Fully efficient robust estimation, outlier detection and variable selection via penalized regression

Dehan Kong[1], Howard Bondell[2] and Yichao Wu[2]

[1] *University of Toronto and* [2] *North Carolina State University*

Abstract: This paper studies the outlier detection and variable selection problem in linear regression. A mean shift parameter is added to the linear model to reflect the effect of outliers, where an outlier has a nonzero shift parameter. We then apply an adaptive regularization to these shift parameters to shrink most of them to zero. Those observations with nonzero mean shift parameter estimates are regarded as outliers. An L1 penalty is added to the regression parameters to select important predictors. We propose an efficient algorithm to solve this jointly penalized optimization problem and use the extended Bayesian information criteria tuning method to select the regularization parameters, since the number of parameters exceeds the sample size. Theoretical results are provided in terms of high breakdown point, full efficiency, as well as outlier detection consistency. We illustrate our method with simulations and data. Our method is extended to high-dimensional problems with dimension much larger than the sample size.

Key words and phrases: Adaptive, Breakdown Point, least trimmed squares, Outliers, Penalized regression, Robust Regression, Variable Selection

1 Introduction

Occuring frequently in data collection, outliers are observations that deviate markedly from the rest. In the presence of outliers, likelihood-based inference can be unreliable, for instance, ordinary least squares regression is very sensitive to outliers. To this end, robust estimation and outlier detection are critical in statistical learning. We consider the mean shift linear regression model $y_i = \alpha + X_i\beta + \gamma_i + \epsilon_i$, where X_i is a p dimensional predictor, β is a p dimensional parameter, and γ_i is an observation-specified mean shift parameter that is nonzero when the corresponding observation is an outlier. This model was previously used by Gannaz (2006); McCann and Welsch (2007); She and Owen (2011), and represents the general notion that the response can be arbitrary due to an outlier.

In this article, we are interested in variable selection as well as robust coefficient estimation together with the task of outlier detection based on this mean shift model. Popular methods for variable selection are penalized regression methods such as LASSO (Tibshirani (1996)), Smoothly Clipped Absolute Deviation Penalty (Fan and Li (2001)) and adaptive LASSO (Zou (2006)). These penalized regression methods can be used not only for variable selection but for outlier detection as

well. For example, McCann and Welsch (2007) used an L1 regression while She and Owen (2011) imposed a nonconvex penalty function on γ_i's to avoid the trivial estimate $\hat{\gamma}_i = y_i$ and $\hat{\beta} = 0$, and achieved a sparse solution in terms of the shift parameter. If the estimate of γ_i was nonzero, the ith observation was identified as an outlier.

Our method is based on this mean shift model, but we use an adaptive penalty which depends on the residuals from some robust initial fit. Meanwhile, we add an L1 penalty to the regression coefficients to achieve variable selection simultaneously. Our work differs from the work of McCann and Welsch (2007) and She and Owen (2011). By judicious choice of penalty function, we can attain high breakdown. Our method enjoys a breakdown point of 1/2, while the breakdown point of She and Owen (2011) is at most $1/(p+1)$ and that of McCann and Welsch (2007) is $1/n$. As shown in our simulation studies, when the proportion of the outliers is large or the outliers are more extreme, the estimates in McCann and Welsch (2007) and She and Owen (2011) break down and cannot detect the outliers correctly, while our methods still perform well. We fully develop the theoretical properties of our approach in contrast to McCann and Welsch (2007) and She and Owen (2011).

In the literature, the asymptotic efficiency and the breakdown point are two criteria to evaluate a robust regression technique. They represent the typical trade-off in efficiency for robustness. It is ideal to achieve full asymptotic efficiency of the true model compared to ordinary least squares while maintaining a high breakdown point of 1/2. Typical robust regression methods do not manage this. Ordinary least squares, which is fully efficient under normality, has a breakdown point of $1/n$, and hence even a single outlier can render the estimate arbitrarily bad. The M-estimates (Huber (1981)) also have a breakdown point of $1/n$ while Generalized M-estimates (Mallows (1975)) can have a breakdown point of only $1/(p+1)$ (Maronna et al. (1979); Donoho and Huber (1983)). While neither of them enjoys full efficiency. There are several methods which enjoy a high breakdown point of 1/2, such as the least median of squares estimates (Hampel (1975); Rousseeuw (1984)), the least trimmed squares estimates (Rousseeuw (1984)), S-estimates (Rousseeuw and Yohai (1984)), MM-estimates (Yohai (1987)) and the Schweppe one-step Generalized M-estimates (Coakley and Hettmansperger (1993)). These methods are not fully efficient. There have been some methods introduced achieving both properties, for example the robust and efficient weighted least squares estimators (Gervini and Yohai (2002)) and the generalized empirical likelihood method (Bondell and Stefanski (2013)).

The proposed method achieves both full efficiency and high breakdown, while also performing variable selection simultaneously. Specifically, our method is robust to outliers and enjoys a breakdown point that can be as high as 1/2. When there are no outliers, our estimator can enjoy full asymptotic efficiency compared to the LASSO estimator. We define outlier detection consistency in the robust regression context, and show that our method correctly detects outliers with probability tending to 1. In particular, we assume the number of outliers is of constant proportion of the sample size. In the context of the mean shift model, this corresponds to the case when the number of nonzero components in γ_i's is of the same order as the sample size. In addition to these properties,

we propose an efficient algorithm for our method when the total number of unknown parameters, $n + p$, is larger than the sample size. The extended Bayesian information criteria (EBIC) (Chen and Chen (2008, 2012)) is adopted to select the tuning parameters that control outlier detection and variable selection. Our method can be extended to the high-dimensional setting when the dimension of the covariate p is diverging at the exponential rate of the sample size, without loss of its properties.

The rest of this paper is organized as follows. In Section 2, we introduce our robust regression method and its implementation. In Section 3, we include the theoretical results for our method, including fully efficiency, high breakdown, outlier detection consistency, and the equivariance property of our estimator. Numerical simulations are provided to evaluate the proposed method in Section 4. In Section 5, we apply our method to the Boston Housing dataset. We extend our method to the high-dimensional setting in Section 6, including the theoretical properties in the diverging p case. The proofs and technical details are in the supplementary materials.

2 Methodology

Let $y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $X = (X_1^{\mathrm{T}}, \ldots, X_n^{\mathrm{T}})^{\mathrm{T}}$ be an $n \times p$ design matrix, $\gamma = (\gamma_1, \ldots, \gamma_n)^{\mathrm{T}}$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$. Our model can be written as

$$y \;=\; \alpha 1 + X\beta + \gamma + \epsilon,$$

where α is the intercept, 1 is a $n \times 1$ vector of ones, and the error term ϵ_i's are independent and identically distributed with $E(\epsilon_i) = 0$. The mean shift parameters γ_i's serve as indicators of the outliers in the regression of $y_i \mid X_i$. If the ith subject is an outlier, $\gamma_i \neq 0$. Another type of outlier may still occur in the covariate space, i.e. high leverage points, while having $\gamma_i = 0$, but these leverage points do not result in the breakdown of the estimator. We are interested in both outlier detection and variable selection for this model. To achieve these goals, it is natural to devise a selection method via shrinkage. We impose penalties on the γ_i's to encourage them to shrink to zero and identify observations with nonzero γ_i's as outliers. Meanwhile, we add penalties on the coefficient β to achieve variable selection. Specifically, we solve the minimization problem

$$\min_{\alpha,\beta,\gamma} Q_n(\alpha, \beta, \gamma) = \min_{\alpha,\beta,\gamma} ||y - \alpha 1 - X\beta - \gamma||_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j| + \mu_n \sum_{i=1}^{n} |\gamma_i|/|\widetilde{\gamma}_i|, \tag{1.1}$$

where $\widetilde{\gamma}_i$'s are residuals of an initial robust regression fit. Here $\widetilde{\gamma}_i$ is the weight we put on penalty, which plays a similar role as that of the weight put on the adaptive Lasso. For those outliers, we expect $\widetilde{\gamma}_i$ to be larger, and we shrink less for the mean shift parameters, while for the "good points", $\widetilde{\gamma}_i$ is smaller, and we would shrink more for the mean shift parameters. Here λ_n and μ_n

are regularization parameters controlling the variable selection and outlier detection, respectively. If we do not want to perform variable selection, we can set $\lambda_n = 0$, a special case of our method.

