Stahel-Donoho Estimator Based on Huberized Outlyingness

S. Van Aelst[a,*], E. Vandervieren[b], G. Willems[a]

[a]*Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium*
[b]*Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium*

Abstract

The Stahel-Donoho estimator is defined as a weighted mean and covariance, where the weight of each observation depends on a measure of its outlyingness. In high dimensions, it can easily happen that an amount of outlying measurements is present in such a way that the majority of the observations is contaminated in at least one of its components. In these situations, the Stahel-Donoho estimator has difficulties in identifying the actual outlyingness of the contaminated observations. An adaptation of the Stahel-Donoho estimator is presented where the data are huberized before the outlyingness is computed. It is shown that the huberized outlyingness better reflects the actual outlyingness of each observation towards the non-contaminated observations. Therefore, the resulting adapted Stahel-Donoho estimator can better withstand large amounts of outliers. It is demonstrated that the Stahel-Donoho estimator based on huberized outlyingness works especially well when the data are heavily contaminated.

Key words: robust multivariate estimators, outlier identification, outlyingness, huberization.

1. Introduction

The Stahel-Donoho (SD) estimator, proposed independently in [21] and [8], is a well-known robust estimator of multivariate location and scatter. It was the first affine equivariant estimator with breakdown point (i.e., the maximum fraction of outliers that the estimator can withstand) close to 50% for any dimension. It has excellent robustness properties as shown in [18, 11, 23, 7], which make the estimator useful for multivariate outlier detection (see [9, 4, 22]).

The SD estimator weighs the observations depending on a measure of their "outlyingness". This measure is based on the one-dimensional projection in

[*] Corresponding author. Email: Stefan.VanAelst@UGent.be Phone: +32-9-2644908. Fax: +32-9-2644995.

which the observation is most outlying. The underlying idea is that every multivariate outlier must be a univariate outlier in some projection. Hence, observations with large outlyingness receive a small weight. Recent applications of the Stahel-Donoho outlyingness measure can be found in [3, 15, 14, 6].

To study the performance of robust estimators, contamination or mixture models are used. Most multivariate contamination models assume that the majority of the observations comes from a nominal distribution such as a multivariate normal distribution, while the remainder comes from another distribution that generates outliers (see e.g. [12, 17]). Unfortunately, such models are not realistic for many large multivariate data sets. In high dimensions, it can easily happen that outlying measurements are present in such a way that the majority of the observations is contaminated in at least one of the components. To handle this case [2] proposed a new contamination model that can allow for instance contamination that appears in each of the variables independently. In such situations, the SD estimator has difficulties in identifying the actual outlyingness of the contaminated observations if the percentage of outlying observations approaches or exceeds its breakdown point. Indeed, the SD estimator considers all one-dimensional projections of the data and selects the projection in which the observation is most outlying. An observation that is contaminated in only one of its components is visible in the projection along the component in which the observation is contaminated. However, observations which are contaminated in more than one component usually have an outlyingness that far exceeds the componentwise outlyingnesses. This outlyingness is then achieved in a direction which is a linear combination of the various components. Now, if there is a majority of outlying observations in the data, then every such direction will produce a majority of projected outliers and this may prevent the detection of large outlyingness in that direction due to masking.

To overcome the masking effect when measuring outlyingness in this setting, we pull the outliers back to the bulk of the data, componentwise, before computing the outlyingness of an observation. This is done by using a lower and an upper bound for the various components of the observations. Extreme low and high values are set equal to the lower respectively upper bound as proposed in [13]. This componentwise shrinking of the extreme data is called *huberization* or *winsorization* (see e.g. [1, 16]) and makes it possible to determine more reliably the outlyingness of each observation by reducing the effect of other outliers.

In this paper we investigate to what extent the adapted SD estimator based on huberized outlyingness can withstand large amounts of contamination. Section 2 provides a review of the contamination model introduced in [2] with an emphasis on the case of independent contamination in the components. In Section 3 we discuss the standard calculation of outlyingness and the resulting SD estimator. In Section 4 we present our proposal for adapting the outlyingness by using huberization of the data. A simulation study is performed in Section 5, which investigates to what extent our proposal succeeds in giving larger outlyingness to contaminated observations, while Section 6 concludes.

2. Contamination models for high dimensions

The standard contamination model assumes that a majority of the observations are regular (outlier-free) observations following some underlying model distribution, while the remaining minority (outliers) can be anything. In high dimensional data, outlying measurements can come from different sources for various reasons. As a result, it is often unrealistic to assume that there exists a majority of completely uncontaminated data, but likely that most observations are contaminated in some of their measurements. Alternative contamination models are needed to handle this case. Therefore, [2] proposed the following flexible contamination model to study robustness properties at high dimensional data. Let X, Y and Z be p-dimensional random vectors, where Y follows some regular distribution F with mean μ and scatter matrix $\mathbf{\Sigma}$ and Z has an arbitrary distribution that generates the outliers. Then, the observed random variable X follows the model

$$X = (\mathbf{I} - \mathbf{B})Y + \mathbf{B}Z, \tag{1}$$

where $\mathbf{B} = diag\,(B_1, B_2, ..., B_p)$ is a diagonal matrix and $B_1, B_2, ..., B_p$ are Bernoulli random variables with $P\,(B_i = 1) = \epsilon_i$. As in [2] we consider the case where Y, Z and $\mathbf{B}$ are independent.

Different contamination models can now be obtained as special cases of (1) by making different assumptions about the joint distribution of $B_1, B_2, ..., B_p$. For example, the standard contamination model corresponds to the assumption $P\,(B_1 = B_2 = \cdots = B_p) = 1$, that is full dependence. In this case an observation is considered to be either completely contaminated or completely clean. Such contaminated observations are called *structural outliers*. Note that the fraction of contaminated observations in this model equals $\epsilon_1 = \cdots = \epsilon_p = \epsilon$ and this fraction remains fixed under affine transformations.

Another interesting case at the other end of the spectrum is the (fully) independent contamination model which corresponds to the assumption that $B_1, B_2, ..., B_p$ are independent. Hence, this model assumes that contamination in each variable is independent from the other variables, which leads to *componentwise outliers*. If $P\,(B_i = 1) = \epsilon$ for all components $(1 \leq i \leq p)$, then each variable contains on average a fraction $(1 - \epsilon)$ of clean measurements, but the probability that an observation is completely uncontaminated is only $(1 - \epsilon)^p$ under this model. Even for moderate fractions ϵ this probability quickly exceeds 50% if the dimension p increases, such that there is no majority of outlier-free observations anymore. It is important to note that contrary to the fully dependent case, the independent contamination model does not support affine transformations anymore. Indeed, while each of the components contains on average a fraction $1 - \epsilon$ of outlier-free measurements, linear combinations of these components may contain a much lower fraction of clean measurements. This phenomenon is called *outlier propagation*. As a consequence, estimation methods that are affine equivariant are not very robust (low breakdown point) under the independent contamination model as shown in [2] for well-known estimators such as the Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) estimators [20] and S-estimators [5]. It was empirically illustrated in [2] that also the SD lacks robustness in the independent contamination model. Moreover, using a similar argument as for the coordinatewise median in [2] it can easily be shown that the SD has the same low breakdown point in this model as other affine equivariant methods. In the next sections we provide more empirical evidence that the SD is heavily biased if the data contains a substantial fraction of componentwise contamination. Hence, estimation methods that aim to be robust under the independent contamination model need to give up on affine equivariance and resort to coordinatewise procedures.

