Multivariate Location and Scatter Matrix Estimation Under Cellwise and Casewise Contamination

Andy Leung[a,*], Victor Yohai[b], Ruben Zamar[a]

[a] *Department of Statistics, University of British Columbia, 3182-2207 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada*

[b] *Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 1, 1426, Buenos Aires, Argentina*

Abstract

This paper considers the problem of multivariate location and scatter matrix estimation when the data contain cellwise and casewise outliers. A two-step approach was proposed to deal with this problem: first, apply a univariate filter to remove cellwise outliers and second, apply a generalized S-estimator to downweight casewise outliers. This paper improves this proposal in three main directions. First, a consistent bivariate filter is introduced to be used in combination with the univariate filter in the first step. Second, a new fast subsampling procedure is proposed to generate starting points for the generalized S-estimator in the second step. Third, a non-monotonic weight function for the generalized S-estimator is proposed to better deal with casewise outliers in high dimension. A simulation study and real data example show that, unlike the original two-step procedure, the modified two-step approach performs and scales well for high dimension. Moreover, the modified procedure outperforms the original one and other state-of-the-art robust procedures under cellwise and casewise data contamination.

Keywords:
Multivariate location and scatter, Robust estimation, Cellwise outliers, Componentwise contamination
2010 MSC: 62G35, 62G05, 62G20

1. Introduction

In this paper, we address the problem of robust estimation of multivariate location and scatter matrix under cellwise and casewise contamination.

[*] Corresponding author
Email address: andy.leung@stat.ubc.ca (Andy Leung)

Traditional robust estimators assume a casewise contamination model for the data where the majority of the cases are assumed to be free of contamination. Any case that deviates from the model distribution is then flagged as an outlier. In situations where only a small number of cases are contaminated this approach works well. However, if a small fraction of cells in a data table are contaminated but in such a way that a large fraction of cases are affected, then traditional robust estimators may fail. This problem, referred to as propagation of cellwise outliers, has been discussed by Alqallaf et al. (2009). Moreover, as pointed out by Agostinelli et al. (2015b) both types of data contamination, casewise and cellwise, may occur together.

Naturally, when data contain both cellwise and casewise outliers, the problem becomes more difficult. To address this problem, Agostinelli et al. (2015b) proposed a two-step procedure: first, apply a univariate filter (UF) to the data matrix $\mathbb{X}$ and set the flagged cells to missing values, NA's; and second, apply the generalized S-estimator (GSE) of Danilov et al. (2012) to the incomplete data set. Here, we call this two-step procedure UF-GSE. It was shown in Agostinelli et al. (2015b) that UF-GSE is simultaneously robust against cellwise and casewise outliers. However, this procedure has three limitations, which are addressed in this paper:

- The univariate filter does not handle well moderate-size cellwise outliers.
- The GSE procedure used in the second step loses robustness against casewise outliers for $p > 10$.
- The initial estimator EMVE used in the second step does not scale well to higher dimensions ($p > 10$).

Rousseeuw and Van den Bossche (2015) pointed out that to filter the variables based solely on their value may be too limiting as no correlation with other variables is taken into account. A not-so-large contaminated cell that passes the univariate filter could be flagged when viewed together with other correlated components, especially for highly correlated data. To overcome this deficiency, we introduce a consistent bivariate filter and use it in combination with UF and a new filter developed by Rousseeuw and Van den Bossche (2016) in the first step of the two-step procedure.

Maronna (2015) made a remark that UF-GSE, which uses a fixed loss function ρ in the second step, cannot handle well high-dimensional casewise outliers. S-estimators with a fixed loss function exhibit an increased Gaussian efficiency when p increases, but at the same time lose their robustness (see Rocke, 1996). Such curse of dimensionality has also been observed for UF-GSE in our simulation study. To overcome this deficiency, we constructed a new robust estimator called *Generalized Rocke S-estimator* or *GRE* to replace GSE in the second step.

The first step of filtering is generally fast, but the second step is slow due to the computation of the extended minimum volume ellipsoid (EMVE), used as initial estimate by the generalized S-estimator. The standard way to compute EMVE is by subsampling, which requires an impractically large number of subsamples when p is large, making the computation extremely slow. To reduce the high computational cost of the two-step approach in high dimension, we introduce a new subsampling procedure based on clustering. The initial estimator computed in this way is called EMVE-C.

The rest of the paper is organized as follows. In Section 2, we describe some existing filters and introduce a new consistent bivariate filter. By consistency, we mean that, when n tend to infinity and the data do not contain outliers, the proportion of data points flagged by the filter tends to zero. We also show in Section 2 how the bivariate filter can be used in combination with the other filters in the first step. In Section 3, we introduce the GRE to be used in place of GSE in the second step. In Section 4, we discuss the computational issues faced by the initial estimator, EMVE, and introduce a new cluster-based-subsampling procedure called EMVE-C. In Section 5 and 6, we compare the original and modified two-step approaches with several state-of-the-art robust procedures in an extensive simulation study. We also give there a real data example. Finally, we conclude in Section 7. The Appendix contains all the proofs. We also give a separate document called "Supplementary Material", which contains further details, simulation results, and other related material.

2. Univariate and Bivariate Filters

Consider a random sample of $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^t$, where $\boldsymbol{X}_i$ are first generated from a central parametric distribution, H_0, and then some cells, that is, some entries in $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^t$, may be independently contaminated. A *filter* $\mathcal{F}$ is a procedure that flags cells in a data table and replaces them by NA's. Let f_n be the fraction of cells in the data table flagged by the filter. A *consistent filter* for a given distribution H_0 is one that asymptotically will not flag any cell if the data come from H_0. That is, $lim_{n\to\infty} f_n = 0$ a.s. $[H_0]$.

Remark 1. *Given a collection of filters $\mathcal{F}_1, ..., \mathcal{F}_k$ they can be combined in several ways: (i) they can be united to form a new filter, $\mathcal{F}_U = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_k$, so that the resulting filter, $\mathcal{F}_U$, will flag all the cells flagged by at least one of them; (ii) they can be intersected, so that the resulting filter, $\mathcal{F}_I = \mathcal{F}_1 \cap \cdots \cap \mathcal{F}_k$, will only flag the cells identified by all of them; and (iii) a filter, $\mathcal{F}$, can be conditioned to yield a new filter, $\mathcal{F}_C$, so that $\mathcal{F}_C$ will only filter the cells filtered by $\mathcal{F}$ which satisfy a given condition C.*

Remark 2. *It is clear that $\mathcal{F}_U$ is a consistent filter provided all the filters $\mathcal{F}_i$, $i = 1, \ldots, k$ are consistent filters. On the other hand, $\mathcal{F}_I$ is a consistent filter provided at least one of the filters $\mathcal{F}_i$, $i = 1, \ldots, k$ is a consistent filter. Finally, it is also clear that if $\mathcal{F}$ is a consistent filter, so is $\mathcal{F}_C$.*

We describe now three basic filters, which will be later combined to obtain a powerful consistent filter for use in the first step of our two-step procedure.

2.1. A Consistent Univariate Filter (UF)

This is the initial filter introduced in Agostinelli et al. (2015b). Let $X_1, \ldots, X_n$ be a random (univariate) sample of observations. Consider a pair of initial location and dispersion estimators, T_{0n} and S_{0n}, such as the median and median absolute deviation (MAD) as adopted in this paper. Denote the standardized sample by $Z_i = (X_i - T_{0n})/S_{0n}$. Let F be a chosen reference distribution for Z_i. Here, we use the standard normal distribution, $F = \Phi$.

Let F_n^+ be the empirical distribution function for the absolute standardized value, that is,

$$F_n^+(t) = \frac{1}{n}\sum_{i=1}^{n} I(|Z_i| \leq t).$$

The proportion of flagged outliers is defined by

$$d_n = \sup_{t \geq \eta} \left\{F^+(t) - F_n^+(t)\right\}^+, \tag{1}$$

where $\{a\}^+$ represents the positive part of a, F^+ is the distribution of $|Z|$ when $Z \sim F$, and $\eta = (F^+)^{-1}(\alpha)$ is a large quantile of F^+. We use $\alpha = 0.95$ for univariate filtering as the aim is to detect large outliers, but other choices could be considered. Then, we flag $\lfloor n d_n \rfloor$ observations with the largest absolute standardized value, $|Z_i|$, as cellwise outliers and replace them by NA's.

The following proposition states this is a consistent filter. That is, even when the actual distribution is unknown, asymptotically, the univariate filter will not flag outliers when the tail of the chosen reference distribution is heavier than (or equal to) the tail of the actual distribution.

Proposition 1 (Agostinelli et al., 2015b)**.** *Consider a random variable $X \sim F_0$ with F_0 continuous. Also, consider a pair of location and dispersion estimators T_{0n} and S_{0n} such that $T_{0n} \to \mu_0 \in \mathbb{R}$ and $S_{0n} \to \sigma_0 > 0$ a.s. [F_0]. Let $F_0^+(t) = P_{F_0}(|\frac{X-\mu_0}{\sigma_0}| \leq t)$. If the reference distribution F^+ satisfies the inequality*

$$\max_{t \geq \eta} \left\{F^+(t) - F_0^+(t)\right\} \leq 0, \tag{2}$$

then

$$\frac{n_0}{n} \to 0 \text{ a.s.},$$

where

$$n_0 = \lfloor n d_n \rfloor.$$

We define the global univariate filter, UF, as the union of all the consistent filters described above, applied to each variable in $\mathbb{X}$. By Remarks 1 and 2, it is clear that UF is a consistent filter.

