

# Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination

Claudio Agostinelli<sup>1</sup>, Andy Leung<sup>2</sup>, Victor J. Yohai<sup>3</sup>,  
Ruben H. Zamar<sup>2</sup>

<sup>1</sup>Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari di Venezia, San Giobbe, Cannaregio 873, 30121 Venezia

<sup>2</sup>Department of Statistics, University of British Columbia, 3182-2207 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada

<sup>3</sup>Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 1, 1426, Buenos Aires, Argentina

June 24, 2014

## Abstract

Multivariate location and scatter matrix estimation is a cornerstone in multivariate data analysis. We consider this problem when the data may contain independent cellwise and casewise outliers. Flat data sets with a large number of variables and a relatively small number of cases are common place in modern statistical applications. In these cases global down-weighting of an entire case, as performed by traditional robust procedures, may lead to poor results. We highlight the need for a new generation of robust estimators that can efficiently deal with cellwise outliers and at the same time show good performance under casewise outliers.

## 1 Introduction

*Outliers* are a common problem for data analysts because they may have a big detrimental effect on estimation, inference and prediction. On the other

hand, outliers could be of main interest to data analysts because they may represent interesting rare cases such as rocks with an unusual composition of chemical compounds and exceptional athletes in a major league. The main goal in this article is robust estimation of multivariate location and scatter matrix in the presence of outliers. The estimation of these parameters is a corner stone in many applications such as principal component analysis, factor analysis, and multiple linear regression. Alqallaf, Van Aelst, Yohai, and Zamar (2009) introduced a new contamination model where traditional robust and affine equivariant estimators fail. To handle this new type of outliers, we propose a new method that involves two steps: a first step of outliers filtering, i.e., detection and replacement by missing values denoted by NA's, and a second step of robust estimation.

## Classical contamination model

To fix ideas, suppose that a multivariate data set is organized in a table with rows as cases and columns as variables, that is,  $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)'$ , with  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$ . The vast majority of procedures for robust analysis of multivariate data are based on the classical Tukey-Huber contamination model (THCM), where a small fraction of rows in the data table may be contaminated. In THCM the contamination mechanism is modeled as a mixture of two distributions: one corresponding to the nominal model and the other corresponding to the outliers. More precisely, THCM considers the following family of distributions:

$$\mathcal{H}_\epsilon = \{H = (1 - \epsilon)H_0 + \epsilon\tilde{H} : \tilde{H} \text{ is any distribution on } \mathbb{R}^p\} \quad (1)$$

where  $H_0$  is a central parametric distribution such as the multivariate normal  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and  $\tilde{H}$  is an unspecified outlier generating distribution. We then assume a case follows a distribution from the above family, that is  $\mathbf{X}_i \sim H$  where  $H \in \mathcal{H}_\epsilon$ . The key feature of this model is that when  $\epsilon$  is small we have  $\mathbf{X}_i \sim H_0$  most of the time, therefore detection and down-weighting of outlying cases makes sense and works well in practice. High breakdown point affine equivariant estimators such as MVE (Rousseeuw, 1985), MCD (Rousseeuw, 1985), S (Davies, 1987), MM (Tatsuoka and Tyler, 2000) and Stahel-Donoho estimators (Stahel, 1981; Donoho, 1982) proceed in this general way.

## Independent contamination model

In many applications, however, the contamination mechanism may be different in that individual components (or cells) in  $\mathbb{X}$  are independently contami-

nated. This is particularly so in the case of high dimensional data where variables are often measured separately and/or obtained from different sources. For instance, pathology and treatment information of a patient can be obtained from the cancer registry while epidemiological information on the patients are normally obtained through a survey. The cellwise contamination mechanism may in principle seem rather harmless, but in fact it has far reaching consequences including the possible breakdown of classical high breakdown point estimators.

The new contamination framework, called independent contamination model (ICM), was presented and formalized in Alqallaf et al. (2009). In the ICM framework we consider a different family of distribution:

$$\mathcal{J}_\epsilon = \{H : H \text{ is the distribution of } \mathbf{X} = (\mathbf{I} - \mathbf{B}_\epsilon)\mathbf{X}_0 + \mathbf{B}_\epsilon\tilde{\mathbf{X}}\}, \quad (2)$$

where  $\mathbf{X}_0 \sim H_0$ ,  $\tilde{\mathbf{X}} \sim \tilde{H}$ , and  $\mathbf{B}_\epsilon = \text{diag}(B_1, \dots, B_p)$ , where the  $B_j$  are independent  $\text{Bin}(1, \epsilon)$ . In other words, each component of  $\mathbf{X}$  has a probability  $\epsilon$  of being independently contaminated. Furthermore, the probability  $\bar{\epsilon}$  that at least one component of  $\mathbf{X}$  is contaminated is now

$$\bar{\epsilon} = 1 - (1 - \epsilon)^p.$$

This implies that even if  $\epsilon$  is small,  $\bar{\epsilon}$  could be large for large  $p$ , and could exceed the 0.5 breakdown point of highly robust affine equivariant estimators under THCM. For example, if  $\epsilon = 0.1$  and  $p = 10$ , then  $\bar{\epsilon} = 0.65$ ; if  $\epsilon = 0.05$  and  $p = 20$ , then  $\bar{\epsilon} = 0.64$  and if  $\epsilon = 0.01$  and  $p = 100$ , then  $\bar{\epsilon} = 0.63$ .

Alqallaf et al. (2009) showed that for this type of contamination the breakdown point of all the traditional 0.5 breakdown point and affine equivariant location estimators is  $1 - 0.5^{1/p} \rightarrow 0$  as  $p \rightarrow \infty$ . It can be shown that the same holds for robust and affine equivariant scatter estimators. Hence we have a new manifestation of the *curse of dimensionality*: when  $p$  is large, traditional robust estimators break down for a rather small fraction of independent contamination.

To remedy this problem, some researchers have proposed to Winsorize potential outliers for each variable separately. For instance, Alqallaf, Konis, Martin, and Zamar (2002) revisited Huberized Pairwise Covariance (Huber and Ronchetti, 1981), which is constructed by using transformed correlation coefficients calculated separately on Huberized data as basic building blocks. *Huberization* is a form of Winsorization. Although pairwise robust estimators show some robustness under ICM, they cannot deal with THCM outliers and finely shaped multivariate data. Another approach to deal with ICM outliers was proposed in Van Aelst, Vandervieren, and Willems (2012). They modified the Stahel-Donoho (SD) estimator (Stahel, 1981; Donoho, 1982)

by calculating the SD-outlyingness measure and weights on Huberized data instead of the raw data. In our simulation study this estimator performs very well under THCM but is not sufficiently robust under ICM.

An alternative approach, called *snipping* in a recent paper by Farcomeni (2014), consists of replacing cellwise outliers by NA. An interesting idea introduced in Farcomeni (2014) is the notion of optimizing over the snipping set. The use of snipping to fend against cellwise contamination has also been suggested by other authors (e.g., Danilov, 2010; Van Aelst et al., 2012). Farcomeni (2014) gives a procedure for clustering multivariate data where each cluster has an unknown location and scatter matrix. This framework can be adapted to our setting by fixing the number of clusters to one. Farcomeni (2014) suggested to first fix the proportion of cells in the data table to be snipped and then to use a maximum likelihood based procedure to obtain an optimal set of snipped cells (of the same size) together with an estimate of the location and scatter matrix for each cluster. In our simulation study this estimator performs very well under ICM but is not sufficiently robust under THCM.

