The Canadian Journal of Statistics
Vol. 45, No. 2, 2017, Pages 164–184
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*
[2] *CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*

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 subsamples 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 data set. 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* 45: 164–184; 2017 © 2017 Statistical Society of Canada

Résumé: Les auteurs développent une méthode robuste de régression multivariée basée sur des estimateurs robustes de la localisation et de la covariance conjoints pour la variable réponse et les covariables, et qui sont fondés sur la notion de profondeur. L'estimateur de la régression multivariée s'avère affine-équivariant, offre le meilleur point de rupture de tous les estimateurs affine-équivariants, et possède une fonction d'influence bornée par les variables explicatives et réponse. Afin d'accroître l'efficacité de l'estimateur, les auteurs proposent un estimateur repondéré selon une distance de Mahalanobis robuste appliquée au vecteur des résidus. D'un point de vue pratique, leur méthode s'avère plus stable que les méthodes existantes construites en utilisant des sous-échantillons des données. En plus d'être réalisable numériquement, la technique de régression multivariée proposée offre une performance supérieure à plusieurs méthodes de régression multivariée robustes quand elle est utilisée sur un échantillon étalon réel et sur des échantillons simulés. Lorsque la dimension des données est élevée par rapport à la taille d'échantillon, il est possible de mettre à profit des notions de profondeur pour construire un estimateur robuste dans une situation clairsemée. *La revue canadienne de statistique* 45: 164–184; 2017 © 2017 Société statistique du Canada

1. INTRODUCTION

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

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

Additional supporting information may be found in the online version of this article at the publisher's website.

* *Author to whom correspondence may be addressed.*
E-mail: duttas@iitk.ac.in, tijahbus@gmail.com

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

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

with $\boldsymbol{\Sigma}_{\mathbf{YX}} = \boldsymbol{\Sigma}_{\mathbf{XY}}^\top$. 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 $\mathbf{z}_i = (\mathbf{y}_i^\top, \mathbf{x}_i^\top)^\top$, where $\mathbf{y}_i$ is the response vector and $\mathbf{x}_i$ is the vector of covariates for $1 \le i \le n$. Let $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ denote the estimators of $\boldsymbol{\mu}_{\mathbf{Z}}$ and $\boldsymbol{\Sigma}_{\mathbf{Z}}$, respectively. The resulting estimates of $\mathbf{B}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma}_{\mathbf{e}}$ are

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

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

The usual method of moments leads to estimators identical to those obtained using the least squares method. However it is well-known that moment-based estimators are extremely sensitive to outliers. Thus a common practice is 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, & Singh, 1999; Serfling, 2006). The use of depth in building such estimators is quite natural and simple because depth has a "centre outward ordering." In other words depth has this appealing property that it is maximized at the centre of the data cloud, and decreases along any ray from that centre. 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 article 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 the minimum covariance determinant (MCD) (Rousseeuw et al., 2004), multivariate least trimmed squares (MLTS) (Agulló, Croux, & Van Aelst, 2008), S estimators (S) (Van Aelst & Willems, 2005), τ estimator (TAU) (García Ben, Martínez, & Yohai, 2006), and modified M estimators (MM) (Yohai, 1987; Kudraszow & Maronna, 2011). Regression depth (RD), introduced by Rousseeuw & 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 the multivariate data points.

The rest of the article 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 contains some 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 (Zuo, Cui, & He, 2004; Zuo, Cui, & Young, 2004; Serfling, 2006), respectively, as follows:

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

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 nondecreasing, nonnegative weight functions. The two weight functions w_1 and w_2 may not necessarily be the same. Consider the following weight functions:

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

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, & 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 larger depth, and this balances efficiency with robustness. The other half of the points with smaller depth could be viewed as outliers, so a lower weight is assigned. One could also consider other weight functions satisfying appropriate properties (Zuo & 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 in Equation (2) are then constructed using projection depth (PD) (Zuo & 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 $O(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1}\{\mathbf{u}^\top\mathbf{x} - \mu(F_u)\}/\sigma(F_u)$ and F_u is the distribution function of $\mathbf{u}^\top\mathbf{X}$. Here $\mu(F_u)$ and $\sigma(F_u)$ are univariate location and scale functionals, respectively, corresponding to $\mathbf{u}^\top\mathbf{X}$. Proofs of the results in Sections 2.1–2.4 are provided in the Appendix.

2.1. Affine Equivariance

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

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 to 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* $\mathbf{Z} = (\mathbf{Y}^\top, \mathbf{X}^\top)^\top$ *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* $(\mathbf{B}^\top, \boldsymbol{\alpha})$.

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

2.3. Breakdown Point

The finite sample breakdown point (BP) (Donoho & 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 defined in Equation (3), that are also discussed by Zuo, Cui, & He (2004). The following result 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, median absolute deviation [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 BP of the median and the MAD (Zuo & Serfling, 2000). To compare alternatives 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 $\min(n-h+1, h-p-q+1)/n$, (Agulló, Croux, & Van Aelst, 2008, p. 315). The BP of the MM estimator is at least $\min\{\text{BP of initial estimator}, (\lfloor n/2 \rfloor - k_n)/n\}$, where $k_n \geq p+q-1$ (Kudraszow & Maronna, 2011, Theorem 3). Let $k(\mathbf{Z}_n)$ 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 nonnegative, symmetric, and nondecreasing function on $[0, c]$ and constant on $[c, \infty)$ for some constant c with $b = E[\rho(\cdot)]$, and assume $k(\mathbf{Z}_n) < \lceil n - nr \rceil$ holds. For S estimators the BP is $\min\{nr, \lceil n - nr \rceil - k(\mathbf{Z}_n)\}/n$ (Van Aelst & Willems, 2005, p. 984). A lower bound for the BP of the τ estimator is $\min\{(1-\eta) - (h/n), \eta\}$; see García Ben, Martínez, & Yohai (2006, p. 1605) for more details concerning 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 centre-outward ordering using PD, whereas the usual trimming is based on ranks. See also the discussion in the first paragraph on 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 $\mathbf{Z} = (\mathbf{Y}^\top, \mathbf{X}^\top)^\top$ 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 Equation (3). The IFs of $\hat{\mathbf{B}}$, $\hat{\boldsymbol{\alpha}}$, and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ based on PD with (μ, σ) = (Median, MAD) are*

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

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

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

Expressions for c_0, t_1, t_2, and K_0 are given in the Appendix.

