \hat{\boldsymbol{\mu}}_{Sh(CCM)} = (1 - \eta)\hat{\boldsymbol{\mu}}_{CCM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e}, \end{array} \tag{7}$$

where $\|\mathbf{x}\|_2^2 = \sum_{j=1}^{p} x_j^2$.

Proposition 1 *The solution of the problem in* (7) *is:*

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{CCM}\mathbf{e}}{p}, \quad \hat{\eta} = \frac{E\left[\left\|\hat{\boldsymbol{\mu}}_{CCM} - \boldsymbol{\mu}\right\|_2^2\right]}{E\left[\left\|\hat{\boldsymbol{\mu}}_{CCM} - \hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\right\|_2^2\right]}\,. \tag{8}$$

See the proof in Section 1.1 from the Supplementary Material. Note that the denominator in the above expression (8) is estimable, but the numerator is not straightforward

because $\boldsymbol{\mu}$ is unknown. Then, it is necessary to provide another expression for the numerator. Chu (1955) investigated the distribution for the sample median estimator and obtained the following result about the variance in presence of normality. Fix j, for $j \in \{1, ..., p\}$:

$$\sigma^2_{\hat{\boldsymbol{\mu}}_{CCM\,j}} = Var(\hat{\boldsymbol{\mu}}_{CCM\,j}) = \frac{\pi}{2n}\sigma^2_{\mathbf{x_j}}\,. \tag{9}$$

Therefore, the numerator in the expression (8) for determining the $\hat{\eta}$ in Proposition 1 is:

$$\begin{aligned} E\left[\left\|\hat{\boldsymbol{\mu}}_{CCM} - \boldsymbol{\mu}\right\|_2^2\right] &= E\left[\sum_{j=1}^{p}(\hat{\boldsymbol{\mu}}_{CCM\,j} - \boldsymbol{\mu}_j)^2\right] \\ &= \sum_{j=1}^{p}\sigma^2_{\hat{\boldsymbol{\mu}}_{CCM\,j}} = \frac{\pi}{2n}\sum_{j=1}^{p}\sigma^2_{\mathbf{x}_j}\,. \end{aligned} \tag{10}$$

We need to estimate $\sigma^2_{\mathbf{x}_j}$ robustly and we will do so as explained in the next subsection with property (22).

On the other hand, consider $v_{\boldsymbol{\mu}}\mathbf{e}$ again as the target estimator $\hat{T}$ and consider $\hat{\boldsymbol{\mu}}_{MM}$ as the sample estimator $\hat{E}$, in (3). Then, the shrinkage estimator over the multivariate L_1-median is:

$$\hat{\boldsymbol{\mu}}_{Sh(MM)} = (1-\eta)\hat{\boldsymbol{\mu}}_{MM} + \eta v_{\boldsymbol{\mu}}\mathbf{e}\,. \tag{11}$$

The scaling factor $v_{\boldsymbol{\mu}}$ and the intensity η should minimize the expected quadratic loss:

$$\begin{aligned} \min_{v_{\boldsymbol{\mu}},\eta} \quad & E\left[\left\|\hat{\boldsymbol{\mu}}_{Sh(MM)} - \boldsymbol{\mu}\right\|_2^2\right] \\ \text{s.t.} \quad & \hat{\boldsymbol{\mu}}_{Sh(MM)} = (1-\eta)\hat{\boldsymbol{\mu}}_{MM} + \eta v_{\boldsymbol{\mu}}\mathbf{e}, \end{aligned} \tag{12}$$

where $\|\mathbf{x}\|_2^2 = \sum_{j=1}^{p} x_j^2$.

Proposition 2 *The solution of the problem in* (12) *is:*

$$\hat{v}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{MM}\mathbf{e}}{p}, \quad \eta = \frac{E\left[\left\|\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}\right\|^2\right]}{E\left[\left\|\hat{\boldsymbol{\mu}}_{MM} - \hat{v}_{\boldsymbol{\mu}}\mathbf{e}\right\|^2\right]}\,. \tag{13}$$

The proof is in Section 1.2 from the Supplementary Material. As in the previous case, the denominator in the η expression (13) can be described as:

$$E\left[\left\|\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}\right\|_2^2\right] = E\left[\sum_{j=1}^{p}(\hat{\boldsymbol{\mu}}_{MM\,j} - \boldsymbol{\mu}_j)^2\right] = \sum_{j=1}^{p}\sigma^2_{\hat{\boldsymbol{\mu}}_{MM\,j}}\,.$$

Bose and Chaudhuri (1993), Bose (1995) and Möttönen et al. (2010) investigated the asymptotic distribution for the L_1-median. In page 184, Section 3, from Möttönen et al. (2010), the authors describe the necessity of the following two assumptions, for $\mathbf{x}$ a p-variate random vector with cdf F, density function f and $p > 1$:

(C1) The p-variate density function of $\mathbf{x}$ is continuous and bounded.
(C2) The spatial median of the distribution of $\mathbf{x}$ is zero and unique.

According to Theorem 2, page 185, Section 3 in Möttönen et al. (2010), under assumptions C1 and C2, $\sqrt{n}\hat{\boldsymbol{\mu}}_{MM} \rightarrow_d N_p(\mathbf{0}, A^{-1}BA^{-1})$, where $\hat{\boldsymbol{\mu}}_{MM}$ is the observed spatial median, and A and B are the following:

$$A(\mathbf{x}) = \frac{1}{||\mathbf{x}||}\left[I_p - \frac{\mathbf{x}\mathbf{x}^t}{||\mathbf{x}||^2}\right] \qquad B(\mathbf{x}) = \frac{\mathbf{x}\mathbf{x}^t}{||\mathbf{x}||^2} \tag{14}$$

In Section 4, page 185 from Möttönen et al. (2010), the authors also provide an estimation for the asymptotic covariance matrix $A^{-1}BA^{-1}$ of the spatial median. They are assuming the true value $\hat{\boldsymbol{\mu}}_{MM} = \mathbf{0}$ is zero (condition C2). Then they write $\hat{A} = ave\{A(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MM})\}$ and $\hat{B} = ave\{B(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MM})\}$ and prove that under C1 and C2: $\hat{A} \rightarrow_P A$ and $\hat{B} \rightarrow_P B$, which means that the estimators converge in probability to the population values A and B, respectively. This result is Theorem 3, Section 4, page 185 from Möttönen et al. (2010). According to the authors (stated in page 186), Theorems 2 and 3 suggest that the distribution of $\hat{\boldsymbol{\mu}}_{MM}$ can be approximated by $N_p\left(\boldsymbol{\mu}, \frac{1}{n}\hat{A}^{-1}\hat{B}\hat{A}^{-1}\right)$, where $\hat{A}(\mathbf{x}_i) = \frac{1}{||\mathbf{x}_i||_2}\left(I_p - \frac{\mathbf{x}_i\mathbf{x}_i^T}{||\mathbf{x}_i||_2^2}\right)$ and $\hat{B}(\mathbf{x}_i) = \frac{\mathbf{x}_i\mathbf{x}_i^T}{||\mathbf{x}_i||_2^2}$, with $\mathbf{x}_i \in \mathbb{R}^p$, for each $i = 1, ..., n$.

The asymptotic result is also given in pages 9–11 of Becker et al. (2014) as well as the estimate for the approximate covariance matrix in page 11. The assumptions in that paper are analogous, but it can be seen that C2 assumption about the spatial median being zero is not necessary, only that it is unique and the density function f is bounded and continuous at μ [(Section 1.4, page 9 from Becker et al. (2014)]. The difference is that when approximating the covariance matrix, the data should be centered around the estimated spatial median.

