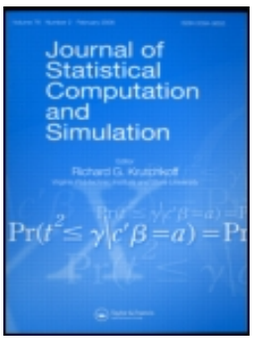

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Stahel-Donoho estimators with cellwise weights

S. Van Aelst , E. Vandervieren & G. Willems

To cite this article: S. Van Aelst , E. Vandervieren & G. Willems (2011) Stahel-Donoho estimators with cellwise weights, Journal of Statistical Computation and Simulation, 81:1, 1-27, DOI: 10.1080/00949650903103873

To link to this article: https://doi.org/10.1080/00949650903103873

Published online: 09 Dec 2009.

Journal of Statistical Computation and Simulation
Vol. 81, No. 1, January 2011, 1–27

Taylor & Francis
Taylor & Francis Group

Stahel-Donoho estimators with cellwise weights

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

(*Received 16 September 2008; final version received 9 June 2009*)

The Stahel–Donoho estimator is defined as a weighted mean and covariance, where each observation receives a weight which depends on a measure of its outlyingness. Therefore, all variables are treated in the same way whether they are responsible for the outlyingness or not. We present an adaptation of the Stahel–Donoho estimator, where we allow separate weights for each variable. By using cellwise weights, we aim to only downweight the contaminated variables such that we avoid losing the information contained in the other variables. The goal is to increase the precision and possibly the robustness, of the estimator. We compare several variants of our proposal and show to what extent they succeed in identifying and downweighting precisely those variables which are contaminated. We further demonstrate that in many situations the mean-squared error of the adapted estimators is lower than that of the original Stahel–Donoho estimator and that this results in better outlier detection capabilities. We also consider some real data examples.

Keywords: robust multivariate estimators; contamination models; outlier identification

1. Introduction

The Stahel–Donoho estimator, proposed independently in [1,2], is a well-known robust estimator of multivariate location and scatter. It was the first affine equivariant estimator with a breakdown point (i.e., the maximum proportion of outliers that the estimator can withstand) close to 50% for any dimension. It has some excellent robustness properties as shown in [3–5], and is in this respect comparable to other popular high-breakdown estimators such as the minimum covariance determinant (MCD) estimator [6] or S-estimators [7].

The Stahel–Donoho estimator is defined as a weighted mean and covariance, where each observation receives a weight which depends on a measure of its 'outlyingness'. This measure is based on the one-dimensional projection in which the observation is most outlying, the underlying idea of which is that every multivariate outlier must be a univariate outlier in some projection. Hence, observations with large outlyingness then receive small weights. Recent applications of the Stahel–Donoho outlyingness measure can be found in [8–10].

*Corresponding author. Email: stefan.vanaelst@ugent.be

ISSN 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/00949650903103873
http://www.informaworld.com

Whether a large outlyingness is due to an aberrant value in one or more specific variables (*componentwise* outliers) or to a deviating covariance structure involving several variables (*structural* outliers), there is no difference in how the weighting scheme treats the outlying point. That is, the entire observation is either downweighted or not. Or still, all components of an observation are treated in the same way whether they are 'responsible' for its outlyingness or not. This kind of weighting is of course intrinsic to all affine equivariant robust estimators. Indeed, once we start to treat the components of an observation differently, we have to give up equivariance. However, by making this sacrifice, we can make a distinction between the contaminated and non-contaminated components of an observation. It thus allows us to downweight the contaminated components only, such that we avoid losing the information contained in the other components. We can then obtain an estimator with increased precision.

The Stahel–Donoho estimator, like other high-breakdown estimators, has been studied primarily in the context of the Tukey–Huber contamination model [11,12]. This model assumes that, on average, a large fraction $(1-\varepsilon)$ of the data (e.g. $0 \leq \varepsilon < 0.25$) is generated from a classical model, whereas the remaining data can be affected by abnormal noise. In other words, the data come from a mixture distribution with a fully described dominant component (e.g. a Gaussian random variable) and an unspecified minority component. The general idea of robust procedures in this context is to conduct inference on the dominant component of the mixture, by limiting the influence of observations that resulted from the other component. Identifying and subsequently downweighting such 'harmful' observations makes perfect sense for this purpose.

The Tukey–Huber model has some limitations, especially in high dimensions. The main criticisms are that it assumes that a majority of the points are perfectly free of contamination, and that downweighting entire points may be inefficient. In [13], the Tukey–Huber model was extended to a large family of contamination models. One of these models, the independent contamination model, assumes that contamination in each variable is independent of the other variables, leading in particular to componentwise outliers. Furthermore, the models now merely assume that there is a majority of outlier-free *values* in each variable, but not necessarily a majority of outlier-free *observations*. Affine equivariant robust estimators do not perform well under such models because they require a majority of clean observations. The Stahel–Donoho estimator may seem somewhat better protected against a possible majority of outlying points than estimators such as MCD or S-estimators, as it is in principle allowed to downweight more than half of the points. Nevertheless, the estimator is surely not particularly suitable for these models since it cannot treat the variables separately. Downweighting a large number of points may lead to a considerable waste of information, especially in high dimensions, and may cause severe instability.

In this paper, we investigate an adaptation of the Stahel–Donoho estimator where we allow separate weights for each component of an observation. The idea is to start from the outlyingness of the observation as measured in the original Stahel–Donoho procedure. Subsequently, for each observation, we attempt to identify to what extent each variable contributes to the outlyingness and we use this information to adjust the original Stahel–Donoho weights in a cellwise manner. In particular, whenever an observation has a considerable outlyingness and hence a small Stahel–Donoho weight, the 'clean' components should be restored to some extent by adjusting the corresponding weight upwards. By adapting the estimator in this way (using cellwise weights), we give up affine equivariance but we can gain precision. Moreover, the estimator becomes more suitable for use in the context of the larger family of contamination models considered in [13]. For instance, in the case of independent contamination, the adaptation not only boosts the precision, but may also increase the robustness.

The rest of the paper is organized as follows. In Section 2, we discuss the Stahel–Donoho estimator and focus on a real data set. In Section 3, we present our proposal for adapting the estimator in a componentwise manner. A simulation study is performed in Section 4, in which we compare several variants of our proposal and see to what extent they succeed in restoring the

weights of the clean components of an observation while leaving the weights of the contaminated components unaffected. Section 5 investigates, through a second simulation study, specifically how the cellwise weights affect the precision and robustness of the estimator. In Section 6, we then examine how our adapted estimators perform in the context of outlier detection based on robust distances. We continue with real data examples in Section 7, while Section 8 concludes.

2. Stahel–Donoho estimator

Suppose $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ is a set of n observations. Let μ be a shift and scale equivariant univariate location statistic and let σ be a shift invariant and scale equivariant univariate scale statistic. Then, for any $y \in \mathbb{R}^p$, the Stahel–Donoho *outlyingness* is defined as

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

with $S_p = \{a \in \mathbb{R}^p : ||a|| = 1\}$. This outlyingness measure is based on the idea that for any multivariate outlier, one can always find a one-dimensional projection for which the observation is a univariate outlier.

The Stahel–Donoho estimator of location and scatter (T_{SD}, S_{SD}) is defined as

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

and

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

where $w_i = w(r(x_i, X))$ and $w : \mathbb{R}^+ \to \mathbb{R}^+$ is a weight function so that observations with large outlyingness get small weights [1,2].

For w, we use the Huber-type weight function as advocated in [3]. It is defined as

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

where c is a threshold which we choose here as $c = \min(\sqrt{\chi^2_p}(0.50), 4)$.

In order to attain maximum breakdown [3,14], the univariate location statistic μ is set equal to the median (MED) and the scale statistic σ is chosen to be the modified MAD, defined as

$$\mathrm{MAD}^*(a'X) = \frac{|a'X - \mathrm{MED}(a'X)|_{\lceil (n+p-1)/2 \rceil : n} + |a'X - \mathrm{MED}(a'X)|_{(\lfloor (n+p-1)/2 \rfloor + 1) : n}}{2\,\beta}, \tag{3}$$

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 $x_{i:n}$ denotes the ith order statistic of the data set.

As an example, we consider the Philips data [15]. For each of 677 diaphragm parts for TV sets, nine characteristics were measured at the beginning of a new production line. In Figure 1, the Stahel–Donoho outlyingnesses of the Philips data are shown. It can be seen that observations 491–565 get a very large outlyingness and hence seem to be strongly deviating from the majority of the observations. Consequently, these aberrant observations will get a small weight in the computation of the Stahel–Donoho estimator, so that they influence the location and scatter estimate in a limited way. By giving a small weight to an outlying observation, every component of this observation will be downweighted.

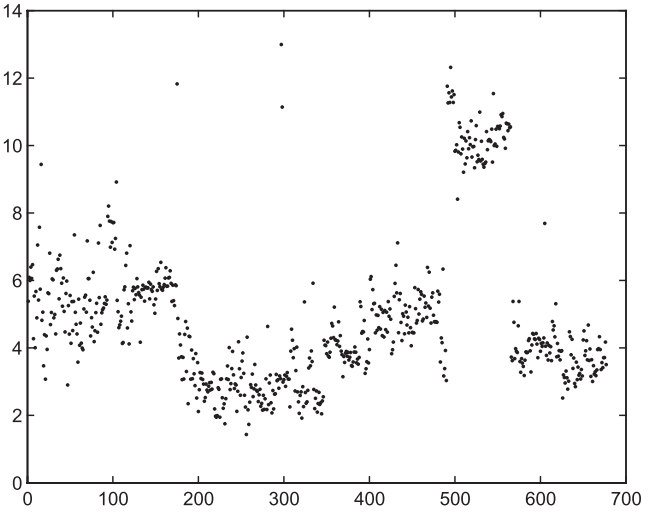

Figure 1. Stahel–Donoho outlyingnesses of the Philips data.

In the Philips data, observations 491–565, among other points, are assigned a small weight in the Stahel–Donoho estimator. To assess whether it is truly required to downweight each component of these observations, let us first look at their outlyingness within each individual component. Figure 2 shows univariate scatter plots for each of the nine components (V1, …, V9), standardized by their median and modified MAD. Some random jitter is added on the horizontal axis. Observations 491–565 are marked in dark plus signs. It can be seen that these observations are somewhat outlying within component V2, for example. On the other hand, their values do not seem particularly suspicious within many of the other components, such as V1, V3, V4 or also V6.

Obviously, unsuspected behaviour within individual components does not mean that the information contained in these components is harmless and can safely be used in the estimation of location and scatter of the data. Indeed, a large outlyingness could be due to the combined values of several components, rather than those of individual components. For example, Figure 3 shows scatter plots depicting component V6 versus V5 and V6 versus V9. Observations 491–565 are again marked in dark plus signs. It can clearly be seen that these observations are outlying with respect to the correlation between V6 and V5, as well as that between V6 and V9. We may, therefore, conclude that component V6 bears a considerable responsibility for the large outlyingness of observations 491–565 and hence should be downweighted (even if these observations were not particularly outlying within the individual component V6).

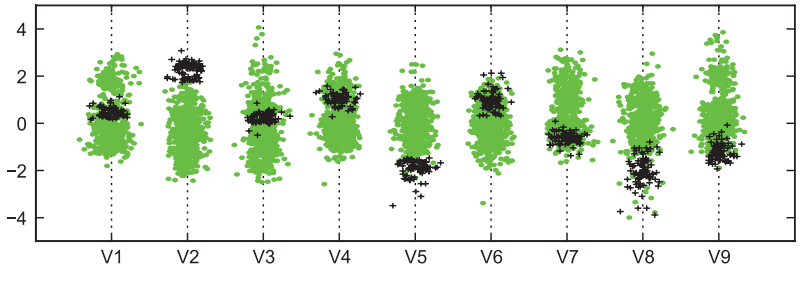

Figure 2. Philips data, componentwise scatter: each component is standardized by its median and modified MAD. Observations 491–565 are marked in dark plus signs.

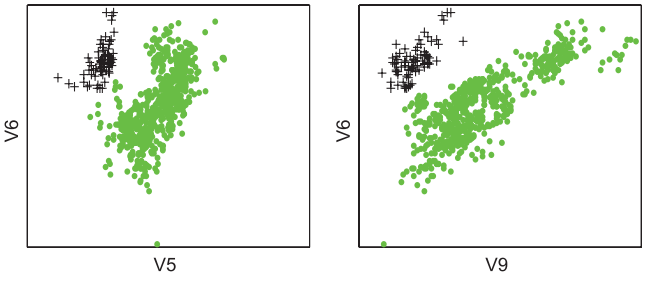

Figure 3. Philips data scatter plots: variables V6 versus V5, and V6 versus V9. Observations 491–565 are marked in dark plus signs.