It is quite easy to derive the IFs 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, & He (2004) gives the IF of $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$, and Theorem 3.3 of Zuo & Cui (2005) gives the IF of $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$. By combining these two expressions we can 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,

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

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

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

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

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

Here $\boldsymbol{\Sigma}$ is the covariance matrix of the residual vector. The related constants are defined in Theorem 4 of Kudraszow & Maronna (2011). Assuming the joint density to be unimodal and spherically symmetric the IF for S estimators is

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

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

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

Here $\boldsymbol{\Sigma}$ is the covariance matrix of the residual vector. The related constants are identified on pp. 1606–1607 of García Ben, Martínez, & Yohai (2006).

The expressions in Proposition 4 are all bounded (see Lemma 1 in the Appendix). By the submultiplicative property of a matrix norm we get $\|\mathbf{x}\mathbf{y}^\top\| \leq \|\mathbf{x}\|\|\mathbf{y}^\top\| \leq \|\mathbf{z}\|^2$. 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 should 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, whereas the expressions for the IFs of the last four methods require the finiteness of $E_H[\mathbf{x}\mathbf{x}^\top]^{-1}$ and the IF function for the MM estimator remains unbounded.

When the estimators $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\alpha}}$ are constructed using spatial depth (SPD) (Vardi & Zhang, 2000; Serfling, 2002) they are referred to 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 $\text{SPD}(\mathbf{x}, F) = 1 - \left\| E_F\{\mathbf{u}(\mathbf{x} - \mathbf{X})\}\right\|$, 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, nonoutlier). Now the masking BP of SPD is $\lceil n(1-\lambda)/2\rceil/n$ (Theorem 3.5 of Dang & 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, & Zhou (2009). Furthermore Dang, Serfling, & Zhou (2009) stated a general result for depth-weighted location estimators and the IF has a very complicated expression for general depth functions. By 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 $\hat{\mathbf{y}}_i = \hat{\boldsymbol{\alpha}} + \hat{\mathbf{B}}^\top\mathbf{x}_i$. Consider the standardized residuals $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}^{-1/2}\hat{\mathbf{e}}_i$, and without confusion we denote them again by $\hat{\mathbf{e}}_i$. Recall the expression for $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ given in Equation (2), which is a robust estimator for $\boldsymbol{\Sigma}_{\mathbf{e}}$. For these $\hat{\mathbf{e}}_i$ good leverage points will have a small value of $\|\hat{\mathbf{e}}_i\|$, whereas vertical outliers and bad leverage points will have a large value of $\|\hat{\mathbf{e}}_i\|$. So the main objective is to retain observations about the point $\mathbf{0}$, and seek to discard the remaining points. In other words we calculate the Mahalanobis distances (Mahalanobis, 1936) for the residual vectors with $\mathbf{0}$ as the centre 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 centre and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ as the scatter matrix for $1 \leq i \leq n$; $d_{(1)} \leq \cdots \leq d_{(n)}$ denote their values in ascending

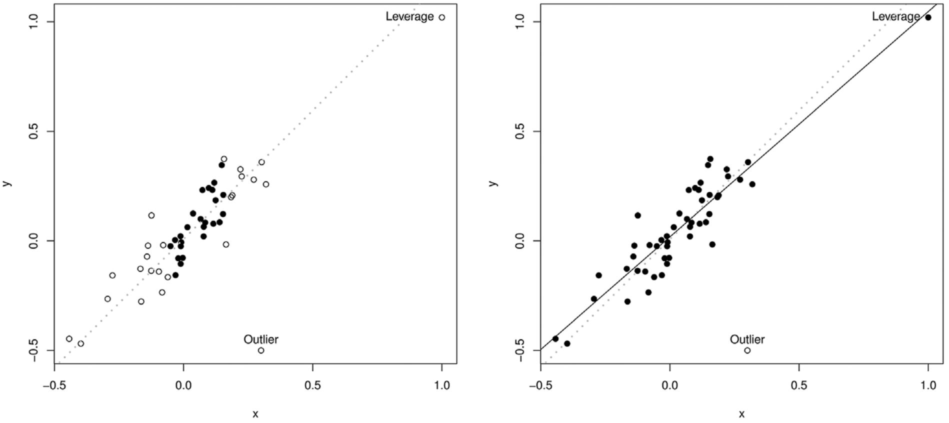

FIGURE 1: A plot illustrating our two-step procedure. In the left panel, the black points are the observed data with the largest depth and the dotted line represents the fitted line based on $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\alpha}}$. In the right panel, after re-weighting, we indicate the newly fitted line in bold together with those data that contribute to this new fit based on $\hat{\mathbf{B}}^R$ and $\hat{\boldsymbol{\alpha}}^R$.

order. Define $i_0 = \max\{i : d^2_{(i)} < \chi^2_{q,1-\alpha}\}$ for a fixed α (say, 0.025) and let $\alpha_n = \max_{i>i_0}\{G_q(d^2_{(i)}) - (i-1)/n\}_+$ with $\{\cdot\}_+$ denoting the positive part, and G_q the cumulative distribution function of χ^2_q, the chi-squared distribution with q degrees of freedom (df). Here $\chi^2_{q,1-\alpha}$ represents the $1-\alpha$ quantile of the χ^2_q distribution. Now the observations corresponding to the largest $\lfloor n\alpha_n \rfloor$ distances are identified as outliers, whereas the remaining points are labelled nonoutliers. For the observations selected as nonoutliers, 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 because $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ is an affine equivariant estimator of $\boldsymbol{\Sigma}_{\mathbf{e}}$.

