Consistent Selection of Tuning Parameters via Variable Selection Stability

Wei Sun   SUN244@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Junhui Wang   JUNHUI@UIC.EDU
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607, USA

Yixin Fang   YIXIN.FANG@NYUMC.ORG
Departments of Population Health and Environmental Medicine
New York University
New York, NY 10016, USA

Editor: Xiaotong Shen

Abstract

Penalized regression models are popularly used in high-dimensional data analysis to conduct variable selection and model fitting simultaneously. Whereas success has been widely reported in literature, their performances largely depend on the tuning parameters that balance the trade-off between model fitting and model sparsity. Existing tuning criteria mainly follow the route of minimizing the estimated prediction error or maximizing the posterior model probability, such as cross validation, AIC and BIC. This article introduces a general tuning parameter selection criterion based on variable selection stability. The key idea is to select the tuning parameters so that the resultant penalized regression model is stable in variable selection. The asymptotic selection consistency is established for both fixed and diverging dimensions. Its effectiveness is also demonstrated in a variety of simulated examples as well as an application to the prostate cancer data.

Keywords: kappa coefficient, penalized regression, selection consistency, stability, tuning

1. Introduction

The rapid advance of technology has led to an increasing demand for modern statistical techniques to analyze data with complex structure such as the high-dimensional data. In high-dimensional data analysis, it is generally believed that only a small number of variables are truly informative while others are redundant. An underfitted model excludes truly informative variables and may lead to severe estimation bias in model fitting, whereas an overfitted model includes the redundant uninformative variables, increases the estimation variance and hinders the model interpretation. Therefore, identifying the truly informative variables is regarded as the primary goal of the high-dimensional data analysis as well as its many real applications such as health studies (Fan and Li, 2006).

Among other variable selection methods, penalized regression models have been popularly used, which penalize the model fitting with various regularization terms to encourage model sparsity, such as the lasso regression (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD, Fan and Li, 2001), the adaptive lasso (Zou, 2006), and the truncated l_1-norm regression (Shen et al., 2012). In the penalized regression models, tuning parameters are often employed to balance the trade-off between model fitting and model sparsity, which largely affects the numerical performance and the asymptotic behavior of the penalized regression models. For example, Zhao and Yu (2006) showed that, under the irrepresentable condition, the lasso regression is selection consistent when the tuning parameter converges to 0 at a rate slower than $O(n^{-1/2})$. Analogous results on the choice of tuning parameters have also been established for the SCAD, the adaptive lasso, and the truncated l_1-norm regression. Therefore, it is of crucial importance to select the appropriate tuning parameters so that the performance of the penalized regression models can be optimized.

In literature, many classical selection criteria have been applied to the penalized regression models, including cross validation (Stone, 1974), generalized cross validation (Craven and Wahba, 1979), Mallows' C_p (Mallows, 1973), AIC (Akaike, 1974) and BIC (Schwarz, 1978). Under certain regularity conditions, Wang et al. (2007) and Wang et al. (2009) established the selection consistency of BIC for the SCAD, and Zhang et al. (2010) showed the selection consistency of generalized information criterion (GIC) for the SCAD. Most of these criteria follow the route of minimizing the estimated prediction error or maximizing the posterior model probability. To the best of our knowledge, few criteria has been developed directly focusing on the selection of the informative variables.

This article proposes a tuning parameter selection criterion based on variable selection stability. The key idea is that if multiple samples are available from the same distribution, a good variable selection method should yield similar sets of informative variables that do not vary much from one sample to another. The similarity between two informative variable sets is measured by Cohen's kappa coefficient (Cohen, 1960), which adjusts the actual variable selection agreement relative to the possible agreement by chance. Similar stability measures have been studied in the context of cluster analysis (Ben-Hur et al., 2002; Wang, 2010) and variable selection (Meinshausen and Bühlmann, 2010). Whereas the stability selection method (Meinshausen and Bühlmann, 2010) also follows the idea of variable selection stability, it mainly focuses on selecting the informative variables as opposed to selecting the tuning parameters for any given variable selection methods. The effectiveness of the proposed selection criterion is demonstrated in a variety of simulated examples and a real application. More importantly, its asymptotic selection consistency is established, showing that the variable selection method with the selected tuning parameter would recover the truly informative variable set with probability tending to one.

The rest of the article is organized as follows. Section 2 briefly reviews the penalized regression models. Section 3 presents the idea of variable selection stability as well as the proposed kappa selection criterion. Section 4 establishes the asymptotic selection consistency of the kappa selection criterion. Simulation studies are given in Section 5, followed by a real application in Section 6. A brief discussion is provided in Section 7, and the Appendix is devoted to the technical proofs.

2. Penalized Least Squares Regression

Given that $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are independent and identically distributed from some unknown joint distribution, we consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \sum_{j=1}^{p} \beta_j \mathbf{x}_{(j)} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T = (\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(p)})$ with $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ or $\mathbf{x}_{(j)} = (x_{1j}, \cdots, x_{nj})^T$, and $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. When p is large, it is also assumed that only a small number of β_j's are nonzero, corresponding to the truly informative variables. In addition, both $\mathbf{y}$ and $\mathbf{x}_{(j)}$'s are centered, so the intercept can be omitted in the regression model.

The general framework of the penalized regression models can be formulated as

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm, and $p_\lambda(|\beta_j|)$ is a regularization term encouraging sparsity in $\boldsymbol{\beta}$. Widely used regularization terms include the lasso penalty $p_\lambda(\theta) = \lambda\theta$ (Tibshirani, 1996), the SCAD penalty with $p'_\lambda(\theta) = \lambda(I(\theta \le \lambda) + \frac{(\gamma\lambda - \theta)_+}{(\gamma - 1)\lambda} I(\theta > \lambda))$ (Fan and Li, 2001), the adaptive lasso penalty $p_\lambda(\theta) = \lambda_j\theta = \lambda\theta/|\hat{\beta}_j|$ (Zou, 2006) with $\hat{\beta}_j$ being some initial estimate of β_j, and the truncated l_1-norm penalty $p_\lambda(\theta) = \lambda \min(1, \theta)$ (Shen et al., 2012).

With appropriately chosen λ_n, all the aforementioned regularization terms have been shown to be selection consistent. Here a penalty term is said to be selection consistent if the probability that the fitted regression model includes only the truly informative variables is tending to one, and λ is replaced by λ_n to emphasize its dependence on n in quantifying the asymptotic behaviors. In particular, Zhao and Yu (2006) showed that the lasso regression is selection consistent under the irrepresentable condition when $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$; Fan and Li (2001) showed that the SCAD is selection consistent when $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$; Zou (2006) showed that the adaptive lasso is selection consistent when $n\lambda_n \to \infty$ and $\sqrt{n}\lambda_n \to 0$; and Shen et al. (2012) showed that the truncated l_1-norm penalty is also selection consistent when λ_n satisfies a relatively more complex constraint.

Although the asymptotic order of λ_n is known to assure the selection consistency of the penalized regression models, it remains unclear how to appropriately select λ_n in finite sample so that the resultant model in (1) with the selected λ_n can achieve superior numerical performance and attain asymptotic selection consistency. Therefore, it is in demand to devise a tuning parameter selection criterion that can be employed by the penalized regression models so that their variable selection performance can be optimized.

3. Tuning via Variable Selection Stability

This section introduces the proposed tuning parameter selection criterion based on the concept of variable selection stability. The key idea is that if we repeatedly draw samples from the population and apply the candidate variable selection methods, a desirable method should produce the informative variable set that does not vary much from one sample to another. Clearly, variable selection stability is assumption free and can be used to tune any penalized regression model.

3.1 Variable Selection Stability

For simplicity, we denote the training sample as z^n. A base variable selection method $\Psi(z^n;\lambda)$ with a given training sample z^n and a tuning parameter λ yields a set of selected informative variables $\mathcal{A} \subset \{1,\cdots,p\}$, called the active set. When Ψ is applied to various training samples, different active sets can be produced. Supposed that two active sets $\mathcal{A}_1$ and $\mathcal{A}_2$ are produced, the agreement between $\mathcal{A}_1$ and $\mathcal{A}_2$ can be measured by Cohen's kappa coefficient (Cohen, 1960),

$$\kappa(\mathcal{A}_1,\mathcal{A}_2) = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}. \tag{2}$$

Here the relative observed agreement between $\mathcal{A}_1$ and $\mathcal{A}_2$ is $Pr(a) = (n_{11}+n_{22})/p$, and the hypothetical probability of chance agreement $Pr(e) = (n_{11}+n_{12})(n_{11}+n_{21})/p^2 + (n_{12}+n_{22})(n_{21}+n_{22})/p^2$, with $n_{11} = |\mathcal{A}_1 \cap \mathcal{A}_2|$, $n_{12} = |\mathcal{A}_1 \cap \mathcal{A}_2^c|$, $n_{21} = |\mathcal{A}_1^c \cap \mathcal{A}_2|$, $n_{22} = |\mathcal{A}_1^c \cap \mathcal{A}_2^c|$, and $|\cdot|$ being the set cardinality. Note that $-1 \le \kappa(\mathcal{A}_1,\mathcal{A}_2) \le 1$, where $\kappa(\mathcal{A}_1,\mathcal{A}_2) = 1$ when $\mathcal{A}_1$ and $\mathcal{A}_2$ are in complete agreement with $n_{12} = n_{21} = 0$, and $\kappa(\mathcal{A}_1,\mathcal{A}_2) = -1$ when $\mathcal{A}_1$ and $\mathcal{A}_2$ are in complete disagreement with $n_{11} = n_{22} = 0$ and $n_{12} = n_{21} = p/2$. For degenerate cases with $\mathcal{A}_1 = \mathcal{A}_2 = \emptyset$ or $\mathcal{A}_1 = \mathcal{A}_2 = \{1,\ldots,p\}$, we set $\kappa(\emptyset,\emptyset) = \kappa(\{1,\ldots,p\},\{1,\ldots,p\}) = -1$ under the assumption that the true model is sparse and containing at least one informative variable. As a consequence, the kappa coefficient in (2) is not suitable for evaluating the null model with no informative variable and the complete model with all variables. Based on (2), the variable selection stability is defined as follows.

