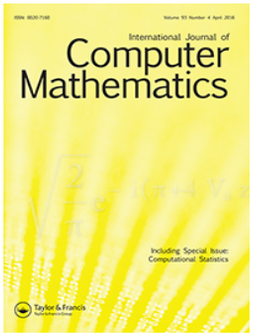

International Journal of Computer Mathematics

ISSN: 0020-7160 (Print) 1029-0265 (Online) Journal homepage: https://www.tandfonline.com/loi/gcom20

Stahel–Donoho estimation for high-dimensional data

S. Van Aelst

To cite this article: S. Van Aelst (2016) Stahel–Donoho estimation for high-dimensional data, International Journal of Computer Mathematics, 93:4, 628-639, DOI: 10.1080/00207160.2014.933815

To link to this article: https://doi.org/10.1080/00207160.2014.933815

International Journal of Computer Mathematics, 2016
Vol. 93, No. 4, 628–639, http://dx.doi.org/10.1080/00207160.2014.933815

Taylor & Francis
Taylor & Francis Group

Stahel–Donoho estimation for high-dimensional data

S. Van Aelst[a,b*]

[a]*Department of Mathematics, Section of Statistics, KULeuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium;* [b]*Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium*

(*Received 12 November 2013; revised version received 27 May 2014; accepted 3 June 2014*)

We discuss two recently proposed adaptations of the well-known Stahel–Donoho estimator of multivariate location and scatter for high-dimensional data. The first adaptation adjusts the calculation of the outlyingness of the observations while the second adaptation allows to give separate weights to each of the components of an observation. Both adaptations address the possibility that in higher dimensions most observations can be contaminated in at least one of its components. We then combine the two approaches in a new method and investigate its performance in comparison to the previously proposed methods.

Keywords: high-dimensional data; robustness; location and scatter estimation; outlyingness; cellwise weights

2010 AMS Subject Classifications: 62F35; 62H12

1. Introduction

The Stahel–Donoho estimator (SDE) is a highly robust estimator of multivariate location and scatter [7,19]. SDE measures the outlyingness of each observation and weighs the observations accordingly. The larger the outlyingness of an observation, the lower its weight. Recent applications of the Stahel–Donoho outlyingness measure can be found in [3,5,12,13]. The SDE has excellent robustness properties [6,10,17,23] according to the standard Tukey–Huber contamination model. In particular, the estimator has a breakdown point (i.e. the maximum fraction of outliers that an estimator can withstand) close to 50% in any dimension which makes the SDE useful for multivariate outlier detection [4,8].

The Tukey–Huber model is the most often used contamination model to evaluate the robustness of estimators. This model assumes that the majority of the observations are generated from a nominal distribution while the remaining observations come from an arbitrary distribution-generating outliers. This model thus assumes that there is a majority of observations that is completely *clean*, i.e. none of their components are contaminated while the outlying observations are assumed to be completely spoiled. Therefore, robust estimators such as the SDE give equal weight to all components of an observation. However, for high-dimensional data sets such a contamination model is less realistic. If all of the variables have some chance of being contaminated, then it can easily happen that the majority of the observations are contaminated. On the other hand, most of

*Email: stefan.vanaelst@wis.kuleuven.be

the observations will not be contaminated in all measured components. To handle this situation, a more flexible contamination model has been developed in [2]. An interesting particular case of this general contamination model is the independent contamination model in which each variable contains some fraction of contamination, independently of the other variables. This leads to *cellwise* contamination in a data set. Note that in practice next to cellwise outliers also *structural outliers* which affect the correlation structure of the data can occur.

In the independent contamination model the fraction of observations that is contaminated in at least one of its cells only depends on the dimension p of the data. In case that all the variables have the same probability ϵ of being contaminated, the probability that an observation is contaminated in at least one of its cells equals $1 - (1 - \epsilon)^p$. If the dimension p increases, this probability easily exceeds the breakdown point of the SDE. For example with $\epsilon = 0.05$ the SDE breaks down for $p \geq 14$ and with $\epsilon = 0.01$ breakdown occurs for $p \geq 69$. In practice, the outliers in the data may well be an unknown combination of cellwise and structural contamination. The breakdown point of the SDE then depends on both the dimension p and the sample size n, and the above numbers only give an upper bound. Hence, for low quality data dimension $p \geq 10$ can already be considered large in this setting.

The SDE encounters two problems in high-dimensional data. The first problem occurs when determining the outlyingness of each observations. This outlyingness is based on the one-dimensional projection in which the observation is most outlying. If a majority of the observations contains contaminated components, then many of these projections contain a majority of projected outliers. In such directions, both *masking* and *swamping* effects can occur. Masking implies that a projected outlier is not recognized as such in this projection because the outlyingness measure is affected too heavily by the majority of outliers. Swamping occurs if the majority of outliers have such a large effect on the outlyingness measure that the minority of regular observations are incorrectly regarded as outliers. These masking and swamping effects thus make it impossible to accurately measure the outlyingness of each observation. The second problem occurs when weighing the observations. If only few components of an observation are contaminated, then a lot of useful information is wasted if the entire observation is downweighted. This is especially the case if only few observations are completely outlier-free.

To overcome the first problem, i.e. the masking and swamping effects in directions with a majority of outliers, Van Aelst *et al.* [21] proposed to shrink extreme values in each of the components towards the bulk of the data before computing the outlyingness of an observation. This shrinking, called *huberization* or *winsorization* [1,11,14], reduces the effect of cellwise outliers which makes it possible to determine more reliably the outlyingness of each observation.

To overcome the second problem a cellwise weighting scheme has been introduced by Van Aelst *et al.* [20]. In this scheme all components of an observation are given different weights. These componentwise weights take into account the componentwise outlyingness of each component as well as the relevance of the component in the direction with maximal outlyingness for the observation. With this weighting scheme structural outliers will have most or all of their components downweighted while cellwise outliers will have only one or a few components downweighted.