In Sections 2-5, we focus on the case that p is a fixed constant and $p < n$. We discuss how to extend our method to the high-dimensional case when p is much larger than n and diverges at an exponential rate of n in Section 6.

2.1 Robust Initial Estimator

We impose an adaptive penalty on γ that relies on the weights depending on an initial robust fit. The weight plays a similar role as the weight used in the adaptive LASSO problem (Zou (2006)), but it is based on the residuals rather than the parameter estimates. The incorporation of an adaptive lasso type of penalty function on the mean shift parameters for detecting outliers is what yields the breakdown theory results, while the use of the residuals from an initial high breakdown fit provides the high breakdown point to the estimator. Any methods can be used for this initial step, for example least trimmed squares, S-estimates, and MM-estimators, among others. The breakdown point of our new method is no less than the breakdown point of the robust method we use for the initial fit.

In this article, we use the least trimmed squares method to obtain the initial robust estimates. We show that the least trimmed squares initial fit carries over the high breakdown point of our estimator. Meanwhile, full efficiency compared to the LASSO estimator, outlier detection, and variable selection consistency can be achieved by using this initial estimator.

Let $r_i^2 = (y_i - \alpha - X_i\beta)^2$. The least trimmed squares method solves $\min_\beta \sum_{i=1}^h r_{(i)}^2$, where $r_{(i)}^2$'s are the order statistics of r_i^2, $r_{(1)}^2 \leq r_{(2)}^2 \leq \ldots \leq r_{(n)}^2$. The number of included residuals, h, is chosen to determine the breakdown point of the estimator. In particular, the breakdown point can be shown to be $(n-h+1)/n$. In our simulation study and in data applications, we use the truncation number $h = \lfloor 3n/4 \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less than x, although this does not yield the maximum breakdown point. We have also tried other truncation numbers; the results were similar and hence are not shown. For implementation, the R function "ltsReg" is adopted to obtain the initial estimates $\widetilde{\beta}_f = (\widetilde{\alpha}, \widetilde{\beta}^{\mathrm{T}})^{\mathrm{T}}$. In particular, we implement the fast minimum determinant estimator algorithm (Rousseeuw and Driessen (1999)), which is computationally quick. For details of the algorithm, we refer readers to Section 4 of Rousseeuw and Driessen (1999). After we get the initial least trimmed squares estimates $\widetilde{\beta}_f$, and the initial residuals are $\widetilde{\gamma}_i = y_i - \widetilde{\alpha} - X_i\widetilde{\beta}$.

2.2 Algorithm

The optimization problem in (1.1) is an L1 penalized least squares and can be easily transformed to a quadratic programming problem. A more efficient way uses the Least Angle Regression algorithm (Efron, Hastie, Johnstone, and Tibshirani (2004)). If $\rho_n = \mu_n/\lambda_n$, then the optimization problem

in (1.1) is

$$\min_{\alpha,\beta,\gamma} Q_n(\alpha,\beta,\gamma) = \min_{\alpha,\beta,\gamma} ||y-\alpha 1 - X\beta - \gamma||_2^2 + \lambda_n\{\sum_{j=1}^{p}|\beta_j| + \rho_n\sum_{i=1}^{n}|\gamma_i|/|\widetilde{\gamma}_i|\}.$$

For a fixed ρ_n, we can reparametrize to $\gamma_i^* = \rho_n\gamma_i/|\widetilde{\gamma}_i|$ so that the problem is

$$\min_{\alpha,\beta,\gamma} Q_n(\alpha,\beta,\gamma) = \min_{\alpha,\beta,\gamma} ||y-\alpha 1 - X\beta - B\gamma^*||_2^2 + \lambda_n\{\sum_{j=1}^{p}|\beta_j| + \sum_{i=1}^{n}|\gamma_i^*|\}, \tag{1.2}$$

with $B = \mathrm{diag}(|\widetilde{\gamma}_1|/\rho_n, \ldots, |\widetilde{\gamma}_n|/\rho_n)$ and $\gamma^* = (\gamma_1^*, \ldots, \gamma_n^*)^{\mathrm{T}}$. Problem (1.2) is a typical LASSO problem, and can be solved easily by the R package "lars", providing the whole solution path of (2) as a function of λ_n.

2.3 Tuning parameter selection

The minimization (1.1) involves tuning λ_n and μ_n, which is equivalent to tuning λ_n and ρ_n together. Since the number of parameters is $n+p$ and larger than the sample size, we use the EBIC (Chen and Chen (2012)) due to its selection consistency properties for high-dimensional problems. Suppose $\hat{\beta}$ and $\hat{\gamma}$ are the estimates when the tuning parameters are set as λ_n and ρ_n. Let $e_i^2 = (y_i - \hat{\alpha} - X_i^*\hat{\beta} - \hat{\gamma}_i)^2$, and define the residual sum of squares as $\mathrm{RSS} = \sum_{i=1}^{n} e_i^2$. The EBIC is defined as

$$\mathrm{EBIC} = n\log(\mathrm{RSS}/n) + k\{\log n + c\log(n+p)\},$$

where k is the degree of freedom, the number of nonzero components of $(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$, and c is a constant that must be specified. In our case, we have $p+n$ parameters with order $O(n)$. By Theorem 1 of Chen and Chen (2012), when $c > 1$, the EBIC can select the tuning parameter consistently if the number of parameters is on the order of n. Toward this end, we set $c = 1+\varepsilon$ with ε being a small positive number to meet the requirement of their theoretical results. Based on our preliminary numerical experience, we have found that the results are not sensitive to the choice of small ε. Consequently, we set $c = 1.01$ for convenience. We set two-dimensional grids for ρ_n and λ_n to find the combination that minimizes the EBIC. Specifically, we first choose a dense grid on ρ_n and, for each ρ_n, we use Least Angle Regression algorithm to obtain the solution paths of the problem in (1.2). We pick the grid of λ_n on each point that the degree of freedom changes. For high-dimensional problems with the number of parameters exceeding the sample size, we get a perfect fit if the degree of freedom is large enough, which makes the EBIC small as the residual sum of squares goes to zero. This gives the wrong selection of λ_n, because it tends to select the λ_n that gives a perfect fit. Consequently, we search over the λ_n that lead to $k \le \lfloor 0.5n \rfloor$ as we assume that the number of outliers is less than half of the sample size.

3 Theoretical results

In this section, we investigate some asymptotic results including outlier detection consistency and variable selection consistency, in the first and third subsection respectively, and we consider the high breakdown point in the second subsection. As far as we know, this is the first time that outlier detection consistency has been formulated in the statistical literature.

Without loss of generality, we only show the results for the case that there is no intercept. The results, as well as the proofs, for the case with an intercept follow in a similar manner.

