Nonparametric Depth-Based Multivariate Outlier Identifiers, and Masking Robustness Properties

Xin Dang[1]

University of Mississippi

and

Robert Serfling[2]

University of Texas at Dallas

May, 2009

(Revision for *Journal of Statistical Planning and Inference*)

[1]Department of Mathematics, University of Mississippi, University, MS 38677-1848, USA. Email: xdang@olemiss.edu.

[2]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75083-0688, USA. Email: serfling@utdallas.edu. Website: www.utdallas.edu/~serfling. Support by NSF Grants DMS-0103698, CCF-0430366, and DMS-0805786, and by NSA Grant H98230-08-1-0106, is gratefully acknowledged.

Abstract

In extending univariate outlier detection methods to higher dimension, various issues arise: limited visualization methods, inadequacy of marginal methods, lack of a natural order, limited parametric modeling, and, when using Mahalanobis distance, restriction to ellipsoidal contours. To address and overcome such limitations, we introduce *nonparametric* multivariate outlier identifiers based on multivariate *depth functions*, which can generate contours following the shape of the data set. Also, we study *masking robustness*, that is, robustness against misidentification of outliers as nonoutliers. In particular, we define a masking breakdown point (MBP), adapting to our setting certain ideas of Davies and Gather (1993) and Becker and Gather (1999) based on the Mahalanobis distance outlyingness. We then compare four affine invariant outlier detection procedures, based on Mahalanobis distance, halfspace or Tukey depth, projection depth, and "Mahalanobis spatial" depth. For the goal of threshold type outlier detection, it is found that the Mahalanobis distance and projection procedures are distinctly superior in performance, each with very high MBP, while the halfspace approach is quite inferior. When a moderate MBP suffices, the Mahalanobis spatial procedure is competitive in view of its contours not constrained to be elliptical and its computational burden relatively mild. A small sampling experiment yields findings completely in accord with the theoretical comparisons. While these four depth procedures are relatively comparable for the purpose of *robust affine equivariant location estimation*, the halfspace depth is not competitive with the others for the quite different goal of *robust setting of an outlyingness threshold*.

AMS 2000 Subject Classification: Primary 62G10 Secondary 62H99.

Key words and phrases: multivariate analysis; nonparametric; robust; outlier identification; depth functions.

1 Introduction

Of fundamental importance in nonparametric multivariate location inference is identification of "outliers" in the data. These are observations far from, or inconsistent with, the main body of data points. Such cases may be of interest in themselves, or their presence can very adversely impact the performance of estimators or testing procedures. For excellent background, see Hawkins [9], Barnett and Lewis [1], and Gnanadesikan [8].

Identification of outliers by *visualization* is limited to dimension 3 or lower. Also, mere *marginal* outlier checking is inadequate, for an outlier can be nonoutlying in each coordinate. *Algorithmic* approaches that take *underlying geometry* into account are needed. One may formulate a suitable *outlyingness function* and set a *threshold*. A popular choice is the highly tractable *Mahalanobis distance outlyingness function*, which, however, is constrained to have *elliptical* contours of equal outlyingness, regardless of whether the underlying model is elliptically symmetric.

Here we introduce a general *nonparametric* approach based on *depth functions*, which provide center-outward orderings of multidimensional data. Higher depth represents higher "centrality", lower depth greater "outlyingness". One can associate with any depth function an equivalent outlyingness function. For suitable choices of depth function, the contours of equal outlyingness follow the actual geometric structure and shape of the given data.

An outlier identifier must, of course, be itself robust in the presence of the outliers it is supposed to identify. As a key relevant robustness criterion, we introduce the *masking breakdown point* (MBP), which measures the fraction of sample allowed to be contaminants without some extreme outlier becoming "masked", i.e., misidentified as a nonoutlier. We use *replacement* contamination. Our approach adapts a notion introduced by Davies and Gather [4] and Becker and Gather [2] using Mahalanobis distance outlyingness with the contaminated normal model and *addition* type contamination. (While not identical, replacement and addition breakdown points are equivalent as measures of robustness performance, although differing in intuitive appeal. See Zuo [27] and Serfling [20] for results and discussion.)

In particular, we derive and compare MBPs for four affine invariant outlyingness functions, based on the well-established *Mahalanobis distance*, *halfspace* (or *Tukey*), and *projection* depths, and on a new "*Mahalanobis spatial*" depth recently treated in Serfling [21]. The latter has a transformation-retransformation representation in terms of the well-known "*spatial*" outlyingness, which is only orthogonally invariant. We define these precisely in Section 2, which provides preliminaries on depth functions.

In Section 3, we formulate our notion of MBP, develop a general lemma on evaluation of MBP, and derive the MBPs of the four outlyingness functions under study. Further, these procedures are then compared within the framework of a contamination model, balancing MBP versus false positive rate. The findings are that, for *robust identification of outliers using a threshold*, both the *Mahalanobis distance* and the *projection* approaches are superior: *they can simultaneously maintain a low false positive rate and a high MBP*. In contrast, even though associated with a robust affine equivariant location estimator, the *halfspace* procedure imposes a severe and unacceptable trade-off between MBP and false positive rate. In cases with anticipated contamination level low enough that a modest MBP suffices, the *Mahalanobis spatial* approach is competitive in view of its contours

not constrained to be elliptical and its computational burden relatively mild. In Section 4, a small sampling experiment corroborates these theoretical conclusions.

The four depth functions under consideration are relatively comparable for the purpose of *robust location estimation*. However, for *robust setting of an outlyingness threshold*, a quite different type of goal, the halfspace depth is not competitive with the others.

2 Depth and outlyingness functions

Let F be a probability distribution on $\mathbb{R}^d$. An associated *depth function* $D(\boldsymbol{x}, F)$ provides a *center-outward* ordering of points $\boldsymbol{x} \in \mathbb{R}^d$, higher values representing higher "centrality" of $\boldsymbol{x}$, with nested contours of equal depth. The set of points of *maximal depth* constitutes the "center". For $D(\boldsymbol{x}, F)$ normalized to have range $[0, 1]$, the function $O(\boldsymbol{x}, F) = 1 - D(\boldsymbol{x}, F)$ gives an equivalent *outlyingness* function. One can also start with a center-outward $O(\boldsymbol{x}, F)$ and generate $D(\boldsymbol{x}, F)$. For a data set $\mathbb{X}_N = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$, we will denote sample versions by $D(\boldsymbol{x}, \mathbb{X}_N)$ and $O(\boldsymbol{x}, \mathbb{X}_N)$.

Quite a number of multivariate depth functions have been formulated. For *location inference* in $\mathbb{R}^d$, as considered here, depth is defined on the sample space. See Liu, Parelius and Singh [12], Zuo and Serfling [30], and Serfling [18], [21] for general treatments and discussion of connections with related multivariate quantile and centered rank functions. For other inference situations, depth is defined on the relevant parameter space. See Zhang [26], Müller [15] and Serfling [19].

We now introduce the four affine invariant outlyingness functions considered here, normalized to take values in $[0, 1)$. Affine invariance assures that a point classified as an "outlier" or not in one coordinate system remains similarly classified under affine transformation to another coordinate system (see Serfling [21] for discussion and the role of standardization).

Mahalanobis distance outlyingness. Perhaps the oldest notion of outlyingness in $\mathbb{R}^d$, $d \geq 2$, is that based on the distance introduced by Mahalanobis [14]. For location and scatter measures $\boldsymbol{m}(F)$ and nonsingular $\boldsymbol{S}(F)$, and with $\|\cdot\|$ the Euclidean norm, the corresponding Mahalanobis distance $\mathrm{MD}(\boldsymbol{x}, F) = \|\boldsymbol{S}(F)^{-1/2}(\boldsymbol{x} - \boldsymbol{m}(F))\|$, $\boldsymbol{x} \in \mathbb{R}^d$, is widely used as an outlyingness function, taking values in $[0, \infty)$. Equivalently, here we use as "*Mahalanobis distance outlyingness*"

$$O_{\mathrm{MD}}(\boldsymbol{x}, F) = \frac{\mathrm{MD}(\boldsymbol{x}, F)}{1 + \mathrm{MD}(\boldsymbol{x}, F)}, \ \boldsymbol{x} \in \mathbb{R}^d.$$

Halfspace or Tukey outlyingness. We take as "*halfspace outlyingness*"

$$O_{\mathrm{H}}(\boldsymbol{x}, F) = 1 - 2D_{\mathrm{H}}(\boldsymbol{x}, F), \ \boldsymbol{x} \in \mathbb{R}^d,$$

where

$$D_{\mathrm{H}}(\boldsymbol{x}, F) = \inf\{F(H) : H \text{ a closed halfspace containing } \boldsymbol{x}\}, \ \boldsymbol{x} \in \mathbb{R}^d,$$

the "halfspace" or "Tukey" depth introduced by Tukey [23] and generally regarded as the first notion of "depth function". In particular, the *sample* halfspace depth of $\boldsymbol{x}$ is the minimum fraction of data points in any closed halfspace containing $\boldsymbol{x}$.

PROJECTION OUTLYINGNESS. With $\mu(\cdot)$ and $\sigma(\cdot)$ any univariate location and scale measures, a "projection outlyingness" and related depth is defined by

$$\widetilde{O}_{\mathrm{P}}(\boldsymbol{x}, F) = \sup_{\|\boldsymbol{u}\| = 1} \left| \frac{\boldsymbol{u}'\boldsymbol{x} - \mu(F_{\boldsymbol{u}'\boldsymbol{X}})}{\sigma(F_{\boldsymbol{u}'\boldsymbol{X}})} \right|.$$

See Liu [11], Zuo and Serfling [30], and Zuo [28]. Here we take as "*projection outlyingness*"

$$O_{\mathrm{P}}(\boldsymbol{x}, F) = \frac{\widetilde{O}_{\mathrm{P}}(\boldsymbol{x}, F)}{1 + \widetilde{O}_{\mathrm{P}}(\boldsymbol{x}, F)},\ \boldsymbol{x} \in \mathbb{R}^d.$$

SPATIAL AND MAHALANOBIS SPATIAL OUTLYINGNESS. The "*spatial outlyingness*" corresponds to the *spatial depth* introduced by Vardi and Zhang [25] and is given by $O_{\mathrm{S}}(\boldsymbol{x}, F) = \|E\boldsymbol{S}(\boldsymbol{x} - \mathbf{X})\|$, where $\mathbf{X} \sim F$ and

$$\boldsymbol{S}(\boldsymbol{x}) = \begin{cases} \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, & \text{if } \boldsymbol{x} \neq \mathbf{0}, \\ \mathbf{0}, & \text{if } \boldsymbol{x} = \mathbf{0}, \end{cases}$$

the vector *sign function* in $\mathbb{R}^d$. It is only *orthogonally invariant*. To obtain an affine invariant modification, we *standardize* using any *weak covariance functional* or *shape functional*, i.e., any symmetric positive definite $d \times d$ matrix-valued functional $\boldsymbol{C}(F)$ defined on distributions F on $\mathbb{R}^d$ and satisfying *weak covariance equivariance*,

$$\boldsymbol{C}(F_{\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}}) = k(\boldsymbol{A}, \boldsymbol{b}, F_{\boldsymbol{X}})\boldsymbol{A}\,\boldsymbol{C}(F_{\boldsymbol{X}})\,\boldsymbol{A}',$$

for any nonsingular $d \times d$ $\boldsymbol{A}$ and any $\boldsymbol{b}$, with $k(\boldsymbol{A}, \boldsymbol{b}, F_{\boldsymbol{X}})$ a positive scalar function. For any such $\boldsymbol{C}(F)$, an associated affine invariant "*Mahalanobis spatial outlyingness function*" (Serfling [21]) is given by

$$O_{\mathrm{MS}}(\boldsymbol{x}, F_{\boldsymbol{X}}) = O_{\mathrm{S}}\left(\boldsymbol{C}(F_{\boldsymbol{X}})^{-1/2}\boldsymbol{x}, F_{\boldsymbol{C}(F_{\boldsymbol{X}})^{-1/2}\boldsymbol{X}}\right) = \left\|E\boldsymbol{S}(\boldsymbol{C}(F_{\boldsymbol{X}})^{-1/2}(\boldsymbol{x} - \boldsymbol{X}))\right\|.$$

