Consistent selection of tuning parameters via variable selection stability

Wei Sun
Department of Statistics
Purdue University
West Lafayette, IN 47907
Email: sun244@purdue.edu

Junhui Wang
Department of Mathematics, Statistics,
and Computer Science
University of Illinois at Chicago
Chicago, IL 60607
Email: jwang@math.uic.edu

Yixin Fang
Division of Biostatistics
Department of Population Health
New York University School of Medicine
New York, NY 10016
Email: Yixin.Fang@nyumc.org

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 a novel concept of 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. The effectiveness of the proposed criterion is also demonstrated in a variety of simulated examples as well as an application to the prostate cancer data.

Key words: 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 the 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), BIC (Schwarz, 1978). For instances, 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) also 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 general tuning parameter selection criterion based on a novel concept of variable selection stability. Similar stability measures have been studied in the context of clustering (Ben-Hur et al., 2002; Wang, 2010) and variable selection (Meinshausen and Buhlmann, 2010). 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. The effectiveness of the proposed selection criterion is demonstrated in a variety of simulated examples and real applications. 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. Section 7 provides a direct extension of the proposed kappa selection criterion. A brief discussion is provided in Section 8, and the Appendix is devoted to the technical proofs.

3

2 Penalized least squares regression

Consider the linear regression model

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

where $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T = (\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(p)})$ with $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ or $\mathbf{x}_{(j)} = (x_{1j}, \cdots, x_{nj})^T$, $\beta = (\beta_1, \cdots, \beta_p)^T$, $E(\epsilon) = \mathbf{0}$ and $\text{cov}(\epsilon) = \Sigma$. 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

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

where $\|\cdot\|$ is the Euclidean norm, and $p_\lambda(|\beta_j|)$ is a regularization term encouraging sparsity in β. Widely used regularization terms include the lasso penalty $p_\lambda(\theta) = \lambda\theta$ (Tibshirani, 1996), the SCAD penalty with $p'_\lambda(\theta) = \lambda\big(I(\theta \le \lambda) + \frac{(\gamma\lambda - \theta)_+}{(\gamma - 1)\lambda}I(\theta > \lambda)\big)$ (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 are 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 penalty is selection consistent when $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$;

Zou (2006) showed that the adaptive lasso penalty 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 a novel 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 cardinality of a set. Note that $-1 \leq \kappa(\mathcal{A}_1, \mathcal{A}_2) \leq 1$, 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 $n_{12} = n_{21} = p/2$. Based on (2), the variable selection stability is defined as follows.

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

$$s(\Psi, \lambda, n) = E\Big(\kappa(\Psi(Z_1^n; \lambda), \Psi(Z_2^n; \lambda))\Big), \tag{3}$$

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, $-1 \leq s(\Psi, \lambda, n) \leq 1$, 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 averaged estimated variable selection stability over all splittings is computed. The selected λ is then the one maximizing the averaged estimated variable selection stability. The details of the proposed kappa selection criterion are given 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 the variable selection stability of $\Psi(\cdot; \lambda)$ in the b-th splitting is estimated as

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

Step 3. Repeat *Steps 1-2* for B times. The averaged 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 $\hat{s}(\Psi, \lambda_t, m)$ for a sequence of λ_t's, and set $\hat{\lambda} = \min_{\lambda_t}\{\lambda_t : \lambda_t \in \widehat{\Lambda}_n\}$ with

$$\widehat{\Lambda}_n = \left\{\lambda : \frac{\hat{s}(\Psi, \lambda, m)}{\max_{\lambda_t} \hat{s}(\Psi, \lambda_t, m)} \geq 1 - \alpha_n\right\}.$$

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 suggests that α_n has very little effect on the selection performance when it varies in 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. Furthermore, since the true model is assumed to be sparse and containing at least some informative variables, any λ leading to an active set with all variables or no variable will be excluded from the comparison by setting the corresponding variable selection stability as -1.

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 $p_0 < p$ variables 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 r_n and s_n such that the base variable selection method is selection consistent if $r_n \prec \lambda_n \prec s_n$. That is,

$$P(\widehat{\mathcal{A}}_{\lambda_n} = \mathcal{A}_T) \geq 1 - \epsilon_n, \text{ for some } \epsilon_n \to 0.$$

Assumption 2: For r_n in Assumption 1, if $\lambda_n \preceq r_n$, the base variable selection method is overfitted in that $P(\mathcal{A}_T \subseteq \widehat{\mathcal{A}}_{\lambda_n}) \to 1$ and there exists a constant $c_0 > 0$ such that for sufficiently large n,

$$P\Big(\mathcal{A}_T \cup \{j\} \subseteq \widehat{\mathcal{A}}_{\lambda_n}\Big) \geq c_0, \text{ for any } j \in \mathcal{A}_T^c. \tag{4}$$

In Assumption 1, r_n and s_n specify an asymptotic working interval for λ_n so that the base variable selection method is selection consistent. Assumption 2 is necessary since it implies a natural order of the variable selection stability with respect to λ_n and it excludes the degenerate variable selection methods that always produce the same $\widehat{\mathcal{A}}_{\lambda_n}$ regardless of the training sample. The inequality (4) can be replaced by a slightly stronger assumption that the distribution of $\{X_{(j)}, j \in \mathcal{A}_T^c\}$ is exchangeable and the base variable selection method is no worse than random guessing (Meinshausen and Buhlmann, 2010).

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

Lemma 1 *Assumptions 1 and 2 are satisfied by the lasso regression and the SCAD with $r_n = n^{-1/2}$ and $s_n = o(1)$ under the assumptions in Zhao and Yu (2006) or Fan and Li (2001), and by the adaptive lasso with $r_n = n^{-1}$ and $s_n = n^{-1/2}$ under the assumptions in Zou (2006).*

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 $\alpha_n \succ \epsilon_n$ is selection consistent. That is, as $n \to \infty$,*

$$P(\widehat{\mathcal{A}}_{\hat{\lambda}_n} = \mathcal{A}_T) \longrightarrow 1.$$

Theorem 1 claims the asymptotic selection consistency of the proposed kappa selection criterion when p is fixed, which indicates that, 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. It is worthy pointing out that 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.

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 p_0 may also diverge with n. To accommodate the diverging p_n scenario, the technical assumptions are modified as follows.

Assumption 1a: There exist r_n and s_n such that if $r_n \prec \lambda_n \prec s_n$ the base variable selection method is selection consistent in that

$$P(\widehat{\mathcal{A}}_{\lambda_n} = \mathcal{A}_T) \geq 1 - \epsilon_n,$$

where $\epsilon_n \prec p_n^{-1} c_0(p_n)$, and $c_0(p_n)$ is defined as in Assumption 2a.

10

Assumption 2a: For r_n in Assumption 1, if $\lambda_n \preceq r_n$, the base variable selection method is overfitted in that $P\big(\mathcal{A}_T \subseteq \widehat{\mathcal{A}}_{\lambda_n}\big) \to 1$ and for sufficiently large n,

$$P\Big(\mathcal{A}_T \cup \{j\} \subseteq \widehat{\mathcal{A}}_{\lambda_n}\Big) \geq c_0(p_n) > 0, \text{ for any } j \in \mathcal{A}_T^c, \tag{5}$$

where $c_0(p_n)$ is allowed to converge to 0 as p_n diverges.

Compared with the previous assumptions in Section 4.1, Assumption 1a is slightly stronger than Assumption 1 in that it requires the base variable selection method to be selection consistent at a rate faster than $p_n^{-1}c_0(p_n)$, and Assumption 2a is weaker than Assumption 2 as $c_0(p_n)$ is allowed to converge to 0.