3.1 Asymptotic theory when there are no outliers

We discuss our main results for outlier detection consistency and variable selection consistency when no outlier exists. Then outlier detection consistency reveals that the resulting estimator is asymptotically equivalent to the simple L1-penalized regression, and thus shares its asymptotic efficiency properties. Without loss of generality, we assume that the first q components of β_0 are nonzero, denoted by $\beta_0(1)$, and the remaining $p-q$ components are zero, denoted by $\beta_0(2)=0$. Let $\psi = n^{-1/2}\gamma$, and take $\theta = (\beta(1)^{\mathrm{T}}, \beta(2)^{\mathrm{T}}, \psi^{\mathrm{T}})^{\mathrm{T}} = (\theta(1)^{\mathrm{T}}, \theta(2)^{\mathrm{T}}, \theta(3)^{\mathrm{T}})^{\mathrm{T}} = (\theta_1, \ldots, \theta_{p+n})^{\mathrm{T}}$ with $\theta(1) = \beta(1)$, $\theta(2) = \beta(2)$ and $\theta(3) = \psi$. Let $X_{a,b}$ be the submatrix consisting of the ath to bth column of the matrix X. We take

$$A = \begin{pmatrix} X_{1,q} & X_{q+1,p} & n^{1/2} I_n, \end{pmatrix}$$

$$C = n^{-1} A^{\mathrm{T}} A = \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33}, \end{pmatrix}$$

with $C_{11} = n^{-1} X_{1,q}^{\mathrm{T}} X_{1,q}$, $C_{21} = n^{-1} X_{q+1,p}^{\mathrm{T}} X_{1,q}$ and $C_{31} = n^{-1/2} X_{1,q}$.

Our method is equivalent to solving

$$\min_{\theta} ||y - A\theta||_2^2/2 + \lambda_n \sum_{j=1}^{p} |\theta_j| + n^{1/2} \mu_n \sum_{j=p+1}^{p+n} |\theta_j| / |\widetilde{\gamma}_{j-p}|.$$

Suppose a_1 and a_2 are two column vectors with same dimension and write $a_1 \leq a_2$ if the inequality holds elementwise. We let $|a_1|$ be the vector with same dimension as a_1, where each element of $|a_1|$ is the absolute value of the corresponding element in a_1.

We need some conditions; they may not be the weakest conditions but help to simplify the proof.

(A) The error ϵ_i's are independent and identically distributed with $E(\epsilon_i^{2k}) < \infty$ for some positive integer k.

(B1) There exists a constant vector C such that $|C_{21}C_{11}^{-1}\text{sign}(\theta_0(1))| \leq 1-C$, , where the 1 on the righthand side of the inequality is a vector of ones.
There exists $0 < d \leq 1$ and $M_1, M_2, M_3 > 0$ so that
(B2) $n^{-1}X_{,j}^{\mathrm{T}}X_{,j} \leq M_1$ for any $1 \leq j \leq p$, where $X_{,j}$ denotes the jth column of X,
(B3) $\alpha^{\mathrm{T}}C_{11}\alpha^{\mathrm{T}} \geq M_2$ for any $||\alpha|| = 1$,
(B4) $n^{(1-d)/2}\min_{j=1,\ldots,q}|\beta_{j0}| \geq M_3$ for some $0 < d \leq 1$.
Conditions (B1)-(B4) are related to n, and we require Conditions (B1)-(B4) to hold for all sufficient large n. Condition (B1) was introduced by Zhao and Yu (2006) to guarantee the selection consistency for LASSO. Condition (B2) can be achieved by normalizing the covariates. Condition (B3) is trivial and only requires the smallest eigenvalue of the matrix C_{11} be nonzero. Condition (B4) quantifies the smallest signal of the coefficient β_{j0}, and we could identify the signal on the order of $O(n^{(d-1)/2})$ for some $0 < d \leq 1$. In particular, when $d = 1$, β_0 can be some fixed value that does not depend on n. In other cases, we allow the magnitude of β_0 to decay as the sample size increases.

Define $a_1 =_s a_2$ if the signs of these two vectors a_1 and a_2 are the same elementwise.

Theorem 1 *Under conditions (A) and (B1)-(B4), for* $\lambda_n = o(n^{(d+1)/2})$ *and* $n^{-1/2}\lambda_n \to \infty$, *and* $\mu_n n^{-1/(2k)-d/2} \to \infty$, *we have* $pr(\hat{\theta} =_s \theta_0) \to 1$ *as* $n \to \infty$.

Thus we can select the important predictors consistently when no outliers exist.

Remark: If a robust estimator can have the same efficiency compared with a non-robust procedure when no outliers exist, we call the robust estimator fully efficient compared to the non-robust procedure. Since $pr(\hat{\gamma} = 0) \to 1$ from Theorem 1, our method is equivalent to the LASSO problem, and thus our estimator is fully efficient compared to the LASSO estimator. If we do not impose any penalty on β, $\lambda_n = 0$, our estimator is fully efficient compared to the ordinary least squares.

3.2 High breakdown point

Let the $n \times (p+1)$ matrix $Z = (X, y)$ denote the sample, and $\widetilde{Z}_m$ denote the contaminated sample by replacing m data points by arbitrary values. The finite sample breakdown point for the regression $\hat{\beta}$ is defined as

$$BP(\hat{\beta}, Z) = \min\{m/n : \sup_{\widetilde{Z}_m} ||\hat{\beta}(\widetilde{Z}_m)||_2 = \infty\},$$

where $\hat{\beta}(\widetilde{Z}_m)$ denotes the estimate of the regression parameter using the contaminated sample $\widetilde{Z}_m$.

We assume the general position condition that is typical in high breakdown point proofs. Suppose G is the set containing all good points (X_i, y_i); for any $p \times 1$ vector $v \neq 0$, $\{(X_i, y_i) : (X_i, y_i) \in G, \text{ and } X_i v = 0\}$ contains at most $p-1$ points.

In contrast, our method would have a breakdown point of at least $(n-h+1)/n$, which is shown by the following theorem.

Theorem 2 *If we use the least trimmed squares with truncation number h as the initial estimator, then under the general position condition the breakdown point of our estimator satisfies that $BP(\hat{\beta}, Z) \geq \min\{(n-h+1)/n, \lfloor (n-p)/2 \rfloor / n\}$.*

Remark: The least trimmed squares with truncation number h has a breakdown point of $\min\{(n-h+1)/n, \lfloor (n-p)/2 \rfloor / n\}$, see Rousseeuw and Leroy (1987) for example. This theorem provides a lower bound of the breakdown point of the proposed method, which performs at least as well as the least trimmed squares initial estimator in terms of high breakdown point. Typically, we choose $h < n/2$ so that the breakdown point cannot exceed $1/2$ since we aim for model to fit the majority of the data.

