Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(7) 2210–2226

DOI: 10.1177/0962280218757560
journals.sagepub.com/home/smm
SAGE

Efficient robust doubly adaptive regularized regression with applications

Rohana J Karunamuni, Linglong Kong and Wei Tu

Abstract

We consider the problem of estimation and variable selection for general linear regression models. Regularized regression procedures have been widely used for variable selection, but most existing methods perform poorly in the presence of outliers. We construct a new penalized procedure that simultaneously attains full efficiency and maximum robustness. Furthermore, the proposed procedure satisfies the oracle properties. The new procedure is designed to achieve sparse and robust solutions by imposing adaptive weights on both the decision loss and the penalty function. The proposed method of estimation and variable selection attains full efficiency when the model is correct and, at the same time, achieves maximum robustness when outliers are present. We examine the robustness properties using the finite-sample breakdown point and an influence function. We show that the proposed estimator attains the maximum breakdown point. Furthermore, there is no loss in efficiency when there are no outliers or the error distribution is normal. For practical implementation of the proposed method, we present a computational algorithm. We examine the finite-sample and robustness properties using Monte Carlo studies. Two datasets are also analyzed.

Keywords

Regularized regression, variable selection, efficiency, robustness

1 Introduction

The need for robust procedures in statistical inference is widely recognized. The importance of robust procedures has also been stressed for regularization methods. They have been widely used for simultaneous variable selection and parameter estimation by the identification of a subset of variables that are associated with a response. Effective variable selection can also lead to parsimonious models with better prediction accuracy and easier interpretation. Most existing methods, such as penalized least-squares and penalized likelihood, are not designed for heavy-tailed distributions and are not robust to model misspecification and in the presence of outliers.

Consider a random sample of observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ following the linear regression model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n \tag{1}$$

where $\mathbf{X}_i \in \mathbb{R}^p$ is a vector of p predictor variables, Y_i is the univariate response variable, $\boldsymbol{\beta} \in \Theta \subseteq \mathbb{R}^p$ is the unknown regression parameter, and the e_i error terms are independent of the $\mathbf{X}_i$'s. An intercept term is included if the first elements of all the $\mathbf{X}_i$'s are 1. We assume that the unobservable e_i's are iid with an unknown distribution $F_0(./\sigma)$ for some scale parameter σ and that F_0 is symmetric about 0. We consider the problem of simultaneous estimation and variable selection in model (1). A penalty function generally facilitates variable selection in regression, and various penalized regression methods have been proposed in this context. In particular, bridge regression,[1] LASSO,[2] SCAD,[3] adaptive LASSO,[4] elastic-net,[5] adaptive elastic-net,[6] and MCP[7] are well known. Fan and Li[3] showed that the SCAD enjoys the *oracle properties*; that is, it simultaneously achieves the variable selection consistency and the optimal estimation error.

Department of Mathematical and Statistical Sciences, University of Alberta, Alberta, Canada

Corresponding author:
Rohana J Karunamuni, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton T6G 2G1, Alberta, Canada.
Email: R.J.Karunamuni@ualberta.ca

The above methods are based on the l_2 estimation loss (the squared error loss), and the lack of robustness of the l_2 loss is well known. Outlying values of $\mathbf{X_i}$ (known as "leverage points") or extreme values of $(\mathbf{X}_i, Y_i)$ (known as "influence points") can jointly have an arbitrarily large influence on l_2-loss-based estimators, which are therefore not robust. Fan and Li[3] examined a general class of penalized robust regression estimators of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \phi(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}) + n \sum_{j=1}^{p} p_{\lambda_{nj}}(|\beta_j|) \right\} \tag{2}$$

where $\phi(.)$ is the Huber function[8] and $p_{\lambda_{nj}}(|\beta_j|)$ is a penalty function on β_j. Since then, various penalized robust regression estimators have been proposed based on various loss and penalty functions. For instance, Wang et al.[9] proposed the LAD-LASSO with $\phi(t) = |t|$ and $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj}|\beta_j|$. See also literature.[10,11] Arslan[12] provided a weighted version of the LAD-LASSO estimator that is more robust to leverage points. Wu and Liu[13] and Wang et al.[14] investigated penalized quantile regression where $\phi(t) = t\{\tau - I(t < 0)\}$, $0 \le \tau \le 1$ and $p_{\lambda_{nj}}(|\beta_j|)$ is either the SCAD or the adaptive LASSO penalty. Kai et al.[15] examined variable selection in the semiparametric varying-coefficient partially linear model via a penalized composite quantile loss.[16] Johnson and Peng[17] studied rank-based variable selection, and Wang and Li[18] proposed a weighted Wilcoxon-type SCAD method for robust variable selection. Leng[19] investigated variable selection via regularized rank regression, and Chen et al.[20] proposed weighted l_2 and l_1 loss functions. Bradic et al.[21] studied the penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. Wang et al.[22] implemented a bounded loss function of the form $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$ with a tuning parameter γ. Fan et al.[23] introduced penalized quantile regression with a weighted l_1-penalty. Alfons et al.[24] and Öllerer et al.[25] studied a sparse least-trimmed squares estimator. Smucler and Yohai[26] recently proposed a robust l_1-penalized MM-estimator[27] with an adaptive l_1-penalty. Loh[28] investigated high-dimensional robust M-estimators.

As many authors have pointed out,[22,23,26] the robustness of regularization methods has not yet been thoroughly studied and well understood. Approaches based on the influence curve, such as optimal bounded influence regression,[29,30] are inherently local and usually force a compromise between efficiency and robustness. The main approach to global robustness in recent years has centered around the construction of high-breakdown-point estimators.[31] It has been observed that methods that use a tuning parameter for high efficiency will be accompanied by an increase in bias as an unpleasant side-effect. Furthermore, they will never achieve maximum asymptotic efficiency and high robustness with a high breakdown point simultaneously.[32] To the best of our knowledge, a penalized regression procedure that achieves maximum robustness and full efficiency simultaneously is not yet available in the literature. This paper may thus be viewed as an attempt to address this issue.

We construct a new regularized regression procedure that attains full efficiency and maximum robustness simultaneously. Furthermore, the proposed procedure satisfies the oracle properties. The down-weighting of outliers is a well-known technique for achieving robustness, and many of the authors mentioned above have successfully implemented this. However, it usually suffers from a loss of efficiency if the chosen model is the true one. For important increases in efficiency, it is necessary to down-weight the outliers *adaptively*. Our procedure down-weights both the leverage and influence observations adaptively using an adaptive weight function. The method is designed to attain the oracle properties with full (oracle) efficiency when the model is correct and to simultaneously achieve maximum robustness when outliers are present and against model misspecification. We examine the robustness properties of the resulting estimator using the finite-sample breakdown point and an influence function. We show that our estimator attains the maximum breakdown point. In summary, the proposed estimator can achieve full efficiency and maximum robustness at the same time in sparse estimation.

The rest of this paper is organized as follows. In Section 2, we construct the proposed penalized regression procedure and investigate its asymptotic properties. In Section 3, we study its robustness properties. In Section 4, we present numerical studies that compare our method with some existing methods. In Section 5, we illustrate the proposed method using two contemporary examples: the ADHD-200 data and the pollution dataset. Section 6 provides concluding remarks. The proofs of the main results are given in the supplementary material.

2 Doubly adaptive penalized regression

2.1 Proposed method

We first define a weight function to adaptively down-weight extreme observations. Let $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma} > 0$ be initial robust estimators of the regression parameter $\boldsymbol{\beta}$ and the scale parameter σ, and let $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ be initial robust

location and scatter estimators of the covariates $\mathbf{X}$, all based on the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. Then we down-weight the extreme observations by a weight function of the form

$$w_n(\mathbf{x}, y) = \varphi_1\left(\tilde{\eta}\left|(y - \mathbf{x}^T\tilde{\boldsymbol{\beta}})/\tilde{\sigma}\right|\right)\varphi_2(\tilde{\eta}d(\mathbf{x})) \tag{3}$$

Here $\varphi_i : [0, \infty) \to (0, 1], i = 1, 2$ are nonincreasing continuous functions that are right-continuous at 0 and satisfy $\varphi_i(0) \underset{P}{=} 1$; $d^2(\mathbf{x}) = (\mathbf{x} - \tilde{\boldsymbol{\mu}})^T(\tilde{\Sigma})^{-1}(\mathbf{x} - \tilde{\boldsymbol{\mu}})$ is the Mahalanobis distance of $\mathbf{x}$ at $(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$; and $\tilde{\eta} > 0$ is a statistic such that $\tilde{\eta} \underset{P}{\to} 0$ under model (1). More details on appropriate statistics for $\tilde{\eta}$ are given below.

The motivation for the weight function (3) is that the least squares estimator satisfies the estimating equation $\sum(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})\mathbf{X}_i = 0$. Thus, leverage or influence points can jointly have an arbitrarily large influence on the value of the least-squares estimator, which is therefore not robust. If $|(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})/\tilde{\sigma}|$, the standardized residual of $(\mathbf{x}, y)$, is large then $(\mathbf{x}, y)$ is an outlier. Similarly, a large standardized leverage point given by $d(\mathbf{x})$ indicates an outlying value of $\mathbf{x}$. The above weight function down-weights the leverage and influence points simultaneously.

A robust penalized regression estimator of $\boldsymbol{\beta}$ is then defined by

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n} w_n(\mathbf{X}_i, Y_i)(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{p} p_{\lambda_{nj}}(|\beta_j|)\right\} \tag{4}$$

where $p_{\lambda_{nj}}(|\beta_j|)$ is a penalty function on β_j.

The role of $w_n(.,.)$ in equation (4) is to discount observations that would otherwise have an undue influence on the proposed estimator $\hat{\boldsymbol{\beta}}_n$. When an outlier value of $(Y_i, \mathbf{X}_i)$ is present, $|(Y_i - \mathbf{X}_i^T\tilde{\boldsymbol{\beta}})/\tilde{\sigma}|$ and $d(\mathbf{X}_i)$ are large and hence the corresponding weights given by $\varphi_1(\tilde{\eta}|(Y_i - \mathbf{X}_i^T\tilde{\boldsymbol{\beta}})/\tilde{\sigma}|)$ and $\varphi_2(\tilde{\eta}d(\mathbf{X}_i))$, respectively, drop to zero quickly. The same situation occurs when $\tilde{\eta}$ is large, i.e. when the sample contains a few observations that simply appear to be inconsistent with the model. Thus, $\hat{\boldsymbol{\beta}}_n$ can be expected to be highly robust to outliers, leverage points, and any departures from the assumed model. If the data appear to be consistent with the model, then $w_n(\mathbf{x}, y) \underset{P}{\to} 1$ for each pair $(\mathbf{x}, y)$ because $\tilde{\eta} \underset{P}{\to} 0$ under the model. This will result in $\hat{\boldsymbol{\beta}}_n$ being asymptotically equivalent to the penalized least squares estimator obtained by minimizing the objective function $\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{p} p_{\lambda_{nj}}(|\beta_j|)$. For example, if $np_{\lambda_{nj}}(|\beta_j|)$ is equal to the adaptive penalty function,[4] then $\hat{\boldsymbol{\beta}}_n$ is asymptotically equivalent to the adaptive LASSO estimator.[4] Indeed, we will show that the estimator $\hat{\boldsymbol{\beta}}_n$ simultaneously attains full *oracle* efficiency and maximum robustness when the initial estimators and the statistic $\tilde{\eta}$ are suitably chosen.

