

# Computationally Easy Outlier Detection via Projection Pursuit with Finitely Many Directions

Satyaki Mazumder<sup>1</sup> and Robert Serfling<sup>2</sup>

*Indian Institute of Science, Education and Research*  
and  
*University of Texas at Dallas*

December, 2011

<sup>1</sup>Indian Institute of Science, Education and Research, Kolkata, India

<sup>2</sup>Department of Mathematics, University of Texas at Dallas, Richardson, Texas 75080-3021, USA. Email: [serfling@utdallas.edu](mailto:serfling@utdallas.edu). Website: [www.utdallas.edu/~serfling](http://www.utdallas.edu/~serfling). Support by NSF Grants DMS-0805786 and DMS-1106691 and NSA Grant H98230-08-1-0106 is gratefully acknowledged.

## Abstract

Outlier detection methods are fundamental to all of data analysis. They are desirably robust, affine invariant, and computationally easy in any dimension. The powerful projection pursuit approach yields the “projection outlyingness”, which is affine invariant and highly robust and does not impose ellipsoidal contours like the Mahalanobis distance approach. However, it is highly computationally intensive, being obtained by taking suprema of univariate scaled deviation outlyingness over all projections of the data onto lines. Here we introduce several outlyingness functions based on a vector of scaled deviations taken over only finitely many directions approximately uniform over the unit hypersphere. A preliminary transformation of the data to a strong invariant coordinate system makes such vectors affine invariant. We establish useful foundational theory for finite vectors of scaled deviations on projections. Also, using artificial and real data sets, we compare our affine invariant outlyingness functions with the usual projection outlyingness and with robust Mahalanobis distance outlyingness.

*AMS 2000 Subject Classification:* Primary 62H99. Secondary 62G99

*Key words and phrases:* Outlier detection; Projection pursuit; Nonparametric; Multivariate; Affine invariance; Robustness

# 1 Introduction

Outlier identification is fundamental to multivariate statistical analysis and data mining. Typically, based on a selected outlyingness function defined on the sample or input space, “outliers” are those points whose outlyingness value exceeds a specified threshold. An outlyingness function should have the following basic properties: (i) robustness against the presence of outliers, (ii) weak affine invariance (i.e., transformation to other coordinates should not affect relative “outlyingness” rankings and comparisons), and (iii) computational efficiency in any practical dimension. It is also desirable to avoid imposing elliptical contours when not justified. Here we develop outlyingness functions based on projection pursuit that are very favorable with respect to (i), (ii), and (iii).

Projection pursuit plays a leading role in multivariate data analysis. Such techniques were originally proposed and experimented with by Kruskal (1969, 1972). Related ideas occur in Switzer (1970) and Switzer and Wright (1971). A key implementation is due to Friedman and Tukey (1974). Recently, “projection depth” has received significant attention in the literature (Liu, 1992; Zuo and Serfling, 2000b; Zuo, 2003). The corresponding *projection pursuit outlyingness function* extends the univariate scaled deviation type outlyingness function of the form  $O(x) = (|x - \text{median}|)/\text{MAD}$ , where MAD is the median absolute deviation from the median, to a multivariate outlyingness function through projection pursuit. The supremum of the projected outlyingness of a data point over all projections defines the “projection outlyingness” (Liu, 1992; Zuo and Serfling, 2000b; Zuo, 2003; Serfling, 2004; Dang and Serfling, 2010). However, this outlyingness function is computationally intensive.

Dang and Serfling (2010) compared four affine invariant (or weakly affine invariant) outlyingness functions, the *halfspace*, the *projection*, the *Mahalanobis distance*, and the *Mahalanobis spatial*, employing masking breakdown point as robustness criterion. The projection and Mahalanobis distance versions performed best. The Mahalanobis spatial version was moderately competitive, and the halfspace version became eliminated. However, the projection pursuit outlyingness is computationally intensive, the Mahalanobis distance outlyingness imposes elliptical contours, and the Mahalanobis spatial outlyingness trades off masking breakdown point against false positive rate.

Mazumder and Serfling (2012) developed a more robust version of Mahalanobis spatial outlyingness using a “spatial trimming” method. Here we apply spatial trimming and related methods in a different way, developing computationally easy projection pursuit type outlyingness functions based on *only finitely many projections*.

Let us make this precise. Let  $\mathbf{X}$  have distribution  $F_{\mathbf{X}}$  on  $\mathbb{R}^d$  and, for any unit vector  $\mathbf{u} = (u_1, \dots, u_d)'$  in  $\mathbb{R}^d$ , let  $F_{\mathbf{u}'\mathbf{X}}$  denote the induced univariate distribution of  $\mathbf{u}'\mathbf{X}$ . With  $\mu(\cdot)$  and  $\sigma(\cdot)$  any univariate location and scale measures, which we assume are *equivariant* in the usual sense, and with the notation

$$g(\mathbf{x}, \mathbf{u}, F_{\mathbf{X}}) = \frac{\mathbf{u}'\mathbf{x} - \mu(F_{\mathbf{u}'\mathbf{X}})}{\sigma(F_{\mathbf{u}'\mathbf{X}})},$$

the associated well-known *projection outlyingness* (which we denote “**SUP**”) is

$$O_P(\mathbf{x}, F_{\mathbf{X}}) = \sup_{\|\mathbf{u}\|=1} |g(\mathbf{x}, \mathbf{u}, F_{\mathbf{X}})|, \quad \mathbf{x} \in \mathbb{R}^d. \quad (1)$$

This represents the worst case scaled deviation outlyingness of projections of  $\mathbf{x}$  onto lines. With  $\mu$  and  $\sigma$  given by Median and MAD,  $O_P(\mathbf{x}, F_{\mathbf{X}})$  is affine invariant and its sample version is highly robust (Dang and Serfling, 2010). However, it is highly computational.

To overcome the computational burden, we take only  $s$  projections, for some choice of  $s \geq d$ . In particular, we take a set  $\Delta = \{\mathbf{u}_1, \dots, \mathbf{u}_s\}$  of  $s$  unit vectors approximately uniformly distributed on the unit sphere in  $\mathbb{R}^d$  but lying on distinct diameters, using algorithms of Fang and Wang (1994). Associated with  $\Delta$  we define the vector function

$$\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, F_{\mathbf{X}}) = (g(\mathbf{x}, \mathbf{u}_1, F_{\mathbf{X}}), \dots, g(\mathbf{x}, \mathbf{u}_s, F_{\mathbf{X}}))', \quad \mathbf{x} \in \mathbb{R}^d,$$

whose components give (signed) scaled deviation outlyingness values for the projections of a point  $\mathbf{x}$  onto the lines represented by  $\Delta$ . For a  $d$ -dimensional data set  $\mathbb{X}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ , the sample version is denoted by  $\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$ .

In Section 2, under minimal regularity conditions, we establish the asymptotic normality of  $\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$ , a result of general interest. Pan, Fung, and Fang (2000) also study  $\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$  and construct a quadratic form outlyingness function based on  $\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n) - \tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, F_{\mathbf{X}})$ . Although it has a convenient asymptotic chi-square distribution as  $n \rightarrow \infty$  with  $s$  fixed, it is not affine invariant and requires a bootstrap step to estimate  $\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, F_{\mathbf{X}})$ . We will use the vectors  $\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$  in a different way and obtain affine invariant outlyingness functions computable without bootstrap steps.

Section 3 addresses the fact that, while  $O_P(\mathbf{x}, F)$  is affine invariant as is easily checked, a *single* scaled deviation  $g(\mathbf{x}, \mathbf{u}, F_{\mathbf{X}})$  is *not*, and hence neither is the vector  $\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, F_{\mathbf{X}})$ . It is not even orthogonally invariant. However, if we first standardize  $\mathbf{X}$  using a *strong invariant coordinate system (SICS) functional*  $\mathbf{D}(F_{\mathbf{X}})$  (Serfling, 2010; Ilmonen, Oja, and Serfling, 2011), then weak affine invariance of  $\tilde{\boldsymbol{\eta}}(\mathbf{D}(F_{\mathbf{X}})\mathbf{x}, \Delta, F_{\mathbf{D}(F_{\mathbf{X}})\mathbf{X}})$  holds. Likewise,  $\tilde{\boldsymbol{\eta}}(\mathbf{D}(\mathbb{X}_n)\mathbf{x}, \Delta, \mathbf{D}(\mathbb{X}_n)\mathbb{X}_n)$  is weakly affine invariant, with  $\mathbf{D}(\mathbb{X}_n)$  a sample SICS functional. Following a method of Serfling (2010) for construction of sample SICS functionals, and using an *inner set* of observations indexed by  $\mathbb{J}$  as obtained by spatial trimming of observations (Mazumder and Serfling, 2012), we introduce an easily computable and robust SICS functional to be used for the purposes of this paper.

The number  $s$  of projections is chosen large enough to capture sufficient information, for example  $s = 4d$ . Then, via robust principal component analysis, we eliminate the singularity (redundancy) in the  $s$ -vector  $\tilde{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$  by transforming to a  $t$ -vector  $\mathbf{V}(\mathbf{x}, \Delta, \mathbb{X}_n)$  for some  $t < s$ . This is carried out in Section 4, along with proof that the transformation is affine invariant. Based on the  $t$ -vector  $\mathbf{V}(\mathbf{x}, \Delta, \mathbb{X}_n)$ , in Section 5 we construct a *Mahalanobis distance outlyingness function on projected scaled deviations*, denoted by “**MDP**”. Under a *normality* assumption on the parent population  $F_{\mathbf{X}}$ , the population MDP has a chi-square distribution. Likewise, in Section 6, we apply a spatial trimming technique on the reduced

vector  $\mathbf{V}(\mathbf{x}, \Delta, \mathbb{X}_n)$  to construct a *robust transformation-retransformation spatial outlyingness function on projected scaled deviations*, denoted by “**RTRP**”. Both MDP and RTRP are affine invariant and easily computable.

In Section 7, we compare MDP, RTRP and SUP using artificial bivariate data sets for visual comparisons and using two higher-dimensional actual data sets, Stackloss Data ( $n = 21$ ,  $d = 4$ ), and Air Pollution and Mortality Data ( $n = 59$ ,  $d = 13$ ), which are studied extensively in the literature. It is seen that SUP is outclassed by MDP and RTRP. Comparison with a robust version of Mahalanobis distance outlyingness (**MD**) is also made. Regarding the most extreme outliers, all methods agree, while differing on the intermediate structure and thus the identification of the moderate levels of outlyingness. Overall, RTRP is especially recommended as requiring less computational burden while remaining competitive in terms of robustness and while not imposing ellipsoidal contours.

Besides the work of Pan, Fung, and Fang (2000), some other authors have developed approaches using only finitely many projections, in some cases data-driven choices, notably Peña and Prieto (2001) and Filzmozer, Maronna, and Werner (2008), for example. However, these either require elliptical contours or give up affine invariance. See Maronna *et al.* (2006) and Serfling (2010) for some discussion.

## 2 Asymptotic Properties of Vectors of Projections

As general foundation, we establish key asymptotic results for  $\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$ , where

$$\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n) = (\widehat{\eta}(\mathbf{x}, \mathbf{u}_1, \mathbb{X}_n), \dots, \widehat{\eta}(\mathbf{x}, \mathbf{u}_s, \mathbb{X}_n))',$$

with

$$\widehat{\eta}(\mathbf{x}, \mathbf{u}_i, \mathbb{X}_n) = \frac{\mathbf{u}_i' \mathbf{x} - \widehat{\nu}(\mathbf{u}_i' \mathbb{X}_n)}{\widehat{\zeta}(\mathbf{u}_i' \mathbb{X}_n)}, \quad 1 \leq i \leq s,$$

where  $\widehat{\nu}$  and  $\widehat{\zeta}$  denote any of the various sample versions of the *median* and *MAD* functionals  $\nu$  and  $\zeta$ , respectively. For notational convenience, denote  $\nu(F_{\mathbf{u}_j' \mathbf{X}})$  and  $\zeta(F_{\mathbf{u}_j' \mathbf{X}})$  by  $\nu_j$  and  $\zeta_j$ , respectively, and  $\widehat{\nu}(\mathbf{u}_j' \mathbb{X}_n)$  and  $\widehat{\zeta}(\mathbf{u}_j' \mathbb{X}_n)$  by  $\widehat{\nu}_{jn}$  and  $\widehat{\zeta}_{jn}$ , respectively. Further, Let  $F_j$  denote the distribution function of  $\mathbf{u}_j' \mathbf{X}$ , for  $1 \leq j \leq s$ . We adopt the following assumption.

- (A)  $F_j$  is continuous in neighborhoods of  $\nu_j \pm \zeta_j$  and differentiable at  $\nu_j$  and  $\nu_j \pm \zeta_j$ , with  $F'(\nu_j) > 0$  and  $G'(\zeta_j) = F'(\nu_j - \zeta_j) + F'(\nu_j + \zeta_j) > 0$ , for  $1 \leq j \leq s$ .

Now, with

$$A_j = F_j'(\nu_j + \zeta_j) + F_j'(\nu_j - \zeta_j), C_j = F_j'(\nu_j - \zeta_j) - F_j'(\nu_j + \zeta_j), \quad 1 \leq j \leq s,$$

we have the following lemma.

**Lemma 1** *Under Assumption (A) we have*

$$\sqrt{n} \left( (\widehat{\nu}_{1n} - \nu_1), (\widehat{\zeta}_{1n} - \zeta_1), \dots, (\widehat{\nu}_{sn} - \nu_s), (\widehat{\zeta}_{sn} - \zeta_s) \right)' \xrightarrow{d} N_{2s}(\mathbf{0}, \boldsymbol{\Sigma}_{2s}), \quad (2)$$

where  $\Sigma_{2s} = (\Sigma_{ij}^{2 \times 2})_{s \times s}$  with  $\Sigma_{ii}^{2 \times 2} = \begin{pmatrix} \sigma_1^{\mathbf{u}_i} & \sigma_{12}^{\mathbf{u}_i} \\ \sigma_2^{\mathbf{u}_i} & \sigma_2^{\mathbf{u}_i} \end{pmatrix}$  and, for  $1 \leq i < j \leq s$ ,  $(\Sigma_{ji}^{2 \times 2})' = \Sigma_{ij}^{2 \times 2} = \begin{pmatrix} \sigma_{11}^{\mathbf{u}_i \mathbf{u}_j} & \sigma_{12}^{\mathbf{u}_i \mathbf{u}_j} \\ \sigma_{21}^{\mathbf{u}_i \mathbf{u}_j} & \sigma_{22}^{\mathbf{u}_i \mathbf{u}_j} \end{pmatrix}$ , where

$$\begin{aligned} \sigma_1^{\mathbf{u}_i} &= \frac{1}{(F_i'(\nu_i))^2} \text{Var}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}), \\ \sigma_{12}^{\mathbf{u}_i} &= \frac{1}{F_i'(\nu_i)A_i} \text{Cov}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}, I\{|\mathbf{u}_i' \mathbf{X} - \nu_i| \leq \zeta_i\}) \\ &\quad + \frac{C_i^2}{A_i^2(F_i'(\nu_i))^2} \text{Var}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}), \\ \sigma_2^{\mathbf{u}_i} &= \frac{1}{A_i^2} \text{Var}(I\{|\mathbf{u}_i' \mathbf{X} - \nu_i| \leq \zeta_i\}) + \frac{C_i^2}{A_i^2(F_i'(\nu_i))^2} \text{Var}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}) \\ &\quad + \frac{2C_i}{A_i^2 F_i'(\nu_i)} \text{Cov}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}, I\{|\mathbf{u}_i' \mathbf{X} - \nu_i| \leq \zeta_i\}), \\ \sigma_{11}^{\mathbf{u}_i \mathbf{u}_j} &= \frac{1}{F_i'(\nu_i)F_j'(\nu_j)} \text{Cov}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}, I\{\mathbf{u}_j' \mathbf{X} \leq \nu_j\}), \\ \sigma_{12}^{\mathbf{u}_i \mathbf{u}_j} &= \frac{1}{F_i'(\nu_i)A_j} \text{Cov}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}, I\{|\mathbf{u}_j' \mathbf{X} - \nu_j| \leq \zeta_j\}) \\ &\quad + \frac{C_j}{A_j F_i'(\nu_i)F_j'(\nu_j)} \text{Cov}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}, I\{\mathbf{u}_j' \mathbf{X} \leq \nu_j\}), \\ \sigma_{21}^{\mathbf{u}_i \mathbf{u}_j} &= \frac{1}{F_j'(\nu_j)A_i} \text{Cov}(I\{\mathbf{u}_j' \mathbf{X} \leq \nu_j\}, I\{|\mathbf{u}_i' \mathbf{X} - \nu_i| \leq \zeta_i\}) \\ &\quad + \frac{C_i}{A_i F_j'(\nu_j)F_i'(\nu_i)} \text{Cov}(I\{\mathbf{u}_j' \mathbf{X} \leq \nu_j\}, I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}), \\ \sigma_{22}^{\mathbf{u}_i \mathbf{u}_j} &= \frac{1}{A_i A_j} \text{Cov}(I\{|\mathbf{u}_i' \mathbf{X} - \nu_i| \leq \zeta_i\}, I\{|\mathbf{u}_j' \mathbf{X} - \nu_j| \leq \zeta_j\}) \\ &\quad + \frac{C_j}{A_i A_j F_j'(\nu_j)} \text{Cov}(I\{\mathbf{u}_j' \mathbf{X} \leq \nu_j\}, I\{|\mathbf{u}_i' \mathbf{X} - \nu_i| \leq \zeta_i\}) \\ &\quad + \frac{C_i}{A_j A_i F_i'(\nu_i)} \text{Cov}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}, I\{|\mathbf{u}_j' \mathbf{X} - \nu_j| \leq \zeta_j\}) \\ &\quad + \frac{C_i C_j}{A_j A_i F_i'(\nu_i)F_j'(\nu_j)} \text{Cov}(I\{\mathbf{u}_i' \mathbf{X} \leq \nu_i\}, I\{\mathbf{u}_j' \mathbf{X} \leq \nu_j\}). \end{aligned}$$

