JOURNAL OF APPLIED STATISTICS
2021, VOL. 48, NO. 2, 234–246
https://doi.org/10.1080/02664763.2020.1722079

Taylor & Francis
Taylor & Francis Group

Outlier detection and robust variable selection via the penalized weighted LAD-LASSO method

Yunlu Jiang[a], Yan Wang[a], Jiantao Zhang[a], Baojian Xie[b], Jibiao Liao[c] and Wenhui Liao[d]

[a]Department of Statistics, College of Economics, Jinan University, Guangzhou, People's Republic of China; [b]College of Economics, Jinan University, Guangzhou, People's Republic of China; [c]Office of Educational Administration, Dongguan Open University, Dongguan, People's Republic of China; [d]School of Financial Mathematics and Statistics, Guangdong University of Finance, Guangzhou, People's Republic of China

ABSTRACT

This paper studies the outlier detection and robust variable selection problem in the linear regression model. The penalized weighted least absolute deviation (PWLAD) regression estimation method and the adaptive least absolute shrinkage and selection operator (LASSO) are combined to simultaneously achieve outlier detection, and robust variable selection. An iterative algorithm is proposed to solve the proposed optimization problem. Monte Carlo studies are evaluated the finite-sample performance of the proposed methods. The results indicate that the finite sample performance of the proposed methods performs better than that of the existing methods when there are leverage points or outliers in the response variable or explanatory variables. Finally, we apply the proposed methodology to analyze two real datasets.

ARTICLE HISTORY
Received 11 March 2019
Accepted 21 January 2020

KEYWORDS
Outlier detection; robust regression; penalized weighted least absolute deviation; LASSO; variable selection

1. Introduction

Chatterjee and Hadi [6] pointed out that the presence of outliers could lead to biased parameter estimation, and inappropriate predictions for the classical methods. Meanwhile, they thought that outliers can be occurred in the response variable, in the explanatory variable, or in both the response and explanatory variables. Hampel *et al.* [17] pointed out that a real dataset maybe contain 1–10% outliers. Therefore, the detection of outliers is an important issue in regression analysis. Many methods have been developed to detect outliers in the multiple linear regression [16,20,25,31–33]. Specially, robust regression is an important and effective estimation technique for analyzing data that are contaminated with outliers. Several robust regression techniques have been developed for outlier detection problems, such as S-estimator [19], the least median of squares method [34], the MM-estimator [45], the τ-estimator [46], the robust and efficient weighted least squares estimators [14], and so on. It is well known that LAD regression is lack of robustness when the data include outliers in the explanatory variables (i.e. there exist leverage points). In the framework of LAD regression, the robustness can also be improved by down-weighting those leverage points

CONTACT Jibiao Liao ✉ Liaojibiao20010512@163.com; Wenhui Liao ✉ lwhgduf@sina.com
This article has been republished with minor changes. These changes do not impact the academic content of the article.

which are detected in advance. For example, Giloni *et al.* [15] proposed a weighted LAD (WLAD) regression to improve the robustness of LAD regression. However, as pointed out by [43], the robustness of WLAD estimator can be also significantly deteriorated by a high percentage of outliers. Gao and Feng [13] proposed a penalized weighted least absolute deviation regression (PWLAD) method for both outlier identification and robust estimation. It leads to estimators with high robustness, and obtains the observations' outlying information. She and Owen [36] considered outlier detection using non-convex penalized regression on the mean shift parameters for the mean shift linear regression model.

In practice, many covariates usually are introduced at the initial stage of modeling. However, it is very difficult to interpret models including all the covariates while irrelevant variables maybe increase the variance. Therefore, selection of important covariates is one of the most important problems in data analysis. Popular methods for variable selection are penalized regression methods such as LASSO [37], Smoothly Clipped Absolute Deviation (SCAD) [10] and adaptive LASSO [47], and so on. It is important to note that many of those methods are closely related to the least squares technique. It is well-known that the least squares method is not robust. Therefore, the outliers can present serious problems for the least-squares-based methods in variable selection. It is necessary to study the robust variable selection method. There exist many robust variable selection method in the literature, such as the least absolute deviation (LAD)-LASSO [40], which is robust for heavy-tailed error, WLAD-LASSO [2], the weighted Wilcoxon-type SCAD method [38], the variable selection via regularized rank regression [42], the non-concave penalized M-estimation method [30], the Hubers criterion and adaptive lasso penalty [29], the variable selection in quantile regression [42], the quantile regression for analyzing heterogeneity in ultra-high dimension [41], the variable selection in the semiparametric varying-coefficient partially linear model via a penalized composite quantile loss [26], the composite quantile regression (CQR) [49], the penalized composite quasi-likelihood for ultrahigh-dimensional variable selection [5], the variable selection with exponential squared loss [21–24,39], the weighted robust Lasso (WR-Lasso) [9], the sparse least trimmed squares regression [1], the efficient robust doubly adaptive regularized regression [27], and so on.

In this paper, we will simultaneously study the outlier detection and robust variable selection problem in the linear regression model. For the mean shift linear regression model, Kong *et al.* [28] apply an adaptive regularization to these shift parameters to shrink most of them to zero, and can simultaneously achieve the outlier detection and robust variable selection. However, they focused only on the mean shift linear regression model to achieve robustness to additive outliers. In this article, we are interested to achieve robustness to outliers, and detect outliers in the explanatory variable, in the explanatory variable, or in both the response and explanatory variables. PWLAD associates each observation with a weight and obtain both the weight and regression coefficient estimates simultaneously using a lasso-type penalty on the weight vector. Even under the scenario where the data include both contaminated observations (in both y and x directions) and the random errors may not have finite variance, the PWLAD is still able to detect corresponding outliers and provide robust regression coefficient estimates. But the PWLAD can't detect additive outliers effectively. An optimized outlier detection method is proposed to solve the this problem. Advocated by the PWLAD, we propose a penalized weighted LAD-LASSO(PWLAD-LASSO) to simultaneously achieve outlier detection and robust variable selection by combining the penalized weighted least absolute deviation

(PWLAD) regression estimation method and the adaptive LASSO. Monte Carlo studies indicate that the finite sample performance of the proposed methods performs better than that of the existing methods when there are outliers in the explanatory variable, or in both the response and explanatory variables.