A new generation of global-robust estimators that can simultaneously deal with cellwise and casewise outliers is needed. In Section 2, we introduce a global-robust estimator of multivariate location and scatter. In Section 3, we show that our estimation procedure is strongly consistent. That is, the multivariate location estimator converges a.s. to the true location and the scatter matrix estimator converges a.s. to a scalar multiple of the true scatter matrix, for a general elliptical distribution. Moreover, for a normal distribution the scalar factor is equal to one. In Section 4, we report the result of an extensive Monte Carlo simulation study. In Section 5, we analyze a real data set using the proposed and several competing estimators. In Section 6, we conclude with some remarks. Section 7 is an Appendix containing all the proofs and some additional numerical results.

## 2 Global-robust estimation under THCM and ICM

The main goal of this paper is to emphasize the need for robust estimation under ICM *and* THCM, that is, to define robust estimators that can deal with cellwise and casewise outliers.

When preprocessing multivariate data, one could try to detect cellwise outliers by applying, for instance, the “3-sigma” rule, and replace the flagged cells by NA’s. Then, an estimate of multivariate location and scatter could

be obtained using the EM-algorithm to deal with the artificially created incomplete data. One reason why this obvious preprocessing step is not routinely employed in multivariate robust estimation might be the lack of consistency of this procedure. Another reason might be that this approach is incapable of dealing with casewise outliers. These two limitations are addressed in our procedure by using an adaptive univariate filter (Gervini and Yohai, 2002) followed by Generalized S-estimator (GSE) (Danilov, Yohai, and Zamar, 2012).

More precisely, our procedure has two steps:

**Step I.** *Eliminating large cellwise outliers.* We flag cellwise outliers and replace them by NA's (this operation was called snipping in Farcomeni (2014)). In our case, this step prevents cellwise contaminated cases from having large robust Mahalanobis distances in the second step. See Section 2.1.

**Step II.** *Dealing with high-dimensional casewise outliers.* We apply GSE, which has been specifically designed to deal with incomplete multivariate data with casewise outliers, to the filtered data coming from Step I. See Section 2.2.

Full account of these steps is provided in the remaining of this section.

## 2.1 Step I: Eliminating large cellwise outliers

Consider a random sample of  $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)'$ , where  $\mathbf{X}_i$  follows a distribution from  $\mathcal{S}_\epsilon$  in (2). In addition, consider a pair of initial location and dispersion estimator,  $\mathbf{T}_{0n} = (T_{0n,1}, \dots, T_{0n,p})$  and  $\mathbf{S}_{0n} = (S_{0n,1}, \dots, S_{0n,p})$ . A common choice for  $\mathbf{T}_{0n}$  and  $\mathbf{S}_{0n}$  that are also adopted in this paper are the coordinate-wise median and median absolute deviation (mad).

Instead of a fixed cutoff value, we introduce an adaptive cutoff (Gervini and Yohai, 2002) which is asymptotically "correct", meaning that for clean data the fraction of flagged outliers tends to zero as the sample size  $n$  tends to infinity. We identify potential outliers on each variable separately using the following GY-univariate filter.

We first fix a variable  $(X_{1j}, X_{2j}, \dots, X_{nj})$  and denote the standardized version of  $X_{ij}$  by  $Z_{ij} = (X_{ij} - T_{0n,j})/S_{0n,j}$ . Let  $F_j$  be a chosen reference distribution for  $Z_{ij}$ . An ideal choice for a reference distribution would be  $F_{0j}$ , the actual distribution of  $(X_{ij} - \mu_{0j})/\sigma_{0j}$ . Unfortunately, the actual distribution of  $Z_{ij}$  is never known in practice. Thus, we use the standard normal,  $F_j = \Phi$ , as a good approximation.

The adaptive cutoff values are defined as follows. Let  $\widehat{F}_{n,j}^+$  be the empirical distribution function for absolute standardized value, that is,

$$\widehat{F}_{n,j}^+(t) = \frac{1}{n} \sum_{i=1}^n I(|Z_{ij}| \leq t).$$

The proportion of flagged outliers is defined by

$$\begin{aligned} d_{n,j} &= \sup_{t \geq \eta_j} \left\{ F_j^+(t) - \widehat{F}_{n,j}^+(t) \right\}^+ \\ &= \max_{i > i_0} \left\{ F_j^+(|Z|_{(i)j}) - \frac{(i-1)}{n} \right\}^+, \end{aligned} \quad (3)$$

where in general  $\{a\}^+$  represents the positive part of  $a$  and  $F^+$  is the distribution of  $|Z|$  when  $Z \sim F$ . Here  $|Z|_{(i)j}$  is the order statistics of  $|Z_{ij}|$ ,  $i_0 = \max\{i : |Z|_{(i)j} < \eta_j\}$ , and  $\eta_j = (F_j^+)^{-1}(\alpha)$  is a large quantile of  $F^+$ . We use  $\alpha = 0.95$  throughout this paper, but other choices could be considered. Then we flag  $\lfloor nd_{n,j} \rfloor$  observations with the largest standardized value as cell-wise outliers and replace them by NA's (here  $\lfloor a \rfloor$  is the largest integer less than or equal to  $a$ ). Finally, the resulting adaptive cutoff value for  $Z_{ij}$ 's is

$$t_{n,j} = \min \left\{ t : \widehat{F}_{n,j}^+(t) \geq 1 - d_{n,j} \right\}, \quad (4)$$

that is,  $t_{n,j} = Z_{(i_{n,j})j}$  with  $i_{n,j} = n - \lfloor nd_{n,j} \rfloor$ . Equivalently, we flag the  $X_{ij}$ 's with  $|Z_{ij}| \geq t_{n,j}$ .

The following proposition states that even when the actual distribution is unknown, asymptotically, the filter will not wrongly flag an outlier provided the tail of the chosen reference distribution is heavier (or equal) than that of the actual distribution.

**Proposition 2.1.** *Consider a (univariate) variable  $X$  and a pair of location and dispersion estimator  $T_{0n}$  and  $S_{0n}$ . Suppose that  $X \sim F_0((x - \mu)/\sigma)$  with  $F_0$  continuous. If the reference distribution  $F^+$  satisfies:*

$$\max_{u \geq \eta} \{ F^+(u) - F_0^+(u) \} \leq 0, \quad (5)$$

$T_{0n} \rightarrow \mu$  and  $S_{0n} \rightarrow \sigma > 0$  a.s., then

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

where

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

**Proof:** See the Appendix.

## 2.2 Step II: Dealing with high-dimensional casewise outliers

This second step introduces robustness against casewise outliers that went undetected in Step I. Data that emerges from Step I has *holes* (i.e., NA's) 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 as briefly reviewed below.

Let  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})'$ ,  $1 \leq i \leq n$  be  $p$ -dimensional i.i.d. random vectors that follow a distribution in an elliptical family  $\mathcal{E}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$  with density

$$f_{\mathbf{X}}(\mathbf{x}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \frac{1}{|\boldsymbol{\Sigma}_0|} f_0(d(\mathbf{x}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)) \quad (6)$$

where  $|A|$  is the determinant of  $A$ ,  $f_0$  is non-increasing and strictly decreasing at 0, and

$$d(\mathbf{x}, \mathbf{m}, \mathbf{C}) = (\mathbf{x} - \mathbf{m})' \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m}) \quad (7)$$

is the squared Mahalanobis distance. We also use the normalized squared Mahalanobis distances

$$d^*(\mathbf{x}, \mathbf{m}, \mathbf{C}) = d(\mathbf{x}, \mathbf{m}, \mathbf{C}^*), \quad (8)$$

where  $\mathbf{C}^* = \mathbf{C}/|\mathbf{C}|^{1/p}$ , so  $|\mathbf{C}^*| = 1$ .