3 Nonparametric outlier identification method

3.1 The nonparametric outlier identification problem

Relative to a given outlyingness function $O(\boldsymbol{x}, F)$, points whose outlyingness values exceed some specified threshold λ are considered "outliers" of F. That is, a point $\boldsymbol{x}$ inside (resp., outside) the region

$$out(\lambda, F) = \{\boldsymbol{x} : O(\boldsymbol{x}, F) > \lambda\}$$

is called a λ *outlier* (resp., *nonoutlier*) of F. Relative to a data set $\mathbb{X}_N = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ in $\mathbb{R}^d$ and a specified choice of outlyingness threshold λ_N, the practical goal is to correctly classify each point $\boldsymbol{x} \in \mathbb{R}^d$ as an F-based λ_N outlier or not, i.e., as belonging to the (unknown) outlier region $out(\lambda_N, F)$ or not. For this purpose, the region $out(\lambda_N, F)$ is *estimated* by an $\mathbb{X}_N$-based *outlier identifier* (or *sample outlier region*)

$$\mathrm{OR}(\lambda_N, \mathbb{X}_N) = \{\boldsymbol{x} : O(\boldsymbol{x}, \mathbb{X}_N) > \lambda_N\}.$$

(This is more demanding than simply *ranking* points by outlyingness values.)

Unfortunately, the outliers in $\mathbb{X}_N$ can exacerbate the situation by adversely influencing the performance of $\mathrm{OR}(\lambda_N, \mathbb{X}_N)$ as an estimator of $out(\lambda_N, F)$. Thus the chosen outlyingness function itself should be *robust*. In particular, *masking* occurs if points which are $O(\cdot, F)$-based λ_N outliers of F are misidentified by $\mathrm{OR}(\lambda_N, \mathbb{X}_N)$ as sample λ_N *nonoutliers*. If, relative to threshold λ_N, points of *arbitrarily extreme* $O(\cdot, F)$*-outlyingness* can be misidentified so, then *masking breakdown* of $\mathrm{OR}(\lambda_N, \mathbb{X}_N)$ occurs, in which case $\mathrm{OR}(\lambda_N, \mathbb{X}_N)$ is grossly unreliable. The minimal fraction of contaminants in $\mathbb{X}_N$ sufficient for masking breakdown to occur provides a useful robustness criterion, the *masking breakdown point*, which we formulate precisely in Section 3.2. General results on evaluating masking breakdown points are developed in Section 3.3, and specific results for our four target outlyingness functions are derived in Section 3.4.

One must specify, of course, the "outlier" threshold λ_N, which depends upon the choice of outlyingness function $O(\cdot, F)$ and possibly the sample size N. One approach, which we follow here, is based on a *contamination model* for F,

$$F = (1 - \varepsilon)G + \varepsilon H, \tag{1}$$

with G a known "ideal" model distribution and H an unknown source of "contaminants" tending to have high outlyingness relative to G. Extreme observations from G are "false positives" and those from H "true outliers". We desire the threshold λ_N high enough to yield a small *false positive rate*

$$(1 - \varepsilon)P_G(O(\boldsymbol{X}, G) > \lambda_N) \approx P_G(O(\boldsymbol{X}, G) > \lambda_N)$$

while also low enough for "true outliers" to be identified with high probability

$$\varepsilon P_H(O(\boldsymbol{X}, H) > \lambda_N) \approx \varepsilon.$$

We thus adopt $\alpha_N = P_G(O(\boldsymbol{X}, G) > \lambda_N)$ and ε as the (approximate) false positive and true positive rates, respectively, quantities which can be specified. Given α_N, the threshold λ_N is determined by

$$\lambda_N = F^{-1}_{O(\boldsymbol{X}, G)}(1 - \alpha_N), \tag{2}$$

based on the quantile function of the distribution of $O(\boldsymbol{X}, G)$ under the ideal distribution G. In turn, α_N, should be selected relative to ε. Scenarios for these choices are discussed in Section 3.5.

For a *fixed choice of* α_N, the four outlier identifiers of form $\mathrm{OR}(\lambda_N, \mathbb{X}_N)$ for different outlyingness functions and corresponding thresholds may be compared in terms of the associated masking breakdown points. These results are derived in Section 3.6, and conclusions and comparisons are provided in Section 3.7.

In comparison with Davies and Gather [4] and Becker and Gather [2], we find it more flexible and convenient to index our outlier regions by the threshold λ instead of by a false positive rate α under an assumed contamination model.

3.2 Masking breakdown point

Masking of some γ outliers of F can occur with k contaminants if there exists a choice of k replacements $\mathbb{Y}_k$, changing $\mathbb{X}_N$ to $\mathbb{X}_{N,k}$, such that

$$\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})} \cap out(\gamma, F) \neq \emptyset, \tag{3}$$

where $\overline{A}$ denotes the complement of A. Note that (3) holds if and only if *some γ outliers of F are included among sample λ_N nonoutliers*, relative to the altered sample $\mathbb{X}_{N,k}$. The lack of robustness of $\mathrm{OR}(\lambda_N, \mathbb{X}_N)$ in the presence of k contaminants in $\mathbb{X}_N$ may be measured by the largest value of γ for which (3) holds for some choice of $\mathbb{Y}_k$, and thus we define the quantity

$$\gamma_M(\lambda_N, \mathbb{X}_N, k) =$$
$$\sup\{\gamma > 0 : \exists \text{ a choice of } k \text{ replacements } \mathbb{Y}_k, \text{ changing } \mathbb{X}_N \text{ to } \mathbb{X}_{N,k}, \text{ such that (3) holds}\}.$$

For and only for $\gamma > \gamma_M(\lambda_N, \mathbb{X}_N, k)$, (3) fails for every choice of $\mathbb{Y}_k$ and thus all γ outliers of F in any altered data set $\mathbb{X}_{N,k}$ are indeed identified as sample λ_N outliers. The *worst case* is that $\gamma_M(\lambda_N, \mathbb{X}_N, k) = 1$ and represents *masking breakdown due to k replacements*: some points with arbitrarily large outlyingness $O(\cdot, F)$ can fail, by suitable choices of $\mathbb{X}_{N,k}$, to be identified by $\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})$ as sample λ_N outliers. Noting that $\gamma_M(\lambda_N, \mathbb{X}_N, k) \leq \gamma_M(\lambda_N, \mathbb{X}_N, k+1) \leq \gamma_M(\lambda_N, \mathbb{X}_N, N) = 1$, a useful robustness criterion is thus the *minimal number* $k_M(\lambda_N, \mathbb{X}_N) = \min\{k : \gamma_M(\lambda_N, \mathbb{X}_N, k) = 1\}$ of sample contaminants necessary to cause masking breakdown, or, equivalently, the *masking breakdown point* (MBP)

$$\varepsilon_M(\lambda_N, \mathbb{X}_N) = \frac{k_M(\lambda_N, \mathbb{X}_N)}{N}.$$

3.3 Evaluation of the masking breakdown point

Evaluation of $\varepsilon_M(\lambda, \mathbb{X}_N)$ is carried out not by solving the equation $\gamma_M(\lambda_N, \mathbb{X}_N, k) = 1$ for successive k, but rather via tools such as the following result.

Theorem 3.1 *Let F be continuous with supp(F) $= \mathbb{R}^d$. Suppose that $O(\cdot, F)$ satisfies $O(\boldsymbol{x}, F) < 1$, all $\boldsymbol{x}$, and*

$$O(\boldsymbol{x}, F) \to 1 \text{ if and only if } \|\boldsymbol{x}\| \to \infty. \tag{4}$$

Then $\gamma_M(\lambda_N, \mathbb{X}_N, k) = 1$ (masking breakdown with replacement of k sample values) if and only if

$$\sup_{\mathbb{X}_{N,k}} \sup_{\boldsymbol{y} \in \overline{OR(\lambda_N, \mathbb{X}_{N,k})}} \|\boldsymbol{y}\| = \infty. \tag{5}$$

Proof. Suppose that $\gamma_M(\lambda_N, \mathbb{X}_N, k) = 1$. Then, for any $\gamma < 1$, there exists $\mathbb{X}_{N,k}$ such that (3) holds. For a sequence $\gamma_n \uparrow 1$, let $\boldsymbol{y}_n$ belong to the intersection in (3) corresponding to $\gamma = \gamma_n$. Then $\gamma_n < O(\boldsymbol{y}_n, F) \uparrow 1$ and by (4) we have $\|\boldsymbol{y}_n\| \to \infty$. Then

$$\sup_{\mathbb{X}_{N,k}} \sup_{\boldsymbol{y} \in \overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}} \|\boldsymbol{y}\| \geq \sup_n \|\boldsymbol{y}_n\| = \infty,$$

and so (5) holds.

Now assume (5). By (4), we have

$$\sup_{\mathbb{X}_{N,k}} \sup_{\boldsymbol{y} \in \overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}} O(\boldsymbol{y}, F) = 1. \tag{6}$$

Let $\gamma_n \uparrow 1$ and select $\mathbb{X}_{N,k}^{(n)}$ such that

$$\sup_{\boldsymbol{y} \in \overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k}^{(n)})}} O(\boldsymbol{y}, F) > \gamma_n.$$

Then there exists $\boldsymbol{y}^{(n)} \in \overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k}^{(n)})}$ with $O(\boldsymbol{y}^{(n)}, F) > \gamma_n$, i.e., with $\boldsymbol{y}^{(n)} \in out(\gamma_n, F)$ and hence satisfying

$$\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k}^{(n)})} \cap out(\gamma_n, F) \supset \{\boldsymbol{y}^{(n)}\} \neq \emptyset.$$

Thus $\gamma_M(\lambda_N, \mathbb{X}_N, k) = 1$. □

Remark 3.1 (a) We note that, under (4), conditions (5) and (6) are equivalent.

(b) If $\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}$ can be made to grow along some direction, then arbitrarily large outliers become elements of $\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}$ and classifed as nonoutliers ("*masking*"). On the other hand, if a diameter of $\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}$ can be made to shrink to the "center" and hence $\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}$ degenerate to a $(d-1)$-dimensional structure, then nonoutliers arbitrarily close to the center will become elements of $\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})$ and classifed as outliers ("*swamping*"). Our concern in the present paper is the first case. □

3.4 Masking breakdown points for selected outlyingness functions

We now derive MBPs, or bounds on them, for the outlyingness functions defined in Section 2. In some cases the results will depend upon the usual (replacement) BPs of relevant location and scatter statistics. For these, we use standard definitions [6], [13], as follows.