PROOF OF LEMMA 1. Under Assumption **(A)** and using the Ghosh (1971) weak Bahadur representation for the sample median and the Mazumder and Serfling (2009) weak Bahadur

representation for the sample MAD, we have, for  $1 \leq j \leq s$ ,

$$\begin{aligned}\widehat{\nu}_{jn} &= \nu_j + \frac{\frac{1}{2} - \widehat{F}_{jn}(\nu_j)}{F'_j(\nu_j)} + R_{jn}, \\ \widehat{\zeta}_{jn} &= \zeta_j + \frac{\frac{1}{2} - \widehat{G}_{jn}(\zeta_j)}{A_j} + \frac{\frac{1}{2} - \widehat{F}_{jn}(\nu_j)}{F'_j(\nu_j)} \frac{C_j}{A_j} + U_{jn},\end{aligned}$$

where  $\widehat{F}_{jn}(\nu_j) = \frac{1}{n} \sum_{i=1}^n I\{\mathbf{u}'_j \mathbf{X}_i \leq \nu_j\}$ ,  $\widehat{G}_{jn}(\zeta_j) = \frac{1}{n} \sum_{i=1}^n I\{|\mathbf{u}'_j \mathbf{X}_i - \nu_j| \leq \zeta_j\}$ , and the  $R_{jn}$ 's and  $U_{jn}$ 's are  $o_p(n^{-1/2})$ . To prove (2) we show that  $\sum_{j=1}^s \{\lambda_j \widehat{\nu}_{jn} + \gamma_j \widehat{\zeta}_{jn}\}$  is asymptotically normal (AN) for any scalars  $\lambda_1, \lambda_2, \dots, \lambda_s, \gamma_1, \dots, \gamma_s$ .

$$\begin{aligned}\sqrt{n} \sum_{j=1}^s \left\{ \lambda_j (\widehat{\nu}_{jn} - \nu_j) + \gamma_j (\widehat{\zeta}_{jn} - \zeta_j) \right\} \\ = -\sqrt{n} \sum_{j=1}^s \left\{ \left( \left( \lambda_j + \frac{C_j \gamma_j}{A_j} \right) \frac{\widehat{F}_{jn}(\nu_j)}{F'_j(\nu_j)} + \frac{\gamma_j}{A_j} \widehat{G}_{jn}(\zeta_j) \right) - \frac{1}{2} \left( \left( \lambda_j + \frac{C_j \gamma_j}{A_j} \right) \frac{1}{F'_j(\nu_j)} + \frac{\gamma_j}{A_j} \right) \right\} \\ + \sqrt{n} \sum_{j=1}^s \{ \lambda_j R_{jn} + \gamma_j U_{jn} \} \\ = -\sqrt{n} \sum_{j=1}^s \left\{ (\alpha_j \widehat{F}_{jn}(\nu_j) + \beta_j \widehat{G}_{jn}(\zeta_j)) - \frac{1}{2} (\alpha_j + \beta_j) \right\} + o_p(1),\end{aligned}$$

where  $\alpha_j = (\lambda_j + \frac{C_j \gamma_j}{A_j}) \frac{1}{F'_j(\nu_j)}$ ,  $\beta_j = \frac{\gamma_j}{A_j}$ ,  $1 \leq j \leq s$ , and  $\sqrt{n} \sum_{j=1}^s \{ \lambda_j R_{jn} + \gamma_j U_{jn} \} = o_p(1)$ . Now

$$\begin{aligned}\sqrt{n} \sum_{j=1}^s \left\{ \alpha_j \widehat{F}_{jn}(\nu_j) + \beta_j \widehat{G}_{jn}(\zeta_j) - \frac{1}{2} (\alpha_j + \beta_j) \right\} \\ = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n W_i - \frac{\sum_{j=1}^s \{ \alpha_j + \beta_j \}}{2} \right),\end{aligned}$$

where  $W_i = \sum_{j=1}^s \{ \alpha_j I\{\mathbf{u}'_j \mathbf{X}_i \leq \nu_j\} + \beta_j I\{|\mathbf{u}'_j \mathbf{X}_i - \nu_j| \leq \zeta_j\} \}$  are i.i.d. for  $1 \leq i \leq n$ , with mean  $E(W_i) = \frac{\sum_{j=1}^s \{ \alpha_j + \beta_j \}}{2}$  and variance

$$\begin{aligned}\sigma_W^2 &= \sum_{1 \leq k, \ell \leq s} \sum_{1 \leq k, \ell \leq s} \alpha_\ell \alpha_k \text{Cov}(I\{\mathbf{u}'_\ell \mathbf{X} \leq \nu_\ell\}, I\{\mathbf{u}'_k \mathbf{X} \leq \nu_k\}) \\ &\quad + \sum_{1 \leq k, \ell \leq s} \sum_{1 \leq k, \ell \leq s} \alpha_\ell \beta_k \text{Cov}(I\{\mathbf{u}'_\ell \mathbf{X} \leq \nu_\ell\}, I\{|\mathbf{u}'_k \mathbf{X} - \nu_k| \leq \zeta_k\}) \\ &\quad + \sum_{1 \leq k, \ell \leq s} \sum_{1 \leq k, \ell \leq s} \beta_\ell \beta_k \text{Cov}(I\{|\mathbf{u}'_\ell \mathbf{X} - \nu_\ell| \leq \zeta_\ell\}, I\{|\mathbf{u}'_k \mathbf{X} - \nu_k| \leq \zeta_k\}) < \infty.\end{aligned}$$

Hence, by the classical Central Limit Theorem,

$$\sqrt{n} \sum_{j=1}^s \left\{ \alpha_j \hat{F}_{jn}(\nu_j) + \beta_j \hat{G}_{jn}(\zeta_j) - \frac{1}{2}(\alpha_j + \beta_j) \right\} \xrightarrow{d} N(0, \sigma_W^2). \quad (3)$$

Then, using (3) and Slutsky's Lemma, we have

$$\sqrt{n} \sum_{j=1}^s \left\{ \lambda_j (\hat{\nu}_{jn} - \nu_j) + \gamma_j (\hat{\zeta}_{jn} - \zeta_j) \right\} \xrightarrow{d} N(0, \sigma_W^2),$$

and thus (2) follows. Calculation of the covariance matrix follows in a direct fashion from the weak Bahadur representations for the sample median and the sample MAD.  $\square$

Using Lemma 1 we obtain the following result on asymptotic normality of  $\hat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$ .

**Theorem 2** *Asymptotic Normality of  $\hat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$ . Assume (A). Then  $\hat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$  is AN with mean  $\boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}})$  and covariance matrix  $\frac{1}{n} \boldsymbol{\Sigma}^*$ , where  $\boldsymbol{\Sigma}^* = (\sigma_{ij}^*)_{s \times s}$  with*

$$\sigma_{ij}^* = \frac{\sigma_{11}^{\mathbf{u}_i \mathbf{u}_j}}{\zeta_i \zeta_j} + \frac{(\mathbf{u}'_i \mathbf{x} - \nu_i)}{\zeta_i^2 \zeta_j} \sigma_{21}^{\mathbf{u}_i \mathbf{u}_j} + \frac{(\mathbf{u}'_j \mathbf{x} - \nu_j)}{\zeta_j^2 \zeta_i} \sigma_{12}^{\mathbf{u}_i \mathbf{u}_j} + \frac{(\mathbf{u}'_i \mathbf{x} - \nu_i)(\mathbf{u}'_j \mathbf{x} - \nu_j)}{\zeta_i^2 \zeta_j^2} \sigma_{22}^{\mathbf{u}_i \mathbf{u}_j}.$$

**PROOF OF THEOREM 2.** Define a function  $\mathbf{g} : \mathbb{R}^{2s} \mapsto \mathbb{R}^s$  as

$$\mathbf{g}(v_1, v_2, \dots, v_{2s}) = \left( \frac{\mathbf{u}'_1 \mathbf{x} - v_1}{v_2}, \dots, \frac{\mathbf{u}'_s \mathbf{x} - v_{2s-1}}{v_{2s}} \right).$$

The derivative of the  $i$ th element of  $\mathbf{g}$  with respect to  $\mathbf{v}_{2s \times 1}$  at  $\mathbf{v} = \boldsymbol{\mu}_{2s \times 1}$  ( $= (\nu_1, \zeta_1, \dots, \nu_s, \zeta_s)'$ ) is  $(0, \dots, 0, -\frac{1}{\zeta_i}, -\frac{(\mathbf{u}'_i \mathbf{x} - \nu_i)}{\zeta_i^2}, 0, \dots, 0)$  (all the components are 0 except the  $(2i-1)^{th}$  and the  $(2i)^{th}$ ). Let

$$\mathbf{S} = \left( \frac{\partial g_i}{\partial v_j} \right)_{s \times 2s} \Big|_{\mathbf{v}=\boldsymbol{\mu}},$$

where  $\frac{\partial g_i}{\partial v_{2i-1}} \Big|_{\mathbf{v}=\boldsymbol{\mu}} = -\frac{1}{\zeta_i}$ ,  $\frac{\partial g_i}{\partial v_{2i}} \Big|_{\mathbf{v}=\boldsymbol{\mu}} = -\frac{(\mathbf{u}'_i \mathbf{x} - \nu_i)}{\zeta_i^2}$ , for  $1 \leq i \leq s$ , and all the other elements of the matrix  $\mathbf{S}$  are 0. Then, by applying Theorem 3.3A of Serfling (1980),  $\mathbf{g}(\hat{\boldsymbol{\mu}}_n) = \hat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$  is AN with mean  $\mathbf{g}(\boldsymbol{\mu}) = \boldsymbol{\eta}$  and covariance matrix  $\frac{1}{n} \mathbf{S} \boldsymbol{\Sigma} \mathbf{S}^T$ , where the  $(i, j)$  element of  $\mathbf{S} \boldsymbol{\Sigma} \mathbf{S}^T$ ,  $1 \leq i, j \leq s$ , is

$$\begin{aligned} & (0, \dots, 0, -\frac{1}{\zeta_i}, -\frac{(\mathbf{u}'_i \mathbf{x} - \nu_i)}{\zeta_i^2}, 0, \dots, 0) \boldsymbol{\Sigma} (0, \dots, 0, -\frac{1}{\zeta_j}, -\frac{(\mathbf{u}'_j \mathbf{x} - \nu_j)}{\zeta_j^2}, 0, \dots, 0)^T \\ &= \frac{\sigma_{11}^{\mathbf{u}_i \mathbf{u}_j}}{\zeta_i \zeta_j} + \frac{(\mathbf{u}'_i \mathbf{x} - \nu_i)}{\zeta_i^2 \zeta_j} \sigma_{21}^{\mathbf{u}_i \mathbf{u}_j} + \frac{(\mathbf{u}'_j \mathbf{x} - \nu_j)}{\zeta_j^2 \zeta_i} \sigma_{12}^{\mathbf{u}_i \mathbf{u}_j} + \frac{(\mathbf{u}'_i \mathbf{x} - \nu_i)(\mathbf{u}'_j \mathbf{x} - \nu_j)}{\zeta_i^2 \zeta_j^2} \sigma_{22}^{\mathbf{u}_i \mathbf{u}_j} \\ &= \sigma_{ij}^*. \end{aligned}$$

Hence  $\mathbf{S} \boldsymbol{\Sigma} \mathbf{S}^T = \boldsymbol{\Sigma}^*$ , and the result follows.  $\square$

**Corollary 3** *Under the assumptions of Theorem 2, we have*

$$n(\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n) - \boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}}))' \boldsymbol{\Sigma}^{*-1} (\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n) - \boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}})) \xrightarrow{d} \chi_s^2. \quad (4)$$

PROOF OF COROLLARY 3. The proof follows by standard results on asymptotic distributions of quadratic forms in asymptotically normal random vectors (Serfling, 1980).  $\square$

Pan, Fung, and Fang (2000) obtain Corollary 3 allowing only one version of sample MAD and imposing the more restrictive assumptions (i)  $F_j(t)$  is 2nd order differentiable, for all  $t \in \mathbb{R}$ ; (ii)  $\nu_j$  and  $\zeta_j$  are continuous functions of  $\mathbf{u}_j$ ,  $1 \leq j \leq s$ ; (iii)  $C_j = 0$ ,  $1 \leq j \leq s$ . Our result allows use of certain modified versions of sample MAD, we assume only 1st order differentiability of  $F_j$  at  $\nu_j$  and  $\nu_j \pm \zeta_j$ , we do not need the continuity assumption on  $\nu_j$  and  $\zeta_j$ , and we permit  $C_j \neq 0$ ,  $1 \leq j \leq s$ .

Pan, Fung, and Fang (2000) use the above quadratic form as an outlyingness function. However, there are several issues. First, the quadratic form defined in (4) involves the population parameter  $\boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}})$  and the derivative  $F'$  at  $\nu_j$  and  $\nu_j \pm \zeta_j$ , through  $\boldsymbol{\Sigma}^*$ . Therefore, to use (4) in practice one has to use the bootstrap to find  $\nu_j$  and  $\nu_j \pm \zeta_j$ , and one has to use a kernel density estimator to estimate  $F'$  at  $\nu_j$  and  $\nu_j \pm \zeta_j$ . All this becomes highly computational. Also, their outlyingness function suffers from nonsingularity issues with the sample covariance matrix of the vector  $\boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}})$ . Moreover, it is not affine invariant because  $\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Delta, \mathbb{X}_n)$  is not. Therefore, (4) as an outlyingness function presents serious difficulties motivating us to consider variations.

## 3 Affine Invariance of Finite Vectors of Projections

### 3.1 $\boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}})$ is not affine invariant

In general, by *weak affine invariance* of a functional  $T(F)$  is meant that

$$T(\mathbf{A}\mathbf{x} + \mathbf{b}, F_{\mathbf{A}\mathbf{X} + \mathbf{b}}) = cT(\mathbf{x}, F_{\mathbf{X}}),$$

where  $\mathbf{A}_{d \times d}$  is nonsingular,  $\mathbf{b}$  is any vector in  $\mathbb{R}^d$  and  $c = c(\mathbf{A}, \mathbf{b}, F_{\mathbf{X}})$  is a constant. We now comment in particular on the functional  $T(\mathbf{x}, F_{\mathbf{X}}) = \boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}})$ .

**Remark 4**  $\boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}})$  is not weakly affine invariant (not even orthogonally invariant).

PROOF OF REMARK 4. Let  $\mathbf{X} \mapsto \mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ , with  $\mathbf{A}$  nonsingular and  $\mathbf{b}$  any vector in  $\mathbb{R}^d$ . Then, for  $1 \leq j \leq s$ ,

$$\frac{\mathbf{u}'_j \mathbf{y} - \text{med}(\mathbf{u}'_j \mathbf{Y})}{\text{MAD}(\mathbf{u}'_j \mathbf{Y})} = \frac{\mathbf{u}'_j \mathbf{A}\mathbf{x} - \text{med}(\mathbf{u}'_j \mathbf{A}\mathbf{X})}{\text{MAD}(\mathbf{u}'_j \mathbf{A}\mathbf{X})}.$$

Now we provide counterexamples showing that this transformation need not be invariant even under orthogonal transformation. For convenience, we present this for sample versions.

Suppose  $\mathbf{x}_1 = (1, 2)'$ ,  $\mathbf{x}_2 = (5, 4)'$ ,  $\mathbf{x}_3 = (3, 11)'$ ,  $\mathbf{x}_4 = (8, 1)'$  and  $\mathbf{x}_5 = (13, 18)'$  are a random sample from some bivariate distribution function  $F$ . Let  $\mathbf{u} = \frac{1}{\sqrt{2}}(1, 1)'$ . Note that  $\|\mathbf{u}\| = 1$ . Then  $\mathbf{u}'\mathbf{x}_1 = 3/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{x}_2 = 9/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{x}_3 = 14/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{x}_4 = 9/\sqrt{2}$  and  $\mathbf{u}'\mathbf{x}_5 = 31/\sqrt{2}$ . Hence  $\text{med}(\mathbf{u}'\mathbb{X}_5) = 9/\sqrt{2}$ . Thus

$$\begin{aligned}\mathbf{u}'\mathbf{x}_1 - \text{med}(\mathbf{u}'\mathbb{X}_5) &= -6/\sqrt{2}, & |\mathbf{u}'\mathbf{x}_2 - \text{med}(\mathbf{u}'\mathbb{X}_5)| &= 0, \\ \mathbf{u}'\mathbf{x}_3 - \text{med}(\mathbf{u}'\mathbb{X}_5) &= 5/\sqrt{2}, & \mathbf{u}'\mathbf{x}_4 - \text{med}(\mathbf{u}'\mathbb{X}_5) &= 0, \\ \mathbf{u}'\mathbf{x}_5 - \text{med}(\mathbf{u}'\mathbb{X}_5) &= 22/\sqrt{2}.\end{aligned}$$

Therefore,  $\text{MAD}(\mathbf{u}'\mathbb{X}_5) = 5/\sqrt{2}$  and we obtain

$i$	$\eta(\mathbf{x}_i, \mathbf{u}, \mathbb{X}_5)$
1	$-6/5$
2	0
3	1
4	0
5	$22/5$

Now we transform  $\mathbb{X}_5 \mapsto \mathbb{Y}_5$  by  $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$  for  $1 \leq i \leq 5$ , where  $\mathbf{A}$  is positive definite. Here we consider three cases of  $\mathbf{A}$ .