Definition 1 *The variable selection stability of $\Psi(\cdot;\lambda)$ is defined as*

$$s(\Psi,\lambda,n) = E\Big(\kappa(\Psi(Z_1^n;\lambda),\Psi(Z_2^n;\lambda))\Big),$$

where the expectation is taken with respect to Z_1^n and Z_2^n, two independent and identically training samples of size n, and $\Psi(Z_1^n;\lambda)$ and $\Psi(Z_2^n;\lambda)$ are two active sets obtained by applying $\Psi(\cdot;\lambda)$ to Z_1^n and Z_2^n, respectively.

By definition, $-1 \le s(\Psi,\lambda,n) \le 1$, and large value of $s(\Psi,\lambda,n)$ indicates a stable variable selection method $\Psi(\cdot;\lambda)$. Note that the definition of $s(\Psi,\lambda,n)$ relies on the unknown population distribution, therefore it needs to be estimated based on the only available training sample in practice.

3.2 Kappa Selection Criterion

This section proposes an estimation scheme of the variable selection stability based on cross validation, and develops a kappa selection criterion to tune the penalized regression models by maximizing the estimated variable selection stability. Specifically, the training sample z^n is randomly partitioned into two subsets z_1^m and z_2^m with $m = \lfloor n/2 \rfloor$ for simplicity. The base variable selection method $\Psi(\cdot;\lambda)$ is applied to two subsets separately, and then two active sets $\widehat{\mathcal{A}}_{1\lambda}$ and $\widehat{\mathcal{A}}_{2\lambda}$ are obtained, and $s(\Psi,\lambda,m)$ is estimated as $\kappa(\widehat{\mathcal{A}}_{1\lambda},\widehat{\mathcal{A}}_{2\lambda})$. Furthermore, in order to reduce the estimation variability due to the splitting randomness, multiple data splitting can be conducted and the average estimated variable selection stability over all splittings is computed. The selected λ is then the one obtaining upper α_n quartile of the average estimated variable selection stability. The proposed kappa selection criterion is present as follows.

Algorithm 1 (kappa selection criterion) :

Step 1. Randomly partition $(\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ into two subsets $z_1^{*b} = (\mathbf{x}_1^{*b}, \cdots, \mathbf{x}_m^{*b})^T$ and $z_2^{*b} = (\mathbf{x}_{m+1}^{*b}, \cdots, \mathbf{x}_{2m}^{*b})^T$.

Step 2. Obtain $\widehat{\mathcal{A}}_{1\lambda}^{*b}$ and $\widehat{\mathcal{A}}_{2\lambda}^{*b}$ from $\Psi(z_1^{*b}, \lambda)$ and $\Psi(z_2^{*b}, \lambda)$ respectively, and estimate the variable selection stability of $\Psi(\cdot; \lambda)$ in the b-th splitting by

$$\hat{s}^{*b}(\Psi, \lambda, m) = \kappa(\widehat{\mathcal{A}}_{1\lambda}^{*b}, \widehat{\mathcal{A}}_{2\lambda}^{*b}).$$

Step 3. Repeat *Steps 1-2* for B times. The average estimated variable selection stability of $\Psi(\cdot; \lambda)$ is then

$$\hat{s}(\Psi, \lambda, m) = B^{-1} \sum_{b=1}^{B} \hat{s}^{*b}(\Psi, \lambda, m).$$

Step 4. Compute $\hat{s}(\Psi, \lambda, m)$ for a sequence of λ's, and select

$$\hat{\lambda} = \min \left\{ \lambda : \frac{\hat{s}(\Psi, \lambda, m)}{\max_{\lambda'} \hat{s}(\Psi, \lambda', m)} \geq 1 - \alpha_n \right\}.$$

Note that the treatment in Step 4 is necessary since some informative variables may have relatively weak effect compared with others. A large value of λ may produce an active set that consistently overlooks the weakly informative variables, which leads to an underfitted model with large variable selection stability. To assure the asymptotic selection consistency, the thresholding value α_n in Step 4 needs to be small and converges to 0 as n grows. Setting $\alpha_n = 0.1$ in the numerical experiments yields satisfactory performance based on our limited experience. Furthermore, the sensitivity study in Section 5.1 suggests that α_n has very little effect on the selection performance when it varies in a certain range. In Steps 1-3, the estimation scheme based on cross-validation can be replaced by other data re-sampling strategies such as bootstrap or random weighting, which do not reduce the sample size in estimating $\widehat{\mathcal{A}}_{1\lambda}^{*b}$ and $\widehat{\mathcal{A}}_{2\lambda}^{*b}$, but the independence between $\widehat{\mathcal{A}}_{1\lambda}^{*b}$ and $\widehat{\mathcal{A}}_{2\lambda}^{*b}$ will no longer hold.

The proposed kappa selection criterion shares the similar idea of variable selection stability with the stability selection method (Meinshausen and Bühlmann, 2010), but they differ in a number of ways. First, the stability selection method is a competitive variable selection method, which combines the randomized lasso regression and the bootstrap, and achieves superior variable selection performance. However, the kappa selection criterion can be regarded as a model selection criterion that is designed to select appropriate tuning parameters for any variable selection method. Second, despite of its robustness, the stability selection method still requires a number of tuning parameters. The authors proposed to select the tuning parameters via controlling the expected number of falsely selected variables. However, this criterion is less applicable in practice since the expected number of falsely selected variables can only be upper bounded by an expression involving various unknown quantities. On the contrary, the kappa selection criterion can be directly applied to select the tuning parameters for the stability selection method.

4. Asymptotic Selection Consistency

This section presents the asymptotic selection consistency of the proposed kappa selection criterion. Without loss of generality, we assume that only the first p_0 variables with $0 < p_0 < p$ are informative, and denote the truly informative variable set as $\mathcal{A}_T = \{1, \cdots, p_0\}$ and the uninformative variable set

as $\mathcal{A}_T^c = \{p_0+1, \cdots, p\}$. Furthermore, we denote $r_n \prec s_n$ if r_n converges to 0 at a faster rate than s_n, $r_n \sim s_n$ if r_n converges to 0 at the same rate as s_n, and $r_n \preceq s_n$ if r_n converges to 0 at a rate not slower than s_n.

4.1 Consistency with Fixed p

To establish the asymptotic selection consistency with fixed p, the following technical assumptions are made.

Assumption 1: There exist positive r_n and s_n such that the base variable selection method is selection consistent if $r_n \prec \lambda_n \prec s_n$. Let λ_n^* be such a tuning parameter with $r_n \prec \lambda_n^* \prec s_n$, then $P(\widehat{\mathcal{A}}_{\lambda_n^*} = \mathcal{A}_T) \geq 1 - \varepsilon_n$ for some $\varepsilon_n \to 0$. In addition, for any positive constant λ_0, there exists positive $c_0(\lambda_0)$ such that, when n is sufficiently large,

$$P\left(\bigcap_{\lambda_0 r_n \leq \lambda_n \leq \lambda_n^*} \{\widehat{\mathcal{A}}_{\lambda_n} = \mathcal{A}_T\} \right) \geq 1 - c_0(\lambda_0), \tag{3}$$

where $c_0(\lambda_0)$ converges to 0 as $\lambda_0 \to \infty$.

Assumption 1 specifies an asymptotic working interval for λ_n within which the base variable selection method is selection consistent. Here the consistent rate ε_n is defined for λ_n^* only, and needs not hold uniformly over all λ_n with $r_n \prec \lambda_n \prec s_n$. Furthermore, (3) establishes an uniform lower bound for the probability of selecting the true model when λ_n is within the interval $(\lambda_0 r_n, \lambda_n^*)$.

Assumption 2: Given r_n in Assumption 1, for any positive constant λ_0, there exist ζ_n, $c_1(\lambda_0)$ and $c_2(\lambda_0)$ such that, when n is sufficiently large,

$$\min_{j \in \mathcal{A}_T} P\left(\bigcap_{r_n^{-1}\lambda_n \leq \lambda_0} \{j \in \widehat{\mathcal{A}}_{\lambda_n}\} \right) \geq 1 - \zeta_n, \tag{4}$$

$$\min_{j \in \mathcal{A}_T^c} P\left(\bigcap_{r_n^{-1}\lambda_n \leq \lambda_0} \{j \in \widehat{\mathcal{A}}_{\lambda_n}\} \right) \geq c_1(\lambda_0), \tag{5}$$

$$\max_{j \in \mathcal{A}_T^c} P\left(\bigcap_{r_n^{-1}\lambda_n \geq \lambda_0} \{j \notin \widehat{\mathcal{A}}_{\lambda_n}\} \right) \geq c_2(\lambda_0), \tag{6}$$

where $\zeta_n \to 0$ as $n \to \infty$, $c_1(\lambda_0)$ and $c_2(\lambda_0)$ are positive and only depend on λ_0, and $c_1(\lambda_0) \to 1$ as $\lambda_0 \to 0$.