For a *location* estimator $T(\mathbb{X}_N)$ in $\mathbb{R}^d$, we say that *breakdown* occurs with k points of $\mathbb{X}_N$ replaced if

$$\sup_{\mathbb{X}_{N,k}} \|T(\mathbb{X}_N) - T(\mathbb{X}_{N,k})\| = \infty,$$

with $\mathbb{X}_{N,k}$ as defined previously. With $k(T(\mathbb{X}_N))$ denoting the minimum k such that $T(\mathbb{X}_N)$ breaks down due to k replacements, the *replacement breakdown point* of $T(\mathbb{X}_N)$ is given by

$$RBP(T(\mathbb{X}_N)) = k(T(\mathbb{X}_N))/N.$$

For a positive definite matrix-valued *scatter* estimator $\boldsymbol{S}(\mathbb{X}_N)$, *explosion breakdown* of $\boldsymbol{S}(\mathbb{X}_N)$ occurs with k points of $\mathbb{X}_N$ replaced if

$$\sup_{\mathbb{X}_{N,k}} \|e_{\max}(\boldsymbol{S}(\mathbb{X}_N)) - e_{\max}(\boldsymbol{S}(\mathbb{X}_N))\| = \infty,$$

and *implosion breakdown* if

$$\sup_{\mathbb{X}_{N,k}} \|1/e_{\min}(\boldsymbol{S}(\mathbb{X}_N)) - 1/e_{\min}(\boldsymbol{S}(\mathbb{X}_N))\| = \infty,$$

where $e_{\max}(\boldsymbol{S}(\mathbb{X}_N))$ and $e_{\min}(\boldsymbol{S}(\mathbb{X}_N)) \geq 0$ denote, respectively, the maximum and minimum eigenvalues of $\boldsymbol{S}(\mathbb{X}_N)$. With obvious notation, the corresponding replacement BPs are given by

$$RBP_{\exp}(\boldsymbol{S}(\mathbb{X}_N)) = k_{\exp}(\boldsymbol{S}(\mathbb{X}_N))/N, \quad RBP_{\text{imp}}(\boldsymbol{S}(\mathbb{X}_N)) = k_{\text{imp}}(\boldsymbol{S}(\mathbb{X}_N))/N.$$

We will see that only the explosion case is relevant here.

3.4.1 Mahalanobis distance outlier identifier

For the Mahalanobis distance outlier identifier using $O_{\text{MD}}(\boldsymbol{x}, F)$, we establish bounds on the MBP in terms of $RBP(\boldsymbol{m}(\mathbb{X}_N))$ and $RBP_{\exp}(\boldsymbol{S}(\mathbb{X}_N))$. Of greatest interest for our purposes is the upper bound $RBP(\boldsymbol{m}(\mathbb{X}_N))$, which is attained by the MBP in the case that $RBP(\boldsymbol{m}(\mathbb{X}_N)) \leq RBP_{\exp}(\boldsymbol{S}(\mathbb{X}_N))$.

Theorem 3.2 *Using $O_{\text{MD}}(\boldsymbol{x}, F)$ with threshold λ_N, we have*

$$\min\{RBP(\boldsymbol{m}(\mathbb{X}_N)), RBP_{\exp}(\boldsymbol{S}(\mathbb{X}_N))\} \leq \varepsilon_M^{\text{MD}}(\lambda_N, \mathbb{X}_N) \leq RBP(\boldsymbol{m}(\mathbb{X}_N)).$$

Remark 3.2 (a) The above bounds do not depend upon the threshold λ_N, in which case this threshold may be chosen to achieve a desired false positive rate without entailing a tradeoff with MBP, as will be discussed in Section 3.5.

(b) Theorem 3.2, with our notion of MBP and replacement contamination, is an analogue of Theorems 1 and 2 of Becker and Gather [2] with their MBP and addition contamination.

(c) Since for $O_{\text{MD}}(\boldsymbol{x}, F)$ the regions $\overline{\text{OR}(\lambda_N, \mathbb{X}_{N,k})}$ are ellipsoidal, we see from Remark 3.1(b) that only the explosion case of breakdown of $\boldsymbol{S}(\mathbb{X}_N)$ is relevant, our concern here being *masking*.

(d) As discussed by Becker and Gather [2], but utilizing an improved bound given by Zuo [28], we may state that *lower bounds* for optimal BPs for $\boldsymbol{m}(\mathbb{X}_N)$ among affine equivariant location estimators and for $\boldsymbol{S}(\mathbb{X}_N)$ among affine equivariant scatter estimators are $(N - d + 2)/2N$ and $\lfloor (N - d + 1)/2) \rfloor / N$, respectively, the latter under the condition that $N \geq d + 1$ and the sample be in general position. Also, the latter is an *upper* bound for $RBP(\boldsymbol{m}(\mathbb{X}_N))$ among affine equivariant location estimators [5]. Using in $O_{\text{MD}}(\boldsymbol{x}, F)$ choices of $\boldsymbol{m}(\cdot)$ and $\boldsymbol{S}(\cdot)$ that attain these lower bounds, we thus have

$$\lfloor (N - d + 1)/2) \rfloor / N \leq \varepsilon_M^{\text{MD}}(\lambda_N, \mathbb{X}_N) \leq (N - d + 1)/2N. \tag{7}$$

The Minimum Covariance Determinant (MCD) estimator of Rousseeuw [16] attains the above lower bound while retaining full affine equivariance, but at some sacrifice of computational ease. On the other hand, a fast algorithm "Fast-MCD" constructed by Rousseeuw and Van Driessen [17] approximates the MCD and is implemented in the R packages *MASS*, *rrcov*, and *robustbase*, as well as in other software packages. Other well-known covariance functionals also attain the lower bound in (7), again at the expense of computational complexity. □

PROOF OF THEOREM 3.2. In an obvious notation, we have

$$\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})} = \left\{ \boldsymbol{x} : \mathrm{MD}(\boldsymbol{x}, \mathbb{X}_{N,k}) \leq \frac{\lambda_N}{1-\lambda_N} \right\},$$

an ellipsoid having center $\boldsymbol{m}(\mathbb{X}_{N,k})$ and contained in the sphere $\mathbb{S}_d(r(\mathbb{X}_{N,k}))$ with radius

$$r(\mathbb{X}_{N,k}) = \|\boldsymbol{m}(\mathbb{X}_{N,k})\| + \left(\frac{\lambda_N}{1-\lambda_N} \right) \sqrt{e\max(\boldsymbol{S}(\mathbb{X}_{N,k}))}.$$

Now (5) is equivalent to {(a) and/or (b)}, with

(a) $\boldsymbol{m}(\mathbb{X}_{N,k}) \to \infty$, with suitable choice of $\mathbb{X}_{N,k}$,

and

(b) volume $\left(\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})} \right) \to \infty$, with suitable choice of $\mathbb{X}_{N,k}$.

Let $RBP(\boldsymbol{m}(\mathbb{X}_N)) = k_1/N$ and $RBP_{\exp}(\boldsymbol{S}(\mathbb{X}_N)) = k_2/N$. By (a) with $k = k_1$, we obtain (5) with $k = k_1$ and thus $\gamma_M(\lambda_N, \mathbb{X}_N, k_1) = 1$. Hence $\varepsilon_M^{\mathrm{MD}}(\lambda_N, \mathbb{X}_N) \leq k_1/N$. Now putting $\varepsilon_M^{\mathrm{MD}}(\lambda_N, \mathbb{X}_N) = k_M^{\mathrm{MD}}(\lambda_N, \mathbb{X}_N)/N$, we note that (5) with $k = k_M^{\mathrm{MD}}(\lambda_N, \mathbb{X}_N)$ holds and so either (a) or (b) with $k = k_M^{\mathrm{MD}}(\lambda_N, \mathbb{X}_N)$ holds. Hence $\varepsilon_M^{\mathrm{MD}}(\lambda_N, \mathbb{X}_N) \geq \min\{k_1, k_2\}/N$. □

3.4.2 Halfspace depth outlier identifier

For the outlier identifier using halfspace outlyingness $O_{\mathrm{H}}(\boldsymbol{x}, F)$, $\boldsymbol{x} \in \mathbb{R}^d$, and with $\boldsymbol{m}_{\mathrm{H}}(\mathbb{X}_N)$ the *halfspace median*, as treated in Donoho and Gasko [5], we establish an upper bound for the MBP.

Theorem 3.3 *Using $O_{\mathrm{H}}(\boldsymbol{x}, F)$ with threshold λ_N,*

$$\varepsilon_M^{\mathrm{H}}(\lambda_N, \mathbb{X}_N) = \min \left\{ RBP(\boldsymbol{m}_{\mathrm{H}}(\mathbb{X}_N)), N^{-1} \left\lceil \left(\frac{1-\lambda_N}{2} \right) N \right\rceil \right\}.$$

PROOF. Subject to $\boldsymbol{m}_{\mathrm{H}}(\mathbb{X}_N)$ *not breaking down*, i.e, subject to $\boldsymbol{m}_{\mathrm{H}}(\mathbb{X}_{N,k})$ remaining within the convex hull $\mathrm{CH}(\mathbb{X}_N)$ over all replacements $\mathbb{Y}_k$, we explore when (5) may or may not hold. We have

$$\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})} = \left\{ \boldsymbol{x} : D_{\mathrm{H}}(\boldsymbol{x}, \mathbb{X}_{N,k})) \geq \frac{1-\lambda_N}{2} \right\}.$$

Choose $\Delta > \sup_{\boldsymbol{x} \in \mathrm{CH}(\mathbb{X}_N)} \|\boldsymbol{x}\|$ and let $\boldsymbol{x}^*$ satisfy $\|\boldsymbol{x}^*\| > \Delta$. In order to achieve $D_{\mathrm{H}}(\boldsymbol{x}^*, \mathbb{X}_{N,k})) \geq (1-\lambda_N)/2$ by replacing k points of $\mathbb{X}_N$ with points in a halfspace containing $\boldsymbol{x}^*$ but not intersecting $\mathrm{CH}(\mathbb{X}_N)$, we need $k = \left\lceil \left(\frac{1-\lambda_N}{2} \right) N \right\rceil$. This can be accomplished for each arbitrarily large Δ and thus (5) follows. By standard arguments analogous to those in the treatment of halfspace depth in Donoho and Gasko [5], no other choice of $\mathbb{X}_{N,k}$ corresponding to a smaller k suffices. □

Remark 3.3 (a) The expression given in Theorem 3.3 depends upon the threshold λ_N.

