Article

Variable Selection and Regularization in Quantile Regression via Minimum Covariance Determinant Based Weights

Edmore Ranganai [1,]* and Innocent Mudhombo [2]

1 Department of Statistics, University of South Africa, Florida Campus, Private Bag X6, Florida Park, Roodepoort 1710, South Africa
2 Department of Accountancy, Vaal University of Technology, Vanderbijlpark Campus, Vanderbijlpark 1900, South Africa; innocentm@vutcloud.onmicrosoft.com
* Correspondence: rangae@unisa.ac.za; Tel.: +27-7380-703-30 or +27-1167-092-57

Abstract: The importance of variable selection and regularization procedures in multiple regression analysis cannot be overemphasized. These procedures are adversely affected by predictor space data aberrations as well as outliers in the response space. To counter the latter, robust statistical procedures such as quantile regression which generalizes the well-known least absolute deviation procedure to all quantile levels have been proposed in the literature. Quantile regression is robust to response variable outliers but very susceptible to outliers in the predictor space (high leverage points) which may alter the eigen-structure of the predictor matrix. High leverage points that alter the eigen-structure of the predictor matrix by creating or hiding collinearity are referred to as collinearity influential points. In this paper, we suggest generalizing the penalized weighted least absolute deviation to all quantile levels, i.e., to penalized weighted quantile regression using the RIDGE, LASSO, and elastic net penalties as a remedy against collinearity influential points and high leverage points in general. To maintain robustness, we make use of very robust weights based on the computationally intensive high breakdown minimum covariance determinant. Simulations and applications to well-known data sets from the literature show an improvement in variable selection and regularization due to the robust weighting formulation.

Keywords: weighted quantile regression; RIDGE penalty; LASSO penalty; elastic net penalty; high leverage points; collinearity influential points; minimum covariance determinant

1. Introduction

Variable selection and robust estimation procedures are an important consideration of multiple regression analysis in the presence of predictor space data aberrations (high leverage points (outliers in the X-space) and multicollinearity) as well as response variable (Y-space) outliers. It is well known that the least squares (LS) are susceptible to both data aberrations in the predictor space and response variable (Y-space) outliers. To counter the influence of Y-space outliers, alternative robust procedures have been developed in the literature. One such attractive robust procedure is quantile regression (QR) [1]. In addition to being robust, QR is more versatile than the LS. This is due to the fact that, while the LS procedure models the conditional mean ($E(Y|X)$) (center of distribution), the QR procedure is able to detect heterogeneous effects of predictors at different quantile levels of the outcome as it models the conditional quantiles ($Q_{Y|X}(\tau)$ for $0 < \tau < 1$) of the response variable Y given the predictors X over the entire range of quantiles in $(0, 1)$ [2]. Regression quantiles (RQs) are optimal solutions to an optimization problem obtained using linear programming (LP) algorithms [3]. An RQ at the $\tau = 0.5$ quantile level corresponds to the well-known least absolute deviation (LAD) estimator, i.e., the ℓ_1 estimator. Although RQs are robust to Y-space outliers, on the other hand, they are susceptible to high leverage points as their influence function are bounded in the Y-space but unbounded in the X-space. Amongst the numerous sources of multicollinearity, some high-leverage observations tend

to influence the eigen-structure of the predictor matrix thereby creating multicollinearity or hiding it [4]. Such high leverage points are referred to as collinearity-influential points. However, not all high leverage observations are collinearity influential points. As remedies to high leverage points influences, weighted LAD (WLAD) [5] and weighted QR (WQR) have been suggested in the literature [6].

In both variable selection and regularization, in order to enhance the prediction accuracy and interpretability of statistical models, the RIDGE type [7], the least absolute shrinkage and selection operator (LASSO) type [8–10] penalties and a hybrid of these two penalties, viz., elastic net (E-NET) penalty [11,12] as well as the smoothly clipped absolute deviation (SCAD) penalty [13,14] have been suggested in the literature. Notable extensions of the LASSO penalty in the literature are the adaptive LASSO proposed by [15], fused LASSO [16], and group LASSO [17]. In high-dimensional sparse models where ordinary quantile regression is not consistent, [18] proposed the ℓ_1-penalized QR. The author of [14] developed, upon the procedure of [19] on model selection in composite quantile regression (CQR) and suggested weighted CQR (WCQR), a procedure based on data driven efficient weights, as the former researcher's equal weight property procedure lacked optimality. To mitigate against the undesirable effects of high leverage points in variable selection in the ℓ_1 estimator ($Q_{Y|X}(0.5)$), the weighted LAD-LASSO (WLAD-LASSO) procedure has been suggested [20,21]. Few variable selection procedures based on WQR have been suggested in the QR regression framework in different settings. We generalize the WLAD-LASSO approach to penalized WQR to mitigate against collinearity influential points and high leverage points in general, and maintain robustness via use of very robust weights.

In summary, the motivations of this study are premised on the following:

- The generalization of WLAD-LASSO [20,21] (in addition, we also include the RIDGE and the E-NET penalties) procedure to the QR framework, i.e., to penalized WQR as each RQ (including the LAD estimator) is a local measure, unlike the LS estimator, which is a global one.

- Rather than carrying an "omnibus" study of penalized WQR as in [20,21], we carry out a detailed study by distinguishing different types of high leverage points viz.,

    - Collinearity influential points which comprise collinearity inducing and collinearity hiding points.
    - High leverage points which are not collinearity influential.

- Taking advantage of high computing power, we make use of the very robust weights based on the computationally intensive high breakdown minimum covariance determinant (MCD) method rather than the well-known classical Mahalanobis distance or any LS based weights as in [20] which are amenable to outliers.

The remainder of this article is structured as follows. Some preliminaries on QR and variable selection in QR are discussed in Section 2. Variable selection in WQR as well as motivation for our choice of weights are detailed in Section 3 while simulation studies are detailed in Section 4. In Section 5, applications to two well-known data sets from the literature are detailed. Lastly, Section 6 concludes the paper.

2. Preliminaries

2.1. Quantile Regression

Consider the linear regression model given by

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \ \ for \ \ i = 1, 2, 3, ..., n \tag{1}$$

where y_i denotes the value of the response variable vector $\mathbf{Y}$ for the ith observation, $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$ denotes the vector of p predictor variables from the $n \times p$ design matrix $\mathbf{X}$ excluding the intercept, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ denotes a $p \times 1$ vector of yet to be estimated unknown regression coefficients (parameters) and ϵ_i denotes the value of the ith random error term, with cumulative distribution function F ($\epsilon_i \sim F$).

QR is based on an optimization problem which can be solved by linear programming techniques, viz.,

$$\widehat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \Sigma_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)), \quad i = 1, 2, ..., n \tag{2}$$

where $\rho_\tau(u) = u[\tau - I(u < 0)] \equiv u[\tau.I(u \geq 0) + (\tau - 1).I(u < 0)]$ and $\widehat{\boldsymbol{\beta}}(\tau)$ denotes the τth RQ.

2.2. Variable Selection in Quantile Regression

In this section, we discuss QR variable selection. We specifically present variable selection using LS-RIDGE [7], LASSO [9], adaptive LASSO [15], and E-NET [11] in a QR scenario. Firstly, we consider QR penalized with the RIDGE penalty [7] denoted by QR-RIDGE. The QR-RIDGE is given by the minimization problem

$$\widehat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \Sigma_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)) + \lambda \Sigma_{j=1}^{p} \beta_j^2, j = 1, 2, ..., p, \ i = 1, 2, ..., n \tag{3}$$

where λ is a positive ridge parameter in the range $0 < \lambda < 1$ and other terms are as defined in Equation (3). Many variations of λ have been used in literature (see [7,22–26]). QR with the RIDGE (ℓ_2-squared) penalty has been proposed as a remedy to the multicollinearity problem [25]. The presence of multicollinearity results in undue large sample variance resulting in unreliable inference and prediction.

Secondly, we consider QR variable selection procedure which uses the LASSO (ℓ_1) penalty [9] denoted by QR-LASSO. The QR-LASSO is then given by the minimization problem

$$\widehat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \Sigma_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)) + n\lambda \Sigma_{j=1}^{p} |\beta_j|, j = 1, 2, ..., p, \ i = 1, 2, ..., n \tag{4}$$

where λ is the tuning parameter that shrinks some beta coefficients towards zero, the second term is the penalty term, and other terms are as defined in Equation (3). This ℓ_1-penalized QR may be superior to the ℓ_2-squared penalized QR in Equation (3) in some instances.

Considering the more recent adaptive LASSO penalty by [15] in a penalized QR scenario, the tuning parameter is no longer constant (λ) but λ_j for $j = 1, 2, ..., p$. The minimization problem becomes

$$\mathbf{x}\widehat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \Sigma_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)) + n\Sigma_{j=1}^{p} \lambda_j |\beta_j|, j = 1, 2, ..., p, \ i = 1, 2, ..., n \tag{5}$$

where λ_j is the jth tuning parameter that shrinks some predictor variables to zero and other unknowns are as defined in Equation (4).

We also present a penalized QR procedure that uses the elastic NET penalty from [11] (E-NET-penalized QR which is best suitable for applications with unidentified groups of predictors). The E-NET penalized QR is given by

$$\widehat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta} \in R^p}{\operatorname{argmin}} \Sigma_{i=1}^{n} \rho_\tau |y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)| + \alpha\lambda \Sigma_{j=1}^{p} |\beta_j| + (1-\alpha)\lambda \Sigma_{j=1}^{p} \beta_j^2, j = 1, 2, ..., p, \ i = 1, 2, ..., n \tag{6}$$

where $\alpha \in [0,1]$ and λ is the tuning parameter for the second and third terms, respectively. Note that, for $\alpha = 0$, the E-NET penalty reduces to the RIDGE penalty while, for $\alpha = 1$, it reduces to the LASSO penalty. In some instances, the E-NET based procedure performs better than its ridge and LASSO counterparts [11]. Since QR is prone to outliers in the predictor space, weighted QR has been proposed as a remedy to high leverage points [6]. In the subsequent section, we motivate the choice of weights used in robust variable selection in QR framework.

