

Propagation of Outliers in Multivariate Data

Fatemah Alqallaf,

*Department of Statistics and Operations Research
Kuwait University
Kuwait
e-mail: fatemah@kuc01.kuniv.edu.kw*

Stefan Van Aelst*,

*Department of Applied Mathematics and Computer Science
Krijgslaan 281 S9
B-9000 Gent
Belgium
e-mail: Stefan.VanAelst@UGent.be*

Victor J. Yohai,

*Department of Mathematics
Ciudad Universitaria, Pabellón 1
1426, Buenos Aires
Argentina
e-mail: vyohai@dm.uba.ar*

and

Ruben H. Zamar[†]

*Department of Statistics
6356 Agricultural Road
Vancouver, BC
Canada V6T 1Z2
e-mail: ruben@stat.ubc.ca*

Abstract: We investigate the performance of robust estimates of multivariate location under non-standard data contamination models such as componentwise outliers (i.e. contamination in each variable is independent from the other variables). This model brings up a possible new source of statistical error that we call “propagation of outliers”. This source of error is unusual in the sense that it is generated by the data processing itself and takes place *after* the data has been collected. We define and derive the influence function of robust multivariate location estimates under flexible contamination models and use it to investigate the effect of propagation of outliers. Furthermore, we show that standard high-breakdown affine equivariant estimators propagate outliers and therefore show poor breakdown behavior under componentwise contamination when the dimension d is high.

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

[†]Research supported by NSERC

AMS 2000 subject classifications: Primary 62F35; secondary 62H12.

Keywords and phrases: Breakdown point, Contamination model, Independent contamination, Influence function, Robustness.

1. Introduction

Most statistical methods are built in the context of a given model and therefore are designed to perform well (e.g. be optimal) for this model. Models are also natural “testing grounds” for statistical procedures and therefore have a profound influence in the way data are processed and analyzed.

Classical models assume that data are affected by “normal” noise: small scale fluctuations arising from measurement errors, item-to-item differences and other sources of “well behaved” randomness, e.g. Gaussian random variables, Gamma random variables, Poisson processes and other “nice” random disturbances. Contamination models, on the other hand, assume that the data may also be affected by abnormal noise: large scale fluctuations that arise from data contamination, uneven data quality, mixed populations, gross errors, etc. Several contamination models have been proposed in the statistical literature. A nice discussion can be found in Barnett and Lewis (1994).

The best known and most important contamination model is the Tukey-Huber model (Tukey (1962) and Huber (1964)). This model assumes that, on average, a large fraction $(1 - \epsilon)$ of the data is generated from a classical, normal-error-only model. The remaining data, however, can be affected by abnormal noise. In other words, the Tukey-Huber model assumes a mixture distribution with a fully described dominant component and an unspecified minority component. The mixture fraction ϵ is a loosely specified nuisance parameter (e.g. $0 \leq \epsilon < 0.25$). The goal of a robust statistical analysis is to conduct inference on the dominant part of the mixture, filtering out possible abnormal noise generated by the minority component. The Tukey-Huber contamination model had a profound influence in the general strategy underlying most robust statistical procedures: identify outlying *cases* – those coming from the minority mixture component – and downweight their influence. This model also inspired the definition of key robustness concepts such as influence function, gross-error-sensitivity, maxbias and breakdown point.

The Tukey-Huber contamination model

$$X = (1 - B)Y + BZ,$$

was first introduced in the univariate location-scale setup. The unobservable variables Y, Z and B are independent, $Y \sim F$ (a well behaved location-scale distribution such as $N(\mu, \sigma^2)$), $Z \sim G$ (an unspecified outlier generating distribution) and $B \sim \text{Binomial}(1, \epsilon)$ (a random contamination indicator). Consequently, the observed variable X has the mixture distribution $(1 - \epsilon)F + \epsilon G$. The model was later extended and used in other settings including regression and multivariate location-scatter models. See for example Martin et al. (1989) and He et al. (1990).

The rest of the paper is organized as follows. In Section 2 we introduce a family of contamination models that includes the Tukey-Huber and componentwise contamination models as particular cases. In Section 3 we define and derive the influence function of robust multivariate location estimates under non-standard contamination models. In Section 4 we discuss propagation of outliers and show that standard high breakdown point (BP) robust estimates propagate outliers. In Section 5 we investigate the breakdown properties under componentwise contamination of robust estimates of multivariate location.

2. Alternative Contamination Frameworks

The *multivariate* Tukey-Huber model, where \mathbf{X} , \mathbf{Y} , \mathbf{Z} are d -dimensional vectors, may be appropriate for small dimensions but has serious limitations in higher dimensions. A main criticism concerns the assumption that the majority of the cases is free of contamination. Another criticism concerns the downweighting of contaminated cases. When d is large, the fraction of perfectly observed cases can be rather small and the downweighting of an entire case may be inconvenient in the case of “fat and short” data tables where the number of variables (columns) is much larger than the number of cases (rows).

We wish to investigate the robustness properties of classical robust estimates of multivariate location under different contamination models. Suppose that the random vector \mathbf{Y} has density

$$f_{\mathbf{Y}}(\mathbf{y}) = h((\mathbf{y} - \mu_0)' \Sigma_0^{-1} (\mathbf{y} - \mu_0)) \quad (1)$$

and we are interested in estimating the multivariate location vector μ_0 . However, we cannot observe \mathbf{Y} directly but we observe the random vector

$$\mathbf{X} = (\mathbf{I} - \mathbf{B}) \mathbf{Y} + \mathbf{BZ} \quad (2)$$

where $\mathbf{B} = \text{diag}(B_1, B_2, \dots, B_d)$ is a diagonal matrix, B_1, B_2, \dots, B_d are Bernoulli random variables with $P(B_i = 1) = \epsilon_i$ and the vector \mathbf{Z} has an arbitrary and unspecified outlier generating distribution.

Note that in principle, the *contamination indicator matrix*, \mathbf{B} , in model (2) could depend on the vector of uncontaminated observations, \mathbf{Y} . Likewise, the contamination vector, \mathbf{Z} , could depend on both, the contamination indicator matrix, \mathbf{B} , and the uncontaminated vector, \mathbf{Y} . In this paper, however, we restrict attention to the simpler case where \mathbf{Y} , \mathbf{B} and \mathbf{Z} are independent.

Different assumptions regarding the joint distribution of B_1, B_2, \dots, B_d give rise to different contamination models. For example, if B_1, B_2, \dots, B_d are fully dependent, that is, $P(B_1 = B_2 = \dots = B_d) = 1$, then model (2) reduces to the classical *fully dependent contamination model (FDCM)* which underlies most of the existing robustness theory. An important feature of this model is that the majority of the cases - rows in the data table - are assumed to be perfectly observed and free of contamination. Therefore, it is natural that robust methods designed to perform well under FDCM check for the possible existence of a

minority of contaminated cases to downweight their influence. Downweighting the influence of suspicious cases is a good strategy when d is relatively small but becomes less attractive when d is large. For example, downweighting an entire case may be unacceptably wasteful if d is very large and n is relatively small.

Another interesting case is the *fully independent contamination model (FICM)* where B_1, B_2, \dots, B_d are independent. Consider the case $P(B_1 = 1) = \dots = P(B_d = 1) = \epsilon$, then the probability that a case is perfectly observed under this model is given by $(1 - \epsilon)^d$. Clearly, this probability quickly decreases and becomes less than the critical value $1/2$ as d increases ($d \geq 14$ when $\epsilon = .05$ and $d \geq 69$ when $\epsilon = .01$). Another feature of FICM is its lack of affine equivariance and its “repulsion” for linear transformations of the original data because linear transformations have the potential to propagate outliers. While each column in the data table has on average $(1 - \epsilon)$ 100% clean data values, the same cannot be said of linear combinations of these columns. Outlier propagation will be discussed further in Section 4.

An intermediate situation between FDCM and FICM can be obtained by assuming that there is a certain probability $1 - \alpha(\epsilon)$ that the case is free of contamination (as in the FDCM), but otherwise the different cells are independently contaminated with probability $\beta(\epsilon)$. This model will be called the *partially clean independent contamination model (PCICM)*. That is,

$$P(B_1 = b_1, \dots, B_d = b_d) = \begin{cases} 1 - \alpha(\epsilon) + \alpha(\epsilon)(1 - \beta(\epsilon))^d, & \sum_{i=1}^d b_i = 0 \\ \alpha(\epsilon)\beta(\epsilon)^{\sum_{i=1}^d b_i}(1 - \beta(\epsilon))^{d - \sum_{i=1}^d b_i}, & \sum_{i=1}^d b_i > 0 \end{cases}$$

where $0 < \alpha(\epsilon) < 1$, $0 < \beta(\epsilon) < 1$, and $\beta(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$. To fix ideas we will consider two possible choices for the functions $\alpha(\epsilon)$ and $\beta(\epsilon)$: (i) $\alpha(\epsilon) = \gamma$ and $\beta(\epsilon) = \epsilon/\gamma$, for some $0 < \gamma < 1$, and (ii) $\alpha(\epsilon) = \beta(\epsilon) = \sqrt{\epsilon}$.