The numerator $E\left[\left\|\hat{\boldsymbol{\mu}}_{MM} - \boldsymbol{\mu}\right\|^2\right]$ from the expression (13) can be approximated with $trace\left(\frac{1}{n}\hat{A}^{-1}\hat{B}\hat{A}^{-1}\right)$.

Then the estimators for $\hat{\nu}$ and $\hat{\eta}$ in Eq. 13 would be:

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{MM}\mathbf{e}}{p} \quad \text{and} \quad \hat{\eta} = \frac{trace\left(\frac{1}{n}\hat{A}^{-1}\hat{B}\hat{A}^{-1}\right)}{\left\|\hat{\boldsymbol{\mu}}_{MM} - \hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\right\|^2}.$$

2.2 Dispersion parameter

Based on the median concept, which is a robust measure of location, one can define a robust measure of dispersion for a random variable X, which is the *Median Absolute*

Deviation (MAD) from the data's median:

$$MAD(X) = median(|X - median(X)|)\,. \tag{15}$$

Falk (1997) showed the following relation, assuming normality, between the MAD and the standard deviation σ_X:

$$MAD(X) = \sigma_X \Phi^{-1}(3/4), \tag{16}$$

where Φ denotes the standard normal cdf. Taking the square in (16) we obtain a relation between the variance σ_X^2 and $MAD^2(X)$:

$$\sigma_X^2 = 2.198 \cdot MAD^2(X)\,. \tag{17}$$

Extending the idea of the MAD, a robust measure of dependence between two random variables X and Y is the *comedian* (Falk 1997):

$$COM\,(X, Y) = med\,((X - med\,(X))\,(Y - med\,(Y)))\,. \tag{18}$$

The comedian generalizes the MAD, because $COM(X, X) = MAD^2(X)$, and also has the highest possible breakdown point (Falk 1997). An important fact is that the comedian parallels the covariance, but the latter requires the existence of the first two moments of the two random variables, whereas the comedian always exists. Other known properties of the comedian are that it is symmetric, location invariant and scale equivariant. Furthermore, Hall and Welsh (1985) discussed about the strong consistency and asymptotic normality of the MAD, and Falk (1997) established similar results for the comedian.

Finally, a comedian matrix can be defined based on a multivariate version of (18). Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ be the $n \times p$ data matrix with n being the sample size and p the number of variables. Then the comedian matrix is defined as:

$$COM(\mathbf{x}) = \left(COM(\mathbf{x}_j, \mathbf{x}_t)\right), \quad j, t = 1, ..., p\,. \tag{19}$$

Note that from relation described in (17), one can consider the adjusted comedian:

$$\hat{S}_{CCM} = 2.198 \cdot COM(\mathbf{x})\,. \tag{20}$$

Note that $\hat{S}_{CCM}$ is a robust alternative for the covariance matrix, but in general it is not positive (semi-) definite (see Falk 1997). Since we need this property for inverting the covariance matrix in a Mahalanobis distance, we propose a shrinkage over $\hat{S}_{CCM}$, because of its advantage of providing always a positive definite and well-conditioned matrix. Therefore, if a shrinkage estimator is considered in (3) for the dispersion parameter:

$$\hat{\Sigma}_{Sh} = (1 - \eta)\hat{E} + \eta\hat{T}, \tag{21}$$

we propose to use in (21), the estimator $\hat{E} = \hat{S}_{CCM}$.

Recall the previous Sect. 2.1 in which we needed to provide a robust estimator for $\sigma^2_{\mathbf{x}_j}$, for each $j = 1, ..., p$ (Eq. 10) note that, because of the relation in (17):

$$\begin{aligned} trace(\hat{S}_{CCM}) &= \sum_{j=1}^{p} 2.198 \cdot COM(\mathbf{x}_j, \mathbf{x}_j) \\ &= \sum_{j=1}^{p} 2.198 \cdot MAD^2(\mathbf{x}_j) = \sum_{j=1}^{p} \sigma^2_{\mathbf{x}_j} . \end{aligned} \tag{22}$$

Thus, when considering a shrinkage estimator of the component-wise median, in order to estimate the variance of $\hat{\boldsymbol{\mu}}_{CCM}$ needed in the expression (8) for the shrinkage intensity $\hat{\eta}$, and according to the relation (10), we propose to estimate $\sum_{j=1}^{p} \sigma^2_{\mathbf{x}_j}$ using the $trace(\hat{S}_{CCM})$. Therefore, the estimates for $\hat{\nu}_{\boldsymbol{\mu}}$ and $\hat{\eta}$ in expression (8) can be calculated as:

$$\hat{\nu}_{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_{CCM}\mathbf{e}}{p} \quad \text{and} \quad \hat{\eta} = \frac{(\pi/2n)trace(\hat{S}_{CCM})}{\left\| \hat{\boldsymbol{\mu}}_{CCM} - \hat{\nu}_{\boldsymbol{\mu}}\mathbf{e} \right\|_2^2} .$$

Back to the dispersion parameter and its shrinkage estimator, several choices for the shrinkage target $\hat{T}$ have been proposed in the literature. For example, Ledoit and Wolf (2003b) proposed a weighted average of the sample covariance matrix and a single-index covariance matrix. Ledoit and Wolf (2003a) proposed selecting the shrinkage target as a "constant correlation matrix", whose correlations are set equal to the average of all sample correlations. Finally, Ledoit and Wolf (2004) proposed to use a multiple of the identity matrix as the shrinkage target. The authors proved that the resulting shrinkage covariance matrix is well-conditioned, even if the sample covariance matrix is not. There is also another approach introduced by DeMiguel et al. (2013). The authors proposed a shrinkage estimator both for the covariance matrix and its inverse. The estimators were constructed as a convex combination of the sample covariance matrix or its inverse, respectively, and a scaled shrinkage target, which they consider the scaled identity matrix as Ledoit and Wolf (2004). Therefore, we propose to use as shrinkage target $\hat{T} = \nu_{\Sigma} I$. Thus (21) results in:

$$\hat{\Sigma}_{Sh(CCM)} = (1 - \eta)\hat{S}_{CCM} + \eta\nu_{\Sigma} I . \tag{23}$$

Finally, the scaling parameter ν_{Σ} and the shrinkage intensity parameter η in (23) need to be estimated. They both are chosen to minimize the expected quadratic loss as in Ledoit and Wolf (2004):

$$\begin{aligned} \min_{\nu_{\Sigma},\eta} \quad & E\left[\left\| \hat{\Sigma}_{Sh} - \Sigma \right\|^2 \right] \\ \text{s.t.} \quad & \hat{\Sigma}_{Sh} = (1 - \eta)\hat{S}_{CCM} + \eta\nu_{\Sigma} I, \end{aligned} \tag{24}$$

where $\|A\|^2 = trace(AA^T)/p$.

Table 1 Combinations of location and dispersion

Name	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
$\hat{\mu}_R$	$\hat{\mu}_{CCM}$	$\hat{\mu}_{Sh(CCM)}$	$\hat{\mu}_{Sh(CCM)}$	$\hat{\mu}_{MM}$	$\hat{\mu}_{Sh(MM)}$	$\hat{\mu}_{Sh(MM)}$
$\hat{\Sigma}_R$	$\hat{\Sigma}_{Sh(CCM)}$	$\hat{\Sigma}_{Sh(CCM)}$	$\hat{\Sigma}_{Sh(Sh(CCM))}$	$\hat{\Sigma}_{Sh(MM)}$	$\hat{\Sigma}_{Sh(MM)}$	$\hat{\Sigma}_{Sh(Sh(MM))}$

Proposition 3 *The solution of the problem* (24) *is:*

$$\nu_\Sigma = trace(\hat{S}_{CCM})/p, \quad \eta = \frac{E\left[\left\|\hat{S}_{CCM} - \Sigma\right\|^2\right]}{E\left[\left\|\hat{S}_{CCM} - \nu_\Sigma I\right\|^2\right]}.$$

The proof can be found in Section 1 from the Supplementary Material. In practice, we propose to estimate the numerator of the expression for η as Ledoit and Wolf (2003a), Ledoit and Wolf (2003b) and Ledoit and Wolf (2004), but considering $\hat{S}_{CCM}$ instead of the sample covariance matrix, as the estimator of Σ.