2.2. A Consistent Bivariate Filter (BF)

Let $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, with $\boldsymbol{X}_i = (X_{i1}, X_{i2})^t$, be a random sample of bivariate observations. Consider also a pair of initial location and scatter estimators,

$$\boldsymbol{T}_{0n} = \begin{pmatrix} T_{0n,1} \\ T_{0n,2} \end{pmatrix} \quad \text{and} \quad \boldsymbol{C}_{0n} = \begin{pmatrix} C_{0n,11} & C_{0n,12} \\ C_{0n,21} & C_{0n,22} \end{pmatrix}.$$

Similar to the univariate case we use the coordinate-wise median and the bivariate Gnanadesikan-Kettenring estimator with MAD scale (Gnanadesikan and Kettenring, 1972) for $\boldsymbol{T}_{0n}$ and $\boldsymbol{C}_{0n}$, respectively. More precisely, the initial scatter estimators are defined by

$$C_{0n,jk} = \frac{1}{4}\left(\mathrm{MAD}(\{X_{ij} + X_{ik}\})^2 - \mathrm{MAD}(\{X_{ij} - X_{ik}\})^2\right),$$

where $\mathrm{MAD}(\{Y_i\})$ denotes the MAD of $Y_1, \ldots, Y_n$. Note that $C_{0n,jj} = \mathrm{MAD}(\{X_j\})^2$, which agrees with our choice of the coordinate-wise dispersion estimators. Now, denote the pairwise (squared) Mahalanobis distances by $D_i = (\boldsymbol{X}_i - \boldsymbol{T}_{0n})^t \boldsymbol{C}_{0n}^{-1} (\boldsymbol{X}_i - \boldsymbol{T}_{0n})$. Let G_n be the empirical distribution for pairwise Mahalanobis distances,

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(D_i \leq t).$$

Finally, we filter outlying points $\boldsymbol{X}_i$ by comparing $G_n(t)$ with $G(t)$, where G is a chosen reference distribution. In this paper, we use the chi-squared distribution with two degrees of freedom, $G = \chi^2_2$. The proportion of flagged bivariate outliers is defined by

$$d_n = \sup_{t \geq \eta} \{G(t) - G_n(t)\}^+ . \tag{3}$$

Here, $\eta = G^{-1}(\alpha)$, and we use $\alpha = 0.85$ for bivariate filtering since we now aim for moderate outliers, but other choices of α can be considered. Then, we flag $\lfloor n d_n \rfloor$ observations with the largest pairwise Mahalanobis distances as outlying bivariate points. Finally, the following proposition states the consistency property of the bivariate filter.

Proposition 2. *Consider a random vector $\boldsymbol{X} = (X_1, X_2)^t \sim H_0$. Also, consider a pair of bivariate location and scatter estimators $\boldsymbol{T}_{0n}$ and $\boldsymbol{C}_{0n}$ such that $\boldsymbol{T}_{0n} \to \boldsymbol{\mu}_0 \in \mathbb{R}^2$ and $\boldsymbol{C}_{0n} \to \boldsymbol{\Sigma}_0 \in \mathrm{PDS}(2)$ a.s. $[H_0]$ ($\mathrm{PDS}(q)$ is the set of all positive definite symmetric matrices of size q). Let $G_0(t) = P_{H_0}((\boldsymbol{X} - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_0) \leq t)$ and suppose that G_0 is continuous. If the reference distribution G satisfies:*

$$\max_{t \geq \eta} \{G(t) - G_0(t)\} \leq 0, \tag{4}$$

then

$$\frac{n_0}{n} \to 0 \ a.s.,$$

where

$$n_0 = \lfloor nd_n \rfloor.$$

In the next section, we will define the global univariate-and-bivariate filter, UBF, using UF and BF as building blocks.

2.3. A Consistent Univariate and Bivariate Filter (UBF)

We first apply the univariate filter from Agostinelli et al. (2015b) to each variable in $\mathbb{X}$ separately using the initial location and dispersion estimators, $\boldsymbol{T}_{0n} = (T_{0n,1}, \dots, T_{0n,p})$ and $\boldsymbol{S}_{0n} = (S_{0n,1}, \dots, S_{0n,p})$. Let $\mathbb{U}$ be the resulting auxiliary matrix of zeros and ones with zeros indicating the filtered entries in $\mathbb{X}$. We next iterate over all pairs of variables in $\mathbb{X}$ to identify outlying bivariate points which helps filtering the moderately contaminated cells.

Fix a pair of variables, (X_{ij}, X_{ik}) and set $\boldsymbol{X}_i^{(jk)} = (X_{ij}, X_{ik})$. Let $\boldsymbol{C}_{0n}^{(jk)}$ be an initial pairwise scatter matrix estimator for this pair of variables, for example, the Gnanadesikan-Kettenring estimator. Note that pairwise scatter matrices do not ensure positive definiteness of $\boldsymbol{C}_{0n}$, but this is not necessary in this case because only bivariate scatter matrix, $\boldsymbol{C}_{0n}^{(jk)}$, is required in each bivariate filtering. We calculate the pairwise Mahalanobis distances $D_i^{(jk)} = (\boldsymbol{X}_i^{(jk)} - \boldsymbol{T}_{0n}^{(jk)})^t (\boldsymbol{C}_{0n}^{(jk)})^{-1} (\boldsymbol{X}_i^{(jk)} - \boldsymbol{T}_{0n}^{(jk)})$ and perform the bivariate filtering on the pairwise distances with no flagged components from the univariate filtering: $\{D_i^{(jk)} : U_{ij} = 1, U_{ik} = 1\}$. We apply this procedure to all pairs of variables $1 \leq j < k \leq p$. Let

$$J = \left\{ (i,j,k) : D_i^{(jk)} \text{ is flagged as bivariate outlier} \right\},$$

be the set of triplets which identify the pairs of cells flagged by the bivariate filter in rows $i = 1, ..., n$. It remains to determine which cells (i,j) in row i are to be flagged as cellwise outliers. For each cell (i,j) in the data table, $i = 1, \dots, n$ and $j = 1, \dots, p$, we count the number of flagged pairs in the i-th row where cell (i,j) is involved:

$$m_{ij} = \# \left\{ k : (i,j,k) \in J \right\}.$$

Cells with large m_{ij} are likely to correspond to univariate outliers. Suppose that observation X_{ij} is not contaminated by cellwise contamination. Then m_{ij} approximately follows the binomial distribution, $Bin(\sum_{k \neq j} U_{ik}, \delta)$, under ICM, where δ is the overall proportion of cellwise outliers that were not detected by the univariate filter. We flag observation X_{ij} if

$$m_{ij} > c_{ij}, \tag{5}$$

where c_{ij} is the 0.99-quantile of $Bin(\sum_{k \neq j} U_{ik}, \delta)$. In practice we obtained good results (in both simulation and real data example) using the conservative choice $\delta = 0.10$, which is adopted in this paper.

The filter obtained as the combination of all the univariate and the bivariate filters described above is called UBF. The following argument shows that UBF is a consistent filter.

By Remarks 1 and 2, the union of all the bivariate consistent filters (from Proposition 2) is a consistent filter. Next, applying the condition described in (5) to the union of these bivariate consistent filters yields another consistent filter. Finally, the union of this with UF results in the consistent filter, UBF.

2.4. The DDC Filter

Recently, Rousseeuw and Van den Bossche (2016) proposed a new procedure to filter and impute cellwise outliers, called *DetectDeviatingCells* (DDC). DDC is a sophisticated procedure that uses correlations between variables to estimate the expected value for each cell, and then flags those with an observed value that greatly deviates from this expected value. The DDC filter exhibited a very good performance when used in the first step in our two-step procedure in our simulation. However, the DDC filter is not shown to be consistent, as needed to ensure the overall consistency of our two-step estimation procedure.

In view of that, we propose a new filter made by intersecting UBF and DDC (denoted here as UBF-DDC). By Remarks 1 and 2, UBF-DDC is consistent. Moreover, we will show in Section 5 and in Appendix B that UBF-DDC is very effective, yielding the best overall performances when used as the first step in our two-step estimation procedure.

3. Generalized Rocke S-estimators

The second step of the procedure introduces robustness against casewise outliers that went undetected in the first step. Data that emerged from the first step has missing values that correspond to potentially contaminated cells. To estimate the multivariate location and scatter matrix from that data, we use a recently developed estimator called GSE, briefly reviewed below.

3.1. Review of Generalized S-estimators

Related to $\mathbb{X}$ denote $\mathbb{U}$ the auxiliary matrix of zeros and ones, with zeros indicating the corresponding missing entries. Let $p_i = p(\boldsymbol{U}_i) = \sum_{j=1}^{p} U_{ij}$ be the actual dimension of the observed part of $\boldsymbol{X}_i$. Given a p-dimensional vector of zeros and ones $\boldsymbol{u}$, a p-dimensional vector $\boldsymbol{m}$ and a $p \times p$ matrix $\boldsymbol{A}$, we denote by $\boldsymbol{m}^{(\boldsymbol{u})}$ and $\boldsymbol{A}^{(\boldsymbol{u})}$ the sub-vector of $\boldsymbol{m}$ and the sub-matrix of $\boldsymbol{A}$, respectively, with columns and rows corresponding to the positive entries in $\boldsymbol{u}$.

Define

$$D(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{C}) = (\boldsymbol{x} - \boldsymbol{m})^t \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{m})$$

the squared Mahalanobis distance and

$$D^*(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{C}) = D(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{C}^*)$$

the normalized squared Mahalanobis distances, where $\boldsymbol{C}^* = \boldsymbol{C}/|\boldsymbol{C}|^{1/p}$, so $|\boldsymbol{C}^*| = 1$, and where $|A|$ is the determinant of A.