Similarly, the *partially spoiled independent contamination model (PSICM)* can be defined by assuming that there is a certain probability $\alpha(\epsilon)$ that the case is fully spoiled (as in the FDCM), but otherwise the cells are independently contaminated with probability $\beta(\epsilon)$. That is,

$$P(B_1 = b_1, \dots, B_d = b_d) = \begin{cases} \alpha(\epsilon) + (1 - \alpha(\epsilon))\beta(\epsilon)^d, & \sum_{i=1}^d b_i = d \\ (1 - \alpha(\epsilon))\beta(\epsilon)^{\sum_{i=1}^d b_i}(1 - \beta(\epsilon))^{d - \sum_{i=1}^d b_i}, & \sum_{i=1}^d b_i < d \end{cases}$$

In this case we assume that $\alpha(\epsilon)$ and $\beta(\epsilon)$ are of order $O(\epsilon)$ [that is, $\alpha(\epsilon)/\epsilon \rightarrow K_1 > 0$ and $\beta(\epsilon)/\epsilon \rightarrow K_2 > 0$ as $\epsilon \rightarrow 0$]. Again, to fix ideas we take $\alpha(\epsilon) = \epsilon/(2 - \epsilon)$ and $\beta(\epsilon) = \epsilon/2$.

In all the models described above the choices of the functions $\alpha(\epsilon)$ and $\beta(\epsilon)$ in PCICM (i) and (ii) and PSICM are such that the probability of contamination of a single cell is $P(B_i = 1) = \epsilon$ for all i . In the less obvious case of PSICM we have $P(B_i = 1) = \alpha(\epsilon) + (1 - \alpha(\epsilon))\beta(\epsilon) = \epsilon/(2 - \epsilon) + (1 - \epsilon/(2 - \epsilon))(\epsilon/2) = \epsilon$. Therefore, meaningful sensitivity analysis can be performed by letting $\epsilon \rightarrow 0$. Another important simplifying feature of these contamination models is that the probability that a case has exactly k contaminated cells is the same for all possible configurations. These probabilities - denoted $\delta_k(\epsilon)$ and summarized in

TABLE 1
Probability δ_k of exactly k contaminated cases for the different contamination models.

	FDCM	FICM	PCICM (i)	PCICM (ii)	PSICM
δ_0	$1 - \epsilon$	$(1 - \epsilon)^d$	$1 - \gamma + \gamma \left(1 - \frac{\epsilon}{\gamma}\right)^d$	$1 - \sqrt{\epsilon} + \sqrt{\epsilon}(1 - \sqrt{\epsilon})^d$	$(1 - \epsilon) \left(1 - \frac{\epsilon}{2}\right)^{d-1}$
δ_1	0	$\epsilon(1 - \epsilon)^{d-1}$	$\epsilon \left(1 - \frac{\epsilon}{\gamma}\right)^{d-1}$	$\epsilon(1 - \sqrt{\epsilon})^{d-1}$	$\left(\frac{\epsilon}{2}\right) (1 - \epsilon) \left(1 - \frac{\epsilon}{2}\right)^{d-2}$
δ_2	0	$\epsilon^2(1 - \epsilon)^{d-2}$	$\gamma \left(\frac{\epsilon}{\gamma}\right)^2 \left(1 - \frac{\epsilon}{\gamma}\right)^{d-2}$	$\epsilon^{3/2}(1 - \sqrt{\epsilon})^{d-2}$	$\left(\frac{\epsilon}{2}\right)^2 (1 - \epsilon) \left(1 - \frac{\epsilon}{2}\right)^{d-3}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
δ_{d-1}	0	$\epsilon^{d-1}(1 - \epsilon)$	$\gamma \left(\frac{\epsilon}{\gamma}\right)^{d-1} \left(1 - \frac{\epsilon}{\gamma}\right)$	$\epsilon^{d/2}(1 - \sqrt{\epsilon})$	$\left(\frac{\epsilon}{2}\right)^{d-1} (1 - \epsilon)$
δ_d	ϵ	ϵ^d	$\gamma \left(\frac{\epsilon}{\gamma}\right)^d$	$\epsilon^{(d+1)/2}$	$\frac{\epsilon}{2-\epsilon} \left[1 + (1 - \epsilon) \left(\frac{\epsilon}{2}\right)^{d-1}\right]$

Table 1 - have an important role in the derivation of the generalized influence function defined below. Note that, for all the considered models except FDCM, $\delta_1 = O(\epsilon)$ and $\delta_i = o(\epsilon)$ for $1 < i < d$.

3. The Influence Function

The influence function (IF) is a key robustness tool. The IF of robust multivariate location estimates has only been defined under the classical FDCM. We wish to extend the definition so that it can be derived under other contamination models.

To fix ideas, we consider the class of M-estimates of multivariate location (see e.g. Tatsuoka and Tyler, 2000) defined as

$$\mu(H) = \arg \min_{\mathbf{m}} E_H \{ \rho[d^2(\mathbf{X}, \mathbf{m}, \Sigma(H))] \} \quad (3)$$

where

$$d^2(\mathbf{X}, \mathbf{m}, \Sigma) = (\mathbf{X} - \mathbf{m})' \Sigma^{-1} (\mathbf{X} - \mathbf{m})$$

and $\Sigma(H)$ is a Fisher consistent, preliminary or simultaneous, estimating functional of multivariate scatter. In Lemma 3 of the Appendix we show that when \mathbf{X} has an elliptical distribution, then $\mu(H)$ is Fisher consistent under mild regularity conditions. Moreover it is easy to show that $\mu(H)$ satisfies the first order condition:

$$E_H \{ \psi[d^2(\mathbf{X}, \mu(H), \Sigma(H))] (\mathbf{X} - \mu(H)) \} = \mathbf{0} \quad (4)$$

where $\psi = \rho'$. Note that equation (4) is satisfied by large classes of estimators for multivariate location such as M-estimators (Maronna 1976), S-estimators (Davies 1987, Lopuhaä 1989), CM-estimators (Kent and Tyler 1996), MM-estimators (Tatsuoka and Tyler 2000, Tyler 2002), and τ -estimators (Lopuhaä 1991).

Suppose that, when there is no outlier contamination, we observe a vector \mathbf{Y} with distribution H_0 given by (1). Consider a family of distributions G_ϵ of

(B_1, \dots, B_d) such that $P(B_i = 1) = \epsilon$, and let $\mathbf{z} = (z_1, \dots, z_d) \in R^d$ be fixed. The distribution of the contaminated vector

$$\mathbf{X} = (\mathbf{I} - \mathbf{B})\mathbf{Y} + \mathbf{B}\mathbf{z},$$

where (B_1, \dots, B_d) has distribution G_ϵ and \mathbf{Y} has distribution H_0 will be denoted by $H(\epsilon, \mathbf{z})$. Then the influence function of the estimating functional $\mu(H)$ given in (3) for the contamination pattern G_ϵ is defined by

$$\text{IF}(\mu, \mathbf{z}) = \left. \frac{\partial}{\partial \epsilon} \mu(H(\epsilon, \mathbf{z})) \right|_{\epsilon=0}. \quad (5)$$

Observe that $H(\epsilon, \mathbf{z})$ and $\text{IF}(\mu, \mathbf{z})$ also depends on the family of distributions G_ϵ and on H_0 but we do not make this dependence explicit in our notation.

Suppose now that under G_ϵ we have that $P(B_1 = j_1, \dots, B_d = j_d) = \delta_k(\epsilon)$, where $k = \sum_{i=1}^d j_i$. Let us write

$$g(H, \mathbf{m}, \Sigma) = E_H \{ \psi(d^2(\mathbf{X}, \mathbf{m}, \Sigma)) (\mathbf{X} - \mathbf{m}) \}. \quad (6)$$

Following from (4) we have that

$$g(H(\epsilon, \mathbf{z}), \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z}))) = \mathbf{0}$$

or equivalently

$$\delta_0(\epsilon) g(H_0, \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z}))) +$$

$$\sum_{k=1}^d \delta_k(\epsilon) \sum_{I \in \mathcal{I}_k} g(H(I, \mathbf{z}), \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z}))) = \mathbf{0} \quad (7)$$

where $\mathcal{I}_k = \{I = \{i_1, \dots, i_k\} : i_1 < \dots < i_k, 1 \leq k \leq d\}$ and where $H(I, \mathbf{z})$ is the distribution function of $\mathbf{X} = (X_1, \dots, X_d)$ where $X_i = z_i$ if $i \in I$ and $X_i = Y_i$ if $i \notin I$. In particular, $\mathcal{I}_d = \{\{1, 2, \dots, d\}\}$ and $H(\{1, 2, \dots, d\}, \mathbf{z}) = \delta_{\mathbf{z}}$, a point mass distribution at \mathbf{z} .

To calculate the influence function (5) by differentiating (7) at $\epsilon = 0$, we assume that the functional $\Sigma(H)$ is Fisher Consistent at the core model H_0 (so $\Sigma(H(0, \mathbf{z})) = \Sigma_0$) and that $\Sigma(H(\epsilon, \mathbf{z}))$ is differentiable with respect to ϵ at $\epsilon = 0$.

First, note that

$$g(H_0, \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z})))|_{\epsilon=0} = g(H_0, \mu_0, \Sigma_0) = \mathbf{0}. \quad (8)$$

In the Appendix we show that under elliptically symmetric distributions,

$$\left. \frac{\partial}{\partial \epsilon} g(H_0, \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z}))) \right|_{\epsilon=0} = -A_\psi \text{IF}(\mu, \mathbf{z}) \quad (9)$$

where A_ψ is a constant which does not depend on μ_0 and Σ_0 . The last three equations set the stage for the derivation of the IF of μ under the different contamination frameworks considered in the previous section.