Theorem 2 *Under Assumptions 1a and 2a, any variable selection method in (1) with $\hat{\lambda}_n$ as selected in Algorithm 1 with $\alpha_n \to 0$ and $\epsilon_n/\alpha_n \prec p_n^{-1}c_0(p_n)$ is selection consistent.*

Theorem 2 shows the asymptotic selection consistency of the proposed kappa selection criterion with diverging p_n, where the diverging speed of p_n is bounded as in $p_n^{-1}c_0(p_n) \succ \epsilon_n$ and depends on the base variable selection method. For example, the exchangeability assumption in Meinshausen and Buhlmann (2010) implies Assumption 2a with $c_0(p_n) \geq p_n^{-1}$, and thus $p_n^{-1} \succ \epsilon_n^{1/2}$ is sufficient for Assumption 1a. In addition, Zhao and Yu (2006) showed that Assumption 1a is satisfied by the lasso regression with $r_n = n^{k/2}p_n^{1/2}$, $s_n = n^{(1-g_1+g_2)/2}$ and $\epsilon_n = O(p_n n^k \lambda_n^{-2k})$, where the error term is assumed to have finite $2k$-th moment and $p_n = o(n^{(g_2-g_1)k})$ with $0 \leq g_1 < g_2 \leq 1$. However, it is relatively difficult to verify Assumption 1a for other variable selection methods with diverging p_n as their convergence rate ϵ_n's are not explicitly specified (Fan and Peng, 2004; Huang et al., 2008).

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,

$$C_p(\lambda) \;=\; \frac{SSE}{n\hat{\sigma}^2} \;+\; \frac{2\widehat{df}}{n}, \tag{6}$$

$$BIC(\lambda) \;=\; \frac{SSE}{n\hat{\sigma}^2} \;+\; \frac{\log(n)\widehat{df}}{n}, \tag{7}$$

$$CV(\lambda) \;=\; \sum_{s=1}^{10} \sum_{(y_k,x_k)\in T^{-s}} \left(y_k - \mathbf{x}_k^T \hat{\beta}^{(s)}(\lambda)\right)^2, \tag{8}$$

$$GCV(\lambda) \;=\; \frac{SSE}{n(1 - \widehat{df}/n)^2}, \tag{9}$$

where $SSE = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2$, $\hat{\sigma}^2$ is an estimated σ^2 based on the saturated model, and $\widehat{df}$ is estimated as the number of nonzero variables in $\hat{\beta}(\lambda)$ (Zou et al., 2007). In (8), 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 estimate $\hat{\beta}(\hat{\lambda})$ 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 $RPE = E\left(\mathbf{x}^T\hat{\beta}(\hat{\lambda}) - \mathbf{x}^T\beta\right)^2/\sigma^2$ (Zou, 2006).

5.1 Scenario I: fixed p

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

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

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\mathbf{x}_{(j)}$ 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 grid points $\{10^{-2+4l/99}; \ l = 0, \ldots, 99\}$. The number of splittings for the kappa selection criterion is $B = 20$. Each simulation is replicated 100 times, and the percentage of selecting the true active set, the averaged number of correctly selected zeros (C) and incorrectly selected zeros (I), and the relative prediction error (RPE) are summarized in Tables 1-2 and Figure 1.

Tables 1-2 and Figure 1 about here

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 (Ks) 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 confirms the selection consistency in Section 4.

13

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 for the lasso regression. 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.

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 percentage of selecting the truly informative variables, and the sparsity is defined as the percentage of excluding the truly uninformative variables. In Figure 2, it is clear that there is a positive relevance 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. In specific, we apply the kappa selection criterion on the lasso regression for $\alpha_n = \{\frac{l}{100}; \ l = 0, \ldots, 30\}$ and compute the corresponding averaged RPE over 100 replications. As shown in the last panel of Figure 2, the averaged RPEs are almost the same for $\alpha_n \in (0, 0.13)$, which confirms the theoretical results in Section 4.

Figure 2 about here

14

5.2   Scenario II: diverging p_n

Next we compare all the selection criteria in the scenario with diverging p_n with a similar simulation model as in Scenario I, except that $\beta = (5, 4, 3, 2, 1, 0, \cdots, 0)^T$ and $p_n = [\sqrt{n}]$. More specifically, four cases are examined: $n = 100$, $p_n = 10$; $n = 200$, $p_n = 14$; $n = 400$, $p_n = 20$; and $n = 800$, $p_n = 28$. A similar simulation example is also studied in Tibshirani (1996). The percentage of selecting the true active set, the averaged number of correctly selected zeros (C) and incorrectly selected zeros (I), and the relative prediction error (RPE) are summarized in Tables 3-4 and Figure 3.

Tables 3-4 and Figure 3 about here

The proposed kappa selection criterion still 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 the adaptive lasso and the SCAD, and for the lasso regression with $n \geq 400$. Furthermore, as shown in Figure 3, the kappa selection criterion yields the smallest relative prediction error across all cases.

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 dataset 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 (*pgg45*).

The dataset 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 $\{10^{-2+4l/99};\ l = 0, \ldots, 99\}$. 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 relative prediction errors (RPE) on the test data in Table 5.

<div align="center">Table 5 about here</div>

As shown in Table 5, the proposed kappa selection criterion yields the sparsest model and achieves the smallest relative prediction errors for the lasso regression and the SCAD, while the relative prediction 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   Extended selection criterion

In this section, we present a direct extension by combining the kappa selection criterion and the conventional cross-validation, which does not require the pre-specified thresholding value α_n in Algorithm 1.

To compute the cross-validation error, for $Z_1^* = \{(y_1^*, x_1^*), \cdots, (y_m^*, x_m^*)\}$ and $Z_2^* =$

$\{(y_{m+1}^*, x_{m+1}^*), \cdots, (y_n^*, x_n^*)\}$, we define

$$CV(Z_1^*, Z_2^*; \lambda) = n^{-1} \left(\sum_{i=1}^{m} (y_i^* - x_i^{*\prime} \widehat{\beta}_{2\lambda})^2 + \sum_{i=m+1}^{n} (y_i^* - x_i^{*\prime} \widehat{\beta}_{1\lambda})^2 \right), \tag{10}$$

where $\widehat{\beta}_{1\lambda}$ and $\widehat{\beta}_{2\lambda}$ are obtained based on Z_1^* and Z_2^*, respectively. The details of the extended selection criterion proceed as follows.

Algorithm 2 (extended selection criterion):

Steps 1-2. The same as those in Algorithm 1.

Step 3. Calculate $CV(Z_1^{*b}, Z_2^{*b}; \lambda)$ as in (10).

Step 4. Repeat Steps 1-3 for B times and obtain the following ratio,

$$\widehat{es}(\lambda) = \sum_{b=1}^{B} \kappa(\widehat{\mathcal{A}}_{1\lambda}^{*b}, \widehat{\mathcal{A}}_{2\lambda}^{*b}) \Big/ \sum_{b=1}^{B} CV(Z_1^{*b}, Z_2^{*b}; \lambda). \tag{11}$$

Step 5. Compute $\widehat{es}(\lambda)$ for a sequence of λ's and select $\widehat{\lambda} = \arg\max_\lambda \widehat{es}(\lambda)$.

The criterion (11) does not require the thresholding value α_n since it will get small when λ deviates from the true value. In specific, small λ leads to small variable selection stability as discussed in Section 3, whereas large λ over-penalizes the model and may exclude some truly informative variables, and thus leads to large cross-validation error. To demonstrate the effectiveness of the extended selection criterion, we repeat the simulated example Scenario I for $n = 40$ on the lasso regression. The percentage of selecting the true active set, the averaged number of correctly selected zeros (C) and incorrectly selected zeros (I), and the averaged RPE are summarized in Table 6. Figure 4 reports the results of detection and sparsity for various λ's as well as the extended selection criterion in (11) on the same sample.

