Journal of the Korean Statistical Society 48 (2019) 1–12

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Review

A systematic review on model selection in high-dimensional regression

Eun Ryung Lee [a], Jinwoo Cho [a], Kyusang Yu [b,*]

[a] Sungkyunkwan University, Republic of Korea
[b] Konkuk University, Republic of Korea

ARTICLE INFO

Article history:
Received 26 September 2018
Accepted 19 October 2018
Available online 12 November 2018

AMS 2000 subject classifications:
62J99
62F12

Keywords:
model selection
Penalized methods
LASSO
SCAD
High dimensional regression models
General convex loss
Quadratic margin condition
High level conditions
Model selection consistency
Oracle property

ABSTRACT

High dimensional models are getting much attention from diverse research fields involving very many parameters with a moderate size of data. Model selection is an important issue in such a high dimensional data analysis. Recent literature on theoretical understanding of high dimensional models covers a wide range of penalized methods including LASSO and SCAD. This paper presents a systematic overview of the recent development in high dimensional statistical models. We provide a brief review on the recent development of theory, methods, and guideline on applications of several penalized methods. The review includes appropriate settings to be implemented and limitations along with potential solution for each of the reviewed method. In particular, we provide a systematic review of statistical theory of the high dimensional methods by considering a unified high-dimensional modeling framework together with high level conditions. This framework includes (generalized) linear regression and quantile regression as its special cases. We hope our review helps researchers in this field to have a better understanding of the area and provides useful information to future study.

Contents

1. Introduction

Consider high dimensional linear regression

$$Y_i = \beta_1^0 X_i^{(1)} + \cdots + \beta_p^0 X_i^{(p)} + \epsilon_i,\ i = 1, \ldots, n,$$

where p, the number of covariates is allowed to increase with n, sometimes $n \ll p$, even though the dependence on n is suppressed in the notation. Suppose that $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^\top$ is sparse, i.e., the number of nonzero β_j^0, $1 \le j \le p$ is

* Corresponding author.
E-mail address: kyusangu@konkuk.ac.kr (K. Yu).

https://doi.org/10.1016/j.jkss.2018.10.001
1226-3192/

relatively small. When $p > n$, the ordinary least square estimator does not work. Instead, a type of penalized methods

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

e.g., using L_1 penalty $p_\lambda(|z|) = \lambda|z|$, (Tibshirani, 1996, Least Absolute Shrinkage and Selection Operator, LASSO), nonconvex penalties such as the smoothly clipped absolute deviation (SCAD) by Fan and Li (2001) and the minimum concave penalty (MCP) functions by Zhang (2010) has been developed. Here, the SCAD is a nonconvex penalty function p_λ with $p_\lambda(0) = 0$ defined by

$$p'_\lambda(|z|) = \lambda I(|z| \le \lambda) + \frac{(a\lambda - |z|)_+}{a-1} I(|z| > \lambda), \quad |z| > 0, \tag{1.1}$$

for some constant $a > 2$ and MCP is $p_\lambda(|z|) = \lambda \int_0^{|z|}(1 - x/(a\lambda))_+ dx$. They have been popularly used for variable selection and estimation. A huge body of mathematical theory was developed for the understanding of the each penalized methods in high dimension. See Bühlmann and van de Geer (2011), Fan and Lv (2010) and references therein. In a nutshell, the LASSO estimator results in so-called oracle inequality (e.g. Bickel, Ritov, & Tsybakov, 2009; van de Geer, 2007, 2008), which implies its statistical accuracy in prediction and estimation errors are almost as good as an infeasible case when one knew which coefficients β_j are nonzero. However, in terms of model selection, it is generally inconsistent and it requires quite restrictive conditions on a design matrix such as irrepresentable condition in order to achieve the selection consistency, (see Zhao & Yu, 2006, for example). Nonconvex penalized estimators such as the SCAD and the MCP enjoy the model selection consistency under less restrictive conditions (see Fan & Peng, 2004; Kim, Choi, & Oh, 2012; Kwon & Kim, 2012; Wang, Wu, & Li, 2012; Zhang, Li, & Tsai, 2010, for example). But, the theory guarantees only the existence of a local minimum which has the oracle property including the selection consistency. It is generally difficult to check if the computed estimator, depending on a choice of optimization algorithm and a specific initial value, is same as the oracle estimator because there exist potentially multiple local minima. See Kim and Kwon (2012), Wang, Kim, and Li (2013) and Zhang et al. (2010) for example. Additionally, the computation is generally difficult because the problem is nonconvex.

As one way of circumventing such difficulties, a one-step SCAD penalty was proposed and studied in fixed dimensional linear regression, $p < n$, by Zou and Li (2008). Since the ordinary least square estimator $\beta_j^{(0)} = \hat{\beta}_j^{ols}$ is available as a good initial value (close to the true regression coefficients), the paper suggested a local linear approximation of the SCAD penalty

$$p_\lambda(|\beta_j|) \approx p'_\lambda(|\beta_j^{(0)}|) + p'_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|)$$

near $\beta_j \approx \beta_j^{(0)}$. Then the resulting estimator is given as

$$\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + n\sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^{ols}|)|\beta_j|. \tag{1.2}$$

Compared to the (original) SCAD estimator, it has some advantages. First, the computation is rather easy and the estimator is defined as the unique minimum of the penalized loss since the problem (1.2) is convex. Second, it has the oracle property in fixed dimensional models. Additionally, it can be regarded as a weighted LASSO estimator (with different pre-specified penalty parameters $w_j = p'_\lambda(|\hat{\beta}_j^{ols}|)$ instead of single parameter λ for all the coefficients). A different type of weights was also proposed for model selection, see Zou (2006).

Even though it is not straightforward from the literature, this idea of one-step SCAD is still applicable to high-dimensional cases with $p > n$ if a uniformly consistent initial estimator is available, e.g., LASSO estimator for high dimension. We will provide more detailed and systematic analysis of this situation. First, we take one unified framework including quantile and logistic regressions as well as linear regression in order to consider general settings and loss functions simultaneously. Then, we discuss about theoretical results of the LASSO estimator under this unified framework from the literature. This will be used to show its uniform consistency, that is, the LASSO estimator can be used for an initial estimator of the one-step SCAD estimator in high dimensional settings. Further, we suggest the one-step SCAD estimator and provide model selection consistency of the estimator in theory under high-level conditions in this general framework.

2. Settings, methods and theory

For this study, we take the general set-up of van de Geer (2008). van de Geer (2008) illustrated this set-up with examples of quadratic loss, negative log-likelihood, hinge loss and developed oracle inequalities for the LASSO. Here, we newly illustrate that the set-up include quantile regression as a special case, which is one setting of our interests. In this section, we first introduce the set-up that generalize high dimensional regression and theoretical results of the LASSO in the literature (e.g., Bühlmann & van de Geer, 2011; van de Geer, 2008). Then, we will consider a one-step SCAD penalized estimation and establish theory about model selection consistency of the estimator by providing high level conditions under this unified framework.

For a data $\{X_i, Y_i\}_{i=1}^n \subset \chi \times \mathbb{R}$, define the empirical and the theoretical means of a function $g : \chi \times \mathbb{R} \to \mathbb{R}$ as

$$P_n(g) := \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i)$$

$$P(g) := \frac{1}{n}\sum_{i=1}^{n} E[g(X_i, Y_i)|X_i],$$

respectively. Let $(\mathcal{F}, \|\cdot\|)$ be a normed space of functions on χ. Consider a loss function of f at x with y

$$\rho(f)(x, y) \equiv \rho^*(f(x), y) \text{ where } \rho^* : \mathbb{R}^2 \to \mathbb{R}.$$

We assume $\rho^*(\cdot, y)$ is convex for each y. Note that $\rho(\cdot)(x, y)$ can be regarded as a function from $\mathcal{F}$ to $\mathbb{R}$.

