

Nonparametric Outlier Identification in \mathbb{R}^d

Robert Serfling¹

University of Texas at Dallas

Joint Statistical Meetings 2008 – Denver

¹www.utdallas.edu/~serfling

Construction of Outlyingness Functions on \mathbb{R}^d

On Mahalanobis Type Outlyingness Functions

On Addition Versus Replacement Breakdown Points

Nonparametric Outlier Identifiers in \mathbb{R}^d , and Comparison by a Masking Breakdown Point Criterion

Outliers Have Been Discussed for Centuries

- ▶ *Francis Bacon, 1620:*
“Whoever knows the ways of Nature will more easily notice her deviations ... ”
- ▶ *Daniel Bernoulli, 1777:*
“I do not condemn in every case the principle of rejecting an observation, indeed I approve it wherever ... ”
- ▶ *B. Pierce, 1852:*
“In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error ... ”

Including, More Recently, the Multivariate Case

- ▶ *Barnett and Lewis, 1995:*
“As Gnanadesikan and Kettenring (1972) remark, a multivariate outlier no longer has a simple manifestation as an observation which ‘sticks out at the end’ of the sample. The sample has no ‘end’! But, notably in bivariate data, we may still perceive an observation as suspiciously aberrant from the data mass ...”
- ▶ *In higher dimension, the first problem is identification of outliers. For this we need algorithmic approaches, i.e., we need to define and use outlyingness functions.*

Finding Outliers via SPSS (For Example)

And SPSS offers algorithms for you!!!

SPSS's advertisement in *Amstat News*, 2007:

Quickly find multivariate outliers

Prevent outliers from skewing analyses when you use the Anomaly Detection Procedure. This procedure searches for unusual cases based on deviations from similar cases and gives reasons for such deviations. You can flag outliers by creating a new variable. Once you have identified unusual cases, you can further examine them and determine if they should be included in your analyses.

But ... Let's Define Outlyingness Functions Formally, for Purposes of Analysis and Comparison

Given a cdf F on \mathbb{R}^d , an outlyingness function $O(\mathbf{x}, F)$ provides an associated center-outward ordering of points \mathbf{x} in \mathbb{R}^d with *higher* values representing greater “outlyingness”.

Three Related Functions

- ▶ Given a cdf F on \mathbb{R}^d , a depth function $D(\mathbf{x}, F)$ provides an associated *center-outward ordering* of points \mathbf{x} in \mathbb{R}^d .
- ▶ A quantile function $\mathbf{Q}(\mathbf{u}, F)$ in \mathbb{R}^d attaches to each point \mathbf{x} a “quantile representation” indexed by \mathbf{u} in the unit ball $\mathbb{B}^{d-1}(\mathbf{0})$, with *nested* contours

$$\{\mathbf{Q}(\mathbf{u}, F) : \|\mathbf{u}\| = c\}, \quad 0 \leq c < 1.$$

- ▶ A centered rank function $\mathbf{R}(\mathbf{x}, F)$ in \mathbb{R}^d takes values in the unit ball, with the origin of the ball assigned to a selected multivariate median \mathbf{M}_F . Thus $\mathbf{R}(\mathbf{x}, F) = \mathbf{0}$ for $\mathbf{x} = \mathbf{M}_F$ and for other \mathbf{x} gives a “directional rank” in $\mathbb{B}^{d-1}(\mathbf{0})$.

The D-O-Q-R Paradigm

Depth, outlyingness, quantiles, and ranks in \mathbb{R}^d are all equivalent.

- ▶ $D(\mathbf{x}, F)$ and $O(\mathbf{x}, F)$ are equivalent (inversely).
- ▶ $\mathbf{Q}(\mathbf{u}, F)$ and $\mathbf{R}(\mathbf{x}, F)$ are equivalent (inversely).
- ▶ These are linked by
 - a) $O(\mathbf{x}, F) = \|\mathbf{R}(\mathbf{x}, F)\|$ ($= \|\mathbf{u}\|$),
 - b) $D(\mathbf{x}, F)$ induces a corresponding $\mathbf{Q}(\mathbf{u}, F)$.