The initial robust location and scatter estimators $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ of continuous covariates can be easily computed using the minimum covariance determinant (MCD) method. The MCD was one of the first affine equivariant and highly robust estimators of multivariate location and scatter; for more recent robust estimates see Maronna and Yohai.[33] For $\tilde{\eta}$ employed in (3), a measure of the goodness-of-fit of the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ to model (1) would be more appropriate. A goodness-of-fit test statistic $\tilde{\eta}$ can be constructed as follows. Let $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma} > 0$ be initial (robust) estimators of $\boldsymbol{\beta}$ and the scale σ, respectively, based on the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. Define the standardized residuals

$$\tilde{R}_i = \frac{Y_i - \mathbf{X}_i^T\tilde{\boldsymbol{\beta}}}{\tilde{\sigma}}, \quad i = 1, \ldots, n \tag{5}$$

Then a large value of $|\tilde{R}_i|$ suggests that $(\mathbf{X}_i, Y_i)$ is an outlier. Let the empirical distribution function of the standardized absolute residuals be

$$F_n^+(t) = \frac{1}{n}\sum_{i=1}^{n} I\left(\left|\tilde{R}_i\right| \le t\right), \quad t > 0 \tag{6}$$

Then a measure of the goodness-of-fit of the data to the assumed regression model is

$$\tilde{\eta} = \sup_{t>0}\left|F_n^+(t) - F_0^+(t)\right| \tag{7}$$

where $F_0^+(t)$ denotes the distribution function of the absolute errors under the model. The distribution F_n^+ can be used as a diagnostic tool to detect outlying observations as well: compare $F_n^+(t)$ with the distribution function of the absolute errors under the model $F_0^+(t)$. If $F_n^+(t) < F_0^+(t)$, then the sample proportion of absolute residuals that

exceed t is greater than the theoretical proportion. If this happens for a large t, then outliers are present in the sample. It can be shown that $\tilde{\eta}$ is $\sqrt{n}$-consistent to 0 under the model with mild conditions.[32]

The actual distribution of the errors, F_0, is likely to be unknown in practice, and thus equation (7) cannot be directly used as a measure of goodness-of-fit. Instead, a hypothetical distribution F must be used in place of F_0 in applications. Typically, the standard normal distribution $N(0, 1)$ is chosen.[32] This choice is reasonable since $\tilde{\eta}$ is robust to small departures from the true distribution F_0. Other measures can also be used: a number of goodness-of-fit statistics for the regression model (1) are discussed in Christensen and Sun.[34] For asymptotic results, we only require that $\tilde{\eta}$ is any $\sqrt{n}$-consistent statistic of 0 under the model. Thus, any positive real-valued $\sqrt{n}$-consistent statistic of 0 under the model can be employed. However, a goodness-of-fit statistic that is sensitive to fairly arbitrary departures from the model would be more appropriate as far as the global robustness properties are concerned.

A *discretized version* of $\tilde{\eta}$, say $\hat{\eta}$, is defined as follows. Let δ be an arbitrary positive number. Starting with the origin as an endpoint, cover the positive real axis with disjoint semiclosed intervals of length $n^{-1/2}\delta$. Set $\hat{\eta}$ equal to the center of the interval that contains $\tilde{\eta}$. Then the tightness of $\{n^{1/2}\tilde{\eta}\}$ implies that of $\{n^{1/2}\hat{\eta}\}$. Also note that $\hat{\eta} \geq n^{-1/2}\delta/2$. When n is large, $\hat{\eta}$ will fall into a compact set in which there are only a finite number of cubes with sides of length $n^{-1/2}\delta$. Hence, $\hat{\eta}$ may be considered deterministic for the purpose of applying asymptotic theory. The discretization device[35] is mostly of theoretical interest, but it has been used in the likelihood literature to relax regularity conditions on the asymptotics.[36,37]

2.2 Oracle and efficiency properties

Under various conditions, we have shown that the proposed penalized estimator $\hat{\boldsymbol{\beta}}_n$ defined by equation (4) is an oracle procedure; that is, it simultaneously achieves the variable selection consistency and the optimal estimation error. See Theorems 1 to 3 in Appendix 1. The proofs of these theorems are given in the Supplementary Material.

Theorem 2 shows that the asymptotic behavior of our estimator is equivalent to that of the adaptive LASSO estimator.[4] Thus, the adaptive weight function on the decision loss has not compromised the efficiency of the estimator. Indeed, our method produces an oracle procedure; that is, it can perform as well as the oracle if the penalization parameter is appropriately chosen. We will show in the next section that our estimator is fully robust as well. Note that the estimator in Theorem 2 is doubly adaptive: it is decision as well as penalty adaptive.

Theorem 3 discusses the performance of a *one-step SCAD type estimator*[38] in the present context. Observe that the extra bias term $(\Sigma_{11} + \Sigma_n)^{-1}\nu_n$ appearing in Theorem 1 part (ii) has been eliminated in Theorems 2 and 3, and the proposed estimator possesses the oracle properties and is asymptotically as good as the least squares estimator for estimating $(\boldsymbol{\beta}^*)_{\mathcal{A}}$ given $(\boldsymbol{\beta}^*)_{\mathcal{A}^c} = 0$.

From Theorem 1 we note that the asymptotic covariance of $\hat{\boldsymbol{\beta}}_n$ is $\frac{1}{n}(\Sigma_{11} + \Sigma_n)^{-1}(\sigma^2\Sigma_{11})(\Sigma_{11} + \Sigma_n)^{-1}$. This is similar to the expression obtained in Fan and Li[3] for their penalized likelihood estimator, derived based on a parametric model assumption on the distribution of $(\mathbf{X}, Y)$. The penalty function assumptions of Theorem 1 are similar to those used in Fan and Li,[3] and they are essentially the same as those used in Wang et al.[22] Note that the hard and SCAD thresholding penalty functions satisfy the condition $a_n = 0$ if $\max_{1\leq j\leq p}\lambda_{nj} = o(1)$. Further, they satisfy the condition that $p^{(1)}_{\lambda_n}(|t|) = 0$ for large $|t|$, which is a sufficient condition for the *unbiasedness* condition defined in Fan and Li.[3] Thus, the extra bias term $(\Sigma_{11} + \Sigma_n)^{-1}\nu_n$ appearing in Theorem 1 part (ii) is negligible for the SCAD penalty. To select the regularization parameter λ_{nj}, a BIC-type criterion can be implemented as in Wang et al.[22]

In Theorems 1 to 3, the $\sqrt{n}$-consistency assumption on $\tilde{\eta}$ may be weakened to $\tilde{\eta} = o_P(n^{-1/4})$, provided the functions φ_i, $i = 1, 2$, have bounded second derivatives. Possible choices of φ_i satisfying conditions in the theorems include: $\varphi_i(x) = \min\{1, x^{-1}\}$; $\varphi_i(x) = \max\{(1 - x^2)^2, 0\}$; $\varphi_i(x) = x^{-1}\sin(x)$, for $i = 1, 2$.

2.3 Computational algorithm

The proposed doubly adaptive penalized estimator in (4) is essentially a convex optimization problem with an L_1 constraint. Therefore, it can be solved using the algorithms used for LASSO, such as the popular LARS algorithm[39] and the coordinate descent algorithm.[40] For our experiments, we use the alternating direction method of multipliers (ADMM) algorithm[41] to compute the proposed estimator. A number of authors have recently used this algorithm to compute penalized regression estimators.[41–43] We give the details of implementing the ADMM algorithm to compute the proposed estimator with the LASSO (L_1) penalty function.

Denote $\mathbf{W} = \text{diag}\{w_n(\mathbf{X}_1, Y_1), \ldots, w_n(\mathbf{X}_n, Y_n)\}$, where $w_n(\mathbf{x}, y)$ is as defined in equation (3). Let $\breve{\mathbf{X}} = (X_{ij})_{n\times p}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be the predictor matrix and the response vector, respectively. Then the optimization problem in equation (4) is equivalent to

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \| \mathbf{W}^{\frac{1}{2}}(\mathbf{Y} - \breve{\mathbf{X}}^T\boldsymbol{\beta}) \|_2^2 + \lambda \| \boldsymbol{\beta} \|_1$$

and the ADMM solves an equivalent formulation

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p,\, \boldsymbol{\gamma}\in\mathbb{R}^p} \| \mathbf{W}^{\frac{1}{2}}(\mathbf{Y} - \breve{\mathbf{X}}^T\boldsymbol{\beta}) \|_2^2 + \lambda \| \boldsymbol{\gamma} \|_1 \quad \text{subject to } \boldsymbol{\beta} - \boldsymbol{\gamma} = 0$$

by alternatively updating the primal variables $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the associated dual variable $\boldsymbol{\alpha}$

$$\boldsymbol{\beta}^{k+1} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big(\| \mathbf{W}^{\frac{1}{2}}(\mathbf{Y} - \breve{\mathbf{X}}^T\boldsymbol{\beta}) \|_2^2 + \frac{\rho}{2} \| \boldsymbol{\beta} - \boldsymbol{\gamma}^k + \rho^{-1}\boldsymbol{\alpha}^k \|_2^2 \Big) \tag{8a}$$

$$\boldsymbol{\gamma}^{k+1} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \lambda \| \boldsymbol{\gamma} \|_1 + \frac{\rho}{2} \| \boldsymbol{\beta}^{k+1} - \boldsymbol{\gamma} + \rho^{-1}\boldsymbol{\alpha}^k \|_2^2 \tag{8b}$$

$$\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^k + \rho(\boldsymbol{\beta}^{k+1} - \boldsymbol{\gamma}^{k+1}) \tag{8c}$$