The remainder of the paper is organized as follows. In Section 2, we introduce the PWLAD-LASSO estimator and its corresponding algorithm, and study the problem of initial value and tuning parameter selection. In Section 3, simulation studies are conducted to evaluate the finite sample performance of the proposed methods. In Section 4, two real datasets are analyzed to compare the proposed methods with the existing methods. A discussion is given in Section 5.

2. Methodology

2.1. PWLAD

Consider the linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $y_i \in R$ is the response variable, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is the p-dimensional covariate vector, $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)'$ is the true coefficients and $\{\varepsilon_i, i = 1, \ldots, n\}$ are independent random errors with unknown distribution G. We further assume that G is symmetric about 0, which is often assumed in the literature, see [39].

Gao and Feng [13] introduced the PWLAD method by minimizing the following penalized objective function,

$$(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{w}})(\lambda) = \underset{\boldsymbol{\beta}, \mathbf{w}}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} w_i^2 |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \lambda \sum_{i=1}^{n} \varpi_i |1 - w_i| \right\}. \tag{2}$$

where $\mathbf{w} = (w_1, \ldots, w_n)'$, $0 < w_i \le 1$ represent the weights quantifying the outlying effects for each observation. For $1 \le i \le n$, the weights $0 < w_i \le 1$ represent the heterogeneity of the errors with $w_i < 1$ representing an outlier and $w_i = 1$ representing a 'normal' observation. $\lambda \sum_{i=1}^{n} \varpi_i |1 - w_i|$ is a penalty shrinking all weight to the direction of 1. ϖ_i's include some prior information on the outlying status of all observation, and λ is a tuning parameter in $(0, \infty)$. If some outlying information is incorporated into ϖ_i, the outlier detection accuracy can be significantly improved. For example, suppose an initial value on weight $w_i^{(0)}$ is obtained, we can set $\varpi_i = 1/|\log(w_i^{(0)})|$. A larger $0 < w_i^{(0)} \le 1$ produces a larger penalty $\lambda\varpi_i$ on $|1 - w_i|$, which pushes $\widehat{w}_i$ more to 1. If $w_i^{(0)} = 1$, then we have $\varpi_i = \infty$, which would leads to $\widehat{w}_i = 1$. On the other hand, when $w_i^{(0)} \to 0$, $\widehat{w}_i$ is usually much smaller than 1 since $\varpi_i \to 0$ leads to very small penalty being imposed for the ith observation. Meanwhile, the non-differentiability of penalty $\varpi_i|1 - w_i|$ at $w_i = 1$ implies that some of the components of $\widehat{\mathbf{w}}$ may be exactly equal to one, the corresponding observations of which are called 'normal' observations. The other observations with $\widehat{w}_i < 1$ are 'abnormal', with possible outlying in the x and/or y direction. In this regards, the estimated weights provide an automatic way to conduct outlier detection. When λ is sufficiently large, all observations have $w_i = 1$, and no outlier is claimed. However, all w_i are close to zero for sufficiently small λ. Ideally, this penalty term $\lambda \sum_{i=1}^{n} \varpi_i |1 - w_i|$ is expected to generate small weights for those outliers in the explanatory variable or response variable and large weight for normal observations.

2.2. PWLAD-LASSO: method and implementation

According to [40], the LAD-LASSO can obtain robust variable selection. According to [13], the PWLAD method can detect the outliers in the dataset. to simultaneously achieve outlier detection and robust variable selection, we propose a following penalized weighted LAD-LASSO (PWLAD-LASSO) estimator,

$$(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{w}})(\lambda, \rho) = \arg\min_{\boldsymbol{\beta}, \mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{n} w_i^2 |y_i - \mathbf{x}_i' \boldsymbol{\beta}| + \lambda \sum_{i=1}^{n} \varpi_i |1 - w_i| + \rho \sum_{j=1}^{p} \mu_j |\beta_j| \right\}, \quad (3)$$

where $\{\rho\mu_j, j = 1, 2, \ldots, p\}$ are the tuning parameters in the adaptive LASSO objective function and will be estimated from the data.

2.2.1. Model implementation

The PWALD-LASSO estimator can be computed by augmenting the data [2,40] and using the algorithms that will be given later. First, define $(y_i^*, \mathbf{x}_i^{*'})$ for $i = 1, 2, \ldots, n, n+1, \ldots, n+p$, where $(y_i^*, \mathbf{x}_i^{*'}) = \left(\frac{1}{2} w_i^2 y_i, \frac{1}{2} w_i^2 \mathbf{x}_i'\right)$ for $i = 1, 2, \ldots, n$ and $(y_i^*, \mathbf{x}_i^{*'}) = (0, \rho\mu_j \mathbf{e}_j')$ for $i = n+1, \ldots, n+p, j = 1, 2, \ldots, p$ and $\mathbf{e}_j$ is a p-dimensional vector with the jth term equals to 1 and all others equal to zero. Then the optimization problem in (2) can be rewritten as

$$(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{w}})(\lambda, \rho) = \arg\min_{\boldsymbol{\beta}, \mathbf{w}} \left\{ \sum_{i=1}^{n+p} |y_i^* - \mathbf{x}_i^{*'} \boldsymbol{\beta}| + \lambda \sum_{i=1}^{n} \varpi_i |1 - w_i| \right\}. \quad (4)$$

Let $\widehat{\boldsymbol{\beta}}$ be a root n-consistent estimator of $\boldsymbol{\beta}$, then, for any $\gamma > 0$, μ_j can be estimated using the relation $\widehat{\mu}_j = 1/(|\widehat{\beta}_j|^{\gamma})$ for $j = 1, 2, \ldots, p$. In this paper, we set $\gamma = 1$. Then $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{w}})$ can be solved alternatively via the following Algorithm 1 for the dataset $(y_i^*, \mathbf{x}_i^{*'})$.