Thus each of D , O , \mathbf{Q} , and \mathbf{R} can generate the others. We will think of these as interchangeable.

Examples of Univariate Outlyingness Functions

- ▶ “Tail probability” approach. Starting with $D(x, F) = \min\{F(x), 1 - F(x)\}$, we obtain

$$O(x, F) = \frac{1}{2} |2F(x) - 1|.$$

- ▶ “Sign function” approach. With $S(\cdot)$ the sign function,

$$O(x, F) = |ES(x - X)| = |2F(x) - 1|.$$

- ▶ “Scaled deviation” approach. With $\mu(F)$ and $\sigma(F)$ given location and spread measures,

$$O(x, F) = \left| \frac{x - \mu(F)}{\sigma(F)} \right|.$$

One Way to Extend to \mathbb{R}^d : Projection Pursuit

- ▶ Given a *univariate* outlyingness function $O_1(\cdot, \cdot)$ and $\mathbf{X} \sim F$ on \mathbb{R}^d , define a multivariate extension by projection pursuit:

$$O_d(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1} O_1(\mathbf{u}'\mathbf{x}, F_{\mathbf{u}'\mathbf{X}}), \quad \mathbf{x} \in \mathbb{R}^d.$$

- ▶ For both the *tail-probability* and *sign* $O_1(\cdot, \cdot)$, this leads to the halfspace (Tukey) depth and outlyingness.
- ▶ For the *scaled deviation* $O_1(\cdot, \cdot)$ based on the *Median* and *MAD* [Mosteller and Tukey, 1977], this gives the projection outlyingness function [Donoho and Gasko, 1992, Zuo, 2003].

Another Way to Extend to \mathbb{R}^d : Substitution

- ▶ In the *univariate sign* $O_1(\cdot, \cdot)$, substitute the d -dimensional sign function $\mathbf{S}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$, $\mathbf{x} \in \mathbb{R}^d$, and obtain the spatial outlyingness function,

$$O_d(\mathbf{x}, F) = \|\mathbf{E}\mathbf{S}(\mathbf{x} - \mathbf{X})\|, \quad \mathbf{x} \in \mathbb{R}^d,$$

with $\|\cdot\|$ the Euclidean norm.

- ▶ In the *univariate scaled deviation* $O_1(\cdot, \cdot)$, substitute multivariate location and spread measures and obtain the usual Mahalanobis distance outlyingness function,

$$O_d(\mathbf{x}, F) = \|\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\|.$$

“Mahalanobis Distance” Outlyingness Function

- ▶ The widely used *Mahalanobis outlyingness function* is

$$\tilde{O}(\mathbf{x}, F) = (\mathbf{x} - m(F))' S(F)^{-1} (\mathbf{x} - m(F)),$$

defined for any choice of location measure $m(F)$ and matrix-valued dispersion measure $S(F)$.

- ▶ The outlyingness contours are necessarily *ellipsoidal*, regardless of the shape of F .
- ▶ We use the equivalent form $O = \tilde{O}/(1 + \tilde{O})$ taking values in $(0, 1)$.
- ▶ To distinguish from another version introduced next, we call this the “Mahalanobis distance” outlyingness.

"Mahalanobis Quantile" Outlyingness Function

- ▶ By "Mahalanobis quantile" outlyingness we mean the outlyingness function $O(\mathbf{x}, F)$ that corresponds via the D-O-Q-R paradigm to the *Mahalanobis quantile function* $Q_M(\mathbf{u}, F)$, which we now define.
- ▶ First, let us recall the *spatial* quantile function: $Q_S(\mathbf{u}, F)$ is the point $\boldsymbol{\theta}$ in \mathbb{R}^d minimizing $E\{\Phi(\mathbf{u}, \mathbf{X} - \boldsymbol{\theta})\}$, where $\Phi(\mathbf{u}, \mathbf{t}) = \|\mathbf{t}\| + \mathbf{u}'\mathbf{t}$ [Chaudhuri, 1996].
- ▶ The corresponding spatial outlyingness function is $O_S(\mathbf{x}, F) = \|E\mathbf{S}(\mathbf{x} - \mathbf{X})\|$, $\mathbf{x} \in \mathbb{R}^d$, noted earlier.