Related to  $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)'$  we form the auxiliary data table of zeros and ones  $\mathbb{U} = (\mathbf{U}_1, \dots, \mathbf{U}_n)'$ . For  $1 \leq i \leq n$ ,  $\mathbf{U}_i = (U_{i1}, \dots, U_{ip})'$  is a  $p$ -dimensional random vector of zeros and ones, with ones indicating the observed entries of  $\mathbf{X}_i$ . Let  $p_i = p(\mathbf{U}_i) = \sum_{j=1}^p U_{ij}$  be the actual dimension of the observed part of  $\mathbf{X}_i$ . Given a  $p$ -dimensional vector of zeros and ones  $\mathbf{u}$ , a  $p$ -dimensional vector  $\mathbf{m}$  and a  $p \times p$  matrix  $\mathbf{A}$ , we denote by  $\mathbf{m}^{(\mathbf{u})}$  and  $\mathbf{A}^{(\mathbf{u})}$  the sub-vector of  $\mathbf{m}$  and the sub-matrix of  $\mathbf{A}$ , respectively, with columns and rows corresponding to the positive entries in  $\mathbf{u}$ .

Let  $\hat{\boldsymbol{\Omega}}$  be a  $p \times p$  positive definite initial estimator for  $\boldsymbol{\Sigma}_0$ . Given the location vector  $\boldsymbol{\mu} \in \mathbb{R}^p$  and a  $p \times p$  positive definite matrix  $\boldsymbol{\Sigma}$ , we define the generalized M-scale,  $s_{GS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \hat{\boldsymbol{\Omega}}, \mathbb{X}, \mathbb{U})$ , as the solution in  $s$  to the following equation:

$$\sum_{i=1}^n c_{p(\mathbf{U}_i)} \rho \left( \frac{d^* \left( \mathbf{X}_i^{(\mathbf{U}_i)}, \boldsymbol{\mu}^{(\mathbf{U}_i)}, \boldsymbol{\Sigma}^{(\mathbf{U}_i)} \right)}{s c_{p(\mathbf{U}_i)} \left| \hat{\boldsymbol{\Omega}}^{(\mathbf{U}_i)} \right|^{1/p(\mathbf{U}_i)}} \right) = b \sum_{i=1}^n c_{p(\mathbf{U}_i)} \quad (9)$$

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

$$E_{\Phi} \left( \rho \left( \frac{\|\mathbf{X}\|^2}{c_k} \right) \right) = b, \quad \mathbf{X} \sim N_k(\mathbf{0}, \mathbf{I}), \quad (10)$$

to ensure consistency under the multivariate normal. We consider the Tukey's bisquare rho function,  $\rho(u) = \min(1, 1 - (1 - u)^3)$ , and  $b = 0.5$  throughout this paper.

The inclusion of  $\widehat{\Omega}$  in (9) is needed to re-normalize the distances  $d^*$  to achieve robustness. A heuristic argument for the inclusion of  $\widehat{\Omega}$  is as follows. Suppose that  $\widehat{\mu} \approx \mu_0$  and  $\widehat{\Sigma} \approx \widehat{\Omega} \approx \Sigma_0$ . Then given  $\mathbf{U} = \mathbf{u}$ ,

$$\frac{d^*(\mathbf{X}^{(\mathbf{u})}, \widehat{\mu}^{(\mathbf{u})}, \widehat{\Sigma}^{(\mathbf{u})})}{c_{p(\mathbf{u})} \left| \widehat{\Omega}^{(\mathbf{u})} \right|^{1/p(\mathbf{u})}} \approx \frac{d^*(\mathbf{X}^{(\mathbf{u})}, \mu_0^{(\mathbf{u})}, \Sigma_0^{(\mathbf{u})})}{c_{p(\mathbf{u})} \left| \Sigma_0^{(\mathbf{u})} \right|^{1/p(\mathbf{u})}} \sim \frac{\|\mathbf{Y}^{(\mathbf{u})}\|^2}{c_{p(\mathbf{u})}}$$

where  $\mathbf{Y}^{(\mathbf{u})}$  is a  $p(\mathbf{u})$  dimensional random vector with an elliptical distribution. Hence,  $\|\mathbf{Y}^{(\mathbf{u})}\|^2/c_{p(\mathbf{u})}$  has M-scale of 1 for the given  $\rho$  function if  $\mathbf{Y}$  is normal, and large Mahalanobis distances can be down-weighted accordingly. Here, we use extended minimum volume ellipsoid (EMVE) for  $\widehat{\Omega}$  as suggested in Danilov et al. (2012).

Generalized S-estimator is then defined by

$$(\widehat{\mu}_{GS}, \widehat{\Sigma}_{GS}) = \arg \min_{\mu, \Sigma} s_{GS}(\mu, \Sigma, \widehat{\Omega}, \mathbb{X}, \mathbb{U}) \quad (11)$$

subject to the constraint

$$s_{GS}(\mu, \Sigma, \Sigma, \mathbb{X}, \mathbb{U}) = 1. \quad (12)$$

Under mild regularity assumptions, in the case of elliptical data with  $\mathbf{U}_i$  independent of  $\mathbf{X}_i$  (missing completely at random assumption) any solution to (11) is a consistent estimator for the shape of the scatter matrix. Moreover, in the case of normal data, any solution to (11) satisfying (12) is consistent in shape and size for the *true* covariance matrix. Proofs of these claims, as well as the formulas and the derivations of the estimating equation for GSE, can be found in Danilov et al. (2012).

Finally our two-step location and scatter estimator is defined by

$$\begin{aligned} \mathbf{T}_{1n} &= \widehat{\mu}_{GS}(\mathbb{X}, \mathbb{U}(\mathbf{t}_n)) \\ \mathbf{C}_{1n} &= \widehat{\Sigma}_{GS}(\mathbb{X}, \mathbb{U}(\mathbf{t}_n)) \end{aligned} \quad (13)$$

where  $\mathbf{t}_n = (t_{n,1}, \dots, t_{n,p})$  ( $t_{n,j}$  is defined in (4)) and

$$U_{ij}(t_{n,j}) = I \left( \left| \frac{X_{ij} - T_{0n,j}}{S_{0n,j}} \right| < t_{n,j} \right).$$



### 3 Consistency of GSE on filtered data

The missing data created in Step I is not missing at random because the missing data indicator,  $\mathbb{U}$ , depends on the original data  $\mathbb{X}$  (univariate outliers are declared missing). Therefore, the consistency of our two-step estimator cannot be directly derived from Danilov et al. (2012). However, as shown in Theorem 3.1 below, our procedure is consistent at the central model provided the fraction of missing data converges to zero. We need the following assumptions:

**Assumption 3.1.** *The function  $\rho$  is (i) non-decreasing in  $|t|$ , (ii) strictly increasing at 0, (iii) continuous, and (iv)  $\rho(0) = 0$  and (v)  $\lim_{v \rightarrow \infty} \rho(v) = 1$  (e.g. Tukey's bisquare rho function).*

**Assumption 3.2.** *The random vector  $\mathbf{X}$  follows a distribution,  $H_0$ , in the elliptical family defined by (6).*

**Assumption 3.3.** *Let  $H_0$  be the distribution of  $\mathbf{X}$  and denote  $\sigma(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  the solution in  $\sigma$  to the following equation*

$$E_{H_0} \left( \rho \left( \frac{d(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{c_p \sigma} \right) \right) = b,$$

and consider the minimization problem,

$$\min_{|\boldsymbol{\Sigma}|=1} \sigma(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{14}$$

We assume that (14) has a unique solution,  $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$ , where  $\boldsymbol{\Sigma}_{00}$  is positive definite. We also put  $\sigma_0 = \sigma(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$ .