Where '$\mathbf{a}\cdot\mathbf{B}$' in Algorithm 1 is a special product between vector $\mathbf{a}$ and matrix $\mathbf{B}$. In particular, if $\mathbf{a} = (a_1, \ldots, a_n)$ is a vector and $\mathbf{B}$ is a $n \times p$ matrix with $\mathbf{b}_i$ being its ith row, then '$\mathbf{a}\cdot\mathbf{B}$' is obtained by multiplying each element of $\mathbf{b}_i$ by a_i for $1 \leq i \leq n$.

Remark 2.1: By using the similar idea in [40] and [2], we can translate (3) to (4), and apply the algorithm proposed by Gao and Feng [13] to solve (4). In fact, we can also apply majorize-minimize (MM) algorithm to solve (3) directly. According to Hunter and Lange [18], we can construct a surrogate function for $|y_i - x_i'\beta|$. For the two penalized function $\lambda \sum_{i=1}^{n} \varpi_i |1 - w_i|$ and $\rho \sum_{j=1}^{p} \mu_j |\beta_j|$, we can apply local quadratic approximation (LQA) algorithm [48] to construct the surrogate function. We will apply MM algorithm to solve (3) directly as future work.

2.2.2. Choice of initial weight

Before we use Algorithm 1 to compute the PWLAD-LASSO estimator, we have to find the initial estimates $\boldsymbol{\beta}^{(0)}$, $\mathbf{w}^{(0)}$ and ϖ. In the later numerical study, in order to compare our

Algorithm 1 PWLAD-LASSO Solution for given λ, ρ

```
Given initial estimate $\boldsymbol{\beta}^{(0)}$, $\varpi$ and $\mathbf{w}^{(0)}$
  Let $j = 1$ and $\lambda_i = \lambda\varpi_i$
While not converged do
  [Update $\boldsymbol{\beta}$]
    $\mathbf{y}^{*adj} = \mathbf{w}^{(j-1)} \cdot \mathbf{w}^{(j-1)} \cdot \mathbf{y}^*$,
    $\mathbf{X}^{*adj} = \mathbf{w}^{(j-1)} \cdot \mathbf{w}^{(j-1)} \cdot \mathbf{X}^*$,
    let $\boldsymbol{\beta}^{(j)} = \arg\min_{\boldsymbol{\beta}}\{\| \mathbf{y}^{*adj} - \mathbf{X}^{*adj}\boldsymbol{\beta} \|_1\}$
  [Update $\mathbf{w}$]
    $\mathbf{r}^{(j)} = \mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^j$,
    If $|r_i^{(j)}| > \lambda_i$, let $\mathbf{w}_i^{(j)} \leftarrow \lambda_i/|r_i^{(j)}|$, otherwise $\mathbf{w}_i^{(j)} \leftarrow 1$
    converged $\leftarrow \| \mathbf{w}_i^{(j)} - \mathbf{w}_i^{(j-1)} \|_\infty < \epsilon$
  $j \leftarrow j + 1$
end while
output $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(j)}$ and $\widehat{\mathbf{w}} = \mathbf{w}^{(j)}$.
```

method with [28], we use the same way to get the initial estimate. We use the least trimmed squares method to obtain the initial robust estimates. The least trimmed squares initial fit carries over the high robustness of our proposed estimator [28]. For implementation, the R function ‘ltsReg’ is adopted to obtain the initial estimates $\boldsymbol{\beta}^{(0)}$.

First, we compute the leverage values from a clean data set [3] and the corresponding robust Mahalanobis distance based on MCD estimator [11]. Before computing the leverage value, we first define the scaled design matrix as $\check{\mathbf{Z}} = (\check{\mathbf{y}}, \check{\mathbf{X}}) = (\check{\mathbf{z}}_1, \dots, \check{\mathbf{z}}_n)'$, where $\check{\mathbf{X}} = (\check{\mathbf{x}}_1, \dots, \check{\mathbf{x}}_n)'$, $\check{\mathbf{y}} = (y_1, \dots, y_n)'$, $\check{\mathbf{x}}_i$ is the ith predictor and y_i is the response. $\check{\mathbf{y}}$,$\check{\mathbf{x}}_i$ is required to be scaled to be between 0 and 1 by subtracting its minimum value and then dividing by its maximum value, and letting $\check{\mathbf{d}}$ be the vector consists of median value of each of $p + 1$ columns in $\check{\mathbf{Z}}$. We compute the robust Mahalanobis distances s_i between $\check{\mathbf{z}}_i$ and $\check{\mathbf{d}}$, the clean subset S consists of all m observations with the smallest s_is. Then we compute the leverage values $h_i = \check{\mathbf{z}}_i'(\check{\mathbf{Z}}_S'\check{\mathbf{Z}}_S)^{-1}\check{\mathbf{z}}_i$ for observation i relative to the clean subset. Let $w_i^{(0)} = w_0 \ll 1$ for all those $(n - m)$ observations with the smallest h_is and $w_i^{(0)} = 1$ for the rest observations. Boudt *et al.* [4] pointed out that a major restriction of the MCD estimator is that the dimension p must satisfy $p < h$ for the covariance matrix of any h-subset to be non-singular. For accuracy of the MCD estimator, it is often recommended to take $n > 5p$.