"Mahalanobis Quantile" Outlyingness Function

- ▶ Next, we define weak covariance functional: any matrix-valued functional $C(F)$ satisfying

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

with $k(F_{\mathbf{X}}, A)$ a positive scalar function of $F_{\mathbf{X}}$ and A .

- ▶ The case $k(F_{\mathbf{X}}, A) = 1$ gives the usual definition of covariance functional.
- ▶ FACT: For any weak covariance functional with A nonsingular, the matrix

$$k(F_{\mathbf{X}}, A)^{-1/2} C(F_{\mathbf{X}})^{-1/2} A^{-1} C(F_{A\mathbf{X}+\mathbf{b}})^{1/2}$$

is *orthogonal*.

"Mahalanobis Quantile" Outlyingness Function

- ▶ For a given weak covariance functional $C(\cdot)$, we define the corresponding Mahalanobis quantile function at \mathbf{u} , $\mathbf{Q}_M(\mathbf{u}, F)$, as $\boldsymbol{\theta}$ minimizing $E\{\Phi(\mathbf{u}, C(F_{\mathbf{X}})^{-1/2}(\mathbf{X} - \boldsymbol{\theta}))\}$, analogous to the definition of spatial quantile but using a standardized deviation.
- ▶ This quantile function is *fully affine equivariant*: for $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ with A nonsingular,

$$\mathbf{Q}_M(\tilde{A}\mathbf{u}, F_{\mathbf{Y}}) = A\mathbf{Q}_M(\mathbf{u}, F_{\mathbf{X}}) + \mathbf{b},$$

with $\tilde{A} = (A C(F_{\mathbf{X}}) A')^{1/2} (A')^{-1} C(F_{\mathbf{X}})^{-1/2}$, which (by the "FACT") is *orthogonal*.

"Mahalanobis Quantile" Outlyingness Function

- ▶ A "transformation-retransformation" (TR) representation connects the Mahalanobis and spatial quantile functions:

$$\mathbf{Q}_M(\mathbf{u}, F_{\mathbf{X}}) = C(F_{\mathbf{X}})^{1/2} \mathbf{Q}_S(\mathbf{u}, F_{C(F_{\mathbf{X}})^{-1/2} \mathbf{X}}).$$

- ▶ Advocates of $\mathbf{Q}_S(\mathbf{u}, F)$ reject the idea of using $\mathbf{Q}_M(\mathbf{u}, F)$ on the grounds that it lacks a geometric interpretation.
- ▶ Advocates of $\mathbf{Q}_S(\mathbf{u}, F)$ compensate for its lack of full affine equivariance by using a *TR sample version*.
- ▶ If, however, we ask what the TR sample version actually estimates, the answer is:

the corresponding Mahalanobis quantile function!

Another "Mahalanobis" Outlyingness Function

- ▶ Let us call the outlyingness function corresponding to $\mathbf{Q}_M(\mathbf{u}, F)$ the "Mahalanobis quantile" outlyingness.
- ▶ By the TR representation, it is given by

$$O_M(\mathbf{x}, F_{\mathbf{X}}) = O_S(C(F_{\mathbf{X}})^{-1/2}\mathbf{x}, F_{C(F_{\mathbf{X}})^{-1/2}\mathbf{X}})$$

and is *affine invariant*.

- ▶ The Mahalanobis quantile outlyingness is *not* constrained to have ellipsoidal contours.