3. Variable Selection and Regularization in Weighted Quantile Regression

3.1. Choice of Weights for Downweighing High Leverage Observations Motivation

In the LS case, statistics of the hat (projection) matrix $h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$ [4] have been used as standard tools to generate weights for weighted LS (WLS) estimation. Although this approach is both mathematically and practically tractable, such estimators have a breakdown point of only $1/n$ so a single high leverage point can completely dominate the ensuing estimates. Furthermore, such weights may suffer from the masking and swamping effect associated with the LS. Permitting contamination in both the predictor and response variables results in the breakdown point of LAD (and hence QR) estimator being the same as that of LS, $1/n$ (see, e.g., [27]). To circumvent the undesirable effects of both outliers and high leverage points, Ref. [5] proposed a weighted version of an LAD (WLAD) estimator. The weights

$$\omega_j = min\left(1, \frac{p}{RD(\mathbf{x}_j)^2}\right), \ \ j = 1, 2, ..., n \tag{7}$$

of this estimator are based on the computationally intensive high breakdown Minimum Covariance Determinant (MCD) method [28]. $RD(\mathbf{x}_j) = \sqrt{(\mathbf{x}_j - \widehat{\boldsymbol{\mu}})'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_j - \widehat{\boldsymbol{\mu}})}$ is the robust distance (a modification of the classical Mahalanobis distance), $\widehat{\boldsymbol{\mu}}$ is the center of the smallest ellipsoid (whose classical covariance matrix has the lowest possible determinant) containing half (or h observations as defined by the user) of the observations of the design matrix $\mathbf{X}$, and $\widehat{\boldsymbol{\Sigma}}$ is their covariance matrix multiplied by a consistency factor [29]. To side step the huge computational load associated with this weight, Ref. [30] suggested $\omega_i = \sqrt{min_j(h_j/h_{c_i})}$, where $h_{c_i} = \mathbf{x}_i(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{x}_i'$ is the ith leverage point relative to the clean subset $\mathbf{X}_c$ (without high leverage points). However, in this study, we make use of $RD(\mathbf{x}_j)$ due to the existing computer power (efficient algorithms) as well as its robustness to generalize the WLAD concept to the whole set of weighted conditional quantiles, i.e., weighted regression quantiles (WRQs).

Generalizing the [20] WLAD regression estimator and making use of the MCD based weights, we suggest a *WQR* estimator as a minimization problem given by

$$\widehat{\widetilde{\boldsymbol{\beta}}}(\tau) = \underset{\boldsymbol{\beta} \in R^p}{argmin} \Sigma_{i=1}^n \omega_i \rho_\tau |y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)| \ \ i = 1, 2, ..., n, \tag{8}$$

where ω_i and $\rho_\tau(.)$ are as defined in Equations (2) and (7), respectively.

3.2. Penalized Weighted Quantile Regression

Building on Equation (8), we suggest a penalized weighted QR (WQR) variable selection procedure using MCD based weight 7 to counter the undesirable influences of high leverage observations. We achieve robustness of WQR in the *X*-space due to the robustness of the this MCD based weights chosen appropriately as in the LS case (see also [5,30]). The then suggested penalized WQR variable selection procedures are based on three penalty functions, viz., the RIDGE, LASSO, and the E-NET penalties bringing about the WQR-RIDGE, WQR-LASSO, and WQR-E-NET estimators, respectively.

Let ω_i be a robust weight. We use ω_i in our proposed quantile variable selection procedures. The incorporation of WLAD-LASSO in the procedure inherits *X*-space robustness property as in [20]. The weight ω_i was discussed in Section 3. First, we propose the WQR-RIDGE given by

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \underset{\widetilde{\boldsymbol{\beta}} \in R^p}{argmin} \Sigma_{i=1}^n \omega_i \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)) + \lambda \Sigma_{j=1}^p \beta_j^2, j = 1, 2, ..., p, \ i = 1, 2, ..., n \tag{9}$$

where the terms are already defined. The tuning parameter λ shrinks the coefficients of the predictor variables towards zero.

We take advantage of the properties of WLAD-LASSO [20] and propose a weighted quantile variable selection procedure called WQR-LASSO. The WQR-LASSO procedure is expected to be robust and superior to its penalized QR counterpart, QR-LASSO. The WQR-LASSO is the solution of a minimization problem given by

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \underset{\tilde{\boldsymbol{\beta}} \in R^p}{\text{argmin}} \, \Sigma_{i=1}^n \omega_i \rho_\tau (y_i - \mathbf{x}_i' \boldsymbol{\beta}(\tau)) + n\lambda \Sigma_{j=1}^p |\beta_j|, j = 1, 2, ..., p, \, i = 1, 2, ..., n \qquad (10)$$

where tuning parameter λ is constant. In the literature, this procedure is found to perform better than the WQR-RIDGE procedure under deviations from the Normality assumptions.

Lastly, we apply the E-NET penalty to WQR variable selection to bring about the WQR-E-NET procedure. This weighted penalized QR procedure has both the LASSO and RIDGE penalties properties inherent in it [11]. The WQR-E-NET is the solution to a minimization problem given by

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \underset{\tilde{\boldsymbol{\beta}} \in R^p}{\text{argmin}} \, \Sigma_{i=1}^n \omega_i \rho_\tau (y_i - \mathbf{x}_i' \boldsymbol{\beta}(\tau)) + \alpha \Sigma_{j=1}^p |\beta_j| + (1 - \alpha)\lambda \Sigma_{j=1}^p \beta_j^2, j = 1, 2, ..., p, \, i = 1, 2, ..., n \qquad (11)$$

where α and λ are as in Equation (6).

3.3. Asymptotic Properties

We conveniently decompose the regression coefficient as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1' = (\beta_1, ..., \beta_{p_0})'$ and $\boldsymbol{\beta}_2' = (\beta_{p_0+1}, ..., \beta_p)'$; $\mathbf{x}_i' = (\mathbf{x}_{i1}', \mathbf{x}_{i2}')'$, where $\mathbf{x}_{i1}' = (x_{i1}, ..., x_{ip_0})'$ and $\mathbf{x}_{i2}' = (x_{i(p_0+1)}, ..., x_{ip})'$ such that

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i \equiv \mathbf{x}_{i1}' \boldsymbol{\beta}_1 + \mathbf{x}_{i2}' \boldsymbol{\beta}_2 + \epsilon_i, \ \ for \ i = 1, 2, 3, ..., n, \qquad (12)$$

with true regression coefficient $\boldsymbol{\beta}_1$ corresponding to non zero coefficients and $\boldsymbol{\beta}_2 = \mathbf{0}$.

To establish asymptotic normality, suppose for a suitable choice of λ_n that we now assume the two theoretical conditions to be true as stated in [13], as follows:

(i)   The regression errors ϵ_i's are *i.i.d.*, with τth quantile zero and continuous, positive density $f(.)$ in a neighborhood of zero (see [31]).

(ii)   Let $\mathbf{W} = diag(\omega_1, \omega_2, ..., \omega_n)$, where ω_i for $i = 1, 2, ..., n$ are known positive values that satisfy $max\{\omega_i\} = O(1)$.

(iii)   There exists a positive definite matrix $\boldsymbol{\Sigma}$: $lim_{n \to \infty} \frac{\Sigma_{i=1}^n \omega_i \mathbf{x}_i' \mathbf{x}_i}{n} = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ denote the $p_0 \times p_0$ and $(p - p_0) \times (p - p_0)$ top-left and right-bottom submatrices of $\boldsymbol{\Sigma}$, respectively.

We first give the following Theorem 1 (oracle property) for the *i.i.d.* error terms case ($\mathbf{W} = I_n$) before we consider the non *i.i.d.* error terms case which concerns this study.

Theorem 1. *Consider a sample* $\{(\mathbf{x}_i, y_i), \ i = 1, ..., n\}$ *from model (12) satisfying conditions (i) and (iii) (with* $\mathbf{W} = I_n$*). If* $\sqrt{n}\lambda_n \to 0$ *and* $n^{(\gamma+1)/2}\lambda_n \to \infty$*, then we have*
(1) Sparsity: $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$*;*
(2) Asymptotic normality: $\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \xrightarrow{d} N\left(\mathbf{0}, \frac{\tau(1-\tau)\boldsymbol{\Sigma}_{11}^{-1}}{f(0)^2}\right)$.

In order to extend the conclusions of the *i.i.d.* error terms case to the non *i.i.d.* error terms case, we invoke the following assumptions by [32]:

(K1)   As $n \to \infty \, max_{i=1,...,n}\{\mathbf{x}_i'\mathbf{x}_i/n\} \to 0$.

(K2)   The random error terms ϵ_i's are independent with $F_i(t) = P(\epsilon_i \leq t)$ the distribution function of ϵ_i. We assume $F_i(.)$ is locally linear near zero (with a positive slope) and $F_i(0) = \tau$.

(K3)   Assume that, for each u, $(1/n) \sum_{i=1}^{n} \psi_{ni}(u', x_i) \to \zeta$, where $\zeta(.)$ is a strictly convex function taking values in $[0, \infty)$ with $\psi_{ni} = \int_{0}^{t} \sqrt{n}(F_i(s/\sqrt{n}) - F_i(0))ds$ denoting a convex function for each n and i.