Filzmoser *et al.* [12] proposed a computationally fast procedure PCOut for identifying outliers in high-dimensional data. This algorithm utilizes simple properties of principal components to identify outliers in the transformed space, leading to significant computational advantages for high-dimensional data. Therefore, we use the program PCOut to calculate the distances s_i between $\check{\mathbf{z}}_i$ and $\check{\mathbf{d}}$ in high dimensions. The *R* code for the procedure PCOut is available as the function ‘pcout’ in the *R* library ‘mvoutlier’. Then let $w_i^{(0)} = w_0 \ll 1$ for all those $(n - m)$ observations with the smallest s_i’s and $w_i^{(0)} = 1$ for the rest observations. In this paper, we set $m = [0.6n]$ and $w_0 = 0.01$. To improve the robustness of PWLAD-LASSO regarding the outliers, we suggest to choose $\varpi_i = 1/|\log(w_i^{(0)})|$ in (4).

Remark 2.2: In order to implement the proposed PWLAD-LASSO method, we need to give the initial estimates $\mathbf{w}^{(0)}$. According to the similar idea in [13], we obtain the initial estimates $\mathbf{w}^{(0)}$. However, this initial estimates need to use the robust Mahalanobis distance based on MCD estimator, which is accurate for $n > 5p$ [4]. For the more high-dimensional data, we apply the procedure PCOut [12] to obtain the initial estimates $\mathbf{w}^{(0)}$. Unfortunately, it is efficient for $p \leq n$. We will study the proposed PWLAD-LASSO method for $p > n$ as future work.

2.2.3. Tuning parameter selection

To implement the above proposed Algorithm 1, we need to obtain a proper tuning parameter λ and ρ in the process of computation. The selection of tuning parameters λ and ρ play an important role in the performance of both outlier identification and parameter estimation. In general, there are many methods to select these parameters, such as cross-validation (CV), generalized cross-validation, Akaike information criterion, and Bayesian information criterion (BIC). In this paper, we apply the BIC to select the tuning parameters. Suppose $\widehat{\boldsymbol{\beta}}_{\lambda,\rho}$ are the estimates when the tuning parameters are set as λ and ρ. Let $e_i^2 = \{w_i^2(y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\lambda,\rho})\}^2$, and define the residual sum of squares as $\mathrm{RSS} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{e}_{\mathrm{i}}^2$. The BIC is defined as

$$\mathrm{BIC}(\lambda, \rho) = n \log(\mathrm{RSS}/n) + k \log n, \tag{5}$$

where k is the degree of freedom, the number of non-zero components of $\widehat{\boldsymbol{\beta}}$ and $1 - \widehat{\mathbf{w}}$. We set two-dimensional grids for λ and ρ to find the combination that minimizes $\mathrm{BIC}(\lambda, \rho)$. Specifically, we first choose a dense grid on ρ, and then, for each ρ, we use the algorithm proposed in the precious section to obtain the solution paths of the problem in (3). We pick the grid of λ on each point that the degree of freedom changes.

3. Simulation studies

Example 3.1: In this example, we assess the finite sample performance of the proposed method via Monte Carlo simulations. The covariance matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' = \mathbf{U\Sigma}$, where $\mathbf{U} = (u_{jk})_{x \times p}$ with $u_{jk} \overset{i.i.d}{\sim} \mathit{Unif}(-5, 5)$ and $\boldsymbol{\Sigma} = (\Sigma_{km})_{p \times p}$ with $\Sigma_{km} = 0.5^{|k-m|}$. The true coefficient was set as $\boldsymbol{\beta}^* = (4, 2, 1, \ldots, 0)'$ with $q = 3$ non-zero components and the remaining $(p - q)$ elements zero. The random error was simulated independently from $\varepsilon \sim N(0, 0.25)$. The data were generated from $y_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \varepsilon_i$ for $1 \leq i \leq n$. We contaminated the first $[cn]$ observations by setting $\mathbf{x}_i^o$ where $x_{i1}^o = x_{i1} + L, x_{i6}^o = x_{i6} + L$ or $y_i^o = y_i + V$ for $1 \leq i \leq cn$ with parameters L and V given later. Thus the first $[cn]$ observations were outliers and the remainder normal points.

To investigate the finite sample performance of PWLAD-LASSO, we compute the following eight measures:

M: the masking probability (fraction of undetected true outliers);
S: the swamping probability (fraction of good points labeled as outliers);
JD: the joint outlier detection rate (fraction of simulations with 0 masking);

FZR: the false zero rate (fraction of non-zero coefficients that are estimated as zero)

$$\mathrm{FZR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1,\ldots,p\} : \widehat{\beta}_j = 0 \wedge \beta_j \neq 0\}|/|\{j \in \{1,\ldots,p\} : \beta_j \neq 0\}| \qquad (6)$$

where $|S|$ denotes the size of the set S;
FPR: the false positive rate (fraction of zero coefficients that are estimated as non-zero)

$$\mathrm{FPR}(\widehat{\boldsymbol{\beta}}) = |\{j \in \{1,\ldots,p\} : \widehat{\beta}_j \neq 0 \wedge \beta_j = 0\}|/|\{j \in \{1,\ldots,p\} : \beta_j = 0\}| \qquad (7)$$

SR: the correct selection rate (fraction of identifying both non-zeros and zeros of β);
CR: the correct coverage rate (fraction of identifying non-zeros of β).
MSE: the mean square error of the parameters

$$\mathrm{MSE} = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' E\left(\mathbf{x}'\mathbf{x}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right), \qquad (8)$$

where $\mathbf{x}$ is the uncontaminated covariates. For performance in terms of outlier detection, M and S should be as small as possible while JD should be as large as possible. For sparse estimator of β, FZR and FPR should be as small as possible and SR and CR as large as possible. With respect to the estimation accuracy of $\boldsymbol{\beta}$, MSE should be as small as possible.

We compared our proposed method with the WLAD-LASSO method [2] and Kong's method (PM) [28]. We generate outliers using a mean shift model. We used different (n, p, L, V, c) combinations. For each combination, we ran Monte Carlo studies with 200 replicates. The corresponding results are given in Table 1.