Assumption 2 implies that if λ_n converges to 0 faster than r_n, the base variable selection method will select all the variables asymptotically, and when λ_n converges to 0 at the same rate of r_n, the base variable selection method will select any noise variable with an asymptotically positive probability. The inequalities (4)-(6) also establish uniform lower bounds for various probabilities of selecting informative variables or noise variables.

Assumptions 1 and 2 are mild in that they are satisfied by many popular variable selection methods. For instance, Lemma 2 in the online supplementary material shows that Assumptions 1 and 2 are satisfied by the lasso regression, the adaptive lasso, and the SCAD. The assumptions can also be verified for other methods such as the elastic-net (Zou and Hastie, 2005), the adaptive elastic net (Zou and Zhang, 2009), the group lasso (Yuan and Lin, 2006), and the adaptive group lasso (Wang and Leng, 2008).

Given that the base variable selection method is selection consistent with appropriately selected λ_n's, Theorem 1 shows that the proposed kappa selection criterion is able to identify such λ_n's.

Theorem 1 *Under Assumptions 1 and 2, any variable selection method in (1) with $\hat{\lambda}_n$ selected as in Algorithm 1 with $\alpha_n \succ \varepsilon_n$ is selection consistent. That is,*

$$\lim_{n\to\infty}\lim_{B\to\infty} P(\widehat{\mathcal{A}}_{\hat{\lambda}_n} = \mathcal{A}_T) = 1.$$

Theorem 1 claims the asymptotic selection consistency of the proposed kappa selection criterion when p is fixed. That is, with probability tending to one, the selected active set by the resultant variable selection method with tuning parameter $\hat{\lambda}_n$ contains only the truly informative variables. As long as α_n converges to 0 not too fast, the kappa selection criterion is guaranteed to be consistent. Therefore, the value of α_n is expected to have little effect on the performance of the kappa selection criterion, which agrees with the sensitivity study in Section 5.1.

4.2 Consistency with Diverging p_n

In high-dimensional data analysis, it is of interest to study the asymptotic behavior of the proposed kappa selection criterion with diverging p_n, where size of truly informative set p_{0n} may also diverge with n. To accommodate the diverging p_n scenario, the technical assumptions are modified as follows.

Assumption 1a: There exist positive r_n and s_n such that the base variable selection method is selection consistent if $r_n \prec \lambda_n \prec s_n$. Let λ_n^* be such a tuning parameter with $r_n \prec \lambda_n^* \prec s_n$, then $P(\widehat{\mathcal{A}}_{\lambda_n^*} = \mathcal{A}_T) \geq 1 - \varepsilon_n$ for some $\varepsilon_n \to 0$. In addition, for any positive constant λ_0, there exists positive $c_{0n}(\lambda_0)$ such that, when n is sufficiently large,

$$P\left(\bigcap_{\lambda_0 r_n \leq \lambda_n \leq \lambda_n^*} \{\widehat{\mathcal{A}}_{\lambda_n} = \mathcal{A}_T\} \right) \geq 1 - c_{0n}(\lambda_0), \tag{7}$$

where $\lim_{\lambda_0\to\infty}\lim_{n\to\infty} c_{0n}(\lambda_0) \to 0$.

Assumption 2a: Given r_n in Assumption 1a, for any positive constant λ_0, there exist ζ_n, $c_{1n}(\lambda_0)$ and $c_{2n}(\lambda_0)$ such that, when n is sufficiently large,

$$\min_{j\in\mathcal{A}_T} P\left(\bigcap_{r_n^{-1}\lambda_n \leq \lambda_0} \{j \in \widehat{\mathcal{A}}_{\lambda_n}\} \right) \geq 1 - \zeta_n, \tag{8}$$

$$\min_{j\in\mathcal{A}_T^c} P\left(\bigcap_{r_n^{-1}\lambda_n \leq \lambda_0} \{j \in \widehat{\mathcal{A}}_{\lambda_n}\} \right) \geq c_{1n}(\lambda_0), \tag{9}$$

$$\max_{j\in\mathcal{A}_T^c} P\left(\bigcap_{r_n^{-1}\lambda_n \geq \lambda_0} \{j \notin \widehat{\mathcal{A}}_{\lambda_n}\} \right) \geq c_{2n}(\lambda_0), \tag{10}$$

where ζ_n satisfies $p_n\zeta_n \to 0$ as $n \to \infty$, $c_{1n}(\lambda_0)$ and $c_{2n}(\lambda_0)$ are positive and may depend on n and λ_0, and $\lim_{\lambda_0\to 0}\lim_{n\to\infty} c_{1n}(\lambda_0) = 1$.

Theorem 2 *Under Assumptions 1a and 2a, any variable selection method in (1) with $\hat{\lambda}_n$ selected as in Algorithm 1 with $\min\left(p_n(1 - \tilde{c}_{1n}), p_n^{-1}c_{1n}c_{2n}\right) \succ \alpha_n \succ \varepsilon_n$ is selection consistent, where $\tilde{c}_{1n} = \sup_{\lambda_0} c_{1n}(\lambda_0)$, $c_{1n} = \inf_{\lambda_0} c_{1n}(\lambda_0)$, and $c_{2n} = \inf_{\lambda_0} c_{2n}(\lambda_0)$.*

Theorem 2 shows the asymptotic selection consistency of the proposed kappa selection criterion with satisfied α_n for diverging p_n, where the diverging speed of p_n is bounded as in Theorem 2 and

depends on the base variable selection method. Lemma 3 in the online supplementary material shows that (7)-(10) in Assumptions 1a and 2a are satisfied by the lasso regression. However, it is generally difficult to verify Assumptions 1a and 2a for other popular variable selection algorithms (Fan and Peng, 2004; Huang and Xie, 2007; Huang et al., 2008), as the convergence rates in both assumptions are not explicitly specified.

5. Simulations

This section examines the effectiveness of the proposed kappa selection criterion in simulated examples. Its performance is compared against a number of popular competitors, including Mallows' C_p (C_p), BIC, 10-fold cross validation (CV), and generalized cross validation (GCV). Their formulations are given as follows,

$$C_p(\lambda) = \frac{SSE_\lambda}{\hat{\sigma}^2} - n + 2\widehat{df}, \tag{11}$$

$$BIC(\lambda) = \log\left(\frac{SSE_\lambda}{n}\right) + \frac{\log(n)\widehat{df}}{n},$$

$$CV(\lambda) = \sum_{s=1}^{10} \sum_{(y_k,x_k)\in T^{-s}} \left(y_k - \mathbf{x}_k^T \hat{\beta}^{(s)}(\lambda)\right)^2, \tag{12}$$

$$GCV(\lambda) = \frac{SSE_\lambda}{n(1 - \widehat{df}/n)^2},$$

where $SSE_\lambda = \|\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda)\|^2$, $\widehat{df}$ is estimated as the number of nonzero variables in $\hat{\beta}(\lambda)$ (Zou et al., 2007), and $\hat{\sigma}^2$ in (11) is estimated based on the saturated model. In (12), T^s and T^{-s} are the training and validation sets in CV, and $\hat{\beta}^{(s)}(\lambda)$ is the estimated β using the training set T^s and tuning parameter λ. The optimal $\hat{\lambda}$ is then selected as the one that minimizes the corresponding $C_p(\lambda)$, $BIC(\lambda)$, $CV(\lambda)$, or $GCV(\lambda)$, respectively.

To assess the performance of each selection criterion, we report the percentage of selecting the true model over all replicates, as well as the number of correctly selected zeros and incorrectly selected zeros in $\hat{\beta}(\hat{\lambda})$. The final estimator $\hat{\beta}(\hat{\lambda})$ is obtained by refitting the standard least squares regression based only on the selected informative variables. We then compare the prediction performance through the relative prediction error $RPE = E(\mathbf{x}^T \hat{\beta}(\hat{\lambda}) - \mathbf{x}^T \beta)^2/\sigma^2$ (Zou, 2006).

5.1 Scenario I: Fixed p

The simulated data sets $(\mathbf{x}_i, y_i)_{i=1}^n$ are generated from the model

$$y = \mathbf{x}^T \beta + \sigma\varepsilon = \sum_{j=1}^{8} \mathbf{x}_{(j)}\beta_j + \sigma\varepsilon,$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\sigma = 1$, $\mathbf{x}_{(j)}$ and ε are generated from standard normal distribution, and the correlation between $\mathbf{x}_{(i)}$ and $\mathbf{x}_{(j)}$ is set as $0.5^{|i-j|}$. This example has been commonly used in literature, including Tibshirani (1996), Fan and Li (2001), and Wang et al. (2007).