(b) Donoho and Gasko [5] show that if $\mathbb{X}_N$ is in general position, then the *addition* BP of $\boldsymbol{m}_{\mathrm{H}}(\mathbb{X}_N)$ is $\geq 1/(d+1)$, $d \geq 2$. Further, Donoho and Gasko [5] and Chen [3] show that if the underlying probability measure is absolutely continuous and angularly symmetric, then this BP has almost sure limit $1/3$, $N \to \infty$. From results of Serfling [20] we thus conclude that $RBP(\boldsymbol{m}_{\mathrm{H}}(\mathbb{X}_N))$ $\geq 1/(d+1)$, for $d \geq 2$ and $\mathbb{X}_N$ in general position, and that $\limsup_{N\to\infty} RBP(\boldsymbol{m}_{\mathrm{H}}(\mathbb{X}_N)) \leq 1/3$ if also the underlying probability measure is absolutely continuous and angularly symmetric. It is thus reasonable to use in practice as a heuristic guideline the upper bound

$$\varepsilon_M^{\mathrm{H}}(\lambda_N, \mathbb{X}_N) \leq \min\left\{ N^{-1}\left\lceil \left(\frac{1-\lambda_N}{2}\right) N \right\rceil, \frac{1}{3} \right\} \approx \min\left\{ \frac{1-\lambda_N}{2}, \frac{1}{3} \right\}. \tag{8}$$

□

3.4.3 Projection depth outlier identifier

We take projection outlyingness $O_{\mathrm{P}}(\boldsymbol{x}, F)$, with $(\mu(\cdot), \sigma(\cdot)) = (\mathrm{Med}, \mathrm{MAD})$, and for the sample MAD we use the MAD_{d-1} in the case $d \geq 2$, where MAD_m is the modified version of sample MAD (see Tyler [24], Gather and Hilker [7], and Zuo [28]) given by

$$\mathrm{MAD}_m(\mathbb{Y}_N) = \mathrm{Med}_m\{|Y_1 - \mathrm{Med}(\mathbb{Y}_N)|, \ldots, |Y_N - \mathrm{Med}(\mathbb{Y}_N)|\}$$

with

$$\mathrm{Med}_m(\mathbb{Z}_N) = \frac{1}{2}\left(Z_{(\lfloor \frac{N+m}{2} \rfloor)} + Z_{(\lfloor \frac{N+m+1}{2} \rfloor)} \right),\ 1 \leq m \leq N.$$

The case $m = 1$ gives the usual MAD. We establish an exact result for the corresponding MBP.

Theorem 3.4 *For $O_{\mathrm{P}}(\boldsymbol{x}, F)$ with (μ, σ) given by (Med, MAD), and using sample version MAD_{d-1} for $d \geq 2$, and for $\mathbb{X}_N$ in general position with $N \geq 2(d-1)^2 + d$, we have for threshold λ_N*

$$\varepsilon_M^{\mathrm{P}}(\lambda_N, \mathbb{X}_N) = N^{-1}\left\lceil \frac{N-d+2}{2} \right\rceil.$$

Proof. It is shown in [24] and [7] that, for $d \geq 2$,

$$(\mathrm{Med}, \mathrm{MAD}_{d-1}) \text{ has } \textit{explosion } \mathrm{RBP}^{**} \ = N^{-1}\left\lceil \frac{N-d+2}{2} \right\rceil, \tag{9}$$

where RBP^{**} represents *uniform* RBP, that is, breakdown with respect to the maximum bias taken over all projections $\boldsymbol{u}$.

We now show that if (5) holds, then $k \geq \lceil (N-d+2)/2 \rceil$. For suppose that $k < \lceil (N-d+2)/2 \rceil$. Then $\mathrm{Med}(\boldsymbol{u}'\mathbb{X}_{N,k})$ and $\mathrm{MAD}_{d-1}(\boldsymbol{u}'\mathbb{X}_{N,k})$ remain uniformly bounded above with respect to all $\boldsymbol{u}$ and all choices of $\mathbb{X}_{N,k}$. Let $B_1(\mathbb{X}_N)$ and $B_2(\mathbb{X}_N)$ denote such bounds, respectively. Then

$$\left| \frac{\boldsymbol{u}'\boldsymbol{x} - \mathrm{Med}(\mathbb{X}_{N,k})}{\mathrm{MAD}_{d-1}(\mathbb{X}_{N,k})} \right| \geq \frac{|\boldsymbol{u}'\boldsymbol{x}| - B_1(\mathbb{X}_N)}{B_2(\mathbb{X}_N)}$$

and then

$$O(\boldsymbol{x}, \mathbb{X}_{N,k}) \geq \frac{\sup_{\|\boldsymbol{u}\|=1} |\boldsymbol{u}'\boldsymbol{x}| - B_1(\mathbb{X}_N)}{B_2(\mathbb{X}_N)} \geq \frac{\max_{1 \leq i \leq d} |x_i| - B_1(\mathbb{X}_N)}{B_2(\mathbb{X}_N)}.$$

Therefore, for all sufficiently large $\|\boldsymbol{x}\|$, the point $\boldsymbol{x}$ cannot belong to $\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}$ for any $\mathbb{X}_{N,k}$, i.e., (5) fails to hold. Hence (5) implies $k \geq \lceil (N-d+2)/2 \rceil$.

Now we show the converse. Zuo [28] establishes that, for $d \geq 2$ and $N \geq 2(d-1)^2 + d$, $k = \lceil (N-d+2)/2 \rceil$ contaminants suffice to break down the projection median $\mathrm{PM}(\mathbb{X}_N)$ with $(\mu, \sigma) = (\mathrm{Med}, \mathrm{MAD}_{d-1})$. Thus $\mathrm{PM}(\mathbb{X}_{N,k})$ minimizes $O(\boldsymbol{x}, \mathbb{X}_{N,k})$ but can $\to \infty$ with some sequence $\{\mathbb{X}_{N,k}^{(i)}\}$. Since $\mathrm{PM}(\mathbb{X}_{N,k}) \in \overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}$, (5) holds. □

Remark 3.4 (a) The expression in Theorem 3.4 does not depend upon the threshold λ_N.

(b) With this choice of (μ, σ), the above MBP equals the RBP of $\mathrm{PM}(\mathbb{X}_N)$. □

3.4.4 Spatial and Mahalanobis spatial outlier identifiers

We now consider the outlier identifiers based on the spatial outlyingness $O_{\mathrm{S}}(\boldsymbol{x}, F)$ and, for a weak covariance functional $\boldsymbol{C}(\cdot)$, the associated Mahalanobis spatial outlyingness $O_{\mathrm{MS}}(\boldsymbol{x}, F_{\boldsymbol{X}})$. We obtain for the *spatial* case an exact MBP result, which yields for the *Mahalanobis spatial* case an upper bound that can serve as a heuristic practical guideline.

Theorem 3.5 *For the $O_{\mathrm{S}}(\boldsymbol{x}, F)$ with threshold λ_N,*

$$\varepsilon_M^{\mathrm{S}}(\lambda_N, \mathbb{X}_N) = N^{-1} \left\lceil \left(\frac{1-\lambda_N}{2} \right) N \right\rceil.$$

Corollary 3.1 *Using $O_{\mathrm{MS}}(\boldsymbol{x}, F_{\boldsymbol{X}})$ with threshold λ_N, we have*

$$\varepsilon_M^{\mathrm{MS}}(\lambda_N, \mathbb{X}_N) \leq \min \left\{ RBP_{\exp}(\boldsymbol{C}(\mathbb{X}_N)), N^{-1} \left\lceil \left(\frac{1-\lambda_N}{2} \right) N \right\rceil \right\}.$$

Remark 3.5 As for the halfspace case, the above MBP results depend upon the threshold λ_N. □

Proof of Theorem 3.5. We explore (5) with respect to $O_{\mathrm{S}}(\boldsymbol{x}, F) = \|E\boldsymbol{S}(\boldsymbol{x} - \mathbf{X})\|$ and its sample analogue

$$O_{\mathrm{S}}(\boldsymbol{x}, \mathbb{X}_N) = \left\| N^{-1} \sum_{i=1}^{N} \boldsymbol{S}(\boldsymbol{x} - \mathbf{X}_i) \right\|.$$

First, suppose that (5) fails for some k, and let

$$\sup_{\mathbb{X}_{N,k}} \sup_{\boldsymbol{y} \in \overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})}} \|\boldsymbol{y}\| = B < \infty.$$

Choose $\boldsymbol{x}^*$ with $\|\boldsymbol{x}^*\| > B$. Then $O_{\mathrm{S}}(\boldsymbol{x}^*, \mathbb{X}_{N,k}) > \lambda_N$, each choice of $\mathbb{X}_{N,k}$. In particular, replace $\boldsymbol{X}_{N-k+1}, \ldots, \boldsymbol{X}_N$ by some choice of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$, forming $\mathbb{X}_{N,k}(\boldsymbol{x}^*)$. To choose $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$, first put

$$\boldsymbol{y}^* = \boldsymbol{S} \left(\sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_i) \right).$$

Then

$$\sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_i) = \left\| \sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_i) \right\| \boldsymbol{y}^*.$$

Now let $\boldsymbol{Y}_j = \boldsymbol{Y}_0$, $j = 1, \ldots, k$, with $\boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{Y}_0) = -\boldsymbol{y}^*$, so that $\sum_{j=1}^{k} \boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{Y}_j) = -k\boldsymbol{y}^*$. Write

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_i) &= \frac{\boldsymbol{x}^* - \boldsymbol{X}_1}{\|\boldsymbol{x}^* - \boldsymbol{X}_i\|} + \frac{\boldsymbol{X}_1 - \boldsymbol{X}_i}{\|\boldsymbol{x}^* - \boldsymbol{X}_i\|} \\ &= \frac{\boldsymbol{x}^* - \boldsymbol{X}_1}{\|\boldsymbol{x}^* - \boldsymbol{X}_1\|} \times \frac{\|\boldsymbol{x}^* - \boldsymbol{X}_1\|}{\|\boldsymbol{x}^* - \boldsymbol{X}_i\|} + \frac{\boldsymbol{X}_1 - \boldsymbol{X}_i}{\|\boldsymbol{x}^* - \boldsymbol{X}_i\|}. \end{aligned}$$

Since, as $\|\boldsymbol{x}^*\| \to \infty$,

$$\frac{\boldsymbol{X}_1 - \boldsymbol{X}_i}{\|\boldsymbol{x}^* - \boldsymbol{X}_i\|} \to \boldsymbol{0} \text{ and } \frac{\|\boldsymbol{x}^* - \boldsymbol{X}_1\|}{\|\boldsymbol{x}^* - \boldsymbol{X}_i\|} \to 1,$$

we have $\boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_i) = \boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_1)\,(1 \pm o(1))$ uniformly in $i = 1, \ldots, N-k$, and thus

$$\left\| \sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_i) \right\| = \|\boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_1)\,(N-k)\,(1 \pm o(1))\| = (N-k)\,(1 \pm o(1)).$$

It follows that

$$\begin{aligned} \lambda_N < O_{\mathrm{S}}(\boldsymbol{x}^*, \mathbb{X}_{N,k}(\boldsymbol{x}^*)) &= N^{-1} \left| \left\| \sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_i) \right\| - k \right| \times \|\boldsymbol{y}^*\| \\ &= N^{-1}((N-2k) \pm (N-k)o(1)), \text{ as } \|\boldsymbol{x}^*\| \to \infty, \end{aligned}$$

yielding $k \le \left\lfloor \left(\frac{1-\lambda_N}{2} \right) N \right\rfloor$. Equivalently, if $k \ge \left\lceil \left(\frac{1-\lambda_N}{2} \right) N \right\rceil$, then (5) holds.