Note that the comedian matrix depends on centered data considering the componentwise median $\hat{\boldsymbol{\mu}}_{CCM}$. A special case of comedian matrix can be defined if the data are centered using a different location estimator. We propose to center the data using the other location estimators described in Sect. 2.1, i.e. the multivariate L_1-median $\hat{\boldsymbol{\mu}}_{MM}$, and the shrinkage estimators $\hat{\boldsymbol{\mu}}_{Sh(CCM)}$ and $\hat{\boldsymbol{\mu}}_{Sh(MM)}$. We will consider shrinkages over those special comedian matrices.

1. $\hat{\Sigma}_{Sh(MM)} = (1-\eta)\hat{S}_{MM} + \eta\nu_\Sigma I$, with for $j, t = 1, ..., p$:
   $\hat{S}_{MM} = 2.198 \cdot COM_{MM}(\mathbf{x}) = 2.198 \cdot (med((\mathbf{x}_j - (\hat{\boldsymbol{\mu}}_{MM})_j)(\mathbf{x}_t - (\hat{\boldsymbol{\mu}}_{MM})_t))$.
2. $\hat{\Sigma}_{Sh(Sh(CCM))} = (1-\eta)\hat{S}_{Sh(CCM)} + \eta\nu_\Sigma I$, with for $j, t = 1, ..., p$:
   $\hat{S}_{Sh(CCM)} = 2.198 \cdot COM_{Sh(CCM)}(\mathbf{x}) = 2.198 \cdot (med((\mathbf{x}_j - (\hat{\boldsymbol{\mu}}_{Sh(CCM)})_j)(\mathbf{x}_t - (\hat{\boldsymbol{\mu}}_{Sh(CCM)})_t))$.
3. $\hat{\Sigma}_{Sh(Sh(MM))} = (1-\eta)\hat{S}_{Sh(MM)} + \eta\nu_\Sigma I$, with for $j, t = 1, ..., p$:
   $\hat{S}_{Sh(MM)} = 2.198 \cdot COM_{Sh(MM)}(\mathbf{x}) = 2.198 \cdot (med((\mathbf{x}_j - (\hat{\boldsymbol{\mu}}_{Sh(MM)})_j)(\mathbf{x}_t - (\hat{\boldsymbol{\mu}}_{Sh(MM)})_t))$.

The optimal expression for the parameters η and ν_Σ in the above cases is analogous to the Proposition 3, but considering in each case the sample estimator as the corresponding special comedian matrix.

2.3 Proposed robust Mahalanobis distances

A robust Mahalanobis distance can be defined as in (2), for each of the following 6 possible combinations for the location and the dispersion estimators (see Table 1). Note that they are meaningful combinations because the shrinkage estimator of dispersion is made upon a special comedian matrix closely based on the location estimator jointly considered for defining the *RMD*.

For all our proposed combinations, the threshold considered to detect the outliers is the $\chi^2_{p;0.975}$ quantile.

3 Simulation results

3.1 Normal distribution

A simulation study is performed considering a p-dimensional random variable X following a contaminated multivariate normal distribution given as a mixture of normals of the form $(1-\alpha)N(\mathbf{0}, I) + \alpha N(\delta\mathbf{e}, \lambda I)$, where $\mathbf{e}$ denotes the p-dimensional vector of ones. This model is analogous to the one used by Rousseeuw and Driessen (1999), Peña and Prieto (2001), Filzmoser et al. (2005), Peña and Prieto (2007), Maronna and Zamar (2002) and Sajesh and Srinivasan (2012). This experiment has been conducted for different values of the sample-space dimension $p = 5, 10, 30, 50$, and the chosen sample size in relation to the dimension was $n = 100, 100, 500, 1000$, respectively. The contamination levels were $\alpha = 0, 0.1, 0.2, 0.3$, the distance of the outliers $\delta = 5$ and 10 and the concentration of the contamination $\lambda = 0.1$ and 1. For each set of values, 100 random sample repetitions have been generated.

For the methods mentioned in previous sections some measures are studied: the true positive rate (TPR) and the false positive rate (FPR). If we call NO the real number of not outlying observations and TO the real number of outliers, then:

$$TPR = \frac{TP}{TO} \quad \text{and} \quad FPR = \frac{FP}{NO},$$

where TP means true positives and are the outliers correctly identified by the method, while FP means false positives and are the observations incorrectly detected as outliers by the method. The TPR is also equal to $1 - FNR = 1 - \frac{FN}{TO}$, where FN means False Negatives and are the outliers that the method fails to identify as such.

Then the two measures TPR and FPR are selected to study the performance of the methods. The method *MCD* refers to the *RMD* based on the *MCD* estimator and with the classical threshold, the method *Adj MCD* refers to the latter distance considering the adjusted quantile of Filzmoser et al. (2005), the method *Kurtosis* refers to the Peña and Prieto (2007) approach, the method *OGK* refers to the Orthogonalized Gnanadesikan–Kettenring method proposed by Maronna and Zamar (2002) and *COM* is the Comedian method proposed by Sajesh and Srinivasan (2012). We have also presented the results for the collection *RMDv1–RMDv6* proposed in Table 1. All simulations were performed in Matlab.

In the Supplementary Material, Section 2.1 shows the Tables 1–5 corresponding to all simulation scenarios with normal data. Here we show only the most significant and representative results. Nevertheless, the tables show general outcomes. For example, *Adj MCD*, actually improves *MCD* with respect to the FPR, lowering it, and in most cases maintaining the same TPR. Although, in other cases it also slightly lowers the TPR. On the other hand, the FPR in case of no contamination are sufficiently low for all methods, but our proposed collection shows the lowest values especially in

high dimension, actually here the best performance is observed for *RMDv6*. With certain percent of contamination, the worst behavior of our proposed methods is when dimension is low and the highest percentage of outliers are considered to be near the center of the data. This matter can be seen in Table 2 which corresponds to the TPR. When the outliers are near the center of the data ($\delta = 5$), in case of low dimension ($p = 5$), with 30% of outliers, *Kurtosis* has better performance. This happens also with $p = 10$, but in all other cases *MCD*, *Adj MCD*, *Kurtosis* and *OGK* are the ones with the worst behavior about the TPR. Meanwhile, *COM* is a good alternative, but the overall best performance is made by *RMDv6* especially in high dimension and even with large contamination.

Another situation is when outliers are far from the center of the data, i.e. $\delta = 10$. This scenario is shown in Table 3.

It is clear from Table 3 that when outliers are far from the center, our proposed methods lead to the best performance, achieving 100% of TPR, for all dimension and percentage of contamination considered. OGK and COM are good alternatives in case of high dimension $p = 30, 50$. Other tables about the TPR can be found in the Supplementary Material, as well as the FPR tables, which show that in the vast majority of cases our proposal have an FPR value equal to zero and when not, a value very close to zero, which is what is desirable.

3.2 t_3-distribution

In order to check the behavior of the methods when the distribution deviates from normality, a simulation study is performed considering a p-dimensional random variable X following a contaminated multivariate t-distribution with 3 degrees of freedom of the form $(1-\alpha)T_3(\mathbf{0}, I) + \alpha T_3(\delta\mathbf{e}, \lambda I)$. The first parameter of the notation of $T_3(\cdot, \cdot)$ refers to the mean and the second one to the covariance matrix. The parameters for the contamination are the same considered above and the same measures TPR and FPR are studied. All the results can be found in the Tables 9–13 from Section 2.2 in the Supplementary Material. It should be noted the unsatisfactory behavior of the alternative methods with respect to the TPR especially in high dimension or with large contamination level, meanwhile in most cases we attain a 100% TPR. With respect to the FPR value, all methods show non-zero FPR values, and the best performance is showed by *COM* and our proposed methods.

