Journal of Multivariate Analysis 171 (2019) 94–111

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A generalized spatial sign covariance matrix

Jakob Raymaekers, Peter Rousseeuw *

Department of Mathematics, KU Leuven, Belgium

ARTICLE INFO

Article history:
Received 3 May 2018
Available online 24 November 2018

AMS 2010 subject classifications:
primary 62H12
secondary 62H86

Keywords:
Orthogonal equivariance
Outliers
Robust location and scatter

ABSTRACT

The well-known spatial sign covariance matrix (SSCM) carries out a radial transform which moves all data points to a sphere, followed by computing the classical covariance matrix of the transformed data. Its popularity stems from its robustness to outliers, fast computation, and applications to correlation and principal component analysis. In this paper we study more general radial functions. It is shown that the eigenvectors of the generalized SSCM are still consistent and the ranks of the eigenvalues are preserved. The influence function of the resulting scatter matrix is derived, and it is shown that its asymptotic breakdown value is as high as that of the original SSCM. A simulation study indicates that the best results are obtained when the inner half of the data points are not transformed and points lying far away are moved to the center.

1. Introduction

Robust estimation of the covariance (scatter) matrix is an important and challenging problem. Over the last decades, many robust estimators for the covariance matrix have been developed. Many of them possess the attractive property of affine equivariance, meaning that when the data are subjected to an affine transformation the estimator will transform accordingly.

However, all highly robust affine equivariant scatter estimators have a combinatorial time complexity. Other estimators possess the less restrictive property of orthogonal equivariance. This means that the estimators commute with orthogonal transformations, which are characterized by orthogonal matrices and include rotations and reflections.

The most well-known orthogonally equivariant scatter estimator is the spatial sign covariance matrix (SSCM) proposed independently in [20,31] and studied in more detail in [8,11,19], among others. The estimator computes the regular covariance matrix on the *spatial signs* of the data, which are the projections of the location-centered datapoints on the unit sphere. Somewhat surprisingly, this transformation yields a consistent estimator of the eigenvectors of the true covariance matrix [20] under relatively general conditions on the underlying distribution. Of course the eigenvalues are different from the eigenvalues of the true covariance matrix, but it was shown in [31] that the order of the eigenvalues is preserved. We build on this idea by illustrating that the SSCM is part of a larger class of orthogonally equivariant estimators, all of which estimate the eigenvectors of the true covariance matrix and preserve the order of the eigenvalues.

The SSCM is easy to compute, and has been used extensively in several applications. The most common use of the SSCM is probably in the context of (functional) *spherical PCA* as developed in [5,17,30,32]. Like classical PCA, spherical PCA aims to find a lower dimensional subspace that captures most of the variability in the data. After centering the data, spherical PCA projects the data onto the unit (hyper)sphere before searching for the directions of highest variability. This projection gives all data points the same weight in the estimation of the subspace, thereby limiting the influence of potential outliers. The

* Corresponding author.
E-mail address: peter@rousseeuw.net (P. Rousseeuw).

https://doi.org/10.1016/j.jmva.2018.11.010

directions ('loadings') of spherical PCA thus correspond to the eigenvectors of the SSCM scatter matrix. The corresponding scores are usually taken to be the inner products of the loading vectors with the original (centered) data points, not with the projections of the data points on the sphere. Some concrete applications of spherical PCA are about the shape of the cornea in ophthalmology as analyzed in [17], and for multichannel signal processing as illustrated in [31].

In addition to spherical PCA, there also has been a lot of recent research on the use of the SSCM for constructing robust correlation estimators [7,9,10]. The main focus of this work is on results including asymptotic properties, the eigenvalues, and the influence function which measures robustness. A third application of the SSCM is its use as an initial estimate for more involved robust scatter estimators [4,16]. The SSCM is particularly well-suited for this task as it is very fast and highly robust against outlying observations and therefore often yields a reliable starting value. Another application of the SSCM is to testing for sphericity [29], which uses the asymptotic properties of the SSCM in order to assess whether the underlying distribution of the data deviates substantially from a spherical distribution. Serneels et al. [28] use the spatial sign transform as an initial preprocessing step in order to obtain a robust version of partial least squares regression. Finally, Boente et al. [1] study SCCM as an operator for functional data analysis.

The next section introduces a generalization of the SSCM and studies its properties. Section 3 compares the performance of several members of this class in a small simulation study. Section 4 applies the method to a real data example, and Section 5 concludes. All proofs can be found in the Appendix.

2. Methodology

2.1. Definition

Definition 1. Let X be a p-variate random variable and μ a vector serving as its center. Define the *generalized spatial sign covariance matrix* (GSSCM) of X by

$$S_{g_X}(X) = \mathrm{E}_{F_X}\{g_X(X-\mu)g_X(X-\mu)^\top\}, \tag{1}$$

where the function g_X is of the form

$$g_X(t) = t\,\xi_X(\|t\|), \tag{2}$$

where we call $\xi_X : \mathbb{R}^+ \to \mathbb{R}^+$ the radial function and $\|\cdot\|$ is the Euclidean norm.

Note that the form of g_X in (2) precisely characterizes an orthogonally equivariant data transformation as shown in [13], p. 276. Also note that the regular covariance matrix corresponds to $\xi_X(r) = 1$, and that $\xi_X(r) = 1/r$ yields the SSCM.

For a finite data set $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ the GSSCM is given by

$$S_{g_{\boldsymbol{X}}}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} \xi^2_{\boldsymbol{X}}\{\|x_i - T(\boldsymbol{X})\|\}\,\{x_i - T(\boldsymbol{X})\}\{x_i - T(\boldsymbol{X})\}^\top, \tag{3}$$

where T is a location estimator. Note that the SSCM gives the x_i with $\|x_i - T(\boldsymbol{X})\| < 1$ a weight higher than 1, but in general this is not required. In fact, the other functions we will propose satisfy $\xi_{\boldsymbol{X}}(r) \leqslant 1$ for all r.

In the above definitions, we added the subscript X or $\boldsymbol{X}$ to the functions g and ξ to indicate that they can depend on X or $\boldsymbol{X}$. In what follows we will drop these subscripts to ease the notational burden. We will study the following functions ξ:

1. Winsorizing (Winsor):

$$\xi(r) = \begin{cases} 1 & \text{if } r \leqslant Q_2, \\ Q_2/r & \text{if } Q_2 < r. \end{cases} \tag{4}$$

2. Quadratic Winsor (Quad):

$$\xi(r) = \begin{cases} 1 & \text{if } r \leqslant Q_2, \\ Q_2^2/r^2 & \text{if } Q_2 < r. \end{cases} \tag{5}$$

3. Ball:

$$\xi(r) = \begin{cases} 1 & \text{if } r \leqslant Q_2, \\ 0 & \text{if } Q_2 < r. \end{cases} \tag{6}$$

4. Shell:

$$\xi(r) = \begin{cases} 0 & \text{if } r < Q_1, \\ 1 & \text{if } Q_1 \leqslant r \leqslant Q_3, \\ 0 & \text{if } Q_3 < r. \end{cases} \tag{7}$$

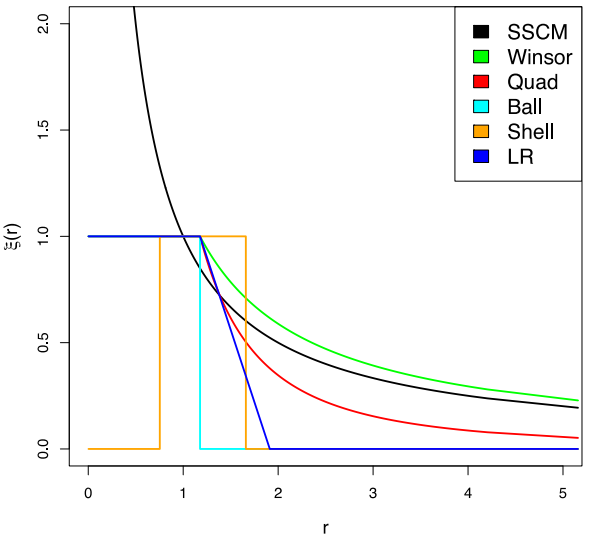

Fig. 1. Radial functions ξ in Eq. (2).

5. Linearly Redescending (LR):

$$\xi(r) = \begin{cases} 1 & \text{if } r \leqslant Q_2, \\ (Q_3^* - r)/(Q_3^* - Q_2) & \text{if } Q_2 < r \leqslant Q_3^*, \\ 0 & \text{if } Q_3^* < r. \end{cases} \tag{8}$$

The cutoffs Q_1, Q_2, Q_3 and Q_3^* depend on the Euclidean distances $\|x_i - T(\boldsymbol{X})\|$ by

$$\begin{aligned} Q_1 &= \left[\text{hmed}_i\{\|x_i - T(\boldsymbol{X})\|^{2/3}\} - \text{hmad}_i\{\|x_i - T(\boldsymbol{X})\|^{2/3}\}\right]^{3/2}, \\ Q_2 &= \left[\text{hmed}_i\{\|x_i - T(\boldsymbol{X})\|^{2/3}\}\right]^{3/2} = \text{hmed}_i\{\|x_i - T(\boldsymbol{X})\|\}, \\ Q_3 &= \left[\text{hmed}_i\{\|x_i - T(\boldsymbol{X})\|^{2/3}\} + \text{hmad}_i\{\|x_i - T(\boldsymbol{X})\|^{2/3}\}\right]^{3/2}, \\ Q_3^* &= \left[\text{hmed}_i\{\|x_i - T(\boldsymbol{X})\|^{2/3}\} + 1.4826 \times \text{hmad}_i\{\|x_i - T(\boldsymbol{X})\|^{2/3}\}\right]^{3/2}, \end{aligned}$$

where hmed and hmad are variations on the median and median absolute deviation given by the order statistic $\text{hmed}(y_1, \ldots, y_n) = y_{(h)}$ and $\text{hmad}(y_1, \ldots, y_n) = \text{hmed}_i|y_i - \text{hmed}_j(y_j)|$ where $h = \lfloor (n + p + 1)/2 \rfloor$. The $2/3$ power in these formulas is the Wilson–Hilferty transformation [33] to near normality. In Appendix A.1 it is verified that this transformation brings the above cutoffs close to the theoretical ones, which are quantiles of a convolution of Gamma random variables with different scale parameters.

Fig. 1 shows the above functions ξ and that of the SSCM for distances whose square follows the χ^2_2 distribution. The ξ of the SSCM is the only one which upweights observations close to the center. The Winsor ξ and its square have a similar shape, but the latter goes down faster. The Ball and Shell ξ functions are both designed to give a weight of 1 to half (in fact, h) of the data points and 0 to the remainder, to make them comparable. Ball does this by giving a weight of 1 to the h points with the smallest distances. Shell is inspired by the idea of Rocke to downweight observations both with very high and very low distances from the center [25]. The Linearly Redescending ξ is a compromise between the Ball and the Quad ξ functions.

2.2. Preservation of the eigenstructure

In what follows, we assume that the distribution F_X of X has an elliptical density with center zero and that its covariance matrix $\Sigma = \text{E}_{F_X}(XX^\top)$ exists. Therefore, X can be written as $X = UDZ$, where U is a $p \times p$ orthogonal matrix, D is a $p \times p$ diagonal matrix with strictly positive diagonal elements, and Z is a p-variate random variable which is spherically symmetric, i.e., its density is of the form $f_Z(z) \sim w(\|z\|)$, where w is a decreasing function. Assume without loss of generality that the covariance matrix of Z is I_p. The following proposition says that $S_g(X)$ has the same eigenvectors as Σ and preserves the ranks of the eigenvalues.

