Published in final edited form as:

Technometrics. 2019 ; 61(2): 154–164. doi:10.1080/00401706.2018.1513380.

Assessing Tuning Parameter Selection Variability in Penalized Regression

Wenhao Hu^a, Eric B. Laber^a, Clay Barker^b, Leonard A. Stefanski^a

^aDepartment of Statistics, NC State University, Raleigh, NC

^bSAS Institute, Cary, NC

Abstract

Penalized regression methods that perform simultaneous model selection and estimation are ubiquitous in statistical modeling. The use of such methods is often unavoidable as manual inspection of all possible models quickly becomes intractable when there are more than a handful of predictors. However, automated methods usually fail to incorporate domain-knowledge, exploratory analyses, or other factors that might guide a more interactive model-building approach. A hybrid approach is to use penalized regression to identify a set of candidate models and then to use interactive model-building to examine this candidate set more closely. To identify a set of candidate models, we derive point and interval estimators of the probability that each model along a solution path will minimize a given model selection criterion, for example, Akaike information criterion, Bayesian information criterion (AIC, BIC), etc., conditional on the observed solution path. Then models with a high probability of selection are considered for further examination. Thus, the proposed methodology attempts to strike a balance between algorithmic modeling approaches that are computationally efficient but fail to incorporate expert knowledge, and interactive modeling approaches that are labor intensive but informed by experience, intuition, and domain knowledge. Supplementary materials for this article are available online.

Keywords

Conditional distribution; Lasso; Prediction sets

1. Introduction

Penalized estimation is a popular means of regression model fitting that is quickly becoming a standard tool among quantitative researchers working across nearly all areas of science. Examples include the Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), Elastic Net (Zou

CONTACT Eric B. Laber ebllaber@ncsu.edu.

Color versions of one or more of the figures in this article are available online at www.tandfonline.com/r/TECH.

Supplementary materials for this article are available online. Please go to <http://www.tandfonline.com/r/TECH>

Supplementary Materials

Simulation results: Simulation results for $\tau = 0.1$ and 0.2 are presented in the online supplement to this article.

Proofs and technical details: Detailed proofs are provided in the online supplement to this article.

R package: R package for proposed methods.

and Hastie 2005), and the adaptive Lasso (Zou 2006). One appealing feature of these methods is that they perform simultaneous model selection and estimation, thereby automating model-building at least partially. This is especially beneficial in settings where the number of predictors is large, precluding manual inspection of all possible models. However, a consequence is that the analyst becomes increasingly dependent on an estimation algorithm that has neither the subject-matter knowledge nor the intuition that might guide a less automated and more interactive model-building process (Henderson and Velleman 1981; Cox 2001). A hybrid approach is to use penalized estimation to construct a small subset of models, for example, the sequence of models occurring on a solution path, and then to apply interactive model-building techniques to choose a model from among these. We develop and advocate such a hybrid approach wherein a set of candidate models are identified using a solution path, and then models along this path are prioritized using their conditional probability of selection according to one or more tuning parameter selection methods. We envision this approach as being useful in at least two ways: (i) it facilitates interactive, expert-knowledge-driven exploration of high-quality candidate models even when the initial pool of models is large; and (ii) it provides valid conditional prediction sets for a data-driven tuning parameter given the observed design matrix and solution path, which is applicable for a large class of tuning parameter selection methods.

There is a vast literature on tuning parameter selection methods. Classical methods include Mallows's C_p (Mallows 1973), Akaike information criterion (AIC; Akaike 1974), Bayesian information criterion (BIC; Schwarz 1978), cross-validation, and generalized cross-validation (Golub, Heath, and Wahba 1979). More recent work on tuning parameter selection, driven by interest in high-dimensional data, includes new information-theoretic selection methods (Chen and Chen 2008; Wang, Li, and Leng 2009; Zhang, Li, and Tsai 2010; Wang and Zhu 2011; Kim, Kwon, and Choi 2012; Fan and Tang 2013; Hui, Warton, and Foster 2015) as well as resampling-based approaches (Hall, Lee, and Park 2009; Meinshausen and Bühlmann 2010; Feng and Yu 2013; Sun, Wang, and Fang 2013; Shah and Samworth 2013). The foregoing methods select a single tuning parameter and hence a single fitted model. Our goal is to quantify the stability of these methods by constructing conditional prediction sets for data-driven tuning parameters and to use these prediction sets to prioritize models for further, expert-guided exploration. Given one or more tuning parameter selection methods, we identify all models with sufficiently large conditional probability of being selected given the design matrix and observed solution path.

In Section 2, we review penalized linear regression. In Section 3, we derive exact and asymptotic estimators of the sampling distribution of a data-driven tuning parameter. We examine the performance of the proposed methods through simulation studies in Section 4. In Section 5, we illustrate the proposed methods using two data examples. A concluding discussion is given in Section 6. Technical details are relegated to the supplement materials.

2. Penalized Linear Regression

We assume that the data are generated according to the linear model $Y_i = \mathbf{X}_i^T \boldsymbol{\beta}_0 + \epsilon_i$, for $i = 1, \dots, n$, where $\epsilon_1, \dots, \epsilon_n$ are independent, identically distributed errors with expectation zero,

$\beta_0 = (\beta_{01}, \dots, \beta_{0p})^\top$, and $\mathbf{X}_1, \dots, \mathbf{X}_n$ are predictors that can be regarded as either fixed or random. Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ be the vector of responses and $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)^\top$ the design matrix with the first column equal to $\mathbf{1}^{n \times 1}$. Let $\mathbb{P}_n$ denote the empirical distribution.

We consider penalized least-square estimators

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=2}^p f_j(\beta_j; \mathbb{P}_n) \right\},$$

where $f_j(\cdot), j = 2, \dots, p$ are penalty functions. For example, $f_j(\beta_j; \mathbb{P}_n) = |\beta_j|$ corresponds to the Lasso, and $f_j(\beta_j; \mathbb{P}_n) = |\beta_j| |\hat{\beta}_{\text{ols},j}|^{-\gamma}$ corresponds to the adaptive Lasso, where $\hat{\beta}_{\text{ols},j}$ is the ordinary least-square estimator and $\gamma > 0$ is a constant.

For any $\Lambda \subseteq [0, \infty)$ define the solution path along Λ as $\hat{S}(\Lambda) = \{\hat{\beta}(\lambda) : \lambda \in \Lambda\}$; we write $\hat{S}$ to denote $\hat{S}([0, \infty))$. While the solution path along Λ may contain a continuum of coefficient vectors, it is commonly viewed as containing a finite set of unique models corresponding to each unique combination of nonzero elements of coefficients in $\hat{S}(\Lambda)$, that is, the set of models $\mathcal{M}\{\hat{S}(\Lambda)\} = \{\mathbf{M} \in \{0, 1\}^p : \mathbf{M} = \mathbb{1}_{\hat{\beta}(\lambda) \neq 0}, \text{ for some } \lambda \in \Lambda\}$. The number of models in $\mathcal{M}\{\hat{S}\}$ is typically much smaller, for example, $O_p\{\min(n, p)\}$, than the set of all of 2^p possible models. Thus, the set of models along the solution path are a natural and computationally manageable subset of models for further investigation. Standard practice is to choose a single value of the tuning parameter, say $\hat{\lambda}$, that optimizes some prespecified criterion and subsequently a single model $\mathcal{M}[\hat{S}(\{\hat{\lambda}\})]$. However, the selected tuning parameter is a random variable and there may be multiple models along the solution path where the support of the selected tuning parameter is large, for example, $\mathcal{M}\{\hat{S}(\mathcal{L}_\tau)\}$ where $\mathcal{L}_\tau$ is a τ upper-level set of the conditional distribution of $\hat{\lambda}$ given $\hat{S}$ and $\mathbf{X}$. If these models can be identified from the observed data, then they can be reported as potential candidate models or a single model can be chosen from among them using expert judgment and other factors not captured in the estimation algorithm. Also, unlikely models can be ruled out. To formalize this procedure, we consider selection methods within the framework of generalized information criterion.

Define the generalized information criterion as

$$\text{GIC}_\lambda = \log(\hat{\sigma}_\lambda^2) + w_n \hat{df}_\lambda, \quad (1)$$

where $\hat{\sigma}_\lambda^2 = n^{-1} \sum_{i=1}^n \{Y_i - \mathbf{X}_i^\top \hat{\beta}(\lambda)\}^2$, $\hat{df}_\lambda = \sum_{j=1}^p \mathbb{1}_{|\hat{\beta}_j(\lambda)| > 0}$, and w_n is a sequence of positive constants, with $w_n = \log(n)/n$ and $w_n = 2/n$ yielding BIC and AIC, respectively. We consider data-driven tuning parameters of the form $\hat{\lambda}_{\text{GIC}} = \underset{\lambda}{\operatorname{argmin}} \{\log(\hat{\sigma}_\lambda^2) + w_n \hat{df}_\lambda\}$. We focus primarily on the setting where $n > p$ as the GIC is not well-defined if $p \geq n$. However, we provide an illustrative example in Section 5 where $p > n$ wherein our method is applied.

after an initial screening step; this two-stage procedure is in line with our vision for using automated methods to identify a small set of candidate models for further consideration. We also present extensions of key distributional approximations to the setting where p diverges with n in the Appendix.

3. Estimating the Conditional Distribution of $\hat{\lambda}_{\text{GIC}}$

In this section, we characterize and derive estimators of the conditional distribution of $\hat{\lambda}_{\text{GIC}}$ given $\hat{S}$ and $\mathbf{X}$. We first show that conditioning on $\hat{S}$ and $\mathbf{X}$ is equivalent to conditioning on $\mathbf{X}^\top \mathbf{Y}$ and $\mathbf{X}$. We then show that $\hat{\lambda}_{\text{GIC}}$ is a nondecreasing function of the sum of squares error of the full model $\hat{\sigma}_0^2 = n^{-1} \sum_{i=1}^n (Y_i - \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}_{\text{ols}})^2$. Therefore, the conditional distribution of $\hat{\lambda}_{\text{GIC}}$ is completely determined by the conditional distribution of $\hat{\sigma}_0^2$.

3.1 Conditioning on the Solution Path

We assume that $f_j(\boldsymbol{\beta}_j; \mathbb{P}_n)$, $j = 2, \dots, p$ depends on the observed data only through $\mathbf{X}^\top \mathbf{Y}$ and $\mathbf{X}^\top \mathbf{X}$; this assumption is natural as $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}^\top \mathbf{Y}$ are sufficient statistics for the conditional mean of Y given $\mathbf{X}$ under the assumed linear model. Under this assumption, $\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \{ \frac{1}{2} \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} - \mathbf{Y}^\top \mathbf{X} \boldsymbol{\beta} + \lambda \sum_{j=2}^p f_j(\boldsymbol{\beta}_j; \mathbb{P}_n) \}$, from which it can be seen that the solution path is completely determined by $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}^\top \mathbf{Y}$. On the other hand, given $\hat{S}$ and $\mathbf{X}$, we can recover $\mathbf{X}^\top \mathbf{Y}$ using $\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}}(0) = \mathbf{X}^\top \mathbf{X} \{ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \} = \mathbf{X}^\top \mathbf{Y}$. Therefore, conditioning on solution path and design matrix is equivalent to conditioning on $\mathbf{X}^\top \mathbf{Y}$ and $\mathbf{X}$ (see Lemma C.1 in the Appendix).