Table 6 and Figure 4 about here

As expected the extended selection criterion is more conservative in variable selection

17

than the kappa selection criterion because of the influence of cross-validation. It performs slightly worse than the kappa selection criterion, but much better than other criterions.

8   Discussion

This article proposes a novel 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 simulated examples and real applications. 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).

Appendix: technical proofs

Proof of Lemma 1: We prove Lemma 1 for (1) the lasso regression, (2) the adaptive lasso, and (3) the SCAD, respectively.

(1): The lasso regression. The proof follows immediately after some existing results in literature. When $n^{1/2}\lambda_n \to \infty$ and $\lambda_n \to 0$, Assumption 1 is satisfied by the lasso regression under the irrepresentable condition following Zhao and Yu (2006) and Yuan and Lin (2006), and Assumption 2 is satisfied by the lasso regression following Zou (2006) and Bach (2008).

(2): The adaptive lasso. First, Zou (2006) showed that the adaptive lasso is selection consistent when $n\lambda_n \to \infty$ and $\sqrt{n}\lambda_n \to 0$, so Assumption 1 is satisfied.

To verify Assumption 2, we denote β^* as the true coefficient, $\beta = \beta^* + \frac{u}{\sqrt{n}}$, and

$$\Psi_n(u) = \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_{(j)} \left(\beta_j^* + \frac{u_j}{\sqrt{n}} \right) \right\|^2 + n\lambda_n \sum_{j=1}^{p} \frac{\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right|}{|\hat{\beta}_j^L|}.$$

where $\hat{\beta}_j^L$ is the estimator from the lasso regression. Let $\hat{u}_n = \arg\min \Psi_n(u)$, $\hat{\beta}_n = \beta^* + \frac{\hat{u}_n}{\sqrt{n}}$, and $V_n(u) = \Psi_n(u) - \Psi_n(0)$ with

$$V_n(u) = u^T \left(\frac{\mathbf{X}^T \mathbf{X}}{n} \right) u - \frac{2\epsilon^T \mathbf{X}}{\sqrt{n}} u + \sqrt{n}\lambda_n \sum_{j=1}^{p} \frac{\sqrt{n}\left(\left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| - |\beta_j^*| \right)}{|\hat{\beta}_j^L|}.$$

Note that $\frac{\mathbf{X}^T \mathbf{X}}{n} \to \mathbf{C}$, $\frac{\epsilon^T \mathbf{X}}{\sqrt{n}} \xrightarrow{d} W^T \sim N(0, \Sigma\mathbf{C})$, and $n\lambda_n \to a$ with $0 \leq a < \infty$ implies $\sqrt{n}\lambda_n \to 0$. Following similar treatment as in Zou (2006), $\sqrt{n}\lambda_n \sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|)/|\hat{\beta}_j^L| \xrightarrow{p} 0$ when $\beta_j^* \neq 0$, and $\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) = |u_j|$ when $\beta_j^* = 0$.

If $a = 0$, the asymptotic normality of $\hat{\beta}_j^L$ implies that $\frac{n\lambda_n}{|\sqrt{n}\hat{\beta}_j^L|} \xrightarrow{p} 0$ when $\beta_j^* = 0$, and then it follows from the Slutsky's theorem that

$$V_n(u) \xrightarrow{d} u^T \mathbf{C} u - 2W^T u.$$

Therefore, $\hat{u}_n \xrightarrow{d} \mathbf{C}^{-1} W$, which implies that $P(j \in \widehat{\mathcal{A}}_{\lambda_n}) \to 1$ for all $j \in \{1, \cdots, p\}$, and thus Assumption 2 is satisfied.

If $0 < a < \infty$, the asymptotic normality of $\hat{\beta}_n$ still holds, which implies that $P(\mathcal{A}_T \subseteq \widehat{\mathcal{A}}_{\lambda_n}) \to 1$. It then suffices to consider the event $j \notin \widehat{\mathcal{A}}_{\lambda_n}$ for any $j \in \mathcal{A}_T^c$. Note that when $j \notin \widehat{\mathcal{A}}_{\lambda_n}$, the Karush-Kuhn-Tucker (KKT) conditions imply that

$$\left| 2\mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\hat{\beta}_n) \right| \leq n \frac{\lambda_n}{|\hat{\beta}_j^L|}.$$

In addition,

$$\frac{2\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\hat{\beta}_n)}{\sqrt{n}} = \frac{\mathbf{x}_{(j)}^T\mathbf{X}\sqrt{n}(\beta^* - \hat{\beta}_n)}{n} + \frac{2\mathbf{x}_{(j)}^T\epsilon}{\sqrt{n}},$$

By the asymptotic normality of $\hat{\beta}_n$ and $\frac{\mathbf{X}^T\mathbf{X}}{n} \to \mathbf{C}$, the Slutsky's theorem implies that $2\mathbf{x}_{(j)}^T\mathbf{X}\sqrt{n}(\beta^* - \hat{\beta}_n)/n \xrightarrow{d} N(0, \Delta_1)$ for some Δ_1, and $2\mathbf{x}_{(j)}^T\epsilon/\sqrt{n} \xrightarrow{d} N(0, 4\|\mathbf{x}_{(j)}\|^2\Sigma_{jj}^2)$. Therefore, as $n\lambda_n \to a$ with $0 < a < \infty$,

$$\begin{aligned} P\Big(j \notin \widehat{\mathcal{A}}_{\lambda_n}\Big) &\leq P\Big(\Big|2\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\hat{\beta}_n)\Big| \leq n\frac{\lambda_n}{|\hat{\beta}_j^L|}\Big) \\ &= P\Big(\Big|\frac{2\mathbf{x}_{(j)}^T\mathbf{X}\sqrt{n}(\beta^* - \hat{\beta}_n)}{n} + \frac{2\mathbf{x}_{(j)}^T\epsilon}{\sqrt{n}}\Big||\sqrt{n}\hat{\beta}_j^L| \leq n\lambda_n\Big) \leq 1 - c_1, \end{aligned}$$

for some constant c_1. Therefore, Assumption 2 is satisfied with $c_0 < c_1$.

(3): The SCAD. First, Fan and Li (2001) showed that the SCAD is selection consistent when $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$, so Assumption 1 is satisfied.

Next, we show that the SCAD will be overfitted when $\sqrt{n}\lambda_n \to a$ with $0 \leq a < \infty$. By Theorem 1 of Fan and Li (2001), $\hat{\beta}_n$ is a $\sqrt{n}$-consistent estimate of β^* when $\lambda_n \to 0$, and hence that $P(\mathcal{A}_T \subseteq \widehat{\mathcal{A}}_{\lambda_n}) \to 1$. It then suffices to consider the event $j \notin \widehat{\mathcal{A}}_{\lambda_n}$ for any $j \in \mathcal{A}_T^c$. In fact, the SCAD minimizes

$$Q(\beta) = \Big\|\mathbf{y} - \sum_{j=1}^p \mathbf{x}_{(j)}\beta_j\Big\|^2 + n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|), \tag{12}$$

where the penalty term satisfies $p'_\lambda(\theta) = \lambda\big(I(\theta \leq \lambda) + \frac{(\gamma\lambda - \theta)_+}{(\gamma - 1)\lambda}I(\theta > \lambda)\big)$ for some $\gamma > 2$ and $\theta > 0$. For any $\beta \in \{\beta : \|\sqrt{n}(\hat{\beta}_n - \beta)\| \leq c_2\}$, then

$$\begin{aligned} \frac{\partial Q(\beta)}{\partial \beta_j} &= -2\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\beta) + np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) \\ &= -n\lambda_n\left(\frac{\frac{2\mathbf{x}_{(j)}^T\mathbf{X}\sqrt{n}(\beta^* - \beta)}{n} + \frac{2\mathbf{x}_{(j)}^T\epsilon}{\sqrt{n}}}{\sqrt{n}\lambda_n} - \frac{p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)}{\lambda_n}\right), \end{aligned}$$