In the next section, we study the outlyingness measure and corresponding SD estimator in the context of the independent contamination model. By using huberization we then reduce the effect of componentwise contamination on the SD outlyingness. Note that in practice componentwise outliers and structural outliers can occur simultaneously as discussed in [2]. Therefore, we do not completely restrict ourselves to coordinatewise procedures to avoid lack of robustness against the possible presence of structural outliers in the data.

3. Stahel-Donoho estimator

Let $\mathbf{X}$ be an $n \times p$ data matrix that contains 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 $y \in \mathbb{R}^p$, the Stahel-Donoho *outlyingness* is defined as

$$r(y, \mathbf{X}) = \sup_{a \in S_p} \frac{|y'a - \mu(\mathbf{X}a)|}{\sigma(\mathbf{X}a)}, \tag{2}$$

with $S_p = \{a \in \mathbb{R}^p : ||a|| = 1\}$. From now on, we will denote $r(x_i, \mathbf{X})$ by r_i.

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

$$T_{SD} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i},$$

and

$$\mathbf{S}_{SD} = \frac{\sum_{i=1}^n w_i \ (x_i - T_{SD})(x_i - T_{SD})'}{\sum_{i=1}^n w_i},$$

where $w_i = w(r_i)$ and $w : \mathbb{R}^+ \to \mathbb{R}^+$ is a weight function so that observations with large outlyingness get small weights (see [21, 8]).

Following [18], we use for w the Huber-type weight function, defined as

$$w(r) = I_{(r \leq c)} + (c/r)^2 I_{(r > c)}, \tag{3}$$

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. Several choices for the threshold c have been proposed in the literature, see e.g. [17]. [19] argues that a small value of the threshold c is needed to obtain robust estimates in high dimensions. Following their proposals for c, we choose the threshold as $c = \min(\sqrt{\chi^2_p(0.50)}, 4)$.

To attain maximum breakdown point (see e.g. [18, 10]) the univariate location statistic μ is taken to be the median (MED) and the scale statistic σ is chosen to be the modified MAD, defined as

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

where $\beta = \Phi^{-1}(\frac{1}{2}(\frac{n+p-1}{2n} + 1))$, $\lceil x \rceil$ 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.

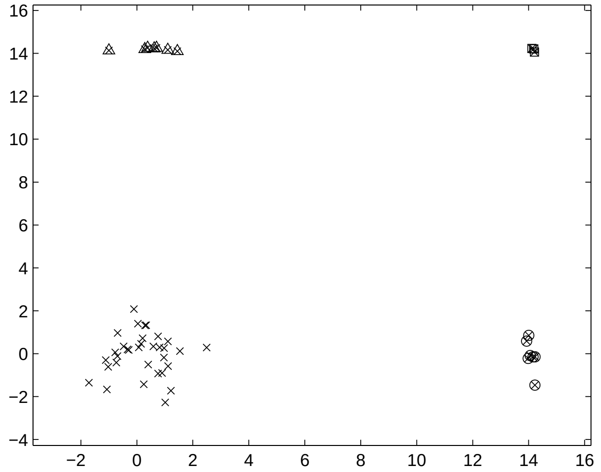
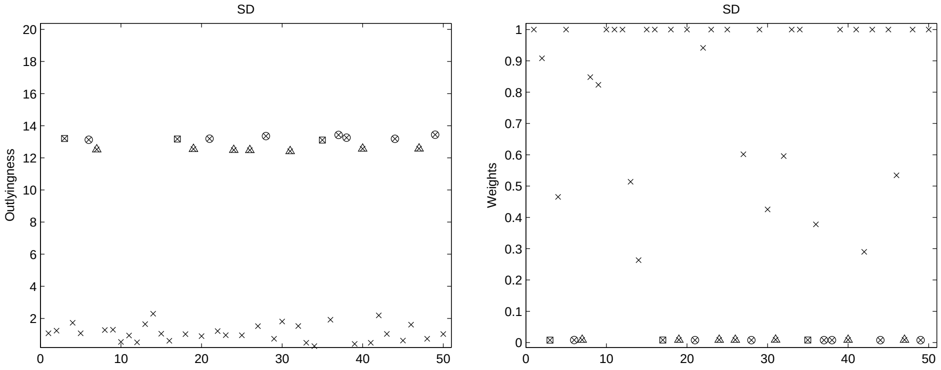


Figure 1: Data set of size $n = 50$ from a two-dimensional standard normal distribution with 20% componentwise outliers independently in both components.

As an example, we generated a data set of size $n = 50$ from a bivariate standard normal distribution and introduced componentwise outliers in both components independently, by randomly replacing 20% of its values with values generated from a normal distribution with mean $10\sqrt{2}$ and standard deviation 0.1. The resulting data are shown in Figure 1. To ease interpretation later on, we use squares to mark observations that are outlying in both components. Furthermore, circles correspond to observations that are only outlying in the first component and triangles are used for observations that are only contaminated in the second component.

Since exact computation of the supremum in the outlyingness (2) is impractical (see [24]), usually a random search algorithm based on subsampling is used. We use a Matlab implementation of the algorithm in [18]. The number of random directions considered by the algorithm is a trade-off between computational feasibility and quality of the obtained approximate solution, see [18] for an extensive discussion. For bivariate data, the computations are performed quickly,



(a) (b)

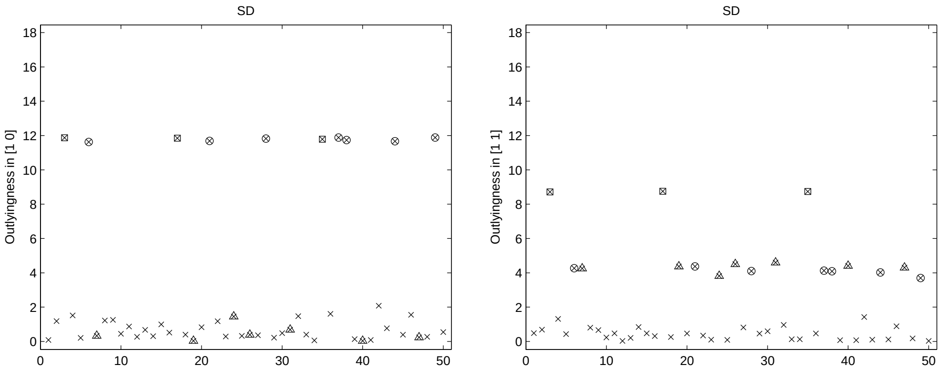
Figure 2: Plot of (a) the SD outlyingnesses and (b) the corresponding weights of the observations in Figure 1.

so we used 10000 random directions. This is most likely much more than needed, but we want to obtain a very good approximation to the SD outlyingnesses.