Consider regression settings, where $Z_i := (X_i, Y_i)$ with response $Y_i \subset \mathbb{R}$ and predictor $X_i \in \chi$ and f is a regression function. The consideration includes the following examples:

1. Least Squares Regression: it uses quadratic loss
$$\rho(f)(\cdot, y) = (y - f(\cdot))^2 \text{ where } \rho^*(a, y) = (y - a)^2.$$
2. Quantile Regression: it uses check loss
$$\rho(f)(\cdot, y) = u_\tau(y - f(\cdot)) \text{ where } \rho^*(a, y) = u_\tau(y - a)$$
and $u_\tau(z) = z(\tau - I(z < 0))$.
3. Logistic Regression: it uses logistic loss
$$\rho(f)(\cdot, y) = -yf(\cdot) + \log(1 + \exp[f(\cdot)]) \text{ where } \rho^*(a, y) = -ya + \log(1 + \exp(a)).$$

Let $\{\psi_j\}_{j=1}^p$ be a collection of (dictionary) functions on χ. For example, $\psi_j(x_i) = x_i^{(j)}$ in (conventional) linear regression. Consider a linear subspace

$$\mathcal{F}_0 := \left\{ f_\beta(\cdot) = \sum_{j=1}^{p} \beta_j \psi_j(\cdot); \boldsymbol{\beta} \in \mathbb{R}^p \right\}$$

and assume

$$f^0 = \arg\min_{f \in \mathcal{F}} P(\rho(f)), \quad f^0 = f_{\boldsymbol{\beta}^0}$$

with sparse $\boldsymbol{\beta}^0$. Note that for a given f, $\rho(f)$ is a real valued function on $\chi \times \mathbb{R}$. We also define the excess risk as

$$\epsilon(f) := P(\rho(f) - \rho(f^0))$$

for $f \in \mathcal{F}$. Note that $\epsilon(f) \geq 0$ for all $f \in \mathcal{F}$. Given a vector $\boldsymbol{b} = (b_1, \ldots, b_p)^\top \in \mathbb{R}^p$ and any $S \subset \{1, 2, \ldots, p\}$, we write $\boldsymbol{b}_S = (b_j I(j \in S) : j = 1, 2, \ldots, p)^\top$ and define $|S|$ as the cardinality of S. Denote the active set as $S_0 = \{j : \beta_j^0 \neq 0\}$.

We make following assumptions.

A1. "Parametric (generalized) linear models": $\mathcal{F}_0$, a linear subspace of $\mathcal{F}$, contains the true minimizer f_0 (among $f \in \mathcal{F}$).
A2. "Sparsity": the f_0 has a sparse representation $\boldsymbol{\beta}^0$ using the dictionary ψ_j, $1 \leq j \leq p$, that is, the number $|S_0|$ of nonzero coefficients is relatively small compared to p.
A3. "compatibility condition": there exists a constant $\Phi_{comp}^2 \equiv \Phi(L, S_0) > 0$ such that
$$\Phi_{comp}^2 \|\boldsymbol{\beta}_{S_0}\|_1^2 \leq \|f_\beta\|^2 |S_0|$$
for all $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\|\boldsymbol{\beta}_{S_0^c}\|_1 \leq L\|\boldsymbol{\beta}_{S_0}\|_1$.

For simplicity, we assume A1. i.e., the true f^0 lies in $\mathcal{F}_0$. However, the analysis and results could be extended to misspecified cases where $f^0 \notin \mathcal{F}_0$, as long as an approximation error of the minimizer f_0 by a sparse representation in terms of ψ_j is relatively small. Regarding A2, more detailed technical conditions on $|S_0|$ will be given below.

The compatibility condition in A3 is known as a quite weak assumption on design matrix compared to those for LASSO such as the restricted isometry condition (Candès & Tao, 2005, 2007) and the restricted eigenvalue conditions (Bickel et al., 2009). See Fig. 1 of van de Geer and Bühlmann (2009) and Section 6.13 of Bühlmann and van de Geer (2011). Without loss of generality, define $\Phi^2(L, S) = \min\{\|f_\beta\|^2 |S| / \|\boldsymbol{\beta}_S\|_1^2 : \|\boldsymbol{\beta}_{S^c}\|_1 \leq L\|\boldsymbol{\beta}_S\|_1\}$. As the true active set S_0 is involved in the definition of the compatibility constant Φ_{comp}^2, it is infeasible to validate the assumption A3 in practice. However, the strict positiveness of Φ_{comp}^2 is an irrestrictive and natural assumption to be imposed in the problem. This is required for the true regression coefficient in order to be identifiable. Moreover, consider linear regression problem with $\psi_j = x_i^{(j)}$ and further

assume $\hat{\Sigma}_{SS} = I$ and $\hat{\Sigma}_{S^cS^c} = I$ for simplicity. Here, we let $\hat{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}/n$ with $n \times p$ design matrix $\boldsymbol{X} = [x_{ij}]$. Given any subset $S \subset \{1, \ldots, p\}$, denote the $n \times |S|$ design matrix $[x_{ij} : 1 \le i \le n,\ j \in S]$ using only covariates $X_j,\ j \in S$ as $\boldsymbol{X}_S$. And define $\hat{\Sigma}_{SS} = \boldsymbol{X}_S^\top \boldsymbol{X}_S/n$, $\hat{\Sigma}_{S^cS^c} = \boldsymbol{X}_{S^c}^\top \boldsymbol{X}_{S^c}/n$ and $\hat{\Sigma}_{SS^c} = \boldsymbol{X}_S^\top \boldsymbol{X}_{S^c}/n$. Then,

$$\Phi^2(L, S) = \min\{\boldsymbol{\beta}^\top \hat{\Sigma} \boldsymbol{\beta} \cdot |S| : \|\boldsymbol{\beta}_S\|_1 = 1 \ \ \|\boldsymbol{\beta}_{S^c}\|_1 \le L\}.$$

Observe that

$$\begin{aligned}\boldsymbol{\beta}^\top \hat{\Sigma} \boldsymbol{\beta} \cdot |S| &= |S|(\boldsymbol{\beta}_S^\top \hat{\Sigma}_{SS} \boldsymbol{\beta}_S + \boldsymbol{\beta}_{S^c}^\top \hat{\Sigma}_{S^cS^c} \boldsymbol{\beta}_{S^c} + 2\boldsymbol{\beta}_S^\top \hat{\Sigma}_{SS^c} \boldsymbol{\beta}_{S^c}) \\ &\ge |S|(\|\boldsymbol{\beta}_S\|_2^2 + \|\boldsymbol{\beta}_{S^c}\|_2^2 - 2\rho\|\boldsymbol{\beta}_S\|_2\|\boldsymbol{\beta}_{S^c}\|_2) \\ &\ge (1 - \rho^2),\end{aligned}$$

where $\|\cdot\|_2$ is the Euclidean norm of a vector and $\rho = \max\{\boldsymbol{\beta}_S^\top \boldsymbol{X}_S^\top \boldsymbol{X}_{S^c} \boldsymbol{\beta}_{S^c} : \|\boldsymbol{X}_S \boldsymbol{\beta}_S\|_2 = 1, \ \ \|\boldsymbol{X}_{S^c} \boldsymbol{\beta}_{S^c}\|_2 = 1\}$ is the multiple correlation between $\boldsymbol{X}_S$ and $\boldsymbol{X}_{S^c}$. The last inequality follows from the facts that $\|\boldsymbol{\beta}_S\|_2^2 + \|\boldsymbol{\beta}_{S^c}\|_2^2 - 2\rho\|\boldsymbol{\beta}_S\|_2\|\boldsymbol{\beta}_{S^c}\|_2 \ge (1 - \rho^2)\|\boldsymbol{\beta}_S\|_2^2$ and $\|\boldsymbol{\beta}_S\|_1^2 \le |S|\|\boldsymbol{\beta}_S\|_2^2$ by Cauchy–Schwarz inequality. Therefore, the compatibility assumption A3 holds provided that the multiple correlation ρ with $S = S_0$ is bounded away from one, that is, intuitively, a linear combination of the active (relevant) variables X_{S_0} cannot be recovered from the non-active (irrelevant) variables $X_{S_0^c}$.