From Table 1, under the FDCM, we have $\delta_i(\epsilon) = \delta'_i(\epsilon) = 0$ for $i = 1, \dots, d-1$, $\delta_d(\epsilon) = \epsilon$ and $\delta'_d(0) = 1$. Therefore, using (8) and (9) we obtain

$$\left. \frac{\partial}{\partial \epsilon} g[H(\epsilon, \mathbf{z}), \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z}))] \right|_{\epsilon=0} = -A_\psi \text{IF}(\mu, \mathbf{z}) + g(\Delta_{\mathbf{z}}, \mu_0, \Sigma_0) = \mathbf{0},$$

where $\Delta_{\mathbf{z}}$ is a point mass distribution at \mathbf{z} , and so

$$\text{IF}(\mu, \mathbf{z}) = \frac{1}{A_\psi} g(\Delta_{\mathbf{z}}, \mu_0, \Sigma_0) = \frac{1}{A_\psi} \psi(d^2(\mathbf{z}, \mu_0, \Sigma_0)) (\mathbf{z} - \mu_0). \quad (10)$$

Also from Table 1 we have that under FICM, and PCICM (i) and (ii), $\delta_0(0) = \delta'_1(0) = 1$, $\delta_1(0) = 0$, and $\delta_i(0) = \delta'_i(0) = 0$ (for $i \geq 2$). So, under FICM, and PCICM (i) and (ii), we have

$$\left. \frac{\partial}{\partial \epsilon} g[H(\epsilon, \mathbf{z}), \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z}))] \right|_{\epsilon=0} = -A_\psi \text{IF}(\mu, \mathbf{z}) + \sum_{k=1}^d g(H(I_k, \mathbf{z}), \mu_0, \Sigma_0) = \mathbf{0},$$

where $I_k = \{k\}$. Therefore, under these three contamination models

$$\text{IF}(\mu, \mathbf{z}) = \frac{1}{A_\psi} \sum_{k=1}^d g(H(I_k, \mathbf{z}), \mu_0, \Sigma_0). \quad (11)$$

In the case of the PSICM, $\delta_0(0) = 1$, $\delta'_1(0) = \delta'_d(0) = 1/2$, and $\delta_i(\epsilon) = o(\epsilon)$ (for $2 \leq i \leq d-1$). Therefore, under this contamination model we obtain

$$\left. \frac{\partial}{\partial \epsilon} g[H(\epsilon, \mathbf{z}), \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z}))] \right|_{\epsilon=0} = -A_\psi \text{IF}(\mu, \mathbf{z}) + \frac{1}{2} \left[\sum_{k=1}^d g(H(I_k, \mathbf{z}), \mu_0, \Sigma_0) + g(\Delta_{\mathbf{z}}, \mu_0, \Sigma_0) \right] = \mathbf{0},$$

which yields

$$\text{IF}(\mu, \mathbf{z}) = \frac{1}{2A_\psi} \left[\sum_{k=1}^d g(H(I_k, \mathbf{z}), \mu_0, \Sigma_0) + g(\Delta_{\mathbf{z}}, \mu_0, \Sigma_0) \right]. \quad (12)$$

Remark 1. It is worth noticing that under all the considered models the corresponding influence functions can be interpreted as directional (Gateaux) derivatives. It is well known that in the FDCM case the derivative is in the direction of $\Delta_{\mathbf{z}}$, a point mass distribution at \mathbf{z} . In the case of FICM, and PCICM (i) and (ii), the derivative is in the direction of

$$\frac{1}{d} \sum_{k=1}^d H(I_k, \mathbf{z})$$

where $H(I_k, \mathbf{z})$ is the distribution of the random vector $\mathbf{Y} \sim H_0$ but with its k^{th} component replaced by the constant z_k . Finally, in the case of PCICM, the derivative is in the direction of the mixture

$$\frac{1}{2} \left[\Delta_{\mathbf{z}} + \frac{1}{d} \sum_{k=1}^d H(I_k, \mathbf{z}) \right].$$

To illustrate the influence functions we consider the case where \mathbf{Y} is bivariate normal with mean zero, variance 1 and correlation r and use the M-estimator based on Tukey's bisquare loss function with $c^2 = 6$. From Figure 1 we see that the influence functions are fully redescending for FDCM (panel a) as is well-known. However, for FICM the influence functions are not redescending (panels b and c). Therefore, a vanishingly small fraction of large coordinatewise contamination may have a persistent influence on the location M-estimate. Since FICM is not affine equivariant, as discussed further in the next sections, the influence function changes with the amount of correlation as illustrated in panel c where $r = 0.9$. Contrary to the classical contamination case, this change is not just a linear transformation by $\Sigma_0^{1/2}$. Note that if $r = 0$ the components are almost solely influenced by contamination in the corresponding component while in the correlated case they are influenced by contamination in both components. To compare, we also consider the coordinatewise Tukey bisquare M-estimator. Figure 2 shows the influence function of the coordinatewise M-estimator. This influence function is the same under FDCM and FICM and does not change with correlation. Note that the components of the coordinatewise M-estimator are only influenced by contamination in the corresponding component as can be expected. Moreover, the influence function of the coordinatewise M-estimator resembles the influence function of the bivariate M-estimator under FICM with $r = 0$ (Figure 1b).

To investigate the combined effect of correlation and a small fraction of contamination we calculated the gross-error sensitivity for FDCM and FICM for several values of r . Figure 3 shows that the gross-error sensitivities for the FICM coordinatewise contamination and the FDCM structural contamination are similar for low correlation. However, for high correlation, structural outliers have a high effect while coordinatewise contamination has a very small effect on the bivariate location M-estimator. Note that the GES for the coordinatewise estimator is independent of r .

Figure 4 shows the effect of the dimension d on the GES of multivariate location estimators under FDCM and FICM (derived from (10) and (11) respectively), when the core model is multivariate standard normal. We compared the affine equivariant multivariate S-location estimator with Tukey loss function and the corresponding coordinatewise S-estimator. Under FDCM the truncation parameter in the Tukey bisquare loss function determines the breakdown point of the S-estimator (see e.g. Lopuhaä 1989). For the multivariate location S-estimator we have chosen the value of the truncation parameter that yields a 50% breakdown point under FDCM. Similarly, for the coordinatewise S-estimator we have selected the value of the truncation parameter that yields a

50% breakdown point for each coordinate separately. Note that FDCM severely underestimates the influence of a vanishingly small fraction of contamination when d is large. The coordinatewise estimator has the same GES under FDCM and FICM and considerably smaller GES than its multivariate counterpart under FICM.

4. Propagation of Outliers

FDCM is translation-scale equivariant and affine equivariant. Therefore, if a random vector \mathbf{X} follows this model, then an affine transformation $\tilde{\mathbf{X}} = \mathbf{A}\mathbf{X} + \mathbf{b}$ will also follow the model, for any invertible matrix \mathbf{A} and vector \mathbf{b} . In particular, if \mathbf{X} has a probability ϵ of contamination, the same probability holds for $\tilde{\mathbf{X}}$. On the other hand, the independent contamination model is *not affine equivariant*. In fact, suppose that the random vector \mathbf{X} follows the FICM and \mathbf{A} is an invertible $d \times d$ matrix, then the transformed vector

$$\tilde{\mathbf{X}} = \mathbf{A}\mathbf{X} + \mathbf{b} = \mathbf{A}(\mathbf{I} - \mathbf{B})\mathbf{Y} + \mathbf{A}\mathbf{B}\mathbf{Z} + \mathbf{b}$$

is in general different from $(\mathbf{I} - \mathbf{B})\mathbf{A}\mathbf{Y} + \mathbf{A}\mathbf{B}\mathbf{Z} + \mathbf{b}$, unless $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$ (i.e. \mathbf{A} is diagonal). Therefore, $\tilde{\mathbf{X}}$ does not follow the independent contamination model.

The lack of affine equivariance of FICM causes a phenomenon that we call “*outlier propagation*”. FICM assumes that each column in the data table contains an average fraction ϵ of contamination. Since affine transformations linearly combine the columns, the independent contamination property is lost.

To illustrate this, we generated a small two-dimensional data set of size $n = 20$. Both components come from a standard Gaussian distribution and we added independent contamination to each component with a contamination probability of 30%. The contaminated data come from a Gaussian distribution with mean 10 and variance 1. Histograms of the original components X_1 and X_2 are shown in the top panels of Figure 5. Both histograms show a clear majority of clean data with approximately 1/3 of outlying points on the right. The thick vertical lines indicate the medians, 0.22 for X_1 and 0.95 for X_2 . The medians are slightly affected by the heavy contamination but still summarize well the majority of the data. Now, we consider an affine transformation: $L_1 = 0.64X_1 + 0.77X_2$ and $L_2 = 0.78X_1 + 0.62X_2$. Histograms of the components L_1 and L_2 are shown in the bottom panels of Figure 5. From these histograms it is clear that both components now contain a majority of contaminated cells and hence do not satisfy FICM with 30% contamination anymore. In fact, we have three distinct groups in each dimension consisting of 49%, 42% and 9% of the data. Note that the medians of L_1 and L_2 no longer reflect the location of the clean data.

Data following FICM and other non-affine equivariant versions of model (2) can severely upset standard, affine equivariant robust procedures. To illustrate this, we consider the following example. We generated 100 observations from a 15 dimensional standard Gaussian distribution and added independent contamination to each column with a contamination probability of 15%. The contamination is obtained by adding a constant value. Propagation of outliers can

have a devastating effect on affine equivariant robust estimators as illustrated in Figure 6. In this Figure we varied the size of the contamination constant from 0 to 100. We calculated the multivariate location of the data using (i) Minimum Volume Ellipsoid (MVE), (ii) Minimum Covariance Determinant (MCD) both proposed by Rousseeuw (1984), (iii) sample mean, (iv) coordinatewise median, and (v) Stahel-Donoho estimator independently proposed by Stahel (1981) and Donoho (1982). We plotted the maximum of the entries of the location estimate against the size of the contamination. Note that both MVE and MCD increase without bound. The Stahel-Donoho estimator as implemented in Splus increases even faster and crashes when the contamination constant exceeds 7. The behavior of the three affine equivariant robust estimates shows clear signs of breaking down. Not surprisingly, so does the sample mean. On the other hand, the coordinatewise median is hardly affected by the outliers in each component. This example clearly shows that robust affine equivariant methods are not robust against propagation of outliers. (A more rigorous treatment of this claim will be given in the next section.) Hence, these methods are not well suited for situations where the contamination regime operates on individual variables (columns) rather than individual cases (rows).