The outlyingnesses and the corresponding weights of the observations are shown in Figures 2(a) and 2(b) respectively. The majority of observations receive a low outlyingness r_i and hence their weight w_i is close to 1. The large outlyingnesses correspond to the outliers in the data set and lead to weights w_i that are approximately zero. The intermediate weights shown in Figure 2(b), stem from the most remote regular observations.

For more insight in the computation of the outlyingnesses r_i, we look at the directional outlyingnesses $\frac{|x_i'a-\mu(\mathbf{X}a)|}{\sigma(\mathbf{X}a)}$ $(i = 1, \ldots, n)$ for some specific directions $a \in S_p$. In Figure 3(a) we depict the outlyingness for $a = (1, 0)'$. Clearly, the largest values correspond to observations indicated with a square or a circle. Indeed, only these observations have a contaminated first component and hence they are more aberrant than the other observations in this direction. Figure 3(b) shows the results for direction $a = (1, 1)'$. The squared observations now have the largest outlyingness as both of their components are contaminated. The circles and triangles have a somewhat smaller outlyingness here because only one of their components is contaminated. Finally, the majority of the observations have small outlyingness, because both components were non-contaminated.

It can be seen from Figures 3(a) and 3(b) that the outlyingnesses of the squared points in direction (1,1) are lower than the outlyingnesses in direction (1,0), even though their distance to the uncontaminated points is clearly maximized in a projection close to (1,1) (as can be seen from Figure 1). The large number of other outliers, mainly the circles and triangles, have affected the projection in direction (1,1) and somewhat masked the outlyingness of the squares. In this example the impact of this masking effect is small, since all contaminated observations obtained sufficiently high outlyingnesses in other directions than



(a) (b)

Figure 3: Simulated data in Figure 1: plot of the directional outlyingnesses for (a) direction (1,0) and (b) direction (1,1).

(1,1), and all non-contaminated points have a relatively small SD outlyingness.

Our next example illustrates a more severe case of masking. Moreover, this example also illustrates that a swamping effect can occur. That is, some regular observations receive a large outlyingness and therefore are incorrectly downweighted. Figure 4 again shows 50 bivariate standard normal observations (using the same symbols as before), but now the amount of componentwise outliers in both components is increased to 40%. The corresponding SD outlyingnesses and weights of the observations are shown in Figures 5(a) and 5(b) respectively. Clearly, the SD does not succeed in identifying the outliers. Indeed, while the outliers do receive a fairly low weight, the same holds for the non-contaminated points and hence the estimator of the mean and covariance will be severely affected by the outliers.

Figures 6(a) and 6(b) show the directional outlyingnesses in directions (1,0) and (1,1) respectively. In direction (1,0), the squared and circled observations have the largest outlyingness, as expected. In direction (1,1) however, the uncontaminated observations are among the points with highest outlyingness. In fact, the outlyingness of these observations is such that they will be considered as outlying by the SD, an effect known as *swamping*. This swamping effect is caused by the large amount of outliers. Indeed, from Figure 4 it can be seen that the observations indicated with a triangle or circle constitute the bulk of the projected data in direction (1,1) on which the median and MAD will be based. As a result, the SD is severely affected by the outliers in the data.

4. Stahel-Donoho estimator with huberized outlyingness

To avoid the masking and swamping effects explained in the previous section, we propose an adaptation of the SD outlyingness by first huberizing the

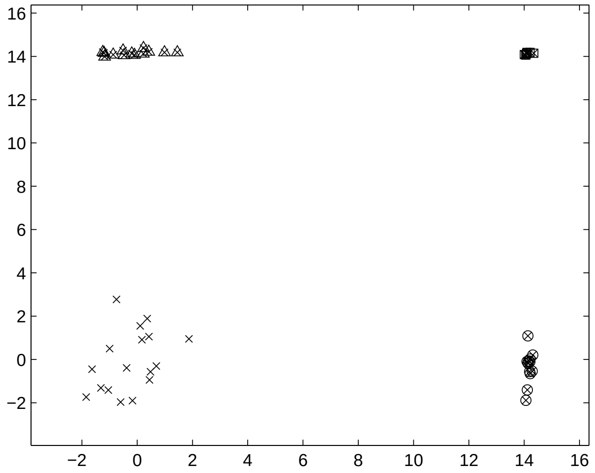


Figure 4: Data set of size $n = 50$ from a two-dimensional standard normal distribution with 40% componentwise outliers independently in both components.

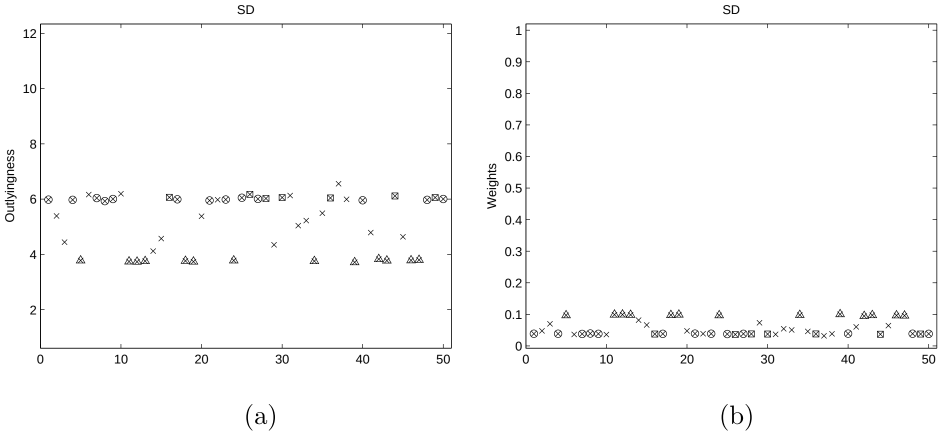
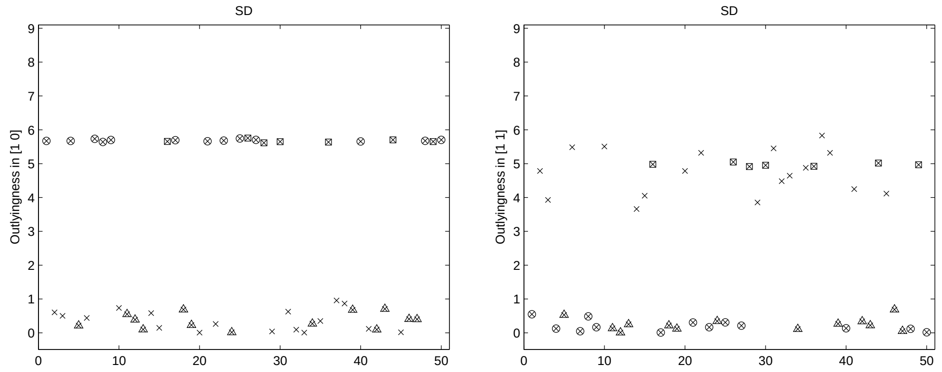


