The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–25
La revue canadienne de statistique

Depth-Weighted Robust Multivariate Regression with Application to Sparse Data

Subhajit Dutta[1]* and Marc G. Genton[2]

[1]*Department of Mathematics and Statistics, Indian Institute of Technology, Kanpur - 208016, India. E-mail: duttas@iitk.ac.in, tijahbus@gmail.com*
[2]*CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. E-mail: marc.genton@kaust.edu.sa*

Key words and phrases: Data depth; LASSO; Projection depth; Spatial depth; Weighted averages.

MSC 2010: Primary 62J05; secondary 62G35

Abstract: A robust method for multivariate regression is developed based on robust estimators of the joint location and scatter matrix of the explanatory and response variables using the notion of data depth. The multivariate regression estimator possesses desirable affine equivariance properties, achieves the best breakdown point of any affine equivariant estimator, and has an influence function which is bounded in both the response as well as the predictor variable. To increase the efficiency of this estimator, a re-weighted estimator based on robust Mahalanobis distances of the residual vectors is proposed. In practice, the method is more stable than existing methods that are constructed using sub-samples of the data. The resulting multivariate regression technique is computationally feasible, and turns out to perform better than several popular robust multivariate regression methods when applied to various simulated data as well as a real benchmark dataset. When the data dimension is quite high compared to the sample size, it is still possible to use meaningful notions of data depth along with the corresponding depth values to construct a robust estimator in a sparse setting. *The Canadian Journal of Statistics* xx: 1–25; 20?? © 20?? Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–25; 20?? © 20?? Société statistique du Canada

1. INTRODUCTION

Suppose that we have a p-dimensional predictor vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ and a q-dimensional response vector $\mathbf{Y} = (Y_1, \ldots, Y_q)^T$ for $p \geq 1$ and $q \geq 1$. The multivariate regression model is

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{B}^T\mathbf{X} + \mathbf{e},$$

where $\mathbf{B}$ is the $p \times q$ slope matrix, $\boldsymbol{\alpha}$ is the q-dimensional intercept vector and the error, $\mathbf{e}$, is independent and identically distributed (i.i.d.) with mean $\mathbf{0}$ and

* *Author to whom correspondence may be addressed.*
E-mail: Insert your email address here only after your paper has been accepted

covariance matrix $\mathbf{\Sigma_e}$. We denote the location and the scatter matrix of the joint variable, $\mathbf{Z} = (\mathbf{Y}^T, \mathbf{X}^T)^T$ as $\boldsymbol{\mu}_\mathbf{Z}$ and $\mathbf{\Sigma_Z}$, respectively. There is a corresponding partition:

$$\boldsymbol{\mu}_\mathbf{Z} = (\boldsymbol{\mu}_\mathbf{Y}^T, \boldsymbol{\mu}_\mathbf{X}^T)^T \text{ and } \mathbf{\Sigma_Z} = \begin{bmatrix} \mathbf{\Sigma_{YY}} & \mathbf{\Sigma_{YX}} \\ \mathbf{\Sigma_{XY}} & \mathbf{\Sigma_{XX}} \end{bmatrix} \tag{1}$$

with $\mathbf{\Sigma_{YX}} = \mathbf{\Sigma_{XY}}^T$. Our method is well suited for both fixed and random designs. However, for our theoretical analysis we will assume that $\mathbf{Z}$ has a joint multivariate probability distribution in $\mathbb{R}^{p+q}$.

In a regression problem, we have data $\mathbf{z}_i = (\mathbf{y}_i^T, \mathbf{x}_i^T)^T$, where $\mathbf{y}_i$ is the response vector and $\mathbf{x}_i$ is the vector of co-variates for $1 \leq i \leq n$. Let $\hat{\boldsymbol{\mu}}_\mathbf{Z}$ and $\hat{\mathbf{\Sigma}}_\mathbf{Z}$ denote the estimators of $\boldsymbol{\mu}_\mathbf{Z}$ and $\mathbf{\Sigma_Z}$, respectively. The resulting estimates of $\mathbf{B}$, $\boldsymbol{\alpha}$ and $\mathbf{\Sigma_e}$ are

$$\hat{\mathbf{B}} = \hat{\mathbf{\Sigma}}_\mathbf{XX}^{-1}\hat{\mathbf{\Sigma}}_\mathbf{XY},\ \hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}_\mathbf{Y} - \hat{\mathbf{B}}^T\hat{\boldsymbol{\mu}}_\mathbf{X} \text{ and } \hat{\mathbf{\Sigma}}_\mathbf{e} = \hat{\mathbf{\Sigma}}_\mathbf{YY} - \hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}_\mathbf{XX}\hat{\mathbf{B}}, \tag{2}$$

where $\hat{\mathbf{\Sigma}}_\mathbf{XX}$ is assumed to be invertible.

The usual method of moments leads to estimators identical to those obtained from the least squares method. However, it is well-known that moment-based estimators are extremely sensitive to outliers. A common practice is thus to use robust estimators of location and scatter. One of the popular ways to construct robust estimators for multivariate data is to use the notion of data depth (e.g., Liu, Parelius and Singh, 1999; Serfling, 2006). The use of depth in building such estimators is quite natural and simple because depth has a '*center outward ordering*'. In other words, depth has this appealing property that it is maximized at the center of the data cloud, and decreases along any ray from that center. Points that are outlying with respect to a data cloud will be naturally down-weighted by depth. Measures based on data depth have nice theoretical properties as well. However, data depth has not been studied much in the context of multivariate regression. Robustness of the regression estimate depends critically on the robustness of the notion of depth that is used. In this paper, our aim is to use depth-based estimates to construct regression estimates and to investigate their performance with respect to existing estimators.

Robust estimators of location and scale for multivariate data have been studied by several authors. Popular methods of robust multivariate regression include estimators constructed using minimum covariance determinant [MCD] (Rousseeuw et al., 2004), multivariate least trimmed squares [MLTS] (Agulló, Croux and Van Aelst, 2008), S estimators [S] (Van Aelst and Willems, 2005), τ estimator [TAU] (García Ben, Martínez and Yohai, 2006), and modified M estimators [MM] (Yohai, 1987; Kudraszow and Maronna, 2011). Regression depth [RD] (introduced by Rousseeuw and Hubert (1999)) yields an alternative robust approach to estimate the regression surface in a linear regression problem. This

method is defined as the fit with the largest RD relative to the data. However, it has a breakdown value that converges almost surely to 1/3 (which is lower than several existing methods) for any dimension, and the response variable is assumed to be univariate only. Moreover, RD is somewhat different from other notions of data depth for multivariate data because it assigns depth to a fitted line and not directly to multivariate data points.

The rest of the paper is organized as follows. Section 2 describes the basic methodology of robust regression using data depth, and states related theoretical properties of the proposed estimator. The re-weighting scheme is discussed in Section 3. We perform a comparative numerical study among several competing estimators in Section 4 to assess the efficiency and robustness of the proposed methods, and we analyze a benchmark data set in Section 5. The case of robust regression for sparse data is developed and studied in Section 6. Section 7 has concluding remarks. Proofs of the mathematical statements are given in the Appendix.

2. ROBUST REGRESSION AND DATA DEPTH

We use robust estimators of $\boldsymbol{\mu}_{\mathbf{Z}}$ and $\boldsymbol{\Sigma}_{\mathbf{Z}}$ constructed from depth-based estimators of location and scatter (see Serfling, 2006; Zuo, Cui and He, 2004; Zuo, Cui and Young, 2004), respectively, as follows:

$$\hat{\boldsymbol{\mu}}_{\mathbf{Z}} = \frac{\sum_{i=1}^{n} w_1\{\delta(\mathbf{z}_i)\}\mathbf{z}_i}{\sum_{i=1}^{n} w_1\{\delta(\mathbf{z}_i)\}} \text{ and } \hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} = \frac{\sum_{i=1}^{n} w_2\{\delta(\mathbf{z}_i)\}(\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\mathbf{Z}})(\mathbf{z}_i - \hat{\boldsymbol{\mu}}_{\mathbf{Z}})^T}{\sum_{i=1}^{n} w_2\{\delta(\mathbf{z}_i)\}}.$$

Here, $\delta(\mathbf{z}_i)$ denotes the depth of $\mathbf{z}_i$ with respect to the entire data cloud for $1 \leq i \leq n$, and w_1 and w_2 are non-decreasing, non-negative weight functions. The two weight functions w_1 and w_2 may not necessarily be the same. Consider the following weight functions:

$$w_j(r) = \frac{\exp[-k\{1-(r/c)^{2j}\}^{2j}] - \exp(-k)}{1-\exp(-k)} I(0 < r < c) + I(c < r < 1), \tag{3}$$

where $I(\cdot)$ denotes the indicator function, $0 < c < 1$ and $k > 0$ for $j = 1, 2$. These are continuous surrogates of the 0-1 indicator function, and the constant k controls the degree of approximation. Following the recommendation of Zuo, Cui and He (2004), we consider a consistent estimate of c which is set to be the *median of the depth values*, and k is taken to be 100. The weight functions now assign weight 1 to half of the points with higher depth, and this balances efficiency with robustness. The other half of the points with lower depth could be viewed as outliers, so a lower weight is assigned. One could also consider other weight functions satisfying appropriate properties (see Zuo and Cui, 2005).

