The Annals of Statistics
2016, Vol. 44, No. 6, 2467–2496
DOI: 10.1214/15-AOS1413

CONSISTENT MODEL SELECTION CRITERIA FOR QUADRATICALLY SUPPORTED RISKS

BY YONGDAI KIM[1] AND JONG-JUNE JEON

Seoul National University and University of Seoul

In this paper, we study asymptotic properties of model selection criteria for high-dimensional regression models where the number of covariates is much larger than the sample size. In particular, we consider a class of loss functions called the *class of quadratically supported risks* which is large enough to include the quadratic loss, Huber loss, quantile loss and logistic loss. We provide sufficient conditions for the model selection criteria, which are applicable to the class of quadratically supported risks. Our results extend most previous sufficient conditions for model selection consistency. In addition, sufficient conditions for pathconsistency of the Lasso and nonconvex penalized estimators are presented. Here, pathconsistency means that the probability of the solution path that includes the true model converges to 1. Pathconsistency makes it practically feasible to apply consistent model selection criteria to high-dimensional data. The data-adaptive model selection procedure is proposed which is selection consistent and performs well for finite samples. Results of simulation studies as well as real data analysis are presented to compare the finite sample performances of the proposed data-adaptive model selection criterion with other competitors.

1. Introduction. High-dimensional data, where the number of covariates greatly exceeds the sample size, arise frequently in modern applications in biology, chemometrics, economics, neuroscience and other scientific fields. To facilitate analysis, it is often useful and reasonable to assume that only a small number of covariates are relevant for modeling the response variable. In this situation, model selection is a fundamental and important task. For high dimensional data, however, classical model selection criteria such as the Akaike information criterion or AIC [1], Bayesian information criterion or BIC [25] and cross validation or generalized cross validation [8, 27] are known to select too many variables than necessary. See, for example, [3] and [5].

Various information criteria such as the modified BIC of [28], extended BIC of [6], corrected risk inflation criterion (CRIC) of [32], generalized information criterion (GIC) of [18] and the high dimensional BIC (HBIC) of [29] have been

Received April 2015; revised November 2015.

[1]Supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A2A2A01004496 and NRF-2014R1A4A1007895).

MSC2010 subject classifications. 60K35.

Key words and phrases. Generalized information criteria, high dimension, model selection, quadratically supported risks, selection consistency.

proposed and proven to be consistent for high dimensional data. Here, the consistency of a model selection criterion means that the probability of the selected model being equal to the true model converges to 1 (see Section 2 for a rigorous definition).

However, most of the aforementioned results for selection consistency use the quadratic loss with the (sub)-Gaussian error distribution, and hence the results are not applicable to other problems such as quantile regression, robust regression and generalized linear models. In this vein, Lee, Noh and Park [20] proposed the extended BIC for quantile regression and proved selection consistency, while Chen and Chen [7] provided sufficient conditions for the GIC to be consistent for generalized linear models.

In this paper, we propose a unified framework for selection consistency that can be applied to various regression models including the linear regression with (sub)-Gaussian errors, generalized linear regression, robust regression and quantile regression. We consider a class of loss functions called *quadratically supported risks* (QSR). This class of loss functions includes all those loss functions used in the aforementioned regression models. We then provide a set of sufficient conditions for a given GIC to be consistent for high dimensional models. Our sufficient conditions are general enough to cover and extend most the previous results of selection consistency.

One problem with using the GIC for high dimensional models is computation since calculating the GIC values for all possible submodels is almost impossible. In practice, one may find a solution path of a penalized estimator such as the Lasso (least absolute shrinkage and selection operator) or SCAD estimator, and apply the GIC for submodels that corresponds to the solution path. This approach is consistent if the solution path includes the true model and the GIC is consistent. Pathconsistency (i.e., the probability that the solution path includes the true model converges to 1) for linear models is well known. For instance, Bühlmann and van de Geer [4] proved that the thresholded Lasso estimator is path-consistent. Zhang [31] and Kim, Kwon and Choi [18] also showed the pathconsistency of a nonconvex penalized least square estimator, while Fan and Tang [11] studied the pathconsistency of a nonconvex penalized maximum likelihood estimator. In this paper, we prove the pathconsistency of certain thresholded penalized estimators with loss functions in the class of QSR.

It turns out that the class of consistent GICs is large and finite sample performances are quite different. Thus, it is important to choose a GIC that works well with moderate sample sizes. We propose a data-adaptive model selection procedure which is selection consistent and performs well for finite samples.

The remainder of this paper is organized as follows. In Section 2, we introduce the GIC and propose the class of QSR. In Section 3, we provide sufficient conditions for a given GIC to be consistent with the class of QSR. In Section 4, by using the sufficient conditions given in Section 3, we prove that the GIC is consistent for several loss functions including the quadratic loss, logistic loss, Huber loss

and quantile loss functions. In Section 5, we explain how to use a solution path of a penalized estimator such as the Lasso and nonconvex penalized estimators for model selection and offer theoretical justifications. A data-adaptive model selection procedure is proposed in Section 6. Simulation results are given in Section 7 and concluding remarks follow in Section 8.

2. GICs and QSR. Let $\mathcal{L} = \{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$ be a given data set of pairs of response and covariates, where $y_i \in R$ and $\mathbf{x}_i \in R^{p_n}$. For a given loss function $l : R \times R \to [0, \infty)$, we consider estimating the regression coefficient β by minimizing the risk function $R_n(\beta) = \frac{1}{n}\sum_{i=1}^n l(y_i, \mathbf{x}_i'\beta)$. Let β^* be the true regression coefficient vector. Suppose $\beta_j^* \neq 0$ for $j \leq q_n$ and $\beta_j^* = 0$ for $j > q_n$.

In this paper, we are concerned with model selection problems when p_n is much larger than the sample size n. When p_n is large, it would not be feasible to search all possible subsets of $\{1, \ldots, p_n\}$. Instead, we set an upper bound on the number of covariates in the candidate submodels, say s_n, and search the optimal model among the candidate submodels that have no more than s_n covariates. Chen and Chen [6] and Kim, Kwon and Choi [18] considered a similar model selection procedure. Let $|\beta|_0 = \sum_{j=1}^{p_n} I(\beta_j \neq 0)$ and $\mathcal{M}_{s_n} = \{\beta \in R^{p_n} : |\beta|_0 \leq s_n\}$.

For a given subset π of $\{1, \ldots, p_n\}$, let

$$\hat{\beta}_\pi = \operatorname*{argmin}_{\beta:\beta_j=0,\, j\in\pi^c} R_n(\beta).$$

A sequence of positive numbers $\{\lambda_n\}$ is called GIC_{λ_n}, if it gives a sequence of random subsets $\hat{\pi}_{\lambda_n}$ defined as

$$\hat{\pi}_{\lambda_n} = \operatorname*{argmin}_{\pi\subset\{1,\ldots,p_n\},|\pi|\leq s_n} R_n(\hat{\beta}_\pi) + \lambda_n|\pi|,$$

where $|\pi|$ is the cardinality of π. When the loss is the quadratic loss, that is, $l(y, \mathbf{x}'\beta) = (y - \mathbf{x}'\beta)^2$, the AIC corresponds to $\lambda_n = 2/n$; the BIC to $\lambda_n = \log n/(n)$; the RIC of Foster and George [12] to $\lambda_n = \log p_n/n$; and the RIC of Zhang and Shen [32] to $\lambda_n = (\log p_n + \log\log p_n)/n$; Shao [26] studied the asymptotic properties of the GIC focusing on the AIC and BIC.

We say that the GIC_{λ_n} is consistent if

$$\Pr(\hat{\pi}_{\lambda_n} = \pi^*) \to 1$$

as $n \to \infty$, where $\pi^* = \{1, \ldots, q_n\}$. Kim et al. [18] provided sufficient conditions for the consistency of the GIC when the quadratic loss is used. The aim of this paper is to prove the consistency of the GIC for a wide class of loss functions including the logistic loss, Huber loss, quantile loss and the quadratic loss functions.

We say that the risk function $R_n(\beta)$ is quadratically supported if there exist sequences of p_n-dimensional random vectors $\{a_n\}$ and $\{\tilde{\beta}_n\}$, sequences of nonnegative random variables $\{\delta_n\}$ and $\{\eta_n\}$ and a positive real number b such that

$\Pr(\mathcal{A}_n) \to 1$ as $n \to \infty$, where

$$\mathcal{A}_n = \Big\{ R_n(\beta) - R_n(\tilde{\beta}_n) \geq a_n'(\beta - \tilde{\beta}_n) + \frac{b}{2}\|\beta - \tilde{\beta}_n\|^2 - \delta_n|\beta|_0 \text{ for all } \beta \in \Theta_n \Big\} \tag{2.1}$$

and $\Theta_n = \mathcal{M}_{s_n} \cap \{\beta : \|\beta - \tilde{\beta}_n\| \leq \eta_n\}$. Here, $\|\cdot\|$ is the Euclidean norm. We say that a given loss belongs to the QSR if the corresponding risk is quadratically supported. In Section 3, we provide sufficient conditions for the consistency of the GIC when the risk is quadratically supported, while in Section 4, we show that various loss functions such as the quadratic loss, logistic loss, Huber loss and quantile loss functions are in the class of the QSR.

Let $\hat{\beta}^o = \hat{\beta}_{\pi^*}$ be the oracle estimator. Suppose $\tilde{\beta}_n = \hat{\beta}^o$, which is a typical choice in many cases. Condition (2.1) for the QSR essentially means that the risk function around the oracle estimator behaves like a quadratic function asymptotically. An interesting result is that a loss function which is linear around $\hat{\beta}^o$ (e.g., the quantile loss) can also belong to the QSR.

For a given $p_n \times p_n$ symmetric matrix $\mathbf{A}$ and a given $\pi \subset \{1, \ldots, p_n\}$, let $\mathbf{A}_\pi$ be the $|\pi| \times |\pi|$ sub-matrix of $\mathbf{A}$ formed by those rows and columns of $\mathbf{A}$ whose indices are in π. Similarly, for a given p_n dimensional vector $\mathbf{v}$, let $|\mathbf{v}|_\infty = \max_j |v_j|$ and $\mathbf{v}_\pi = (v_j, j \in \pi)$. For a given β, let $\sigma(\beta) = \{j : \beta_j \neq 0\}$.

Let $\mathbf{X}$ be the $n \times p_n$ dimensional matrix whose jth column is $X^j = (x_{1j}, \ldots, x_{nj})'$ with $\|X^j\|^2 = n$. We assume that there exist positive constants ρ_* and ρ^* such that

$$\rho_* \leq \min_{\pi : |\pi| \leq 2s_n} \lambda_{\min}\{(\mathbf{X}'\mathbf{X}/n)_\pi\} \leq \max_{\pi : |\pi| \leq 2s_n} \lambda_{\max}\{(\mathbf{X}'\mathbf{X}/n)_\pi\} \leq \rho^*,$$

where $\lambda_{\min}\{(\mathbf{X}'\mathbf{X}/n)_\pi\}$ and $\lambda_{\max}\{(\mathbf{X}'\mathbf{X}/n)_\pi\}$ are the smallest and largest eigenvalues of $(\mathbf{X}'\mathbf{X}/n)_\pi$, respectively. This assumption is called the sparse Riesze condition (SRC), which a standard one for model selection with high dimensional models (see, e.g., [6] and [18]).

3. Sufficient conditions for the consistency of the GIC. In this section, we prove that $\hat{\pi}_{\lambda_n}$ is consistent under the following regularity conditions. For given two sequences $\{u_n\}$ and $\{v_n\}$ of positive real numbers, we write $u_n \gg v_n$ if $u_n/v_n \to \infty$ as $n \to \infty$. To simplify notation, we assume hereafter that all the equalities, inequalities and convergences are understood to hold for all sufficiently large n in probability whenever random quantities are involved. For example, $\lambda_n - |a_n|_\infty^2/2 \geq \lambda_n/2$ means that $\Pr\{\lambda_n - |a_n|_\infty^2/2 \geq \lambda_n/2\} \to 1$ as $n \to \infty$.

(C1) $\sigma(\tilde{\beta}_n) = \{1, \ldots, q_n\}$.
(C2) $\lambda_n \gg |a_n|_\infty^2$.
(C3) $\min_{j \in \pi^*} |\tilde{\beta}_{nj}| \gg \sqrt{\lambda_n}$.

(C4) $\lambda_n \gg q_n \max\{|a_{n,\pi^*}|_\infty^2, \delta_n\}$.
(C5) $s_n \geq q_n$.
(C6) $\eta_n^2 \gg \lambda_n s_n$.

The key conditions are (C2) and (C3), which require that $|a_n|_\infty$ is sufficiently small and $\min_{j\in\pi^*}|\tilde{\beta}_{nj}|$ is sufficiently large. Most of the conditions for the consistency of the GIC for various loss functions such as in [6, 7, 18, 20] satisfy (C2) and (C3). The other conditions are easily satisfied, in particular when $q_n = O(1)$.