Under these settings, the LASSO estimator is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\{P_n \rho(f_{\boldsymbol{\beta}}) + \lambda\|\boldsymbol{\beta}\|_1\}, \tag{2.1}$$

employing L_1 penalty $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$. And write $\hat{f} = f_{\hat{\boldsymbol{\beta}}}$. We introduce some concepts and review results of van de Geer (2008) for the LASSO defined in (2.1), which will be also used later. Consider a "local neighborhood" $\mathcal{F}_{local} \subset \mathcal{F}$ of f^0. It is said that the margin condition holds (with G) if there exists a strictly convex function G such that for all $f \in \mathcal{F}_{local}$, $\epsilon(f) \ge G(\|f - f^0\|)$. Given a strictly convex function G on $[0, \infty)$ with $G(0) = 0$, define the convex conjugate H of G by $H(v) = \sup_t\{tv - G(t)\}$ for $v \ge 0$. Define $Z_M = \max_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_1 \le M} |(P_n - P)(\rho(f_\beta) - \rho(f^0))|$ and $T(\lambda_{0,n}) = \{Z_{M_0} \le \lambda_{0,n} M_0\}$ with $M_0 = H\left(\frac{4\lambda\sqrt{s_0}}{\Phi_{comp}}\right)/\lambda_{0,n}$ and $s_0 = |S_0|$. Then, for $\lambda \ge B\lambda_{0,n}$ with some constant B depending only on L, the LASSO estimator $\hat{f}$ satisfies

$$\epsilon(\hat{f}) + \lambda\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 \le 4H\left(\frac{4\lambda\sqrt{s_0}}{\Phi_{comp}}\right) \tag{2.2}$$

on the set $T(\lambda_{0,n})$ provided that the assumptions A1–A3 hold. (see Bühlmann & van de Geer, 2011, for example). This is non-asymptotic results for a certain range of (fixed) λ.

In this paper, we focus "quadratic margin conditions", with a quadratic function G. This corresponds to the case of least square regression. Further, we will illustrate that it is satisfied in the considered examples of the paper under regular conditions. For examples of general margin conditions, refer to Tsybakov and van de Geer (2005) and van de Geer (2008).

Remark 1 (*Quadratic Margin Condition*)**.**

- For a quadratic margin function, G, say $G(t) = t^2/2$ such as in least square regression, the convex conjugate $H(v) = v^2/2$ is quadratic.
- We suppose that for some strictly positive function Γ on χ.

$$P(\rho(f) - \rho(f^0))(\cdot) \ge \Gamma(\cdot)|f(\cdot) - f^0(\cdot)|^2, \ \forall\|f - f^0\|_\infty \le \eta \tag{2.3}$$

Assume that $\Gamma(\cdot) \ge 1/K$ for some constant K. Then, it follows that for all $\|f - f^0\|_\infty \le \eta$,

$$\epsilon(f) \ge c\|f - f^0\|^2,$$

with $c = 1/K$. i.e. $G(t) = ct^2$, so that $H(v) = v^2/(4c)$.

Remark 2 (*Quadratic Approximation*)**.** We illustrate two important examples of which excessive risks allow quadratic approximations. This approximation will be used later in a discussion about quadratic margin condition.

1. In quantile regression, consider

$$f = f_{\boldsymbol{\beta}}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}, \quad f^0 = f_{\boldsymbol{\beta}^0}(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}^0.$$

Let $g_{Y|X=x}$ be the conditional density of Y conditioned on $X = x$ with the corresponding (conditional) distribution function $G_{Y|X=x}$. Observe that

$$\begin{aligned} E(u_\tau(Y-\alpha)|X=x) &= \tau\int_\alpha^\infty y\, g_{Y|X=x}(y)dy - (1-\tau)\int_{-\infty}^\alpha y\, g_{Y|X=x}(y)dy \\ &\quad - \tau\alpha(1-G_{Y|X=x}(\alpha)) + (1-\tau)\alpha G_{Y|X=x}(\alpha) \\ &= \tau\int_{-\infty}^\infty y\, g_{Y|X=x}(y)dy - \int_{-\infty}^\alpha y\, g_{Y|X=x}(y)dy + \alpha(G_{Y|X=x}(\alpha)-\tau). \end{aligned}$$

This gives

$$\begin{aligned} &\frac{\partial}{\partial\alpha}E(u_\tau(Y-\alpha)|X=x) = G_{Y|X=x}(\alpha)-\tau \\ &\frac{\partial^2}{\partial\alpha^2}E(u_\tau(Y-\alpha)|X=x) = g_{Y|X=x}(\alpha). \end{aligned}$$

Therefore, the excessive risk can be computed as

$$\epsilon(f) = P(\rho(f)-\rho(f^0)) = E[(\rho(f)-\rho(f^0))(X,Y)|X=\cdot] \approx g_{Y|X}(f^0)(f-f^0)^2$$

for a function f near f_0.

2. In logistic regression, recall that

$$\begin{aligned} &\rho(f) := -yf + \log(1+\exp(f)) \\ &f^0 = \log\left(\frac{\pi^0}{1-\pi^0}\right),\ \pi^0(x) = P(Y=1|X=x). \end{aligned}$$

Then, for any function f near f_0, the excessive risk is given as

$$\begin{aligned} \epsilon(f) = E\left((\rho(f)-\rho(f^0))(X,Y)|X=\cdot\right) &= -\pi(f-f^0) + \log(1+e^f) - \log(1+e^{f^0}) \\ &\approx \frac{1}{2}(f-f^0)^2\pi^0(1-\pi^0). \end{aligned}$$

Under the quadratic margin condition when G is a quadratic function e.g., ct^2 for some $c > 0$, the nonasymptotic result (2.2) for the LASSO reduces to

$$\epsilon(\hat{f}) + \lambda\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 \le \frac{16\lambda^2|S_0|}{c\Phi_{comp}^2} \tag{2.4}$$

In other words, provided that

I. there exists a random sequence $\lambda_{0,n}$ such that $Z_M \le \lambda_{0,n}M$ for any $M > 0$ and $\lambda_{0,n} = O_p(r_n)$ with $r_n \to 0$;
II. the quadratic margin condition holds,

one has (2.4) in the sparse parametric (generalized) linear models with A1–A3. Later, this result will be used to prove $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 = O_p(r_n|S_0|/\Phi_{comp}^2)$ with $\lambda \asymp \lambda_{0,n}$, also see Remark 3. This further implies consistency of the LASSO if $r_n|S_0|/\Phi_{comp}^2 \to 0$ as $n \to \infty$. Typical stochastic order r_n in I is $n^{-1/2}$ up to a log-factor of p so that the l_1 norm $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$ of the LASSO estimator has the convergence rate of $O_p(n^{-1/2}|S_0|)$ (up to the log-factor) provided that Φ_{comp}^2 is bounded away from zero. The rate order $n^{-1/2}|S_0|$ could be obtained if the true active set S_0 were known.