In general, our proposals in the next section aim to determine, for each observation, exactly which components bear some kind of responsibility for its Stahel–Donoho outlyingness and which components can be considered clean. The responsibility of the component can either be through its individual outlying value or as part of a combination of two or more components. Those components of the observation that are deemed clean, should then be awarded a weight increase.

3. Adapted Stahel–Donoho estimators

As mentioned before, a considerable amount of information could be lost if we only have one scalar weight to control the influence of an observation on the Stahel–Donoho estimator. Therefore, we propose an adaptation that assigns a *vector* of weights to each observation, i.e. every component of an observation receives its own weight.

Suppose $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ is a set of n observations. The adapted Stahel–Donoho estimator of location and scatter $(T_{\mathrm{SD}^*}, S_{\mathrm{SD}^*})$ is defined as

$$T_{\mathrm{SD}^*,j} = \frac{\sum_{i=1}^{n} w_{ij} x_{ij}}{\sum_{i=1}^{n} w_{ij}} \tag{4}$$

and

$$S_{\mathrm{SD}^*,jk} = \frac{\sum_{i=1}^{n} \sqrt{w_{ij}}\sqrt{w_{ik}}(x_{ij} - T_{\mathrm{SD}^*,j})(x_{ik} - T_{\mathrm{SD}^*,k})}{\sum_{i=1}^{n} \sqrt{w_{ij}}\sqrt{w_{ik}}} \tag{5}$$

for $j, k = 1, \ldots, p$.

The weight matrix $W = (w_{ij})_{ij}$ is defined as

$$w_{ij} = w(r_{ij}), \tag{6}$$

where the weight function w is as before and the adapted outlyingness r_{ij} is defined as

$$r_{ij} = \alpha_{ij} r_i + (1 - \alpha_{ij}) c_{ij}, \tag{7}$$

where α_{ij} is a weighting parameter, r_i the Stahel–Donoho outlyingness of x_i (i.e. $r_i = r(x_i, X)$) and c_{ij} the outlyingness of x_i in the direction of component j, given by

$$c_{ij} = \frac{|x_{ij} - \mathrm{MED}(X_j)|}{\mathrm{MAD}^*(X_j)}. \tag{8}$$

Here, $X_j = \{x_{1j}, \ldots, x_{nj}\}$, the set of all values of the jth component of X, MED is the median and MAD* is defined by Equation (3).

The weighting parameter α_{ij} in Equation (7) will be chosen in $[0, 1]$, such that the adapted outlyingness r_{ij} is a weighted average of r_i and c_{ij}. Since the componentwise directions are a subset of the directions considered in Equation (1), we always have that $r_i \geq c_{ij}$ and hence $r_i \geq r_{ij}$. That is, the adapted outlyingness is smaller than the original outlyingness. The idea is to reduce the outlyingness, and hence increase the weight, of those components which had limited influence on the global outlyingness r_i (i.e. those components for which a weight increase is 'justified'), while largely keeping the original outlyingness and weight for the other components.

We could choose α_{ij} as a constant, i.e. $\alpha_{ij} = \alpha$, in which case the reduction in outlyingness associated with r_{ij} is linearly related to c_{ij}, the outlyingness in the jth component. This would make sense if we only suspect componentwise outliers but does not seem appropriate for structural outliers. Indeed, if x_i is such a structural outlier, it may have components with small c_{ij} which nevertheless share a responsibility for the large r_i, and hence should not be awarded any considerable reduction in outlyingness. Now, in order to account for structural outliers as well, α_{ij} would preferably represent the extent to which the jth component contributes to r_i. If this contribution is assessed to be large, r_{ij} should remain close to r_i, otherwise, r_{ij} is allowed to decrease largely toward c_{ij}.

In this paper, we consider the following choices for α_{ij}:

(1) $\alpha_{ij} = \alpha = 1/2$. Hence the adapted outlyingness r_{ij} always equals the average of r_i and c_{ij}. As explained above, a constant weighting parameter may not account appropriately for structural outliers. Another disadvantage is that the choice of $\alpha = 1/2$ is arbitrary and may be far from optimal in some situations. We do not expect this choice of weighting parameter to perform well, but consider it a benchmark.

(2) $\alpha_{ij} = (\max_{k=1}^{p} c_{ik})^{-1} c_{ij}$ with c_{ij} as defined in Equation (8). It follows that α_{ij} is large whenever c_{ij} is large, relative to the outlyingnesses in the direction of the other components. This weighting parameter α_{ij} compares to a constant α as follows: (1) when observation x_i is a componentwise outlier, the contrast between the r_{ij} of contaminated and non-contaminated components is now stronger; (2) when x_i is a structural outlier involving multiple components, all of the c_{ij} may be relatively small and then this choice of α_{ij} is more conservative and may avoid unwarranted reduction of outlyingness for those components sharing responsibility for the large outlyingness r_i.

(3) $\alpha_{ij} = (\max_{k=1}^{p} |u_{ik}|)^{-1} |u_{ij}|$, where $u_i = (u_{i1}, \ldots, u_{ip})$ denotes the direction that maximizes r_i. Hence, α_{ij} is large whenever the jth component has a relatively large coefficient in the maximizing direction u_i. The underlying idea is that the magnitude of the coefficients in u_i reflects the extent to which the respective components are responsible for the outlyingness r_i, both in the case of componentwise and structural outliers.

We will refer to these methods as (1) SDH, (2) SDC and (3) SDM (respectively, from *Half*, *Components* and *Maximizing*). The resulting location and scatter estimates (4) and (5) will sometimes be denoted by $(T_{\mathrm{SDH}}, S_{\mathrm{SDH}})$, $(T_{\mathrm{SDC}}, S_{\mathrm{SDC}})$ and $(T_{\mathrm{SDM}}, S_{\mathrm{SDM}})$, respectively. Note further that when $\alpha_{ij} = 1$, we obtain the original Stahel–Donoho estimator $(T_{\mathrm{SD}}, S_{\mathrm{SD}})$.

Remark 1 when applying the SDM method, it would be desirable to rescale each variable first by dividing it componentwise by its MAD, ensuring that the variances of the components are within the same range. This would avoid the maximizing directions u_i being attracted to the variables with the smallest scales. Note that rescaling the variables by the componentwise MAD before computing the maximizing directions u_i, however, is equivalent to adjusting the directions u_i by multiplying its coefficients by the respective MADs (without rescaling the variables). In the rest of the paper, adjusted u_i vectors will be used in the SDM method.

4. Simulation study

In this section, we investigate through simulation to what extent our methods succeed in reducing the outlyingness, and thus increasing the weight, of precisely those components for which it would be justified.

Samples $\{x_1, \ldots, x_n\}$ of size $n = 100$ in $p = 3$ dimensions were generated from a standard normal distribution $N(0, \Sigma)$. For each sample, $100\epsilon\%$ of the data were shifted over a distance of $k\mathbf{m}$. (Additionally, the variance was reduced for those components with non-zero coefficient in the outlying direction $\mathbf{m}$, by multiplying their standard deviation by 0.1).

We considered two choices for Σ, respectively, corresponding to uncorrelated and partly correlated data: (1) $\Sigma_u = I_3$ (i.e. the identity matrix), and (2) $\Sigma_c = [1 - 0.9\,0; -0.9\,1\,0; 0\,0\,1]$. In the case of Σ_u we considered outliers in directions $\mathbf{m} = (1, 0, 0), (1, 1, 0)$ and $(1, 1, u)$ with $u \geq 1$. For the correlated data corresponding to Σ_c we looked at outliers in directions $\mathbf{m} = (1, 0, 0), (0, 0, 1), (1, 1, 0), (1, 0, 1), (1, 1, u)$ and $(1, u, 1)$, again with $u \geq 1$. The distance k and the proportion ϵ of the outliers were varied.

For each situation, 100 samples were generated. For each sample, and for each observation x_i, we computed the Stahel–Donoho outlyingness r_i and subsequently the adapted outlyingnesses r_{ij} based on the methods SDH, SDC and SDM. The corresponding weights w_i and w_{ij} are based on the weight function (2). When the Stahel–Donoho procedure is sufficiently robust against the outliers generated as above, the weight w_i will be close to 0 for the outliers and close to 1 for the regular observations. The adaptations SDH, SDC or SDM then can be said to work appropriately if the following more or less hold:

(1) if x_i was generated as an outlier, and if the jth component has a zero coefficient in the outlying direction $\mathbf{m}$, then w_{ij} is high (i.e. the weight of the jth component should be awarded an increase compared with the SD weight);
(2) if x_i was generated as an outlier, and if the jth component has a non-zero coefficient in the outlying direction $\mathbf{m}$, then w_{ij} is close to w_i (i.e. the weight of the jth component should be similar to the original SD weight);
(3) if x_i was generated as a regular observation, then w_{ij} is high (i.e. the weights of all components should be as close as possible to 1).

(Recall that by definition $w_{ij} \geq w_i$.) When reporting our results we will divide the components x_{ij} into groups corresponding to these three situations, which we will also, respectively, refer to as (1) non-contaminated components of outliers; (2) contaminated components of outliers; and (3) components of regular observations. In particular, we will report the mean weights w_i and w_{ij} for each of the three groups over all generated samples in each of the outlier situations described above.

For the data in this simulation study, we would say that we are dealing with *componentwise* outliers as soon as the distance k is sufficiently large (such that c_{ij} is large for each contaminated component j). On the other hand, we would refer to *structural* outliers for those cases where each c_{ij} is relatively small but the global outlyingness r_i is large. This especially occurs in the case of covariance matrix Σ_c and outliers in the direction (1,1,0) or (1,1,u) at a relatively small distance k. However, we do not attempt to make a strict distinction between the two types of outliers in this study. In fact we make the general assumption that those components that were not contaminated, i.e. that were actually generated according to the standard normal distribution, are justified to have their weight increased. The other components are not.

Note that exact computation of the supremum in the Stahel–Donoho outlyingness (1) is impractical and typically a random search algorithm based on subsampling is used to obtain an approximation. In this paper, we applied a Matlab implementation of the Gauss-algorithm

used in [3]. The number of random directions in the algorithm was taken equal to 1000, which should be sufficient for p as small as in this simulation study [3].

We first consider the simple case of $\Sigma = \Sigma_u$, such that none of the components are correlated. In this situation all outliers may be regarded as componentwise outliers. The top panel in Figure 4 presents the results for $\mathbf{m} = (1, 0, 0)$, while the bottom panel represents the case $\mathbf{m} = (1, 1, u)$. The results for $\mathbf{m} = (1, 1, 0)$ are omitted as they are very similar to the case $\mathbf{m} = (1, 0, 0)$. Here, and in the following, the left plots correspond to the group of non-contaminated components of outliers, the middle plots to the contaminated components, and the right plots to the components of

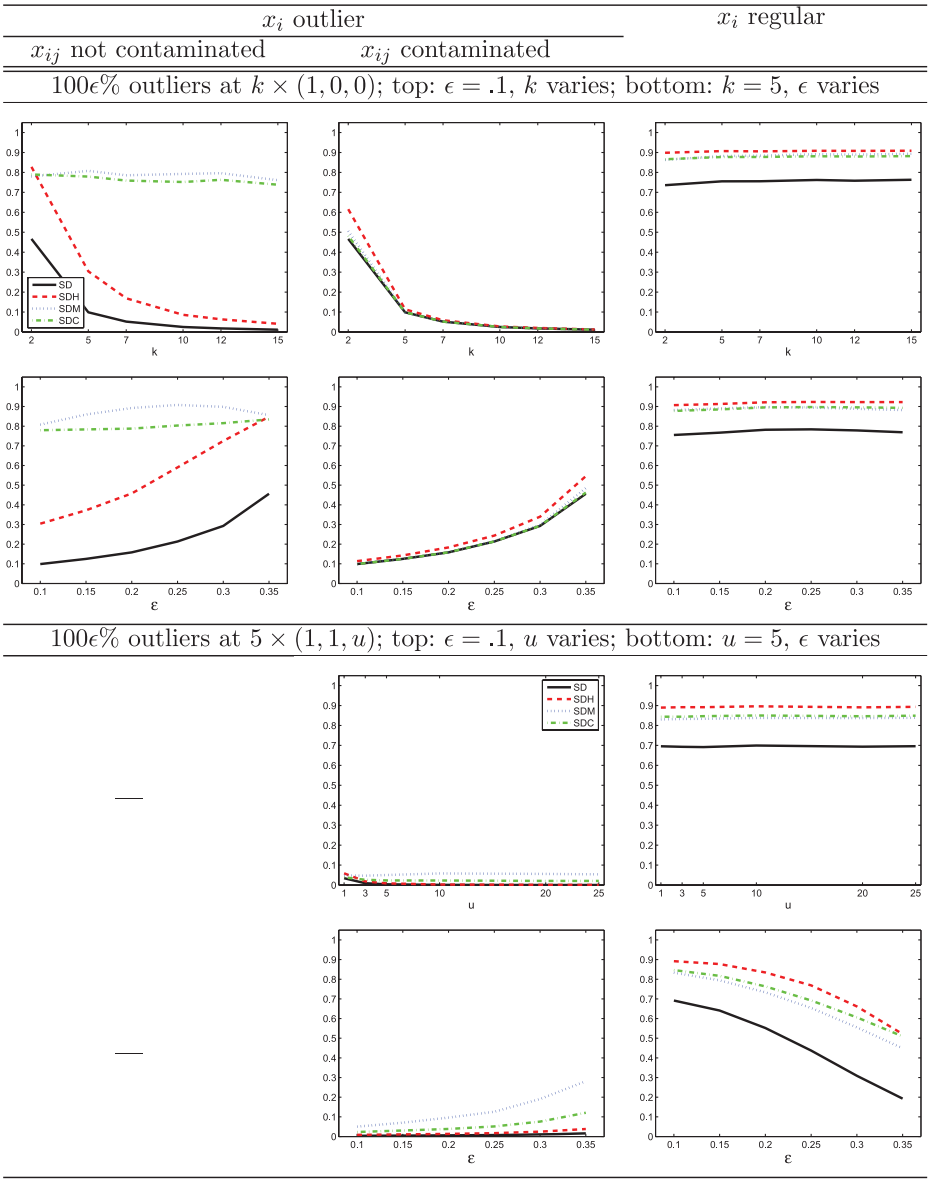

Figure 4. Simulation results uncorrelated data: mean weights for (left) non-contaminated components of outliers, (middle) contaminated components of outliers, (right) components of regular observations; various outlier configurations as indicated.