The $\boldsymbol{\beta}$-update (8a) is essentially a ridge regression, and it has a closed-form solution. The $\boldsymbol{\gamma}$-update (8b) can be solved using a thresholding operator. Therefore, the ADMM algorithm reduces to

$$\boldsymbol{\beta}^{k+1} = (\breve{\mathbf{X}}\mathbf{W}\breve{\mathbf{X}}^T + \rho\mathbf{I})^{-1}(\mathbf{W}^{\frac{1}{2}}\breve{\mathbf{X}}^T\mathbf{W}^{\frac{1}{2}}\mathbf{Y} + \rho(\boldsymbol{\gamma}^k - \boldsymbol{\alpha}^k)) \tag{9a}$$

$$\boldsymbol{\gamma}^{k+1} = S_{\lambda/\rho}(\boldsymbol{\beta}^{k+1} + \boldsymbol{\alpha}^k) \tag{9b}$$

$$\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^k + \rho(\boldsymbol{\beta}^{k+1} - \boldsymbol{\gamma}^{k+1}) \tag{9c}$$

where $S_a(v) = (v-a)_+ - (-v-a)_+$. To calculate the initial weight $w_i(\mathbf{X}_i, Y_i)$ in equation (3), we need an initial estimator $\tilde{\boldsymbol{\beta}}$; more details on computing an initial weight will be given in Section 4.1. We can also use our initial estimator $\tilde{\boldsymbol{\beta}}$ as the initial γ^0. Using a good initial estimator usually gives a good approximation to the result in fewer iterations than if we start at some default initialization. The initial value of α^0 can be chosen as 0, and ρ is a tuning parameter, which can be chosen around 1.[41] Then an estimator of $\boldsymbol{\beta}$ can be obtained by repeating equations (9a) to (9c) until convergence.

3 Robustness properties

An understanding of the robustness properties of any estimator is important from a practical point of view. Various methods have been developed to measure robustness. For instance, bounded influence functions are used to describe the robustness of an estimator in Hampel et al.[44] Another important measure is the breakdown point, which is a global measure of the robustness of outliers. Roughly speaking, the breakdown point of an estimator is the proportion of incorrect observations (i.e. arbitrary values) an estimator can handle before giving an arbitrarily large result.[30,45,46] The asymptotic robustness properties of an estimator can be analyzed by using the maximum bias function and subsequently computing the asymptotic breakdown point. On the other hand, the finite-sample breakdown point reflects the finite-sample robustness properties. For more discussion of these concepts, see Maronna et al.[31] and the references therein.

3.1 Finite-sample breakdown point

The finite-sample breakdown point of an estimator is defined as the largest fraction of data values that can be corrupted (made arbitrarily bad) with the estimates remaining bounded.[30,46] For a random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ satisfying model (1), let $Z_i = (\mathbf{X}_i, Y_i), i = 1, \ldots, n$. Define $\mathbf{Z}_{(n)} = \{Z_1, \ldots, Z_n\}$. For $m \le n$, let $\mathbb{Z}_m$ be the set of all corrupted samples $\mathbf{Z}^*_{(n)} = \{Z_1^*, \ldots, Z_n^*\}$ obtained after replacing m data points of $\mathbf{Z}_{(n)}$ with arbitrary values (i.e. at least $n-m$ elements in $\mathbf{Z}^*_{(n)}$ are common with $\mathbf{Z}_{(n)}$ and at most m are corrupted). Then the finite-sample breakdown point of an estimator $\boldsymbol{\beta}_n$ of $\boldsymbol{\beta}$ based on $\mathbf{Z}^*_{(n)}$ is defined as

$$\varepsilon_n^*(\boldsymbol{\beta}_n, \mathbf{Z}_{(n)}) = \max\left\{\frac{m}{n} : \sup_{\mathbf{Z}^*_{(n)}\in\mathbb{Z}_m} \left\| \boldsymbol{\beta}_n(\mathbf{Z}^*_{(n)}) \right\| < \infty \right\} \tag{10}$$

that is, the smallest fraction of outliers that can carry the estimator beyond all bounds. We assume that $\mathbf{Z}_{(n)}$ is in the *general position*. Recall that $\mathbf{Z}_{(n)}$ is said to be in the general position if no hyperplane in $\mathbb{R}^p$ can contain more than p points of the sample.[47] Given $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$, a scale M-estimator of σ^2 is defined as (see Maronna et al.[31])

$$S_n(\mathbf{v}) = \inf\left\{ s > 0 : n^{-1} \sum_{i=1}^{n} \rho(v_i/s) \leq b \right\} \tag{11}$$

where ρ is even, nonnegative, and nondecreasing for $v \geq 0$ with $\rho(0) = 0$. Usually, b is chosen to be $E(\rho)$ to make S_n consistent for σ when $v_1, \ldots, v_n$ is a random sample from the $N(0, \sigma^2)$ distribution.[32]

Let X_m be the set of all corrupted samples $\mathbf{X}^*_{(n)} = \{\mathbf{X}^*_1, \ldots, \mathbf{X}^*_n\}$ obtained after replacing m data points of $\mathbf{X}_{(n)} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ with arbitrary vectors. Further, let $\boldsymbol{\mu}_n(\mathbf{X})$ and $\Sigma_n(\mathbf{X})$ be the location and scatter estimators of the covariate vector $\mathbf{X}$. Let $\boldsymbol{\beta}_{0n} = \boldsymbol{\beta}_{0n}(\mathbf{Z}^*_{(n)})$ and $\beta_{1n} = \beta_{1n}(\mathbf{Z}^*_{(n)})$, respectively, be an initial robust estimator and the proposed estimator of $\boldsymbol{\beta}$ based on $\mathbf{Z}^*_{(n)} \in \mathbb{Z}_m$. Further, let $\boldsymbol{\mu}_{0n} = \boldsymbol{\mu}_{0n}(\mathbf{X}^*_{(n)})$ and $\Sigma_{0n} = \Sigma_{0n}(\mathbf{X}^*_{(n)})$ be the robust location and scatter estimators based on $\mathbf{X}^*_{(n)} \in X_m$, and let $d^2(\mathbf{X}^*_i)$ be the Mahalanobis distance of $\mathbf{X}^*_i$ based on $\boldsymbol{\mu}_{0n}$ and Σ_{0n}, i.e. $d^2(\mathbf{X}^*_i) = (\mathbf{X}^*_i - \boldsymbol{\mu}_{0n})^T(\Sigma_{0n})^{-1}(\mathbf{X}^*_i - \boldsymbol{\mu}_{0n})$, $i = 1, \ldots, n$. We assume that $\boldsymbol{\mu}_{0n}$ and Σ_{0n} are chosen such that the $d(\mathbf{X}^*_i)$'s satisfy condition **S0** (for **S0**, see Supplementary Material). Robust estimators of location and scatter satisfying condition **S0** are readily available in the literature.[33,48] We assume that a discretized version of $\tilde{\eta}$ (see circa (7) above) is used to define β_{1n}. Then we obtain the following lower bounds for $\varepsilon^*_n(\boldsymbol{\beta}_{1n}, \mathbf{Z}_{(n)})$ as a function of $\varepsilon^*_n(\boldsymbol{\beta}_{0n}, \mathbf{Z}_{(n)})$.

Under the conditions of Theorem 3.1 given in the Supplementary Material, we have

$$\varepsilon^*_n(\boldsymbol{\beta}_{1n}, \mathbf{Z}_{(n)}) \geq \min\{\varepsilon^*_n(\boldsymbol{\beta}_{0n}, \mathbf{Z}_{(n)}),\ b/a,\ 1 - b/a - p/n\}$$

and under the conditions of Theorem 3.2 given in the Supplementary Material, we have

$$\varepsilon^*_n(\boldsymbol{\beta}_{1n}, \mathbf{Z}_{(n)}) \geq \varepsilon^*_n(\breve{\boldsymbol{\beta}}_{0n}, \mathbf{Z}_{(n)})$$

From the preceding results, we obtain $\varepsilon^*_n(\boldsymbol{\beta}_{1n}, \mathbf{Z}_{(n)}) \geq \min\{\varepsilon^*_n(\boldsymbol{\beta}_{0n}, \mathbf{Z}_{(n)}),\ (n-p)/2n\}$ if $b/a = (n-p)/2n$. Furthermore, if $\boldsymbol{\beta}_{0n}$ is an equivariant estimator of $\boldsymbol{\beta}$ then we have $\varepsilon^*_n(\boldsymbol{\beta}_{0n}, \mathbf{Z}_{(n)}) \leq (\lfloor (n-p)/2 \rfloor + 1)/n$, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.[49] Thus, if $\boldsymbol{\beta}_{0n}$ is chosen as an equivariant estimator of $\boldsymbol{\beta}$ then we have $\varepsilon^*_n(\boldsymbol{\beta}_{1n}, \mathbf{Z}_{(n)}) \geq \varepsilon^*_n(\boldsymbol{\beta}_{0n}, \mathbf{Z}_{(n)}) - 1/n$. This shows that $\boldsymbol{\beta}_{1n}$ attains the maximum breakdown point asymptotically if the initial estimator $\boldsymbol{\beta}_{0n}$ is properly chosen.

The tuning parameter λ_n used in Theorem 3.2 of the Supplemetary Material is not data-dependent. In practice, however, the tuning parameters of penalized estimators are generally chosen using a data-driven method, such as cross-validation or an AIC or BIC criterion. The breakdown points of such penalized estimators may be affected, as observed in Theorem 3.1 of the Supplemetary Material. It is well known that the role of a penalty function is the variable selection, not the robustness of the estimator. However, the penalty function of Theorem 3.2 of the Supplemetary Material plays a major role in deciding the breakdown point of the estimator $\boldsymbol{\beta}_{1n}$. In Theorem 3.1 of the Supplemetary Material on the other hand, the breakdown point of $\boldsymbol{\beta}_{1n}$ is mainly decided by the decision component of the objective function (4), and this is more appropriate from a statistical point of view.

The penalty function employed in Theorem 3.1 of the Supplemetary Material covers many commonly used penalty functions, including LASSO, the L_q-penalty with $q > 0$, the logarithm penalty, the elastic-net penalty, and the seamless L_0-penalty in which g is an increasing unbounded function. Furthermore, Theorem 3.1 of the Supplemetary Material covers bounded penalties such as SCAD[3] and MCP.[7]

Another important measure of the global robustness of an estimator is the asymptotic breakdown point, which is defined in terms of the maximum asymptotic bias of the estimator over an ε-contamination neighborhood of the target model. Naturally, this bias increases with ε and eventually becomes infinite. The smallest value of ε for which the maximum asymptotic bias is infinite is called the asymptotic breakdown point of the estimator.[44,50] A theorem similar to Theorem 3.1 of the Supplemetary Material can be stated for the asymptotic breakdown point of the proposed procedure.