Let $\mathbf{\Omega}_{0n}$ be a $p \times p$ positive definite initial estimator. Given the location vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and a $p \times p$ positive definite matrix $\mathbf{\Sigma}$, we define the generalized M-scale, $s_{GS}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Omega}_{0n}, \mathbb{X}, \mathbb{U})$, as the solution in s to the following equation:

$$\sum_{i=1}^{n} c_{p(\boldsymbol{U}_i)} \rho \left(\frac{D^* \left(\boldsymbol{X}_i^{(\boldsymbol{U}_i)}, \boldsymbol{\mu}^{(\boldsymbol{U}_i)}, \mathbf{\Sigma}^{(\boldsymbol{U}_i)} \right)}{s \, c_{p(\boldsymbol{U}_i)} \left| \mathbf{\Omega}_{0n}^{(\boldsymbol{U}_i)} \right|^{1/p(\boldsymbol{U}_i)}} \right) = b \sum_{i=1}^{n} c_{p(\boldsymbol{U}_i)} \tag{6}$$

where $\rho(t)$ is an even, non-decreasing in $|t|$ and bounded loss function. The tuning constants c_k, $1 \leq k \leq p$, are chosen such that

$$E_{\Phi} \left(\rho \left(\frac{||\boldsymbol{X}||^2}{c_k} \right) \right) = b, \quad \boldsymbol{X} \sim N_k(\mathbf{0}, \boldsymbol{I}), \tag{7}$$

to ensure consistency under the multivariate normal. A common choice of ρ is the Tukey's bisquare rho function, $\rho(u) = \min(1, 1 - (1-u)^3)$, and $b = 0.5$, as also used in this paper.

A generalized S-estimator is then defined by

$$(\boldsymbol{T}_{GS}, \boldsymbol{C}_{GS}) = \arg\min_{\boldsymbol{\mu}, \mathbf{\Sigma}} s_{GS}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Omega}_{0n}, \mathbb{X}, \mathbb{U}) \tag{8}$$

subject to the constraint

$$s_{GS}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Sigma}, \mathbb{X}, \mathbb{U}) = 1. \tag{9}$$

3.2. Generalized Rocke S-estimators

Rocke (1996) showed that if the weight function $W(x) = \rho'(x)/x$ in S-estimators is non-increasing, the efficiency of the estimators tends to one when $p \to \infty$. However, this gain in efficiency is paid for by a decrease in robustness. Not surprisingly, the same phenomenon has been observed for generalized S-estimators in simulation studies. Therefore, there is a need for new generalized S-estimators with controllable efficiency/robustness trade off.

Rocke (1996) proposed that the ρ function used to compute S-estimators should change with the dimension to prevent loss of robustness in higher dimensions. The Rocke-ρ function is constructed based on the fact that for large p the scaled squared Mahalanobis distances for normal data

$$\frac{D(\boldsymbol{X}, \boldsymbol{\mu}, \mathbf{\Sigma})}{\sigma} \approx \frac{Z}{p} \quad \text{with} \quad Z \sim \chi_p^2,$$

and hence that D/σ are increasingly concentrated around one. So, to have a high enough, but not too high, efficiency, we should give a high weight to the values of D/σ near one and downweight the cases where D/σ is far from one.

Let

$$\gamma = \min \left(\frac{\chi^2(1-\alpha)}{p} - 1, 1 \right), \tag{10}$$

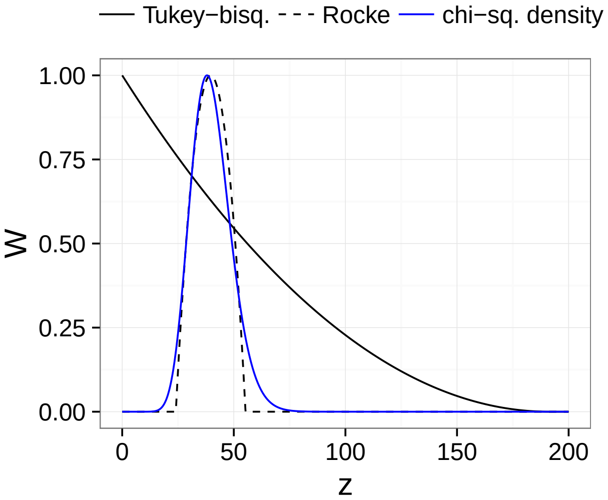

Figure 1: Weight functions of the Tukey-bisquare and the Rocke for $p = 40$. Chi-square density functions are also plotted in blue for comparison. All the functions are scaled so that their maximum is 1 to facilitate comparison.

where $\chi^2(\beta)$ is the β-quantile of χ^2_p. In this paper, we use a conventional choice of $\alpha = 0.05$ that gives an acceptable efficiency of the estimator. We have also explored smaller values of α according to Maronna and Yohai (2015), but we have seen some degree of trade-offs between efficiency and casewise robustness (see the supplementary material). Maronna et al. (2006) proposed a modification of the Rocke-ρ function, namely

$$\rho(u) = \begin{cases} 0 & \text{for} \quad 0 \le u \le 1-\gamma \\ \left(\frac{u-1}{4\gamma}\right)\left[3 - \left(\frac{u-1}{\gamma}\right)^2\right] + \frac{1}{2} & \text{for} \quad 1-\gamma < u < 1+\gamma \\ 1 & \text{for} \quad u \ge 1+\gamma \end{cases} \tag{11}$$

which has as derivative the desired weight function that vanishes for $u \notin [1-\gamma, 1+\gamma]$

$$W(u) = \frac{3}{4\gamma}\left[1 - \left(\frac{u-1}{\gamma}\right)^2\right] I(1-\gamma \le u \le 1+\gamma).$$

Figure 1 compares the Rocke-weight function, $W_{Rocke}(z/c_p)$, and the Tukey-bisquare weight function, $W_{Tukey}(z/c_p)$, for $p = 40$, where c_p as defined in (7). The chi-square density function is also plotted in blue for comparison. When p is large the tail of the Tukey-bisquare weight function greatly deviates from the tail of the chi-square density function and inappropriately assigns high weights to large distances. On the other hand, the Rocke-weight function can resemble the shape of the chi-square density function and is capable of assigning low weights to large distances.

Finally, we define the generalized Rocke S-estimators or GRE by (8) and (9) with the ρ-function in (6) replaced by the modified Rocke-ρ function in (11). We compared GRE with GSE via simulation and found that GRE has a substantial better performance in dealing with casewise outliers when p is large (e.g., $p > 10$). Results from this simulation study are provided in the supplementary material.

4. Computational Issues

The generalized S-estimators described above are computed via iterative re-weighted means and covariances, starting from an initial estimate. We now discuss some computing issues associated with this iterative procedure.

4.1. Computation of the Initial Estimator

For the initial estimate, the extended minimum volume ellipsoid (EMVE) has been used, as suggested by Danilov et al. (2012). The EMVE is computed with a large number of subsamples (> 500) to increase the chance that at least one clean subsample is obtained. Let ε be the proportion of contamination in the data and m be the subsample size. The probability of having at least one clean subsample of size m out of M subsamples is

$$q = 1 - \left[1 - \left(\begin{array}{c} n \cdot (1-\varepsilon) \\ m \end{array} \right) / \left(\begin{array}{c} n \\ m \end{array} \right)\right]^M . \tag{12}$$

For large p, the number of subsamples M required for a large q, say $q = 0.99$, can be impractically large, dramatically slowing down the computation. For example, suppose $m = p$, $n = 10p$, and $\varepsilon = 0.50$. If $p = 10$, then $M = 7758$; if $p = 30$, then $M = 2.48 \times 10^{10}$; and if $p = 50$, then $M = 4.15 \times 10^{16}$. Therefore, there is a need for a faster and more reliable starting point for large p.

Alternatively, pairwise scatter estimators could be used as fast initial estimator (e.g., Alqallaf et al., 2002). Previous simulation studies have shown that pairwise scatter estimators are robust against cellwise outliers, but they perform not as well in the presence of casewise outliers and finely shaped multivariate data (Danilov et al., 2012; Agostinelli et al., 2015b).

4.1.1. Cluster-Based Subsampling

Next, we introduce a cluster-based algorithm for faster and more reliable subsampling for the computation of EMVE. The EMVE computed with the cluster-based subsampling is called called EMVE-C throughout the paper.

High-dimensional data have several interesting geometrical properties as described in Hall et al. (2005). One such property that motivated the Rocke-ρ function, as well as the following algorithm, is that for large p the p-variate standard normal distribution $N_p(\mathbf{0}, \boldsymbol{I})$ is concentrated "near" the spherical shell with radius $\sqrt{p}$. So, if outliers have a slightly different covariance structure from clean data, they would appear geometrically different. Therefore, we could apply a clustering algorithm to first separate the outliers from the clean data. Subsampling from a big cluster, which in principle is composed of mostly clean cases, should be more reliable and require fewer number of subsamples.

Given $\mathbb{X}$ and $\mathbb{U}$. The following steps describe our clustering-based subsampling:

1. Standardize the data $\mathbb{X}$ with some initial location and dispersion estimator T_{0j} and S_{0j}. Common choices for T_{0j} and S_{0j} that are also adopted in this paper are the coordinate-wise median and MAD. Denote the standardized data by $\mathbb{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^t$, where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^t$ and $Z_{ij} = (X_{ij} - T_{0j})/S_{0j}$.
2. Compute a simple robust correlation matrix estimate $\boldsymbol{R} = (R_{jk})$. Here, we use the Gnanadesikan-Kettenring estimator (Gnanadesikan and Kettenring, 1972), where
$$R_{ij} = \frac{1}{4}(S_{0jk+}^2 - S_{0jk-}^2),$$
and where S_{0jk+} is the dispersion estimate for $\{Z_{ij} + Z_{ik} | U_{ij} = 1, U_{ik} = 1\}$ and S_{0jk-} the estimate for $\{Z_{ij} - Z_{ik} | U_{ij} = 1, U_{ik} = 1\}$. We use Q_n (Rousseeuw and Croux, 1993) for the dispersion estimate.
3. Compute the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and eigenvectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ of the correlation matrix estimate
$$\boldsymbol{R} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^t,$$
where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $\boldsymbol{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p)$. Let p_+ be the largest dimension such that $\lambda_j > 0$ for $j = 1, \ldots, p_+$. Retain only the eigenvectors $\boldsymbol{E}_0 = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{p_+})$ with a positive eigenvalue.
4. Complete the standardized data $\mathbb{Z}$ by replacing each missing entry, as indicated by $\mathbb{U}$, by zero. Then, project the data onto the basis eigenvectors $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}\boldsymbol{E}_0$, and then standardize the columns of $\tilde{\boldsymbol{Z}}$, or so called principal components, using coordinate-wise median and MAD of $\tilde{\boldsymbol{Z}}$.
5. Search for a "clean" cluster C in the standardized $\tilde{\boldsymbol{Z}}$ using a hierarchical clustering framework by doing the following. First, compute the dissimilarity matrix for the principal components using the Euclidean metric. Then, apply classical hierarchical clustering (with any linkage of choice). A common choice is the Ward's linkage, which is adopted in this paper. Finally, define the "clean" cluster by the smallest sub-cluster C with a size at least $n/2$. This can be obtained by cutting the clustering tree at various heights from the top until all the clusters have size less than $n/2$.
6. Take a subsample of size n_0 from C.