regular observations. Each plot shows the mean values of w_i and w_{ij} of all concerned components in all generated samples. The top row of a panel always presents the results as a function of the outlier location (determined by k and/or u) for fixed proportion of outliers ϵ. In the bottom row the outlier location (k and/or u) is fixed and the plots show the weights in function of the fraction of outliers ϵ. The solid lines correspond to the initial SD weights w_i, while the dashed, dotted and dash-dotted lines, respectively, represent the SDH, SDM and SDC weights w_{ij}.

Consider the choice $\mathbf{m} = (1, 0, 0)$, i.e. outliers are situated in the first component only, in the top panel of Figure 4. In the top row, where $\epsilon = 0.1$ and the distance k varies, we see that the SD weights for the outliers are close to 0, except when k is small, in which case the Stahel–Donoho estimator could consequently not withstand the outliers. For larger k we see in the left plot that the SDC and SDM methods are both successful in awarding an increase in weight for the non-contaminated components of the outliers. SDH succeeds in only a limited increase. From the middle plot it is clear that all methods leave the weight of the contaminated component almost untouched, which is as desired. The right plot, representing the regular observations, indicates that all methods yield a slight gain in precision by somewhat increasing the (already high) average weights of those observations. Note that increasing the weight of clean components may constitute a gain in robustness, rather than a gain in precision, if it offsets a possible bias induced by contamination in the corresponding components from other observations (see the next section).

In the bottom row of the top panel in Figure 4 the distance of the outliers is fixed at $k = 5$ and the proportion ϵ is varied from 10% to 35%. It is again clear that the non-contaminated components obtain a considerable weight increase when w_i is low, while the contaminated components do not. We may conclude that when $\mathbf{m} = (1, 0, 0)$, the adapted outlyingnesses succeed nicely in boosting precision while largely preserving robustness.

Now for the choice $\mathbf{m} = (1, 1, u)$, which is shown in the bottom panel of Figure 4, each component of the outliers is contaminated and hence we have only two plots in each row of this panel. Our interest is now primarily in the components of the outliers, which preferably should not be awarded an increase in weight. The main difficulty for our methods here (especially SDC and SDM) is that the contamination in components 1 and 2 would be dwarfed by the much more severe contamination in component 3, and that this may lead to the first two components obtaining a somewhat unjustified gain in weight. However, because it always holds that $r_i \geq c_{ij}$ it follows from Equation (7) that the adapted outlyingness r_{ij} is also bounded below by c_{ij}. Hence, whenever an observation is outlying in the direction of the jth component, any decrease in outlyingness awarded to that component is bounded by the large componentwise outlyingness c_{ij}. Therefore, any gain in weight w_{ij} is kept limited. This ensures that our methods to a reasonable extent are conservative, which indeed can be seen in the results in Figure 4.

The top row again takes $\epsilon = 0.1$, fixes additionally $k = 5$, and shows the mean weights for a grid of increasing u. The bottom row sets $k = 5$ and $u = 5$ and varies the proportion of outliers. First, the SDH method has a fixed weighting parameter α_{ij} and that is the reason why it fares reasonably well in this situation. Indeed, its weights w_{ij} are closest to the SD weights w_i on average. Both SDM and SDC yield a somewhat larger increase for the weights of the contaminated components, which can be entirely attributed to the components 1 and 2, for which α_{ij} is relatively low compared with component 3. Nevertheless, the increase in weight is very limited, for reasons explained above, and it is not deemed harmful in the sense that at worst it would yield a slight bias increase in the adapted Stahel–Donoho estimates. In the next section, we will examine the issue of robustness and possible bias increase of the presented estimators more thoroughly.

Next, we consider the partly correlated data generated with $\Sigma = \Sigma_c$. The situation is now more complicated since the high negative correlation between components 1 and 2 could produce far outliers for which none of the components needs to be severely contaminated. Or, in other words, the SD outlyingness r_i is potentially much higher than any of the componentwise c_{ij}.

Figure 5 contains the results for $\mathbf{m} = (1, 0, 0)$ and $(1, 1, 0)$, while Figure 7 contains those for $\mathbf{m} = (1, 0, 1)$ and $(1, 1, u)$. Results for $\mathbf{m} = (0, 0, 1)$ and $(1, u, 1)$ are omitted (these results were found to be comparable to, respectively, $\mathbf{m} = (1, 0, 0)$ and $(1, 1, u)$ in the case of uncorrelated data, as considered above).

Let us first look at the results for $\mathbf{m} = (1, 0, 0)$ in the top panel of Figure 5. The two rows again, respectively, fix the outlier proportion at $\epsilon = 0.1$ and the distance at $k = 5$. We see that the behaviour of the methods looks similar to that in the corresponding uncorrelated setting that

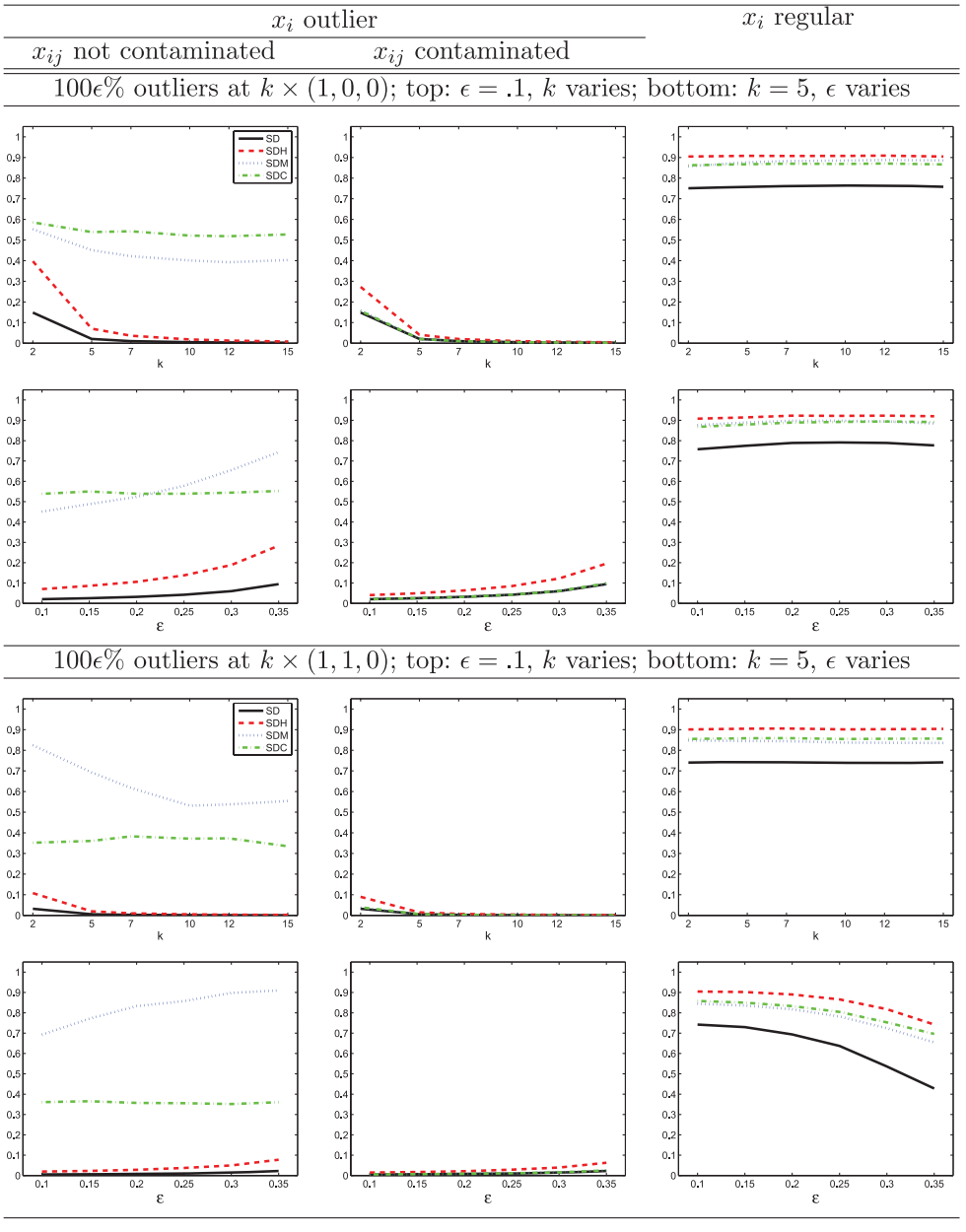

Figure 5. Simulation results correlated data: mean weights for (left) non-contaminated components of outliers, (middle) contaminated components of outliers, (right) components of regular observations; various outlier configurations as indicated.

was shown in Figure 4, top panel, except for the following differences. First, note that the SD weights are generally lower than in Figure 4 because the outliers now are more pronounced and hence easier to detect. Indeed, although the outliers were generated in the first component only, because of the high correlation, the outlyingness r_i becomes considerably higher (and is maximized) in a direction which additionally involves the second component. This is also the reason why, as seen in the left plots, the SDM method behaves somewhat more conservative for the non-contaminated components than in the uncorrelated setting. In fact, the dashed curves are rather low here because the second component is considered by the SDM α_{ij} parameter as partly responsible for the high r_i and hence does not receive a high weight. The third component, on the other hand, does generally receive a gain in weight and so the SDM curves still end up somewhere in the middle of 0 and 1. The fact that SDM wrongly attributes some responsibility to the second component is understandable and difficult to avoid, since the data correlation implies that these particular outliers could equally easily have been produced by contaminating both components 1 and 2, as shown in Figure 6. Regarding SDC, the method comes out on top but its performance for non-contaminated components is also somewhat worse than in Figure 4. The reason is that r_i is now much higher, due to the correlation, while the weighting parameter α_{ij} remains the same. For the contaminated components shown in the middle plots, the methods perform very well, except for SDH, which seems overly liberal here. Hence, we may again conclude that SDC and SDM best preserve the outlier resistance while increasing the precision.

The bottom panel in Figure 5 corresponds to outliers in the direction $\mathbf{m} = (1, 1, 0)$, which means that the outliers are now even more pronounced than in the previous setting, given k, resulting in even lower SD weights w_i. The only non-contaminated component of the outliers is (the uncorrelated) component 3, and we see in the left plots that SDM performs best in awarding this component a higher weight. The SDC method clearly has difficulties in overcoming the large r_i in view of the small c_{ij} (and the small differences among the c_{ij}), and therefore turns out to be more conservative here. The SDH method, finally, completely fails as it is unable to distinguish properly between contaminated and non-contaminated components (especially when k is small).

Next, consider the case $\mathbf{m} = (1, 0, 1)$ in the top panel of Figure 7, where the only non-contaminated component is the second one, which is of course highly negatively correlated with the first one. We see in the left plots that the SDM method hardly increases the weight of the second component, because it deems the latter partly responsible for the outlyingness of these observations. The current situation is comparable to the one corresponding to $\mathbf{m} = (1, 0, 0)$ in Figure 5, top panel, where we had additionally component 3 as non-contaminated (which ensured that the SDM curves on average still indicated an increase). The SDC method here again performs well, similarly to the $\mathbf{m} = (1, 0, 0)$ case.