Example 1. We will illustrate that the above two conditions I and II are satisfied under some regular conditions in the each example considered.

1. (Quantile regression) For simpler illustration, consider $p = O(n^\alpha)$ for some $0 \le \alpha < 1/2$. Then, one has Lemma A.1 and the fact (A.10) in Lee, Noh, and Park (2014b) under some regular conditions. This gives

$$\begin{aligned} Z_M &= \max_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_1\le M} |(P_n - P)(\rho(f_\beta) - \rho(f^0))| \\ &\le \max_{1\le j\le p}\left|\frac{2}{n}\sum_{i=1}^n \psi_j(x_i)\underbrace{I(\tau - I(Y_i - f^0(x_i) < 0))}_{:=\epsilon_i}\right| \times M, \end{aligned}$$

for any $\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}^0\| \le M$, and take $\lambda_{0,n} = 2\max_{1\le j\le p}\left|\frac{1}{n}\sum_{i=1}^n \psi_j(x_i)\epsilon_i\right|$. Notice that $E(\epsilon_i) = 0$. The stochastic order r_n of $\lambda_{0,n}$ can be obtained under regular conditions, e.g., $r_n = \sqrt{\log p/n}$, applying standard concentration inequalities. And as $\ell(a,x) = g_{Y|X=x}(f^0(x))$ where $g_{Y|X=x}$ is the conditional density of Y given x, the quadratic margin condition, that is, II holds provided that $g_{Y|X=x}(f^0(x)) > 0$.

2. (Logistic regression) Let

$$\rho(f)(\cdot, y) = -yf(\cdot) + \log(1 + \exp[f(\cdot)]).$$

Define $\pi(\cdot) := P(Y = 1|X = \cdot)$. Then,

$$\ell(a, x) = E[-ya + \log(1 + \exp(a))|X = \cdot] = -a\pi(x) + \log(1 + \exp(a)).$$

Hence,

$$\arg\min_a \ell(a, x) = \log\left(\frac{\pi^0(x)}{1 - \pi^0(x)}\right) := f^0(x).$$

The first condition I with choice $\lambda_{0,n} = \max_{1\le j\le p}|\sum_{i=1}^n \psi_j(x_i)\epsilon_i|/n$ holds because of

$$Z_M = \max_{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|_1\le M} |(P_n - P)(\rho(f_\beta) - \rho(f^0))| \le \max_{1\le j\le p}|\sum_{i=1}^n \underbrace{(y_i - \pi^0(x_i))}_{:=\epsilon_i}\psi_j(x_i)|/n \times M.$$

By a similar calculation as the above quantile case, one can get the order r_n of $\lambda_{0,n}$. Moreover, we have $\ddot{\ell}(a, x) = \left(\frac{e^a}{1+e^a}\right) \times \left(1 - \frac{e^a}{1+e^a}\right)$. We find for some constant $c' > 0$,

$$\Gamma(\cdot) \ge \frac{(1 - \pi^0(x)) \wedge \pi^0(x)}{c'}.$$

Assume $\pi^0(x)$ is uniformly bounded away from zero and 1, so that $\Gamma(x) \ge c$ for all x and some $c > 0$. Therefore, the quadratic margin with $G(t) = ct^2$ holds.

Remark 3 (*Uniform Consistency of the LASSO*)**.** From (2.4), observe that

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}\right\|_1 = \sum_{j=1}^p |\hat{\beta}_j - \beta_j^0| \le \frac{16\lambda|S^0|}{c\Phi_{comp}^2}.$$

If $r_n|S^0|/c\Phi_{comp}^2 \to 0$ with choosing $\lambda \asymp \lambda_{0,n} = O_p(r_n)$, then thus, the LASSO estimator is consistent uniformly for $1 \le j \le p$ and $\max_{j=1,\dots,p}|\hat{\beta}_j - \beta_j^0| = O_p(r_n|S^0|/\Phi_{comp}^2)$.

As long as a consistent estimator $\tilde{\beta}_j$ uniformly for $1 \le j \le p$ is given for a pilot estimator e.g. the LASSO estimator for high dimensional cases, then we propose the following one-step SCAD estimator for consistent model selection:

$$\hat{\boldsymbol{\beta}} = \arg\min P_n\rho(f_{\boldsymbol{\beta}}) + \sum_{j=1}^p w_{n,j}|\beta_j| \tag{2.5}$$

where $w_{n,j} = P_\lambda(|\tilde{\beta}_j|)$ and P_λ is the derivative of the SCAD penalty function at (1.1) with penalty parameter λ.

Remark 4. When the loss is the quadratic loss in least square regression, the estimator (2.5) is equal to the SCAD penalized estimator computed from the calibrated CCCP algorithm proposed by Wang et al. (2013). In this sense, our proposal is an extension of Wang et al. (2013) to these general set-ups and our study will give some high-level conditions for the estimator to be working in the set-ups.

For simplicity, assume $S_0 = \{1, 2, \dots, s_0\}$. Define the oracle estimator $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}^{ora}, \mathbf{0})$ by taking the unpenalized estimator $\hat{\boldsymbol{\beta}}^{ora}$ using only the relevant $X_j,\ j = 1, 2, \dots, s_0$ for its first s_0 components and setting zeros for the remaining ones. That is,

$$\hat{\boldsymbol{\beta}}^{ora} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^{s_0}} P_n\rho(f_{S_0,\boldsymbol{b}}), \tag{2.6}$$

where we define $f_{S_0,\boldsymbol{b}} = \sum_{j=1}^{s_0}\psi_j b_j$ for any $\boldsymbol{b} = (b_1, \dots, b_{s_0})^\top$. And let $\hat{f}^* = f_{\hat{\boldsymbol{\beta}}^*}$ and $\delta = \inf_{j\in S_0}|\beta_j^0|$. Define $C_n(\boldsymbol{\beta}) = P_n\rho(\boldsymbol{\beta})$. Note that the function C is convex by convexity of the loss function ρ^*. Let $\partial C_n(\boldsymbol{\beta}) = \{\boldsymbol{t} : C_n(\boldsymbol{b}) \ge C_n(\boldsymbol{\beta}) + (\boldsymbol{b}-\boldsymbol{\beta})^\top \boldsymbol{t},\ \text{for any } \boldsymbol{b}\}$. Then, any vector $\boldsymbol{s}(\boldsymbol{\beta}) \in \partial C_n(\boldsymbol{\beta})$ is a sub-gradient of $C_n(\cdot)$ at a point $\boldsymbol{\beta}$. In order to enable model selection consistency for our proposal (2.5), we make one additional assumption:

III. $\max_{1\le j\le p}|s_j(\hat{\boldsymbol{\beta}}^*)| = O_p(b_n)$ for any $\boldsymbol{s}(\hat{\boldsymbol{\beta}}^*) = (s_1(\hat{\boldsymbol{\beta}}^*), s_2(\hat{\boldsymbol{\beta}}^*), \dots, s_p(\hat{\boldsymbol{\beta}}^*))^\top \in \partial C_n(\hat{\boldsymbol{\beta}}^*)$,

where $b_n \to 0$ is some decreasing sequence as n tends to infinity.