3.3 Exponential distribution

We considered also a p-dimensional random variable X following a contaminated multivariate exponential distribution given as a mixture $(1-\alpha)Exp(\mathbf{0}) + \alpha Exp(\delta\mathbf{e})$. The parameter of the notation $Exp(\cdot)$ refers to the mean. This case is analogous to the previous ones, with the difference that only the schemes associated with the distance of the outliers are considered. Tables 14–16 in Section 2.3 of the Supplementary Material show all the results and it can be seen that our proposed methods achieve 100% of TPR in the majority of cases. The highest value of TPR is also achieved by $Kurtosis$, OGK and COM, when dimension is high. When dimension is low, the TPR is high

Table 2 True positive rates, with Normal distribution

$\delta = 5$	$\lambda = 0.1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	0.9000	1	1	1	1	1	1	1	1
	0.2	0.8700	0.8700	0.5100	0.9500	0.9941	1	1	1	1	1	1
	0.3	0.0600	0.0600	0.9800	0.1500	0.5719	0.8766	0.8782	0.8782	0.9146	0.9090	0.9130
10	0.1	0.9900	0.9900	0.8600	1	1	1	1	1	1	1	1
	0.2	0.2800	0.2800	0.4600	0.9416	1	1	1	1	1	1	1
	0.3	0	0	0.9900	0.1612	0.7205	0.8774	0.8747	0.8750	0.9711	0.9672	0.9711
30	0.1	0.1900	0.1900	1	1	1	1	1	1	1	1	1
	0.2	0	0	0.1000	1	1	1	1	1	1	1	1
	0.3	0	0	0.6100	0.0100	0.9407	0.5308	0.5275	0.5286	0.9990	0.9988	0.9991
50	0.1	0	0	1	1	1	1	1	1	1	1	1
	0.2	0	0	0	1	1	1	1	1	1	1	1
	0.3	0	0	0	0	0.9839	0.5021	0.5000	0.5000	0.9939	0.9932	0.9942

Table 3 True positive rates, with Normal distribution

$\delta = 10$	$\lambda = 1$										
p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMDv1	RMDv2	RMDv3	RMDv4	RMDv5	RMDv6
5	0.1	1	1	1	1	1	1	1	1	1	1	1
	0.2	0.8480	0.8465	0.9900	1	1	1	1	1	1	1	1
	0.3	0.2190	0.1976	0.9307	0.9591	0.9991	1	1	1	1	1	1
10	0.1	1	1	0.9800	1	1	1	1	1	1	1	1
	0.2	0.8623	0.8548	0.6558	1	1	1	1	1	1	1	1
	0.3	0.2280	0.2046	0.4618	0.9911	1	1	1	1	1	1	1
30	0.1	1	1	0.8919	1	1	1	1	1	1	1	1
	0.2	0.4879	0.4654	0.0125	1	1	1	1	1	1	1	1
	0.3	0.0810	0.0509	0.1087	1	1	1	1	1	1	1	1
50	0.1	1	1	0.6017	1	1	1	1	1	1	1	1
	0.2	0.2695	0.2348	0.0017	1	1	1	1	1	1	1	1
	0.3	0.0643	0.0378	0.0006	1	1	1	1	1	1	1	1

in most situations for all $Kurtosis$, OGK, COM and $RMDv1 - -RMDv6$, but in the majority of cases our proposed method's TPR is higher. MCD and $AdjMCD$ decreases their TPR value with the increase of dimension or contamination level. With respect to the FPR value of $Kurtosis$ and OGK their FPR is high in most cases. COM, MCD and $AdjMCD$ have low FPR values. $RMDv1 - -RMDv6$ also have low FPR values in the majority of cases, except in some cases when the level of contamination is the lowest. On the other hand, in case of no contamination, MCD and $AdjMCD$ show more or less the same FPR value than our proposed methods, while the other alternatives $Kurtosis$, OGK and COM show higher values than them. Considering both the TPR and the FPR, the best overall performance is showed by *RMDv6*.

3.4 Summary and selection of one of our proposed distances

In the simulation study, for each contamination scheme we have also calculated a measure called F-score (Goutte and Gaussier 2005; Sokolova et al. 2006; Powers 2011), often used in Engineering, which is a measure of a test's accuracy. Its expression is F-score$= 2PR/(P + R)$, where P is called precision and R is known as the recall. The precision P is the number of correctly detected outliers divided by the total number of detected outliers, and the recall R is the number of correctly detected outliers divided by the real total number of outliers. The recall coincides with the TPR.

$$P = \frac{TP}{TP + FP} \text{ and } R = \frac{TP}{TP + FN}.$$

These measure provides a trade-off between the two desired outcomes: a high rate of correctly identified outliers and a low rate of observations mislabel as outliers. The results are not included in the paper for avoiding large extension, but the method with the overall classification between the top 3 best positions ranking with respect to the F-score, is method *RMDv6*.

It is clear the out-performance of our proposed methods with Gaussian data, especially in high dimension and even when we deviate from the normality assumption, for example when considering heavy-tailed and skewed distributions like the multivariate t_3-distribution and the multivariate exponential distribution. From all of our six proposed robust distances, the one that shows the best results in the vast majority of cases is *RMDv6*. Thus, we decided to select it as the best one in the matter of performance, and from now on we will refer to it as *RMD-S*.

4 Properties of the estimator

In this section some properties like the behavior under correlated data, the affine equivariance, the breakdown value, and the computational times are studied.

4.1 Correlation and affine equivariance

Consider $X = \{\mathbf{x_1}, ..., \mathbf{x_n}\}$ and a pair of multivariate location and covariance estimators (m, S). In general, these estimators are called affine equivariant if for any nonsingular matrix A it holds that:

$$m_A = m(X_A) = Am(X) \quad \text{and} \quad S_A = S(X_A) = S(X)A^T . \tag{25}$$

The affine transformation of X is $X_A = \{A\mathbf{x_1}, ..., A\mathbf{x_n}\}$. Affine equivariance implies that the estimator transforms well under any nonsingular reparametrization of the space of the $\mathbf{x}_i$. The data might for instance be rotated, translated or rescaled (for example through a change of the measurement units).

The method *RMD-S* is ultimately based on not affine equivariant estimators which are the L_1-median (Lopuhaa and Rousseeuw 1991) and the comedian matrix. However, the L_1-median is orthogonal equivariant, i.e. it satisfies Eq. (25) with A any orthogonal matrix ($A' = A^{-1}$). This implies that the L_1-median transforms appropriately under all transformations that preserve Euclidean distances (such as translations, rotations and reflections). About the comedian matrix, which always exists, it is symmetric, location invariant and scale equivariant (Falk 1997), i.e. $COM(X, \mathbf{a}Y+\mathbf{b}) = \mathbf{a}COM(X, Y) = \mathbf{a}COM(Y, X)$. Since the proposed method is not affine equivariant, it is important to investigate the behavior under correlated data. Devlin et al. (1981) used a correlation matrix P for generating Monte Carlo data from different distributions of moderate dimension $p = 6$. The matrix has the form:

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix} \tag{26}$$

$$P_1 = \begin{bmatrix} 1 & 0.95 & 0.3 \\ 0.95 & 1 & 0.1 \\ 0.3 & 0.1 & 1 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1 & -0.499 & -0.499 \\ -0.499 & 1 & -0.499 \\ -0.499 & -0.499 & 1 \end{bmatrix}. \tag{27}$$

The reason for the selection of the matrix P is because the dimension is large enough to study multivariate estimators and the range of correlation values is large. This way the differences in the abilities of the methods to detect outliers from highly correlated data can be observed. For the simulations, $n = 100$ observations were generated from a mixture of Normals $(1 - \alpha)N(\mathbf{0}, P) + \alpha N(5\mathbf{e}, P)$. The contamination level $\alpha = 10\%, 20\%, 30\%$.