This gives us the initial set of estimators, $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\alpha}}$. We next state theoretical results for the initial depth-weighted regression (DWR) estimates $\hat{\mathbf{B}}$ and

$\hat{\boldsymbol{\alpha}}$. The estimators $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\alpha}}$ defined above are then constructed using projection depth (PD) (see Zuo and Serfling, 2000), and we call this method DWR-PD. The PD of a point $\mathbf{x} \in \mathbb{R}^d$ with respect to a distribution function F of $\mathbf{X}$ on $\mathbb{R}^d$ is defined as $PD(\mathbf{x}, F) = 1/(1 + O(\mathbf{x}, F))$, where the outlyingness $O(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1} \{\mathbf{u}^T\mathbf{x} - \mu(F_u)\}/\sigma(F_u)$ and F_u is the distribution function of $\mathbf{u}^T\mathbf{X}$. Here $\mu(F_u)$ and $\sigma(F_u)$ are univariate location and scale functionals, respectively, corresponding to $\mathbf{u}^T\mathbf{X}$. Proofs of the results in Sections 2.1-2.4 are given in the Appendix.

2.1. Affine equivariance

Define $\mathbf{T}_n^{PD}(\mathbf{z})$ to be the matrix $(\hat{\mathbf{B}}^T, \hat{\boldsymbol{\alpha}})$ based on DWR-PD, where $\mathbf{z} = (\mathbf{y}^T, \mathbf{x}^T)^T$.

Proposition 1 *The multivariate regression estimator* $\mathbf{T}_n^{PD}(\mathbf{z})$ *is regression,* $\mathbf{y}$*-affine and* $\mathbf{x}$*-affine equivariant.*

Popular robust regression estimators like MCD, MLTS, MM, S and TAU are all affine equivariant. We expect this property to hold in a multivariate regression method because it ensures that affine transformations of the data are reflected appropriately in the corresponding estimators. Furthermore, this also helps simplify mathematical calculations related to robustness properties of the estimator such as its influence function.

2.2. Consistency

Proposition 2 *Assume that the joint distribution of* $\mathbf{Z} = (\mathbf{Y}^T, \mathbf{X}^T)^T$ *is centrally symmetric about* $\mathbf{0}$ *and* $E(\|\mathbf{Z}\|^2) < \infty$*. Here* $\|\cdot\|$ *denotes the usual Euclidean, or* l_2 *norm. Then,* $\mathbf{T}_n^{PD}(\mathbf{z})$ *is Fisher consistent, and consistent in probability for* $(\mathbf{B}^T, \boldsymbol{\alpha})$.

Propositions 1 and 2 also hold for other depth functions which are affine invariant, i.e., $\delta(\mathbf{AZ} + \mathbf{b}) = \delta(\mathbf{Z})$ for any non-singular $(p + q) \times (p + q)$ matrix $\mathbf{A}$ and vector $\mathbf{b} \in \mathbb{R}^{p+q}$.

2.3. Breakdown point

The finite sample breakdown point (BP) (Donoho and Huber, 1983) of $\mathbf{T}_n(\mathbf{z})$ at the data set $\mathbf{Z}_n$ is defined as the smallest fraction of observations that need to be replaced by arbitrary points to carry $\mathbf{T}_n(\mathbf{z})$ beyond all bounds. We assume $\mathbf{Z}_n$ to be a set of n $(\geq p + q + 1)$ observations from a continuous distribution F in a general position, and consider the weight functions (3), also discussed by Zuo, Cui and He (2004). The result below gives the finite sample BP of $\mathbf{T}_n(\mathbf{z})$.

Proposition 3 *The multivariate regression estimator* $\mathbf{T}_n^{PD}(\mathbf{z})$ *based on PD with* $(\mu, \sigma) =$ *(Median, MAD) has a BP of* $\lfloor (n - p - q + 1)/2 \rfloor / n$*, where* $\lfloor x \rfloor$ *represents the largest integer less than or equal to* x.

The main idea of the proof relies on the breakdown point of the median and the MAD (see Zuo and Serfling, 2000). To compare, we state the BP for the existing procedures. For estimators of location and scatter based on MCD, the BP is $n\lceil\gamma\rceil/n$, with $\lceil x\rceil$ denoting the smallest integer greater than or equal to x. Here $\gamma = (n-h)/n \leq \{n-(p+q)\}/(2n)$, and h is the size of a subset used for estimation. The BP of the multivariate regression estimator based on MCD is therefore:

$$n\lceil\gamma\rceil/n$$

(Rousseeuw et al., 2004, Theorem 2, p. 300). For the estimator based on MLTS, the BP is given as:

$$\min(n-h+1, h-p-q+1)/n$$

(Agulló, Croux and Van Aelst, 2008, p. 315). The BP of the MM estimator is greater than or equal to:

$$\min\{\text{BP of initial estimator}, (\lfloor n/2\rfloor - k_n)/n\},$$

where $k_n \geq p+q-1$ (Kudraszow and Maronna, 2011, Theorem 3). Let $k(Z_n)$ denote the maximum number of observations lying on the same hyperplane of $\mathbb{R}^{p+q}$. Define $r = b/\rho(\infty)$, where ρ is a non-negative, symmetric, and non-decreasing function on $[0, c]$ and constant on $[c, \infty)$ for some constant c with $b = E[\rho(\cdot)]$, and assume $k(Z_n) < \lceil n - nr\rceil$ holds. For S estimators, the BP is as follows:

$$\min\{nr, \lceil n-nr\rceil - k(Z_n)\}/n$$

(Van Aelst and Willems, 2005, p. 984). A lower bound for the BP of the τ estimator is given by

$$\min\{(1-\eta) - (h/n), \eta\}$$

(see García Ben, Martínez and Yohai, 2006, p. 1605 for more details on the constants η and h).

It is clear from all of these expressions that the BP of all the existing methods depends on the assumed maximum proportion of contamination, and this has to be tuned appropriately. The BP of our regression estimator, $\mathbf{T}_n^{PD}(\mathbf{z})$ based on PD achieves the optimal asymptotic breakdown of 50% as we have used the median of the PD values. In the case of DWR-PD, trimming is done based on the center-outward ordering using PD, while usual trimming is based on ranks (also see discussion in the first paragraph of p. 2234 in Zuo (2006)).

2.4. Influence function

The influence function (IF) (Hampel et al., 1986) of an estimator $\mathbf{T}(\mathbf{z})$ at a general distribution H measures the effect of an infinitesimal contamination at a single point on $\mathbf{T}(\mathbf{z})$. We first state conditions for calculating the IF. Assume that the joint distribution of $\mathbf{Z} = (\mathbf{Y}^T, \mathbf{X}^T)^T$ is spherically symmetric (H); without loss of generality assume that $MAD(Z_1) = m_0$ and f is the continuous density of Z_1 satisfying $f(0)f(m_0) > 0$. Here, Z_1 is the first component of the vector $\mathbf{Z}$.

Proposition 4 *Consider the weight functions w_1 and w_2 defined in (3). The IFs of $\hat{\mathbf{B}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ based on PD with $(\mu, \sigma) =$ (Median, MAD) are given as follows:*

$$IF(\mathbf{z}; \hat{\mathbf{B}}, H) = \frac{t_1(\|\mathbf{z}\|)}{c_0} \frac{\mathbf{x}\mathbf{y}^T}{\|\mathbf{z}\|^2};$$

$$IF(\mathbf{z}; \hat{\boldsymbol{\alpha}}, H) = \frac{K_0(\mathbf{y}/\|\mathbf{z}\|) + w_1\{(1+\|\mathbf{z}\|)^{-1}\}\mathbf{y}}{\int w_1\{(1+\|\mathbf{u}\|)^{-1}\}\, dH(\mathbf{u})};$$

$$IF(\mathbf{z}; \hat{\boldsymbol{\Sigma}}_{\mathbf{e}}, H) = \frac{t_1(\|\mathbf{z}\|)\mathbf{y}\mathbf{y}^T/\|\mathbf{z}\|^2 + t_2(\|\mathbf{z}\|)I_p}{c_0}.$$

The expressions of c_0, t_1, t_2 and K_0 are given in the Appendix.

It is quite easy to derive the IF for these estimators under elliptic symmetry by using an affine transformation of $\mathbf{z}$. The IFs based on PD can also be derived for more general continuous distributions, but the expressions are fairly complicated. For general joint distributions, Theorem 2.1 of Zuo, Cui and He (2004) gives the IF of $\hat{\boldsymbol{\mu}}_{\mathbf{z}}$, and Theorem 3.3 of Zuo and Cui (2005) gives the IF of $\hat{\boldsymbol{\Sigma}}_{\mathbf{z}}$. By combining these two expressions, we may obtain a version of Proposition 4 for more general multivariate distributions.

We now state the IF of $\hat{\mathbf{B}}$ for the existing procedures and compare them with the IF of DWR-PD. For the usual MCD estimators, assuming ellipticity of the joint distribution:

$$IF(\mathbf{z}; \hat{\mathbf{B}}, H) = \frac{-1}{c} I(\|\mathbf{z}\|^2 \le q_\alpha)\mathbf{x}\mathbf{y}^T,$$

where the constant c and q_α depend on the specific elliptic distribution (Theorem 1 of Croux and Haesbroeck, 1999). Under spherical symmetry and $E(\|\mathbf{X}\|^2) < \infty$, the IF for MLTS is given by:

$$IF(\mathbf{z}; \hat{\mathbf{B}}, H) = E_H[\mathbf{x}\mathbf{x}^T]^{-1} \frac{\mathbf{x}\mathbf{y}^T}{-2c_2} I(\|\mathbf{y}\|^2 \le q_\alpha),$$

where c_2 and q_α are constants depending on the joint distribution (Agulló, Croux and Van Aelst, 2008, p. 319). The IF for the MM estimator is given by:

$$IF(\mathbf{z}; \hat{\mathbf{B}}, H) = cW\Big[\frac{\{(\mathbf{y}-\mathbf{B}^T\mathbf{x})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{B}^T\mathbf{x})\}^{1/2}}{\boldsymbol{\Sigma}}\Big]\boldsymbol{\Sigma}(\mathbf{y}-\mathbf{B}^T\mathbf{x})\mathbf{x}^T E_H[\mathbf{x}\mathbf{x}^T]^{-1}.$$

The related constants are defined in Theorem 4 of Kudraszow and Maronna (2011). Assuming the joint density to be unimodal and spherically symmetric, the IF for S estimators is:

$$IF(\mathbf{z}; \hat{\mathbf{B}}; H) = \frac{E_H[\mathbf{x}\mathbf{x}^T]^{-1}\rho(\|\mathbf{y}\|)}{\beta}\frac{\mathbf{x}\mathbf{y}^T}{\|\mathbf{y}\|},$$

where β and ρ are defined on p. 985 of Van Aelst and Willems (2005). Under appropriate conditions, the IF for the τ estimator is:

$$IF(\mathbf{z}; \hat{\mathbf{B}}, H) = c_0 w^*\Big[\frac{\{(\mathbf{y}-\mathbf{B}^T\mathbf{x})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{B}^T\mathbf{x})\}^{1/2}}{k_0}\Big]E_H[\mathbf{x}\mathbf{x}^T]^{-1}\mathbf{x}(\mathbf{y}-\mathbf{B}^T\mathbf{x})^T.$$

The related constants are given on p. 1606 of García Ben, Martínez and Yohai (2006).