5. Affine equivariance and independent contamination

For simplicity we will keep the Section 3 assumption that the marginal probabilities of a contaminated cell are equal for all components, that is $P(B_1 = 1) = \dots = P(B_d = 1) = \epsilon$. However, with obvious modifications the results hold for the general case as well.

For each distribution G_0 on R with finite first moment, let $\mathcal{G}_h(G_0)$ be the set of distribution functions G on R^d with marginal distributions which are all stochastically larger than $G_0(x - h)$. For each $\delta > 0$ set

$$\mathcal{F}_{\delta h}(G_0) = \{H = (1/2 - \delta)H_0 + (1/2 + \delta)G, \quad G \in \mathcal{G}_h(G_0)\}$$

Definition 1. Let $\mathbf{T} = (T_1, \dots, T_d)$ be an equivariant multivariate location estimating functional on R^d . We say that \mathbf{T} is δ -consistent at infinity, when the central model is H_0 , if for any distribution G_0

$$\lim_{h \rightarrow \infty} \inf_{H \in \mathcal{F}_{\delta, h}(G_0)} T_i(H) = +\infty, \quad 1 \leq i \leq d.$$

In other words, δ -consistent estimates have the property that if at least $1/2 + \delta$ of the mass goes to infinity for all the coordinates, then all the coordinates of the estimate go to infinity too. Note that $\delta_2 > \delta_1$ implies $\mathcal{F}_{\delta_2, h}(G_0) \subset \mathcal{F}_{\delta_1, h}(G_0)$, thus if \mathbf{T} is δ_1 -consistent, then it is also δ_2 -consistent.

Let us introduce the following notation. Given a distribution H_0 on R^d denote by \mathcal{F}_ϵ^I its FICM contamination neighborhood of size ϵ that contains all

the distributions of $\mathbf{X} = (\mathbf{I}-\mathbf{B})\mathbf{Y} + \mathbf{B}\mathbf{Z}$ where \mathbf{Y}, \mathbf{B} and \mathbf{Z} are independent, \mathbf{Y} has distribution H_0 , \mathbf{B} is a diagonal matrix where the diagonal elements B_1, \dots, B_d are independent Bernoulli variables such that $P(B_i = 1) = \epsilon$ and \mathbf{Z} has an arbitrary distribution H^* . We denote by \mathcal{F}_ϵ^D its FDCM contamination neighborhood that contains the distributions of the form

$$H = (1 - \epsilon)H_0 + \epsilon H^*$$

where H^* is arbitrary.

We can now define the breakdown point under FICM, $\varepsilon_{\text{FICM}}^*$, of a multivariate location estimator $\mathbf{T}(H)$ as the smallest probability ϵ of contamination in each of the components that is needed to make $\|\mathbf{T}(H)\|$ arbitrary large. That is,

$$\varepsilon_{\text{FICM}}^*(\mathbf{T}, H_0) = \inf\{\epsilon > 0; \sup_{H \in \mathcal{F}_\epsilon^I} \|\mathbf{T}(H)\| = +\infty\}.$$

Theorem 1 shows that the FICM breakdown point of any equivariant estimate of location which is δ -consistent at infinity under a FICM model is at most $1 - (1/2 - \delta)^{1/d}$. Hence, if δ is independent of d , the FICM breakdown point tends to 0.

Theorem 1. Let $\mathbf{T}(H)$ be an affine equivariant multivariate location estimator which is δ -consistent at infinity for the central distribution H_0 , with finite first moments. If

$$\epsilon > \epsilon_0 = 1 - (1/2 - \delta)^{1/d},$$

then

$$\sup_{H \in \mathcal{F}_\epsilon^I} \|\mathbf{T}(H)\| = +\infty.$$

Hence,

$$\varepsilon_{\text{FICM}}^*(\mathbf{T}, H_0) \leq 1 - (1/2 - \delta)^{1/d}.$$

Proof. Consider the linear transformation $\mathbf{U} = \mathbf{A}\mathbf{X}$ with

$$A = \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 2 \end{pmatrix}.$$

Note that A is invertible since its eigenvalues are 2 with multiplicity one and 1 with multiplicity $(d - 1)$.

Let $H_h \in \mathcal{F}_\epsilon^I$ where $\epsilon > \epsilon_0$ and $\mathbf{Z} \sim \delta_h$ with δ_h the point mass at $(h, \dots, h) \in R^d$. It follows that with probability $(1 - \epsilon)^d = 1/2 - \delta^*$, with $\delta^* > \delta$ the vector \mathbf{X} comes from H_0 , and thus with probability $1/2 + \delta^*$ at least one component of \mathbf{X} is equal to h .

Let \tilde{H}_h and \tilde{H}_0 be the distributions of \mathbf{U} when \mathbf{X} has distribution H_h and H_0 respectively. Then

$$\tilde{H}_h = (1 - \delta^*)\tilde{H}_0 + \delta^*G_h,$$

where G_h is the distribution of \mathbf{U} when \mathbf{X} has distribution H_h conditionally on $\sum_{i=1}^d B_i > 0$. Therefore all the marginals of G_h are stochastically larger than $G_0(u-h)$ where G_0 is the distribution of $-2\sum_{j=1}^d |Y_j|$ with $\mathbf{Y} \sim H_0$. Since \mathbf{T} is δ^* -consistent at infinity, we then have

$$\lim_{h \rightarrow \infty} \|\mathbf{T}(\tilde{H}_h)\| = +\infty.$$

Since A is invertible and \mathbf{T} is affine equivariant,

$$\lim_{h \rightarrow \infty} \|\mathbf{T}(H_h)\| = \lim_{h \rightarrow \infty} \left\| A^{-1} \mathbf{T}(\tilde{H}_h) \right\| = +\infty,$$

proving the theorem. \square

It is obvious that a scatter estimate breaks down whenever the multivariate location estimate it is using to center the data breaks down. Therefore, although Theorem 1 is stated for multivariate location, it has clear implications for the companion scatter estimates. The following lemma will be used to show that many of the well-known affine equivariant robust estimators of multivariate location are δ -consistent at infinity.

Lemma 1. Suppose that $\mathbf{T}(H)$ is location-scale equivariant and can be represented as a weighted average, that is, it can be written as

$$\mathbf{T}(H) = E_H(\mathbf{X}w(H, \mathbf{X})). \quad (13)$$

where the weight function $w(H, \mathbf{x})$ satisfies: (i) $w(H, \mathbf{x}) \geq 0$, (ii) there exists K such that $w(H, \mathbf{x}) \leq K$ and (iii) there exists $\eta > 0$ such that $P_H(w(H, \mathbf{x}) > \eta) > 1/2 - \delta_0$ for some $\delta_0 > 0$. Then T is δ -consistent at infinity when the central model distribution H_0 has finite first moments, for all $\delta > \delta_0$.

Proof. Suppose that $\mathbf{T}(H)$ is not δ -consistent at infinity for some $\delta > \delta_0$. Then there exists a distribution G_0 on R , with finite first moment and a sequence of distributions G_h with marginals $G_{hi}(x_i) \leq G_0(x_i - h)$, such that if we call $H_h = (1/2 - \delta)H_0 + (1/2 + \delta)G_h$, then $T_i(H_h) \leq c$ for some c , for all $h > 0$. Let

$$H_h^*(\mathbf{x}) = H_h(\sqrt{h}\mathbf{x}) = (1/2 - \delta)H_0(\sqrt{h}\mathbf{x}) + (1/2 + \delta)G_h(\sqrt{h}\mathbf{x}).$$

Then, by scale equivariance of $\mathbf{T}(H)$

$$T_i(H_h^*) \leq c/\sqrt{h} \rightarrow 0, \quad \text{as } h \rightarrow \infty. \quad (14)$$

Observe that $M_h(\mathbf{x}) = H_0(\sqrt{h}\mathbf{x})$ converges weakly to the point mass distribution at zero, as $h \rightarrow \infty$.

Moreover

$$T_i(H_h^*) = (1/2 - \delta) \int x_i w(H_h^*, \mathbf{x}) dM_h(\mathbf{x}) + (1/2 + \delta) \int x_i w(H_h^*, \mathbf{x}) dG_h^*(\mathbf{x}),$$

where $G_h^*(\mathbf{x}) = G_h(\sqrt{h}\mathbf{x})$. Since $w(H_h^*, \mathbf{x}) \leq K$ and $E_{H_0}(|X_i|) < \infty$, we have

$$\left| \int x_i w(H_h^*, \mathbf{x}) dM_h(\mathbf{x}) \right| \leq K \int |x_i| dM_h(\mathbf{x}) = K \int \left| \frac{x_i}{\sqrt{h}} \right| dF_0(\mathbf{x}) \rightarrow 0, \quad \text{as } h \rightarrow \infty.$$

On the other hand, if $A_h = \{\mathbf{x} = (x_1, \dots, x_d) : x_i \geq h, 1 \leq i \leq d\}$ then

$$\lim_{h \rightarrow \infty} P_{G_h^*}(A_{\sqrt{h}}) = 1$$

and therefore

$$P_{H_h^*}(A_{\sqrt{h}}) \geq 1/2 + \delta.$$

Note that by assumptions (iii), $P_{H_h^*}(w(H_h^*, \mathbf{x}) > \eta) > 1/2 - \delta_0$. Set

$$B_h = A_{\sqrt{h}} \cap \{\mathbf{x} : w(H_h^*, \mathbf{x}) > \eta\},$$

then

$$\begin{aligned} \lim_{h \rightarrow \infty} P_{H_h^*}(B_h) &\geq \lim_{h \rightarrow \infty} P_{H_h^*}(\{x : w(H_h^*, \mathbf{x}) > \eta\}) - \lim_{h \rightarrow \infty} P_{H_h^*}(A_{\sqrt{h}}^c) \\ &\geq 1/2 - \delta_0 - (1/2 - \delta) \\ &= \delta - \delta_0 > 0. \end{aligned}$$