For comparison, we set $n = 40, 60$ or 80 and implement the lasso regression, the adaptive lasso and the SCAD as the base variable selection methods. The lasso regression and the adaptive lasso

are implemented by package 'lars' (Efron et al., 2004) and the SCAD is implemented by package 'ncvreg' (Breheny and Huang, 2011) in R. The tuning parameter λ's are selected via each selection criterion, optimized through a grid search over 100 points $\{10^{-2+4l/99}; l = 0, \ldots, 99\}$. The number of splittings B for the kappa selection criterion is 20. Each simulation is replicated 100 times, and the percentages of selecting the true active set, the average numbers of correctly selected zeros (C) and incorrectly selected zeros (I), and the relative prediction errors (RPE) are summarized in Tables 1-2 and Figure 1.

n	Penalty	Ks	Cp	BIC	CV	GCV
	Lasso	**0.63**	0.16	0.26	0.09	0.16
40	Ada lasso	**0.98**	0.53	0.72	0.63	0.52
	SCAD	**0.98**	0.55	0.78	0.76	0.52
	Lasso	**0.81**	0.16	0.32	0.14	0.17
60	Ada lasso	**0.99**	0.52	0.84	0.65	0.52
	SCAD	**1**	0.58	0.86	0.76	0.56
	Lasso	**0.89**	0.16	0.38	0.08	0.16
80	Ada lasso	**0.99**	0.56	0.86	0.77	0.56
	SCAD	**0.99**	0.62	0.89	0.75	0.61

Table 1: The percentages of selecting the true active set for various selection criteria in simulations of Section 5.1. Here 'Ks', 'Cp', 'BIC', 'CV' and 'GCV' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.

		Ks	Ks	Cp	Cp	BIC	BIC	CV	CV	GCV	GCV
n	Penalty	C	I	C	I	C	I	C	I	C	I
	Lasso	**4.58**	0.01	3.26	0	3.60	0	2.66	0	3.25	0
40	Ada lasso	**4.98**	0	4.16	0	4.54	0	4.25	0	4.15	0
	SCAD	**4.99**	0.01	4.11	0	4.59	0	4.39	0	4.06	0
	Lasso	**4.80**	0	3.12	0	3.91	0	2.85	0	3.13	0
60	Ada lasso	**4.99**	0	4.17	0	4.80	0	4.35	0	4.17	0
	SCAD	**5**	0	4.15	0	4.79	0	4.37	0	4.12	0
	Lasso	**4.88**	0	3.01	0	4.02	0	2.66	0	3	0
80	Ada lasso	**4.99**	0	4.19	0	4.80	0	4.49	0	4.19	0
	SCAD	**4.99**	0	4.23	0	4.83	0	4.45	0	4.22	0

Table 2: The average numbers of correctly selected zeros (C) and incorrectly selected zeros (I) for various selection criteria in simulations of Section 5.1. Here 'Ks', 'Cp', 'BIC', 'CV' and 'GCV' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.

Evidently, the proposed kappa selection criterion delivers superior performance against its competitors in terms of both variable selection accuracy and relative prediction error. As shown in Table 1, the kappa selection criterion has the largest probability of choosing the true active set and consistently outperforms other selection criteria, especially when the lasso regression is used as the base variable selection method. As the sample size n increases, the percentage of selecting the true active set is also improving, which supports the selection consistency in Section 4.

Table 2 shows that the kappa selection criterion yields the largest number of correctly selected zeros in all scenarios, and it yields almost perfect performance for the adaptive lasso and the SCAD. In addition, all selection criteria barely select any incorrect zeros, whereas the kappa selection criterion is relatively more aggressive in that it has small chance to shrink some informative variables to zeros when sample size is small. All other criteria tend to be conservative and include some uninformative variables, so the numbers of correctly selected zeros are significantly less than 5.

Besides the superior variable selection performance, the kappa selection criterion also delivers accurate prediction performance and yields small relative prediction error as displayed in Figure 1. Note that other criteria, especially C_p and GCV, produce large relative prediction errors, which could be due to their conservative selection of the informative variables.
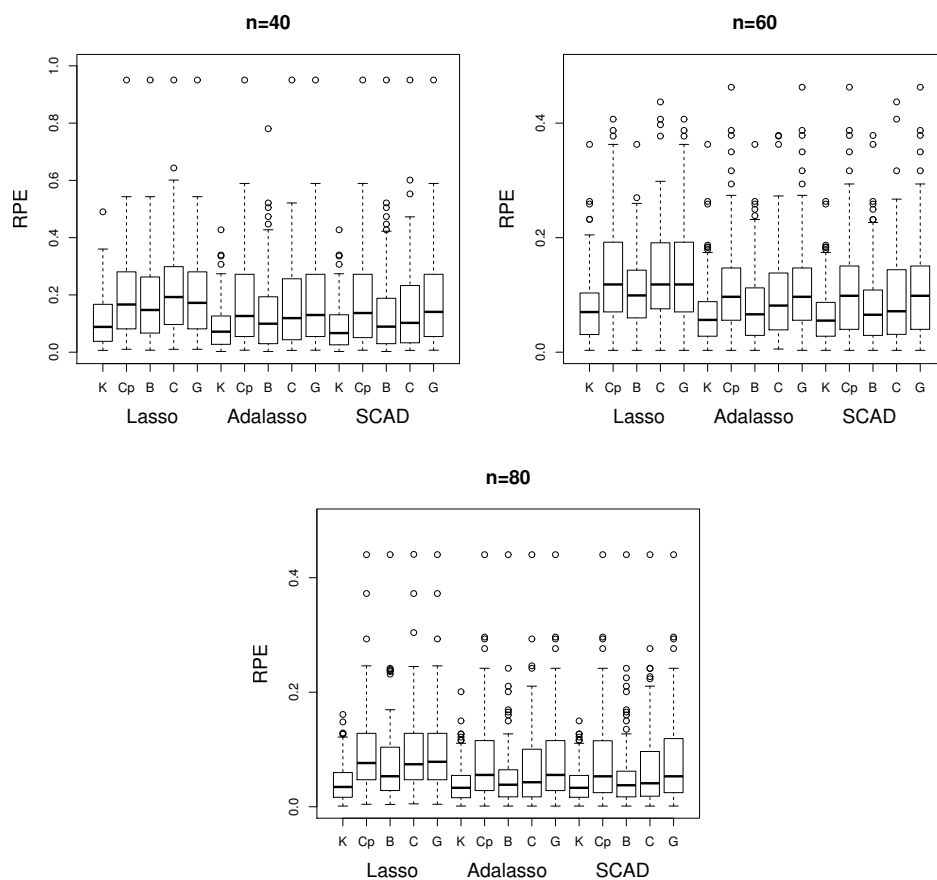

Figure 1: Relative prediction errors (RPE) for various selection criteria in simulations of Section 5.1. Here 'K', 'Cp', 'B', 'C' and 'G' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.

To illustrate the effectiveness of the kappa selection criterion, we randomly select one replication with $n = 40$ and display the estimated variable selection stability as well as the results of detection and sparsity for various λ's for the lasso regression. The detection is defined as the per-
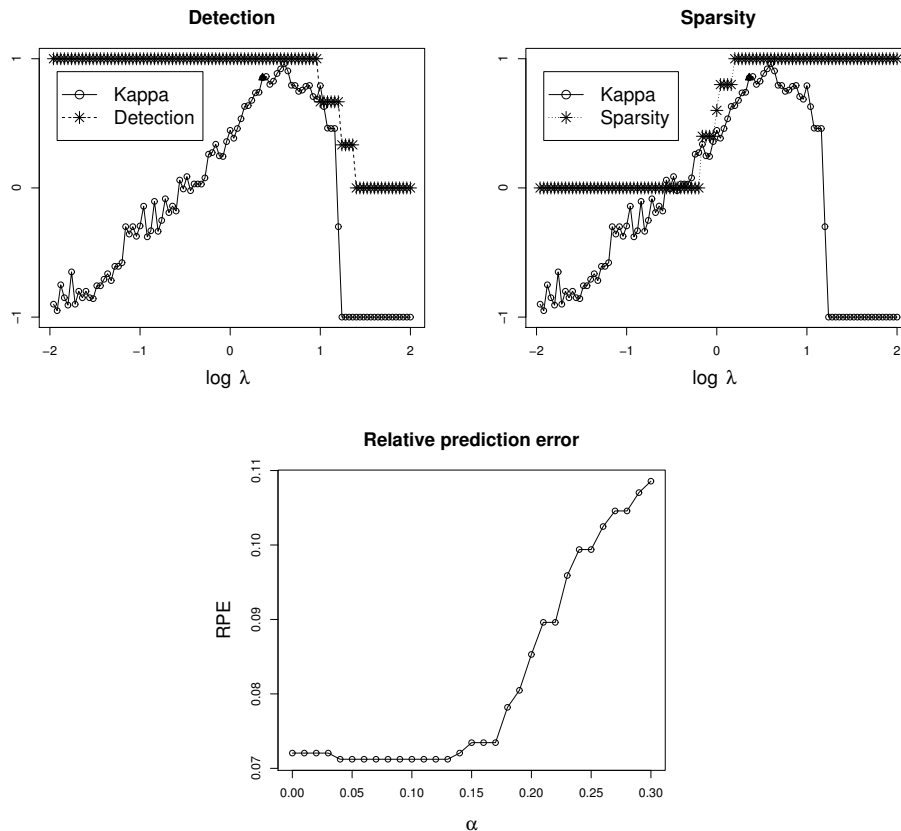
Figure 2: The detection and sparsity of the lasso regression with the kappa selection criterion are shown on the top and the sensitivity of α to the relative prediction error is shown on the bottom. The optimal $\log(\lambda)$ selected by the kappa selection criterion is denoted as the filled triangle in the detection and sparsity plots.

centage of selecting the truly informative variables, and the sparsity is defined as the percentage of excluding the truly uninformative variables. Figure 2 illustrates the clear relationship between the variable selection stability and the values of detection and sparsity. More importantly, the selection performance of the kappa selection criterion is very stable against α_n when it is small. Specifically, we apply the kappa selection criterion on the lasso regression for $\alpha_n = \{\frac{l}{100}; \ l = 0, \ldots, 30\}$ and compute the corresponding average RPE over 100 replications. As shown in the last panel of Figure 2, the average RPE's are almost the same for $\alpha_n \in (0, 0.13)$, which agrees with the theoretical result in Section 4.