To get a better understanding of how the re-weighting step works we constructed two plots. We first generated a data set of size 48, where the regressors came from a normal distribution with mean 0 and variance 0.2, whereas the response was obtained by adding an error term which also had a standard normal distribution with variance 0.1. We next added a good leverage point at (1, 1) and an outlier at (0.3, −0.5), making the total sample size 50.

The panel on the left in Figure 1 shows the simulated points and the fitted line based on the PD-weighted estimators with the 0–1 weight function; the cutoff is set to be the median of the PD values, which ensures that the half of the points with the highest PDs are selected. The black points are the data that have a weight equal to 1, and they contribute significantly to the regression estimate. We see that the leverage point as well as the outlier in the data set have been omitted because our method is based on depth. The re-weighting step rotates the line clockwise—see the bold line in the right panel of Figure 1—and includes the good leverage point as a part of the fit.

3.2. Practical Aspects

We consider PD together with the median and MAD as univariate robust estimators (Zuo & Serfling, 2000). The theoretical version of PD has good robustness properties, but its computation poses an additional difficulty (Liu & Zuo, 2014). For data in $\mathbb{R}^d$ PD involves calculating a supremum in 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 function `zdepth` developed by Wilcox (2012) for computing PD. The algorithm is based on the well-known Nelder–Mead or "downhill simplex" method (Nelder & Mead, 1965) for solving the maximization problem in the outlyingness function $O(\mathbf{x}, F)$ in $\mathbb{R}^d$.

The notion of SPD eases this computational burden. SPD can be calculated exactly because 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 weights are assigned to all those unit vectors. We implement both methods in our procedure, and present a comparative study in Section 4.

4. NUMERICAL WORK

Our numerical study is motivated by the examples considered by Agulló, Croux, & Van Aelst (2008). We performed a study of the efficiency and robustness of the overall procedure. We used `R` (R Core Team, 2015) code from the `robustbase` package (Rousseeuw et al., 2015) for MCD, and the `FRB` package (Van Aelst & Willems, 2013) for both S and MM. Code for MLTS/RMLTS is available at http://www.econ.kuleuven.be/public/NDBAE06/programs/mlts/mlts.r.txt, whereas the necessary functions for TAU were obtained from Prof. Victor J. Yohai. We have made our `R` code available at Section 8. For the sake of comparison we also studied the performance of our method based on SPD (labelled DWR-SPD). We calculated the mean squared error (MSE) of the slope matrix $\hat{\mathbf{B}}$ by computing an average over the MSE of each element of $\hat{\mathbf{B}}$ over the random realizations of the data. The matrix norm used here is the usual component-wise l_2 norm. In our experiments we observed occasional instances of singularity (the weights became zero in the calculation of the weighted covariance) for S estimators; therefore, we have not reported 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 labelled 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, & 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 for 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 them 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 accounting for the intercept term. The remaining $p - 1$ explanatory variables were generated from the following distributions:

(i) The multivariate standard normal distribution;
(ii) The multivariate standard Cauchy distribution;
(iii) The multivariate uniform distribution on $(-1, 1)^p$.

TABLE 1: Estimated mean squared errors (with corresponding estimated standard errors) of various robust multivariate regression estimators of $\hat{\mathbf{B}}$, for three choices of error distribution.

Error	Explanatory variable	Estimation method								
distribution	distribution	DWR-PD	RDWR-PD	DWR-SPD	RDWR-SPD	MLTS	RMLTS	MM	RMCD	TAU
Gaussian	Gaussian	8.5965	3.6662	8.5965	2.5833	8.1382	3.2503	2.1851[a]	2.8920	2.2177[b]
		(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[a]	1.6228	0.0799[b]
		(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[a]	7.4889	6.4733[b]
		(0.1043)	(0.0947)	(0.1043)	(0.0877)	(0.1547)	(0.0975)	(0.0795)	(0.0864)	(0.0803)
Cauchy	Gaussian	6.8035	4.7128	6.8035	4.1929[b]	5.2219	5.0848	5.0376	6.5701	4.1126[a]
		(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[b]	0.2272	2.5505	0.3000[a]
		(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[a]	15.2026	14.6114	15.1219	19.4509	12.3017[b]
		(0.1133)	(0.1138)	(0.1133)	(0.1095)	(0.1232)	(0.1208)	(0.1229)	(0.1394)	(0.1108)
Uniform	Gaussian	5.0259	1.5050	5.0259	0.8124[a]	4.9035	1.6965	0.8229[b]	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[a]	0.5426	0.0262[b]
		(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[a]	15.2485	5.5557	2.5225[b]	2.9651	2.5831
		(0.0800)	(0.0601)	(0.0800)	(0.0499)	(0.1234)	(0.0745)	(0.0502)	(0.0544)	(0.0508)

[a] Denotes the smallest estimated MSE.
[b] The second smallest value.