Proposition 1. *Let $X = UDZ$ be a p-variate random variable as described above, with $D = \text{diag}(\delta_1, \ldots, \delta_p)$ where $\delta_1 \geqslant \cdots \geqslant \delta_p > 0$. Assume that the covariance matrix $S_g = \text{E}_{F_X}\left\{g(X)g(X)^\top\right\}$ of $g(X)$ exists. Then Σ and S_g can be diagonalized as*

$$\Sigma = U\Lambda U^\top \text{ and } S_g = U\Lambda_g U^\top$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ *with* $\lambda_j = \delta_j^2$ *and* $\Lambda_g = \mathrm{diag}(\lambda_{g,1}, \ldots, \lambda_{g,p})$ *with* $\lambda_{g,1} \geqslant \cdots \geqslant \lambda_{g,p} > 0$ *and* $\lambda_j = \lambda_{j+1} \Leftrightarrow \lambda_{g,j} = \lambda_{g,j+1}$.

This proposition justifies the generalized SSCM approach.

2.3. Location estimator

So far we have not specified any location estimator T. For the SSCM the most often used location estimator is the *spatial median*; see, e.g., [2] and [12], which we denote by T_0. The spatial median of a dataset $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ is defined as

$$T_0(\boldsymbol{X}) = \arg\min_{\theta} \sum_{i=1}^{n} \|x_i - \theta\|.$$

In order to improve its robustness against a substantial fraction of outliers, we propose to use the *k-step least trimmed squares (LTS) estimator*. The LTS method was originally proposed in regression [26], and for multivariate location it becomes

$$T_{\mathrm{LTS}}(\boldsymbol{X}) = \arg\min_{\theta} \sum_{i=1}^{h} \|x_\bullet - \theta\|_{(i)}^2,$$

where the subscript (i) stands for the ith smallest squared distance. Without the square this becomes the least trimmed absolute distance estimator studied in [3]. For the multivariate location LTS the C-step of [27] simplifies to

Definition 2 (*C-step*)**.** Fix $h = \lfloor (n+1)/2 \rfloor$. Given a location estimate $T_{j-1}(\boldsymbol{X})$, we take the set $I_j = \{i_1, \ldots, i_h\} \subset \{1, \ldots, n\}$ such that $\{\|x_i - T_{j-1}(\boldsymbol{X})\| : i \in I_j\}$ are the h smallest distances in the set $\{\|x_i - T_{j-1}(\boldsymbol{X})\| : i = 1, \ldots, n\}$. The C-step then yields

$$T_j(\boldsymbol{X}) = \frac{1}{h} \sum_{i \in I_j} x_i.$$

The C-step is fast to compute, and guaranteed to lower the LTS objective. The k-step LTS is then the result of k successive C-steps starting from the spatial median $T_0(\boldsymbol{X})$.

It is also possible to avoid the estimation of location altogether, by calculating the GSSCM on the $O(n^2)$ pair-wise differences of the data points. This approach is called the "symmetrization" of an estimator, but is more computationally intensive. Visuri et al. [31] studied the symmetrized SSCM and called it Kendall's τ covariance matrix.

2.4. Robustness properties

A major reason for the SSCM's popularity is its robustness against outliers. Robustness can be quantified by the influence function and the breakdown value. We will study both for the GSSCM.

The influence function [13] quantifies the effect of a small amount of contamination on a statistical functional T. Consider the contaminated distribution $F_{\varepsilon,z} = (1-\varepsilon)F + \varepsilon \Delta(z)$, where $\Delta(z)$ is the distribution that puts all its mass in z. The influence function of T at F is then given by

$$\mathrm{IF}(z, T, F) = \lim_{\varepsilon \to 0} \frac{T(F_{\varepsilon,z}) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(F_{\varepsilon,z}) \bigg|_{\varepsilon=0}.$$

For the generalized SSCM class we obtain the following result:

Proposition 2. *Denote* $S_g(F) = \Xi_g$ *and let* $\mu = 0$ *in* (1)*. The influence function of* S_g *at the distribution* F *is given by*

$$IF(z, S_g, F) = \frac{\partial}{\partial \varepsilon} S_g(F_{\varepsilon,z}) \bigg|_{\varepsilon=0} = g(z) g(z)^\top - \Xi_g + \frac{\partial}{\partial \varepsilon} \int g_\varepsilon(X) g_\varepsilon(X)^\top dF(X) \bigg|_{\varepsilon=0}. \tag{9}$$

If g does not depend on F, the last term of (9) vanishes. For example, for $g(t) = t$, we retrieve the IF of the classical covariance matrix $\mathrm{IF}(z, \Sigma, F) = zz^\top - \Sigma$, and for $g(t) = t/\|t\|$ we obtain $\mathrm{IF}(z, \mathrm{SSCM}, F) = (z/\|z\|)\,(z/\|z\|)^\top - \mathrm{SSCM}(F)$ in line with the findings of [5]. For the GSSCM estimators defined by the functions (4)–(8) the last term of (9) remains, and the expressions of their IF can be found in Appendix A.3.

In order to visualize the influence function we consider the bivariate standard normal case, i.e., $F = \mathcal{N}(0, I_2)$. We put contamination at (z, z) or $(z, 0)$ for different values of z and plot the IF for the diagonal elements and the off-diagonal element. Note that we cannot compare the raw IFs directly as $S_g(F) = \Xi_g = c_g I$, where $c_g = \int g_1(X)^2 dF(X)$; hence Ξ_g is only equal to I_2 up to a factor. In order to make the estimators consistent for this distribution, we can divide them by c_g, and so we plot $\mathrm{IF}(z, S_g, F)/c_g$ in Fig. 2.

The rows in Fig. 2 correspond to the IF of the first diagonal element S_{11} (top), the off-diagonal element S_{12} (middle) and the element S_{22} (bottom). Let us first consider the left part of the figure, which contains the IFs for an outlier in (z, z). By symmetry, the IFs of the diagonal elements S_{11} and S_{22} are the same here. In the regions where the function ξ is 1 the IF is

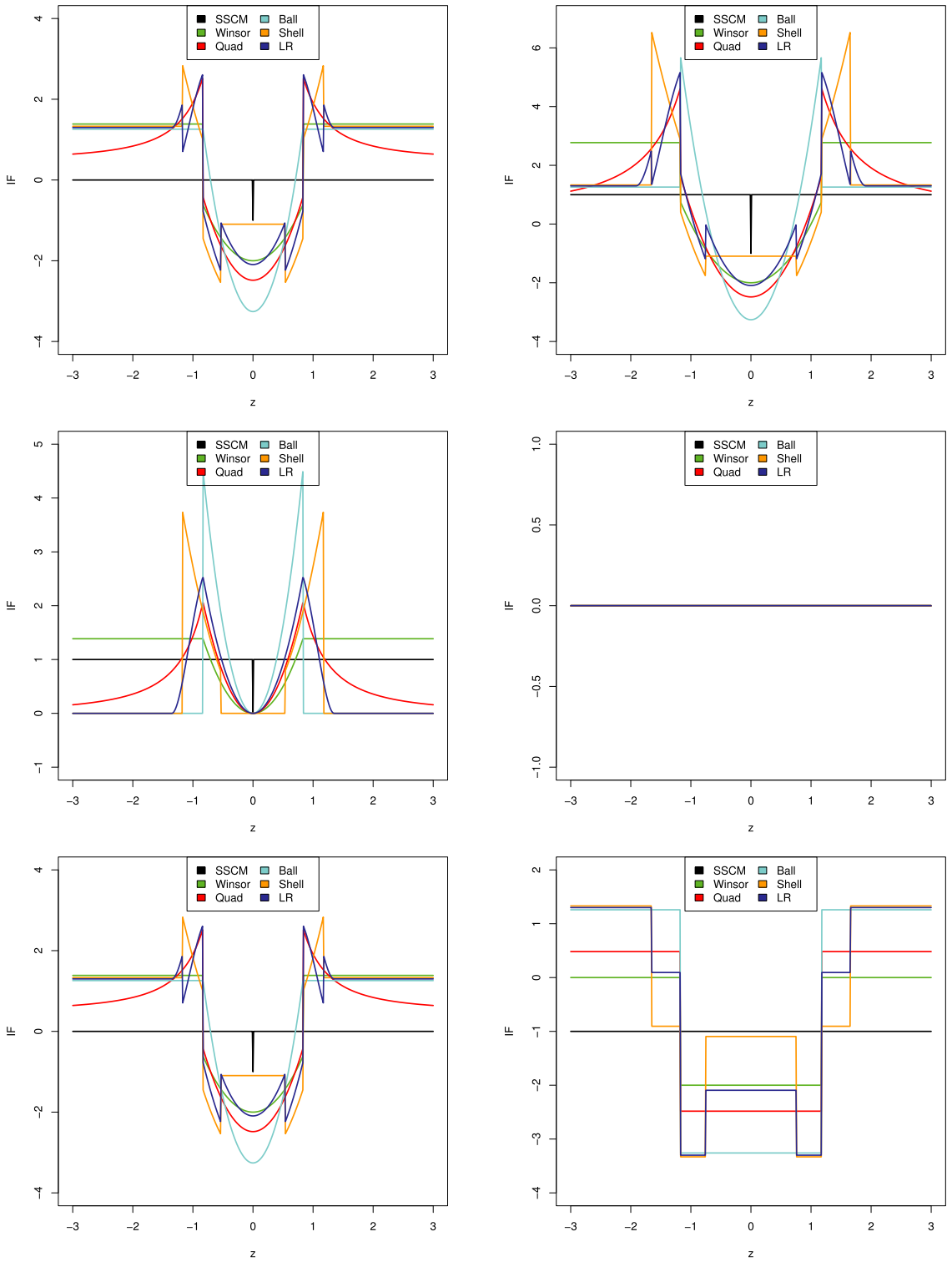

Fig. 2. Influence functions of the GSSCM at the bivariate standard normal distribution for contamination at (z, z) (left) and $(z, 0)$ (right). The rows correspond to the first diagonal element S_{11} (top), the off-diagonal element S_{12} (middle), and S_{22} (bottom).

quadratic, like that of the classical covariance. The diagonal elements of the IF of the SSCM are zero, except at $z = 0$ where it takes the value -1. The Quad IF is the only one which redescends as $|z|$ increases, whereas the others are also bounded but

stabilize at a value around 1.3. The shape of the IF of the Ball estimator resembles that of the univariate Huber M-estimator of scale.

For the IF of the off-diagonal element S_{12}, the picture is very different. All are redescending except for the SSCM and Winsor. Here it is Winsor whose IF resembles that of Huber's M-estimator of scale. Note that the IFs of the Ball and Shell estimators have large jumps at their cutoff values. The discontinuities in the IFs are due to the fact that the cutoffs depend on the median and the MAD of the distances $\|X\|^{2/3}$, as both the median and the MAD have jumps in their IF.

The right panel of Fig. 2 shows the influence functions for an outlier in $(z, 0)$. In this case the IFs of the diagonal elements S_{11} and S_{22} are no longer the same, as the symmetry is broken. The IFs of S_{11} are again quadratic where $\xi = 1$, with jumps at the cutoffs. Note that these cutoffs are now located at different values of z, as $\|(z, 0)\| \neq \|(z, z)\|$. The IF of the off-diagonal element is constant at 0, indicating that S_{12} remains zero even when there is an outlier at $(z, 0)$. Finally, for the second diagonal element S_{22} the IF of the SSCM is -1. This is because adding ε of contamination at $(z, 0)$ reduces the mass of the remaining part of F by ε which lowers the estimated scatter in the vertical direction. For the other estimators there is an additional effect of $(z, 0)$ on the cutoffs, which causes the discontinuities.

A second tool for quantifying the robustness of an estimator is the finite-sample breakdown value [6]. For a multivariate location estimator T and a dataset $\boldsymbol{X}$ of size n, the breakdown value is the smallest fraction of the data that needs to be replaced by contamination to make the resulting location estimate lie arbitrarily far away from the original location $T(\boldsymbol{X})$. More precisely,

$$\varepsilon^*(T, \boldsymbol{X}) = \min\left\{ m/n : \sup_{\boldsymbol{X}^*_m} \|T(\boldsymbol{X}^*_m) - T(\boldsymbol{X})\| = \infty \right\},$$

where $\boldsymbol{X}^*_m$ ranges over all datasets obtained by replacing any m points of $\boldsymbol{X}$ by arbitrary points.