3.3 Outlier detection consistency

In this subsection, we consider the case when there are outliers in the conditional distribution of $y \mid X$, and show that we can identify these outliers consistently. We assume that the fraction of outliers in the data remains nonzero as more data are collected, otherwise we are in the trivial case. Denote s_n as the number of outliers, and assume $s_n = O(n)$ and $s_n < n/2$. We take $\psi = n^{-1/2}\gamma$. Without loss of generality, we assume that the first s_n components $\psi_0(1)$ are outliers while the remaining $n - s_n$ components $\psi_0(2) = 0$ correspond to the normal data points. We take $\eta = (\psi(1)^{\mathrm{T}}, \beta^{\mathrm{T}}, \psi(2)^{\mathrm{T}})^{\mathrm{T}} = (\eta(1)^{\mathrm{T}}, \eta(2)^{\mathrm{T}}, \eta(3)^{\mathrm{T}})^{\mathrm{T}}$. With $X_{a:b}$ as the ath to bth row of the matrix X and $X_{a:b,c:d}$ as the sub matrix of X with ath to bth row and cth to dth column, the design matrix is

$$B = \begin{pmatrix} B_1 & B_2 & B_3 \end{pmatrix},$$

where $B_1 = (n^{1/2} I_{s_n}, 0_{s_n \times (n-s_n)})^{\mathrm{T}}$, $B_2 = X$ and $B_3 = (0_{(n-s_n)\times s_n}, n^{1/2} I_{n-s_n})^{\mathrm{T}}$. Let

$$D = n^{-1} B^{\mathrm{T}} B = \begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{pmatrix}$$

with $D_{11} = I_{s_n}$, $D_{21} = n^{-1/2} X_{1:s_n}^{\mathrm{T}}$, $D_{31} = 0$, $D_{22} = n^{-1} X^{\mathrm{T}} X$.

The estimator is the solution to

$$\min_{\eta} ||y - B\eta||_2^2/2 + n^{1/2}\mu_n \sum_{j=1}^{s_n} |\eta_j|/|\widetilde{\gamma}_j| + \lambda_n \sum_{j=s_n+1}^{s_n+p} |\eta_j|$$
$$+ n^{1/2}\mu_n \sum_{j=p+s_n+1}^{p+n} |\eta_j|/|\widetilde{\gamma}_{j-p}|.$$

As $s_n = O(n)$, our problem is a weighted L1 regression with the number of nonzero components on

the order of $O(n)$. It is different from the traditional high-dimensional sparse regression problem, for example Zhao and Yu (2006); they deal with the case when the number of nonzero components is on the order of $O(n^a)$ with some $a < 1$.

Let $\pi_n = \min_{i=1,\ldots,s_n} |\gamma_{i0}|$. We need some further conditions:

(C1) $\pi_n n^{-1/(2k)} \to \infty$ as $n \to \infty$.
(C2) The number of outliers $s_n \leq n - h$, and $s_n = O(n)$.
(C3) There exists a constant vector C such that $|D_{21}\text{sign}(\eta_0(1))| \leq 1 - C$, where 1 on the righthand side of the inequality is a vector of ones.

Condition (C1) requires that the minimum signal of the outliers diverges with the sample size. To see this is needed, consider that $y_i = \gamma_i + \epsilon_i$, where $\epsilon_i \sim N(0,1)$, $\gamma_i = d^* > 0$ for $1 \leq i \leq s_n$ and $\gamma_i = 0$ for $s_n + 1 \leq i \leq n$. Then $\pi_n = d^*$. As the support of the distribution is the real line, there cannot be a fixed d^* that defines an "outlier" in this distribution as n grows. Hence it must be assumed that π_n diverges sufficiently fast in order to distinguish the "outlier" from a random variable coming from the true distribution. (C2) requires that the number of data points used in least trimmed squares cannot be smaller than the number of the good points in the data, guaranteeing a robust initial estimate. Condition (C3) is parallel to Condition (B1) in Section 3.1 when no outliers exist.

Theorem 3 *Under conditions (A), (B2), (C1)-(C3), for* $\mu_n = o(\pi_n^2)$, $\mu_n^k n^{-1} \to \infty$ *and* $\lambda_n n^{-1} \mu_n^{-1} \pi_n \to \infty$, *we have* $pr(\hat{\gamma} =_s \gamma_0) \to 1$ *as* $n \to \infty$.

3.4 Equivariance properties of our estimator

In this subsection, we discuss the equivariance properties of our estimator. There are three types of equivariance properties: regression, scale, and affine equivariance; see Rousseeuw and Leroy (1987) for example.

An estimator T is called regression equivariant if

$$T(\{X_i, y_i + X_i v), i = 1, \ldots n\}) = T(\{X_i, y_i), i = 1, \ldots n\}) + v$$

where v is any column vector.

An estimator T is called scale equivariant if

$$T(\{X_i, cy_i), i = 1, \ldots n\}) = cT(\{X_i, y_i), i = 1, \ldots n\})$$

for any constant c.

An estimator T is called affine equivariant if

$$T(\{X_i A, y_i), i = 1, \ldots n\}) = A^{-1} T(\{X_i, y_i), i = 1, \ldots n\})$$

for any nonsingular square matrix A.

We focus on the equivariance of our estimator from (1.1) in terms of the parameter β. We consider the two cases: λ_n is nonzero, and zero. As our method depends on the residuals from the initial estimator, the equivariance properties of our estimators depend on the equivariance properties of the initial estimators.

If no penalty is imposed on β, $\lambda_n = 0$, then our estimator inherits the equivariance properties of the initial estimator. In particular, for the least trimmed squares estimator we use, it has been shown that it is regression, scale, and affine equivariant (Rousseeuw and Leroy (1987)). Thus, our estimator is regression, scale, and affine equivariant.

If λ_n is nonzero, to obtain equivariance properties for the estimator, we center and scale the X_i's and y_i's at the beginning of our two-step procedure, typical in penalized regularization problems, then scale the estimate of β back. In this case, our procedure enjoys regression equivariance, scale equivariance, and partial affine equivariance. In particular, this equivariance is with respect to scale change transformations only, not general affine transformations. This is the situation with all penalized regressions as other affine transformations no longer preserve the desired coordinate system for variable selection.

As mentioned in Section 2.1, we can use such other initial estimators as least median of squares, S-estimator, and MM-estimator as long as they have a high breakdown point. For least median of squares and S-estimator, they are regression, scale and affine equivariant (Rousseeuw and Leroy (1987)). For the MM-estimator, it is scale equivariant. As MM-estimator depends on an initial high breakdown estimator, if the initial estimator is regression and/or affine equivariant, the resulting estimator is as well (Yohai (1987)). Consequently, when $\lambda_n = 0$, our estimator inherits the equivariance properties of the initial estimator. When λ_n is nonzero, if we use the standardization procedure, our estimator still inherits the scale and regression equivariance properties of the initial estimator. If the initial estimator is equivariant with respect to scale change transformations, our estimator is as well.

4 Simulation Studies

In this section, we report on our method using simulation examples. The covariate X_i was generated independently and identically from a multivariate normal distribution with zero mean and covariance matrix Σ, with the jkth element of the matrix $\Sigma_{jk} = 0.5^{|j-k|}$. The true coefficient was set as $\beta_0 = (4, 2, 1, 0.5, 0.2, 0, \ldots, 0)^{\mathrm{T}}$ with $q = 5$ nonzero components and the remaining $(p - q)$ elements zero. The random error was simulated independently from $\epsilon_i \sim N(0, 0.25)$. The data were generated from $y_i = X_i\beta_0 + \epsilon_i$ for $1 \leq i \leq n$. We contaminated the first cn observations by setting $X_i^* = X_i + L$ and $y_i^* = y_i + V$ 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.

We investigated the numerical performance of our method by using the following 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 nonzero coefficients that are estimated as zero)

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

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

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

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

$$\mathrm{MSE} \quad = \quad (\hat{\beta}_f - \beta_{f,0})^{\mathrm{T}} E(X_f^{\mathrm{T}} X_f)(\hat{\beta}_f - \beta_{f,0}),$$

where α is the estimated intercept, $\beta_f = (\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}$, $\beta_{f,0}$ represents the true value of β_f, and $X_f = (1, X)$ with X being 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 β, MSE should be as small as possible.