Note that the independent contamination model, and contamination models with cellwise contamination in general lack affine equivariance and suffer from the *outlier propagation* effect [2]. That is, linear transformations of the data can largely increase the number of outlying cells so that a majority of the cells becomes contaminated. Therefore, estimators that can handle cellwise contamination have to give up on affine equivariance, because (near) affine equivariant estimators can only produce reliable results if a majority of the observations is completely clean (regardless of the linear transformation applied to the data).

In this paper we combine the two previous solutions to simultaneously address both of the above problems. We compare the performance of the combined method to the two previously proposed methods that address only one of the issues. In Section 2 we review the standard SDE and its

adaptations using adjusted outlyingness [21] or cellwise weights [20]. We then explain the new combined method. Section 3 shows the results of simulation studies to compare the performance of the Stahel–Donoho adaptations and Section 4 summarizes our conclusions.

2. Methods

Let $\mathbf{X}$ be an $n \times p$ data matrix corresponding to n observations $x_1, \ldots, x_n$ in $\mathbb{R}^p$. Let μ and σ be shift and scale equivariant univariate location and scale statistics. Then, for any $y \in \mathbb{R}^p$ its Stahel–Donoho *outlyingness* w.r.t. $\mathbf{X}$ is defined as

$$r(y, \mathbf{X}) = \sup_{a \in S_p} \frac{|y'a - \mu(\mathbf{X}a)|}{\sigma(\mathbf{X}a)}, \tag{1}$$

with $S_p = \{a \in \mathbb{R}^p : ||a|| = 1\}$. In particular, for the observations x_i we denote the corresponding outlyingness $r(x_i, \mathbf{X})$ by r_i. Note that in practice S_p is a finite set of (randomly) selected directions. In our algorithm these directions are obtained as the directions orthogonal to the hyperplane spanned by p randomly selected points from the data.

The Stahel–Donoho estimator of location and scatter $(T_{\mathrm{SD}}, \mathbf{S}_{\mathrm{SD}})$ is defined as

$$T_{\mathrm{SD}} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad \text{and} \quad \mathbf{S}_{\mathrm{SD}} = \frac{\sum_{i=1}^n w_i (x_i - T_{\mathrm{SD}})(x_i - T_{\mathrm{SD}})'}{\sum_{i=1}^n w_i}, \tag{2}$$

where $w_i = w(r_i)$ with $w : \mathbb{R}^+ \to \mathbb{R}^+$ a weight function so that observations with large outlyingness get small weights [7,19].

Following Maronna and Yohai [17], we use for w the Huber-type weight function, defined as

$$w(r) = I_{(r \le c)} + \left(\frac{c}{r}\right)^2 I_{(r > c)}, \tag{3}$$

for some threshold c. The choice of the threshold c is a trade-off between robustness and efficiency. Small values of c quickly start to downweigh observations with increasing outlyingness while larger values of c only downweigh observations with extreme outlyingness value. Following Maronna and Zamar [18] we set the threshold equal to $c = \min(\sqrt{\chi^2_p(0.50)}, 4)$ to obtain robust estimates in high dimensions.

To attain maximum breakdown point [9,17], the univariate location statistic μ is taken to be the median (MED) and the scale statistic σ is chosen to be the modified median absolute deviation (MAD), defined as

$$\mathrm{MAD}^*(\mathbf{X}a) = \frac{|\mathbf{X}a - \mathrm{MED}(\mathbf{X}a)|_{\lceil (n+p-1)/2 \rceil : n} + |\mathbf{X}a - \mathrm{MED}(\mathbf{X}a)|_{(\lfloor (n+p-1)/2 \rfloor + 1) : n}}{2\beta}, \tag{4}$$

where $\beta = \Phi^{-1}(\frac{1}{2}((n+p-1)/2n+1))$, $\lceil x \rceil$ and $\lfloor x \rfloor$ indicate the ceiling and the floor of x, respectively, and $\mathrm{v}_{i:n}$ denotes the ith order statistic of the data vector v.

If the observations are projected onto a direction with a majority of outliers, then masking and swamping effects may occur as mentioned before. Masking occurs when an outlier is considered a regular observation because the median and modified MAD (i.e. MAD^* defined in Equation (4)) of the projected data have been adversely affected by the high amount of outliers. Swamping occurs when a regular observation is falsely identified as an outlier because the median and modified MAD are dominated by the large group of outliers. To reduce these effects, Van Aelst *et al.* [21] proposed to huberize the data matrix $\mathbf{X}$ and calculate the outlyingness of the observations with respect to the huberized data matrix.

Let $X_1, \ldots, X_p$ denote the columns of the data matrix $\mathbf{X}$. Then, the huberized observations $x_{1,H}, \ldots, x_{n,H}$ that form the data matrix $\mathbf{X}_H$ are defined as

$$x_{ij,H} = \begin{cases} \text{MED}(X_j) - c_H \, \text{MAD}^*(X_j) & \text{if } c_{ij} < -c_H, \\ x_{ij} & \text{if } -c_H \leq c_{ij} \leq c_H, \\ \text{MED}(X_j) + c_H \, \text{MAD}^*(X_j) & \text{if } c_{ij} > c_H, \end{cases}$$

where $x_{ij,H}$ denotes the jth component of $x_{i,H}$ and

$$c_{ij} = \frac{x_{ij} - \text{MED}(X_j)}{\text{MAD}^*(X_j)}. \tag{5}$$

The cutoff parameter c_H determines the amount of shrinkage which is a trade-off between robustness and efficiency. The idea is that unusually large components are shrunken towards the centre of the data distribution. We choose $c_H = \Phi^{-1}(0.975)$, i.e. the 97.5% quantile of a standard normal distribution, which is a standard choice for univariate outlier identification.