For a multivariate estimator of scale S, the breakdown value is defined as the smallest fraction of contamination needed to make an eigenvalue of S either arbitrarily large or arbitrarily close to zero. We denote the eigenvalues of $S(\boldsymbol{X})$ by $\lambda_1\{S(\boldsymbol{X})\} \geqslant \cdots \geqslant \lambda_p\{S(\boldsymbol{X})\}$. The breakdown value of S is then given by

$$\varepsilon^*(S, \boldsymbol{X}) = \min\left\{ m/n : \sup_{\boldsymbol{X}^*_m} \max[\lambda_1\{S(\boldsymbol{X}^*_m)\}, \lambda_p^{-1}\{S(\boldsymbol{X}^*_m)\}] = \infty \right\}.$$

For the results on breakdown we assume the following conditions on the function ξ:

1. The function ξ takes values in $[0, 1]$.
2. For any dataset $\boldsymbol{X}$, one has $\#\{x_i : \xi\{\|x_i - T(\boldsymbol{X})\|\} = 1\} \geqslant \lfloor (n + p + 1)/2 \rfloor$.
3. For any vector t, one has $\|g(t)\| = \|t\|\, \xi(\|t\|) \leqslant \text{hmed}_i(d_i) + 1.4826 \times \text{hmad}_i(d_i)$.

Note that all functions ξ proposed in (4)–(8) satisfy these assumptions. The following proposition gives the breakdown value of the GSSCM scatter estimator S_g.

Proposition 3. *Let $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ be a p-dimensional dataset in general position, meaning that no $p + 1$ points lie on the same hyperplane. Also assume that the location estimator T has a breakdown value of at least $\lfloor (n - p + 1)/2 \rfloor / n$. Then $\varepsilon^*(S_g, \boldsymbol{X}) = \lfloor (n - p + 1)/2 \rfloor / n$.*

As we would like the GSSCM scatter estimator to attain this breakdown value, we have to use a location estimator whose breakdown value is at least $\lfloor (n - p + 1)/2 \rfloor / n$. The following proposition verifies that the k-step LTS estimator satisfies this, and even attains the best possible breakdown value for translation equivariant location estimators.

Proposition 4. *The k-step LTS estimator T_k satisfies $\varepsilon^*(T_k, \boldsymbol{X}) = \lfloor (n + 1)/2 \rfloor / n$ at any p-variate dataset $\boldsymbol{X} = \{x_1, \ldots, x_n\}$. When the C-steps are iterated until convergence ($k \to \infty$), the breakdown value remains the same.*

3. Simulation study

We now perform a simulation study comparing the GSSCM versions (4)–(8). As the estimators are orthogonally equivariant, it suffices to generate diagonal covariance matrices. We generate $m = 1000$ samples of size $n = 100$ from the multivariate Gaussian distribution of dimension $p = 10$ with center $\mu = \mathbf{0}$ and covariance matrices $\Sigma_1 = I_p$ ('constant eigenvalues'), $\Sigma_2 = \text{diag}(10, 9, \ldots, 1)$ ('linear eigenvalues'), and $\Sigma_3 = \text{diag}(10^2, 9^2, \ldots, 1)$ ('quadratic eigenvalues'). To assess robustness we also add 20% and 40% of contamination in the direction of the last eigenvector, at the point $(0, \ldots, 0, \gamma)$ for several values of γ. For the location estimator T in (3) we used the k-step LTS with $k = 5$.

For measuring how much the estimated $\widehat{\Sigma}$ deviates from the true Σ we use the Kullback–Leibler divergence (KLdiv) given by

$$\text{KLdiv}(\widehat{\Sigma}, \Sigma) = \text{trace}(\widehat{\Sigma}\Sigma^{-1}) - \ln\{\det(\widehat{\Sigma}\Sigma^{-1})\} - p.$$

We also consider the shape matrices $\widehat{\Gamma} = \{\det(\widehat{\Sigma})\}^{-1/p}\widehat{\Sigma}$ and $\Gamma = \{\det(\Sigma)\}^{-1/p}\Sigma$ which have determinant 1, and compute $\text{KLdivshape}(\widehat{\Sigma}, \Sigma) = \text{KLdiv}(\widehat{\Gamma}, \Gamma)$. Both the KLdiv and the KLdivshape are then averaged over the $m = 1000$ replications.

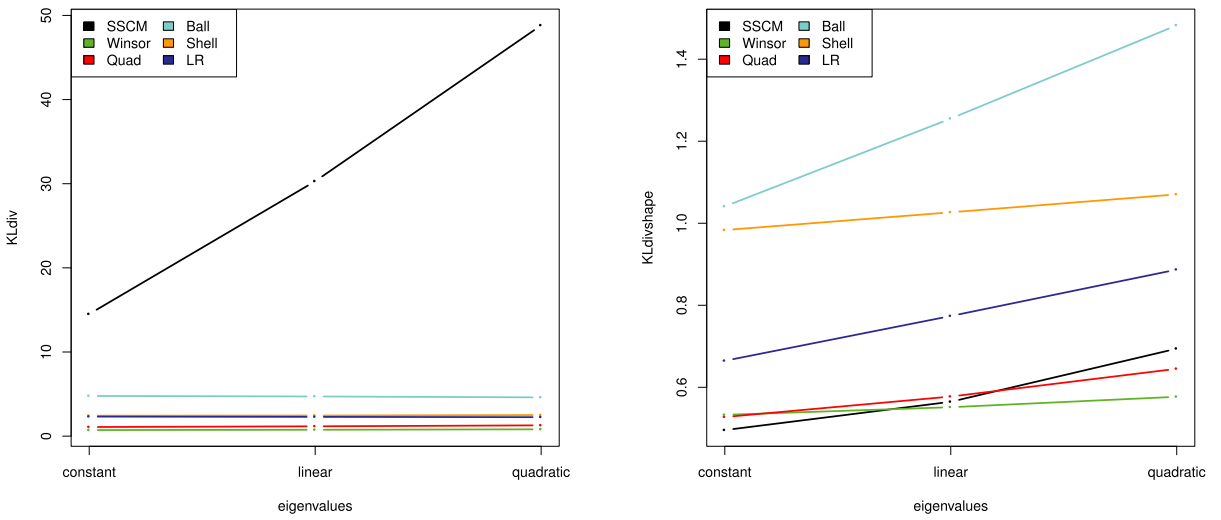

Fig. 3. Simulation results: KLdiv (left) and KLdivshape (right) for the uncontaminated normal distribution, with constant, linear and quadratic eigenvalues.

Fig. 3 shows the simulation results on the uncontaminated data. Looking at KLdiv (left panel), we note that the SSCM deviates the most from the true covariance matrix Σ. Among the other choices, Winsor and Quad have the lowest bias, followed by LR, Shell, and Ball. When looking only at the shape component (right panel), SSCM performs the best when the distribution is spherical (constant eigenvalues), in line with Remark 3.1 in [19]. However, it loses this dominant performance once the distribution deviates from sphericity. Among the other GSSCM methods Winsor performs the best, followed by its quadratic counterpart, LR, Shell, and finally Ball.

The result for the simulation with 20% of point contamination is presented in Fig. 4. All plots are as a function of γ, which indicates the position of the outliers. In the left panel (KLdiv), the SSCM has a large bias. The Winsor GSSCM, which did very well in the uncontaminated setting, now has a disappointing performance when the eigenstructure becomes more challenging with linear or quadratic eigenvalues. Quad performs a lot better, but also suffers under quadratic eigenvalues. LR and Shell perform the best here, followed by Ball. Their redescending nature helps them for far outliers. The conclusions for the shape component (right panel) are largely similar, except that Winsor and especially Ball look worse here.

The simulation results for 40% of contamination are shown in Fig. 5. The KLdiv plots on the left indicate that the SSCM performs poorly for constant and linear eigenvalues, and looks better for quadratic eigenvalues but not when γ is large (far outliers). Winsor performs badly for linear and quadratic eigenvalues, whereas Quad does much better. Ball looks okay except for relatively small γ. LR and Shell perform the best for both small and large γ, and are okay for intermediate γ. When estimating the shape component (right panels) SSCM and Winsor have the worst performance overall, whereas Ball also does poorly for small to intermediate γ. LR and Shell are the best picks here. Quad does almost as well, but redescends more slowly.

4. Application: Principal component analysis

We analyze a multivariate dataset from a study by Reaven and Miller [24]. The dataset contains five numerical variables for 109 subjects, consisting of 33 overt diabetes patients and 76 healthy people. The variables are body weight, fasting plasma glucose, area under the plasma glucose curve, area under the plasma insulin curve, and steady state plasma glucose response. These data were previously analyzed in [22] in the context of clustering using statistical data depth, and are available in the R package `ddalpha` [23] under the tag *chemdiab_2vs3*. Here we analyze the data by principal component analysis. We first standardize the data, as the variables have quite different scales. Denote the standardized observations by z_i for $i \in \{1, \ldots, 109\}$.

We consider the diabetes patients as outliers and would like the PCA subspace to model the variability within the healthy patients. For classical PCA, the PCA subspace corresponds to the linear span of the k eigenvectors (also called 'loadings') of the covariance matrix which correspond with the k largest eigenvalues. In similar fashion we can perform PCA based on the GSSCM with the LR radial function (8), by considering the linear span of its k first eigenvectors. We take $k = 3$ components, thereby explaining more than 95% of the variance.

Fig. 6 shows the scores with respect to the first 3 loadings for classical PCA and GSSCM PCA. The scores s_i are the projections of the observations z_i onto the PCA subspace, i.e., $s_{i,j} = z_i^\top v_j$ where v_j denotes the jth eigenvector. From these plots, it is clear that the first eigenvector of the classical PCA is heavily attracted by the diabetes patients. As a result, the outliers are only