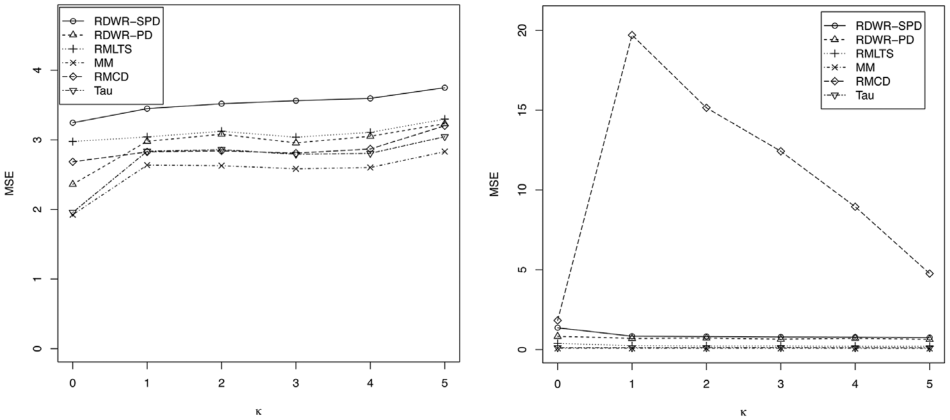

FIGURE 2: Estimated MSEs of $\hat{\mathbf{B}}$ for normal (left panel) and Cauchy (right panel) distributed explanatory variables when $\lambda = 0$.

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

From Table 1 it is clear that the TAU estimator yielded the best overall performance, closely followed by the MM estimator. The re-weighted version of the SPD-based estimator, labelled RDWR-SPD, also led to competitive performance in some scenarios when both the response and the explanatory variables were uniformly distributed. This improved performance may be due to the fact that SPD led to a "centre outward ordering" of observations from a uniform distribution; see also 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 smaller estimated values of MSE compared to the corresponding estimates obtained using RMCD.

4.2. Finite Sample Robustness

To study the finite sample robustness of our proposed estimators we carried out simulations with contaminated data sets. The parameters were chosen to be $\boldsymbol{\alpha} = 1$ 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 $p - 1$ explanatory variables were generated according to $\mathrm{N}(\lambda\sqrt{\chi^2_{p-1,0.99}}, 1.5)$, whereas the new q response variables were generated from $\mathrm{N}(\kappa\sqrt{\chi^2_{q,0.99}}, 1.5)$. Here $\chi^2_{r,0.99}$ represents the 0.99 quantile of the chi-squared distribution with r df for $r = p - 1$ and q. We considered λ and κ values from $\{0, 1, 2, 3, 4, 5\}$. 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 the TAU estimators are uniformly more efficient than all other methods. For data with normal explanatory variables RDWR-PD exhibited substantial improvement over RDWR-SPD. As RMLTS led to performance that was comparable to RDWR-PD, the former method did have an edge in some situations. In the case when $\lambda = 0$ and the explanatory variables followed a

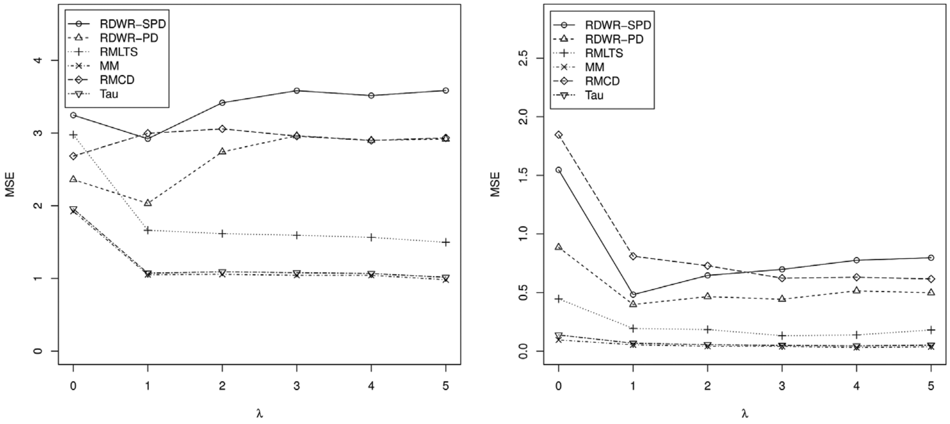

FIGURE 3: Estimated MSEs of $\hat{\mathbf{B}}$ for normal (left panel) and Cauchy (right panel) distributed explanatory variables when $\kappa = 0$.

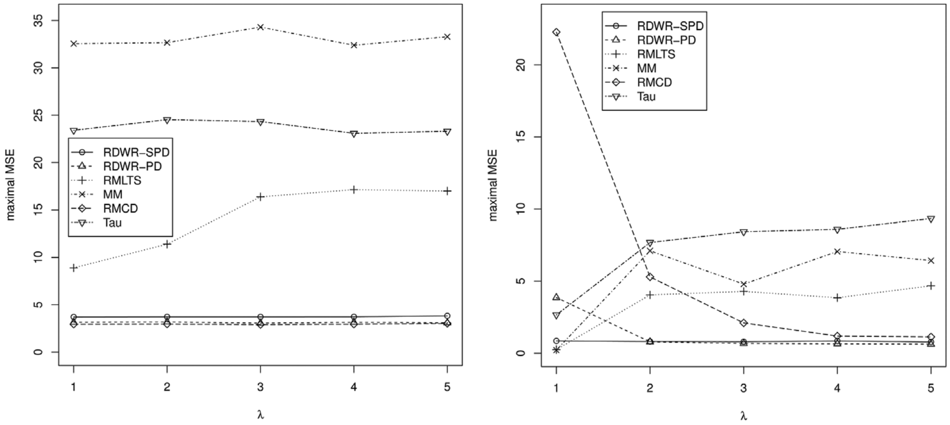

FIGURE 4: Maximal estimated MSEs of $\hat{\mathbf{B}}$ in the contaminated data for normal (left panel) and Cauchy (right panel) distributed explanatory variables, as a function of λ.