Table 4 shows that the TPR and FPR of *MCD*, *Adj MCD*, *Kurtosis* and *OGK* are worse than that of our proposal . On the other hand, *COM* show more or less the same behavior in case of 10% and 20% of contamination, and slightly worse than our proposal when the contamination level increases to 30%. The methods *MCD*, *Adj MCD* and *Kurtosis* are affine equivariant, while *OGK* and *COM* are not. Hence, the proposed procedure *RMD-S* is more efficient than other affine and not affine equivariant methods in case of correlated datasets. Also the FPR is very low even in this case of presence of correlation.

Table 4 Simulation results for correlated data

α	MCD		Adj MCD		Kurtosis		OGK		COM		RMD-S	
	c	f	c	f	c	f	c	f	c	f	c	f
0.1	1	0.0397	1	0.0226	1	0.0371	1	0.0736	1	0.0025	1	0.0128
0.2	0.8659	0.0127	0.8565	0.0062	0.8771	0.0453	0.9792	0.0533	1	0.0011	1	0.0013
0.3	0.1504	0.0762	0.1238	0.0614	0.8186	0.0443	0.4780	0.0460	0.8302	0.0001	0.9274	0

Affine equivariance of the estimators is equivalent to say that the robust Mahalanobis distance is affine invariant:

$$RMD(\mathbf{Ax_i}, \mathbf{m_A}) = (\mathbf{Ax_i} - \mathbf{m_A})^T \mathbf{S_A^{-1}} (\mathbf{Ax_i} - \mathbf{m_A}) = RMD(\mathbf{x_i}, \mathbf{m}) \,.$$

Maronna and Zamar (2002) and Sajesh and Srinivasan (2012) proposed to investigate the lack of equivariance with transformed data, by simulations. We study the same for our proposal. They propose to generate random matrices as $A = TD$, where T is a random orthogonal matrix and $D = diag(u_1, ..., u_p)$, where the u_j's are independent and uniformly distributed in (0, 1). Then, the proposed simulations consist on affinely transform each generated data matrix X in each repetition, by applying the random matrix of transformation A to X, in order to obtain X_A. The contamination scheme consist in data generated from a mixture of normals $(1-\alpha)N(\mathbf{0}, I)+\alpha N(\delta \mathbf{e}, \lambda I)$. The dimension $p = 5, 10, 30, 50$, with sample size $n = 100, 100, 500, 1000$ respectively, the contamination level $\alpha = 10, 20, 30\%$, the distance of the outliers $\delta = 5$ and 10, and the concentration of the contamination $\lambda = 0.1$ and 1. Table 5 shows the obtained results about the TPR and FPR.

As it can be observed, even under affine transformations, *RMD-S* is able to detect all the outliers, except for a few cases (in bold type) that corresponds to large contamination level (30%) in case of outliers close to the center of the distribution. However, it can be noted that these cases improve in performance when dimension increases.

4.2 Breakdown value

For an estimator, the maximum proportion of outliers that it can safely tolerate is known as the breakdown value. Usually, the definition of finite sample breakdown value is used (Donoho and Huber 1983; Hubert and Debruyne 2009). Given any sample $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, where $\mathbf{x}_i$ is of dimension $1 \times p$, for all $i = 1, ..., n$, denote by $T(\mathbf{x})$ an estimate of a parameter. Let $\widetilde{\mathbf{x}}$ be the corrupted sample where any m of the original points of $\mathbf{x}$ are replaced by arbitrary outliers. Then the finite sample breakdown value γ^* is defined as:

$$\gamma^*(T, \mathbf{x}) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\widetilde{x}} ||T(\widetilde{\mathbf{x}}) - T(\mathbf{x})|| = \infty \right\} . \tag{28}$$

Table 5 True positive rates and false positive rates of RMD-S for transformed data

	p	α	$\delta = 5$		$\delta = 10$	
			TPR	FPR	TPR	FPR
$\lambda = 0.1$	5	0.1	1	0.0454	1	0.0455
		0.2	1	0.0155	1	0.0165
		0.3	**0.9709**	0.0034	1	0.0023
	10	0.1	1	0.0328	1	0.0252
		0.2	1	0.0088	1	0.0062
		0.3	**0.9844**	0.0023	1	0.0009
	30	0.1	1	0.0089	1	0.0074
		0.2	1	0.0006	1	0.0003
		0.3	1	0	1	0
	50	0.1	1	0.0008	1	0.0004
		0.2	1	0.0002	1	0.0001
		0.3	1	0	1	0
$\lambda = 1$	5	0.1	1	0.0451	1	0.0400
		0.2	1	0.0189	1	0.0120
		0.3	**0.9344**	0.0039	1	0.0046
	10	0.1	1	0.0282	1	0.0279
		0.2	1	0.0113	1	0.0071
		0.3	**0.9872**	0.0020	1	0.0010
	30	0.1	1	0.0093	1	0.0072
		0.2	1	0.0006	1	0.0004
		0.3	1	0	1	0
	50	0.1	1	0.0009	1	0.0002
		0.2	1	0.0002	1	0.0001
		0.3	1	0	1	0

where $||\cdot||$ is the Euclidean norm. The asymptotic breakdown value is understood as the limit of the finite sample breakdown value when n goes to infinity. Intuitively, the maximum possible asymptotic breakdown value is $1/2$ because if more than half of the observations are contaminated, it is not possible to distinguish between the background data and the contamination (Leroy and Rousseeuw 1987).

For an outlier detection method, the breakdown value can be defined as the maximum m^* outliers that the procedure can successfully detect, so that if the data is contaminated with m outliers and $m > m^*$ the method will fail to identify most of the true outliers and it will falsely detect many inliers, reducing drastically the true positive rate and inflating the false positive rate (Sajesh and Srinivasan 2012). Thus, it is necessary to use the true positive and the false positive rates for studying the breakdown value of the outlier detection procedure. Analogously as in Sajesh and Srinivasan (2012), n observations were generated from a p-dimensional $N(\mathbf{0}, I)$ and two forms of contamination are considered: α percent symmetric, for which the ith observation is multiplied by $100i$, and α percent asymmetric, for which the ith obser-

Table 6 Simulation results for breakdown value

$n = 1000$		Symmetric		Asymmetric	
p	α	TPR	FPR	TPR	FPR
10	0.1	1	0.0055	1	0.0047
	0.2	1	0.0001	1	0.0002
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0
30	0.1	1	0.0002	1	0.0002
	0.2	1	0	1	0
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0
50	0.1	1	0	1	0
	0.2	1	0	1	0
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0
80	0.1	1	0	1	0
	0.2	1	0	1	0
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0
100	0.1	1	0	1	0
	0.2	1	0	1	0
	0.3	1	0	1	0
	0.4	1	0	1	0
	0.45	1	0	1	0

vation is replaced by $(100i)\mathbf{e}$, $i = 1, ..., n\alpha$, where $\mathbf{e} = (1, ..., 1)$. In the first case the outliers are symmetrically distributed, and asymmetrically in the second case. The dimensions considered are $p = 10, 30, 50, 80, 100$ and the sample size $n = 1000$. The contamination level $\alpha = 10, 20, 30, 40, 45\%$. Table 6 gives the resulting TPR and FPR for both forms of contamination. It can be seen that the TPR is not affected and the FPR is zero for most cases or it is very reduced and near zero, then RMD-S can successfully detect the outliers even when there is large contamination and even in high dimension, without falsely detect many inliers.