20

where $\mathbf{X}^T\mathbf{X}/n \to \mathbf{C}$, $\|\sqrt{n}(\beta^*-\beta)\| \leq \|\sqrt{n}(\beta^*-\hat{\beta}_n)\|+\|\sqrt{n}(\hat{\beta}_n-\beta)\|$ is bounded in probability, and $2\mathbf{x}_{(j)}^T\epsilon/\sqrt{n} \overset{d}{\to} N(0, 4\|\mathbf{x}_{(j)}\|^2\Sigma_{jj}^2)$. In addition, $p'_{\lambda_n}(|\beta_j|)/\lambda_n = I(\theta \leq \lambda_n) + \frac{(\gamma\lambda_n-\theta)_+}{(\gamma-1)\lambda_n}I(\theta > \lambda_n) \leq 1$. Therefore, as $\sqrt{n}\lambda_n \to a$ with $0 \leq a < \infty$,

$$P\left(\left|\frac{\frac{2\mathbf{x}_{(j)}^T\mathbf{X}\sqrt{n}(\beta^*-\beta)}{n} + \frac{2\mathbf{x}_{(j)}^T\epsilon}{\sqrt{n}}}{\sqrt{n}\lambda_n}\right| > \left|\frac{p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)}{\lambda_n}\right|\right)$$

$$= P\left(\left|\frac{2\mathbf{x}_{(j)}^T\mathbf{X}\sqrt{n}(\beta^*-\beta)}{n} + \frac{2\mathbf{x}_{(j)}^T\epsilon}{\sqrt{n}}\right| > \sqrt{n}\lambda_n\frac{p'_{\lambda_n}(|\beta_j|)}{\lambda_n}\right) \to \begin{cases} c_2, & \text{if } a > 0 \\ 1, & \text{if } a = 0, \end{cases}$$

for some constant $c_2 > 0$. Therefore, if $a > 0$, there exists a constant $c_0 \geq 0$ such that with a positive probability c_0,

$$\frac{\partial Q(\beta)}{\partial \beta_j} < 0 \text{ when } 0 < \beta_j < Mn^{-1/2}; \tag{13}$$

$$\frac{\partial Q(\beta)}{\partial \beta_j} > 0 \text{ when } -Mn^{-1/2} < \beta_j < 0, \tag{14}$$

with M sufficient large such that $P\left(\sup_{\|u\|=M} Q(\beta^* + (n^{-1/2} + a_n)u) > Q(\beta^*)\right) \to 1$ and $a_n = \max\{p'_{\lambda_n}(|\beta_j^*|) : \beta_j^* \neq 0\}$, which implies that $P(\hat{\beta}_j \neq 0) \geq c_0$ for sufficiently large n. If $a = 0$, with probability tending to 1,

$$\frac{\partial Q(\beta)}{\partial \beta_j} < 0 \text{ when } -Mn^{-1/2} < \beta_j < Mn^{-1/2}, \tag{15}$$

and hence $P(\hat{\beta}_j \neq 0) \to 1$. Therefore, Assumptions 2 is satisfied by the SCAD with $r_n = n^{-1/2}$ and $s_n = o(1)$. This ends the proof of Lemma 1. ∎

Additional Notations: Note that any variable selection method is trivially stable if it always selects the complete variable set or the empty variable set, however it violates the assumption that the true active set is neither the complete set nor the empty set. In Algorithm 1, the variable selection stabilities of such trivial methods are set as -1 and thus will

never be selected. Therefore, it suffices to focus on the set of λ's that lead to non-degenerate variable selection methods. Specifically, for some fixed constant $\delta > 0$, define

$$\Lambda_n = \left\{ \lambda : P(\widehat{\mathcal{A}}_\lambda \neq \emptyset) \geq \delta \text{ and } P(\widehat{\mathcal{A}}_\lambda \neq \{1, \ldots, p\}) \geq \delta \right\},$$

and a set of λ's that lead to non-degenerate stable variable selection methods as

$$\widetilde{\Lambda}_n = \left\{ \lambda \in \Lambda_n : P\big(\hat{s}(\Psi, \lambda, m) \geq 1 - \eta_n\big) \geq 1 - \xi_n \text{ for some } \eta_n \to 0 \text{ and } \xi_n \to 0 \right\}, \quad (16)$$

where $m = \lfloor \frac{n}{2} \rfloor$ and the probability P is taken with respect to the training sample.

Lemma 2 *For λ_n defined as in Assumption 1, the resultant variable selection method is selection consistent in that $P(\widehat{\mathcal{A}}_{\lambda_n} = \mathcal{A}_T) \geq 1 - \epsilon_n$ for some $\epsilon_n \to 0$, then for any $\eta_n \succ \epsilon_n$,*

$$P\big(\hat{s}(\Psi, \lambda_n, m) \geq 1 - \eta_n\big) \geq 1 - 2\epsilon_n/\eta_n,$$

and hence that $\lambda_n \in \widetilde{\Lambda}_n$.

Proof of Lemma 2: For clarity, we denote λ_n satisfying Assumption 1 as λ_n^*, and then the selection consistency implies that $P(\widehat{\mathcal{A}}_{\lambda_n^*} = \mathcal{A}_T) \geq 1 - \epsilon_n$ for some $\epsilon_n \to 0$. We further 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

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

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

$$E\big(\hat{s}(\Psi, \lambda_n^*, m)\big) = E\Big(B^{-1} \sum_{b=1}^B \hat{s}^{*b}(\Psi, \lambda_n^*, m)\Big) = E\big(\hat{s}^{*b}(\Psi, \lambda_n^*, m)\big) \geq 1 - 2\epsilon_n.$$

In addition, since $0 \leq \hat{s}(\Psi, \lambda_n^*, n) \leq 1$, and the Markov inequality yields that

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

which implies the desired result immediately. ∎

Lemma 2 shows that if a variable selection method is selection consistent, its variable selection stability converges to 1 in probability. It also assures that there always exists λ_n such that the resultant variable selection method is stable and non-degenerate.

Proof of Theorem 1: Let $r_n \prec \lambda_n^* \prec s_n$, Assumption 1 implies that $P(\widehat{\mathcal{A}}_{\lambda_n^*} = \mathcal{A}_T) \geq 1 - \epsilon_n$ for some $\epsilon_n \to 0$, and Lemma 2 implies that $\lambda_n^* \in \widetilde{\Lambda}_n$. Denote $\tilde{\lambda}_n = \widetilde{\min}_\lambda\{\lambda : \lambda \in \widetilde{\Lambda}_n\}$ with $\widetilde{\min}$ representing minimization up to a constant, and hence that $\tilde{\lambda}_n \preceq \lambda_n^*$. Then we prove Theorem 1 in two steps. Step 1 shows that the variable selection method with $\tilde{\lambda}_n$ is selection consistent, and step 2 assures that $P(\tilde{\lambda}_n \sim \hat{\lambda}_n) \to 1$ with $\hat{\lambda}_n$ being defined as in Algorithm 1. The desired result follows immediately after these two steps.