With good clustering results, we can draw fewer subsamples, and equally important, we can use a larger subsample size. The current default choices in GSE are $M = 500$ subsamples of size $n_0 = (p + 1)/(1 - \alpha_{mis})$ as suggested in Danilov et al. (2012), where α_{mis} is the fraction of missing data (α_{mis} = number of missing entries $/(np)$). For the new clustering-based subsampling, we choose $M = 50$ and $n_0 = 2(p+1)/(1-\alpha_{mis})$ in view of their overall good performance in our simulation study. However, using equation (12), a more formal procedure for the choice of M and n_0 could be considered. M and n_0 could be chosen as a function of the cluster size C, the expected remaining fraction of contamination δ, and a desired level of confidence. In such case, n and ε in equation (12) should be replaced by to the size of the cluster C and the value of δ, respectively. Without clustering, ε would be chosen fairly large (e.g. $\varepsilon = 0.50$) for conservative reasons. However, with clustering, ε can be made smaller (e.g., $\varepsilon \leq 0.10$).

In general, p is the primary driver of computational time, but the procedure could also be time-consuming for large n because the number of operations required by hierarchical clustering is of order n^3. As an alternative, one may bypass the hierarchical clustering step and sample directly from the data points with the smallest Euclidean distances to the origin calculated from $\tilde{\boldsymbol{Z}}$. This is because the Euclidean distances, in principle, should approximate the Mahalanobis distances to the mean of the original data. However, our simulations show that the hierarchical clustering step is essential for the excellent performance of the estimates, and that this step entails only a small increase in real computational time, even for large n.

A recent simulation study (Maronna and Yohai, 2015) has shown that Rocke estimator starting from the the "kurtosis plus specific direction" (KSD) estimator (Peña and Prieto, 2001) estimator can attain high efficiency and high robustness for large p. The KSD estimator uses a multivariate outlier detection procedure based on finding directions that maximize or minimize the kurtosis coefficient of the respective projections. The "clean" cases that were not flagged as outliers are then used for estimating multivariate location and scatter matrix. Unfortunately, KSD is not implemented for incomplete data. The study of the adaption of KSD for incomplete data would be of interest and worth of future research.

4.2. Other Computational Issues

There is no formal proof that the recursive algorithm decreases the objective function at each iteration for the case of generalized S-estimators with a monotonic weight function (Danilov et al., 2012). This also the case for generalized S-estimators with a non-monotonic weight function. For Rocke estimators with complete data, Maronna et al. (2006, see Section 9.6.3) described an algorithm that ensures attaining a local minimum. We have adapted this algorithm for the generalized counterparts. Although we cannot provide a formal proof, we have seen so far in our experiments that the descending property of the recursive algorithms always holds.

5. Two-Step Estimation and Simulation Results

The original two-step approach for global–robust estimation under cellwise and casewise contamination is to first flag outlying cells in the data table and to replace them by NA's using a univariate filter only (shortened to UF). In the second step, the generalized S-estimator is then applied to this incomplete data. Our new version of this is to replace UF in the first step by the proposed combination of univariate-and-bivariate filter and DDC (shortened to UBF-DDC) and to replace GSE in the second step by GRE-C (i.e., GRE starting from EMVE-C). We call the new two-step procedure UBF-DDC-GRE-C. The new procedure will be made available in the `TSGS` function in the `R` package `GSE` (Leung et al., 2015).

We now conduct a simulation study similar to that in Agostinelli et al. (2015b) to compare the two-step procedures, UF-GSE as introduced in Agostinelli et al. (2015b) and UBF-DDC-GRE-C, as well as the classical correlation estimator (MLE) and several other robust estimators that showed a competitive performance under

- Cellwise contamination: SnipEM (shortened to Snip) introduced in Farcomeni (2014)
- Casewise contamination: Rocke S-estimator as recently revisited by Maronna and Yohai (2015) and HSD introduced by Van Aelst et al. (2012)
- Cellwise and casewise contamination: DetMCDScore (shortened to DMCDSc) introduced by Rousseeuw and Van den Bossche (2015)

We also considered the different variations of the two-step procedures using different first steps, including UBF-GRE-C and DDC-GRE-C. However, UBF-DDC-GRE-C generally performs better in simulations than UBF-GRE-C and DDC-GRE-C. Therefore, we present only the results of UBF-DDC-GRE-C here. The complete results of UBF-GRE-C and DDC-GRE-C can be found in Appendix B.

We consider clean and contaminated samples from a $N_p(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$ distribution with dimension $p = 10, 20, 30, 40, 50$ and sample size $n = 10p$. The simulation mechanisms are briefly described below.

Since the contamination models and the estimators considered in our simulation study are location and scale equivariant, we can assume without loss of generality that the mean, $\boldsymbol{\mu}_0$, is equal to $\mathbf{0}$ and the variances in diag$(\boldsymbol{\Sigma}_0)$ are all equal to $\mathbf{1}$. That is, $\boldsymbol{\Sigma}_0$ is a correlation matrix.

Since the cellwise contamination model and the estimators are not affine-equivariant, we consider the two different approaches to introduce correlation structures:

- Random correlation as described in Agostinelli et al. (2015b) and
- First order autoregressive correlation.

The random correlation structure generally has small correlations, especially with increasing p. For example, for $p = 10$, the maximum correlation values have an average of 0.49, and for $p = 50$, the average maximum is 0.28. So, we consider the first order autoregressive correlation (AR1) with higher correlations, in which the correlation matrix has entries

$$\Sigma_{0,jk} = \rho^{|j-k|},$$

with $\rho = 0.9$.

We then consider the following scenarios:

- Clean data: No further changes are done to the data.
- Cellwise contamination: We randomly replace a ϵ of the cells in the data matrix by $X_{ij}^{cont} \sim N(k, 0.1^2)$, where $k = 1, 2, \ldots, 10$.
- Casewise contamination: We randomly replace a ϵ of the cases in the data matrix by $\boldsymbol{X}_i^{cont} \sim 0.5N(c\boldsymbol{v}, 0.1^2\boldsymbol{I}) + 0.5N(-c\boldsymbol{v}, 0.1^2\boldsymbol{I})$, where $c = \sqrt{k(\chi^2)_p^{-1}(0.99)}$ and $k = 1, 2, \ldots, 20$ and $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}_0$ with length such that $(\boldsymbol{v} - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{v} - \boldsymbol{\mu}_0) = 1$. Experiments show that the placement of outliers in this way is the least favorable for the proposed estimator.

We consider $\epsilon = 0.02, 0.05$ for cellwise contamination, and $\epsilon = 0.10, 0.20$ for casewise contamination. The number of replicates in our simulation study is $N = 500$.

The performance of a given scatter estimator $\boldsymbol{\Sigma}_n$ is measured by the Kulback–Leibler divergence between two Gaussian distribution with the same mean and covariances $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$:

$$D(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0) = \mathrm{trace}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}) - \log(|\boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}|) - p.$$

This divergence also appears in the likelihood ratio test statistics for testing the null hypothesis that a multivariate normal distribution has covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. We call this divergence measure the likelihood ratio test distance (LRT). Then, the performance of an estimator $\boldsymbol{\Sigma}_n$ is summarized by

$$\overline{D}(\boldsymbol{\Sigma}_n, \boldsymbol{\Sigma}_0) = \frac{1}{N} \sum_{i=1}^{N} D(\hat{\boldsymbol{\Sigma}}_{n,i}, \boldsymbol{\Sigma}_0)$$

where $\hat{\boldsymbol{\Sigma}}_{n,i}$ is the estimate at the i-th replication. Finally, the maximum average LRT distances over all considered contamination values, k, is also calculated.

Table 1 shows the maximum average LRT distances under cellwise contamination. UBF-DDC-GRE-C and UF-GSE perform similarly under random correlation, but UBF-DDC-GRE-C outperforms UF-GSE under AR1(0.9). When correlations are small, like in random correlation, the bivariate filter fails to filter moderate cellwise outliers (e.g., $k = 2$) because there is not enough information about the bivariate correlation structure in the data. Therefore, the bivariate filter gives similar results as the univariate filter. However, when correlations are large, like in AR1(0.9), the bivariate filter can filter moderate cellwise outliers and therefore, outperforms the univariate filter. This is demonstrated, for example, in Figure 2 which shows the average LRT distance behaviors for various cellwise contamination values, k.

Table 1: Maximum average LRT distances under cellwise contamination. The sample size is $n = 10p$.