*Case 1.* Let  $\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$ , a diagonal matrix. Then  $\mathbf{y}_1 = (3, 4)'$ ,  $\mathbf{y}_2 = (15, 8)'$ ,  $\mathbf{y}_3 = (9, 22)'$ ,  $\mathbf{y}_4 = (24, 2)'$ ,  $\mathbf{y}_5 = (39, 36)'$ , and hence  $\mathbf{u}'\mathbf{y}_1 = 7/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{y}_2 = 23/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{y}_3 = 31/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{y}_4 = 26/\sqrt{2}$ , and  $\mathbf{u}'\mathbf{y}_5 = 75/\sqrt{2}$ . Therefore,  $\text{med}(\mathbf{u}'\mathbb{Y}_5) = 26/\sqrt{2}$ . Thus

$$\begin{aligned}\mathbf{u}'\mathbf{y}_1 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= -19/\sqrt{2}, & \mathbf{u}'\mathbf{y}_2 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= -3/\sqrt{2}, \\ \mathbf{u}'\mathbf{y}_3 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= 5/\sqrt{2}, & \mathbf{u}'\mathbf{y}_4 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= 0, \\ \mathbf{u}'\mathbf{y}_5 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= 49/\sqrt{2}.\end{aligned}$$

Hence  $\text{MAD}(\mathbf{u}'\mathbb{Y}_5) = 5/\sqrt{2}$  and we obtain

$i$	$\eta(\mathbf{y}_i, \mathbf{u}, \mathbb{Y}_5)$
1	$-19/5$
2	$-3/5$
3	1
4	0
5	$49/5$

Note that the values of  $\eta(\mathbf{y}_i, \mathbf{u}, \mathbb{Y}_5)$  are completely different from the values of  $\eta(\mathbf{x}_i, \mathbf{u}, \mathbb{X}_5)$ .

*Case 2.* Let  $\mathbf{A} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$ , an orthogonal matrix. Then  $\mathbf{y}_1 = (3/\sqrt{2}, 1/\sqrt{2})'$ ,

$\mathbf{y}_2 = (9/\sqrt{2}, -1/\sqrt{2})'$ ,  $\mathbf{y}_3 = (14/\sqrt{2}, 8/\sqrt{2})'$ ,  $\mathbf{y}_4 = (9/\sqrt{2}, -7/\sqrt{2})'$ ,  $\mathbf{y}_5 = (31/\sqrt{2}, 5/\sqrt{2})'$ , and  $\mathbf{u}'\mathbf{y}_1 = 2$ ,  $\mathbf{u}'\mathbf{y}_2 = 4$ ,  $\mathbf{u}'\mathbf{y}_3 = 11$ ,  $\mathbf{u}'\mathbf{y}_4 = 1$ , and  $\mathbf{u}'\mathbf{y}_5 = 18$ . Therefore,  $\text{med}(\mathbf{u}'\mathbb{Y}_5) = 4$ , and thus

$$\begin{aligned}\mathbf{u}'\mathbf{y}_1 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= -2, & |\mathbf{u}'\mathbf{y}_2 - \text{med}(\mathbf{u}'\mathbb{Y}_5)| &= 0, \\ \mathbf{u}'\mathbf{y}_3 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= 7, & \mathbf{u}'\mathbf{y}_4 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= -3, \\ \mathbf{u}'\mathbf{y}_5 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= 14.\end{aligned}$$

Therefore,  $\text{MAD}(\mathbf{u}'\mathbb{Y}_5) = 3$  and we obtain

$i$	$\boldsymbol{\eta}(\mathbf{y}_i, \mathbf{u}, \mathbb{Y}_5)$
1	-2/3
2	0
3	7/3
4	-1
5	14/3

Note that the values of  $\boldsymbol{\eta}(\mathbf{y}_i, \mathbf{u}, \mathbb{Y}_5)$  are completely different from the values of  $\boldsymbol{\eta}(\mathbf{x}_i, \mathbf{u}, \mathbb{X}_5)$ .

*Case 3.* Let  $\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 1 & 1 \end{pmatrix}$ . Note that  $\mathbf{A}$  is neither symmetric nor orthogonal. Then  $\mathbf{y}_1 = (7, 3)'$ ,  $\mathbf{y}_2 = (23, 9)'$ ,  $\mathbf{y}_3 = (31, 14)'$ ,  $\mathbf{y}_4 = (26, 9)'$ ,  $\mathbf{y}_5 = (75, 31)'$ , and  $\mathbf{u}'\mathbf{y}_1 = 10/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{y}_2 = 32/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{y}_3 = 45/\sqrt{2}$ ,  $\mathbf{u}'\mathbf{y}_4 = 35/\sqrt{2}$ , and  $\mathbf{u}'\mathbf{y}_5 = 106/\sqrt{2}$ . Therefore,  $\text{med}(\mathbf{u}'\mathbb{Y}_5) = 35/\sqrt{2}$  and thus

$$\begin{aligned}\mathbf{u}'\mathbf{y}_1 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= -25/\sqrt{2}, & \mathbf{u}'\mathbf{y}_2 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= -3/\sqrt{2}, \\ \mathbf{u}'\mathbf{y}_3 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= 10/\sqrt{2}, & \mathbf{u}'\mathbf{y}_4 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= 0, \\ \mathbf{u}'\mathbf{y}_5 - \text{med}(\mathbf{u}'\mathbb{Y}_5) &= 71/\sqrt{2}.\end{aligned}$$

Therefore,  $\text{MAD}(\mathbf{u}'\mathbb{Y}_5) = 10/\sqrt{2}$  and hence

$i$	$\boldsymbol{\eta}(\mathbf{y}_i, \mathbf{u}, \mathbb{Y}_5)$
1	-5/2
2	-3/10
3	1
4	0
5	71/10

Again, the values of  $\boldsymbol{\eta}(\mathbf{y}_i, \mathbf{u}, \mathbb{Y}_5)$  are completely different from those of  $\boldsymbol{\eta}(\mathbf{x}_i, \mathbf{u}, \mathbb{X}_5)$ .  $\square$

Therefore, to construct a weakly affine invariant outlyingness function using only finitely many projections, one needs an appropriate preliminary transformation. In particular, to make  $\boldsymbol{\eta}(\mathbf{x}, \Delta, F_{\mathbf{X}})$  weakly affine invariant, we use a strong invariant coordinate system (SICS) functional (Serfling, 2010) as treated in the next section.

### 3.2 Standardization using a SICS Functional

**Definition 5** (Serfling, 2010). A positive definite matrix-valued functional  $D(F)$  is a *strong invariant coordinate system (SICS) functional* if, for  $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ ,

$$D(F_{\mathbf{Y}}) = k_3 D(F_{\mathbf{X}}) \mathbf{A}^{-1},$$

where  $\mathbf{A}_{d \times d}$  is nonsingular,  $\mathbf{b}$  is any vector in  $\mathbb{R}^d$ , and  $k_3 = k_3(\mathbf{A}, \mathbf{b}, F_{\mathbf{X}})$  is a scalar.

See discussion in Tyler, Critchley, Dümbgen and Oja, 2009, Serfling, 2010, and Ilmonen, Oja, and Serfling, 2011.

Following the method of construction of sample SICS functionals in Serfling (2010), here we give a family of SICS functionals and provide two easily computable sample SICS functionals. Let  $\mathbb{Z}_N$  be a subset of  $\mathbb{X}_n$  of size  $N$  obtained through some affine invariant and permutation invariant procedure. Next, for  $m = \lfloor N/(d+1) \rfloor$ , form  $d+1$  means  $\bar{\mathbf{Z}}_1, \dots, \bar{\mathbf{Z}}_{d+1}$  based, respectively, on consecutive blocks of size  $m$  from  $\mathbb{Z}_N$ . Define the matrix

$$\mathbf{W}(\mathbb{X}_n) = [(\mathbf{Z}_2 - \mathbf{Z}_1), \dots, (\mathbf{Z}_{d+1} - \mathbf{Z}_1)]_{d \times d}.$$

Then the matrix  $D(\mathbb{X}_n) = \mathbf{W}(\mathbb{X}_n)^{-1}$  is a sample SICS functional.

The robustness and computational burden rest on the method of choosing  $\mathbb{Z}_N$ . One implementation is to let  $\mathbb{Z}_N$  be the set of the observations selected and used in computing the well-known Fast-MCD covariance matrix  $\hat{\Sigma}$  with  $N = \alpha n$ , where  $\alpha = 0.5$  or  $0.75$ . This uses all the observations in selecting  $\mathbb{Z}_N$  and all those observations in defining  $W(\mathbb{X}_n)$ . Another sample SICS functional is based on  $\mathbb{Z}_N$  consisting of the “inner” observations obtained in *spatial trimming*. These implementations will be treated in detail later. They trade off robustness versus computational burden.

We now establish that the vector  $\boldsymbol{\eta}(\mathbf{D}(F_{\mathbf{X}})\mathbf{x}, \Delta, F_{\mathbf{D}(F_{\mathbf{X}})\mathbf{X}})$  is weakly affine invariant.

**Theorem 6** *Let  $\mathbf{X}$  have distribution  $F_{\mathbf{X}}$  on  $\mathbb{R}^d$  and let  $\mathbf{X}^* = \mathbf{D}(F_{\mathbf{X}})\mathbf{X}$ , where  $\mathbf{D}(F)$  is a SICS functional. Then  $\boldsymbol{\eta}(\mathbf{x}^*, \Delta, F_{\mathbf{X}^*})$  is weakly affine invariant.*

**PROOF OF THEOREM 6.** Suppose  $\mathbf{X} \mapsto \mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ , where  $\mathbf{A}$  is a  $d \times d$  nonsingular matrix and  $\mathbf{b}$  is any vector in  $\mathbb{R}^d$ . Now, transform  $\mathbf{Y} \mapsto \mathbf{Y}^* = \mathbf{B}(F_{\mathbf{Y}})\mathbf{Y}$ . Then, using Definition 5, we have  $\mathbf{Y}^* = \mathbf{B}(F_{\mathbf{A}\mathbf{X} + \mathbf{b}})(\mathbf{A}\mathbf{X} + \mathbf{b}) = k \mathbf{B}(F_{\mathbf{X}})\mathbf{X} + k \mathbf{B}(F_{\mathbf{X}})\mathbf{A}^{-1}\mathbf{b} = k \mathbf{X}^* + \mathbf{c}$ , where  $k = k(\mathbf{b}, \mathbf{A}, F_{\mathbf{X}})$ , and  $\mathbf{c} = k \mathbf{B}(F_{\mathbf{X}})\mathbf{A}^{-1}\mathbf{b}$ . Hence, for  $1 \leq j \leq s$ ,

$$\begin{aligned} & \frac{\mathbf{u}'_j \mathbf{y}^* - \text{med}(\mathbf{u}'_j \mathbf{Y}^*)}{\text{MAD}(\mathbf{u}'_j \mathbf{Y}^*)} \\ &= \frac{\mathbf{u}'_j (k \mathbf{x}^* + \mathbf{c}) - \text{med}(\mathbf{u}'_j (k \mathbf{X}^* + \mathbf{c}))}{\text{MAD}(\mathbf{u}'_j (k \mathbf{X}^* + \mathbf{c}))} \\ &= \text{sgn}(k) \frac{\mathbf{u}'_j \mathbf{x}^* - \text{med}(\mathbf{u}'_j \mathbf{X}^*)}{\text{MAD}(\mathbf{u}'_j \mathbf{X}^*)}. \end{aligned}$$

Thus  $\boldsymbol{\eta}(\mathbf{y}^*, F_{\mathbf{Y}^*}) = \text{sgn}(k) \boldsymbol{\eta}(\mathbf{x}^*, F_{\mathbf{X}^*})$  and the result follows.  $\square$

Now we revisit the example of Remark 4 and see that pre-transformation by a SICS functional corrects the lack of weak affine invariance. For the given data set, it can be seen that

$$\mathbf{D}(\mathbb{X}_5) = \begin{pmatrix} -0.1259 & 0.0370 \\ 0.0519 & -0.0741 \end{pmatrix}$$

is SICS. We first transform  $\mathbb{X}_5 \mapsto \mathbb{X}_5^*$  by  $\mathbf{x}_i^* = (\mathbf{D}(\mathbb{X}_5)) \mathbf{x}_i$ ,  $1 \leq i \leq 5$ , and obtain  $\mathbf{u}'\mathbf{x}_1^* = -0.1048$ ,  $\mathbf{u}'\mathbf{x}_2^* = -0.3666$ ,  $\mathbf{u}'\mathbf{x}_3^* = -0.4452$ ,  $\mathbf{u}'\mathbf{x}_4^* = -0.4452$ , and  $\mathbf{u}'\mathbf{x}_5^* = -1.1523$ . Therefore,  $\text{med}(\mathbf{u}'\mathbb{X}_5^*) = -0.4452$  and  $\text{MAD}(\mathbf{u}'\mathbb{X}_5^*) = 0.0786$  and hence

$i$	$\boldsymbol{\eta}(\mathbf{x}_i^*, \mathbf{u}, \mathbb{X}_5^*)$
1	4.333
2	1
3	0
4	0
5	-9

Now, we transform  $\mathbb{X}_5 \mapsto \mathbb{Y}_5$  by  $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$ ,  $1 \leq i \leq 5$ , with  $\mathbf{A}$  positive definite. For each of the three cases of  $\mathbf{A}$  considered previously, we obtain

$i$	$\boldsymbol{\eta}(\mathbf{y}_i^*, \mathbf{u}, \mathbb{Y}_5^*)$
1	4.333
2	1
3	0
4	0
5	-9

Note that the values are exactly as obtained above for  $\mathbf{x}^*$ .  $\square$

## 4 Obtaining a Reduced Vector of Projections

Let  $\mathbb{X}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$  be a random sample from some  $F$  on  $\mathbb{R}^d$  and let  $\mathbf{D}(\mathbb{X}_n)$  be a sample SICS functional. We describe in Algorithm A below the steps for transforming  $\mathbb{X}_n$  to a set of  $t$  vectors  $\mathbb{V}_n = \{\mathbf{V}_1, \dots, \mathbf{V}_n\}$  by a projection pursuit approach using  $\mathbf{u}_i$ ,  $1 \leq i \leq s$ , combined with a SICS preliminary transformation and a PCA dimension reduction step. We also provide key properties of  $\mathbb{V}_n$ . First, however, we provide some preliminaries needed for the steps of Algorithm A that involve “spatial trimming”, the Dümbgen-Tyler TR functional, and the Fast-MCD scatter matrix. (See Serfling, 2010, and Mazumder and Serfling, 2012, for more details.)

TRANSFORMATION-RETRANSFORMATION (TR) MATRICES. Any inverse square root of a

scatter or shape matrix is called a “*transformation-retransformation*” (TR) matrix and is given by any  $\mathbf{M}(\mathbb{X}_n)$  satisfying

$$\mathbf{A}'\mathbf{M}(\mathbb{Y}_n)'\mathbf{M}(\mathbb{Y}_n)\mathbf{A} = k_2\mathbf{M}(\mathbb{X}_n)'\mathbf{M}(\mathbb{X}_n) \quad (5)$$

for  $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$ , and with  $k_2 = k_2(\mathbf{A}, \mathbf{b}, \mathbb{X}_n)$  a positive scalar function of  $\mathbf{A}$ ,  $\mathbf{b}$ , and  $\mathbb{X}_n$ .

**THE MCD AND FAST-MCD SCATTER ESTIMATORS.** A highly robust scatter estimator is the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985). It is the covariance matrix corresponding to a specified fraction (e.g., 50%) of observations such that this covariance matrix has minimal determinant. Typically used is the computationally more efficient *Fast-MCD*, which *approximates* the MCD (Rousseeuw and Van Driessen, 1999) and is implemented in the R packages *MASS*, *rrcov*, and *robustbase*, for example, as well as in other software packages. We denote the TR standardization matrix corresponding to Fast-MCD by “**MCD**”. Also, we denote by  $\mathbb{J}_{\text{MCD}}$  the indices of the corresponding “inner” set of observations selected by Fast-MCD.

**THE TYLER (1987) SCATTER ESTIMATOR.** A useful scatter matrix was introduced by Tyler (1987) to achieve certain favorable theoretical properties in the elliptical model. It is easily computed in any dimension, making it an alternative to Fast-MCD when Fast-MCD is computationally prohibitive, although it is not as robust as Fast-MCD. With respect to a specified location functional  $\theta(\mathbb{X}_n)$ , the Tyler matrix is defined as  $\mathbf{V}(\mathbb{X}_n) = (\mathbf{M}_s(\mathbb{X}_n)'\mathbf{M}_s(\mathbb{X}_n))^{-1}$ , with  $\mathbf{M}_s(\mathbb{X}_n)$  the TR matrix defined as the unique symmetric square root of  $\mathbf{V}^{-1}$  obtained through the M-estimation equation

$$n^{-1} \sum_{i=1}^n \left\{ \left( \frac{\mathbf{M}_s(\mathbb{X}_n)(\mathbf{X}_i - \theta(\mathbb{X}_n))}{\|\mathbf{M}_s(\mathbb{X}_n)(\mathbf{X}_i - \theta(\mathbb{X}_n))\|} \right) \left( \frac{\mathbf{M}_s(\mathbb{X}_n)(\mathbf{X}_i - \theta(\mathbb{X}_n))}{\|\mathbf{M}_s(\mathbb{X}_n)(\mathbf{X}_i - \theta(\mathbb{X}_n))\|} \right)' \right\} = d^{-1} \mathbf{I}_d. \quad (6)$$

An iterative algorithm using Cholesky factorizations to compute  $\mathbf{M}_s(\mathbb{X}_n)$  quickly in any practical dimension is given in Tyler (1987). Another solution of (6) is given by the upper triangular square root  $\mathbf{M}_t(\mathbb{X}_n)$  of  $\mathbf{V}^{-1}$  and is computed by a similar algorithm. In Tyler (1987), the quantity  $\theta(\mathbb{X}_n)$  is specified as a known constant. For inference situations when  $\theta(\mathbb{X}_n)$  is not known or specified, a symmetrized version of  $\mathbf{M}_s(\mathbb{X}_n)$  eliminating the need of a location measure is given by Dümbgen (1998):  $\mathbf{M}_{s1}(\mathbb{X}_n) = \mathbf{M}_s(\mathbb{X}_n \ominus \mathbb{X}_n)$ , where  $\mathbb{X}_n \ominus \mathbb{X}_n$  denotes the set of differences  $\mathbf{X}_i - \mathbf{X}_j$ . Convenient R-packages (e.g., ICSNP) are available for computation of these estimators. See Tyler (1987) and Serfling (2010) for detailed discussion. We denote the “Dümbgen-Tyler” TR matrix  $\mathbf{M}_{s1}(F)$  by “**DT**”.