Remark 5. If the (convex) loss ρ is differentiable as in least squares and logistic regression, then a sub-gradient $\boldsymbol{s}(\hat{\boldsymbol{\beta}}^*) \in \partial C_n(\hat{\boldsymbol{\beta}}^*)$ is the gradient of $C_n(\cdot)$ at $\hat{\boldsymbol{\beta}}^*$. Then, $s_j(\hat{\boldsymbol{\beta}}^*) = -2\sum_{i=1}^n(Y_i - \hat{f}^*)\psi_j(x_i)/n$ in least square regression and

$\sum_{i=1}^{n}\psi_j(x_i)(y_i-\hat{\pi}^*(x_i))/n$ with $\hat{\pi}^* = \exp(\hat{f}^*)/(1+\exp(\hat{f}^*))$. Notice that $\max_{1\le j\le p}|s_j(\boldsymbol{\beta}^0)|$ is equal to $\lambda_{0,n}$ defined in Example 1, which is of $O_p(r_n)$, $r_n \to 0$ as $n \to \infty$ under the condition I. Thus, the condition III is satisfied provided that the oracle estimator is close to the true i.e., $\hat{f}^* \approx f^0$ (in a certain sense).

Consider the case of quantile regression, where the check loss ρ is non-differentiable. Any sub-differential $\boldsymbol{s}(\hat{\boldsymbol{\beta}}^*) \in \partial C_n(\hat{\boldsymbol{\beta}}^*)$ of this loss C_n has of the form

$$s_j(\hat{\boldsymbol{\beta}}^*) = -\frac{\tau}{n}\sum_{i=1}^{n}\psi_j(x_i)I(Y_i-\hat{f}^*>0) + \frac{1-\tau}{n}\sum_{i=1}^{n}\psi_j(x_i)I(Y_i-\hat{f}^*<0) - \frac{1}{n}\sum_{i=1}^{n}\psi_j(x_i)\nu_i I(Y_i-\hat{f}^*=0), \quad 1\le j\le p$$

for $\nu_i \in [\tau-1, \tau]$. Then, there exists a decreasing sequence $b_n \to 0$ such that $\max_{1\le j\le p}|s_j(\hat{\boldsymbol{\beta}}^*)| = O_p(b_n)$ (under some technical conditions), e.g., by using similar techniques as in the proof of Lemma 2.3 in Wang et al. (2012).

Theorem 1. *Assume the condition III holds and that the pilot estimator $\tilde{\boldsymbol{\beta}}$ used for the estimator at (2.5) is consistent uniformly for $1\le j\le p$ with $\max_{1\le j\le p}|\tilde{\beta}_j-\beta_j^0| = O_p(a_n)$ for some decreasing sequence $a_n \to 0$. And if $\max\{a_n, b_n\}/\lambda \to 0$ and $\lambda/\delta \to 0$, then the following properties hold:*

i. $\hat{\beta}_j = 0$ *for* $j \notin S_0$;
ii. *its nonzero part* $\hat{\boldsymbol{\beta}}_{S_0}$ *is same as* $\hat{\boldsymbol{\beta}}^{ora}$,

with probability tending to one as n tends to infinity.

Proof. From the assumptions that $\max_{1\le j\le p}|\tilde{\beta}_j-\beta_j^0| = O_p(a_n)$, $a_n/\lambda \to 0$ and $\lambda/\delta \to 0$, one has $\max_{j\notin S_0}|\tilde{\beta}_j| \ll \lambda$ and $\min_{j\in S^0}|\tilde{\beta}_j| > \delta/2$ with a probability tending to 1. Therefore, without loss of generality, we assume that

$$w_{n,j} = \lambda \quad \text{for } j\notin S_0 \text{ and } w_{n,j} = 0 \quad \text{for } j\in S_0. \tag{2.7}$$

From (2.6), there exists a vector $\boldsymbol{s}^* = (s_1^*, \ldots, s_p^*) \in \partial C(\hat{\boldsymbol{\beta}}^*)$ such that $s_j^* = 0$ for $1\le j\le s_0$ by convex optimization theory. This further implies

$$C_n(\hat{\boldsymbol{\beta}}) - C_n(\hat{\boldsymbol{\beta}}^*) \ge \sum_{j=s_0+1}^{p} s_j^*(\hat{\beta}_j - \hat{\beta}_j^*) \tag{2.8}$$

From the facts (2.7)–(2.8), one gets

$$\begin{aligned} 0 &\ge P_n\rho(f_{\hat{\boldsymbol{\beta}}}) - P_n\rho(f_{\boldsymbol{\beta}^*}) + \lambda\sum_{j\notin S_0}|\hat{\beta}_j| \\ &\ge (-\max_j|s_j^*| + \lambda)\sum_{j\notin S_0}|\hat{\beta}_j|. \end{aligned}$$

This together with the condition III and $b_n \ll \lambda$ implies that with a probability approaching to one,

$$\hat{\beta}_j = 0, \quad j\notin S_0,$$

consequently, the second property in ii holds because of (2.7). □

Remark 6. Theorem 1 requires a slight stronger property than (uniform) consistency for an initial estimator $\tilde{\boldsymbol{\beta}}$. As long as $\tilde{\boldsymbol{\beta}}$ has some (decreasing) order of the rate of convergence, our proposed method works in theory. In cases with slowly increasing p where the unpenalized estimator is consistent and its order of the convergence rate is given, the unpenalized estimator is one choice for initial estimator. And in high dimensional cases where the unpenalized estimator is not working, the LASSO estimator (2.1) discussed previously can be used.

Tuning parameter selection is also important for performance of penalized estimators such as (2.5). Bayesian information Criterion (BIC) has been used as a criterion for a traditional problem of subset selection in regression. Recently, some studies suggested that the BIC is applicable to a problem of selecting penalty parameter for penalized methods, which guarantees theoretical consistency in model selection in fixed dimensional regression (Wang & Leng, 2007; Wang, Li, & Tsai, 2007; Zhang et al., 2010). However, it was observed that this ordinary BIC tends to overfit in high dimensional cases, thus, there have been several remedies with some high dimensional adjustments proposed for such cases (e.g., Chen & Chen, 2008, 2012; Lee, Noh, & Park, 2014a; Wang, Li, & Leng, 2009). We call them high-dimensional BIC. They can be also seen as a generalized information criterion for model selection (see Kim, Kwon, & Choi, 2012, for example). The current theory of BICs discusses about a model selection consistency for the selection of the penalty parameter $\hat{\lambda}$ by the BIC. This means that the selected set $\hat{S}(\hat{\lambda}) = \{j : \hat{\beta}_j(\hat{\lambda}) \neq 0\}$ by $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ is equal to S_0 with a probability tending to one as $n \to \infty$ under some regularity conditions.

3. Numerical evidences

In order to investigate the finite sample performance of the one-step SCAD estimator defined at (2.5), we conducted some simulations. For this, we consider the following simulated scenarios:

- **Scenario 1 (LS regression model)**
  $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n$ with $\epsilon_i \sim N(0, 1)$;
- **Scenario 2 (Logistic regression model)**
  $Y_i|\boldsymbol{X}_i \sim Bernoulli(\pi(\boldsymbol{X}_i)), \quad i = 1, \ldots, n$ with $\log\left(\frac{\pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}\right) = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ i.e., $\pi(\boldsymbol{X}_i) = \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta})/(1 + \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}))$;
- **Scenario 3 (Quantile regression model)**
  $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n$ with $\epsilon_i \sim t(2)$.