From Table 1, we can observe that our proposed PWLAD-LASSO method perform better than PM method in terms of parameter estimation, outlier detection and variable selection. PWLAD-LASSO can detect all outliers correctly, but PM method loses its outlier detection ability almost completely across all settings when there are outliers in x directions or in both x and y directions. When there are outliers in y directions, the PM method also has very good performance in terms of parameter estimation, outlier detection and variable selection. For the WLAD-LASSO method, it obtains very good performance for variable selection. However, it is based on a pre-assigned weight depending on a pre-selected clean subset. As pointed out by Čížek [7] and Gao and Feng [13], if a large amount of leverage points exist, the clean subset and the produced weight assignment can be misleading, which cause the WLAD estimator to be severely biased. Therefore, the WLAD-LASSO method obtains higher MSE than the PWLAD-LASSO method. Furthermore, our proposed PWLAD-LASSO method is more efficient in estimating $\boldsymbol{\beta}$ than other two methods since our proposed method has smaller MSE than other two methods.

Example 3.2: In Example 3.1, we only consider a very small number of parameters $p = 15$. Therefore, in this example, we will illustrate that our proposed method can also work for high-dimensional data. We use the same setting as in Example 3.1, except that $\boldsymbol{\beta}^* = (4, 2, 1, 0.5, 0.2, \ldots, 0)'$, $(n, p, L, V, c) = (200, 50, 10, 0, 0.1)$ or $(200, 50, 10, 5, 0.1)$. The corresponding results are given in Table 2. From Table 2, we can find that our proposed PWLAD-LASSO method performs better than PM method with regard to outliers detection and variable selection.

Table 1. Simulation results for Example 3.1.

(n, p, L, V, c)	Method	M	S	JD	FZR	FPR	SR	CR	MSE
$(100, 15, 0, 5, 0.1)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0.004(0.0002)	0.97	1	0.004(0.0001)
	WLAD-LASSO	–	–	–	0(0)	0(0)	1	1	0.027(0.0001)
	PM	0(0)	0.004(0.0001)	1	0(0)	0.060(0.0006)	0.63	1	0.014(0.0001)
$(100, 15, 0, 5, 0.2)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0.003(0.0001)	0.95	1	0.005(0.0001)
	WLAD-LASSO	–	–	–	0(0)	0(0)	1	1	0.041(0.0003)
	PM	0(0)	0.004(0.0001)	1	0(0)	0.061(0.0006)	0.56	1	0.021(0.0001)
$(100, 15, 10, 0, 0.1)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0.007(0.0002)	0.94	1	0.004(0.0001)
	WLAD-LASSO	–	–	–	0.020(0.0008)	0(0)	0.94	0.94	2.646(0.0169)
	PM	0.040(0.0006)	0.849(0.0006)	0.79	0.067(0.0013)	0.300(0.0016)	0	0.8	4.160(0.0540)
$(100, 15, 10, 0, 0.2)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0.003(0.0002)	0.92	1	0.005(0.0001)
	WLAD-LASSO	–	–	–	0.047(0.0012)	0(0)	0.86	0.86	18.302(0.0245)
	PM	0.188(0.0017)	0.810(0.0009)	0	0.300(0.0024)	0.138(0.0015)	0	0.12	17.891(0.0488)
$(100, 15, 10, 5, 0.1)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0.006(0.0003)	0.94	1	0.003(0.0001)
	WLAD-LASSO	–	–	–	0.003(0.0003)	0(0)	0.99	0.99	2.487(0.0164)
	PM	0.056(0.0021)	0.827(0.0013)	0.74	0.053(0.0024)	0.300(0.0030)	0	0.84	3.519(0.0864)
$(100, 15, 10, 5, 0.2)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0.008(0.0003)	0.91	1	0.004(0.0001)
	WLAD-LASSO	–	–	–	0.010(0.0006)	0(0)	0.97	0.97	14.179(0.0191)
	PM	0.215(0.0018)	0.800(0.0011)	0.02	0.240(0.0030)	0.146(0.0021)	0.06	0.28	13.540(0.0530)
$(100, 15, 0, 0, 0)$	PWLAD-LASSO	–	0(0)	–	0(0)	0.016(0.0052)	0.90	1	0.002(0.0001)
	WLAD-LASSO	–	–	–	0(0)	0(0)	1	1	0.021(0.0001)
	PM	–	0(0)	–	0(0)	0.042(0.0010)	0.76	1	0.008(0.0001)
$(200, 15, 0, 5, 0.1)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0(0)	1	1	0.002(0.0001)
	WLAD-LASSO	–	–	–	0(0)	0(0)	1	1	0.026(0.0001)
	PM	0(0)	0.001(0.0001)	1	0(0)	0.045(0.0007)	0.64	1	0.007(0.0001)
$(200, 15, 0, 5, 0.2)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0(0)	1	1	0.002(0.0001)
	WLAD-LASSO	–	–	–	0(0)	0(0)	1	1	0.037(0.0002)
	PM	0(0)	0.003(0.0001)	1	0(0)	0.047(0.0005)	0.62	1	0.013(0.0001)
$(200, 15, 10, 0, 0.1)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0(0)	0.99	1	0.002(0.0001)
	WLAD-LASSO	–	–	–	0.007(0.0005)	0(0)	0.98	0.98	2.094(0.0149)
	PM	0.159(0.0017)	0.707(0.0013)	0.14	0.200(0.0033)	0.293(0.0019)	0	0.40	10.832(0.0691)
$(200, 15, 10, 0, 0.2)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0(0)	1	1	0.002(0.0001)
	WLAD-LASSO	–	–	–	0.010(0.0006)	0(0)	0.97	0.97	17.965(0.0192)
	PM	0.255(0.0012)	0.736(0.0008)	0	0.273(0.0025)	0.260(0.0018)	0	0.18	15.191(0.0451)
$(200, 15, 10, 5, 0.1)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0(0)	0.99	1	0.002(0.0001)
	WLAD-LASSO	–	–	–	0(0)	0(0)	1	1	1.901(0.0144)
	PM	0.166(0.0020)	0.691(0.0013)	0.16	0.153(0.0034)	0.310(0.0018)	0	0.54	8.132(0.0669)
$(200, 15, 10, 5, 0.2)$	PWLAD-LASSO	0(0)	0(0)	1	0(0)	0(0)	0.99	1	0.002(0.0001)
	WLAD-LASSO	–	–	–	0.007(0.0005)	0(0)	0.98	0.98	13.996(0.0140)
	PM	0.304(0.0013)	0.681(0.0008)	0	0.253(0.0028)	0.276(0.0019)	0	0.24	11.807(0.0345)
$(200, 15, 0, 0, 0)$	PWLAD-LASSO	–	0(0)	–	0(0)	0.002(0.0001)	0.97	1	0.001(0.0001)
	WLAD-LASSO	–	–	–	0(0)	0(0)	1	1	0.017(0.0001)
	PM	–	0(0)	–	0(0)	0.105(0.0039)	0.77	1	0.005(0.0001)