Corr.	p	ϵ	MLE	Rocke	HSD	Snip	DMCDSc	UF-GSE	UBF-DDC-GRE-C
Random	10	0	0.6	1.2	0.8	5.0	1.5	0.8	1.0
		0.02	114.8	1.2	2.3	6.9	1.6	1.2	1.1
		0.05	285.4	3.6	11.2	7.5	3.2	4.5	2.5
	20	0	1.1	2.0	1.2	11.5	2.0	1.3	1.8
		0.02	146.1	2.7	10.6	13.9	2.6	4.0	2.5
		0.05	375.9	187.2	57.1	15.5	9.3	11.0	7.3
	30	0	1.6	2.8	1.7	16.7	2.6	1.9	3.3
		0.02	179.0	23.1	22.6	18.5	4.4	5.8	5.0
		0.05	475	380.5	123.1	20.8	13.7	14.2	13.3
	40	0	2.1	3.6	2.3	20.7	3.2	2.4	5.8
		0.02	215.1	121.3	38.9	22.6	6.0	7.3	8.8
		0.05	>500	>500	212.4	25.8	17.9	16.6	18.6
	50	0	2.7	4.4	2.8	25.4	3.8	2.9	4.9
		0.02	249.0	192.8	58.7	27.1	8.1	9.1	12.1
		0.05	>500	>500	298.7	29.7	20.7	19.6	23.8
AR1(0.9)	10	0	0.6	1.1	0.8	4.3	1.4	0.7	1.0
		0.02	149.8	1.2	0.9	4.9	1.5	0.9	1.0
		0.05	383.8	2.6	2.8	7.0	3.1	2.1	1.3
	20	0	1.1	1.9	1.2	7.8	2.1	1.2	1.7
		0.02	311.3	2.5	3.9	10.5	2.6	2.1	1.9
		0.05	>500	>500	31.3	14.3	12.3	9.3	2.5
	30	0	1.6	2.8	1.8	9.4	2.7	1.7	3.2
		0.02	475.9	71.1	10.7	13.9	5.4	4.0	3.3
		0.05	>500	>500	103.3	19.8	22.6	20.3	3.6
	40	0	2.1	3.6	2.2	10.9	3.4	2.3	5.5
		0.02	>500	222.1	22.7	16.2	8.9	6.7	5.6
		0.05	>500	>500	259.9	23.7	34.8	31.4	5.9
	50	0	2.7	4.4	2.8	13.0	4.0	2.8	5.0
		0.02	>500	>500	43.3	18.9	12.8	9.7	7.8
		0.05	>500	>500	>500	28.9	46.5	42.8	8.9

Table 2 shows the maximum average LRT distances under casewise contamination. Overall, UBF-DDC-GRE-C outperforms UF-GSE. This is because the Rocke ρ function in GRE in UBF-DDC-GRE-C is more capable of downweighting moderate casewise outliers (e.g., $10 < k < 20$) than the Tukey-bisquare ρ function in GSE in UF-GSE. Therefore, UBF-DDC-GRE-C outperforms UF-GSE under moderate casewise contamination and gives overall better results. This is demonstrated, for example, in Figure 3 which shows the average LRT distance behaviors for various casewise contamination values, k.

Table 3 shows the finite sample relative efficiency under clean samples with random correlation for the considered robust estimates, taking the MLE average LRT distances as the baseline. The results for the AR1(0.9) correlation are very similar and not shown here. As expected, UF-GSE show an increasing efficiency as p increases while UBF-DDC-GRE-C have lower efficiency. Improvements can be achieved by us-

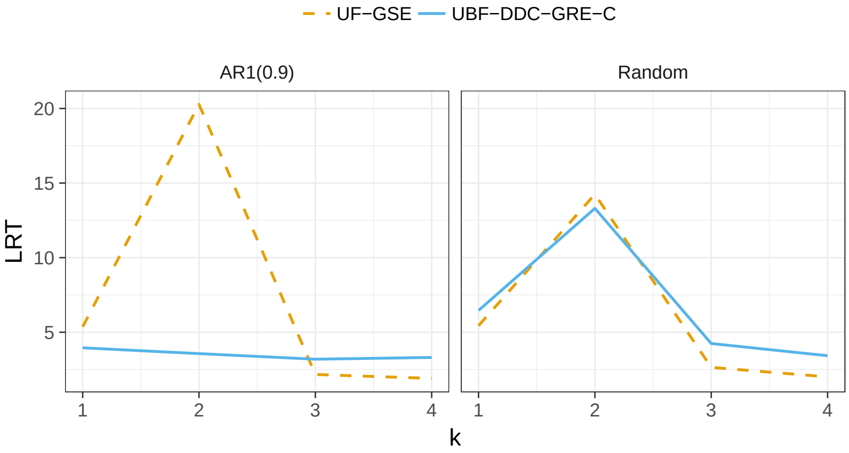

Figure 2: Average LRT distance behaviors for various contamination values, k, of UF-GSE and UBF-DDC-GSE for random and AR1(0.9) correlations under 5% cellwise contamination. The dimension is $p = 30$ and the sample size is $n = 10p$. The results remain the same for larger values of k; thus, they are not included in the figure.

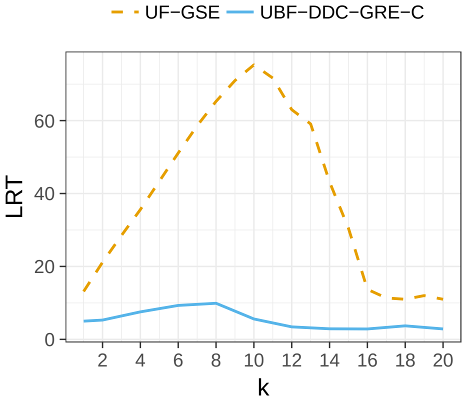

Figure 3: Average LRT distance behaviors for various contamination values, k, of UF-GSE and UBF-DDC-GRE-C for random correlations under 10% casewise contamination. The dimension is $p = 30$ and the sample size is $n = 10p$.

ing smaller α in the Rocke ρ function with some trade-off in robustness. Results from this experiment are provided in the supplementary material.

Finally, we compare the computing times of the two-step procedures. Table 4 shows the average computing times over all contamination settings for various dimensions and for $n = 10p$. The computing times for the two-step procedure have been substantially improved with the implementation of the faster initial estimator, EMVE-C.

Table 2: Maximum average LRT distances under casewise contamination. The sample size is $n = 10p$.

Corr.	p	ϵ	MLE	Rocke	HSD	Snip	DMCDSc	UF-GSE	UBF-DDC-GRE-C
Random	10	0	0.6	1.2	0.8	5.0	1.5	0.8	1.0
		0.10	43.1	2.8	3.9	44.4	4.9	9.7	7.7
		0.20	89.0	4.7	21.8	110.3	123.6	91.8	23.7
	20	0	1.1	2.0	1.2	11.5	2.0	1.3	1.8
		0.10	77.0	3.4	13.4	76.9	37.8	29.7	9.1
		0.20	146.7	5.6	95.9	166.5	187.6	291.8	17.4
	30	0	1.6	2.8	1.7	16.7	2.6	1.9	3.3
		0.10	100.0	4.3	26.1	82.3	118.6	75.3	9.9
		0.20	200.7	7.4	297.7	220.9	268.4	415.5	16.9
	40	0	2.1	3.6	2.3	20.7	3.2	2.4	5.8
		0.10	125.9	5.2	46.3	101.6	130.6	140.2	16.2
		0.20	252.4	9.1	>500	186.2	340.1	>500	19.5
	50	0	2.7	4.4	2.8	25.4	3.8	2.9	4.9
		0.10	150.3	5.9	80.0	121.9	139.5	258.1	17.6
		0.20	303.1	10.0	>500	224.3	407.7	>500	23.0
AR1(0.9)	10	0	0.6	1.1	0.8	4.3	1.4	0.7	1.0
		0.10	43.1	2.8	1.7	20.2	2.9	3.7	2.9
		0.20	88.9	4.8	8.7	49.7	29.7	50.8	6.9
	20	0	1.1	1.9	1.2	7.8	2.1	1.2	1.7
		0.10	77.0	2.8	4.7	43.8	14.8	12.9	3.3
		0.20	146.6	5.3	35.3	113.0	87.6	260.5	6.0
	30	0	1.6	2.8	1.8	9.4	2.7	1.7	3.2
		0.10	98.9	3.4	8.9	66.1	32.2	31.3	4.1
		0.20	200.5	8.2	155.5	144.8	122.9	372.7	6.8
	40	0	2.1	3.6	2.2	10.9	3.4	2.3	5.5
		0.10	124.9	4.3	15.6	83.7	49.2	69.1	6.4
		0.20	253.0	9.2	430.3	151.9	209.3	477.6	8.7
	50	0	2.7	4.4	2.8	13.0	4.0	2.8	5.0
		0.10	150.2	5.1	26.5	103.3	64.4	148.2	7.9
		0.20	302.6	10.1	>500	188.5	276.0	>500	8.8

Table 3: Finite sample efficiency for random correlations. The sample size is $n = 10p$.

p	MLE	Rocke	HSD	Snip	DMCDSc	UF-GSE	UBF-DDC-GRE-C
10	1.00	0.50	0.73	0.12	0.41	0.75	0.57
20	1.00	0.57	0.92	0.09	0.56	0.83	0.61
30	1.00	0.58	0.93	0.10	0.63	0.87	0.50
40	1.00	0.60	0.94	0.10	0.68	0.89	0.40
50	1.00	0.60	0.94	0.11	0.70	0.91	0.58

6. Real data example: small-cap stock returns data

In this section, we consider the weekly returns from 01/08/2008 to 12/28/2010 for a portfolio of 20 small-cap stocks from Martin (2013).

Table 4: Average "CPU time" – in seconds of a 2.8 GHz Intel Xeon – evaluated using the `R` command, `system.time`. The sample size is $n = 10p$.

p	UF-GSE	UBF-DDC-GRE-C
10	0.7	0.2
20	7.7	1.7
30	34.5	6.4
40	120.5	17.1
50	278.4	37.8

The purpose of this example is fourfold: first, to show that the classical MLE and traditional robust procedures perform poorly on data affected by propagation of cellwise outliers; second, to show that the two-step procedures (e.g., UF-GSE) can provide better estimates by filtering large outliers; third, that the bivariate-filter version of the two-step procedure (e.g., UBF-GSE) provides even better estimates by flagging additional moderate cellwise outliers; and fourth, that the two-step procedures that use GRE-C (e.g., UBF-GRE-C) can more effectively downweight some high-dimensional casewise outliers than those that use GSE (e.g., UBF-GSE), for this 20-dimensional dataset. Therefore, UBF-GRE-C provides the best results for this dataset.