LEMMA 1. *When* $\|\beta - \tilde{\beta}_n\| > \eta_n$ *with* $|\beta|_0 \leq s_n$,

$$R_n(\beta) - R_n(\tilde{\beta}_n) \geq \frac{b}{16}\eta_n^2$$

on $\mathcal{A}_n$ *under the regularity conditions*.

PROOF. For a given $\beta \in \mathcal{M}_{s_n} \cap \{\beta : \|\beta - \tilde{\beta}_n\| > \eta_n\}$, let $\beta_h = \tilde{\beta}_n + h(\beta - \tilde{\beta}_n)$, and let $\phi(h) = R_n(\beta_h) - R_n(\tilde{\beta}_n)$. Note that (C4) and (C6) implies $\eta_n \gg \max\{|a_n|_\infty\sqrt{s_n}, \sqrt{\delta_n s_n}\}$. Hence, when $\eta_n/(2\|\beta - \tilde{\beta}_n\|) \leq h \leq \eta_n/\|\beta - \tilde{\beta}_n\|$, we have

$$\begin{aligned}\phi(h) &\geq -|a_n|_\infty\sqrt{s_n}\|\beta_h - \tilde{\beta}_n\| + \frac{b}{2}\|\beta_h - \tilde{\beta}_n\|^2 - \delta_n|\beta_h|_0 \\ &\geq -|a_n|_\infty\sqrt{s_n}\eta_n + \frac{b}{8}\eta_n^2 - \delta_n s_n \\ &\geq \frac{b}{16}\eta_n^2.\end{aligned}$$

Since ϕ is convex, when $\|\beta - \tilde{\beta}_n\| > \eta_n$, we have

$$R_n(\beta) - R_n(\tilde{\beta}_n) \geq R_n(\beta_h) - R_n(\tilde{\beta}_n) \geq \frac{b}{16}\eta_n^2 > 0$$

for $h = \eta_n/\|\beta - \tilde{\beta}_n\|$. □

THEOREM 1. *Under the regularity conditions*,

$$\Pr(\hat{\pi}_{\lambda_n} = \pi^*) \to 1.$$

PROOF. Let

$$\hat{\beta} = \operatorname*{argmin}_{\beta\in\mathcal{M}_{s_n}} R_n(\beta) + \lambda_n|\beta|_0.$$

When $\beta \in \mathcal{M}_{s_n} \cap \{\beta : \|\beta - \tilde{\beta}_n\| > \eta_n\}$, Lemma 1 and condition (C6) imply

$$R_n(\beta) + \lambda_n|\beta|_0 - R_n(\tilde{\beta}_n) - \lambda_n|\tilde{\beta}_n|_0 \geq \frac{b}{16}\eta_n^2 - \lambda_n q_n > 0$$

on $\mathcal{A}_n$, and hence $\|\hat{\beta} - \tilde{\beta}_n\| \leq \eta_n$ on $\mathcal{A}_n$.

To complete the proof, we show that

$$\Pr\bigl(\sigma(\hat{\beta}) = \pi^*\bigr) \to 1 \tag{3.1}$$

as $n \to \infty$. Let $Q(\beta) = \sum_j w_j$ where

$$w_j = a_{nj}(\beta_j - \tilde{\beta}_{nj}) + b(\beta_j - \tilde{\beta}_{nj})^2/2 + \lambda_n\bigl\{I(\beta_j \neq 0) - I(\tilde{\beta}_{nj} \neq 0)\bigr\} - \delta_n|\beta_j|_0.$$

To prove (3.1), it suffices to show that $Q(\beta) > 0$ on $\mathcal{A}_n$ unless $\sigma(\beta) = \pi^*$. When $\beta_j = 0$ and $\tilde{\beta}_{nj} = 0$, then $w_j = 0$. When $\beta_j \neq 0$ and $\tilde{\beta}_{nj} = 0$,

$$w_j \geq -\frac{|a_n|_\infty^2}{2b} + \lambda_n - \delta_n \geq \lambda_n/2,$$

due to (C2) and (C4). When $\beta_j = 0$ and $\tilde{\beta}_{nj} \neq 0$, then

$$w_j \geq -|a_n|_\infty|\tilde{\beta}_{nj}| + \frac{b}{2}\tilde{\beta}_{nj}^2 - \lambda_n \geq \frac{b}{2}\lambda_n\left\{1 - \frac{|a_n|_\infty}{\sqrt{\lambda_n}} - \frac{\lambda_n}{\tilde{\beta}_{nj}^2}\right\} \geq \frac{b}{4}\lambda_n > 0,$$

due to (C2) and (C3). When $\beta_j \neq 0$ and $\tilde{\beta}_{nj} \neq 0$, then $w_j > -a_{nj}^2/2b - \delta_n$. To sum up,

$$Q(\beta) \geq \bigl|\sigma(\beta) \cap \pi^{*c}\bigr|\lambda_n/2 + \bigl|\sigma(\beta)^c \cap \pi^*\bigr|\frac{b\lambda_n}{4} - q_n\left(\frac{|a_{n,\pi^*}|_\infty^2}{2b} + \delta_n\right).$$

Since $\lambda_n \gg q_n \max\{|a_{n,\pi^*}|_\infty^2, \delta_n\}$ according to (C4), $Q(\beta) > 0$ unless $\sigma(\beta) = \pi^*$, which completes the proof. □

4. Examples. In this section, we show that various loss functions belong to the QSR and the corresponding GICs are consistent. Let $\zeta_n = \max_{ij}|x_{ij}|$.

4.1. *Smooth losses*. Suppose that $R_n(\beta)$ has the first and second derivatives denoted by $R_n^{(1)}(\beta) = \partial R_n(\beta)/\partial\beta$ and $R_n^{(2)}(\beta) = \partial^2 R_n(\beta)/\partial\beta\partial\beta'$, respectively. If we let $\tilde{\beta}_n = \hat{\beta}^o$, Taylor's expansion yields that

$$R_n(\beta) - R_n\bigl(\hat{\beta}^o\bigr) = R_n^{(1)'}\bigl(\hat{\beta}^o\bigr)\bigl(\beta - \hat{\beta}^o\bigr) + \tfrac{1}{2}\bigl(\beta - \hat{\beta}^o\bigr)' R_n^{(2)}\bigl(\beta^\dagger\bigr)\bigl(\beta - \hat{\beta}^o\bigr) \tag{4.1}$$

for some $\beta^\dagger = (\beta_1^\dagger, \ldots, \beta_{p_n}^\dagger)'$ with $\beta_j^\dagger \in (\min\{\beta_j, \hat{\beta}_j^o\}, \max\{\beta_j, \hat{\beta}_j^o\})$. Hence, the loss belongs to the QSR with $a_n = R^{(1)}(\hat{\beta}^o)$, $b = \min_{\beta\in\Theta_n}\lambda_{\min}\{R^{(2)}(\beta^\dagger)_{\sigma(\beta)\cup\pi^*}\}$ and $\delta_n = 0$, provided $b > 0$.

4.1.1. *Linear regression with the quadratic loss*. Suppose that the true model is $y_i = \mathbf{x}_i'\beta^* + \varepsilon_i$, where ε_i are independent random variables whose common distribution has a sub-Gaussian tail. That is, there is some $\nu > 0$ such that for every $t \in R$ one has

$$\mathrm{E}\bigl(e^{t\varepsilon_i}\bigr) \leq e^{\nu^2 t^2}/2,$$

which implies that there exist positive constants c_ε and d_ε such that

$$\Pr\left(\left|\sum_{i=1}^n a_i\varepsilon_i\right| > t\right) \le c_\varepsilon \exp\left(-\frac{d_\varepsilon t^2}{\sum_{i=1}^n a_i^2}\right) \tag{4.2}$$

for all $(a_1,\dots,a_n)' \in R^n$ and $t>0$. When $l(y,\mathbf{x}'\beta) = (y-\mathbf{x}'\beta)^2/2$, we have $\tilde{\beta}_n = \hat{\beta}^o$, $a_{nj} = X^{j'}(Y-\hat{Y})$, $b=\rho_*$, $\delta_n = 0$ and $\eta_n = \infty$, where $\hat{Y} = (\mathbf{x}_i'\hat{\beta}^o,\dots,\mathbf{x}_n'\hat{\beta}^o)'$. While $a_{nj}=0$ for $j\in\pi^*$, for $j\in\pi^{*c}$ it is proven in the proof of Theorem 2 of Kim et al. [18] that there exist $\mathbf{h}_{nj}\in R^n$ such that

$$a_{nj} = \mathbf{h}_{nj}'\boldsymbol{\varepsilon}/n \tag{4.3}$$

for all j and $\sup_j \|\mathbf{h}_{nj}\|^2 \le n$, where $\boldsymbol{\varepsilon} = (\varepsilon_1,\dots,\varepsilon_n)'$. Hence, (4.2) implies that for any $\kappa_n\to\infty$

$$\begin{aligned}
&\Pr\Big\{\max_{j\in\pi^{*c}}|a_{nj}| > \sqrt{\kappa_n \log p_n/(d_\varepsilon n)}\Big\}\\
&\quad\le \sum_{j\in\pi^{*c}} \Pr\big\{|a_{nj}| > \sqrt{\kappa_n\log p_n/(d_\varepsilon n)}\big\}\\
&\quad\le c_\varepsilon \exp\big(-(\kappa_n-1)\log p_n\big)\to 0,
\end{aligned}$$

and so we have $|a_n|_\infty = O_p(\sqrt{\log p_n/n})$. Therefore, the GIC_{λ_n} with $\lambda_n \gg \log p_n/n$ is consistent provided $\min_{j\in\pi^*}|\hat{\beta}_j^o| \gg \sqrt{\lambda_n}$. This result coincides with that of [18]. For example, when $p_n = \exp(an^d)$ for $0\le d<1$, the GIC with $\lambda_n = n^{c-1}$ for $d<c<1$ is consistent provided $\min_{j\in\pi^*}|\beta_j^*| \gg n^{(c-1)/2}$.

4.1.2. *Logistic regression.* When $y\in\{0,1\}$, let $l(y,\mathbf{x}'\beta)$ be the logistic loss defined as

$$l\big(y,\mathbf{x}'\beta\big) = -y\mathbf{x}'\beta + \log\big(1+\exp(\mathbf{x}'\beta)\big),$$

which is the negative log-likelihood of the logistic regression model. Suppose that the true regression model is

$$\Pr(y_i=1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i'\beta^*)}{1+\exp(\mathbf{x}_i'\beta^*)}. \tag{4.4}$$

Note that $R_n^{(2)}(\beta) = \mathbf{X}'\mathbf{W}(\beta)\mathbf{X}/n$, where $\mathbf{W}(\beta)$ is the $n\times n$ diagonal matrix whose diagonal elements are $p(\mathbf{x}_i'\beta)\{1-p(\mathbf{x}_i'\beta)\}$ with $p(a) = \exp(a)/(1+\exp(a))$. Suppose that there exists a constant $\rho>0$ such that

$$\min_{\pi:|\pi|\le 2s_n} \lambda_{\min}\big\{R_n^{(2)}\big(\hat{\beta}^o\big)_\pi\big\} > \rho \tag{4.5}$$

for all sufficiently large n. Note that $R_n^{(2)}(\beta^\dagger) - R_n^{(2)}(\hat{\beta}^o) = \mathbf{X}'\{\mathbf{W}(\beta^\dagger)-\mathbf{W}(\hat{\beta}^o)\}\mathbf{X}$. Taylor's expansion yields

$$\big|\mathbf{W}\big(\beta^\dagger\big)_{ii} - \mathbf{W}\big(\hat{\beta}^o\big)_{ii}\big| \le \big|\mathbf{x}_i'\big(\beta^\dagger-\hat{\beta}^o\big)\big| \le \zeta_n\sqrt{2s_n}\big\|\beta^\dagger-\hat{\beta}^o\big\|. \tag{4.6}$$

Hence, if $\|\beta^\dagger - \hat{\beta}^o\| \le \rho/(2\rho^*\zeta_n\sqrt{2s_n})$, then $\lambda_{\min}\{R_n^{(2)}(\beta^\dagger)_\pi\} > \rho/2$, where $\pi = \sigma(\beta) \cup \pi^*$. Thus, the risk is quadratically supported with $\tilde{\beta}_n = \hat{\beta}^o$, $a_n = R_n^{(1)}(\hat{\beta}^o)$, $b = \rho/2$, $\eta_n = \rho/(2\rho^*\zeta_n\sqrt{2s_n})$ and $\delta_n = 0$.