Figure 5: Plot of (a) the SD outlyingnesses and (b) the corresponding weights of the observations in Figure 4.

data before the SD outlyingness is computed. These adapted outlyingness measures should yield a better approximation of what could be perceived as the outlyingness of an observation. The calculation of the huberized outlyingness of an observation x_i can be summarized as follows:

(i) Huberize the data to obtain the modified data matrix $\mathbf{X}_H$ (i.e. component-



(a) (b)

Figure 6: Simulated data in Figure 4: plot of the directional outlyingnesses for (a) direction (1,0) and (b) direction (1,1).

wise winsorize the data)

(ii) For each direction a that is considered, compute the corresponding projection of $\mathbf{X}_H$ and the accompanying median and (modified) MAD.

(iii) Compute the outlyingness of each original observation x_i with respect to $\mathbf{X}_H$ (that is, using the medians and MADs obtained in step (ii).

The huberized observations $x_{1,H}, \ldots, x_{n,H}$ of the data matrix $\mathbf{X}_H$ in step (i) 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 j-th component of $x_{i,H}$, $c_{ij} = \frac{x_{ij} - \text{MED}(X_j)}{\text{MAD}(X_j)}$, $\text{MED}(X_j)$ is the median of $X_j = \{x_{1j}, \ldots, x_{nj}\}$ and $\text{MAD}(X_j) = \text{MED}(|X_j - \text{MED}(X_j)|)$. The cutoff parameter c_H determines the amount of huberization. This cutoff is again a trade-off between robustness and efficiency. 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. While small changes in the value of c_H do not have much effect on the resulting outlyingnesses and corresponding weights in our experience, large changes in c_H (e.g. 1 instead of almost 2) have an impact on the properties of the resulting estimator.

For any $y \in I\!R^p$, the huberized Stahel-Donoho outlyingness in step (iii) is now defined as

$$r_H(y, \mathbf{X}) = \sup_{a \in S_p} \frac{|y'a - \mu(\mathbf{X}_H a)|}{\sigma(\mathbf{X}_H a)}. \qquad (5)$$

For each direction a, $\mu(\mathbf{X}_H a)$ is the median of the projected data obtained in step (ii) and $\sigma(\mathbf{X}_H a)$ is the corresponding modified MAD of the projected data as defined in (4). Note that in practice S_p is a finite set of selected directions.

By huberizing the data in step (i), we reduce the effect of the outliers on the location and scale estimates in the projections when searching for the maximal outlyingness of an observation. Consequently, $\mu(\mathbf{X}_H a)$ and $\sigma(\mathbf{X}_H a)$ better reflect the location and scale of the uncontaminated projected data. Note that steps (i) and (ii) need to be performed only once. An alternative would be to huberize all observations except x_i in step (i). However, this would require steps (i) and (ii) to be repeated for each observation x_i and this may become a computational burden. Both alternatives yield very similar results.

After computing the huberized Stahel-Donoho outlyingnesses $r_H(x_i, \mathbf{X})$, denoted by $r_{i,H}$ in the rest of the paper, the huberized Stahel-Donoho (HSD) estimator of location and scatter $(T_{HSD}, \mathbf{S}_{HSD})$ is defined as

$$T_{HSD} = \frac{\sum_{i=1}^n w_{i,H} x_i}{\sum_{i=1}^n w_{i,H}},$$

and

$$\mathbf{S}_{HSD} = \frac{\sum_{i=1}^n w_{i,H}\ (x_i - T_{HSD})(x_i - T_{HSD})'}{\sum_{i=1}^n w_{i,H}},$$

where $w_{i,H} = w(r_{i,H})$ and w is the Huber-type weight function in (3) as before.

To illustrate the effect of huberization, we focus again on the examples from the previous section. In Figure 7, the data set from Figure 1 is shown after huberizing the observations. Clearly, all outliers (squares, circles and triangles) have been pulled back, componentwise, to the bulk of the data. This can be very

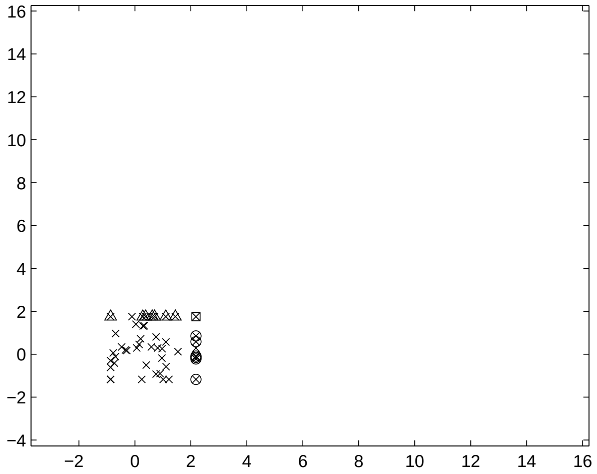
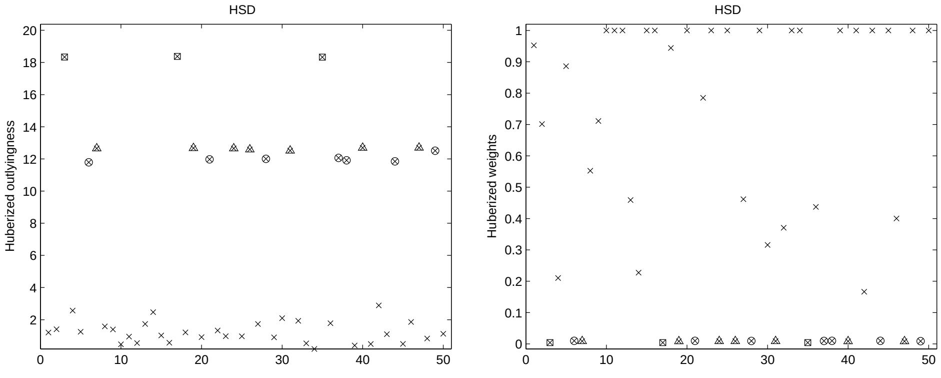