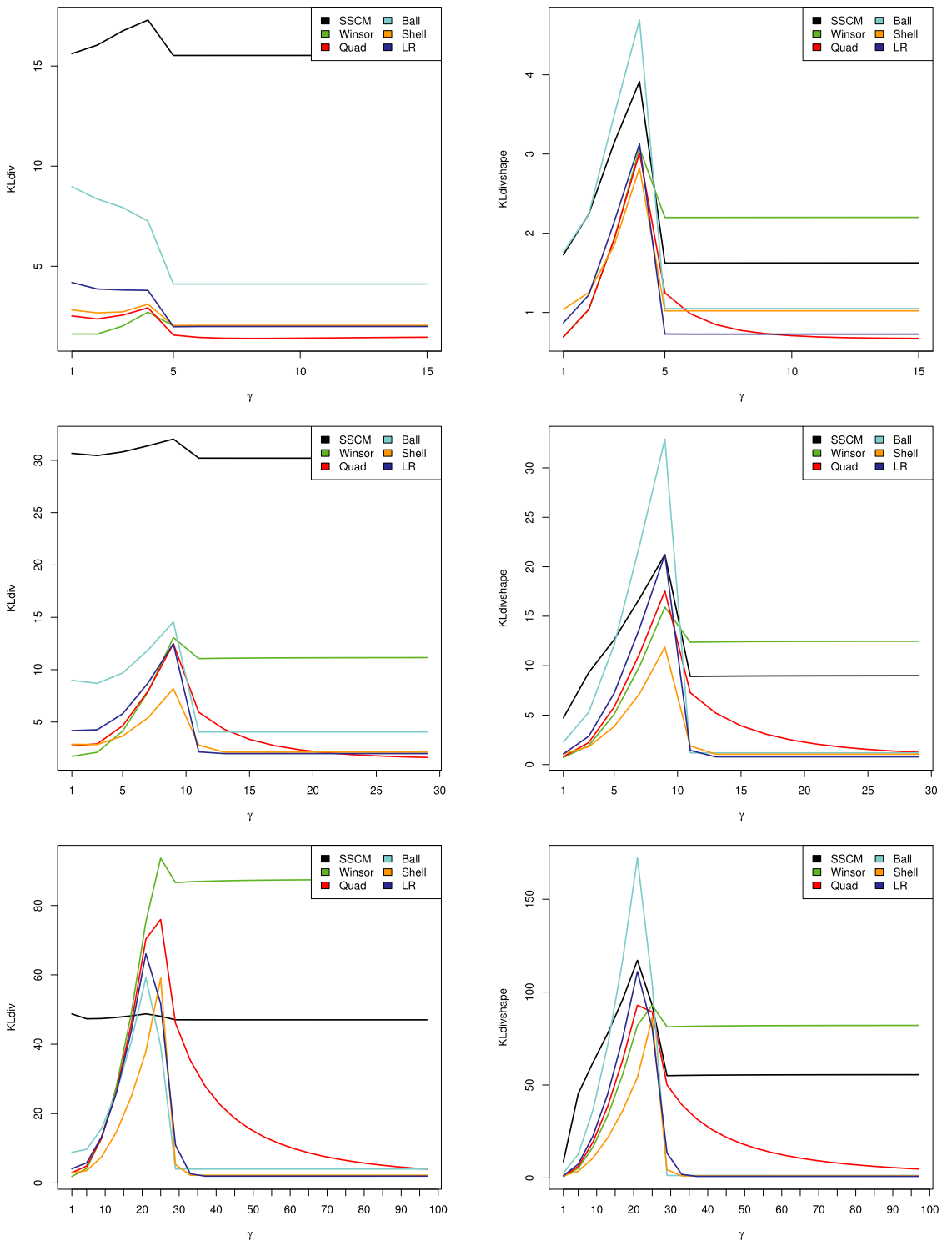

Fig. 4. Simulation results: KLdiv (left) and KLdivshape (right) for the normal distribution with constant (top), linear (middle) and quadratic (bottom) eigenvalues and 20% of contamination. The outliers were placed at the point $(0, \ldots, 0, \gamma)$.

distinguishable in their scores with respect to the first principal component. This is very different for the GSSCM PCA, where the principal components seem to fit the healthy patients better, resulting in outlying scores for the diabetes patients with respect to several principal components.

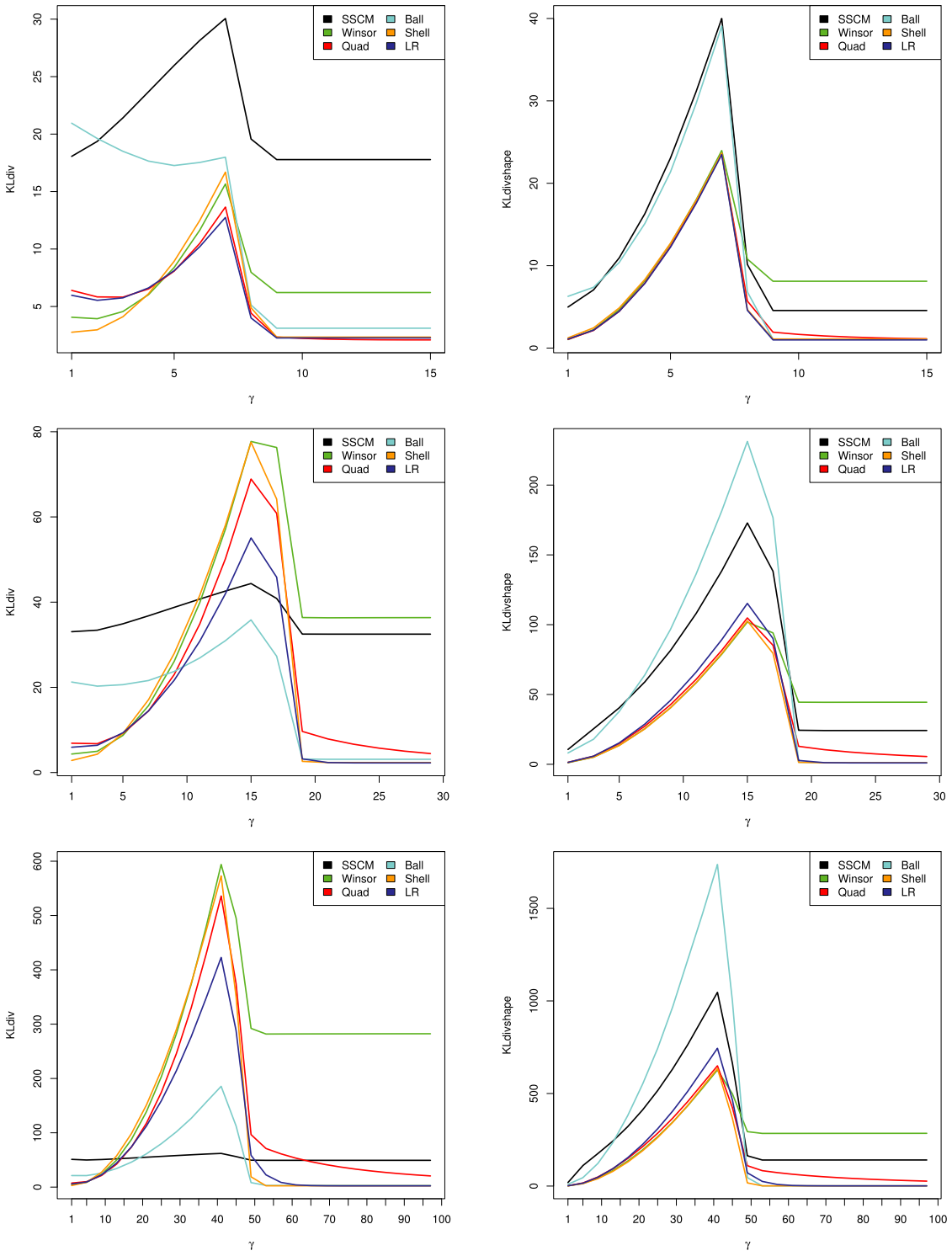

Fig. 5. Simulation results: KLdiv (left) and KLdivshape (right) for the normal distribution with constant (top), linear (middle) and quadratic (bottom) eigenvalues and 40% of point contamination.

In addition to the scores plots, the PCA outlier map of [15] can serve as a diagnostic tool for identifying outliers. It plots the orthogonal distance OD_i against the score distance SD_i for every observation z_i in the dataset. The score distance

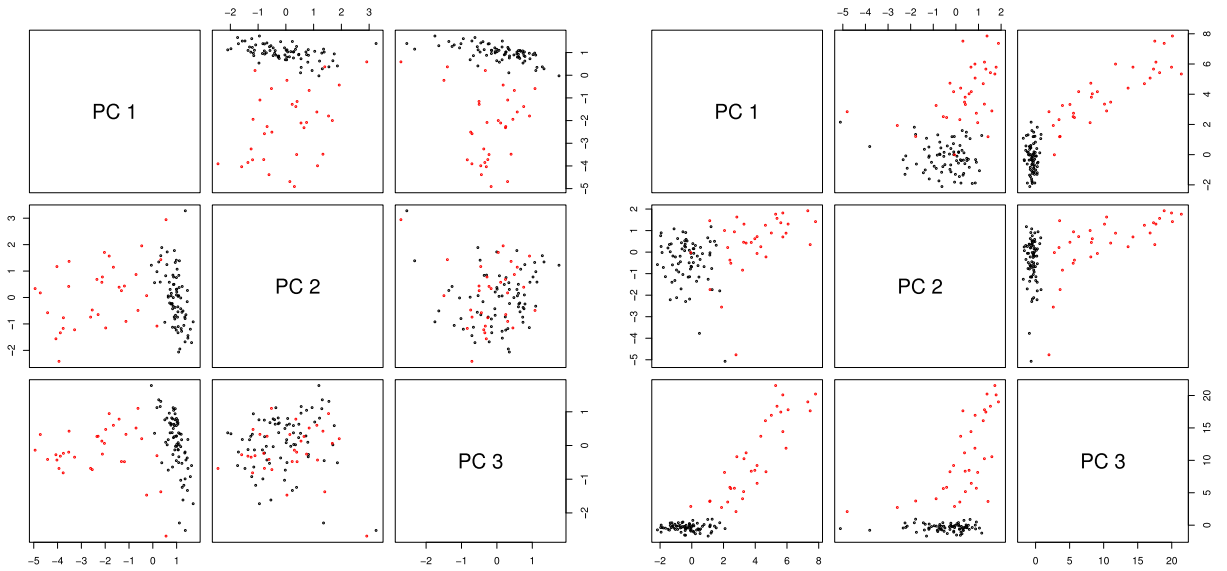

Fig. 6. Scores from the 3 first loading vectors of classical PCA (left) and GSSCM PCA (right).

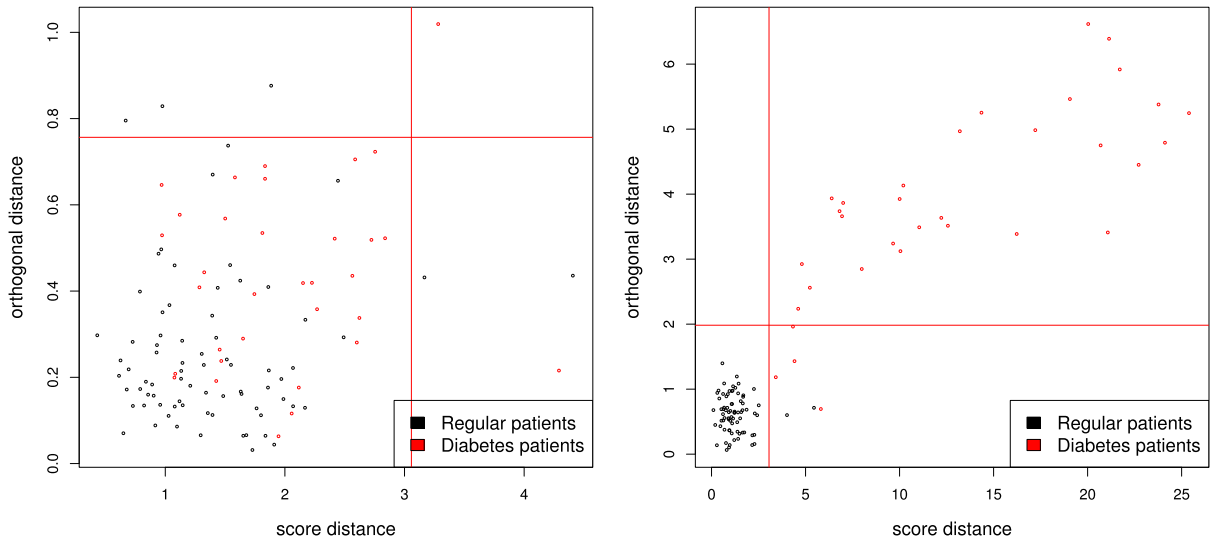

Fig. 7. Outlier maps based on classical PCA (left) and GSSCM PCA (right).

of observation i captures the distance between the observation and the center of the data within the PCA subspace. It is given by

$$\mathrm{SD}_i = \sqrt{\sum_{j=1}^{3} (s_{ij}/\hat{\sigma}_j)^2},$$

where $\hat{\sigma}_j$ denotes the scale of the jth scores. For classical PCA $\hat{\sigma}_j$ is their standard deviation, whereas for GSSCM PCA we take their median absolute deviation. The orthogonal distance to the PCA subspace is given by $\mathrm{OD}_i = \|z_i - Vs_i\|$, where V is the 5×3 matrix containing the three eigenvectors in its columns. Both the score distances and the orthogonal distances have cutoffs, described in [15]. Fig. 7 shows the outlier maps resulting from the classical PCA and the GSSCM PCA. Classical PCA clearly fails to distinguish the diabetes patients from the healthy subjects. In contrast, GSSCM PCA flags most of the diabetes patients as having both an abnormally high orthogonal distance to the PCA subspace as well as having a projection in the PCA subspace far away from those of the healthy subjects.