For any $y \in \mathbb{R}^p$ the adjusted outlyingness measure $r_H(y, \mathbf{X})$ then calculates the outlyingness of y with respect to the huberized data matrix, i.e.

$$r_H(y, \mathbf{X}) = r(y, \mathbf{X}_H) = \sup_{a \in S_p} \frac{|y'a - \mu(\mathbf{X}_H a)|}{\sigma(\mathbf{X}_H a)}. \tag{6}$$

For the observations x_i, the corresponding huberized outlyingness $r_H(x_i, \mathbf{X})$ is denoted by $r_{i,H}$. By calculating the outlyingness of observations with respect to the huberized data matrix, the effect of componentwise outliers in the data can be reduced which makes the outlyingness values and corresponding direction of maximal outlyingness more reliable as illustrated in [21]. The huberized SDE considered in [21] is obtained by using the huberized outlyingnesses $r_{i,H}$ in the weight function (3) when calculating the weighted mean and covariance in Equation (2).

To avoid unnecessary downweighting of a complete observation as soon as one of its components is contaminated, Van Aelst *et al.* [20] introduced a more flexible, the cellwise weighting method. The cellwise weighted Stahel–Donoho estimator of location and scatter is defined as

$$T_{SDc,j} = \frac{\sum_{i=1}^n w_{ij} x_{ij}}{\sum_{i=1}^n w_{ij}} \tag{7}$$

and

$$\mathbf{S}_{SDc,jk} = \frac{\sum_{i=1}^n \sqrt{w_{ij}}\sqrt{w_{ik}}(x_{ij} - T_{SDc,j})(x_{ik} - T_{SDc,k})}{\sum_{i=1}^n \sqrt{w_{ij}}\sqrt{w_{ik}}} \tag{8}$$

for $j, k = 1, \ldots, p$. The weight matrix $W = (w_{ij})_{ij}$ is defined by $w_{ij} = w(r_{ij})$ where w is the weight function w in Equation (3). Here, r_{ij} is a cellwise outlyingness which is defined as

$$r_{ij} = \alpha_{ij} r_i + (1 - \alpha_{ij})|c_{ij}|, \tag{9}$$

where α_{ij} is a weighting parameter, r_i is the Stahel–Donoho outlyingness of x_i (i.e. $r_i = r(x_i, X)$) and c_{ij} is given in Equation (5). Hence, $|c_{ij}|$ is the outlyingness of x_i in the direction of component j. The weighting parameter α_{ij} in Equation (9) will be chosen in $[0, 1]$, such that the cellwise outlyingness r_{ij} is a convex combination of the global outlyingness r_i of the observation and the outlyingness $|c_{ij}|$ of its jth component. Note that $r_i \geq r_{ij}$ because $r_i \geq |c_{ij}|$.

The idea behind the cellwise outlyingness is to largely keep the global outlyingness r_i for those components that are responsible for it, while allowing to reduce the outlyingness (and increase the weight) of those components which contributed little to the global outlyingness r_i. To achieve

this goal, taking into account that the data may contain cellwise as well as structural outliers, the following two choices for the parameters α_{ij} have shown good performance in [20]:

(1) $\alpha_{ij} = (\max_{k=1}^{p} |c_{ik}|)^{-1}|c_{ij}|$ It follows that α_{ij} is large whenever c_{ij} is large, relative to the outlyingnesses in the direction of the other components. The effect of this weighting parameter α_{ij} can be explained as follows: (1) If observation x_i contains cellwise contamination then this choice yields a strong contrast between the r_{ij} of contaminated and non-contaminated components. If x_i is a structural outlier then all of the c_{ij} may be relatively small and this choice avoids reduction of outlyingness for those components responsible for the large outlyingness r_i. This cellwise Stahel–Donoho estimator is denoted by $(T_{\text{SDC}}, \mathbf{S}_{\text{SDC}})$.

(2) $\alpha_{ij} = (\max_{k=1}^{p} |u_{ik}|)^{-1}|u_{ij}|$ where $u_i = (u_{i1}, \ldots, u_{ip})$ denotes the direction that maximizes r_i. Hence, the parameters α_{ij} are proportional to the coefficients in the maximizing direction u_i. The magnitude of these coefficients reflect to what extent each component contributes to the outlyingness r_i, both in the case of componentwise and structural outliers. Note that the direction u_i depends on the scales of the variables X_j. Therefore, the components X_j are rescaled using $\text{MAD}^*(X_j)$. This cellwise Stahel–Donoho estimator is denoted by $(T_{SDM}, \mathbf{S}_{SDM})$.

To handle all types of outliers in high-dimensional data, we now combine the huberized outlyingness in Equation (6) with the cellwise outlyingness in Equation (9). That is, we replace Equation (9) by cellwise huberized outlyingnesses, given by

$$r_{ij,H} = \alpha_{ij} r_{i,H} + (1 - \alpha_{ij})|c_{ij,H}|, \tag{10}$$

where $c_{ij,H}$ is given by $(x_{ij} - \text{MED}(X_{j,H}))/\text{MAD}^*(X_{j,H})$ with $X_{1,H}, \ldots, X_{p,H}$ the columns of $\mathbf{X}_H$. The corresponding modified Stahel–Donoho estimators for high-dimensional data are then obtained by using the corresponding cellwise weights $w_{ij} = w(r_{ij,H})$ in Equations (7) and (8). The idea is that by using huberized outlyingness measures the maximal outlyingness of each observation and its corresponding direction can be determined more accurately in the presence of cellwise contamination. The cellwise weights then allow to recover the information from those components that are not contaminated. If the parameters α_{ij} are derived from the componentwise outlyingnesses $|c_{ij,H}|$ (case 1 above), then we denote the corresponding high-dimensional Stahel–Donoho estimator by $(T_{\text{HSDC}}, \mathbf{S}_{\text{HSDC}})$. If the parameters α_{ij} are derived from the maximizing direction u_i (case 2 above), then we denote the corresponding high-dimensional Stahel–Donoho estimator by $(T_{\text{HSDM}}, \mathbf{S}_{\text{HSDM}})$.