Table 2. Simulation results for Example 3.2.

(n, p, L, V, c)	Method	M	S	JD	FZR	FPR	SR	CR	MSE
$(200, 50, 10, 0, 0.1)$	PWLAD-LASSO	0.006(0.0004)	0(0)	0.9	0.016(0.0010)	0.055(0.0018)	0.80	0.92	0.057(0.0030)
	PM	0.173(0.0040)	0.710(0.0029)	0.10	0.182(0.0055)	0.113(0.0024)	0	0.20	11.065(0.1841)
$(200, 50, 10, 5, 0.1)$	PWLAD-LASSO	0.012(0.0007)	0(0)	0.86	0.004(0.0005)	0.067(0.0020)	0.74	0.98	0.081(0.0061)
	PM	0.210(0.0041)	0.667(0.0030)	0.05	0.110(0.0047)	0.104(0.0025)	0	0.45	8.673(0.1093)

Example 3.3: In this example, we will illustrate that our proposed PWLAD-LASSO method can detect almost all outliers correctly. We use the same setting as in Example 3.1, except that $\boldsymbol{\beta}^* = (4, 2, 1, 0.5, 0.2, \ldots, 0)'$, $(n, p, L, V, c) = (100, 15, 0, 5, 0.1)$, $(100, 15, 10, 0, 0.1)$,$(100, 15, 10, 5, 0.1)$, $(200, 50, 10, 0, 0.1)$ or $(200, 50, 10, 5, 0.1)$. We obtain the corresponding estimator for $w_i < 1$. The results are shown in Table 3, where $\widehat{w}_i$

Table 3. Simulation results for in Example 3.3.

(n,p,L,V,c)	$\widehat{w}_1$	$\widehat{w}_2$	$\widehat{w}_3$	$\widehat{w}_4$	$\widehat{w}_5$	$\widehat{w}_6$	$\widehat{w}_7$	$\widehat{w}_8$	$\widehat{w}_9$	$\widehat{w}_{10}$
$(100,15,0,5,0.1)$	0.438	0.394	0.424	0.376	0.406	0.364	0.371	0.392	0.413	0.416
$(100,15,10,0,0.1)$	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.018	0.019	0.019
$(100,15,10,5,0.1)$	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.042	0.043
$(200,50,10,0,0.1)$	0.040	0.040	0.039	0.040	0.039	0.041	0.040	0.041	0.039	0.041
	$\widehat{w}_{11}$	$\widehat{w}_{12}$	$\widehat{w}_{13}$	$\widehat{w}_{14}$	$\widehat{w}_{15}$	$\widehat{w}_{16}$	$\widehat{w}_{17}$	$\widehat{w}_{18}$	$\widehat{w}_{19}$	$\widehat{w}_{20}$
	0.040	0.040	0.040	0.039	0.039	0.039	0.040	0.039	0.040	0.040
$(200,50,10,5,0.1)$	0.049	0.048	0.048	0.048	0.048	0.048	0.049	0.05	0.048	0.049
	$\widehat{w}_{11}$	$\widehat{w}_{12}$	$\widehat{w}_{13}$	$\widehat{w}_{14}$	$\widehat{w}_{15}$	$\widehat{w}_{16}$	$\widehat{w}_{17}$	$\widehat{w}_{18}$	$\widehat{w}_{19}$	$\widehat{w}_{20}$
	0.048	0.049	0.049	0.048	0.048	0.05	0.049	0.049	0.05	0.049

is an estimator of weight w_i for ith observation. From Table 3, we can observe that PWLAD-LASSO can detect all outliers correctly.

4. Real data analysis

As an illustration, we apply the proposed procedure to analyze the two real dataset. The first data set is the modified wood gravity data [34], and include 5 covariates and 20 observations. The raw data came from [8] and were used to determine the influence of anatomical factors on wood specific gravity. In order to illustrate the resistance of least median of squares (LMS) Regression estimation, Rousseeuw [34] contaminated it by replacing a few observations. The detained modified procedure can be found in [8] and [34]. The standardized residuals based on the LMS fit make it easy to spot the four outliers with indices 4, 6, 8, and 19. Subsequently, Giloni *et al.* [15] and Gao and Feng [13] also analyzed the modified wood gravity data to illustrate the robustness of the WLAD method and PWLAD method, respectively. Therefore, the modified wood gravity data is commonly used as an example for robust regression and outlier detection.

Next, we applied our proposed method and PM method to analyze this data set to perform the robust estimation and the outlier detection. The 20 samples are used as the training set, the 16 samples without outliers are used as the test set. Those four observations 4, 6, 8 and 19 are identified as outliers using PWLAD-LASSO, with weights being estimated as 0.2, 0.18, 0.19 and 0.16, respectively. However, none of these four observations was found using PM method.

