Please see the online version of this article for supplementary materials.

Regularization Parameter Selections via Generalized Information Criterion

Yiyun ZHANG, Runze LI, and Chih-Ling TSAI

We apply the nonconcave penalized likelihood approach to obtain variable selections as well as shrinkage estimators. This approach relies heavily on the choice of regularization parameter, which controls the model complexity. In this paper, we propose employing the generalized information criterion, encompassing the commonly used Akaike information criterion (AIC) and Bayesian information criterion (BIC), for selecting the regularization parameter. Our proposal makes a connection between the classical variable selection criteria and the regularization parameter selections for the nonconcave penalized likelihood approaches. We show that the BIC-type selector enables identification of the true model consistently, and the resulting estimator possesses the oracle property in the terminology of Fan and Li (2001). In contrast, however, the AIC-type selector tends to overfit with positive probability. We further show that the AIC-type selector is asymptotically loss efficient, while the BIC-type selector is not. Our simulation results confirm these theoretical findings, and an empirical example is presented. Some technical proofs are given in the online supplementary material.

KEY WORDS: Akaike information criterion; Bayesian information criterion; Least absolute shrinkage and selection operator; Nonconcave penalized likelihood; Smoothly clipped absolute deviation.

1. INTRODUCTION

Using the penalized likelihood function to simultaneously select variables and estimate unknown parameters has received considerable attention in recent years. To avoid the instability of the classical subset selection procedure, Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator (LASSO). In the same spirit as LASSO, Fan and Li (2001) introduced a unified approach via a nonconcave penalized likelihood function, and demonstrated its usefulness for both linear and generalized linear models. Subsequently, Fan and Li (2002) employed the nonconcave penalized likelihood approach for Cox models, while Fan and Li (2004) used it on semiparametric regression models. In addition, Fan and Peng (2004) investigated the theoretical properties of the nonconcave penalized likelihood when the number of parameters tends to infinity as the sample size increases. Recently, more researchers applied penalized approaches to study variable selections (e.g., Park and Hastie 2007; Wang and Leng 2007; Yuan and Lin 2007; Zhang and Lu 2007; Li and Liang 2008).

In employing the nonconcave penalized likelihood in regression analysis, we face two challenges. The first hurdle is to compute the nonconcave penalized likelihood estimate. This issue has been carefully studied in the recent literature: (i) Fan and Li (2001) proposed the local quadratic approximation (LQA) algorithm, which was further analyzed by Hunter and Li (2005); (ii) Efron et al. (2004) introduced the LARS algorithm, which can be used for the adaptive LASSO (Zou 2006; Wang, Li, and Tsai 2007a; Zhang and Lu 2007). With the aid of local linear approximation (LLA) algorithm (Zou and Li 2008), LARS can be adopted to solve optimization problems of nonconcave penalized likelihood functions. However, the above computational procedures rely on the regularization parameter. Hence, the selection of this parameter becomes the second challenge, and is therefore the primary aim of our paper.

In the literature, selection criteria are usually classified into two categories: consistent [e.g., the Bayesian information criterion (BIC), Schwarz 1978] and efficient [e.g., the Akaike information criterion (AIC), Akaike 1974; the generalized cross-validation (GCV), Craven and Wahba 1979]. A consistent criterion identifies the true model with a probability that approaches 1 in large samples when a set of candidate models contains the true model. An efficient criterion selects the model so that its average squared error is asymptotically equivalent to the minimum offered by the candidate models when the true model is approximated by a family of candidate models. Detailed discussions on efficiency and consistency can be found in Shibata (1981, 1984), Li (1987), Shao (1997), and McQuarrie and Tsai (1998). In the context of linear and generalized linear models (GLIM) with a nonconcave penalized function, Fan and Li (2001) proposed applying GCV to choose the regularization parameter. Recently, Wang, Li, and Tsai (2007b) found that the resulting model selected by GCV tends to overfit, while BIC is able to identify the finite-dimensional true linear and partial linear models consistently. Wang, Li, and Tsai (2007b) also indicated that GCV is similar to AIC. However, they only studied the penalized least squares function with the smoothly clipped absolute deviation (SCAD) penalty. This motivated us to study the issues of regularization parameter selection for penalized likelihood-based models with a nonconcave penalized function.