3.2 Influence function

The influence function approach is another useful method for evaluating the robustness properties of estimators; it describes the local stability of an estimator in the presence of a single outlier. For the theory behind

influence functions of estimators and for a characterization of robustness based on influence functions, see Hampel et al.[44]

For a fixed point $(\mathbf{x}_0, y_0) \in \mathbb{R}^{p+1}$, let $\triangle_{(\mathbf{x}_0, y_0)}$ denote the corresponding point-mass distribution, and let $H_\varepsilon = (1-\varepsilon)H_0 + \varepsilon\triangle_{(\mathbf{x}_0, y_0)}$ denote the mixture distribution between H_0 and $\triangle_{(\mathbf{x}_0, y_0)}$, where $(\mathbf{X}, Y) \sim H_0 = H_0(\mathbf{x}, y) = G_0(\mathbf{x})F_0((y - \mathbf{x}^T\boldsymbol{\beta})/\sigma)$ with $\mathbf{X} \sim G_0$ and $0 < \varepsilon < 1$. Assume that when $\varepsilon \to 0$ the goodness-of-fit statistic $\eta_0(H_\varepsilon)$ converges to a constant $\eta_0 \geq 0$ and the regularization parameter $\lambda_j(H_\varepsilon)$ has a limit point $\lambda_{0j} > 0, j = 1, \ldots, p$. Let

$$\bar{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ E_{H_0}(w_0(\mathbf{X}, Y)(Y - \mathbf{X}^T\boldsymbol{\beta})^2) + \sum_{j=1}^{p} p_{\lambda_{0j}}(|\beta_j|) \right\} \tag{12}$$

where $w_0(\mathbf{X}, Y) = \varphi_1(\eta_0|(Y - \mathbf{X}^T\boldsymbol{\beta})|/\sigma)\varphi_2(\eta_0 d^2(\mathbf{X}))$ with $\varphi_i, i = 1, 2,$ as in equation (3) and $d^2(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_0)^T (\Sigma_0)^{-1}(\mathbf{X} - \boldsymbol{\mu}_0)$. Clearly, $\bar{\boldsymbol{\beta}}$ is a shrinkage of the true value of $\boldsymbol{\beta}$ to 0. It follows that $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \bar{\boldsymbol{\beta}}$ under the assumptions of Theorem 1. Let

$$\bar{\boldsymbol{\beta}}(H_\varepsilon) = \arg\min_{\boldsymbol{\beta}} \left\{ E_{H_\varepsilon}(w_{H_\varepsilon}(\mathbf{X}, Y)(Y - \mathbf{X}^T\boldsymbol{\beta})^2) + \sum_{j=1}^{p} p_{\lambda_{0j}}(|\beta_j|) \right\} \tag{13}$$

where $w_{H_\varepsilon}(\mathbf{X}, Y) = \varphi_1(\eta_0(H_\varepsilon)|Y - \mathbf{X}^T\boldsymbol{\beta}_0(H_\varepsilon)|/S_0(H_\varepsilon))\varphi_2(\eta_0(H_\varepsilon)d^2(H_\varepsilon))$ with $d^2(H_\varepsilon) = (\mathbf{X} - \boldsymbol{\mu}_0(H_\varepsilon))^T(\Sigma_0(H_\varepsilon))^{-1}$ $(\mathbf{X} - \boldsymbol{\mu}_0(H_\varepsilon))$. We then define the influence function of estimators of type (4) as

$$IF_{\bar{\boldsymbol{\beta}}}(\mathbf{x}_0, y_0) = \lim_{\varepsilon 0} \frac{1}{\varepsilon}\{\bar{\boldsymbol{\beta}}(H_\varepsilon) - \bar{\boldsymbol{\beta}}\}$$

assuming that the limit exists. Then it can be shown that (see Theorem 3.3 of the Supplementary Material)

$$\mathrm{IF}_{\bar{\boldsymbol{\beta}}}(\mathbf{x}_0, y_0) = (\kappa\Sigma + \Sigma_1)^{-1}\{\varphi_1(\eta_0|(y_0 - \mathbf{x}_0^T\boldsymbol{\beta})|/\sigma)\varphi_2(\eta_0 d_0^2)\mathbf{x}_0(y_0 - \mathbf{x}_0^T\bar{\boldsymbol{\beta}}) - \tau - \nu\},$$

where $\nu = (p^{(1)}_{\lambda_{01}}(|\bar{\beta}_1|)\mathrm{sgn}(\bar{\beta}_1), \ldots, p^{(1)}_{\lambda_{0p}}(|\bar{\beta}_p|)\mathrm{sgn}(\bar{\beta}_p))^T, \Sigma_1 = \mathrm{diag}\{p^{(2)}_{\lambda_{01}}(|\bar{\beta}_1|), \ldots, p^{(2)}_{\lambda_{0p}}(|\bar{\beta}_p|)\}, \Sigma = E_{G_0}(\mathbf{X}\mathbf{X}^T), \; d_0^2 = (\mathbf{x}_0 - \boldsymbol{\mu}_0)^T$ $(\Sigma_0)^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_0), \; \kappa = \varphi^*(\eta_0, 0)$, and $\tau = E_{H_0}\{w_0(\mathbf{X}, Y)\mathbf{X}(Y - \mathbf{X}^T\bar{\boldsymbol{\beta}})\}$.

When $\eta_0 = 0$ we have $\kappa = 1$ and $\varphi_1(\eta_0|(y_0 - \mathbf{x}_0^T\boldsymbol{\beta})|/\sigma) = \varphi_2(\eta_0 d_0^2) = 1$. Then the influence function $\mathrm{IF}_{\bar{\boldsymbol{\beta}}}(\mathbf{x}_0, y_0)$ is proportional to $\mathbf{x}_0(y_0 - \mathbf{x}_0^T\bar{\boldsymbol{\beta}})$, which is an unbounded function of $(\mathbf{x}_0, y_0)$. Nevertheless, Theorems 3.1 and 3.2 of the Supplemetary Material show that the proposed estimator can still attain the maximum breakdown point, and thus it is very robust. This behavior is not unusual in the robust estimation context. It has been observed that a broad class of nonpenalized regression estimators, including *S*-estimators, the least median of squares estimator, and certain weighted least-squares estimators, share this characteristic of a positive (even maximum) breakdown point but an unbounded influence function.[32,51]

4 Simulation study

In this section, we conduct a simulation study to evaluate the finite-sample performance of our method, and we compare it with some popular existing methods and the oracle estimator. We considered the regression model that was studied in Alfons et al.[24] We set the sample size n to be 100, and the simulated data sets were obtained from the regression model $Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$, where the coefficient vector $\boldsymbol{\beta} = (\beta_j)_{1 \leq j \leq p}$ with $\beta_1 = \beta_7 = 1.5, \quad \beta_2 = 0.5,$ $\beta_4 = \beta_{11} = 1$, and $\beta_j = 0$ for $j \in \{1, \ldots, p\}\backslash\{1, 2, 4, 7, 11\}$, with $p = 20$. The predictor variable $\mathbf{X}$ is assumed to follow a multivariate normal distribution $N(0, \Sigma_X)$ with covariance matrix $(\Sigma_X)_{i,j} = \rho^{|i-j|}$ for $1 \leq i, j \leq p$. We considered four error distributions:

(i) $\varepsilon \sim N(0, 1)$;
(ii) $\varepsilon \sim 0.8N(0, 1) + 0.2N(10, 6)$;
(iii) a t-distribution with three degrees of freedom;
(iv) a Cauchy distribution.

Table 1. Comparison of estimators with the LASSO penalty.

Case			Oracle	LASSO	ADA	LAD	CQR	ESL	RNE
i		PSR	0.998	0.998	0.992	0.997	0.995	0.993	0.978
		NSR	0.998	0.629	0.936	0.596	0.474	0.352	0.976
	ME	median	0.048	0.114	0.070	0.173	0.179	0.410	0.063
		MAD	0.032	0.059	0.047	0.088	0.084	0.194	0.045
ii		PSR	0.996	0.798	0.598	0.184	0.983	0.984	0.959
		NSR	0.996	0.687	0.865	0.972	0.802	0.806	0.869
	ME	median	0.070	2.346	3.009	7.919	0.249	0.184	0.231
		MAD	0.044	1.201	1.795	1.811	0.150	0.128	0.176
iii		PSR	0.998	0.940	0.864	0.873	0.985	0.987	0.950
		NSR	0.998	0.642	0.881	0.849	0.606	0.473	0.918
	ME	median	0.088	0.578	0.483	0.553	0.291	0.420	0.211
		MAD	0.062	0.410	0.394	0.500	0.171	0.287	0.154
iv		PSR	0.998	0.356	0.250	0.076	0.889	0.930	0.905
		NSR	0.998	0.855	0.940	0.991	0.863	0.753	0.836
	ME	median	0.136	6.573	7.670	8.524	0.403	0.424	0.401
		MAD	0.100	3.620	2.909	1.416	0.317	0.301	0.356

For each setting, we examined six competitors: the oracle method based on the MM-estimator[51]; LASSO[2]; adaptive LASSO,[4] labeled ADA; LAD[9]; CQR, the composite quantile regression with LASSO[16]; ESL[22]; and our method with LASSO penalty, labeled **RNE**. For CQR, we set the quantiles to be $\tau_k = k/10$ for $k = 1, \ldots, 9$. For each method, we obtained its tuning parameter by tenfold cross-validation and selected from a sequence of 100 λ values generated by the R package glmnet using the default setting of LASSO.