4.3 Computational times

Table 7 show the resulting computational times in seconds for the Normal case when outliers are close to the center of the data and they are concentrated.

Table 7 Computational times with normal data, $\delta = 5$ and $\lambda = 0.1$

p	α	MCD	Adj MCD	Kurtosis	OGK	COM	RMD-S
5	0.1	1.0951	0.7670	0.1880	0.1087	0.0228	0.0096
	0.2	0.7619	0.7910	0.0499	0.0203	0.0088	0.0085
	0.3	0.7605	0.8304	0.0266	0.0196	0.0089	0.0074
	Mean	0.8725	0.7961	0.0882	0.0495	0.0135	0.0085
10	0.1	1.3184	0.9970	0.2191	0.1626	0.0247	0.0200
	0.2	1.0329	1.0477	0.1358	0.0793	0.0120	0.0118
	0.3	0.9685	1.0641	0.0482	0.0865	0.0128	0.0108
	Mean	1.1066	1.0363	0.1344	0.1095	0.0165	0.0142
30	0.1	6.2387	6.0934	0.7154	0.8969	0.2000	0.2206
	0.2	5.8676	6.3999	1.4635	0.8158	0.1687	0.1804
	0.3	5.9453	7.0405	1.6572	0.8407	0.1669	0.1674
	Mean	6.0172	6.5113	1.2787	0.8511	0.1785	0.1895
50	0.1	7.3521	7.2307	2.2497	1.2854	0.2174	0.2053
	0.2	7.2501	7.2337	2.2778	1.2678	0.2166	0.2018
	0.3	7.2479	7.2376	2.3753	1.2774	0.2169	0.2099
	Mean	7.2834	7.2340	2.3009	1.2769	0.2169	0.2057

The other tables can be founded in the Supplementary Material. The experiment is carried out on a PC with a 3.40 GHz Intel Core i7 processor with 32GB RAM. On average, the fastest methods are *COM* and *RMD-S* with very similar computational times. Compared to the *MCD* and its adjusted version *Adj MCD*, the latters are much more slower than our proposal. Depending on the dimension of the data, *MCD* and *Adj MCD* are between 31-93 and 34-102 times slower than *RMD-S*, respectively. *Kurtosis* and *OGK* are not as slower as *MCD* and *Adj MCD*, but they show worse computational times than *COM* and *RMD-S*. *Kurtosis* and *OGK* are between 6-11 and 4-8 times slower than our proposal. Thus, *RMD-S* shows competitive computational times as well as *COM*.

5 Real dataset

The proposed *RMD* is applied to a real dataset to evaluate its performance. The following dataset was taken from the *UCI Knowledge Discovery in Databases Archive* (Bay 1999). Specifically, we have chosen the *Breast Cancer Wisconsin (Diagnostic) Data Set* (WDBC). Features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe 30 characteristics of the cell nuclei present in the image, for 569 samples, from which 357 are benign and 212 malign. We propose to study only the 357 benign data. In Maronna and Zamar (2002) the authors analyzed several datasets but they only show the results of four of them. Specifically, in Section 4.5, page 314 the authors mention the data we used and they specify that the dimension was $p = 30$ and the sample size $n = 357$, which means that they selected only the 357

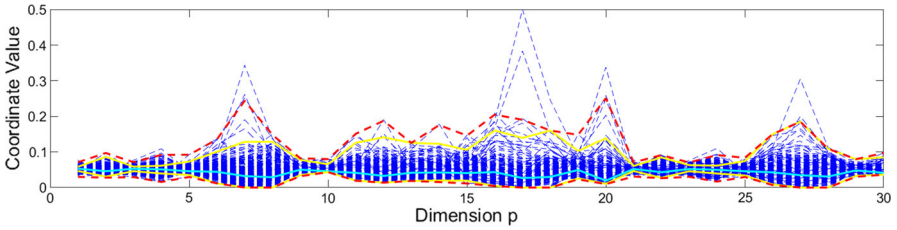

Fig. 1 Standardized data with the "multivariate boxplot"

observations corresponding to benign data. This is the same that they do with the study of Ionospheric data (Section 4.3, page 312), since the classification of the observations is previously known and it makes sense to study only one of the two groups because they come from a different distribution and the observations from the "bad" group are almost half of the entire dataset. The data is available at UCI repository. The archive has 32 columns but the first two are (1) the ID number and (2) Diagnosis (good or bad), which leaves us with 30 features. Therefore, this example has dimension $p = 30$ and sample size $n = 357$. We applied each method for detecting outliers and we retained the results, along with the computational times.

In order to interpret the outcome, we show the standardized data (after the detection) only for better visualization aim. We have also plotted the multivariate L_1 median and a kind of "multivariate boxplot", which is based on the idea from Sun and Genton (2011) method, but for finite dimensional. What the "box" would be is constructed sorting the data according to their L_1 depth value. The corresponding Q_1 and Q_3 "quartiles" delimiting the "box" are in fact the minimum and maximum values for each coordinate taking only into account the 50% of the most central data. Thus, the "fences" can be constructed with the same approach $F_1 = Q_1 - 1.5RI$ and $F_2 = Q_3 + 1.5RI$, where the "interquartile range" is $RI = Q_3 - Q_1$. Then, we can look for each method's result how many detected outliers are inside the "fences" for all their coordinates, and how many are outside the "fences". Figure 1 shows the data in blue color plotted in parallel coordinates (Inselberg and Dimsdale 1990; Wegman 1990; Inselberg 2009), the "box" delimiting the 50% of most central data in yellow color, the "fences" in red and the multivariate L_1-median in cyan.

Table 8a shows the detected outliers by each method. Outside the "fences" there are 3 or 4 for all the methods. Also, the method *Kurtosis* detected 162 outliers out of the 357 data. More or less like *OGK*, which detected 148. Furthermore, our method *RMD-S* is the one that label less amount of data as outliers. Table 8b shows how many outliers belong to the 50% of the most central data, according to the L_1-median.

We can investigate the shape of the detected outliers that are inside the "multivariate box", in order to see if they are similar or near to the median, or if they have a distinct shape. The motivation is that in case of real data we do not know the true outliers, thus we propose to study the shape of these observations in parallel coordinates (similar as in Maronna and Zamar (2002) with Ionospheric data, where they studied the shape of each observation's sequence of coordinates). Then, since the methods detected a large amount of observations as outliers, the multivariate boxplot is used to study the shape of the ones that are closest to the multivariate median, i.e. the ones belonging to the

Table 8 Detected outliers

Method	Inside	Outside	Total
(a) Inside and outside the fences			
MCD	72	4	76
Adj MCD	64	4	68
Kurtosis	158	4	162
OGK	144	4	148
COM	59	4	63
RMD-S	25	3	28
Method	Inside	Total	
(b) Inside the "box" with the 50% of the most central data			
MCD	29	76	
Adj MCD	27	68	
Kurtosis	65	162	
OGK	58	148	
COM	20	63	
RMD-S	7	28	

"box" of the multivariate boxplot. Figure 2 shows the shape of some of the outliers detected by the alternative methods that belong to the 50% of the most central data. In cyan color is the multivariate median, in yellow color the "box" and in blue color the detected outlier. The title of each subplot represent the index of the observation. The three subplots from the first column correspond to observations 236 (detected by MCD, AdjMCD, KUR and OGK), 155 (detected by KUR) and 212 (detected by KUR and OGK). The next three from the second column of subplots correspond to observations 254 (detected by MCD, AdjMCD, KUR and OGK), 182 (detected by KUR and OGK) and 234 (detected by MCD, AdjMCD, KUR, OGK and COM). The observation's sequence of coordinates can be considered similar to the multivariate median.