Finally, the bottom panel of Figure 7 represents outliers in $\mathbf{m} = (1, 1, u)$. The top row fixes $\epsilon = 0.1$ and $k = 2$, and shows the mean weights in function of u. The bottom row varies again

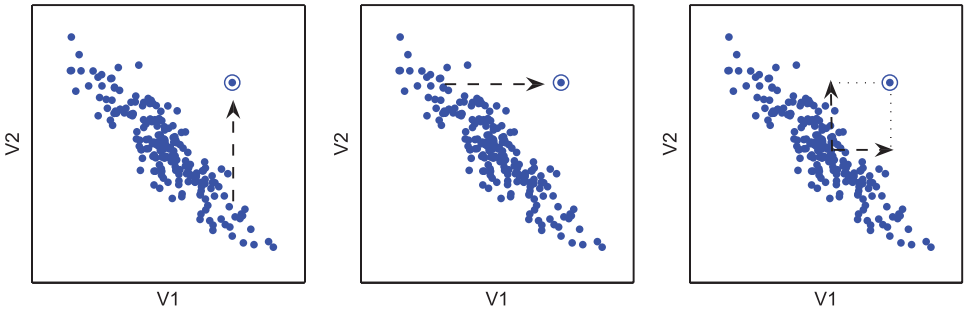

Figure 6. 2-D example with highly correlated data: different causes leading to the same outlier.

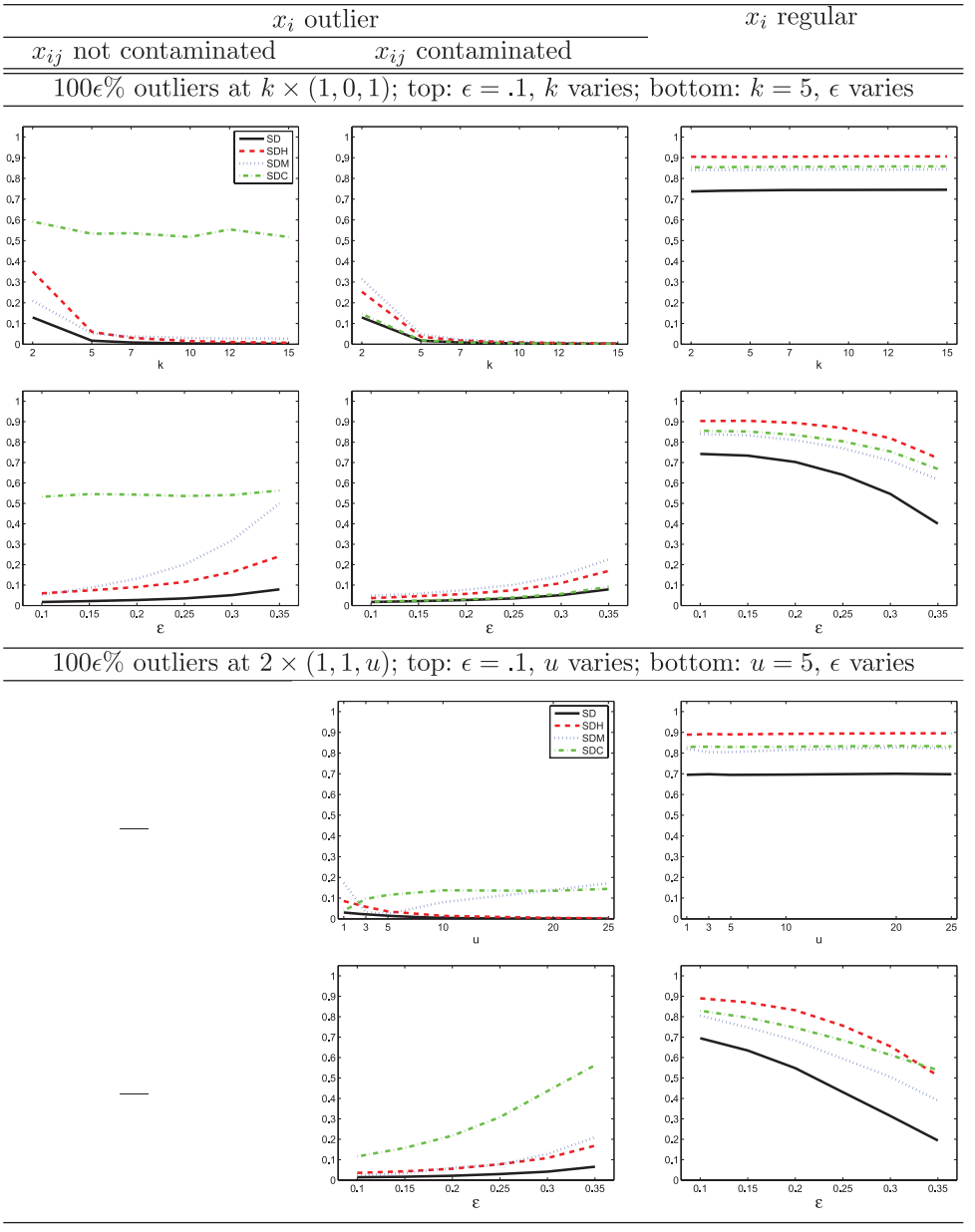

Figure 7. Simulation results correlated data: mean weights for (left) non-contaminated components of outliers, (middle) contaminated components of outliers, (right) components of regular observations; various outlier configurations as indicated.

the proportion of outliers, with $k = 2$ and $u = 5$. Since all components of the outliers are contaminated, we again have only two plots to consider. Note that the small k in combination with the high correlation implies that we are dealing with truly structural outliers. We are concerned here that the components of the outliers should not receive a high weight w_{ij}. We see, however, that SDC is behaving in a way which is somewhat too liberal in this context. The reason is that it believes it should justifiably increase the weights of components 1 and 2 since their c_{ij} is very small compared with that of component 3. The SDM method does better in this regard.

This last situation illustrates the main difficulty of our methods, namely that they may be underestimating the responsibility of some components in the case of certain structural outliers, e.g. in the case of large outlyingness due to correlation combined with other outlying components. In general, we may conclude that both SDC and SDM offer a considerable increase in weights for non-contaminated components for many outlier situations. In some cases this increase in weights also leads to an increased robustness, such as in the case of independent contamination in the different components. However, in other cases the potential gain in precision comes at a cost of increased weights for contaminated components, and thus some loss of robustness, but this cost is limited since the adjusted outlyingnesses r_{ij} are bounded by the componentwise outlyingnesses c_{ij}. In the next section, we examine more closely this possible trade-off between precision and robustness.

5. Precision and robustness

In this section we show the results of a simulation study that investigates how the precision and the robustness of the Stahel–Donoho estimator are affected by our adaptations. In particular we present mean-squared errors (MSEs) of the estimates, and we focus on the location part. Here, *precision* is meant to represent the variance of the estimates, while by *robustness* we mainly refer to the bias.

Recall that our adaptations can only adjust the weights w_{ij} of the observations upward from the Stahel–Donoho weight w_i. Such upward adjustments produce either a reduction or an increase of the MSE for the resulting location estimates $T_{\mathrm{SD}^*,j}$, depending on which of the three groups (as distinguished in the previous section) component x_{ij} belongs to:

(1) non-contaminated components of outliers: *reduction* of the MSE, due to gain in precision or bias reduction;
(2) contaminated components of outliers: *increase* of the MSE, due to a bias increase;
(3) components of regular observations: *reduction* of the MSE, due to gain in precision or bias reduction.

Regarding the reduction of the MSE, we distinguish between gain in precision and bias reduction. The former refers to a variance reduction when the original SD estimate $T_{\mathrm{SD},j}$ was roughly unbiased (either because there are no harmful outliers in the jth component or the outliers have been downweighted by the SD estimate). Bias reduction, on the other hand, corresponds to the situation where $T_{\mathrm{SD},j}$ is biased because of outliers that have not been sufficiently downweighted. In such cases, adjusting the weight w_{ij} for an observation in which the jth component is clean, indeed may reduce the bias of $T_{\mathrm{SD},j}$.

Hence, the interest in this study is essentially to investigate whether the positive MSE effect corresponding to components from groups (1) and (3) outweighs the negative MSE effect of observations from group (2), resulting in an average MSE reduction.

Samples $X = \{x_1, \ldots, x_n\}$ were generated from a p-variate normal distribution, with p taking the values 5 and 10, and size $n = 10p$. Subsequently, contamination was added.

5.1. *Componentwise outliers*

First, we used a standard normal distribution and randomly added univariate outliers independently in each component. In particular, for component X_j (with $j = 1, \ldots, p$), $100\varepsilon\%$ of the observations $\{x_{1j}, \ldots, x_{nj}\}$ were shifted over a distance of $k\,m_j$, with m_j the jth component (after normalization) of the outlying direction $\mathbf{m}$. Again, the variance was reduced for those components with non-zero coefficient in $\mathbf{m}$, by multiplying their standard deviation by 0.1.

We considered various choices of $\mathbf{m}$, with outlying distances $k = 6, 24, 64$ and 160. For each situation, $N = 500$ samples were generated. Then, for each sample $X^{(l)}; l = 1, \ldots, N$ and for each observation x_i in $X^{(l)}$, we computed the Stahel–Donoho outlyingness r_i, the adapted outlyingnesses r_{ij} based on the methods SDH, SDC and SDM, and subsequently the corresponding location estimates $T^{(l)}_{.,j}$. The number of random directions in the algorithm was set at $200p$.

Next, the MSE of the various methods was calculated as

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

To simplify interpretation, the MSE was computed also separately for the variables which contain contaminated observations and the variables without any contamination:

$$\mathrm{MSE}_C(T_.) = \operatorname*{ave}_{j\in C} \left(\operatorname*{ave}_{l=1,\ldots,N} (T^{(l)}_{.,j})^2 \right), \quad \mathrm{MSE}_{NC}(T_.) = \operatorname*{ave}_{j\in NC} \left(\operatorname*{ave}_{l=1,\ldots,N} (T^{(l)}_{.,j})^2 \right)$$

with C and NC, respectively, denoting the set of variables with and without contamination.

We first consider the case $p = 5$ where 20% of independent contamination was added in the outlying direction $\mathbf{m} = (1, 1, 0, 0, 0)$. The top row in Figure 8 presents the MSE for the location

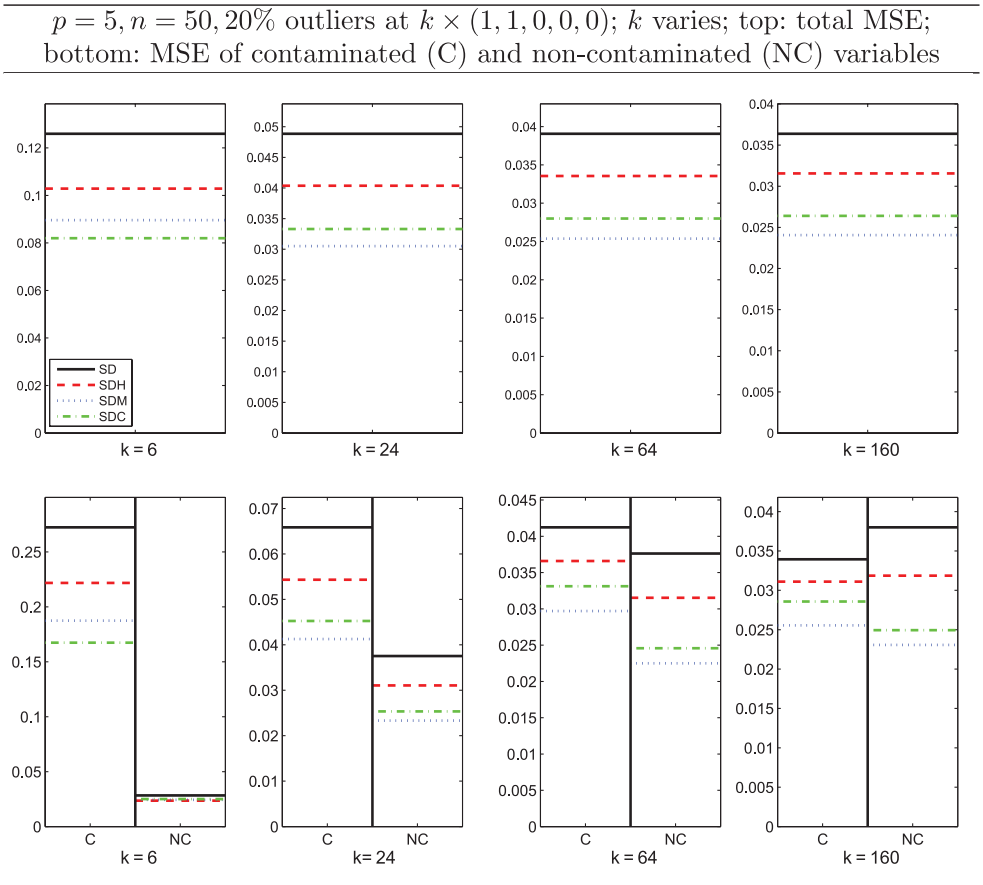

Figure 8. Simulation results univariate outliers: mean squared errors for different k values; various outlier configurations as indicated.

estimates of the original SD (solid) and the adaptations SDH, SDM and SDC (dashed, dotted and dash-dotted, respectively), for different values of the outlying distance k. From these plots, it can be seen that the adjusted methods SDH, SDM and SDC perform well, as their overall MSE is considerably smaller than the MSE of the original Stahel–Donoho estimator.