Two Versions of Finite Sample Breakdown Points

- ▶ *Addition breakdown* of an estimator $T(\mathbb{X}_N)$ defined on a data set \mathbb{X}_N occurs with k contaminants \mathbb{Y}_k added to \mathbb{X}_N if the estimator based on the combined sample can be taken to ∞ for suitable choice of \mathbb{Y}_k .
- ▶ *Replacement breakdown* of $T(\mathbb{X}_N)$ occurs with k replacements of values of \mathbb{X}_N by contaminants \mathbb{Y}_k if the estimator based on the modified sample can be taken to ∞ for suitable choice of \mathbb{Y}_k .
- ▶ The *minimal fraction* of contaminants in the sample needed for breakdown, either $k/(N+k)$ if by addition or k/N if by replacement, is called the *breakdown point*.

Are the ABP and RBP Equal?

Based on relationships between the ABP and RBP in a wide class of situations (Zuo, 2001), and on inequalities relating the two versions in general (Serfling, 2008), we can assert:

Indeed the ABP and RBP are equivalent, in the senses that

- (i) each corresponds to the other, through explicit expressions or by inequalities, although their values can be slightly different, and*
- (ii) their asymptotic limits, whether deterministic or almost sure, agree exactly.*

ABP and RBP: Equivalent Robustness Criteria

- ▶ Therefore, *as measures of robustness of estimators, the ABP and RBP perform equivalently for practical purposes, giving the same value with negligible difference.*
- ▶ This grants a pardon to authors who have perhaps inadvertently committed the crime of comparing one estimator's ABP with another's RBP.

Nonparametric Outlier Identification in \mathbb{R}^d

- ▶ Choose $O(\mathbf{x}, F)$ taking values in $[0, 1]$ (w. l. o. g.).
- ▶ The goal: for specified $\lambda \in (0, 1)$, identify the region

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

associated with F via $O(\mathbf{x}, F)$.

- ▶ The method: Estimate $out(\lambda, F)$ by the sample analogue

$$OR(\mathbb{X}_N, \lambda) = \{\mathbf{x} : O(\mathbf{x}, \hat{F}_N) > \lambda\},$$

based on a data set \mathbb{X}_N with sample df \hat{F}_N .

- ▶ The main issue: $O(\mathbf{x}, \hat{F}_N)$ must itself be *robust*.

Robustness Against Masking is Desired

- ▶ Masking occurs if λ outliers of F are misidentified by the sample as λ *nonoutliers*.
- ▶ Masking breakdown is based on the size of the most extreme outlier of F which is misidentifiable by the sample as a λ nonoutlier.
- ▶ Masking breakdown occurs if contaminants in the sample cause points of *arbitrarily extreme* $O(\cdot, F)$ -outlyingness to be classified as sample λ *nonoutliers*.

Masking Breakdown with k Contaminants

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

$$\overline{OR(\mathbb{X}_{N,k}, \lambda)} \cap out(\gamma, F) \neq \emptyset \quad (1)$$

holds}

- ▶ Note that (1) holds if and only if *some γ outliers of F are included among sample λ nonoutliers.*
- ▶ The worst case, $\gamma_M(\lambda, \mathbb{X}_N, k) = 1$, represents “Masking Breakdown” due to k replacements: some points with arbitrarily large $O(\cdot, F)$ fail to be identified as outliers by $OR(\mathbb{X}_{N,k}, \lambda)$.

Masking Breakdown Point

- ▶ A useful robustness criterion is the *minimal fraction* of sample contaminants necessary for masking breakdown.
- ▶ This masking breakdown point (MBP) is given by

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

where $k_M(\lambda, \mathbb{X}_N) = \min\{k : \gamma_M(\lambda, \mathbb{X}_N, k) = 1\}$.

- ▶ This is the *replacement contamination* version of the *addition contamination* MBP of Davies and Gather (1993) and Becker and Gather (1999).