The general outcomes are that Adj MCD detected the same outliers as MCD except for the observations 266 and 332 which shape can be considered near the median. This makes sense since with the adjusted quantile the false positive rate decreases. Kurtosis and OGK detected a lot of observations as outliers and some of the ones inside the "box" are very similar to the multivariate median in parallel coordinates. Comedian method's detected outliers also have some observations similar to the median. In summary, for all of the alternative methods there seems to be some outliers having a shape very alike to the multivariate median or close to it for all the values of its components, leading us to think that maybe the alternative methods are detecting too much observations as outliers, in other words, the false positive rate is inflated. However, in Fig. 3, we can see that all outliers detected by *RMD-S*, belonging to the "box", are quite different than the multivariate median, in fact, they might be "shape outliers". For a final argument, we can say that all the outliers inside the "box", detected by method *RMD-S*, are actually

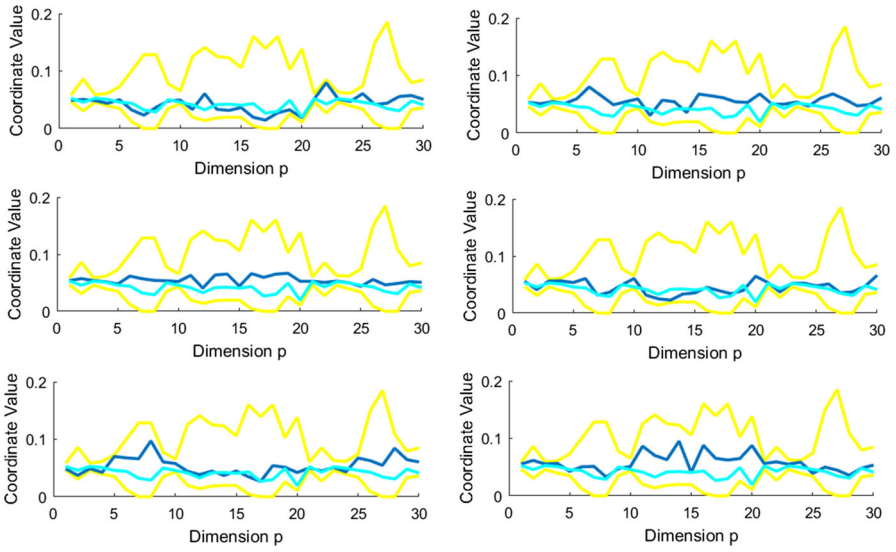

Fig. 2 Some of the alternative methods detected outliers belonging to the 50% of the most central data

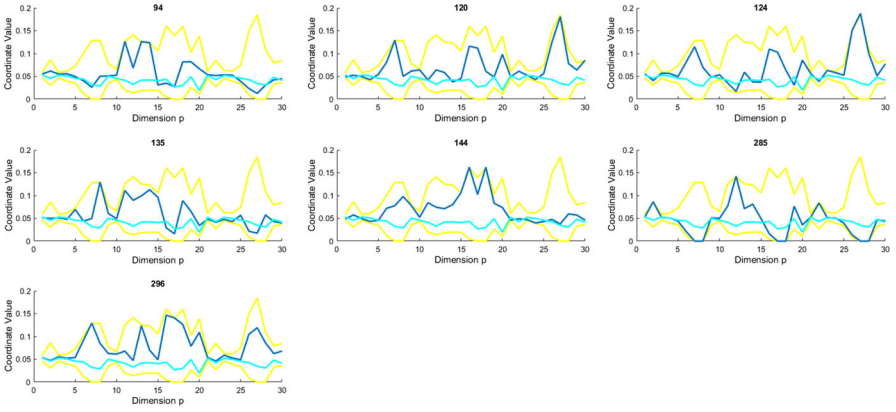

Fig. 3 RMD-S detected outliers that belong to the 50% of the most central data

detected by the alternative methods, so this also makes us think that our proposed method detects just enough.

Table 9 shows the computational times for each method in the task of detecting outliers with this example of real dataset. The results demonstrate that the alternative methods are much more slower than our proposal, except for *COM* which has a similar computational time.

Table 9 Computational times for each methods with the WDBC dataset

Method	MCD	Adj MCD	Kurtosis	OGK	COM	RMD-S
Times (s)	12.155	12.3378	6.3077	3.5325	0.3534	0.3299

6 Conclusions

Correct detection of outliers in the multivariate case is well-known to be a very important task for thorough data analysis. In order to reach that goal properly, it is necessary to consider the shape of the data and its structure in the multivariate space. That is the reason why the Mahalanobis distance approach is frequently used for the task of identifying the outliers. Different robust Mahalanobis distances can be defined according to the selected robust location and dispersion estimators. There are various robust estimators in the literature that have been considered in this paper. A collection of different combinations of robust location and covariance matrix estimators based on the notion of shrinkage is proposed, in order to define with each combination a robust Mahalanobis distance for the outlier detection problem. The performance of the proposed *RMD's* and the others from the literature is shown through a simulation study. It can be concluded that the alternative methods increase their FPR and decrease the TPR in the presence of contamination, especially in high dimension. The proposed *RMD's* have the ability to discover outliers with high TPR and low FPR in the vast majority of cases in the simulations, with Gaussian data and with skewed or heavy-tailed distributions. *RMD-S* is the most competitive version, as the simulation results showed. That is the reason why it is selected and some properties are investigated. The behavior under correlated and transformed data shows that *RMD-S* is approximately affine equivariant. With highly contaminated data it is shown that the approach has high breakdown value even in high dimension. There is also evidence of its inexpensive computational time. A real dataset example is also studied, in which the results bear out with the latter conclusions.

The results presented in this article emphasize the advantages of using shrinkage estimators for the location and dispersion in the definition of a robust Mahalanobis distance. It remains to be examined whether the proposal could be improved by adapting the adjusted quantile to the proposed robust distances. It could also be an interesting matter to study, whether the use of the different definitions of "depth" in the literature (Tukey 1975; Liu et al. 1990; Serfling 2002; Chen et al. 2009; Agostinelli and Romanazzi 2011; Paindaveine and Van Bever 2013), could improve the performance of the approach, as it is known that depth is a robust measure for location. As the referee suggested, a robustified likelihood could be interesting and we will consider it for future work.

Acknowledgements The authors are grateful to the editor and the referees for the constructive and valuable comments. This research was partially supported by Ministerio de Economía, Industria y Competitividad, España, award number: ECO2015-66593-P.

References

Agostinelli C, Romanazzi M (2011) Local depth. J Stat Plan Inference 141(2):817–830

Alqallaf F, Van Aelst S, Yohai VJ, Zamar RH (2009) Propagation of outliers in multivariate data. Ann Stat 37(1):311–331

Bay SD (1999) The UCI KDD archive [http://kdd.ics.uci.edu]. University of California, Irvine. Department of Information and Computer Science, vol 404, p 405

Becker C, Gather U (1999) The masking breakdown point of multivariate outlier identification rules. J Am Stat Assoc 94(447):947–955

Becker C, Fried R, Kuhnt S (2014) Robustness and complex data structures: festschrift in honour of Ursula Gather. Springer, New York

Bose A (1995) Estimating the asymptotic dispersion of the l1 median. Ann Inst Stat Math 47(2):267–271

Bose A, Chaudhuri P (1993) On the dispersion of multivariate median. Ann Inst Stat Math 45(3):541–550

Brettschneider J, Collin F, Bolstad BM, Speed TP (2008) Quality assessment for short oligonucleotide microarray data. Technometrics 50(3):241–264

Brown B (1983) Statistical uses of the spatial median. J R Stat Soc Ser B (Methodol) 45:25–30

Cerioli A, Riani M, Atkinson AC, Perrotta D, Torti F (2008) Fitting mixtures of regression lines with the forward search. Min Massive Data Sets Secur 19:271