Here, we set the coefficient $\boldsymbol{\beta}$, the number of covariates p and the number M of Monte Carlo replications as follows:

- Coefficients:

  C1 $\beta_1 = 3, \beta_2 = 1.5, \beta_5 = 2$ and $\beta_j = 0$ for $j \notin \{1, 2, 5\}$;
  C2 $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$. Here first two β_j are same as in "C1".

- We set p to 100,200, and 400.
- $\boldsymbol{X}_i = (X_i^1, \ldots, X_i^p)^\top$ are generated from a MVN with mean 0 and covariance matrix $\Sigma = (\sigma_{j_1 j_2})$ with $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$.
- $M = 200$.

The true active set S_0 is $\{1, 2, 5\}$ and $\{1, 2, 3, 4, 5\}$ in our simulation models with the coefficient "C1" and "C2", respectively. When assessing model selection performance, the selected set $\hat{S}$ is called as *correct fit* if it is equal to the true active set, i.e., $\hat{S} = S_0$. It is said to *overfit* (*underfit*) if $\hat{S} \supsetneq S_0$ ($\hat{S} \not\supset S_0$).

We first computed the LASSO estimator (2.1) then obtained the final one-step SCAD estimator (2.5) using the LASSO for an initial estimator. In (2.1) and (2.5), the loss ρ is taken as the quadratic loss, negative logistic likelihood, check loss function with the quantile level $\tau = 0.5$ for Scenarios 1–3, respectively. The numerical implementation for computing the estimators (2.1) and (2.5) was done via the R function 'glmnet' in the first two cases with the sum of squares and logistic likelihood and via the R linear programming solver 'Rglpk_ solve_ LP ' in the last case with check loss. In the simulations, we used the cross validated estimates of prediction errors for tuning parameter selection of (2.1) and we chose the penalty parameter λ for (2.5) by minimizing high dimensional BICs.

In this simulation study, we consider two types of high dimensional BICs. First one is the high-dimensional BIC proposed by Chen and Chen (2008). Precisely, it is

$$\begin{aligned}
\text{H-BIC1}(\lambda) &= \log\left(\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2\right) + \frac{\log n}{n}|\hat{S}(\lambda)| + \frac{2\gamma}{n}\log\binom{p}{|\hat{S}(\lambda)|};\\
&= n^{-1}(-y_i \times \hat{f}_\lambda(x_i) + \log(1 + \exp(\hat{f}_\lambda(x_i)))) + \frac{\log n}{2n}|\hat{S}(\lambda)| + \frac{\gamma}{n}\log\binom{p}{|\hat{S}(\lambda)|};\\
&= \log\left(\sum_{i=1}^{n} u_{0.5}(y_i - \hat{f}_\lambda(x_i))\right) + \frac{\log n}{2n}|\hat{S}(\lambda)| + \frac{\gamma}{n}\log\binom{p}{|\hat{S}(\lambda)|,}
\end{aligned}$$

in Scenarios 1–3, respectively. Here, $\hat{S}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\}$ is the selected set by the penalized estimator $\hat{\beta}_j(\lambda),\ j = 1, \ldots, p$ using the penalty parameter value λ, which is defined at (2.5). When $\gamma = 0$, it becomes the ordinary BIC, named "O-BIC" here. The additional term, which appears last of H-BIC2, is derived from the prior on the collection of submodels using $s(= |\hat{S}(\lambda)|)$ variables. For high dimensional BIC, we tried the same choices $\gamma = 0.5, 1$ as in the simulation study of Chen and Chen (2008) but we reported the results with $\gamma = 1$ for shorten the length. As the value of γ in H-BIC1 increases, the penalty term on model complexity in this high dimensional BIC gets larger so that it tends to select a model of smaller size. So, the H-BIC1 with the choice of $\gamma = 0.5$ gives selection results between the reported results of O-BIC and H-BIC1 with $\gamma = 1$. Additionally, we consider the type considered in Lee et al. (2014a) and Wang et al. (2009). This is defined as

$$\begin{aligned}
\text{H-BIC2}(\lambda) &= \log\left(\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2\right) + C_n\frac{\log n}{n}|\hat{S}(\lambda)|;\\
&= n^{-1}(-y_i \times \hat{f}_\lambda(x_i) + \log(1 + \exp(\hat{f}_\lambda(x_i)))) + C_n\frac{\log n}{2n}|\hat{S}(\lambda)|\\
&= \log\left(\sum_{i=1}^{n} u_{0.5}(y_i - \hat{f}_\lambda(x_i))\right) + C_n\frac{\log n}{2n}|\hat{S}(\lambda)|
\end{aligned}$$

Table 1
Scenario 1: $n = 200$ and $\beta_1 = 3, \beta_2 = 1.5, \beta_5 = 2$ and $\beta_j = 0$ for $j \notin \{1, 2, 5\}$.

p	BIC type	C	O	U	NC	NIC
$p = 100$	O-BIC	43.5	56.5	0.0	3.00	1.09
	H-BIC1	72.0	28.0	0.0	3.00	0.33
	H-BIC2	88.0	12.0	0.0	3.00	0.13
$p = 200$	O-BIC	37.5	62.5	0.0	3.00	1.33
	H-BIC1	74.0	26.0	0.0	3.00	0.34
	H-BIC2	88.0	12.0	0.0	3.00	0.13
$p = 400$	O-BIC	43.0	57.0	0.0	3.00	1.08
	H-BIC1	83.5	16.5	0.0	3.00	0.18
	H-BIC2	93.5	6.5	0.0	3.00	0.07

Table 2
Scenario 2: $n = 200$ and $\beta_1 = 3, \beta_2 = 1.5, \beta_5 = 2$ and $\beta_j = 0$ for $j \notin \{1, 2, 5\}$.

p	BIC type	C	O	U	NC	NIC
$p = 100$	O-BIC	35.5	64.0	0.5	3.00	1.27
	H-BIC1	65.5	33.5	1.0	3.00	0.43
	H-BIC2	81.0	10.0	9.0	2.87	0.12
$p = 200$	O-BIC	31.5	68.5	0.0	3.00	1.40
	H-BIC1	73.5	26.0	0.0	3.00	0.33
	H-BIC2	86.5	1.5	12.0	2.81	0.02
$p = 400$	O-BIC	43.5	56.5	0.0	3.00	0.97
	H-BIC1	81.0	17.5	1.5	3.00	0.20
	H-BIC2	80.0	3.0	17.0	2.72	0.03

Table 3
Scenario 3: $n = 200$ and $\beta_1 = 3, \beta_2 = 1.5, \beta_5 = 2$ and $\beta_j = 0$ for $j \notin \{1, 2, 5\}$.

p	BIC type	C	O	U	NC	NIC
$p = 100$	O-BIC	89.0	11.0	0.0	3.00	0.19
	H-BIC1	96.0	4.0	0.0	3.00	0.05
	H-BIC2	99.0	1.0	0.0	3.00	0.01
$p = 200$	O-BIC	87.0	13.0	0.0	3.00	0.15
	H-BIC1	96.5	3.5	0.0	3.00	0.04
	H-BIC2	99.0	1.0	0.0	3.00	0.01
$p = 400$	O-BIC	67.5	32.5	0.0	3.00	42.07
	H-BIC1	92.5	7.5	0.0	3.00	9.81
	H-BIC2	99.0	1.0	0.0	3.00	0.01

Table 4
Scenario 1: $n = 200$ and $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$.