**Assumption 3.4.** *The proportion of fully observed entries,*

$$q_n = \#\{i, 1 \leq i \leq n : p_i = p(\mathbf{U}_i(\mathbf{t}_n)) = p\}/n,$$

tends to one a.s. as  $n$  tends to infinity. Recall that  $\mathbf{t}_n$  is the vector of cutoff values and  $\mathbf{U}_i(\mathbf{t}_n)$  is the corresponding indicator of observed entries in  $\mathbf{X}_i$ .

**Remark 3.1.** *Davies (1987) showed that Assumption 3.2 implies Assumption 3.3 with  $\boldsymbol{\Sigma}_{00} = \boldsymbol{\Sigma}_0/|\boldsymbol{\Sigma}_0|$ .*

**Remark 3.2.** *By Proposition 2.1, the procedure described in Step I satisfies Assumption 3.4, provided that the marginal distributions for the distribution that generated the data have tails which are lighter than or equally light to those of the reference distribution. That is, they satisfy equation (5).*

**Theorem 3.1.** *Let  $\mathbf{X}_1, \dots, \mathbf{X}_n$  be a random sample from  $H_0$  and  $\mathbf{U}_1, \dots, \mathbf{U}_n$  be as described in Section 2.2. Suppose Assumptions 3.1–3.4 hold. Let  $(\hat{\boldsymbol{\mu}}_{GS}, \hat{\boldsymbol{\Sigma}}_{GS})$  be the GSE defined by (11)–(13). Then*

(i)  $\hat{\boldsymbol{\mu}}_{GS} \rightarrow \boldsymbol{\mu}_0$  a.s. and

(ii)  $\hat{\boldsymbol{\Sigma}}_{GS} \rightarrow \sigma_0 \boldsymbol{\Sigma}_{00}$  a.s..

(iii) When  $\mathbf{X} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ , we have  $\sigma_0 \boldsymbol{\Sigma}_{00} = \boldsymbol{\Sigma}_0$ .

**Proof:** See the Appendix.

## 4 Monte Carlo results

We conduct a Monte Carlo simulation study to assess the performance of the proposed scatter estimator. We consider contaminated samples from a  $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$  distribution. The contamination mechanisms are described below. The sample sizes are  $n = 100$  for dimension  $p = 10$  and  $n = 200$  for dimension  $p = 20$ .

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  $\text{diag}(\boldsymbol{\Sigma}_0)$  are all equal to  $\mathbf{1}$ . That is,  $\boldsymbol{\Sigma}_0$  is a correlation matrix. To account for the lack affine equivariance of the proposed estimator we consider different correlation structures. For each sample in our simulation we create a different random correlation matrix with condition number fixed at  $CN = 100$ . Correlation matrices with high condition number are less favorable for our proposed estimator. We use the following procedure to obtain random correlations with a fixed condition number  $CN$ :

1. For a fixed condition number  $CN$ , we first obtain a diagonal matrix  $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ ,  $[\lambda_1 < \lambda_2 < \dots < \lambda_p]$  with smallest eigenvalue  $\lambda_1 = 1$  and largest eigenvalue  $\lambda_p = CN$ . The remaining eigenvalues  $\lambda_2, \dots, \lambda_{p-1}$  are  $p - 2$  sorted independent random variables with a uniform distribution in the interval  $(1, CN)$ .
2. We first generate a random  $p \times p$  matrix  $\mathbf{Y}$ , which elements are independent standard normal random variables. Then we form the symmetric matrix  $\mathbf{Y}'\mathbf{Y} = \mathbf{U}\mathbf{V}\mathbf{U}'$  to obtain a random orthogonal matrix  $\mathbf{U}$ .
3. Using the results of 1 and 2 above, we construct the random covariance matrix by  $\boldsymbol{\Sigma}_0 = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$ . Notice that the condition number of  $\boldsymbol{\Sigma}_0$  is equal to the desired  $CN$ .

4. Convert the covariance matrix  $\Sigma_0$  into the correlation matrix  $\mathbf{R}_0$  as follows:

$$\mathbf{R}_0 = \mathbf{D}^{-1/2} \Sigma_0 \mathbf{D}^{-1/2}$$

where

$$\mathbf{D} = \text{diag}(\sigma_1, \dots, \sigma_p).$$

5. After the conversion to correlation matrix in step 4 above, the condition number of  $\mathbf{R}_0$  is no longer necessarily equal to  $CN$ . To remedy this problem, we consider the eigenvalue diagonalization of  $\mathbf{R}_0$

$$\mathbf{R}_0 = \mathbf{U}_0 \mathbf{\Lambda}_0 \mathbf{U}_0'. \quad (15)$$

where

$$\mathbf{\Lambda}_0 = \text{diag}(\lambda_1^{R_0}, \dots, \lambda_p^{R_0}), \quad \lambda_1^{R_0} < \lambda_2^{R_0} \dots < \lambda_p^{R_0}.$$

is the diagonal matrix formed using the eigenvalues of  $\mathbf{R}_0$ . We now re-establish the desired condition number  $CN$  by redefining

$$\lambda_p^{R_0} = CN \times \lambda_1^{R_0}$$

and using the modified eigenvalues in (15).

6. Repeat 4 and 5 until the condition number of  $\mathbf{R}_0$  is within a tolerance level (or until we reach some maximum iterations). In our Monte Carlo study convergence was reached after a few iteration in all the cases.

Two types of outliers are considered: (i) generated by THCM and (ii) generated by ICM. When the outliers are generated using THCM, we randomly replace 5% or 10% of the cases in the data matrix by  $k\mathbf{v}$ , where  $k = 1, 2, \dots, 100$  and  $\mathbf{v}$  is the eigenvector corresponding to the smallest eigenvalue of  $\Sigma_0$  with length such that  $(\mathbf{v} - \boldsymbol{\mu}_0)' \Sigma_0^{-1} (\mathbf{v} - \boldsymbol{\mu}_0) = 1$ . Monte Carlo experiments show that the placement of outliers in this direction,  $\mathbf{v}$ , is the least favorable for the proposed estimator. When the outliers are generated using ICM, we randomly replace 5% or 10% of the cells in the data matrix by the value  $k$  where  $k = 1, 2, \dots, 100$ . The number of replicates in our simulation study is  $N = 500$ .

The performance of a given scatter estimator  $\widehat{\Sigma}$  is measured by the Kulback-Leibler divergence between two Gaussian distribution with the same mean and covariances  $\Sigma$  and  $\Sigma_0$ :

$$D(\Sigma, \Sigma_0) = \text{trace}(\Sigma \Sigma_0^{-1}) - \log(|\Sigma \Sigma_0^{-1}|) - p.$$

This divergence also appears in the likelihood ratio test statistics for testing the null hypothesis that a multivariate normal distribution has covariance

matrix  $\Sigma = \Sigma_0$ . We call this divergence measure the likelihood ratio test distance (LRT). Then the performance of an estimator  $\widehat{\Sigma}$  is summarized by

$$\overline{D}(\widehat{\Sigma}, \Sigma_0) = \frac{1}{N} \sum_{i=1}^N D(\widehat{\Sigma}_i, \Sigma_0)$$

where  $\widehat{\Sigma}_i$  is the estimate at the  $i$ -th replication.