Corollary 1. *Under assumptions (ii), (iii), and (K1), Theorem 1 holds provided the non i.i.d. error terms satisfy (K2) and (K3).*

Remark 1. *The proofs of Theorem 1 and Corollary 1 are outlined in* [13] *(online supplement materials).*

4. Simulation Study

In this section, we perform a simulation study to investigate the finite-sample performance of penalized WQR under the RIDGE, the LASSO, and the E-NET penalty (for $\alpha = 0.5$) functions in terms of the variable selection and regularization of the regression parameters making use of Equations (9)–(11) and the robust MCD based weight (ω_j) given in Equation (7) in comparison to their unweighted versions. For brevity, we consider $\tau = 0.5$ (the LAD estimator) and $\tau = 0.25$ RQ levels. We summarize the simulation results in terms of the percentage of correctly estimated regression models, the average correct zero coefficients (β_3, β_4, β_6, β_7, and β_8) and the average incorrect zero coefficients along with the median of the test error and its respective measure of dispersion,

$$MAD = 1.4826 \, median_{i=1,\ldots,n}|\epsilon_i - median_{k=1,\ldots,n}(\epsilon_k)|,$$

where constant 1.4826 is a correction factor which makes the MAD consistent at Normal distributions (see e.g., [29]). The simulation study is designed according to the following scenarios with the predictor matrices of size $n \times p$, $p = 8$ and $n = 50$, 100 (but we only give results for $n = 50$ for brevity);

* $D1-$ This predictor matrix is obtained from orthogonalization such that $\mathbf{X}'\mathbf{X} = n\mathbf{I}$. Using singular value decomposition (SVD), we solve $\mathbf{W} = \mathbf{UDV}'$, where $w_{ij} \sim N(0,1)$ for $i = 1, ..., n$ and $j = 1, ..., p$; $\mathbf{U}$ and $\mathbf{V}$ are orthogonal with the diagonal entries of $\mathbf{D}$ giving the singular (eigen) values of $\mathbf{W}$. Then, $\mathbf{X} = \sqrt{n}\mathbf{U}$ is such that $\mathbf{X}'\mathbf{X} = n\mathbf{I}$ due to orthogonality of $\mathbf{U}$.

* $D2-$ has a collinearity inducing point which is $D1$, but with observation having the largest Euclidean distance from the center of the design space moved 10 units in the X-space.

* $D3-$ has collinearity hiding point which is $D1$, but with observations having the largest (second largest) Euclidean distance from the center of the design space moved 10 units in the X-space.

* $D4-$ has a collinearity inducing point which is $D1$, but with observation having the largest Euclidean distance from the center of the design space moved 100 units in the X-space.

* $D5-$ has collinearity hiding point which is $D1$, but with observations having the largest (second largest) Euclidean distance from the center of the design space moved 100 units in the X-space.

* $D6-$ has ($m = 5$) $(n - m) \times p$ correlated $\mathbf{X}_1$ and $m \times p$ leverage contaminated $\mathbf{X}_2$ sub matrices of $D6$, i.e., $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, where $\mathbf{X}_1 \sim N(\boldsymbol{\mu_1}, \boldsymbol{V})$ with $\boldsymbol{\mu_1} = (0,0,0,0,0,0,0,0)'$ and $v_{ij} = 0.5^{|j-i|}$ (0.5 controls the degree of correlation), $i, j = 1, 2, 3, 4, 5, 6, 7, 8$, is the $(ij)^{th}$ entry of the covariance matrix $\boldsymbol{V}$), is the $(n - m) \times p$ correlated sub matrix of $D6$; $\mathbf{X}_2 \sim N(\boldsymbol{\mu_2}, \boldsymbol{I})$ with $\boldsymbol{\mu_2} = (1,1,1,1,1,1,1,1)'$ is the $m \times p$ leverage contaminated sub matrix of $D6$.

The predictor matrices $D1$–$D5$ are constructed as in [33], while $D6$ is constructed as in [20].

The regression coefficients with zero intercept, i.e., $\beta_0 = 0$ are

- $\boldsymbol{\beta_1} = (3.5, 2, 0, 0, 2.5, 0, 0, 0)'$, $\boldsymbol{\beta_2} = (2, 1, 0, 3, 1.5, 0, 1, 0)'$.

We consider the following error term distribution scenarios;

- $\epsilon \sim N(\mu, \sigma^2)$, with (μ, σ) choices $(0, 1)$ and $(0, 3)$.
- $\epsilon \sim t_d$ with choices $d = 1, 6$.

The design matrix $D1$ is used as a baseline, and a schematic representation of $D2$–$D5$ departures from it is shown in Figure 1; $D2$ and $D4$ have a high leverage point that induces collinearity while $D3$ and $D5$ each have a pair of high leverage points that hide collinearity, viz., one for inducing the collinearity and the other for hiding it. On the other hand, $D6$ has both multicollinearity due the covariance structure, V of the sub matrix $\mathbf{X}_1$, and high leverage points due to the mean shift in $\mathbf{X}_2$.

$D1$–$D6$ (see, e.g., the response variable $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2})'$ is generated as in [20], i.e., $\mathbf{Y_1} = \mathbf{X_1}'\boldsymbol{\beta_1} + \sigma\epsilon$, $\epsilon \sim N(0, \sigma^2)$; $\sigma = 1$, 3 and $\mathbf{Y_2} = \mathbf{X_2}'\boldsymbol{\beta_2}$;
$\mathbf{Y_1} = \mathbf{X_1}'\boldsymbol{\beta_1} + \sigma\epsilon$, $\epsilon \sim t_d$; $\sigma = 0.5$, 1 and $\mathbf{Y_2} = \mathbf{X_2}'\boldsymbol{\beta_2}$.

The number of simulation runs is 200 while the 100-fold cross-validations are employed to obtain the tuning parameters λ. Instead of using cross-validation error metrics independent tuning data sets and testing data sets of size n and $100n$ were generated in the exact way the training data sets were generated (see, e.g., [13]). This simulation study explores these different scenarios in both QR and WQR variable selection procedures using the R-add-on package hqreg. In this package, the regularization parameter lambda is fit using a semismooth Newton coordinate descent algorithm. See [34] for details.

A schematic representation of collinearity influential points is given in Figure 1 below showing the scatter plots representations of predictor matrices $D2$–$D5$. The extreme observation in $D2$ and $D4$ creates collinearity while the second extreme observation $D3$ and $D5$ obscures it. We only consider the Normal distribution at the well-behaved orthogonal predictor matrix D1, and from thereon we only consider the t distribution as QR being designed to handle distributions heavier than the Normal one which is handled best by the LS.

4.1. Results

D1 SCENARIO

As a point of departure, we consider the well-behaved predictor matrix D1 under the Normal distribution as contrasted with the D1 under the t distribution on 1 $d.f.$ (implying outliers). The results given in Table 1 are as expected. At D1 under the Normal distribution, variable (model) selection performs best under the LASSO penalty followed by the E-NET penalty across all models with no marked differences in the median and MAD test error measures indicating the robustness of the QR-LASSO procedure under the t distribution. This is further illustrated graphically in Figures 2 and 3 where the QR-LASSO procedure correctly shrinks to zero the zero coefficients $\{\beta_3, \beta_4, \beta_6, \beta_7, \beta_8\}$ more often than not.
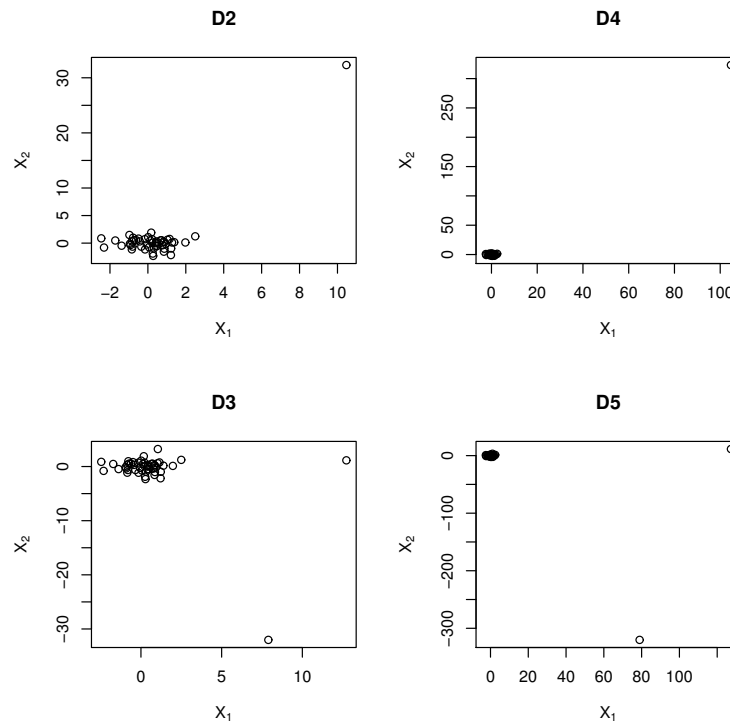
Figure 1. First Row Panel: Comprise Collinearity Influential Points that create collinearity in D2 and D4; Second Row Panel: Comprise Collinearity Influential Points that hide collinearity in D3 and D5.

Table 1. Quantile regression at D1 (at quantile levels $\tau = 0.5$ and 0.25) for $n = 50$.