5. Conclusions

The spatial sign covariance matrix (SSCM) can be seen as a member of a larger class called Generalized SSCM (GSSCM) estimators in which other radial functions are allowed. It turns out that the GSSCM estimators are still consistent for the true eigenvectors while preserving the ranks of the eigenvalues. Their computation is as fast as the SSCM. We have studied five GSSCM methods with intuitively appealing radial functions, and shown that their asymptotic breakdown values are as high as that of the original SSCM. We also derived their influence functions and carried out a simulation study.

The radial function of the SSCM is $\xi(r) = 1/r$ which implies that points near the center are given a very high weight in the covariance computation. Our alternative radial functions give these points a weight of at most 1, which yields better performance at uncontaminated Gaussian data (Fig. 3) as well as contaminated data (Figs. 4 and 5). In particular, Winsor is the most similar to SSCM since its $\xi(r)$ is 1 for the central half of the data and $1/r$ for the outer half. It performs best for uncontaminated data, but still suffers when far outliers are present. It is almost uniformly outperformed by Quad, whose $\xi(r)$ is 1 in the central half and $1/r^2$ outside it. The influence of outliers on Quad smoothly redescends to zero. The other three estimators are hard redescenders whose $\xi(r) = 0$ for large enough r. Among them, the linear redescending (LR) radial function performed best overall.

A potential topic for further research is to investigate principal component analysis based on a GSSCM covariance matrix.

Software availability

R-code for computing these estimators and an example script are available from the website wis.kuleuven.be/stat/robust/software.

Acknowledgments

This research was supported by projects of Internal Funds KU Leuven, Belgium.

Appendix

A.1. Distribution of Euclidean distances

Exact distribution. The exact distribution of the squared Euclidean distances $\|X\|^2$ of a multivariate Gaussian distribution with general covariance matrix is given by the following result:

Proposition 5. *Let $X \sim \mathcal{N}(0, \Sigma)$, and suppose the eigenvalues of Σ are given by $\lambda_1, \ldots, \lambda_p$. Then*

$$\|X\|^2 \sim \sum_{i=1}^{p} \Gamma\,(1/2, 2\lambda_i).$$

For $p \to \infty$ we have $\|X\|^2 \rightsquigarrow \mathcal{N}(\sum_{i=1}^{\infty} \lambda_i, 2\sum_{i=1}^{\infty} \lambda_i^2)$.

Proof. We can write $X = UDZ$, where U is an orthogonal matrix, D is the diagonal matrix with elements $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}$, and Z follows the p-variate standard Gaussian distribution. Note that $\|X\|^2 = \|UDZ\|^2 = \|DZ\|^2 = \sum_{i=1}^{p} \lambda_i Z_i^2$, where $Z_i^2 \sim \chi^2(1)$. Therefore, $\lambda_i Z_i^2 \sim \Gamma(1/2, 2\lambda_i)$ so the distribution of $\|X\|^2$ is a sum of iid gamma distributions with a constant shape of $1/2$ and varying scale parameters equal to twice the eigenvalues of the covariance matrix. As $p \to \infty$, one has

$$\|X\|^2 \rightsquigarrow \mathcal{N}\left(\sum_{i=1}^{\infty} \lambda_i, 2\sum_{i=1}^{\infty} \lambda_i^2\right)$$

by the Lyapunov Central Limit Theorem. □

Approximate distribution of a sum of Gamma variables. Proposition 5 gives the exact distribution of the squared Euclidean distances $\|X\|^2$. The distribution of a sum of gamma distributions has been studied in [21]. Quantiles of this distribution can be computed by the R package `coga` [14] for **co**nvolutions of **ga**mma distributions. However, this computation requires the knowledge of the eigenvalues $\lambda_1, \ldots, \lambda_p$ that we are trying to estimate. Therefore we need a transformation of the Euclidean distances such that the transformed distances have an approximate distribution whose quantiles do not require knowing $\lambda_1, \ldots, \lambda_p$.

In the simplest case $\lambda_1 = \cdots = \lambda_p$ (constant eigenvalues), and then $\|X\|^2/\lambda_1$ follows a χ_p^2 distribution. It is known that when p increases the distribution of $\|X\|^2$ tends to a Gaussian distribution, but this also holds for some other powers of $\|X\|$. Wilson and Hilferty [33] found that the best transformation of this type was $\|X\|^{2/3}$ in the sense of coming closest to a

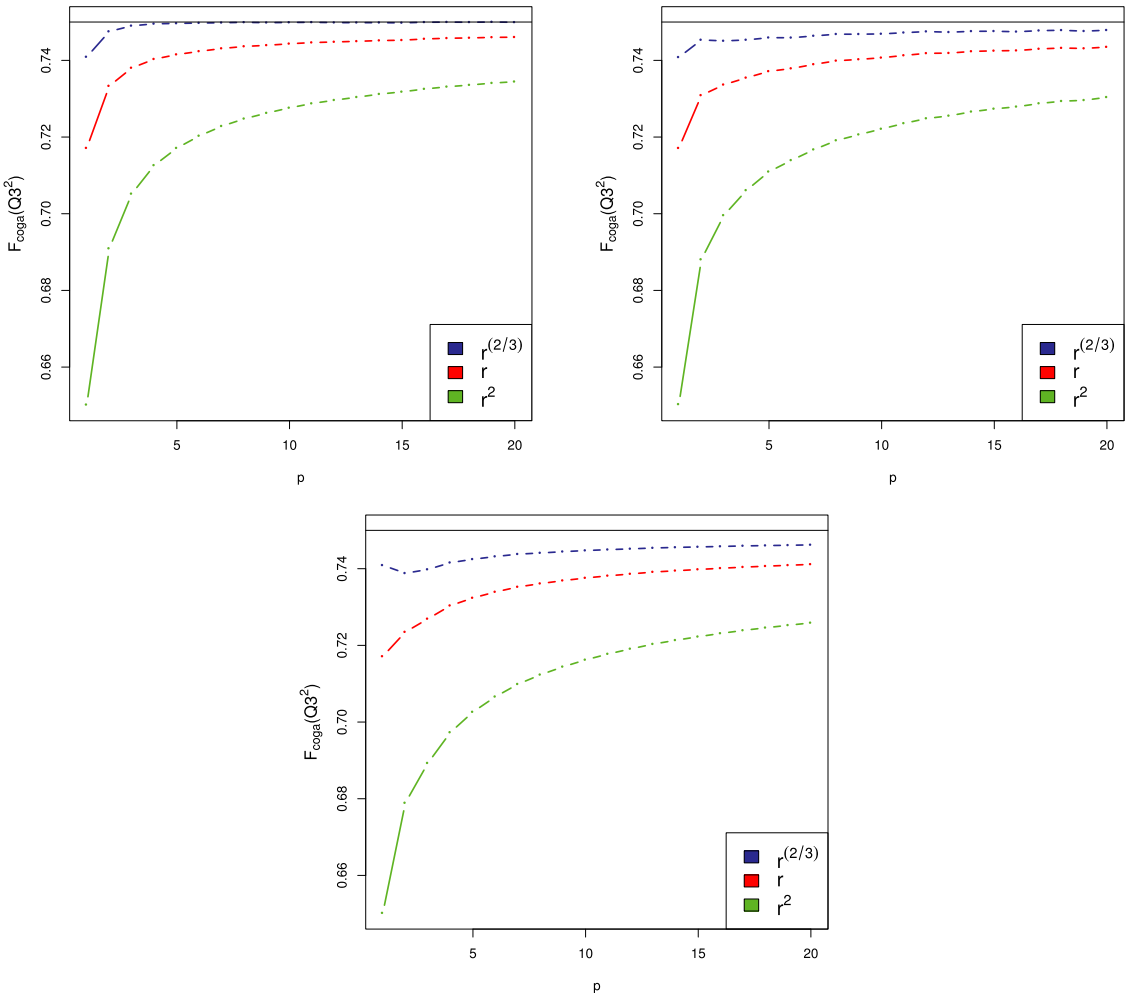

Fig. 8. Approximation of the third quartile of a coga distribution for dimensions $p \in \{1, \ldots, 20\}$ when the eigenvalues are constant (top left), linear (top right), or quadratic (bottom), using three different normalizing transforms.

Gaussian distribution. The quantiles q_α of a Gaussian distribution are easier to compute and can then be transformed back to $q_\alpha^{3/2}$.

It turns out that the same Wilson–Hilferty transformation also works quite well in the more general situation where the eigenvalues $\lambda_1, \ldots, \lambda_p$ need not be the same. We came to this conclusion by a simulation study, a part of which is illustrated here. The dimension p ranged from 1 to 20 by steps of 1. For each p, we generated $n = 10^6$ observations $y_1, \ldots, y_n$ from the coga distribution with shape parameters $(0.5, \ldots, 0.5)$. The scale parameters had three settings: constant $(2, \ldots, 2)$, linear $(p, p-1, \ldots, 1)$, and quadratic $(p^2, (p-1)^2, \ldots, 1)$, after which the scale parameters were further standardized in order to sum to $2p$. These correspond to the distribution of the squared Euclidean norms of a multivariate normal distribution where the covariance matrix has eigenvalues that are constant or proportional to $(p, p-1, \ldots, 1)$ (linear eigenvalues) or to $(p^2, (p-1)^2, \ldots, 1)$ (quadratic eigenvalues). Denote the unsquared Euclidean norms as $r_i = \sqrt{y_i}$. Then we estimate quantiles, e.g., Q_3 by assuming normality of the transformed values $h_1(r_i) = r_i^2$ (square), $h_2(r_i) = r_i$ (Fisher), and $h_3(r_i) = r_i^{2/3}$ (Wilson–Hilferty), by computing the third quartile of a Gaussian distribution with $\hat{\mu} = \text{median}_i\{h(r_i)\}$ and $\hat{\sigma} = \text{mad}_i\{h(r_i)\}$. Finally, we have evaluated the cumulative distribution function of the coga distribution in $\hat{Q}_3^2$. Ideally, we would like to obtain $F_{\text{coga}}(\hat{Q}_3^2) = 0.75$. The result of this experiment is shown in Fig. 8. We clearly see that the Wilson–Hilferty transform brings the approximate quantile closest to its target value. The results for the first quartile Q1 (not shown) are very similar.

A.2. Proof of Proposition 1

Part 1: Preservation of the eigenvectors. First note that g is orthogonally equivariant, i.e., $g(HX) = Hg(X)$ for any orthogonal matrix H. Therefore $S_g = \text{E}_{F_X}\{g(X)g(X)^\top\}$ implies $\text{E}_{F_X}\{g(HX)g(HX)^\top\} = HS_gH^\top$.

The distribution of Z is spherically symmetric hence invariant to reflections along a coordinate axis, which are described by diagonal matrices R with an entry of -1 and all other entries $+1$. For every reflection matrix R it thus holds that $\mathrm{E}_{F_Z}\{g(DZ)g(DZ)^\top\} = \mathrm{E}_{F_Z}\{g(DRZ)g(DRZ)^\top\} = \mathrm{E}_{F_Z}\{g(RDZ)g(RDZ)^\top\} = R\mathrm{E}\{g(DZ)g(DZ)^\top\}R^\top$, where the second equality holds because $DR = RD$ as both D and R are diagonal, and the last equality because R is orthogonal. Therefore $\mathrm{E}_{F_Z}\{g(DZ)g(DZ)^\top\}$ is a diagonal matrix, which we can denote as $\Lambda_g = \mathrm{diag}(\lambda_{g,1}, \ldots, \lambda_{g,p})$.