The expressions in Proposition 4 are all bounded (see Lemma 1 in the Appendix). By the sub-multiplicative property of a matrix norm, we get $\|\mathbf{x}\mathbf{y}^T\| \leq \|\mathbf{x}\|\|\mathbf{y}^T\| \leq \|\mathbf{z}\|^2$. This implies that the IF of $\hat{\mathbf{B}}$ is bounded in both variables $\mathbf{x}$ and $\mathbf{y}$ for DWR-PD. The IF of the slope matrix based on MCD is also bounded in both variables. However, the IFs of the slope matrix based on re-weighted MCD, MLTS and the S estimator are all bounded in $\mathbf{y}$ but unbounded in $\mathbf{x}$. This suggests that all these methods safeguard the procedure against 'vertical outliers' and 'bad leverage' points. Moreover, the expression for the IF corresponding to the first method is derived under normality of the joint variables, while the expressions of IF for the last four methods require the finiteness of $E_H[\mathbf{x}\mathbf{x}^T]^{-1}$ and it remains unbounded for the MM estimator.

The estimators $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\alpha}}$ when constructed using spatial depth (SPD) (Vardi and Zhang, 2000; Serfling, 2002) are referred as DWR-SPD. The SPD of an observation $\mathbf{x} \in \mathbb{R}^d$ with respect to a distribution function F on $\mathbb{R}^d$ is defined as $\text{SPD}(\mathbf{x}, F) = 1 - \big\|E_F\{\mathbf{u}(\mathbf{x}-\mathbf{X})\}\big\|$, where $\mathbf{X} \sim F$. Here $\mathbf{u}(\cdot)$ is the multivariate sign function defined as $\mathbf{u}(\mathbf{x}) = \|\mathbf{x}\|^{-1}\mathbf{x}$ if $\mathbf{x} \neq \mathbf{0}_d$, and $\mathbf{u}(\mathbf{0}_d) = \mathbf{0}_d$, with $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{0}_d$ is the d-dimensional vector of zeros. The theoretical results mentioned above for DWR-PD hold only partially for DWR-SPD. Note that SPD is invariant under orthogonal transformations, and fails to be affine equivariant. The estimators based on DWR-SPD are consistent only when $\mathbf{Z}$ is spherically symmetric. Fix a constant λ with $0 < \lambda < 1$, and consider the set $\{\mathbf{x} : \text{SPD}(\mathbf{x}, F) < 1 - \lambda\}$. If we have an observation lying inside (respectively, outside) this set, then it is called a λ outlier (respectively, non-outlier). Now, the masking BP of SPD is

$\lceil n(1-\lambda)/2 \rceil / n$ (Theorem 3.5 of Dang and Serfling, 2010), and the resulting BP for DWR-SPD also depends on this trimming factor λ. An expression for the IF of SPD has been calculated and shown to be bounded by Dang, Serfling and Zhou (2009). Furthermore, Dang, Serfling and Zhou (2009) stated a general result for depth-weighted location estimators and the IF has a very complicated expression for general depth functions. Combining these two expressions, one may try to obtain a final expression for the IF of DWR-SPD. However, we have not been able to derive this result which is still an open problem.

3. RE-WEIGHTED MULTIVARIATE REGRESSION

3.1. Re-weighting based on robust Mahalanobis distances

The use of a depth-weighted estimator invokes robustness, although the overall method loses efficiency. For example, our procedure clearly puts very low weight on the 'good leverage' points. This is the case because the point is outlying with respect to such a data cloud and hence has a low depth value. A 'bad leverage' point is a point for which both $\|\mathbf{x}\|$ and $\|\mathbf{e}\|$ have high values; and a 'vertical outlier' is a point for which $\|\mathbf{x}\|$ has a low value and $\|\mathbf{e}\|$ has a high value.

Define the estimated residual vector as $\hat{\mathbf{e}}_i = \mathbf{y}_i - \hat{\mathbf{y}}_i$ for all $1 \leq i \leq n$, where $\hat{\mathbf{y}}_i = \mathbf{y}_i - \hat{\boldsymbol{\alpha}} - \hat{\mathbf{B}}^T \mathbf{x}_i$. Consider the standardized residuals $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}^{-1/2} \hat{\mathbf{e}}_i$, and without confusion we write them $\hat{\mathbf{e}}_i$ again. Recall the expression of $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ given by (2), which is a robust estimator for $\boldsymbol{\Sigma}_{\mathbf{e}}$. For these $\hat{\mathbf{e}}_i$, good leverage points will have a low value of $\|\hat{\mathbf{e}}_i\|$; while vertical outliers and bad leverage points will have a high value of $\|\hat{\mathbf{e}}_i\|$. So, the main objective is to retain observations about the point $\mathbf{0}$, and try to discard the remaining points. In other words, we calculate the Mahalanobis distances (Mahalanobis, 1936) for the residual vectors with $\mathbf{0}$ as the center and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ as the scatter matrix. We then use the adaptive re-weighting scheme of Gervini (2003, pp. 118-119) to identify outliers in the cloud of residuals. The details are as follows.

Let d_i denote the Mahalanobis distances for the residual vectors $\hat{\mathbf{e}}_i$ with $\mathbf{0}$ as the center and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ as the scatter matrix for $1 \leq i \leq n$, and $d_{(1)} \leq \cdots \leq d_{(n)}$ be their values in ascending order. Define $i_0 = \max\{i : d_{(i)}^2 < \chi_{q,1-\alpha}^2\}$ for a fixed α (say, 0.025) and $\alpha_n = \max_{i>i_0}\{G_q(d_{(i)}^2) - (i-1)/n\}_+$ with $\{\cdot\}_+$ denoting the positive part, and G_q the cumulative distribution function of χ_q^2 (the chi-square distribution with q degrees of freedom (df)). Here $\chi_{q,1-\alpha}^2$ denotes the upper α point of the χ_q^2 distribution. Now, the observations corresponding to the largest $\lfloor n\alpha_n \rfloor$ distances are considered as outliers, while the remaining are non-outliers. For the observations selected as non-outliers, we update the corresponding weights to be 1. Next, we do a round of weighted least squares (WLS) regression with this 'new' set of weights which gives us the final set of regression estimators, namely, $\hat{\mathbf{B}}^R$ and $\hat{\boldsymbol{\alpha}}^R$. This re-weighted estimator is affine equivariant as $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ is an affine equivariant estimator of $\boldsymbol{\Sigma}_{\mathbf{e}}$.

To get a better understanding of how the re-weighting step works, we construct two plots. We first generate a data set of size 48, where the regressors come from a normal distribution with mean 0 and variance 0.2, while the response is obtained by adding an error term which also has a standard normal distribution with variance 0.1. We next add a good leverage point at $(1, 1)$ and an outlier at $(0.3,\text{-}0.5)$, and this makes the total sample size 50. The panel on the left in Figure 1 shows the fitted line based on the PD-weighted estimators with the 0-1 weight function, and the points with the cut-off is set to be the median of the PD values (which ensures the first 50% points with the highest PD to be selected). The black points are the data points with their weight equal to 1, and they contribute significantly in the regression estimate. We see that the leverage point as well as the outlier have been omitted because our method is based on depth. The re-weighting step rotates the line clockwise (the bold line in the right panel of Figure 1) and includes the good leverage point as a part of the fit.

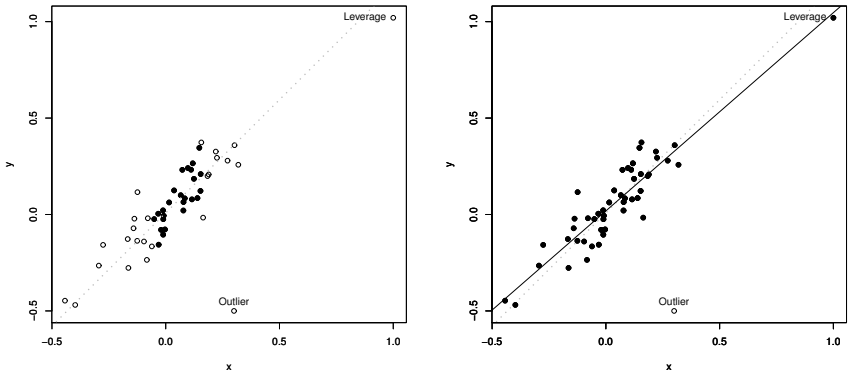

FIGURE 1: An illustrative plot of our two step procedure. In the left panel, the black points are the points with highest depth and the dotted line represents the fitted line based on $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\alpha}}$. After re-weighting, in the right panel, we plot the new fitted line as a bold line and the black points that contribute to this new fit based on $\hat{\mathbf{B}}^R$ and $\hat{\boldsymbol{\alpha}}^R$.