	Parameters	Method	Correctly Fitted	No. of Zeros Correct	No. of Zeros Incorrect	Med (MAD) Test Error	Optimal λ
D1 under the Normal Distribution	$\sigma = 1, \tau = 0.25$	QR-RIDGE	0	2.27	0	1.28(1.97)	0.12
		QR-LASSO	**67.5**	**4.56**	0	**0.71(1.20)**	0.04
		QR-E-NET	18.5	3.59	0	0.72(1.25)	0.04
	$\sigma = 1, \tau = 0.5$	QR-RIDGE	1.5	2.33	0	-0.03(1.99)	0.14
		QR-LASSO	**62**	**4.49**	0	**0.00(1.15)**	0.05
		QR-E-NET	24	3.6	0	0.01(1.19)	0.04
	$\sigma = 3, \tau = 0.25$	QR-RIDGE	9	3.07	0.03	2.70(4.32)	0.12
		QR-LASSO	**39.5**	**4.52**	0.38	**2.03(3.60)**	0.04
		QR-E-NET	30.5	4	0.2	2.18(3.69)	0.04
	$\sigma = 3, \tau = 0.5$	QR-RIDGE	2.5	2.37	0.01	-0.04(4.06)	0.12
		QR-LASSO	**40**	**4.57**	0.32	**0.01(3.45)**	0.05
		QR-E-NET	31	3.9	0.11	0.00(3.55)	0.04
D1 under the *t* Distribution	$d = 1, \sigma = 0.5,$ $\tau = 0.25$	QR-RIDGE	3.00	2.33	0.02	2.17(3.21)	0.11
		QR-LASSO	**64.00**	**4.92**	0.72	**1.24(2.16)**	0.04
		QR-E-NET	36.50	4.42	0.62	1.44(2.41)	0.03
	$d = 1, \sigma = 0.5,$ $\tau = 0.5$	QR-RIDGE	1.50	2.56	0.01	0.02(2.94)	0.13
		QR-LASSO	**64.50**	**4.94**	0.67	**0.02(1.72)**	0.04
		QR-E-NET	32.50	4.33	0.58	-0.01(1.94)	0.03
	$d = 1, \sigma = 1,$ $\tau = 0.25$	QR-RIDGE	2.50	2.37	0.03	3.05(4.30)	0.11
		QR-LASSO	**30.50**	**4.95**	1.57	**2.44(3.80)**	0.04
		QR-E-NET	26.00	4.78	1.49	2.62(4.04)	0.03
	$d = 1, \sigma = 1,$ $\tau = 0.5$	QR-RIDGE	3.50	2.49	0.02	0.02(4.15)	0.12
		QR-LASSO	**33.50**	**4.95**	1.38	**0.02(3.34)**	0.04
		QR-E-NET	25.50	4.66	1.27	0.02(3.58)	0.03
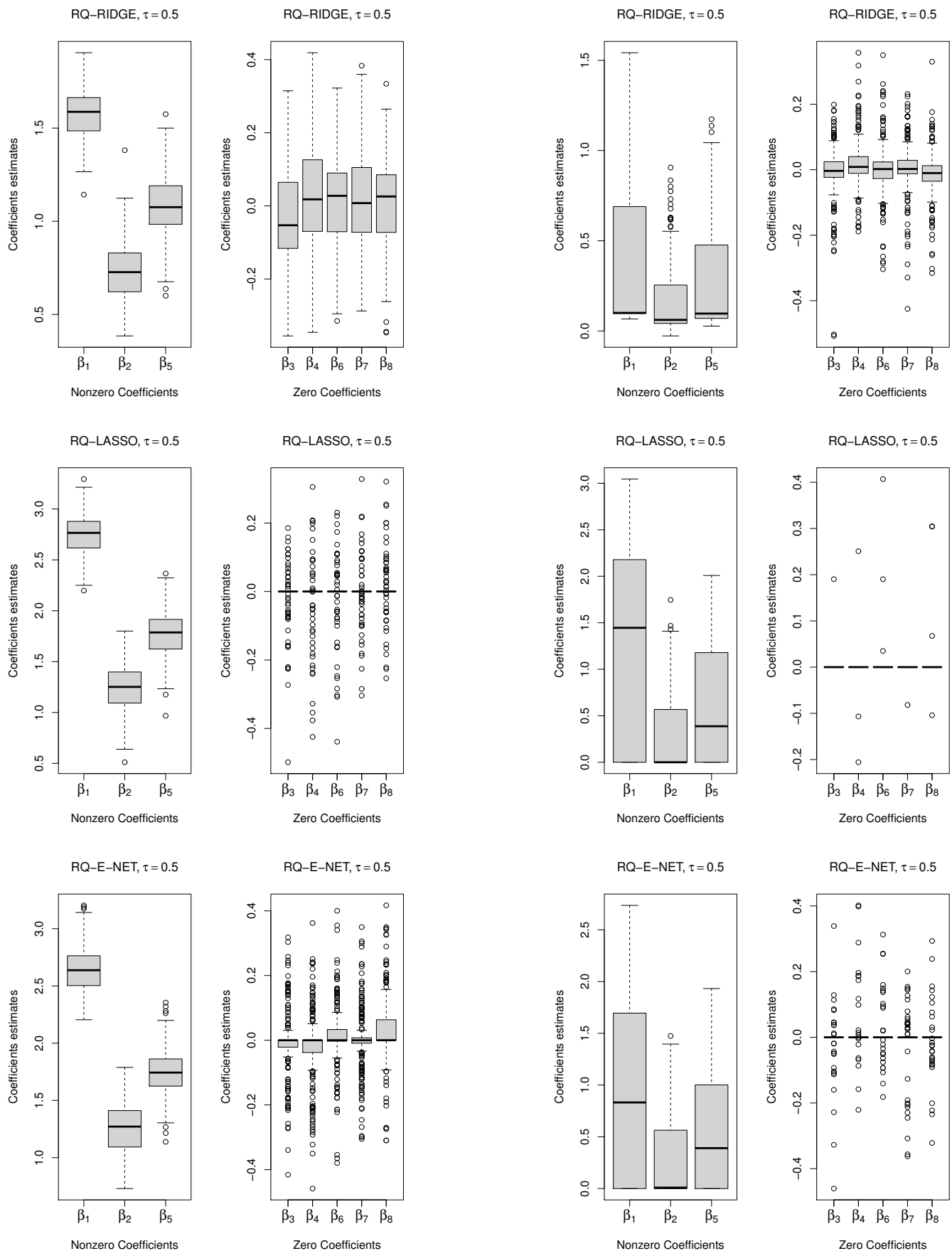
Figure 2. Box Plots at D1 for RQ: Left panel; under the Normal distribution with $\sigma = 1$, Right panel; Under the *t*-distribution with $\sigma = 1$ and $d = 1$, $\tau = 0.5$.
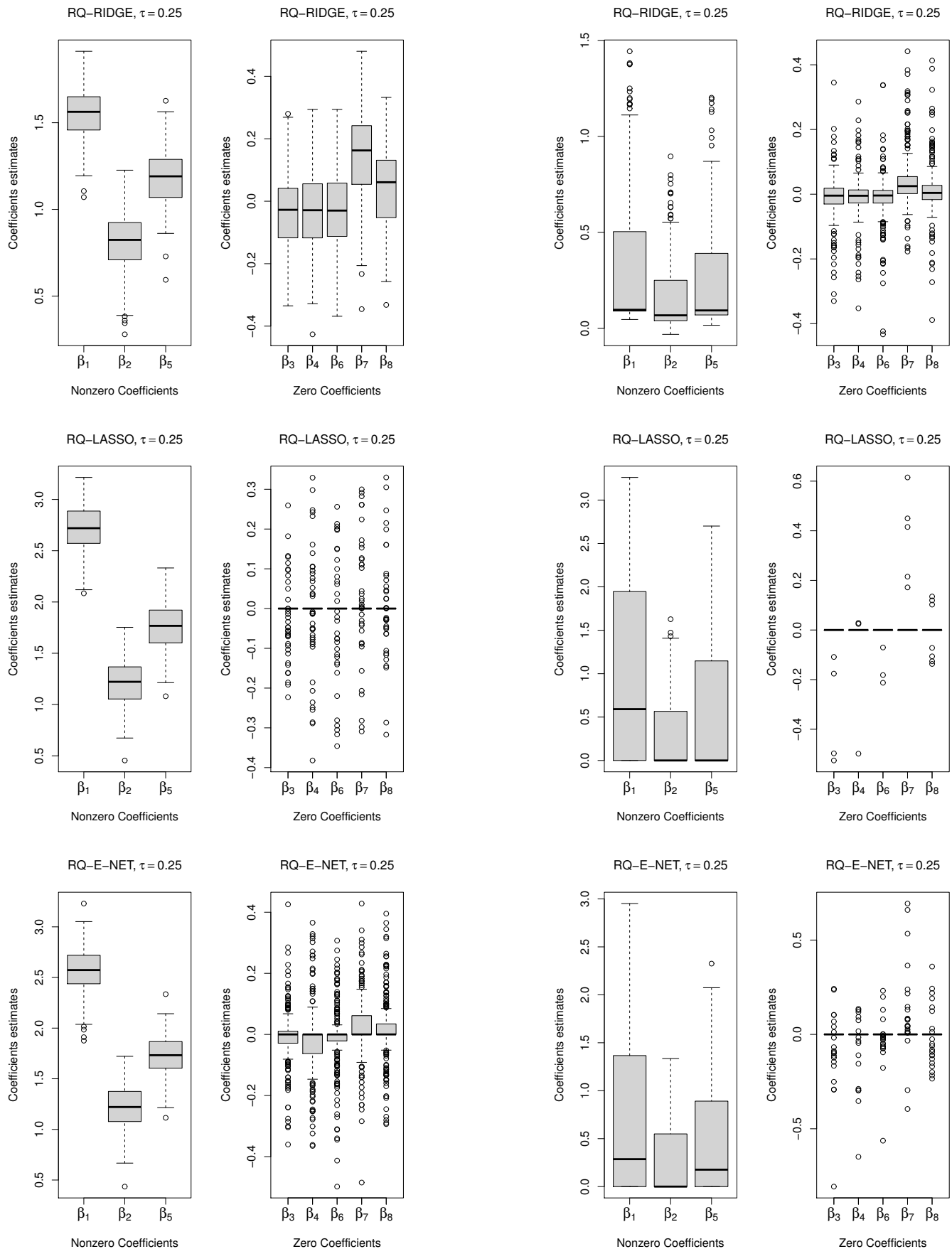
Figure 3. Box Plots at D1 for RQ: Left panel; under the Normal distribution with $\sigma = 1$, Right panel; under the *t*-distribution with $\sigma = 1$ and $d = 1$, $\tau = 0.25$.

Remark 2. *The five zero coefficients correspond to the set $\{\beta_3, \beta_4, \beta_6, \beta_7, \beta_8\}$, hence the maximum average of correctly/incorrectly selected (shrunk) coefficients is 5 while the set of correctly selected models is given as a proportion, i.e., a %.*

D2 AND D4 SCENARIOS

We consider the introduction of collinearity inducing points in both D2 and D4. The RIDGE penalty performs the worst in every scenario in both penalized QR and penalized WQR in model/variable selection. The model/variable selection pattern at D2 and D4 under both the t_6 and t_1 distributions is shown in Figure 4. The dominance of WQR-LASSO followed by WQR-E-NET is clearly depicted. However, the performance of WQR-E-NET is generally better under the t_1 distribution than it is under the t_6 distribution. In Figure 4 (lower panels), both the median absolute test error and its MAD measure (in the line graph) show that generally the unweighted versions outperform the weighted ones.