In the case of adaptive Lasso, we assume that $\mathbf{X}$ is full column rank so that $f_j(\boldsymbol{\beta}_j; \mathbb{P}_n)$, which depends on $\hat{\boldsymbol{\beta}}_{\text{ols}, j}$, is well-defined. It can be seen that if $\mathbf{X}$ is full column rank, then the entire solution path is determined by $\mathbf{X}^\top \mathbf{X}$ and $\hat{\boldsymbol{\beta}}_{\text{ols}}$. Conditioning on the solution path is also practically relevant because it is consistent with the common practice wherein an analyst is presented with a full solution path and then proceeds to identify a model as a point along this path.

3.2 Exact Distribution of $\hat{\lambda}_{\text{GIC}} \mid (\hat{S}, \mathbf{X})$

We assume that the models along the solution path are determined by the sequence of tuning parameters $\hat{\lambda}_{(1)} < \hat{\lambda}_{(2)} < \dots < \hat{\lambda}_{(\hat{m})}$, so that $\hat{m}$ is the total number of tuning parameters to be considered. The following lemma characterizes the conditional distribution of $\hat{\lambda}_{\text{GIC}}$.

Lemma 1.—The selected tuning parameter, $\hat{\lambda}_{\text{GIC}}$, is completely determined by $(\hat{S}, \mathbf{X}, \hat{\sigma}_0^2)$.

Furthermore, assume $\| \mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\lambda) \|^2$ is a nondecreasing function of λ , write $\hat{\lambda}_{\text{GIC}} = \lambda(\hat{S}, \mathbf{X}, \hat{\sigma}_0^2)$, then for each fixed $\hat{S} = s$ and $\mathbf{X} = \mathbf{x}$, the map $\sigma^2 \mapsto \lambda(s, \mathbf{x}, \sigma^2)$ is nondecreasing.

The assumption that $\|Y - X\hat{\beta}(\lambda)\|^2$ is nondecreasing function of λ holds under mild conditions, for example, if $\sum_{j=2}^p f_j\{\hat{\beta}(\lambda); \mathbb{P}_n\}$ is a decreasing function of λ and the original penalized problem can be recast as constrained minimization problem of the form minimize $\|Y - X\beta\|^2$ subject to the constraint $\sum_{j=2}^p f_j(\beta_j; \mathbb{P}_n) \leq \sum_{j=2}^p f_j\{\hat{\beta}(\lambda); \mathbb{P}_n\}$. It is well known that Lasso satisfies this property. If the error is normally distributed, then $(n\hat{\sigma}_0^2) / \sigma_0^2$ is independent of $(\hat{S}, \mathbf{X})$ and follows a chi-square distribution with $n - p$ degrees of freedom. Therefore, the preceding lemma shows that, under normal errors, the conditional distribution of $\hat{\lambda}_{\text{GIC}}$ given $(\hat{S}, \mathbf{X})$ is a nondecreasing function of a chi-square random variable. The remaining results stated in this section do not require the assumption that $\|Y - X\hat{\beta}(\lambda)\|^2$ is nondecreasing in λ ; rather, the results are stated in terms of a finite but arbitrary sequence of tuning parameter values.

Define $\hat{D}_\lambda = \{\hat{\beta}_{\text{ols}} - \hat{\beta}(\lambda)\}^\top \mathbf{X}^\top \mathbf{X} \{\hat{\beta}_{\text{ols}} - \hat{\beta}(\lambda)\}$. For $k = 1, \dots, \hat{m}$, define $\hat{A}_k = \{i: \hat{d}f_{\hat{\lambda}(i)} < \hat{d}f_{\hat{\lambda}(k)}\}$, $\hat{B}_k = \{i: \hat{d}f_{\hat{\lambda}(i)} > \hat{d}f_{\hat{\lambda}(k)}\}$, $\hat{C}_k = \{i: i \neq k, \text{ and } \hat{d}f_{\hat{\lambda}(i)} = \hat{d}f_{\hat{\lambda}(k)}\}$, and

$$\hat{\ell}_{i,k} = \frac{\hat{D}_{\hat{\lambda}(k)} \exp\{w_n(\hat{d}f_{\hat{\lambda}(k)} - \hat{d}f_{\hat{\lambda}(i)})\} - \hat{D}_{\hat{\lambda}(i)}}{1 - \exp\{w_n(\hat{d}f_{\hat{\lambda}(k)} - \hat{d}f_{\hat{\lambda}(i)})\}},$$

for $1 \leq i, k \leq \hat{m}$,

where w_n is from Equation (1). The quantities in the foregoing definitions are all measurable with respect to $\mathbf{X}$ and $\hat{S}$ and thus, for probability statements conditional on $\mathbf{X}$ and $\hat{S}$, they are regarded as constants.

The following proposition gives the exact conditional distribution of $\hat{\lambda}_{\text{GIC}}$ given $\hat{S}$ and $\mathbf{X}$.

Proposition 1.—Define $\hat{I}_k = \mathbb{1}(\hat{D}_{\hat{\lambda}(k)} < \hat{D}_{\hat{\lambda}(i)})$, for all $i \in \hat{C}_k$ with the convention that $\hat{I}_k = 1$ if $\hat{C}_k$ is empty, and $p_k = P(\max_{i \in \hat{B}_k} \hat{\ell}_{i,k} \leq n\hat{\sigma}_0^2 \leq \min_{i \in \hat{A}_k} \hat{\ell}_{i,k} \mid \hat{S}, \mathbf{X})$. Then,

$$P(\hat{\lambda}_{\text{GIC}} = \hat{\lambda}_{(k)} \mid \hat{S}, \mathbf{X}) = \min(p_k, \hat{I}_k).$$

Provided that the conditional distribution of $\hat{\sigma}_0^2$ given $(\hat{S}, \mathbf{X})$ is known or can be consistently estimated, the preceding proposition can be used to construct conditional prediction sets for $\hat{\lambda}_{\text{GIC}}$. A $(1 - \alpha) \times 100\%$ conditional prediction set is $\{\hat{\lambda}_{(i)}: i \in \Gamma\}$, where $\sum_{i \in \Gamma} P(\hat{\lambda}_{\text{GIC}} = \hat{\lambda}_{(i)} \mid \hat{S}, \mathbf{X}) \geq 1 - \alpha$. Alternatively, as discussed previously, one can construct the τ upper-level set $\mathcal{L}_\tau = \{\hat{\lambda}_{(i)}: P(\hat{\lambda}_{\text{GIC}} = \hat{\lambda}_{(i)} \mid \hat{S}, \mathbf{X}) > \tau\}$, for any $\tau \in (0, 1)$.

Define $\hat{a}_k = \min_{i \in \hat{A}_k} \hat{\epsilon}_{i,k}$ and $\hat{b}_k = \max_{i \in \hat{B}_k} \hat{\epsilon}_{i,k}$. If the errors are normally distributed then

$$p_k = F_{\chi_{n-p}^2} \left(\frac{\hat{a}_k}{\sigma_0^2} \right) - F_{\chi_{n-p}^2} \left(\frac{\hat{b}_k}{\sigma_0^2} \right), \text{ for } \hat{a}_k \geq \hat{b}_k. \quad (2)$$

Plugging $\hat{\sigma}_0^2$ into this expression yields an estimator $\hat{p}_k$ for p_k .

Define $g_k(t) = F_{\chi_{n-p}^2}(\hat{a}_k/t) - F_{\chi_{n-p}^2}(\hat{b}_k/t)$. Then a $(1 - \alpha) \times 100\%$ projection confidence interval (Berger and Boos 1994) for p_k (Equation (2)) is

$$\left(\inf_{t \in \mathcal{C}} g_k(t), \sup_{t \in \mathcal{C}} g_k(t) \right), \quad (3)$$

Where $C = (n\hat{\sigma}_0^2/\chi_{\alpha/2, n-p}^2, n\hat{\sigma}_0^2/\chi_{1-\alpha/2, n-p}^2)$ is a $(1-\alpha) \times 100\%$ confidence interval for σ_0^2 . Thus, an estimator of $\mathcal{L}_\tau$ is

$$\widehat{\mathcal{L}}_\tau = \left\{ \hat{\lambda}_{(k)} : \sup_{t \in \mathcal{C}} g_k(t) > \tau \right\}. \quad (4)$$

Remark 1. The assumption that $\mathbf{X}$ is full rank is not necessary for Proposition 1. Note that the conclusions depend only on the quantities $\mathbf{X}\hat{\beta}_{\text{ols}}$, $\mathbf{X}\hat{\beta}(\lambda)$, and $\hat{\sigma}_0^2$, which are computable even when $\mathbf{X}$ is not full rank.

3.3 Limiting Conditional Distribution of $\hat{\sigma}_0^2$

As discussed above, if the errors are assumed to be normally distributed then exact distribution theory for $\hat{\lambda}_{\text{GIC}}$ is possible using a transformed chi-square random variable.

Here, we consider asymptotic approximations that apply more generally.

Denote the third and fourth moment of ϵ as $\mu_{3,\epsilon}$ and $\mu_{4,\epsilon}$ respectively. Define

$$\Sigma = \begin{pmatrix} \sigma_0^2 C^{-1} & \mu_{3,\epsilon} \\ \mu_{3,\epsilon}^T & \mu_{4,\epsilon} - \sigma_0^4 \end{pmatrix},$$

where $C = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T$. And write $\Phi_{p+1}(\mathbf{t})$ to denote the cumulative distribution function of a standard $(p+1)$ -dimensional multivariate normal distribution evaluated at $\mathbf{t}$. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{p+1}$ write $\mathbf{u} \leq \mathbf{v}$ to mean component-wise inequality. The following are standard

results from ordinary linear regression under common regularity conditions summarized in Section C in the Appendix (see the proof of Proposition 2).