To get more insight in this MSE reduction, we split up the MSE by looking at the contaminated and non-contaminated variables separately. As contamination was added in direction (1, 1, 0, 0, 0) here, only variables 1 and 2 are contaminated. The average MSE for these two components of the location estimate, MSE_C, is shown in the left column (C) of the plots in the bottom row of Figure 8. The MSE of the remaining components, MSE_NC, is shown in the right column (NC).

First, it is clear that the original SD estimate suffers from a bias due to the contamination in the first two components, which corresponds to MSE_C being larger than MSE_NC. The bias gradually disappears when the outlier distance k increases, as the estimate succeeds better in downweighting the outliers. For small k values, we see that the observed overall MSE reduction of our adaptations was mainly due to a reduction of MSE_C. This particular reduction results from observations in groups (1) and (3) above receiving considerably higher weights and hence reducing the bias caused by the observations from group (2). The largest effect can be seen for SDC and SDM, which indicates that these methods succeed well in distinguishing contaminated from non-contaminated components. Regarding the non-contaminated components, where MSE_NC roughly corresponds to the variance, we see that SDC and SDM yield a relatively large gain in precision. For small

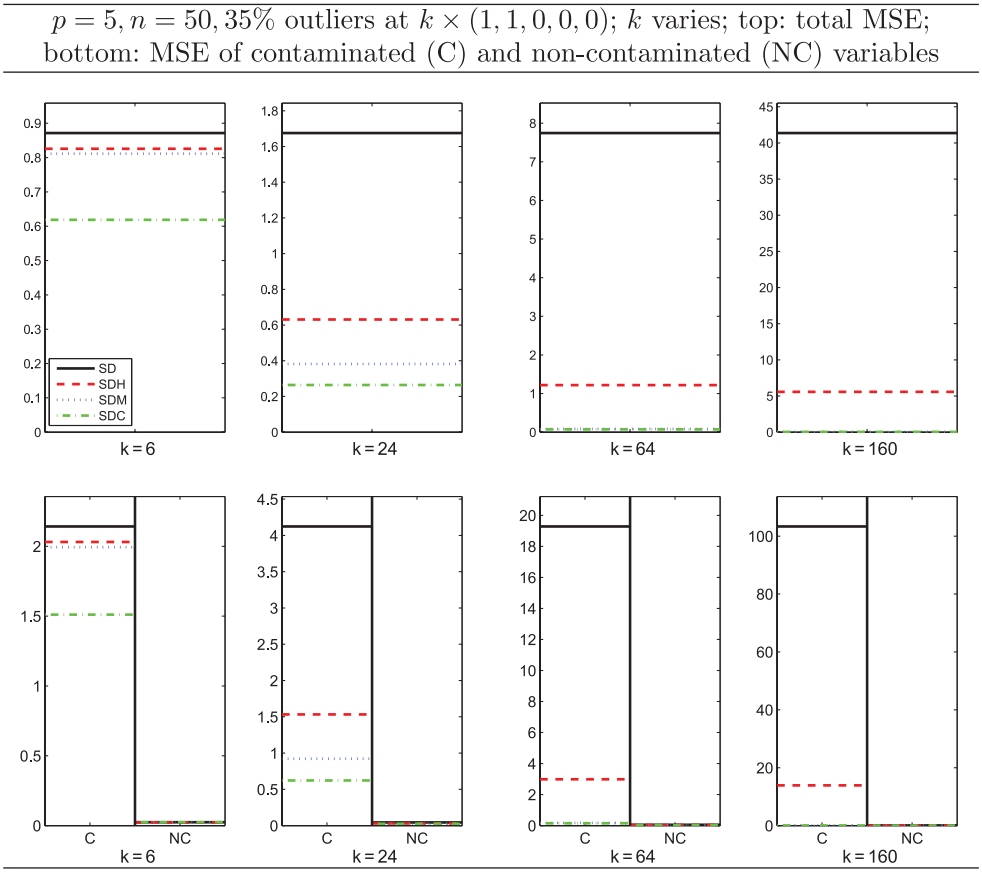

Figure 9. Simulation results univariate outliers: mean squared errors for different k values; various outlier configurations as indicated.

values of k however, this effect is somewhat dwarfed by the bias reduction in the contaminated components. For 20% of contamination in direction (1, 5, 0, 0, 0) (results not shown), we obtained similar conclusions.

Figure 9 presents the results for 35% of independent contamination in the outlying direction (1, 1, 0, 0, 0). We see that the original SD estimate is not able to resist this large amount of contamination, resulting in a very high MSE. However, the MSE of the adaptations SDH, SDM and SDC is rather low and hence offers a large improvement. Method SDC especially, performs very well, with a huge bias reduction for the contaminated components. For the non-contaminated components, the gains in precision are similar to those in the previous situations, but these gains are obviously negligible compared with the gain in robustness in the other components. These results indicate that the cellwise weighted SD estimates can potentially cope with much larger amounts of contamination than the original SD.

Let us now increase the dimension to $p = 10$ and add 10% of independent contamination in the outlying direction $\mathbf{m} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$. The results in Figure 10 again show that both SDC and SDM considerably improve the MSE of the location estimates, either due to a bias reduction of the contaminated components or due to a gain in precision. We further see that SDC and SDM perform similarly, except for contamination close to the regular observations (small k) where SDC outperforms SDM. Results for 10% of contamination in the outlying direction (1, 3, 5, 7, 9, 0, 0, 0, 0, 0) were similar to those in Figure 10 and are omitted here.

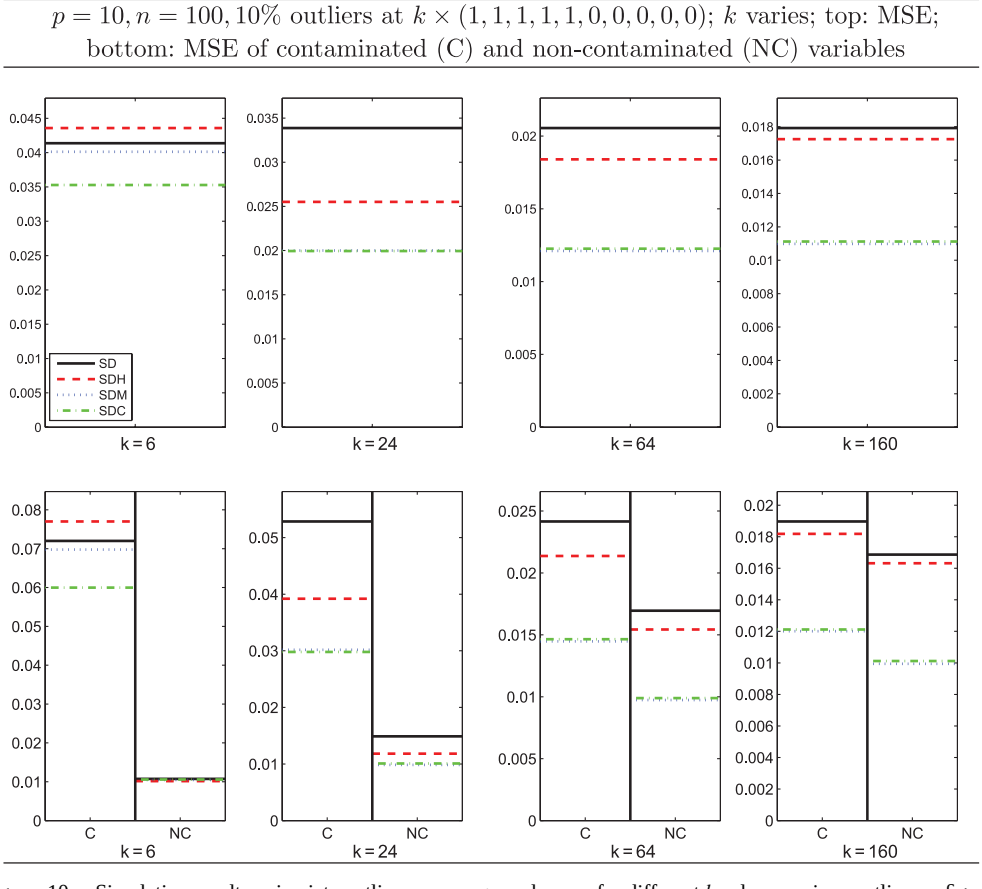

Figure 10. Simulation results univariate outliers: mean squared errors for different k values; various outlier configurations as indicated.

In Figure 11, the amount of independent contamination was increased to 20%, again in direction $\mathbf{m} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$. It can be seen that the SD location estimate has severe bias problems, which slowly improve, though for increasing outlying distance k. For $k = 6$, or contamination close to the regular observations, the cellwise weighted adaptations are not very effective at reducing the MSE. For contamination further away, however, we again observe quite a large improvement by methods SDC and SDM, mainly due to bias reduction and thus gain in robustness, but also because of gain in precision. Similar results were found for 35% of contamination.

5.2. *Structural outliers*

We also investigated the performance of the cellwise adaptations in the case of correlation outliers. The main difference with (independent) componentwise outliers is that now, we have more observations which are contaminated in several components at once. Consequently, the adverse effect of bias increase due to observations in group (2) above is more difficult to avoid.

The correlated data were generated (similarly as in [16]) by first drawing samples $\{y_1, \ldots, y_n\}$ of size $n = 100$ in $p = 10$ dimensions from a standard normal distribution and then setting $x_i = Ry_i$, with

$$R = \begin{pmatrix} A & \mathbf{0} \\ \mathbf{0} & I_{p-\tilde{p}} \end{pmatrix},$$

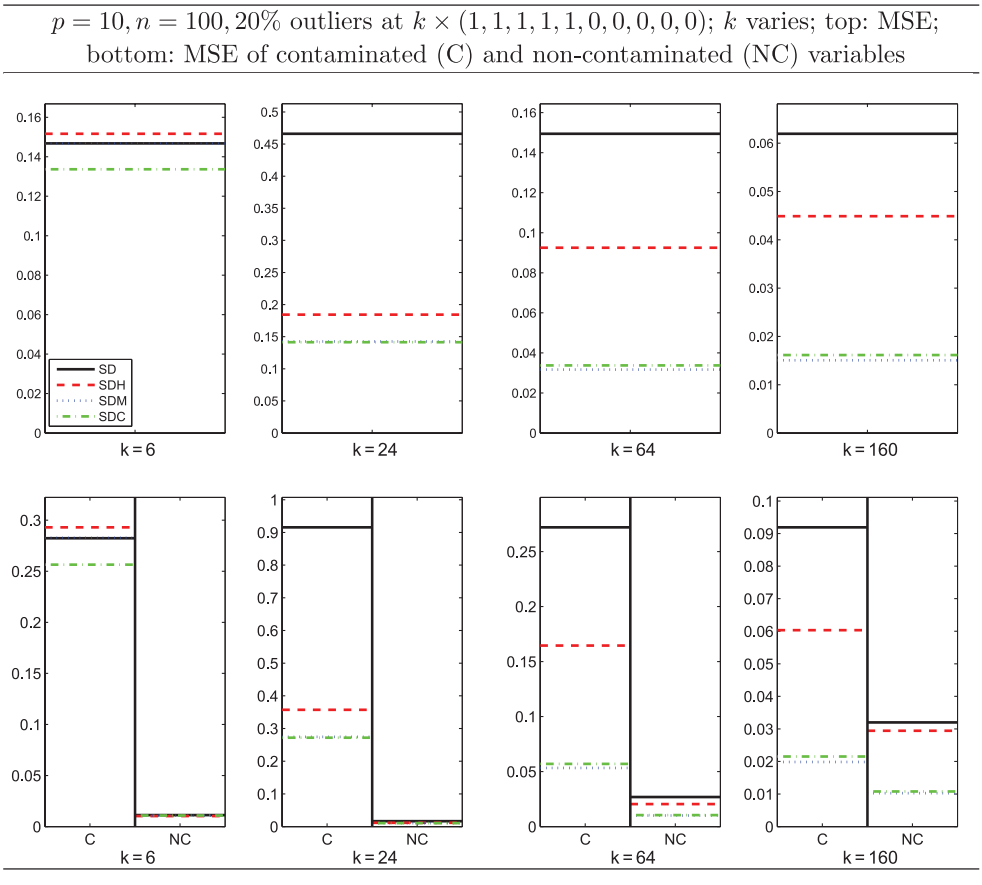

Figure 11. Simulation results univariate outliers: mean squared errors for different k values; various outlier configurations as indicated.

where $\mathbf{0}$ represents a zero matrix and $I_{p-\tilde{p}}$ is the identity matrix of dimension $p-\tilde{p}$, with $\tilde{p}$ the number of correlated components. Matrix $A \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ is defined by

$$A_{jj} = 1 \quad \text{and} \quad A_{jk} = \rho \quad \text{for } j \neq k.$$

Parameter ρ was chosen so that the multiple correlation ρ_{mult} between any coordinate of $\tilde{X} = (X_1, \ldots, X_{\tilde{p}})$ and all of the others, took on chosen values. If ρ_{mult} is high, then $\tilde{X}$ is concentrated around the line with direction $\mathbf{e} = (1, \ldots, 1) \in \mathbb{R}^{\tilde{p}}$.