Figure 7: Plot of the huberized data corresponding to the data in Figure 1.

helpful when computing a measure of outlyingness. Indeed, the huberized SD outlyingness replaces the original data set $\mathbf{X}$ by the huberized data set $\mathbf{X}_H$ when computing the univariate measures of location and scale in each of the directions considered. By doing so, the influence of the outliers is reduced. Consequently, the huberized outlyingness much better reflect how outlying a given observation is with respect to the noncontaminated data. This is confirmed by Figures 8(a) and 8(b) where the HSD outlyingnesses and weights are shown respectively. Clearly, the outlying observations correspond to large HSD outlyingnesses and hence small weights. This indicates that HSD succeeds in detecting the outliers. As shown in Figures 2(a) and 2(b), SD also succeeded in identifying the outliers in the data, but Figures 3(a) and 3(b) indicated that some SD outlyingnesses were not as large as expected due to masking in directions close to (1,1).



(a)

(b)

Figure 8: Based on the simulated data in Figure 1: Plot of (a) the HSD outlyingnesses and (b) the corresponding HSD weights of the observations.

Figures 9(a) and 9(b) show the directional outlyingnesses $\frac{|x_i' a - \mu(\mathbf{X}_H a)|}{\sigma(\mathbf{X}_H a)}$ in direction $a = (1,0)'$ and $a = (1,1)'$ respectively. First note that Figure 9(a) is very similar to Figure 3(a). That is, the outlyingnesses in direction (1,0) are largely unaffected by the huberization. On the other hand, in Figure 9(b) we see that the squared observations, which are outlying in both components, now have very high HSD outlyingness in direction (1,1), which is in fact higher than in direction (1,0). As HSD makes use of the huberized data for the computation of μ and σ in each direction, its directional outlyingnesses were less influenced by outliers. This result is what would be expected and hence, we can say that the squared observations now receive the outlyingness they “deserve”.

Now, let us apply the HSD to the highly contaminated data set in Figure 4. By huberizing the data, the outliers have been pulled back towards the center of the data as shown in Figure 10. The HSD outlyingnesses and weights are shown in Figures 11(a) and 11(b). As opposed to the SD estimator, HSD succeeds in

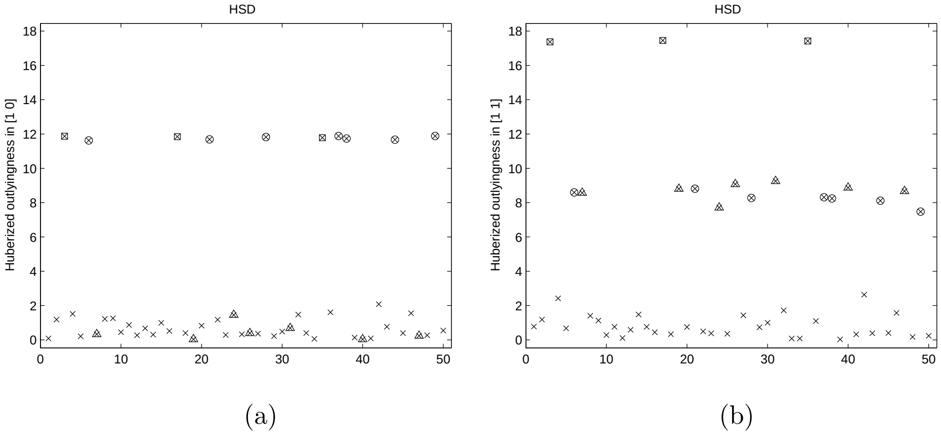


Figure 9: Based on the simulated data in Figure 1: plot of the directional HSD outlyingnesses for (a) direction (1,0) and (b) direction (1,1).

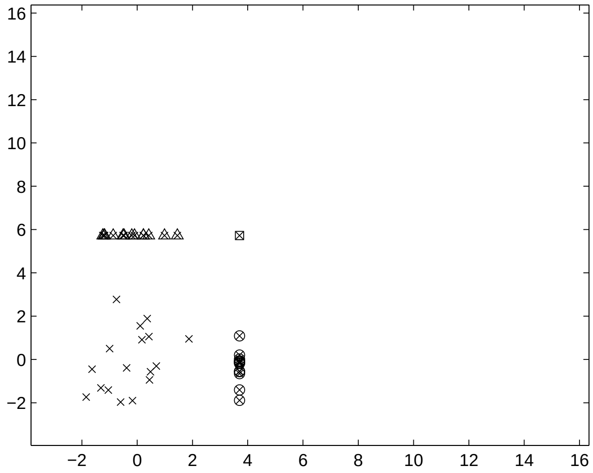
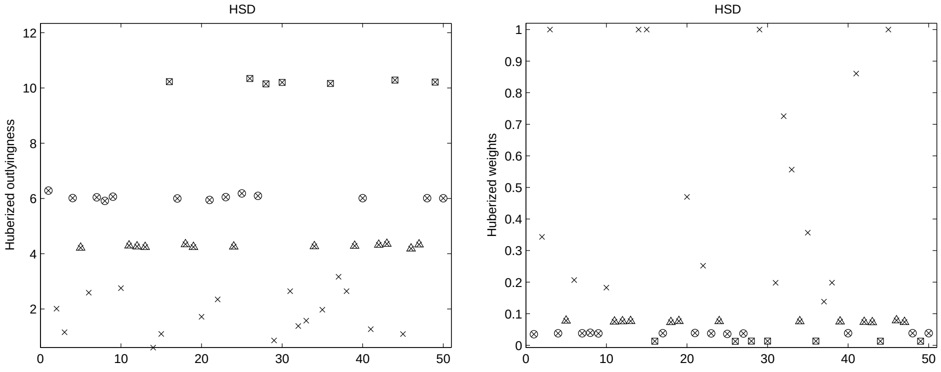


Figure 10: Plot of the huberized data corresponding to the data in Figure 4.