5.2 Scenario II: Diverging p_n

To investigate the effects of the noise level and the dimensionality, we compare all the selection criteria in the diverging p_n scenario with a similar simulation model as in Scenario I, except that $\beta = (5, 4, 3, 2, 1, 0, \cdots, 0)^T$, $p_n = [\sqrt{n}]$, and $\sigma = 1$ or 6. More specifically, 8 cases are examined:

$n = 100$, $p_n = 10$; $n = 200$, $p_n = 14$; $n = 400$, $p_n = 20$; and $n = 800$, $p_n = 28$, with $\sigma = 1$ or 6 respectively. Note that when $\sigma = 6$, the truly informative variables are much more difficult to detect due to the increased noise level. The percentages of selecting the true active set, the average numbers of correctly selected zeros (C) and incorrectly selected zeros (I), and the relative prediction errors (RPE) are summarized in Tables 3-4 and Figures 3-4.

n	p_n	Penalty	Ks	Cp	BIC	CV	GCV
					$\sigma = 1$		
		Lasso	**0.98**	0.17	0.43	0.10	0.17
100	10	Ada lasso	**0.99**	0.48	0.86	0.74	0.47
		SCAD	**0.97**	0.47	0.92	0.82	0.47
		Lasso	**1**	0.11	0.49	0.07	0.11
200	14	Ada lasso	**1**	0.38	0.90	0.66	0.38
		SCAD	**1**	0.46	0.93	0.73	0.47
		Lasso	**1**	0.09	0.53	0.04	0.09
400	20	Ada lasso	**1**	0.34	0.93	0.73	0.33
		SCAD	**1**	0.43	0.98	0.75	0.43
		Lasso	**1**	0.11	0.51	0.04	0.11
800	28	Ada lasso	**1**	0.30	0.96	0.74	0.29
		SCAD	**1**	0.46	0.99	0.71	0.46
					$\sigma = 6$		
		Lasso	**0.35**	0.14	0.31	0.11	0.15
100	10	Ada lasso	**0.21**	0.15	0.18	0.11	0.15
		SCAD	**0.17**	0.07	0.12	0.12	0.07
		Lasso	**0.52**	0.10	0.39	0.08	0.09
200	14	Ada lasso	**0.40**	0.18	0.30	0.16	0.18
		SCAD	**0.24**	0.09	0.15	0.13	0.09
		Lasso	**0.77**	0.10	0.47	0.04	0.10
400	20	Ada lasso	0.53	0.22	**0.57**	0.24	0.19
		SCAD	**0.40**	0.13	0.30	0.13	0.13
		Lasso	**0.82**	0.07	0.51	0.04	0.06
800	28	Ada lasso	**0.68**	0.20	0.66	0.37	0.20
		SCAD	**0.46**	0.21	0.39	0.17	0.21

Table 3: The percentages of selecting the true active set for various selection criteria in simulations of Section 5.2. Here 'Ks', 'Cp', 'BIC', 'CV' and 'GCV' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.

In the low noise case with $\sigma = 1$, the proposed kappa selection criterion outperforms other competitors in both variable selection and prediction performance. As illustrated in Tables 3-4, the kappa selection criterion delivers the largest percentage of selecting the true active set among all the selection criteria, and achieves perfect variable selection performance for all the variable selection methods when $n \geq 200$. Furthermore, as shown in Figure 3, the kappa selection criterion yields the smallest relative prediction error across all cases.

As the noise level increases to $\sigma = 6$, the kappa selection criterion still delivers the largest percentage of selecting the true active set among all scenarios except for the adaptive lasso with $n = 400$, where the percentage is slightly smaller than that of BIC. As shown in Table 4, the kappa

n	p_n	Penalty	Ks C	Ks I	Cp C	Cp I	BIC C	BIC I	CV C	CV I	GCV C	GCV I
							$\sigma = 1$					
		Lasso	**5**	0.02	3.25	0	4.20	0	2.95	0	3.25	0
100	10	Ada lasso	**5**	0.01	4.23	0	4.84	0	4.48	0	4.21	0
		SCAD	**5**	0.03	4.12	0	4.91	0	4.67	0	4.15	0
		Lasso	**9**	0	6.18	0	8.26	0	5.62	0	6.18	0
200	14	Ada lasso	**9**	0	7.50	0	8.87	0	8.24	0	7.50	0
		SCAD	**9**	0	7.43	0	8.91	0	8.26	0	7.47	0
		Lasso	**15**	0	11.29	0	14.23	0	10.56	0	11.29	0
400	20	Ada lasso	**15**	0	12.93	0	14.92	0	14.28	0	12.91	0
		SCAD	**15**	0	12.67	0	14.98	0	14.21	0	12.64	0
		Lasso	**23**	0	18.49	0	22.27	0	18.20	0	18.63	0
800	28	Ada lasso	**23**	0	20.31	0	22.94	0	22.07	0	20.23	0
		SCAD	**23**	0	20.21	0	22.99	0	21.95	0	20.21	0
							$\sigma = 6$					
		Lasso	**4.76**	0.57	3.27	0.24	4.31	0.35	3.09	0.20	3.28	0.24
100	10	Ada lasso	4.57	0.77	3.81	0.54	**4.62**	0.85	3.31	0.49	3.84	0.54
		SCAD	**4.88**	1.22	3.63	0.56	4.37	0.94	3.52	0.58	3.65	0.56
		Lasso	**8.93**	0.43	6.20	0.08	8.32	0.21	5.79	0.07	6.22	0.08
200	14	Ada lasso	**8.72**	0.55	7.28	0.32	8.69	0.56	7.34	0.37	7.26	0.32
		SCAD	**9**	0.95	7.07	0.37	8.37	0.63	7.25	0.44	7.07	0.37
		Lasso	**14.98**	0.21	11.46	0.03	14.21	0.07	10.60	0.03	11.45	0.03
400	20	Ada lasso	**14.88**	0.40	12.24	0.09	14.80	0.30	12.93	0.15	12.16	0.09
		SCAD	**15**	0.67	11.97	0.13	14.65	0.51	12.66	0.23	11.88	0.12
		Lasso	**22.99**	0.17	18.65	0.01	22.27	0.01	18.14	0.01	18.68	0.01
800	28	Ada lasso	**22.96**	0.29	19.84	0.02	22.71	0.16	21.19	0.04	19.71	0.02
		SCAD	**23**	0.55	19.55	0.04	22.73	0.37	20.42	0.11	19.47	0.04

Table 4: The average numbers of correctly selected zeros (C) and incorrectly selected zeros (I) for various selection criteria in simulations of Section 5.2. Here 'Ks', 'Cp', 'BIC', 'CV' and 'GCV' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.

selection criterion yields the largest number of correctly selection zeros. However, it has relatively higher chance of shrinking the fifth informative variable to zero, while the chance is diminishing as n increases. This phenomenon is also present for BIC. Considering the smaller relative prediction errors achieved by the kappa selection criterion and BIC, these two criteria tend to produce sparser models with satisfactory prediction performance. In practice, if false negatives are of concern, one can increase the thresholding value α_n in the kappa selection criterion, to allow higher tolerance of instability and hence decrease the chance of claiming false negatives. In addition, as shown in Figure 4, the kappa selection criterion yields the smallest relative prediction error for the lasso regression and the adaptive lasso among all scenarios, whereas the advantage is considerably less significant for the SCAD. This is somewhat expected as the SCAD is sensitive to the noise level (Zou, 2006), which may lead to inaccurate estimation of the variable selection stability.
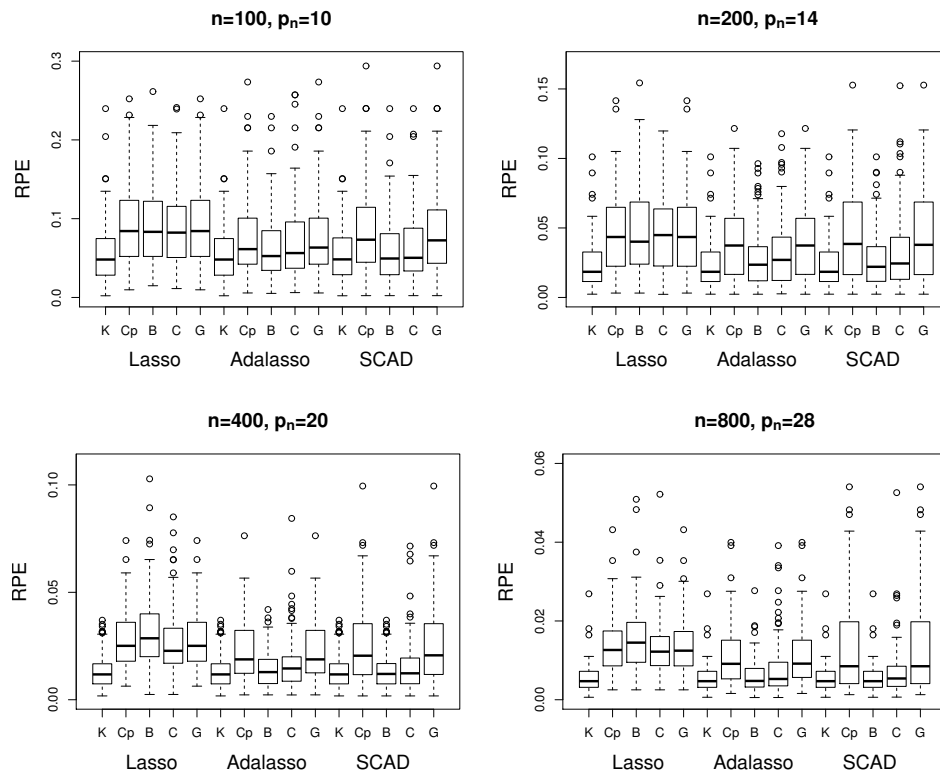
Figure 3: Relative prediction errors (RPE) for various selection criteria in Scenario 2 with $\sigma = 1$. Here 'K', 'Cp', 'B', 'C' and 'G' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.