REMARK 1. A simple sufficient condition for (4.5) is that there exists $c > 0$ such that $\sup_i |\mathbf{x}_i'\hat{\beta}^o| \le c$ because $\min_i \mathbf{W}(\hat{\beta}^o)_{ii} > p(c)(1 - p(c))$ in such a case. Another sufficient condition is that there exist $c > 0$ and $\rho > 0$ such that

$$\min_{\pi:|\pi|\le s_n} \lambda_{\min}\left\{\left(\frac{1}{n}\sum_{i:|\mathbf{x}_i'\hat{\beta}^o|\le c} \mathbf{x}_i\mathbf{x}_i'\right)_\pi\right\} \ge \rho.$$

To derive the convergence rate of $|a_n|_\infty$, note that

$$R_n^{(1)}(\hat{\beta}^o)_j - R_n^{(1)}(\beta^*)_j = \frac{1}{n}\sum_{i=1}^n x_{ij}\{p(\mathbf{x}_i'\hat{\beta}^o) - p(\mathbf{x}_i'\beta^*)\}.$$

Let $\dot{p}(a) = dp(a)/da$. The mean value theorem implies there exists $c_i \in R$ such that

$$p(\mathbf{x}_i'\hat{\beta}^o) - p(\mathbf{x}_i'\beta^*) = \dot{p}(c_i)\mathbf{x}_i'(\hat{\beta}^o - \beta^*).$$

Since $|\dot{p}(c_i)| \le 1$, we have

$$\begin{aligned} &|R_n^{(1)}(\hat{\beta}^o)_j - R_n^{(1)}(\beta^*)_j| \\ &\quad\le \sqrt{(\hat{\beta}^o - \beta^*)'(\mathbf{X}'\mathbf{X}/n)(\hat{\beta}^o - \beta^*)} \le \sqrt{\rho^*}\|\hat{\beta}^o - \beta^*\|, \end{aligned} \tag{4.7}$$

where the first inequality is due to the Cauchy–Schwarz inequality. Moreover, we have

$$\mathrm{E}\{R_n^{(1)}(\beta^*)_j\} = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[x_{ij}\{y_i - p(\mathbf{x}_i'\beta^*)\}] = 0$$

and

$$\mathrm{E}\{R_n^{(1)}(\beta^*)_j^2\} = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[x_{ij}^2\{y_i - p(\mathbf{x}_i'\beta^*)\}^2] \le 1,$$

and hence Theorem 2 in the Appendix implies

$$\max_j |R_n^{(1)}(\beta^*)_j| = O(\sqrt{\log p_n/n}), \tag{4.8}$$

provided $\zeta_n^2 \log p_n/n \to 0$. Combining (4.7) and (4.8), we have

$$\max_{j\in\pi^*} |R_n^{(1)}(\hat{\beta}^o)_j| = O_p(\sqrt{\max\{\log p_n/n, q_n/n\}}),$$

provided $\|\hat{\beta}^o - \beta^*\| = O_p(\sqrt{q_n/n})$ (see, e.g., [10]). Therefore, the GIC_{λ_n} is consistent if

$$\lambda_n \gg \max\{\log p_n/n, q_n/n\}, \tag{4.9}$$

provided $\zeta_n^2 s_n^2 \lambda_n \to 0$ and $\min_{j\in\pi^*} |\beta_j^*| \gg \sqrt{\lambda_n}$. To sum up, we have proven the following theorem.

THEOREM 2. *Suppose that the true conditional distribution of y_i given $\mathbf{x}_i$ is (4.4) and (4.5) holds. Then the GIC_{λ_n} with $\lambda_n \gg \max\{\log p_n/n, q_n/n\}$ is consistent for the logistic loss provided that $\|\hat{\beta}^o - \beta^*\| = O_p(\sqrt{q_n/n})$, $\zeta_n^2 s_n^2 \lambda_n \to 0$ and $\min_{j\in\pi^*} |\hat{\beta}_j^o| \gg \lambda_n$.*

REMARK 2. For the consistency of GIC, ζ_n should be sufficiently small that $\zeta_n^2 s_n^2 \lambda_n \to 0$. For example, when $p_n > n$ and $\mathbf{x}_i$ are independent realizations of a Gaussian random vector with bounded variances, we have $\zeta_n = O_p(\sqrt{\log p_n})$, and hence the GIC $\lambda_n = \alpha_n \log p_n/n$ with $\alpha_n = \sqrt{n}/(s_n \log p_n)$ is consistent provided $(s_n \log p_n)^2/n \to 0$ and $\log p_n \geq q_n$ because $\zeta_n^2 s_n^2 \lambda_n \to 0$.

REMARK 3. The condition $\zeta_n^2 s_n^2 \lambda_n \to 0$ is required for the logistic loss but not for the quadratic loss. This difference comes partly from the fact that the Hessian matrix of the logistic loss depends on β.

Suppose $\zeta_n = O(1)$. Then the GIC_{λ_n} is consistent provided $\lambda_n \gg \max\{\log p_n, q_n\}/n, s_n^2 \log p_n/n \to 0$ and $\min_{j\in\pi^*} |\beta_j^*| \gg \sqrt{\lambda_n}$. For example, the GIC_{λ_n} with $\lambda_n = n^{b-1}$ for $0 < b < 1$ is consistent when $p_n = \exp(an^d)$ and $s_n = n^c$ for some positive constants a, c and d with $\max\{d, c\} < b < 1$ and $d + 2c < 1$ provided $\min_{j\in\pi^*} |\beta_j^*| \gg n^{(d-1)/2}$. This result extends the result of in [7] which requires $d + c < 1/3$.

4.2. *Huber loss.* The Huber loss, which is a robust version of the quadratic loss, is defined as

$$l_d(y, z) = \begin{cases} \frac{1}{2}(y - z)^2, & \text{for } |y - z| < d, \\ d(|y - z| - d/2), & \text{for } |y - z| \geq d \end{cases}$$

for some $d > 0$, where $z = \mathbf{x}'\beta$. Suppose that the true model is

$$y_i = \mathbf{x}_i'\beta^* + \varepsilon_i, \tag{4.10}$$

where ε_i are independent random variables whose distributions are symmetric at 0. Let $l_d^{(1)}(y, z) = \frac{\partial}{\partial z} l_d(y, z)$. Since $l_d(y, z)$ is convex,

$$l_d(y, z) \geq l_d(y, \hat{z}) + l_d^{(1)}(y, \hat{z})(z - \hat{z}), \tag{4.11}$$

where $\hat{z} = \mathbf{x}'\hat{\beta}^o$. Moreover, if $|y - z| < d$ and $|y - \hat{z}| < d$,

$$l_d(y, z) = l_d(y, \hat{z}) + l_d^{(1)}(y, \hat{z})(z - \hat{z}) + \tfrac{1}{2}(z - \hat{z})^2. \tag{4.12}$$

Let $J_\beta = \{i : |\mathbf{x}_i'(\beta - \beta^*) + \varepsilon_i| < d, |\mathbf{x}_i'(\hat{\beta}^o - \beta^*) + \varepsilon_i| < d\}$. Then (4.11) and (4.12) imply

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n l_d(y_i, \mathbf{x}_i'\beta) \\ &\quad \geq \frac{1}{n}\sum_{i=1}^n l_d(y_i, \mathbf{x}_i'\hat{\beta}^o) + \frac{1}{n}\sum_{i=1}^n l_d^{(1)}(y_i, \mathbf{x}_i'\hat{\beta}^o)\mathbf{x}_i'(\beta - \hat{\beta}^o) \\ &\qquad + \frac{1}{2n}\sum_{i \in J_\beta} (\beta - \hat{\beta}^o)'\mathbf{x}_i\mathbf{x}_i'(\beta - \hat{\beta}^o). \end{aligned} \tag{4.13}$$

Assume that $\|\hat{\beta}^o - \beta^*\| = O_p(\sqrt{q_n/n})$ (see, e.g., [22]) and $\zeta_n^2 s_n^2/n \to 0$. Since $|\mathbf{x}_i'(\beta - \beta^*)| \leq \zeta_n\sqrt{2s_n}\|\beta - \beta^*\|$ for $\beta \in \mathcal{M}_{s_n}$, we have $\sup_i |\mathbf{x}_i'(\hat{\beta}^o - \beta^*)| \to 0$. Similarly, we can show $\sup_i |\mathbf{x}_i'(e\beta - \beta^*)| \leq d/2$ if $\|\beta - \hat{\beta}^o\| < d/(4\zeta_n\sqrt{2s_n})$. Since $(\beta - \hat{\beta}^o)'\mathbf{x}_i\mathbf{x}_i'(\beta - \hat{\beta}^o) \geq 0$ for all i, the inequality

$$\begin{aligned} &\frac{1}{2n}\sum_{i \in J_\beta} (\beta - \hat{\beta}^o)'\mathbf{x}_i\mathbf{x}_i'(\beta - \hat{\beta}^o) \\ &\qquad \geq (\beta - \hat{\beta}^o)'\left\{\frac{1}{2n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' I\big(|\varepsilon_i| \leq d/2\big)\right\}(\beta - \hat{\beta}^o) \end{aligned}$$

holds when $\beta \in \mathcal{M}_{s_n}$ with $\|\beta - \hat{\beta}^o\| < d/(4\zeta_n\sqrt{2s_n})$. Let $\mathbf{A} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' I(|\varepsilon_i| < d/2)$. When $\zeta_n^2 \log p_n/n \to 0$, the Bernstein inequality implies

$$\sup_{jk}\left|\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}\big\{I\big(|\varepsilon_i| < d/2\big) - P\big(|\varepsilon_i| < d/2\big)\big\}\right| = O_p(\zeta_n\sqrt{\log p_n/n}),$$

which in turn implies that the smallest eigenvalue of $\mathbf{A}_\pi$ is no less than $\Pr(|\varepsilon_1| > d/2)\rho_* - O_p(\zeta_n s_n\sqrt{\log p_n/n})$ for all $\pi \subset \{1, \ldots, p_n\}$ with $|\pi| \leq 2s_n$. Hence, when $\zeta_n^2 s_n^2 \log p_n/n \to 0$, there exists $\rho > 0$ such that

$$(\beta - \hat{\beta}^o)'\left\{\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik} I\big(|\varepsilon_i| < d/2\big)\right\}(\beta - \hat{\beta}^o) \geq \rho\|\beta - \hat{\beta}^o\|^2.$$

Therefore, the risk is quadratically supported with $\tilde{\beta}_n = \hat{\beta}^o$, $a_n = \sum_{i=1}^n l_d^{(1)}(y_i, \mathbf{x}_i'\hat{\beta}^o)\mathbf{x}_i'/n$, $\delta_n = 0$, $\eta_n = d/(4\zeta_n\sqrt{2s_n})$ and $b = \rho$ provided that $\zeta_n^2 s_n^2 \log p_n/n \to 0$.

To derive the convergence rate of $|a_n|_\infty$, note that $|l_d^{(1)}(y,\hat{z}) - l_d^{(1)}(y,z^*)| \le |\hat{z} - z^*|$, and hence we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}(l_d^{(1)}(y_i,\mathbf{x}_i'\hat{\beta}^o) - l_d^{(1)}(y_i,\mathbf{x}_i'\beta^*))x_{ij}\right| \\ \le \frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i'(\hat{\beta}^o-\beta^*)x_{ij}| \\ \le \sqrt{\rho^*}\|\hat{\beta}^o-\beta^*\| = O_p(\sqrt{q_n/n})\end{aligned} \tag{4.14}$$

similarly to (4.7). In turn, since the distributions of ε_i are symmetric at 0, $\mathrm{E}(l_d^{(1)}(y_i,\mathbf{x}_i'\beta^*)) = 0$. In addition, $\sum_{i=1}^n x_{ij}^2 = n$ and $l_d^{(1)}(y_i,\mathbf{x}_i'\beta^*)$ are bounded and thus Proposition 2 in the Appendix implies that

$$\max_{j\in\pi^*}\left|\sum_{i=1}^{n} l_d^{(1)}(y_i,\mathbf{x}_i'\beta^*)x_{ij}\Big/n\right| = O_p(\sqrt{\log p_n/n}), \tag{4.15}$$

provided $\zeta_n^2\log p_n/n \to 0$. Hence, the GIC_{λ_n} is consistent if $\lambda_n \gg \max\{\log p_n, q_n\}/n$ provided $\zeta_n^2 s_n^2\lambda_n \to 0$ and $\min_{j\in\pi^*}|\hat{\beta}^o| \gg \sqrt{\lambda_n}$. These conditions are the same as those for the logistic regression. We have proven the following theorem.

THEOREM 3. *Suppose that the true conditional distribution of y_i given $\mathbf{x}_i$ is* (4.10)*, where ε_i are independent random variables whose distributions are symmetric at* 0*. Then the GIC_{λ_n} with $\lambda_n \gg \max\{\log p_n, q_n\}/n$ is consistent for the Huber loss provided $\|\hat{\beta}^o - \beta^*\| = O_p(\sqrt{q_n/n})$, $\zeta_n^2 s_n^2\lambda_n \to 0$ and $\min_j|\hat{\beta}^o| \gg \lambda_n$.*

4.3. *Quantile regression.* Quantile loss is given as $l(y,\mathbf{x}'\beta) = \rho_\tau(y-\mathbf{x}'\beta)$, where $\rho_\tau(u) = u(2\tau - 2I(u<0))$. The quantile loss is popularly used for quantile regression (see, e.g., [30] and references therein).