**SPATIAL TRIMMING.** An affine invariant “*TR spatial outlyingness function*” or “*Mahalanobis spatial outlyingness function*” based on a selected TR matrix  $\mathbf{M}(\mathbb{X}_n)$  is given by

$$O_S^{(\text{TR})}(\mathbf{x}, \mathbb{X}_n) = \left\| n^{-1} \sum_{i=1}^n \mathbf{S}(\mathbf{M}(\mathbb{X}_n)(\mathbf{x} - \mathbf{X}_i)) \right\|, \quad \mathbf{x} \in \mathbb{R}^d, \quad (7)$$

based on the *spatial sign function* (or *unit vector function*),

$$\mathbf{S}(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|}, & \mathbf{x} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0}, & \mathbf{x} = \mathbf{0}. \end{cases}$$

“Spatial trimming” with specified threshold  $\lambda_0$  (possibly depending on  $n$  and/or  $d$ ) consists of selecting those “inner” observations satisfying  $O_S^{(\text{TR})}(\mathbf{X}_i, \mathbb{X}_n) \leq \lambda_0$ . For  $\mathbf{M}(\mathbb{X}_n)$  given by DT, we denote by  $\mathbb{J}_{\text{MS}_{\text{DT}}}$  the set of indices of the “inner” observations.

#### 4.1 Algorithm A, for $\mathbb{V}_n$

1. Standardize  $\mathbf{X}_i$ ,  $1 \leq i \leq n$ , with the sample SICS functional  $\mathbf{D}(\mathbb{X}_n)_{d \times d}$  via  $\mathbf{X}_i^* = \mathbf{D}(\mathbb{X}_n) \mathbf{X}_i$ ,  $1 \leq i \leq n$ , and put  $\mathbb{X}_n^* = [\mathbf{X}_1^* \cdots \mathbf{X}_n^*]$ .
2. Choose the number directions  $s$  and select  $s$  unit vectors  $\Delta = \{\mathbf{u}_1, \dots, \mathbf{u}_s\}$  uniformly distributed on the unit sphere in  $\mathbb{R}^d$  but lying on distinct diameters, following the algorithm of Fang and Wang (1994).
3. Calculate the projections  $\mathbf{u}'_i \mathbf{X}_j^*$ ,  $1 \leq i \leq s$ ,  $1 \leq j \leq n$ , and put  $\mathbf{L}_{s \times n} = \mathbf{U}' \mathbb{X}_n^* = [\mathbf{u}'_i \mathbf{X}_j^*]_{s \times n}$ , and denote by  $\mathbf{l}'_i = (\mathbf{u}'_i \mathbf{X}_1^*, \dots, \mathbf{u}'_i \mathbf{X}_n^*)_{1 \times n}$  the  $i$ -th row of  $\mathbf{L}$ ,  $1 \leq i \leq s$ .
4. Compute the vector  $\mathbf{med} = (\widehat{\nu}(\mathbf{u}'_1 \mathbb{X}_n), \dots, \widehat{\nu}(\mathbf{u}'_s \mathbb{X}_n))'_{s \times 1}$ , such that  $\widehat{\nu}(\mathbf{u}'_i \mathbb{X}_n) = \text{med}(\mathbf{l}'_i)$ ,  $1 \leq i \leq s$ .
5. Let  $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_s]'_{s \times n}$ , with  $\mathbf{h}'_i = (|\mathbf{u}'_i \mathbf{X}_1^* - \widehat{\nu}(\mathbf{u}'_i \mathbb{X}_n)|, \dots, |\mathbf{u}'_i \mathbf{X}_n^* - \widehat{\nu}(\mathbf{u}'_i \mathbb{X}_n)|)_{1 \times n}$ ,  $1 \leq i \leq s$ .
6. Compute the vector  $\mathbf{MAD} = (\widehat{\zeta}(\mathbf{u}'_1 \mathbb{X}_n), \dots, \widehat{\zeta}(\mathbf{u}'_s \mathbb{X}_n))'_{s \times 1}$ , with  $\widehat{\zeta}(\mathbf{u}'_i \mathbb{X}_n) = \text{med}(\mathbf{h}_i)$ ,  $1 \leq i \leq s$ .
7. Put  $\widehat{\boldsymbol{\eta}}_n = [\widehat{\boldsymbol{\eta}}_{1n} \cdots \widehat{\boldsymbol{\eta}}_{nn}]_{s \times n}$ , where  $\widehat{\boldsymbol{\eta}}_{in} = (\widehat{\eta}_n(i, 1), \dots, \widehat{\eta}_n(i, s))'_{s \times 1}$ ,  $1 \leq i \leq n$ , with

$$\widehat{\eta}_n(i, j) = \frac{\mathbf{u}'_j \mathbf{X}_i^* - \widehat{\nu}(\mathbf{u}'_j \mathbb{X}_n)}{\widehat{\zeta}(\mathbf{u}'_j \mathbb{X}_n)}, \quad 1 \leq j \leq s, \quad 1 \leq i \leq n.$$

The initial  $d \times n$  data matrix  $\mathbb{X}_n$  now has been reduced to a new data matrix  $\widehat{\boldsymbol{\eta}}_n$  of dimension  $s \times n$ , i.e., consisting of  $n$   $s$ -vectors.

8. This step and the next calculate the sample covariance matrix of the  $s \times 1$  dimensional vector

$$\boldsymbol{\eta}_i = \widehat{\boldsymbol{\eta}}(\mathbb{J}_{\text{MS}_{\text{DT}}}(i)) \quad (\text{or } \widehat{\boldsymbol{\eta}}(\mathbb{J}_{\text{MCD}}(i))), \quad 1 \leq i \leq K,$$

where  $K$  denotes the cardinality of  $\mathbb{J}_{\text{MS}_{\text{DT}}}$  (or  $\mathbb{J}_{\text{MCD}}$ ). Form the matrix

$$\widehat{\boldsymbol{\eta}} = [\boldsymbol{\eta}_1 \cdots \boldsymbol{\eta}_K]_{s \times K}.$$

9. Put  $\bar{\boldsymbol{\eta}} = \frac{1}{K} \sum_{i=1}^K \boldsymbol{\eta}_i$ , the sample mean of the  $\boldsymbol{\eta}_i$ , and

$$\widehat{\boldsymbol{\Sigma}}_{1n} = \left[ \frac{1}{K} \sum_{i=1}^K (\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})' \right],$$

the sample covariance matrix of the  $\boldsymbol{\eta}_i$ ,  $1 \leq i \leq K$ .

10. We now reduce the dimension of the vectors  $\boldsymbol{\eta}_i$  from  $s$  to  $t < s$ , in order to eliminate redundancy. This step and those to Step 13 accomplish this by transforming the vectors  $\tilde{\boldsymbol{\eta}}_i$  to new vectors  $\mathbf{V}_i$  having dimension  $t$ .

Calculate the eigenvalues  $\lambda_1 \geq \dots \geq \lambda_s$  of  $\widehat{\boldsymbol{\Sigma}}_{1n}$  and the orthogonal matrix  $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_s]_{s \times s}$  containing the corresponding eigenvectors  $\mathbf{p}_1, \dots, \mathbf{p}_s$  as column vectors. Put  $\boldsymbol{\Lambda}_{s \times s} = \text{diag}(\lambda_1, \dots, \lambda_s)$ .

11. Find the number  $t$  of eigenvalues whose values are greater than  $10^{-6}$ .
12. Define the  $s \times 1$  dimensional vectors  $\tilde{\mathbf{V}}_i = \mathbf{P}' \tilde{\boldsymbol{\eta}}_{in}$ ,  $1 \leq i \leq n$ .
13. Calculate the  $t \times 1$  vectors  $\mathbf{V}_i$  which contain the first  $t$  components of the vectors  $\tilde{\mathbf{V}}_i$ , i.e., calculate  $\mathbf{V}_i = (\tilde{\mathbf{V}}(1, i), \dots, \tilde{\mathbf{V}}(t, i))'$ ,  $1 \leq i \leq n$ , where, for  $1 \leq j \leq t$ ,  $\tilde{\mathbf{V}}(j, i)$  represents the  $j$ th element of the vector  $\tilde{\mathbf{V}}_i$ ,  $1 \leq i \leq n$ . Put

$$\mathbb{V}_n = [\mathbf{V}_1 \cdots \mathbf{V}_n]_{t \times n}.$$

14. Calculate the  $t \times t$  matrix  $\boldsymbol{\Lambda}$  containing the 1st  $t$  rows and the first  $t$  columns of  $\boldsymbol{\Lambda}_{s \times s}$ . This is the covariance matrix of each  $\mathbf{V}(i)$ ,  $1 \leq i \leq n$ , as will be proved below.

## 4.2 Properties of the Reduced Vector $\mathbb{V}_n$

First we show that the above algorithm is affine invariant, in the sense that if we transform  $\mathbf{X}_i \rightarrow \mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \mathbf{b}$ ,  $1 \leq i \leq n$ , where  $\mathbf{A}$  is any  $d \times d$  dimensional nonsingular matrix and  $\mathbf{b}$  is any vector in  $\mathbb{R}^d$ , then the matrix of vectors  $\mathbf{V}_i$ ,  $1 \leq i \leq n$ , will remain the same up to a global sign change.

**Lemma 7** *Let  $\mathbf{X}_1, \dots, \mathbf{X}_n$  be a random sample from  $d$  dimensional distribution function  $F$ . Suppose  $\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \mathbf{b}$ , where  $\mathbf{A}$  is any  $d \times d$  dimensional nonsingular matrix and  $\mathbf{b}$  is any vector in  $\mathbb{R}^d$ . Put  $\mathbb{Y}_n = [\mathbf{Y}_1 \cdots \mathbf{Y}_n]$ . Further, assume that  $\mathbb{V}_n(\mathbb{X}_n)$  is obtained using **Algorithm A** starting with the data matrix  $\mathbb{X}_n$ , and that  $\mathbb{V}_n(\mathbb{Y}_n)$  is obtained using **Algorithm A** starting with the data matrix  $\mathbb{Y}_n$ . Then*

$$\mathbb{V}_n(\mathbb{Y}_n) = \text{sgn}(k) \mathbb{V}_n(\mathbb{X}_n), \tag{8}$$

where  $k = k(\mathbf{A}, \mathbf{b}, \mathbb{X}_n)$ .

PROOF OF LEMMA 7. Let  $\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i$ ,  $1 \leq i \leq n$ , where  $\mathbf{A}$  is positive definite. Define  $\mathbb{Y}_n = \{\mathbf{Y}_1 \cdots \mathbf{Y}_n\}$ . The index set  $\mathbb{J}_{\text{MCD}}$  or  $\mathbb{J}_{\text{MS}_{\text{DT}}}$  for observations  $\mathbf{Y}_i$ ,  $1 \leq i \leq n$ , will be the same as that for  $\mathbf{X}_i$ ,  $1 \leq i \leq n$ , i.e.,  $\mathbb{J}_{\text{MS}_{\text{DT}}}(\mathbb{Y}_n) = \mathbb{J}_{\text{MS}_{\text{DT}}}(\mathbb{X}_n)$  or  $\mathbb{J}_{\text{MCD}}(\mathbb{Y}_n) = \mathbb{J}_{\text{MCD}}(\mathbb{X}_n)$ , because ‘‘spatial trimming’’ method and the minimum covariance determinant method both are affine invariant.

Now we are in a position to follow the steps of **Algorithm A** starting with the data matrix  $\mathbb{Y}_n$ . In Step 1 we find the  $\mathbf{D}(\mathbb{Y}_n)$  SICS functional for  $\mathbb{Y}_n$ . Then, in Step 2, for  $1 \leq i \leq n$ , we standardize  $\mathbf{Y}_i \rightarrow \mathbf{Y}_i^*$  by

$$\mathbf{Y}_i^* = \mathbf{D}(\mathbb{Y}_n) \mathbf{Y}_i = k \mathbf{D}(\mathbb{X}_n) \mathbf{X}_i = k \mathbf{X}_i^*, \quad (9)$$

with  $k = k(\mathbf{A}, \mathbf{b}, \mathbb{X}_n)$ . In Step 3, we calculate the projections  $\mathbf{u}'_i \mathbf{Y}_j^*$ ,  $1 \leq i \leq s$ ,  $1 \leq j \leq n$ . Using equation (9),  $\mathbf{u}'_i \mathbf{Y}_j^*$  becomes  $k \mathbf{u}'_i \mathbf{X}_j^*$ . Therefore,  $\text{med}(\mathbf{u}'_i \mathbf{Y}_j^*, 1 \leq j \leq n) = k \text{med}(\mathbf{u}'_i \mathbf{X}_j^*, 1 \leq j \leq n)$ , which is done in Step 4. A similar argument shows that  $\text{MAD}(\mathbf{u}'_i \mathbf{Y}_j^*, 1 \leq j \leq n) = |k| \text{MAD}(\mathbf{u}'_i \mathbf{X}_j^*, 1 \leq j \leq n)$ , which is done in Steps 5 and 6. Hence, for  $1 \leq i \leq n$ , we have

$$\begin{aligned} \widehat{\boldsymbol{\eta}}_{in}(\mathbf{Y}_i^*, \mathbb{Y}_n^*) &= \left( \frac{\mathbf{u}'_j \mathbf{Y}_i^* - \text{med}(\mathbf{u}'_j \mathbf{Y}_i^*, 1 \leq i \leq n)}{\text{MAD}(\mathbf{u}'_j \mathbf{Y}_i^*, 1 \leq i \leq n)}, 1 \leq j \leq s \right)'_{s \times 1} \\ &= \left( \frac{k \mathbf{u}'_j \mathbf{X}_i^* - k \text{med}(\mathbf{u}'_j \mathbf{X}_i^*, 1 \leq i \leq n)}{|k| \text{MAD}(\mathbf{u}'_j \mathbf{X}_i^*, 1 \leq i \leq n)}, 1 \leq j \leq s \right)'_{s \times 1} \\ &= \text{sgn}(k) \left( \frac{\mathbf{u}'_j \mathbf{X}_i^* - \text{med}(\mathbf{u}'_j \mathbf{X}_i^*, 1 \leq i \leq n)}{\text{MAD}(\mathbf{u}'_j \mathbf{X}_i^*, 1 \leq i \leq n)}, 1 \leq j \leq s \right)'_{s \times 1} \\ &= \text{sgn}(k) \widetilde{\boldsymbol{\eta}}_{in}(\mathbf{X}_i^*, \mathbb{X}_n^*). \end{aligned}$$

This follows Step 7. Now, noting that  $\mathbb{J}_{\text{MS}_{\text{DT}}}(\mathbb{Y}_n) = \mathbb{J}_{\text{MS}_{\text{DT}}}(\mathbb{X}_n)$  (or  $\mathbb{J}_{\text{MCD}}(\mathbb{Y}_n) = \mathbb{J}_{\text{MCD}}(\mathbb{X}_n)$ ), and following Step 8, we have

$$\begin{aligned} \boldsymbol{\eta}_i(\mathbf{Y}_i^*, \mathbb{Y}_n^*) &= \widetilde{\boldsymbol{\eta}}_{\mathbb{J}(i)}(\mathbf{Y}_{\mathbb{J}(i)}^*, \mathbb{Y}_n^*) \\ &= \text{sgn}(k) \widetilde{\boldsymbol{\eta}}_{\mathbb{J}(i)}(\mathbf{X}_{\mathbb{J}(i)}^*, \mathbb{X}_n^*) \\ &= \text{sgn}(k) \boldsymbol{\eta}_i(\mathbf{X}_i^*, \mathbb{X}_n^*), \quad 1 \leq i \leq K, \end{aligned}$$

with  $\mathbb{J}(i)$  the  $i$ th observation of  $\mathbb{J}_{\text{MS}_{\text{DT}}}$  (or  $\mathbb{J}_{\text{MCD}}$ ) and  $K$  the cardinality of  $\mathbb{J}_{\text{MS}_{\text{DT}}}$  (or  $\mathbb{J}_{\text{MCD}}$ ). Now, following Step 9, we have  $\boldsymbol{\Sigma}_1(\mathbb{Y}_n^*) = (\text{sgn}(k))^2 \text{Cov}(\boldsymbol{\eta}_i(\mathbf{X}_i^*, \mathbb{X}_n^*)) = \boldsymbol{\Sigma}_1(\mathbb{X}_n^*)$ . Therefore, following Step 10, the eigenvalues and eigenvectors of the covariance matrix  $\widehat{\boldsymbol{\Sigma}}_{1n}(\mathbb{Y}_n^*)$  will be the same as that for  $\widehat{\boldsymbol{\Sigma}}_{1n}(\mathbb{X}_n^*)$ . Let the  $s \times s$  diagonal matrix  $\boldsymbol{\Lambda}$  contain the eigenvalues of  $\widehat{\boldsymbol{\Sigma}}_{1n}(\mathbb{Y}_n^*)$  ( $= \widehat{\boldsymbol{\Sigma}}_{1n}(\mathbb{X}_n^*)$ ) in the diagonal in increasing order, and let  $\mathbf{P}$  contain the corresponding eigenvectors of  $\boldsymbol{\Sigma}_1(\mathbb{Y}_n^*)$  ( $= \boldsymbol{\Sigma}_1(\mathbb{X}_n^*)$ ). Now in Step 11 we find the number  $t$  of eigenvalues whose values are greater than  $10^{-6}$ . Note that the number  $t$  does not change when we use