For the converse implication, suppose that (5) holds for some k. Then there exists $\{\boldsymbol{x}_n\}$ with $\|\boldsymbol{x}_n\| \to \infty$ satisfying $\boldsymbol{x}_n \in \overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k}(\boldsymbol{x}_n))}$ for some choice of $\mathbb{X}_{N,k}$, say $\mathbb{X}_{N,k}(\boldsymbol{x}_n)$, i.e., we have

$$O_{\mathrm{S}}(\boldsymbol{x}_n, \mathbb{X}_{N,k}(\boldsymbol{x}_n)) \le \lambda_N.$$

Denote the *unreplaced* observations in $\mathbb{X}_N$ by $\boldsymbol{X}_1^{(n)}, \ldots, \boldsymbol{X}_{N-k}^{(n)}$. Now, using similar arguments as above, we have

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{x}_n - \boldsymbol{X}_i^{(n)}) &= \frac{\boldsymbol{x}_n}{\|\boldsymbol{x}_n\|} \times \frac{\|\boldsymbol{x}_n\|}{\|\boldsymbol{x}_n - \boldsymbol{X}_i^{(n)}\|} - \frac{\boldsymbol{X}_i^{(n)}}{\|\boldsymbol{X}_i^{(n)}\|} \times \frac{\|\boldsymbol{X}_i^{(n)}\|}{\|\boldsymbol{x}_n - \boldsymbol{X}_i^{(n)}\|} \\ &= \boldsymbol{S}(\boldsymbol{x}_n)\,(1 + o(1)),\ n \to \infty. \end{aligned} \tag{10}$$

Moreover, (10) holds uniformly over $\mathbb{X}_N$.

Now let $\varepsilon > 0$ be given, small enough that $\left\lceil \left(\frac{1-\lambda_N-\varepsilon}{2}\right) N \right\rceil = \left\lceil \left(\frac{1-\lambda_N}{2}\right) N \right\rceil$. Then there exists $n(\varepsilon)$, which may depend upon $\mathbb{X}_N$, such that for $n > n(\varepsilon)$,

$$\left\| (N-k)^{-1} \sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{x}_n - \boldsymbol{X}_i^{(n)}) \right\| = \|\boldsymbol{S}(\boldsymbol{x}_n)\| \, (1 + o(1)) > 1 - \varepsilon.$$

Also, the replacements $\boldsymbol{Y}_1^{(n)}, \ldots, \boldsymbol{Y}_k^{(n)}$ must satisfy

$$\left\| k^{-1} \sum_{j=1}^{k} \boldsymbol{S}(\boldsymbol{x}_n - \boldsymbol{Y}_j^{(n)}) \right\| < 1 + \varepsilon.$$

Then

$$\begin{aligned} O_{\mathrm{S}}(\boldsymbol{x}_n, \mathbb{X}_{N,k}(\boldsymbol{x}_n)) &= N^{-1} \left\| \sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{x}_n - \boldsymbol{X}_i^{(n)}) + \sum_{j=1}^{k} \boldsymbol{S}(\boldsymbol{x}_n - \boldsymbol{Y}_j^{(n)}) \right\| \\ &\geq N^{-1} \left(\left\| \sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{x}_n - \boldsymbol{X}_i^{(n)}) \right\| - \left\| \sum_{j=1}^{k} \boldsymbol{S}(\boldsymbol{x}_n - \boldsymbol{Y}_j^{(n)}) \right\| \right) \\ &\geq N^{-1}((N-k)(1-\varepsilon) - k(1+\varepsilon)), \end{aligned}$$

yielding $N^{-1}((N-k)(1-\varepsilon) - k(1+\varepsilon)) \leq \lambda_N$, in which case $k \geq \left(\frac{1-\lambda_N-\varepsilon}{2}\right) N$, or, equivalently, $k \geq \left\lceil \left(\frac{1-\lambda_N-\varepsilon}{2}\right) N \right\rceil = \left\lceil \left(\frac{1-\lambda_N}{2}\right) N \right\rceil$, completing the proof. □

Proof (sketch) of Corollary 3.1. For the sample Mahalanobis quantile outlyingness,

$$O_{\mathrm{MS}}(\boldsymbol{x}, \mathbb{X}_N) = \left\| N^{-1} \sum_{i=1}^{N} \boldsymbol{S}(\boldsymbol{C}(\mathbb{X}_N)^{-1/2}(\boldsymbol{x} - \mathbf{X}_i)) \right\|,$$

and thus with

$$\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_{N,k})} = \left\{ \boldsymbol{y} : \left\| N^{-1} \sum_{i=1}^{N} \boldsymbol{S}(\boldsymbol{C}(\mathbb{X}_{N,k})^{-1/2}(\boldsymbol{y} - \mathbf{X}_i)) \right\| > \lambda_N \right\},$$

we explore (5). It is straightforward to see that explosion breakdown of $\boldsymbol{C}(\mathbb{X}_N)$ suffices to cause masking breakdown of $\overline{\mathrm{OR}(\lambda_N, \mathbb{X}_N)}$, yielding

$$\varepsilon_M^{\mathrm{MS}}(\lambda_N, \mathbb{X}_N) \leq RBP_{\exp}(\boldsymbol{C}(\mathbb{X}_N)). \tag{11}$$

To obtain the other upper bound,

$$\varepsilon_M^{\mathrm{MS}}(\lambda_N, \mathbb{X}_N) \leq N^{-1} \left\lceil \left(\frac{1-\lambda_N}{2}\right) N \right\rceil, \tag{12}$$

we need to extend the first part of the proof of Theorem 3.5. First, with $RBP_{\exp}(\boldsymbol{C}(\mathbb{X}_N)) = k_0/N$, suppose that (5) fails for some $k < k_0$, and let B and $\boldsymbol{x}^*$ be defined as previously. Then $O_{\mathrm{MS}}(\boldsymbol{x}^*, \mathbb{X}_{N,k}) > \lambda_N$, each choice of $\mathbb{X}_{N,k}$. Again, in particular, replace $\boldsymbol{X}_{N-k+1}, \ldots, \boldsymbol{X}_N$ by some choice of $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$, forming $\mathbb{X}_{N,k}(\boldsymbol{x}^*)$. However, for choosing $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_k$, the previous first step would now take the form of defining $\boldsymbol{y}^*$ by

$$\boldsymbol{y}^* = \boldsymbol{S}\left(\sum_{i=1}^{N-k} \boldsymbol{S}(\boldsymbol{C}(\mathbb{X}_{N,k}(\boldsymbol{x}^*))^{-1/2}(\boldsymbol{x}^* - \boldsymbol{X}_i))\right).$$

It is quickly clear that this is problematic, so let us first substitute $\boldsymbol{C}(\mathbb{X}_N)$ for $\boldsymbol{C}(\mathbb{X}_{N,k}(\boldsymbol{x}^*))$, since $\mathbb{X}_{N,k}(\boldsymbol{x}^*)$ does not break down $\boldsymbol{C}(\mathbb{X}_N)$. With $\boldsymbol{S}(\boldsymbol{C}(\mathbb{X}_{N,k}(\boldsymbol{x}^*))^{-1/2}(\boldsymbol{x}^* - \boldsymbol{X}_i))$ substituted for $\boldsymbol{S}(\boldsymbol{x}^* - \boldsymbol{X}_i)$, etc., the previous steps of proof go through readily, establishing

$$\lambda_N < N^{-1}((N-2k) \pm (N-k)o(1)), \text{ as } \|\boldsymbol{x}^*\| \to \infty,$$

yielding (12) under the substitution of $\boldsymbol{C}(\mathbb{X}_N)$ for $\boldsymbol{C}(\mathbb{X}_{N,k}(\boldsymbol{x}^*))$. With similar but much more cumbersome steps, the same result without this substitution can be obtained. □

3.5 Selection of the outlier threshold λ_N

Based on the discussion of the contamination model in Section 3.1, we will choose the threshold λ_N to be the $(1-\alpha_N)$th quantile of the distribution of $O(\boldsymbol{X}, G)$ under the ideal distribution G, i.e., $\lambda_N = F^{-1}_{O(\boldsymbol{X},G)}(1-\alpha_N)$, where α_N is a selected value for the approximate false positive rate. In choosing α_N, it is desired that this rate be small relative to the approximate true positive rate, ε_N. That is, we desire that

$$\delta = \alpha_N/\varepsilon_N$$

be small. In terms of ε_N and δ, the threshold λ_N is given by

$$\lambda_N = F^{-1}_{O(\boldsymbol{X},G)}(1-\delta\varepsilon_N). \tag{13}$$

For example, with $\delta = 0.10$, we might choose $\varepsilon_N = 0.25$ to allow for a substantial fraction of outliers, yielding $\lambda_N = F^{-1}_{O(\boldsymbol{X},G)}(0.975)$, or a more moderate $\varepsilon_N = 0.15$, yielding $\lambda_N = F^{-1}_{O(\boldsymbol{X},G)}(0.985)$, or a very modest $\varepsilon_N = 0.02$, yielding $\lambda_N = F^{-1}_{O(\boldsymbol{X},G)}(0.998)$. In such cases with ε_N fixed and not depending upon N, the (approximate) expected number of true outliers $N\varepsilon_N$ grows as $O(N)$.

An alternative approach (Jaeckel [10]) to specification of ε_N argues that the contamination fraction should decrease with increasing sample size N. In this spirit, we might assume

$$\varepsilon_N = \frac{c}{\sqrt{N}}, \tag{14}$$

for some constant c (to be determined). In this case, the (approximate) expected number of true outliers $N\varepsilon_N = c\sqrt{N}$ still is an increasing function of N but grows as $o(N)$. In terms of c and δ, the threshold λ_N is given by

$$\lambda_N = F^{-1}_{O(\boldsymbol{X},G)}(1-c\delta/\sqrt{N}). \tag{15}$$

For given choice of δ, say $\delta = 0.1$, we calibrate the function in (14) by choosing c. For example, relative to a sample of size $N = 100$, we might like the "feel" of allowing for up to 15 outliers, giving $c = 15/\sqrt{100} = 1.5$ and thus $F^{-1}_{O(\boldsymbol{X},G)}(1-0.15/\sqrt{N})$ and thresholds of $F^{-1}_{O(\boldsymbol{X},G)}(0.985)$, $F^{-1}_{O(\boldsymbol{X},G)}(0.993)$, and $F^{-1}_{O(\boldsymbol{X},G)}(0.995)$, respectively, for $N = 100$, 500, and 1000. If, on the other hand, we prefer the "feel" of allowing for 2 outliers in a sample of size $N = 20$, we would obtain $c = 2/\sqrt{20} = 0.45$ and thus $F^{-1}_{O(\boldsymbol{X},G)}(1 - 0.045/\sqrt{N})$ and thresholds of $F^{-1}_{O(\boldsymbol{X},G)}(0.9955)$, $F^{-1}_{O(\boldsymbol{X},G)}(0.9980)$, and $F^{-1}_{O(\boldsymbol{X},G)}(0.9986)$, respectively, for $N = 100$, 500, and 1000.

Other scenarios for setting ε_N are possible. Clearly, in any case the threshold λ_N needs to be a relatively high quantile of $F_{O(\boldsymbol{X},G)}$.

3.6 λ_N and MBP with G multivariate normal

We compare values of MBP for different outlier identifiers within the framework of a common contamination model. First, values of δ and ε_N are fixed, and then for each outlyingness function the corresponding threshold λ_N is determined via (13). Here we select ε_N using (14) with a fixed choice of c, so that λ_N is given by (15). For the "MD", "H", "P", and "MS" outlier identifiers, we carry out this approach with G *multivariate normal*. Since these four procedures are based on affine invariant outlyingness functions, it suffices without loss of generality to take G to be *standard d-variate normal*, $G_0 = N(\mathbf{0}, \boldsymbol{I}_d)$.

Our first step is to obtain the distribution $F_{O(\boldsymbol{X},G_0)}$ for each outlyingness function under consideration. We denote by χ^2_ν a random variable having the chi-square distribution with ν degrees of freedom.