We compared our method with the sparse least trimmed squares method, Least Absolute Deviation-LASSO (Wang, Li, and Jiang (2007)), the robust and efficient weighted least squares estimators (Gervini and Yohai (2002)) and the estimator proposed by She and Owen (2011) with L1 penalty, which is just the method in McCann and Welsch (2007). We also compared with the LASSO and adaptive LASSO methods to see how non-robust methods perform under different scenarios. For those scenarios that have outliers, we pretended that we knew the outliers in advance and fitted LASSO and adaptive LASSO only on the true good points. We also compared our method with the oracle LASSO and oracle adaptive LASSO estimator, which served as benchmarks. For the LASSO and adaptive LASSO procedures involved in the comparisons, we used BIC for parameter tuning. For the adaptive LASSO, we chose the weights as the reciprocal of the ordinary least squares estimates (Zou (2006)).

The sparse least trimmed squares method gives a binary weight to each observation. If $w_i = 1$, we identified the ith observation as a normal point and if $w_i = 0$, we regarded it as an outlier. The truncation number was chosen the same as our initial least trimmed squares fit, $h = \lfloor 0.75n \rfloor$. For the Least Absolute Deviation-LASSO method and LASSO, they cannot be used for outlier detection, we only report the MSE, FZR, FPR, SR, and CR. For the oracle LASSO, since we only fit the LASSO on the true good points, we only report the MSE, FZR, FPR, SR and CR. For the robust and efficient weighted least squares estimators method, we also used an initial least trimmed

squares fit with the truncation number $h = \lfloor 0.75n \rfloor$. Since the robust and efficient weighted least squares estimators method and She and Owen (2011)'s method do not perform variable selection, we only report the M, S, JD and MSE.

We used different (n, p, V, L, c) combinations. For each combination, we ran Monte Carlo studies with 1000 replicates. We report the average over 1000 replicates in terms of the aforementioned performance criteria. The standard errors of these quantities are given in the corresponding parentheses. We use PM, SLTS, LL, REWLS, She, LASSO, ALASSO, Oracle, and Aoracle to denote our proposed method, the sparse least trimmed squares, the Least Absolute Deviation-LASSO method, the robust and efficient weighted least squares estimators, She and Owen (2011)'s method, LASSO, adaptive LASSO, the oracle LASSO, and the oracle adaptive LASSO, respectively. When no outliers exist, the LASSO and Oracle, ALASSO and Aoracle are exactly the same. Thus, we do not report the results of Oracle and Aoracle when $c = 0$. We have included the results when $n = 100$ in Table 1.1, and $n = 200$ in Table 1.2.

From Tables 1.1 and 1.2, we can see that our method, sparse least trimmed squares, and robust and efficient weighted least squares estimators perform quite well in terms of outlier detection. For She and Owen (2011)'s method, it works well when $(V, L, c) = (4, 0, 0.1)$, but fails for the other scenarios when outliers exist. With the contamination rate of 0.2, our method performs slightly worse than the sparse least trimmed squares and robust and efficient weighted least squares estimators methods in terms of outlier detection. For the LASSO estimator, when the proportion of outliers increases, and L increases, LASSO performs worse in terms of both variable selection and parameter estimation as the MSE can be very large, which shows the benefits of using a robust method when outliers exist. The adaptive LASSO estimator performs even worse than LASSO because it uses the ordinary least squares estimator as an initial fit. Our method has a much higher selection accuracy than the sparse least trimmed squares and the Least Absolute Deviation-LASSO methods for the parameter β, no matter whether we have contamination or not, although the correct coverage rates of the three methods are comparable. The oracle adaptive LASSO method performs better than the oracle LASSO in terms of false positive rate, but worse than oracle LASSO in terms of false zero rate, resulting in higher correct selection rate but lower correct coverage rate. For the MSE, oracle adaptive LASSO is slightly better because it is asymptotically unbiased compared with oracle LASSO.

When there is no contamination, our method is more efficient in estimating β than the sparse least trimmed squares and the Least Absolute Deviation-LASSO method, as expected, because our method is fully efficient compared with the LASSO estimator while the above comparison methods are not. Our method is comparable with LASSO estimator in terms of MSE when no outliers exist. The adaptive LASSO method performs better than LASSO in terms of false positive rate, but worse than LASSO in terms of false zero rate, which results in higher correct selection rate but lower correct coverage rate. For the MSE, the adaptive LASSO is slightly better because it is asymptotically unbiased compared with LASSO. Our estimator is also more efficient than the robust and efficient weighted least squares estimators in the finite sample scenario. Although

Table 1.1: Simulation results for our methods PM compared with the SLTS, LL, REWLS, She, LASSO, ALASSO, Oracle and Aoracle methods when $n = 100$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(100,15,4,0,0.1)	PM	0(0)	0.01(0)	1	0.02(0.002)	0.04(0.002)	0.58	0.91	0.2(0.002)
	SLTS	0(0)	0.04(0.001)	1	0.02(0.002)	0.28(0.007)	0.13	0.9	0.26(0.002)
	LL	*	*	*	0.04(0.003)	0.37(0.01)	0.14	0.8	0.22(0.002)
	REWLS	0(0)	0.01(0)	1	*	*	*	*	0.22(0.001)
	She	0(0)	0.08(0.002)	0.999	*	*	*	*	0.3(0.002)
	LASSO	*	*	*	0.09(0.003)	0.09(0.004)	0.23	0.55	0.57(0.003)
	ALASSO	*	*	*	0.17(0.003)	0.05(0.003)	0.12	0.22	3.65(0.009)
	Oracle	*	*	*	0.01(0.001)	0.11(0.004)	0.36	0.97	0.17(0.001)
	Aoracle	*	*	*	0.04(0.003)	0.05(0.003)	0.51	0.78	0.16(0.001)
(100,15,4,0,0.2)	PM	0(0.001)	0.02(0.001)	0.998	0.02(0.002)	0.06(0.003)	0.49	0.89	0.22(0.002)
	SLTS	0(0)	0.01(0)	1	0.01(0.002)	0.27(0.007)	0.17	0.94	0.22(0.002)
	LL	*	*	*	0.02(0.002)	0.7(0.008)	0.02	0.89	0.29(0.003)
	REWLS	0(0)	0(0)	0.995	*	*	*	*	0.23(0.001)
	She	0.44(0.016)	0.08(0.003)	0.552	*	*	*	*	0.72(0.01)
	LASSO	*	*	*	0.13(0.004)	0.08(0.003)	0.17	0.41	0.96(0.003)
	ALASSO	*	*	*	0.22(0.004)	0.04(0.003)	0.06	0.12	3.75(0.009)
	Oracle	*	*	*	0.01(0.001)	0.12(0.004)	0.36	0.95	0.18(0.002)
	Aoracle	*	*	*	0.05(0.003)	0.05(0.003)	0.47	0.76	0.17(0.002)
(100,15,4,4,0.1)	PM	0(0)	0(0)	1	0.03(0.002)	0.06(0.003)	0.5	0.87	0.22(0.002)
	SLTS	0(0)	0.04(0.001)	1	0.02(0.002)	0.29(0.007)	0.16	0.89	0.25(0.002)
	LL	*	*	*	0.03(0.003)	0.95(0.003)	0	0.87	2.46(0.006)
	REWLS	0(0)	0.01(0)	1	*	*	*	*	0.22(0.001)
	She	0.93(0.005)	0.03(0.002)	0	*	*	*	*	2.36(0.005)
	LASSO	*	*	*	0.32(0.005)	0.74(0.004)	0	0.09	2.42(0.006)
	ALASSO	*	*	*	0.23(0.005)	0.56(0.005)	0	0.24	16.78(0.947)
	Oracle	*	*	*	0.01(0.001)	0.11(0.004)	0.36	0.97	0.17(0.001)
	Aoracle	*	*	*	0.04(0.003)	0.05(0.003)	0.51	0.78	0.16(0.001)
(100,15,4,4,0.2)	PM	0(0.001)	0.01(0)	0.999	0.04(0.003)	0.09(0.003)	0.36	0.8	0.25(0.003)
	SLTS	0(0)	0.01(0)	1	0.01(0.002)	0.27(0.007)	0.17	0.94	0.22(0.002)
	LL	*	*	*	0.03(0.003)	0.95(0.003)	0	0.88	2.62(0.006)
	REWLS	0(0.001)	0(0)	0.999	*	*	*	*	0.23(0.003)
	She	0.96(0.003)	0.03(0.002)	0	*	*	*	*	2.47(0.006)
	LASSO	*	*	*	0.33(0.005)	0.75(0.004)	0	0.08	2.54(0.006)
	ALASSO	*	*	*	0.26(0.006)	0.59(0.005)	0	0.19	15.03(0.93)
	Oracle	*	*	*	0.01(0.001)	0.12(0.004)	0.36	0.95	0.18(0.002)
	Aoracle	*	*	*	0.05(0.003)	0.05(0.003)	0.47	0.76	0.17(0.002)
(100,15,0,0,0)	PM	*	0(0)	*	0.01(0.002)	0.03(0.002)	0.68	0.94	0.18(0.002)
	SLTS	*	0.08(0.001)	*	0.03(0.002)	0.28(0.007)	0.11	0.86	0.27(0.003)
	LL	*	*	*	0.04(0.003)	0.3(0.009)	0.2	0.81	0.2(0.002)
	REWLS	*	0.02(0.001)	*	*	*	*	*	0.22(0.001)
	She	*	0.01(0.001)	*	*	*	*	*	0.2(0.001)
	LASSO	*	*	*	0(0.001)	0.11(0.004)	0.41	0.98	0.16(0.001)
	ALASSO	*	*	*	0.04(0.002)	0.05(0.003)	0.54	0.82	0.15(0.001)