Step 1 is proved by contradiction. If the variable selection method with $\tilde{\lambda}_n$ is not selection consistent, then by Assumption 1 we have $\tilde{\lambda}_n \not\succ r_n$ or $\tilde{\lambda}_n \not\prec s_n$. Without loss of generality, we assume that the limits of $r_n^{-1}\tilde{\lambda}_n$ and $s_n^{-1}\tilde{\lambda}_n$ exist (where the limit of $s_n^{-1}\tilde{\lambda}_n$ can be infinity), since otherwise we can focus on the corresponding convergent subsequences $r_{n_m}^{-1}\tilde{\lambda}_{n_m}$ and $s_{n_m}^{-1}\tilde{\lambda}_{n_m}$. Then $\tilde{\lambda}_n \not\succ r_n$ implies that (1) $r_n^{-1}\tilde{\lambda}_n \to a \geq 0$, and $\tilde{\lambda}_n \not\prec s_n$ implies that (2) $s_n^{-1}\tilde{\lambda}_n \to b > 0$, where b can be infinity. We now show that both (1) and (2) will lead to contradictions.

If case (2) occurs, $\tilde{\lambda}_n \succeq s_n \succ \lambda_n^*$, which contradicts with the fact that $\tilde{\lambda}_n \preceq \lambda_n^*$.

If case (1) occurs, by Assumption 2, there exists a constant $c_0 > 0$ such that for any $j \in \mathcal{A}_T^c$, $P\left(\mathcal{A}_T \cup \{j\} \subseteq \widehat{\mathcal{A}}_{\tilde{\lambda}_n}\right) \geq c_0$ for sufficiently large n. In addition, there also exists $j_1 \in \mathcal{A}_T^c$ such that $P\left(j_1 \notin \widehat{\mathcal{A}}_{\tilde{\lambda}_n}\right) \geq c_3 > 0$ when n is sufficiently large, since otherwise

$P\big(\widehat{\mathcal{A}}_{\tilde{\lambda}_n} = \{1,\ldots,p\}\big) \to 1$ which contradicts with the fact that $\tilde{\lambda}_n \in \Lambda_n$. Therefore,

$$P\Big(\widehat{\mathcal{A}}^{*b}_{1\tilde{\lambda}_n} \neq \widehat{\mathcal{A}}^{*b}_{2\tilde{\lambda}_n}\Big) \geq P\Big(j_1 \notin \widehat{\mathcal{A}}^{*b}_{1\tilde{\lambda}_n}, j_1 \in \widehat{\mathcal{A}}^{*b}_{2\tilde{\lambda}_n}\Big) \geq c_0 c_3,$$

for sufficiently large n, where the last inequality follows from the fact that the two sub-samples are independent.

Since $\widehat{\mathcal{A}}^{*b}_{1\tilde{\lambda}_n} \neq \widehat{\mathcal{A}}^{*b}_{2\tilde{\lambda}_n}$ implies that $\hat{s}^{*b}(\Psi, \tilde{\lambda}_n, m) \leq c_4$ with $c_4 = \max_{\mathcal{A}_1 \neq \mathcal{A}_2} \kappa(\mathcal{A}_1, \mathcal{A}_2) \leq \frac{p-1}{p}$ where $\mathcal{A}_1, \mathcal{A}_2 \subset \{1, \cdots, p\}$, we have for sufficiently large n,

$$P\Big(\hat{s}^{*b}(\Psi, \tilde{\lambda}_n, m) \leq c_4\Big) \geq c_0 c_3.$$

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

$$E\Big(\hat{s}(\Psi, \tilde{\lambda}_n, m)\Big) = E\Big(B^{-1} \sum_{b=1}^{B} \hat{s}^{*b}(\Psi, \tilde{\lambda}_n, m)\Big) = E\Big(\hat{s}^{*1}(\Psi, \tilde{\lambda}_n, m)\Big) \leq 1 - c_0 c_3(1 - c_4),$$

which is a constant strictly less than 1. By the Markov inequality, for any $\eta_n \to 0$,

$$P\Big(\hat{s}(\Psi, \tilde{\lambda}_n, m) \geq 1 - \eta_n\Big) \leq \frac{E\Big(\hat{s}(\Psi, \tilde{\lambda}_n, m)\Big)}{1 - \eta_n} \leq \frac{1 - c_0 c_3(1 - c_4)}{1 - \eta_n} \to 1 - c_0 c_3(1 - c_4). \quad (17)$$

This contradicts with the fact that $P\big(\hat{s}(\Psi, \tilde{\lambda}_n, m) \geq 1 - \eta_n\big) \geq 1 - \xi_n$ for some $\xi_n \to 0$. This ends the proof of step 1.

Next we show that $P(\tilde{\lambda}_n \sim \hat{\lambda}_n) \to 1$. On one hand, setting $\alpha_n \to 0$ and $\alpha_n \succ \epsilon_n$ in Algorithm 1 yields that

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

Then by Lemma 2 and the fact that $\alpha_n \succ \epsilon_n$, $P\Big(\hat{s}(\Psi, \lambda_n^*, m) \geq 1 - \alpha_n\Big) \geq 1 - \frac{2\epsilon_n}{\alpha_n}$. Therefore,

$$P\Big(\hat{s}(\Psi, \hat{\lambda}_n, m) \geq (1 - \alpha_n)(1 - \alpha_n)\Big) \geq 1 - \frac{2\epsilon_n}{\alpha_n}.$$

Since $(1 - \alpha_n)^2 \geq 1 - 2\alpha_n$, $\alpha_n \to 0$ and $\epsilon_n/\alpha_n \to 0$, we have $P(\hat{\lambda}_n \in \widetilde{\Lambda}_n) \to 1$, which implies that $P(\hat{\lambda}_n \succeq \tilde{\lambda}_n) \to 1$.

On the other hand, since $\tilde{\lambda}_n = \widetilde{\min}_\lambda\{\lambda : \lambda \in \widetilde{\Lambda}_n\}$ and $\alpha_n \to 0$,

$$
\begin{aligned}
P\Big(\frac{\hat{s}(\Psi, \tilde{\lambda}_n, m)}{\max_\lambda \hat{s}(\Psi, \lambda, m)} \geq 1 - \alpha_n\Big) &= P\Big(\hat{s}(\Psi, \tilde{\lambda}_n, m) \geq (1 - \alpha_n) \max_\lambda \hat{s}(\Psi, \lambda, m)\Big) \\
&\geq P\Big(\hat{s}(\Psi, \tilde{\lambda}_n, m) \geq 1 - \alpha_n\Big) \to 1,
\end{aligned}
$$

and hence that $P(\tilde{\lambda}_n \in \widehat{\Lambda}_n) \to 1$, which implies that $P(\tilde{\lambda}_n \succeq \hat{\lambda}_n) \to 1$. Therefore, step 2 is proved, and Theorem 1 follows immediately after steps 1 and 2. ∎

Proof of Theorem 2: In the diverging p_n case, we denote the set of λ's that lead to non-degenerate stable variable selection methods as

$$\widetilde{\Lambda}_{p_n} = \Big\{\lambda \in \Lambda_n : P\big(\hat{s}(\Psi, \lambda, m) \geq 1 - \eta_n\big) \geq 1 - \xi_n \text{ for some } \eta_n \to 0 \text{ and } \epsilon_n \prec \xi_n \prec p_n^{-1} c_0(p_n)\Big\},$$

where $\widetilde{\Lambda}_{p_n}$ depends on the dimension p_n. We further denote $\tilde{\lambda}_{p_n} = \widetilde{\min}_\lambda\{\lambda : \lambda \in \widetilde{\Lambda}_{p_n}\}$ with $\widetilde{\min}$ representing minimization up to a constant.