$\mathbb{Y}_n^*$  instead of  $\mathbb{X}_n^*$ . In Step 12, we calculate,

$$\begin{aligned}\tilde{\mathbf{V}}_i(\mathbf{Y}_i, \mathbb{Y}_n^*) &= \mathbf{P}' \widehat{\boldsymbol{\eta}}_{in}(\mathbf{Y}_i^*, \mathbb{Y}_n^*) \\ &= \text{sgn}(k) \mathbf{P}' \widehat{\boldsymbol{\eta}}_{in}(\mathbf{X}_i^*, \mathbb{X}_n^*) \\ &= \text{sgn}(k) \tilde{\mathbf{V}}_i(\mathbf{X}_i, \mathbb{X}_n), \quad 1 \leq i \leq n.\end{aligned}$$

Finally, in Step 13, we calculate the  $t \times t$  vector

$$\begin{aligned}\mathbf{V}_i(\mathbf{Y}_i^*, \mathbb{Y}_n^*) &= \text{vector containing the 1st } t \text{ components of } \tilde{\mathbf{V}}_i(\mathbf{Y}_i, \mathbb{Y}_n^*) \\ &= \text{vector containing the 1st } t \text{ components of } \text{sgn}(k) \tilde{\mathbf{V}}_i(\mathbf{X}_i, \mathbb{X}_n^*) \\ &= \text{sgn}(k) \tilde{\mathbf{V}}_i(\mathbf{X}_i^*, \mathbb{X}_n^*), \quad 1 \leq i \leq n.\end{aligned}\tag{10}$$

We conclude the proof with the observation that, clearly, the  $\boldsymbol{\Lambda}$  matrix does not change since  $\boldsymbol{\Sigma}_{1n}(\mathbb{Y}_n^*) = \boldsymbol{\Sigma}_{1n}(\mathbb{X}_n^*)$  does not.  $\square$

Next we show that the population covariance matrix for  $\mathbf{V}$  is the diagonal matrix  $\boldsymbol{\Lambda}_{t \times t}$ .

**Lemma 8** *Let  $F_{\mathbf{X}}$  be the distribution function of  $\mathbf{X}$  and  $\tilde{\boldsymbol{\eta}}$  be defined as earlier. Further assume that  $\text{cov}(\tilde{\boldsymbol{\eta}}(\mathbf{X}, F_{\mathbf{X}})) = \boldsymbol{\Sigma}_{s \times s}$ , and  $\boldsymbol{\Lambda}_{s \times s} = \text{diag}(\lambda_1 \geq \dots \geq \lambda_s)$  with  $\lambda_1 \geq \dots \geq \lambda_s$  the eigenvalues of  $\boldsymbol{\Sigma}_{s \times s}$ . Let the orthogonal matrix  $\mathbf{P}_{s \times s} = [\mathbf{p}_1 \cdots \mathbf{p}_s]$  contain the corresponding eigenvectors. Define*

$$\tilde{\mathbf{V}}(\mathbf{X}, F_{\mathbf{X}}) = \mathbf{P}' \tilde{\boldsymbol{\eta}}(\mathbf{X}, F_{\mathbf{X}})$$

*and let  $t \leq s$  be the number of positive eigenvalues of  $\boldsymbol{\Sigma}$ . Let  $\mathbf{V}_{t \times 1}$  be the  $t$ -vector taking only the first  $t$  components of  $\tilde{\mathbf{V}}(\mathbf{X}, F_{\mathbf{X}})$ . Then*

$$\text{cov}(\mathbf{V}) = \text{diag}(\lambda_1 \geq \dots \geq \lambda_t).\tag{11}$$

**PROOF OF LEMMA 8.** Note that  $\boldsymbol{\Sigma} \mathbf{p}_i = \lambda_i \mathbf{p}_i$ ,  $1 \leq i \leq s$ , which implies  $\boldsymbol{\Sigma} \mathbf{P} = \mathbf{P} \boldsymbol{\Lambda}$ . Therefore,  $\mathbf{P}' \boldsymbol{\Sigma} \mathbf{P} = \boldsymbol{\Lambda}$  and hence  $\text{cov}(\tilde{\mathbf{V}}) = \mathbf{P}' \boldsymbol{\Sigma} \mathbf{P} = \boldsymbol{\Lambda}$ . Now  $\mathbf{V}_{t \times 1}$  contains the first  $t$  components of  $\tilde{\mathbf{V}}$ , and hence the covariance matrix for  $\tilde{\mathbf{V}}$  is the diagonal matrix containing first  $t$  eigenvalues, i.e.,  $\boldsymbol{\Lambda}_{t \times t} = \text{diag}(\lambda_1 \geq \dots \geq \lambda_t)$ .  $\square$

## 5 A Quadratic Form Outlyingness Function using $\mathbb{V}_n$

Here we construct a quadratic form outlyingness function based on the vector  $\mathbf{V}_{t \times 1}$ . It has been shown that the covariance matrix of  $\mathbf{V}$  is  $\boldsymbol{\Lambda}_{t \times t}$ . Therefore, our new quadratic form outlyingness function based on the reduced projected vector  $\mathbf{V}$  is given by

$$O_{\text{MDP}}(\mathbf{x}, F_{\mathbf{X}}) = \mathbf{V}(\mathbf{x}, F_{\mathbf{X}})' \boldsymbol{\Lambda}^{-1}(F_{\mathbf{X}}) \mathbf{V}(\mathbf{x}, F_{\mathbf{X}}).$$

The corresponding sample version is given by

$$O_{\text{MDP}}(\mathbf{x}, \mathbb{X}_n) = \mathbf{V}(\mathbf{x}, \mathbb{X}_n)' \boldsymbol{\Lambda}^{-1}(\mathbb{X}_n) \mathbf{V}(\mathbf{x}, \mathbb{X}_n),$$

where  $\mathbf{V}(\mathbf{x}, \mathbb{X}_n)$  and  $\mathbf{\Lambda}^{-1}(\mathbb{X}_n)$  are obtained following **Algorithm A**. This ‘‘Mahalanobis distance outlyingness function on reduced projected vectors’’ is denoted by **MDP**. Depending on which set of inner observations we use to find the sample SICS functional and the covariance matrix  $\mathbf{\Sigma}_1$  (in Step 9), i.e., indexed by  $\mathbb{J}_{\text{MCD}}$  or  $\mathbb{J}_{\text{MS}_{\text{DT}}}$ , we call our quadratic outlyingness functions **MDP(MCD)** or **MDP(DT)**, respectively.

**Lemma 9**  $O_{\text{MDP}}(\mathbf{x}, \mathbb{X}_n)$  is affine invariant, in the sense that if we transform  $\mathbf{X}_i \rightarrow \mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \mathbf{b}$ , then

$$O_{\text{MDP}}(\mathbf{y}, \mathbb{Y}_n) = O_{\text{MDP}}(\mathbf{x}, \mathbb{X}_n),$$

where  $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ .

**PROOF OF LEMMA 9.** Since  $\mathbf{V}(\mathbf{x}, \mathbb{X}_n)$  and  $\mathbf{\Lambda}(\mathbb{X}_n)$  are affine invariant (Lemma 7),  $O_{\text{MDP}}(\mathbf{x}, \mathbb{X}_n) = \mathbf{V}(\mathbf{x}, \mathbb{X}_n)' \mathbf{\Lambda}^{-1}(\mathbb{X}_n) \mathbf{V}(\mathbf{x}, \mathbb{X}_n)$  is also affine invariant.  $\square$

Our next result gives a distributional property of  $O_{\text{MDP}}(\mathbf{X}, F_{\mathbf{X}})$ .

**Lemma 10** Suppose  $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Then

$$O_{\text{MDP}}(\mathbf{X}, N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \mathbf{V}(\mathbf{X}, N(\boldsymbol{\mu}, \boldsymbol{\Sigma}))' \mathbf{\Lambda}^{-1}(N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \mathbf{V}(\mathbf{X}, N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) \sim \chi_t^2, \quad (12)$$

i.e.,  $O_{\text{MDP}}(\mathbf{X}, N(\boldsymbol{\mu}, \boldsymbol{\Sigma}))$  follows a chi-square distribution with  $t$  degrees of freedom.

To prove the above lemma, we first introduce the following basic lemma and its corollary.

**Lemma 11** Let  $\mathbf{X}_{d \times 1}$  have a distribution  $F_{\mathbf{X}}$  which is halfspace symmetric (Zuo and Serfling, 2000a) about a point  $\boldsymbol{\theta}$ . Then

$$\tilde{\boldsymbol{\eta}}(\mathbf{X}^*, \boldsymbol{\Delta}, F_{\mathbf{X}^*}) = \mathbf{B}(\boldsymbol{\Delta}_1, F_{\mathbf{X}}) (\mathbf{X} - \boldsymbol{\theta}),$$

where  $\mathbf{X}^* = \mathbf{D}(F_{\mathbf{X}})\mathbf{X}$ ,  $\boldsymbol{\Delta}_1 = [\mathbf{v}_1 \cdots \mathbf{v}_s]$ , and

$$\mathbf{B}(\boldsymbol{\Delta}_1, F_{\mathbf{X}})_{s \times d} = \left[ \frac{\mathbf{v}_1}{\text{med}|\mathbf{v}_1'(\mathbf{X} - \boldsymbol{\theta})|} \cdots \frac{\mathbf{v}_s}{\text{med}|\mathbf{v}_s'(\mathbf{X} - \boldsymbol{\theta})|} \right]',$$

with  $\mathbf{D}(F_{\mathbf{X}})$  a SICS functional and  $\mathbf{v}_i' = \frac{\mathbf{u}_i' \mathbf{D}(F_{\mathbf{X}})}{\|\mathbf{u}_i' \mathbf{D}(F_{\mathbf{X}})\|}$ .

**PROOF OF LEMMA 11.** With  $\mathbf{v}_i' = \frac{\mathbf{u}_i' \mathbf{D}(F_{\mathbf{X}})}{\|\mathbf{u}_i' \mathbf{D}(F_{\mathbf{X}})\|}$ , note that, for  $1 \leq i \leq s$ ,

$$\begin{aligned} \text{MAD}(\mathbf{u}_i' \mathbf{X}^*) &= \text{med}|\mathbf{u}_i' \mathbf{X}^* - \text{med}(\mathbf{u}_i' \mathbf{X}^*)| \\ &= \text{med}(\|\mathbf{u}_i' \mathbf{D}(F_{\mathbf{X}})\| |\mathbf{v}_i' \mathbf{X} - \mathbf{v}_i' \boldsymbol{\theta}|) \\ &= \|\mathbf{u}_i' \mathbf{D}(F_{\mathbf{X}})\| \text{med}(|\mathbf{v}_i' \mathbf{X} - \mathbf{v}_i' \boldsymbol{\theta}|). \end{aligned}$$

Therefore, for  $1 \leq i \leq s$ , we have

$$\begin{aligned} \frac{\mathbf{u}'_i \mathbf{X}^* - \text{med}(\mathbf{u}'_i \mathbf{X}^*)}{\text{MAD}(\mathbf{u}'_i \mathbf{X}^*)} &= \frac{(\mathbf{v}'_i \mathbf{X} - \mathbf{v}'_i \boldsymbol{\theta}) \|\mathbf{u}'_i \mathbf{D}(F_{\mathbf{X}})\|}{\text{med}(|\mathbf{v}'_i (\mathbf{X} - \boldsymbol{\theta})|) \|\mathbf{u}'_i \mathbf{D}(F_{\mathbf{X}})\|} \\ &= \frac{\mathbf{v}_i}{\text{med}(|\mathbf{v}'_i (\mathbf{X} - \boldsymbol{\theta})|)} (\mathbf{X} - \boldsymbol{\theta}). \end{aligned}$$

Hence the result follows.  $\square$

**Corollary 12** *Assume that  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Then*

$$\Phi^{-1}(3/4) \tilde{\boldsymbol{\eta}}(\mathbf{X}^*, F_{\mathbf{X}}) \sim N_s \left( \mathbf{0}, \boldsymbol{\Sigma}_0 \left( (\mathbf{D}(F_{\mathbf{X}}) \boldsymbol{\Sigma} \mathbf{D}'(F_{\mathbf{X}}))^{1/2} \right) \right),$$

where  $\boldsymbol{\Sigma}_0 \left( (\mathbf{D}(F_{\mathbf{X}}) \boldsymbol{\Sigma} \mathbf{D}'(F_{\mathbf{X}}))^{1/2} \right) = (\sigma_{ij}^0 \left( (\mathbf{D}(F_{\mathbf{X}}) \boldsymbol{\Sigma} \mathbf{D}'(F_{\mathbf{X}}))^{1/2} \right))_{s \times s}$ , with

$$\begin{aligned} \sigma_{ij}^0 \left( (\mathbf{D}(F_{\mathbf{X}}) \boldsymbol{\Sigma} \mathbf{D}'(F_{\mathbf{X}}))^{1/2} \right) &= \frac{\mathbf{u}'_i \mathbf{D}(F_{\mathbf{X}}) \boldsymbol{\Sigma} \mathbf{D}'(F_{\mathbf{X}}) \mathbf{u}_j}{\sqrt{\mathbf{u}'_i \mathbf{D}(F_{\mathbf{X}}) \boldsymbol{\Sigma} \mathbf{D}'(F_{\mathbf{X}}) \mathbf{u}_i \mathbf{u}'_j \mathbf{D}(F_{\mathbf{X}}) \boldsymbol{\Sigma} \mathbf{D}'(F_{\mathbf{X}}) \mathbf{u}_j}} \\ &= \mathbf{S}(\boldsymbol{\Sigma}^{1/2} \mathbf{D}'(F_{\mathbf{X}}) \mathbf{u}_i)' \mathbf{S}(\boldsymbol{\Sigma}^{1/2} \mathbf{D}'(F_{\mathbf{X}}) \mathbf{u}_j). \end{aligned}$$

PROOF OF COROLLARY 12. Note that here  $\mathbf{X}$  is halfspace symmetric about  $\boldsymbol{\mu}$  in the sense of Zuo and Serfling (2000a). Moreover, we have

$$\text{MAD}(\mathbf{v}'_i \mathbf{X}) = \Phi^{-1}(3/4) \sqrt{\mathbf{v}'_i \boldsymbol{\Sigma} \mathbf{v}_i}, \quad 1 \leq i \leq s.$$

Therefore, using Lemma 11 we have

$$\tilde{\boldsymbol{\eta}}(\mathbf{X}, \boldsymbol{\Delta}_1, N(\boldsymbol{\mu}, \boldsymbol{\Sigma}))_{s \times 1} = \mathbf{B}(\boldsymbol{\Delta}_1, N(\boldsymbol{\mu}, \boldsymbol{\Sigma}))(\mathbf{X} - \boldsymbol{\mu}),$$

where

$$\mathbf{B}(\boldsymbol{\Delta}_1, N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \left[ \frac{\mathbf{v}_1}{\Phi^{-1}(3/4) \sqrt{\mathbf{v}'_1 \boldsymbol{\Sigma} \mathbf{v}_1}} \cdots \frac{\mathbf{v}_s}{\Phi^{-1}(3/4) \sqrt{\mathbf{v}'_s \boldsymbol{\Sigma} \mathbf{v}_s}} \right]'$$

Hence the result follows.  $\square$

PROOF OF LEMMA 10. The proof follows directly using Corollary 12 and Lemma 8.  $\square$

## 6 Robust TR Outlyingness using $\mathbb{V}_n$

We develop two robust sample Mahalanobis spatial outlyingness functions based on the vectors  $\mathbf{V}_i$ ,  $1 \leq i \leq n$ . The key idea is to implement spatial trimming on the reduced projected vectors  $\mathbf{V}_i$ . This yields two robust transformation-retransformation (TR) outlyingness functions based on whether we are using the DT approach (**RTRP(DT)**) or the MCD approach

(RTRP(MCD)) for standardization prior to forming the spatial outlyingness function. Below we provide extensive algorithms for RTRP(DT) and RTRP(MCD).

**Algorithm B, for RTRP(DT)**

After transforming the given data  $\mathbb{X}_n = [\mathbf{X}_1 \cdots \mathbf{X}_n]_{d \times n}$  to  $\mathbb{V}_n = [\mathbf{V}_1 \cdots \mathbf{V}_n]_{t \times n}$ , using  $\mathbb{J}_{\text{MS}_{\text{DT}}}$ , we follow the steps below to form RTRP(DT).

1. Let  $\mathbf{M}_{T_{t \times t}}^{\mathbf{V}}$  be the Tyler TR functional for the data matrix  $\mathbb{V}_{t \times n}$ .
2. Calculate  $\mathbf{V}_i^* = \mathbf{M}_{T_{t \times t}}^{\mathbf{V}} \mathbf{V}_i$ ,  $1 \leq i \leq n$ .
3. For each  $i = 1, \dots, n$ , calculate  $\mathbf{S}_{t \times n}^i = [\mathbf{S}_1^i \cdots \mathbf{S}_n^i]$ , where

$$\mathbf{S}_j^i = \mathbf{S}(\mathbf{V}_i^* - \mathbf{V}_j^*), \quad 1 \leq j \leq n,$$

and  $\mathbf{S}(\cdot)$  is the sign function.