For RNE, we obtained an initial estimator $\tilde{\boldsymbol{\beta}}$ for equation (3) via the penalized LTS estimator.[24] For the goodness-of-fit statistic $\tilde{\eta}$ in equation (3), we employed equation (7) with F_0 as the standard normal distribution. In the weight function $w_n(\mathbf{x}_i, y_i) = \varphi_1(\tilde{\eta}|(y_i - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})/\tilde{\sigma}|)\varphi_2(\tilde{\eta}d_i^2)$, the φ_i functions control the decay rate of the weight according to the initial residual and leverage. During our simulation study, we studied many different φ_i's and observed that the performance of our estimator is not very sensitive to the choice of the φ_i's, as long as they satisfy the regularity conditions. For the results reported here, we have taken $\varphi_i(t) = (1 + t^2)^{-5}$, $i = 1, 2$. Furthermore, to calculate the Mahalanobis distances, $d^2(\mathbf{x}_i) = (\mathbf{x}_i - \tilde{\mu})^T(\tilde{\Sigma})^{-1}(\mathbf{x}_i - \tilde{\mu})$, we used the robust location and scatter estimators $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ of the $\mathbf{X}_i$'s defined in Rousseeuw and Driessen.[52]

To compare the variable-selection performance of each method, we computed the positive selection rate (PSR) and noncausal selection rate (NSR). The PSR is the proportion of causal features selected by any method, and the NSR is the proportion of the true zero coefficients not selected by any method. For both PSR and NSR, a value close to 1 is desired. We report the median and the median absolute deviation (MAD) of the model error, as advocated in Fan and Li[3]; this error is defined as

$$\mathrm{ME} = (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T \mathrm{E}[\mathbf{X}\mathbf{X}^T](\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$$

Table 1 gives our simulation results based on 1000 replications for $\rho = 0.5$. For case (i) with a normal error distribution, all the estimators have PSR close to 1, and RNE and ADA have the best performance for the model error and the NSR. Whereas, CQR and ESL have a small NSR and large model error. For case (ii) with a mixed normal error distribution, ADA and LAD perform poorly with a small PSR and large model error, and the robust estimators CQR, ESL and RNE perform reasonably well. For case (iii) where the error distribution follows a t-distribution with three degrees of freedom, RNE has the best performance for the NSR and the model error. For case (iv) where the Cauchy error distribution has an infinite variance, LASSO, ADA, and LAD perform poorly for the PSR and the model error, and RNE has decent performance for both the variable selection and the model error.

We have also carried out a simulation study with ($p = 50, n = 100, \rho = 0.75$) and ($p = 100$, $n = 150$, $\rho = 0.75$). The results for those two cases are reported in the Supplementary Material, as suggested by one of the reviewers; see Tables 1 and 2 in the Supplementary Material for detail. Again, we observed a similar performance among estimators as observed above. In particular, the proposed estimator, RNE, has a good performance for both the variable selection and the model error. ESL also shows good performance in PSR, but it has a poor performance in

Table 2. Characteristics of ADHD subjects and healthy controls.

Characteristics	ADHD ($n = 78$)	Controls ($n = 42$)	p-value
Gender (female/male)	41/37	27/15	0.297[a]
Age (year)	9.75 ± 0.37	10.93 ± 0.55	<0.0005[b]
Verbal IQ	107.37 ± 2.31	114.31 ± 3.11	<0.0005[b]
Performance IQ	110.85 ± 3.31	106.81 ± 4.07	0.1255[b]
Full-scale IQ	110.14 ± 2.56	111.69 ± 3.39	0.4656[b]

[a]The p value was obtained by χ^2 test.
[b]The p value was obtained by two-sample two-tailed t-test.

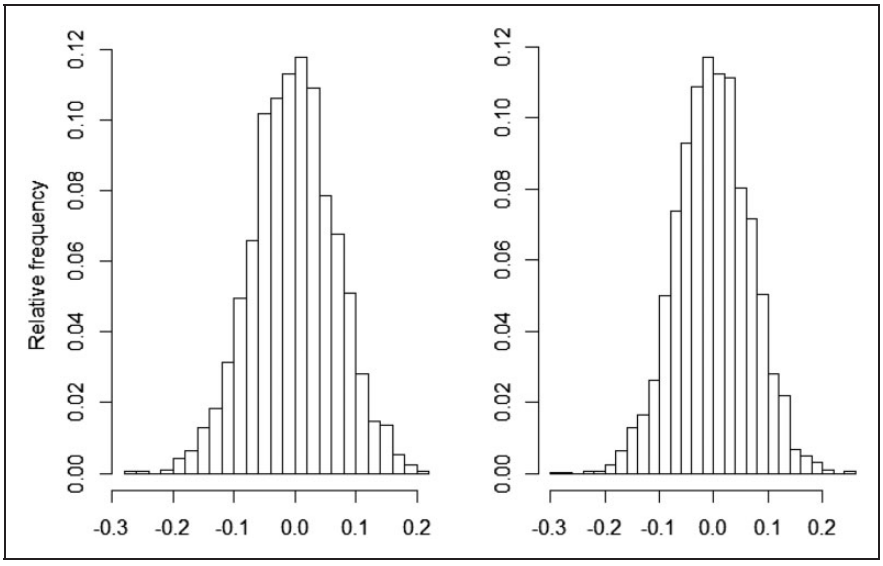

Figure 1. Histograms of mean partial correlations between ROIs (left: control, right: ADHD subjects).

NSR. This means that the method usually selects too many active variables. The non-robust estimators, such as LASSO and ADA, perform poorly in situations when there are outliers, especially in the Cauchy-error case.

We have also explored the connection between collinearity and robustness of estimators. The design matrix is assumed to follow a multivariate normal distribution $N(0, \Sigma_X)$ with covariance matrix $(\Sigma_X)_{i,j} = \rho^{|i-j|}$ for $1 \leq i, j \leq p$, and the level of collinearity increases as the ρ increases. We have looked at the relationship between the performance of estimators and the collinearity level ρ. Figures 1 to 12 given in the Supplementary Material show the affect of ρ on the estimators with $p = 20$, $n = 50$ and under various error models. We observed that most methods actually perform quite stable with the increase of collinearity level. However, we noted that the performance of most estimators gets worse when ρ is very large, such as 0.9. The proposed estimator, RNE, performs reasonably stable in almost all cases considered.

In order to study the effectiveness of estimators under other penalty functions, we have examined the performance of estimators under the SCAD penalty function. Table 3 given in the Supplementary Material presents the results for the proposed estimator and the other six competitors under the SCAD penalty function with tuning parameter $a = 3.7$, as suggested in Fan and Li.[3] To make the results comparable with Table 1, here we considered the same settings with $n = 100$, $p = 20$ and $\rho = 0.5$. From Table 3 of the Supplementary Material, we see that the performance of the proposed estimator, RNE, and the other six competitors is similar to that of Table 1. In particular, RNE achieves good performance in variable selection and model error under all four cases, especially in NSR, with a value close to 1 under all considered scenarios. This means that RNE fared well in screening out the non-signals in the model. The non-robust estimators, such as SCAD and ADA, break down under contaminated models as expected. The ESL estimator also shows good robustness properties; however, the

Table 3. Estimated coefficients for the ADHD-200 data.

Region 1	Classification	Region 2	Classification	MM	ADA	LAD	CQR	LASSO	ESL	RNE
IFGoperc.R	Frontal lobe	ROL.L	Central region	−0.031	−0.028	−0.034	0.038	−0.026	−0.052	0.039
ORBsup.L	Frontal lobe	ROL.R	Central region	−0.034	−0.028	−0.038	0.039	−0.026	−0.04	−0.017
IFGtriang.L	Frontal lobe	OLF.L	Frontal lobe	−0.034	−0.033	−0.018	0.037	−0.028	0	0
SFGdor.R	Frontal lobe	INS.R	Insula	0.004	0	−0.012	0	0.003	0	0
SFGdor.R	Frontal lobe	PCG.L	Limbic lobe	0.018	0.005	0.024	0.02	0.011	0.031	0.036
INS.R	Insula	FFG.L	Occipital lobe	0.015	0.012	0.016	0	0.016	0	0
AMYG.R	Subcortical gray nuclei	PoCG.R	Central region	0.025	0.021	0.031	0.027	0.021	0.061	−0.004
IOG.L	Occipital lobe	SPG.R	Parietal lobe	0.006	0	0	0	0	0	0
INS.L	Insula	IPL.L	Parietal lobe	−0.02	−0.017	−0.015	0	−0.018	0	0
MOG.R	Occipital lobe	IPL.R	Parietal lobe	−0.021	−0.031	−0.018	0.029	−0.025	−0.019	0.041
SMG.L	Parietal lobe	ANG.L	Parietal lobe	−0.016	−0.005	−0.029	0.016	−0.013	−0.022	0.018
OLF.R	Frontal lobe	ANG.R	Parietal lobe	−0.005	0	−0.008	0.011	−0.004	0	0
PHG.R	Limbic lobe	PCUN.R	Parietal lobe	−0.044	−0.043	−0.036	0.041	−0.039	0	−0.031
OLF.R	Frontal lobe	PCL.R	Frontal lobe	−0.014	−0.009	−0.015	0.016	−0.012	0	−0.046
ROL.R	Central region	PUT.L	Subcortical gray nuclei	−0.047	−0.04	−0.041	0.042	−0.037	−0.024	−0.036
INS.L	Insula	PUT.L	Subcortical gray nuclei	0.028	0.027	0.03	0.028	0.024	0.021	0.033
ACG.R	Limbic lobe	PAL.L	Subcortical gray nuclei	0	0	0.002	0	0	0	0
ACG.R	Limbic lobe	THA.L	Subcortical gray nuclei	0.031	0.029	0.021	0.035	0.026	0.014	0.04
ANG.L	Parietal lobe	STG.L	Temporal lobe	0.008	0	0.01	0	0	0.017	0
LING.R	Occipital lobe	STG.R	Temporal lobe	−0.029	−0.017	−0.021	0.024	−0.02	0	0.024
MFG.R	Frontal lobe	MTG.L	Temporal lobe	−0.012	−0.007	−0.019	0.01	−0.011	−0.016	0
HIP.L	Limbic lobe	TPOmid.L	Limbic lobe	0.012	0.013	0	0	0.015	0	0
PCUN.R	Parietal lobe	ITG.L	Temporal lobe	−0.016	0	0	0.013	0	0	0
ANG.R	Parietal lobe	CER45.R	Cerebellum	0.014	0	0.005	0	0.009	0	0
SOG.R	Occipital lobe	CER6.L	Cerebellum	−0.006	0	−0.012	0.009	−0.003	−0.01	0
ORBsup.L	Frontal lobe	CER6.R	Cerebellum	0.03	0.023	0.018	0.031	0.021	0.049	0.028
IFGoperc.R	Frontal lobe	CER7.R	Cerebellum	−0.039	−0.025	−0.048	0.035	−0.028	−0.044	−0.025
CER7.R	Cerebellum	CER8.L	Cerebellum	0.02	0.03	0.021	0.024	0.025	0.029	0.04
PCG.L	Limbic lobe	CER8.R	Cerebellum	0	0	0	0	0	0	0
PCL.R	Frontal lobe	CER9.L	Cerebellum	0.025	0.023	0.009	0.026	0.02	0.01	0
CER8.R	Cerebellum	VER3	Vermis	0.023	0.023	0.027	0.02	0.023	0.018	0
STG.R	Temporal lobe	VER45	Vermis	−0.012	−0.005	−0.022	0.018	−0.011	−0.025	0
TPOsup.L	Limbic lobe	VER9	Vermis	−0.008	0	0.003	0	0	0	0
REC.L	Frontal lobe	VER10	Vermis	0.018	0.005	0.018	0.008	0.013	0	0

NSR is too low for ESL in some cases, especially in case (i). Further, we also noticed that when p / n gets larger, the performance of ESL drops quickly.