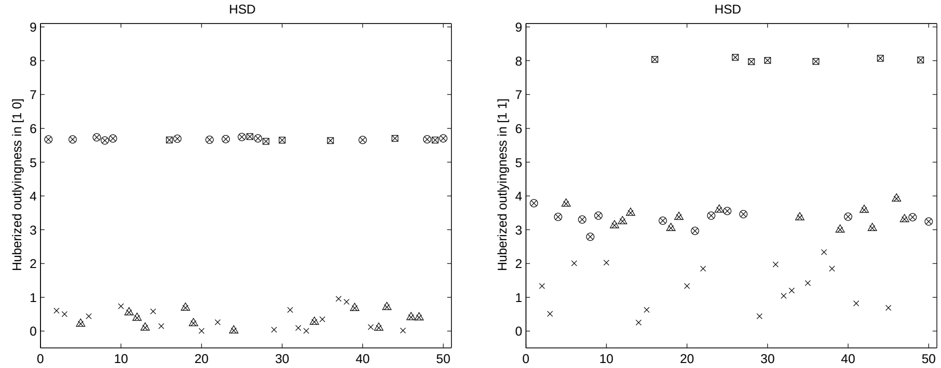
detecting the outliers in the data set. Every contaminated observation has a large HSD outlyingness $r_{i,H}$ and a small weight $w_{i,H}$. Furthermore, the non-contaminated points did not obtain a high HSD outlyingness, which means that the swamping effect has disappeared.

Figure 12(a) shows that the HSD outlyingnesses in direction (1,0) are naturally very similar to those in Figure 6(a). From Figures 12(b) we see that in



(a) (b)

Figure 11: Based on the simulated data in Figure 4: Plot of (a) the HSD outlyingnesses and (b) the corresponding HSD weights of the observations.



(a) (b)

Figure 12: Based on the simulated data in Figure 4: plot of the directional HSD outlyingnesses for (a) direction (1,0) and (b) direction (1,1).

direction (1,1) the squared observations have the largest HSD outlyingness, in accordance to the fact that they are outlying in both components. Note that there is only a slight difference between the HSD outlyingnesses of the remaining outliers and the uncontaminated points (as opposed to Figure 9(b)). This is due to the large amount of outliers in the data (66% of the observations are contaminated). Hence, in direction $a = (1, 1)'$ the majority of points is contaminated, which leads to the median of the projected points being shifted upwards. This results in a lower HSD outlyingness for the circled and triangled observations in this direction. However, this is not a problem because these points achieve their actual outlyingness in other projections.

5. Simulation study

In this section we investigate through simulation the effect of our adaptation of the outlyingness measure on the precision and the robustness of the corresponding HSD estimator. Here, *precision* is meant to represent the variance of the estimators, while by *robustness* we mainly refer to the bias. In particular, we compare the mean squared errors (MSEs) of the SD and HSD estimators. The componentwise huberization that is used to calculate the adjusted outlyingness does not take correlation among variables into account. Therefore, we also investigate how effective the HSD is for correlated data.

We generated correlated normal data as in [19]. That is, samples $\mathcal{X} = \{x_1, \ldots, x_n\}$ were generated from a p-variate normal distribution with mean zero and covariance matrix $\mathbf{R}^2$ where $\mathbf{R}$ is a matrix with elements $\mathbf{R}_{jj} = 1$ $(j = 1, \ldots, p)$ and $\mathbf{R}_{kl} = \rho$ $(1 \leq k \neq l \leq p)$. Following [19] we considered values of ρ 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. In this paper, we report the results for data with $R^2 = 0$ and $R^2 = 0.9$. These two cases are quite representative and show well the effect of correlation on the performance of the estimators. For the dimension p we considered the values $5, 7$ and 10, and the sample size was $n = 50$ (for $p = 5$) and $n = 100$ (for $p = 7, 10$). Subsequently, in the first d components $(d \leq p)$, we independently introduced a fraction ϵ of univariate outliers. We consider the cases $d = 2\,(p = 5), d = 5\,(p = 7, 10)$, and $d = 7\,(p = 7)$. For each contaminated component, the outlying values were generated from a univariate normal distribution with mean $k/\sqrt{d}$ and standard deviation 0.1. For several combinations of ϵ and d in the simulations, the fraction of outlying observations can exceed the breakdown point (50%) of the SD, so we expect that the SD has a large MSE in these cases. The main purpose of these simulation settings is to see to what extent the HSD can withstand these amounts of contamination and thus avoids the adverse effect on the SD.

We considered outlying distances $k = 6, 24, 64$ and 160. For each situation, $N = 500$ samples were generated. Then, for each sample $\mathcal{X}^{(l)}; l = 1, \ldots, N$ and for each observation $x_i^{(l)}$ in $\mathcal{X}^{(l)}$, we computed the SD outlyingness $r_i^{(l)}$ and the HSD outlyingness $r_{i,H}^{(l)}$, and subsequently the corresponding location and scatter estimates $(T_{SD}^{(l)}, \mathbf{S}_{SD}^{(l)})$ and $(T_{HSD}^{(l)}, \mathbf{S}_{HSD}^{(l)})$. The MSE for the location estimators of both methods was calculated as

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

For both methods 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. The number of random directions in the SD/HSD algorithms was set equal to $200p$ which corresponds with the choice of [19] for data sets in higher dimensions.

			$\rho = 0$	$\rho = 0$	$\rho = 0.9$	$\rho = 0.9$
	ϵ	Comp	$k = 6$	$k = 64$	$k = 6$	$k = 64$
center	0.20	All	0.85	1.09	1.04	1.07
center	0.20	Cont	0.82	1.29	1.06	1.10
center	0.35	All	0.90	0.02	1.07	0.03
center	0.35	Cont	0.90	0.02	1.11	0.02
diag	0.20	All	0.66	0.70	1.16	1.21
diag	0.20	Cont	0.63	0.68	1.14	1.14
diag	0.35	All	0.91	0.02	1.22	0.01
diag	0.35	Cont	0.91	0.02	1.22	0.01
offdiag	0.20	All	0.90	0.61	1.09	1.19
offdiag	0.20	1 cont	0.92	1.00	1.08	1.18
offdiag	0.20	2 cont	0.81	0.38	1.04	1.15
offdiag	0.35	all	0.95	0.01	1.14	0.02
offdiag	0.35	1 cont	0.99	0.07	1.13	0.30
offdiag	0.35	2 cont	0.92	0.01	1.17	0.01

Table 1: MSE ratios of HSD vs SD for data in 5 dimensions with $\epsilon = 20\%$ or $\epsilon = 35\%$ of independent contamination in the first two components for $k = 6$ or $k = 64$. 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 first consider the case $p = 5$ and $d = 2$. The fraction of independent contamination in each of the first two components was taken equal to $\epsilon = 20\%$ and $\epsilon = 35\%$. In Table 1 we show the MSE ratio $\text{MSE}(T_{HSD})/\text{MSE}(T_{SD})$ for the location and similar ratios for the diagonal and off-diagonal elements of the scatter matrix. Table 1 contains the overall MSE ratio for the various settings when all components are taken into account, as well as the MSE ratio when only the contaminated components are taken into account when calculating the MSE. The latter provides information about the difference in bias between the two estimators 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).