Table 1.2: Simulation results for our methods PM compared with the SLTS, LL, REWLS, She, LASSO, ALASSO, Oracle and Aoracle methods when $n = 200$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(200,15,4,0,0.1)	PM	0(0)	0.01(0)	1	0(0)	0.03(0.002)	0.71	1	0.15(0.001)
	SLTS	0(0)	0.02(0)	1	0(0.001)	0.25(0.006)	0.12	0.98	0.18(0.002)
	LL	*	*	*	0.02(0.002)	0.2(0.007)	0.34	0.91	0.16(0.001)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.15(0.001)
	She	0(0)	0.06(0.001)	1	*	*	*	*	0.22(0.001)
	LASSO	*	*	*	0.05(0.003)	0.07(0.003)	0.39	0.74	0.5(0.002)
	ALASSO	*	*	*	0.12(0.003)	0.03(0.002)	0.28	0.38	3.61(0.006)
	Oracle	*	*	*	0(0)	0.08(0.003)	0.5	1	0.12(0.001)
	Aoracle	*	*	*	0(0.001)	0.02(0.002)	0.79	0.98	0.1(0.001)
(200,15,4,0,0.2)	PM	0.03(0.005)	0.02(0)	0.974	0.01(0.001)	0.05(0.002)	0.63	0.97	0.19(0.004)
	SLTS	0(0)	0(0)	1	0(0)	0.27(0.006)	0.09	1	0.15(0.001)
	LL	*	*	*	0.01(0.001)	0.61(0.007)	0.01	0.97	0.23(0.002)
	REWLS	0(0)	0(0)	0.996	*	*	*	*	0.16(0.001)
	She	0.38(0.015)	0.07(0.002)	0.624	*	*	*	*	0.57(0.009)
	LASSO	*	*	*	0.08(0.003)	0.06(0.003)	0.33	0.6	0.88(0.002)
	ALASSO	*	*	*	0.16(0.003)	0.03(0.002)	0.15	0.24	3.69(0.006)
	Oracle	*	*	*	0(0)	0.09(0.003)	0.46	1	0.13(0.001)
	Aoracle	*	*	*	0.01(0.001)	0.03(0.002)	0.75	0.96	0.11(0.001)
(200,15,4,4,0.1)	PM	0(0)	0(0)	1	0(0.001)	0.06(0.002)	0.57	0.98	0.17(0.001)
	SLTS	0(0)	0.02(0)	1	0(0.001)	0.28(0.006)	0.12	0.98	0.17(0.002)
	LL	*	*	*	0.05(0.003)	0.94(0.003)	0	0.8	2.47(0.004)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.15(0.001)
	She	0.97(0.003)	0.01(0.001)	0	*	*	*	*	2.44(0.003)
	LASSO	*	*	*	0.28(0.005)	0.87(0.003)	0	0.1	2.46(0.004)
	ALASSO	*	*	*	0.14(0.004)	0.71(0.005)	0	0.44	97.6(45.208)
	Oracle	*	*	*	0(0)	0.08(0.003)	0.5	1	0.12(0.001)
	Aoracle	*	*	*	0(0.001)	0.02(0.002)	0.79	0.98	0.1(0.001)
(200,15,4,4,0.2)	PM	0(0)	0.01(0)	1	0.01(0.001)	0.12(0.004)	0.32	0.94	0.22(0.002)
	SLTS	0(0)	0(0)	1	0(0)	0.3(0.007)	0.13	1	0.15(0.001)
	LL	*	*	*	0.05(0.003)	0.95(0.003)	0	0.81	2.65(0.004)
	REWLS	0(0)	0(0)	1	*	*	*	*	0.16(0.001)
	She	0.98(0.002)	0.01(0.001)	0	*	*	*	*	2.57(0.004)
	LASSO	*	*	*	0.28(0.005)	0.88(0.003)	0	0.11	2.58(0.004)
	ALASSO	*	*	*	0.16(0.005)	0.73(0.005)	0	0.37	35.06(5.304)
	Oracle	*	*	*	0(0)	0.09(0.003)	0.46	1	0.13(0.001)
	Aoracle	*	*	*	0.01(0.001)	0.03(0.002)	0.75	0.96	0.11(0.001)
(200,15,0,0,0)	PM	*	0(0)	*	0(0)	0.02(0.002)	0.78	1	0.12(0.001)
	SLTS	*	0.05(0.001)	*	0.01(0.001)	0.22(0.005)	0.14	0.97	0.19(0.002)
	LL	*	*	*	0.02(0.002)	0.07(0.004)	0.6	0.9	0.14(0.001)
	REWLS	*	0.01(0)	*	*	*	*	*	0.15(0.001)
	She	*	0(0)	*	*	*	*	*	0.14(0.001)
	LASSO	*	*	*	0(0)	0.07(0.003)	0.52	1	0.11(0.001)
	ALASSO	*	*	*	0(0.001)	0.02(0.002)	0.83	0.99	0.09(0.001)

the robust and efficient weighted least squares estimator is also fully efficient asymptotically, its finite-sample efficiency can be relative low, as noted by Gervini and Yohai (2002), for example. Our estimator is also more efficient than She and Owen (2011)'s method in finite sample scenario, which indicates that the adaptive weights not only help for outlier detection but also for estimation of the parameter.