Figure 4. Performance at D2 and D4 under the t_6 and t_1 distributions with the RIDGE and E-NET on the LHS and RHS of LASSO, respectively; Upper panel: Model/Variable selection showing the proportion of correct models and the average of correct/incorrect βs selected; Lower panel: Prediction metrics.

D3 AND D5 SCENARIO

We consider the introduction of collinearity hiding points in both D3 and D5. The performances under the t_6 and t_1 distributions are shown graphically in Figure 5. In model/variable selection, throughout all the scenarios, the weighted penalized versions outperform the penalized unweighted versions under the LASSO and E-NET penalties. Amongst the penalized weighted versions, WQR-RIDGE performs the worst while WQR-LASSO performs better than WQR-E-NET. The prediction pattern under the t distributions exhibited under D3 and D5 in Figure 5 (lower panel) are different to that exhibited at D2 and D4 in that the *MAD* error measure is more erratic (but the absolute *median* error is less erratic) at D3 and D5 but the absolute. In fact, the prediction picture of QR-LASSO and WQR-LASSO at $\tau = 0.5$ based on the absolute *median* error are similar at D3 and D5, whereas, at D2 and D4, WQR-LASSO performs better with respect to this measure.
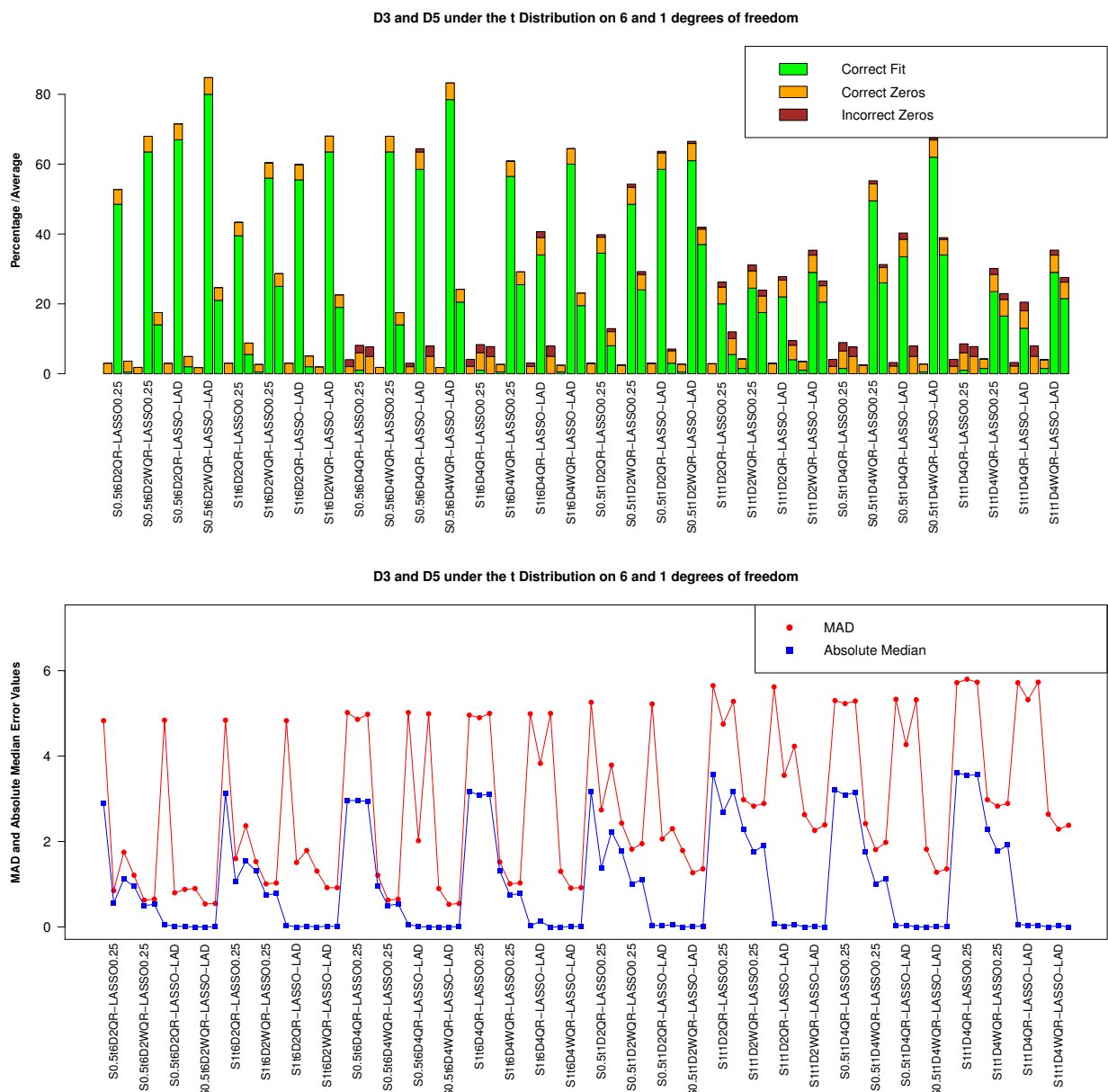
Figure 5. Performance at D3 and D5 under the t_6 and t_1 distributions distributions with the RIDGE and E-NET on the LHS and RHS, respectively; Upper panel: Model/Variable selection showing the proportion of correct models and the average of correct/incorrect βs selected; Lower panel: Prediction metrics.

D6 SCENARIO

The D6-scenario model/variable selection performance outcomes under the t distribution is given in Figure 6. In Figure 6, the dominance of the LASSO penalty in both QR and weighted QR is clearly depicted with WQR-LASSO far outperforming QR-LASSO in model/variable selection. Overall, on average, the zero βs are most incorrectly selected under the t_1 distribution and at $\sigma = 1$. The prediction pattern under the t distribution exhibited under D6 in Figure 6 (lower panel) is slightly poorer under the t_1 distribution compared to that exhibited under the t_6 distribution.
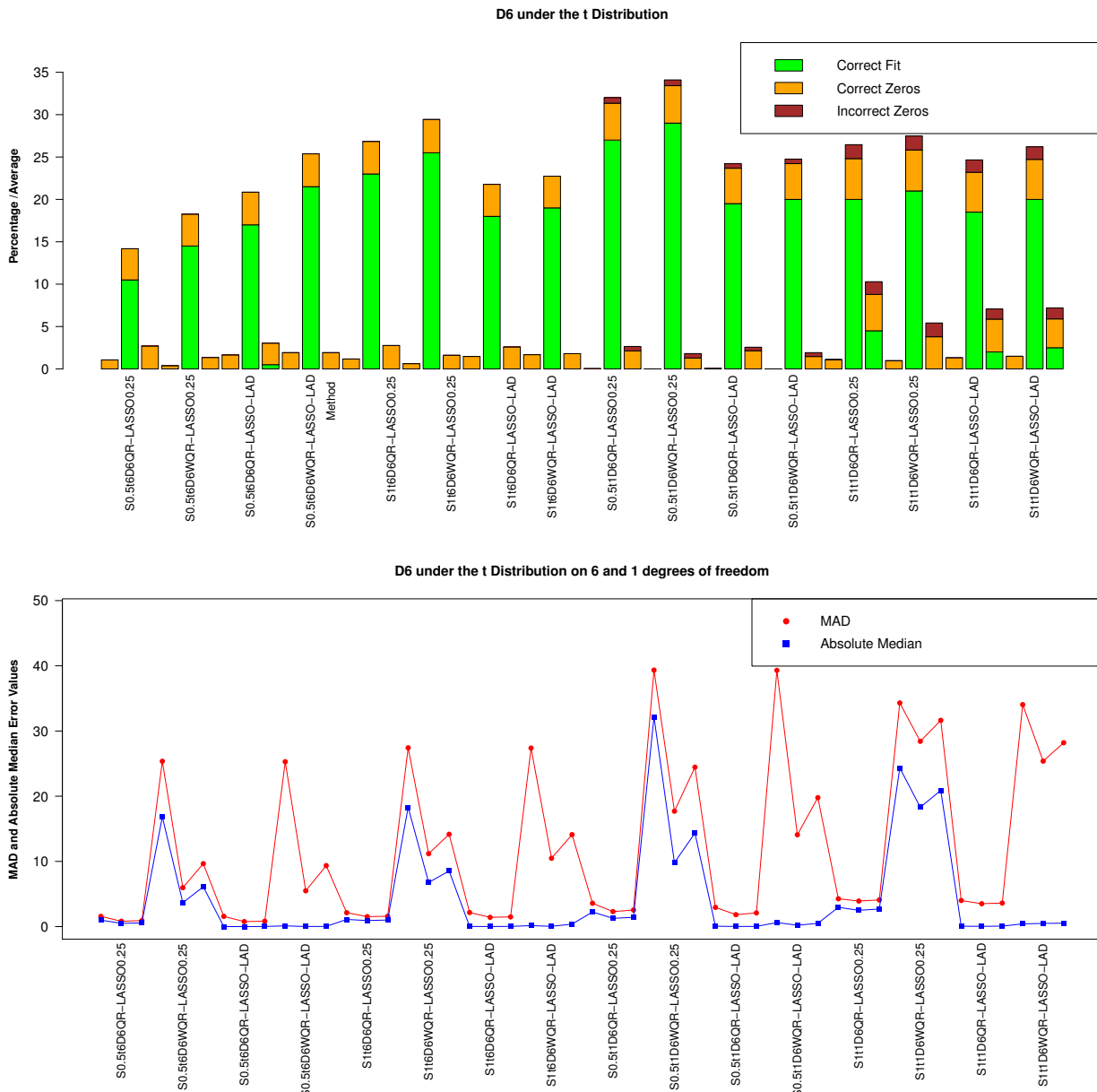
Figure 6. Performance at D6 under the t_6 and t_1 distributions with the RIDGE and E-NET on the LHS and RHS of LASSO, respectively; Model/Variable selection showing the proportion of correct models and the average of correct/incorrect βs selected; Lower panel: Prediction metrics.