We compare the following estimators:

- (a) MVE-S, the estimator proposed by Maronna, Martin, and Yohai (2006, Section 6.7.5). It is an S-estimator with bisquare  $\rho$  function that uses as initial value of the iterative algorithm, an MVE estimator. The MVE estimator is computed by subsampling with concentration step. Once the estimator of location and covariance corresponding to one subsample are computed, the concentration step consist in computing the sample mean and sample covariance of the  $[n/2]$  observations with smallest Mahalanobis distance. MVE-S is implemented in the R package `rrcov`, function `CovSest`, option `method="bisquare"`;
- (b) FS, the S-estimator with bisquare  $\rho$  function, computed with an iterative algorithm similar to the Fast S-estimator for regression proposed by Salibián-Barrera and Yohai (2006). FS is implemented in the R package `rrcov`, function `CovSest`, option `method="sfast"`;
- (c) MCD, the fast Minimum Covariance Determinant proposed by Rousseeuw and Van Driessen (1999) ( see also Maronna et al. (2006, Section 6.7.5) ). MCD is implemented in the R package `rrcov`, function `CovMcd`;
- (d) HSD, Stahel-Donoho estimator with Huberized outlyingness proposed by Van Aelst et al. (2012). We use a `MATLAB` code kindly provided by the authors. The number of subsamples used in HSD is  $200 \times p$ ;
- (e) SnipEM, the procedure proposed in Farcomeni (2014). We use the R code kindly provided by the author. This method requires an initial specification of the position of the snipped cells in the form of a binary data table. We compared (using simulation) several possible choices for this initial set including: (a) snipping the largest 10% of the absolute standardized values for each variable; (b) snipping the largest 15% of the absolute standardized values for each variable; and (c) snipping the standardized values that are more than 1.5 times the interquartile range less the first quartile or more than 1.5 times the interquartile range plus the third quartile, for each variable. We only report the results from case (b) as it yields the best performances.

- (f) 2SGS, the two-step procedure proposed in Section 2. This estimator is available as the `TSGS` function in the R package `GSE`.

The tuning parameters for the high breakdown-point estimators MVE-S, FS, and MCD are chosen to attain 0.5 breakdown point under THCM. We have also considered pairwise scatter estimator obtained by combining bivariate S-estimator and found that this approach did not perform well in our settings (not shown here).

Table 1 shows the maximum average LRT distances from the true correlation matrices among the considered contamination sizes and both contamination models. The average LRT distances behavior for different contamination sizes  $k$  are displayed in Figures 1 and 2. We notice 2SGS has the best performance under ICM. Not surprisingly, MVE-S has the best behavior under THCM. However, 2SGS has an acceptable performance, comparable with that of main stream high breakdown point estimators designed for good performance under THCM.

Table 1: Maximum average LRT distances. Sample size is  $10 \times p$ . Results are based on 500 replicates.

Estimator	ICM				THCM			
	Dim 10		Dim 20		Dim 10		Dim 20	
	5%	10%	5%	10%	5%	10%	5%	10%
MLE	>500	>500	>500	>500	>500	>500	>500	>500
MCD	368.4	>500	>500	>500	1.8	10.0	5.8	130.9
FS	>500	>500	>500	>500	1.2	8.7	7.2	204.8
MVE-S	>500	>500	>500	>500	1.2	3.3	3.4	7.9
HSD	11.6	64.7	75.5	>500	1.4	4.5	4.1	14.8
SnipEM	7.4	10.2	14.2	18.3	13.9	30.9	34.8	61.4
2SGS	4.6	15.5	10.8	24.0	2.5	8.7	7.4	22.3

Table 2 shows the finite sample relative efficiency under clean samples for the considered robust estimates, taking the MLE average LRT distances as the baseline. Results for larger sample sizes, not reported here, show an identical pattern, except for MCD which efficiency increases with the sample size.

We also consider the barrow wheel contamination setting (Stahel and Maechler, 2009; Vakili, Hubert, and Rousseeuw, 2012) as suggested by an anonymous referee. The barrow wheel outliers are generated from a distribution that could create a large shape bias. The performance 2SGS is similar to the performance of the THCM high breakdown point estimators. The

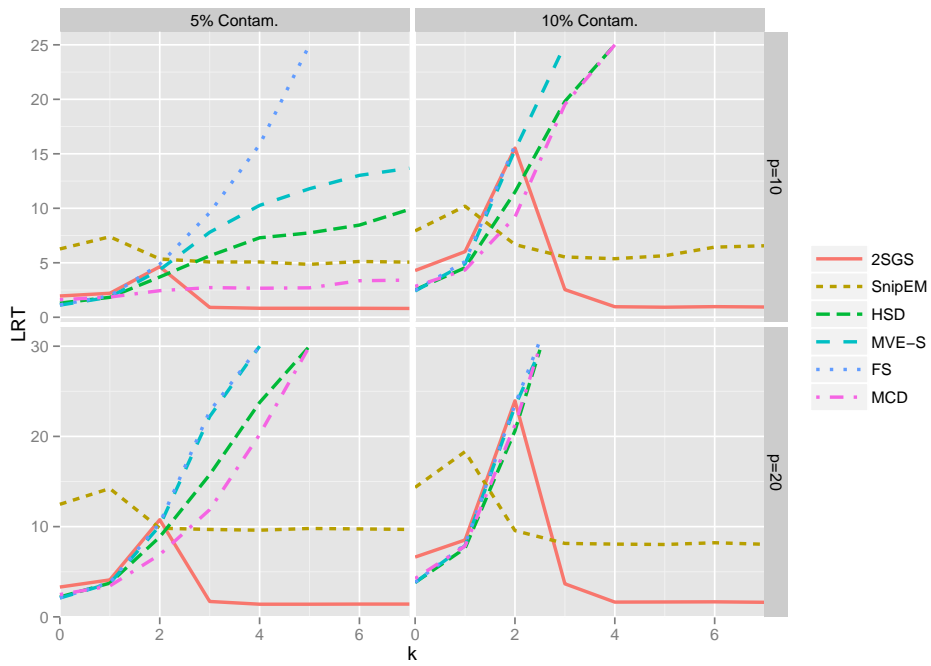


Figure 1: Average LRT distances for various contamination values,  $k$ , from ICM.

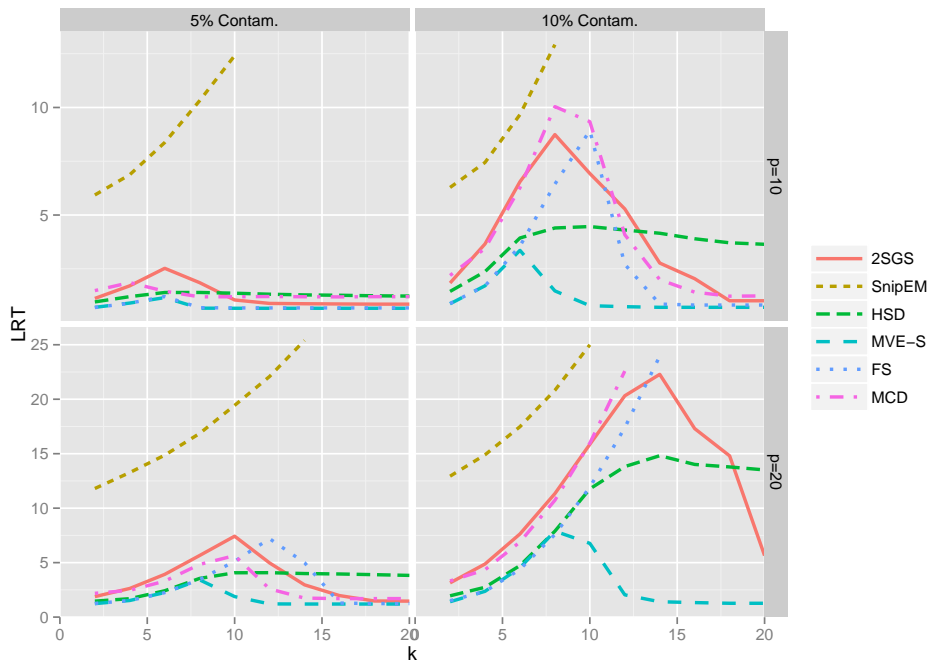


Figure 2: Average LRT distances for various contamination values,  $k$ , from THCM.