Since $\lim_{h \rightarrow \infty} P_{M_h}(B_h) = 0$, we have

$$\begin{aligned} 0 < \delta - \delta_0 &\leq \lim_{h \rightarrow \infty} P_{F_h^*}(B_h) = (1/2 - \delta) \lim_{h \rightarrow \infty} P_{M_h}(B_h) + (1/2 + \delta) \lim_{h \rightarrow \infty} P_{G_h^*}(B_h) \\ &= (1/2 + \delta) \lim_{h \rightarrow \infty} P_{G_h^*}(B_h). \end{aligned}$$

Therefore,

$$\lim_{h \rightarrow \infty} P_{G_h^*}(B_h) \geq \gamma = (\delta - \delta_0) / (1/2 + \delta).$$

Then

$$\begin{aligned} \lim_{h \rightarrow \infty} \int x_i w(H_h^*, \mathbf{x}) dG_h^*(\mathbf{x}) &\geq \sqrt{h} \eta \lim_{h \rightarrow \infty} \int_{B_h} dG_h^*(\mathbf{x}) + \lim_{h \rightarrow \infty} \int_{x_i < 0} x_i w(H_h^*, \mathbf{x}) dG_h^*(\mathbf{x}) \\ &= \sqrt{h} \eta \lim_{h \rightarrow \infty} \int_{B_h} dG_h^*(\mathbf{x}) + \lim_{h \rightarrow \infty} \int_{x_i < 0} \frac{x_i}{\sqrt{h}} w\left(H_h^*, \frac{\mathbf{x}}{\sqrt{h}}\right) dG_h(\mathbf{x}) \\ &\geq \sqrt{h} \eta \lim_{h \rightarrow \infty} \int_{B_h} dG_h^*(\mathbf{x}) + K \lim_{h \rightarrow \infty} \int_{x_i < 0} \frac{x_i}{\sqrt{h}} dG_h(\mathbf{x}). \end{aligned} \tag{15}$$

Now, regarding the first term we have of the right hand side of (15) we get

$$\eta \lim_{h \rightarrow \infty} \sqrt{h} \int_{B_h} dG_h^*(\mathbf{x}) = \eta \lim_{h \rightarrow \infty} \sqrt{h} P_{G_h^*}(B_h) \geq \gamma \eta \lim_{h \rightarrow \infty} \sqrt{h} = \infty. \tag{16}$$

Regarding the second term, first note that the distribution of x_i ($x_i < 0$) under $G_h(\mathbf{x})$ is also stochastically larger than the corresponding distribution under $G_0(x-h)$. Then, by the change of variable $y = x-h$, and using this stochastic inequality we have

$$\begin{aligned} \int_{x_i < 0} x_i dG_h(\mathbf{x}) &\geq \int_{x_i < 0} x_i dG_0(x_i - h) = \int_{y_i < -h} (y_i + h) dG_0(y_i) \\ &= \int_{y_i < -h} y_i dG_0(y_i) + h \int_{y_i < -h} dG_0(y_i). \end{aligned} \quad (17)$$

The first term is uniformly bounded as follows

$$\left| \int_{y_i < -h} y_i dG_0(y_i) \right| \leq \int_{y_i < -h} |y_i| dG_0(y_i) \leq \int_{-\infty}^{\infty} |y_i| dG_0(y_i) < \infty. \quad (18)$$

The second term tends to zero because G_0 has finite first moments and so

$$h \int_{y_i < -h} dG_0(y_i) = hP_{G_0}(y_i < -h) \rightarrow 0. \quad (19)$$

By (16) the first term in (15) tends to $+\infty$. By (17), (18) and (19) the second term in (15) is uniformly bounded. Therefore,

$$\lim_{h \rightarrow \infty} T_i(H_h^*) = \lim_{h \rightarrow \infty} \int x_i w(H_h^*, \mathbf{x}) dG_h^*(\mathbf{x}) = +\infty,$$

contradicting (14). \square

In the next section we will show that many well-known affine equivariant high-breakdown estimators of multivariate location are δ -consistent for any $\delta > 0$. Table 2 shows the fraction ϵ of contamination that must independently affect each of the variables to break down such estimators. We see that for higher dimensions ($d \geq 10$) a small amount of contamination in each variable suffices to make standard affine equivariant high-breakdown estimators useless.

TABLE 2
Minimal fraction of independent contamination that causes breakdown of δ -consistent, affine equivariant estimators.

	dimension								
	1	2	3	4	5	10	15	20	100
ϵ	0.50	0.29	0.21	0.16	0.13	0.07	0.05	0.03	0.01

5.1. Examples of δ -consistent at infinity estimators

Sample mean. A simple example is the sample mean which satisfies (13) with weights $w(H, \mathbf{x}) = 1$, hence the assumptions of Lemma 1 hold.

Coordinatewise median. Another simple example is the coordinatewise median. Although this estimator does not satisfy the assumptions of Lemma 1, a simple argument shows that it is δ -consistent at infinity for all $\delta > 0$. Using the notation introduced in the proof of Lemma 1, we have that $P_{G_h}(X_i \leq \sqrt{h}) \rightarrow 0$ and so

$$\lim_{h \rightarrow \infty} [(1/2 - \delta)H_{0i}(\sqrt{h}) + (1/2 + \delta)G_{hi}(\sqrt{h})] < 1/2.$$

Therefore, $\lim_{h \rightarrow \infty} \text{Med}(F_{hi}) = \infty$. Note, however, that the coordinatewise median is not affine equivariant and thus Theorem 1 does not apply in this case.

Next we show that several well-known affine equivariant high-breakdown point estimators of multivariate location are δ -consistent at infinity.

Minimum Covariance Determinant. The Minimum Covariance Determinant estimator (MCD) of multivariate location introduced by Rousseeuw (1984) is defined as a scaled weighted mean $\mathbf{T}_{MCD}(H)$ with weight $w(\mathbf{x}, H)$ equal to $1/2$ if \mathbf{x} belongs to a subset $A^* \subset R^d$ and is zero elsewhere. Let us denote by

$$\mu(H, A) = \frac{\int_A \mathbf{x} dH(\mathbf{x})}{P_H(A)}$$

the mean associated to any subset $A \subset R^d$. Then, the subset A^* is selected such that its covariance matrix $\Sigma(H, A^*) = \int_{A^*} (\mathbf{x} - \mu(H, A^*))(\mathbf{x} - \mu(H, A^*))' dH(\mathbf{x})$ has smallest determinant among all subsets A such that $P_H(A) \geq 1/2$. Therefore,

$$\mathbf{T}_{MCD}(H) = E(w(\mathbf{X}, H) \mathbf{X}),$$

where

$$w(\mathbf{x}, H) = \begin{cases} 1/P_H(A^*) & \text{if } \mathbf{x} \in A^* \\ 0 & \text{if } \mathbf{x} \notin A^*. \end{cases}$$

Clearly, the weights are nonnegative and bounded. Moreover, since $1 \leq w(\mathbf{x}, H) \leq 2$ for all $\mathbf{x} \in A^*$, we have that $P(w(H, \mathbf{x}) > \eta) > 1/2 - \delta_0$ for any $\eta < 1$ and $\delta_0 > 0$. Then \mathbf{T}_{MCD} satisfies the assumptions of Lemma 1 and is δ -consistent at infinity for any $\delta > 0$.

S-estimators. S-estimators of multivariate location and scatter $(\mathbf{T}(H), S(H))$ are defined as follows. Consider a function $\rho : R \rightarrow R^+$ that satisfies the following assumptions:

A1. ρ is even, bounded and nondecreasing on $[0, \infty)$ with $\rho(0) = 0$. Without loss of generality we will take $\rho(\infty) = 1$.

A2. ρ is differentiable, $\psi(t) = \rho'(t)$ is differentiable at 0, and $u(t) = \psi(t)/t$ is non-increasing on $[0, \infty)$. We will also assume that $\rho(u) < 1$ implies $\psi(u) > 0$. Then $(\mathbf{T}(H), S(H))$ is defined by the values (μ, Σ) satisfying

$$(\mathbf{T}(H), S(H)) = \arg \min_{\mu, \Sigma} \det(\Sigma)$$

subject to

$$E_H(\rho(d(\mathbf{x}, \mu, \Sigma)/s_0)) = b.$$

When the central model is multivariate normal, one often uses $s_0 = \sqrt{d}$, $b = E\left(\rho\left(\sqrt{V}/s_0\right)\right)$ with $V \sim \chi_d^2$. The asymptotic breakdown point of these estimates is $\min\{b, 1-b\}$.

It can be shown (see for example Davies (1987)) that if ρ is differentiable with $\psi(t) = \rho'(t)$, then $\mathbf{T}(H)$ satisfies the following equation

$$\mathbf{T}(H) = E_H(w(\mathbf{X}, H) \mathbf{X}) \quad (20)$$

with

$$w(\mathbf{x}, H) = \frac{u(d(\mathbf{x}, \mathbf{T}(H)), S(H))}{E_H(u(d(\mathbf{X}, \mathbf{T}(H)), S(H)))} \quad (21)$$

and

$$u(\mathbf{x}, H) = \frac{\psi(d(\mathbf{x}, \mathbf{T}(H)), S(H))}{d(\mathbf{x}, \mathbf{T}(H)), S(H)}. \quad (22)$$

Lemma 2. Suppose A1 and A2 are satisfied. Then the weight function $w(\mathbf{x}, H)$ associated with the S-location estimate $\mathbf{T}(H)$ with $b = 1/2$ satisfies assumptions (i), (ii) and (iii) of Lemma 1, for any $\delta > 0$.