3. Simulation studies

We consider simulation settings similar as in [20,21] to investigate the performance of the newly proposed combined method in comparison to the previous solutions. We examine the performance of the methods through their mean-squared error (MSE). Since huberization ignores any correlation among the variables, we consider data with different levels of correlation to investigate the effect of correlation on the huberized outlyingness.

As in [21], data sets were generated from a p-variate normal distribution with mean zero and covariance matrix $\mathbf{R}^2$ where $\mathbf{R}$ is a matrix whose diagonal elements are equal to 1 and all off-diagonal elements have the same value ρ. The value of ρ is chosen such that the multiple correlation coefficient R^2 between any component of the p-variate distribution and all the other components takes the values 0, 0.5, 0.7, 0.9 or 0.999. The dimension p took the values 5 and 10, while the sample size n is equal to 50 for $p = 5$ and equal to 100 for $p = 10$. A fraction ϵ of univariate outliers was then introduced in the first d components ($d \leq p$) with $d = 2$ ($p = 5$) or $d = 5$ ($p = 10$). The outlying values were generated from a univariate normal distribution with mean $k/\sqrt{d}$ and

standard deviation 0.1 where the outlying distances $k = 6, 24, 64$ and 160 were used. For each situation, $N = 500$ samples were generated and the outlyingnesses were calculated by setting the number of random directions equal to $200p$, as advocated in [17] for data in higher dimensions. It has been shown in [15,16] that exact calculation of the SDE requires a number of data-dependent directions that is of order $O(n^{4p})$ which is too high for practical purposes [22]. Therefore, a smaller number of data-dependent random directions is usually considered. The choice of the number of random directions is a trade-off between computational efficiency and accuracy of the SDE approximation. Using $200p$ gave good results in our experiments, but if sufficient computer power is available this number can be increased to obtain more accurate results.

Then, for each of the $N = 500$ generated data sets in each setting we computed six different estimates. These are the original SD location and scatter estimates, its cellwise adaptations SDC and SDM, the huberized Stahel–Donoho (HSD) estimates and its cellwise adaptations: the huberized Stahel-Donoho estimator using componentwise weights (HSDC) and the huberized Stahel-Donoho estimator using maximal direction weights (HSDM). The MSE for the location estimators of each of these methods was calculated as

$$\mathrm{MSE}(T_\cdot) = \operatorname*{ave}_{j=1,\ldots,p} \left(\operatorname*{ave}_{l=1,\ldots,N} (T_\cdot^{(l)})_j^2 \right).$$

Table 1. MSE ratios of adjusted Stahel–Donoho estimators vs. original SDE for data in 5 dimensions with $\epsilon = 20\%$ of independent contamination in the first two components with $k = 6$ or $k = 64$.

	Rsq	k	comp	SDC	SDM	HSD	HSDC	HSDM
Centre	0	6	All	0.544	0.611	0.960	0.611	0.667
Centre	0	6	Cont	0.444	0.567	0.954	0.564	0.634
Centre	0	64	All	0.802	0.816	1.088	0.795	0.893
Centre	0	64	Cont	0.672	0.695	1.241	0.950	0.994
Diag	0	6	All	0.626	0.710	0.882	0.456	0.577
Diag	0	6	Cont	0.619	0.691	0.870	0.387	0.530
Diag	0	64	All	0.433	0.459	0.915	0.319	0.489
Diag	0	64	Cont	0.428	0.451	0.910	0.292	0.473
Offdiag	0	6	All	0.963	1.117	0.984	0.794	1.137
Offdiag	0	6	1 Cont	0.994	1.135	0.998	0.811	1.157
Offdiag	0	6	2 Cont	0.901	1.095	0.949	0.782	1.131
Offdiag	0	64	All	1.318	1.366	0.782	0.959	1.009
Offdiag	0	64	1 Cont	2.360	1.948	1.024	1.779	1.675
Offdiag	0	64	2 Cont	0.927	1.152	0.611	0.643	0.674
Centre	90	6	All	1.333	0.699	0.997	1.017	0.814
Centre	90	6	Cont	1.899	0.737	1.008	1.187	0.860
Centre	90	64	All	1.092	0.837	1.020	0.894	0.870
Centre	90	64	Cont	1.692	0.685	1.036	1.110	0.829
Diag	90	6	All	1.193	1.236	1.028	1.454	1.316
Diag	90	6	Cont	1.019	2.507	1.067	0.913	2.255
Diag	90	64	All	1.219	1.800	1.044	1.494	1.619
Diag	90	64	Cont	1.015	5.128	1.027	0.910	3.950
Offdiag	90	6	All	1.549	0.729	1.000	1.235	0.819
Offdiag	90	6	1 Cont	1.614	0.733	0.998	1.132	0.780
Offdiag	90	6	2 Cont	2.471	0.947	0.994	0.946	0.917
Offdiag	90	64	All	1.514	0.698	1.044	1.231	0.773
Offdiag	90	64	1 Cont	1.576	0.632	1.044	1.121	0.722
Offdiag	90	64	2 Cont	2.322	1.156	1.045	0.915	0.968

Notes: Both uncorrelated data and correlated data ($R^2 = 0.9$) are considered. The ratio of the overall MSE averages (all) are shown as well as the ratio of the MSE averages of the contaminated components (Cont). For the off-diagonal elements, we further differentiate between elements with only one contaminated component (1 cont) and elements with both components contaminated (2 cont).