Now take U an arbitrary orthogonal matrix and let $X = UDZ$. Then

$$S_g = \mathrm{E}_{F_Z}\{g(UDZ)g(UDZ)^\top\} = U\mathrm{E}_{F_Z}\{g(DZ)g(DZ)^\top\}U^\top = U\Lambda_g U^\top.$$

For the plain covariance matrix Σ of X we have $\Sigma = \mathrm{E}_{F_Z}\{UDZ(UDZ)^\top\} = U\Lambda U^\top$, where $\Lambda = DD^\top = \mathrm{diag}(\delta_1^2, \ldots, \delta_p^2)$. Therefore, the same matrix U orthogonalizes both Σ and S_g, hence S_g and Σ have the same eigenvectors.

Part 2: Preservation of the ranks of the eigenvalues. Let $i > j$ and suppose that $\delta_i > \delta_j$. We will show that $\lambda_{g,i} > \lambda_{g,j}$. Note that

$$\lambda_{g,i} = \int g(DZ)_i^2 f_Z(Z)dZ = \int \delta_i^2 z_i^2 \xi(\|DZ\|)^2 f_Z(Z)dZ,$$

where f_Z is the density of Z. Similarly, we have

$$\lambda_{g,j} = \int g(DZ)_j^2 f_Z(Z)dZ = \int \delta_j^2 z_j^2 \xi(\|DZ\|)^2 f_Z(Z)dZ.$$

This means that $\lambda_{g,i} > \lambda_{g,j}$ is equivalent to

$$\int (\delta_i^2 z_i^2 - \delta_j^2 z_j^2)\xi(\|DZ\|)^2 f_Z(Z)dZ > 0. \tag{A.1}$$

As Z is spherically symmetric, i.e., $f_Z(Z) \sim w(\|Z\|)$, we can write (A.1) as

$$\int (\delta_i^2 z_i^2 - \delta_j^2 z_j^2)\xi(\|DZ\|)^2 w(\|Z\|)dZ > 0. \tag{A.2}$$

Note that we can change the variable of integration as follows. Let $y_k = \delta_k z_k$ and write $Y = (y_1, \ldots, y_p)$. Then (A.2) is equivalent to

$$\frac{1}{\delta_1 \cdots \delta_p}\left\{\int (y_i^2 - y_j^2)\xi(\|Y\|)^2 w\left(\sqrt{\sum_{k=1}^p \frac{y_k^2}{\delta_k^2}}\right)dY\right\} > 0. \tag{A.3}$$

We can ignore the positive constant $1/(\delta_1 \cdots \delta_p)$ and split the integral over the domains $A = \{x \in \mathbb{R}^d : |x_i| > |x_j|\}$ and $B = \{x \in \mathbb{R}^d : |x_i| < |x_j|\}$, yielding

$$\begin{aligned}
\int (y_i^2 - y_j^2)\xi(\|Y\|)^2 w\left(\sqrt{\sum_{k=1}^p \frac{y_k^2}{\delta_k^2}}\right)dY &= \int_A (y_i^2 - y_j^2)\xi(\|Y\|)^2 w\left(\sqrt{\sum_{k=1}^p \frac{y_k^2}{\delta_k^2}}\right)dY \\
&\quad + \int_B (y_i^2 - y_j^2)\xi(\|Y\|)^2 w\left(\sqrt{\sum_{k=1}^p \frac{y_k^2}{\delta_k^2}}\right)dY \\
&= \int_A (y_i^2 - y_j^2)\xi(\|Y\|)^2 w\left(\sqrt{\sum_{k=1}^p \frac{y_k^2}{\delta_k^2}}\right)dY \\
&\quad + \int_A (y_j^2 - y_i^2)\xi(\|Y\|)^2 w\left(\sqrt{\sum_{k=1}^p \frac{y_k^2}{\delta_k^2} + \Delta_{ij}}\right)dY \\
&= \int_A (y_i^2 - y_j^2)\xi(\|Y\|)^2 \left\{w\left(\sqrt{\sum_{k=1}^p \frac{y_k^2}{\delta_k^2}}\right) - w\left(\sqrt{\sum_{k=1}^p \frac{y_k^2}{\delta_k^2} + \Delta_{ij}}\right)\right\}dY
\end{aligned}$$

where in the second equality we have changed the variables of the integration over B by replacing (y_i, y_j) by $(-y_j, y_i)$ which has Jacobian 1. The Δ_{ij} in that step is the correction term

$$\Delta_{ij} = y_i^2/\delta_j^2 + y_j^2/\delta_i^2 - y_i^2/\delta_i^2 - y_j^2/\delta_j^2 = (y_i^2 - y_j^2)/\delta_j^2 - (y_i^2 - y_j^2)/\delta_i^2 = (y_i^2 - y_j^2)(1/\delta_j^2 - 1/\delta_i^2).$$

Note that on A it holds that $|y_i| > |y_j|$ hence $y_i^2 - y_j^2 > 0$ so $\Delta_{ij} > 0$. Since w is a decreasing function, it follows that

$$w\left(\sqrt{\sum_{k=1}^{p}\frac{y_k^2}{\delta_k^2}}\right) - w\left(\sqrt{\sum_{k=1}^{p}\frac{y_k^2}{\delta_k^2} + \Delta_{ij}}\right) > 0$$

which implies (A.3) so $\lambda_{g,i} > \lambda_{g,j}$. In contrast if δ_i and δ_j are tied, i.e., $\delta_i = \delta_j$, it follows that $\Delta_{ij} = 0$ hence $\lambda_{g,i} = \lambda_{g,j}$. This concludes the proof of Proposition 1. □

A.3. Influence functions

Proof of Proposition 2. Consider the contaminated distribution $F_{\varepsilon,z} = (1-\varepsilon)F + \varepsilon\Delta_z$, where $z \in \mathbb{R}^p$ and $\varepsilon \in [0, 1]$. We then have

$$S_g(F_{\varepsilon,z}) = \mathrm{E}_{F_{\varepsilon,z}}\{g(X)g(X)^\top\} = (1-\varepsilon)\int g_\varepsilon(X)g_\varepsilon(X)^\top dF(X) + \varepsilon\int g_\varepsilon(X)g_\varepsilon(X)^\top d\Delta_z.$$

If we take the derivative with respect to ε and evaluate it in $\varepsilon = 0$, we get

$$\frac{\partial}{\partial\varepsilon}S_g(F_{\varepsilon,z})\bigg|_{\varepsilon=0} = g(z)g(z)^\top - \Xi_g + \frac{\partial}{\partial\varepsilon}\int g_\varepsilon(X)g_\varepsilon(X)^\top dF(X)\bigg|_{\varepsilon=0}.$$

Calculation of the IF. While the expression of the influence function might seem relatively simple, its (numerical) calculation is rather involved. We can write

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\int g_\varepsilon(X)g_\varepsilon(X)^\top dF(X)\bigg|_{\varepsilon=0} &= \int \frac{\partial}{\partial\varepsilon}\left\{g_\varepsilon(X)\right\}g_\varepsilon(X)^\top + g_\varepsilon(X)\frac{\partial}{\partial\varepsilon}\left\{g_\varepsilon(X)^\top\right\}dF(X)\bigg|_{\varepsilon=0}\\ &= \int\left\{\frac{\partial}{\partial\varepsilon}g_\varepsilon(X)\bigg|_{\varepsilon=0}\right\}g(X)^\top + g(X)\left\{\frac{\partial}{\partial\varepsilon}g_\varepsilon(X)^\top\bigg|_{\varepsilon=0}\right\}dF(X).\end{aligned}$$

So the term we need to determine is $\partial g_\varepsilon(X)/\partial\varepsilon|_{\varepsilon=0}$. Recalling that $g(t) = t\xi(\|t\|)$ we have $g_\varepsilon(t) = t\xi_\varepsilon(\|t\|)$. This means that the contamination affects g because it affects the radial function ξ. Therefore we have to compute $\partial g_\varepsilon(X)/\partial\varepsilon|_{\varepsilon=0} = X\,\partial\xi_\varepsilon(\|X\|)/\partial\varepsilon|_{\varepsilon=0}$ for the functions g given by (4)–(8).

In these functions ξ depends on F_X though the distribution of $\|X\|^{2/3}$. Suppose that $\|X\|^{2/3} \sim G$ when $X \sim F$, so G is a univariate distribution. For $X_\varepsilon \sim F_{\varepsilon,z} = (1-\varepsilon)F + \varepsilon\Delta_z$ we then have $\|X_\varepsilon\|^{2/3} \sim G_{\varepsilon,\|z\|^{2/3}} = (1-\varepsilon)G + \varepsilon\Delta_{\|z\|^{2/3}}$. For uncontaminated data the density of $\|X\|^{2/3}$ is given by

$$f_G(t) = f_{\mathrm{coga}}(t^3)|3t^2|,$$

where f_{coga} is the density of the **co**nvolution of **ga**mma distributions. We need this density to evaluate the influence functions of their median and mad. The cutoffs in the paper are

$$\begin{aligned}Q_1 &= (\mathrm{hmed}\|X\|^{2/3} - \mathrm{hmad}\|X\|^{2/3})^{3/2}, \quad Q_2 = (\mathrm{hmed}\|X\|^{2/3})^{3/2},\\ Q_3 &= (\mathrm{hmed}\|X\|^{2/3} + \mathrm{hmad}\|X\|^{2/3})^{3/2}, \quad Q_3^* = (\mathrm{hmed}\|X\|^{2/3} + 1.4826\times\mathrm{hmad}\|X\|^{2/3})^{3/2},\end{aligned}$$

and we can compute their influence functions, viz.

$$\begin{aligned}\mathrm{IF}(z,Q_1,F) =& \frac{3}{2}\sqrt{\mathrm{median}(G) - \mathrm{mad}(G)}\{\mathrm{IF}(\|z\|^{2/3},\mathrm{median},G) - \mathrm{IF}(\|z\|^{2/3},\mathrm{mad},G)\},\\ \mathrm{IF}(z,Q_2,F) =& \frac{3}{2}\sqrt{\mathrm{median}(G)}\mathrm{IF}(\|z\|^{2/3},\mathrm{median},G),\\ \mathrm{IF}(z,Q_3,F) =& \frac{3}{2}\sqrt{\mathrm{median}(G) + \mathrm{mad}(G)}\{\mathrm{IF}(\|z\|^{2/3},\mathrm{median},G) + \mathrm{IF}(\|z\|^{2/3},\mathrm{mad},G)\},\\ \mathrm{IF}(z,Q_3^*,F) =& \frac{3}{2}\sqrt{\mathrm{median}(G) + 1.4826\times\mathrm{mad}(G)}\{\mathrm{IF}(\|z\|^{2/3},\mathrm{median},G)\\ &+ 1.4826\times\mathrm{IF}(\|z\|^{2/3},\mathrm{mad},G)\}.\end{aligned}$$

The Winsor GSSCM is given by $\xi(r) = \mathbf{1}_{r\leqslant Q_2} + Q_2/r\,\mathbf{1}_{r>Q_2}$. For the contaminated case this becomes $\xi_\varepsilon(r) = \mathbf{1}_{r\leqslant Q_{2,\varepsilon}} + Q_{2,\varepsilon}/r\,\mathbf{1}_{r>Q_{2,\varepsilon}}$. We then have

$$\frac{\partial}{\partial\varepsilon}\xi_\varepsilon(r) = \frac{\partial}{\partial\varepsilon}\left\{\mathbf{1}_{[0,Q_{2,\varepsilon}]}(r) + \frac{Q_{2,\varepsilon}}{r}\mathbf{1}_{(Q_{2,\varepsilon},\infty)}(r)\right\} = \delta(r-Q_{2,\varepsilon})Q'_{2,\varepsilon} + \frac{Q'_{2,\varepsilon}}{r}\mathbf{1}_{(Q_{2,\varepsilon},\infty)}(r) - \frac{Q_{2,\varepsilon}}{r}\delta(r-Q_{2,\varepsilon})Q'_{2,\varepsilon},$$

where $\delta(x-y)$ denotes the distributional derivative of $\mathbf{1}_{(-\infty,x]}(y) = \mathbf{1}_{[y,\infty)}(x)$ with respect to x. Evaluation in $\varepsilon = 0$ gives

$$\begin{aligned}&\delta(r-Q_2)IF(z,Q_2,F) + \frac{\mathrm{IF}(z,Q_2,F)}{r}\,\mathbf{1}_{(Q_2,\infty)}(r) - \frac{Q_2}{r}\delta(r-Q_2)IF(z,Q_2,F)\\ &\quad= \left(1-\frac{Q_2}{r}\right)\delta(r-Q_2)IF(z,Q_2,F) + \frac{\mathrm{IF}(z,Q_2,F)}{r}\,\mathbf{1}_{(Q_2,\infty)}(r).\end{aligned}$$

As $(1-Q_2/r)\,\delta(r-Q_2)$ is 0 everywhere, we only need to integrate the last term. This yields

$$\left.\frac{\partial}{\partial\varepsilon}g_\varepsilon(X)\right|_{\varepsilon=0} = \frac{X}{\|X\|}\,\mathrm{IF}(z,Q_2,F)\mathbf{1}_{(Q_2,\infty)}(\|X\|).$$

The influence function of S_g is thus given by

$$\begin{aligned}\mathrm{IF}(z,S_g,F) =& g(z)g(z)^\top - \Xi_g(F)\\ &+ \int\left\{\frac{X}{\|X\|}\mathrm{IF}(z,Q_2,F)\mathbf{1}_{(Q_2,\infty)}(\|X\|)\right\}g(X)^\top dF(X) + \int g(X)\left\{\frac{X}{\|X\|}\mathrm{IF}(z,Q_2,F)\mathbf{1}_{(Q_2,\infty)}(\|X\|)\right\}^\top dF(X).\end{aligned}$$

Note that the last two terms in the sum are each other's transpose. The integration is done numerically.