3.2. Practical aspects

We consider PD together with the median and the median absolute deviation about the median (MAD) as univariate robust estimators (Zuo and Serfling, 2000). The theoretical version of PD has good robustness properties, however its computation poses an additional difficulty (Liu and Zuo (2014)). For data in $\mathbb{R}^d$, PD involves calculating a supremum in the d-dimensional space. Practically, it is not possible to compute this supremum exactly in arbitrary dimensions and one usually uses approximation algorithms to calculate PD. We have used the R func-

tion `zdepth` developed by Wilcox (2012) for computing PD. The algorithm is based on the well-known Nelder-Mead or *downhill simplex* method (Nelder and Mead, 1965) for solving the maximization problem in the outlyingness function $O(\mathbf{x}, F)$ in $\mathbb{R}^d$.

The notion of spatial depth (SPD) eases this computational burden. SPD can be calculated exactly as it is an average of unit vectors $\mathbf{x}_i/\|\mathbf{x}_i\|$ constructed from the n data points. However, one loses a bit of robustness because equal weight are assigned to all those unit vectors. We implement both methods in our procedure, and present a comparative study.

4. NUMERICAL WORK

Our numerical study is motivated by the examples considered by Agulló, Croux and Van Aelst (2008). We perform a study of the efficiency and robustness of the overall procedure. We have used R (www.r-project.org) codes from the `robustbase` package (Rousseeuw et al., 2015) for MCD, and the `FRB` package (Van Aelst and Willems, 2013) for both S as well as MM. Codes for MLTS/RMLTS are available at `http://www.econ.kuleuven.be/public/NDBAE06/programs/mlts/mlts.r.txt`, while those for TAU were obtained from Prof. Victor J. Yohai. We have our codes in R, and they are available as Supplementary Material. For the sake of comparison, we also study the performance of our method based on SPD (called DWR-SPD). We calculate the mean square error (MSE) of the slope matrix $\hat{\mathbf{B}}$ by computing an average over the MSE of each element of this matrix over several random realizations of the data. The matrix norm used here is the usual component-wise l_2 norm. In our experiments, we observe issues of singularity (the weights become zero in the calculation of weighted covariance) for S estimators on several occasions, and have not considered this estimator as one of our competitors.

The tuning parameters for MCD and MLTS were set to be $\alpha = 0.50$ and $\gamma = 0.50$, respectively. The re-weighted versions of MLTS and MCD are denoted as RMLTS and RMCD, respectively. For the TAU estimator, we set $N =$ (sample size)$/2$, while the constants c_1, c_2 and k_a were chosen based on Tables 1 and 2 of García, Martínez and Yohai (2006) to attain 95% efficiency. We did not have to set any default parameters for the MM estimator. The `FRB` package uses Tukey's bi-weight function. In the first step (the S estimate), this function was first tuned to obtain 50% BP, and in the second step (the M estimate) it was tuned again to ensure 95% efficiency at the normal model.

4.1. Finite sample performance

In this section, we report on our investigation of the finite sample performance of our estimators and compare the estimators with other robust multivariate regression estimators. We generated $m = 500$ regression data sets each of size $n = 100$. For this study, we considered $p = q = 3$ with the first regressor ac-

TABLE 1: MSE of the estimated slope matrix for different error distributions with standard error in brackets

					Gaussian response				
Expl.	DWR-PD	RDWR-PD	DWR-SPD	RDWR-SPD	MLTS	RMLTS	MM	RMCD	TAU
Gaussian	8.5965	3.6662	8.5965	2.5833	8.1382	3.2503	2.1851∗	2.8920	2.2177†
	(0.0767)	(0.0569)	(0.0767)	(0.0508)	(0.0902)	(0.0570)	(0.0467)	(0.0537)	(0.0470)
Cauchy	5.1099	1.2143	5.1099	0.7893	0.6137	0.2942	0.0593∗	1.6228	0.0799†
	(0.0651)	(0.0332)	(0.0651)	(0.0280)	(0.0247)	(0.0171)	(0.0077)	(0.0402)	(0.0089)
Uniform	16.2207	10.0745	16.2207	7.7058	24.0619	9.5513	6.3473∗	7.4889	6.4733†
	(0.1043)	(0.0947)	(0.1043)	(0.0877)	(0.1547)	(0.0975)	(0.0795)	(0.0864)	(0.0803)
					Cauchy response				
Expl.	DWR-PD	RDWR-PD	DWR-SPD	RDWR-SPD	MLTS	RMLTS	MM	RMCD	TAU
Gaussian	6.8035	4.7128	6.8035	4.1929†	5.2219	5.0848	5.0376	6.5701	4.1126∗
	(0.0765)	(0.0678)	(0.0765)	(0.0647)	(0.0722)	(0.0713)	(0.0709)	(0.0810)	(0.0641)
Cauchy	4.3427	1.2998	4.3427	1.0523	0.3383	0.3036†	0.2272	2.5505	0.3000∗
	(0.0637)	(0.0356)	(0.0637)	(0.0323)	(0.0183)	(0.0174)	(0.0150)	(0.0504)	(0.0173)
Uniform	14.2377	13.1315	14.2377	12.0071∗	15.2026	14.6114	15.1219	19.4509	12.3017†
	(0.1133)	(0.1138)	(0.1133)	(0.1095)	(0.1232)	(0.1208)	(0.1229)	(0.1394)	(0.1108)
					Uniform response				
Expl.	DWR-PD	RDWR-PD	DWR-SPD	RDWR-SPD	MLTS	RMLTS	MM	RMCD	TAU
Gaussian	5.0259	1.5050	5.0259	0.8124∗	4.9035	1.6965	0.8229†	1.1139	0.8427
	(0.0552)	(0.0360)	(0.0552)	(0.0284)	(0.0699)	(0.0411)	(0.0286)	(0.0333)	(0.0290)
Cauchy	2.8044	0.5173	2.8044	0.2876	0.3652	0.1479	0.0191∗	0.5426	0.0262†
	(0.0475)	(0.0217)	(0.0475)	(0.0169)	(0.0190)	(0.0121)	(0.0043)	(0.0232)	(0.0051)
Uniform	10.4807	4.2928	10.4807	2.4930∗	15.2485	5.5557	2.5225†	2.9651	2.5831
	(0.0800)	(0.0601)	(0.0800)	(0.0499)	(0.1234)	(0.0745)	(0.0502)	(0.0544)	(0.0508)

Here ∗ denotes the best result and † denotes the second best result in each row.

counting for the intercept term. The remaining $p-1$ explanatory variables were generated from the following distributions:

[1] The multivariate standard normal distribution;

[2] The multivariate standard Cauchy distribution;

[3] The multivariate uniform distribution on $(-1,1)^p$.

The multivariate uniform distribution was generated by considering componentwise univariate uniform distributions on the symmetric interval $(-1,1)$. Without loss of generality, we took $\mathbf{B}=\mathbf{0}$ in the multivariate regression model. The response variables were generated from each of these three distributions.

From Table 1 it is clear that the TAU estimator yields the best overall performance, closely followed by the MM estimator. The re-weighted version of the SPD based estimator, RDWR-SPD also led to competitive performance in some scenarios when the response as well as the explanatory variable was uniformly distributed. This improved performance may be due to the fact that SPD led to a 'center outward ordering' of observations from a uniform distribution (also see p.5 of Serfling, 2006). Generally, we also observed that the re-weighted version of DWR-SPD was more efficient than RDWR-PD. The estimator RMLTS resulted in lower MSE as compared to those obtained from RMCD.

4.2. Finite sample robustness

To study the finite sample robustness of the proposed estimators, we carried out simulations with contaminated data sets. The parameters were chosen to be $\boldsymbol{\alpha}=\mathbf{1}$ and $\mathbf{B}=\mathbf{0}$ as in Section 4.1. We first simulated $m=500$ data sets of size $n=100$ with $p=q=3$ and assumed the errors to be Gaussian. The predictor variables were generated from Gaussian as well as Cauchy distributions. To generate contaminated data sets, we replaced 20% of the data with new observations. The new $p-1$ independent variables were generated according to $\mathrm{N}(\lambda\sqrt{\chi^2_{p-1,0.99}},1.5)$, while the new q dependent variables were generated from $\mathrm{N}(\kappa\sqrt{\chi^2_{q,0.99}},1.5)$. Here $\chi^2_{r,\alpha}$ represents the upper α point of the chi-squared distribution with r df for $r=p-1$ and q, and $\alpha=0.99$. We considered λ and κ values to be in the range $\{0,1,2,3,4,5\}$. If $\lambda=0$ and $\kappa>0$, we obtained 'vertical outliers'. On the other hand, if $\lambda>0$ and $\kappa=0$ we obtained 'good leverage' points.

From Figures 2 and 3, it is clear that both the MM and TAU estimators are uniformly more efficient than all other methods. For data with normal predictors, RDWR-PD exhibited substantial improvement over RDWR-SPD. While RMLTS lead to comparable performance with RDWR-PD, the former did have an edge in some situations. Surprisingly, RMCD led to very high MSE for the first case ($\lambda=0$) with Cauchy explanatory variables (see right panel of Figure 2).