4. For each  $i = 1, \dots, n$ , calculate

$$O_{\text{MSP}}(i) = \sqrt{\left( \frac{1}{n} \sum_{j=1}^n \mathbf{S}_j^i \right)' \left( \frac{1}{n} \sum_{j=1}^n \mathbf{S}_j^i \right)}.$$

5. Form the set of indices  $\mathbb{J}_{\text{MS}_T}^{\mathbf{V}}$  such that

$$O_{\text{MSP}}(j) \leq \frac{t}{t+2}, \quad \forall j \in \mathbb{J}_{\text{MS}_T}^{\mathbf{V}}.$$

6. Let  $K_1 = |\mathbb{J}_{\text{MS}_T}^{\mathbf{V}}|$ , where  $|\cdot|$  denotes the cardinality of a set.
7. Form a matrix  $\mathbb{W}_{t \times K_1} = [\mathbf{W}_1 \cdots \mathbf{W}_{K_1}]$  of dimension  $t \times K_1$  such that

$$\mathbf{W}_i = \mathbf{V} \left( \mathbb{J}_{\text{MS}_T}^{\mathbf{V}}(i) \right), \quad 1 \leq i \leq K_1,$$

where  $\mathbb{J}_{\text{MS}_T}^{\mathbf{V}}(i)$  denotes the  $i$ th element of  $\mathbb{J}_{\text{MS}_T}^{\mathbf{V}}$ .

8. Put  $\overline{\mathbf{W}}_{t \times 1} = \frac{1}{K_1} \sum_{i=1}^{K_1} \mathbf{W}_i$  and

$$\Sigma_{2_{t \times t}} = \left[ \frac{1}{K_1} \sum_{i=1}^{K_1} (\mathbf{W}_i - \overline{\mathbf{W}})(\mathbf{W}_i - \overline{\mathbf{W}})' \right],$$

the sample covariance matrix of the  $\mathbf{W}_i$ ,  $1 \leq i \leq K_1$ .

9. Calculate the eigenvalues  $\gamma_1, \dots, \gamma_t$ , and the corresponding orthonormal eigenvectors  $\mathbf{q}_1, \dots, \mathbf{q}_t$  of  $\Sigma_2$ . Put  $\mathbf{\Gamma}_{t \times t} = \text{diag}(\gamma_1, \dots, \gamma_t)$  and  $\mathbf{Q}_{t \times t} = [\mathbf{q}_1 \cdots \mathbf{q}_t]$ .

10. Calculate the  $t \times t$  matrix  $\Sigma_2^{-1/2} = \mathbf{Q}\mathbf{\Gamma}^{-1/2}\mathbf{Q}'$ , where  $\mathbf{\Gamma}^{-1/2} = \text{diag}(1/\gamma_1^{1/2}, \dots, 1/\gamma_t^{1/2})$ .
11. Calculate  $\mathbf{W}_i^* = \Sigma_2^{-1/2}\mathbf{V}_i$ ,  $1 \leq i \leq n$ .
12. For each  $i = 1, \dots, n$ , calculate  $\tilde{\mathbf{S}}_{t \times n}^i = [\tilde{\mathbf{S}}_1^i \cdots \tilde{\mathbf{S}}_n^i]$  where

$$\tilde{\mathbf{S}}_j^i = \mathbf{S}(\mathbf{W}_i^* - \mathbf{W}_j^*), \quad 1 \leq j \leq n.$$

13. For each  $i = 1, \dots, n$ , calculate

$$O_{\text{MSPR}}(i) = \sqrt{\left( \frac{1}{K_1} \sum_{j \in \mathbb{J}_{\text{MS}_T}^{\mathbf{V}}} \tilde{\mathbf{S}}_j^i \right)' \left( \frac{1}{K_1} \sum_{j \in \mathbb{J}_{\text{MS}_T}^{\mathbf{V}}} \tilde{\mathbf{S}}_j^i \right)}.$$

### Algorithm C, for RTRP(MCD)

Here we give the algorithm for RTRP(MCD), after transforming the data  $\mathbb{X}_n = [\mathbf{X}_1 \cdots \mathbf{X}_n]_{d \times n}$  to  $\mathbb{V}_n = [\mathbf{V}_1 \cdots \mathbf{V}_n]_{t \times n}$  using the MCD set of observations,  $\mathbb{J}_{\text{MCD}}$ .

1. Let  $\mathbb{J}_{\text{MCD}_V}$  contain the indices of the MCD observations and  $\Sigma_{\text{MCD}_V}$  be the MCD covariance matrix of dimension  $t \times t$  obtained based on  $\mathbb{V}_{t \times n}$  with  $\alpha_{\text{MCD}} = 0.5$ .
2. Calculate the eigenvalues  $\gamma_1, \dots, \gamma_t$ , and the corresponding orthonormal eigenvectors  $\mathbf{q}_1, \dots, \mathbf{q}_t$  of  $\Sigma_{\text{MCD}_V}$ . Put  $\mathbf{\Gamma}_{t \times t} = \text{diag}(\gamma_1, \dots, \gamma_t)$  and  $\mathbf{Q}_{t \times t} = [\mathbf{q}_1 \cdots \mathbf{q}_t]$ .
3. Calculate the  $t \times t$  matrix  $\Sigma_{\text{MCD}_V}^{-1/2} = \mathbf{Q}\mathbf{\Gamma}^{-1/2}\mathbf{Q}'$ , where  $\mathbf{\Gamma}^{-1/2} = \text{diag}(1/\gamma_1^{1/2}, \dots, 1/\gamma_t^{1/2})$ .
4. Calculate  $\mathbf{V}_i^* = \Sigma_{\text{MCD}_V}^{-1/2}\mathbf{V}_i$ ,  $1 \leq i \leq n$ .
5. For each  $i = 1, \dots, n$ , calculate  $\mathbf{S}_{t \times n}^i = [\mathbf{S}_1^i \cdots \mathbf{S}_n^i]$ , where

$$\mathbf{S}_j^i = \mathbf{S}(\mathbf{V}_i^* - \mathbf{V}_j^*), \quad 1 \leq j \leq n.$$

6. For each  $i = 1, \dots, n$ , calculate

$$O_{\text{MSP}}(i) = \sqrt{\left( \frac{1}{n} \sum_{j=1}^n \mathbf{S}_j^i \right)' \left( \frac{1}{n} \sum_{j=1}^n \mathbf{S}_j^i \right)}.$$

7. Form the set of indices  $\mathbb{J}_{\text{MS}_{\text{MCD}}}^{\mathbf{V}}$  such that

$$O_{\text{MSP}}(j) \leq \frac{t}{t+2}, \quad \forall j \in \mathbb{J}_{\text{MS}_{\text{MCD}}}^{\mathbf{V}}.$$

8. Let  $K_1 = |\mathbb{J}_{\text{MSMCD}}^{\mathbf{V}}|$ , where  $|\cdot|$  denotes the cardinality of a set.
9. Form a matrix  $\mathbb{W}_{t \times K_1} = [\mathbf{W}_1 \cdots \mathbf{W}_{K_1}]$  of dimension  $t \times K_1$  such that

$$\mathbf{W}_i = \mathbf{V} \left( \mathbb{J}_{\text{MSMCD}}^{\mathbf{V}}(i) \right), \quad 1 \leq i \leq K_1,$$

where  $\mathbb{J}_{\text{MSMCD}}^{\mathbf{V}}(i)$  denotes the  $i$ th element of  $\mathbb{J}_{\text{MSMCD}}^{\mathbf{V}}$ .

10. Put  $\overline{\mathbf{W}}_{t \times 1} = \frac{1}{K_1} \sum_{i=1}^{K_1} \mathbf{W}_i$  and

$$\Sigma_{2t \times 2t} = \left[ \frac{1}{K_1} \sum_{i=1}^{K_1} (\mathbf{W}_i - \overline{\mathbf{W}})(\mathbf{W}_i - \overline{\mathbf{W}})' \right],$$

the sample covariance matrix of the  $\mathbf{W}_i$ ,  $1 \leq i \leq K_1$ .

11. Calculate the eigenvalues  $\gamma_1, \dots, \gamma_t$ , and the corresponding orthonormal eigenvectors  $\mathbf{q}_1, \dots, \mathbf{q}_t$  of  $\Sigma_2$ . Put  $\mathbf{\Gamma}_{t \times t} = \text{diag}(\gamma_1, \dots, \gamma_t)$   $\mathbf{Q}_{t \times t} = [\mathbf{q}_1 \cdots \mathbf{q}_t]$ .
12. Calculate the  $t \times t$  matrix  $\Sigma_2^{-1/2} = \mathbf{Q}\mathbf{\Gamma}^{-1/2}\mathbf{Q}'$ , where  $\mathbf{\Gamma}^{-1/2} = \text{diag}(1/\gamma_1^{1/2}, \dots, 1/\gamma_t^{1/2})$ .
13. Calculate

$$\mathbf{W}_i^* = \Sigma_2^{-1/2} \mathbf{V}_i, \quad 1 \leq i \leq n.$$

14. For each  $i = 1, \dots, n$ , calculate  $\tilde{\mathbf{S}}_{t \times n}^i = [\tilde{\mathbf{S}}_1^i \cdots \tilde{\mathbf{S}}_n^i]$  where

$$\tilde{\mathbf{S}}_j^i = \mathbf{S}(\mathbf{W}_i^* - \mathbf{W}_j^*), \quad 1 \leq j \leq n.$$

15. For each  $i = 1, \dots, n$ , calculate

$$O_{\text{MSPR}}(i) = \sqrt{\left( \frac{1}{K_1} \sum_{j \in \mathbb{J}_{\text{MSMCD}}^{\mathbf{V}}} \tilde{\mathbf{S}}_j^i \right)' \left( \frac{1}{K_1} \sum_{j \in \mathbb{J}_{\text{MSMCD}}^{\mathbf{V}}} \tilde{\mathbf{S}}_j^i \right)}.$$

Next we show that the sample RTRP outlyingness function is affine invariant.

**Lemma 13**  $O_{\text{RTRP}}(\mathbf{x}, \mathbb{X}_n)$  is affine invariant, in the sense that if we transform  $\mathbf{X}_i \rightarrow \mathbf{Y}_i = \mathbf{A}\mathbf{X}_i + \mathbf{b}$ , where  $\mathbf{A}_{d \times d}$  is nonsingular and  $\mathbf{b}$  is any vector in  $\mathbb{R}^d$ , then

$$O_{\text{RTRP}}(\mathbf{y}, \mathbb{Y}_n) = O_{\text{RTRP}}(\mathbf{x}, \mathbb{X}_n),$$

with  $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ .

PROOF OF LEMMA 13. Using Lemma 7 we get

$$\mathbb{V}_n(\mathbb{Y}_n) = \text{sgn}(k)\mathbb{V}_n(\mathbb{X}_n),$$

where  $k = k(\mathbf{A}, \mathbf{b}, \mathbb{X}_n)$ . Now using the affine invariance property of  $O_S^{(\text{TR})}$  the index set  $\mathbb{J}^{\mathbf{V}}$  is affine invariant, and this implies that  $O_{\text{RTRP}}$  is affine invariant.  $\square$

## 7 Comparison of MDP, RTRP, SUP, and MD

### 7.1 Artificial Data

Using artificially created *bivariate* data sets, we provide *visual illustrations* of important differences among SUP, MDP(DT), MDP(MCD), RTRP(DT), and RTRP(MCD). Besides identifying the more extreme outliers, an outlyingness function also has the role of providing a *structural description* of the data set. In effect, this provides a *quantile-based description*. Our plots exhibit 50%, 75%, and 90% outlyingness contours (based on the given outlyingness function and enclosing 50%, 75%, and 90% of the observations, respectively). The minimum outlyingness point, which represents a notion of “median” or “center”, is indicated by an asterisk in the plots. We make the following points, illustrated with Figures 1-8.

1. *The new projection outlyingness functions are better than SUP.* It is seen that SUP is dominated by MDP(DT), RTRP(MCD), and RTRP(DT) convincingly. In Figure 1, these and MDP(MCD) are compared with a triangular data set free of extreme outliers. It is seen that the contours of SUP are strange looking, the contours of MDP(MCD) are very circular, whereas the inner 50% and 75% contours of RTRP(MCD) and RTRP(DT) are describing the shape of the data better than others. Although the contours of MDP(DT) are ellipsoidal, still they are better oriented to the data than MDP(MCD). The difference is even more pronounced with the bivariate Pareto data (Figure 2). The 90% contour of SUP lies unnecessarily far apart from the data structure, whereas those of MDP(DT), RTRP(MCD), and RTRP(DT) describe the data structure better. Note that here also the 90% contour of MDP(MCD) is little apart from data set (however, it is better than SUP’s). In case of Normal data with no outliers (Figure 3), all these outlyingness functions describe the data reasonably well.
2. *The robustness properties of all five methods are comparable.* Figure 4 compares SUP, MDP(MCD), MDP(DT), RTRP(MCD), and RTRP(DT), for the triangular data with extreme outliers added, including a cluster. Here we see that all methods perform equally well in detecting outliers. Interestingly, the contours of all these methods are dragged a little bit by the outlier “C”. However, the performance of SUP for bivariate Pareto data with added extreme outliers (Figure 5) is relatively poor in comparison, in the sense that the inner 75% contour of SUP is unduely influenced by the outlier “C”. A similar kind of observation holds for MDP(MCD), where the 75% ellipse is little elongated by the presence of outlier “C”, however, less severely than for SUP. For the bivariate Normal example with replacement outliers, we exhibit two outlier scenarios, a cluster outlier structure, and a lower dimensional outlier structure, in Figures 6 and 7, respectively. From Figure 6, we observe that all five methods including SUP are performing equally well with the cluster outlier structure. In case of the lower dimensional outlier structure, i.e., in Figure 7, we observe that SUP, MDP(MCD) and RTRP(MCD) are more robust than MDP(DT) and RTRP(DT) with this particular kind of contamination. Overall, we can say that all these methods

are equivalent in terms of robustness except in some particular cases. However, our MDP(MCD), MDP(DT), RTRP(MCD), and RTRP(DT) are much less computational than SUP. Among these, MDP(DT) and RTRP(DT) are least computational, offering an equivalent level of robustness with much less computational burden.

For completeness, we apply MDP(DT) and RTRP(DT) on correlated bivariate Normal data structure, without and with added cluster outliers. Figure 8 shows that MDP(DT) and RTRP(DT) successfully describe the correlated nature of the data and detect the outliers.

## 7.2 Actual Data

For two well-studied higher-dimensional data sets, we examine the performance of MDP(DT) and RTRP(DT). They are seen to be closely competitive in robustness with other methods including MD, corroborating the findings obtained in the previous part with the artificial data sets.

### Stackloss Data

The stackloss data set of Brownlee (1965) has been much studied as a test of outlier detection methods. It consists of  $n = 21$  observations in dimension  $d = 4$ . See Rousseeuw and Leroy (1987) and Becker and Gather (1999) for the data set, for references to many studies, and for general discussion. In particular, the latter authors use robust Mahalanobis distance with S-estimates for location and covariance based on Tukey’s biweight (BW) function. We denote their outlyingness function as **MD(BW)**. All robust methods cited in the literature, plus MD(BW), MDP(DT) and RTRP(DT) agree on **1**, **3**, **4**, and **21** as the top 4 in outlyingness, with little difference in order. All of MD(BW), MDP(DT) and RTRP(DT) rank **2** as **5th**. Now MD(BW) and RTRP(DT) rank **13** as **6th** and **17** as **7th**, whereas MDP(DT) ranks them as **8th** and **6th** respectively. The differences show up in the region of *moderate outliers* between the extreme outlier region and the middle part of the data. This reflects how the outlyingness functions are *structurally descriptive of the data*. All these are equivalent in detecting extreme outliers, but our methods are more computationally attractive.

### Pollution and Mortality Data

Becker and Gather (1999) study in detail a 13-dimensional data set available at Data and Story Library (<http://lib.stat.cmu.edu/DASL/>) which provides information on social and economic conditions, climate, air pollution, and mortality for 60 Standard Metropolitan Statistical Areas (a standard U. S. Census Bureau designation of the region around a major city) in the United States. They omit a case with incomplete data and rank the remaining  $n = 59$  cases in dimension  $d = 13$  by MD(BW) as described above. Here we compare MDP(DT), RTRP(DT), and MD(BW). All agree on the extreme observations indexed **28**, **47**, **46**, **48** and ranked in this order. Also, they agree on 11 cases as among the next 12 cases, although with somewhat differing ranks. The exceptions are that MD(BW) ranks

observation **36** as **14**th and observation **38** as **22**nd, while MDP(DT) and RTRP(DT) ranks these as **17**th and **16**th, respectively. The difference in ranks 16 versus 22 for observation **38** raises the question of whether observation **36** (New Orleans) in comparison with observation **38** (Philadelphia) should rank far apart, 14th versus 22nd as per MD(BW), or closely, 17th versus 16th as per MDP(DT) or RTRP(DT). Coordinatewise dotplots of all 13 variables for these cases reveals that these points overall are not very outlying, except that **36** is moderate to extreme in outlyingness for mortality and moderate for SO<sub>2</sub> pollution, and that **38** is moderate to strong in outlyingness for population size, moderate for population density, and nearly moderate for SO<sub>2</sub> pollution. On this basis we regard **36** and **38** as comparable and moderate in outlyingness, corroborating the opinion of MDP(DT) and RTRP(DT) over that of MD(BW).