Lemma 3.1 *(i) For Mahalanobis distance outlyingness with mean $\boldsymbol{\mu}(F)$ and covariance $\boldsymbol{\Sigma}(F)$ as the location and dispersion measures (= $\mathbf{0}$ and $\boldsymbol{I}_d$ for G_0),*

$$F_{O_{\mathrm{MD}}(\boldsymbol{X},G_0)}(\lambda) = P\left(\chi^2_d \le \left(\frac{\lambda}{1-\lambda}\right)^2\right),\ 0 \le \lambda < 1. \tag{16}$$

(ii) For halfspace outlyingness,

$$\begin{aligned} F_{O_{\mathrm{H}}(\boldsymbol{X},G_0)}(\lambda) &= P\left(\chi^2_d \le \left[\Phi^{-1}\left(\frac{1+\lambda}{2}\right)\right]^2\right) \\ &\left[= P\left(\chi^2_d \le \left[\Phi^{-1}\left(\frac{1-\lambda}{2}\right)\right]^2\right)\right],\ 0 \le \lambda < 1. \end{aligned} \tag{17}$$

(iii) For projection outlyingness with $(\mu, \sigma) = (Med, MAD)$ and using sample version MAD_{d-1} for MAD,

$$F_{O_{\mathrm{P}}(\boldsymbol{X},G_0)}(\lambda) = P\left(\chi^2_d \le \left[\Phi^{-1}\left(\frac{3}{4}\right)\frac{\lambda}{1-\lambda}\right]^2\right),\ 0 \le \lambda < 1. \tag{18}$$

Remark 3.6 (a) The location and scatter measures for Mahalanobis distance outlyingness in (i) and for projection outlyingness in (iii) accommodate straightforward derivations of the distributions $F_{O(\boldsymbol{X},G_0)}(\lambda)$. Other choices would yield somewhat different distributions in (i) and (iii). Thus the results in Lemma 3.1 serve merely as benchmarks.

(b) We lack an explicit result for $F_{O_{\rm MS}(\boldsymbol{X},G_0)}$ for $G_0 = N(\boldsymbol{0}, \boldsymbol{I}_d)$, even with covariance $\boldsymbol{\Sigma}(F)$ as the scatter measure. However, as may be needed in any application, this distribution can be determined numerically, of course.

(c) Furthermore, since in practice we actually are interested in thresholds pertaining to the cdf of $O(\boldsymbol{X}, \widehat{G}_0)$ based on a sample estimate of G_0, empirical thresholds based on appropriate sampling experiments are more apropos and accommodate any choices of location and dispersion measures. We use such an approach in the numerical study in Section 4. □

Proof of Lemma 3.1. (i) We have

$$\begin{aligned} P(O_{\rm MD}(\boldsymbol{X}, G_0) \le \lambda) &= P\left(\mathrm{MD}(\boldsymbol{X}, G_0) \le \frac{\lambda}{1-\lambda}\right) \\ &= P\left(\|\boldsymbol{\Sigma}(G_0)^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}(G_0))\| \le \frac{\lambda}{1-\lambda}\right), \end{aligned}$$

yielding (16).

(ii) From Donoho and Gasko [5] we have $D_{\rm H}(\boldsymbol{x}, G_0) = \Phi(-\|\boldsymbol{x}\|)$, and hence $O_{\rm H}(\boldsymbol{x}, F) = 1 - 2\Phi(-\|\boldsymbol{x}\|)$, from which the equality in (17) readily follows. The second inequality in (ii) follows from $\Phi^{-1}((1+\lambda)/2) = -\Phi^{-1}((1-\lambda)/2)$.

(iii) From Zuo [28] we have $D_{\rm P}(\boldsymbol{x}, G_0) = \Phi^{-1}(3/4)/(\Phi^{-1}(3/4) + \|\boldsymbol{x}\|)$, and hence

$$O_{\rm P}(\boldsymbol{x}, G_0) = \frac{\|\boldsymbol{x}\|}{\Phi^{-1}(3/4) + \|\boldsymbol{x}\|},$$

leading to (18). □

With the notation

$$Q(d, \alpha) = \sqrt{\left(\chi^2_d\right)^{-1}(1-\alpha)},$$

the formula (15) for λ_N as a function of specified false positive rate $\alpha_N = c\delta/\sqrt{N}$ yields via Lemma 3.1 the following threshold values.

Corollary 3.2 *(i) For Mahalanobis distance outlyingness with mean $\boldsymbol{\mu}(F)$ and covariance $\boldsymbol{\Sigma}(F)$ as the location and dispersion measures (= $\boldsymbol{0}$ and $\boldsymbol{I}_d$ for G_0),*

$$\lambda_N = \frac{Q(d, c\delta/\sqrt{N})}{1 + Q(d, c\delta/\sqrt{N})}. \tag{19}$$

(ii) For halfspace outlyingness,

$$\lambda_N = 2\Phi(Q(d, c\delta/\sqrt{N})) - 1. \tag{20}$$

(iii) For projection outlyingness with $(\mu, \sigma) = (Med, MAD)$ *and using sample version* MAD_{d-1} *for MAD,*

$$\lambda_N = \frac{Q(d, c\delta/\sqrt{N})}{(\Phi^{-1}(3/4) + Q(d, c\delta/\sqrt{N})}. \tag{21}$$

It turns out that the range λ_N does not vary a lot over typical values of N, d, c, and δ.

Example 3.1 For $c = 1.5$ and $\delta = 0.1$, i.e., $\alpha_N = \delta c/\sqrt{N} = 0.15/\sqrt{N}$, and for $N =$ **100**, **500**, and **1000** and dimension $d =$ **2**, **5**, **10**, **15**, and **20**, the solutions λ_N given in Corollary 3.2 range tightly:

- For Mahalanobis distance outlyingness, $\mathbf{0.74} \leq \lambda_{\mathbf{N}} \leq \mathbf{0.86}$.
- For halfspace outlyingness, $\mathbf{0.996} \leq \lambda_N \leq \mathbf{1.00}$.
- For projection outlyingness, $\mathbf{0.81} \leq \lambda_N \leq \mathbf{0.90}$.

For halfspace outlyingness, the upper bound $(1 - \lambda_N)/2$ for the MBP is very small for the above range of λ_N. For Mahalanobis distance and projection outlyingness, however, the range of λ_N imposes no restriction on the MBP. □

3.7 Conclusions and Comments

For *classifying points as "outliers" or not*, using a threshold λ_N determined by a G-based false positive rate, the Mahalanobis distance and projection outlyingness allow choices of λ_N with both *high MBP* and *low false positive rate*. For Mahalanobis spatial outlyingness, we lack an explicit theoretical result connecting MBP and false positive rate, but the numerical study in Section 4 shows that with this outlyingness one can set a low false positive rate and still have the MBP at levels often acceptable. Although its MBP is not as high as for the Mahalanobis distance and projection outlier identifiers, the Mahalanobis spatial outlier identifier remains competitive because its contours are not constrained to be elliptical and its computational burden is not intensive. Thus the Mahalanobis distance, projection, and Mahalanobis spatial identifiers offer *satisfactory masking protection*. The halfspace outlier identifier, however, entails a *severe and unacceptable tradeoff* between MBP and false positive rate.

On the other hand, all four of these outlyingness functions can be used for purposes such as *robust outlyingness ranking* of points in $\mathbb{X}_N$, or *robust location estimation*. These goals are quite different from that of setting an *outlyingness threshold*.

The Mahalanobis distance and projection approaches succeed especially well perhaps due to requiring only a limited objective, *robust estimation of location and scale parameters*, which then determine the outlyingness function. The halfspace and Mahalanobis spatial approaches entail a wider and more challenging inference objective, *robust nonparametric estimation of the outlyingness function*.

4 A brief numerical experiment

Here we provide a brief but illustrative numerical study to explore the qualitative findings of Section 3 by comparing the four outlier identifiers empirically. A detailed and comprehensive study is of considerable interest but beyond the scope of the present paper.

4.1 The simulation plan

The data $\mathbb{X}_N$ consists of a sample of size $N = 100$ from the bivariate standard normal distribution, and we consider a contamination model with $c = 1.5$ and $\delta = 0.1$ as in Example 3.1, so that the approximate true positive rate due to contamination becomes $\varepsilon_{100} = 0.15$ and the approximate false positive rate under no contamination is $\alpha_{100} = 0.015$. In fact, taking account of the discrete sample size $N = 100$, we shall use $\alpha_{100} = 0.01$ and thus expect the (uncontaminated) data $\mathbb{X}_{100}$ to contain one or two observations with outlyingness beyond the threshold value λ_{100} determined by $\alpha_{100} = 0.01$ for the particular outlyingness function under consideration. For reasons as discussed in Remark 3.6, however, we use sample-based thresholds consistent with $\alpha_{100} = 0.01$, namely *sample* α_{100} = the *largest* observation in the *uncontaminated* sample of size 100.

As evident in the proof of Lemma 3.1, each of the outlyingness functions $O_{\mathrm{MD}}(\boldsymbol{x}, G_0)$, $O_{\mathrm{H}}(\boldsymbol{x}, G_0)$, and $O_{\mathrm{P}}(\boldsymbol{x}, G_0)$, evaluated at G_0, is an increasing function of $\|\boldsymbol{x}\|$. For convenience later, we index the data points $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{100}$ in order of increasing $\|\boldsymbol{X}_i\|$.

Six affine invariant sample outlier identifiers are considered:

- *Classical Mahalanobis distance (CMD)*: $O_{\mathrm{MD}}(\boldsymbol{x}, \mathbb{X}_N)$ with the *classical* location and covariance estimators, $\overline{\boldsymbol{X}}$ and $\boldsymbol{S}$. This, of course, is nonrobust.
- *Robust Mahalanobis distance (RMD)*: $O_{\mathrm{MD}}(\boldsymbol{x}, \mathbb{X}_N)$ with *robust* location and covariance estimators, using the minimum covariance determinant (MCD) method of Rousseeuw [16] and Rousseeuw and Van Driessen [17], as computed in the package *robust* in R.
- *Halfspace (H)*: $O_{\mathrm{H}}(\boldsymbol{x}, \mathbb{X}_N)$.
- *Classical Mahalanobis spatial (CMS)*: $O_{\mathrm{MS}}(\boldsymbol{x}, \mathbb{X}_N)$ with $\boldsymbol{S}$. This is nonrobust.
- *Robust Mahalanobis spatial (RMS)*: $O_{\mathrm{MS}}(\boldsymbol{x}, \mathbb{X}_N)$ with the MCD covariance estimator.
- *Projection (P)*: $O_{\mathrm{P}}(\boldsymbol{x}, \mathbb{X}_N)$ with $(\mathrm{Med}, \mathrm{MAD}_{d-1})$ as sample (μ, σ).

We explore the masking robustness of CMD, RMD, H, CMS, RMS, and P, with respect to two scenarios for replacement of the 15 most outlying points of $\mathbb{X}_{100}$ by extreme outliers.