Proposition 2.—The asymptotic joint distribution of $\hat{\beta}_{ols} - \beta_0$ and $\hat{\sigma}_0^2 - \sigma_0^2$ is multivariate normal with mean zero and covariance Σ , that is,

$$\sup_{\mathbf{t} \in \mathbb{R}^{p+1}} \left| P \left[\sqrt{n} \Sigma^{-1/2} \begin{pmatrix} \hat{\beta}_{ols} - \beta_0 \\ \hat{\sigma}_0^2 - \sigma_0^2 \end{pmatrix} \leq \mathbf{t} \right] - \Phi_{p+1}(\mathbf{t}) \right| \rightarrow 0.$$

Because we assume that $\mathbf{X}$ is full column rank, conditioning on $(\hat{S}, \mathbf{X})$ is equivalent to conditioning on $(\hat{\beta}_{ols}, \mathbf{X})$ (in the sense that they generate the same σ -algebra). Therefore to approximate the conditional distribution of $\hat{\sigma}_0^2$ given $(\hat{S}, \mathbf{X})$, we construct an estimator of Σ , say $\hat{\Sigma}$, and then use the above proposition to form a plug-in estimator of the distribution of $\hat{\sigma}_0^2$ given $(\hat{\beta}_{ols}, \mathbf{X})$. Define

$$\hat{e}_i = Y_i - \mathbf{X}_i^T \hat{\beta}_{ols}, \quad i = 1, 2, \dots, n, \quad (5)$$

and subsequently $\hat{\sigma}_0^2 = n^{-1} \sum_{i=1}^n \hat{e}_i^2$, $\hat{\mu}_{3,\epsilon} = n^{-1} \sum_{i=1}^n \hat{e}_i^3$, $\hat{\mu}_{4,\epsilon} = n^{-1} \sum_{i=1}^n \hat{e}_i^4$, $\hat{\mu}_x = n^{-1} \sum_{i=1}^n \mathbf{X}_i$, and $\hat{C} = n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T$. The estimated conditional distribution of $\hat{\sigma}_0^2$ is

$$N \left[\hat{\sigma}_0^2, \frac{1}{n} \left(\hat{\mu}_{4,\epsilon} - \hat{\sigma}_0^4 \right) - \frac{\hat{\mu}_{3,\epsilon}^2}{\hat{\sigma}_0^2} \hat{\mu}_x^T \hat{C} \hat{\mu}_x \right]. \quad (6)$$

This approximation, coupled with Proposition 1, can be used to approximate the conditional distribution of $\hat{\lambda}_{GIC}$ when a chi-squared approximation is not feasible.

Henceforth, we assume that the errors are symmetric about zero, in which case the third moment of ϵ_i , $\mu_{3,\epsilon}$ is zero, which implies $\hat{\sigma}_0^2$ is asymptotically independent of $\hat{\beta}_{ols}$.

Therefore,

$$\begin{aligned} p_k &= P(\hat{b}_k \leq n \hat{\sigma}_0^2 \leq \hat{a}_k \mid \hat{S}, \mathbf{X}) \\ &= \Phi \left(\frac{\sqrt{n}(\hat{a}_k / n - \sigma_0^2)}{\sqrt{\mu_{4,\epsilon} - \sigma_0^4}} \right) - \Phi \left(\frac{\sqrt{n}(\hat{b}_k / n - \sigma_0^2)}{\sqrt{\mu_{4,\epsilon} - \sigma_0^4}} \right) \\ &\quad + o_p(1), \text{ for } \hat{a}_k \geq \hat{b}_k, \end{aligned} \quad (7)$$

where $\mu_{4,\epsilon}$ is the fourth moment of ϵ_i . Define

$$h_k(t_1, t_2) = \Phi\left(\frac{\sqrt{n}(\hat{a}_k / n - t_1)}{\sqrt{t_2}}\right) - \Phi\left(\frac{\sqrt{n}(\hat{b}_k / n - t_1)}{\sqrt{t_2}}\right).$$

Suppose that $\mathcal{E}_y$ is a $(1 - \alpha) \times 100\%$ asymptotic confidence region for $\mu_{4,\epsilon} - \sigma_0^4$ and σ_0^2 , then

$$\left[\inf_{(t_1, t_2) \in \mathcal{E}_y} h_k(t_1, t_2), \sup_{(t_1, t_2) \in \mathcal{E}_y} h_k(t_1, t_2) \right], \quad (8)$$

is an approximate $(1 - \alpha) \times 100\%$ projection confidence interval for p_k (see Proposition C.1 in Section C of the Appendix.).

We construct the confidence set $\mathcal{E}_y$ using Wald confidence region:

$$\mathcal{E}_y = \left\{ (t_1, t_2) : \begin{pmatrix} t_1 - \hat{\sigma}_0^2 \\ t_2 - \hat{\mu}_{4,\epsilon} + \hat{\sigma}_0^4 \end{pmatrix}^\top \times \hat{V}^{-1} \begin{pmatrix} t_1 - \hat{\sigma}_0^2 \\ t_2 - \hat{\mu}_{4,\epsilon} + \hat{\sigma}_0^4 \end{pmatrix} \leq \chi_{1-\alpha, 2}^2 \right\},$$

where $\hat{V}$ is the estimated covariance matrix of $(\hat{\sigma}_0^2, \hat{\mu}_{4,\epsilon} - \hat{\sigma}_0^4)^\top$. Then the optimization problem in Equation (9) is solved using an augmented Lagrangian method (Bertsekas 2014). An estimator of $\mathcal{L}_\tau$ is

$$\widehat{\mathcal{L}}_\tau = \left\{ \hat{\lambda}_{(k)} : \sup_{(t_1, t_2) \in \mathcal{E}_y} h_k(t_1, t_2) > \tau \right\}. \quad (9)$$

Proposition 2 is stated in terms of fixed p and diverging n . We show in the Appendix that the approximation in Equation (7) remains valid in the setting $p = o(\sqrt{n})$ as well as the setting where $p = O\{\exp(n^c)\}$ provided that an appropriate screening step is applied as a first step. We illustrate this screening approach with a high-dimensional example in our simulation experiments.

3.4 Bootstrap Approximation to the Distribution of $\hat{\lambda}_{\text{GIC}} | (\hat{S}, \mathbf{X})$

In small samples, it may be preferable to estimate the conditional distribution of $\hat{\sigma}_0^2$ using the bootstrap. Let $\boldsymbol{\gamma}^{(b)} = (\gamma_1^{(b)}, \dots, \gamma_n^{(b)})^\top$ be a sample drawn with replacement from $\{\hat{e}_1, \dots, \hat{e}_n\}$. Define $\mathbf{Y}^{(b)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ols}} + (\mathbf{I} - \mathbf{P}_\mathbf{X})\boldsymbol{\gamma}^{(b)}$, where $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. This bootstrap method differs from the usual residual bootstrap in ordinary linear regression because our

goal is to estimate the conditional distribution of $\hat{\sigma}_0^2$. We accomplish this by multiplying the error vector by $(\mathbf{I} - \mathbf{P}_X)$, which ensures that $\hat{\beta}_{\text{ols}}^{(b)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}^{(b)} = \hat{\beta}_{\text{ols}}$ so that $\mathbf{Y}^{(b)}$ produces the same solution path as the original sample $\mathbf{Y}$. The conditional distribution of the tuning parameter is estimated by generating $b = 1, \dots, B$ bootstrap samples and calculating the corresponding tuning parameter for each bootstrap sample. See Proposition C.2 in Section C of the Appendix for a statement of the asymptotic equivalence between the proposed bootstrap method and the normal approximation given in Equation (6).

4. Simulation Studies

In this section, we investigate the finite-sample performance of the proposed methods using a series of simulation experiments. We focus on the Lasso tuned using BIC. Simulated datasets are generated from the model $Y_i = \mathbf{X}_i^T \beta_0 + \epsilon_i$, where ϵ_i , $i = 1, \dots, n$ are generated independently from a standard normal distribution and X_i , $i = 1, \dots, n$ are generated independently from a multivariate normal distribution with mean zero and autoregressive covariance structure, $C_{j,k} = \rho^{|j-k|}$, with $\rho = 0$ or 0.5 and $1 \leq j, k \leq 20$, or 200 . For the regression coefficients β_0 , we consider the following four settings:

Model 1: $\beta_0 = c_1 \times (1, 1, 1, 1, 0, 0, 0, 0, \dots, 0)^T$;

Model 2: $\beta_0 = c_2 \times (1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, \dots, 0)^T$;

Model 3: $\beta_0 = c_3 \times (3, 2, 1, 0, 0, 0, 0, \dots, 0)^T$;

Model 4: $\beta_0 = c_4 \times (3, 2, 1, 0, 0, 0, 0, 0, 3, 2, 1, 0, \dots, 0)^T$;

where $c_1, \dots, c_4$ are constants chosen so that the population R^2 of each model is 0.5 under the definition $R^2 = 1 - \text{var}(Y|X)/\text{var}(Y)$. For each combination of parameter settings, 10,000 datasets were generated; the bootstrap estimator was constructed using 5000 bootstrap replications.

For estimating the τ upper-level set, $\mathcal{L}_\tau = \{\lambda: P(\hat{\lambda}_{\text{GIC}} = \lambda \mid \hat{S}, \mathbf{X}) > \tau\}$, we consider:

1. (AsympNor) the plug-in estimator based on the normal approximation to the distribution of p_k ;
2. (Bootstrap) the estimator based on the bootstrap approximation to the sampling distribution of p_k as described in Section 3.4;
3. (UP1) the estimator based on a 90% projection confidence set as in Equation (4);
4. (UP2) the estimator based on a 90% projection confidence set as in Equation (10);

5. (Akaike) the estimator based on Akaike weights:

$$\widehat{\mathcal{L}}_{\tau} = \{\lambda_i: \frac{\exp(-0.5n\text{GIC}_{\lambda_i})}{\sum_{i=1}^{\widehat{m}} \exp(-0.5n\text{GIC}_{\lambda_i})} > \tau\}, \text{ with } w_n = 2/n \text{ (Burnham and Anderson$$

2003);

6. (ApproxPost) the estimator based on approximate posterior distribution:

$$\widehat{\mathcal{L}}_{\tau} = \{\lambda_i: \frac{\exp(-0.5n\text{GIC}_{\lambda_i})}{\sum_{i=1}^{\widehat{m}} \exp(-0.5n\text{GIC}_{\lambda_i})} > \tau\}, \text{ with } W_n = \log(n)/n \text{ (Burnham and$$

Anderson 2003).

We define the performance of these estimators in terms of their true and false discovery rates. Provided that $\mathcal{L}_{\tau}$ is nonempty, define the true discovery rate of an estimator $\widehat{\mathcal{L}}_{\tau}$ as

$$\text{TDR}(\widehat{\mathcal{L}}_{\tau}) = \mathbb{E} \left[\frac{\#\{\widehat{\mathcal{L}}_{\tau} \cap \mathcal{L}_{\tau}\}}{\#\widehat{\mathcal{L}}_{\tau}} \right],$$

where # denotes the number of elements in a set. Provided $\widehat{\mathcal{L}}_{\tau}$ is nonempty with probability one, define the false discovery rate of an estimator $\widehat{\mathcal{L}}_{\tau}$ as

$$\text{FDR}(\widehat{\mathcal{L}}_{\tau}) = 1 - \mathbb{E} \left[\frac{\#\{\widehat{\mathcal{L}}_{\tau} \cap \mathcal{L}_{\tau}\}}{\#\widehat{\mathcal{L}}_{\tau}} \right].$$

Here, we present results for $\tau = 0.05$; results for $\tau = 0.1$ and $\tau = 0.2$ are presented in the supplemental materials. The results for $p = 20$, $n = 50$ and $p = 100$, $n = 200$ are presented in Tables 1 and 2, respectively. AsympNor and the Bootstrap perform similarly with a TDR above 0.9 and an FDR below 0.10. As expected, methods based on the upper bound of the confidence interval achieve higher TDR but at the price of higher FDR. Methods based on Akaike weights and approximate posterior have the worst performances in terms of discovery rate of $\mathcal{L}_{\tau}$. The poor performance is not surprising as these methods were not designed for conditional inference.