Proof. The weights $w(\mathbf{x}, H)$ in (21) are clearly nonnegative. Moreover, $u(t)$ is bounded because by assumption (A2), $u(t) \leq u(0) = \psi'(0) = \kappa < \infty$. By the definition of the S-estimate, we have

$$\frac{1}{2} = E_H(\rho(d(\mathbf{x}, \mathbf{T}(H)), S(H))/s_0). \quad (23)$$

For any $0 < t < 1$, let

$$B_t = \{\mathbf{x} : \rho(d(\mathbf{x}, \mathbf{T}(H)), S(H))/s_0 \leq t\}.$$

It follows from (23) that

$$\frac{1}{2} \geq \int_{(B_t)^c} \rho(d(\mathbf{x}, \mathbf{T}(H)), S(H))/s_0) dH(\mathbf{x}) \geq (1 - P(B_t)) t$$

and so

$$P(B_t) \geq 1 - \frac{1}{2t}. \quad (24)$$

Let t_0 be such that

$$1 - \frac{1}{2t_0} = \frac{1}{2} - \delta_0, \quad (25)$$

that is,

$$t_0 = \frac{1}{1 + 2\delta_0}.$$

Note that $0 < \delta_0 \leq 1/2$ implies that $1/2 \leq t_0 < 1$. Moreover, combining (24) and (25) yields

$$P(B_{t_0}) \geq \frac{1}{2} - \delta_0. \quad (26)$$

Let $r_0 = \rho^{-1}(t_0)$, then we can write

$$B_{t_0} = \{\mathbf{x} : d(\mathbf{x}, \mathbf{T}(H), S(H))/s_0 \leq r_0\}.$$

Then, by the monotonicity of $u(t)$ we have that for all \mathbf{x} in B_{t_0} , $u(d(\mathbf{x}, \mathbf{T}(H), S(H))/s_0) \geq u(r_0)$. Put $\zeta = u(r_0)$, since $1/2 \leq t_0 < 1$, using the assumption that $\psi(u) > 0$ when $\rho(u) < 1$ we obtain that $\zeta > 0$. Then we can write

$$u(d(\mathbf{x}, \mathbf{T}(H), S(H))/s_0) \geq \zeta, \quad \text{for } \mathbf{x} \in B_{t_0}$$

and then

$$\kappa \geq E(u(d(\mathbf{X}, \mathbf{T}(H), S(H))/s_0)) \geq \zeta P(B_{t_0}) \geq \zeta/4.$$

Therefore, $w(\mathbf{x}, H) \leq 4\kappa/\zeta$ for all \mathbf{x} and $w(\mathbf{x}, H) > \zeta/\kappa$ for $\mathbf{x} \in B_{t_0}$. Together with (26) this means that assumptions (i)-(iii) of Lemma 1 are satisfied. \square

Remark. Although the Minimum Volume Ellipsoid (MVE) is an S-estimator, it doesn't satisfy assumptions A1 and A2. However, a simple reasoning shows that MVE would break down under independent contamination. The multivariate location MVE estimate minimizes the median of the Mahalanobis distances using a scatter matrix that yields the ellipsoid of smallest volume containing 50% of the data. If the MVE scatter matrix diverges, the whole MVE procedure breaks down. So, we can assume without loss of generality that the scatter matrix remains bounded. Since the median Mahalanobis distance must be largest than the smallest distance to points from G_h , the MVE location must accommodate such points and therefore must tend to infinity when h tends to infinity.

τ -estimates. Lopuhaä (1991) introduced τ -estimates of multivariate location as follows. Given μ and Σ with $\det(\Sigma) = 1$, find $s(\mu, \Sigma)$ such that

$$E_H[\rho_1(d(\mathbf{x}, \mu, \Sigma)/s(\mu, \Sigma))] = 1/2. \quad (27)$$

Then, define

$$\tau^2(\mu, \Sigma) = s^2(\mu, \Sigma) E_H[\rho_2(d(\mathbf{x}, \mu, \Sigma)/s(\mu, \Sigma))].$$

The τ -estimates of location and scatter $(\mathbf{T}(H), S(H))$ are defined by

$$(\mathbf{T}(H), S(H)) = \arg \min_{\mu, \Sigma} \det(\Sigma) \quad \text{subject to } \tau^2(\mu, \Sigma) = \tau_0^2,$$

where τ_0 is chosen so that $S(H)$ is consistent for some given function h , for example the function h corresponding to the multivariate normal distribution.

We assume that both ρ_1 and ρ_2 satisfy assumptions (A1) and (A2). Furthermore, we assume

A3. (a) $\psi_1(t) = \rho_1'(t) > 0$ implies that $\psi_2(t) = \rho_2'(t) > 0$.

(b) ρ_2 is continuously differentiable and $2\rho_2(t) - \psi_2(t)t \geq 0$.

It can be shown that τ -estimates satisfy (20) and (22) with

$$\psi(t) = W_H(t)\psi_1(t) + \psi_2(t) \quad (28)$$

where $W_H(t) \geq 0$ and bounded for all H provided (A3)b is satisfied. Then assumptions (i) and (ii) of Lemma 1 follow directly. Assumption (iii) holds as well because the τ -estimator of multivariate location satisfies equation (27) which corresponds to equation (23) in the proof of Lemma 2. Similar arguments as in Lemma 2 can then be used to show that there exists a bounded, closed set B with $P(\mathbf{x} \in B) > 1/2 - \delta$ and $\psi_1(d(\mathbf{x}, \mathbf{T}(H), S(H))) > \eta_1$ for all $\mathbf{x} \in B$ and some $\eta_1 > 0$. It follows from assumption (A3)a, the compactness of B , and the continuity of ψ_2 that also $\psi_2(d(\mathbf{x}, \mathbf{T}(H), S(H))) > \eta_2$ for all $\mathbf{x} \in B$ and some $\eta_2 > 0$. Therefore, we obtain from (28) that for all $\mathbf{x} \in B$, $\psi(d(\mathbf{x}, \mathbf{T}(H), S(H))) > \eta_2$.

Note that the set B is bounded and closed because it is of the form $B_t = \{\rho(d(\mathbf{x}, \mathbf{T}(H), S(H))/s_0) \leq t\}$ with $0 < t < 1$. If $\|\mathbf{x}\| \rightarrow \infty$, then also $d(\mathbf{x}, \mathbf{T}(H), S(H)) \rightarrow \infty$ which implies that $\rho(d(\mathbf{x}, T(H), S(H))/s_0) \rightarrow 1$. The set is closed because of the continuity of ρ_1 .

6. Concluding Remarks

FDCM assumes that the majority of the cases is clean and follows the underlying model. Robust methods developed for this model exploit the fact that the fraction of clean cases remains constant under affine transformations and concentrate on identifying and downweighting the minority of outlying cases. In fact, the maximal breakdown point of any affine equivariant robust estimator can not exceed 50% as shown by Lopuhaä and Rousseeuw (1991). On the other hand, in Section 4 we have shown that the fraction of outlying cells in FICM can drastically change under affine transformations. Consequently, data following FICM and other non-affine equivariant versions of model (2) can severely upset standard robust procedures, as demonstrated in Section 5.

Protection against outliers propagation can be achieved by using coordinate-wise procedures such as the (coordinatewise) median that only operate on one column at the time. However, a well-known weakness of coordinatewise methods is their lack of robustness against structural outliers. This type of outliers can only be handled by robust affine equivariant methods. Therefore, we have a trade-off between two robustness goals. One possible way to address this trade-off is to apply robust affine equivariant methods to subsets of columns at the time and combine the results. With larger subset size more protection against structural outliers is assured, but less protection against outliers propagation is obtained and vice versa.

Truly robust methods should be resistant against all kind of outliers, not only rowwise (cases) contamination. A first attempt in this direction has been made in the seminal paper of Croux et al. (2003) in the context of factor models. We hope that a new generation of robust methods that are resistant against both structural and independent outliers will emerge in the near future.

Appendix A: APPENDIX

A.1. Derivation of (9)

Since $g(H_0, \mu_0, \Sigma) = \mathbf{0}$ for all positive definite matrices Σ and elliptically symmetric distributions H_0 , we have that

$$\left. \frac{\partial}{\partial \Sigma} g(H_0, \mu_0, \Sigma) \right|_{\Sigma=\Sigma_0} = \mathbf{0}.$$

Hence,

$$\left. \frac{\partial}{\partial \epsilon} g(H_0, \mu(H(\epsilon, \mathbf{z})), \Sigma(H(\epsilon, \mathbf{z}))) \right|_{\epsilon=0} = \left. \frac{\partial}{\partial \mu} g(H_0, \mu, \Sigma_0) \right|_{\mu=\mu_0} \left. \frac{\partial}{\partial \epsilon} \mu(H(\epsilon, \mathbf{z})) \right|_{\epsilon=0} \quad (29)$$

We also have

$$\begin{aligned} \left. \frac{\partial}{\partial \mu} g(H_0, \mu, \Sigma_0) \right|_{\mu=\mu_0} &= -2E_{H_0}(\psi'(d^2(\mathbf{Y}, \mu_0, \Sigma_0))(\mathbf{Y} - \mu_0)(\mathbf{Y} - \mu_0)')\Sigma_0^{-1} \quad (30) \\ &\quad - E_{H_0}(\psi(d^2(\mathbf{Y}, \mu_0, \Sigma_0))\mathbf{I}). \end{aligned}$$

Let $\mathbf{w} = \Sigma_0^{-1/2}(\mathbf{Y} - \mu_0)$. Then \mathbf{w} has density given by (1) with $\mu_0 = \mathbf{0}$ and $\Sigma_0 = \mathbf{I}$. Since \mathbf{w} has a spherical distribution, it holds that

$$E(\psi'(\|\mathbf{w}\|^2) \mathbf{w}\mathbf{w}') = \frac{1}{d} E(\psi'(\|\mathbf{w}\|^2) \|\mathbf{w}\|^2) \mathbf{I}. \quad (31)$$

From (30) and (31) we get

$$\left. \frac{\partial}{\partial \mu} g(H_0, \mu, \Sigma_0) \right|_{\mu=\mu_0} = -A_\psi \mathbf{I},$$

where the constant $A_\psi = (2/d) E(\psi'(\|\mathbf{w}\|^2) \|\mathbf{w}\|^2) + E(\psi(\|\mathbf{w}\|^2))$ is independent of μ_0 and Σ_0 . Finally, from (29) we get (9).