Suppose that the true model is $y_i = \mathbf{x}_i'\beta^* + \varepsilon_i$, where ε_i are independent random variables with $\Pr(\varepsilon_i \le 0) = \tau$. Under this model, $\mathbf{x}_i'\beta^*$ is the conditional τth quantile of y_i given $\mathbf{x}_i$. Let F_i and f_i be the distribution and density function of ε_i, respectively. We assume that there exist constants $\alpha_1 > 0$ and $\alpha_2 > 0$ such that

$$0 < \alpha_1 < \inf_i \inf_{|u|\le\alpha_2} f_i(u).$$

For technical reasons, we assume that $q_n = O(1)$ and $\zeta_n = O(1)$ for the quantile loss in the remainder of this paper unless otherwise stated. The following proposition, the proof of which is in the Appendix, proves that the risk corresponding to the quantile loss is quadratically supported.

PROPOSITION 1. *Under the regularity conditions, the risk function corresponding to the quantile loss is quadratically supported with* $\tilde{\beta}_n = \beta^*$,

$$a_{nj} = -\sum_{i=1}^{n} x_{ij}\big(2\tau - 2I(\varepsilon_i < 0)\big),$$
$$b = 4\alpha_1\rho_*$$

provided $\sqrt{s_n}\eta_n \le \alpha_2/\zeta_n$, $\eta_n/(\delta_n s_n^{1/2}) \ll p_n$ *and*

$$\frac{n\delta_n^2}{\eta_n^2} \gg s_n \log p_n. \tag{4.16}$$

Let $\delta_n = 1/\sqrt{n}$ and $\eta_n = n^{-1/4}$. Note that $|a_n|_\infty = O_p(\sqrt{\log p_n/n})$ and $|a_{n,\pi^*}|_\infty = O_p(\sqrt{\log q_n/n})$ according to Hoeffding's inequality. Hence, the regularity conditions as well as (4.16) are satisfied with $\lambda_n \gg \log p_n/n$ provided $s_n \ll \sqrt{n}$, $p_n \gg n^{1/4}$, $s_n \log p_n \ll n^{1/2}$ and $\min_{j\in\pi^*}|\beta_j^*| \gg \sqrt{\lambda_n}$, and thus the corresponding GIC is consistent. We summarize the results in the following theorem.

THEOREM 4. *Suppose that* $y_i = \mathbf{x}_i'\beta^* + \varepsilon_i$, *where* ε_i *are independent random variables with* $\Pr(\varepsilon_i \le 0) = \tau$. *Assume that* $q_n = O(1)$ *and* $\zeta_n = O(1)$. *Then the GIC*$_{\lambda_n}$ *with* $\lambda_n \gg \log p_n/n$ *is consistent for the quantile loss provided* $s_n \ll \sqrt{n}$, $p_n \gg n^{1/4}$, $s_n \log p_n \ll n^{1/2}$ *and* $\min_{j\in\pi^*}|\beta_j^*| \gg \sqrt{\lambda_n}$.

Theorem 4 implies that the GIC$_{\lambda_n}$ is consistent when $p_n = \exp(an^d)$ and $s_n = n^c$ for some positive constants a, c and d with $d + c < 1/2$. This extends the results of [20] which only considered polynomially increasing p_n.

4.4. *Linear regression with AR*(1) *errors*. Suppose that the true model is $y_i = \mathbf{x}_i'\beta^* + \varepsilon_i$. We assume that ε_i are an AR(1) process defined as $\varepsilon_i = \theta\varepsilon_{i-1} + \nu_i$, where $\theta \in [0, 1)$ and ν_i are independent random variables whose common distribution has a sub-Gaussian tail. In this section, we investigate how the dependency in errors affects the variable selection. For technical simplicity, we let $q_n = O(1)$.

By replacing $\hat{\beta}^o$ in (4.1) by β^*, it can be shown that the quadratic loss belongs to the QSR with $\tilde{\beta}_n = \beta^*$, $a_n = -\mathbf{X}'\boldsymbol{\varepsilon}/n$, $b = \rho_*$, $\eta_n = \infty$ and $\delta_n = 0$, where $\mathbf{h}_{nj}$ are defined in (4.3). Note that $a_{nj} = -\sum_{i=1}^n \omega_{ij}\nu_i/n$, where $\omega_{ij} = \sum_{k=i}^n h_{njk}\theta^{k-i}$. Let $\gamma_{nj} = \sum_{i=1}^n \omega_{ij}^2/n$ and let $\gamma_n^* = \max_j \gamma_{nj}$. Then (4.2) implies that for any $\kappa_n \to \infty$

$$\begin{aligned}
&\Pr\Big\{\max_j |a_{nj}| > \sqrt{\kappa_n\gamma_n^*\log p_n/(d_\varepsilon n)}\Big\} \\
&\quad\le \sum_j \Pr\big\{|a_{nj}| > \sqrt{\kappa_n\gamma_n^*\log p_n/(d_\varepsilon n)}\big\} \\
&\quad\le c_\varepsilon \exp\big(-(\kappa_n - 1)\log p_n\big) \to 0,
\end{aligned}$$

which leads $|a_n|_\infty = O_p(\sqrt{\gamma_n^* \log p_n/n})$. Hence, the GIC_{λ_n} with $\lambda_n \gg \gamma_n^* \times \log p_n/n$ is consistent provided $\min_{j\in\pi^*} |\beta_j^*| \gg \sqrt{\lambda_n}$.

For γ_n^*, we have

$$\gamma_{nj} \le \zeta_n^2 \frac{\sum_{i=1}^n (\sum_{k=i}^n \theta^{k-i})^2}{n} \uparrow \frac{\zeta_n^2}{(1-\theta)^2},$$

as $n \to \infty$, and hence $\gamma_n^* \le \zeta_n^2/(1-\theta)^2$ for all sufficiently large n. Note that the upper bound of γ_n^* increases as $\theta \to 1$, which implies that we need the larger value of $\min_{j\in\pi^*} |\beta_j^*|$ for the consistency of the GIC when the correlation between errors becomes larger.

5. Pathconsistency with QSR. When p_n is large, the cardinality of $\mathcal{M}_{s_n}$ is so large that all possible search over $\mathcal{M}_{s_n}$ is almost impossible. A practical solution is to construct a sequence of submodels M_ξ indexed by $\xi \in (0,\infty)$, and choose the optimal $\hat{\xi}$ by minimizing $R_n(\hat{\beta}_{M_\xi}) + \lambda_n |\hat{\beta}_{M_\xi}|_0$ with respect to ξ, where λ_n is a consistent GIC. We state that $\{M_\xi, \xi \in (0,\infty)\}$ is path-consistent if there exists ξ^* such that $M_{\xi^*} = \pi^*$. Note that whenever M_ξ is path-consistent, $M_{\hat{\xi}} = \pi^*$ asymptotically.

When the quadratic loss is used, the solution path of the Lasso estimator can be used to construct M_ξ. Let $\hat{\beta}(\xi)^{\mathrm{lasso}}$ be the Lasso estimator with the tuning parameter ξ. That is,

$$\hat{\beta}(\xi)^{\mathrm{lasso}} = \operatorname*{argmin}_\beta R_n(\beta) + \xi|\beta|_1,$$

where $|\beta|_1 = \sum_{j=1}^p |\beta_j|$. Let $M_\xi^{\mathrm{lasso}} = \{j : |\hat{\beta}(\xi)_j^{\mathrm{lasso}}| \ge \tau\}$ for some $\tau > 0$. Then, under the regularity conditions, M_ξ^{lasso} is path-consistent when $\tau = O(\sqrt{q_n \log p_n/n})$ and $\min_{j\in\pi^*} |\beta_j^*| \gg \sqrt{q_n \log p_n/n}$ (see, e.g., Chapter 7 of [4]). A similar result with nonconvex penalties was obtained by [18]. In this section, we develop similar procedures for the class of QSR, which produce path-consistent submodels $\{M_\xi, \xi \in (0,\infty)\}$.

5.1. *Pathconsistency with Lasso estimators.* Let $\hat{\beta}(\xi)$ be the minimizer of $R_n(\beta) + \xi|\beta|_1$ on $\mathcal{M}_{s_n}$. We show that there exists a positive constant c such that $M_\xi = \{j : |\hat{\beta}(\xi)_j| > c\sqrt{q_n}\xi\}$ is path-consistent with the QSR when we replace the regularity conditions (C3) and (C4) with

(C3′) $\min_{j\in\pi^*} |\tilde{\beta}_{nj}| \gg \sqrt{q_n \lambda_n}$.
(C4′) $\lambda_n \gg s_n \delta_n / q_n$.

REMARK 4. Condition (C3′) requires a larger $\min_{j\in\pi^*} |\tilde{\beta}_{nj}|$ than (C3), which is a disadvantage that derives from using the Lasso penalty instead of the l_0 penalty. Condition (C4′) is weaker than (C4) when $s_n < q_n^2$.

THEOREM 5. *Suppose that there exists a sequence of positive numbers* $\{\lambda_n\}$ *satisfying the regularity conditions with* (C3) *and* (C4) *being replaced by* (C3′) *and* (C4′). *Then there exists* $c > 0$ *such that* M_ξ *is path-consistent*.

PROOF. We show that $M_{\sqrt{\lambda_n}} = \pi^*$ with probability tending to 1. Let $\xi_n = \sqrt{\lambda_n}$ and let $\beta_n = \hat{\beta}(\xi_n)$. Moreover, let

$$Q(\beta) = R_n(\beta) - R_n(\tilde{\beta}_n) + \xi_n|\beta|_1 - \xi_n|\tilde{\beta}_n|_1.$$

We first show that $\|\beta_n - \tilde{\beta}_n\| \le \eta_n$. Suppose $\|\beta - \tilde{\beta}_n\| > \eta_n$ with $|\beta|_0 \le s_n$. Since $|\beta - \tilde{\beta}_n|_1 \le \sqrt{s_n}\|\beta - \tilde{\beta}_n\|$ and $\xi_n|\beta|_1 - \xi_n|\tilde{\beta}_n|_1 \ge -\xi_n|\beta - \tilde{\beta}_n|_1$, we have

$$Q(\beta) \ge -(|a_n|_\infty + \xi_n)\sqrt{s_n}\|\beta - \tilde{\beta}_n\| + \frac{b}{2}\|\beta - \tilde{\beta}_n\|^2 - \delta_n s_n,$$

which is positive eventually according to condition (C6). Since $Q(\beta)$ is convex, $\|\beta_n - \tilde{\beta}_n\| \le \eta_n$.

Let

$$w_j = a_{nj}(\beta_{nj} - \tilde{\beta}_{nj}) + \frac{b}{2}(\beta_{nj} - \tilde{\beta}_{nj})^2 + \xi_n|\beta_{nj}| - \xi_n|\tilde{\beta}_{nj}|.$$

Note that $Q(\beta_n) \ge \sum_{j=1}^{p_n} w_j - \delta_n s_n$.

For $j \in \pi^*$, since $|\beta_{nj}| - |\tilde{\beta}_{nj}| \ge -|\beta_{nj} - \tilde{\beta}_{nj}|$, we have

$$w_j \ge -(|a_{n,\pi^*}|_\infty + \xi_n)|\beta_{nj} - \tilde{\beta}_{nj}| + \frac{b}{2}(\beta_{nj} - \tilde{\beta}_{nj})^2. \tag{5.1}$$

Hence,

$$\begin{aligned}\sum_{j\in\pi^*} w_j &\ge -(|a_{n,\pi^*}|_\infty + \xi_n)\sqrt{q_n}\|\beta_{n,\pi^*} - \tilde{\beta}_{n,\pi^*}\| + \frac{b}{2}\|\beta_{n,\pi^*} - \tilde{\beta}_{n,\pi^*}\|^2 \\ &\ge -\frac{q_n(|a_{n,\pi^*}|_\infty + \xi_n)^2}{2b} \ge -2q_n\xi_n^2/b\end{aligned} \tag{5.2}$$

since $\xi_n \gg |a_{n,\pi^*}|_\infty$ by (C2).