p	H-BIC	C	O	U	NC	NI	X_1	X_2	X_3	X_4	X_5
$p = 100$	O-BIC	13.5	86.5	0.0	5.00	4.64	1.00	1.00	1.00	1.00	1.00
	H-BIC1	23.0	77.0	0.0	5.00	2.74	1.00	1.00	1.00	1.00	1.00
	H-BIC2	49.5	50.5	0.0	5.00	0.95	1.00	1.00	1.00	1.00	1.00
$p = 200$	O-BIC	18.0	82.0	0.0	5.00	5.65	1.00	1.00	1.00	1.00	1.00
	H-BIC1	35.0	65.0	0.0	5.00	2.14	1.00	1.00	1.00	1.00	1.00
	H-BIC2	57.0	43.0	0.0	5.00	0.75	1.00	1.00	1.00	1.00	1.00
$p = 400$	O-BIC	17.5	82.5	0.0	5.00	6.19	1.00	1.00	1.00	1.00	1.00
	H-BIC1	37.0	63.0	0.0	5.00	2.23	1.00	1.00	1.00	1.00	1.00
	H-BIC2	61.5	38.5	0.0	5.00	0.68	1.00	1.00	1.00	1.00	1.00

for Scenarios 1–3. The (positive) constant $C_n \to \infty$ is an adjustment constant for high dimension, i.e., it is increasing with n. Notice that H-BIC2 with $C_n = 1$ corresponds to O-BIC. In the simulations, we took $C_n = \log p$ as the practical choice of Lee et al. (2014a).

First, consider the setting with the coefficient "C1". This choice of coefficients is same as in Examples 1–2 of Zou and Li (2008). We set the sample size $n = 200$ and $p = 100, 200, 400$. Tables 1–3 summarize model selection results for the penalized estimator (2.5) in all the considered scenarios. They report the percentage (over $M = 200$ Monte Carlo replications) whether the final set $\hat{S}(\hat{\lambda})$ with the selected parameter $\hat{\lambda}$ (by the BIC) is correct fit (C), overfits (O), and underfits (U), respectively. And they show the averaged numbers of correctly and incorrectly selected covariates, i.e., $\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0,\ \beta_j \neq 0)$ and $\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0,\ \beta_j = 0)$, over 200 iterations, named NC and NIC, respectively. From the tables, it is easily seen that the use of O-BIC for our penalized estimator (2.5) results in overfitting in all the high dimensional scenarios, as also

Table 5
Scenario 2: $n = 200$ and $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$.

p	H-BIC	C	O	U	NC	NI	X_1	X_2	X_3	X_4	X_5
$p = 100$	O-BIC	17.5	33.5	49.0	4.47	1.08	1.00	1.00	1.00	0.91	0.57
	H-BIC1	25.5	8.0	66.5	4.19	0.22	1.00	1.00	0.96	0.86	0.37
	H-BIC2	8.0	0.0	92.0	3.30	0.01	1.00	0.94	0.76	0.49	0.12
$p = 200$	O-BIC	15.5	26.5	58.0	4.38	0.90	1.00	1.00	0.97	0.91	0.51
	H-BIC1	18.0	13.0	69.0	4.15	0.36	1.00	1.00	0.94	0.83	0.39
	H-BIC2	2.0	0.0	98.0	2.86	0.00	0.97	0.88	0.60	0.36	0.06
$p = 400$	O-BIC	11.5	21.5	67.0	4.23	0.94	1.00	1.00	0.97	0.83	0.44
	H-BIC1	9.5	2.0	88.5	3.67	0.07	1.00	1.00	0.85	0.63	0.19
	H-BIC2	0.0	0.0	100.0	2.35	0.00	0.90	0.76	0.47	0.19	0.02

Table 6
Scenario 3: $n = 200$ and $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$.

p	H-BIC	C	O	U	NC	NI	X_1	X_2	X_3	X_4	X_5
$p = 100$	O-BIC	52.0	41.0	7.0	4.93	0.73	1.00	1.00	1.00	0.99	0.94
	H-BIC1	64.0	11.5	24.5	4.75	0.16	1.00	1.00	1.00	0.98	0.77
	H-BIC2	48.5	2.0	49.5	4.42	0.03	1.00	1.00	0.95	0.92	0.55
$p = 200$	O-BIC	44.0	41.5	14.5	4.85	0.72	1.00	1.00	1.00	1.00	0.86
	H-BIC1	62.0	9.0	29.0	4.69	0.11	1.00	1.00	0.98	0.99	0.72
	H-BIC2	43.0	1.0	56.0	4.30	0.01	1.00	1.00	1.00	1.00	0.86
$p = 400$	O-BIC	34.5	50.5	15.0	4.85	43.11	1.00	1.00	1.00	0.98	0.88
	H-BIC1	57.0	9.0	34.0	4.64	5.89	1.00	1.00	0.99	0.96	0.70
	H-BIC2	39.0	0.5	60.5	4.21	0.01	1.00	1.00	1.00	0.83	0.45

Table 7
Scenario 1: $n = 400$ and $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$.

p	H-BIC	C	O	U	NC	NI	X_1	X_2	X_3	X_4	X_5
$p = 100$	O-BIC	20.5	79.5	0.0	5.00	2.21	1.00	1.00	1.00	1.00	1.00
	H-BIC1	24.5	75.5	0.0	5.00	1.70	1.00	1.00	1.00	1.00	1.00
	H-BIC2	42.5	57.5	0.0	5.00	0.93	1.00	1.00	1.00	1.00	1.00
$p = 200$	O-BIC	14.5	85.5	0.0	5.00	3.59	1.00	1.00	1.00	1.00	1.00
	H-BIC1	30.5	69.5	0.0	5.00	2.10	1.00	1.00	1.00	1.00	1.00
	H-BIC2	53.5	46.5	0.0	5.00	0.84	1.00	1.00	1.00	1.00	1.00
$p = 400$	O-BIC	16.5	83.5	0.0	5.00	5.12	1.00	1.00	1.00	1.00	1.00
	H-BIC1	36.5	63.5	0.0	5.00	2.08	1.00	1.00	1.00	1.00	1.00
	H-BIC2	61.5	38.5	0.0	5.00	0.56	1.00	1.00	1.00	1.00	1.00

Table 8
Scenario 2: $n = 400$ and $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$.

p	H-BIC	C	O	U	NC	NI	X_1	X_2	X_3	X_4	X_5
$p = 100$	O-BIC	27.0	62.5	10.5	4.90	1.55	1.00	1.00	1.00	0.99	0.91
	H-BIC1	55.0	30.0	15.0	4.85	0.46	1.00	1.00	1.00	0.98	0.87
	H-BIC2	63.5	0.02	34.5	4.62	0.02	1.00	1.00	1.00	0.94	0.69
$p = 200$	O-BIC	22.5	65.0	12.5	4.88	1.54	1.00	1.00	1.00	0.98	0.90
	H-BIC1	50.0	18.0	32.0	4.68	0.24	1.00	1.00	1.00	0.97	0.71
	H-BIC2	42.5	1.5	56.0	4.33	0.02	1.00	1.00	0.98	0.87	0.48
$p = 400$	O-BIC	23.0	58.0	19.0	4.81	1.36	1.00	1.00	1.00	0.99	0.83
	H-BIC1	54.0	11.5	34.5	4.64	0.18	1.00	1.00	1.00	0.95	0.70
	H-BIC2	34.5	0.0	65.5	4.13	0.00	1.00	1.00	0.96	0.78	0.39

reported in the literature (e.g., Chen & Chen, 2008; Lee et al., 2014a). In contrast, the penalized estimator (2.5) with penalty parameter selection by the high dimensional BICs seem to work quite well in terms of model selection.