Let us now reanalyze the modified wood gravity data using the robust Mahalanobis distance based on the MCD estimator. The MCD estimator looks for the h observations (out of n) whose classical covariance matrix has the lowest possible determinant. The MCD estimate of location is then the average of these h points, whereas the MCD estimate of scatter is a multiple of their covariance matrix. For each observation $\mathbf{x}_i$, we now compute the robust Mahalanobis distance [35] given by

$$\mathrm{RD}(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})'\hat{\Sigma}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})}, \tag{9}$$

where $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ are the MCD location and scatter estimates. If $\mathrm{RD}(\mathbf{x}_i) > \sqrt{\chi^2_{p,0.975}}$, then the observation $\mathbf{x}_i$ could be an outlier, where p is the dimension of $\mathbf{x}_i$. The robust Mahalanobis

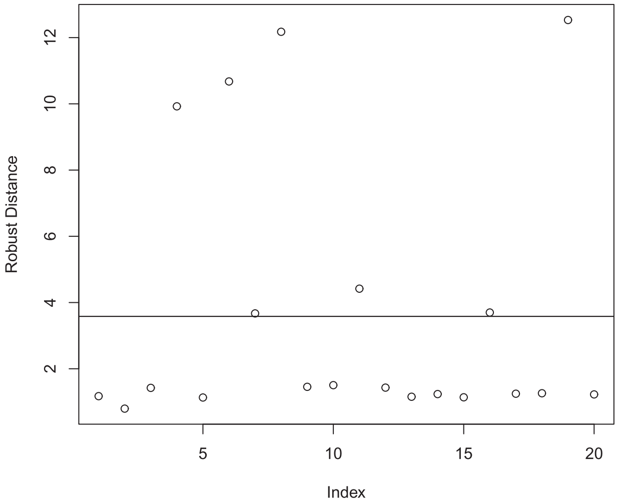

Figure 1. Robust distances of the modified wood gravity data.

Table 4. Results for two real dataset.

data1	Method		data2	Method	
Variable	PWLAD-LASSO	PM	Variable	PWLAD-LASSO	PM
x_1	0.211	0.227	Cement	0.093	0.083
x_2	0	0	Slag	0.026	0.008
x_3	0	0	Fly ash	0.089	0.084
x_4	0	0.377	Water	−0.016	−0.001
x_5	0.448	0.200	SP	0	0.146
–	–	–	Coarse Aggr	0	0
–	–	–	Fine Aggr	0.022	0.004
MAE	0.022(0.0010)	0.031(0.0009)	MAE	2.310(0.0047)	2.864(0.0043)

distances in Figure 1 shows a strongly deviating group of outliers, index 4, 6, 8 and 19, where the horizontal line is at the usual cutoff value $\sqrt{\chi^2_{5,0.975}} = 3.58$. Meanwhile, these outliers were also detected using the PWLAD-LASSO method.

The second data set is the concrete slump test data [44]. The data set of 78 records, each containing seven components of the input vector and one output value, 28-day Compressive Strength (Mpa). To evaluate the predictive performance of the PWLAD-LASSO and the the PM method, we apply a 4-fold cross-validation to calculate the predictive mean absolute error MAE $= \frac{1}{|S|}\sum_{i\in S}|y_i - \mathbf{x}_i'\boldsymbol{\beta}|$, where S denotes the index of the test data set. From Table 4, we can find that our proposed method has smaller MAE than PM method.

5. Discussion

In this article, we have presented the penalized weighted LAD-LASSO method. The proposed PWLAD-LASSO method combines the ideas of PWLAD regression method and LASSO method to perform simultaneous outlier detection, robust variable selection, even when the random error is both heterogenous and heavy tailed. The merits of our methodology were illustrated via simulation studies and application. According to numerical simulations, our proposed method performs better than PM method when there are outliers in the response or explanatory variables. By analyzing the modified wood gravity data data and the concrete slump test data, our proposed method had better the ability of detecting outliers and predictive performance. Furthermore, we will investigate the large sample properties of the proposed PWLAD-LASSO method as future work. In fact, it is very important to select the proper weights. Therefore, it warrants further effort to investigate the choice of weights.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Jiang's research is partially supported by the Natural Science Foundation of Guangdong and the National Science Foundation of Guangdong (No. 2018A030313171, No. 2019A1515011830).

References

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

[2] O. Arslan, *Weighted lad-lasso method for robust parameter estimation and variable selection in regression*, Comput. Stat. Data Anal. 56 (2012), pp. 1952–1965.

[3] N. Billor, A.S. Hadi and P.F. Velleman, *Bacon: Blocked adaptive computationally efficient outlier nominators*, Comput. Stat. Data Anal. 34 (2000), pp. 279–298.

[4] K. Boudt, P.J. Rousseeuw, S. Vanduffel, and T. Verdonck, *The minimum regularized covariance determinant estimator*, Stat. Comput. 30 (2020), pp. 113–128.

[5] J. Bradic, J. Fan, and W. Wang, *Penalized composite quasi-likelihood for ultrahigh dimensional variable selection*, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (2011), pp. 325–349.

[6] S. Chatterjee and A.S. Hadi, *Sensitivity Analysis in Linear Regression*, 327, John Wiley & Sons, New York, 2009.

[7] P. Čížek, *Least trimmed squares in nonlinear regression under dependence*, J. Stat. Plan. Inference 136 (2006), pp. 3967–3988.

[8] N.R. Draper and H. Smith, *Applied Regression Analysis*, Vol. 1, John Wiley & Sons, New York, 1966.

[9] J. Fan, Y. Fan, and E. Barut, *Adaptive robust variable selection*, Ann. Stat. 42 (2014), pp. 324.

[10] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc. 96 (2001), pp. 1348–1360.

[11] P. Filzmoser and K. Hron, *Outlier detection for compositional data using robust methods*, Math. Geosci. 40 (2008), pp. 233–248.

[12] P. Filzmoser, R. Maronna, and M. Werner, *Outlier identification in high dimensions*, Comput. Stat. Data Anal. 52 (2008), pp. 1694–1711.

[13] X. Gao and Y. Feng, *Penalized weighted least absolute deviation regression*, Stat. Interface 11 (2018), pp. 79–89.