Cauchy distribution RMCD led to a surprisingly large estimated MSE; see the right panel of Figure 2.

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

In the left panel of Figure 4 we observe a lack of robustness for both the MM and TAU estimators. The performance of RMLTS is also quite poor. Clearly both our depth-based methods as well as RMCD led to uniformly smaller estimated values of MSE. The situation improved when the explanatory variables had a Cauchy distribution (the right panel); the MSE decreased considerably for both the MM and TAU estimators, but depth-based methods were clearly more robust. In fact both RDWR-PD and RDWR-SPD yielded MSE estimates that were close to 0 for all values of λ.

5. DATA ANALYSIS

We analyzed a benchmark data set to illustrate the usefulness of the methods. The diagnostic plots in Figure 5 show combined 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 were calculated using projection outlyingness. We plotted the robust distance of the residuals versus the robust distance of the predictor variables. The horizontal and vertical lines on each plot indicate the square roots of the 0.975 quantiles of χ^2_p and χ^2_q, respectively.

5.1. School Data

The aim of this study conducted by Charnes, Cooper, & Rhodes (1981) was to explain the scores observed on three tests written at 70 schools using five explanatory variables. The three test scores were (i) the total reading score; (ii) the total mathematics score; and (iii) the Coopersmith Self-Esteem Inventory; both the reading and the mathematics scores were measured by the Metropolitan achievement test. The explanatory variables included the education level of the mother, measured as the percentage of high school graduates among female parents, the highest observed occupation rating 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. For these data $p = 5$, $q = 3$, and $n = 70$.

In these data the RDWR-PD and RDWR-SPD methods performed on par with the four alternative robust estimators; all methods uniformly classified observation 59 as a “bad leverage” point; see Table 2 and Figure 5. The MM and RMLTS estimators identified observation number 44 as an additional “vertical outlier,” whereas RDWR-PD and RDWR-SPD highlighted observation 47. RMCD failed to identify observation numbers 12 and 35 as outliers, whereas TAU overlooked the latter.

To understand the relative importance of the vertical outliers we prepared pairwise plots of the response variables for observations 12, 21, 35, 44, and 47; see Figure 6. The influence of observations 44 and 47 is quite evident.

TABLE 2: Index numbers of “bad” points in the school data.

Estimation 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

6. ROBUST MULTIVARIATE REGRESSION FOR SPARSE DATA

6.1. The Re-Weighted LASSO Estimator

With the modern advances in statistical methods data that have a dimension greater 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 & Chaudhuri (2014)