For $j \in \pi^{*c}$, note that

$$w_j \ge (\xi_n - |a_n|_\infty)|\beta_{nj}| + \frac{b}{2}\beta_{nj}^2.$$

Since $\xi_n \gg |a_n|_\infty$, we have $\min_{j\in\pi^{*c}} w_j \ge 0$ for all sufficiently large n. Let c_1 be a positive constant that satisfies $bc_1^2/2 - 2/b - 1 > 0$. If $|\beta_{nj}| > c_1\sqrt{q_n}\xi_n$ for $j \in \pi^{*c}$, $w_j > bc_1^2 q_n\xi_n^2/2$ for sufficiently large n. Note that condition (C4′) and (5.2) imply

$$\sum_{j\in\pi^*} w_j - s_n\delta_n \ge -2q_n\xi_n^2/b - q_n\xi_n^2.$$

Hence, if there exists $j \in \pi^{*c}$ such that $|\beta_{nj}| > c_1\sqrt{q_n}\xi_n$, then $Q(\beta_n) \geq q_n\xi_n^2(bc_1^2/2 - 2/b - 1) > 0$, which is impossible. Therefore, we conclude that

$$\max_{j\in\pi^{*c}} |\beta_{nj}| \leq c_1\sqrt{q_n}\xi_n. \tag{5.3}$$

On the other hand, since $\min_{j\in\pi^{*c}} w_j \geq 0$, we have

$$Q(\beta_n) \geq \sum_{j\in\pi^*} w_j - s_n\delta_n.$$

Choose $c_2 > 0$ such that $c_2^2 b/2 - 2c_2 - 2/b - 1 > 0$. Suppose that there exists $j \in \pi^*$ such that $|\beta_{nj} - \tilde{\beta}_{nj}| > c_2\sqrt{q_n}\xi_n$. Then (5.1) and (C4) imply

$$w_j \geq -2\xi_n c_2\sqrt{q_n} + \frac{b}{2}c_2^2 q_n\xi_n^2,$$

and thus

$$Q(\beta_n) \geq \sum_{l\in\pi^*-\{j\}} w_l - 2\xi_n c_2\sqrt{q_n} + \frac{b}{2}c_2^2 q_n\xi_n^2 - s_n\delta_n.$$

Similarly to (5.2), we can show $\sum_{l\in\pi^*-\{j\}} w_l \geq -2q_n\xi_n^2/b$. Hence, (C4′) implies

$$Q(\beta_n) \geq q_n\xi_n^2\bigl(c_2^2 b/2 - 2/b - 2c_2 - 1\bigr) > 0,$$

which is impossible, and thus we conclude that

$$\max_{j\in\pi^*} |\beta_{nj} - \tilde{\beta}_{nj}| \leq c_2\sqrt{q_n}\xi_n.$$

Since $|\tilde{\beta}_{nj}| \gg \sqrt{q_n}\xi_n$ by (C3′), we have

$$\min_{j\in\pi^*} |\beta_{nj}| \gg \sqrt{q_n}\xi_n. \tag{5.4}$$

By combining (5.3) and (5.4), we conclude that $\{j : |\beta_{nj}| > \max\{c_1, c_2\}\sqrt{q_n}\xi_n\} = \pi^*$ with probability tending to 1. □

The regularity condition (C3′) is suboptimal since it requires larger $\min_{j\in\pi^*} |\tilde{\beta}_{nj}|$ than the one in condition (C3), in particular when q_n diverges. This is a disadvantage of using the Lasso estimator for model selection. In the next subsection, we show that this disadvantage disappears when we use a nonconvex penalty.

5.2. *Pathconsistency for nonconvex penalties.* Let $\hat{\beta}(\xi)$ be the minimizer of

$$R_n(\beta) + \sum_{j=1}^{p_n} J_\xi\bigl(|\beta_j|\bigr)$$

over $\mathcal{M}_{s_n}$, where J_ξ is a nonconvex penalty such that (1) J'_ξ is nonnegative, nonincreasing and continuous on $(0,\infty)$ and (2) there exists $a>1$ such that $\lim_{t\to 0+} J'_\xi(t)=\xi$, $J'_\xi(t)\geq \xi - t/a$ for $t\in(0,a\xi)$, and $J'_\xi(t)=0$ for $t\geq a\xi$. Here, $J'_\xi(t)=dJ_\xi(t)/dt$. This class of nonconvex penalties includes the SCAD penalty [9] and MCP [31].

Let $M_\xi=\{j:|\hat\beta(\xi)_j|>\tau_n\}$, where $\tau_n=\sqrt{q_n}|a^*_{n,\pi}|_\infty/b+\sqrt{2\delta_n s_n/b}$. The next theorem proves that M_ξ is path-consistent.

THEOREM 6. *Let $\xi_n=\sqrt{\lambda_n}$. Under the regularity conditions with* (C5) *being replaced by*

$$\lambda_n \gg \max\{q_n|a_{n,\pi^*}|_\infty, s_n\delta_n\}, \tag{5.5}$$

$\Pr(M_{\xi_n}=\pi^*)\to 1$ *as* $n\to\infty$.

PROOF. Let $\beta_n=\hat\beta(\xi_n)$ and

$$Q(\beta)=R_n(\beta)+\sum_{j=1}^{p_n} J_{\xi_n}(|\beta_j|)-R_n(\tilde\beta_n)-\sum_{j=1}^{p_n} J_{\xi_n}(|\tilde\beta_{nj}|).$$

First, we show that $\|\beta_n-\tilde\beta_n\|\leq\eta_n$. Since $|\tilde\beta_{nj}|\gg\xi_n$ for $j\in\pi^*$, we have $J_{\xi_n}(|\tilde\beta_{nj}|)\leq a\xi_n^2$. Hence, when $\|\beta-\tilde\beta_n\|>\eta_n$ with $|\beta|_0\leq s_n$, Lemma 1 implies

$$Q(\beta)\geq R_n(\beta)-R_n(\tilde\beta_n)-q_n a\xi_n\geq \frac{b}{16}\eta_n^2-q_n a\xi_n^2>0$$

eventually according to condition (C6), which implies $\|\beta_n-\tilde\beta_n\|\leq\eta_n$.

Let $\pi_1=\{j:\beta_{nj}\neq 0,\tilde\beta_{nj}=0\}$, $\pi_2=\{j:|\beta_{nj}|>a\xi_n,\tilde\beta_{nj}\neq 0\}$ and $\pi_3=\{j:|\beta_{nj}|\leq a\xi_n,\tilde\beta_{nj}\neq 0\}$. Then

$$\begin{aligned} &R_n(\beta_n)+\sum_{j=1}^{p_n} J_{\xi_n}(|\beta_{nj}|)-R_n(\tilde\beta_n)-\sum_{j=1}^{p_n} J_{\xi_n}(|\tilde\beta_{nj}|)\\ &\quad\geq \sum_{j\in\pi_1} w_j+\sum_{j\in\pi_2} w_j+\sum_{j\in\pi_3} w_j-\delta_n s_n, \end{aligned}$$

where

$$w_j=a_{nj}(\beta_{nj}-\tilde\beta_{nj})+\frac{b}{2}(\beta_{nj}-\tilde\beta_{nj})^2+J_{\xi_n}(|\beta_{nj}|)-J_{\xi_n}(|\tilde\beta_{nj}|).$$

For $j\in\pi_1$, we can choose $0<c<a$ such that $J_{\xi_n}(|\beta_{nj}|)\geq\xi_n|\beta_{nj}|/2$ for $|\beta_{nj}|\leq c\xi_n$. Then $J_{\xi_n}(|\beta_{nj}|)\geq\min\{c\xi_n^2/2,\xi_n|\beta_{nj}|/2\}$, and hence

$$w_j\geq\min\left\{-\frac{|a_n|_\infty^2}{2b}+c\xi_n^2/2,(\xi_n/2-|a_n|_\infty)|\beta_{nj}|+\frac{b}{2}\beta_{nj}^2\right\}>0$$

eventually since $\xi_n\gg|a_n|_\infty$.

For $j \in \pi_2$, $J_{\xi_n}(|\beta_{nj}|) = J_{\xi_n}(|\tilde{\beta}_{nj}|)$. Since $\sum_{j\in\pi_2} |\beta_{nj} - \tilde{\beta}_{nj}| \le \sqrt{q_n}\|\beta_{n,\pi_2} - \tilde{\beta}_{n,\pi_2}\|$, we have

$$\begin{aligned}\sum_{j\in\pi_2} w_j &\ge -|a_{n,\pi^*}|_\infty\sqrt{q_n}\|\beta_{n,\pi_2} - \tilde{\beta}_{n,\pi_2}\| + \frac{b}{2}\|\beta_{n,\pi_2} - \tilde{\beta}_{n,\pi_2}\|^2 \\ &\ge \frac{-|a_{n,\pi^*}|_\infty^2 q_n}{2b}.\end{aligned} \tag{5.6}$$

For $j \in \pi_3$, note that $J_{\xi_n}(|\beta_{nj}|) - J_{\xi_n}(|\tilde{\beta}_{nj}|) \ge -\xi_n|\beta_{nj} - \tilde{\beta}_{nj}|$, and hence

$$w_j \ge -\big(|a_{n,\pi^*}|_\infty + \xi_n\big)|\beta_{nj} - \tilde{\beta}_{nj}| + \frac{b}{2}(\beta_{nj} - \tilde{\beta}_{nj})^2.$$

Therefore, if $|\beta_{nj} - \tilde{\beta}_{nj}| \gg \xi_n$, $w_j > |a_{n,\pi^*}|_\infty^2 q_n/(2b) + \delta_n s_n$ according to condition (C4′), and hence $Q(\beta_n) > 0$ which is impossible. However, $|\beta_{nj} - \tilde{\beta}_{nj}| \gg \xi_n$ for all $j \in \pi_3$ since $|\tilde{\beta}_{nj}| \gg \xi_n$ and $|\beta_{nj}| \le a\xi_n$. Therefore, $\pi_3 = \varnothing$ and $Q(\beta_n) \ge -|a_{n,\pi^*}|_\infty^2 q_n/(2b)$.

To complete the proof, it suffices to show that $|\beta_{nj}| \le \tau_n$ for all $j \in \pi_1$. Let c be a positive constant such that $J_{\xi_n}(|\beta_{nj}|) \ge \xi_n|\beta_{nj}|/2$ for all $|\beta_{nj}| \le c\xi_n$. If $|\beta_{nj}| > c\xi_n$, we have $w_j \ge -|a_n|_\infty^2/(2b) + c\xi_n^2$, and hence $Q(\beta_n) > 0$, which is a contradiction. So $|\beta_{nj}| \le c\xi_n$, in which case $w_j \ge \xi_n|\beta_{nj}|/2 + b\beta_{nj}^2/2$ since $\xi_n \gg |a_n|_\infty$. Hence, if $|\beta_{nj}| > \tau_n$, it can be shown that $w_j > |a_{n,\pi^*}|_\infty^2 q/(2b) + \delta_n s_n$ and so $Q(\beta) > 0$ which is a contradiction. Thus, $|\beta_{nj}| \le \tau_n$ for all $j \in \pi_1$, and the proof is complete. □

REMARK 5. Condition (5.5) is the same as (C4) for the quadratic, logistic and Huber losses since $\delta_n = 0$. For the quantile loss, (5.5) holds when $\log p_n \gg s_n$ and $\lambda_n \gg \log p_n/n$ as well as (C4) holds. In particular, when $s_n = O(1)$, (5.5) and (C4) are equivalent.

REMARK 6. Note that $\tau_n = 0$ for the quadratic, logistic and Huber losses. Theorem 6 implies that the non-covex penalized least square estimator is path-consistent provided $\min_{j\in\pi^*}|\hat{\beta}_j^o| \gg \sqrt{\log p_n/n}$ and the error distribution has a sub-Gaussian tail. This result concurs with those of [31] and [17]. Similarly, Theorem 6 shows that the nonconvex penalized estimator for the logistic or Huber loss is path-consistent if $\min_{j\in\pi^*}|\hat{\beta}_j^o| \gg \max\{\sqrt{\log p_n/n}, \sqrt{q_n/n}\}$, provided that $\|\hat{\beta}^o - \beta^*\| = O_p(\sqrt{q_n/n})$ and $\zeta_n^2 s_n^2 \log p_n/n \to 0$.

6. Adaptive model selection with Huber loss. Any GIC with $\lambda_n = \alpha_n \times \log p_n/n$ is consistent as long as $\alpha_n \to 0$ and $\lambda_n \to 0$, when $q_n = O(1)$ and $\min_{j\in\pi^*}|\hat{\beta}_j^o| > c$ for some $c > 0$. That is, the class of consistent GIC is large and it is important to choose a GIC that performs well with finite samples. In this section,

we propose a data-adaptive selection of λ_n with the Huber loss. As a by-product, we give a guidance of choosing d data-adaptively. Even if we consider the Huber loss only, the proposed procedure in this section can be extended for other convex losses without much modification.

We assume the following conditions:

(A1) $\zeta_n^2 s_n^2 \log p_n/n \to 0$,
(A2) $\|\hat{\beta}^o - \beta^*\| = O_p(\sqrt{q_n/n})$,
(A3) $\log p_n \gg q_n$ and $\log p_n \to \infty$,
(A4) ε_i have the common distribution F that is symmetric and continuous,
(A5) for a given class $\mathcal{G} \subset \mathcal{M}_{s_n}$ and $d > 0$, there exists $\rho_{\mathcal{G},d} > 0$ such that $\min_{\pi:\pi\in\mathcal{G}} b_n(\pi, \pi^*, d) > \rho_{\mathcal{G},d}$, where

$$b_n(\pi_1, \pi_2, d) = \lambda_{\min}\left\{\left(\frac{1}{n}\sum_{i\in J(\pi_1,\pi_2,d)} \mathbf{x}_i\mathbf{x}_i'\right)_{\pi_1\cup\pi_2}\right\}$$

with $J(\pi_1, \pi_2, d) = \{i : |y_i - \mathbf{x}_i'\hat{\beta}_{\pi_1}| \le d, |y_i - \mathbf{x}_i'\hat{\beta}_{\pi_2}| \le d\}$.