Next, correlation outliers were added as follows. For each sample, $100\varepsilon\%$ of the data were shifted over a distance of $k\mathbf{m}$, where $\mathbf{m}$ is a unit vector. Define $\mathbf{b} \in \mathbb{R}^{\tilde{p}}$ by $b_j = (-1)^j$, and set $\mathbf{a} = \mathbf{b} - (\mathbf{b}'\mathbf{e}/\tilde{p})\mathbf{e}$, which is orthogonal to $\mathbf{e}$. Then, the outlying direction $\tilde{\mathbf{m}}$ is chosen as $\tilde{m}_j = a_j$ (for $j \leq \tilde{p}$) and $\tilde{m}_j = 0$ (for $j > \tilde{p}$). After normalizing $\tilde{\mathbf{m}}$ to unit norm, we obtain $\mathbf{m}$. Again the variance was reduced for the contaminated components, by multiplying their standard deviation by 0.1. For each situation, 500 samples were generated.

Similarly as before, we consider the overall MSE as well as MSE_{C} and MSE_{NC}, as defined above, where MSE_{C} represents the first $\tilde{p}$ components.

We first consider the case of $\rho_{\text{mult}} = 0$, such that none of the components are correlated. Figure 12 presents the results for 20% of outliers and $\tilde{p} = 4$. As before, the original SD suffers from some bias which diminishes as the distance k increases. In the top row, we see that

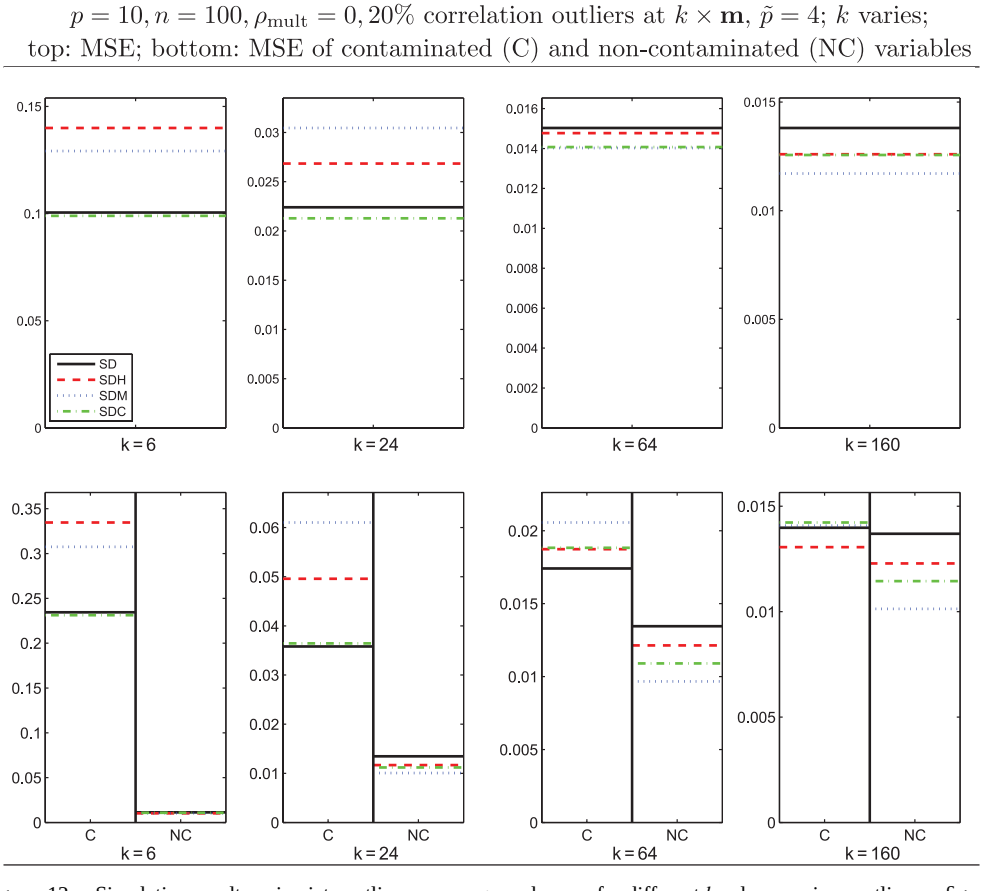

Figure 12. Simulation results univariate outliers: mean squared errors for different k values; various outlier configurations as indicated.

for small k, methods SDH and SDM do not succeed in reducing the MSE of the original SD estimator. In this case the higher weights of the clean observations of groups (1) and (3) above were not sufficient to offset the bias increase due to higher weights for contaminated observations from group (2). On the other hand, method SDC overall slightly improves on the SD. For larger outlying distance k, as the bias decreases, the MSE becomes dominated by the variance and SDM turns out to yield the largest gain in precision in these cases, although SDC also performs well.

Results for the correlated case with $\rho_{\text{mult}} = 0.9$ are shown in Figure 13. Due to the correlation, the outliers are relatively further from the regular observations than in the previous setting, and hence the bias of the original SD is limited, even for small k. However, the cellwise adaptations increase this bias and therefore increase the overall MSE for small k. Methods SDH and SDM especially, perform poorly. For larger distances k, we again observe that SDM offers a nice gain in precision, while the other adaptations also have a positive effect on the overall MSE. Similar conclusions were found for other simulation settings involving structural outliers, which are not reported here.

We may summarize the simulation results as follows.

- When the original SD estimates are roughly unbiased, then the cellwise adaptations have been shown to offer a relatively large reduction in MSE, corresponding to a gain in precision.
- When the original SD estimates are not able to fully resist the contamination and suffer from bias, then it is as follows.

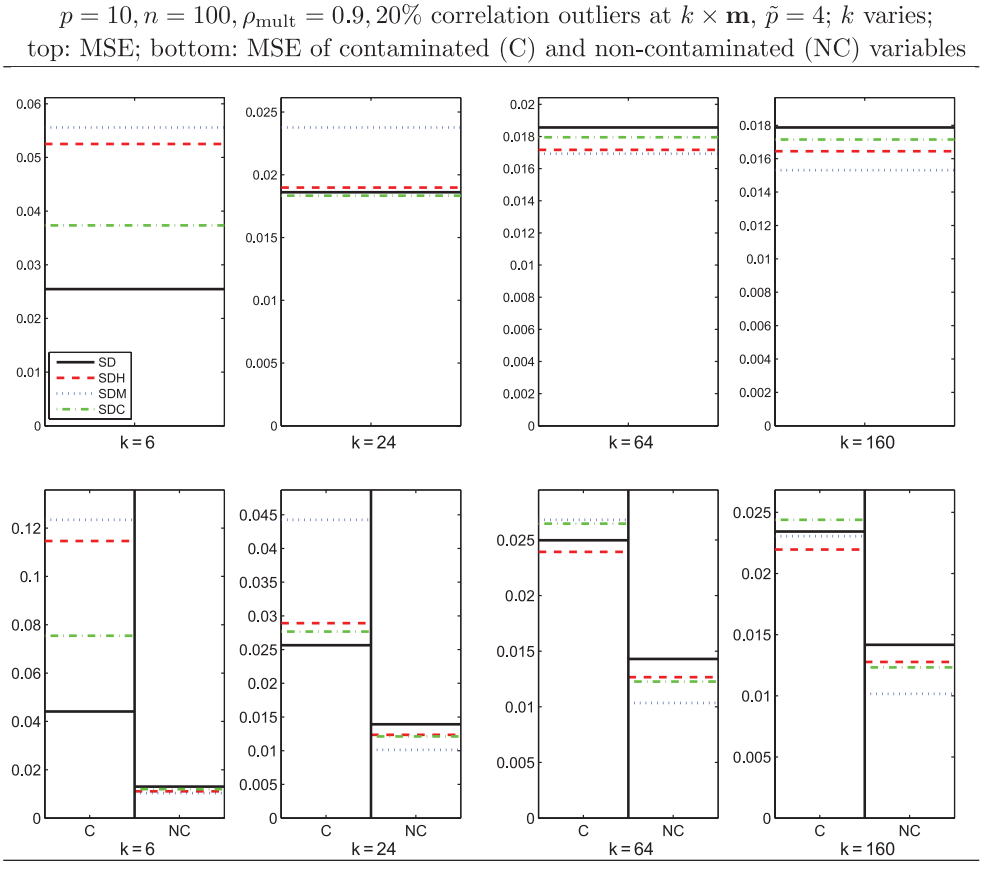

Figure 13. Simulation results correlation outliers: mean squared errors for different k values; various outlier configurations as indicated.

- In the case of independent componentwise contamination: the cellwise adaptations SDC and SDM are quite capable of identifying the correct contaminated component(s), hence prohibiting these components from obtaining an increased weight w_{ij}. The bias is therefore generally not increased, and on the contrary reduced, by increased weights for many clean observations. Additionally to this gain in robustness, a gain in precision is obtained in non-contaminated components.
- In the case of structural outliers: the cellwise adaptations SDC and especially SDM give evidence of difficulties identifying *all* of the contaminated components of an observation. Therefore some of these components may obtain a significant weight increase and this increases the bias. This loss of robustness in its turn may outweigh the gain in precision obtained in non-contaminated components.

We may conclude that the cellwise weighted adaptations in many situations offer a nice reduction of MSE with respect to the original Stahel–Donoho estimator. However, in some cases with structural outliers the methods do not perform well, with SDC being more reliable in this regard than SDM. Very positive results are obtained in the setting of independent componentwise contamination, in terms of both improving the precision and the robustness. With respect to the latter, it has been observed that both SDC and SDM can cope with larger amounts of contamination than the SD estimate itself.

6. Outlier detection

When indeed the cellwise adaptations increase the precision and robustness of the SD estimates, we may expect that this corresponds to better outlier detection capabilities for the adapted estimates. Outlier detection in a multivariate sample $X = \{x_1, \ldots, x_n\}$ is often based on the robust Mahalanobis distances

$$d_i = \sqrt{(x_i - T_{\mathrm{R}})^t S_{\mathrm{R}}^{-1}(x_i - T_{\mathrm{R}})}, \quad i = 1, \ldots, n,$$

where $(T_{\mathrm{R}}, S_{\mathrm{R}})$ are robust estimates of location and scatter. Those observations with large distances, for example exceeding a given quantile of the χ_p distribution, are considered outliers. The SD estimates of location and scatter are quite suitable for outlier detection based on robust distances. To investigate whether SDC and SDM can improve upon the original SD in this matter, we here present some additional simulation results.

We consider the same simulation design as in Section 5.1. That is, observations are generated from a standard normal distribution and a (random) proportion ϵ in each component j is subsequently shifted over a distance $k\,m_j$ for some outlying direction $\mathbf{m}$. Hence, a total number of n_{c} observations (with $n_{\mathrm{c}} \geq \epsilon\, n$) will be contaminated in one or more of its components. The performance of SD, SDC, and SDM in this setting can be examined in terms of *sensitivity* and *specificity*. Here, the sensitivity is the proportion of the n_{c} contaminated observations with distances exceeding the cutoff, whereas the specificity is the proportion of the $n - n_{\mathrm{c}}$ non-contaminated points with distances below the cutoff.