- *Scenario A.* The points $\boldsymbol{X}_{86}, \ldots, \boldsymbol{X}_{100}$ are replaced, respectively, by $K\boldsymbol{X}_{86}, \ldots, K\boldsymbol{X}_{100}$ for some inflation factor K. We shall use $K = 5$. Each outlier lies along the ray in the direction of the replaced point from the origin. Denote the modified data set by $\mathbb{X}_{100}^{(\mathrm{A})}$.
- *Scenario B.* The points $\boldsymbol{X}_{86}, \ldots, \boldsymbol{X}_{100}$ are replaced by $K_1\boldsymbol{X}_{100}, \ldots, K_{15}\boldsymbol{X}_{100}$ for some respective inflation factors $K_1, \ldots, K_{15}$. We shall use $K_1 = 1.25$, $K_2 = 1.50$, $K_3 = 1.75$, $K_4 = 2.00$,

$K_5 = 2.25, \ldots, K_{15} = 4.75$. These outliers are spread along the single ray in the direction of $\boldsymbol{X}_{100}$ from the origin. Denote the modified data set by $\mathbb{X}_{100}^{(\mathrm{B})}$.

4.2 The simulation results

Figure 1 displays the original data set $\mathbb{X}_{100}$ consisting of 100 observations from the bivariate standard normal distribution. Also indicated are the modified data sets $\mathbb{X}_{100}^{(\mathrm{A})}$ and $\mathbb{X}_{100}^{(\mathrm{B})}$ resulting from contamination under Scenarios A and B, respectively, for replacement of the 15 original sample points that have the greatest Euclidean distance $\|\boldsymbol{x}\|$ from the origin.

The case of no contamination

For the 25 sample cases with uppermost $\|\boldsymbol{x}\|$, labeled with row index i corresponding to order of increasing $\|\boldsymbol{x}\|$, the sample outlyingness values in the case of *no contamination* are listed in Table 1 for each of CMD, H, CMS, RMD, RMS, and P. For each, the *largest* outlyingness value is indicated in boldface. Reflecting a 1% false positive rate, these values will serve as the relevant thresholds for outlier detection under the contamination scenarios A and B. Although, for at least CMD, H, and P, the population outlyingness values $O(\boldsymbol{x}, G_0)$ are monotone increasing with $\|\boldsymbol{x}\|$, the corresponding sample versions need not strictly follow such monotonicity, of course.

For CMD, H, and P, we can compare the respective sample-based thresholds **0.76**, **0.98**, and **0.86** with the population-based thresholds determined by Corollary 3.2 for $\alpha_{100} = 0.01$, i.e., with $Q(2, 0.01) = 9.21$. For CMD, (19) yields $\lambda_{100} = \sqrt{9.21}/(1 + \sqrt{9.21}) = 3.03/(1 + 3.03) = \mathbf{0.75}$. For H, (20) yields $\lambda_{100} = 2\Phi(3.03) - 1 = \mathbf{0.9976}$, and, for P, (21) yields $\lambda_{100} = \mathbf{0.82}$. The population and outlyingness values agree fairly well for all 25 cases listed, although for H, however, the step function character of the sample halfspace depth results in minimum sample depth $1/N = 1/100 = 0.01$ for all points on the boundary of the convex hull of the data, yielding maximum possible sample outlyingness **0.98**. Therefore, sample outlyingness values (using H) cannot reach the theoretical threshold of **0.9976** without a much larger sample size N, although then the relevant threshold would become even closer to **1.00**.

The case of contamination and nonrobust identifiers

For the 25 cases of Table 1, the performance of two nonrobust outlier identifiers, CMD and CMS, is illustrated in Table 2 under Scenarios A and B for replacement contamination of the 15 cases 86-100 by outliers. For purposes of comparison, the unaltered cases 76-85 are retained. For all 25 cases, the original sample outlyingness values and those under Scenarios A and B are shown.

For CMD, it is seen that under Scenario A most (12 out of 15) outliers are detected (and the other 3 are almost detected), whereas in Scenario B only the most extreme 3 cases are detected leaving the other 12 outliers masked (although 2 of these are almost detected). This is not surprising, since in Scenario B the sample mean is pulled in the direction of the 15 outliers. We also note that under Scenario A the outlyingness values of cases 76-85 change considerably, becoming far below the threshold 0.76. For CMS, 9 of the outliers are detected under Scenario A (with 3 more almost detected), while under Scenario B only the 2 most extreme are detected and one *nonoutlier*, case 82,

is misidentified as an outlier. Of course, a good identifier should perform well under both Scenario A and Scenario B. As expected, neither CMD nor CMS meets this criterion.

The case of contamination and weakly robust identifiers

Table 3 illustrates the performance of two weakly masking robust outlier identifiers, H and RMS, under Scenarios A and B for replacement of 15 sample points by outliers.

For H, the 9 cases that met the threshold without contamination are also detected under Scenario A, while under Scenario B only the most extreme case among the created outliers is detected, the others being masked. Also, 6 among the 10 *nonoutliers* are misidentified as outliers, indicating a serious masking problem with H.

For RMS, 6 of the created outliers are detected under Scenario A and 3 under Scenario B, but none of the nonoutliers are misidentified as outliers.

Of course, this weak masking performance is anticipated in the present case of a high threshold combined with a high level of contamination. The associated masking breakdown points, MBP $= (1 - \lambda)/\lambda)$, are $.02/.98 = .02$ and $.04/.96 = .04$ for H and CMS, respectively, whereas the contamination level is 0.15. For a contamination level of 0.03, however, we see that CMS (but not H) performs well in detecting the 3 most extreme outliers under either scenario.

The case of contamination and strongly robust identifiers

Table 4 illustrates the performance of two strongly masking robust outlier identifiers, RMD and P, under Scenarios A and B for replacement of 15 sample points by outliers. Both have excellent performance, each identifying all 15 outliers and only these, under both scenarios. Of course, this is expected from the high masking breakdown points, independent of the threshold, possessed by these procedures. The identifiers RMD and P are competitive with each other, with RMD more favorable computationally but P not constrained to follow elliptical outlyingness contours.

4.3 Practical recommendations

The findings of this sampling study are consistent with the general conclusions of Section 3.7 based on theoretical MBPs considered relative to a low false positive rate. Under two quite different scenarios for creation of 15 outliers by replacement in a sample of size 100 from standard bivariate normal, the outlyingness functions RMD and P are considerably superior in performance. At the other extreme are CMD, CMS, and H. In between falls RMS, which is competitive in the case of a small level of contamination.

Acknowledgments

The authors highly value an anonymous reviewer's very pertinent and constructive suggestions, which led to significant improvements. We also thank Satyaki Mazumder for a careful reading of an early draft, leading to several corrections. Support by NSF Grants DMS-0103698, CCF-0430366, and DMS-0805786, and by NSA Grant H98230-08-1-0106, is gratefully acknowledged.

References

[1] Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd edition. Wiley.

[2] Becker, C. and Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association* **94** 947–955.

[3] Chen, Z. (1995). Bounds for the breakdown point of the simplicial median. *Journal of Multivariate Analysis* **55** 1–13.

[4] Davies, L. and Gather, U. (1993). The identification of multiple outliers (with discussion). *Journal of the American Statistical Association* **88** 782–801.

[5] Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estinates based on halfspace depth and projected outlyingness. *Annals of Statistics* **20** 1803–1827.

[6] Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift foe Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) pp. 157-184, Wadsworth, Belmont, California.

[7] Gather, U. and Hilker, T. (1997). A note on Tyler's modification of the MAD for the Stahel-Donoho estimator. *Annals of Statistics* **25** 2024-2026.

[8] Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd edition. Wiley, New York.

[9] Hawkins, D.M. (1980). *Identification of Outliers*. Chapman and Hall.

[10] Jaeckel, L. A. (1971). Robust estimates of location: symmetry and asymmetric contamination. *Annals of Mathematical Statistics* **42** 102O–1034.

[11] Liu, R. Y. (1992). Data depth and multivariate rank tests. In L_1-*Statistics and Related Methods* (Y. Dodge, ed.) 279–294. North-Holland, Amsterdam.

[12] Liu, R.Y., Parelius, J.M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). *Annals of Statistics* **27** 783–858.

[13] Lopuhaä, H. P. and Rousseeuw, J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Annals of Statistics* **19** 229-248.

[14] Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India* **12** 49–55.

[15] Müller, C. H. (2005). Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis* **95** 153–181.

[16] Rousseeuw P. (1985). Multivariate estimation with high breakdown point. In: W. Grossmann, G. Pflug, and W. Wertz (Eds.), *Mathematical Statistics and Applications, Vol. B*, Reidel Publishing, Dordrecht, 283–297.

[17] Rousseeuw, P. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.

[18] Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica* **56** 214–232.

[19] Serfling, R. (2006). Depth functions in nonparametric multivariate analysis. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications* (R. Y. Liu, R. Serfling, D. L. Souvaine, eds.), pp. 1–16. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Volume 72, American Mathematical Society.

[20] Serfling, R. (2009a). Inequalities relating addition and replacement type finite sample breakdown points. *International Journal of Statistical Sciences*, to appear (Special Issue on Nonparametric Statistics in honor of Professor A. K. Md. Ehsanes Saleh)

[21] Serfling, R. (2009b). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. In review.

[22] Silverman, B. W. (1986). *Density Estimation*. Chapman and Hall, London.

[23] Tukey, J.W. (1975). Mathematics and picturing data. In *Proceedings of the 1974 International Congress of Mathematicians* (R. James, ed.) **2** 523–531. Canadian Math. Congress.

[24] Tyler, D.E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics. *Annals of Statistics* **22** 1024–1044.

[25] Vardi, Y. and Zhang, C.H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of National Academy of Science USA* **97** 1423–1426.

[26] Zhang, J. (2002). Some extensions of Tukey depth function. *Journal of Multivariate Analysis* **82** 134–165.

[27] Zuo, Y. (2001). Some quantitative relationships between two types of finite sample breakdown point. *Statistics & Probability Letters* **51** 369–375.

[28] Zuo, Y. (2003). Projection-based depth functions and associated medians. *Annals of Statistics* **31** 1460–1490.

[29] Zuo, Y., Cui, H. and Young, D. (2004). Influence function and maximum bias of projection depth based estimators. *Annals of Statistics* **32** 189–218.

[30] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics* **28** 461–482.