We also calculated the MSE for the diagonal elements of the covariance matrix $\mathbf{R}^2$ as

$$\mathrm{MSE}(\mathbf{S}_{\cdot}^{\mathrm{diag}}) = \underset{j=1,\ldots,p}{\mathrm{ave}} \left(\underset{l=1,\ldots,N}{\mathrm{ave}} [(\mathbf{S}_{\cdot}^{(l)})_{jj} - (\mathbf{R}^2)_{jj}]^2 \right)$$

and similarly for the MSE for the off-diagonal elements.

Note that for most combinations of ϵ and d in the simulations, the fraction of contaminated observations exceeds the breakdown point (50%) of the SDE. Our purpose is to see to what extent the adjusted Stahel–Donoho estimators can better withstand such amounts of contamination.

We first consider the case $p = 5$ and $d = 2$. Table 1 shows the results when the first two components contain 20% of independent contamination. This table contains for each of the adjusted methods the ratio of its MSE with respect to the original SDE. MSE ratios are shown for the location estimator as well as for the diagonal and off-diagonal elements of the scatter matrix estimator. For the various settings, Table 1 shows the overall MSE ratio which takes all $p = 5$ components into account, as well as the MSE ratio based on only the two contaminated components. The latter provides information about the difference in bias between the adjusted estimator and the original SDE due to the contamination in these components. For the off-diagonal elements, we further differentiate between elements related to two contaminated components (2 cont) and elements related to a contaminated and an uncontaminated component (1 cont).

The fraction of contaminated observations in the settings of Table 1 is well below 50% and thus the original SDE should not be highly affected by it. However, since the SDE can only downweight complete observations, a large amount of useful information may be wasted and we examine to what extend the adjusted methods are able to recover some of that information. The top half of Table 1 contains the results for uncorrelated data with close-by ($k = 6$) and further away ($k = 64$) contamination. We can see that huberizing the data (HSD) has little effect on the performance of the estimator in this case. For the estimates of the centre this is illustrated in more detail in the top panel of Figure 1 which contains the boxplots of the absolute errors of the estimates for the components of the centre. Separate boxplots are shown for the contaminated

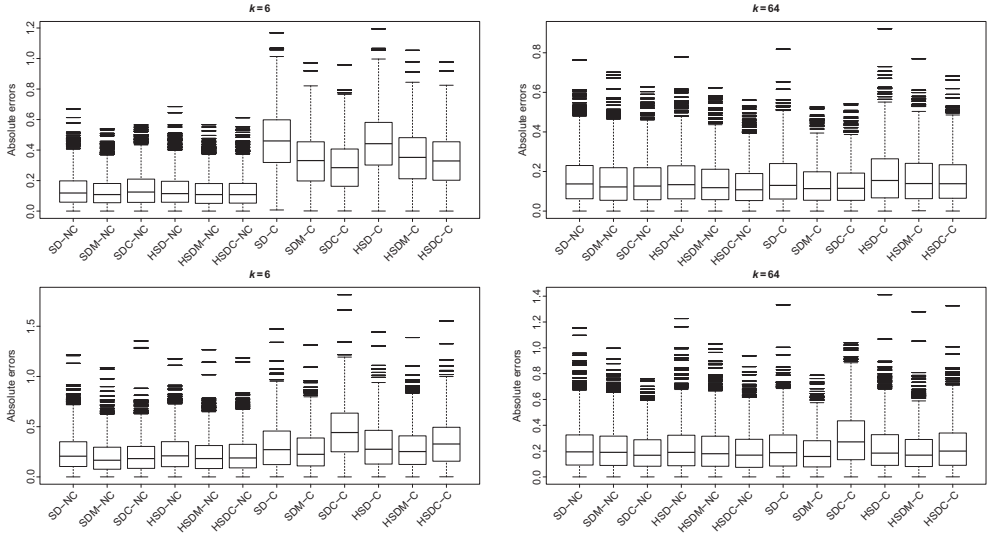

Figure 1. Boxplots of absolute errors of the estimates of the uncontaminated and contaminated components of the centre. Data were generated in five dimensions with $\epsilon = 20\%$ of independent contamination in the first two components. The top panels show the results for uncorrelated data and the bottom panels contain the results for correlated data ($R^2 = 0.9$). The left plots correspond to $k = 6$ and the right plots to $k = 64$.

and uncontaminated components. The boxplots of the SDE and HSD estimates look very similar, both for the contaminated and uncontaminated components. Using cellwise outlyingnesses and weights on the other hand does have an effect on the performance of the estimators. The (H)SDC and (H)SDM estimators improve on the SDE, especially for the estimators of the centre and diagonal elements of the scatter. For the off-diagonal elements, only the HSDM succeeds in yielding a better overall MSE. The boxplots in Figure 1 reveal that the gain in performance is mainly due to a bias reduction in the contaminated components.

To show the effect of increasing correlation, the bottom half of Table 1 contains the results for (highly) correlated data ($R^2 = 0.9$). Huberization does not take into account any correlation among the variables. As a result, we can see that the HSD cannot improve on the SDE in the presence of strong correlations. Of the cellwise adaptations, the (H)SDC relies most on the componentwise outlyingness which is more difficult to measure in case of highly correlated data. As a consequence the outlying components may receive a too high weight as can be seen from Figure 1. It follows that the estimates become biased leading to a worse performance as can be seen in Table 1. The (H)SDM estimators focuses on the direction in which each observation is most outlying and pays

Table 2. MSE ratios of adjusted Stahel–Donoho estimators vs. original SDE for data in five dimensions with $\epsilon = 35\%$ of independent contamination in the first two components with $k = 6$ or $k = 64$.