We have run more comprehensive simulations by considering different correlation structures for the design matrix, different combination of p and q, different settings of β, different settings of γ instead of a constant. The results are in Section 1 and Tables 1-6 in the supplementary materials. The findings are quite similar as the findings from this simulation except in Setting III (Tables 5 and 6) in the supplementary, the oracle adaptive LASSO outperforms oracle LASSO in terms of lower false positive rate, higher correct selection rate, and lower MSE. For the false zero rate and correct coverage rate, the methods perform the same.

5 Data application

We applied our method to the Boston housing data, which originated with Harrison and Rubinfeld (1978) and was corrected by Pace and Gilley (1997). The dataset consists of median values of owner-occupied housing and various predictors. We have listed them in the Table 1.3.

Table 1.3: Boston Housing Data variables and descriptions.

Name	Description
CMEDV	Corrected median values of owner-occupied housing
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $\$10,000$
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	lower status of the population
LON	Geographical longitude
LAT	Geographical lattitude

We used exactly the same model as Belsley, Kuh, and Welsch (1980) and Pace and Gilley

(1997), which contains 18 candidate predictors:

$$\begin{aligned}\log(CMEDV) &= \beta_0 + \beta_1 CRIM + \beta_2 ZN + \beta_3 INDUS + \beta_4 CHAS \\ &\quad + \beta_5 NOX^2 + \beta_6 RM^2 + \beta_7 AGE + \beta_8 \log(DIS) \\ &\quad + \beta_9 \log(RAD) + \beta_{10} TAX + \beta_{11} PTRATIO + \beta_{12} B \\ &\quad + \beta_{13} \log(LSTAT) + \beta_{14} LAT + \beta_{15} LON + \beta_{16} LAT \cdot LON \\ &\quad + \beta_{17} LAT^2 + \beta_{18} LON^2.\end{aligned}$$

We standardized all variables in Table 1.3 except DIS, RAD, and $LSTAT$. For these three variables, we first took logs and then standardized them. Since the model involves quadratic terms and interaction term, we standardized these terms after taking squares and multiplication.

We applied our method to the data and selected predictors with indices $(1, 4, 5, 6, 10, 11, 12, 13, 15, 17)$. We also detected six data points with subject indices $(372, 373, 381, 410, 419, 490)$ as the outliers. We compared our method with sparse least trimmed squares, which selects all the predictors. It also detected fifty data points as outliers that contained all six outliers we detected using our method. It has been seen in our simulation studies that the sparse least trimmed squares method often overselects the number of significant predictors because it has a much larger false positive rate. Thus, it is reasonable that sparse least trimmed squares selects more predictors than our method. For the robust and efficient weighted least squares method that does not perform variable selection, we only report the outlier detection result. The robust and efficient weighted least squares method identifies thirty-two data points as outliers, which contains all six outliers found by our method. We also applied the method of She and Owen (2011), which does not identify any data points as outliers.

6 Extension to the high-dimensional case

In this section, we study the problem when the number of parameters p diverges with n, written by p_n in this section. We can extend our outlier detection and variable selection procedures to the case in which p_n is much larger than n. Under this scenario, especially when $p_n > h$, the least trimmed squares cannot be used as an initial fit because it leads to overfitting. We therefore use the sparse least trimmed squares (Alfons et al. (2013)) as the initial fit and get the fitted residuals $\widetilde{\gamma}$. Rather than minimizing the trimmed squares, we find

$$\min_{H,\beta} Q(H, \beta) = \min_{H,\beta} \{ \sum_{i \in H} (y_i - X_i \beta)^2 + h \lambda_s \sum_{j=1}^{p_n} |\beta_j| \},$$

with $H \subseteq \{1, \ldots, n\}$ and $|H| = h$. For a fixed subsample H, if $\hat{\beta}_H = \text{argmin}_\beta Q(H, \beta)$ and

$$H_{opt} = \text{argmin}_{H \subseteq \{1,\ldots,n\}, |H|=h} Q(H, \hat{\beta}_H),$$

the sparse least trimmed squares estimator is $\hat{\beta}_{H_{opt}}$. To choose the tuning parameter λ_s, we use the root trimmed mean squared prediction error criterion (Alfons et al. (2013)). After obtaining the initial fit, we apply the analogous reparameterization and solve (1.2) using the "lars" package in R.

6.1 Theoretical Results

We give the results for diverging p_n. In particular, when p_n diverges at an exponential rate of the sample size n, Corollary 1 shows that our method can select the important predictors consistently and identify all the data as good points with probability tending to 1 when there are no outliers. When outliers exist, Corollary 2 shows that our method still enjoys a high breakdown point. Corollary 3 shows that our method enjoys outlier detection consistency.

A random variable Z is subgaussian if there exists some $C > 0$ such that for every $t \in R$, $E(\{\exp(tZ)\}) \leq \exp(Ct^2/2)$. We need conditions to guarantee reasonable estimates for the initial residual $\widetilde{\gamma}$. Without loss of generality, we assume the first s_n observations are outliers. Let $G = \{s_n + 1, s_n + 2, \ldots, n\}$ denote the indices corresponding to the good points.

(D1) The error ϵ_i's are independent and identically subgaussian distributed.
(D2) The λ_s used in the sparse least trimmed squares satisfy $\lambda_s \to \infty$.
(D3) $E(y_i^2) < \infty$ for all $i \in G$ and $||\beta_0||_2 < \infty$.
Condition (D1) is a stronger condition than (A). It is commonly used in high-dimensional literature that allows the dimension of the covariates p_n to diverge at the exponential rate in n.

We first show variable selection consistency when no outliers exist. We still need the conditions (B1)-(B4), only needing to change p into p_n and q into q_n in them. We still refer to the modified conditions as (B1) to (B4) when p and q are replaced by p_n and q_n.
(B5) $q_n = O(n^{c_1})$ for some constant $0 < c_1 < d$.

Write $a_n \equiv O(b_n)$ if a_n and b_n have the same order.

Corollary 1 *Under conditions (B1)-(B5), (D1)-(D3), if there exists* $0 < c_2 < d - c_1$ *for which* $p_n = O(e^{n^{c_2}})$, *for* $\lambda_n \equiv O(n^{(1+c_3)/2})$ *and* $\mu_n n^{-1-c_1-c_3/2} \to \infty$ *such that* $0 < c_2 < c_3 < d - c_1$, *we have* $pr(\hat{\theta} =_s \theta_0) \to 1$ *as* $n \to \infty$

Corollary 1 allows p_n to diverge at an exponential rate of n, and under this scenario, we can still identify all the data as good points and select the important predictors consistently.

Corollary 2 *If we use the sparse least trimmed squares method with truncation number* h, *under the general position condition the breakdown point of our estimator* $BP(\hat{\beta}, Z) \geq (n - h + 1)/n$.

Since Alfons, Croux, and Gelper (2013) showed that the sparse least trimmed squares has breakdown point $(n-h+1)/n$, our method performs at least as well as the sparse least trimmed squares initial estimator in terms of high breakdown point.

Our method still enjoys outlier detection consistency when p_n diverges at an exponential rate of sample size. We still need the conditions (B2), (C2) and (C3) only needing to change p into p_n and q into q_n in them. With a little abuse of the notation, we still call the modified conditions (B2)(C2)(C3) when p is replaced by p_n. We also need conditions (B5), (D1)-(D3) used in corollary 1.

(C4) $\pi_n n^{-1/2}(\log n)^{-1/4} \to \infty$.

Condition (C4) is parallel to condition (C1), but requires a faster diverging rate of π_n due to the high dimensionality of X.