5. Examples

In this section, we consider two data sets often used to illustrate the efficacy of robust methodologies in mitigating against high leverage points in general as well as collinearity

influential points in particular, viz. the [35] and the [36] data sets. In both data sets, the response is generated based on the t_1 error term distribution in line with testing the efficacy of robust procedures like QR as in [20].

Remark 3. *The LS procedure is adversely affected by both high leverage points and outliers in the Y-space, hence it consistently gives the worst performance as expected. On the other hand, QR is not affected by the latter data aberrations. Consequently, we mainly focus on the efficacy of penalized WQR at quantile levels $\tau = 0.5$ and $\tau = 0.25$ in addressing the problem of high leverage points with the intercept $= F^{-1}(\tau) + \beta_0$ corresponding to 0 and -1 under the t_1 error term distribution, respectively.*

5.1. Hawkins, Bradu, and Kass Data Set

We firstly consider the [35] which consists of 75 observations with three predictor variables. The first 14 observations of the 75 observations are high leverage points with the first 13 observations inducing the collinearity while the 14th observation greatly affects the collinearity structure on its own although it is also a collinearity inducing point (see [37]). Figure 7 shows the leverage structure for the predictor variables of this data set based on the robust MCD based distance as well as the classical one. We give results for the full data set and reduced data sets without observations 1–14.

The response variable is generated as $\mathbf{Y}_2 = \mathbf{X}_2' \boldsymbol{\beta}_2$ for the first 10 observations and $\mathbf{Y}_1 = \mathbf{X}_1' \boldsymbol{\beta}_1 + \epsilon$, $\epsilon \sim t_1$ for the remainder of the data, where $\boldsymbol{\beta}_2 = (2, 2, 0)'$ and $\boldsymbol{\beta}_1 = (1, 1, 0)'$ such that $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)'$

The results for the full data set are given in Table 2.

Table 2. Results for the full [35] data set.

	β	NONE-BIASED		RIDGE		LASSO		E-NET	
					UNWEIGHTED				
		$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*
λ		0.00		0.00		0.00		0.00	
RQ $\tau = 0.5$	*intercept*	2.27	−2.27	2.39	−2.39	2.39	−2.39	2.39	−2.39
	X_1	1.39	0.61	1.45	0.55	1.45	0.55	1.45	0.55
	X_2	1.87	0.13	1.79	0.21	1.79	0.21	1.79	0.21
	X_3	−0.78	0.78	−0.74	0.74	−0.74	0.74	−0.74	0.74
λ		0.00		0.00		0.00		0.00	
RQ $\tau = 0.25$	*intercept*	1.09	−2.09	1.32	−2.32	1.32	−2.32	1.32	−2.32
	X_1	1.59	0.41	1.48	0.52	1.48	0.52	1.48	0.52
	X_2	1.94	0.06	1.80	0.20	1.80	0.20	1.80	0.20
	X_3	−0.88	0.88	−0.76	0.76	−0.76	0.76	−0.76	0.76
					WEIGHTED				
	β	NONE-BIASED		RIDGE		LASSO		E-NET	
		$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*
λ		0.00		0.00		0.06		0.00	
RQ $\tau = 0.5$	*intercept*	2.27	−2.27	0.11	−0.11	0.00	0.00	0.11	−0.11
	X_1	1.39	0.61	1.93	0.07	1.93	0.07	1.93	0.07
	X_2	1.87	0.13	2.01	−0.01	1.97	0.03	2.01	−0.01
	X_3	−0.78	0.78	−0.09	0.09	0.00	0.00	−0.09	0.09
λ		0.00		0.50		0.50		0.50	
RQ $\tau = 0.25$	*intercept*	1.09	−2.09	0.18	−1.18	0.29	−1.29	0.29	−1.29
	X_1	1.59	0.41	0.35	1.65	0.00	2.00	0.00	2.00
	X_2	1.94	0.06	0.39	1.61	0.00	2.00	0.00	2.00
	X_3	−0.88	0.88	0.38	−0.38	0.00	0.00	0.00	0.00

Note: In the β column: RQ $\tau = 0.5$ rows have intercept 0.00, X_1 2.00, X_2 2.00, X_3 0.00; RQ $\tau = 0.25$ rows have intercept −1.00, X_1 2.00, X_2 2.00, X_3 0.00.
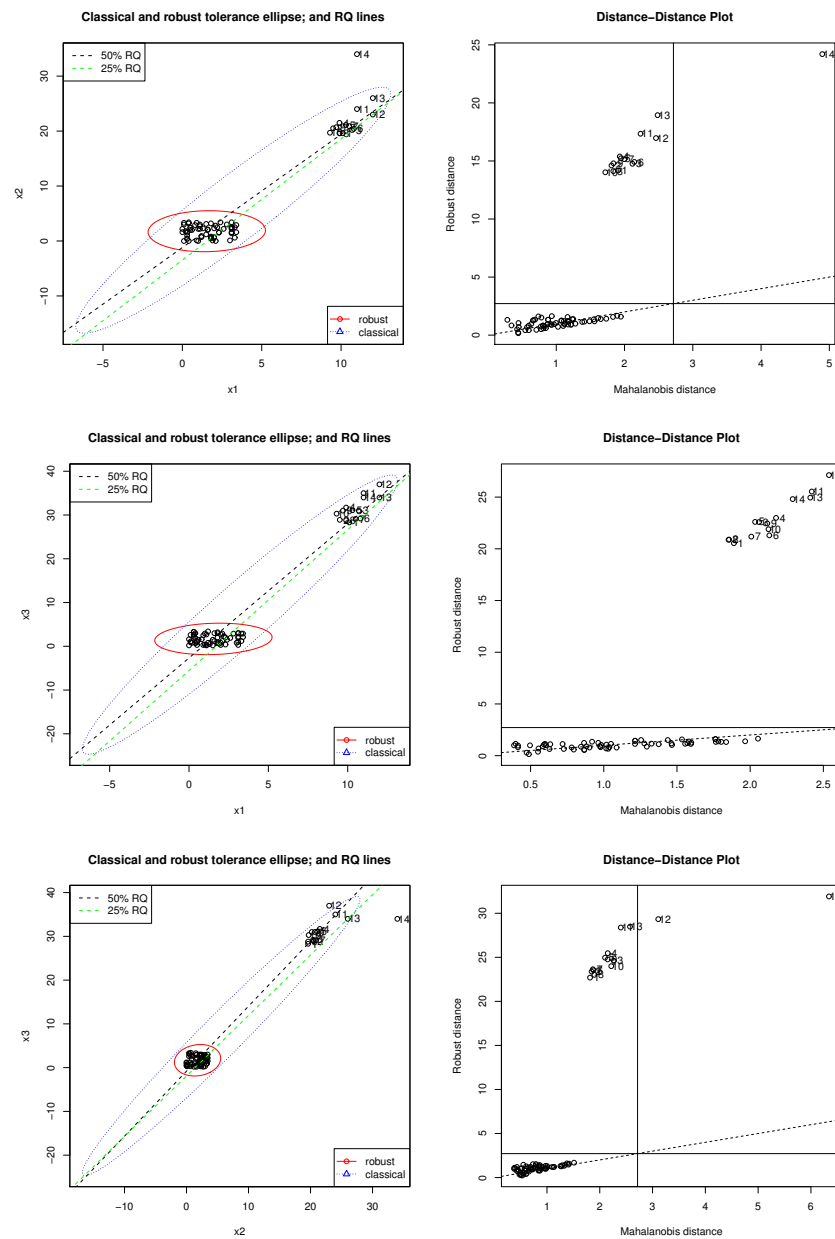
Figure 7. Tolerance ellipses and distance–distance plots for the [35] data set.

Similar poor performances of penalized QR at both τ-levels are exhibited across all penalty functions. However, penalized WQR exhibits a drastic improvement at $\tau = 0.5$ where penalized WQR is equivalent to unpenalized WQR under both the RIDGE and the E-NET penalties as the $\lambda = 0$. At $\tau = 0.25$, where $\lambda \neq 0$, WQR tends to be too "greedy" under LASSO and E-NET penalties where all the parameters are shrunk to zero.

Results for the reduced [35] data set with a "clean" predictor matrix (without observations 1–14) are given in Table 3. As expected, both QR and WQR select unpenalized models (models with tuning parameter $\lambda = 0$) except WQR at $\tau = 0.25$. QR performs considerably well at both τ levels and across penalty functions. However, there is a marginal improvement in using penalized WQR at $\tau = 0.5$ while, at $\tau = 0.25$, the LASSO and E-NET penalties are too "greedy".

Table 3. Results for [35]; without observations 1–14.