	Rsq	k	comp	SDC	SDM	HSD	HSDC	HSDM
Centre	0	6	All	0.661	0.902	0.998	0.692	0.933
Centre	0	6	Cont	0.647	0.899	0.998	0.687	0.931
Centre	0	64	All	0.012	0.015	0.035	0.011	0.015
Centre	0	64	Cont	0.008	0.011	0.029	0.008	0.012
Diag	0	6	All	0.891	0.991	0.994	0.840	0.968
Diag	0	6	Cont	0.893	0.991	0.993	0.832	0.965
Diag	0	64	All	0.010	0.016	0.039	0.008	0.013
Diag	0	64	Cont	0.010	0.016	0.039	0.008	0.013
Offdiag	0	6	All	0.752	1.046	1.005	0.710	1.062
Offdiag	0	6	1 Cont	0.833	1.094	1.005	0.764	1.097
Offdiag	0	6	2 Cont	0.675	1.001	1.005	0.663	1.031
Offdiag	0	64	Al l	0.034	0.055	0.052	0.020	0.032
Offdiag	0	64	1 Cont	0.494	0.632	0.114	0.322	0.490
Offdiag	0	64	2 Cont	0.032	0.052	0.052	0.018	0.030
Centre	90	6	All	0.747	0.693	1.024	0.875	0.754
Centre	90	6	Cont	0.821	0.762	1.034	0.924	0.832
Centre	90	64	All	0.022	0.022	0.035	0.018	0.020
Centre	90	64	Cont	0.012	0.011	0.020	0.012	0.012
Diag	90	6	All	1.010	1.518	1.059	1.208	1.558
Diag	90	6	Cont	1.020	1.668	1.064	1.077	1.675
Diag	90	64	All	0.001	0.011	0.009	0.001	0.009
Diag	90	64	Cont	0.001	0.011	0.009	0.001	0.009
Offdiag	90	6	All	1.313	0.912	1.025	1.241	1.007
Offdiag	90	6	1 Cont	1.261	0.847	1.024	1.138	0.926
Offdiag	90	6	2 Cont	1.967	1.443	1.028	1.078	1.468
Offdiag	90	64	All	0.016	0.125	0.014	0.018	0.041
Offdiag	90	64	1 Cont	0.664	0.507	0.384	0.454	0.367
Offdiag	90	64	2 Cont	0.003	0.118	0.005	0.005	0.034

Notes: Both uncorrelated data and correlated data ($R^2 = 0.9$) are considered. The ratio of the overall MSE averages (all) are shown as well as the ratio of the MSE averages of the contaminated components (Cont). For the off-diagonal elements, we further differentiate between elements with only one contaminated component (1 cont) and elements with both components contaminated (2 cont).

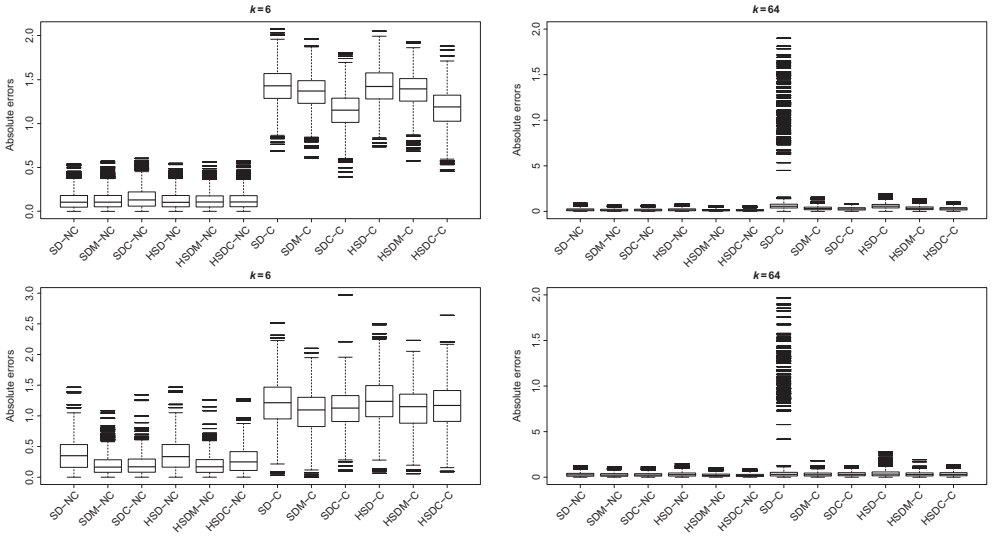

Figure 2. Boxplots of absolute errors of the estimates of the uncontaminated and contaminated components of the centre. Data were generated in five dimensions with $\epsilon = 35\%$ of independent contamination in the first two components. The top panels show the results for uncorrelated data and the bottom panels contain the results for correlated data ($R^2 = 0.9$). The left plots correspond to $k = 6$ and the right plots to $k = 64$.

less attention to the componentwise outlyingness. In this setting this approach leads to better weights for the components and thus results in estimators with a better performance.

Table 2 and Figure 2 show the performance of the estimators when the fraction of contamination in the first two components is increased to 35%. It can be seen from the left plots in Figure 2 that the effect on the SDE is rather small when the contamination is close by ($k = 6$). In this case the HSD estimates again show similar performance. When the contamination lies further away from the bulk of the data, its effect on the SDE becomes much larger. Note that in this setting it can easily happen that a majority of observations is contaminated. This implies that the SD outlyingnesses are not determined accurately anymore because the majority of outliers may lead to masking and swamping effects, i.e. the SD outlyingness of the outliers is underestimated whereas regular observations may even receive an SD outlyingness r_i that is far too large. HSD effectively reduces the effect of the componentwise contamination which resolves the masking and swamping problems and thus the bias on the contaminated components is much smaller for the HSD as shown in the plots on the right of Figure 2. This naturally results in a much lower MSE as can be seen in Table 2. For low correlated data the HSD improves on the SDE as soon as the SDE is seriously affected by the componentwise contamination. For highly correlated data the outliers need to be further away before the HSD succeeds in reducing their effect to a large extent.