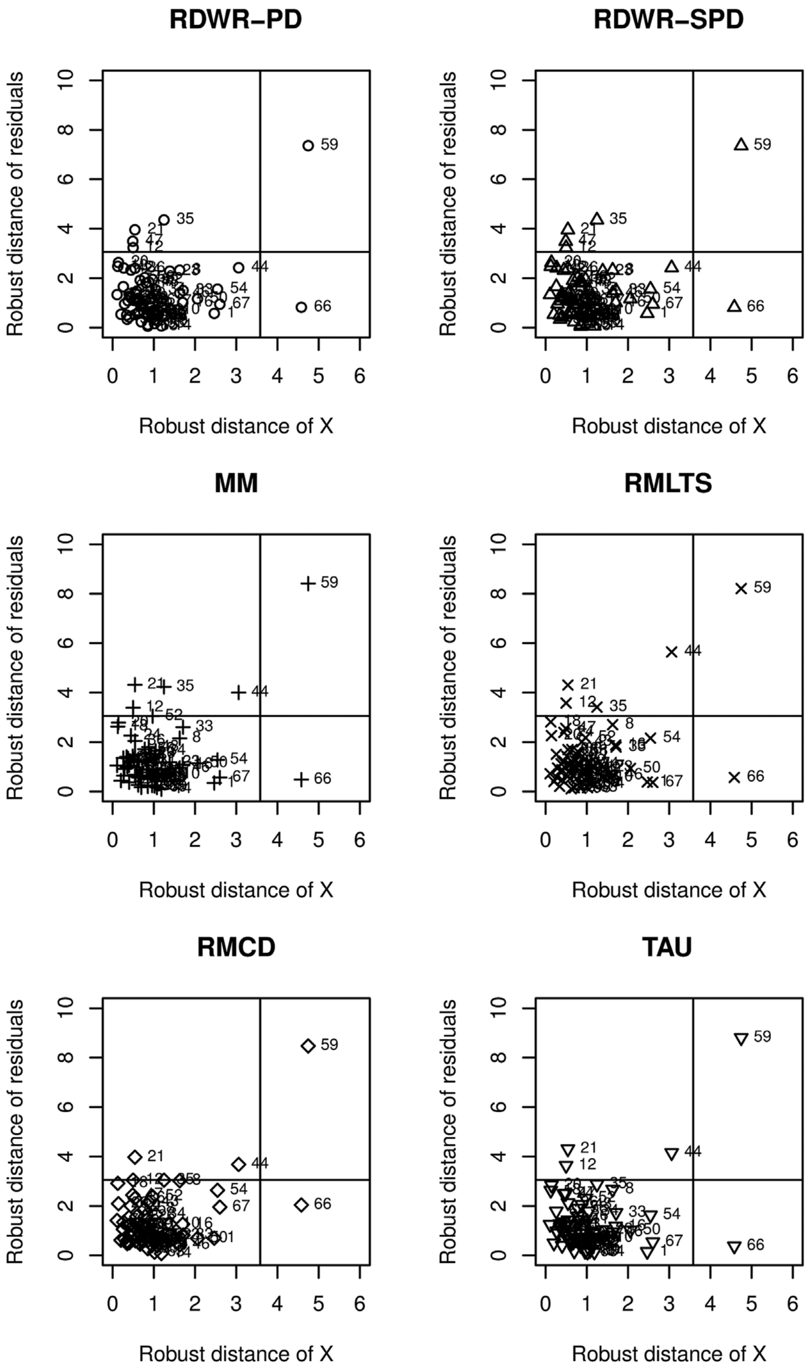

FIGURE 5: Diagnostic plots for various robust estimation methods showing robust residuals versus robust distances of the explanatory variables for the school data. The vertical and horizontal cutoff lines are set at the square roots of the 0.975 quantiles of χ^2_5 and χ^2_3 , respectively.

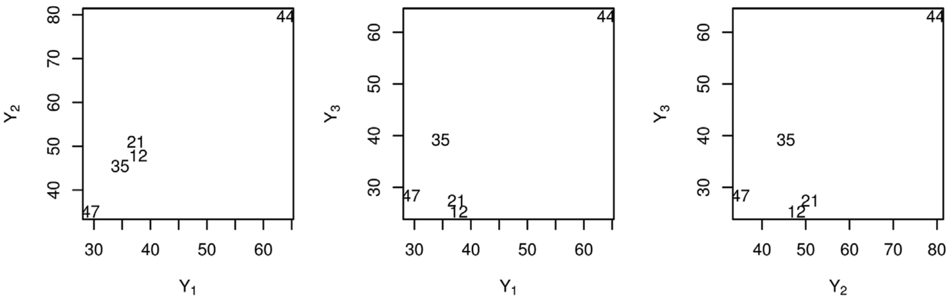

FIGURE 6: Pairwise plots of components of the response variable for observation numbers 12, 21, 35, 44, and 47 in the school data.

showed that for a large class of infinite-dimensional distributions, the notion of SPD transforms all values to 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 explore the area of robust regression for sparse data.

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

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

with appropriate weights w_i for $1 \le i \le n$; see pp. 387–389 of Johnson & Wichern (2007) for a derivation of this least squares minimization problem. In Section 2 we 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 the usual multivariate regression. LASSO allows us to carry out 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 **B** to obtain

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

The constant $\lambda > 0$ controls the effect of the penalty; $\|\mathbf{B}\|_{l_1} = \sum_{kl} |b_{kl}|$ is the l_1 matrix norm of **B**. For details on the formulation of multi-response sparse linear regression and some of its variants see Li, Nan, & Zhu (2015) and Wang, Liang, & Xing (2015). The weights w_i are calculated based on SPD, and the weight functions are specified in Equation (3). This approach, which we call LASSO-SPD, can be used directly for robust regression with sparse data. In fact we are not restricted to a univariate response because LASSO has the necessary 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, & Tibshirani, 2010), or using the mixed co-ordinate descent algorithm (Li, Nan, & Zhu, 2015).

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

TABLE 3: Estimated mean squared errors (with corresponding estimated standard errors) of various robust multivariate regression estimators of $\hat{\mathbf{B}}$ for sparse data.

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

[a] Denotes the smallest estimated MSE.
[b] The second smallest value.

6.2. A Numerical Evaluation

The R code for sparse LTS and LASSO can be found in the packages `robustHD` (Alfons, 2014) and `glmnet` (Friedman, Hastie, & Tibshirani, 2010), respectively. In our implementation of sparse LTS and glmnet we fixed 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 obtained an estimate of $\mathbf{B}$. We then computed the MSE of this estimate over this sequence of values of λ, and the minimum value of MSE observed is reported in Table 3. Each experiment was replicated 100 times.

Following Alfons, Croux, & Gelper (2013) we generated a high-dimensional data set with 20 observations from a p-dimensional distribution with $p = 1{,}000$. Element (i, j) of the variance–covariance matrix was $0.5^{|i-j|}$, which gave rise to correlated predictor variables. The coefficient vector was made sparse by fixing the first 20 components to be 1, and the rest 0. 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 assumed 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 were the p-dimensional constant vectors of 10's and -15's, respectively. For the case when $q = 2$ we added a second coefficient vector by setting the last 20 components equal to 0, and all the rest to be 1. We also considered the same four examples described in this paragraph with a bivariate response.

The values reported in Table 3 indicated 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 lower half of this table RLASSO-SPD outperformed the LASSO considerably, whereas sparse LTS is not applicable to such data. In terms of computational time the average computing time for LASSO and hence for RLASSO-SPD was about 1 s per iteration. The computation was dominated by LASSO because SPD involved only the computation of averages of unit vectors. This calculation was quite fast compared with sparse LTS, which took around 15–20 s 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 chose to select λ by minimizing the estimated prediction error using the idea of cross-validation (e.g., Hastie, Tibshirani, & Friedman, 2009). In the sparse setup n is quite small compared to p and we used 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 observation for prediction. To prevent outliers from affecting the choice of λ a robust prediction error is desirable (e.g., Cantoni & Ronchetti, 2001). For a given value of λ we obtained a set of n prediction error vectors. We used the approach described in Section 3.1 to identify outliers in this set of error vectors, and then computed the mean squared prediction error (MSPE) using only the nonoutlier error vectors. In other words $\mathrm{MSPE}(\lambda) = |I|^{-1} \sum_{i \in I} \mathbf{e}_i^\top \mathbf{e}_i$, where I denotes the subset of nonoutliers in the collection of prediction error vectors and $|I|$ is the cardinality of the set I. Finally we chose the value of λ that minimized $\mathrm{MSPE}(\lambda)$ to be the optimal one.