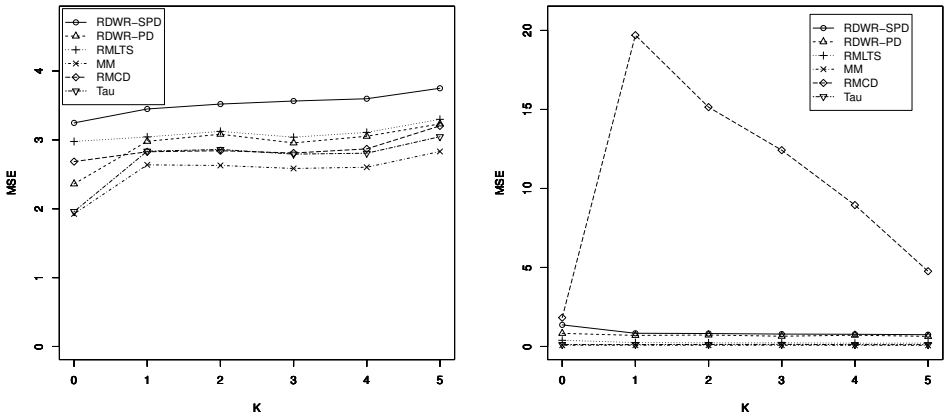

FIGURE 2: MSE of the slopes for normal and Cauchy explanatory variables for $\lambda = 0$

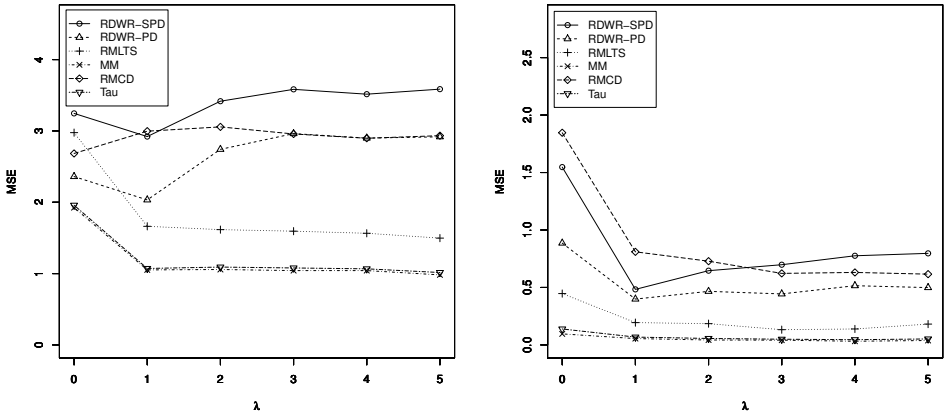

FIGURE 3: MSE of the slopes for normal and Cauchy explanatory variables for $\kappa = 0$

If both $\lambda > 0$ and $\kappa > 0$, we obtain 'bad leverage' points. Large values of λ and κ produce extreme outliers while small values produce intermediate outliers. In Figure 4, for each value of λ, we plot the maximal value of MSE based on the contaminated data over all possible values of κ.

In the left panel of Figure 4, we observe a lack of robustness of both the MM and TAU estimators. The performance of RMLTS is also quite poor. Clearly, both our depth based methods as well as RMCD lead to uniformly lower values of MSE. The situation improves for the case with Cauchy explanatory variables (the right panel) as the MSE decreases considerably for the MM and TAU estimators, but depth based methods are clearly more robust. In fact, both RDWR-PD and RDWR-SPD yield MSE close to 0 over all values of λ.

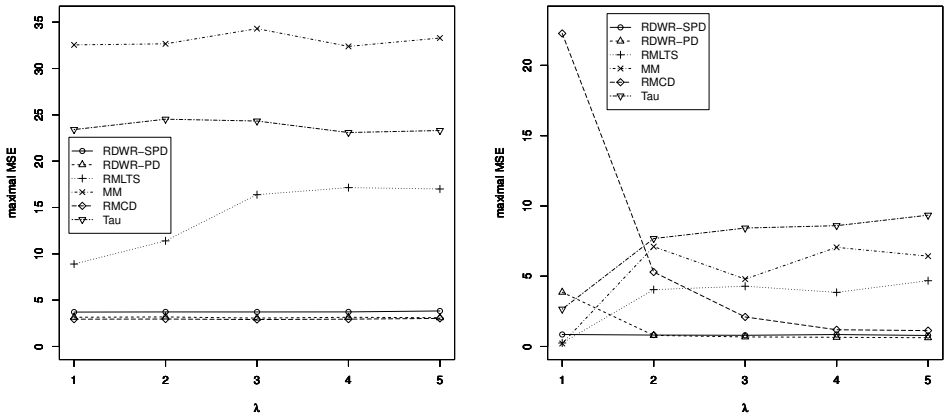

FIGURE 4: Maximal MSE of the slopes for normal and Cauchy explanatory variables

5. DATA ANALYSIS

We analyze a benchmark data set to illustrate the usefulness of the methods. The diagnostic plots shown in Figure 5 combine the information on regression outliers and leverage points, and are more useful than separately analyzing each distance (e.g., Rousseeuw et al., 2004). Here, robust distances are calculated using projection outlyingness. We plot the robust distance of the residuals versus the robust distance of the predictor variables. The horizontal and vertical lines are cut-off values of $\chi^2_{p,0.975}$ and $\chi^2_{q,0.975}$, respectively.

5.1. School data

The aim of the study conducted by Charnes, Cooper and Rhodes (1981) was to explain the scores on three tests from 70 schools by using five explanatory variables. The three test scores were (i) the total reading score, (ii) the total mathematics score (both measured by the Metropolitan achievement test) and (iii) the Coopersmith Self-Esteem Inventory. The explanatory variables include the education level of the mother as measured in terms of percentage of high school graduates among female parents, the highest occupation of a family member according to a pre-established rating scale, a parental visit index indicating the number of visits to the school site, a parent counselling index calculated from data on time spent with the child on school-related topics such as reading together, etc., and the number of teachers at a given site. We have $p = 5$, $q = 3$ and $n = 70$ for this data set.

In this data, RDWR-PD and RDWR-SPD methods perform at par with the existing estimators, with all methods uniformly classifying observation 59 as a 'bad leverage' point (see Table 2 and Figure 5). The MM and RMLTS estimators identify observation number 44 as an additional 'vertical outlier' while RDWR-PD

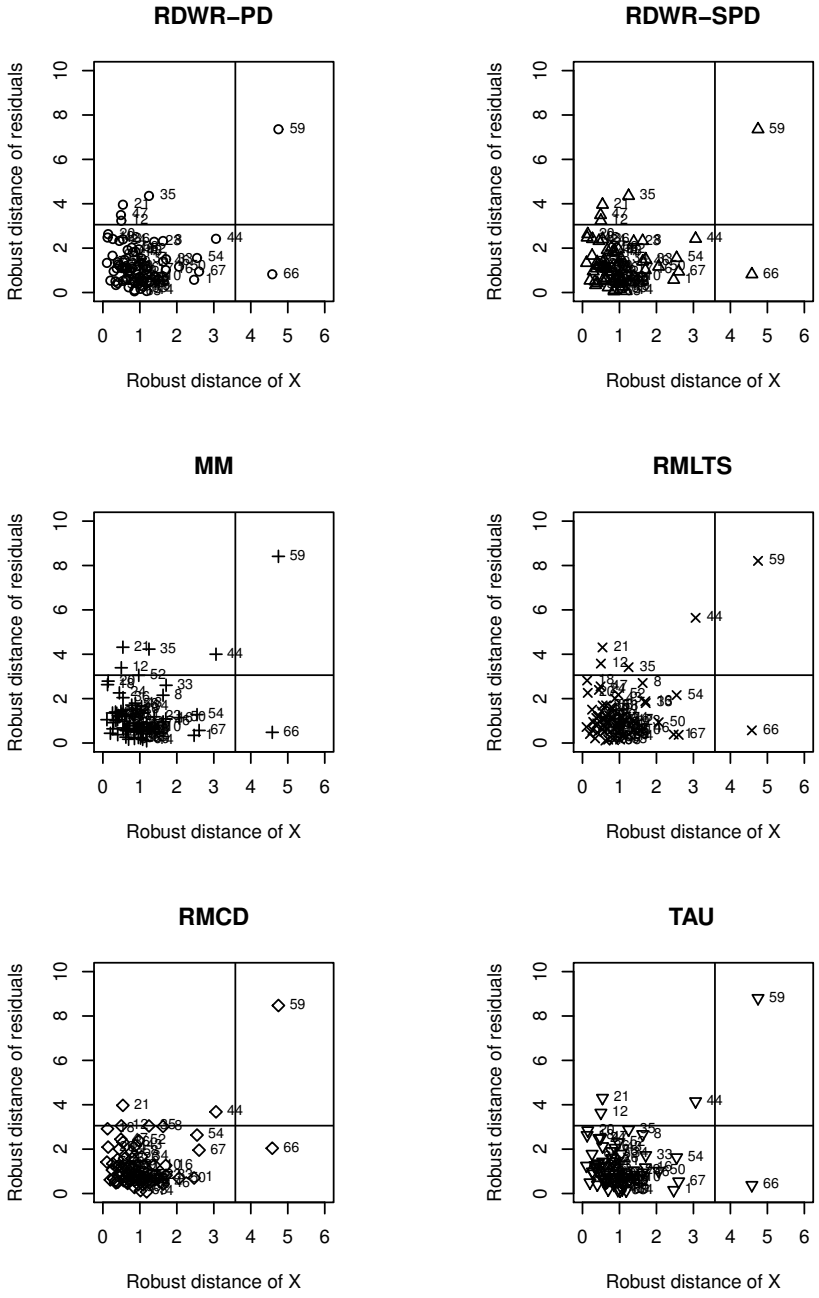

FIGURE 5: Diagnostic plots of various methods showing robust residuals versus robust distances of the explanatory variables for the school data. The cut-off points are set at $\chi^2_{5,0.975}$ and $\chi^2_{3,0.975}$ for the horizontal axis and vertical axis, respectively.

and RDWR-SPD point out observation 47. RMCD fails to identify observation numbers 12 and 35 as outliers, while TAU misses the latter.

TABLE 2: Index of 'bad' points in the school data

Method	Vertical outlier	Bad leverage
RDWR-PD	12, 21, 35, 47	59
RDWR-SPD	12, 21, 35, 47	59
RMLTS	12, 21, 35, 44	59
MM	12, 21, 35, 44	59
RMCD	21, 44	59
TAU	12, 21, 44	59

To understand the relative importance among the vertical outliers, we make a pairwise plot of the response variable for the observations 12, 21, 35, 44 and 47. The influence of observations 44 and 47 is quite clear from Figure 6.