Table 2: Finite sample efficiency for several estimators measured by *relative average LRT distances* taking MLE as baseline. Sample size is  $10 \times p$ . Results are based on 500 replicates.

Estimator	$p = 10$	$p = 20$
MLE	1.00	1.00
MCD	0.47	0.66
FS	0.90	0.96
MVE-S	0.89	0.96
HSD	0.73	0.90
SnipEM	0.11	0.28
2SGS	0.81	0.84

results from this simulation as well as the computing times for our estimator (for several sample sizes and dimensions) are shown in the Appendix

## 5 Application to Chemical data

We use 20 variables from a data set analyzed by Smith, Campbell, and Licheld (1984). These variables measure the contents (in parts per million) for 20 chemical compounds in 53 samples of rocks in Western Australia. We compute several multivariate location and scatter estimates for this data.

Since we suspect the occurrence of independent contamination, we compute the  $N = 53 \times 20 = 1060$  squared standardized cellwise distances and the

$$N = 53 \times 20 \times 19/2 = 10070$$

squared Mahalanobis distances for all the pairs  $(x_{ij}, x_{ik}), i = 1, 2, \dots, 53, 1 \leq j < k \leq 20$  using the different estimates. To account for multiple comparison, cellwise and pairwise distances are compared with the thresholds  $(\chi_1^2)^{-1}(0.99^{1/(np)})$  and  $(\chi_2^2)^{-1}(0.99^{2/(np(p-1))})$ , respectively. To illustrate the phenomenon of outliers propagation, full Mahalanobis distances (using all the variables) are also computed and compared with the threshold  $(\chi_p^2)^{-1}(0.99^{1/n})$ . All distances are computed using the appropriate parts from the multivariate location and scatter matrix estimates. Table 3 shows the proportion of outliers identified using the different approaches. The proportions of identified cellwise, pairwise and casewise outliers are higher for robust estimators in the third generation. In addition, the non-robust MLE flags the smallest proportions of cellwise and pairwise outliers, and zero casewise outliers.

Table 3: Contamination summary in Chemical data based on different estimates

Estimators	Proportion of outliers		
	Cell	Pair	Case
MLE	0.007	0.008	0.000
Tyler	0.016	0.024	0.170
Rocke	0.017	0.027	0.302
MCD	0.016	0.028	0.283
MVE	0.024	0.036	0.283
FS	0.015	0.027	0.170
MVE-S	0.018	0.030	0.208
HSDE	0.025	0.038	0.302
2SGS	0.021	0.033	0.415

## 6 Conclusions

Affine equivariance, a proven asset for achieving THCM robustness, becomes a hindrance under ICM because of outliers propagation.

We advocate the practical and theoretical importance of ICM and point to the perils and drawbacks of relying solely on the THCM paradigm. ICM promotes a less aggressive cellwise down-weighting of outliers and becomes an essential tool for modeling contamination in flat data sets (large in  $p$  but relatively small in  $n$ ). Moreover, many low/moderate dimensional data sets may also be well modeled by ICM.

We introduce a two-step procedure to achieve robustness under ICM and THCM. The first step in our procedure is aimed at reducing the impact of outliers propagation and overcome the curse of dimensionality posed by ICM. The second step is aimed at achieving robustness under THCM. Our procedure is not affine equivariant but nevertheless provides fairly high resistance against both ICM and THCM outliers. Our procedure exhibits some loss of robustness under THCM, when compared with the best performing robust affine equivariant estimators in this setting.

We conjecture that the influence function of our estimator is the same as the influence function of the S-estimator for complete data. This conjecture is based on a similar result in Gervini and Yohai (2002). They showed that the similarly derived robust regression estimator has the same influence function as the least squares estimator (they used a weighted least squares in the second step and showed that the asymptotic weights are equal to one under the central normal model). The derivation in our case seems rather involved because of the added complexity introduced by the independent contamina-



tion model. Moreover, we believe that in general the influence function is not a very informative robustness measure. A bounded influence function is not a necessary nor sufficient condition for robustness under THCM and ICM.

There is a need for further research on these topics.

## 7 Appendix: Proofs

### 7.1 Proof of Proposition 2.1

Let  $\widehat{F}_n^+$  be the empirical distribution  $|Z|$  and  $\widehat{Z}$  as defined by replacing  $\mu$  and  $\sigma$  with  $T_{0n}$  and  $S_{0n}$  respectively in the definition of  $Z$ .

Note that

$$\begin{aligned} |Z - \widehat{Z}| &\leq \left| \frac{X - \mu}{\sigma} - \frac{X - T_{0n}}{S_{0n}} \right| \\ &\leq \left| \frac{X - \mu}{\sigma} - \frac{X - \mu}{S_{0n}} \right| + \frac{|T_{0n} - \mu|}{S_{0n}} \\ &\leq \widehat{A} + \widehat{B} \end{aligned}$$

where  $\widehat{A} \rightarrow 0$  a.s and  $\widehat{B} \rightarrow 0$  a.s.. By the uniform continuity of  $F^+$ , given  $\varepsilon > 0$ , there exists  $\delta > 0$  such that  $|F^+(z(1 - \delta) - \delta) - F^+(z)| \leq \varepsilon/2$ . With probability one there exists  $n_1$  such that  $n \geq n_1$  implies  $|\widehat{A}| < \delta$  and  $|\widehat{B}| < \delta$ . By the Glivenko-Cantelli Theorem, with probability one there exists  $n_2$  such that  $n \geq n_2$  implies that  $\sup_z |\widehat{F}_n^+(z) - F^+(z)| \leq \varepsilon/2$ . Let  $n_3 = \max(n_1, n_2)$ , then  $n \geq n_3$  imply

$$\begin{aligned} \widehat{F}_n^+(z) &\geq \widehat{F}_n^+(z(1 - \delta) - \delta) \\ &= \left( \widehat{F}_n^+(z(1 - \delta) - \delta) - F_0^+(z(1 - \delta) - \delta) \right) \\ &\quad + (F_0^+(z(1 - \delta) - \delta) - F_0^+(z)) + (F_0^+(z) - F^+(z)) + F^+(z) \end{aligned}$$

and then

$$\begin{aligned} \sup_{z > \eta} (F^+(z) - \widehat{F}_n^+(z)) &\leq \sup_{z > \eta} \left| F_0^+(z(1 - \delta) - \delta) - \widehat{F}_n^+(z(1 - \delta) - \delta) \right| \\ &\quad + \sup_{z > \eta} |F_0^+(z(1 - \delta) - \delta) - F_0^+(z)| \\ &\quad + \sup_{z > \eta} (F^+(z) - F_0^+(z)) \\ &\leq \varepsilon \end{aligned}$$

This implies that  $n_0/n \rightarrow 0$  a.s..

## 7.2 Proof of Theorem 3.1

We need the following Lemma proved in Yohai (1985).