First, as in Figure 8, let $n = 50$, $p = 5$, $\epsilon = 0.2$, and $\mathbf{m} = (1, 1, 0, 0, 0)$. We generated $N = 50$ samples for a range of outlying distances k and for each sample we computed the robust distances based on SD, SDC, and SDM. For each of the three methods we computed the sensitivity and specificity for a large range of cutoff values. We can then plot sensitivity versus (1 - specificity), averaged over the N samples, resulting in so-called receiver operator characteristic (ROC) curves. The closer the curve follows the left-hand and then the top borders, the better the distances distinguish between outliers and non-outliers. The top panel in Figure 14 shows the curves for

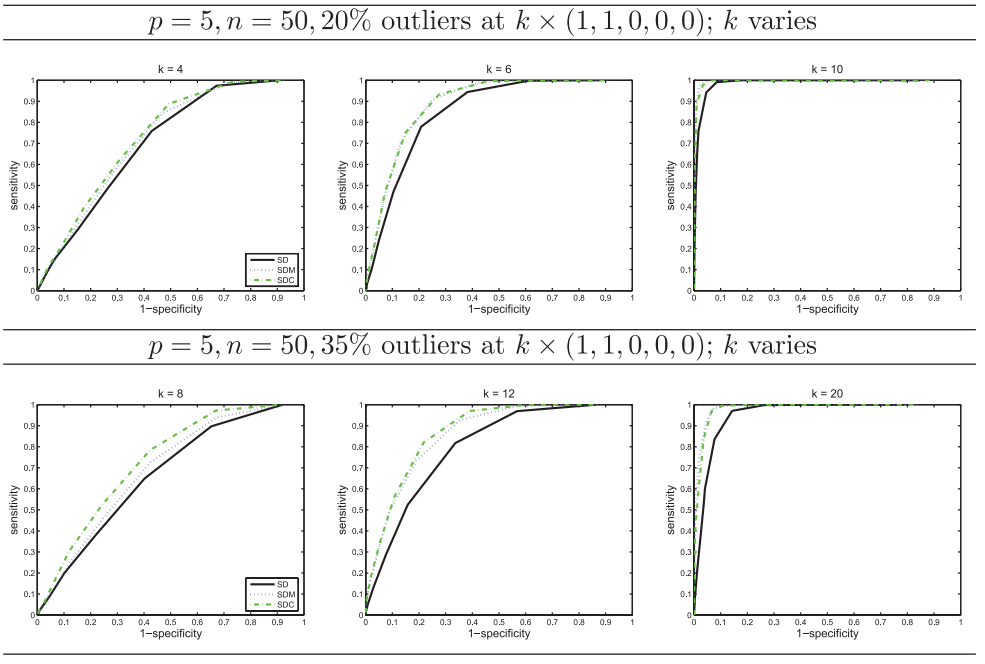

Figure 14. Outlier detection simulation results for univariate outliers with different k values; sensitivity versus (1-specificity)

$k = 4$, 6 and 10. The SD-based outlier rule is represented by the solid line, while SDC and SDM correspond to the dash-dotted and dotted line, respectively. We see that the SDC and SDM curves generally lie above the SD curve, indicating that the distances based on the cellwise adapted estimates better discriminate between outliers and non-outliers. This result is in accordance with the MSE results from the previous section.

The bottom panel in Figure 14 presents the results when the percentage of outliers is increased to 35%, for outlying distances $k = 8$, 12 and 20. The difference between the curves is now more pronounced, in favour of the cellwise methods, in line again with our findings regarding the MSE of the estimates.

Finally, for the same settings as in Figure 14, Figure 15 shows the average distances d_i for the contaminated observations on the one hand and the non-contaminated on the other hand, as a function of the outlier distance k. The horizontal line indicates the square root of the 0.975 quantile of the χ^2_p distribution, which is a common cutoff value. The left panel corresponds to $\epsilon = 0.20$, the right panel to $\epsilon = 0.35$. The average distances of the contaminated points are represented by the increasing curves, and we see that the distances based on SDC or SDM exceed the cutoff line faster than the SD-based distances. For the non-contaminated observations the average distances are decreasing with k, and we notice that the distances based on the cellwise estimates are slightly larger than those based on the SD. Hence, there is a trade-off, primarily because SDC and SDM in general yield smaller-scale estimates than SD. However, it is clear both from Figures 14 and 15 that the cellwise adapted estimates perform better than the original SD.

Other outlier settings gave similar results in the sense that the outlier detection performance is directly related to the MSE of the respective estimates. This means that, as expected, SDC and SDM do not outperform SD in detecting correlation outliers such as those corresponding to Figure 13.

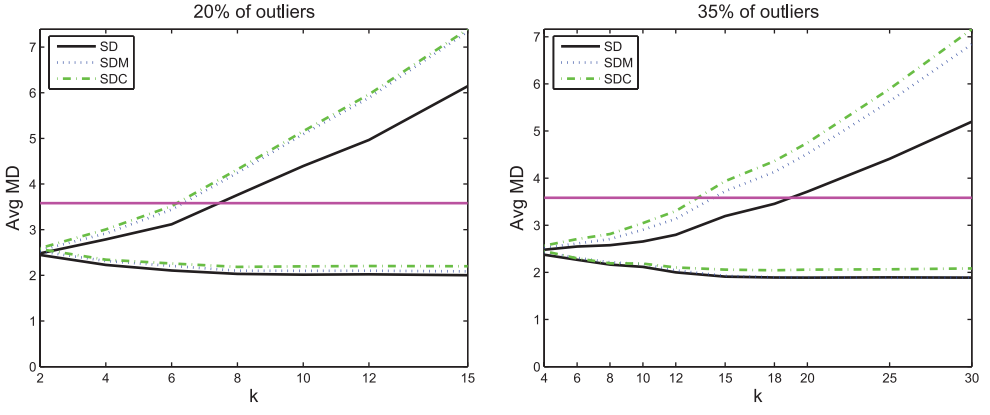

Figure 15. Outlier detection simulation results for univariate outliers with different k values; average robust distance for contaminated and non-contaminated observations separately.

7. Examples

7.1. *Philips data*

Consider again the Philips data ($n = 677$, $p = 9$), for which we have illustrated in Section 2 that the original Stahel–Donoho estimator is potentially ignoring a lot of information by assigning the same weight to each component of an outlier. Let us see now whether our adaptations of the Stahel–Donoho estimator can be useful.

Figure 16 shows the weights w_{ij} for SDH, SDC and SDM, respectively, in the form of heat maps. For each map, the horizontal axis represents the variable j and the vertical axis the observation i. The colours range from dark to light, corresponding to the weights ranging from 0 to 1. The maps have an additional column on the right, the $(p + 1)$th column, which shows the original SD weights w_i. These weights correspond to the outlyingnesses r_i shown in Figure 1. Observations 491–565, for example, can again be identified as a group of strongly outlying points, based on the very low SD weights represented by the dark colours. We now take a closer look at how the componentwise adaptations deal with the outliers.

In the left plot in Figure 16 we see that the weights assigned to observations 491–565 by the SDH method remain low for each of the nine variables. The method takes componentwise differences into account to a limited extent only, and none of the variables is awarded a full weight

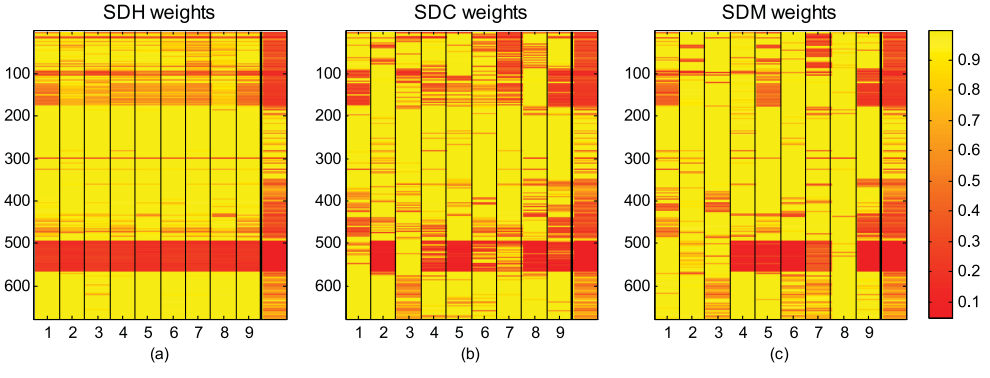

Figure 16. Philips data: adapted SD weights for (a) SDH, (b) SDC and (c) SDM.

repair. On the other hand, for many of the other outliers found by the SD weights, for example observations 1–175, the SDH method assigns full weights to each of the variables, which is not likely to be justified since at least one component should be responsible for the large outlyingness r_i. The reason for these large adapted weights is that apparently none of the componentwise outlyingnesses c_{ij} is reasonably large compared to r_i and SDH is not able to take into account combinations of components.

As can be seen in the middle and right plots in Figure 16, both SDC and SDM yield a much stronger distinction between the nine components than SDH did. The two methods act differently though. Concerning the group of observations 491–565, the SDC method mostly retains the low weight for components V2, V5 and V8, while the weights for components V1, V3 and V7 are considerably increased. On the other hand, the SDM method mainly holds components V4, V5, V6 and V9 responsible while fully repairing the weights of V1, V2, V3 and V8. Some differences between SDC and SDM can be noticed regarding the other outliers as well. All of these differences can, of course, directly be traced back to the behaviour of, respectively, the componentwise outlyingnesses c_{ij} and the maximizing direction coefficients u_{ij}. Figure 17(a) shows the values of $-c_{ij}$ while Figure 17(b) depicts $-|u_{ij}|/\|u_i\|$. The minus was added here to arrange that large values (light colours) be comparable to large values (light colours) in the plots of w_{ij} in Figure 16. We see for example that for observations 491–565 the value of c_{ij} is indeed relatively high in components V2, V5 and V8, while components V4, V5 and especially V6 and V9 exhibit the largest coefficients in the corresponding directions u_i.

Hence, estimators $(T_{\text{SDC}}, S_{\text{SDC}})$ and $(T_{\text{SDM}}, S_{\text{SDM}})$ both take componentwise differences into account. The question then arises which method acts most appropriately here. Let us focus on observations 491–565 and recall Figure 3, which showed some bivariate scatter plots. It could be seen that this group of points was clearly outlying with regard to the correlation between V6 and V9, as well as that between V6 and V5. Therefore we may conclude that at least components V5, V6 and V9 should retain a low weight and in this view SDM performs much better than SDC. Moreover, closer inspection suggests that the SDM results are generally supported by all pairwise scatter plots (not shown here). On the other hand, we have seen in Figure 2 that V2 is the component in which the outlyingness of observations 491–565 seems most extreme. Nevertheless, component V2 of these outliers is assigned weight 1 by the SDM method, while SDC retains a very low weight for this component. This could be seen as a shortcoming of the SDM method, although it is open to interpretation. Indeed, one might as well argue that SDM

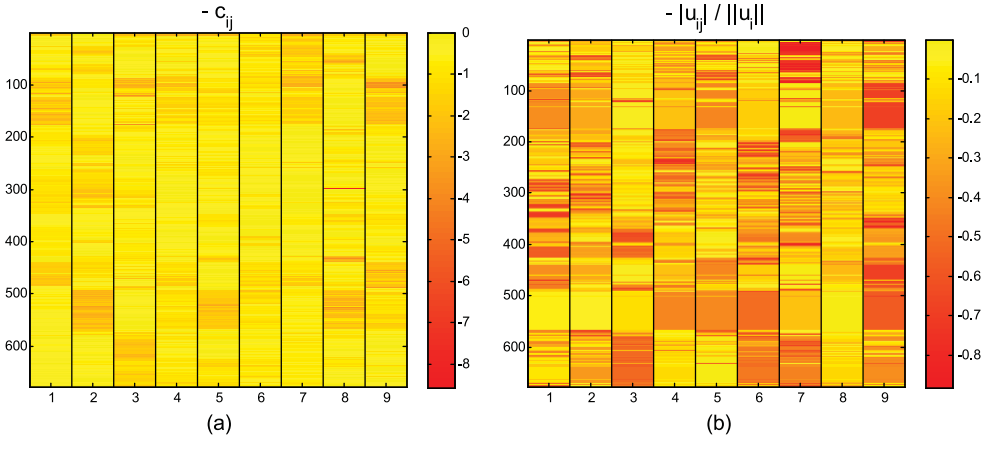

Figure 17. Philips data: (a) (negative) componentwise outlyingnesses c_{ij}, (b) (negative) coefficients in maximizing directions u_i (normalized).

righteously viewed the outlyingness in V2 as negligible and succeeded in identifying instead the truly important components to downweight. To conclude this example, although preceding simulation results indicated that the estimator $(T_{\mathrm{SDM}}, S_{\mathrm{SDM}})$ can be inaccurate in some cases, the SDM method here seems to act appropriately and to outperform the other methods. Given the number of components that were awarded a large weight increase, the adapted $(T_{\mathrm{SDM}}, S_{\mathrm{SDM}})$ estimates clearly use considerably more information than the original $(T_{\mathrm{SD}}, S_{\mathrm{SD}})$ estimates and hence are likely to be more efficient.