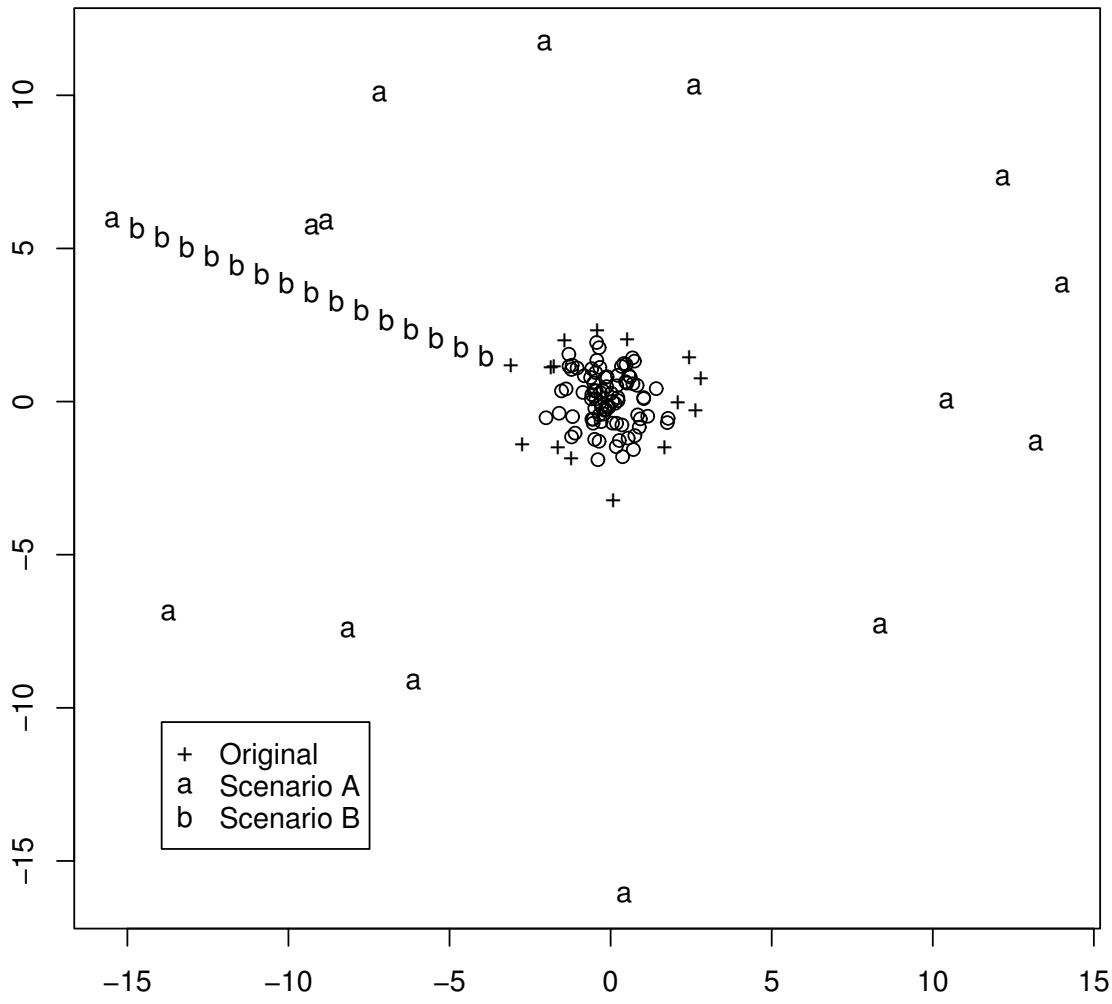

Figure 1: Plot of 100 observations from bivariate standard normal, the 15 most outlying observations represented by +. Scenario A inflates the 15 most outlying observations by the factor $K = 5$, to locations indicated by "a". Scenario B replaces the 15 most outlying observations by more extreme values along a single direction, to locations indicated by "b".

$O(\boldsymbol{x}, \mathbb{X}_{100})$						
i	CMD	H	CMS	RMD	RMS	P
76	0.64	0.92	0.82	0.64	0.80	0.75
77	0.60	0.88	0.76	0.59	0.76	0.73
78	0.62	0.90	0.77	0.59	0.75	0.71
79	0.65	0.96	0.84	0.65	0.81	0.75
80	0.66	0.90	0.82	0.70	0.85	0.79
81	0.66	0.92	0.83	0.70	0.85	0.79
82	0.67	0.96	0.85	0.65	0.82	0.77
83	0.64	0.94	0.83	0.61	0.80	0.74
84	0.64	0.92	0.83	0.63	0.82	0.76
85	0.66	0.94	0.85	0.68	0.87	0.79
86	0.68	0.94	0.85	0.72	0.88	0.82
87	0.65	0.94	0.84	0.66	0.85	0.78
88	0.67	0.96	0.88	0.66	0.85	0.77
89	0.66	0.96	0.85	0.67	0.87	0.78
90	0.69	0.96	0.88	0.69	0.87	0.80
91	0.69	0.96	0.88	0.69	0.88	0.80
92	0.69	0.98	0.88	0.71	0.89	0.80
93	0.69	0.98	0.90	0.66	0.87	0.77
94	0.69	0.98	0.90	0.68	0.89	0.79
95	0.73	0.98	0.92	0.76	0.94	0.85
96	0.74	0.98	0.93	0.77	0.94	0.85
97	0.75	0.98	0.93	**0.78**	0.95	**0.86**
98	0.75	0.98	0.94	0.77	0.95	0.85
99	**0.76**	0.98	**0.95**	0.75	0.93	0.83
100	0.75	**0.98**	0.94	0.77	**0.96**	0.85

Table 1: Sample outlyingness values $O(\boldsymbol{x}, \mathbb{X}_{100})$, for CMD, H, CMS, RMD, RMS, and P. Largest outlyingness value defining relevant sample threshold λ_{100} is indicated in bold.

	CMD			CMS		
	sample $\lambda_{100} =$ **0.76**			sample $\lambda_{100} =$ **0.95**		
i	$O(\boldsymbol{x}, \mathbb{X}_{100})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{A})})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{B})})$	$O(\boldsymbol{x}, \mathbb{X}_{100})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{A})})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{B})})$
76	0.64	0.37	0.64	0.82	0.79	0.84
77	0.60	0.29	0.43	0.76	0.72	0.48
78	0.62	0.32	0.66	0.77	0.75	0.89
79	0.65	0.38	0.69	0.84	0.81	0.92
80	0.66	0.36	0.48	0.82	0.76	0.63
81	0.66	0.36	0.48	0.83	0.77	0.64
82	0.67	0.39	0.72	0.85	0.81	**0.96**
83	0.64	0.35	0.68	0.83	0.78	0.93
84	0.64	0.33	0.56	0.83	0.77	0.71
85	0.66	0.34	0.62	0.85	0.76	0.79
86	0.68	0.73	0.40	0.85	0.92	0.51
87	0.65	0.74	0.47	0.84	0.92	0.57
88	0.67	**0.76**	0.52	0.88	0.94	0.62
89	0.66	0.74	0.57	0.85	0.93	0.67
90	0.69	**0.76**	0.60	0.88	**0.95**	0.70
91	0.69	**0.76**	0.63	0.88	0.94	0.74
92	0.69	**0.76**	0.66	0.88	**0.95**	0.77
93	0.69	**0.78**	0.68	0.90	**0.95**	0.80
94	0.69	**0.78**	0.70	0.90	**0.95**	0.83
95	0.73	**0.78**	0.72	0.92	**0.96**	0.85
96	0.74	**0.79**	0.74	0.93	**0.96**	0.88
97	0.75	**0.79**	0.75	0.93	**0.96**	0.90
98	0.75	**0.80**	**0.76**	0.94	**0.97**	0.93
99	**0.76**	**0.83**	**0.78**	**0.95**	**0.98**	**0.95**
100	0.75	**0.81**	**0.79**	0.94	**0.97**	**0.97**

Table 2: Performance of nonrobust outlier identifiers, CMD and CMS, under Scenarios A and B for replacement of 15 sample points by outliers (cases $i = 86, \ldots, 100$). Outlyingness values at or above relevant sample threshold λ_{100} are indicated in bold.

	H			RMS		
	sample $\lambda_{100} =$ **0.98**			sample $\lambda_{100} =$ **0.96**		
i	$O(\boldsymbol{x}, \mathbb{X}_{100})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{A})})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{B})})$	$O(\boldsymbol{x}, \mathbb{X}_{100})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{A})})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{B})})$
76	0.92	0.86	**0.98**	0.80	0.76	0.85
77	0.88	0.82	0.66	0.76	0.72	0.59
78	0.90	0.82	0.96	0.75	0.70	0.69
79	0.96	0.86	**0.98**	0.81	0.77	0.85
80	0.90	0.88	**0.98**	0.85	0.80	0.93
81	0.92	0.86	**0.98**	0.85	0.80	0.93
82	0.96	0.88	**0.98**	0.82	0.78	0.81
83	0.94	0.86	**0.98**	0.80	0.74	0.72
84	0.92	0.86	0.82	0.82	0.76	0.65
85	0.94	0.88	0.90	0.87	0.80	0.69
86	0.94	0.94	0.70	0.88	0.93	0.70
87	0.94	0.94	0.72	0.85	0.92	0.73
88	0.96	0.96	0.74	0.85	0.93	0.75
89	0.96	0.96	0.76	0.87	0.94	0.77
90	0.96	0.96	0.78	0.87	0.94	0.79
91	0.96	0.96	0.80	0.88	0.94	0.82
92	**0.98**	**0.98**	0.82	0.89	0.94	0.84
93	**0.98**	**0.98**	0.84	0.87	0.94	0.86
94	**0.98**	**0.98**	0.86	0.89	0.94	0.88
95	**0.98**	**0.98**	0.88	0.94	**0.97**	0.90
96	**0.98**	**0.98**	0.90	0.94	**0.97**	0.92
97	**0.98**	**0.98**	0.92	0.95	**0.98**	0.94
98	**0.98**	**0.98**	0.94	0.95	**0.98**	**0.96**
99	**0.98**	**0.98**	0.96	0.93	**0.97**	**0.98**
100	**0.98**	**0.98**	**0.98**	**0.96**	**0.98**	**1.00**

Table 3: Performance of two weakly robust outlier identifiers, H and RMS, under Scenarios A and B for replacement of 15 sample points by outliers (cases $i = 86, \ldots, 100$). Outlyingness values at or above relevant sample threshold λ_{100} are indicated in bold.

	RMD			P		
	sample $\lambda_{100} =$ **0.78**			sample $\lambda_{100} =$ **0.86**		
i	$O(\boldsymbol{x}, \mathbb{X}_{100})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{A})})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{B})})$	$O(\boldsymbol{x}, \mathbb{X}_{100})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{A})})$	$O(\boldsymbol{x}, \mathbb{X}_{100}^{(\mathrm{B})})$
76	0.64	0.64	0.64	0.75	0.75	0.76
77	0.59	0.61	0.61	0.73	0.71	0.69
78	0.59	0.60	0.60	0.71	0.71	0.76
79	0.65	0.64	0.64	0.75	0.75	0.78
80	0.70	0.69	0.69	0.79	0.78	0.79
81	0.70	0.69	0.69	0.79	0.78	0.79
82	0.65	0.65	0.65	0.77	0.76	0.81
83	0.61	0.63	0.63	0.74	0.73	0.77
84	0.63	0.64	0.64	0.76	0.74	0.72
85	0.68	0.69	0.69	0.79	0.79	0.78
86	0.72	**0.92**	**0.81**	0.82	**0.95**	**0.87**
87	0.66	**0.92**	**0.84**	0.78	**0.94**	**0.89**
88	0.66	**0.91**	**0.86**	0.77	**0.94**	**0.90**
89	0.67	**0.92**	**0.87**	0.78	**0.95**	**0.92**
90	0.69	**0.92**	**0.89**	0.80	**0.95**	**0.93**
91	0.69	**0.92**	**0.90**	0.80	**0.95**	**0.93**
92	0.71	**0.92**	**0.91**	0.80	**0.95**	**0.94**
93	0.66	**0.92**	**0.91**	0.77	**0.95**	**0.94**
94	0.68	**0.92**	**0.92**	0.79	**0.95**	**0.95**
95	0.76	**0.94**	**0.93**	0.85	**0.96**	**0.95**
96	0.77	**0.94**	**0.93**	0.85	**0.96**	**0.95**
97	**0.78**	**0.94**	**0.93**	**0.86**	**0.97**	**0.96**
98	0.77	**0.95**	**0.94**	0.85	**0.97**	**0.96**
99	0.75	**0.94**	**0.94**	0.83	**0.96**	**0.96**
100	0.77	**0.95**	**0.94**	0.85	**0.97**	**0.96**

Table 4: Performance of two strongly robust outlier identifiers, RMD and P, under Scenarios A and B for replacement of 15 sample points by outliers (cases $i = 86, \ldots, 100$). Outlyingness values at or above relevant sample threshold λ_{100} are indicated in bold.