As another approach for handling outliers, one of the reviewers suggested the approach of outlier detection first, delete the outliers and then try the usual LASSO and ADA-LASSO. For the outlier detection, the reviewer suggested the IPOD method proposed in She and Owen.[53] The reviewer also suggested to compare our method with the simultaneous outlier detection and variable selection method proposed in Wei[54] as a competitor, which implements the IPOD method of She and Owen[53] for the outlier detection. For our simulation, we applied outlier detection using the IPOD method of She and Owen[53] with $\eta = 2.5,$ and the hard threshold was used. In particular, all observations with outlier indicator $\gamma > 2.5$ were considered as outliers, and then LASSO and adaptive LASSO were applied to the remaining observations; these methods are denoted as IPOD and A-IPOD, respectively, in Table 4 given in the Supplementary Material. Table 4 of the Supplementary Material considers the case with $p = 20$, $n = 100$ and $\rho = 0.5$. For the simultaneous outlier detection and variable selection method proposed in Wei[54] (denoted as Z-log p in Wei[54]), we used the estimator defined by equation (1.5) of Wei[54] and implemented the iterative algorithm proposed in Chapter 6 of Wei.[54]

Table 4 in the Supplementary Material presents the performance of IPOD, A-IPOD and Z-log p, along with other estimators. Compared to LASSO, we see that the outlier detection step using IPOD does improve the performance, and A-IPOD generally has a better performance than IPOD, except in case (iv). A-IPOD also has good performance in case (ii). Z-log p estimator also exhibits a good performance, again except in case (iv), where the error distribution follows a Cauchy distribution. We note that there is a breakdown of model error for

Table 4. Estimated coefficients from pollution dataset.

	Method						
Variable	OLS	MM	ADA	LAD	CQR	ESL	RNE
PREC	0.306	0.326	0.162	0.236	0.152	0	0.051
JANT	−0.317	−0.185	−0.270	−0.061	−0.234	0	−0.285
JULT	−0.237	−0.211	0	0	0	0	0
OVR65	−0.213	−0.257	0	0	0	0	0
POPN	−0.232	−0.163	0	0	0	0	0
EDUC	−0.233	−0.283	−0.184	−0.130	−0.213	−0.277	−0.249
HOUS	−0.054	−0.187	0	0	0	0	0
DENS	0.084	0.249	0.087	0.134	0.169	0.214	0.140
NONW	0.640	0.449	0.609	0.462	0.562	0.414	0.692
WWDRK	−0.014	−0.004	0	0	0	0	0
POOR	−0.011	−0.153	0	0	0	0	0
HC	−0.994	−0.858	−0.990	−0.086	−1.091	−1.428	0
NOX	0.998	0.875	1.074	0	1.130	1.331	0
SOX	0.088	0.074	0	0.310	0	0	0
HUMID	0.009	−0.030	0	0	0	0	0

Table 5. Bootstrap results.

Dataset		ADA	LAD	CQR	ESL	RNE
Pollution	Nonzero	6.800	7.205	6.920	4.736	4.815
	std	1.414	1.225	0.837	0.924	0.447
	MAPE	0.373	0.284	0.370	0.444	0.332
	std	0.145	0.098	0.076	0.060	0.083
	MADPE	0.526	0.427	0.511	0.598	0.461
	std	0.181	0.185	0.154	0.103	0.105
ADHD-200	Nonzero	31.232	23.220	31.395	17.892	15.235
	std	1.213	1.432	0.998	1.124	0.547
	MAPE	0.345	0.354	0.450	0.467	0.322
	std	0.123	0.078	0.058	0.045	0.073
	MADPE	0.634	0.567	0.631	0.578	0.461
	std	0.135	0.135	0.136	0.094	0.045

both IPOD based and Z-log$p$ estimators. The proposed estimator, RNE, performs better than the above three estimators in most cases considered for both variable selection and model error.

In the above simulations, we have generated the co-variates $\mathbf{X}$ from a multivariate normal distribution. A reviewer suggested to include binary variables also as independent variables and the check the performance of estimators. In our model considered above, there are five active predictors. To include binary variables, we changed the first active signal X_1 to follow a Bernoulli(p) distribution with $p = 0.5$. Furthermore, out of the $p - 5$ inactive predictors, we randomly chose a third of them to follow a Bernoulli(p) distribution with $p = 0.5$, and these binary variables were assumed to be independent from each other. The remaining predictors still assumed to follow the multivariate normal distribution used above in the simulation with $p = 20$, $n = 100$ and $\rho = 0.5$. Thus, out of 20 predictors, 6 of them are binary variables. Table 5 given in the Supplementary Material exhibits the performance of the proposed estimator along with the six competitors for the preceding set up of continuous and binary predictors. By comparing Table 1 and Table 5 in the Supplementary Material, we observe a similar performance among estimators. In other words, the inclusion of binary variables has not much affected on the performance of estimators.

Overall, RNE shows good efficiency for clean data situations and has excellent robustness in different settings when there are outliers present or under model misspecification. In summary, RNE is stable and performs reasonably well in comparison to other methods in most cases considered.

5 Applications

In this section, we use two contemporary medically related datasets, the ADHD-200 data and the pollution dataset, to illustrate the performance of our method.

5.1 ADHD data

Attention deficit hyperactivity disorder (ADHD) is one of the most common childhood neuropsychiatric disorders. The psychopathology of ADHD is marked by developmentally inappropriate and pervasive expressions of inattention, overactivity, and impulsiveness, and it often persists into adulthood. The understanding of the underlying pathophysiology of neuropsychiatric illnesses remains insufficient,[55] and clinically useful biomarkers are rarely attained for ADHD.[56] Recent studies have demonstrated the potential of medical imaging such as functional magnetic resonance imaging (fMRI) and diffusion tensor imaging (DTI) in predicting patient outcomes and understanding the underlying pathophysiology of diseases.[57–59]

The data used here is the publicly available resting-state fMRI (rs-fMRI) data from the ADHD-200 Consortium.[60] fMRI is a neuroimaging procedure that measures brain activity by detecting changes associated with blood flow, and rs-fMRI is acquired when a subject is not performing an explicit task. rs-fMRI is useful for exploring the brain's functional organization and determining whether it is altered in neurological or psychiatric diseases. The data set contains 120 subjects ($n = 120$) from the NYU site (New York University Child Study Center) of the ADHD-200 Consortium. The data were preprocessed through the Athena pipeline[61] and are region of interest (ROI) based. The Anatomical Automatic Labeling (AAL) atlas[62] was used for the parcellation. For each subject, there are 172 time courses and the AAL has 116 ROIs. Table 6 given in the Supplementary Material lists the regions and their abbreviations. We use the suffixes .L and .R to differentiate the left and right hemispheres for some bilateral regions. The cerebra include 90 regions (45 in each hemisphere), and the cerebella include 26 regions (nine in each cerebellar hemisphere and eight in the vermis). Of the 120 subjects, 42 are typically developing children and 78 are diagnosed as ADHD. The ADHD group is further separated into ADHD-Combined ($n = 33$) and ADHD-Inattentive ($n = 45$). Table 2 gives the demographic characteristics and neuropsychological scores of the subjects analyzed in this study. The verbal IQ scores measure general knowledge, language, reasoning, and memory skills, while the performance IQ scores measure spatial, sequencing, and problem-solving skills. The verbal and performance IQ scores are summed and converted to obtain the full-scale IQ scores. No significant differences in gender, performance IQ, or full-scale IQ were found, and a significantly higher age and verbal IQ were found in the healthy control subjects.

For each subject, we obtained the mean time series for each of the 116 regions by averaging the fMRI time series over all voxels in the region. We computed partial correlation coefficients between each pair of ROIs. Each partial correlation measures the degree of association between two regions while controlling the effect of the remaining regions. In the variable selection, the goal is to find the significant brain functional connectivities that contribute to the level of ADHD. Each partial correlation is considered to be a predictor, so initially we have $p = (116 \times 116 - 116)/2 = 6670$. Figure 1 shows the histograms of the mean partial correlations between ROIs in ADHD subjects and controls.

Because of the large p small n scenario ($n = 120$, $p = 6670$), we first applied a variable screening to remove some partial correlations that are not significant. We used a Fisher's r-to-z transformation to improve the normality of these partial correlation coefficients. We used a two-tailed t test between the z values of the ADHD group and the control group to determine whether the functional connections are different. The selected significant functional correlations between the ADHD subjects and the control must satisfy two criteria: (1) significantly different z values at the threshold of $p < 0.01$; (2) z values for the correlations that are significantly different from zero in at least one group at the threshold of $p < 0.01$. We did not apply a multiple-testing p-value correction because the screening step is a preliminary step, and a more complicated analysis will be applied using the proposed variable selection algorithms. After the screening, we selected $p = 34$ functional connections for the robust variable selection.

The response variable is the ADHD index, which is a measurement of the overall level of ADHD symptoms. It is a continuous variable ranging from 40 to 90, and typically developing children usually have a score below 50. This variable is more informative than the ordinal ADHD diagnosis result. The ADHD index was log-transformed in the analysis. Robust variable selection is needed for this data set since fMRI signals often suffer from noise and artifacts, and different preprocessing pipelines will also result in slightly different data. A multicollinearity problem is evident among the predictors since the functional networks between regions are often complicated even though we used partial correlations instead of Pearson's correlation.

Table 3 gives the estimated coefficients from the MM, ADA-LASSO, LAD-LASSO, CQR-LASSO, LASSO, ESL-LASSO, and RNE-LASSO. The tuning parameter for each estimator is selected by tenfold cross-validation and selected from a sequence of 100 λ values generated by the R package glmnet. Each row is a partial correlation between Region 1 and Region 2. The selected variables and their coefficients vary widely among the methods tested. ESL-LASSO and RNE-LASSO selected only 18 and 15 variables, respectively, while ADA-LASSO, CQR-LASSO, and LASSO selected at least 24 variables. Of the 15 variables selected by RNE-LASSO, all are shared with the other methods except ESL-LASSO, indicating that our method selected the strongest signals.