MBP with Mahalanobis Distance Outlyingness

- ▶ For the *Mahalanobis distance* outlyingness we have, *independently of choice of threshold λ* ,

$$\min\{\text{RBP}(m(\mathbb{X}_N)), \text{RBP}(S(\mathbb{X}_N))\} \leq \varepsilon_M(\lambda, \mathbb{X}_N) \leq \text{RBP}(m(\mathbb{X}_N)).$$

- ▶ The MBP depends on $\text{RBP}(m(\mathbb{X}_N))$ and $\text{RBP}(S(\mathbb{X}_N))$.
- ▶ With *high-breakdown estimators* for m and S , e.g., MVE, MCD, Stahel-Donoho, S-type, and others, we have the lower bound $\lfloor (N - d + 1)/2 \rfloor / N \sim 1/2$ for $\varepsilon_M(\lambda, \mathbb{X}_N)$.

MBP with Halfspace (Tukey) Outlyingness

- ▶ Take as *halfspace outlyingness* $O = 1 - 2\tilde{D}$, with $\tilde{D}(\mathbf{x}, F) = \inf\{P(H) : H \text{ a closed halfspace containing } \mathbf{x}\}$.
- ▶ With $m_H(\mathbb{X}_N)$ the *halfspace median*, we have

$$\varepsilon_M(\lambda, \mathbb{X}_N) = \min \left\{ \text{RBP}(m_H(\mathbb{X}_N)), N^{-1} \left\lceil \left(\frac{1-\lambda}{2} \right) N \right\rceil \right\},$$

which depends on both λ and $\text{RBP}(m_H(\mathbb{X}_N))$.

- ▶ The almost sure upper bound of $1/3$ for $\text{RBP}(m_H(\mathbb{X}_N))$ for symmetric F suggests for $\varepsilon_M(\lambda, \mathbb{X}_N)$ the practical value $\min\{(1-\lambda)/2, 1/3\}$.

MBP with TR Spatial Outlyingness

- ▶ The *spatial outlyingness* $O(\mathbf{x}, F) = \|\mathbf{E}\mathbf{S}(\mathbf{x} - \mathbf{X})\|$ has sample version $O(\mathbf{x}, \mathbb{X}_N) = \|\mathbf{N}^{-1} \sum_{i=1}^N \mathbf{S}(\mathbf{x} - \mathbf{X}_i)\|$.
- ▶ In practice, however, as discussed earlier, a fully affine invariant *transformation-retransformation (TR)* sample version is used.
- ▶ This is defined by transforming the data \mathbb{X}_N to a new coordinate system using any (weak) covariance functional $C(\cdot)$, i.e., $\mathbf{x} \mapsto C(\mathbb{X}_N)^{-1/2}\mathbf{x} = \mathbf{w}$, say, and taking

$$O^{(\text{TR})}(\mathbf{x}, \mathbb{X}_N) = O(\mathbf{w}, C(\mathbb{X}_N)^{-1/2}\mathbb{X}_N).$$

MBP with TR Spatial Outlyingness

- ▶ Of course, $O^{(\text{TR})}(\mathbf{x}, \mathbb{X}_N)$ estimates not $O(\mathbf{x}, F)$ but rather the *Mahalanobis quantile* outlyingness function based on the chosen weak covariance functional $C(\cdot)$.
- ▶ The MBPs of the TR and non-TR spatial identifiers are equal, with

$$\varepsilon_M(\lambda, \mathbb{X}_N) = \min \left\{ \text{RBP}(C(\mathbb{X}_N)), N^{-1} \left[\left(\frac{1-\lambda}{2} \right) N \right] \right\},$$

which depends on both λ and $\text{RBP}(C(\mathbb{X}_N))$.

- ▶ We can choose high-breakdown $C(\cdot)$, with $\text{RBP}(C(\mathbb{X}_N)) = N^{-1} \left[\left(\frac{N-d+1}{2} \right) N \right] \sim 1/2$.

MBP with Projection Outlyingness

- ▶ The *projection* outlyingness function is $O = \tilde{O}/(1 + \tilde{O})$ with

$$\tilde{O}(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1} \left| \frac{\mathbf{u}'\mathbf{x} - \mu(F_{\mathbf{u}'\mathbf{x}})}{\sigma(F_{\mathbf{u}'\mathbf{x}})} \right|,$$

for univariate location and scale measures $\mu(\cdot)$ and $\sigma(\cdot)$.