First, since $\epsilon_n \prec p_n^{-1} c_0(p_n)$, it implies that there always exists $\eta_n \to 0$ such that $\epsilon_n \prec \epsilon_n/\eta_n \prec p_n^{-1} c_0(p_n)$, and thus $\lambda_n^* \in \widetilde{\Lambda}_{p_n}$ by Lemma 2. Next, we prove Theorem 2 in the same two steps as in the proof of Theorem 1. Step 1 shows that the variable selection method with $\tilde{\lambda}_{p_n}$ is selection consistent, and step 2 assures that $P(\tilde{\lambda}_{p_n} \sim \hat{\lambda}_n) \to 1$ with $\hat{\lambda}_n$ being defined as in Algorithm 1.

Both steps can be shown similarly as in the proof of Theorem 1 after some slight modification. In fact, Step 1 can be showed by deriving similar contradictions, except that in (17),

25

since $c_4 \leq \frac{p_n - 1}{p_n}$,

$$P\Big(\hat{s}(\Psi, \tilde{\lambda}_{p_n}, m) \geq 1 - \eta_n\Big) \leq \frac{1 - c_0(p_n)c_3(1 - c_4)}{1 - \eta_n} \leq \frac{1 - p_n^{-1}c_0(p_n)c_3}{1 - \eta_n},$$

which still leads to contradiction with the fact that $\tilde{\lambda}_{p_n} \in \widetilde{\Lambda}_{p_n}$. Step 2 can be shown similarly by setting $\alpha_n \to 0$ and $\epsilon_n/\alpha_n \prec p_n^{-1}c_0(p_n)$ in Algorithm 1, which yields that $P\Big(\hat{s}(\Psi, \hat{\lambda}_n, m) \geq (1 - \alpha_n)(1 - \alpha_n)\Big) \geq 1 - \frac{2\epsilon_n}{\alpha_n}$, and

$$P\Big(\frac{\hat{s}(\Psi, \tilde{\lambda}_n, m)}{\max_\lambda \hat{s}(\Psi, \lambda, m)} \geq 1 - \alpha_n\Big) \geq P\Big(\hat{s}(\Psi, \tilde{\lambda}_n, m) \geq 1 - \alpha_n\Big) \to 1.$$

This ends the proof of Theorem 2. ∎

References

[1] AKAIKE, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

[2] BACH, F.R. (2008). Bolasso: Model Consistent Lasso Estimation Through the Bootstrap. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[3] BEN-HUR, A., ELISSEEFF, A. AND GUYON, I. (2002). A Stability Based Method for Discovering Structure in Clustered Data. *Pacific Symposium on Biocomputing*, 6-17.

[4] BREHENY, P. AND HUANG, J. (2011). Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *Annals of Applied Statistics*, **5**, 232-253.

[5] COHEN, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**, 37-46.

[6] CRAVEN, P. AND WAHBA, G. (1979). Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik*, **31**, 317-403.

[7] EFRON, B., HASTIE, T., JOHNSTONE, I. AND TIBSHIRANI, R. (2004). Least Angle Regression. *Annals of Statistics*, **32**, 407-451.

[8] 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.

[9] FAN, J. AND LI, R. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. In *Proceedings of the International Congress of Mathematicians*, **3**, 595-622.

[10] FAN, J. AND PENG, H. (2004). Nonconcave Penalized Likelihood with A Diverging Number of Parameters. *Annals of Statistics*, **32**, 928-961.

[11] HUANG, J., MA, S. AND ZHANG, C. H. (2008). Adaptive Lasso for Sparse High-dimensional Regression Models. *Statistica Sinica*, **18**, 1603-1618.

[12] MALLOWS, C. (1973). Some Comments on Cp. *Technometrics*, **15**, 661-675.

[13] MEINSHAUSEN, N. AND BUEHLMANN, P. (2010). Stability Selection. *Journal of the Royal Statistical Society, Series B*, **72**, 414-473.

[14] NISHII, R. (1984). Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression. *Annals of Statistics*, **12**, 758-765.

[15] SCHWARZ, G. E. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464.

[16] Shen, X., Pan, W., Zhu, Y. and Zhou, H. (2012). On L0 Regularization in High-dimensional Regression. *Journal of the American Statistical Association*, to appear.

[17] Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine, E.A. and Yang, N. (1989). Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate: II. Radical Prostatectomy Treated Patients. *Journal of Urology*, **141**, 1076-1083.

[18] Stone, M. (1974). Cross-validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, Series B*, **36**, 111-147.

[19] Sun, W., Wang, J. and Fang, Y. (2012). Regularized K-means Clustering of High-dimensional Data and Its Asymptotic Consistency, *Electronic Journal of Statistics*, **6**, 148-167.

[20] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.

[21] Wang, H. and Leng, C. (2008). A Note on Adaptive Group Lasso. *Computational Statistics and Data Analysis*, **52**, 5277-5286.

[22] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, **94**, 553-568.

[23] 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.

[24] Wang, J. (2010). Consistent Selection of the Number of Clusters via Cross Validation. *Biometrika*, **97**, 893-904.

[25] Xue, L., Qu, A. and Zhou, J. (2010). Consistent Model Selection for Marginal Generalized Additive Model for Correlated Data. *Journal of the American Statistical Association*, **105**, 1518-1530.

[26] YUAN, M. AND LIN, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49-67.

[27] 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.

[28] ZHAO, P. AND YU, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541-2563.

[29] ZOU, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

[30] 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.

[31] ZOU, H., HASTIE, T. AND TIBSHIRANI, R. (2007). On the "Degree of Freedom" of the Lasso. *Annals of Statistics*, **35**, 2173-2192.

[32] ZOU, H. AND ZHANG, H. (2009). On the Adaptive Elastic-net with A Diverging Number of Parameters. *Annals of Statistics*, **37**, 1733-1751.

Table 1: The percentages of selecting the true active set for various selection criteria in simulation 1.

n	Penalty	Ks	Cp	BIC	CV	GCV
	Lasso	**0.63**	0.16	0.29	0.09	0.16
40	Ada lasso	**0.98**	0.53	0.75	0.63	0.52
	SCAD	**0.98**	0.55	0.81	0.76	0.52
	Lasso	**0.81**	0.16	0.35	0.14	0.17
60	Ada lasso	**0.99**	0.52	0.87	0.65	0.52
	SCAD	**1**	0.58	0.88	0.76	0.56
	Lasso	**0.89**	0.16	0.38	0.09	0.16
80	Ada lasso	**0.99**	0.56	0.88	0.77	0.56
	SCAD	**0.99**	0.62	0.89	0.75	0.61

Table 2: The averaged numbers of correctly selected zeros (C) and incorrectly selected zeros (I) for various selection criteria in simulation 1.