The cellwise methods can already improve on the SDE when the contamination is close by ($k = 6$) by recovering information from the clean components of the contaminated observations. This works best in low correlation settings where it is easier to determine which components are responsible for the outlyingness of an observation. For close by outliers the huberized cellwise estimators do not yield any further improvement which is in line with the fact that huberization was not helpful in this case. However, when the outliers lie further away, huberization becomes useful and the huberized cellwise estimators generally perform even better than their non-huberized cellwise counterparts. In low correlation settings the (H)SDC estimators generally outperform the (H)SDM, but for highly correlated data their order reverses.

Table 3. MSE ratios of adjusted Stahel–Donoho estimators vs. original SDE for data in 10 dimensions with $\epsilon = 20\%$ of independent contamination in the first five components with $k = 6$ or $k = 64$.

	Rsq	k	comp	SDC	SDM	HSD	HSDC	HSDM
Centre	0	6	All	0.764	0.972	0.994	0.864	0.993
Centre	0	6	Cont	0.749	0.972	0.994	0.859	0.994
Centre	0	64	All	0.211	0.241	0.676	0.163	0.267
Centre	0	64	Cont	0.190	0.199	0.636	0.137	0.239
Diag	0	6	All	0.971	1.074	0.987	0.891	1.035
Diag	0	6	Cont	0.980	1.082	0.984	0.871	1.035
Diag	0	64	All	0.166	0.171	0.524	0.074	0.156
Diag	0	64	Cont	0.165	0.171	0.524	0.074	0.156
Offdiag	0	6	All	0.916	1.031	0.998	0.875	1.032
Offdiag	0	6	1 Cont	0.937	1.049	1.002	0.886	1.048
Offdiag	0	6	2 Cont	0.878	1.000	0.994	0.858	1.005
Offdiag	0	64	All	0.192	0.201	0.285	0.074	0.148
Offdiag	0	64	1 Cont	1.439	1.525	0.808	0.974	1.464
Offdiag	0	64	2 Cont	0.167	0.174	0.272	0.056	0.121
Centre	90	6	All	0.463	0.304	0.993	0.717	0.331
Centre	90	6	Cont	0.553	0.358	0.993	0.800	0.384
Centre	90	64	All	0.468	0.625	1.071	0.583	0.679
Centre	90	64	Cont	0.524	0.694	1.094	0.698	0.795
Diag	90	6	All	0.857	1.021	1.003	1.321	0.992
Diag	90	6	Cont	1.399	2.336	1.000	0.914	2.034
Diag	90	64	All	0.192	2.336	0.971	0.168	1.832
Diag	90	64	Cont	0.167	2.385	0.969	0.107	1.864
Offdiag	90	6	All	0.831	0.556	0.984	1.229	0.615
Offdiag	90	6	1 Cont	0.757	0.517	0.986	1.226	0.587
Offdiag	90	6	2 Cont	1.067	0.665	0.974	1.196	0.702
Offdiag	90	64	All	0.953	1.498	1.000	1.314	1.343
Offdiag	90	64	1 Cont	1.137	0.870	1.045	1.550	0.862
Offdiag	90	64	2 Cont	0.650	2.905	0.913	0.481	2.391

Notes: Both uncorrelated data and correlated data ($R^2 = 0.9$) are considered. The ratio of the overall MSE averages (all) are shown as well as the ratio of the MSE averages of the contaminated components (Cont). For the off-diagonal elements, we further differentiate between elements with only one contaminated component (1 cont) and elements with both components contaminated (2 cont).

In Table 3, the dimension was increased to $p = 10$ with $\epsilon = 20\%$ of independent contamination in the first five components. Clearly, there again is no majority of contamination free observations. For low correlated data, HSD improves on the SDE, especially if the outliers lie further away. However, in this setting with only componentwise outliers the SDC estimator yields a better improvement of the SDE in low correlation settings and HSDC yields even further improvement. In highly correlated settings huberization is again less successful. Using cellwise weights can largely improve the estimator for the centre, but the results for the scatter are more diverse. For the highly correlated data (H)SDM again generally performs better than (H)SDC.

In practice structural (correlation) outliers and independent contamination can occur simultaneously in a data set. To examine this case, we generated data sets of size $n = 100$ in $p = 10$ dimensions as before. If the correlation ρ is high, then the data are concentrated around the line with direction $\mathbf{e} = (1, \ldots, 1) \in \mathbb{R}^{\tilde{p}}$. Hence, 10% of structural outliers were added by shifting observations over a distance c in direction $c\mathbf{m}$, where $\mathbf{m}$ is a unit vector orthogonal to $\mathbf{e}$. In particular, the direction $\mathbf{m}$ is determined as follows. Take $\mathbf{b} \in \mathbb{R}^p$ with $b_j = (-1)^j$, and set $\mathbf{m} = \mathbf{b} - (\mathbf{b}'\mathbf{e}/p)\mathbf{e}$. Then $\mathbf{m}$ is orthogonal to $\mathbf{e}$ by construction. Finally, rescale $\mathbf{m}$ to unit norm. The distance c was generated uniformly within a range from 2 to 3 standard deviations of the

Table 4. MSE ratios of adjusted Stahel–Donoho estimators vs original SDE for data in 10 dimensions with 10% of structural outliers as well as 10% of independent contamination in all components. Both uncorrelated data and correlated data ($R^2 = 0.9$) are considered.

	Rsq	SDC	SDM	HSD	HSDC	HSDM
Centre	0	0.337	0.363	0.659	0.272	0.317
Centre	90	0.639	0.527	1.047	0.739	0.476
Diag	0	0.194	0.229	0.334	0.051	0.127
Diag	90	0.165	2.763	0.654	0.124	1.792
Offdiag	0	0.430	0.504	0.494	0.251	0.439
Offdiag	90	0.808	1.336	0.844	0.791	1.416

components. Finally, the standard deviation of the outliers was reduced by a factor 10. Next to these structural outliers, also 10% of independent contamination with $k = 64$ was added to each of the components of the remaining regular observations.