		β	UNWEIGHTED							
			NONE-BIASED		**RIDGE**		**LASSO**		**E-NET**	
			$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*
λ			0.00		0.00		0.00		0.00	
RQ $\tau = 0.5$	intercept	0.00	0.74	−0.74	0.50	−0.50	0.50	−0.50	0.50	−0.50
	X_1	2.00	1.86	0.14	1.84	0.16	1.84	0.16	1.84	0.16
	X_2	2.00	1.93	0.07	1.87	0.13	1.87	0.13	1.87	0.13
	X_3	0.00	−0.07	0.07	−0.03	0.03	−0.03	0.03	−0.03	0.03
λ			0.00		0.00		0.00		0.00	
RQ $\tau = 0.25$	intercept	−1.00	0.69	−1.69	−0.09	−0.91	−0.09	0.09	−0.09	−0.91
	X_1	2.00	1.67	0.33	1.73	0.27	1.73	0.27	1.73	0.27
	X_2	2.00	1.90	0.10	1.95	0.05	1.95	0.05	1.95	0.05
	X_3	0.00	−0.08	0.08	−0.10	0.10	−0.10	0.10	−0.10	0.10
		β	WEIGHTED							
			NONE-BIASED		**RIDGE**		**LASSO**		**E-NET**	
			$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*	$\widehat{\beta}$	*Bias*
λ			0.00		0.00		0.00		0.00	
RQ $\tau = 0.5$	intercept	0.00	0.74	−0.74	0.27	−0.27	0.27	−0.27	0.27	−0.27
	X_1	2.00	1.86	0.14	1.86	0.14	1.86	0.14	1.86	0.14
	X_2	2.00	1.93	0.07	1.96	0.04	1.96	0.04	1.96	0.04
	X_3	0.00	−0.07	0.07	−0.06	0.06	−0.06	0.06	−0.06	0.06
λ			0.00		0.50		0.33		0.50	
RQ $\tau = 0.25$	intercept	−1.00	0.69	−1.69	1.74	−2.74	2.22	−3.22	2.20	−3.20
	X_1	2.00	1.67	0.33	0.30	1.70	0.00	2.00	0.00	2.00
	X_2	2.00	1.90	0.10	0.48	1.52	0.00	2.00	0.10	1.90
	X_3	0.00	−0.08	0.08	0.22	−0.22	0.00	0.00	0.00	0.00

5.2. Hocking and Pendleton Data Set

While the [35] set is an example of high leverage points that are collinearity inducing points, the [36] data set is an example of high leverage points that are collinearity hiding points. This data set has 26 observations with three predictor variables, X_1, X_2, and X_3, whereby X_3 is created as a linear combination of X_1 and X_2. The response variable is generated as $\mathbf{Y}_1 = \mathbf{X}_1'\boldsymbol{\beta}_1 + \epsilon$, $\epsilon \sim t_1$ for the first 22 observations and $\mathbf{Y}_2 = \mathbf{X}_2'\boldsymbol{\beta}_2$ for the remainder of the data, where $\boldsymbol{\beta}_1 = (3, -2, 0)'$ and $\boldsymbol{\beta}_2 = (1, 1, 0)'$ such that $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)'$.

Figure 8 shows the leverage structure for the predictor variables of this data set based on the robust MCD based distance as well as the classical one. We give results for the full data set and a reduced data set (without the collinearity hiding observation 24).

The results for the full data set are given in Table 4. Under penalized unweighted RQ, $\boldsymbol{\beta}$ is reasonably best estimated at $\tau = 0.25$ while, under penalized, WQR, it is reasonably best estimated at $\tau = 0.5$ across all penalty functions. Under penalized WQR, the model with tuning parameter $\lambda = 0$ is selected indicating that unpenalized WQR exhibits the optimal model. There is an improvement in adopting unpenalized WQR compared to penalized QR at both τ-levels. However, the best parameter estimation is exhibited under unpenalized WQR at $\tau = 0.5$.
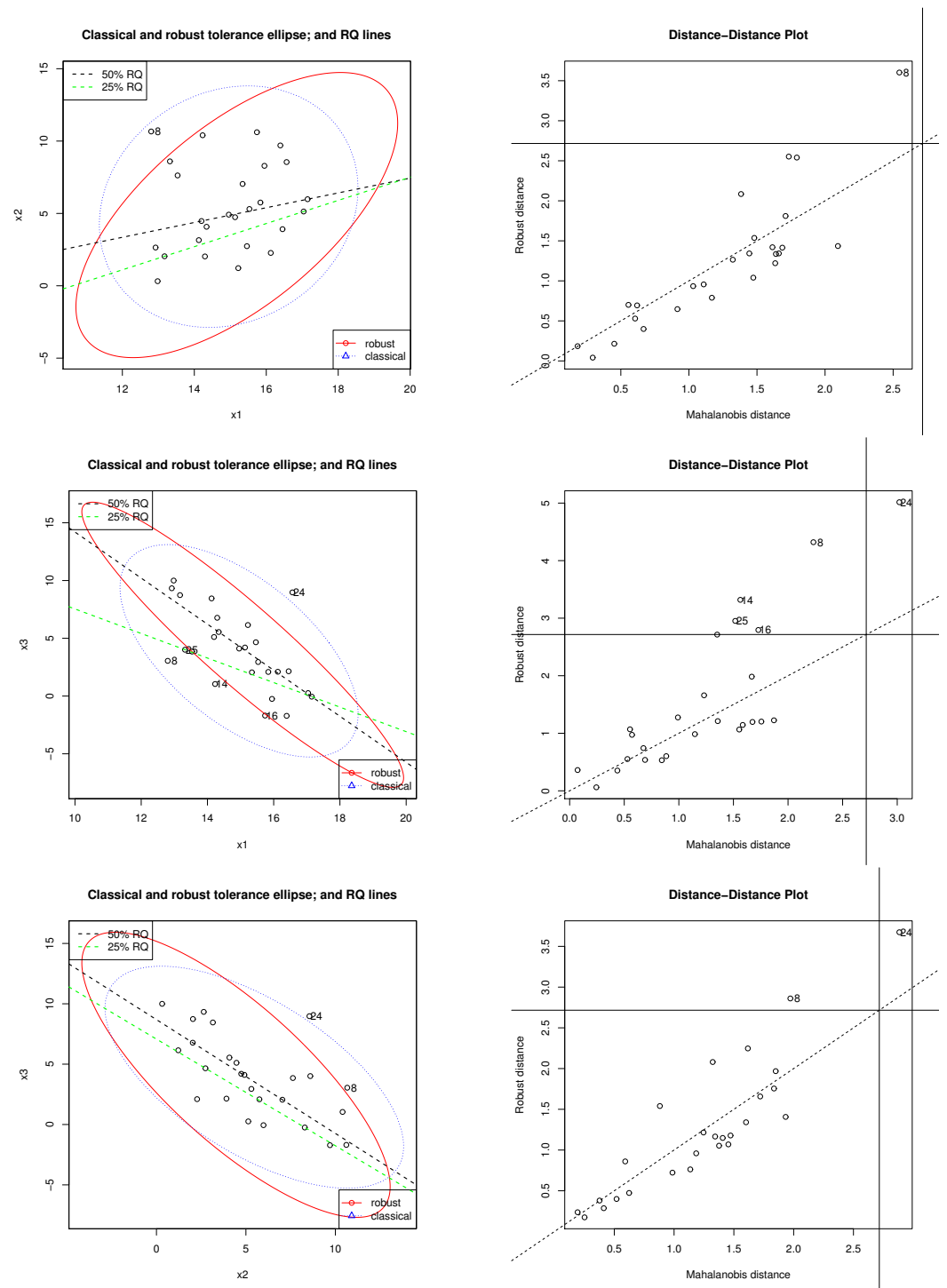
Figure 8. Tolerance Ellipses and Distance-Distance Plots for the [36] data set.

The results for the reduced data set without observation 24 are given in Table 5, leaving only one mild leverage point, observation 8. Cross validation results have consistently chosen unpenalized QR and unpenalized WQR models as the optimal models except for QR-LASSO, QR-E-NET, and WRQ-LASSO at $\tau = 0.5$, where $\lambda \neq 0$. The best results are exhibited under unpenalized WQR at $\tau = 0.5$ followed by unpenalized WQR at $\tau = 0.25$.

Table 4. Results for the full [36] data set.

		β	UNWEIGHTED							
			NONE-BIASED		RIDGE		LASSO		E-NET	
			$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias
λ			0.00		0.11		0.06		0.11	
RQ $\tau = 0.5$	intercept	0.00	25.09	−25.09	24.34	−24.34	27.63	−27.63	23.63	−23.63
	X_1	3.00	1.55	1.45	0.86	2.14	1.28	1.72	1.06	1.94
	X_2	−2.00	−2.30	0.30	−0.86	−1.14	−2.12	0.12	−1.21	−0.79
	X_3	0.00	−0.66	0.66	0.17	−0.17	−0.49	0.49	0.00	0.00
λ			0.00		0.00		0.06		0.06	
RQ $\tau = 0.25$	intercept	−1.00	23.53	−24.53	25.26	−26.26	30.32	−31.32	33.13	−34.13
	X_1	3.00	1.19	1.81	1.09	1.91	0.56	2.44	0.30	2.70
	X_2	−2.00	−1.96	−0.04	−1.98	−0.02	−1.70	−0.30	−1.53	−0.47
	X_3	0.00	−0.15	0.15	−0.16	0.16	0.00	0.00	0.02	−0.02
		β	WEIGHTED							
			NONE-BIASED		WQR-RIDGE		WQR-LASSO		WQR-E-NET	
			$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias
RQ $\tau = 0.5$	intercept	0.00	25.09	−25.09	0.36	−0.36	0.36	−0.36	0.36	−0.36
	X_1	3.00	1.55	1.45	2.94	0.06	2.94	0.06	2.94	0.06
	X_2	−2.00	−2.30	0.30	−2.08	0.08	−2.08	0.08	−2.08	0.08
	X_3	0.00	−0.66	0.66	0.01	−0.01	0.01	−0.01	0.01	−0.01
λ			0.00		0.00		0.00		0.00	
RQ $\tau = 0.25$	intercept	−1.00	23.53	−24.53	−0.08	−0.92	−0.08	−0.92	7.62	−8.62
	X_1	3.00	1.19	1.81	2.95	0.05	2.95	0.05	2.95	0.05
	X_2	−2.00	−1.96	−0.04	−2.47	0.47	−2.47	0.47	−2.47	0.47
	X_3	0.00	−0.15	0.15	−0.03	0.03	−0.03	0.03	−0.03	0.03

Table 5. Results for [36] data set without observation 24.