6. Real Application

In this section, we apply the kappa selection criterion to the prostate cancer data (Stamey et al., 1989), which were used to study the relationship between the level of log(prostate specific antigen) (*lpsa*) and a number of clinical measures. The data set consisted of 97 patients who had received a radical prostatectomy, and eight clinical measures were log(cancer volume) (*lcavol*), log(prostate weight) (*lweight*), *age*, log(benign prostaic hyperplasia amount) (*lbph*), seminal vesicle invasion (*svi*), log(capsular penetration) (*lcp*), Gleason score (*gleason*) and percentage Gleason scores 4 or 5 (*pgg*45).

The data set is randomly split into two halves: a training set with 67 patients and a test set with 30 patients. Similarly as in the simulated examples, the tuning parameter λ's are selected through a grid search over 100 grid points $\{10^{-2+4l/99}; l = 0, \ldots, 99\}$ on the training set. Since it is unknown whether the clinical measures are truly informative or not, the performance of all the selection criteria are compared by computing their corresponding prediction errors on the test set in Table 5.

As shown in Table 5, the proposed kappa selection criterion yields the sparsest model and achieves the smallest prediction error for the lasso regression and the SCAD, while the predic-
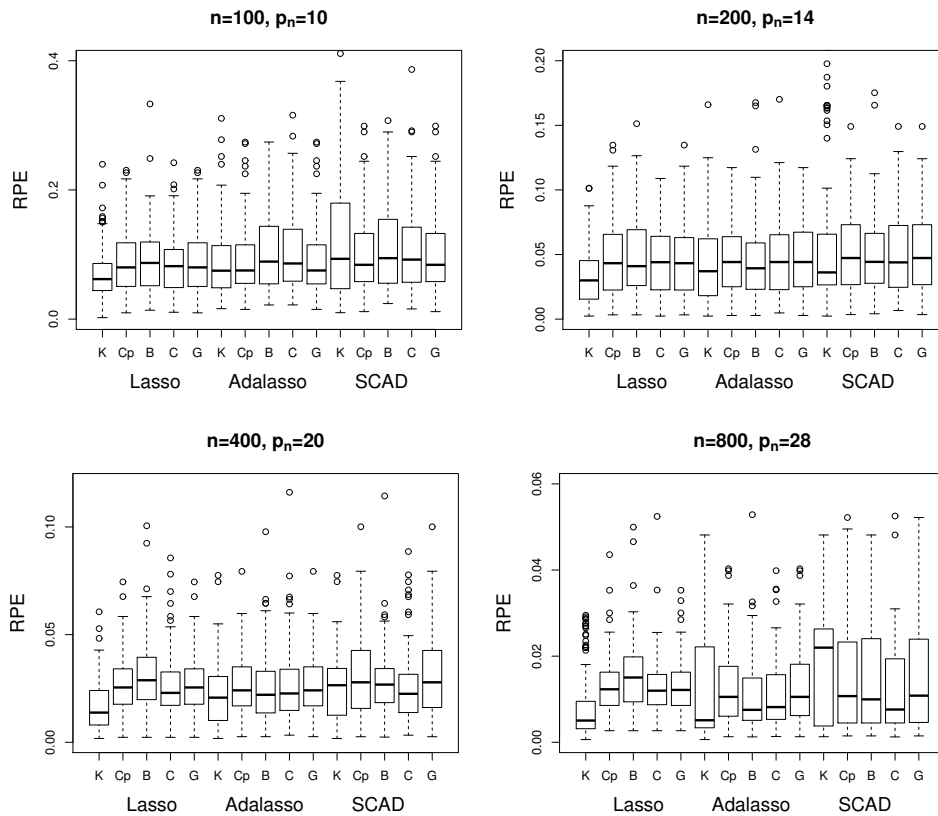
Figure 4: Relative prediction errors (RPE) for various selection criteria in Scenario 2 with $\sigma = 6$. Here 'K', 'Cp', 'B', 'C' and 'G' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.

	Penalty	Ks	Cp	BIC	CV	GCV
Active	Lasso	1,2,4,5	1,2,3,4,5,6,7,8	1,2,4,5	1,2,3,4,5,7,8	1,2,3,4,5,6,7,8
Set	Ada lasso	1,2,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5,6,7,8	1,2,3,4,5
	SCAD	1,2,4,5	1,2,3,4,5	1,2,3,4,5	1,2,3,4,5,6,7,8	1,2,3,4,5
	Lasso	**0.734**	0.797	**0.734**	0.807	0.797
PE	Ada lasso	0.806	0.825	0.825	**0.797**	0.825
	SCAD	**0.734**	0.825	0.825	0.797	0.825

Table 5: The selected active sets and the prediction errors (PE) for various selection criteria in the prostate cancer example. Here 'Ks', 'Cp', 'BIC', 'CV' and 'GCV' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.

tion error for the adaptive lasso is comparable to the minima. Specifically, the lasso regression and the SCAD with the kappa selection criterion include *lcavol*, *lweight*, *lbph* and *svi* as the informative variables, and the adaptive lasso with the kappa selection criterion selects only *lcavol*, *lweight*

and *svi* as the informative variables. As opposed to the sparse regression models produced by other selection criteria, the variable *age* is excluded by the kappa selection criterion for all base variable selection methods, which agrees with the findings in Zou and Hastie (2005).

7. Discussion

This article proposes a tuning parameter selection criterion based on the concept of variable selection stability. Its key idea is to select the tuning parameter so that the resultant variable selection method is stable in selecting the informative variables. The proposed criterion delivers superior numerical performance in a variety of experiments. Its asymptotic selection consistency is also established for both fixed and diverging dimensions. Furthermore, it is worth pointing out that the idea of stability is general and can be naturally extended to a broader framework of model selection, such as the penalized nonparametric regression (Xue et al., 2010) and the penalized clustering (Sun et al., 2012).

8. Supplementary Material

Lemmas 2 and 3 and their proofs are provided as online supplementary material for this article.

Acknowledgments

The authors thank the Action Editor and three referees for their constructive comments and suggestions, which have led to a significantly improved paper.

Appendix A.

The lemma stated below shows that if a variable selection method is selection consistent and $\varepsilon_n \prec \alpha_n$, then its variable selection stability converges to 1 in probability.

Lemma 3 *Let λ_n^* be as defined in Assumption 1. For any α_n,*

$$P\left(\hat{s}(\Psi, \lambda_n^*, m) \geq 1 - \alpha_n \right) \geq 1 - 2\varepsilon_n/\alpha_n.$$

Proof of Lemma 3: We denote $\widehat{\mathcal{A}}_{1\lambda_n^*}^{*b}$ and $\widehat{\mathcal{A}}_{2\lambda_n^*}^{*b}$ as the corresponding active sets obtained from two sub-samples at the b-th random splitting. Then the estimated variable selection stability based on the b-th splitting can be bounded as

$$P\left(\hat{s}^{*b}(\Psi, \lambda_n^*, m) = 1 \right) = P\left(\widehat{\mathcal{A}}_{1\lambda_n^*}^{*b} = \widehat{\mathcal{A}}_{2\lambda_n^*}^{*b} \right) \geq P\left(\widehat{\mathcal{A}}_{1\lambda_n^*}^{*b} = \mathcal{A}_T \right)^2 \geq (1 - \varepsilon_n)^2 \geq 1 - 2\varepsilon_n.$$

By the fact that $0 \leq \hat{s}^{*b}(\Psi, \lambda_n^*, n) \leq 1$, we have

$$E\left(\hat{s}(\Psi, \lambda_n^*, m) \right) = E\left(B^{-1} \sum_{b=1}^{B} \hat{s}^{*b}(\Psi, \lambda_n^*, m) \right) \geq 1 - 2\varepsilon_n,$$

and $0 \leq \hat{s}(\Psi, \lambda_n^*, n) \leq 1$. Finally, Markov inequality yields that

$$P\left(1 - \hat{s}(\Psi, \lambda_n^*, m) \geq \alpha_n\right) \leq \frac{E\left(1 - \hat{s}(\Psi, \lambda_n^*, m)\right)}{\alpha_n} \leq \frac{2\varepsilon_n}{\alpha_n},$$

which implies the desired result immediately. ■

Proof of Theorem 1: We first show that for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P\left(\hat{\lambda}_n > \lambda_n^* \text{ or } r_n^{-1}\hat{\lambda}_n \leq 1/\varepsilon\right) = 0.$$