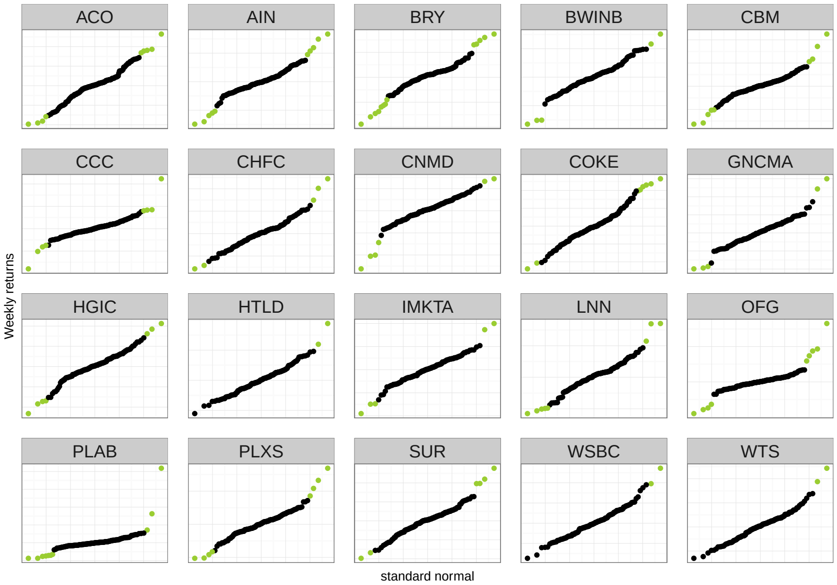

Figure 4: Normal quantile–quantile plots of weekly returns. Weekly returns that are three MAD's away from the coordinatewise-median are shown in green.

Figure 4 shows the normal QQ-plots of the 20 small-cap stocks returns in the portfolio. The bulk of the returns in all stocks seem roughly normal, but large outliers are clearly present for most of these stocks. Stocks with returns lying more than three

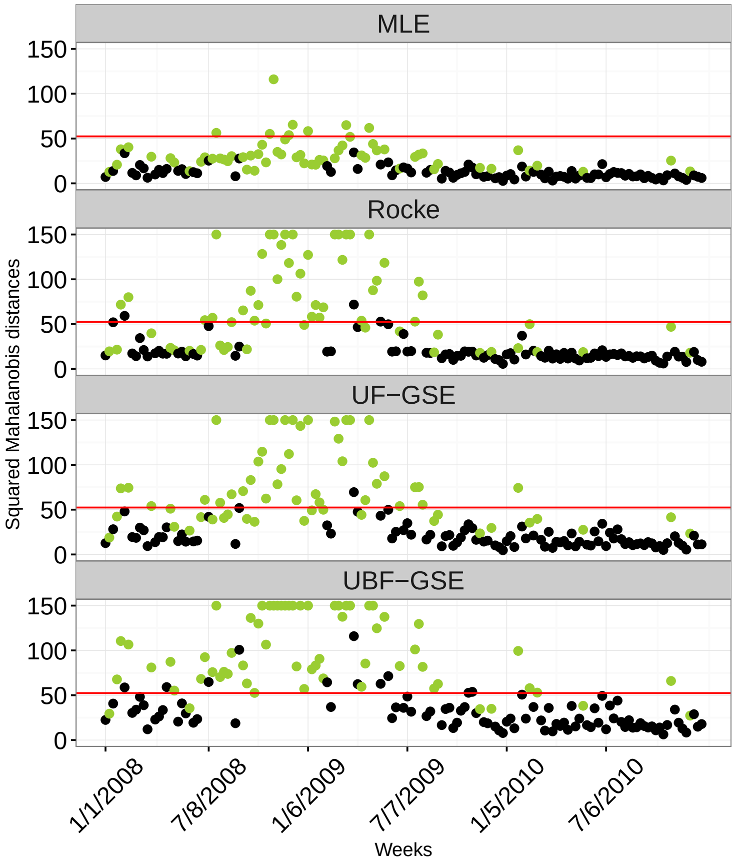

Figure 5: Squared Mahalanobis distances of the weekly observations in the small-cap asset returns data based on the MLE, the Rocke, the UF-GSE, and the UBF-GSE estimates. Weeks that contain one or more asset returns with values three MAD's away from the coordinatewise-median are in green. Large distances are truncated for better visualization.

MAD's away from the coordinatewise-median (i.e., the large outliers) are shown in green in the figure. There is a total of 4.8% large cellwise outliers that propagate to 40.1% of the cases. Over 75% of these weeks correspond to the 2008 financial crisis.

Figure 5 shows the squared Mahalanobis distances of the 157 weekly observations based on four estimates: the MLE, the Rocke-S estimates, the UF-GSE, and the UBF-GSE. Weeks that contain large cellwise outliers (asset returns with values three MAD's away from the coordinatewise-median) are in green. From the figure, we see that the MLE and the Rocke-S estimates have failed to identify many of those weeks as MD outliers (i.e., failed to flag these weeks as having estimated full Mahalanobis distance exceeding the 99.99% quantile chi-squared distribution with 20 degrees of freedom). The MLE misses all but seven of the 59 green cases. The Rocke-S estimate does slightly better but still misses one third of the green cases. This is because it is severely affected by the large cellwise outliers that propagate to 40.1% of the cases. The UF-GSE estimate also does a relatively poor job. This may be due to the

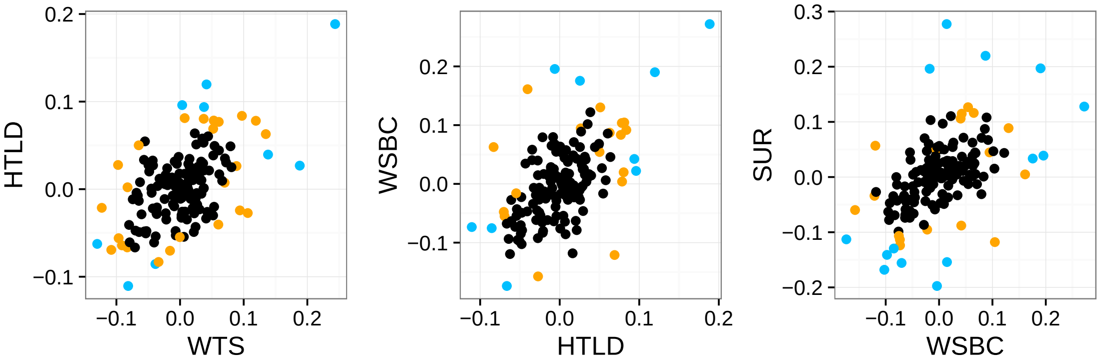

Figure 6: Pairwise scatterplots of the asset returns data for WTS versus HTLD, HTLD versus WSBC, and WSBC versus SUR. Points with components flagged by the univariate filter are in blue. Points with components additionally flagged by the bivariate filter are in orange.

presence of several moderate cellwise outliers. In fact, Figure 6 shows the pairwise scatterplots for WTS versus HTLD, HTLD versus WSBC, and WSBC versus SUR with the results from the univariate and the bivariate filter. The points flagged by the univariate filter are in blue, and those flagged by the bivariate filter are in orange. We see that the bivariate filter has identified some additional cellwise outliers that are not-so-large marginally but become more visible when viewed together with other correlated components. These moderate cellwise outliers account for 6.9% of the cells in the data and propagate to 56.7% of the cases. The final median weight assigned to these cases by UF-GSE and UBF-GSE are 0.50 and 0.65, respectively. By filtering the moderate cellwise outliers, UBF-GSE makes a more effective use of the clean part of these partly contaminated data points (i.e., the 56.7% of the cases). As a result, UBF-GSE successfully flags all but five of the 59 green cases.

Figure 7 shows the squared Mahalanobis distances produced by UBF-GRE-C and UBF-GSE, for comparison. Here, we see that UBF-GRE-C has missed only 3 of the 59 green cases, while UBF-GSE has missed 6 of the 59. UBF-GRE-C has also clearly flagged weeks 36, 59, and 66 (with final weights 0.6, 0.0, and 0.0, respectively) as casewise outliers. In contrast, UBF-GSE gives final weights 0.8, 0.5, and 0.5 to these cases. Consistent with our simulation results, UBF-GSE has difficulty downweighting some high-dimensional outlying cases on datasets of high dimension.

In this example, UBF-GRE-C makes the most effective use of the clean part of the data and has the best outlier detecting performance among the considered estimates.

7. Conclusions

In this paper, we overcome three serious limitations of UF-GSE. First, the estimator cannot deal with moderate cellwise outliers. Second, the estimator shows an incontrollable increase in Gaussian efficiency, which is paid off by a serious decrease in robustness, for larger p. Third, the initial estimator (extended minimum volume

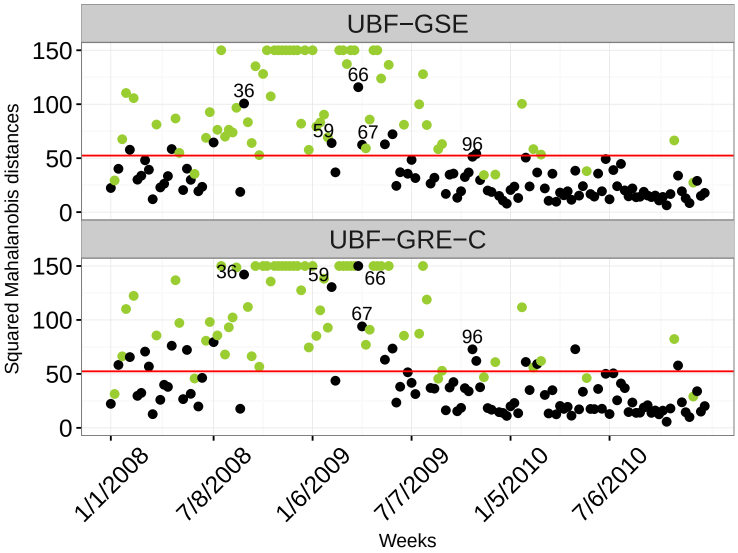

Figure 7: Squared Mahalanobis distances of the weekly observations in the small-cap asset returns data based on the UBF-GSE and the UBF-GRE-C estimates. Weeks that contain one or more asset returns with values three MAD's away from the coordinatewise-median are in green.

ellipsoids, EMVE) used by GSE and UF-GSE does not scale well in higher dimensions because it requires an impractically large number of subsamples to achieve a high breakdown point in larger dimensions.