[14] D. Gervini and V.J. Yohai, *A class of robust and fully efficient regression estimators*, Ann. Stat. 30 (2002), pp. 583–616.

[15] A. Giloni, J.S. Simonoff, and B. Sengupta, *Robust weighted lad regression*, Comput. Stat. Data Anal. 50 (2006), pp. 3124–3140.

[16] Ö. Gürünlü Alma, S. Kurt, and A. Uğur, *Genetic algorithms for outlier detection in multiple regression with different information criteria*, J. Stat. Comput. Simul. 81 (2011), pp. 29–47.

[17] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Vol. 196, John Wiley & Sons, New York, 2011.

[18] D.R. Hunter and K. Lange, *Quantile regression via an mm algorithm*, J. Comput. Graph. Stat. 9 (2000), pp. 60–77.

[19] B. Iglewicz and J. Martinez, *Outlier detection using robust measures of scale*, J. Stat. Comput. Simul. 15 (1982), pp. 285–293.

[20] A.R. Imon and A.S. Hadi, *Identification of multiple high leverage points in logistic regression*, J. Appl. Stat. 40 (2013), pp. 2601–2616.

[21] Y. Jiang, *Robust estimation in partially linear regression models*, J. Appl. Stat. 42 (2015), pp. 2497–2508.

[22] Y. Jiang, *An exponential-squared estimator in the autoregressive model with heavy-tailed errors*, Stat. Interface 9 (2016), pp. 233–238.

[23] Y. Jiang, Q. Ji, and B. Xie, *Robust estimation for the varying coefficient partially nonlinear models*, J. Comput. Appl. Math. 326 (2017), pp. 31–43.

[24] Y. Jiang, G.L. Tian, and Y. Fei, *A robust and efficient estimation method for partially nonlinear models via a new mm algorithm*, Stat. Papers 60 (2019), pp. 2063–2085.

[25] J.M. Jobe and M. Pokojovy, *A cluster-based outlier detection scheme for multivariate data*, J. Amer. Statist. Assoc. 110 (2015), pp. 1543–1551.

[26] B. Kai, R. Li, and H. Zou, *New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models*, Ann. Stat. 39 (2011), pp. 305.

[27] R.J. Karunamuni, L. Kong, and W. Tu, *Efficient robust doubly adaptive regularized regression with applications*, Stat. Methods Med. Res. 28 (2019), pp. 2210–2226.

[28] D. Kong, H. Bondell, and Y. Wu, *Fully efficient robust estimation, outlier detection and variable selection via penalized regression*, Stat. Sin. 28 (2018), pp. 1031–1052.

[29] S. Lambert-Lacroix and L. Zwald, *Robust regression through the hubers criterion and adaptive lasso penalty*, Electron. J. Stat. 5 (2011), pp. 1015–1053.

[30] G. Li, H. Peng, and L. Zhu, *Nonconcave penalized m-estimation with a diverging number of parameters*, Stat. Sin. 21 (2011), pp. 391–419.

[31] A.M. Millar and D.C. Hamilton, *Modern outlier detection methods and thier effect on subsequent inference*, J. Stat. Comput. Simul. 64 (1999), pp. 125–150.

[32] A. Nurunnabi, A.S. Hadi, and A. Imon, *Procedures for the identification of multiple influential observations in linear regression*, J. Appl. Stat. 41 (2014), pp. 1315–1331.

[33] M. Riani, A.C. Atkinson, and A. Cerioli, *Finding an unknown number of multivariate outliers*, J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2009), pp. 447–466.

[34] P.J. Rousseeuw, *Least median of squares regression*, J. Amer. Statist. Assoc. 79 (1984), pp. 871–880.

[35] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, Vol. 1, Wiley Online Library, New York, 1987.

[36] Y. She and A.B. Owen, *Outlier detection using nonconvex penalized regression*, J. Amer. Statist. Assoc. 106 (2011), pp. 626–639.

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

[38] L. Wang and R. Li, *Weighted Wilcoxon-type smoothly clipped absolute deviation method*, Biometrics 65 (2009), pp. 564–571.

[39] X. Wang, Y. Jiang, M. Huang, and H. Zhang, *Robust variable selection with exponential squared loss*, J. Amer. Statist. Assoc. 108 (2013), pp. 632–643.

[40] H. Wang, G. Li, and G. Jiang, *Robust regression shrinkage and consistent variable selection through the lad-lasso*, J. Bus. Econ. Stat. 25 (2007), pp. 347–355.

[41] L. Wang, Y. Wu, and R. Li, *Quantile regression for analyzing heterogeneity in ultra-high dimension*, J. Amer. Statist. Assoc. 107 (2012), pp. 214–222.

[42] Y. Wu and Y. Liu, *Variable selection in quantile regression*, Statist. Sinica 19 (2009), pp. 801.

[43] F. Xue and A. Qu, *Variable selection for highly correlated predictors*, preprint (2017). Available at arXiv:1709.04840.

[44] I.C. Yeh, *Modeling slump flow of concrete using second-order regressions and artificial neural networks*, Cem. Concr. Compos. 29 (2007), pp. 474–480.

[45] V.J. Yohai, *High breakdown-point and high efficiency robust estimates for regression*, Ann. Stat. 15 (1987), pp. 642–656.

[46] V.J. Yohai and R.H. Zamar, *High breakdown-point estimates of regression by means of the minimization of an efficient scale*, J. Amer. Statist. Assoc. 83 (1988), pp. 406–413.

[47] H. Zou, *The adaptive lasso and its oracle properties*, J. Amer. Statist. Assoc. 101 (2006), pp. 1418–1429.

[48] H. Zou and R. Li, *One-step sparse estimates in nonconcave penalized likelihood models*, Ann. Stat. 36 (2008), pp. 1509–1533.

[49] H. Zou and M. Yuan, *Composite quantile regression and the oracle model selection theory*, Ann. Stat. 36 (2008), pp. 1108–1126.