We also evaluated the coverage of the proposed confidence intervals based on assumption of normality assumption as well as the asymptotic approximation. In calculating the coverage probabilities, we restricted calculations to the set $(\lambda : 0.9999 > P(\hat{\lambda}_{\text{GIC}} = \lambda \mid \hat{S}, \mathbf{X}) > 0.0001)$.

Nominal coverage is set at 0.90. The results are presented in Table 3. The confidence intervals based on normality (Equation (3)) achieve nominal coverage in all cases. The confidence intervals based on an asymptotic approximation undercover slightly, though coverage approaches nominal levels as n increases.

5. Illustrative Data Examples

In this section, we apply the proposed methods to two datasets. The first dataset informs the relationship of pollution and other factors related to urban-living, and to age-adjusted mortality (McDonald and Schwing 1973; Luo, Stefanski, and Boos 2006); and the second regards the relationship between gene mutations and drug resistance level (Rhee et al. 2006). We demonstrate that reporting a single model may not be appropriate in these two examples and that the proposed methods have the potential to identify interesting models warranting further examination. As in the simulation experiments, we consider the Lasso estimator tuned using BIC.

5.1 Pollution and Mortality

As our first illustrative example we consider data on mortality rates recorded in 60 metropolitan areas. Prior analyses of these data focused on the regression of age-adjusted mortality on 15 predictors that are grouped into three broad categories: weather, socioeconomic factors, and pollution. A copy of the dataset and a detailed description of each predictor are provided in the supplemental materials.

Ignoring uncertainty in the tuning parameter selection, the LASSO estimator tuned using BIC leads to a model with six variables, *Percent Non-White*, *Education*, *SO₂ Pollution Potential*, *Precipitation*, *Mean January Temperature*, and *Population Per Mile*. However, the estimated conditional sampling distribution of the tuning parameter indicates that a larger model with eight predictors is approximately equally probable. Figure 1 displays the solution path with the estimated selection probabilities using both the asymptotic normal and bootstrap approximations. The estimated conditional distribution of the tuning parameter is displayed in Table 4; 90% confidence intervals based on Equations (3) and (9) are presented. The LASSO coefficient estimates are presented in Table B.1 of the Appendix.

We further investigate the model with eight predictors. The eighth predictor added to the model is *Mean July Temperature*. Fitting a simple linear regression of mortality on *Mean July Temperature* yields a *p*-value of 0.03 compared with a *p*-value of 0.81 in the regression of mortality on *Mean January Temperature*. This is in line with Katsouyanni et al. (1993) who concluded that high temperatures are related to the mortality rate. It can be seen that in this case, reporting a single model may not be appropriate. Rather, it may be more informative to report the two models that contain essentially all of the mass of the conditional distribution of the tuning parameter.

For comparison, results from forward stepwise regression are presented in Figure 2. The smallest BIC value corresponds to Step 5 of the procedure which corresponds to a model with the five predictors: *Percent Non-White*, *Education*, *Mean January Temperature*, *SO₂ Pollution Potential*, and *Precipitation*. This model is smaller than the model selected by LASSO tuned using BIC. This may be because forward stepwise regression is greedy in that at each step it seeks a variable that captures maximum variation in the residuals. Thus, if a candidate variable is correlated with those selected in previous steps, it may be difficult to see the improvement in the fitted model. In such cases, it might be preferable to use the solution path to generate a candidate set of models.

5.2 ATV Drug Resistance

Our second example considers mutations that affect resistance to Atazanavir (ATV), a protease inhibitor for HIV (Rhee et al. 2006; Barber et al. 2015). After preprocessing, the dataset contains 328 observations and 361 predictors of gene mutations. The response is a measure of drug resistance for ATV. Because $p > n$ in this example, we use 100 observations for screening 50 important predictors, ranked by Pearson correlation with response. We then fit a linear model using the Lasso applied to the remaining 228 observations with the 50 important predictors selected at screening.

The estimated conditional distribution of the tuning parameter and 90% confidence intervals are presented in Table 5; the estimated distribution is overlaid on the solution path in Figure 3. It can be seen that the estimated distribution of the tuning parameter mainly favors two models.

Similar to Barber et al. (2015), we evaluate candidate models based on treatment selected mutations (TSM) panels, which provide a surrogate for the true important mutations. The model minimizing BIC contains 15 variables, while two of them correspond to the same mutation. This leads to 14 unique mutation locations, and four locations are potential false discoveries (as assessed by TSM); see Table 5. Therefore, the surrogate-based estimated false discovery rate is $4/14 \approx 0.29$. For tuning parameter $\lambda = 1101.71$, 11 unique positions are identified, and two locations are potential false discoveries. Tuning parameter $\lambda = 1347.92$ leads to nine unique locations with one potential false discovery. The corresponding surrogate-based estimated false discovery rate is $1/9 \approx 0.11$, a decrease of 20% compared with the model minimizing BIC. Thus, in this case it might not be appropriate to report the single model selected by BIC.

6. Conclusion

We proposed two simple procedures for estimating the conditional distribution of a data-driven tuning parameter in penalized regression given the observed solution path and design matrix. Our objective was to quantify the stability of selected models and thereby identify a set of potential models for consideration by domain experts. A plot of the solution path with the estimated selection probabilities or upper confidence bounds overlaid, for example, Figures 1 and 3, is one means of easily conveying uncertainty in the tuning parameter and identifying models that warrant additional investigation. It is noteworthy that in both examples the identified sets of likely models are not contiguous in size. Thus, our methods provide a theoretically motivated, confidence-set-based alternative to the practice of considering models *near in size* to the BIC-chosen LASSO model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

The authors gratefully acknowledge funding from the National Science Foundation (DMS-1555141, DMS-1557733, DMS-1513579) and the National Institutes of Health (P01 CA142538).

Appendix A: Simulation Results

Table A.1.

Discovery rate for $p = 20$, $n = 50$; $\tau = 0.1$.

Model	ρ	AsympNor		Bootstrap		UP1		UP2		Akaike		ApproxPost	
		TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR
1	0	0.90	0.10	0.90	0.10	1.00	0.33	0.99	0.37	0.42	0.80	0.89	0.48
	0.5	0.92	0.08	0.92	0.08	1.00	0.30	0.99	0.33	0.48	0.78	0.93	0.48
2	0	0.87	0.11	0.88	0.11	0.99	0.33	0.99	0.36	0.35	0.81	0.81	0.48
	0.5	0.89	0.10	0.89	0.11	1.00	0.35	0.99	0.36	0.45	0.78	0.87	0.47
3	0	0.91	0.09	0.92	0.09	1.00	0.29	0.99	0.34	0.44	0.80	0.93	0.48
	0.5	0.92	0.07	0.93	0.08	1.00	0.26	0.99	0.30	0.50	0.78	0.96	0.49
4	0	0.89	0.10	0.90	0.10	1.00	0.34	0.99	0.37	0.39	0.80	0.87	0.48
	0.5	0.90	0.10	0.91	0.10	1.00	0.34	0.99	0.35	0.46	0.79	0.91	0.48

Table A.2.

Discovery rate for $p = 100$, $n = 200$; $\tau = 0.1$.

Model	ρ	AsympNor		Bootstrap		UP1		UP2		Akaike		ApproxPost	
		TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR
1	0	0.97	0.03	0.97	0.03	1.00	0.13	1.00	0.14	0.07	0.97	0.99	0.42
	0.5	0.97	0.02	0.97	0.02	1.00	0.10	1.00	0.10	0.15	0.93	1.00	0.40
2	0	0.95	0.04	0.95	0.04	1.00	0.17	1.00	0.18	0.03	0.99	0.98	0.42
	0.5	0.96	0.04	0.96	0.04	1.00	0.14	1.00	0.15	0.07	0.97	0.99	0.42
3	0	0.97	0.03	0.97	0.03	1.00	0.12	1.00	0.13	0.09	0.96	0.99	0.42
	0.5	0.98	0.02	0.98	0.02	1.00	0.09	1.00	0.10	0.17	0.93	1.00	0.40
4	0	0.96	0.04	0.96	0.04	1.00	0.14	1.00	0.15	0.04	0.98	0.99	0.42
	0.5	0.97	0.03	0.97	0.03	1.00	0.12	1.00	0.13	0.09	0.96	0.99	0.42

Table A3.

Discovery rate for $p = 20$, $n = 50$; $\tau = 0.2$.

Model	ρ	AsympNor		Bootstrap		UP1		UP2		Akaike		ApproxPost	
		TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR
1	0	0.89	0.11	0.89	0.11	0.99	0.34	0.99	0.39	0.25	0.69	0.84	0.22
	0.5	0.91	0.08	0.91	0.08	1.00	0.29	0.99	0.34	0.29	0.65	0.90	0.22
2	0	0.86	0.13	0.86	0.13	0.99	0.36	0.99	0.41	0.22	0.73	0.77	0.23
	0.5	0.88	0.11	0.88	0.12	0.99	0.37	0.99	0.40	0.28	0.65	0.83	0.23
3	0	0.91	0.09	0.91	0.09	1.00	0.30	0.99	0.35	0.26	0.68	0.89	0.23
	0.5	0.92	0.07	0.92	0.08	1.00	0.26	0.99	0.31	0.31	0.64	0.92	0.23
4	0	0.88	0.11	0.88	0.11	0.99	0.35	0.99	0.40	0.23	0.71	0.83	0.22

Model	ρ	AsympNor		Bootstrap		UP1		UP2		Akaike		ApproxPost	
		TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR
4	0.5	0.89	0.10	0.90	0.10	0.99	0.34	0.99	0.38	0.28	0.66	0.86	0.23

Table A.4.

Discovery rate for $p = 100$, $n = 200$; $\tau = 0.2$.