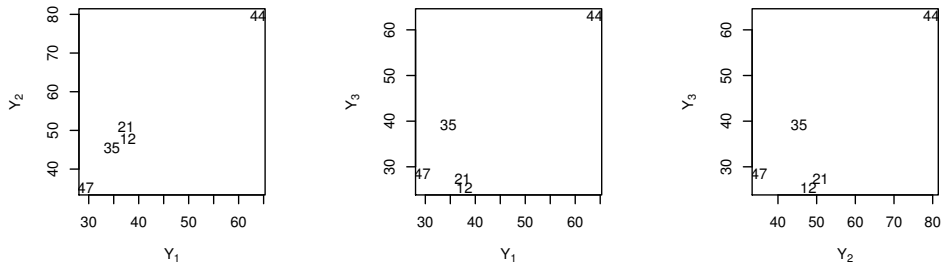

FIGURE 6: Pairwise plots for the response variable.

6. ROBUST MULTIVARIATE REGRESSION FOR SPARSE DATA

6.1. The re-weighted LASSO estimator

With the advancement of scientific techniques, data with dimension higher than the sample size have become quite common in practice. Moreover, an interesting question in such a scenario is the identification of an outlier. A study on data depth by Chakraborty and Chaudhuri (2014) shows that for a large class of infinite-dimensional distributions, the notion of SPD takes all values in the interval $(0, 1)$ (see their Theorems 6 and 7). SPD is therefore still meaningful for data arising from a quite large class of infinite-dimensional distributions. This motivates us to look into the area of robust regression for sparse data.

There is a limited literature for this scenario. An approach by Alfons, Croux and Gelper (2013) combined the idea of LASSO (Tibshirani, 1996) and LTS (Rousseeuw and Leroy, 1987) to construct a new method for sparse data. Regarding our method described in Section 1, the estimates in (2) can also be obtained by minimizing the function

$$(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{B}}^T) = \arg\min_{\boldsymbol{\alpha},\mathbf{B}} \sum_{i=1}^{n} w_i(\mathbf{y}_i - \boldsymbol{\alpha} - \mathbf{B}^T\mathbf{x}_i)^T(\mathbf{y}_i - \boldsymbol{\alpha} - \mathbf{B}^T\mathbf{x}_i),$$

with appropriate weights w_i for $1 \leq i \leq n$. One may refer to Johnson and Wichern (2007, pp. 387-389) for a derivation of this least squares minimization problem. In Section 2, we have described multivariate least squares regression using depth as the weights.

For data with sparsity, we now use LASSO as our method of regression instead of usual multivariate regression. LASSO allows one to do weighted regression, and like DWR-SPD we continue to use SPD as the weights. In the sparse case, one may re-formulate the minimization problem by penalizing the matrix$\mathbf{B}$:

$$\arg\min_{\boldsymbol{\alpha},\mathbf{B}} \sum_{i=1}^{n} w_i(\mathbf{y}_i - \boldsymbol{\alpha} - \mathbf{B}^T\mathbf{x}_i)^T(\mathbf{y}_i - \boldsymbol{\alpha} - \mathbf{B}^T\mathbf{x}_i) + \lambda\|\mathbf{B}\|_{l_1}.$$

The constant $\lambda > 0$ controls the amount of penalization, and $\|\mathbf{B}\|_{l_1} = \sum_{kl} |b_{kl}|$ is the l_1 matrix norm of $\mathbf{B}$. For details on the formulation of multi-response sparse linear regression and some of its variants, see Li, Nan and Zhu (2015) and Wang, Liang and Xing (2015). The weights w_i are calculated based on SPD, and the weight function is determined by (3). This approach (called LASSO-SPD) can be directly used for robust regression with sparse data. In fact, we are not restricted to the response being univariate because LASSO has the flexibility to model data from multivariate responses. The LASSO with a multi-response Gaussian model allows such a fit with a "group-lasso" penalty on the coefficients for each variable (Friedman, Hastie and Tibshirani, 2010), or using the mixed co-ordinate descent algorithm (Li, Nan and Zhu, 2015).

Following the re-weighting step in Section 3 on the data cloud of residual vectors, we carry out an additional step of the LASSO after assigning weight 1 to the new observations to increase the efficiency of our estimates, and call it RLASSO-SPD. The advantage offered by our final estimator RLASSO-SPD is evident in the numerical study below.

6.2. A numerical evaluation

The R codes for sparse LTS and LASSO are available in the packages `robustHD` (Alfons, 2014) and `glmnet` (Friedman, Hastie and Tibshirani, 2010), respectively. In the implementation of sparseLTS and glmnet, we fix a grid of values from 0.05 to 0.50 with an increment of 0.05 for the regularization

parameter λ. For each value of λ, we obtain an estimate of $\mathbf{B}$. We then compute the MSE of this estimate over this sequence of values of λ, and the minimum value of MSE is reported in Table 3. Each experiment is replicated 100 times.

TABLE 3: Average MSE of the estimated slope vector/matrix with different error distributions with standard error in brackets

		Univariate response	$(q = 1)$	
Explanatory	LASSO-SPD	RLASSO-SPD	LASSO	Sparse LTS
Gaussian	0.00081	0.00073†	0.00084	0.00066∗
	(0.00006)	(0.00005)	(0.00005)	(0.00003)
Cauchy	0.00143	0.00136†	3.19593	0.00086∗
	(0.00010)	(0.00009)	(0.01773)	(0.00005)
Gaussian + out	0.00080†	0.00076∗	0.00082	0.00103
	(0.00007)	(0.00005)	(0.00005)	(0.00007)
Cauchy + out	0.00155	0.00151†	0.52499	0.00080∗
	(0.00012)	(0.00011)	(0.00716)	(0.00007)
		Bivariate response	$(q = 2)$	
Explanatory	LASSO-SPD	RLASSO-SPD	LASSO	Sparse LTS
Gaussian	0.58797†	0.38068∗	0.85954	–
	(0.00304)	(0.00608)	(0.00633)	–
Cauchy	0.81918†	0.69451∗	4.08733	–
	(0.00577)	(0.07108)	(0.01901)	–
Gaussian + out	0.68871†	0.68167∗	0.71757	–
	(0.00435)	(0.00434)	(0.00472)	–
Cauchy + out	0.89229†	0.75397∗	2.25875	–
	(0.00636)	(0.00512)	(0.01322)	–

Here, $*$ denotes the best result and † denotes the second best result in each row.

Following Alfons, Croux and Gelper (2013), we generated a high-dimensional dataset with 20 observations from a p-dimensional distribution with $p = 1000$. The (i, j)-th element of the covariance matrix is $0.5^{|i-j|}$, which gives rise to correlated predictor variables. The coefficient vector was made sparse by choosing the first 20 components to be one, and the rest zero. We generated the predictor variables from the multivariate standard Gaussian and Cauchy distributions with the above correlation structure. The response variable was generated according to the regression model, where the error terms followed a standard normal distribution with a standard deviation of 0.5. We then considered a second set of examples where we added five 'vertical outliers' at two locations, namely, $\mathbf{10}$ and $-\mathbf{15}$, which are the p-dimensional vectors of 10's and -15's, respectively. For the case when $q = 2$, we added a second coefficient vector by taking the last 20 components to be zero, and all the rest to be one. The four ex-

amples mentioned above in this paragraph are all re-considered in this case with bivariate response.

The values in Table 3 indicate that RLASSO-SPD is quite competitive with respect to sparse LTS, and improves the classical LASSO. However, the real advantage and usefulness of our method appears when we use it for bivariate responses with outliers and heavy tails. In the second part of this table, RLASSO-SPD outperforms the LASSO considerably, while sparse LTS is not applicable to such data. In terms of computational time, the average computing time for LASSO and hence for RLASSO-SPD is about 1 second per iteration. The computation is dominated by LASSO because SPD involves only the calculation of averages of unit vectors. This is quite fast compared with sparse LTS, which takes around 15 to 20 seconds per iteration.

6.3. Choice of the parameter λ

In practical data analysis, a suitable value of the regularization parameter λ in the LASSO is unknown. We propose to select λ by minimizing the estimated prediction error using the idea of cross-validation (e.g., Hastie, Tibshirani and Friedman, 2009). In the sparse setup, n is quite small compared to p and we use leave one out cross-validation (LOOCV). In LOOCV, each data point is left out once to fit the model and the left out data point is used later as a test data for prediction. To prevent outliers from affecting the choice of λ, a robust prediction error is desirable (e.g., Cantoni and Ronchetti, 2001). For a given value of λ, a set of n prediction error vectors are obtained. We use the approach in Section 3.1 to identify outliers in this set of error vectors, and compute the mean squared prediction error (MSPE) based only on the non-outlier error vectors. In other words, $\text{MSPE}(\lambda) = |I|^{-1} \sum_{i \in I} \mathbf{e}_i^T \mathbf{e}_i$, where I denotes the subset of non-outliers in the set of prediction error vectors and $|I|$ is the cardinality of the set I. The value of λ which minimizes $\text{MSPE}(\lambda)$ is chosen to be the optimal one.