The derivation of the influence function of the Quad GSSCM is entirely similar to that of Winsor. The main difference is that now $\partial g_\varepsilon(X)/\partial\varepsilon|_{\varepsilon=0}$ is given by

$$\left.\frac{\partial}{\partial\varepsilon}g_\varepsilon(X)\right|_{\varepsilon=0} = 2Q_2\mathrm{IF}(z,Q_2,F)\frac{X}{\|X\|^2}\mathbf{1}_{(Q_2,\infty)}(\|X\|).$$

The linearly redescending (LR) method uses a second cutoff, viz.

$$\xi(r) = \begin{cases}1 & \text{if } r \leqslant Q_2,\\ (Q_3^* - r)/(Q_3^* - Q_2) & \text{if } Q_2 < r \leqslant Q_3^*,\\ 0 & \text{if } r > Q_3^*.\end{cases}$$

In the contaminated case we obtain $g_\varepsilon(x) = x\xi_\varepsilon(\|x\|)$ with

$$\xi_\varepsilon(r) = \begin{cases}1 & \text{if } r \leqslant Q_{2,\varepsilon},\\ (Q_{3,\varepsilon}^* - r)/(Q_{3,\varepsilon}^* - Q_{2,\varepsilon}) & \text{if } Q_{2,\varepsilon} < r \leqslant Q_{3,\varepsilon}^*,\\ 0 & \text{if } r > Q_{3,\varepsilon}^*.\end{cases}$$

Taking the derivative with respect to ε yields

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\xi_\varepsilon(r) =\ & \delta(r-Q_{2,\varepsilon}) + \frac{Q_{3,\varepsilon}^* - r}{Q_{3,\varepsilon}^* - Q_{2,\varepsilon}}\left\{\delta(r-Q_{3,\varepsilon}^*) - \delta(r-Q_{2,\varepsilon})\right\}\\ &+\mathbf{1}_{[Q_{2,\varepsilon},Q_{3,\varepsilon}^*]}\frac{Q_{3,\varepsilon}^{*\prime}(Q_{3,\varepsilon}^* - Q_{2,\varepsilon}) - (Q_{3,\varepsilon}^{*\prime} - Q_{2,\varepsilon}^\prime)(Q_{3,\varepsilon}^* - r)}{(Q_{3,\varepsilon}^* - Q_{2,\varepsilon})^2}.\end{aligned}$$

Evaluation in $\varepsilon = 0$ gives

$$\begin{aligned}&\delta(r-Q_2) + \frac{Q_3^* - r}{Q_3^* - Q_2}\left\{\delta(r-Q_3^*) - \delta(r-Q_2)\right\}\\ &+\mathbf{1}_{[Q_2,Q_3^*]}\frac{\mathrm{IF}(z,Q_3^*,F)(Q_3^* - Q_2) - \{\mathrm{IF}(z,Q_3^*,F) - \mathrm{IF}(z,Q_2,F)\}(Q_3^* - r)}{(Q_3^* - Q_2)^2}.\end{aligned}$$

When integrating only the last term plays a role, yielding

$$\begin{aligned}\left.\frac{\partial}{\partial\varepsilon}g_\varepsilon(X)\right|_{\varepsilon=0} &= X\mathbf{1}_{[Q_2,Q_3^*]}(\|X\|)\\ &\qquad\frac{\mathrm{IF}(\|z\|,Q_3^*,F)(Q_3^* - Q_2) - \{\mathrm{IF}(\|z\|,Q_3^*,F) - \mathrm{IF}(\|z\|,Q_2,F)\}(Q_3^* - \|X\|)}{(Q_3^* - Q_2)^2}\\ &= X\mathbf{1}_{[Q_2,Q_3^*]}(\|X\|)\frac{\mathrm{IF}(\|z\|,Q_3^*,F)(\|X\| - Q_2) + \mathrm{IF}(\|z\|,Q_2,F)(Q_3^* - \|X\|)}{(Q_3^* - Q_2)^2}.\end{aligned}$$

For the Ball GSSCM we analogously derive that

$$\left.\frac{\partial}{\partial\varepsilon}g_\varepsilon(X)\right|_{\varepsilon=0} = \delta(\|X\| - Q_2)\mathrm{IF}(z,Q_2,F)X.$$

Finally, for the Shell GSSCM we obtain

$$\frac{\partial}{\partial\varepsilon}g_\varepsilon(X)\bigg|_{\varepsilon=0} = \{\delta(\|X\|-Q_3)\mathrm{IF}(z,Q_3,F)-\delta(\|X\|-Q_1)\mathrm{IF}(z,Q_1,F)\}\,X.$$

This concludes the proof of Proposition 2. □

A.4. Breakdown values

Proof of Proposition 3. Denote by $\mathcal{J}$ the set of all subsets of $\{1,\ldots,n\}$ with $p+1$ elements. For every subset $J\in\mathcal{J}$ we define $\eta_J = \max_{i\in J} d^2(x_i,H_J)$, where H_J is the hyperplane minimizing

$$\sum_{i\in J} d^2(x_i,H)$$

over all possible hyperplanes H and $d(x,H)$ is the Euclidean distance between a point x and a hyperplane H.

Define $\eta_{\boldsymbol{X}} = \min_{J\in\mathcal{J}}\eta_J$. Since the original points $\{x_1,\ldots,x_n\}$ are in general position, no $p+1$ points can lie on the same hyperplane, which ensures that $\eta_{\boldsymbol{X}}>0$. We also put $c_1=\max_i\|x_i-T(\boldsymbol{X})\|<\infty$.

Part 1. We first need to show that $\varepsilon^*\geqslant\lfloor(n-p+1)/2\rfloor/n$.

Let $m<\lfloor(n-p+1)/2\rfloor$ and replace m observations of $\boldsymbol{X}=\{x_1,\ldots,x_n\}$ yielding $\boldsymbol{X}^*$ with location estimate $T(\boldsymbol{X}^*)$. Because m/n is below the breakdown value of T, there is a constant $c_2<\infty$ so that $\|T(\boldsymbol{X}^*)-T(\boldsymbol{X})\|\leqslant c_2$ for all such contaminated datasets $\boldsymbol{X}^*$. By the triangle inequality, $\|x_i-T(\boldsymbol{X}^*)\|\leqslant c_1+c_2<\infty$. This implies $\mathrm{hmed}(d_i^*)\leqslant c_1+c_2$, hence $\mathrm{hmed}(d_i^*)+1.4826\times\mathrm{hmad}(d_i^*)\leqslant 2.4826\times\mathrm{hmed}(d_i^*)\leqslant 2.4826\times(c_1+c_2)$, where $d_i^*=\|x_i^*-T(\boldsymbol{X}^*)\|$. Therefore $\|g(t)\|\leqslant 2.4826\times(c_1+c_2)$ by condition 3.

First we show that the largest eigenvalue of $S_g(\boldsymbol{X}^*)$ is bounded over all such datasets $\boldsymbol{X}^*$. Take any $\boldsymbol{X}^*$, obtained by replacing m points of $\boldsymbol{X}$ by arbitrary points. Then

$$\begin{aligned}\lambda_{\max} &= \sup_{\|u\|=1} u^\top S_g(\boldsymbol{X}^*)u = \sup_{\|u\|=1}\frac{1}{n}\sum_{i=1}^n u^\top g\{x_i^*-T(\boldsymbol{X}^*)\}g\{x_i^*-T(\boldsymbol{X}^*)\}^\top u\\ &= \sup_{\|u\|=1}\frac{1}{n}\sum_{i=1}^n\left[u^\top g\{x_i^*-T(\boldsymbol{X}^*)\}\right]^2 \leqslant \sup_{\|u\|=1}\frac{1}{n}\sum_{i=1}^n\|u\|^2\|g\{x_i^*-T(\boldsymbol{X}^*)\}\|^2 \leqslant \{2.4826\times(c_1+c_2)\}^2<\infty.\end{aligned}$$

Next we show that the smallest eigenvalue of $S_g(\boldsymbol{X}^*)$ has a positive lower bound for all contaminated datasets $\boldsymbol{X}^*$. By condition 2 on ξ we know that $\#\{x_i:\xi\{\|x_i-T(\boldsymbol{X}^*)\|\}=1\}\geqslant\lfloor(n+p+1)/2\rfloor$. Therefore, we have at least $\lfloor(n+p+1)/2\rfloor-(\lfloor(n-p+1)/2\rfloor-1)=p+1$ regular points for which $\xi\{\|x_i-T(\boldsymbol{X}^*)\|\}=1$, let us assume without loss of generality that these are $x_1,\ldots,x_{p+1}$. We can now write

$$\begin{aligned}\lambda_{\min} &= \min_{\|u\|=1} u^\top S_g(\boldsymbol{X}^*)u = \min_{\|u\|=1}\frac{1}{n}\sum_{i=1}^n u^\top g\{x_i^*-T(\boldsymbol{X}^*)\}g\{x_i^*-T(\boldsymbol{X}^*)\}^\top u = \min_{\|u\|=1}\frac{1}{n}\sum_{i=1}^n\left[u^\top g\{x_i^*-T(\boldsymbol{X}^*)\}\right]^2\\ &\geqslant \min_{\|u\|=1}\frac{1}{n}\sum_{i=1}^{p+1}\left[u^\top\{x_i-T(\boldsymbol{X}^*)\}\xi\{x_i-T(\boldsymbol{X}^*)\}\right]^2 = \min_{\|u\|=1}\frac{1}{n}\sum_{i=1}^{p+1}\left[u^\top\{x_i-T(\boldsymbol{X}^*)\}\right]^2\\ &\geqslant \frac{1}{n}\sum_{i=1}^{p+1} d^2(x_i,H_{\{1,\ldots,p+1\}}) \geqslant \eta_{\boldsymbol{X}}>0.\end{aligned}$$

Part 2. It remains to show that $\varepsilon^*\leqslant\lfloor(n-p+1)/2\rfloor/n$. This is the known upper bound for affine equivariant scatter estimators but that result does not apply here, so we need to show it for this case. Take any $m\geqslant\lfloor(n-p+1)/2\rfloor$ and replace the last m points of $\boldsymbol{X}$, keeping the points $x_1,\ldots,x_{n-m}$ unchanged. By location equivariance we can assume without loss of generality that the average of $x_1,\ldots,x_{n-m}$ is zero. For $j\in\{n-m+1,\ldots,n\}$, put $x_j^*=\lambda a_j$, where a_j is such that $\min_{i\in\{n-m+1,\ldots,n\}}\|a_j-a_i\|\geqslant 1$ and such that for all $\lambda>1$ one has $\min_{i\in\{1,\ldots,n-m\}}\|\lambda a_j-x_i\|\geqslant\lambda$. This is possible by placing the a_j outside of the convex hull of $\boldsymbol{X}$ and far enough from each other and $\boldsymbol{X}$.