## 8 Conclusions and Recommendations

*Gaps and ranking* are used to detect outliers in dimension higher than 2. Looking at the rankings and gaps rather than pictures, we make the following basic conclusions about MDP and RTRP based on our simulation studies.

1. MDP and RTRP are both affine invariant, robust, and easily computable in all dimensions.
2. *Triangle*: MDP and RTRP are similar.
3. *Bivariate Pareto*: MDP and RTRP are similar.
4. *Bivariate Normal*: RTRP is better than MDP.
5. Overall, RTRP is closely followed by MDP.
6. The well-known projection outlyingness SUP is outclassed by MDP and RTRP.
7. If ellipsoidal contours are desired then one can simply use a robust version of MD. For more efficient contours, however, RTRP(DT) may be used in any practical dimension and offers reasonably high robustness combined with high computational efficiency.

## Acknowledgments

The authors gratefully acknowledge support under National Science Foundation Grants DMS-0103698 and DMS-0805786 and National Security Agency Grant H98230-08-1-0106.

## References

- [1] Becker, C. and Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association* **94** 947–955.
- [2] Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, 2nd edition. John Wiley & Sons, New York.
- [3] Dang, X. and Serfling, R. (2010). Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference* **140** 198–213.
- [4] Dümbgen, L. (1998). On Tyler’s M-functional of scatter in high dimension. *Annals of the Institute of Statistical Mathematics* **50** 471–491.
- [5] Fang, K.T. and Wang, Y. (1994). *Number Theoretic Methods in Statistics*. Chapman and Hall, London.
- [6] Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis* **52** 1694–1711.
- [7] Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–889.
- [8] Ghosh, J. K. (1971). A new proof of the Bahadur representation of sample quantiles and an application. *Annals of Mathematical Statistics* **42** 1957–1961.
- [9] Ilmonen, P., Oja, H., and Serfling, R. (2011). On invariant coordinate system (ICS) functionals. *International Statistical Review*, in press.
- [10] Kruskal, J. B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new ‘new index of consideration’. In *Statistical Computation*. (R. C. Milton and J. A. Nelder, eds.) Academic, New York.
- [11] Kruskal, J. B. (1972). Linear transformation to multivariate data to reveal clustering. In *Multidimensional scaling: Theory and Application in the Behavioral Science, I, Theory*. Seminar Press, New York and London.
- [12] Liu, R. Y. (1992). Data depth and multivariate rank tests. In *L<sub>1</sub>-Statistics and Related Methods* (Y. Dodge, ed.), pp. 279–294, North-Holland, Amsterdam.
- [13] Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester, England.
- [14] Mazumder, S. and Serfling, R. (2009). Bahadur Representations for the Median Absolute Deviation and Its Modifications. *Statistics and Probability Letters*, 2009, 79, 1774–1783.

- [15] Mazumder, S. and Serfling, R. (2012). A robust sample spatial outlyingness function. In revision for *Journal of Statistical Planning and Inference*.
- [16] Pan, J.-X., Fung, W.-K., and Fang, K.-T. (2000). Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference* **83** 153–167.
- [17] Peña, D. and Prieto, F. J. (2001). Robust covariance matrix estimation and multivariate outlier rejection. *Technometrics* **43** 286–310.
- [18] 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.
- [19] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.
- [20] Rousseeuw, P. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.
- [21] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.
- [22] Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference* **123** 259–278.
- [23] Serfling, R. (2010). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. *Journal of Nonparametric Statistics* **22** 915–936.
- [24] Switzer, P. (1970). Numerical Classification. In *Geostatistics*. Plenum, New York.
- [25] Switzer, P. and Wright, R. M. (1971). Numerical classification applied to certain Jamaican eocene nummulitids. *Math. Geol.* **3** 297-311.
- [26] Tyler, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Annals of Statistics* **15** 234–251.
- [27] Tyler, D. E., Critchley, F., Dümbgen, L. and Oja, H. (2009). Invariant co-ordinate selection. *Journal of the Royal Statistical Society, Series B* **71** 127.
- [28] Zuo, Y. and Serfling, R. (2000a). On the performance of some nonparametric location measures relative to a general notion of multivariate symmetry. *Journal of Statistical Planning and Inference* **84** 55–79.
- [29] Zuo, Y. and Serfling, R. (2000b). General notions of statistical depth function. *Annals of Statistics* **28** 461–482.

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

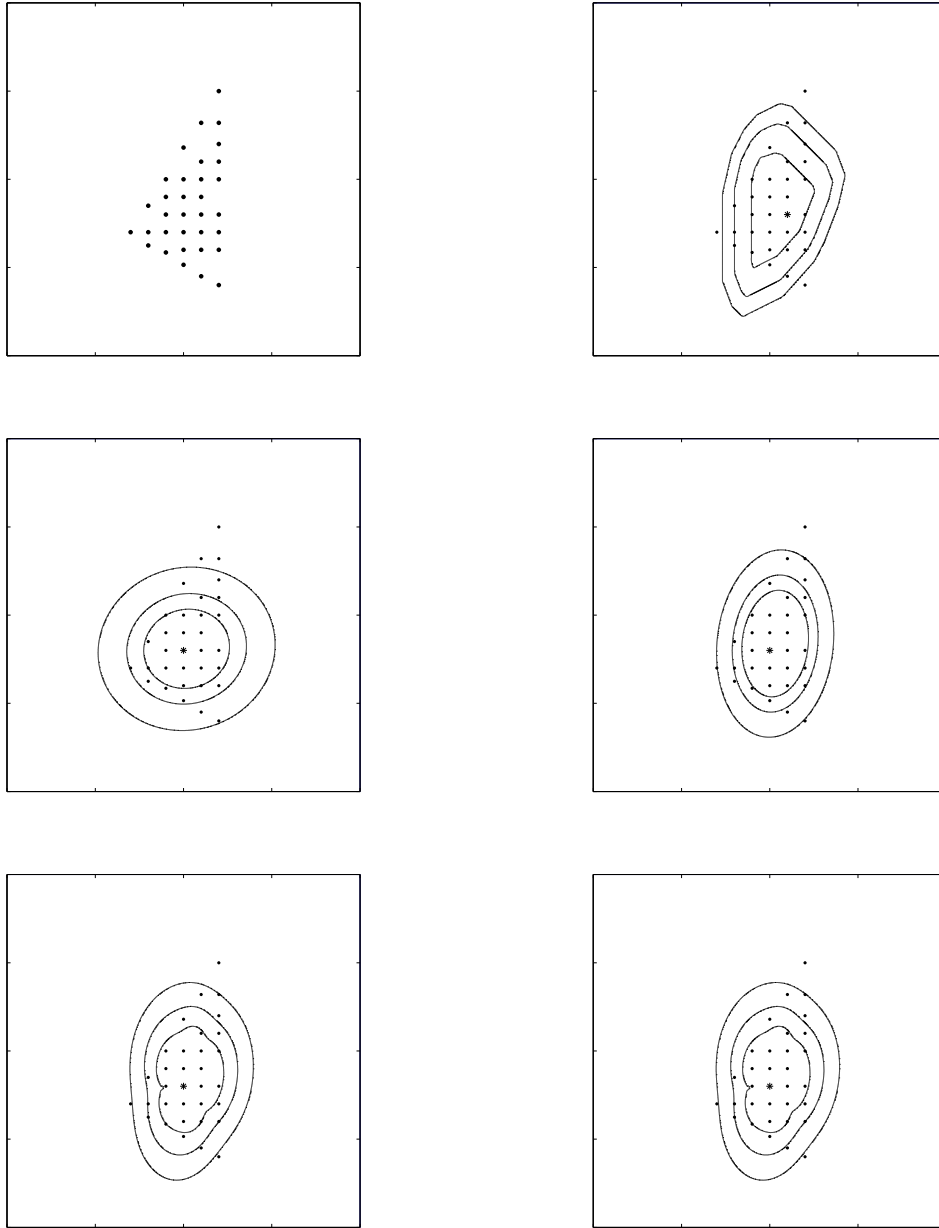


Figure 1: Triangular data set, and 50%, 75%, and 90% outlyingness contours for SUP (upper right), MDP(MCD) (middle left), MDP(DT) (middle right) RTRP(MCD) (lower left), and RTRP(DT) (lower right).

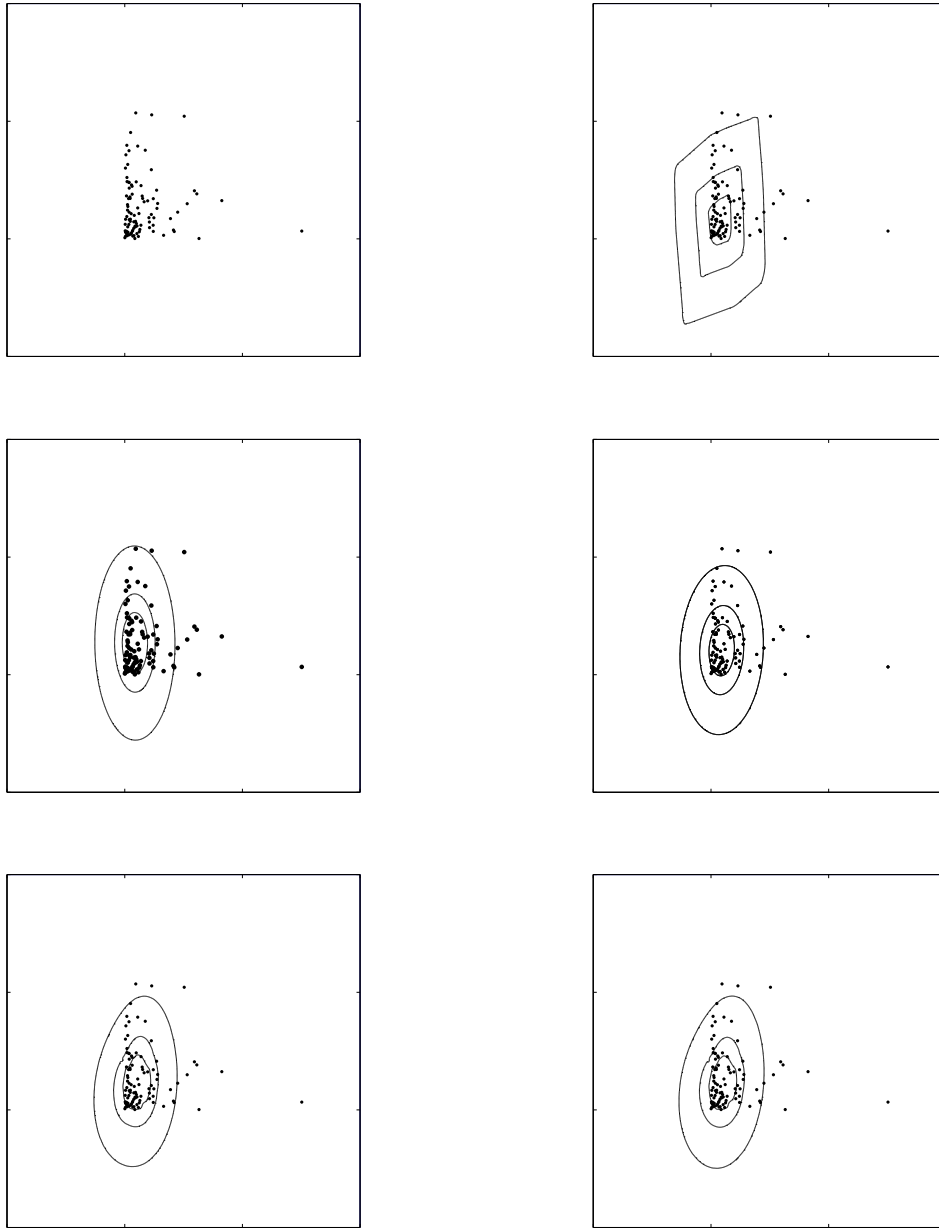


Figure 2: Bivariate Pareto data set, and 50%, 75%, and 90% outlyingness contours for SUP (upper right), MDP(MCD) (middle left), MDP(DT) (middle right) RTRP(MCD) (lower left), and RTRP(DT) (lower right).

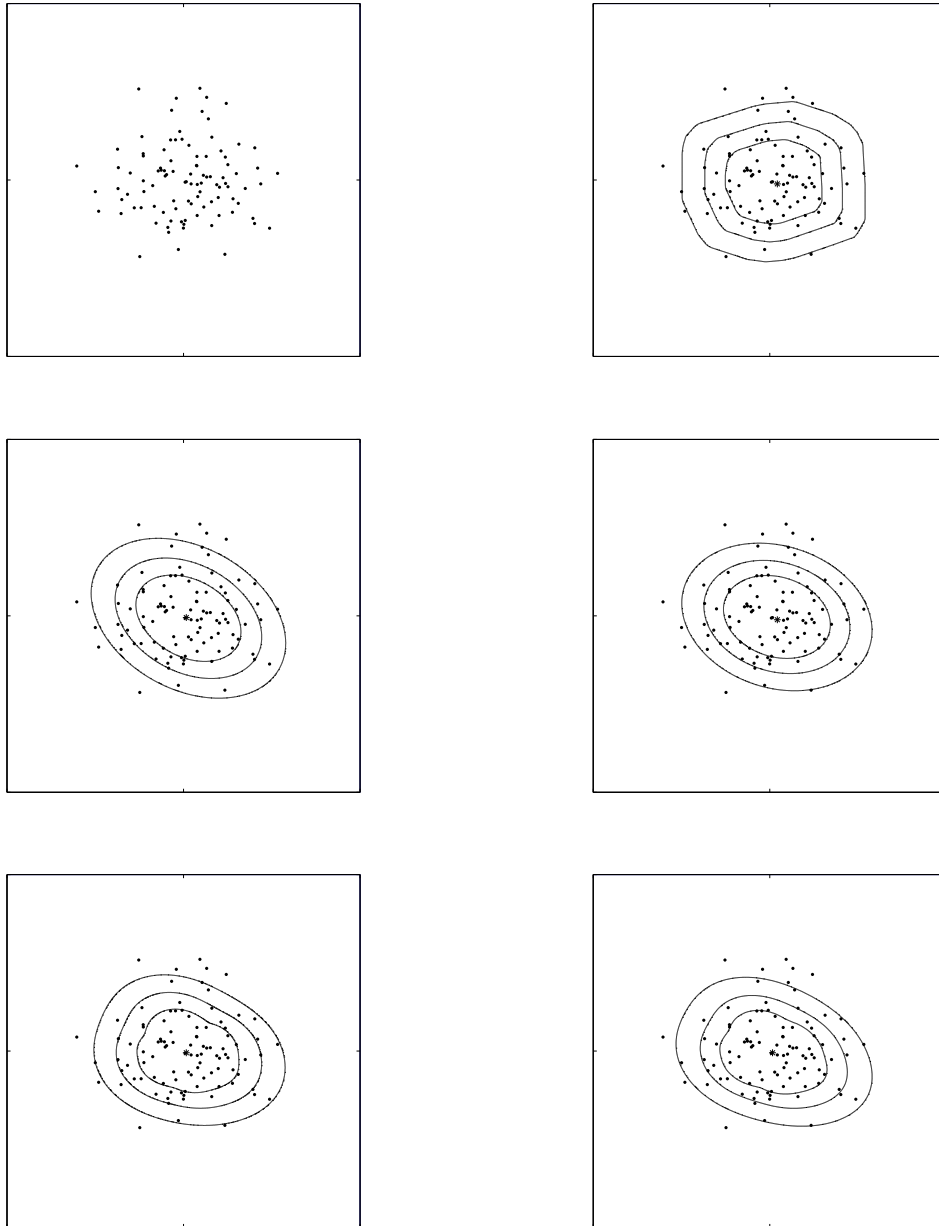


Figure 3: Bivariate Normal data set, and 50%, 75%, and 90% outlyingness contours for SUP (upper right), MDP(MCD) (middle left), MDP(DT) (middle right) RTRP(MCD) (lower left), and RTRP(DT) (lower right).