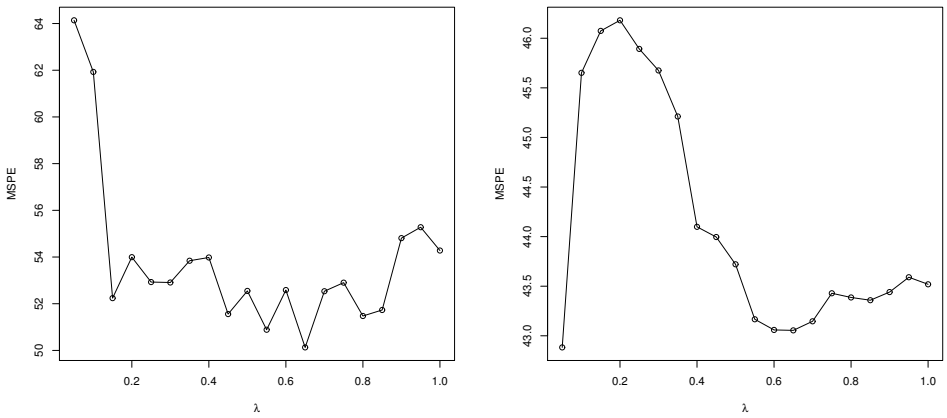

FIGURE 7: Mean squared prediction error (MSPE) with varying values of λ for Gaussian data in the usual sparse case, and the sparse case with outliers, respectively.

To illustrate this approach, we have analyzed a dataset generated from the example with Gaussian distributions in Section 6.2 for the case when $q = 2$. We have considered both cases, namely, without and with outliers. The results for varying values of λ using our method RLASSO-SPD over the grid of values from 0.05 to 1 with an increment of 0.05 are plotted below. The plot on the left panel (respectively, the right panel) of Figure 7 corresponds to the case when there are no outliers (respectively, there are outliers) in the data. The optimum value of λ for the two cases are 0.65 and 0.05, respectively.

7. CONCLUSIONS

In this paper, we proposed and studied a new method for robust multivariate regression based on data depth, and we investigated related theoretical properties. The method yielded competitive results in the numerical examples, and was found to be computationally quite stable because the estimator had contribution from all the observations instead of sub-samples from the data. With the LASSO, the method can also be used to carry out regressions for sparse data. Overall, the method based on the notion of data depth appears to be a novel approach, and applicable to a large variety of data sets.

ACKNOWLEDGEMENTS

We are thankful to the Editor, Associate Editor and two anonymous referees for their useful comments which led to an improvement of the method, and the paper.

BIBLIOGRAPHY

Agulló, J., Croux, C. and Van Aelst, S. (2008) The multivariate least-trimmed squares estimator. *J. Multivariate Anal.*, **99**, 311-338.

Alfons, A., Croux, C. and Gelper, S. (2013) Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Statist.*, **7**, 226-248.

Alfons, A. (2014) robustHD: Robust methods for high-dimensional data.

Cantoni, E. and Ronchetti, E. (2001) Resistant selection of the smoothing parameter for smoothing splines. *Statist. and Comput.*, **11**, 141-146.

Chakraborty, A. and Chaudhuri, P. (2014) On data depth in infinite dimensional spaces. *Ann. Inst. Statist. Math.*, **66**, 303-324.

Charnes, A., Cooper, W. W. and Rhodes, E. (1981) Evaluating program and managerial efficiency: An application of data envelopment analysis to program follow through. *Management Science*, **27**, 668-697.

Croux, C., and Haesbroeck, G. (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *J. Multivariate Anal.*, **71**, 161-190.

Dang, X., Serfling, R., and Zhou, W. (2009) Influence functions of some depth functions, and application to depth-weighted L-statistics. *J. Nonparametric Statist.*, **21**, 49-66.

Dang, X. and Serfling, R. (2010) Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties, *J. Statist. Plan. Inf.*, **140**, 198-213.

Donoho, D. L., and Huber, P. J. (1983) The notion of breakdown point. *In A Festschrift for Erich Lehmann, eds. P. J. Bickel, K. A. Doksum, and J. L. Hodges, Belmont, CA*, 157-184.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softw.*, **33**, 1-22.

García Ben, M., Martínez, E. and Yohai, V. J. (2006) Robust estimation for the multivariate linear model based on a τ-scale. *J. Multivariate Anal.*, **97**, 1600-1622.

Gervini, D. (2003) A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *J. Multivariate Anal.*, **84**, 116-144.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986) *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *Elements of Statistical Learning Theory*. Wiley, New York.

Johnson, R. A., and Wichern, D. W. (2007) *Applied Multivariate Statistical Analysis*. Prentice-Hall, New Jersey.

Kudraszow, N. L. and Maronna, R. A. (2011) Estimates of MM type for the multivariate linear model. *J. Multivariate Anal.*, **102**, 1280-1292.

Li, Y., Nan, B. and Zhu, J. (2015) Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, **71**, 354-363.

Liu, R. Y., Parelius, J. M. and Singh, K. (1999) Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.*, **27**, 783-840.

Liu, X. and Zuo, Y. (2014) Computing projection depth and its associated estimators. *Statist. Comput.*, **24**, 51-63.

Mahalanobis, P. C. (1936) On the generalized distance in statistics. *Proc. Natl. Inst. Sci. (Calcutta)*, **2**, 49-55.

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization *Computer J.*, **7**, 308-313.

Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection.* Wiley, New York.

Rousseeuw, P. J. and Hubert, M. (1999) Regression depth. *J. Amer. Statist. Assoc.*, **94**, 388-402.

Rousseeuw, P. J., Van Aelst, S., Van Driessen, K. and Agulló, J. (2004) Robust multivariate regression. *Technometrics*, **46**, 293-305.

Rousseeuw, P. J., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M. and Maechler, M. (2015) robustbase: Basic Robust Statistics.

Serfling, R. (2002) A depth function and a scale curve based on spatial quantiles. *In Statistical Data Analysis Based on the L1-Norm and Related Methods (Y. Dodge, ed.), Birkhaeuser*, 25-28.

Serfling, R. (2006) Depth functions in nonparametric multivariate inference. *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications (R. Y. Liu, R. Serfling, D. L. Souvaine, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, **72**, 1-16.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **58**, 267-288.

Van Aelst, S. and Willems, G. (2005) Multivariate regression S-estimators for robust estimation and inference. *Statist. Sinica*, **15**, 981-1001.

Van Aelst, S. and Willems, G. (2013) Fast and robust bootstrap for multivariate inference: The R package FRB. *J. Statist. Softw.*, **53**, 1-32.

Vardi, Y. and Zhang, C-H. (2000) The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. (USA)*, **97**, 1423-1426.

Yohai, V. J. (1987) High breakdown-point and high efficiency robust estimators for regression. *Ann. Statist.*, **15**, 642-656.

Wang, W., Liang, Y., and Xing, E. P. (2015) Collective support recovery for multi-design multi-response linear regression. *IEEE Trans. Inf. Theory*, **61**, 513-534.

Wilcox, R. R. (2012) *Introduction to Robust Estimation and Hypothesis Testing. (Third Edition).* Elsevier, New York.

Zuo, Y. and Serfling, R. (2000) General notions of statistical depth function. *Ann. Statist.*, **28**, 461-482.

Zuo, Y., Cui, H., and He, X. (2004) On the Stahel-Donoho estimators and depth weighted means of multivariate data. *Ann. Statist.*, **32**, 167-188.

Zuo, Y., Cui, H., and Young, D. (2004) Influence function and maximum bias of projection depth-based estimators. *Ann. Statist.*, **32**, 189-218.

Zuo, Y. and Cui, H. (2005) Depth weighted scatter estimators. *Ann. Statist.*, **33**, 381-413.

Zuo, Y. (2006) Multidimensional trimming based on projection depth. *Ann. Statist.*, **34**, 2211-2251.

APPENDIX

Proof of Proposition 1: Since PD is affine invariant with $PD(\mathbf{A}\mathbf{z}_i + \mathbf{b}) = PD(\mathbf{z}_i)$ for any non-singular $(p+q) \times (p+q)$ matrix, $\mathbf{A}$ and $\mathbf{b} \in \mathbb{R}^{p+q}$, we obtain that

$$\hat{\boldsymbol{\mu}}_{\mathbf{Az+b}} = \mathbf{A}\hat{\boldsymbol{\mu}}_{\mathbf{Z}} + \mathbf{b} \text{ and } \hat{\boldsymbol{\Sigma}}_{\mathbf{Az+b}} = \mathbf{A}\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}\mathbf{A}^T.$$

The affine equivariance of the depth-based location estimator $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ and the scatter estimator $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ imply the affine equivariance of the estimated regression coefficients $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\alpha}}$ in $\mathbf{T}_n^{PD}(\mathbf{z})$; see Lemma A.1 in Rousseeuw et al. (2004). □

Proof of Proposition 2: We first prove the Fisher consistency of the estimates based on PD. Since both $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ are affine equivariant, we obtain that for a distribution H which is centrally symmetric about $\mathbf{0}$, we have $E(\hat{\boldsymbol{\mu}}_{\mathbf{Z}}) = \mathbf{0}$. This assertion follows by taking $\mathbf{A}$ to be $-\mathbf{I}_{p+q}$ and $\mathbf{b} = \mathbf{0}$. Using a similar line of argument and the fact that $E[w_2\{\delta(\mathbf{Z})\}Z_k Z_{k'}] = 0$ for $k \neq k'$ (which follows from central symmetry of $\mathbf{Z}$), we have $E(\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}) = \kappa Cov(\mathbf{Z})$ (also see Zuo and Cui, 2005).