		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	3.68	0	2.66	0	3.25	0
40	Ada lasso	**4.98**	0	4.16	0	4.59	0	4.25	0	4.15	0
	SCAD	**4.99**	0.01	4.11	0	4.63	0	4.39	0	4.06	0
	Lasso	**4.8**	0	3.12	0	4	0	2.85	0	3.13	0
60	Ada lasso	**4.99**	0	4.17	0	4.84	0	4.35	0	4.17	0
	SCAD	**5**	0	4.15	0	4.84	0	4.37	0	4.12	0
	Lasso	**4.88**	0	3.01	0	4.05	0	2.66	0	3	0
80	Ada lasso	**4.99**	0	4.19	0	4.84	0	4.49	0	4.19	0
	SCAD	**4.99**	0	4.23	0	4.83	0	4.45	0	4.22	0

Table 3: The percentages of selecting the true active set for various selection criteria in simulation 2.

n	p_n	Penalty	Ks	Cp	BIC	CV	GCV
		Lasso	**0.89**	0.11	0.22	0.09	0.11
100	10	Ada lasso	**1**	0.58	0.89	0.70	0.58
		SCAD	**1**	0.58	0.89	0.80	0.57
		Lasso	**0.96**	0.02	0.09	0	0.02
200	14	Ada lasso	**1**	0.41	0.93	0.80	0.42
		SCAD	**1**	0.43	0.91	0.77	0.43
		Lasso	**1**	0.04	0.07	0.01	0.04
400	20	Ada lasso	**1**	0.3	0.87	0.72	0.29
		SCAD	**1**	0.37	0.88	0.72	0.37
		Lasso	**1**	0	0.03	0	0
800	28	Ada lasso	**1**	0.22	0.94	0.77	0.22
		SCAD	**1**	0.34	0.98	0.76	0.34

Table 4: The averaged numbers of correctly selected zeros (C) and incorrectly selected zeros (I) for various selection criteria in simulation 2.

			Ks	Ks	Cp	Cp	BIC	BIC	CV	CV	GCV	GCV
n	p_n	Penalty	C	I	C	I	C	I	C	I	C	I
		Lasso	**4.88**	0	2.80	0	3.37	0	2.60	0	2.80	0
100	10	Ada lasso	**5**	0	4.34	0	4.84	0	4.45	0	4.34	0
		SCAD	**5**	0	4.32	0	4.84	0	4.64	0	4.30	0
		Lasso	**8.96**	0	5.52	0	6.70	0	5.13	0	5.53	0
200	14	Ada lasso	**9**	0	7.71	0	8.92	0	8.37	0	7.73	0
		SCAD	**9**	0	7.59	0	8.89	0	8.37	0	7.58	0
		Lasso	**15**	0	9.52	0	11.78	0	9.24	0	9.52	0
400	20	Ada lasso	**15**	0	12.48	0	14.83	0	14.10	0	12.47	0
		SCAD	**15**	0	12.60	0	14.81	0	14.06	0	12.59	0
		Lasso	**23**	0	16.50	0	19.44	0	16.39	0	16.40	0
800	28	Ada lasso	**23**	0	19.95	0	22.94	0	22.54	0	19.95	0
		SCAD	**23**	0	19.59	0	22.98	0	22.28	0	19.59	0

Table 5: The selected active sets and the relative prediction errors (RPE) for various selection criteria in the prostate cancer example.

	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
RPE	Ada lasso	0.806	0.825	0.825	**0.797**	0.825
	SCAD	**0.734**	0.825	0.825	0.797	0.825

Table 6: The percentage of selecting the true active set, the averaged number of correctly selected zeros (C) and incorrectly selected zeros (I), and the relative prediction error (RPE) of Algorithm 2 (Extended) compared with that of Algorithm 1 (Ks).

Algorithms	Percentage	C	I	RPE (s.d.)
Ks	0.63	4.58	0.01	0.088 (0.021)
Extended	0.45	4.16	0	0.100 (0.012)