Denote $\Omega_1 = \{\lambda_n : \lambda_n > \lambda_n^*\}$, $\Omega_2 = \{\lambda_n : r_n^{-1}\lambda_n \leq \tau\}$ and $\Omega_3 = \{\lambda_n : \tau \leq r_n^{-1}\lambda_n \leq 1/\varepsilon\}$, where $\tau < 1/\varepsilon$, $c_1(\tau) \geq 1 - 1/p$. It then suffices to show that for any $\varepsilon > 0$,

$$P\left(\hat{\lambda}_n \in \Omega_1 \cup \Omega_2 \cup \Omega_3\right) \to 0.$$

First, the definition of $\hat{\lambda}_n$ and Lemma 1 imply that

$$P(\hat{\lambda}_n \leq \lambda_n^*) \geq P\left(\frac{\hat{s}(\Psi, \lambda_n^*, m)}{\max_\lambda \hat{s}(\Psi, \lambda, m)} \geq 1 - \alpha_n\right) \geq P\left(\hat{s}(\Psi, \lambda_n^*, m) \geq 1 - \alpha_n\right) \geq 1 - \frac{2\varepsilon_n}{\alpha_n}.$$

This, together with $\varepsilon_n \prec \alpha_n$, yields that

$$P\left(\hat{\lambda}_n \in \Omega_1\right) = P(\hat{\lambda}_n > \lambda_n^*) \leq \frac{2\varepsilon_n}{\alpha_n} \to 0.$$

Next, the definition of $\hat{\lambda}_n$ implies that

$$\hat{s}(\Psi, \hat{\lambda}_n, m) \geq (1 - \alpha_n) \max_\lambda \hat{s}(\Psi, \lambda, m) \geq (1 - \alpha_n)\hat{s}(\Psi, \lambda_n^*, m).$$

This, together with Lemma 1, leads to

$$P\left(\hat{s}(\Psi, \hat{\lambda}_n, m) \geq 1 - 2\alpha_n\right) \geq P\left(\hat{s}(\Psi, \hat{\lambda}_n, m) \geq (1 - \alpha_n)^2\right)$$
$$\geq P\left(\hat{s}(\Psi, \lambda_n^*, m) \geq 1 - \alpha_n\right) \geq 1 - \frac{2\varepsilon_n}{\alpha_n},$$

and hence when $\varepsilon_n \prec \alpha_n$,
$$P\left(\hat{s}(\Psi, \hat{\lambda}_n, m) \geq 1 - 2\alpha_n\right) \to 1.$$

Therefore, to show $P(\hat{\lambda}_n \in \Omega_2) \to 0$, it suffices to show

$$P\left(\sup_{\lambda_n \in \Omega_2} \hat{s}(\Psi, \lambda_n, m) < 1 - 2\alpha_n\right) \to 1. \tag{13}$$

But Assumption 2 implies that for any $j \in \mathcal{A}_T^c$ and $j_1 \in \mathcal{A}_T$, we have

$$P\left(j \in \bigcap_{\lambda_n \in \Omega_2} \widehat{\mathcal{A}}_{\lambda_n}\right) \geq c_1(\tau) \geq 1 - \frac{1}{p} \quad \text{and} \quad P\left(j_1 \in \bigcap_{\lambda_n \in \Omega_2} \widehat{\mathcal{A}}_{\lambda_n}\right) \geq 1 - \zeta_n.$$

It implies that

$$\lim_{n\to\infty} P\Big(\{1,\dots,p\} \in \bigcap_{\lambda_n\in\Omega_2} \widehat{\mathcal{A}}_{\lambda_n}\Big)$$

$$\geq \lim_{n\to\infty} 1 - \sum_{j\in\mathcal{A}_T^c} P\Big(j \notin \bigcap_{\lambda_n\in\Omega_2} \widehat{\mathcal{A}}_{\lambda_n}\Big) - \sum_{j_1\in\mathcal{A}_T} P\Big(j_1 \notin \bigcap_{\lambda_n\in\Omega_2} \widehat{\mathcal{A}}_{\lambda_n}\Big)$$

$$\geq \lim_{n\to\infty} 1 - \frac{p-p_0}{p} - p_0\zeta_n = \frac{p_0}{p} > 0.$$

Since $\{1,\dots,p\} \in \bigcap_{\lambda_n\in\Omega_2} \widehat{\mathcal{A}}_{\lambda_n}^{*b}$ implies $\sup_{\lambda_n\in\Omega_2} \hat{s}^{*b}(\Psi,\lambda_n,m) = -1$, then

$$\lim_{n\to\infty} E\Big(\sup_{\lambda_n\in\Omega_2} \hat{s}^{*b}(\Psi,\lambda_n,m)\Big) \leq 1 - \lim_{n\to\infty} P\Big(\sup_{\lambda_n\in\Omega_2} \hat{s}^{*b}(\Psi,\lambda_n,m) = -1\Big) \leq 1 - \frac{p_0}{p}.$$

In addition, the strong law of large number for U-statistics (Hoeffding, 1961) implies that

$$B^{-1} \sum_{b=1}^{B} \sup_{\lambda_n\in\Omega_2} \hat{s}^{*b}(\Psi,\lambda_n,m) \xrightarrow{a.s.} E\Big(\sup_{\lambda_n\in\Omega_2} \hat{s}^{*b}(\Psi,\lambda_n,m)\Big) \text{ as } B\to\infty.$$

Note that $\sup_{\lambda_n\in\Omega_2} \hat{s}(\Psi,\lambda_n,m) \leq B^{-1}\sum_{b=1}^{B} \sup_{\lambda_n\in\Omega_2} \hat{s}^{*b}(\Psi,\lambda_n,m)$, it then follows immediately that $P(\sup_{\lambda_n\in\Omega_2} \hat{s}(\Psi,\lambda_n,m) \leq 1 - \frac{p_0}{p}) \to 1$ and hence $P(\sup_{\lambda_n\in\Omega_2} \hat{s}(\Psi,\lambda_n,m) < 1 - 2\alpha_n) \to 1$. Therefore $P(\hat{\lambda}_n \in \Omega_2) \to 0$.

Finally, to show $P(\hat{\lambda}_n \in \Omega_3) \to 0$, it also suffices to show

$$P\Big(\sup_{\lambda_n\in\Omega_3} \hat{s}(\Psi,\lambda_n,m) < 1 - 2\alpha_n\Big) \to 1. \tag{14}$$

Assumption 2 implies that for any $j \in \mathcal{A}_T^c$ and some $j_1 \in \mathcal{A}_T^c$, when n is sufficiently large,

$$P(\cap_{\lambda_n\in\Omega_3}\{j \in \widehat{\mathcal{A}}_{\lambda_n}\}) \geq c_1(1/\varepsilon) > 0 \text{ and } P(\cap_{\lambda_n\in\Omega_3}\{j_1 \notin \widehat{\mathcal{A}}_{\lambda_n}\}) \geq c_2(\tau) > 0.$$

Therefore, it follows from the independence between two sub-samples that

$$P\Big(\bigcap_{\lambda_n\in\Omega_3} \{\widehat{\mathcal{A}}_{1\lambda_n}^{*b} \neq \widehat{\mathcal{A}}_{2\lambda_n}^{*b}\}\Big) \geq P\Big(\bigcap_{\lambda_n\in\Omega_3} \bigcup_{j\in\mathcal{A}_T^c} \{j \notin \widehat{\mathcal{A}}_{1\lambda_n}^{*b}, j \in \widehat{\mathcal{A}}_{2\lambda_n}^{*b}\}\Big)$$

$$\geq P\Big(\bigcap_{\lambda_n\in\Omega_3} \{j_1 \notin \widehat{\mathcal{A}}_{1\lambda_n}^{*b}, j_1 \in \widehat{\mathcal{A}}_{2\lambda_n}^{*b}\}\Big)$$

$$= P\Big(\bigcap_{\lambda_n\in\Omega_3} \{j_1 \notin \widehat{\mathcal{A}}_{1\lambda_n}^{*b}\}\Big) P\Big(\bigcap_{\lambda_n\in\Omega_3} \{j_1 \in \widehat{\mathcal{A}}_{2\lambda_n}^{*b}\}\Big),$$

$$\geq c_1(1/\varepsilon)c_2(\tau).$$

Since the event $\bigcap_{\lambda_n\in\Omega_3}\{\widehat{\mathcal{A}}_{1\lambda_n}^{*b} \neq \widehat{\mathcal{A}}_{2\lambda_n}^{*b}\}$ implies that $\sup_{\lambda_n\in\Omega_3} \hat{s}^{*b}(\Psi,\lambda_n,m) \leq c_3$ with $c_3 = \max_{\mathcal{A}_1\neq\mathcal{A}_2} \kappa(\mathcal{A}_1,\mathcal{A}_2) \leq \frac{p-1}{p}$ where $\mathcal{A}_1,\mathcal{A}_2 \subset \{1,\cdots,p\}$, we have, for sufficiently large n,

$$P\Big(\sup_{\lambda_n\in\Omega_3} \hat{s}^{*b}(\Psi,\lambda_n,m) \leq c_3\Big) \geq c_1(1/\varepsilon)c_2(\tau).$$

Therefore, for sufficiently large n and any $b > 0$,

$$E\left(\sup_{\lambda_n \in \Omega_3} \hat{s}^{*b}(\Psi, \lambda_n, m)\right) \leq 1 - c_1(1/\varepsilon)c_2(\tau)(1 - c_3).$$

Again, by the strong law of large number for U-statistics (Hoeffding, 1961) and the fact that $\sup_{\lambda_n \in \Omega_3} \hat{s}(\Psi, \lambda_n, m) \leq B^{-1} \sum_{b=1}^{B} \sup_{\lambda_n \in \Omega_3} \hat{s}^{*b}(\Psi, \lambda_n, m)$, we have

$$P\left(\sup_{\lambda_n \in \Omega_3} \hat{s}(\Psi, \lambda_n, m) \leq 1 - c_1(1/\varepsilon)c_2(\tau)(1 - c_3)\right) \to 1.$$

For any ε, $c_1(1/\varepsilon)c_2(\tau)(1 - c_3)$ is strictly positive and $\alpha_n \to 0$, we have

$$P\left(\sup_{\lambda_n \in \Omega_3} \hat{s}(\Psi, \lambda_n, m) < 1 - 2\alpha_n\right) \geq P\left(\sup_{\lambda_n \in \Omega_3} \hat{s}(\Psi, \lambda_n, m) \leq 1 - c_1(1/\varepsilon)c_2(\tau)(1 - c_3)\right) \to 1,$$

and hence (14) is verified and $P\left(\hat{\lambda}_n \in \Omega_3\right) \to 0$.