Conditions (A1) and (A2) are assumed in Theorem 3. Conditions (A3) and (A4) are for notational and technical simplicity. Condition (A5) is a data-adaptive version of the SRC for the Huber loss. When $\mathcal{G} = \mathcal{M}_{s_n}$ and $d = \infty$, we have $\rho_{\mathcal{G},d} \ge \rho_*$. The following theorem provides theoretical bases for the proposed adaptive model selection procedure.

THEOREM 7. *For a given class $\mathcal{G}$ of submodels and $d > 0$, suppose that there exists $\delta_0 > 0$ such that*

$$\min_{j\in\pi^*} |\hat{\beta}_j^o| > \frac{1}{\rho_{\mathcal{G},d}}\sqrt{(2\sigma_d^2 + \delta_0)\frac{\log p_n}{n}}, \tag{6.1}$$

where $\rho_{\mathcal{G},d}$ is the constant defined in (A5) *and*

$$\sigma_d^2 = \mathrm{E}\{\varepsilon_i^2 I(|\varepsilon_i| \le d)\} + d^2 \Pr(|\varepsilon_i| > d).$$

Under the conditions (A1) *to* (A5), *for any $0 < \delta < \delta_0$, we have*

$$R_n(\hat{\beta}_\pi) + \hat{\lambda}_n(\pi, \pi^*, d, \delta)|\pi| \ge R_n(\hat{\beta}^o) + \hat{\lambda}_n(\pi, \pi^*, d, \delta)|\pi^*| \tag{6.2}$$

for all $\pi \in \mathcal{G}$ with probability converging to 1, *where*

$$\hat{\lambda}_n(\pi_1, \pi_2, d, \delta) = \frac{(2\hat{\sigma}^2_{d,\pi_1\cup\pi_2} + \delta)}{2b_n(\pi_1, \pi_2, d)}\frac{\log p_n}{n},$$

and

$$\hat{\sigma}^2_{d,\pi} = \frac{1}{n}\sum_{i=1}^n \{(y_i - \mathbf{x}_i'\hat{\beta}_{\pi,i})^2 I(|y_i - \mathbf{x}_i'\hat{\beta}_{\pi,i}| \le d) + d^2 I(|y_i - \mathbf{x}_i'\hat{\beta}_{\pi,i}| > d)\}.$$

Also, (6.2) *holds when $d = \infty$* (*i.e., quadratic loss*) *and the error distribution is Gaussian.*

For given d, $\mathcal{G}$ and $\delta > 0$, we propose a model selection procedure so called the *adaptive information criterion* with d (AdIC_d) which selects $\pi \in \mathcal{G}$ satisfying

$$R_n(\hat{\beta}_\nu) + \hat{\lambda}_n(\nu, \pi, d, \delta)|\nu| \geq R_n(\hat{\beta}_\pi) + \hat{\lambda}_n(\nu, \pi, d, \delta)|\pi| \tag{6.3}$$

for all $\nu \in \mathcal{G}$. If there is no such π in $\mathcal{G}$, we conclude that $\pi^* \notin \mathcal{G}$. Theorem 7 implies that under the conditions (A1) to (A5) the AdIC_d is consistent as long as $\mathcal{G}$ contains π^* and (6.1) holds with probability tending to 1.

A naive choice of $\mathcal{G}$ is $\mathcal{M}_{s_n}$. This choice, however, does not work well since $b_n(\pi, \pi^*, d)$ can be very small and so is $\rho_{\mathcal{G},d}$ when $|\pi|$ and d are small. Thus, it is difficult to satisfy condition (6.1). We propose to construct $\mathcal{G}$ by deleting π from $\mathcal{M}_{s_n}$ when $|J(\pi, d)|$ is small, where $J(\pi, d) = \{i : |y_i - \mathbf{x}_i'\hat{\beta}_\pi| < d\}$. Let $\tilde{\beta}$ be an initial estimator, and for given $\alpha \in (0, 1)$ let $c_\alpha = |\{i : |y_i - \mathbf{x}_i'\tilde{\beta}| \leq \alpha d\}|$. We delete π from $\mathcal{M}_{s_n}$ when $|J(\pi, d)| \leq c_\alpha$. We denote $\tilde{\mathcal{G}}_d$ the class of submodels obtained by this way. In Proposition 3 in the Appendix, we show that $\pi^* \in \tilde{\mathcal{G}}_d$ with probability tending to 1 when $\sigma(\tilde{\beta}) \leq cs_n$ for some $c > 0$ and $\|\tilde{\beta} - \beta^*\| = O_p(\sqrt{s_n \log p_n / n})$. The Lasso estimator with the absolute loss can be used for the initial estimator [2].

For given $d > 0$, let $\mathcal{M}_d$ be the class of submodels obtained by the solution path of the path-consistent penalized estimator given in either Section 5.1 or Section 5.2. To reduce the computational burden, we propose to use the AdIC_d with $\mathcal{G} = \mathcal{G}_d$, where $\mathcal{G}_d = \mathcal{M}_d \cap \tilde{\mathcal{G}}_d$. Since (6.1) is implied by (C3), it is easy to see that this AdIC_d is consistent under the same conditions assumed in either Theorem 5 or Theorem 6 along with the conditions (A1) to (A5).

Condition (6.1) suggests a way of choosing d data-adaptively. Note that the lower bound of $\min_{j \in \pi^*} |\hat{\beta}_j^o|$ in (6.1) is roughly proportional to $\sigma_d / \rho_{\mathcal{G},d}$. We propose to choose d which minimizes $\hat{\sigma}_{d,\hat{\pi}_d} / \hat{\rho}_{\mathcal{G},d}$, where $\hat{\rho}_{\mathcal{G},d} = \min_{\pi \in \mathcal{G}} b_n(\pi, \hat{\pi}_d, d)$ and $\hat{\pi}_d$ is the model selected by the AdIC_d with $\mathcal{G}$. Here, we let $\mathcal{G} = \bigcup_d \mathcal{G}_d$. We write AdIC for the AdIC_d with the data-adaptively selected d.

REMARK 7. The $\rho_{\mathcal{G},d}$ is decreasing as s_n is increasing. Since the lower bound of $\min_{j \in \pi^*} |\hat{\beta}_j^o|$ is reciprocally proportional to $\rho_{\mathcal{G},d}$, it would be desirable to set s_n as small as possible. In practice, we apply the proposed data-adaptive model selection procedure for all integers of s_n less than $s_{\max}$ to get the sequence of models $\hat{\pi}_s, s = 1, \ldots, s_{\max}$, where $\hat{\pi}_s$ is the selected model with $s_n = s$. Then we choose $\hat{\pi}_{\hat{s}}$, where $\hat{s} = \min\{s : \phi(\hat{\pi}_s) = \operatorname{argmax}_r \phi(\hat{\pi}_r)\}$ and $\phi(\pi_s) = |\{r : \hat{\pi}_r = \hat{\pi}_s\}|$. This method works well unless $s_{\max}$ is too large.

7. Numerical studies. In this section, we investigate the finite sample properties of the GIC and AdIC for various loss functions by simulation as well as real data analysis. First, we compare the path consistency of the SCAD solution path for the Huber loss with various values of d. For computation, we use the combination of the CCCP algorithm of [16] and the solution path algorithm of [23].

Second, we examine the performance of the data-adaptively selected d in the AdIC by simulation. Finally, we compare the AdIC with other selection consistent GICs such as the CRIC of [32] and the HIBC of [29] by simulation as well as real data analysis.

7.1. *Simulation*. For the simulation model, we set $p_n = 500$ and $q_n = 5$. We generate $\mathbf{x}_i$ independently from the multivariate normal distribution with mean 0 and variance 1 and covariance $\operatorname{corr}(x_{ij}, x_{ik}) = 0.3^{-|j-k|}$ and normalize them to have $\|X^j\|^2 = n$. We consider the four models for ε_i: (1) $N(0,4)$, (2) $0.9N(0,3/5) + 0.1N(0.25)$, (3) t-distribution with 3 degrees of freedom multiplied by $\sqrt{3/4}$ and (4) the t-distribution with 1 degree of freedom (i.e., Cauchy distribution). Note that variances of the first three distributions are 4. For β^*, we let $\beta_j^* = \tau j \beta_*$ for some $\tau > 0$, where

$$\beta_* = \frac{1}{b^*}\sqrt{\frac{8\log p_n}{n}}.$$

Here, $b_* = \min_{\pi:|\pi|\le 2q_n} \lambda_{\min}(\Sigma_\pi)$, where Σ is the $p_n \times p_n$ matrix whose (j,k) element is the correlation of X^j and X^k. When $\varepsilon_i \sim N(0,4)$, (4.2) implies

$$\Pr\left\{ |a_n^*|_\infty > \sqrt{\frac{(8+\delta)\log p_n}{n}} \right\} \to 0$$

for any $\delta > 0$, where $a_n^* = \sum_{i=1}^n \varepsilon_i \mathbf{x}_i / n$. That is β_* can be considered as a surrogated quantity of the lower bound of $\min_{j\in\pi^*} |\hat{\beta}_j^o|$ in (6.1). Thus, τ measures how much the $\min_{j\in\pi^*} |\beta_j^*|$ is larger than the lower bound. Note that the larger the τ is the easier the model selection is.

7.1.1. *Pathconsistency.* Table 1 shows the frequencies of pathconsistency of among 100 simulated data sets for various values of d when $n \in \{100, 400\}$ and $\tau \in \{1, 1.2, 1.5, 2\}$. In the table, $d_\alpha, \alpha > 0$ are constants satisfying $\Pr(|y - \mathbf{x}'\beta^*| < d_\alpha) = \alpha/100$, and d_0 represents the absolute loss. For the first three error distributions that have finite variances, the results are similar regardless of the choice of d. In contrast, the results with d_{80} and d_{100} are worse than the others for the Cauchy distribution, which indicates that the choice of d is important when the error distribution is heavy tailed.

7.1.2. *Performance of $\hat{d}$.* Table 2 presents the percentages for the AdIC_d with various values of d as well as AdIC to select the true model among the pathconsistent cases. We let $s_n = 10, \delta = 0, \alpha = 1/2$ and $\mathcal{G} = \bigcup_{\alpha=10,20,\ldots,100} \mathcal{G}_{d_\alpha}$. For most cases, the AdIC yields the best performance of selecting the true model.

TABLE 1
Frequencies of pathconsistency

		$n = 100$					$n = 400$				
ε	τ	d_0	d_{20}	d_{50}	d_{80}	d_{100}	d_0	d_{20}	d_{50}	d_{80}	d_{100}
Normal	1.0	89	87	87	92	95	97	97	97	99	99
	1.2	95	94	94	97	97	100	100	100	100	100
	1.5	97	97	97	97	97	100	100	100	100	100
	2.0	97	97	97	97	97	100	100	100	100	100
Mixture	1.0	97	97	97	97	91	100	100	100	100	99
	1.2	97	97	97	97	97	100	100	100	100	100
	1.5	97	97	97	97	97	100	100	100	100	100
	2.0	97	97	97	97	97	100	100	100	100	100
$t(3)$	1.0	95	97	97	97	94	100	100	100	100	97
	1.2	97	97	97	97	95	100	100	100	100	100
	1.5	97	97	97	97	97	100	100	100	100	100
	2.0	97	97	97	97	97	100	100	100	100	100
$t(1)$	1.0	81	83	86	82	5	99	99	99	93	1
	1.2	86	88	89	88	10	99	99	99	96	4
	1.5	94	94	94	91	23	100	100	100	100	17
	2.0	98	98	98	97	42	100	100	100	100	37

TABLE 2
Percentages of selecting the true model among the path-consistent cases for the $AdIC_d$s as well as AdIC (the columns of $\hat{d}$)

		$n = 100$					$n = 400$				
ε	τ	d_{20}	d_{50}	d_{80}	d_{100}	$\hat{d}$	d_{20}	d_{50}	d_{80}	d_{100}	$\hat{d}$
Normal	1.0	2.3	8.0	42.4	83.2	83.2	18.6	39.2	75.8	89.9	86.9
	1.2	8.5	11.7	73.2	95.9	95.9	35.0	70.0	95.0	98.0	95.0
	1.5	12.4	19.6	89.7	100.0	100.0	50.0	97.0	100.0	99.0	98.0
	2.0	9.3	34.0	96.9	100.0	100.0	48.0	100.0	100.0	99.0	99.0
Mixture	1.0	9.3	24.7	78.4	78.0	87.6	48.0	89.0	99.0	90.9	93.0
	1.2	10.3	23.7	93.8	88.7	99.0	47.0	98.0	100.0	99.0	98.0
	1.5	9.3	27.8	94.8	97.9	100.0	41.0	100.0	100.0	100.0	100.0
	2.0	4.1	26.8	90.7	99.0	100.0	27.0	100.0	100.0	100.0	100.0
$t(3)$	1.0	8.2	28.9	83.5	87.2	89.7	55.0	95.0	100.0	92.8	97.0
	1.2	7.2	37.1	92.8	94.7	96.9	47.0	99.0	100.0	93.0	100.0
	1.5	5.2	28.9	93.8	96.9	99.0	37.0	100.0	100.0	98.0	100.0
	2.0	6.2	22.7	92.8	100.0	100.0	21.0	100.0	100.0	100.0	100.0
$t(1)$	1.0	7.2	14.0	24.4	0.0	31.7	52.5	76.8	50.5	0.0	81.8
	1.2	9.1	28.1	50.0	0.0	62.9	50.5	91.9	77.1	0.0	93.9
	1.5	7.4	41.5	82.4	0.0	90.2	35.0	100.0	99.0	0.0	100.0
	2.0	11.2	29.6	93.8	14.3	94.9	24.0	100.0	100.0	0.0	100.0