From the results in Table 1 we can see that for $\epsilon = 20\%$ the HSD is generally somewhat worse than the SD in case of (highly) correlated data. For uncorrelated data the HSD often yields a small improvement over the MSE of the SD, especially for small k. However, in general the difference between the estimators is relatively small here as can be seen from the top panel of Figure 13. In this figure we show boxplots of the absolute errors of the estimates for the com-

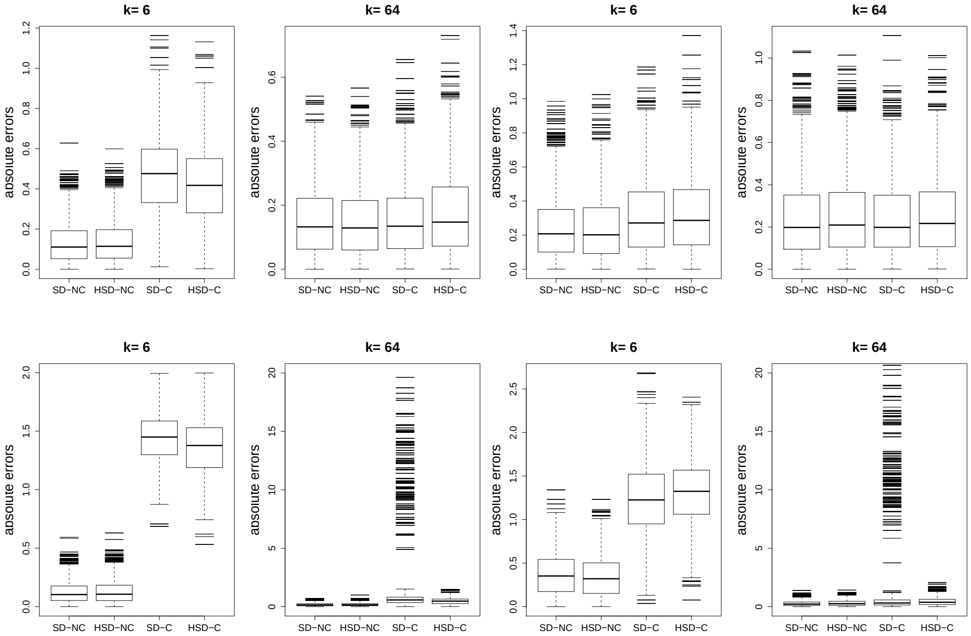


Figure 13: Boxplots of absolute errors of the estimates of the uncontaminated and contaminated components of the center. Data were generated in 5 dimensions with a fraction ϵ of independent contamination in the first two components for different values of k. The top panels correspond to $\epsilon = 20\%$ and the bottom panels correspond to $\epsilon = 35\%$. The left plots shows the results for uncorrelated data and the right plots contain the results for correlated data ($R^2 = 0.9$).

ponents of the center. Separate boxplots are shown for the contaminated and uncontaminated components. The figures for the elements of the scatter matrix are similar and therefore not shown. The top panel of Figure 13 corresponds to $\epsilon = 20\%$. From these plots we see that the absolute errors for both SD and HSD are small. Further examination of the results has shown that both SD and HSD succeed well in identifying the contaminated observations, by assigning a large outlyingness measure to them.

The results in Table 1 and the bottom panel of Figure 13 reveal the effect on the estimators when the fraction of contamination is increased to 35%. When the contamination is close by ($k = 6$) the effect on the SD remains small as can be seen from Figure 13. In this case the huberization only yields a small improvement for uncorrelated data, but has a small adverse effect for highly correlated data. When the contamination lies further away from the bulk of the data, its effect on the SD becomes much larger. The reason is that due to the large amount of componentwise contamination in some samples, the SD suffers from the swamping effect illustrated in Figure 5, i.e. regular observations receive an outlyingness r_i that is far too large, whereas the SD outlyingness of the outliers is underestimated. This effect does not occur in all samples, but it strongly affects the MSE of the SD. In these situations, the HSD succeeds in reducing the effect of the componentwise contamination. Due to the huberization, the HSD outlyingnesses do not suffer from the swamping problem and hence the outliers are recognized and receive a large $r_{i,H}$ whereas regular observations receive a small $r_{i,H}$. Therefore, the bias on the contaminated components is much smaller for the HSD as shown in Figure 13 which results in a much lower MSE as can be seen in Table 1. These results indicate that the HSD can potentially cope with larger amounts of (componentwise) contamination than the original SD.

To illustrate this, let us now increase the dimension to $p = 7$ and introduce 30% of independent contamination in the first five components. This is a severe case with data that contain only a minority of observations that are contamination-free. The results in Table 2 show that HSD considerably improves the MSE of the SD estimates as soon as the contamination is far enough from the majority of the data, both for uncorrelated and correlated data. The bias in the contaminated components contributes most to the MSE of the SD and the results of Table 2 show that HSD yields a large bias reduction at the contaminated components. Indeed, as explained in the previous sections, the SD outlyingnesses may be severely influenced by the outliers in this case. The HSD on the other hand, can better cope with these outliers and attributes large outlyingnesses to them. A more detailed investigation of all the simulation results revealed that with increasing outlier distance the MSE of the HSD decreases most quickly for uncorrelated data. This is illustrated further by the next case.

		$\rho = 0$	$\rho = 0$	$\rho = 0.9$	$\rho = 0.9$
	Comp	$k = 6$	$k = 64$	$k = 6$	$k = 64$
center	All	0.99	0.10	0.99	0.15
center	Cont	0.99	0.10	0.99	0.14
diag	All	0.98	0.11	1.13	0.11
diag	Cont	0.97	0.11	1.09	0.11
offdiag	All	1.00	0.10	1.00	0.21
offdiag	1 cont	0.99	0.24	1.03	0.76
offdiag	2 cont	1.01	0.10	0.97	0.20

Table 2: MSE ratios of HSD vs SD for data in 7 dimensions with $\epsilon = 30\%$ of independent contamination in the first five components for $k = 6$ or $k = 64$. 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 further to $p = 10$ with $\epsilon = 10\%$ and $\epsilon = 20\%$ of independent contamination in the first five components. For $\epsilon = 10\%$ the results are similar to those for the case $p = 5$ with 20% of independent