7.2. *Bank notes data*

For our second example, we consider the Swiss bank notes data [17], consisting of $n = 100$ forged old Swiss 1000 franc bills. The variables correspond to $p = 6$ different measurements, such as the length and height of the bill. The Stahel–Donoho outlyingnesses r_i of the observations are shown in Figure 18. A set of 15 outliers is highlighted. We are now again interested in the componentwise adaptations of the weights of these outliers provided by SDH, SDC and SDM.

Figure 19 depicts the weights w_{ij} for the three respective methods, with the original SD weight w_i again represented by the additional column on the right-hand side of each plot. In these plots, on the vertical axis the observations are now shown in the order of decreasing weight w_i, which

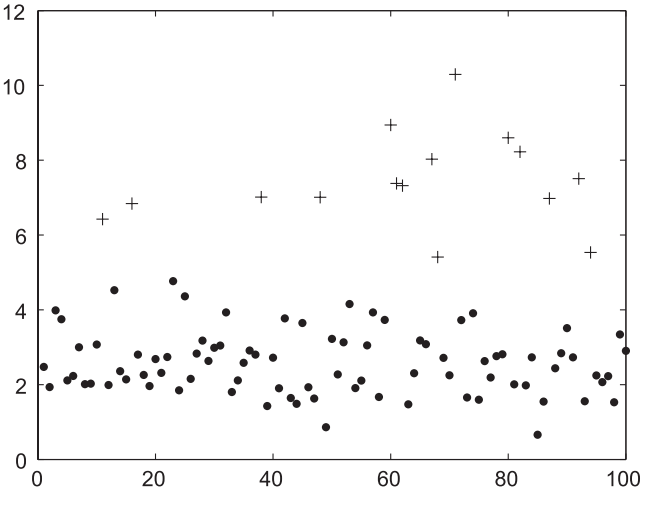

Figure 18. Stahel–Donoho outlyingnesses of the Bank notes data.

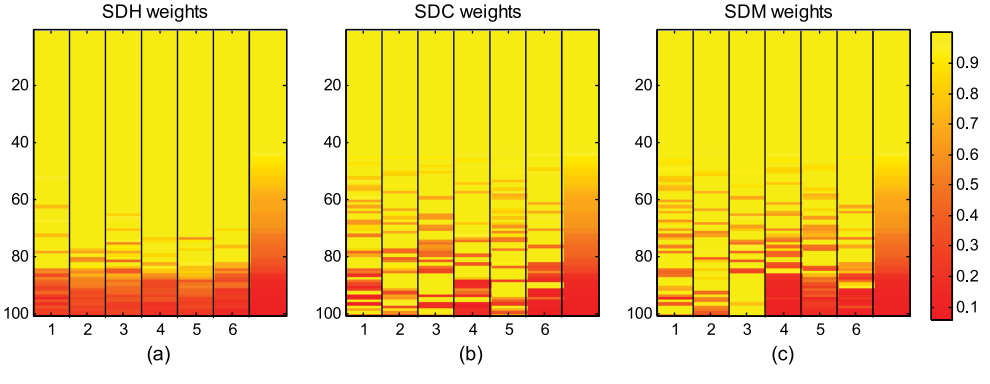

Figure 19. Bank notes data: adapted SD weights for (a) SDH, (b) SDC and (c) SDM.

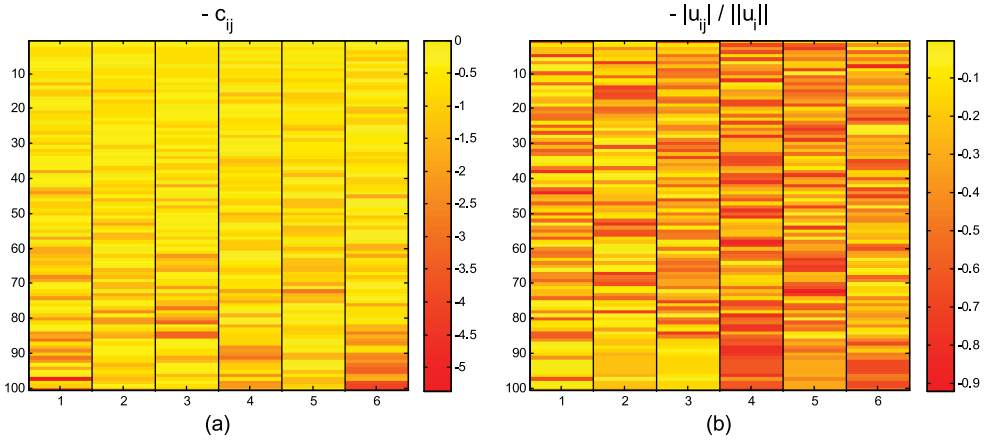

Figure 20. Bank notes data: (a) (negative) componentwise outlyingnesses c_{ij}, (b) (negative) coefficients in maximizing direction u (normalized).

aids the clarity of the results in this example. Figure 20 then again shows the corresponding values of $-c_{ij}$ and $-|u_{ij}|/\|u_i\|$ for comparison.

First, we see in the left plot of Figure 19 that the SDH weights deviate little from the original SD weights. In particular, for the outlying observations, the different components are often treated equally and few components have been restored through a higher weight. The SDC (middle plot) and SDM (right plot), on the other hand, both assign very distinct weights to the various components of the outliers. As in the previous example, the two methods do not fully agree on which components to hold responsible for the large outlyingnesses. The SDC method mainly down-weights components V1, V4 and V6 in accordance with the componentwise outlyingnesses c_{ij} of Figure 20(a), while SDM largely targets components V4, V5 and V6 following the coefficients in the maximizing directions u_i in Figure 20(b).

Figure 21 depicts univariate scatter plots for each component (standardized and with random jitter, analogously to Figure 2). The observations identified as outliers in Figure 18 are marked as dark plus signs. Additionally, in Figure 22 bivariate scatter plots are shown for V6 versus V4, V5 versus V4 and V6 versus V1. We see in Figure 21 that many of the outliers have large values in components V4 and V6, and the correlation between V4 and V6 apparently makes all of them strong bivariate outliers for these components, as seen in the left plot of Figure 22.

In the middle plot of Figure 22, we see a strong correlation between V4 and V5, with the highlighted points as clear bivariate outliers again. In fact, this correlation explains why SDM

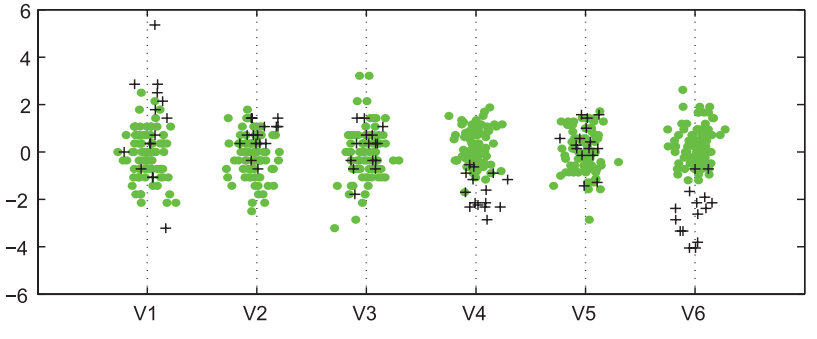

Figure 21. Bank notes data, componentwise scatter: each component is standardized by its median and modified MAD. The outlying observations are marked in dark.

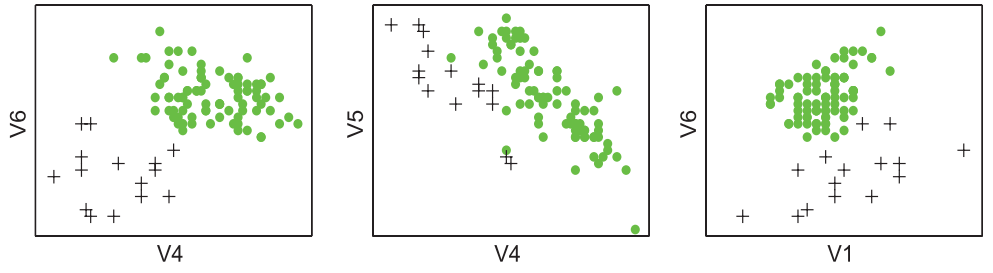

Figure 22. Bank notes data scatter plots: (a) variables 6 versus 4; (b) variables 5 versus 4; (c) variables 1 versus 6. The outlying observations are marked in dark.

targeted V5, even though we may view the outlyingness in this bivariate plot as caused by contamination in V4 only and not in V5 (see also the artificial example in Figure 6). In this sense, by considering V5 to be clean, SDC performed somewhat better than SDM. All other bivariate correlations, such as that between V6 and V1, are rather weak and do not seem to contribute much to the Stahel–Donoho outlyingnesses.

Since both SDC and SDM have recognized the responsibility of V4 and V6, we may conclude that both methods performed satisfactorily. The methods then only differ in the way they emphasized the minor contribution of V1 rather than that of V5 or vice versa. Finally, both SDC and SDM repair the weights of about half of the components of the outliers, which is likely to constitute an appreciable gain in precision for the adapted Stahel–Donoho estimates $(T_{\mathrm{SDC}}, S_{\mathrm{SDC}})$ and $(T_{\mathrm{SDM}}, S_{\mathrm{SDM}})$.

8. Conclusion

We presented an adaptation of the Stahel–Donoho estimator with separate weights for each component of the observations. Three cellwise weighted adaptations, denoted as $(T_{\mathrm{SDH}}, S_{\mathrm{SDH}})$, $(T_{\mathrm{SDC}}, S_{\mathrm{SDC}})$ and $(T_{\mathrm{SDM}}, S_{\mathrm{SDM}})$, were considered. It was shown that both SDC and SDM offer a considerable increase in weights for non-contaminated components for many outlier situations. This gain in precision sometimes comes at a cost of increased weights for contaminated components, and thus loss of robustness. But this cost is limited and a simulation study showed that in many situations the cellwise weighted adaptations enjoy a lower MSE than the original Stahel–Donoho estimator. Especially in the setting of independent componentwise contamination, our methods yield very good results. Finally, the identification of contaminated components in outliers may be of interest in its own right in the context of outlier detection and understanding the cause of outliers.

Acknowledgements

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 no. P6/03 of the Belgian government (Belgian Science Policy).

References

[1] W.A. Stahel, *Breakdown of covariance estimators*, Research Rep. 31, Fachgruppe für Statistik, E.T.H. Zürich, Switzerland, 1981.
[2] D.L. Donoho, *Breakdown properties of multivariate location estimators*, Ph.D. diss., Harvard University, 1982.
[3] R.A. Maronna and V.J. Yohai, *The behavior of the Stahel–Donoho robust multivariate estimator*, J. Amer. Statist. Assoc. 90 (1995), pp. 329–341.

[4] D. Gervini, *The influence function of the Stahel–Donoho estimator of multivariate location and scatter*, Statist. Probab. Lett. 60 (2002), pp. 425–435.
[5] Y. Zuo, H. Cui, and X. He, *On the Stahel–Donoho estimator and depth-weighted means of multivariate data*, Ann. Statist. 32 (2004), pp. 167–188.
[6] P.J. Rousseeuw, *Least median of squares regression*, J. Amer. Statist. Assoc. 79 (1984), pp. 871–880.
[7] P.L. Davies, *Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices*, Ann. Statist. 15 (1987), pp. 1269–1292.
[8] M. Hubert, P.J. Rousseeuw, and K. Vanden Branden, *ROBPCA: a new approach to robust principal component analysis*, Technometrics 47 (2005), pp. 64–79.
[9] M. Hubert and S. Verboven, *A robust PCR method for high-dimensional regressors*, J. Chemom. 17 (2003), pp. 438–452.
[10] M. Debruyne and M. Hubert, *The influence function of the Stahel–Donoho covariance estimator of smallest outlyingness*, preprint (2008), Statist. Probab. Lett. 79 (2009), pp. 275–282.
[11] J.W. Tukey, *The future of data analysis*, Ann. Math. Statist. 33 (1962), pp. 1–67.
[12] P.J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist. 35 (1964), pp. 73–101.
[13] F. Alqallaf, S. Van Aelst, V.J. Yohai, and R.H. Zamar, *Propagation of outliers in multivariate data*, preprint (2008), Ann. Statist. 37 (2009), pp. 311–331.
[14] U. Gather and T. Hilker, *A note on Tyler's modification of the MAD for the Stahel–Donoho estimator*, Ann. Statist. 25 (1997), pp. 2024–2026.
[15] P.J. Rousseeuw and K. Van Driessen, *A fast algorithm for the minimum covariance determinant estimator*, Technometrics 41 (1999), pp. 212–223.
[16] R. Maronna and R.H. Zamar, *Robust estimates of location and dispersion for high-dimensional datasets*, Technometrics 44 (2002), pp. 307–317.
[17] B. Flury and H. Riedwyl, *Multivariate Statistics. A Practical Approach*, Chapman and Hall, London, 1988.