**Lemma 7.1.** *Let  $\{\mathbf{Z}_i\}$  be i.i.d. random vectors taking values in  $\mathbb{R}^k$ , with common distribution  $Q$ . Let  $f : \mathbb{R}^k \times \mathbb{R}^h \rightarrow \mathbb{R}$  be a continuous function and assume that for some  $\delta > 0$  we have that*

$$E_Q \left[ \sup_{\|\lambda - \lambda_0\| \leq \delta} |f(\mathbf{Z}, \lambda)| \right] < \infty.$$

Then if  $\hat{\lambda}_n \rightarrow \lambda_0$  a.s., we have

$$\frac{1}{n} \sum_{i=1}^n f(\mathbf{Z}_i, \hat{\lambda}_n) \rightarrow E_Q [f(\mathbf{Z}, \lambda_0)] \text{ a.s..}$$

*Proof of Theorem 3.1,*

Define

$$(\hat{\boldsymbol{\mu}}_{GS}, \tilde{\boldsymbol{\Sigma}}_{GS}) = \arg \min_{\boldsymbol{\mu}, |\boldsymbol{\Sigma}|=1} s_{GS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \hat{\boldsymbol{\Omega}}). \quad (16)$$

We drop out  $\mathbb{X}$  and  $\mathbb{U}$  in the argument to simplify the notation. Since  $s_{GS}(\boldsymbol{\mu}, \lambda \boldsymbol{\Sigma}, \hat{\boldsymbol{\Omega}}) = s_{GS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \hat{\boldsymbol{\Omega}})$ , to prove Theorem 3.1 it is enough to show

(a)

$$(\hat{\boldsymbol{\mu}}_{GS}, \tilde{\boldsymbol{\Sigma}}_{GS}) \rightarrow (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00}) \text{ a.s.,} \quad \text{and} \quad (17)$$

(b)

$$s_{GS}(\hat{\boldsymbol{\mu}}_{GS}, \tilde{\boldsymbol{\Sigma}}_{GS}, \tilde{\boldsymbol{\Sigma}}_{GS}) \rightarrow \sigma_0 \text{ a.s..} \quad (18)$$

Note that since we have

$$E_{H_0} \left( \rho \left( \frac{d(\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{\sigma_0 c_p} \right) \right) = b,$$

then part (i) of Lemma 6 in the Supplemental Material of Danilov et al. (2012) implies that given  $\varepsilon > 0$ , there exists  $\delta > 0$  such that

$$\underline{\lim}_{n \rightarrow \infty} \inf_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in C_\varepsilon^c, |\boldsymbol{\Sigma}|=1} \frac{1}{n} \sum_{i=1}^n c_p \rho \left( \frac{d(\mathbf{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sigma_0 c_p (1 + \delta)} \right) > (b + \delta) c_p, \quad (19)$$

where  $C_\varepsilon$  is a neighborhood of  $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$  of radius  $\varepsilon$  and if  $A$  is a set, then  $A^C$  denotes its complement. In addition, by part (iii) of the same Lemma we have for any  $\delta > 0$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n c_p \rho \left( \frac{d(\mathbf{X}_i, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})}{\sigma_0 c_p (1 + \delta)} \right) < b c_p. \quad (20)$$

Let

$$Q_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = c_p \rho \left( \frac{d(\mathbf{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sigma_0 c_p (1 + \delta)} \right)$$

and

$$Q_i^{(\mathbf{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = c_{p(\mathbf{U}_i)} \rho \left( \frac{d^* \left( \mathbf{X}_i^{(\mathbf{U}_i)}, \boldsymbol{\mu}^{(\mathbf{U}_i)}, \boldsymbol{\Sigma}^{(\mathbf{U}_i)} \right)}{S c_{p(\mathbf{U}_i)} \left| \widehat{\boldsymbol{\Omega}}^{(\mathbf{U}_i)} \right|^{1/p(\mathbf{U}_i)}} \right),$$

Now if  $|\boldsymbol{\Sigma}| = 1$  and  $S = \sigma_0(1 + \delta)/|\widehat{\boldsymbol{\Omega}}|^{1/p}$ , we have

$$\frac{1}{n} \sum_{i=1}^n Q_i^{(\mathbf{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{n} \sum_{p_i=p} Q_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{n} \sum_{p_i \neq p} Q_i^{(\mathbf{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (21)$$

We also have

$$\frac{1}{n} \sum_{p_i \neq p} Q_i^{(\mathbf{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq c_p (1 - t_n) \quad (22)$$

and therefore by Assumption 3.4 we have

$$\lim_{n \rightarrow \infty} \sup_{\boldsymbol{\mu}, |\boldsymbol{\Sigma}|=1} \frac{1}{n} \sum_{p_i \neq p} Q_i^{(\mathbf{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \text{ a.s.} \quad (23)$$

Similarly we can prove that

$$\lim_{n \rightarrow \infty} \sup_{\boldsymbol{\mu}, |\boldsymbol{\Sigma}|=1} \frac{1}{n} \sum_{p_i \neq p} Q_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \text{ a.s.} \quad (24)$$

and

$$c_p - \frac{1}{n} \sum_{i=1}^n c_{p(\mathbf{U}_i)} \rightarrow 0, \text{ a.s.} \quad (25)$$

Then, from (19) and (21)–(25) we get

$$\underline{\lim}_{n \rightarrow \infty} \inf_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in C_\varepsilon^C, |\boldsymbol{\Sigma}|=1} \frac{1}{n} \sum_{i=1}^n Q_i^{(\mathbf{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) > (b + \delta) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n c_{p(\mathbf{U}_i)} = (b + \delta) c_p \text{ a.s.} \quad (26)$$

Using similar arguments, from (20) we can prove

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Q_i^{(\mathbf{U})}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00}) < b \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n c_{p(\mathbf{U}_i)} = b c_p \text{ a.s.} \quad (27)$$

Equations (26)–(27) imply that

$$\liminf_{n \rightarrow \infty} \inf_{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in C_\varepsilon^C, |\boldsymbol{\Sigma}|=1} s_{GS}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{\boldsymbol{\Omega}}) > S \text{ a.s.}$$

and

$$\lim_{n \rightarrow \infty} s_{GS}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00}, \widehat{\boldsymbol{\Omega}}) < S \text{ a.s..}$$

Therefore, with probability one there exists  $n_0$  such that for  $n > n_0$  we have  $(\widehat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}) \in C_\varepsilon^C$ . Then  $(\widehat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}) \rightarrow (\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$  a.s. proving (a).