To illustrate this approach we analyzed a data set generated from the example with Gaussian predictor variables discussed in Section 6.2 for the case when $q = 2$. We considered both cases namely without and with outliers. For varying values of λ the results of using our method RLASSO-SPD over the grid of values from 0.05 to 1 with an increment of 0.05 are plotted in Figure 7. The plot in the left panel of Figure 7 corresponds to the case when there are no outliers in the data set; for the corresponding plot involving outliers, see the right panel. The optimum values of λ for these two cases were 0.65 and 0.05, respectively.

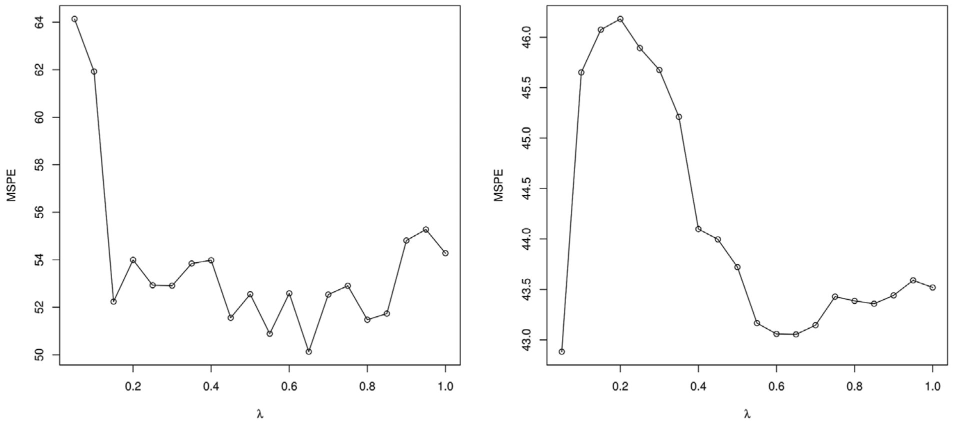

FIGURE 7: Mean squared prediction error (MSPE), as a function of λ, for Gaussian data in the usual sparse case (left panel), and the sparse case with outliers (right panel), respectively, for RLASSO-SPD.

7. CONCLUSIONS

In this article we investigated a new method for robust multivariate regression based on data depth and explored its related theoretical properties. In the numerical examples that we reported in Sections 4–6 our proposed method yielded competitive results compared to alternative robust methods of estimation. In addition we found our approach was quite stable computationally because the estimator involved contributions from all the observations instead of just subsamples from the data. By combining this approach with the LASSO our method can also be used to carry out regressions for sparse data. Overall using this robust approach to estimation based on the notion of data depth appears to be novel, 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 in the method and the article.

APPENDIX

Proof of Proposition 1. *As PD is affine invariant with $PD(\mathbf{Az}_i + \mathbf{b}) = PD(\mathbf{z}_i)$ for any nonsingular $(p+q) \times (p+q)$ matrix $\mathbf{A}$ and $\mathbf{b} \in \mathbb{R}^{p+q}$ we obtain*

$$\hat{\boldsymbol{\mu}}_{\mathbf{Az+b}} = \mathbf{A}\hat{\boldsymbol{\mu}}_{\mathbf{Z}} + \mathbf{b} \;\; \textit{and} \;\; \hat{\boldsymbol{\Sigma}}_{\mathbf{Az+b}} = \mathbf{A}\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}\mathbf{A}^\top.$$

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. As both $\hat{\boldsymbol{\mu}}_{\mathbf{Z}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ are affine equivariant, for a distribution H that 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 the central symmetry of $\mathbf{Z}$) we have $E(\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}) = \kappa Cov(\mathbf{Z})$ (see also Zuo & Cui, 2005).*
From Zuo & 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 $n \to \infty$. Using again the Fisher consistency and the consistency of $\hat{\boldsymbol{\mu}}_{\mathbf{Z}} = (\hat{\boldsymbol{\mu}}_{\mathbf{Y}}^\top, \hat{\boldsymbol{\mu}}_{\mathbf{X}}{}^\top)^\top$ (Zuo, Cui, & Young, 2004) we obtain $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}_{\mathbf{Y}} - \hat{\mathbf{B}}^\top\hat{\boldsymbol{\mu}}_{\mathbf{X}} \xrightarrow{P} \boldsymbol{\alpha}$ 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, & 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 & Cui (2005) gives the BP of $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ to be $\lfloor(n-p-q+1)/2\rfloor/n$ (by choosing $k = p+q$).*
Let $\mathbf{z}_n^*$ denote the new data set obtained by replacing m observations (with $m <$ $\min\{\lfloor(n-p-q+1)/2\rfloor, \lfloor n/2\rfloor\}$*) from the original data set $\mathbf{z}_n$ 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 eigenvalues, respectively. Both values are bounded because $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ does not break down 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 the Fisher consistency of the estimates based on PD from the proof of Proposition 2. Also $IF(\mathbf{z};\hat{\mathbf{B}},H)=IF(\mathbf{z};\hat{\boldsymbol{\Sigma}}_{\mathbf{XY}},H)$, $IF(\mathbf{z};\hat{\boldsymbol{\alpha}},H)=IF(\mathbf{z};\hat{\boldsymbol{\mu}}_{\mathbf{Y}},H)$, 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, & Young (2004), and for $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}}$ it is given in Corollary 3.2 of Zuo & Cui (2005). Thus*