Model	ρ	AsympNor		Bootstrap		UP1		UP2		Akaike		ApproxPost	
		TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR
1	0	0.97	0.03	0.97	0.03	1.00	0.13	1.00	0.13	0.03	0.94	0.98	0.22
	0.5	0.97	0.03	0.97	0.03	1.00	0.10	1.00	0.11	0.07	0.88	0.99	0.21
2	0	0.95	0.05	0.95	0.05	1.00	0.17	1.00	0.18	0.01	0.98	0.96	0.21
	0.5	0.96	0.04	0.96	0.04	1.00	0.14	1.00	0.15	0.03	0.93	0.98	0.21
3	0	0.97	0.03	0.97	0.03	1.00	0.12	1.00	0.12	0.04	0.91	0.99	0.21
	0.5	0.98	0.03	0.98	0.03	1.00	0.09	1.00	0.10	0.09	0.86	0.99	0.20
4	0	0.96	0.04	0.96	0.04	1.00	0.15	1.00	0.16	0.02	0.96	0.97	0.21
	0.5	0.97	0.03	0.97	0.03	1.00	0.12	1.00	0.13	0.04	0.92	0.99	0.21

Appendix B: Additional Results for Real-Data

Table B.1.

Coefficient estimates for pollution and mortality data; columns BIC-8 and BIC-6 correspond to lasso estimates with eight and six predictors, respectively; columns Refit-8 and Refit-6 correspond to refitting coefficient with eight and six predictors; column FS-5 corresponds to model selected by forward stepwise regression.

Variable	Lasso				
	BIC-8	Refit-8	BIC-6	Refit-6	FS-5
Mean annual precipitation	14.95	17.76	11.78	14.85	14.86
Mean January temperature	-12.04	-14.26	-8.80	-16.61	-16.50
Mean July temperature	-5.80	-11.71	0	0	0
Median school years	-8.89	-8.05	-9.99	-9.75	-10.80
Pct of housing units with facilities	-2.55	-4.69	0	0	0
Population per square mile	5.15	7.11	2.62	6.039	0
Pct of non-White	35.38	39.98	30.04	36.98	36.27
Pollution potential of sulfur dioxide	14.47	14.92	13.66	15.51	18.00

Appendix C: Proof and Technical Details

Lemma C.1.

If the penalty function $f_j(\beta_j; \mathbb{P}_n)$ depends on the data only through $\mathbf{X}^\top \mathbf{Y}$ and $\mathbf{X}$, then the distribution of $\hat{\lambda}_{\text{GIC}}$ conditional on $\hat{S}$ and $\mathbf{X}$ is equal to the distribution of $\hat{\lambda}_{\text{GIC}}$ conditional on $\mathbf{X}^\top \mathbf{Y}$ and $\mathbf{X}$.

Proof. $\hat{\beta}(\lambda) = \argmin_{\beta} \{ \frac{1}{2} \beta^\top \mathbf{X}^\top \mathbf{X} \beta - \mathbf{Y}^\top \mathbf{X} \beta + \sum_{j=2}^p f_j(\beta_j; \mathbb{P}_n) \}$, from which it can be seen that the solution path is completely determined by $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}^\top \mathbf{Y}$. On the other hand, given $\hat{S}$ and $\mathbf{X}$, we can recover $\mathbf{X}^\top \mathbf{Y}$ using $\mathbf{X}^\top \mathbf{X} \hat{\beta}(0) = \mathbf{X}^\top \mathbf{X} \{ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \} = \mathbf{X}^\top \mathbf{Y}$. $\square$

Lemma C.2.

Suppose that $b(\cdot)$ and $c(\cdot)$ are nonnegative valued functions defined on $[0, \infty]$ such that $b(\lambda)$ is a nondecreasing for $\lambda \geq 0$ with $b(0) = 0$. For $x \geq 0$, define

$$H(x, \lambda) = \log \{x + b(\lambda)\} + c(\lambda)$$

and

$$\lambda(x) = \argmin_{\lambda} H(x, \lambda).$$

Then $\lambda(x)$ is nondecreasing in $x \geq 0$.

Proof of Lemma C.2. Supposed $x_1 \leq x_2$, we need to show that $\lambda(x_1) \leq \lambda(x_2)$. First, consider the difference of $H(x_2, \lambda)$ and $H(x_1, \lambda)$

$$\begin{aligned} H(x_2, \lambda) - H(x_1, \lambda) &= \log \{x_2 + b(\lambda)\} - \log \{x_1 + b(\lambda)\} \\ &= \log \left\{ 1 + \frac{x_2 - x_1}{x_1 + b(\lambda)} \right\}, \end{aligned}$$

which is nonnegative for every λ and nonincreasing in λ . Therefore if $\lambda(x_1) = \argmin_{\lambda} H(x_1, \lambda)$, it follows that

$$\begin{aligned} \lambda(x_2) &= \argmin_{\lambda} H(x_2, \lambda), \\ &= \argmin_{\lambda} \left[H(x_1, \lambda) + \log \left\{ 1 + \frac{x_2 - x_1}{x_1 + b(\lambda)} \right\} \right] \\ &\geq \lambda(x_1). \end{aligned}$$

The last inequality follows from that $\log \{1 + (x_2 - x_1)/(x_1 + b(\lambda))\}$ is nonnegative and nonincreasing with respect to λ . $\square$

Proof of Lemma 1. Recall that the information criterion can also be expressed as

$$\begin{aligned}
\text{GIC}_{\lambda} &= \log \left(\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)\|^2}{n} \right) + w_n \hat{df}_{\lambda} \\
&= \log \left(\frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ols}} + \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ols}} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)\|^2}{n} \right) + w_n \hat{df}_{\lambda} \\
&= \log \left(\hat{\sigma}_0^2 + \frac{D_{\lambda}}{n} \right) + w_n \hat{df}_{\lambda},
\end{aligned}$$

where $D_{\lambda} = \{\hat{\boldsymbol{\beta}}_{\text{ols}} - \hat{\boldsymbol{\beta}}(\lambda)\}^T \mathbf{X}^T \mathbf{X} \{\hat{\boldsymbol{\beta}}_{\text{ols}} - \hat{\boldsymbol{\beta}}(\lambda)\}$.

Because D_{λ} is a deterministic function of λ conditional on the solution path and design matrix, the only variability in $\hat{\lambda}_{\text{GIC}}$ is due to $\hat{\sigma}_0^2$. Therefore, $\hat{\lambda}_{\text{GIC}} | (\hat{\boldsymbol{\beta}}_{\text{ols}}, \mathbf{X})$ is a function of $\hat{\sigma}_0^2$.

Then monotonicity comes immediately by observing that D_{λ} is a non-decreasing function for $\lambda \geq 0$ with $D(0) = 0$ and invoking Lemma C.2. $\square$

Lemma C.3.

Let $\hat{\mathbf{S}}$ and $\mathbf{X}$ be fixed. If $\hat{df}_{\hat{\lambda}_{(k)}} < \hat{df}_{\hat{\lambda}_{(i)}}$, then $\text{GIC}_{\hat{\lambda}_{(i)}} \leq \text{GIC}_{\hat{\lambda}_{(k)}}$ iff $n\hat{\sigma}_0^2 \leq \hat{\ell}_{i,k}$ and if

$\hat{df}_{\hat{\lambda}_{(k)}} > \hat{df}_{\hat{\lambda}_{(i)}}$, then $\text{GIC}_{\hat{\lambda}_{(i)}} \leq \text{GIC}_{\hat{\lambda}_{(k)}}$ iff $n\hat{\sigma}_0^2 \geq \hat{\ell}_{i,k}$.

Proof of Lemma C.3. Consider the case $\hat{df}_{\hat{\lambda}_{(k)}} < \hat{df}_{\hat{\lambda}_{(i)}}$,

$$\begin{aligned}
p(\text{GIC}_{\hat{\lambda}_{(i)}} &\leq \text{GIC}_{\hat{\lambda}_{(k)}} \mid \hat{S}, \mathbf{X}) \\
&= P \left[\log \left[\hat{\sigma}_0^2 + \frac{D_{\hat{\lambda}_{(i)}}}{n} \right] + w_n \widehat{df}_{\hat{\lambda}_{(i)}} \right. \\
&\quad \left. \leq \log \left[\hat{\sigma}_0^2 + \frac{D_{\hat{\lambda}_{(k)}}}{n} \right] + w_n \widehat{df}_{\hat{\lambda}_{(k)}} \mid \hat{S}, \mathbf{X} \right] \\
&= P \left[\log \left[\hat{\sigma}_0^2 + \frac{D_{\hat{\lambda}_{(i)}}}{n} \right] - \log \left[\hat{\sigma}_0^2 + \frac{D_{\hat{\lambda}_{(k)}}}{n} \right] \right. \\
&\quad \left. \leq w_n \left[\widehat{df}_{\hat{\lambda}_{(k)}} - \widehat{df}_{\hat{\lambda}_{(i)}} \right] \mid \hat{S}, \mathbf{X} \right] \\
&= P \left[\frac{n\hat{\sigma}_0^2 + D_{\hat{\lambda}_{(i)}}}{n\hat{\sigma}_0^2 + D_{\hat{\lambda}_{(k)}}} \leq \exp\{w_n(\widehat{df}_{\hat{\lambda}_{(k)}} - \widehat{df}_{\hat{\lambda}_{(i)}})\} \mid \hat{S}, \mathbf{X} \right] \\
&= P \left(\frac{D_{\hat{\lambda}_{(k)}}}{n\hat{\sigma}_0^2} \leq \frac{\exp\{w_n(\widehat{df}_{\hat{\lambda}_{(k)}} - \widehat{df}_{\hat{\lambda}_{(i)}})\} - D_{\hat{\lambda}_{(i)}}}{1 - \exp\{w_n(\widehat{df}_{\hat{\lambda}_{(k)}} - \widehat{df}_{\hat{\lambda}_{(i)}})\}} \mid \hat{S}, \mathbf{X} \right).
\end{aligned}$$

The case $\widehat{df}_{\hat{\lambda}_{(k)}} > \widehat{df}_{\hat{\lambda}_{(i)}}$ follows by a similar argument. $\square$

Proof of Proposition 1. The proof follows from the fact that

$$\text{GIC}_{\hat{\lambda}_{(k)}} < \text{GIC}_{\hat{\lambda}_{(i)}} \text{ for all } i \in A_k \cup B_k \text{ if and only if } \max_{i \in \hat{B}_k} \hat{\ell}_{i,k} \leq n\hat{\sigma}_0^2 \leq \max_{i \in \hat{A}_k} \hat{\ell}_{i,k}.$$

$\square$

To prove Proposition 2, we assume:

(A1F): under a fixed design $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top = C$, $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{X}_i = \mu_x$, where $C \in \mathbb{R}^{p \times p}$ is nonnegative definite and $\mu_{xx} \in \mathbb{R}^p$,

(A1R): under a random design, with probability one, $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top = C$, and $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{X}_i = \mu_x$, where $C \in \mathbb{R}^{p \times p}$ is nonnegative definite and $\mu_x \in \mathbb{R}^p$,

(A2): $\mathbb{E} \epsilon_i^4 < \infty$.

Under assumptions (A1F) and (A2), we have the following well-known results, which facilitate the proof of Proposition 2.