- ▶ With $(\mu, \sigma) = (\text{Med}, \text{MAD}_{d-1})$ and MAD_m the modified version of MAD [Tyler, 1994, Gather and Hilker, 1997, Zuo, 2003], for $N \geq 2(d-1)^2 + d$ we have, *independently of choice of threshold* λ ,

$$\varepsilon_M(\lambda, \mathbb{X}_N) = N^{-1} \left[\left(\frac{N-d+2}{2} \right) N \right] \sim 1/2,$$

the same as the RBP of the projection median.

Choosing the “Outlier” Threshold λ

- ▶ Consider now a contamination model for the data \mathbb{X}_N ,

$$F = (1 - \varepsilon)G + \varepsilon H$$

with G a known “ideal” distribution and ε the probability of a gross contaminant from an unknown distribution H .

- ▶ For given choice of λ , the goal now is to “identify” the λ outliers of F in the data \mathbb{X}_N on the basis of the criterion $OR(\mathbb{X}_N, \lambda)$, while maintaining a low false positive rate with respect to G .
- ▶ *What choice of threshold λ makes sense?*

Choosing the “Outlier” Threshold λ

- ▶ We want the threshold λ *high enough* to yield a *small false positive rate*

$$p = P_G(O(\mathbf{X}, G) > \lambda)$$

and *low enough* for true contaminants to be identified with high probability, i.e.,

$$P_H(O(\mathbf{X}, H) > \lambda) \approx 1.$$

Choosing the “Outlier” Threshold λ

- ▶ For specified *false positive rate* p , we obtain λ by solving the equation $p = P_G(O(\mathbf{X}, G) > \lambda)$, yielding

$$\lambda = F_{O(\mathbf{X}, G)}^{-1}(1 - p), \quad (2)$$

based on the quantile function of the distribution of $O(\mathbf{X}, G)$ under the ideal distribution G .

- ▶ As a benchmark, we take $G = \underline{\text{multivariate normal}}$.
- ▶ By affine invariance of the outlyingness functions we are considering, we take G to be *standard* d -variate normal, $N(\mathbf{0}, \mathbf{I}_d)$.

Choosing the “Outlier” Threshold λ

- ▶ For *Mahalanobis distance* outlyingness (values in $(0, 1)$),

$$F_{O(\mathbf{x}, G)}(\lambda) = P \left(\chi_d^2 \leq \left(\frac{\lambda}{1 - \lambda} \right)^2 \right).$$

- ▶ For the *halfspace* outlyingness,

$$F_{O(\mathbf{x}, G)}(\lambda) = P \left(\chi_d^2 \leq \left[\Phi^{-1} \left(\frac{1 + \lambda}{2} \right) \right]^2 \right).$$

- ▶ For the *projection* outlyingness,

$$F_{O(\mathbf{x}, G)}(\lambda) = P \left(\chi_d^2 \leq \left[\Phi^{-1} \left(\frac{3}{4} \right) \frac{\lambda}{1 - \lambda} \right]^2 \right).$$

Choosing the “Outlier” Threshold λ

With $Q(d, p) = \sqrt{(\chi_d^2)^{-1} (1 - p)}$, the formula (2) for λ becomes

- ▶ For *Mahalanobis distance* outlyingness,

$$Q(d, p)/(1 + Q(d, p)).$$

- ▶ For *halfspace* outlyingness,

$$2\Phi(Q(d, p)) - 1.$$

- ▶ For *projection* outlyingness,

$$Q(d, p)/(\Phi^{-1}(\frac{3}{4}) + Q(d, p)).$$

Choosing the “Outlier” Threshold λ

- ▶ It has been argued (e.g., Jaeckel, 1971) that ε should decrease with increasing N , e.g., $\varepsilon_N = c/\sqrt{N}$.
- ▶ Desirably, the false positive rate p is small relative to the “true positive rate” ε_N , say $p_N = \delta \varepsilon_N = \delta c/\sqrt{N}$, with δ small, perhaps $\delta = 0.1$.
- ▶ We might choose c by deciding what expected number $N\varepsilon_N = c\sqrt{N}$ of contaminants to protect against in a sample of size $N = 100$, say. For example, for *moderate to strong protection*, set $c\sqrt{100} = 15$, obtaining $c = 1.5$.