Combining the above results, we have for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \lim_{B \to \infty} P\left(r_n/\varepsilon \leq \hat{\lambda}_n \leq \lambda_n^*\right) = 1. \tag{15}$$

Furthermore, since for any $\varepsilon > 0$,

$$
\begin{aligned}
P(\widehat{\mathcal{A}}_{\hat{\lambda}_n} = \mathcal{A}_T) &\geq P\left(\widehat{\mathcal{A}}_{\hat{\lambda}_n} = \mathcal{A}_T,\ r_n/\varepsilon \leq \hat{\lambda}_n \leq \lambda_n^*\right) \\
&\geq P\left(\bigcap_{r_n/\varepsilon \leq \lambda_n \leq \lambda_n^*} \{\widehat{\mathcal{A}}_{\lambda_n} = \mathcal{A}_T\}\right) + P\left(r_n/\varepsilon \leq \hat{\lambda}_n \leq \lambda_n^*\right) - 1.
\end{aligned}
$$

Therefore, the desired selection consistency directly follows from (15) and Assumption 1 by letting $\varepsilon \to 0$. ∎

Proof of Theorem 2: We prove Theorem 2 by similar approach as in the proof of Theorem 1. For any $\varepsilon > 0$, we denote $\Omega_1 = \{\lambda_n : \lambda_n > \lambda_n^*\}$, $\Omega_2 = \{\lambda_n : r_n^{-1}\lambda_n \leq \tau\}$ and $\Omega_3 = \{\lambda_n : \tau \leq r_n^{-1}\lambda_n \leq 1/\varepsilon\}$, where τ is selected so that $\tau < 1/\varepsilon$ and $p_n(1 - c_{1n}(\tau)) \succ \alpha_n$. Then we just need to show that $P(\hat{\lambda}_n \in \Omega_1 \cup \Omega_2 \cup \Omega_3) \to 0$. The probability $P(\hat{\lambda}_n \in \Omega_1) \to 0$ for any $\varepsilon > 0$ can be proved similarly as in Theorem 1.

In addition, Lemma 1 implies that $P(\hat{s}(\Psi, \lambda_n^*, m) \geq 1 - \alpha_n) \geq 1 - 2\varepsilon_n/\alpha_n$, and the definition of $\hat{\lambda}_n$ leads to $P(\hat{s}(\Psi, \hat{\lambda}_n, m) \geq (1 - \alpha_n)(1 - \alpha_n)) \geq 1 - 2\varepsilon_n/\alpha_n$, and hence

$$P\left(\hat{s}(\Psi, \hat{\lambda}_n, m) \geq 1 - 2\alpha_n\right) \geq P\left(\hat{s}(\Psi, \hat{\lambda}_n, m) \geq (1 - \alpha_n)(1 - \alpha_n)\right) \geq 1 - \frac{2\varepsilon_n}{\alpha_n} \to 1.$$

To show $P(\hat{\lambda}_n \in \Omega_2) \to 0$, it suffices to show $P(\sup_{\lambda_n \in \Omega_2} \hat{s}(\Psi, \lambda_n, m) < 1 - 2\alpha_n) \to 1$, which can be verified by slightly modifying the proof of (13). Assumption 2a implies that for any $j \in \mathcal{A}_T^c$ and $j_1 \in \mathcal{A}_T$, we have

$$P\left(j \in \bigcap_{\lambda_n \in \Omega_2} \widehat{\mathcal{A}}_{\lambda_n}\right) \geq c_{1n}(\tau) \quad \text{and} \quad P\left(j_1 \in \bigcap_{\lambda_n \in \Omega_2} \widehat{\mathcal{A}}_{\lambda_n}\right) \geq 1 - \zeta_n.$$

As shown in Theorem 1, it implies that

$$P\left(\{1,\ldots,p\} \in \bigcap_{\lambda_n \in \Omega_2} \widehat{\mathscr{A}}_{\lambda_n}\right) \geq 1 - (p_n - p_{0n})(1 - c_{1n}(\tau)) - p_{0n}\zeta_n,$$

and hence $E(\sup_{\lambda_n \in \Omega_2} \hat{s}^{*b}(\Psi, \lambda_n, m)) \leq 1 - (p_n - p_{0n})(1 - c_{1n}(\tau)) - p_{0n}\zeta_n$. By the strong law of large number for U-statistics,

$$P\left(\sup_{\lambda_n \in \Omega_2} \hat{s}(\Psi, \lambda_n, m) \leq 1 - (p_n - p_{0n})(1 - c_{1n}(\tau)) - p_{0n}\zeta_n\right) \to 1.$$

Therefore, $P(\sup_{\lambda_n \in \Omega_2} \hat{s}(\Psi, \lambda_n, m) < 1 - 2\alpha_n) \to 1$ provided that $p_n(1 - c_{1n}(\tau)) \succ \alpha_n$ and $p_n\zeta_n \to 0$.

To show $P(\hat{\lambda}_n \in \Omega_3) \to 0$, it suffices to show

$$P\left(\sup_{\lambda_n \in \Omega_3} \hat{s}(\Psi, \lambda_n, m) < 1 - 2\alpha_n\right) \to 1. \tag{16}$$

Here (16) follows by modifying the proof of (14). According to $c_4 \leq (p_n - 1)/p_n$, we have

$$E\left(\sup_{\lambda_n \in \Omega_3} \hat{s}^{*b}(\Psi, \lambda_n, m)\right) \leq 1 - p_n^{-1} c_{1n}(1/\varepsilon) c_{2n}(\tau).$$

Therefore, following the same derivation as in Theorem 1, we have

$$P\left(\sup_{\lambda_n \in \Omega_3} \hat{s}(\Psi, \lambda_n, m) \leq 1 - p_n^{-1} c_{1n}(1/\varepsilon) c_{2n}(\tau)\right) \to 1.$$

This, together with the assumptions that $\alpha_n \prec p_n^{-1} c_{1n}(1/\varepsilon) c_{2n}(\tau)$ for any ε and τ, leads to the convergence in (16), which completes the proof of Theorem 2. ∎

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723, 1974.

A. Ben-Hur, A. Elisseeff and I. Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, 6-17, 2002.

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5, 232-253, 2011.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46, 1960.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31, 317-403, 1979.

B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32, 407-451, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360, 2001.

J. Fan and R. Li. Statistical Challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians*, 3, 595-622, 2006.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32, 928-961, 2004.

W. Hoeffding. The strong law of large numbers for U-statistics. *Institute of Statistical Mimeograph Series, No. 302, University of North Carolina*, 1961.

J. Huang, S. Ma and C.H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18, 1603-1618, 2008.

J. Huang and H. Xie. Asymptotic oracle properties of SCAD-penalized least squares estimators. *IMS Lecture Notes - Monograph Series, Asymptotics: Particles, Processes and Inverse Problems*, 55, 149-166, 2007.

C. Mallows. Some comments on Cp. *Technometrics*, 15, 661-675, 1973.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72, 414-473, 2010.

G.E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464, 1978.

X. Shen, W. Pan and Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107, 223-232, 2012.

T.A. Stamey, J.N. Kabalin, J.E. McNeal, I.M. Johnstone, F. Freiha, E.A. Redwine and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. Radical prostatectomy treated patients. *Journal of Urology*, 141, 1076-1083, 1989.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36, 111-147, 1974.

W. Sun, J. Wang and Y. Fang. Regularized K-means clustering of high-dimensional data and its asymptotic consistency, *Electronic Journal of Statistics*, 6, 148-167, 2012.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288, 1996.

H. Wang and C. Leng. A note on adaptive group Lasso. *Computational Statistics and Data Analysis*, 52, 5277-5286, 2008.

H. Wang, R. Li and C.L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553-568, 2007.

H. Wang, B. Li and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B*, 71, 671-683, 2009.

J. Wang. Consistent selection of the number of clusters via cross validation. *Biometrika*, 97, 893-904, 2010.

L. Xue, A. Qu and J. Zhou. Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association*, 105, 1518-1530, 2010.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68, 49-67, 2006.

Y. Zhang, R. Li and C.L. Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105, 312-323, 2010.

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

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, 301-320, 2005.

H. Zou, T. Hastie and R. Tibshirani. On the "Degree of Freedom" of the Lasso. *Annals of Statistics*, 35, 2173-2192, 2007.

H. Zou and H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37, 1733-1751, 2009.