Figure 1: Relative prediction errors (RPE) for various selection criteria in simulation 1, where 'K', 'Cp', 'B', 'C' and 'G' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.
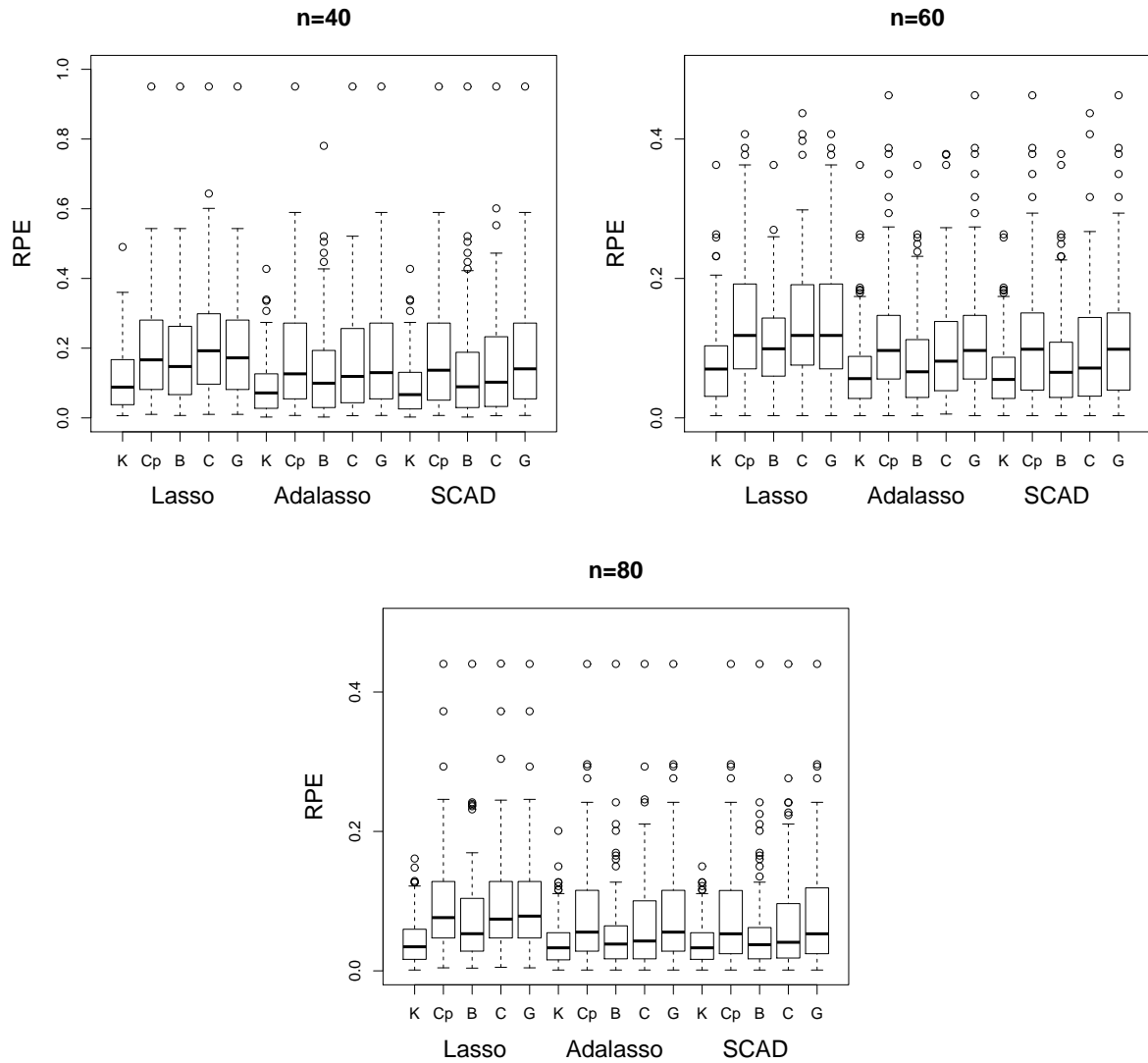
Figure 2: The detection and sparsity of the lasso regression with the kappa selection criterion in simulation 1 are shown on the top, and the sensitivity of α to the relative prediction error is shown on the bottom.
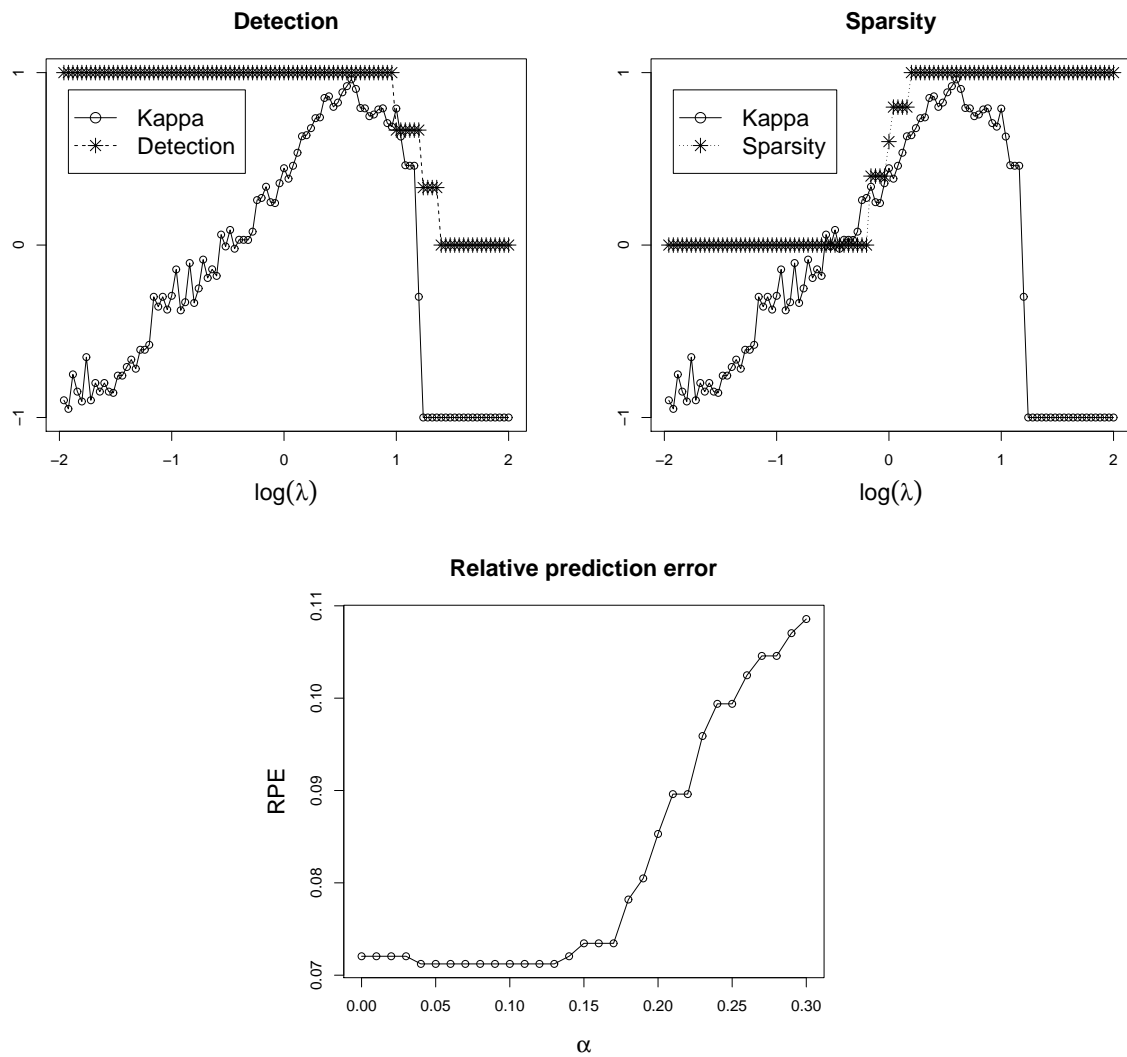
Figure 3: Relative prediction errors (RPE) for various selection criteria in simulation 2, where 'K', 'Cp', 'B', 'C' and 'G' represent the kappa selection criterion, Mallows' C_p, BIC, CV and GCV, respectively.
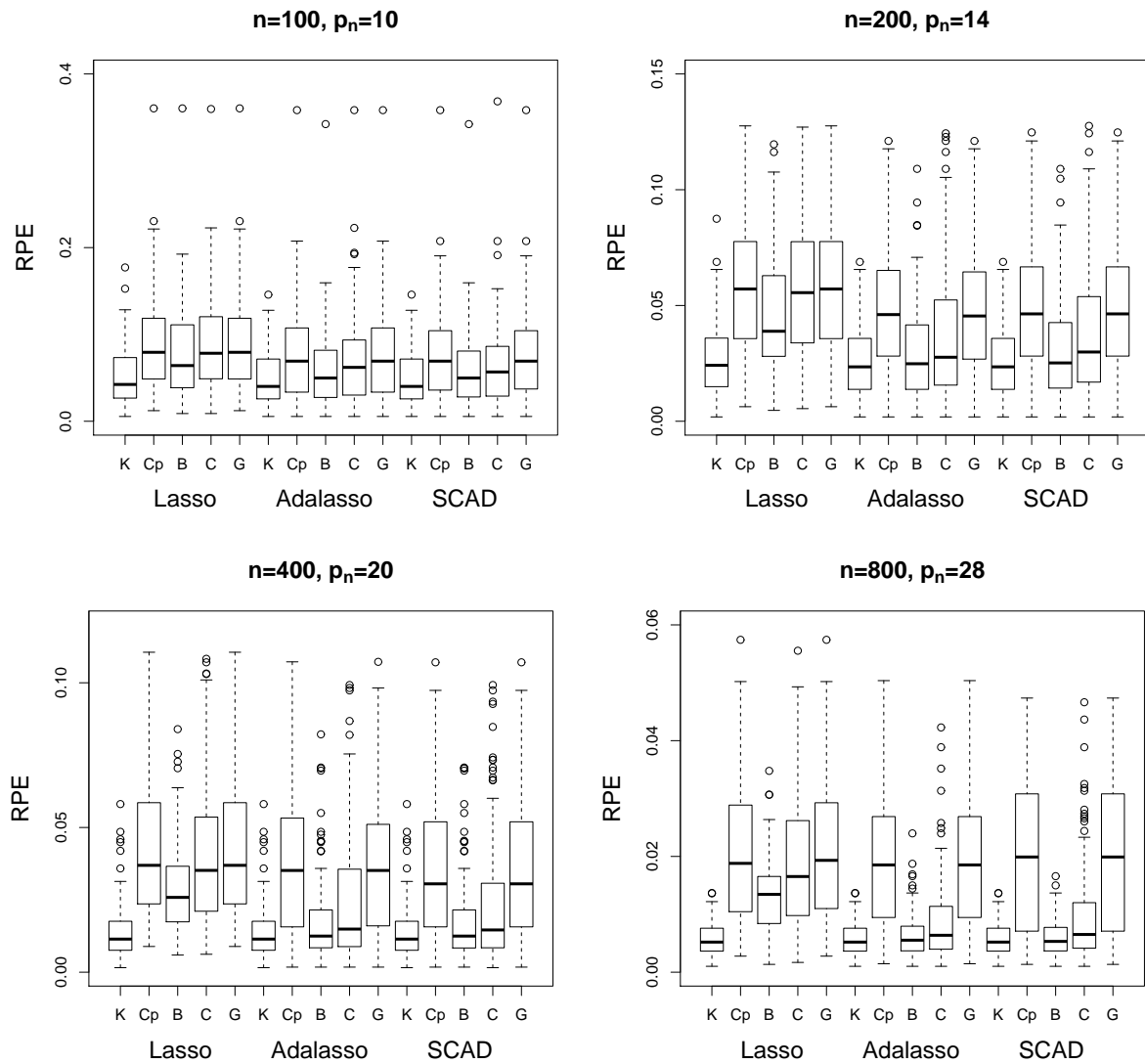
Figure 4: The detection and sparsity of the lasso regression with the extended selection criterion (denoted as Extended) in Algorithm 2 .