		β	UNWEIGHTED							
			NONE-BIASED		RIDGE		LASSO		E-NET	
			$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias
λ			0.00		0.00		0.22		0.08	
RQ $\tau = 0.5$	intercept	0.00	−59.31	59.31	−56.47	56.47	40.67	−40.67	8.77	−8.77
	X_1	3.00	5.78	−2.78	5.65	−2.65	0.00	3.00	2.09	0.91
	X_2	−2.00	−0.22	−1.78	−0.32	−1.68	−1.18	−0.82	−1.37	−0.63
	X_3	0.00	2.13	−2.13	2.05	−2.05	0.00	0.00	0.30	−0.30
λ			0.00		0.00		0.00		0.00	
RQ $\tau = 0.25$	intercept	−1.00	−56.16	55.16	−59.60	58.60	−59.61	58.61	−59.61	58.61
	X_1	3.00	5.67	−2.67	5.80	−2.80	5.80	−2.80	5.80	−2.80
	X_2	−2.00	−0.61	−1.39	−0.48	−1.52	−0.48	−1.52	−0.48	−1.52
	X_3	0.00	1.96	−1.96	2.13	−2.13	2.13	−2.13	2.13	−2.13
		β	WEIGHTED							
			NONE-BIASED		WQR-RIDGE		WQR-LASSO		WQR-E-NET	
			$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias	$\widehat{\beta}$	Bias
λ			0.00		0.00		0.06		0.00	
RQ $\tau = 0.5$	intercept	0.00	−59.31	59.31	0.12	−0.12	−0.24	0.24	0.12	−0.12
	X_1	3.00	5.78	−2.78	2.88	0.12	2.77	0.23	2.88	0.12
	X_2	−2.00	−0.22	−1.78	−1.88	−0.12	−1.44	−0.56	−1.88	−0.12
	X_3	0.00	2.13	−2.13	0.07	−0.07	0.20	−0.20	0.07	−0.07
λ			0.00		0.00		0.00		0.00	
RQ $\tau = 0.25$	intercept	−1.00	−56.16	55.16	−0.37	−0.63	−0.37	−0.63	−0.37	−0.63
	X_1	3.00	5.67	−2.67	3.02	−0.02	3.02	−0.02	3.02	−0.02
	X_2	−2.00	−0.61	−1.39	−2.59	0.59	−2.59	0.59	−2.59	0.59
	X_3	0.00	1.96	−1.96	−0.12	0.12	−0.12	0.12	−0.12	0.12

6. Conclusions

This paper suggested a penalized WQR procedure making use of robust weights based on the computationally intensive high breakdown MCD method rather than the well-known classical Mahalanobis distance or any other LS based weights as in [20] which are amenable to outliers. As penalty functions, the RIDGE, LASSO, and E-NET penalties were used yielding the procedures WQR-RIDGE, WQR-LASSO, and the WQR-E-NET, respectively. The efficacy of these procedures as a remedy to high leverage points and collinearity influential points were investigated via simulations and applications to well-known data sets from the literature.

Simulation studies show that generally penalized versions of robustly WQR performs better than their unweighted counterparts, with the WQR-LASSO generally performing the best although marginally so at D2 and D4 under the Normal distribution. However, there are few exceptions; at D2 and D4 under the Normal distribution, with respect to model/variable selection, WQR-LASSO and WQR-E-NET alternately dominate each other while, with respect to prediction, penalized QR performs better than penalized WQR. The occasional dominance of the WQR-E-NET over WQR-LASSO is expected (see, e.g., [11]).

Applications to well-known data sets from the literature show that, in some cases, applying the MCD based robust weight is adequate, i.e., WQR (tuning parameter $\lambda = 0$) performs better than penalized WQR. The best performance is mostly at a quantile level $\tau = 0.5$ while WQR-LASSO and WQR-E-NET are too "greedy" at $\tau = 0.25$, shrinking all parameters to zero. Thus, overall, simulations and applications to the [35] and the [36] data sets show an improvement in variable selection and regularization due to the robust weighting formulation.

Abbreviations

The following abbreviations are used in this manuscript:

LS	Least Squares
QR	Quantile Regression
RQ	Regression Quantile
$Q_{Y\mid X}(\tau)$	Regression Quantile at quantile level $0 < \tau < 1$
WQR	Weighted Quantile Regression
LAD	Least Absolute Deviation
LP	Linear Programming
LASSO	Least Absolute Shrinkage and Selection Operator
E-NET	Elastic Net
SCAD	Smoothly Clipped Absolute Deviation
CQR	Composite Quantile Regression
MCD	Minimum Covariance Determinant
SVD	Singular value Decomposition

References

1. Koenker, R.W.; Bassett, G. Regression Quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
2. Yu, K.; Lu, Z.; Stander, J. Quantile regression: Applications and current research areas. *J. R. Stat. Soc. Ser. D* **2003**, *52*, 331–350. [CrossRef]
3. Koenker, R. *Econometric Society Monographs: Quantile Regression*; Cambridge University: New York, NY, USA, 2005.
4. Chatterjee, S.; Hadi, A.S. Impact of simultaneous omission of a variable and an observation on a linear regression equation. *Comput. Stat. Data Anal.* **1988**, *6*, 129–144. [CrossRef]
5. Hubert, M.; Rousseeuw, P.J. Robust regression with both continuous and binary regressors. *J. Stat. Plan. Inference* **1997**, *57*, 153–163. [CrossRef]
6. Salibián-Barrera, M.; Wei, Y. Weighted quantile regression with nonelliptically structured covariates. *Can. J. Stat.* **2008**, *36*, 595–611. [CrossRef]
7. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
8. Wang, H.; Li, G.; Jiang, G. Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso. *J. Bus. Econ. Stat.* **2007**, *25*, 347–355. [CrossRef]
9. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
10. Obuchi, T.; Kabashima, Y. Cross validation in {LASSO} and its acceleration. *J. Stat. Mech. Theory Exp.* **2016**, *2016*, 53304. [CrossRef]
11. Zou, H.; Hastie, T. Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
12. Yi, C.; Huang, J. Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression. *J. Comput. Graph. Stat.* **2017**, *26*, 547–557. [CrossRef]
13. Wu, Y.; Liu, Y. Variable selection in quantile regression. *Stat. Sin.* **2009**, *19*, 801–817.
14. Jiang, X.; Jiang, J.; Song, X. Oracle model selection for nonlinear models based on weighted composite quantile regression. *Stat. Sin.* **2012**, *22*, 1479–1506.
15. Zou, H. The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
16. Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 91–108. [CrossRef]
17. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* **2006**, *68*, 49–67. [CrossRef]
18. Belloni, A.; Chernozhukov, V. ℓ_1-penalized quantile regression in high-himensional sparse models. *Ann. Stat.* **2011**, *39*, 82–130. [CrossRef]
19. Zou, H.; Yuan, M. Composite quantile regression and the oracle model selection theory. *Ann. Stat.* **2008**, *36*, 1108–1126. [CrossRef]
20. Arslan, O. Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Comput. Stat. Data Anal.* **2012**, *56*, 1952–1965. [CrossRef]
21. Norouzirad, M.; Hossain, S.; Arashi, M. Shrinkage and penalized estimators in weighted least absolute deviations regression models. *J. Stat. Comput. Simul.* **2018**, *88*, 1557–1575. [CrossRef]
22. Hoerl, A.E.; Kannard, R.W.; Baldwin, K.F. Ridge regression:some simulations. *Commun. Stat.* **1975**, *4*, 105–123. [CrossRef]
23. Lawless, J.F.; Wang, P. A simulation study of ridge and other regression estimators. *Commun. Stat. Theory Methods* **1976**, *5*, 307–323. [CrossRef]
24. Hocking, R.R.; Speed, F.M.; Lynn, M.J. A Class of Biased Estimators in Linear Regression. *Technometrics* **1976**, *18*, 425–437. [CrossRef]
25. Kibria, B.M.G. Performance of Some New Ridge Regression Estimators. *Commun. Stat. Simul. Comput.* **2003**, *32*, 419–435. [CrossRef]
26. Khalaf, G.; Shukur, G. Choosing Ridge Parameter for Regression Problems. *Commun. Stat. Theory Methods* **2005**, *34*, 1177–1182. [CrossRef]
27. Rousseeuw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*; John Wiley & Sons: Hoboken, NY, USA, 2005; Volume 589.
28. Rousseeuw, P. Multivariate Estimation with High Breakdown Point. *Math. Stat. Appl.* **1985**, 283–297. [CrossRef]
29. Rousseeuw, P.J.; Hubert, M. Anomaly detection by robust statistics. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1236. [CrossRef]
30. Giloni, A.; Simonoff, J.S.; Sengupta, B. Robust weighted LAD regression. *Comput. Stat. Data Anal.* **2006**, *50*, 3124–3140. [CrossRef]
31. Pollard, D. Asymptotics for Least Absolute Deviation Regression Estimators. *Econom. Theory* **1991**, *7*, 186–199. [CrossRef]
32. Knight, K. Asymptotics for L1-estimators of regression parameters under heteroscedasticityY. *Can. J. Stat.* **1999**, *27*, 497–507. [CrossRef]
33. Ranganai, E.; Van Vuuren, J.O.; De Wet, T. Multiple case high leverage diagnosis in regression quantiles. *Commun. Stat. Theory Methods* **2014**, *43*, 3343–3370. [CrossRef]
34. Yi, C. *Hqreg: Regularization Paths for Huber Loss Regression and Quantile Regression Penalized by Lasso or Elastic-Net*; R Package Version 1.3; R-CRAN: Repository, 2016. Available online: https://cran.r-project.org/web/packages/hqreg/index.html (accessed on 28 October 2020)
35. Hawkins, D.M.; Bradu, D.; Kass, G.V. Location of Several Outliers in Multiple-Regression Data Using Elemental Sets. *Technometrics* **1984**, *26*, 197–208. [CrossRef]

36. Hocking, R.; Pendleton, O. The regression dilemma. *Commun. Stat. Theory Methods* **1983**, *12*, 497–527. [CrossRef]
37. Bagheri, A.; Habshah, M.; Imon, R. A novel collinearity-influential observation diagnostic measure based on a group deletion approach. *Commun. Stat. Simul. Comput.* **2012**, *41*, 1379–1396. [CrossRef]