Lemma C.4.

$$\hat{\boldsymbol{\beta}}_{\text{ols}} \xrightarrow{\text{as}} \boldsymbol{\beta}_0;$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}_{p \times 1}, C^{-1}).$$

Proof of Proposition 2. First consider the fixed design model. Let

$$\psi(Y_i, \mathbf{X}_i, \boldsymbol{\beta}, \sigma^2) = \begin{pmatrix} (Y_i - \mathbf{X}_i^T \boldsymbol{\beta}) \mathbf{X}_i \\ (Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 - \sigma^2 \end{pmatrix}.$$

Then $(\hat{\boldsymbol{\beta}}_{\text{ols}}^T, \hat{\sigma}^2)^T$ is a solution to the equation

$$\sum_{i=1}^n \psi(Y_i, \mathbf{X}_i, \boldsymbol{\beta}, \sigma^2) = \mathbf{0}.$$

A Taylor series expansion around the true value $(\boldsymbol{\beta}_0, \sigma_0^2)$ results in

$$\begin{aligned} \sum_{i=1}^n \psi(Y_i, \mathbf{X}_i, \hat{\boldsymbol{\beta}}_{\text{ols}}, \hat{\sigma}^2) &= \sum_{i=1}^n \psi(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0, \sigma_0^2) \\ &+ \sum_{i=1}^n \psi'(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0, \sigma_0^2) \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}_0 \\ \hat{\sigma}^2 - \sigma_0^2 \end{pmatrix} + R_n, \end{aligned}$$

where ψ' is the derivative of ψ and

$$R_n = \sum_{i=1}^n \begin{pmatrix} \mathbf{0}_{p \times 1} \\ (\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}_0)^T \mathbf{X}_i \mathbf{X}_i^T (\hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}_0) \end{pmatrix}.$$

Rearranging it leads to

$$\begin{aligned} &\left\{ -\frac{1}{n} \sum_{i=1}^n \psi'(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0, \sigma_0^2) \right\} \sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta}_0 \\ \hat{\sigma}^2 - \sigma_0^2 \end{pmatrix} \\ &= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0, \sigma_0^2) \right\} + R_n / \sqrt{n}. \end{aligned}$$

Because $-\psi'(Y_i, \mathbf{X}_i, \boldsymbol{\beta}_0, \sigma_0^2) = \begin{pmatrix} \mathbf{X}_i \mathbf{X}_i^T & \mathbf{0}_{p \times 1} \\ 2\mathbf{X}_i^T (Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_0) & 1 \end{pmatrix}$, it follows that

$$-\frac{1}{n} \sum_{i=1}^n \psi'(Y_i, \mathbf{X}_i, \beta_0, \sigma_0^2) \xrightarrow{p} \begin{pmatrix} \mathbf{C} & \mathbf{0}_{p \times 1} \\ \mathbf{0}_{1 \times p} & 1 \end{pmatrix}$$

by consistency of $\hat{\beta}_{\text{ols}}$.

Then by the multivariate Lindberg–Feller CLT,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(Y_i, \mathbf{X}_i, \beta_0, \sigma_0^2) \xrightarrow{d} N \left[\begin{pmatrix} 0_{p \times 1} \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 \mathbf{C} & \mu_x \mu_{3,\epsilon} \\ \mu_x \mu_{3,\epsilon} & \mu_{4,\epsilon} - \sigma_0^4 \end{pmatrix} \right].$$

Finally, $R_n / \sqrt{n}$ is $o_p(1)$ as of

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^n (\hat{\beta}_{\text{ols}} - \beta_0)^T \mathbf{X}_i \mathbf{X}_i^T (\hat{\beta}_{\text{ols}} - \beta_0) \\ &= \sqrt{n} (\hat{\beta}_{\text{ols}} - \beta_0)^T \left\{ \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \right\} (\hat{\beta}_{\text{ols}} - \beta_0). \end{aligned}$$

Therefore by Slutsky's theorem,

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_{\text{ols}} - \beta_0 \\ \hat{\sigma}_0^2 - \sigma_0^2 \end{pmatrix} \xrightarrow{d} N \left[\begin{pmatrix} 0_{p \times 1} \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 \mathbf{C}^{-1} & \mu_x \mu_{3,\epsilon} \\ \mu_x \mu_{3,\epsilon} & \mu_{4,\epsilon} - \sigma_0^4 \end{pmatrix} \right]. \quad (\text{C.1})$$

Then, for the random design, because $\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T = \mathbf{C}$ and

$\lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbf{X}_i = \mu_x$ almost surely, assumption A1F holds for almost every sequence $\mathbf{x}_1, \mathbf{x}_2, \dots$. Therefore Equation (C.1) holds for almost every sequence $\mathbf{x}_1, \mathbf{x}_2, \dots$. $\square$

Lemma C.5.

Suppose $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are normally distributed with mean zero and variance σ_0^2 , then

plugging estimator $\hat{p}_k = F_{\chi_{n-p}^2}(\hat{a}_k / \hat{\sigma}_0^2) - F_{\chi_{n-p}^2}(\hat{b}_k / \hat{\sigma}_0^2)$ need not be a consistent estimator

for p_k , where $\hat{a}_k$ and $\hat{b}_k$ are defined in Section 3.2.

Proof. We show that there exist sequences for which $\hat{p}_k$ is not consistent for p_k . Considering the case $\mathbf{X}^T \mathbf{X} = \mathbf{n} \times \mathbf{I}_2$ with $\mathbf{X}_1 = \mathbf{1}^{n \times 1}$ and $\sigma_0^2 = 1$, then

$$\begin{aligned}
 p_1 &= P(\text{GIC}_{+\infty} > \text{GIC}_0 \mid \hat{\beta}_{\text{ols}}, \mathbf{X}) \\
 &= P\left(\frac{\hat{\beta}_{\text{ols},2}^2}{\exp(w_n) - 1} > \hat{\sigma}_0^2 \mid \hat{\beta}_{\text{ols}}, \mathbf{X}\right) \\
 &= \Phi\left(\frac{\hat{a}_1 - n}{\sqrt{2n}}\right) + o_p(1),
 \end{aligned}$$

where $\hat{a}_1 = n\hat{\beta}_{\text{ols},2}^2 / \{\exp(w_n) - 1\}$ and $\hat{\beta}_{\text{ols},2}$ corresponds least-square estimate for $\mathbf{X}_2$. And

$$\hat{p}_1 = \Phi\left(\frac{\hat{a}_1 / \hat{\sigma}_0^2 - n}{\sqrt{2n}}\right) + o_p(1).$$

For the sequence of $\hat{\beta}_{\text{ols},2} = \sqrt{(\exp(w_n) - 1)(1 + cn^{-1/2})}$, where c is any constant, it follows that $\hat{a}_1 = n + c\sqrt{n}$. And then

$$\begin{aligned}
 \frac{\hat{a}_1 - n}{\sqrt{2n}} - \frac{\hat{a}_1 / \hat{\sigma}_0^2 - n}{\sqrt{2n}} &= \frac{\hat{a}_1(\hat{\sigma}_0^2 - 1)}{\sqrt{2n}\hat{\sigma}_0^2} \\
 &= \frac{(n + c\sqrt{n})(\hat{\sigma}_0^2 - 1)}{\sqrt{2n}\hat{\sigma}_0^2},
 \end{aligned}$$

which is not $o_p(1)$. Therefore, $\hat{p}_1$ is not a consistent estimator for p_1 . $\square$

Proposition C.1.

Assume the distribution of ϵ_i , $i = 1, \dots, n$ is symmetric about zero, then for any $\epsilon > 0$,

$$P\left\{\inf_{(t_1, t_2) \in \mathcal{E}_y} |p_k - h_k(t_1, t_2)| > \epsilon\right\} \leq \alpha + o(1), \quad (\text{C.2})$$

where $\mathcal{E}_y$ is an asymptotically $(1 - \alpha) \times 100\%$ confidence region for $\mu_{4,\epsilon} - \sigma_0^4$ and σ_0^2 .

Proof. Denote the event that $(\sigma_0^2, \mu_{4,\epsilon} - \sigma_0^4) \in \mathcal{E}_y$ as A ,

$$\begin{aligned}
 &P\left(\inf_{(t_1, t_2) \in \mathcal{E}_y} |p_k - h_k(t_1, t_2)| > \epsilon\right) \\
 &\leq P\left(\inf_{(t_1, t_2) \in \mathcal{E}_y} |p_k - h_k(t_1, t_2)| > \epsilon \mid A\right)P(A) + P(A^c) \\
 &\leq 0(1 - \alpha) + \alpha + o(1) \\
 &= \alpha + o(1).
 \end{aligned}$$

□

Lemma C.6.

For any $s \geq 1$, assume $n^{-1} \sum_{i=1}^n |X_i|^s = O(1)$, then

$$n^{-1} \sum_{i=1}^n |\hat{\epsilon}_i|^s \xrightarrow{\text{as}} m_s,$$

where $m_s = E|\epsilon_1|^s$.

Proof of Lemma C.6.

$$\begin{aligned} \left| \left(n^{-1} \sum_{i=1}^n |\hat{\epsilon}_i|^s \right)^{1/s} - \left(n^{-1} \sum_{i=1}^n |\epsilon_i|^s \right)^{1/s} \right|^s &\leq n^{-1} \sum_{i=1}^n |\hat{\epsilon}_i - \epsilon_i|^s \\ &= n^{-1} \sum_{i=1}^n \left| X_i^\top (\hat{\beta}_{\text{ols}} - \beta_0) \right|^s \\ &\leq n^{-1} \sum_{i=1}^n \|X_i\|^s \|\hat{\beta}_{\text{ols}} - \beta_0\|^s. \end{aligned}$$

But $\hat{\beta}_{\text{ols}} \xrightarrow{\text{as}} \beta_0$ and $n^{-1} \sum_{i=1}^n |X_i|^s = O(1)$, so that

$(n^{-1} \sum_{i=1}^n |\hat{\epsilon}_i|^s)^{1/s} - (n^{-1} \sum_{i=1}^n |\epsilon_i|^s)^{1/s} \xrightarrow{\text{as}} 0$. Then by Strong Law of Large Numbers,

$n^{-1} \sum_{i=1}^n |\epsilon_i|^s \xrightarrow{\text{as}} E|\epsilon_1|^s$. And thus $n^{-1} \sum_{i=1}^n |\hat{\epsilon}_i|^s \xrightarrow{\text{as}} E|\epsilon_1|^s$. □

Lemma C.7.

Assume (A1F) and (A2), then

$$\frac{1}{\sqrt{n}} \gamma^{(b)\top} \mathbf{P}_X \gamma^{(b)} \xrightarrow{P} 0,$$

conditionally almost surely.