Corollary 3 *Under conditions (B2), (B5), (C2)-(C4), (D1)-(D3), if there exists a constant* d_1 *such that* $\lambda_n n^{-1/2-d_1/2} \to \infty$, *and there exists* $0 < c_2 < d_1 - c_1$ $(c_1 > 0)$ *for which* $p_n = O(e^{n^{c_2}})$, *then for* $\mu_n = o(\pi_n^2)$, $\mu_n n^{-1}(\log n)^{-1/2} \to \infty$ *and* $\lambda_n n^{-1}\mu_n^{-1}\pi_n \to \infty$, *we have* $pr(\hat{\gamma} =_s \gamma_0) \to 1$ *as* $n \to \infty$.

6.2 Simulation Results

We now report on a simulation study with $p > n$. In particular, we set $(n, p) = (100, 500)$. The true coefficient was set as $\beta_0 = (4, 2, 1, 0.5, 0.2, 0, \ldots, 0)^{\mathrm{T}}$ with $q = 5$ nonzero components and the remaining $(p - q)$ elements zero. The other settings are the same as the previous simulation. We compared our methods with SLTS, LASSO, and ORACLE as the other comparison methods are not applicable for the high dimensional case. For tuning methods, we used the root trimmed mean squared prediction error cross-validation criterion. The results are in Table 1.4. From the results, our method has similar performance as SLTS, obtaining smaller MSE but higher swamping probability.

Table 1.4: Simulation results for our methods PM compared with the SLTS, LASSO, ORACLE methods when $(n, p) = (100, 500)$. The $*$ denotes the values that are not applicable.

(n,p,V,L,c)	method	M	S	JD	FZR	FPR	SR	CR	MSE
(100,500,4,0,0.1)	PM	0(0)	0.12(0.004)	0.997	0.02(0.002)	0.07(0.001)	0	0.88	0.33(0.003)
	SLTS	0(0)	0.05(0.001)	0.985	0.08(0.003)	0.02(0.001)	0.04	0.62	0.5(0.004)
	LASSO	*	*	*	0.11(0.004)	0.15(0.002)	0.02	0.47	1.26(0.007)
	ORACLE	*	*	*	0.03(0.002)	0.14(0.002)	0.07	0.87	0.46(0.003)
(100,500,4,4,0.1)	PM	0(0)	0.19(0.005)	1	0.02(0.002)	0.07(0.001)	0	0.88	0.32(0.002)
	SLTS	0(0)	0.05(0.002)	1	0.07(0.003)	0.02(0.001)	0.04	0.64	0.48(0.003)
	LASSO	*	*	*	0.11(0.004)	0.15(0.002)	0.02	0.47	1.26(0.007)
	ORACLE	*	*	*	0.03(0.002)	0.14(0.002)	0.07	0.87	0.46(0.003)
(100,500,0,0,0)	PM	*	0.07(0.003)	*	0.01(0.001)	0.07(0.001)	0	0.94	0.28(0.003)
	SLTS	*	0.07(0.002)	*	0.09(0.003)	0.01(0.001)	0.03	0.56	0.61(0.005)
	LASSO	*	*	*	0.02(0.002)	0.16(0.002)	0.09	0.9	0.46(0.003)

Acknowledgement

The authors thank the Editor, an associate editor, and two referees for their constructive comments and helpful suggestions, which substantially improved the paper. This research was supported by the U.S. National Institutes of Health and National Science Foundation, and by the Natural Science and Engineering Research Council of Canada.

Supplementary material

Supplementary material available online includes additional simulation results, the auxiliary lemmas and the proofs of the lemmas, theorems, and corollaries.

Bibliography

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

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression diagnostics: identifying influential data and sources of collinearity*, John Wiley & Sons, New York-Chichester-Brisbane, wiley Series in Probability and Mathematical Statistics.

Bondell, H. and Stefanski, L. (2013), "Efficient robust regression via two-stage generalized empirical likelihood," *Journal of the American Statistical Association*, 108, 644–655.

Chen, J. and Chen, Z. (2008), "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, 95, 759–771.

— (2012), "Extended BIC for small-n-large-P sparse GLM," *Statistica Sinica*, 22, 555–574.

Coakley, C. W. and Hettmansperger, T. P. (1993), "A bounded influence, high breakdown, efficient regression estimator," *Journal of the American Statistical Association*, 88, 872–880.

Donoho, D. and Huber, P. J. (1983), "The notion of breakdown point," in *A Festschrift for Erich L. Lehmann*, Belmont, CA: Wadsworth, Wadsworth Statist./Probab. Ser., pp. 157–184.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *The Annals of Statistics*, 32, 407–499.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Gannaz, I. (2006), "Robust Estimation and Wavelet Thresholding in Partial Linear Models," Tech. Rep. math.ST/0612066.

Gervini, D. and Yohai, V. J. (2002), "A class of robust and fully efficient regression estimators," *The Annals of Statistics*, 30, 583–616.

Hampel, F. R. (1975), "Beyond location parameters: robust concepts and methods," in *Proceedings of the 40th Session of the International Statistical Institute (Warsaw, 1975), Vol. 1. Invited papers*, vol. 46, pp. 375–382, 383–391 (1976), with discussion.

Harrison, D. and Rubinfeld, D. L. (1978), "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, 5, 81–102.

Huber, P. J. (1981), *Robust statistics*, New York: John Wiley & Sons Inc., wiley Series in Probability and Mathematical Statistics.

Mallows, C. (1975), "On Some Topics in Robustness," *unpublished memorandum, Bell Tel. Laboratories, Murray Hill.*

Maronna, R., Bustos, O., and Yohai, V. (1979), "Bias- and efficiency-robustness of general M-estimators for regression with random carriers," in *Smoothing techniques for curve estimation (Proc. Workshop, Heidelberg, 1979)*, Berlin: Springer, vol. 757 of *Lecture Notes in Math.*, pp. 91–116.

McCann, L. and Welsch, R. E. (2007), "Robust variable selection using least angle regression and elemental set sampling," *Computational Statistics & Data Analysis*, 52, 249–257.

Pace, K. and Gilley, O. W. (1997), "Using the Spatial Configuration of the Data to Improve Estimation," *The Journal of Real Estate Finance and Economics*, 14, 333–40.

Rousseeuw, P. and Yohai, V. (1984), "Robust regression by means of S-estimators," in *Robust and nonlinear time series analysis (Heidelberg, 1983)*, New York: Springer, vol. 26 of *Lecture Notes in Statist.*, pp. 256–272.

Rousseeuw, P. J. (1984), "Least median of squares regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J. and Driessen, K. V. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust regression and outlier detection*, New York, NY, USA: John Wiley & Sons, Inc.

She, Y. and Owen, A. B. (2011), "Outlier detection using nonconvex penalized regression," *Journal of the American Statistical Association*, 106, 626–639.

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

Wang, H., Li, G., and Jiang, G. (2007), "Robust regression shrinkage and consistent variable selection through the LAD-Lasso," *Journal of Business & Economic Statistics*, 25, 347–355.

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

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Dehan Kong
Department of Statistical Sciences
University of Toronto
Toronto, ON, Canada, M5S 3G3
E-mail: (kongdehan@utstat.toronto.edu)

Howard Bondell
Department of Statistics
North Carolina State University
Raleigh, NC, USA, 27695
E-mail: (bondell@stat.ncsu.edu)

Yichao Wu
Department of Statistics
North Carolina State University
Raleigh, NC, USA, 27695
E-mail: (wu@stat.ncsu.edu)