During the rs-fMRI, a network called the default-mode network (DMN) was identified. It is a large and robustly replicable network of brain regions that is associated with task-irrelevant mental processes and mind-wandering. Some of the 15 partial correlations selected by RNE-LASSO are related to DMN regions, such as the posterior cingulate cortex (PCG), anterior cingulate and paracingulate gyri (ACG), and inferior parietal (IPL). Similarly to literature,[63,64] we found that stronger connectivity between the ACG and thalamus (THA) is associated with the ADHD index.

5.2 Pollution data

The pollution dataset is from McDonald and Schwing,[65] where ridge regression was illustrated. The goal is to predict the total age-adjusted mortality rate per 100,000 (MORT) using various pollution factors and other covariates. There are 60 observations and 15 covariates: PREC (average annual precipitation in inches), JANT (average January temperature in degrees F), JULT (same for July), OVR65 (% of 1960 SMSA population aged 65 or older), POPN (average household size), EDUC (median school years completed by those over 22), HOUS (percentage of housing units that are sound and with all facilities), DENS (population per sq. mile in urbanized areas, 1960), NONW (percentage of nonwhite population in urbanized areas, 1960), WWDRK (% employed in white collar occupations), POOR (percentage of families with income $<\$3000$), HC (relative hydrocarbon pollution potential), NOX (same for nitric oxide), SOX (same for sulphur dioxide), and HUMID (annual average percentage relative humidity at 1 pm). Variable selection is needed for this data set since a multicollinearity problem was evident amongst the predictors.

Table 4 gives the estimated coefficients from ordinary least squares (OLS), MM, ADA-LASSO, LAD-LASSO, CQR-LASSO, ESL-LASSO, and RNE-LASSO. The tuning parameter for each estimator is obtained using tenfold cross-validation and selected from a sequence of 100 λ values generated by the R package glmnet. The selected variables and their coefficients vary widely among the methods tested. ADA-LASSO, LAD-LASSO, and CQR-LASSO select 7 of the 15 variables, while ESL-LASSO and RNE-LASSO select only 5. Interestingly, although ESL-LASSO and RNE-LASSO select fewer variables, they select different ones. Of the five variables selected by ESL-LASSO, two are different from those of RNE-LASSO, one is different from those of LAD-LASSO, and all are shared with ADA-LASSO and CQR-LASSO. The variables selected by RNE-LASSO are shared by ADA-LASSO, LAD-LASSO, and CQR-LASSO, and three are shared by ESL-LASSO.

Following Wang et al.,[22] we applied a combination of the bootstrap and cross-validation methods to obtain the standard errors of the estimates for the number of nonzeros and the model errors for the two data sets. In the pollution dataset, for each bootstrap sample, we randomly split the 60 observations into a training sample (40 observations) and a testing set (20 observations). For each variable selection method, the median of the absolute prediction error (MAPE) is based on $|Y_i - \mathbf{X}_i^T\widehat{\boldsymbol{\beta}}|$ and the MAD of the prediction error (MADPE) is based on $Y_i - \mathbf{X}_i^T\widehat{\boldsymbol{\beta}}$ for the test set. We repeated the procedure 200 times, and Table 5 summarizes the average MAPE, the MADPE, and the number of nonzero coefficients, along with their corresponding standard deviations. RNE-LASSO and ESL-LASSO select the fewest nonzeros and have smaller standard deviations, which indicates that they have stability and also sparsity.

6 Concluding remarks

Regularization techniques play an important role in identifying covariates that truly affect the outcome of a response in models containing covariates and a response variable. Variable selection is fundamentally important for knowledge discovery with fixed- and high-dimensional data, and it can greatly enhance the prediction performance of a fitted model. Despite considerable progress on variable selection in various models, the robustness of regularization methods has not been thoroughly studied. Efficiency and robustness are extremely important in statistics, and most existing regularization methods fail to achieve both of these goals simultaneously.

We have investigated sparse model estimation in the linear regression model $Y = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon$. Much research has been done in this area, and some robust procedures have been developed. These methods have had varying degrees of success in dealing with "bad" data, but they all suffer from a loss of efficiency if the chosen model is the correct one. In contrast to most existing procedures, our goal was to simultaneously achieve maximum robustness and full efficiency. We have proposed a doubly adaptive regularized procedure that has an adaptive weight function on the decision loss. Our method of estimation and variable selection is designed to attain full efficiency when the model is correct and, at the same time, to achieve maximum robustness when outliers are present or model misspecification occurs. The appropriate down-weighting of extreme observations has been used in the literature to obtain robust estimators. We have proposed particular adaptive weights in which the magnitude of the lack of fit and the extreme observations determine the down-weighting mechanism. The idea is to capture model deviation (misspecification) when it occurs and to manage potential outliers simultaneously with the down-weighting. Furthermore, there is no loss in efficiency when there are no outliers and no model misspecification. We have also presented a computational algorithm. Through theoretical and simulation results, we have demonstrated the merits of our method. Excellent efficiency and robustness properties make our penalized regression procedure appealing in practical applications.

Inspite of considerable progress on variable selection in various models for high-dimensional data (i.e. when $p > n$ and $p \gg n$), there has been relatively little published work on the global robustness of penalized regression and penalized likelihood procedures. Thus, it would be interesting to extend our results for high-dimensional data problems.

The idea used in this paper can be implemented with a likelihood setup as well. For instance, for generalized linear models[66] with the density function $f_\theta(y|\mathbf{x}, \theta) = h(y)\exp(y\theta - l(\theta))$ and $\theta = \mathbf{x}^T\beta$, we define an estimator similar to (4) as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} w_n(\mathbf{X}_i, Y_i)(-Y_i(\mathbf{X}_i^T\boldsymbol{\beta}) + l(\mathbf{X}_i^T\boldsymbol{\beta})) + n\sum_{j=1}^{p} p_{\lambda_{nj}}(|\beta_j|) \right\}$$

where $w_n(\mathbf{X}_i, Y_i) = \varphi(\tilde{\eta}\|(Y_i - l^{(1)}(\mathbf{X}_i^T\tilde{\boldsymbol{\beta}}_n))\mathbf{X}_i\|)$, with $\tilde{\eta}$ being an appropriate goodness-of-fit statistic. We believe that the idea can also be applied to other well-known models such as partially linear models, varying coefficient models, and semiparametric partially linear varying coefficient models.

Acknowledgements

The authors wish to thank two reviewers for their helpful comments and suggestions that led to substantial improvements in the article. Sincere thanks also go to the Editor-in-Chief, Professor Brian Everitt.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada. Linglong Kong's research was also supported by the Canadian Statistical Sciences Institute.

Supplementary Material

Supplementary material is available for this article online.

ORCID iD

Rohana J Karunamuni http://orcid.org/0000-0003-3612-9247

References

1. Frank I and Friedman J. A statistical view of some chemometrics regression tools. *Technometrics* 1993; **35**: 109–135.
2. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B* 1996; **58**: 267–288.

3. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
4. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.
5. Zou H and Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc Ser B* 2005; **67**: 301–320.
6. Zou H and Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat* 2009; **37**: 1733–1751.
7. Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010; **38**: 894–942.
8. Huber PJ. *Robust statistics*. New York, NY: Wiley, 1981.
9. Wang H, Li G and Jiang G. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J Business Econ Stat* 2007; **25**: 347–355.
10. Li G, Peng H and Zhu L. Nonconcave penalized m-estimation with a diverging number of parameters. *Statistica Sinica* 2011; **21**: 391–419.
11. Lambert-Lacroix S and Zwald L. Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic J Stat* 2011; **5**: 1015–1053.
12. Arslan O. Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computat Stat Data Analys* 2012; **56**: 1952–1965.
13. Wu Y and Liu Y. Variable selection in quantile regression. *Stat Sinica* 2009; **19**: 801–817.
14. Wang L, Wu Y and Li R. Quantile regression for analyzing heterogeneity in ultra-high dimension. *J Am Stat Assoc* 2012; **107**: 214–222.
15. Kai B, Li R and Zou H. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Ann Stat* 2011; **39**: 305–332.
16. Zou H and Yuan M. Composite quantile regression and the oracle model selection theory. *Ann Stat* 2008; **36**: 1108–1126.
17. Johnson B and Peng L. Rank-based variable selection. *J Nonparametric Stat* 2008; **20**: 241–252.
18. Wang L and Li R. Weighted Wilcoxon-type smoothly clipped absolute deviation method. *Biometrics* 2009; **65**: 564–571.
19. Leng C. Variable selection and coefficient estimation via regularized rank regression. *Stat Sinica* 2010; **20**: 167–181.
20. Chen X, Wang J and McKeown A. Asymptotic analysis of robust LASSOs in the presence of noise with large variance. *IEEE Transact Inform Theory* 2010; **56**: 5131–5149.
21. Bradic J, Fan J and Wang W. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J Royal Stat Soc Ser B* 2011; **73**: 325–349.
22. Wang X, Jiang Y, Huang M, et al. Robust variable selection with exponential squared loss. *J Am Stat Assoc* 2013; **108**: 632–643.
23. Fan J, Fan Y and Barut E. Adaptive robust variable selection. *Ann Stat* 2014; **42**: 324–351.
24. Alfons A, Croux C and Gelper S. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann Appl Stat* 2013; **7**: 226–248.
25. Öllerer V, Croux C and Alfons A. The influence function of penalized regression estimators. *Statistics* 2015; **49**: 741–765.
26. Smucler E and Yohai VJ. Robust and sparse estimators for linear regression model. *Computat Stat Data Analys* 2017; **111**: 116–130.
27. Yohai V. High breakdown-point and high efficiency robust estimates for regression. *Ann Stat* 1987; **15**: 642–656.
28. Loh P-L. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann Stat* 2017; **45**: 866–896.
29. Hample F. The influence curve and its role in robust estimation. *J Am Stat Assoc* 1974; **69**: 383–393.
30. Huber PJ. Final sample breakdown of M- and P-estimators. *Ann Stat* 1984; **12**: 119–126.
31. Maronna RA, Martin DR and Yohai VJ. *Robust statistics: theory and methods*. New York, NY: Wiley, 2006.
32. Gervini D and Yohai V. A class of robust and fully efficient regression estimators. *Ann Stat* 2002; **30**: 583–616.
33. Maronna RA and Yohai VJ. Robust and efficient estimation of multivariate scatter and location. *Computat Stat Data Analys* 2017; **109**: 64–75.
34. Christensen R and Sun SK. Alternative goodness-of-fit tests for linear models. *J Am Stat Assoc* 2010; **105**: 291–301.
35. Le Cam L. *Théorie Asymptotique de la Décision Statistique*. Les Presses de l'Université de Montréal, 1969.
36. Bickel PJ, Klaassen C, Ritov Y, et al. *Efficient and adaptive estimation for semiparametric models*. New York, NY: Springer, 1998.
37. van der Vaart AW. *Asymptotic statistics*. Cambridge, UK: Cambridge University Press, 2000.
38. Zou H and Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann Stat* 2008; **36**: 1509–1533.
39. Efron B, Hastie T, Johnstone I, et al. Least angle regression. *Ann Stat* 2004; **32**: 407–499.
40. Friedman J, Hastie T, Höfling H, et al. Pathwise coordinate optimization. *Ann Appl Stat* 2007; **1**: 302–332.
41. Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations Trends Mach Learn* 2011; **3**: 1–122.
42. Arnold TB and Tibshirani RJ. Efficient implementations of the generalized lasso dual path algorithm. *J Computat Graphic Stat* 2016; **25**: 1–27.
43. Zhu Y. An augmented ADMM algorithm with application to the generalized lasso problem. *J Computat Graph Stat* 2017; **26**: 195–204.