Proof of Lemma C.7. Denote $\hat{\beta}_{\text{ols}}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X} \hat{\beta}_{\text{ols}} + \gamma^{(b)})$, then we have

$$\frac{1}{\sqrt{n}} \gamma^{(b)\top} \mathbf{P}_X \gamma^{(b)} = \sqrt{n} (\hat{\beta}_{\text{ols}}^* - \hat{\beta}_{\text{ols}})^\top \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \right) (\hat{\beta}_{\text{ols}}^* - \hat{\beta}_{\text{ols}}).$$

Then by noting $\sqrt{n} (\hat{\beta}_{\text{ols}}^* - \hat{\beta}_{\text{ols}}) \xrightarrow{d} N(0, C^{-1})$ conditionally almost surely (Theorem 2.2 of

Freedman 1981), $\frac{1}{\sqrt{n}} \gamma^{(b)\top} \mathbf{P}_X \gamma^{(b)}$ is $o_P(1)$ conditionally almost surely. □

Proposition C.2.

Under the assumptions (A1F) and (A2), and further assuming that $E|\epsilon_i|^{4+\delta} < \infty$ and $n^{-1} \sum_{i=1}^n |X_i|^{4+\delta} < \infty$ for any $\delta > 0$, then $\sqrt{n}(\hat{\sigma}_*^2 - \hat{\sigma}_0^2) \xrightarrow{d} N(0, \mu_{4,\epsilon} - \sigma_0^4)$ conditionally almost surely. $\square$

Proof of Proposition C.2. Recall

$$\begin{aligned}\hat{\sigma}_*^2 &= \frac{1}{n} \gamma^{(b)\top} (\mathbf{I} - \mathbf{P}_X) \gamma^{(b)} \\ &= \frac{1}{n} \gamma^{(b)\top} \gamma^{(b)} - \frac{1}{n} \gamma^{(b)\top} \mathbf{P}_X \gamma^{(b)}.\end{aligned}$$

Because $n^{-1/2} \gamma^{(b)\top} \mathbf{P}_X \gamma^{(b)} \xrightarrow{p} 0$ almost surely by Lemma C.7, $\hat{\sigma}_*^2$ has the same asymptotic distribution as $n^{-1} \gamma^{(b)\top} \gamma^{(b)}$. Then because the $\gamma_1^{(b)}, \gamma_2^{(b)}, \dots, \gamma_n^{(b)}$ are sampled from different distribution for every n , the Lindberg central limit theorem is used to obtain the asymptotic distribution. The conditional mean of $(\gamma_i^{(b)})^2$ is

$$E(\{\gamma_1^{(b)}\}^2 \mid \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n (\hat{e}_i)^2 = \hat{\sigma}_0^2.$$

The conditional variance is

$$\text{var}(\{\gamma_1^{(b)}\}^2 \mid \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n (\hat{e}_i)^4 - \left\{ \frac{1}{n} \sum_{i=1}^n (\hat{e}_i)^2 \right\}^2.$$

Use Lemma C.6, $n^{-1} \sum_{i=1}^n (\hat{e}_i)^4 \xrightarrow{\text{as}} \mu_{4,\epsilon}$ and $\{n^{-1} \sum_{i=1}^n (\hat{e}_i)^2\}^2 \xrightarrow{\text{as}} \sigma_0^4$. So the conditional variance converges to $\mu_{4,\epsilon} - \sigma_0^4$ almost surely.

Then to verify the Lyapunov condition,

$$\begin{aligned}& \frac{1}{\{\text{var}(\{\gamma_1^{(b)}\}^2 \mid \mathbf{Y})\}^{2+\delta}} \sum_{i=1}^n E \left\{ |\gamma_1^{(b)}|^2 / \sqrt{n} \right\}^{2+\delta} \mid \mathbf{Y} \Bigg\} \\ &= \frac{1}{\{\text{var}(\{\gamma_1^{(b)}\}^2 \mid \mathbf{Y})\}^{2+\delta}} \frac{1}{n^{\delta/2}} E \left\{ |\gamma_1^{(b)}|^{4+2\delta} \mid \mathbf{Y} \right\} \\ &= \frac{1}{\{\text{var}(\{\gamma_1^{(b)}\}^2 \mid \mathbf{Y})\}^{2+\delta}} \frac{1}{n^{1+\delta/2}} \sum_{i=1}^n |\hat{e}_i|^{4+2\delta},\end{aligned}$$

which is $\alpha(1)$ almost surely by invoking Lemma C.6. And thus $\sqrt{n}(\hat{\sigma}_*^2 - \hat{\sigma}_0^2) \xrightarrow{d} N(0, \mu_{4,\epsilon} - \sigma_0^4)$ conditionally almost surely. $\square$

C1. Theoretical Results for High Dimensions

Proposition C.3.

If $p = \alpha(n^{1/2})$, then

$$\sqrt{n}(\hat{\sigma}_0^2 - \sigma_0^2) \xrightarrow{d} N(0, \mu_{4,\epsilon} - \sigma_0^4).$$

Proof of Proposition C.3. We know $\hat{\sigma}_0^2 = n^{-1} \epsilon^T \epsilon - n^{-1} \epsilon^T P_x \epsilon$, where $\sqrt{n}(n^{-1} \epsilon^T \epsilon - \sigma_0^2)$ converges to $N(0, \mu_{4,\epsilon} - \sigma_0^4)$ in distribution. It remains to prove $n^{-1/2} \epsilon^T P_x \epsilon \xrightarrow{p} 0$.

By expectation of quadratic form, we have

$n^{-1/2} E(\epsilon^T P_x \epsilon) = n^{-1/2} \text{tr}(P_x \times I) \leq n^{-1/2} p = o(1)$. Therefore, $n^{-1/2} \epsilon^T P_x \epsilon \xrightarrow{p} 0$. This completes the proof. $\square$

Now we study the distribution of variance estimator after screening. First, we restate Theorem 1 from Fan and Lv (2008) with slight modification. Denote A_0 to be the true index of nonzero regression coefficients, and S to be the screened subset. Assume Conditions 1–4 in Fan and Lv (2008) hold for some $2\kappa + \tau < 1/2$, then we have the following result.

Theorem C.1 (Accuracy of SIS). Under Conditions 1–4 in Fan and Lv (2008), if $2\kappa + \tau < 1/2$, there exists $\theta > 1/2$, we have

$$P(A_0 \in S) = 1 - O(\exp(-Cn^{1-2\kappa} / \log n)),$$

where C is a positive constant, and the size of S is $O(n^{1-\theta})$.

From above result, we have a screening approach to reduce number of predictors from huge scale, $O(\exp(n^c))$, to a smaller scale, $o(\sqrt{n})$. Denote $\mathbf{X} = (\mathbf{X}^{(1)T}, \mathbf{X}^{(2)T})^T$, $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ are corresponding to the first and second half of design matrix, respectively. Similarly, define $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)}$. Then the variance estimator is defined as

$$\hat{\sigma}_0^2 = 1/m \left\{ \mathbf{Y}^{(2)} \right\}^T \left(I - P_{\mathbf{X}_S^{(2)}} \right) \mathbf{Y}^{(2)},$$

where $m = n/2$, and $P_{\mathbf{X}_S^{(2)}}$ is the projection matrix constructed from screened subset S and second half of design matrix, $\mathbf{X}^{(2)}$.

Proposition C.4.

Under Conditions 1–4 in Fan and Lv (2008), if $2\kappa + \tau < 1/2$, then

$$\sqrt{m}(\hat{\sigma}_0^2 - \sigma_0^2) \xrightarrow{d} N(0, \mu_{4,\epsilon} - \sigma_0^4).$$

Proof of Proposition C.4.

$$\begin{aligned}
 \sqrt{m}\hat{\sigma}_0^2 &= 1 / \sqrt{m} \{ \mathbf{Y}^{(2)} \}^T \left(I - P_{\mathbf{X}_S^{(2)}} \right) \mathbf{Y}^{(2)} \\
 &= 1 / \sqrt{m} \{ \mathbf{X}^{(2)} \beta_0 + \epsilon^{(2)} \}^T \left(I - P_{\mathbf{X}_S^{(2)}} \right) \{ \mathbf{X}^{(2)} \beta_0 + \epsilon^{(2)} \} \\
 &= 1 / \sqrt{m} \{ \epsilon^{(2)} \}^T \epsilon^{(2)} - 1 / \sqrt{m} \{ \epsilon^{(2)} \}^T P_{\mathbf{X}_S^{(2)}} \epsilon^{(2)} \\
 &\quad + 1 / \sqrt{m} \{ \mathbf{X}^{(2)} \beta_0 \}^T \left(I - P_{\mathbf{X}_S^{(2)}} \right) \{ \mathbf{X}^{(2)} \beta_0 \} \\
 &\quad + 2 / \sqrt{m} \{ \mathbf{X}^{(2)} \beta_0 \}^T \left(I - P_{\mathbf{X}_S^{(2)}} \right) \{ \epsilon^{(2)} \}.
 \end{aligned}$$

For the first term, we know $\sqrt{m}(1 / m \{ \epsilon^{(2)} \}^T \epsilon^{(2)} - \sigma_0^2)$ converges to $N(0, \mu_{4,\epsilon} - \sigma_0^4)$. It remains to prove the remaining term are $o_p(1)$. For the second term, we know

$$E(1 / \sqrt{m} \{ \epsilon^{(2)} \}^T P_{\mathbf{X}_S^{(2)}} \epsilon^{(2)}) = 1 / \sqrt{m} E(P_{\mathbf{X}_S^{(2)}}) = 1 / \sqrt{m} \times O(\sqrt{n}).$$