To deal with also moderate cellwise outliers, we complement the univariate filter with a combination of bivariate filters (UBF-DDC). To achieve a controllable efficiency/robustness trade off in higher dimensions, we replace the GSE in the second step with the Rocke-type GSE which we called it GRE. Finally, to overcome the high computational cost of the EMVE, we introduce a clustering-based subsampling procedure. The proposed procedure is called UBF-DDC-GRE-C.

As shown by our simulation, UBF-DDC-GRE-C provides reliable results for cellwise contamination when $\epsilon \leq 0.05$ and $p \leq 50$. For larger dimensions ($p > 50$), in our experience, the proposed estimator still performs well unless there is a large fraction of small size cellwise outliers that evade the filter and propagate. Furthermore, UBF-DDC-GRE-C exhibits high robustness against moderate and large cellwise outliers, as well as casewise outliers in higher dimensions (e.g., $p > 10$). We also show via simulation studies that, in higher dimensions, estimators using the proposed subsampling with only 50 subsamples can achieve equivalent performance than the usual uniform subsampling with 500 subsamples.

The proposed two-step procedure still has some limitation. As pointed out in the rejoinder in Agostinelli et al. (2015a), the GSE in the second step does not handle well flat data sets, i.e., $n \approx 2p$. In fact, when $n \leq 2p$, these estimators fail to exist (cannot be computed). This is also the case for GRE-C, and for all the casewise robust estimators with breakdown point 1/2. Our numerical experiments show that the proposed two-step procedure works well when $n \geq 5p$ but not as well when $2p < n < 5p$, depending on the amount of data filtered in the first step. In this situation, if much data are filtered leaving a small fraction of complete data cases, GSE and GRE may fail to converge (Danilov et al., 2012; Agostinelli et al., 2015a). This problem could be remedied by using graphical lasso (GLASSO, Friedman et al., 2008) to improve the conditioning of the estimates.

Appendix A. Proofs of Propositions

Appendix A.1. Proof of Proposition 1

The proof was available in Agostinelli et al. (2015b), but we provide a more detailed proof in the supplementary material for completeness.

Appendix A.2. Proof of Proposition 2

We need the following lemma for the proof.

Lemma 1. *Consider a sample of p-dimensional random vectors* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$*. Also, consider a pair of multivariate location and scatter estimators* $\boldsymbol{T}_{0n}$ *and* $\boldsymbol{C}_{0n}$*. Suppose that* $\boldsymbol{T}_{0n} \to \boldsymbol{\mu}_0$ *and* $\boldsymbol{C}_{0n} \to \boldsymbol{\Sigma}_0$ *a.s.. Let* $D_i = (\boldsymbol{X}_i - \boldsymbol{T}_{0n})^t \boldsymbol{C}_{0n}^{-1} (\boldsymbol{X}_i - \boldsymbol{T}_{0n})$ *and* $D_i = (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)$*. Given* $K < \infty$*. For all* $i = 1, \ldots, n$*, if* $D_{0i} \leq K$ *, then:*

$$D_i \to D_{0i} \quad a.s..$$

Proof of Lemma 1. Note that

$$\begin{aligned}
|D_i - D_{0i}| &= |(\boldsymbol{X}_i - \boldsymbol{T}_{0n})^t \boldsymbol{C}_{0n}^{-1} (\boldsymbol{X}_i - \boldsymbol{T}_{0n}) - (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\
&= |((\boldsymbol{X}_i - \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n}))^t (\boldsymbol{\Sigma}_0^{-1} + (\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1}))((\boldsymbol{X}_i - \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})) \\
&\qquad - (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\
&\leq |(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| + |(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})^t (\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1}) (\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| \\
&\quad + |2(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| + |2(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t (\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1}) (\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| \\
&\qquad + |(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t (\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1}) (\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\
&= A_n + B_n + C_n + D_n + E_n.
\end{aligned}$$

By assumption, there exists n_1 such that for $n \geq n_1$ implies $A_n \leq \varepsilon/5$ and $B_n \leq \varepsilon/5$.

Next, note that

$$\begin{aligned}
&|(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{y}| = |\boldsymbol{y}^t \boldsymbol{\Sigma}_0^{-1/2} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\
&\quad \leq ||\boldsymbol{y}|| ||\boldsymbol{\Sigma}_0^{-1/2} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)|| = ||\boldsymbol{y}|| \sqrt{(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}_0)} \leq ||\boldsymbol{y}|| \sqrt{K}.
\end{aligned}$$

So, there exists n_2 such that $n \geq n_2$ implies

$$\begin{aligned}
C_n &= |2(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| \\
&= |2(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| \\
&\leq 2||\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})||\sqrt{K} \\
&\leq \varepsilon/5.
\end{aligned}$$

Similarly, there exists n_3 such that $n \geq n_3$ implies

$$\begin{aligned}
D_n &= |2(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t (\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| \\
&= |2(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})| \\
&\leq 2||\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\mu}_0 - \boldsymbol{T}_{0n})||\sqrt{K} \\
&\leq \varepsilon/5.
\end{aligned}$$

Also, there exists n_4 such that $n \geq n_4$ implies

$$\begin{aligned}
E_n &= |(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t (\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\
&= |(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{X}_i - \boldsymbol{\mu}_0)| \\
&\leq ||\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{X}_i - \boldsymbol{\mu}_0)||\sqrt{K} \\
&\leq ||(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})||\ ||\boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)||\sqrt{K} \\
&\leq ||(\boldsymbol{C}_{0n}^{-1} - \boldsymbol{\Sigma}_0^{-1})||K \\
&\leq \varepsilon/5.
\end{aligned}$$

Finally, let $n_5 = \max\{n_1, n_2, n_3, n_4\}$, then for all i, $n \geq n_5$ implies

$$|D_i - D_{0i}| \leq \varepsilon/5 + \varepsilon/5 + \varepsilon/5 + \varepsilon/5 + \varepsilon/5 = \varepsilon.$$

□

Proof of Proposition 2. Let $D_{0i} = (\boldsymbol{X}_i - \boldsymbol{\mu}_0)^t \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)$ and $D_i = (\boldsymbol{X}_i - \boldsymbol{T}_{0n})^t \boldsymbol{C}_{0n}^{-1}(\boldsymbol{X}_i - \boldsymbol{T}_{0n})$. Denote the empirical distributions of $D_{01}, \ldots, D_{0n}$ and $D_1, \ldots, D_n$ by

$$G_{0n}(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \leq t\right) \quad \text{and} \quad G_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(D_i \leq t\right).$$

Note that

$$
\begin{aligned}
|G_n(t) - G_{0n}(t)| &= \left| \frac{1}{n}\sum_{i=1}^{n} I\left(D_i \le t\right) - \frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t\right) \right| \\
&= \left| \frac{1}{n}\sum_{i=1}^{n} I\left(D_i \le t\right) I(D_{0i} > K) + \frac{1}{n}\sum_{i=1}^{n} I\left(D_i \le t\right) I(D_{0i} \le K) \right. \\
&\qquad \left. - \frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t\right) I(D_{0i} > K) - \frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t\right) I(D_{0i} \le K) \right| \\
&\le \left| \frac{1}{n}\sum_{i=1}^{n} I\left(D_i \le t\right) I(D_{0i} > K) - \frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t\right) I(D_{0i} > K) \right| \\
&\qquad + \left| \frac{1}{n}\sum_{i=1}^{n} I\left(D_i \le t\right) I(D_{0i} \le K) - \frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t\right) I(D_{0i} \le K) \right| \\
&= |A_n| + |B_n|.
\end{aligned}
$$

We will show that $|A_n| \to 0$ and $|B_n| \to 0$ a.s..