Now consider an unbounded increasing sequence of $\lambda_k>1$. For every λ_k the set $\{x_{n-m+1}^*,\ldots,x_n^*\}$ must contain at least one point for which $w_i=1$, call this point x_b^*. Take another point of $\boldsymbol{X}^*$ for which $w_i=1$, name this x_c^*. Note that x_c^* can be an original data point or a replaced point. We now have that $\|x_b^*-x_c^*\|\geqslant\lambda$ hence $\|x_b^*-T(\boldsymbol{X}^*)\|+\|x_c^*-T(\boldsymbol{X}^*)\|\geqslant\lambda$. Therefore $\|x_b^*-T(\boldsymbol{X}^*)\|^2+\|x_c^*-T(\boldsymbol{X}^*)\|^2\geqslant\lambda^2/2$. We then obtain

$$\begin{aligned}\sum_{j=1}^p\lambda_j\{S_g(\boldsymbol{X}^*)\} &= \mathrm{trace}\{S_g(\boldsymbol{X}^*)\} = \frac{1}{n}\sum_{i=1}^n\mathrm{trace}[\{x_i^*-T(\boldsymbol{X}^*)\}\{x_i^*-T(\boldsymbol{X}^*)\}^\top] = \frac{1}{n}\sum_{i=1}^n\|x_i^*-T(\boldsymbol{X}^*)\|^2\\ &\geqslant \frac{1}{n}\{\|x_b^*-T(\boldsymbol{X}^*)\|^2+\|x_c^*-T(\boldsymbol{X}^*)\|^2\}\geqslant\lambda^2/(2n).\end{aligned}$$

This becomes arbitrarily large and so $S_g(\boldsymbol{X}^*)$ explodes. This concludes the proof of Proposition 3. □

Proof of Proposition 4. Showing that $\varepsilon^*(T, \boldsymbol{X}) \leqslant \lfloor (n+1)/2 \rfloor / n$ is easy, since $\lfloor (n+1)/2 \rfloor / n$ is the upper bound on the breakdown value of all translation equivariant location estimators; see, e.g., [18].

It remains to show that $\varepsilon^*(T, \boldsymbol{X}) \geqslant \lfloor (n+1)/2 \rfloor / n$.

Note that the objective given by the sum of the h smallest squared Euclidean distances is nonincreasing in every C-step. The value of the objective function after step k is

$$\sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}, T_k(\boldsymbol{X})\},$$

where $d_{(j)}\{\boldsymbol{X}, T_k(\boldsymbol{X})\}$ denotes the jth order statistic of the distances $\|x_i - T_k(\boldsymbol{X})\|$, and we have that

$$\sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}, T_k(\boldsymbol{X})\} \leqslant \sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}, T_{k-1}(\boldsymbol{X})\}.$$

Recall that $h = \lfloor (n+1)/2 \rfloor$ and define $c_1 = \max_i \|x_i - T_k(\boldsymbol{X})\| < \infty$. Let $m < n - h$ and replace without loss of generality the last m observations of $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ to obtain $\boldsymbol{X}^* = \{x_1, \ldots, x_{n-m}, x^*_{n-m+1}, \ldots, x^*_n\} = \{x^*_1, \ldots, x^*_n\}$. Since the spatial median T_0 does not yet break down for this m [18], there is a constant c_2 such that $\max_i \|x_i - T_0(\boldsymbol{X}^*)\| \leqslant c_2 < \infty$ for all such datasets $\boldsymbol{X}^*$.

Consider $T_k(\boldsymbol{X}^*)$ and the corresponding objective function $\sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}^*, T_k(\boldsymbol{X}^*)\}$. Since the C-step does not increase the value of the objective function, we have that

$$\sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}^*, T_k(\boldsymbol{X}^*)\} \leqslant \sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}^*, T_{k-1}(\boldsymbol{X}^*)\} \leqslant \cdots \leqslant \sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}^*, T_0(\boldsymbol{X}^*)\}.$$

Note that

$$\sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}^*, T_0(\boldsymbol{X}^*)\} \leqslant \sum_{i=1}^{h} \|x^*_i - T_0(\boldsymbol{X}^*)\|^2 = \sum_{i=1}^{h} \|x_i - T_0(\boldsymbol{X}^*)\|^2 \leqslant \left\{ \sum_{i=1}^{h} \|x_i - T_0(\boldsymbol{X}^*)\| \right\}^2 \leqslant (hc_2)^2 .$$

Since m is at most $\lfloor (n-1)/2 \rfloor$ and $h = \lfloor (n+1)/2 \rfloor$ we have at least $\lfloor (n+1)/2 \rfloor - \lfloor (n-1)/2 \rfloor = 1$ point x_j with $1 \leqslant j \leqslant n - m$ for which $\|x_j - T_k(\boldsymbol{X}^*)\|^2 \leqslant d^2_{(h)}\{\boldsymbol{X}^*, T_k(\boldsymbol{X}^*)\}$. Note that

$$\|x_j - T_k(\boldsymbol{X}^*)\|^2 \leqslant \sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}^*, T_k(\boldsymbol{X}^*)\} \leqslant \sum_{j=1}^{h} d^2_{(j)}\{\boldsymbol{X}^*, T_0(\boldsymbol{X}^*)\}.$$

So for this x_j we can write

$$\|T_k(\boldsymbol{X}^*) - T_0(\boldsymbol{X})\| \leqslant \|T_k(\boldsymbol{X}^*) - x_j\| + \|x_j - T_0(\boldsymbol{X})\| \leqslant hc_2 + c_1 < \infty.$$

Note that this upper bound does not depend on k and therefore remains valid when the procedure is iterated until convergence ($k \to \infty$). This concludes the proof of Proposition 4. □

References

[1] G. Boente, D. Rodriguez, M. Sud, The spatial sign operator: Asymptotic results and applications, J. Multivariate Anal. 170 (2018) (in press).
[2] B.M. Brown, Statistical uses of the spatial median, J. R. Stat. Soc. Ser. B Stat. Methodol. 45 (1983) 25–30.
[3] C. Chatzinakos, L. Pitsoulis, G. Zioutas, Optimization techniques for robust multivariate location and scatter estimation, J. Comb. Optim. 31 (2016) 1443–1460.
[4] C. Croux, C. Dehon, A. Yadine, The k-step spatial sign covariance matrix, Adv. Data Anal. Classif. 4 (2010) 137–150.
[5] C. Croux, E. Ollila, H. Oja, Sign and rank covariance matrices: Statistical properties and application to principal components analysis, in: Y. Dodge (Ed.), Statistical Data Analysis Based on the L_1-Norm and Related Methods, Birkhäuser, Basel, 2002, pp. 257–269.
[6] D. Donoho, P. Huber, The notion of breakdown point, in: P. Bickel, K. Doksum, J. Hodges (Eds.), A Festschrift for Erich Lehmann, Wadsworth, Belmont, CA, pp. 157–184.
[7] A. Dürre, R. Fried, D. Vogel, The spatial sign covariance matrix and its application for robust correlation estimation, Austrian J. Statist. 46 (2017) 13–22.
[8] A. Dürre, D.E. Tyler, D. Vogel, On the eigenvalues of the spatial sign covariance matrix in more than two dimensions, Statist. Probab. Lett. 111 (2016) 80–85.
[9] A. Dürre, D. Vogel, Asymptotics of the two-stage spatial sign correlation, J. Multivariate Anal. 144 (2016) 54–67.
[10] A. Dürre, D. Vogel, R. Fried, Spatial sign correlation, J. Multivariate Anal. 135 (2015) 89–105.
[11] A. Dürre, D. Vogel, D.E. Tyler, The spatial sign covariance matrix with unknown location, J. Multivariate Anal. 130 (2014) 107–117.
[12] J.C. Gower, Algorithm AS 78: The Mediancentre, J. R. Stat. Soc. Ser. C. Appl. Stat. 23 (1974) 466–470.
[13] F. Hampel, E. Ronchetti, P. Rousseeuw, W. Stahel, Robust Statistics: The Approach Based on Influence Functions, Wiley, New York, 1986.
[14] C. Hu, V. Pozdnyakov, J. Yan, Coga: Convolution of Gamma Distributions, University of Connecticut, 2018. R package version 0.2.2.
[15] M. Hubert, P.J. Rousseeuw, K. Vande. Branden, ROBPCA: A new approach to robust principal component analysis, Technometrics 47 (2005) 64–79.
[16] M. Hubert, P.J. Rousseeuw, T. Verdonck, A deterministic algorithm for robust location and scatter, J. Comput. Graph. Statist. 21 (2012) 618–637.
[17] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, K.L. Cohen, Robust principal component analysis for functional data, Test 8 (1999) 1–28.

[18] H. Lopuhaä, P. Rousseeuw, Breakdown points of affine equivariant estimators of multivariate location and covariance matrices, Ann. Statist. 19 (1991) 229–248.
[19] A.F. Magyar, D.E. Tyler, The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions, Biometrika 101 (2014) 673–688.
[20] J.I. Marden, Some robust estimates of principal components, Statist. Probab. Lett. 43 (1999) 349–359.
[21] P.G. Moschopoulos, The distribution of the sum of independent gamma random variables, Ann. Inst. Statist. Math. 37 (1985) 541–544.
[22] P. Mozharovskyi, K. Mosler, T. Lange, Classifying real-world data with the DDα-procedure, Adv. Data Anal. Classif. 9 (2015) 287–314.
[23] O. Pokotylo, P. Mozharovskyi, R. Dyckerhoff, Depth and depth-based classification with R-package `ddalpha`, arXiv:1608.04109, 2016.
[24] G.M. Reaven, R.G. Miller, An attempt to define the nature of chemical diabetes using a multidimensional analysis, Diabetologia 16 (1979) 17–24.
[25] D.M. Rocke, Robustness properties of S-estimators of multivariate location and shape in high dimension, Ann. Statist. 24 (1996) 1327–1345.
[26] P. Rousseeuw, Least median of squares regression, J. Amer. Statist. Assoc. 79 (1984) 871–880.
[27] P. Rousseeuw, K. Van Driessen, A fast algorithm for the Minimum Covariance Determinant estimator, Technometrics 41 (1999) 212–223.
[28] S. Serneels, E. De Nolf, P.J. Van Espen, Spatial sign preprocessing: A simple way to impart moderate robustness to multivariate estimators, J. Chem. Inf. Model. 46 (2006) 1402–1409, PMID: 16711760.
[29] S. Sirkia, S. Taskinen, H. Oja, D.E. Tyler, Tests and estimates of shape based on spatial signs and ranks, J. Nonparametr. Stat. 21 (2009) 155–176.
[30] S. Taskinen, I. Koch, H. Oja, Robustifying principal component analysis with spatial sign vectors, Statist. Probab. Lett. 82 (2012) 765–774.
[31] S. Visuri, V. Koivunen, H. Oja, Sign and rank covariance matrices, J. Statist. Plann. Inference 91 (2000) 557–575.
[32] S. Visuri, H. Oja, V. Koivunen, Subspace-based direction-of-arrival estimation using nonparametric statistics, IEEE Trans. Signal Process. 49 (2001) 2060–2073.
[33] E.B. Wilson, M.M. Hilferty, The distribution of chi-square, Proc. Nat. Acad. Sci. USA 17 (1931) 684–688.