For a further investigation we take settings with the coefficients "C2", where β_j for $1 \leq j \leq 5$ decay at a rate j^{-1}. Note that $\beta_5 = 0.6$, which means that the distinction problem of nonzero coefficient from zeros gets more difficult than in the case with "C1". We tried several values of n. Tables 4–11 describe how the penalized estimator (2.5) works. In addition to the previously reported numbers, we present the proportions (over $M = 200$ replications) of the cases in which X_j is selected, i.e., $\hat{\beta}_j \neq 0$ $(1 \leq j \leq 5)$. With small sample sizes, the performance of the estimator seems not good as in "C1". Additionally, it was observed that how it behaves depend on the type of loss ρ, i.e., which scenario, quite much. When the sample size is relatively small, the estimator (2.5) tends to overfit (even with the high dimensional BICs) in Scenario 1 (LS regression), whereas the proportion of underfit gets inflated in the other Scenarios 2–3 (Logistic and Quantile regression). However, a high dimensional BIC behaves pretty well over all the scenarios as the sample size increases, for example, see Tables 10 and

Table 9
Scenario 3: $n = 400$ and $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$.

p	H-BIC	C	O	U	NC	NI	X_1	X_2	X_3	X_4	X_5
$p = 100$	O-BIC	68.5	31.5	0.0	5.00	0.47	1.00	1.00	1.00	1.00	1.00
	H-BIC1	88.0	11.0	1.0	4.99	0.14	1.00	1.00	1.00	0.99	0.87
	H-BIC2	90.0	0.00	10.0	4.90	0.00	1.00	1.00	1.00	1.00	0.91
$p = 200$	O-BIC	66.5	33.5	0.0	5.00	0.55	1.00	1.00	1.00	1.00	1.00
	H-BIC1	91.5	6.5	2.0	4.98	0.07	1.00	1.00	1.00	1.00	0.99
	H-BIC2	82.5	0.0	17.5	4.83	0.00	1.00	1.00	1.00	0.99	0.84
$p = 400$	O-BIC	66.0	34.0	1.0	5.00	0.48	1.00	1.00	1.00	1.00	1.00
	H-BIC1	84.0	14.0	3.0	4.98	0.16	1.00	1.00	1.00	1.00	0.98
	H-BIC2	80.0	0.0	20.0	4.80	0.00	1.00	1.00	1.00	1.00	0.79

Table 10
Scenario 1: $n = 1000$ and $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$.

p	H-BIC	C	O	U	NC	NI	X_1	X_2	X_3	X_4	X_5
$p = 100$	O-BIC	65.0	35.0	0.0	5.00	0.58	1.00	1.00	1.00	1.00	1.00
	H-BIC1	85.5	14.5	0.0	5.00	0.16	1.00	1.00	1.00	1.00	1.00
	H-BIC2	97.5	2.5	0.0	5.00	0.03	1.00	1.00	1.00	1.00	1.00
$p = 200$	O-BIC	61.5	38.5	0.0	5.00	0.59	1.00	1.00	1.00	1.00	1.00
	H-BIC1	86.5	13.5	0.0	5.00	0.03	1.00	1.00	1.00	1.00	1.00
	H-BIC2	97.0	3.0	0.0	5.00	0.03	1.00	1.00	1.00	1.00	1.00
$p = 400$	O-BIC	65.0	35.0	0.0	5.00	0.55	1.00	1.00	1.00	1.00	1.00
	H-BIC1	85.0	15.0	0.0	5.00	0.16	1.00	1.00	1.00	1.00	1.00
	H-BIC2	93.0	7.0	0.0	5.00	0.08	1.00	1.00	1.00	1.00	1.00

Table 11
Scenario 2: $n = 1000$ and $\beta_j = 3/j$ for $j = 1, 2, \ldots, 5$ and $\beta_j = 0$ for $j > 5$.

p	H-BIC	C	O	U	NC	NI	X_1	X_2	X_3	X_4	X_5
$p = 100$	O-BIC	20.5	79.0	0.5	5.00	1.87	1.00	1.00	1.00	1.00	0.99
	H-BIC1	38.5	61.0	0.5	5.00	0.91	1.00	1.00	1.00	1.00	0.99
	H-BIC2	86.5	12.0	1.5	4.99	0.13	1.00	1.00	1.00	1.00	0.99
$p = 200$	O-BIC	15.5	84.0	0.5	5.00	1.80	1.00	1.00	1.00	1.00	1.00
	H-BIC1	37.0	62.5	0.5	5.00	0.82	1.00	1.00	1.00	1.00	1.00
	H-BIC2	92.5	2.5	5.0	5.00	0.82	1.00	1.00	1.00	1.00	1.00
$p = 400$	O-BIC	21.5	78.0	0.5	5.00	1.94	1.00	1.00	1.00	1.00	1.00
	H-BIC1	47.5	50.5	2.0	4.98	0.67	1.00	1.00	1.00	1.00	0.90
	H-BIC2	93.0	1.5	5.5	4.95	0.02	1.00	1.00	1.00	1.00	0.95

11 when $n = 1000$ (for Scenarios 1–2) and Table 9 when $n = 400$ (for Scenarios 3). In Scenario 3, we do not report results with $n = 1000$ in the paper because it works well enough when $n = 400$. But, this good performance might depend on a choice of BIC. Note that even if considering a type, e.g., H-BIC2, one should choose a value of C_n in practice though a wide range of C_n is known to work in theory. A good choice seems to be dependent on the setting, as also seen in Lee et al. (2014a). This is an important issue for good finite sample results of the estimator (2.5) and it deserves to a future investigation.

Acknowledgments

Research of Eun Ryung Lee and Jinwoo Cho was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2016R1C1B1011874).

References

Bickel, P., Ritov, Y., & Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, *37*, 1705–1732.
Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.
Candès, E., & Tao, T. (2005). Decoding by linear programming. *IEEE Transaction on Information Theory*, *59*, 1207–1223.
Candès, E., & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*, 2313–2351.
Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*, 759–771.
Chen, J., & Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica*, *22*, 555–574.
Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its Oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.
Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*, 101–148.
Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*, 928–961.
Kim, Y., Choi, H., & Oh, H. S. (2012). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, *103*, 1665–1673.
Kim, Y., & Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, *99*, 315–325.

Kim, Y., Kwon, S., & Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research (JMLR)*, *13*, 1037–1057.
Kwon, S., & Kim, Y. (2012). Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica*, *22*, 629–653.
Lee, E. R., Noh, H., & Park, B. U. (2014a). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 216–229.
Lee, E. R., Noh, H., & Park, B. U. (2014b). Supplement to "Model selection via Bayesian information criterion for quantile regression models". *Journal of the American Statistical Association*, 216–229.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *B58*, 267–288.
Tsybakov, A. B., & van de Geer, S. A. (2005). Square root penalty: Adaptation to the margin in classification and in edge estimation. *The Annals of Statistics*, *33*, 1203–1224.
van de Geer, S. (2007). The deterministic lasso. In *JSM proceedings*. American Statistical Association.
van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, *36*, 614–645.
van de Geer, S., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, *3*, 1360–1392.
Wang, L., Kim, Y., & Li, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *The Annals of Statistics*, *41*, 2505–2536.
Wang, H., & Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, *102*, 1418–1429.
Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society B*, *71*, 671–683.
Wang, H., Li, R., & Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*, 553–568.
Wang, L., Wu, Y., & Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, *107*, 214–222.
Zhang, (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.
Zhang, Y., Li, R., & Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*, 312–323.
Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research (JMLR)*, *2541–2563*.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.
Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, *36*, 1509–1533.