Therefore it is $o_p(1)$. For the third term,

$$\begin{aligned}
 &1 / \sqrt{m} E(\{ \mathbf{X}^{(2)} \beta_0 \}^T (I - P_{\mathbf{X}_S^{(2)}}) \{ \mathbf{X}^{(2)} \beta_0 \}) \\
 &= 1 / \sqrt{m} E(\{ \mathbf{X}^{(2)} \beta_0 \}^T (I - P_{\mathbf{X}_S^{(2)}}) \{ \mathbf{X}^{(2)} \beta_0 \} \mid A_0 \in S) P(A_0 \in S) \\
 &\quad + 1 / \sqrt{m} E(\{ \mathbf{X}^{(2)} \beta_0 \}^T (I - P_{\mathbf{X}_S^{(2)}}) \{ \mathbf{X}^{(2)} \beta_0 \} \mid A_0 \notin S) P(A_0 \notin S) \\
 &= 0 + 1 / \sqrt{m} E(\{ \mathbf{X}^{(2)} \beta_0 \}^T (I - P_{\mathbf{X}_S^{(2)}}) \{ \mathbf{X}^{(2)} \beta_0 \} \mid A_0 \notin S) P(A_0 \notin S) \\
 &\leq 1 / \sqrt{m} E(\{ \mathbf{X}^{(2)} \beta_0 \}^T \{ \mathbf{X}^{(2)} \beta_0 \}) P(A_0 \notin S) \\
 &= \sqrt{m} \beta_0^T C \beta_0 P(A_0 \notin S) \\
 &\leq \sqrt{m} \text{var}(Y) P(A_0 \notin S) \\
 &= O(\sqrt{n} \exp(-Cn^{1-2k} / \log n)),
 \end{aligned}$$

where the last inequality follows Condition 3 in Fan and Lv (2008), $\text{var}(Y) = O(1)$. And thus it is $o_p(1)$. For the last term,

$$\begin{aligned}
& \text{var}(2 / \sqrt{m} \{ \mathbf{X}^{(2)} \beta_0 \}^T (I - P_{\mathbf{X}_S^{(2)}}) \{ \epsilon^{(2)} \}) \\
& = 4 / m E(\{ \mathbf{X}^{(2)} \beta_0 \}^T (I - P_{\mathbf{X}_S^{(2)}}) \{ \mathbf{X}^{(2)} \beta_0 \})
\end{aligned}$$

which is $o(1)$. This completes the proof. $\square$

References

- Akaike H (1974), “A New Look at the Statistical Model Identification,” *IEEE Transactions on Automatic Control*, 19, 716–723. [1]
- Barber RF, and Candès EJ, (2015), “Controlling the False Discovery Rate Via Knockoffs,” *The Annals of Statistics*, 43, 2055–2085. [6]
- Berger RL, and Boos DD (1994), “P Values Maximized Over a Confidence Set for the Nuisance Parameter,” *Journal of the American Statistical Association*, 89, 1012–1016. [3]
- Bertsekas DP (2014), *Constrained Optimization and Lagrange Multiplier Methods*, Boston, MA: Academic Press [4]
- Burnham KP, and Anderson D (2003), *Model Selection and Multi-Model Inference: A Practical Information-theoretic Approach*, New York: Springer [4]
- Chen J, and Chen Z (2008), “Extended Bayesian Information Criteria for Model Selection with Large Model Spaces,” *Biometrika*, 95, 759–771. [1]
- Cox DR (2001), “Statistical Modeling: The Two Cultures: Comment,” *Statistical Science*, 16, 216–218. [1]
- Fan J, and Li R (2001), “Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties,” *Journal of the American Statistical Association*, 96, 1348–1360. [1]
- Fan J, and Lv J (2008), “Sure Independence Screening for Ultrahigh Dimensional Feature Space,” *Journal of the Royal Statistical Society, Series B*, 70, 849–911. [10,11]
- Fan Y, and Tang CY (2013), “Tuning Parameter Selection in High Dimensional Penalized Likelihood,” *Journal of the Royal Statistical Society, Series B*, 75, 531–552. [1]
- Feng Y, and Yu Y (2013), “Consistent Cross-validation for Tuning Parameter Selection in High-dimensional Variable Selection.” *arXiv preprint arXiv:1308.5390*. [1]
- Freedman DA (1981), “Bootstrapping Regression Models,” *The Annals of Statistics*, 9, 1218–1228. [10]
- Golub GH, Heath M, and Wahba G (1979), “Generalized Crossvalidation as a Method for Choosing a Good Ridge Parameter,” *Technometrics*, 21, 215–223. [1]
- Hall P, Lee ER, and Park BU (2009), “Bootstrap-based Penalty Choice for the Lasso, Achieving Oracle Performance,” *Statistica Sinica*, 19, 449–471. [1]
- Henderson HV, and Velleman PF (1981), “Building Multiple Regression Models Interactively,” *Biometrics*, 37, 391–411. [1]
- Hui FK, Warton DI, and Foster SD (2015), “Tuning Parameter Selection for the Adaptive Lasso Using Eric,” *Journal of the American Statistical Association*, 110, 262–269. [1]
- Katsouyanni K, Pantazopoulou A, Touloumi G, Tselepidaki I, Moustis K, Asimakopoulos D, Pouloupoulou G, and Trichopoulos D (1993), “Evidence for Interaction Between Air Pollution and High Temperature in the Causation of Excess Mortality,” *Archives of Environmental Health: An International Journal*, 48, 235–242. [6]
- Kim Y, Kwon S, and Choi H (2012), “Consistent Model Selection Criteria on High Dimensions,” *Journal of Machine Learning Research*, 13, 1037–1057. [1]
- Luo X, Stefanski LA, and Boos DD (2006), “Tuning Variable Selection Procedures by Adding Noise,” *Technometrics*, 48, 165–175. [5]
- Mallows CL (1973), “Some Comments on Cp,” *Technometrics*, 15, 661–675. [1]

- McDonald GC, and Schwing RC (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics*, 15, 463–481. [5]
- Meinshausen N, and Bühlmann P (2010), "Stability Selection," *Journal of the Royal Statistical Society, Series B*, 72, 417–473. [1]
- Rhee S-Y, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, and Shafer RW (2006), "Genotypic Predictors of Human Immunodeficiency Virus Type 1 Drug Resistance," *Proceedings of the National Academy of Sciences*, 103, 17355–17360. [5,6]
- Schwarz G (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [1]
- Shah RD, and Samworth RJ (2013), "Variable Selection with Error Control: Another Look at Stability Selection," *Journal of the Royal Statistical Society, Series B*, 75, 55–80. [1]
- Sun W, Wang J, and Fang Y (2013), "Consistent Selection of Tuning Parameters Via Variable Selection Stability," *The Journal of Machine Learning Research*, 14, 3419–3440. [1]
- Tibshirani R (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [1]
- Wang H, Li B, and Leng C (2009), "Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters," *Journal of the Royal Statistical Society, Series B*, 71, 671–683. [1]
- Wang T, and Zhu L (2011), "Consistent Tuning Parameter Selection in High Dimensional Sparse Linear Regression," *Journal of Multivariate Analysis*, 102, 1141–1151. [1]
- Zhang Y, Li R, and Tsai C-L (2010), "Regularization Parameter Selections Via Generalized Information Criterion," *Journal of the American Statistical Association*, 105, 312–323. [1] [PubMed: 20676354]
- Zou H (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1]
- Zou H, and Hastie T (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320. [1]

Figure 1.

The top figure shows LASSO solution path of mortality rates data; The vertical lines above and below x-axis correspond to the distribution estimated by bootstrap and asymptotic normal approximation, respectively. The bottom figure shows BIC values for candidate models along solution path. The solid vertical line corresponds to the model with six variables and smallest BIC value. The dashed vertical line corresponds to a model with eight variables.

Figure 2.

The top figure shows forward stepwise regression solution path of mortality rates data; The vertical line corresponds to the model which minimizes BIC. The bottom figures show BIC values for candidate models along solution path. The solid vertical line corresponds to the smallest BIC value.

Figure 3.

The top figure shows LASSO solution path of ATV drug resistance data; The vertical lines above and below x-axis correspond to the distribution estimated by bootstrap and asymptotic normal approximation, respectively. The bottom figure shows BIC values for candidate models along solution path. The solid vertical line corresponds to the model with 15 variables and smallest BIC value. The dashed vertical lines correspond to model with 12 and 10 variables.

Table 1.Discovery rate for $p = 20$, $n = 50$.

Model	ρ	AsympNor		Bootstrap		UP1		UP2		Akaike		ApproxPost	
		TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR
1	0	0.90	0.10	0.90	0.10	1.00	0.32	0.99	0.36	0.54	0.83	0.90	0.58
	0.5	0.92	0.09	0.92	0.09	1.00	0.29	0.99	0.34	0.61	0.83	0.95	0.58
2	0	0.88	0.11	0.88	0.10	0.99	0.30	0.99	0.34	0.46	0.83	0.82	0.59
	0.5	0.89	0.10	0.90	0.10	1.00	0.35	0.99	0.36	0.55	0.82	0.89	0.59
3	0	0.92	0.09	0.92	0.09	1.00	0.29	0.99	0.34	0.57	0.84	0.94	0.58
	0.5	0.93	0.08	0.93	0.08	1.00	0.27	0.99	0.32	0.64	0.83	0.96	0.57
4	0	0.89	0.11	0.90	0.10	0.99	0.32	0.99	0.36	0.50	0.84	0.88	0.59
	0.5	0.91	0.10	0.91	0.09	0.99	0.33	0.99	0.35	0.58	0.83	0.92	0.58

Table 2.Discovery rate for $p = 100$, $n = 200$.

Model	ρ	AsympNor		Bootstrap		UP1		UP2		Akaike		ApproxPost	
		TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR
1	0	0.97	0.03	0.97	0.03	1.00	0.12	1.00	0.13	0.11	0.98	1.00	0.58
	0.5	0.98	0.03	0.98	0.02	1.00	0.09	1.00	0.10	0.24	0.95	1.00	0.56
2	0	0.95	0.04	0.95	0.04	1.00	0.17	1.00	0.18	0.05	0.99	0.98	0.59
	0.5	0.96	0.03	0.96	0.03	1.00	0.14	1.00	0.15	0.12	0.97	0.99	0.58
3	0	0.97	0.03	0.97	0.03	1.00	0.12	1.00	0.13	0.16	0.97	1.00	0.57
	0.5	0.98	0.02	0.98	0.02	1.00	0.09	1.00	0.10	0.28	0.94	1.00	0.56
4	0	0.96	0.04	0.96	0.04	1.00	0.14	1.00	0.16	0.07	0.98	0.99	0.59
	0.5	0.97	0.03	0.97	0.03	1.00	0.12	1.00	0.13	0.16	0.97	1.00	0.57

Table 3.

Coverage probability. Results are based on 10,000 replicated datasets.

ρ		<u>Approximate</u>		<u>Normality</u>	
		0	0.5	0	0.5
$n = 50$	Model 1	0.86	0.85	0.92	0.91
	Model 2	0.86	0.85	0.91	0.91
	Model 3	0.85	0.85	0.91	0.91
	Model 4	0.86	0.85	0.92	0.91
$n = 200$	Model 1	0.88	0.88	0.91	0.90
	Model 2	0.87	0.88	0.91	0.91
	Model 3	0.87	0.87	0.91	0.90
	Model 4	0.87	0.88	0.91	0.90

Table 4.

Estimated conditional distribution of the tuning parameter for the mortality rates data. Both the asymptotic normal approximation and the bootstrap had only two support points, {124.21, 288.20}.

λ	288.20	124.21
Model size	6	8
Probability mass normal	0.51	0.49
Probability mass bootstrap	0.48	0.52
90% CI based on normality	(0.052, 0.950)	(0.048, 0.948)
90% Approximate CI	(0.009, 0.848)	(0.154, 1.000)

Table 5.

Estimated conditional distribution of the tuning parameter for ATV drug resistance data.

λ	1347.92	1101.71	805.53
Model size	10	12	15
Probability mass normal	0.443	0.045	0.472
Probability mass bootstrap	0.494	0.047	0.454
90% CI based on normality	(0.019, 0.896)	(0.026, 0.127)	(0.059, 0.959)
90% Approximate CI	(0.000, 0.751)	(0.020, 0.064)	(0.020, 1)