A.2. Fisher consistency of $\mu(H)$

The following lemma proves the Fisher Consistency of the multivariate location estimate defined in equation (3).

Lemma 3. Suppose that (1) $\Sigma(H)$ is Fisher consistent, (2) ρ is non-decreasing, continuous and strictly increasing at zero and (3) the function h in (1) is non-increasing, continuous and strictly decreasing at zero. Then, $\mu(H)$ is Fisher consistent.

Proof. We have to show that if \mathbf{Y} has distribution given by (1) and $\mu \neq \mu_0$, then

$$E(\rho((\mathbf{Y} - \mu)\Sigma_0^{-1}(\mathbf{Y} - \mu))) > E(\rho((\mathbf{Y} - \mu_0)\Sigma_0^{-1}(\mathbf{Y} - \mu_0))).$$

It is enough to show this for the case $\mu = \mathbf{0}$ and $\Sigma_0 = \mathbf{I}$. Hence, we have to show

$$E(\rho(\|\mathbf{Y} - \mu\|^2)) > E(\rho(\|\mathbf{Y}\|^2)). \quad (32)$$

Let \mathbf{Y} be a random variable with density $h(\|\mathbf{y}\|^2)$ ($\mu_0 = \mathbf{0}, \Sigma_0 = \mathbf{I}$). We start by showing that for all z

$$P(\|\mathbf{Y} - \mu\|^2 \leq z) \leq P(\|\mathbf{Y}\|^2 \leq z) \quad (33)$$

and if $\mu_0 \neq \mathbf{0}$ there exists $\varepsilon > 0$ so that for $z \leq \varepsilon$ we have

$$P(\|\mathbf{Y} - \mu\|^2 \leq z) < P(\|\mathbf{Y}\|^2 \leq z). \quad (34)$$

Let $A_1 = \{\|\mathbf{Y} - \mu\|^2 \leq z\}$, $A_2 = \{\|\mathbf{Y}\|^2 \leq z\}$, $B = A_1 \cap A_2$. We have $\text{vol}(A_1) = \text{vol}(A_2)$ and $\text{vol}(A_1 - B) = \text{vol}(A_2 - B)$ where vol stands for volume.

To show (33) it is enough to prove that

$$\int_{A_1 - B} h(\|y\|^2) dy \geq \int_{A_2 - B} h(\|y\|^2) dy.$$

Since h is non-increasing, this follows from the fact that $A_1 - B \subset \{\|x\| \leq z\}$ and $A_2 - B \subset \{\|x\| > z\}$. Similarly using the fact that h is strictly decreasing at 0 we obtain (34).

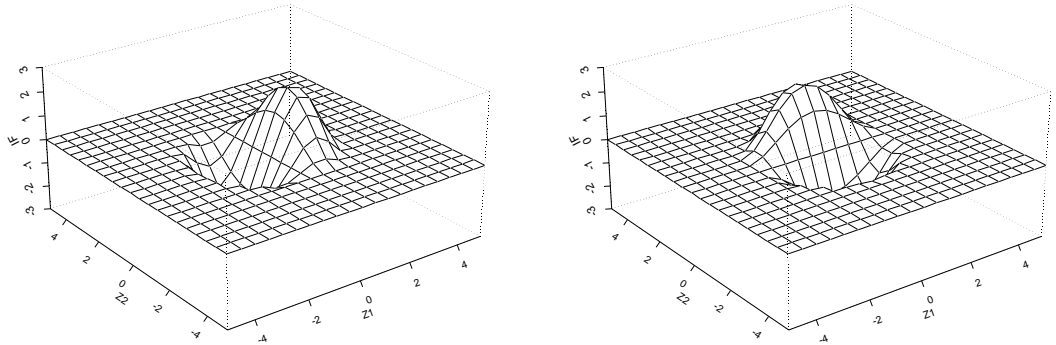
Inequalities (33) (34) and assumptions (2) and (3) imply that $\rho(\|\mathbf{Y} - \mu\|^2)$ is stochastically larger than $\rho(\|\mathbf{Y}\|^2)$ and then (32) follows. \square

References

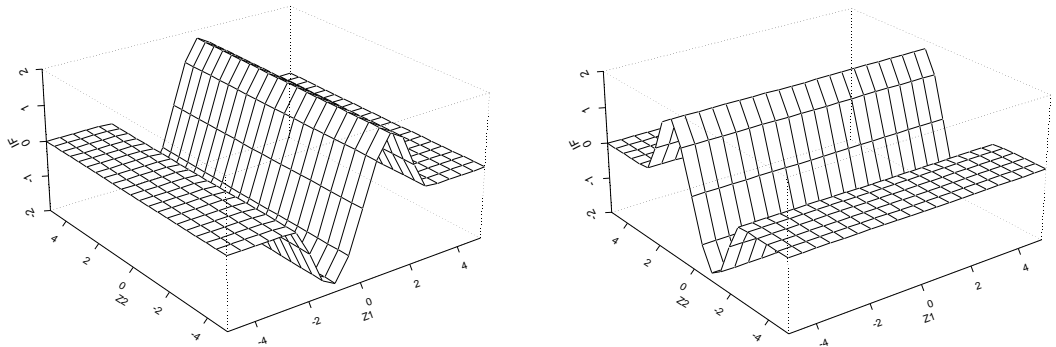
- Barnet, V. and Lewis, T. (1994). "Outliers in Statistical Data," John Wiley and Sons, Inc.
- Croux, C., Filzmoser, P., Pison, G., and Rousseeuw, P.J. (2003). "Fitting multiplicative models by robust alternating regressions," *Statistics and Computing*, 13, 23-36.
- Davies, P.L. (1987). "Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices," *The Annals of Statistics*, 15, 1269-1292.
- Donoho, D.L. (1982). "Breakdown properties of multivariate location estimators," Qualifying paper, Harvard University, Boston.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, John Wiley and Sons, New York.
- He, X., Simpson, D.G., and Portnoy, S. (1990). "Breakdown Robustness of Tests". *Journal of the American Statistical Association*, 85, 446-452.

- Huber, P.J. (1964). "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, 35, 73-101.
- Kent, J.T and Tyler, D.E. (1996). "Constrained M-estimation for multivariate location and scatter," *The Annals of Statistics*, 24, 1346-1370.
- Lopuhaä, H.P. (1989). "On the relation between S-estimators and M-estimators of multivariate location and covariance," *The Annals of Statistics*, 17, 1662-1683.
- Lopuhaä, H.P. (1991). "Multivariate τ -estimators for location and scatter," *Canadian Journal of Statistics*, 19, 307-321.
- Lopuhaä, H.P. and Rousseeuw, P.J. (1991). "Breakdown points of affine equivariant estimators of multivariate location and covariance matrices," *The Annals of Statistics*, 19, 229-248.
- Maronna, R.A. (1976). "Robust M-estimators of multivariate location and scatter," *The Annals of Statistics*, 4, 51-67.
- Martin, R.D., Yohai, V.J., and Zamar, R.H. (1989). "Min-max bias robust regression," *The Annals of Statistics*, 17, 1608-1630.
- Rousseeuw, P.J. (1984). "Least median of squares regression," *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, New York: John Wiley.
- Stahel, W.A., (1981). "Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen," Phd thesis, ETH Zürich.
- Tatsuoka, K.S. and Tyler, D.E. (2000). "The uniqueness of S and M-functionals under non-elliptical distributions," *The Annals of Statistics*, 28, 1219-1243.
- Tukey, J.W. (1962). "The Future of Data Analysis," *The Annals of Mathematical Statistics*, 33, 1-67.
- Tyler, D.E. (2002). "High-Breakdown Point Multivariate M-Estimation," *Estadística*, 54, 213-247.

(a)



(b)



(c)

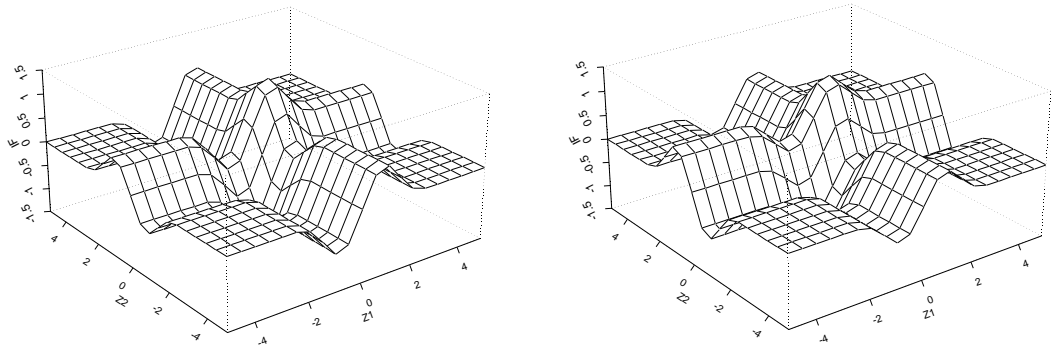


FIG 1. Influence functions for Tukey bisquare M -estimator of bivariate location. Left panels are for the first component, right panels are for the second component. Panel (a) FDCM; Panel (b) FICM with $r = 0$; Panel (c) FICM with $r = 0.9$.