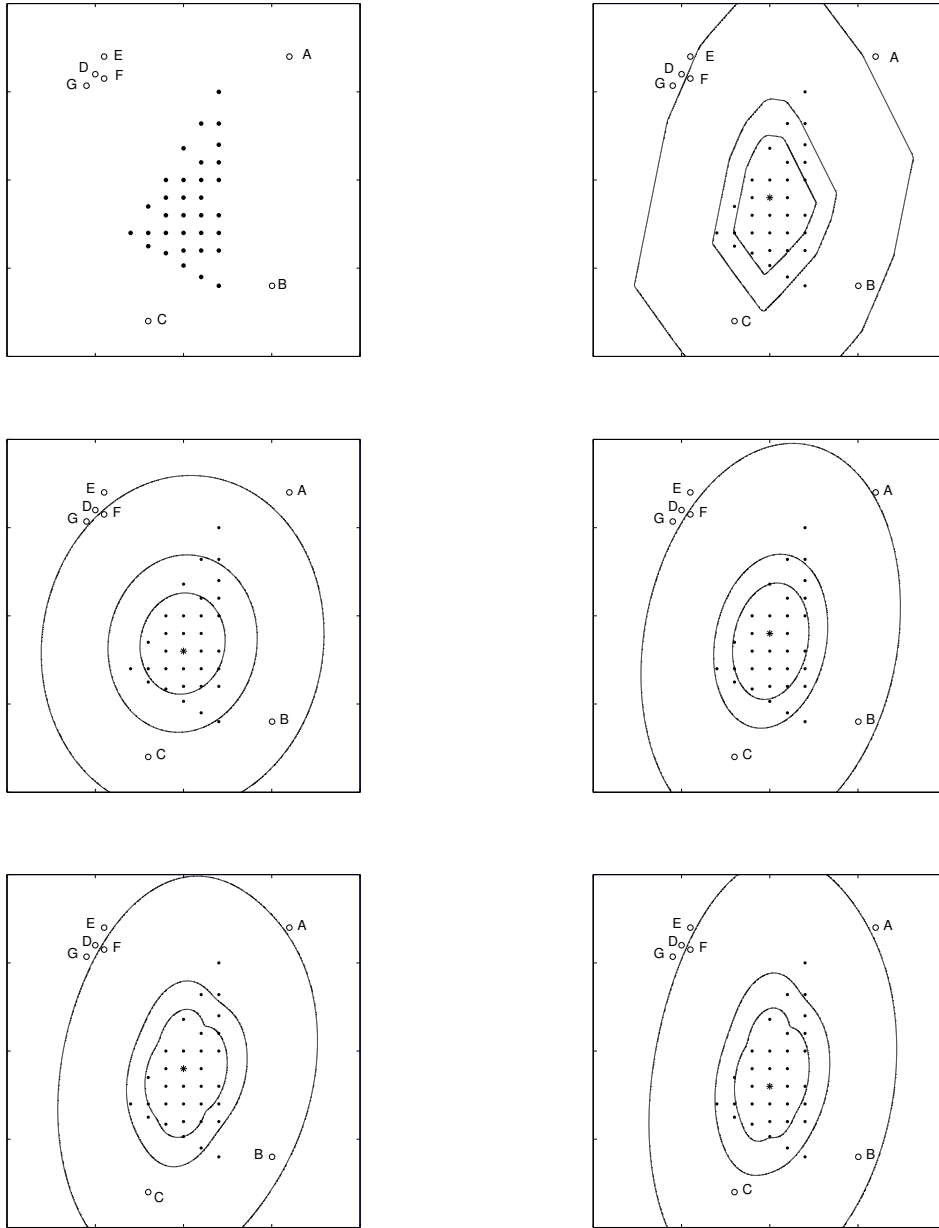


Figure 4: Triangular data set with extreme added outliers, and 50%, 75%, and 90% outlyingness contours for SUP (upper right), MDP(MCD) (middle left), MDP(DT) (middle right) RTRP(MCD) (lower left), and RTRP(DT) (lower right).

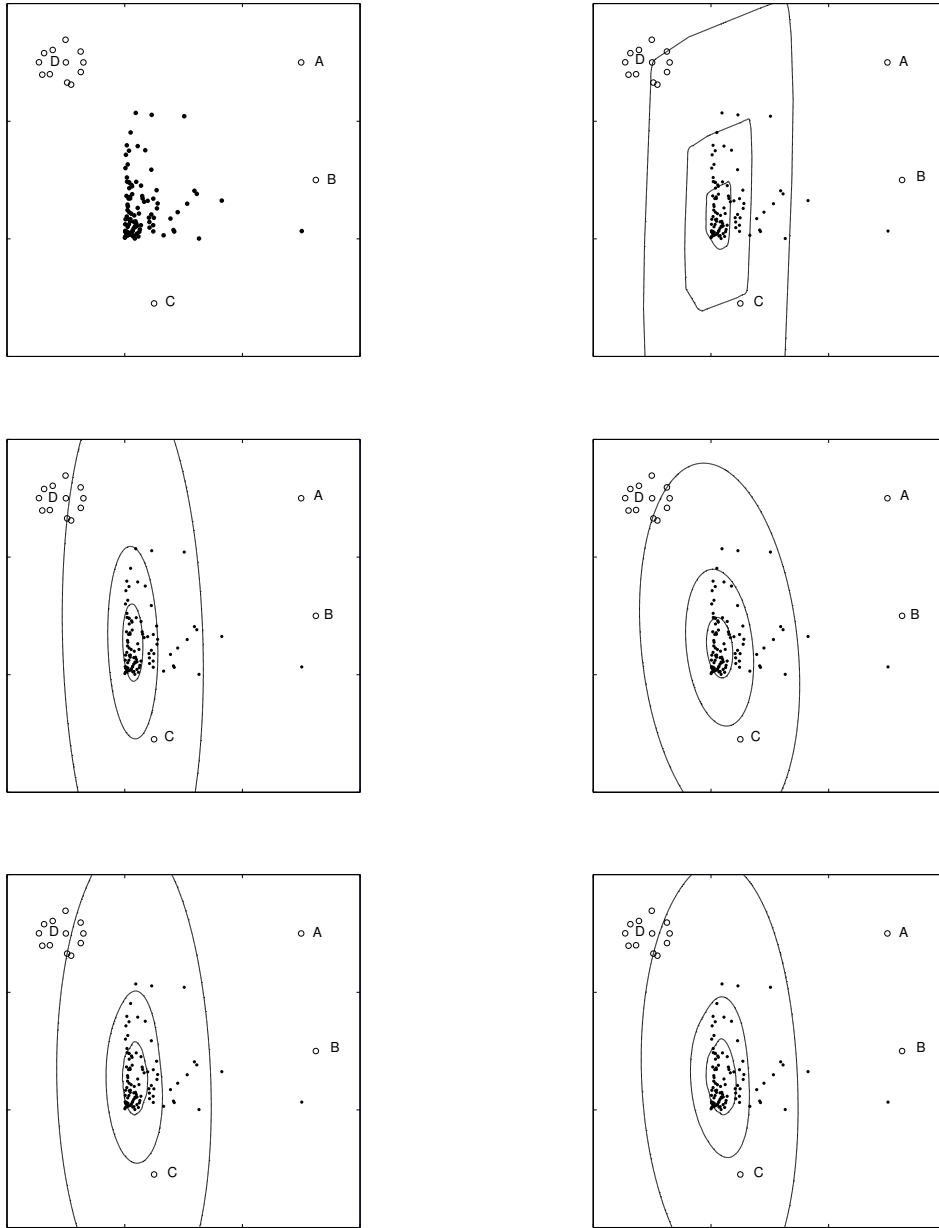


Figure 5: Bivariate Pareto data set with extreme added outliers, and 50%, 75%, and 90% outlyingness contours for SUP (upper right), MDP(MCD) (middle left), MDP(DT) (middle right) RTRP(MCD) (lower left), and RTRP(DT) (lower right).

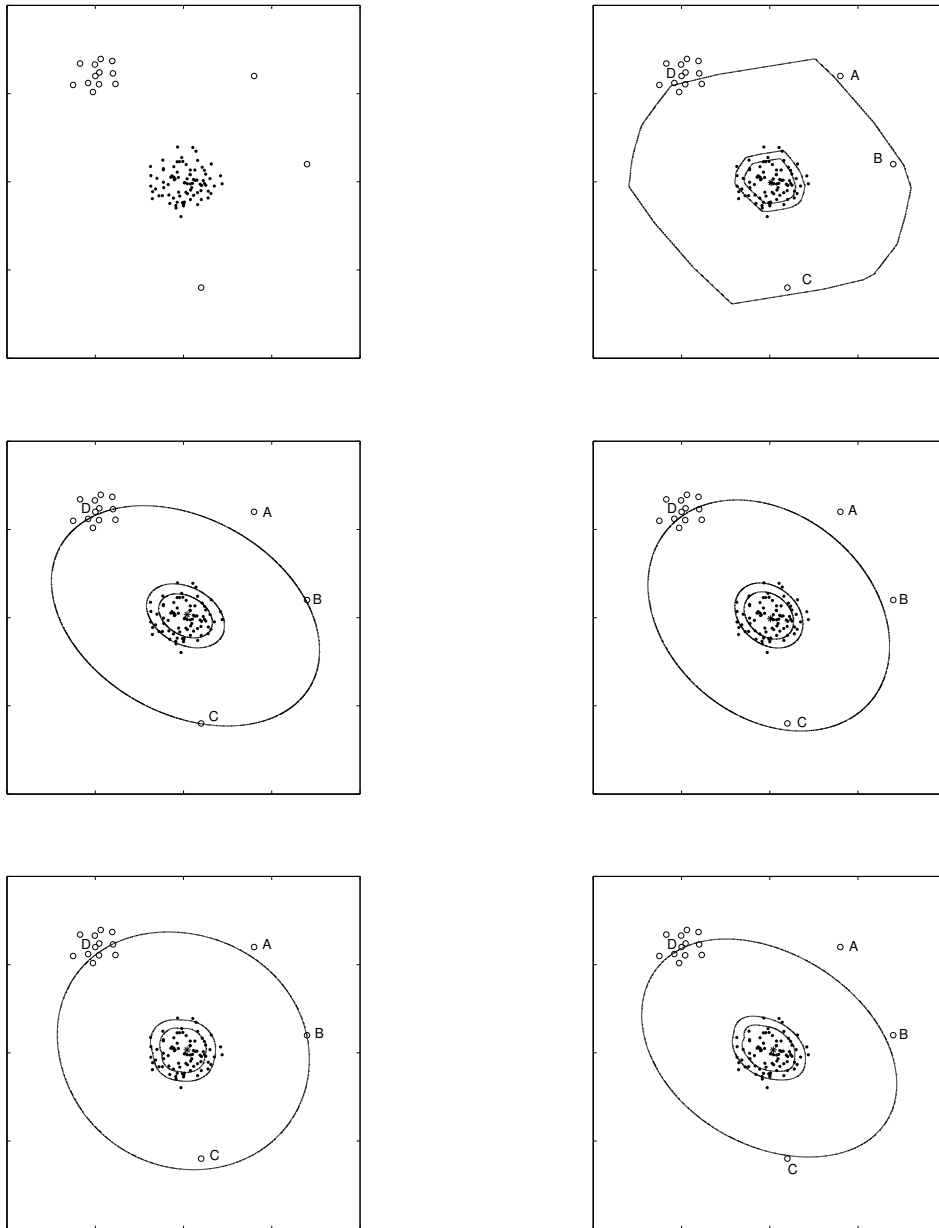


Figure 6: Bivariate Normal data set with extreme replacement outliers including a cluster, and 50%, 75%, and 90% outlyingness contours for SUP (upper right), MDP(MCD) (middle left), MDP(DT) (middle right), RTRP(MCD) (lower left), and RTRP(DT) (lower right).

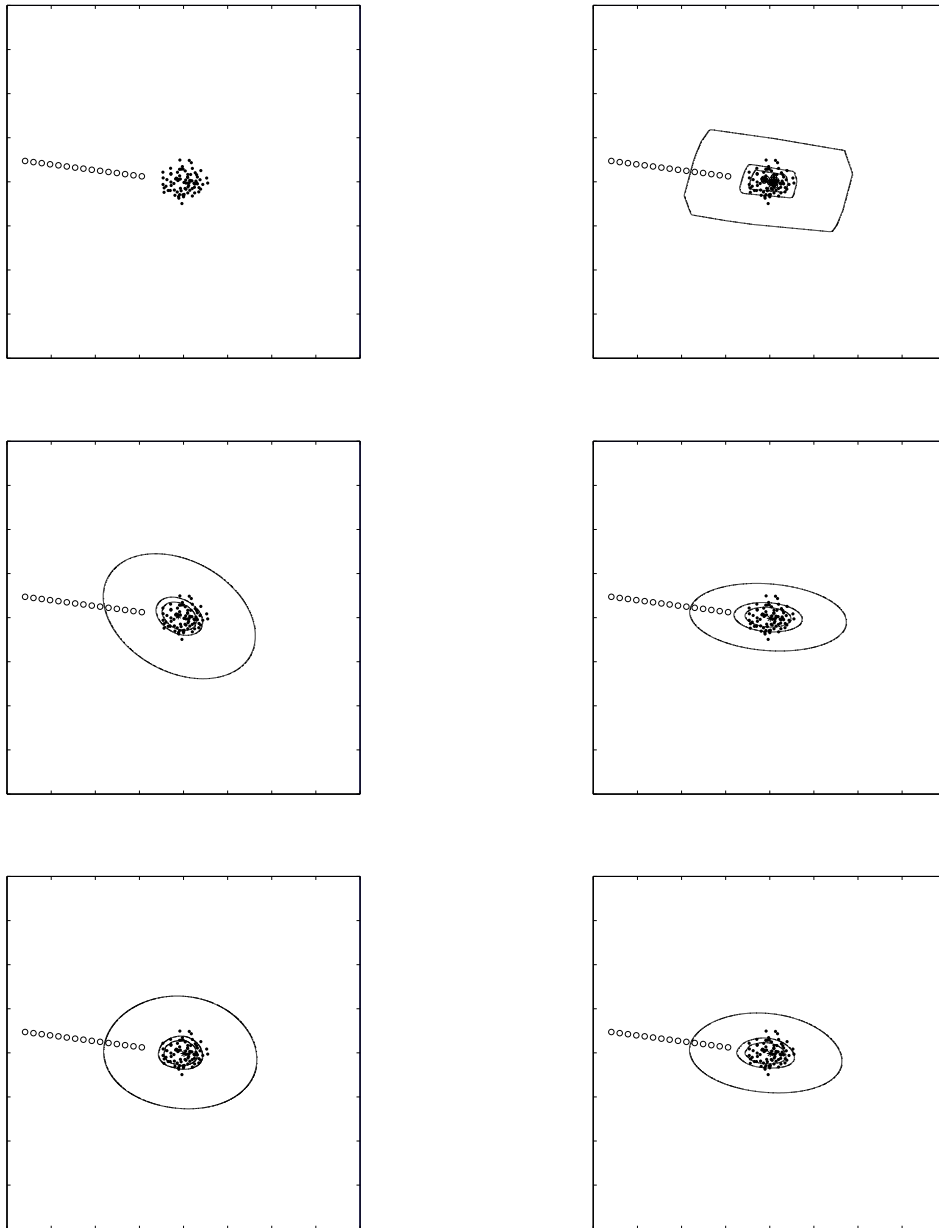


Figure 7: Bivariate Normal data set with replacement outliers in a lower dimensional space, and 50%, 75%, and 90% outlyingness contours for SUP (upper right), MDP(MCD) (middle left), MDP(DT) (middle right), RTRP(MCD) (lower left), and RTRP(DT) (lower right).

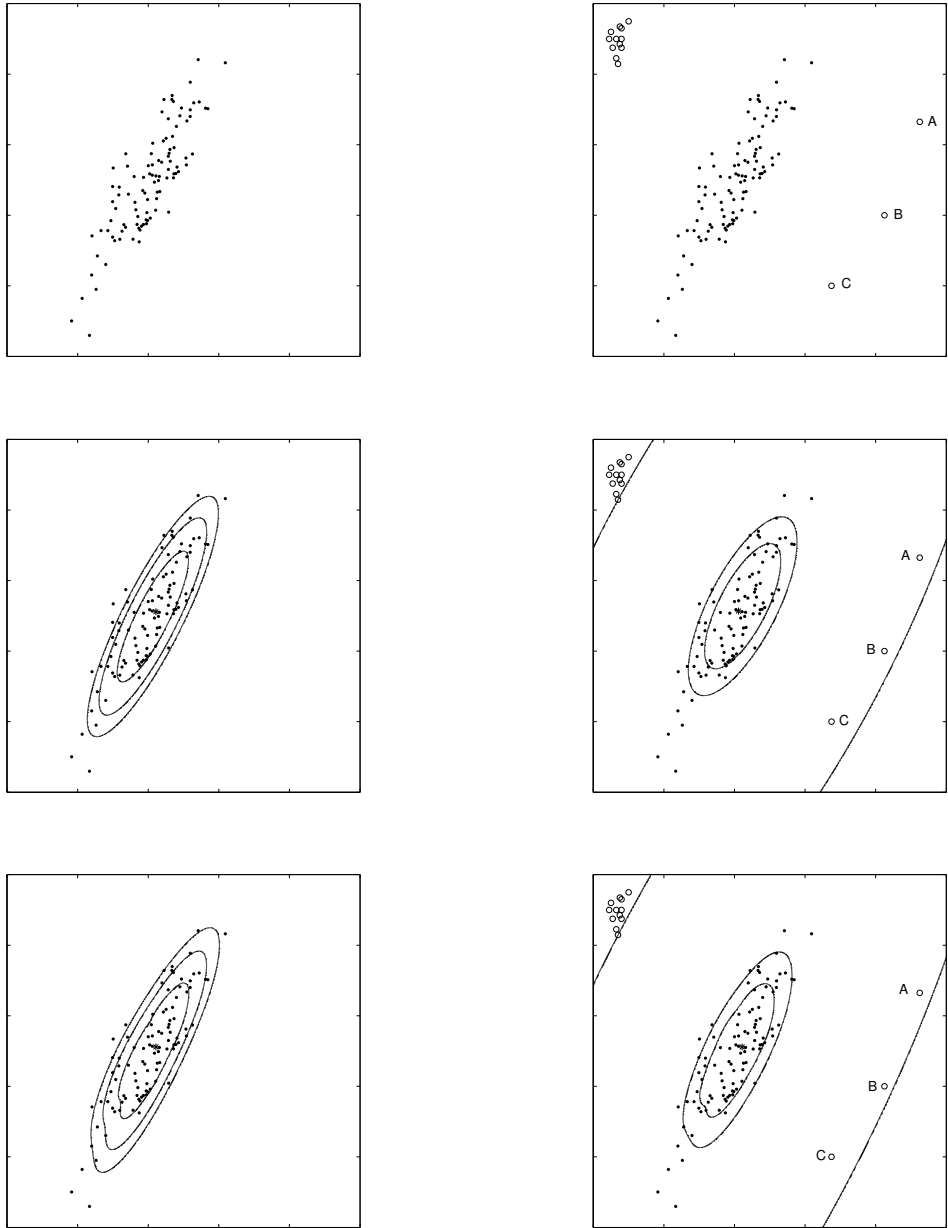


Figure 8: Bivariate correlated Normal data set without and with added outliers including a cluster, and 50%, 75%, and 90% outlyingness contours for MDP(DT) (middle row), and RTRP(DT) (lower row).