From Zuo and Cui (2005), it follows that $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ is a consistent estimator of $\kappa\boldsymbol{\Sigma}_{\mathbf{Z}}$. Using Fisher consistency $\hat{\mathbf{B}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{XY}} \xrightarrow{P} (\kappa\boldsymbol{\Sigma}_{\mathbf{XX}})^{-1}(\kappa\boldsymbol{\Sigma}_{\mathbf{XY}}) = \mathbf{B}$ as $n \to \infty$. Again, using Fisher consistency and consistency of $\hat{\boldsymbol{\mu}}_{\mathbf{Z}} = (\hat{\boldsymbol{\mu}}_{\mathbf{Y}}^T, \hat{\boldsymbol{\mu}}_{\mathbf{X}}^T)^T$ (Zuo, Cui and Young, 2004), we obtain $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}_{\mathbf{Y}} - \hat{\mathbf{B}}^T\hat{\boldsymbol{\mu}}_{\mathbf{X}} \xrightarrow{P} \boldsymbol{\alpha}$ as $n \to \infty$. □

Proof of Proposition 3: First note that the BP of $\mathbf{T}_n^{PD}(\mathbf{z})$ depends on the BP of $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ which are constructed using PD. Theorem 3.1 of Zuo, Cui and Young (2004) gives the BP of $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ to be 1/2. This fact follows quite easily by combining the BP of the median and the MAD (both of which have a BP of 1/2). On the other hand, Theorem 3.7 of Zuo and Cui (2005) gives the BP of $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ to be $\lfloor (n-p-q+1)/2 \rfloor / n$ (with the choice of $k = p+q$).

Let $\mathbf{z}_n^*$ denote the new dataset obtained by replacing m observations (with $m < \min\{\lfloor (n-p-q+1)/2 \rfloor, \lfloor n/2 \rfloor\}$) from the original dataset $\mathbf{z}_n$ by arbitrary values. Note that $\|\hat{\mathbf{B}}(\mathbf{z}_n^*)\| = \|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}(\mathbf{z}_n^*)\boldsymbol{\Sigma}_{\mathbf{XY}}(\mathbf{z}_n^*)\| \leq \|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}(\mathbf{z}_n^*)\|\|\boldsymbol{\Sigma}_{\mathbf{XY}}(\mathbf{z}_n^*)\|$, we have $\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}(\mathbf{z}_n^*)\| = \lambda_{min}(\hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}(\mathbf{z}_n^*))^{-1}$ and $\|\hat{\boldsymbol{\Sigma}}_{\mathbf{XY}}(\mathbf{z}_n^*)\| \leq \|\hat{\boldsymbol{\Sigma}}(\mathbf{z}_n^*)\| \leq \lambda_{max}(\hat{\boldsymbol{\Sigma}}(\mathbf{z}_n^*))$, where λ_{min} and λ_{max} are the minimum and the maximum eigenvalue, respectively. They are both bounded because $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ does not breakdown for $m < \lfloor (n-p-q+1)/2 \rfloor$, and hence $\|\hat{\mathbf{B}}(\mathbf{z}_n^*)\|$ is bounded. Furthermore, note that $\|\hat{\boldsymbol{\alpha}}(\mathbf{z}_n^*)\| \leq \|\hat{\boldsymbol{\mu}}_{\mathbf{Y}}(\mathbf{z}_n^*)\| + \|\hat{\mathbf{B}}(\mathbf{z}_n^*)\|\|\hat{\boldsymbol{\mu}}_{\mathbf{X}}(\mathbf{z}_n^*)\|$, which is bounded because m is also assumed to be less than $\lfloor n/2 \rfloor$. The proof then follows easily by combining all these ideas. □

Proof of Proposition 4: Recall Fisher consistency of the estimates based on PD from the proof of Proposition 2. Note the following: $IF(\mathbf{z}; \hat{\mathbf{B}}, H) = IF(\mathbf{z}; \hat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, H)$, $IF(\mathbf{z}; \hat{\boldsymbol{\alpha}}, H) = IF(\mathbf{z}, \hat{\boldsymbol{\mu}}_{\mathbf{Y}}, H)$ and $IF(\mathbf{z}; \hat{\boldsymbol{\Sigma}}_{\mathbf{e}}, H) = IF(\mathbf{z}; \hat{\boldsymbol{\Sigma}}_{\mathbf{YY}}, H)$. This follows from Lemma A.3 of Rousseeuw et al. (2004). Assuming spherical symmetry of the joint distribution, the expression for IF of $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ is given in Theorem 3.4 of Zuo, Cui and Young (2004), and for $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ it is given in Corollary 3.2 of Zuo and Cui (2005):

$$IF(\mathbf{z}; \hat{\boldsymbol{\mu}}_{\mathbf{Z}}, H) = \frac{K_0(\mathbf{z}/\|\mathbf{z}\|) + w_1\{(1+\|\mathbf{z}\|)^{-1}\}\mathbf{z}}{\int w_1\{(1+\|\mathbf{u}\|)^{-1}\}\, dH(\mathbf{u})}, \text{ and}$$

$$IF(\mathbf{z}; \hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}, H) = \frac{t_1(\|\mathbf{z}\|)\mathbf{z}\mathbf{z}^T/\|\mathbf{z}\|^2 + t_2(\|\mathbf{z}\|)\mathbf{I}_{p+q}}{c_0}.$$

The expressions of IF in this result now follow from these expressions and using the partition given by (1). □

Lemma 1: Under the conditions of Proposition 3, the terms c_0, K_0 and the functions t_1, t_2 are bounded.

Proof of Lemma 1: We first state related expressions and give conditions under which they are bounded. Define $\mathbf{U} = \mathbf{Z}/\|\mathbf{Z}\|$, where $\mathbf{Z} \sim H$, the spherical distribution function; $m_0 = MAD(Z_1)$, and p is the density of Z_1. Without loss of generality, we take m_0 to be 1. We denote the first derivative of the function w_i by $w_i^{(1)}$ for $i = 1, 2$. Then:

- $s_0(z) = 1/(1+z)$, and $0 < s_0(z) \leq 1$ for any z.
- $s_i(z) = E\{U_1^{2i-2}\text{sign}(|U_1|z - m_0)\}$, bounded since $\|\mathbf{U}\| \leq 1$ and $|\text{sign}(u)| \leq 1$ for $i = 1, 2$. Here $\text{sign}(u)$ is the univariate sign function which is -1 or 0 or $+1$ accordingly as $u <$ or $=$ or > 0.
- $c_0 = E[w_2\{s_0(\|\mathbf{Z}\|)\}]$, and bounded in virtue of the facts that $0 < s_0(z) \leq 1$ and w_2 is bounded in the range $(0, 1)$.
- $c_1 = E[\|\mathbf{Z}\|^2 w_2\{s_0(\|\mathbf{Z}\|)\}]/\{(p+q)c_0\}$, and it is bounded because $w^\dagger(z) = z^2 w_2\{s_0(z)\}$ is bounded as we explain next. We take $v = c^{-1}s_0(z)$, and obtain

$$w^\dagger(v) = \begin{cases} \left(\dfrac{1}{vc} - 1\right)^2 \dfrac{1}{vc^2} \dfrac{\exp\{-k(1-v^4)^4\} - \exp(-k)}{1 - \exp(-k)}, & 0 < v < 1, \\ \left(\dfrac{1}{vc} - 1\right)^2 \dfrac{1}{vc^2}, & 1 < v < 1/C. \end{cases}$$

  The function is continuous over the interval $(0, 1/C]$. However, $w^\dagger(v)$ is of the form $0/0$ at $v = 0$. By L'Hôpital's rule, we can argue that $w^\dagger(0) = 0$, and hence it is bounded over the range of v.
- $c_2 = E[\|\mathbf{Z}\| s_0^2(\|\mathbf{Z}\|) w_2^{(1)}\{s_0(\|\mathbf{Z}\|)\}]/\{4p(1)\}$, which is bounded because firstly, we have $0 < z s_0(z) = 1/(1/z + 1) \leq 1$. Moreover, $w_2^{(1)}\{s_0(z)\}$ is of the form $16k/ct(1-t^4)\exp\{-k(1-t^4)^4\}$ for $0 < t < 1$ and $t = c^{-1}s_0(z)$, which is bounded in the unit interval.
- $c_3 = E[\|\mathbf{Z}\|^3 s_0^2(\|\mathbf{Z}\|) w_2^{(1)}\{s_0(\|\mathbf{Z}\|)\}]/\{4p(1)\}$, and is bounded by virtue of the fact that the following function:

$$z w_2^{(1)}\{s_0(z)\} = (16k/c^2)(1-ct)(1-t^4)^3 \exp\{-k(1-t^4)^4\},$$

  where $t = c^{-1}s_0(z)$ for $0 < t < 1$ is continuous in the bounded interval $(0, 1)$, and also note that $z s_0(z) \leq 1$.

Now, we consider the quantities in the expression of the IFs as follows:

- c_0 is stated above, and has been argued to be bounded.
- $K_0 = 1/\{2h(0)\} \int_{\mathbb{R}^d} |x_1| w_1^{(1)}\{(1 + \|\mathbf{x}\|)^{-1}\}/(1 + \|\mathbf{x}\|)^2 \, dH(\mathbf{x})$. Note that $|x_1| w_1^{(1)}\{(1 + \|\mathbf{x}\|)^{-1}\} \leq \|\mathbf{x}\| w_1^{(1)}\{(1 + \|\mathbf{x}\|)^{-1}\}$, and we argue below that $z w_1^{(1)}\{s_0(z)\}$ is bounded.

$$z w_1^{(1)}\{s_0(z)\} = \frac{4k(1-ct)(1-t^2)\exp\{-k(1-t^2)^2\}}{c^2 t^2},$$

  where $t = c^{-1}s_0(z)$ for $0 < t < 1$ is clearly unbounded at $t = 0$. However, the choice of k is in our hands, and we can choose it appropriately to make this quantity arbitrarily close to 0.

- $t_1(z) = c_3\{s_2(z) - \frac{s_2(z)-s_1(z)}{p+q-1}\} + z^2 w_2\{s_0(z)\}$. Recall that c_3 is bounded, and we have argued above that $z^2 w_2\{s_0(z)\}$ is bounded.
- $t_2(z) = c_3 \frac{s_2(z)-s_1(z)}{p+q-1} - c_1 c_2 s_1(z) - c_1 w_2\{s_0(z)\}$ is bounded in view of the facts stated above.

□

Received 9 July 2009
Accepted 8 July 2010