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

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

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

Lemma 1. Under the conditions specified in 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 $0<s_0(z)\le 1$ for any z.*
- *$s_i(z)=E\{U_1^{2i-2}sign(|U_1|z-m_0)\}$ is bounded as $\|\mathbf{U}\|\le 1$ and $|sign(u)|\le 1$ for $i=1,2$. Here $sign(u)$ is the univariate sign function which equals -1, 0, or 1 according as $u<0$, equal to 0 or exceeds 0, respectively.*
- *$c_0=E[w_2\{s_0(\|\mathbf{Z}\|)\}]$, and is bounded by virtue of the fact that $0<s_0(z)\le 1$ and w_2 is bounded on the interval $(0,1)$.*
- *$c_1=E[\|\mathbf{Z}\|^2w_2\{s_0(\|\mathbf{Z}\|)\}]/\{(p+q)c_0\}$, and is bounded because $w^\dagger(z)=z^2w_2\{s_0(z)\}$ is bounded as we now explain. Take $v=c^{-1}s_0(z)$, and define*

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

  The function $w^\dagger(v)$ is continuous over the interval $(0,1/C]$; however, it is of the form $0/0$ at $v=0$. By L'Hôpital's rule we can argue that $w^\dagger(0)=0$, and hence $w^\dagger(v)$ 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 first, we have $0<zs_0(z)=1/(1/z+1)\le 1$. 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.*
- *$c_3=E[\|\mathbf{Z}\|^3s_0^2(\|\mathbf{Z}\|)w_2^{(1)}\{s_0(\|\mathbf{Z}\|)\}]/\{4p(1)\}$, and is bounded by virtue of the fact that*

$$zw_2^{(1)}\{s_0(z)\}=(16k/c^2)(1-ct)(1-t^4)^3\exp\{-k(1-t^4)^4\}\,,$$

  where $t=c^{-1}s_0(z)$ for $0<t<1$ is continuous in the bounded interval $(0,1)$; note also that $zs_0(z)\le 1$.

Now we consider the quantities in the expressions for the IFs:

- c_0 *is defined above and has been shown 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* $|x_1| w_1^{(1)}\{(1+\|\mathbf{x}\|)^{-1}\} \leq \|\mathbf{x}\| w_1^{(1)}\{(1+\|\mathbf{x}\|)^{-1}\}$. *Now*

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

  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.*
- $t_2(z) = c_3 \frac{s_2(z)-s_1(z)}{p+q-1} - c_1c_2s_1(z) - c_1 w_2\{s_0(z)\}$ *is bounded in view of the facts stated above.* □

BIBLIOGRAPHY

Agulló, J., Croux, C., & Van Aelst, S. (2008). The multivariate least-trimmed squares estimator. *Journal of Multivariate Analysis*, 99, 311–338.

Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7, 226–248.

Alfons, A. (2014). *robustHD: Robust methods for high-dimensional data. R package version 0.5.0.* https://CRAN.R-project.org/package=robustHD.

Cantoni, E. & Ronchetti, E. (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, 11, 141–146.

Chakraborty, A. & Chaudhuri, P. (2014). On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66, 303–324.

Charnes, A., Cooper, W. W., & 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. & Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71, 161–190.

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

Dang, X. & Serfling, R. (2010). Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference*, 140, 198–213.

Donoho, D. L. & Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Bickel, P. J., Doksum, K. A., & Hodges, J. L., editors. Wadsworth International Group, Belmont CA, 157–184.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

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

Gervini, D. (2003). A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, 84, 116–144.

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

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

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

Kudraszow, N. L. & Maronna, R. A. (2011). Estimates of MM type for the multivariate linear model. *Journal of Multivariate Analysis*, 102, 1280–1292.

Li, Y., Nan, B., & 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., & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27, 783–840.

Liu, X. & Zuo, Y. (2014). Computing projection depth and its associated estimators. *Statistics and Computing*, 24, 51–63.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49–55.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7, 308–313.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.

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

Rousseeuw, P. J. & Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association*, 94, 388–4026.

Rousseeuw, P. J., Van Aelst, S., Van Driessen, K., & 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., & Maechler, M. (2015). *robustbase: Basic Robust Statistics. R package version 0.92-3*. http://CRAN.R-project.org/package=robustbase

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*, Dodge, Y., editor. Birkhaeuser, Basel, 25–28.

Serfling, R. (2006). Depth functions in nonparametric multivariate inference. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Liu, R. Y., Serfling, R., & Souvaine, D. L., editors. 72, American Mathematical Society, Providence RI, pp. 1–16.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.

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

Van Aelst, S. & Willems, G. (2013). Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software*, 53, 1–32.

Vardi, Y. & Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 1423–1426.

Wang, W., Liang, Y., & Xing, E. P. (2015). Collective support recovery for multi-design multi-response linear regression. *Transactions on Information Theory*, 61, 513–534.

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed., Elsevier, Amsterdam.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimators for regression. *The Annals of Statistics*, 15, 642–656.

Zuo, Y. & Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28, 461–482.

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

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

Zuo, Y. & Cui, H. (2005). Depth weighted scatter estimators. *The Annals of Statistics*, 33, 381–413.

Zuo, Y. (2006). Multidimensional trimming based on projection depth. *The Annals of Statistics*, 34, 2211–2251.

Received 11 January 2016
Accepted 24 December 2016