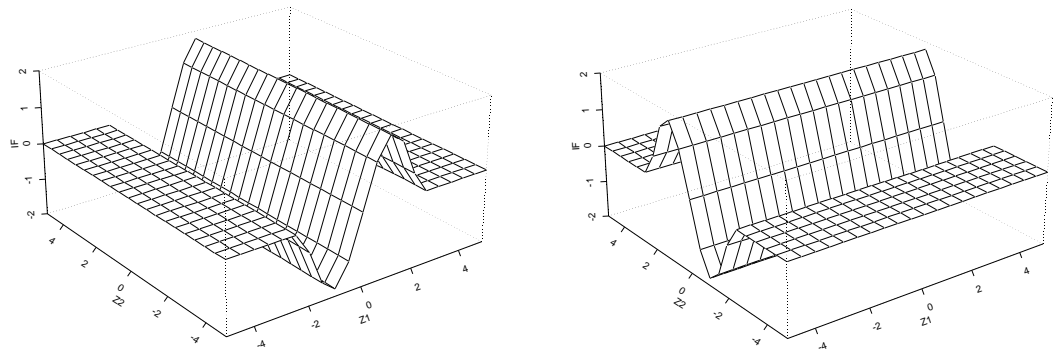


FIG 2. Influence functions for coordinatewise Tukey bisquare M-estimator of location. Left panel is for the first component, right panel is for the second component.

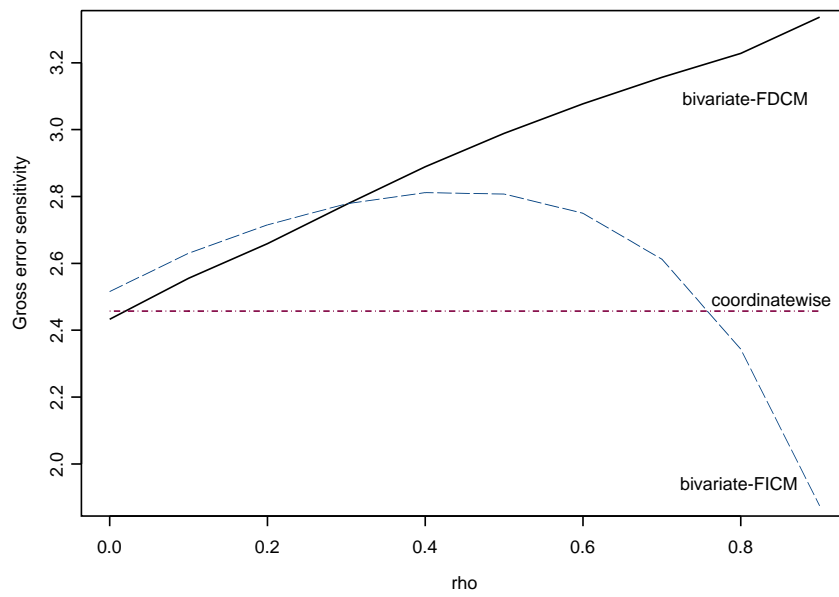


FIG 3. GES for Tukey bisquare M-estimator of bivariate location for different values of the correlation.

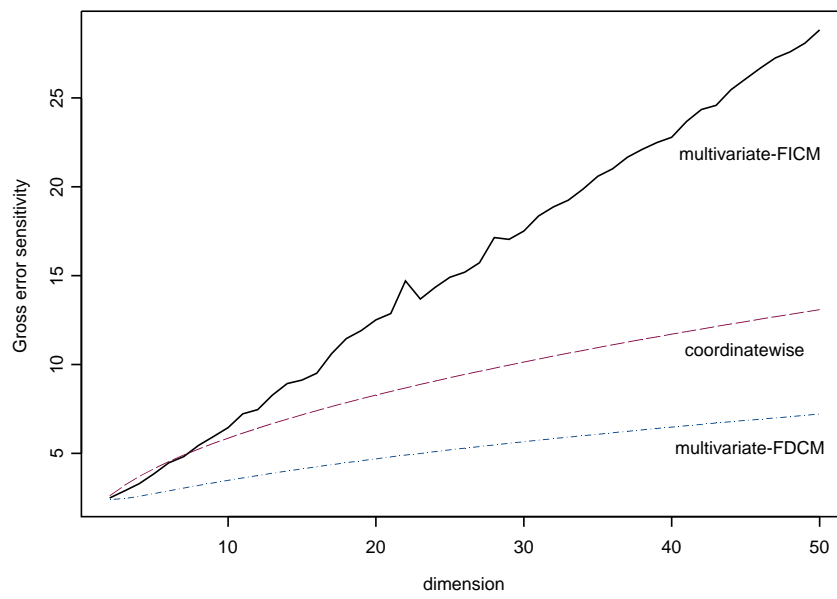


FIG 4. GES for Tukey bisquare M-estimator of multivariate location and corresponding coordinatewise estimator.

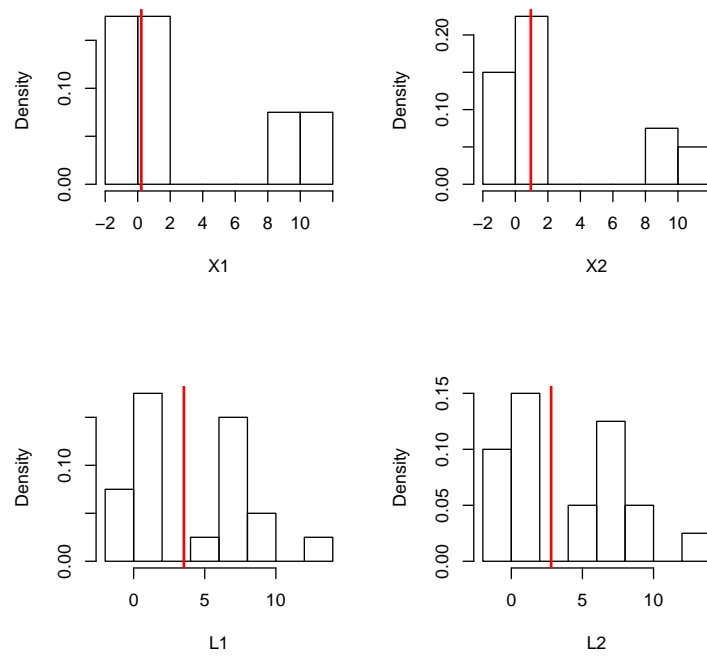


FIG 5. FICM outliers propagated by linear combinations.

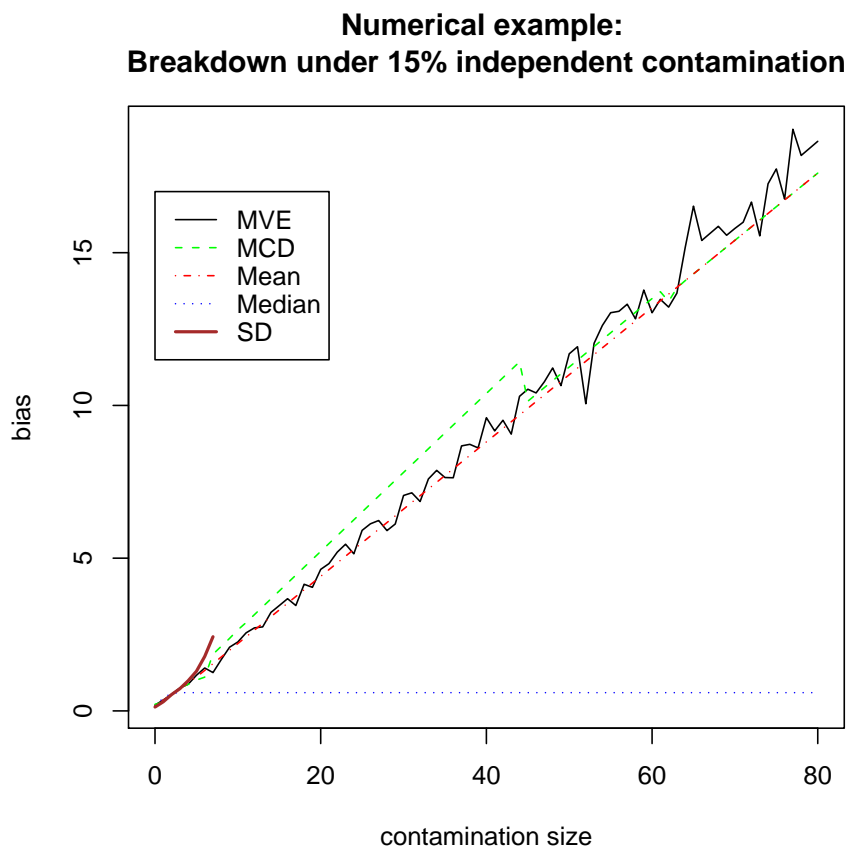


FIG 6. Affine equivariant, high breakdown point estimators try to identify outlying cases and break down when more than 50% of the cases are contaminated which can easily occur with small fractions of independent contamination in the variables when the dimension is moderately large.