	ϵ	Comp	$\rho = 0$ $k = 6$	$\rho = 0$ $k = 64$	$\rho = 0.9$ $k = 6$	$\rho = 0.9$ $k = 64$
center	0.10	All	0.90	1.03	0.89	0.99
center	0.10	Cont	0.87	1.03	0.89	1.00
center	0.20	All	0.93	0.53	0.92	1.05
center	0.20	Cont	0.92	0.47	0.93	1.07
diag	0.10	All	0.76	0.58	1.17	0.85
diag	0.10	Cont	0.67	0.58	1.09	0.59
diag	0.20	All	0.87	0.32	1.15	0.86
diag	0.20	Cont	0.85	0.32	1.05	0.85
offdiag	0.10	All	0.94	0.65	1.12	1.26
offdiag	0.10	1 cont	0.95	0.88	1.12	1.26
offdiag	0.10	2 cont	0.91	0.48	1.05	1.22
offdiag	0.20	all	0.97	0.16	1.06	1.00
offdiag	0.20	1 cont	0.97	0.69	1.07	1.11
offdiag	0.20	2 cont	0.98	0.15	1.00	0.76

Table 3: MSE ratios of HSD vs SD for data in 10 dimensions with $\epsilon = 10\%$ or $\epsilon = 20\%$ of independent contamination in the first five components for $k = 6$ or $k = 64$. 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).

contamination in the first two components. In this setting there is a majority of contamination-free observations. Therefore, the SD has a small MSE and the HSD yields similar results. For uncorrelated data, the HSD often gives small improvements over the SD, but for highly correlated data the effect reverses.

For $\epsilon = 20\%$ there is no majority of contamination free observations anymore. The MSE of the SD increases due to bias problems. As can be seen from Table 3, for uncorrelated data the HSD yields an improvement already for close by outliers with increasing effect if the distance of the outliers increases. For highly correlated data on the other hand, the outliers need to be much further away before the HSD can improve on the SD. Table 3 shows that for $k = 64$ the HSD still cannot improve the MSE of the SD.

We repeated the simulation study for large data sets with $n = 5000$ (for $p = 5$) and $n = 10000$ (for $p = 7, 10$). The results were similar and are omitted here.

6. Conclusion

We presented a huberized version of the SD outlyingness where the outlyingness of the observations is calculated w.r.t. the huberized data set in which outliers are componentwise pulled back to the bulk of the data. The huberization clearly improves the outlyingness measure in settings with independent componentwise contamination. Contamination models that include componentwise outliers are realistic for many high dimensional settings. Such models are not affine equivariant anymore and the overall fraction of contamination can easily exceed 50% in higher dimensions. It was shown that in such cases the HSD suffers less from masking and swamping effects. Hence, HSD can better withstand the outliers as shown in a simulation study. The improvement of HSD over SD is largest for data that are uncorrelated or weakly correlated. For highly correlated data, the contamination needs to lie further from the bulk of the measurements before the HSD can improve on the SD. In some cases the increased variability of HSD makes its MSE worse than for SD, even though the HSD has a smaller bias due to the outliers. A further improvement of the procedure would be desirable to avoid that its performance is worse than that of the SD in such cases. Adapted weight functions for the huberized outlyingnesses may be of interest for this matter, but this requires further research.

As an alternative to our huberization approach, one could consider applying a univariate outlier detection rule to each of the variables separately to remove componentwise outliers from the data. A multivariate outlier identification rule can then be applied to the cleaned data to detect structural outliers. A drawback of such an approach is that the removal of the componentwise outliers in the first step creates several empty cells in the data matrix. For the multivariate outlier detection robust estimates are needed, but multivariate robust estimation procedures in general cannot easily handle data with empty cells. Restricting to the observations without empty cells is often not possible either because the number of complete observations may have become to low (lower than the dimension). Huberization on the other hand does not create empty cells, but modifies the outlying values to make them more regular which allows a more robust multivariate estimation procedure in the second step.

Acknowledgment

Research of the first author is supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen) and by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

References

[1] Alqallaf, F.A., Konis, K.P., Martin, R.D., Zamar, R.H., 2002. Scalable Robust Covariance and Correlation Estimates for Data Mining, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, 14-23.

[2] Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H., 2009. Propagation of Outliers in Multivariate Data, Ann. Statist. 37, 311-331.

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

[4] Cerioli A., Farcomeni, A., 2011. Error rates for multivariate outlier detection, Computat. Statist. Data Anal. 55, 544-553.

[5] Davies, P.L., 1987. Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices, Ann. Stat. 15, 1269-1292.

[6] Debruyne M. 2009. An outlier map for support vector machine classification, Ann. Applied Statist., 3, 1566-1580.

[7] Debruyne M., Hubert, M., 2009. The influence function of the Stahel-Donoho covariance estimator of smallest outlyingness, Statist. Probab. Lett. 79, 275-282.

[8] Donoho, D.L., 1982. Breakdown Properties of Multivariate Location Estimators, Ph.D. diss., Harvard University.

[9] Filzmoser, P., Maronna, R., Werner, M., 2008. Outlier identification in high dimensions, Computat. Statist. Data Anal. 52, 1694-1711.

[10] Gather, U., Hilker, T., 1997. A Note on Tyler's Modification of the MAD for the Stahel-Donoho Estimator, Ann. Statist. 25, 2024-2026.

[11] Gervini, D., 2002. The influence function of the Stahel-Donoho estimator of multivariate location and scatter, Statist. Probab. Lett. 60, 425-435.

[12] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions, John Wiley and Sons, New York.

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

[14] Hubert M., Verboven, S., 2003. A robust PCR method for high-dimensional regressors, J. Chemom. 17, 438-452.

[15] Hubert, M., Rousseeuw, P.J., Vanden Branden, K., 2005. ROBPCA: a new approach to robust principal component analysis, Technometrics 47, 64-79.

[16] Khan, J.A., Van Aelst, S., Zamar, R.H., 2007. Robust linear model selection based on least angle regression, J. Amer. Statist. Assoc. 102, 1289-1299.

[17] Maronna, R.A., Martin, D.R., Yohai, V.J., 2006. Robust Statistics: Theory and Methods, Wiley, New York.

[18] Maronna, R.A., Yohai, V.J., 1995. The behavior of the Stahel-Donoho Robust Multivariate Estimator, J. Amer. Statist. Assoc. 90, 329-341.

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

[20] Rousseeuw, P.J., 1984. Least median of squares regression, J. Amer. Statist. Assoc. 79, 871-880.

[21] Stahel, W.A., 1981. Breakdown of Covariance Estimators, Research Report 31, Fachgruppe für Statistik, E.T.H. Zürich, Switzerland.

[22] Van Aelst, S., Vandervieren, E., Willems, G., 2011. Stahel-Donoho Estimators with Cellwise Weights, J. Stat. Comput. Simul. 81, 1-27.

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

[24] Zuo, Y., Laia, S., 2011. Exact computation of bivariate projection depth and the StahelDonoho estimator, Computat. Statist. Data Anal. 55, 1173-1179.