Choose a large K such that $P_{G_0}(D_0 > K) \le \varepsilon/8$. By law of large numbers, there exists n_1 such that for $n \ge n_1$ implies $|\frac{1}{n}\sum_{i=1}^{n} I(D_{0i} > K) - P_{G_0}(D_0 > K)| \le \varepsilon/8$ and

$$
\begin{aligned}
|A_n| &= \left| \frac{1}{n}\sum_{i=1}^{n} [I\left(D_i \le t\right) - I\left(D_{0i} \le t\right)] I(D_{0i} > K) \right| \\
&\le \frac{1}{n}\sum_{i=1}^{n} \left| I\left(D_i \le t\right) - I\left(D_{0i} \le t\right) \right| I(D_{0i} > K) \\
&\le \frac{1}{n}\sum_{i=1}^{n} I(D_{0i} > K) \\
&\le P_{G_0}(D_0 > K) + \varepsilon/8 \\
&\le \varepsilon/8 + \varepsilon/8 = \varepsilon/4.
\end{aligned}
$$

By assumption, we have from Lemma 1 that $D_i \to D_{0i}$ a.s. for all i where $D_{0i} \le K$. Let $E_i = D_i - D_{0i}$. So, with probability 1, there exists n_2 such that $n \ge n_2$ implies that $-\delta \le E_i \le \delta$ for all i. Then,

$$\begin{aligned}
B_n &= \frac{1}{n}\sum_{i=1}^{n}[I\left(D_i \le t\right) - I\left(D_{0i} \le t\right)]I(D_{0i} \le K) \\
&= \frac{1}{n}\sum_{i:D_{0i}\le K}[I\left(D_i \le t\right) - I\left(D_{0i} \le t\right)] \\
&= \frac{1}{n}\sum_{i:D_{0i}\le K}[I\left(D_{0i} \le t - E_i\right) - I\left(D_{0i} \le t\right)] \\
&\le \frac{1}{n}\sum_{i:D_{0i}\le K}[I\left(D_{0i} \le t + \delta\right) - I\left(D_{0i} \le t\right)] \\
&\le \frac{1}{n}\sum_{i=1}^{n}[I\left(D_{0i} \le t + \delta\right) - I\left(D_{0i} \le t\right)].
\end{aligned}$$

Also,

$$\begin{aligned}
B_n &= \frac{1}{n}\sum_{i:D_{0i}\le K}[I\left(D_{0i} \le t - E_i\right) - I\left(D_{0i} \le t\right)] \\
&\ge \frac{1}{n}\sum_{i:D_{0i}\le K}[I\left(D_{0i} \le t - \delta\right) - I\left(D_{0i} \le t\right)] \\
&\ge \frac{1}{n}\sum_{i=1}^{n}[I\left(D_{0i} \le t - \delta\right) - I\left(D_{0i} \le t\right)]
\end{aligned}$$

Now, by the Gilvenko–Cantelli Theorem, with probability one there exists n_3 such that $n \ge n_3$ implies that $\sup_t |\frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t + \delta\right) - G_0(t+\delta)| \le \varepsilon/16$, $\sup_t |\frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t - \delta\right) - G_0(t-\delta)| \le \varepsilon/16$, and $\sup_t |\frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t\right) - G_0(t)| \le \varepsilon/16$. Also, by the uniform continuity of G_0, there exists $\delta > 0$ such that $|G_0(t+\delta) - G_0(t)| \le \varepsilon/8$ and $|G_0(t-\delta) - G_0(t)| \le \varepsilon/8$. Together,

$$\begin{gathered}
\frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t - \delta\right) - I\left(D_{0i} \le t\right) \le B_n \le \frac{1}{n}\sum_{i=1}^{n} I\left(D_{0i} \le t + \delta\right) - I\left(D_{0i} \le t\right) \\
G_0(t-\delta) - \varepsilon/16 - G_0(t) - \varepsilon/16 \le B_n \le G_0(t+\delta) + \varepsilon/16 - G_0(t) + \varepsilon/16 \\
(G_0(t-\delta) - G(t)) - \varepsilon/8 \le B_n \le (G_0(t+\delta) - G_0(t)) + \varepsilon/8 \\
-\varepsilon/8 - \varepsilon/8 = -\varepsilon/4 \le B_n \le \varepsilon/8 + \varepsilon/8 = \varepsilon/4.
\end{gathered}$$

Finally, note that

$$G(t) - G_n(t) = (G(t) - G_0(t)) + (G_0(t) - G_{0n}(t)) + (G_{0n}(t) - G_n(t)).$$

Let $n_4 = \max\{n_1, n_2, n_3\}$, then $n \ge n_4$ implies

$$\begin{aligned}
\sup_{t>\eta}(G(t) - G_n(t)) &\le \sup_{t>\eta}(G(t) - G_0(t)) + \sup_{t>\eta}(G_0(t) - G_{0n}(t)) + \sup_{t>\eta}(G_{0n}(t) - G_n(t)) \\
&\le (\varepsilon/4 + \varepsilon/4) + \varepsilon/16 + 0 \le \varepsilon.
\end{aligned}$$

□

Appendix B. Additional Tables from the Simulation Study in Section 5

Table B.5: Maximum average LRT distances under cellwise contamination. The sample size is $n = 10p$.

Corr.	p	ϵ	UBF-	DDC-	UBF-DDC-
			GRE-C	GRE-C	GRE-C
Random	10	0	1.3	1.0	1.0
		0.02	1.4	1.1	1.1
		0.05	2.5	2.6	2.5
	20	0	2.0	1.8	1.8
		0.02	3.0	2.5	2.5
		0.05	8.2	7.7	7.3
	30	0	3.9	3.5	3.3
		0.02	5.9	5.3	5.0
		0.05	13.4	14.2	13.3
	40	0	6.2	5.8	5.8
		0.02	10.9	9.5	8.8
		0.05	19.9	18.8	18.6
	50	0	5.3	4.9	4.9
		0.02	12.9	12.5	12.1
		0.05	23.6	24.4	23.8
AR1(0.9)	10	0	1.2	1.1	1.0
		0.02	1.3	1.1	1.0
		0.05	1.4	1.3	1.3
	20	0	1.9	1.8	1.7
		0.02	2.1	2.0	1.9
		0.05	2.8	2.1	2.5
	30	0	3.4	3.6	3.2
		0.02	3.4	3.5	3.3
		0.05	5.5	3.4	3.6
	40	0	5.7	5.8	5.5
		0.02	5.7	6.0	5.6
		0.05	12.4	6.1	5.9
	50	0	5.2	4.6	5.0
		0.02	6.4	6.4	7.8
		0.05	20.4	7.9	8.9

Table B.6: Maximum average LRT distances under casewise contamination. The sample size is $n = 10p$.

Corr.	p	ϵ	UBF- GRE-C	DDC- GRE-C	UBF-DDC- GRE-C
Random	10	0	1.3	1.0	1.0
		0.10	19.1	9.4	7.7
		0.20	53.0	25.3	23.7
	20	0	2.0	1.8	1.8
		0.10	20.9	9.5	9.1
		0.20	49.3	18.0	17.4
	30	0	3.9	3.5	3.3
		0.10	21.8	10.6	9.9
		0.20	47.6	18.7	16.9
	40	0	6.2	5.8	5.8
		0.10	29.5	17.7	16.2
		0.20	52.3	21.2	19.5
	50	0	5.3	4.9	4.9
		0.10	43.4	21.2	17.6
		0.20	64.8	23.7	23.0
AR1(0.9)	10	0	1.2	1.1	1.0
		0.10	3.6	3.0	2.9
		0.20	8.4	6.8	6.9
	20	0	1.9	1.8	1.7
		0.10	4.3	3.3	3.3
		0.20	10.5	6.0	6.0
	30	0	3.4	3.6	3.2
		0.10	5.1	4.2	4.1
		0.20	13.3	6.9	6.8
	40	0	5.7	5.8	5.5
		0.10	7.3	5.8	6.4
		0.20	17.4	8.9	8.7
	50	0	5.2	4.6	5.0
		0.10	8.1	7.5	7.9
		0.20	21.2	10.0	8.8

Appendix C. Supplementary Materials

Additional simulation results and related supplementary material referenced in the article can be found in a separate document, “Supplementary Material”.

Acknowledgement

Ruben Zamar and Andy Leung research were partially funded by the Natural Science and Engineering Research Council of Canada.

References

Agostinelli, C., Leung, A., Yohai, V. J., Zamar, R. H., 2015a. Rejoinder on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST 24 (3), 484–488.

Agostinelli, C., Leung, A., Yohai, V. J., Zamar, R. H., 2015b. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST 24 (3), 441–461.

Alqallaf, F., Van Aelst, S., Yohai, V. J., Zamar, R. H., 2009. Propagation of outliers in multivariate data. Ann Statist 37 (1), 311–331.

Alqallaf, F. A., Konis, K. P., Martin, R. D., Zamar, R. H., 2002. Scalable robust covariance and correlation estimates for data mining. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. KDD ’02. pp. 14–23.

Danilov, M., Yohai, V. J., Zamar, R. H., 2012. Robust estimation of multivariate location and scatter in the presence of missing data. J Amer Statist Assoc 107, 1178–1186.

Farcomeni, A., 2014. Robust constrained clustering in presence of entry-wise outliers. Technometrics 56, 102–111.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.

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

Hall, P., Marron, J., Neeman, A., 2005. Geometric representation of high dimension, low sample size data. J R Stat Soc Ser B Stat Methodol 67, 427–444.

Leung, A., Danilov, M., Yohai, V., Zamar, R., 2015. GSE: Robust Estimation in the Presence of Cellwise and Casewise Contamination and Missing Data. R package version 3.2.3.

Maronna, R. A., 2015. Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST 24 (3), 471–472.

Maronna, R. A., Martin, R. D., Yohai, V. J., 2006. Robust Statistics: Theory and Methods. John Wiley & Sons, Chichister.

Maronna, R. A., Yohai, V. J., 2015. Robust and efficient estimation of high dimensional scatter and location. arXiv:1504.03389 [math.ST].

Martin, R., 2013. Robust covariances: Common risk versus specific risk outliers. Presented at the 2013 R-Finance Conference, Chicago, IL, `www.rinfinance.com/agenda/2013/talk/DougMartin.pdf`, visited 2016-08-24.

Peña, D., Prieto, F. J., 2001. Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43, 286–310.

Rocke, D. M., 1996. Robustness properties of S-estimators of multivariate location and shape in high dimension. Ann Statist 24, 1327–1345.

Rousseeuw, P. J., Croux, C., 1993. Alternatives to the median absolute deviation. J Amer Statist Assoc 88, 1273–1283.

Rousseeuw, P. J., Van den Bossche, W., 2015. Comments on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. TEST 24 (3), 473–477.

Rousseeuw, P. J., Van den Bossche, W., 2016. Detecting deviating data cells. arXiv:1601.07251v2 [stat.ME].

Van Aelst, S., Vandervieren, E., Willems, G., 2012. A Stahel-Donoho estimator based on Huberized outlyingness. Comput Statist Data Anal 56, 531–542.