Choosing the “Outlier” Threshold λ

For $c = 1.5$ and $\delta = 0.1$, i.e., $p_N = \delta c / \sqrt{N} = 0.15 / \sqrt{N}$,
 the solutions λ_N range tightly for $N = \mathbf{100, 500, \text{ and } 1000}$,
 and for dimension $d = \mathbf{2, 5, 10, 15, \text{ and } 20}$, as follows:

- ▶ For *Mahalanobis distance* outlyingness:

$$\mathbf{0.74} \leq \lambda_N \leq \mathbf{0.86}$$

- ▶ For *halfspace* outlyingness:

$$\mathbf{0.996} \leq \lambda_N \leq \mathbf{1.00}$$

(implies very small MBP, $(1 - \lambda_N)/2$)

- ▶ For *projection* outlyingness: $\mathbf{0.81} \leq \lambda_N \leq \mathbf{0.90}$

Choosing the “Outlier” Threshold λ

- ▶ Also, we may think of ε_N as the level of protection desired in terms of the MBP robustness criterion, thus requiring

$$\varepsilon_M(\lambda_N, \mathbb{X}_N) \geq \varepsilon_N. \quad (3)$$

- ▶ For the *Mahalanobis distance* and *projection* identifiers, the MBPs do not depend upon the threshold λ_N , and thus (3) imposes no restriction on the choice of λ_N .

Choosing the “Outlier” Threshold λ

- ▶ For the *Mahalanobis quantile* identifier, however, (3) imposes

$$\lambda_N \leq 1 - 2\varepsilon_N, \quad (4)$$

and for the *halfspace* identifier (3) imposes

$$\max\{\lambda_N, 1/3\} \leq 1 - 2\varepsilon_N \quad (5)$$

(which requires $\varepsilon_N \leq 1/3$).

- ▶ With $\varepsilon_N = 1.5/\sqrt{N}$ ($\leq 1/3$ if $N \geq 21$), the upper bound for λ_N in (4) and (5) takes values **0.4**, **0.7**, and **0.9** (implies very large false positive rate, p_N), for $N = 25$, **100**, and **1000**, respectively.

Comparisons and Conclusions

- ▶ The *Mahalanobis distance* and *projection* outlyingness allow λ_N thresholds with both *high MBP* and *low false positive rate (FPR)*. The *Mahalanobis quantile* outlyingness entails a trade off between MBP and FPR.
- ▶ The *halfspace* outlyingness, however, imposes a severe and unacceptable trade off between MBP and FPR.
- ▶ For classifying points as “outliers” using a *threshold* λ_N , the *Mahalanobis distance*, *projection*, and *Mahalanobis quantile* identifiers are distinctly superior to the *halfspace* identifier for masking protection.
- ▶ On the other hand, all of these outlyingness functions can be used for robust outlyingness *ranking* of points in \mathbb{X}_N .

Explanation: a Bit of a Hint

- ▶ The *Mahalanobis distance* and *projection* approaches succeed perhaps by requiring only *robust estimation of location and scale parameters.*
- ▶ The *Mahalanobis quantile* and *halfspace* approaches entail a wider and more challenging scope, *pointwise robust estimation of the outlyingness function.*

Next: Empirical Studies and Further Criteria

- ▶ Empirical studies to illustrate and to explore these differences are in progress.
- ▶ It is also of interest to compare identifiers with respect to the further criterion of *Swamping Breakdown Point*.

Acknowledgments

The work on masking breakdown points is joint with Xin Dang, University of Mississippi (paper in review).

Support by NSF grants DMS-0103698 and DMS-0805786 is greatly appreciated.