TABLE 3

Percentages of selecting the true model among the path-consistent cases for the AdIC when s_n is selected data-adaptively with $s_{\max} = 30$. The numbers in the parentheses are the mean and standard error of the selected s_n

n	τ	Normal	Mixture	$t(3)$	$t(1)$
100	1.0	84.9	87.6	88.6	30.1
		(4.99, 0.054)	(5.11, 0.091)	(5.20, 0.118)	(5.04, 0.207)
	1.2	95.8	99.0	96.9	58.4
		(4.99, 0.033)	(5.08, 0.036)	(5.04, 0.031)	(5.02, 0.127)
	1.5	100	100	98.9	87.9
		(5.04, 0.031)	(5.13, 0.044)	(5.13, 0.078)	(5.20, 0.134)
	2.0	100	100	100	93.8
		(5.02, 0.028)	(5.17, 0.051)	(5.15, 0.034)	(5.06, 0.121)
400	1.0	90.8	93.0	97.0	81.8
		(5.71, 0.165)	(5.01, 0.026)	(5.04, 0.019)	(5.03, 0.055)
	1.2	98.0	98.0	100	94.9
		(5.38, 0.016)	(5.00, 0.014)	(5.00, 0.000)	(5.04, 0.031)
	1.5	99.0	100	100	100
		(5.15, 0.102)	(5.00, 0.000)	(5.00, 0.000)	(5.00, 0.000)
	2.0	99.0	100	100	100
		(5.02, 0.014)	(5.00, 0.000)	(5.00, 0.000)	(5.0, 0.000)

7.1.3. *Selection of s_n.* We investigate the selection method of s_n proposed in the remark of Section 6. Table 3 presents the percentages for the AdIC to select the true model among the path-consistent cases when s_n is selected by the proposed method in Section 6 with $s_{\max} = 30$. The results are similar to the corresponding results in Table 2 (i.e., the columns of $\hat{d}$) where s_n is fixed at 10, which suggest that the performance of AdIC is not sensitive to the choice of s_n.

7.1.4. *Comparison of the AdIC with other competitors.* We compare the AdIC with the CRIC of [32] and the HBIC of [29] that select the model which minimizes $\log(R_n(\hat{\beta}_\pi)) + \lambda_n|\pi|$, where $\lambda_n = 2(\log p_n + \log\log p_n)/n$ for the CRIC and $\lambda_n = \log\log n \log p_n/n$ for the HBIC, and present the results in Table 4. The Huber loss with the data-adaptively selected d is used for R_n. The performance of the AdIC is similar or slightly better compared to the CRIC and HBIC for the first three error distributions while the AdIC dominates the other competitors for the Cauchy distribution.

REMARK 8. The CRIC and HBIC are developed for the quadratic loss. We investigated the performance of the CRIC and HBIC with the quadratic loss, and we found that the performances for model selection are similar or worse compared to those with the Huber loss.

TABLE 4
Comparison of the AdIC with the CRIC and HBIC—the proportions of selecting the true model among 100 *simulations*

		$n = 100$			$n = 400$		
ε	τ	**CRIC**	**HBIC**	**AdIC**	**CRIC**	**HBIC**	**AdIC**
Normal	1.0	0.88	0.80	0.79	0.93	0.88	0.86
	1.2	0.96	0.81	0.93	0.99	0.92	0.95
	1.5	0.97	0.79	0.97	1.00	0.93	0.98
	2.0	0.97	0.80	0.97	1.00	0.90	0.99
Mixture	1.0	0.93	0.94	0.85	0.99	0.98	0.93
	1.2	0.97	0.92	0.96	1.00	0.98	0.98
	1.5	0.97	0.89	0.97	1.00	0.98	1.00
	2.0	0.97	0.92	0.97	1.00	1.00	1.00
$t(3)$	1.0	0.95	0.91	0.87	1.00	1.00	0.97
	1.2	0.97	0.91	0.94	1.00	1.00	1.00
	1.5	0.97	0.91	0.96	1.00	1.00	1.00
	2.0	0.97	0.90	0.97	1.00	1.00	1.00
$t(1)$	1.0	0.06	0.20	0.26	0.02	0.11	0.81
	1.2	0.10	0.26	0.56	0.06	0.25	0.93
	1.5	0.22	0.53	0.83	0.27	0.58	1.00
	2.0	0.51	0.74	0.93	0.65	0.83	1.00

7.2. *Real data analysis*. We analyze the data set used in [24], which consists of the gene expression levels of 18,975 genes obtained from 120 rats. The main objective of the analysis is to find genes that are correlated with gene TRIM32, which is known to cause Bardet–Biedl syndromes. As carried out by [15], we select the 3000 genes that have the largest absolute correlation with gene TRIM32, and then apply the GIC to select the signal genes among the selected 3000 genes.

To compare variable selectivity, we first divide the data set randomly into training set with 90 observations and test set with 30 observations. Then we select genes using the AdIC as well as the CRIC and HIBC. Then we calculate the Kendall's τ's and mean absolute deviation (MAD) between the predicted and observed response variables on the test set. We repeat this experiment 100 times and summarize the averages of the numbers of selected genes, the Kendall's τ's and MAD for each selection method in Table 5. While the predictive performance of the AdIC and HBIC are similar, the AdIC selects a fewer genes. The CRIC is worst in prediction even though it uses the fewest genes. The results amply show that the AdIC is a useful tool to select covariates without hampering prediction accuracy.

8. Concluding remarks. When using the GIC, we may use different loss functions for estimation and selection. For example, let $\hat{\beta}_{\pi,d_e}$ be the Huber estimator with $d = d_e$. Then we select $\pi \in \mathcal{M}_{s_n}$ by minimizing $R_n(\hat{\beta}_{\pi,d_e}) + \lambda_n|\pi|$, where R_n is obtained with the Huber loss with $d = d_s$. Theorem 3 and Theorem 7

TABLE 5
Results of data analysis (TRIM-32). The numbers in the parentheses are the standard errors

Method	Number of genes	Kendall's τ	MAD
AdIC	1.93 (0.263)	0.428 (0.031)	0.522 (0.020)
HBIC	2.27 (0.228)	0.432 (0.033)	0.523 (0.021)
CRIC	1.53 (0.187)	0.418 (0.032)	0.526(0.021)

are still valid as long as $\|\hat{\beta}_{\pi^*,d_e} - \beta^*\| = O_p(\sqrt{q_n/n})$. However, the choice of d_e is not obvious and we leave this problem as a future work.

We have seen that the results about the class of QSR give a way of choosing a loss function data-adaptively among the class of Huber losses for linear regression models. A similar problem is to choose a loss function for the classification. Examples of the loss function for classification are the exponential loss for boosting and the hinge loss for the support vector machine. Note that the key quantity of the selection consistency of the GIC is the upper bound of $|a_n|_\infty$: the smaller the $|a_n|_\infty$ is, the easier the GIC becomes selection consistent. Since the gradient of the exponential loss is unbounded, we can conjecture that the exponential loss is worse in variable selection than the logistic loss whose gradient is bounded.

The results in this paper can be extended to more complicated models such as correlated errors and link misspecification in generalized linear models. For the quadratic loss with the AR(1) errors, we have illustrated how the correlations in errors affect the conditions for the consistency of the GIC. Similar results could be derived for other losses and other error distributions as long as the convergence rate of $|a_n|_\infty$ is available. For link misspecification in generalized linear models, $\hat{\beta}^o$ may be asymptotically biased, but as long as $\sigma(\hat{\beta}^o)$ is equal to the set of true signal covariates, the GIC can be consistent (see, e.g., [21] for link misspecification).

APPENDIX

A.1. Proof of Proposition 1.

LEMMA 2. *Let*

$$V(\beta) = R_n(\beta) - R_n(\tilde{\beta}_n) + \sum_{i=1}^{n} \mathbf{x}_i'(\beta - \tilde{\beta}_n)\big(2\tau - 2I(\varepsilon_i < 0)\big)/n - \mathrm{E}\big\{R_n(\beta) - R_n(\tilde{\beta}_n)\big\}.$$

Then

$$\Pr\left\{\sup_{\beta\in\Theta_n} \frac{|V(\beta)|}{|\beta|_0} > \delta_n\right\} \le 2\exp\left\{-c_1\frac{n\delta_n^2}{\zeta_n\eta_n^2} + s_n\log p_n + s_n\log\big(1 + c_2\zeta_n\eta_n/(\delta_n s_n^{1/2})\big)\right\} \tag{A.1}$$

for some positive constants c_1 and c_2.

PROOF. The proof can be carried out similar to Lemma 3.2 of He and Shi [14]. From the choice of $\rho_\tau(u)$, it suffices to prove Lemma 2 with $\tau = 1/2$. Let $\mathbf{z}_i = \eta_n \mathbf{x}_i$ and $\theta = \eta_n^{-1}(\beta - \tilde{\beta}_n)$. Let $\Theta_{n,\pi} = \Theta_n \cap \{\beta : \sigma(\beta) = \pi\}$. Note that

$$\sup_{\beta\in\Theta_n} \frac{|V(\beta)|}{|\beta|_0} \le \sup_{\pi:\pi\subset\{1,\dots,p_n\},|\pi|\le s_n} \frac{1}{|\pi|} \sup_{\beta\in\Theta_{n,\pi}} |V(\beta)|.$$

In turn,

$$\sup_{\beta\in\Theta_{n,\pi}} |V(\beta)| \le \sup_{\theta:\sigma(\theta)\subset\pi\cup\pi^*,\|\theta\|\le 1} |W(\theta)|,$$

where $W(\theta) = \sum_{i=1}^n W_i(\theta)/n$ and

$$W_i(\theta) = |\varepsilon_i - \mathbf{z}_i'\theta| - |\varepsilon_i| + \mathbf{z}_i'\theta\big(2\tau - 2I(\varepsilon_i < 0)\big) - \mathrm{E}\big\{|\varepsilon_i - \mathbf{z}_i'\theta| - |\varepsilon_i|\big\}.$$

For given π, let $\pi^+ = \pi \cup \pi^*$. We can cover $\{\mathbf{v} \in R^{|\pi^+|} : \|\mathbf{v}\| \le 1\}$ with the $K_{\pi,n}$ many balls $\Gamma_1, \dots, \Gamma_{K_{\pi,n}}$ of radius $r_n = \delta_n|\pi^+|/(12\zeta_n\eta_n\sqrt{|\pi^+|})$ with centers $\tau_1, \dots, \tau_{K_{\pi,n}}$. According to Lemma 2.5 of van de Geer [13], $K_{\pi,n} \le (1 + 4/r_n)^{|\pi^+|}$. For θ with $\sigma(\theta) \subset \pi^+$, we define θ_j^π, $j = 1, \dots, K_{\pi,n}$ as $\theta_{j,\pi^+}^\pi = \tau_j$ and $\theta_{j,\pi^{+c}}^\pi = 0$. Then we have

$$\begin{aligned}\min_{j=1,\dots,K_{\pi,n}} |W(\theta) - W(\theta_j^\pi)| &\le 3 \max_i \|\mathbf{z}_{i,\pi^+}\| \min_j \|\theta - \theta_j\| \\ &\le 3\sqrt{|\pi^+|}\eta_n\zeta_n r_n \le \frac{\delta_n}{4}|\pi|.\end{aligned}$$

Hence, for the given π,

$$\begin{aligned}&\Pr\Big\{\sup_{\theta:\sigma(\theta)\subset\pi\cup\pi^*,\|\theta\|\le 1} |W(\theta)| \ge \delta_n|\pi|\Big\} \\ &\quad\le \Pr\Big\{\max_{j=1,\dots,K_{\pi,n}} |W(\theta_j^\pi)| \ge \frac{\delta_n|\pi|}{2}\Big\} \\ &\quad\le \sum_{j=1}^{K_{\pi,n}} \Pr\Big\{|W(\theta_j^\pi)| \ge \frac{\delta_n|\pi|}{2}\Big\},\end{aligned}$$

which implies

$$\Pr\Big\{\sup_{\beta\in\Theta_n} \frac{|V(\beta)|}{|\beta|_0} > \delta_n\Big\} \le \sum_{\pi:\pi\subset\{1,\dots,p_n\},|\pi|\le s_n} \sum_{j=1}^{K_{\pi,n}} \Pr\Big\{|W(\theta_j^\pi)| \ge \frac{\delta_n|\pi|}{2}\Big\}. \tag{A.2}$$

Since $|\varepsilon_i - \mathbf{z}_i'\theta| - |\varepsilon_i| \le |\mathbf{z}_i'\theta|$, we have

$$|W_i(\theta_j^\pi)| \le 3|\mathbf{z}_i'\theta_j^\pi| \le 3\zeta_n\sqrt{|\pi^+|}\eta_n.$$

Hoeffding's inequality implies that

$$\Pr\left\{|W(\theta_j^{\pi})| \geq \frac{\delta_n|\pi|}{2}\right\} \leq 2\exp\left(-\frac{2}{36}\frac{n\delta_n^2|\pi|^2}{\zeta_n\eta_n^2|\pi^+|}\right) \leq 2\exp\left(-c_1\frac{n\delta_n^2}{\zeta_n\eta_n^2}\right) \tag{A.3}$$

for some $c_1 > 0$ since $|\pi^+| \leq |\pi| + q_n$ and $q_n = O(1)$. From (A.2) and (A.3), we have

$$\Pr\left\{\sup_{\beta\in\Theta_n}\frac{|V(\beta)|}{|\beta|_0} > \delta_n\right\} \leq 2\exp\left(-c_1\frac{n\delta_n^2}{\zeta_n\eta_n^2} + s_n\log p_n + 2s_n\log(1+4/r_n)\right).$$

Since $r_n \leq \delta_n\sqrt{s_n}/(12\zeta_n\eta_n)$, the proof is complete. □

LEMMA 3. *If* $\sqrt{s_n}\eta_n \leq \alpha_2/\zeta_n$,

$$\mathrm{E}\{R_n(\beta) - R_n(\tilde{\beta}_n)\} \geq b\|\beta - \tilde{\beta}_n\|^2$$

for all $\beta \in \Theta_n$.