Let

$$P_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}, s) = c_p \rho \left( \frac{d(\mathbf{X}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{c_p s} \right)$$

and

$$P_i^{(\mathbf{U})}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, s) = c_{p(\mathbf{U}_i)} \rho \left( \frac{d(\mathbf{X}_i^{(\mathbf{U}_i)}, \boldsymbol{\mu}^{(\mathbf{U}_i)}, \boldsymbol{\Sigma}^{(\mathbf{U}_i)})}{c_{p(\mathbf{U}_i)} s} \right).$$

Since  $|\widetilde{\boldsymbol{\Sigma}}_{GS}| = 1$ , we have that  $s_{GS}(\widehat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS})$  is the solution in  $s$  in the following equation

$$\frac{1}{n} \sum_{i=1}^n P_i^{(\mathbf{U})}(\widehat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, s) = \frac{b}{n} \sum_{i=1}^n c_{p(\mathbf{U}_i)}. \quad (28)$$

Then, to prove (18) it is enough to show that for all  $\varepsilon > 0$

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P_i^{(\mathbf{U})}(\widehat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \sigma_0 + \varepsilon) &< b c_p \text{ a.s.} \quad \text{and} \\ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P_i^{(\mathbf{U})}(\widehat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \sigma_0 - \varepsilon) &> b c_p \text{ a.s.} \end{aligned} \quad (29)$$

Using Assumption 3.4, to prove (29) it is enough to show

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P_i(\widehat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \sigma_0 + \varepsilon) &< b c_p \text{ a.s.} \quad \text{and} \\ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n P_i(\widehat{\boldsymbol{\mu}}_{GS}, \widetilde{\boldsymbol{\Sigma}}_{GS}, \sigma_0 - \varepsilon) &> b c_p \text{ a.s.} \end{aligned} \quad (30)$$

It is immediate that

$$E \left( \rho \left( \frac{d(\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{c_p (\sigma_0 + \varepsilon)} \right) \right) < E \left( \rho \left( \frac{d(\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{c_p \sigma_0} \right) \right) = b$$

and

$$E \left( \rho \left( \frac{d(\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{c_p (\sigma_0 - \varepsilon)} \right) \right) > E \left( \rho \left( \frac{d(\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{c_p \sigma_0} \right) \right) = b.$$

Then equations (30) follow from Lemma 7.1 and part (a). This proves (b).

### 7.3 Investigation on the performance on the barrow wheel outliers

An anonymous referee suggested considering the performance of 2SGS under the barrow wheel contamination setting (Stahel and Maechler, 2009; Vakili et al., 2012). We conduct a Monte Carlo study to compare the performance of 2SGS with three second generation estimators under 5% and 10% of outliers from the barrow wheel distribution. The data is generated using the R package `robustX` with default parameters. The three second generation estimators are: the fast Minimum Covariance Determinant (MCD), the fast S-estimator (FS), and the S-estimator (S), described in Section 4. The sample size are  $n = 10 \times p$ , for  $p = 10$  and 20. The results in terms of the LRT measure are graphically displayed in Figure 3.

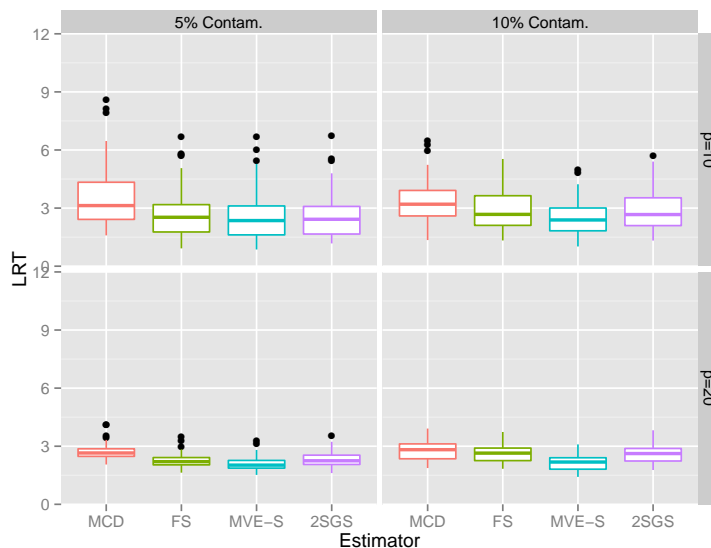


Figure 3: LRT distances under barrow-wheel contamination setting.

## 7.4 Timing experiment

Table 4 shows the mean time needed to compute 2SGS for data with cellwise or casewise outliers as described in Section 5. We consider 10% contamination and several sample sizes and dimensions. We use the random correlation structures as described in Section 4. For each pair of dimension and sample size, we average the computing times over 250 replications for each of the following setups: (a) cellwise contamination with  $k$  generated from  $U(0, 6)$  and (b) casewise contamination with  $k$  generated from  $U(0, 20)$ .

Table 4: Average “CPU time” – in seconds of a 2.8 GHz Intel Xeon – evaluated using the R command, `system.time`.

$p$	$n$	Cellwise	Casewise
5	50	0.03	0.03
	100	0.04	0.04
10	100	0.12	0.10
	200	0.17	0.13
15	150	0.40	0.28
	300	0.60	0.40
20	200	1.03	0.73
	400	1.88	1.06
25	250	2.52	1.62
	500	4.58	2.45
30	300	5.08	3.26
	600	8.47	5.16
35	350	9.30	6.13
	700	15.64	9.79

**Acknowledgement** Victor Yohai research was partially supported by Grants W276 from Universidad of Buenos Aires, PIP 112-2008-01-00216 and 112-2011-01-00339 from CONICET and PICT2011-0397 from ANPCYT, Argentina. Ruben Zamar and Andy Leung research were partially funded by the Natural Science and Engineering Research Council of Canada.

## References

Alqallaf, F., Van Aelst, S., Yohai, V.J., and Zamar, R.H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331.

- Alqallaf, F.A., Konis, K.P., Martin, R.D., and 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, pages 14–23, New York, NY, USA. ACM. ISBN 1-58113-567-X. doi: 10.1145/775047.775050. URL <http://doi.acm.org/10.1145/775047.775050>.
- Danilov, M. (2010). *Robust Estimation of Multivariate Scatter under Non-Affine Equivariant Scenarios*. PhD thesis, University of British Columbia.
- Danilov, M., Yohai, V.J., and Zamar, R.H. (2012). Robust estimation of multivariate location and scatter in the presence of missing data. *Journal of the American Statistical Association*, 107:1178–1186.
- Davies, P. (1987). Asymptotic behaviour of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15: 1269–1292.
- Donoho, D.L. (1982). *Breakdown Properties of Multivariate Location Estimators*. PhD thesis, Harvard University.
- Farcomeni, A. (2014). Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, 56(1):102–111.
- Gervini, D. and Yohai, V.J. (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics*, 30(2):583–616.
- Huber, P.J. and Ronchetti, E.M. (1981). *Robust Statistics (2nd edition)*. John Wiley & Sons, New Jersey.
- Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006). *Robust Statistic: Theory and Methods*. John Wiley & Sons, Chichester.
- Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 8:283–297.
- Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.
- Salibian-Barrera, M. and Yohai, V.J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2):414–427.

- Smith, R.E., Campbell, N.A., and Licheld, A (1984). Multivariate statistical techniques applied to pisolitic laterite geochemistry at Golden Grove, Western Australia. *Journal of Geochemical Exploration*, 22:193–216.
- Stahel, W.A. (1981). Breakdown of covariance estimators. Technical Report 31, Fachgruppe für Statistik, ETH Zürich, Switzerland.
- Stahel, W.A. and Maechler, M. (2009). Comment on “invariant co-ordinate selection”. *Journal of the Royal Statistical Society B* 71, pages 584–586.
- Tatsuoka, K.S. and Tyler, D.E. (2000). On the uniqueness of s-functionals and m-functionals under nonelliptical distributions. *The Annals of Statistics*, 28:1219–1243.
- Vakili, K., Hubert, M., and Rousseeuw, P. (2012). The MCS estimator of location and scatter. In *Proceedings of the twentieth international conference on Computational Statistics*, COMPSTAT ’12, pages 825–834.
- Van Aelst, S., Vandervieren, E., and Willems, G. (2012). A Stahel-Donoho estimator based on huberized outlyingness. *Computational Statistics and Data Analysis*, 56:531–542.
- Yohai, V.J. (1985). High breakdown point and high efficiency robust estimates for regression. Technical Report 66, Department of Statistics, University of Washington. Available at <http://www.stat.washington.edu/research/reports/1985/tr066.pdf>.