44. Hampel FR, Ronchetti EM, Rousseeuw PJ, et al. *Robust statistics: the approach based on influence functions*. New York, NY: Wiley, 1986.
45. Hampel F. A global qualitative definition of robustness. *Ann Math Stat* 1971; **42**: 1887–1895.
46. Donoho D and Huber P. The notion of breakdown point. In: Bickel PJ, Doksum KA and Hodges JL, Jr. (eds) *A Festschrift for E. L. Lehmann*. Belmont, CA: Wadsworth, 1983, pp.157–184.
47. Rousseeuw PJ and Yohai V. Robust regression by means of S-estimators. In: *Robust and nonlinear time series analysis. Lecture Notes in Statistics*. New York, NY: Springer, 1984, pp.256–272.
48. Gervini D. A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *J Multivariate Analys* 2003; **84**: 116–144.
49. Rousseeuw PJ and Leroy A. *Robust regression and outlier detection*. New York, NY: Wiley, 1987.
50. Martin RD, Yohai VJ and Zamar RH. Min-max bias regression. *Ann Stat* 1989; **17**: 1608–1630.
51. Yohai VJ and Zamar RH. A minimax-bias property of the least-quantile estimates. *Ann Stat* 1993; **21**: 1824–1842.
52. Rousseeuw PJ and Driessen KV. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999; **41**: 212–223.
53. She Y and Owen AB. Outlier detection using nonconvex penalized regression. *J Am Stat Assoc* 2011; **106**: 626–639.
54. Wei L. *Simultaneous variable selection and outlier detection using LASSO with applications to aircraft landing data analysis*. PhD Thesis: Rutgers University, NJ, 2012.
55. Linden DE. The challenges and promise of neuroimaging in psychiatry. *Neuron* 2012; **73**: 8–22.
56. Nestler EJ and Hyman SE. Animal models of neuropsychiatric disorders. *Nat Neurosci* 2010; **13**: 1161–1169.
57. Wang L, Zang Y, He Y, et al. Changes in hippocampal connectivity in the early stages of Alzheimer's disease: evidence from resting state fMRI. *Neuroimage* 2006; **31**: 496–504.
58. Greicius MD, Flores BH, Menon V, et al. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol Psychiatr* 2007; **62**: 429–437.
59. Rombouts SA, Damoiseaux JS, Goekoop R, et al. Model free group analysis shows altered BOLD FMRI networks in dementia. *Human Brain Map* 2009; **30**: 256–266.
60. ADHD-200 Consortium. The ADHD-200 consortium: a model to advance the translationa l potential of neuroimaging in clinical neuroscience. *Frontiers Syst Neurosci* 2012; **6**: 62.
61. Bellec P, Chu C, Chouinard-Decorte F, et al. The Neuro bureau ADHD-200 preprocessed repository. *Neuroimage* 2017; **144**: 275–286.
62. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 2002; **15**: 273–289.
63. Tian L, Jiang T, Wang Y, et al. Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. *Neurosci Lett* 2006; **400**: 39–43.
64. Castellanos FX, Margulies DS, Kelly C, et al. Cingulate-precuneus interactions: a new locus of dysfunction in adult attention-deficit/hyperactivity disorder. *Biol Psychiatr* 2008; **63**: 332–337.
65. McDonald GC and Schwing RC. Instabilities of regression estimates relating air pollution to mortality. *Technometrics* 1973; **15**: 463– 482.
66. McCullagh P and Nelder JA. *Generalized linear models*. 2nd ed. London: Chapman and Hall, 1989.

Appendix 1

We state a few theorems here showing that the proposed penalized estimator $\hat{\boldsymbol{\beta}}_n$ is an oracle procedure. We first introduce some notation and assumptions. Let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ be the true regression coefficient of $\boldsymbol{\beta}$. Let $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ and assume that the cardinality of $\mathcal{A}$ is p_0, where $p_0 < p$. Thus, the true model depends only on p_0 predictors. Without loss of generality, assume that $\mathcal{A} = \{1, 2, \ldots, p_0\}$. Let $\mathcal{A}_n = \{j : \hat{\beta}_{nj} \neq 0\}$, where $\hat{\beta}_{nj}$ is the j^{th}-component of $\hat{\boldsymbol{\beta}}_n$, $j = 1, \ldots, p$. Let Σ be partitioned into submatrices $\Sigma_{11}, \Sigma_{12}, \Sigma_{21}$, and Σ_{22} as follows:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where Σ_{11} is a $p_0 \times p_0$ matrix. Let $a_n = \max\{\left|p^{(1)}_{\lambda_{nj}}(\beta_j^*)\right| : \beta_j^* \neq 0\}$ and $b_n = \max\{\left|p^{(2)}_{\lambda_{nj}}(\beta_j^*)\right| : \beta_j^* \neq 0\}$, where $p^{(i)}_{\lambda_{nj}}(.)$ denotes the i^{th}-derivative of $p_{\lambda_{nj}}(.)$, $i = 1, 2$.

We make the following assumptions on the functions φ_i, $i = 1, 2$, used in equation (3), the predictors $\mathbf{X}_i$, and the errors e_i of model (1):

R0. $\varphi_i : [0, \infty) \to (0, 1]$ is a nonincreasing continuous function with right-continuity at 0 such that $\varphi_i(0) = 1$ with a continuous bounded first derivative $\varphi_i^{(1)}$ satisfying $\varphi_i^{(1)}(0+) = 0$, $i = 1, 2$.

R1. $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T \xrightarrow{P} \Sigma$, where Σ is a positive definite matrix.

R2. $E||\mathbf{X}||^4 < \infty$.
R3. $Ee^2 < \infty$.

Theorem 1. Assume that conditions **R0** to **R3** hold and the second derivative of $p_{\lambda_{nj}}(.)$ exists. Assume that $b_n = o_P(1)$. Suppose that $\hat{\boldsymbol{\beta}}_n$ is defined by equation (4) with $\tilde{\eta}$ satisfying $\tilde{\eta} = O_P(n^{-1/2})$. Then there exists a local minimizer $\hat{\boldsymbol{\beta}}_n$ such that $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*|| = O_P(n^{-1/2} + a_n)$. If further $\sqrt{n}a_n = O_P(1)$, $1/\min_{p_0+1\le j\le p}(\sqrt{n}\lambda_{nj}) = o_P(1)$, and $\lim_{n\to\infty}\lim_{t\to 0+}\{\min_{p_0+1\le j\le p}\lambda_{nj}^{-1}p_{\lambda_{nj}}^{(1)}(t)\} > 0$ with probability one, then we have

(i) Sparsity: $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$.
(ii) Asymptotic normality: $\sqrt{n}(\Sigma_{11} + \Sigma_n)\{(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)_{\mathcal{A}} + (\Sigma_{11} + \Sigma_n)^{-1}v_n\} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2\Sigma_{11})$, where $\Sigma_n = \mathrm{diag}\{p_{\lambda_{n1}}^{(2)}(|\beta_1^*|)/2, \ldots, p_{\lambda_{np_0}}^{(2)}(|\beta_{p_0}^*|)/2\}$ and $v_n = (p_{\lambda_{n1}}^{(1)}(|\beta_1^*|)\mathrm{sgn}(\beta_1^*)/2, \ldots, p_{\lambda_{np}}^{(1)}(|\beta_{p_0}^*|)\mathrm{sgn}(\beta_{p_0}^*)/2)^T$.

Theorem 2. Assume that conditions **R0** to **R3** hold. Suppose that the penalty function is the adaptive penalty: $np_{\lambda_{nj}}(|\beta_j|) = \lambda_n|\beta_j|/|\breve{\beta}_j|^\gamma$ for some $\gamma > 0$ and a $\sqrt{n}$-consistent estimator $\breve{\boldsymbol{\beta}} = (\breve{\beta}_1, \ldots, \breve{\beta}_p)^T$ of $\boldsymbol{\beta}^*$. Assume that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$, as $n \to \infty$. Suppose that $\hat{\boldsymbol{\beta}}_n$ is defined by equation (4) with $\tilde{\eta}$ satisfying $\tilde{\eta} = O_P(n^{-1/2})$. Then we have

(i) Sparsity: $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$.
(ii) Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)_{\mathcal{A}} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2\Sigma_{11}^{-1})$ as $n \to \infty$.

Theorem 3. Assume that conditions **R0** to **R3** hold. Suppose that the penalty function is of the form $p_{\lambda_{nj}}(|\beta_j|) = p_{\lambda_n}^{(1)}(|\breve{\beta}_j|)|\beta_j|$, where $p_{\lambda_n}(.)$ is the SCAD penalty function, $p_{\lambda_n}^{(1)}$ is the first derivative of p_{λ_n}, and $\breve{\boldsymbol{\beta}} = (\breve{\beta}_1, \ldots, \breve{\beta}_p)^T$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}^*$. Assume that $\sqrt{n}\lambda_n \to 0$ and $\lambda_n \to 0$, as $n \to \infty$. Suppose that $\hat{\boldsymbol{\beta}}_n$ is defined by equation (4) with $\tilde{\eta}$ satisfying $\tilde{\eta} = O_P(n^{-1/2})$. Then we have

(i) Sparsity: $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$.
(ii) Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)_{\mathcal{A}} \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2\Sigma_{11}^{-1})$ as $n \to \infty$.

In addition, suppose $p_{\lambda_{nj}}(.) = \lambda_n p(.)$, where $p^{(1)}(.)$ is continuous on $(0, \infty)$, and there is some $s > 0$ such that $p^{(1)}(t) = O(t^{-s})$ as $t \to 0+$. Then (i) and (ii) hold, provided $n^{(s+1)/2}\lambda_n \to \infty$ and $\sqrt{n}\lambda_n \to 0$.