PROOF. Knight's identity [19] implies

$$\rho_\tau(u-v) - \rho_\tau(v) = -v\big(2\tau - 2I(u<0)\big) + \int_0^v 2\big(I(u\leq s) - I(u\leq 0)\big)\,ds.$$

Hence, for $\beta \in \Theta_n$,

$$\begin{aligned}\mathrm{E}\{R_n(\beta) - R_n(\tilde{\beta}_n)\} &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}_{\varepsilon_i}\left\{\int_0^{\mathbf{x}_i'(\beta-\tilde{\beta}_n)} 2\big(I(\varepsilon_i\leq s) - I(\varepsilon_i\leq 0)\big)\,ds\right\}\\ &= \frac{1}{n}\sum_{i=1}^n \int_0^{\mathbf{x}_i'(\beta-\tilde{\beta}_n)} 2\big(F_i(s) - F_i(0)\big)\,ds\\ &\geq \Big(\inf_{|u|\leq\alpha_2} f_i(u)\Big)\frac{2}{n}\sum_{i=1}^n(\beta-\tilde{\beta}_n)'\mathbf{x}_i\mathbf{x}_i'(\beta-\tilde{\beta}_n) \geq \frac{b}{2}\|\beta-\tilde{\beta}_n\|^2,\end{aligned}$$

where the last inequality is because of the mean value theorem and $\sup_{\beta\in\Theta_n}\max_i|\mathbf{x}_i(\beta-\tilde{\beta}_n)| \leq \zeta_n\sqrt{s_n}\eta_n$. □

PROOF OF PROPOSITION 1. Lemma 2 implies that the probability of

$$\begin{aligned}&R_n(\beta) - R_n(\tilde{\beta}_n)\\ &\quad\geq \sum_{i=1}^n \mathbf{x}_i'(\beta-\tilde{\beta}_n)\big(2\tau - 2I(\varepsilon_i<0)\big)/n + \mathrm{E}\{R_n(\beta) - R_n(\tilde{\beta}_n)\} - \delta_n|\beta|_0\end{aligned}$$

converges to 1 when $n\delta_n^2/\eta_n^2 \gg s_n\log p_n$ provided $\eta_n/(\delta_n s_n^{1/2}) \ll p_n$, and hence the proof is complete by Lemma 3. □

A.2. Maximum inequality.

PROPOSITION 2.

$$\Pr\left\{\max_j \left|\frac{1}{n}\sum_{i=1}^n l^{(1)}(y_i, \mathbf{x}_i'\beta^*)x_{ij}\right| > \sqrt{\frac{\alpha \log p_n}{n}}\right\} \le 2\exp\left\{-\frac{\log p_n}{2}\left(\frac{\alpha}{\sigma_d^2 + (d/3)\zeta_n\sqrt{\alpha \log p_n/n}} - 2\right)\right\}.$$

PROOF. This is a direct consequence of the Bernstein inequality. □

A.3. Proof of Theorem 7.

LEMMA 4.

$$\sup_{\pi:|\pi|\le s_n, \pi\supset\pi^*} \|\hat\beta_\pi - \hat\beta^o\| = O_p(\sqrt{s_n \log p_n/n}).$$

PROOF. Since for $\pi \supset \pi^*$, $R_n(\hat\beta_\pi) - R_n(\hat\beta^o) < 0$, and so Lemma 1 implies $\|\hat\beta_\pi - \hat\beta^o\| \le \eta_n$. Thus, the QSR representation in Section 4.2 gives

$$0 \ge R_n(\hat\beta_\pi) - R_n(\hat\beta^o) \ge -|a_n|_\infty\sqrt{s_n}\|\hat\beta_\pi - \hat\beta^o\| + \frac{\rho}{2}\|\hat\beta_\pi - \hat\beta^o\|^2.$$

Hence, $\|\hat\beta_\pi - \hat\beta^o\| \le 2\sqrt{s_n}|a_n|_\infty/\rho = O_P(\sqrt{s_n \log p_n/n})$ since $|a_n|_\infty = O_p(\sqrt{\log p_n/n})$ by (4.14) and (4.15). □

LEMMA 5.

$$\sup_{\pi:|\pi|\le s_n, \pi\supset\pi^*} |\hat\sigma_{d,\pi}^2 - \sigma_d^2| = o_p(1).$$

PROOF. Lemma 4 implies

$$\sup_i |\mathbf{x}_i'(\hat\beta_\pi - \hat\beta^o)| \le \zeta_n\sqrt{s_n}\|\hat\beta_\pi - \hat\beta^o\| = O_p(\zeta_n s_n\sqrt{\log p_n/n}) = o_p(1). \tag{A.4}$$

Since $\|\hat\beta^o - \beta^*\| = O_p(\sqrt{q_n/n})$, we have

$$\sup_{\pi:|\pi|\le s_n, \pi\supset\pi^*} \sup_i |\mathbf{x}_i'(\hat\beta_\pi - \beta^*)| = o_p(1),$$

which completes the proof since F is continuous. □

LEMMA 6. *Let* $a_n^o = \sum_{i=1}^n l(y_i, \mathbf{x}_i'\beta^*)\mathbf{x}_i/n$. *Then for any* $\delta > 0$,

$$\Pr\left\{|a_n^o|_\infty < \sqrt{(2\sigma_d^2 + \delta)\frac{\log p_n}{n}}\right\} \to 1$$

as $n \to \infty$, *provided* $\log p_n \to \infty$.

PROOF. This is a direct consequence of Proposition 2. □

PROOF OF THEOREM 7. Since $a_{nj} = 0$ for $j \in \pi^*$, (4.13) implies

$$
\begin{aligned}
&R_n(\hat{\beta}_\pi) - R_n(\hat{\beta}^o) \\
&\geq -\sum_{j \in \pi - \pi^*} \left\{ |a_n|_\infty |\hat{\beta}_{\pi,j}| - \frac{b_n(\pi, \pi^*, d)}{2} \hat{\beta}_{\pi,j}^2 \right\} + \frac{b_n(\pi, \pi^*, d)}{2} \sum_{j \in \pi^* - \pi} \hat{\beta}_j^{o2} \qquad \text{(A.5)} \\
&\geq -|\pi - \pi^*| \frac{|a_n|_\infty^2}{2 b_n(\pi, \pi^*, d)} + \frac{b_n(\pi, \pi^*, d)}{2} \sum_{j \in \pi^* - \pi} \hat{\beta}_j^{o2}.
\end{aligned}
$$

Note that Lemmas 5 and 6 with (4.14) in Section 4.2 and conditions (A2) and (A3) imply that for any $\delta > 0$

$$
\text{(A.6)} \quad \Pr\left\{ |a_n|_\infty < \sqrt{(2\hat{\sigma}_{d,\pi}^2 + \delta) \frac{\log p_n}{n}} \text{ for all } \pi \supset \pi^* \text{ with } |\pi| \leq s_n \right\} \to 1
$$

as $n \to \infty$. Finally, (A.6) with (6.1) implies that for $0 < \delta < \delta_0$, (A.5) is greater than or equal to

$$
-|\pi - \pi^*| \hat{\lambda}(\pi, \pi^*, d, \delta) + |\pi^* - \pi| \hat{\lambda}(\pi, \pi^*, d, \delta)
$$

with probability converging to 1, which completes the proof. □

A.4. Proposition.

PROPOSITION 3. *Let $\tilde{\beta}$ be an estimator such that $|\sigma(\tilde{\beta})| \leq c s_n$ for some $c > 0$ and $\|\tilde{\beta} - \beta^*\| = O_p(\sqrt{s_n \log p_n / n})$. Under the conditions* (A1) *to* (A5), *$|J(\pi^*, d)| > c_\alpha$ for any $\alpha \in (0, 1)$ with probability converging to* 1.

PROOF. (A2) implies $\|\tilde{\beta} - \hat{\beta}^o\| = O_p(\sqrt{s_n \log p_n / n})$, and hence

$$
\begin{aligned}
|y_i - \mathbf{x}_i' \hat{\beta}^o| &\leq |y_i - \mathbf{x}_i' \tilde{\beta}| + \sup_i |\mathbf{x}_i' (\hat{\beta}^o - \tilde{\beta})| \\
&\leq |y_i - \mathbf{x}_i' \tilde{\beta}| + \zeta_n \sqrt{(c+1) s_n} \|\hat{\beta}^o - \tilde{\beta}\| \\
&= |y_i - \mathbf{x}_i' \tilde{\beta}| + o_p(1).
\end{aligned}
$$

Hence, $|y_i - \mathbf{x}_i' \tilde{\beta}| \leq \alpha d$ for any $\alpha \in (0, 1)$ implies $|y_i - \mathbf{x}_i' \hat{\beta}^o| \leq d$, and so the proof is complete. □

REFERENCES

[1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) 267–281. Akadémiai Kiadó, Budapest. MR0483125

[2] BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

[3] BROMAN, K. W. and SPEED, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 641–656. MR1979381

[4] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Series in Statistics*. Springer, Heidelberg. MR2807761

[5] CASELLA, G., GIRÓN, F. J., MARTÍNEZ, M. L. and MORENO, E. (2009). Consistency of Bayesian procedures for variable selection. *Ann. Statist.* **37** 1207–1228. MR2509072

[6] CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

[7] CHEN, J. and CHEN, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statist. Sinica* **22** 555–574. MR2954352

[8] CRAVEN, P. and WAHBA, G. (1978/79). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403. MR0516581

[9] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[10] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

[11] FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. MR3065478

[12] FOSTER, D. and GEORGE, E. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

[13] GEER, S. A. (2000). *Empirical Processes in M-estimation* **6**, Cambridge University Press, Cambridge.

[14] HE, X. and SHI, P. (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *J. Nonparametr. Statist.* **3** 299–308. MR1291551

[15] HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618. MR2469326

[16] KIM, Y., CHOI, H. and OH, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.* **103** 1665–1673. MR2510294

[17] KIM, Y. and KWON, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika* **99** 315–325. MR2931256

[18] KIM, Y., KWON, S. and CHOI, H. (2012). Consistent model selection criteria on high dimensions. *J. Mach. Learn. Res.* **13** 1037–1057. MR2930632

[19] KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

[20] LEE, E. R., NOH, H. and PARK, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.* **109** 216–229. MR3180558

[21] LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052. MR1015136

[22] PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499

[23] ROSSET, S. and ZHU, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35** 1012–1030. MR2341696

[24] SCHEETZ, T. E., KIM, K.-Y. A., SWIDERSKI, R. E., PHILP, A. R., BRAUN, T. A., KNUDTSON, K. L., DORRANCE, A. M., DIBONA, G. F., HUANG, J., CASAVANT, T. L., SHEFFIELD, V. C. and STONE, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci. USA* **103** 14429–14434.

[25] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[26] SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. MR1466682

[27] STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147. MR0356377

[28] WANG, H., LI, B. and LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 671–683. MR2749913

[29] WANG, L., KIM, Y. and LI, R. (2013). Calibrating nonconvex penalized regression in ultra-high dimension. *Ann. Statist.* **41** 2505–2536. MR3127873

[30] WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107** 214–222. MR2949353

[31] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[32] ZHANG, Y. and SHEN, X. (2010). Model selection procedure for high-dimensional data. *Stat. Anal. Data Min.* **3** 350–358. MR2726244

DEPARTMENT OF STATISTICS
SEOUL NATIONAL UNIVERSITY
SEOUL
KOREA 08826
E-MAIL: ydkim0903@gmail.com

DEPARTMENT OF STATISTICS
UNIVERSITY OF SEOUL
SEOUL
KOREA 02505
E-MAIL: jj.jeon@gmail.com