Cerioli A, Riani M, Atkinson AC (2009) Controlling the size of multivariate outlier tests with the mcd estimator of scatter. Stat Comput 19(3):341–353

Chen SX, Qin Y-L et al (2010) A two-sample test for high-dimensional data with applications to gene-set testing. Ann Stat 38(2):808–835

Chen Y, Dang X, Peng H, Bart HL (2009) Outlier detection with the kernelized spatial depth function. IEEE Trans Pattern Anal Mach Intell 31(2):288–305

Chen Y, Wiesel A, Hero AO (2011) Robust shrinkage estimation of high-dimensional covariance matrices. IEEE Trans Signal Process 59(9):4097–4107

Choi HC, Edwards HP, Sweatman CH, Obolonkin V (2016) Multivariate outlier detection of dairy herd testing data. ANZIAM J 57:38–53

Chu JT (1955) On the distribution of the sample median. Ann Math Stat 26:112–116

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(8):3018–3034

Devlin SJ, Gnanadesikan R, Kettenring JR (1981) Robust estimation of dispersion matrices and principal components. J Am Stat Assoc 76(374):354–362

Dodge Y (1987) An introduction to l1-norm based statistical data analysis. Comput Stat Data Anal 5(4):239–253

Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL Jr (eds) A festschrift for Erich L. Lehmann. Wadsworth, Belmont, pp 157–184

Falk M (1997) On mad and comedians. Ann Inst Stat Math 49(4):615–644

Filzmoser P, Garrett RG, Reimann C (2005) Multivariate outlier detection in exploration geochemistry. Comput Geosci 31(5):579–587

Gao X (2016) A flexible shrinkage operator for fussy grouped variable selection. Statistical Papers, pp 1–24

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

Goutte C, Gaussier E (2005) A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: Proceedings of the European Conference on Information Retrieval, pp 345–359. Springer

Gower J (1974) Algorithm as 78: the mediancentre. J R Stat Soc Ser C (Appl Stat) 23(3):466–470

Hall P, Welsh A (1985) Limit theorems for the median deviation. Ann Inst Stat Math 37(1):27–36

Hardin J, Rocke DM (2005) The distribution of robust distances. J Comput Graph Stat 14(4):928–946

Hubert M, Debruyne M (2009) Breakdown value. Wiley Interdiscip Rev Comput Stat 1(3):296–302

Hubert M, Debruyne M (2010) Minimum Covariance Determinant. Wiley Interdiscip Rev Comput Stat 2(1):36–43

Hubert M, Rousseeuw PJ, Van Aelst S (2008) High-breakdown robust multivariate methods. Stat Sci 23:92–119

Inselberg A (2009) Parallel coordinates. Springer, New York

Inselberg A, Dimsdale B (1990) Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: Proceedings of the 1st conference on Visualization'90, pp 361–378. IEEE Computer Society Press
James W, Stein C (1961) Estimation with quadratic loss. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, vol 1, pp 361–379
Lazar N (2008) The statistical analysis of functional MRI data. Springer, New York
Ledoit O, Wolf M (2003a) Honey, i shrunk the sample covariance matrix. UPF economics and business working paper (691)
Ledoit O, Wolf M (2003b) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empir Finance 10(5):603–621
Ledoit O, Wolf M (2004) A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal 88(2):365–411
Leroy AM, Rousseeuw PJ(1987) Robust regression and outlier detection
Lindquist MA (2008) The statistical analysis of FMRI data. Stat Sci 23:439–464
Liu RY et al (1990) On a notion of data depth based on random simplices. Ann Stat 18(1):405–414
Lopuhaa HP, Rousseeuw PJ (1991) Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Ann Stat 19:229–248
Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci (Calcutta) 2:49–55
Marcano L, Fermín W (2013) Comparación de métodos de detección de datos anómalos multivariantes mediante un estudio de simulación. SABER. Revista Multidisciplinaria del Consejo de Investigación de la Universidad de Oriente 25(2):192–201
Maronna RA, Yohai VJ (1976) Robust estimation of multivariate location and scatter. Statistics Reference Online, Wiley StatsRef
Maronna RA, Zamar RH (2002) Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44(4):307–317
Monti MM (2011) Statistical analysis of fmri time-series: a critical review of the glm approach. Front Hum Neurosci 5:28
Möttönen J, Nordhausen K, Oja H et al (2010) Asymptotic theory of the spatial median. In: Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková, pp 182–193. Institute of Mathematical Statistics
Oja H (2010) Multivariate nonparametric methods with R: an approach based on spatial signs and ranks. Springer, New York
Paindaveine D, Van Bever G (2013) From depth to local depth: a focus on centrality. J Am Stat Assoc 108(503):1105–1119
Peña D, Prieto FJ (2001) Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43(3):286–310
Peña D, Prieto FJ (2007) Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. J Comput Graph Stat 16(1):228–254
Perrotta D, Torti F (2010) Detecting price outliers in european trade data with the forward search. In: Data Analysis and Classification, pp 415–423. Springer
Poline J-B, Brett M (2012) The general linear model and fmri: does love last forever? Neuroimage 62(2):871–880
Powers DM (2011) Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation
Reimann C, Filzmoser P (2000) Normal and lognormal data distribution in geochemistry: death of a myth. consequences for the statistical treatment of geochemical and environmental data. Environ Geol 39(9):1001–1014
Rousseeuw PJ (1985) Multivariate estimation with high breakdown point. Math Stat Appl 8:283–297
Rousseeuw PJ, Driessen KV (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223
Rousseeuw PJ, Van Zomeren BC (1990) Unmasking multivariate outliers and leverage points. J Am Stat Assoc 85(411):633–639
Sajesh T, Srinivasan M (2012) Outlier detection for high dimensional data using the comedian approach. J Stat Comput Simul 82(5):745–757
Serfling R (2002) A depth function and a scale curve based on spatial quantiles. In: Statistical data analysis based on the L1-norm and related methods, pp 25–38. Springer, New York
Small CG (1990) A survey of multidimensional medians. Int Stat Rev 58:263–277

Sokolova M, Japkowicz N, Szpakowicz S (2006) Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: Australasian Joint Conference on Artificial Intelligence, pp 1015–1021. Springer, New York

Steland A (2018) Shrinkage for covariance estimation: asymptotics, confidence intervals, bounds and applications in sensor monitoring and finance. Statistical Papers, pp 1–22

Sun R, Ma T, Liu S (2018) Portfolio selection: shrinking the time-varying inverse conditional covariance matrix. Statistical Papers, pp 1–22

Sun Y, Genton MG (2011) Functional boxplots. J Comput Graph Stat 20(2):316–334

Tarr G, Müller S, Weber NC (2016) Robust estimation of precision matrices under cellwise contamination. Comput Stat Data Anal 93:404–420

Templ M, Filzmoser P, Reimann C (2008) Cluster analysis applied to regional geochemical data: problems and possibilities. Appl Geochem 23(8):2198–2213

Tukey JW (1975) Mathematics and the picturing of data. Proc Int Congr Math 2:523–531

Vardi Y, Zhang C-H (2000) The multivariate l1-median and associated data depth. Proc Natl Acad Sci USA 97(4):1423–1426

Vargas JA, Robust N (2003) estimation in multivariate control charts for individual observations. J Qual Technol 35(4):367–376

Verboven S, Hubert M (2005) Libra: a matlab library for robust analysis. Chemometr Intell Lab Syst 75(2):127–136

Wegman EJ (1990) Hyperdimensional data analysis using parallel coordinates. J Am Stat Assoc 85(411):664–675

Zeng Y, Wang G, Yang E, Ji G, Brinkmeyer-Langford CL, Cai JJ (2015) Aberrant gene expression in humans. PLoS Genet 11(1):e1004942