Table 4 shows the MSE ratios for uncorrelated data and correlated data ($R^2 = 0.90$). In this highly contaminated case, all adaptations of SDE successfully improve the SDE performance. HSD is again less successful for correlated data. In this setting with correlation outliers, the cellwise estimators can largely improve over the SDE. The huberization generally further improves the cellwise estimators. HSDM has lower MSE than HSDC for the centre estimators, but performs worse for the scatter matrix in this setting.

4. Conclusion

We reviewed a huberization of the SD outlyingness which calculates the outlyingness of observations with respect to a huberized data set that is obtained by componentwise pulling outliers back to the bulk of the data. This huberization yields more reliable outlyingness measures in settings with independent componentwise contamination. Contamination models that include componentwise outliers are especially realistic for high-dimensional data sets. In these contamination models the overall fraction of contamination can easily exceed 50% so that the original SDE breaks down.

We also reviewed an adaptation of the SDE that uses a more flexible weighting scheme in which a separate weight can be used for each component of the observations. These cellwise weighted adaptations can offer a considerable boost in performance if observations are outlying due to contamination in only a few of their components.

We combined these two approaches that were proposed to adjust the SDE for high-dimensional data. That is, we examined the performance of adjusted SDEs that use cellwise weights based on huberized SD outlyingness. Our empirical study shows that the combined approach often outperforms the HSD estimator and also can outperform the cellwise weighted estimators in high-dimensional situations with a large fraction of componentwise contamination. In low correlation data the HSDC usually outperforms HSDM, but with highly correlated data this order often reverses.

Acknowledgements

This work was supported by a grant of the Fund for Scientific Research-Flanders [grant number G.0.077.11.N.10] and by Interuniversity Attraction Pole (IAP) research network grant of the Belgian government (Belgian Science Policy) [grant number P7/06].

References

[1] F.A. Alqallaf, K.P. Konis, R.D. Martin, and R.H. Zamar, *Scalable robust covariance and correlation estimates for data mining*, in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, 2002, pp. 14–23.

[2] F. Alqallaf, S. Van Aelst, V.J. Yohai, and R.H. Zamar, *Propagation of outliers in multivariate data*, Ann. Stat. 37 (2009), pp. 311–331.

[3] K. Boudt, C. Croux, and S. Laurent, *Outlyingness weighted covariation*, Technical Report, ORSTAT Research Center, KULeuven, Belgium, 2009.

[4] A. Cerioli and A. Farcomeni, *Error rates for multivariate outlier detection*, Comput. Stat. Data Anal. 55 (2011), pp. 544–553.

[5] M. Debruyne, An outlier map for support vector machine classification, *Ann. Appl. Stat.* 3 (2009), pp. 1566–1580.

[6] M. Debruyne and M. Hubert, *The influence function of the Stahel–Donoho covariance estimator of smallest outlyingness*, Stat. Probab. Lett. 79 (2009), pp. 275–282.

[7] D.L. Donoho, *Breakdown properties of multivariate location estimators*, Ph.D. diss., Harvard University, 1982.

[8] P. Filzmoser, R. Maronna, and M. Werner, *Outlier identification in high dimensions*, Comput. Stat. Data Anal. 52 (2008), pp. 1694–1711.

[9] U. Gather and T. Hilker, *A note on Tyler's modification of the MAD for the Stahel–Donoho estimator*, Ann. Stat. 25 (1997), pp. 2024–2026.

[10] D. Gervini, *The influence function of the Stahel–Donoho estimator of multivariate location and scatter*, Stat. Probab. Lett. 60 (2002), pp. 425–435.

[11] P.J. Huber, *Robust Statistics*, Wiley, New York, 1981.

[12] M. Hubert and S. Verboven, *A robust PCR method for high-dimensional regressors*, J. Chemom. 17 (2003), pp. 438–452.

[13] M. Hubert, P.J. Rousseeuw, and K. Vanden Branden, *ROBPCA: A new approach to robust principal component analysis*, Technometrics 47 (2005), pp. 64–79.

[14] J.A. Khan, S. Van Aelst, and R.H. Zamar, *Robust linear model selection based on least angle regression*, J. Amer. Stat. Assoc. 102 (2007), pp. 1289-1299.

[15] X. Liu, and Y. Zuo, *Computing projection depth and its associated estimators*, Stat. Comput. 24 (2014), pp. 51–63.

[16] X. Liu, Y. Zuo, and Z. Wang, *Exactly computing bivariate projection depth contours and median*, Comput. Stat. Data Anal. 60 (2013), pp. 1–11.

[17] R.A. Maronna and V.J. Yohai, *The behavior of the Stahel–Donoho robust multivariate estimator*, J. Amer. Stat. Assoc. 90 (1995), pp. 329–341.

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

[19] W.A. Stahel, *Breakdown of covariance estimators*, Research Report 31, Fachgruppe für Statistik, E.T.H. Zürich, Switzerland, 1981.

[20] S. Van Aelst, E. Vandervieren, and G. Willems, *Stahel–Donoho estimators with cellwise weights*, J. Stat. Comput. Simul. 81 (2011), pp. 1–27.

[21] S. Van Aelst, E. Vandervieren, and G. Willems, *Stahel–Donoho estimator based on huberized outlyingness*, Comput. Stat. Data Anal. 56 (2012), pp. 531–542.

[22] Y. Zuo and S. Lai, *Exact computation of bivariate projection depth and the Stahel–Donoho estimator*, Comput. Stat. Data Anal. 55 (2011), pp. 1173–1179.

[23] Y. Zuo, H. Cui, and X. He, *On the Stahel–Donoho estimator and depth-weighted means of multivariate data*, Ann. Stat. 32 (2004), pp. 167–188.