In this paper, we adopt Nishii's (1984) generalized information criterion (GIC) to choose regularization parameters in nonconcave penalized likelihood functions. This criterion not only contains AIC and BIC as its special cases, but also bridges the connection between the classical variable selection criteria and

Yiyun Zhang is Senior Biostatistician, Novartis Oncology, Florham Park, NJ 07932 (E-mail: *yiyun.zhang@novartis.com*). Runze Li is the correspondence author and Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111 (E-mail: *rli@stat.psu.edu*). Chih-Ling Tsai is Robert W. Glock Chair professor, Graduate School of Management, University of California, Davis, CA 95616-8609 (E-mail: *cltsai@ucdavis.edu*). We are grateful to the editor, the associate editor, and three referees for their helpful and constructive comments that substantially improved an earlier draft. Zhang's research is supported by National Institute on Drug Abuse grants R21 DA024260 and P50 DA10075 as a research assistant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH. Li's research is supported by National Science Foundation grants DMS 0348869 and 0722351.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Theory and Methods
DOI: 10.1198/jasa.2009.tm08013

the nonconcave penalized likelihood methodology. This connection provides more flexibility for practitioners to employ their own favored variable selection criteria in choosing desirable models. When the true model is among a set of candidate models with the GLIM structure, we show that the BIC-type tuning parameter selector enables us to identify the true model consistently, whereas the AIC-type selector tends to yield an overfitted model. On the other hand, if the true model is approximated by a set of candidate models with the linear structure, we demonstrate that the AIC-type tuning parameter selector is asymptotically loss efficient, while the BIC-type selector does not have this property in general. These findings are consistent with the features of AIC (see Shibata 1981; Li 1987) and BIC (see Shao 1997) used in best subset variable selections.

The rest of the paper is organized as follows. Section 2 proposes the GIC under a general nonconcave penalized likelihood setting. Section 3 studies the consistent property of GIC for generalized linear models, while Section 4 investigates the asymptotic loss efficiency of the AIC-type selector for linear regression models. Monte Carlo simulations and an empirical example are presented in Section 5 to illustrate the use of the regularization parameter selectors. Section 6 provides discussions, and technical proofs are given in the Appendix.

2. NONCONCAVE PENALIZED LIKELIHOOD FUNCTION

2.1 Penalized Estimators and Penalty Conditions

Consider the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ being collected identically and independently, where y_i is the response from the ith subject, and $\mathbf{x}_i$ is the associated d dimensional predictor variable. Let $\ell(\boldsymbol{\beta})$ be the log-likelihood-based (or loss) function of d dimensional parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$. Then, adopting Fan and Li's (2001) approach, we define a penalized likelihood to be

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{1}$$

where $p_\lambda(\cdot)$ is a penalty function with regularization parameter λ. Several penalty functions have been proposed in the literature. For example, the L_q $(0 < q < 2)$ penalty, namely $p_\lambda(|\beta|) = q^{-1}\lambda|\beta|^q$, leads to the bridge regression (Frank and Friedman 1993). In particular, the L_1 penalty yields the LASSO estimator (Tibshirani 1996). Fan and Li (2001) proposed the nonconcave penalized likelihood method and advocated the use of the SCAD penalty, whose first derivative is given by

$$p'_\lambda(|\beta_j|) = \lambda\left\{I(|\beta_j| \le \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I(|\beta_j| > \lambda)\right\},$$

with $a = 3.7$ and $p_\lambda(0) = 0$. With properly chosen regularization parameter λ from 0 to its upper limit, $\lambda_{\max}$, the resulting penalized estimator is sparse and therefore suitable for variable selections.

To investigate the asymptotic properties of the regularization parameter selectors, we present the penalty conditions given below:

(C1) Assume that $\lambda_{\max}$ depends on n and satisfies $\lambda_{\max} \to 0$ as $n \to \infty$.

(C2) There exists a constant m such that the penalty $p_\lambda(\zeta)$ satisfies $p'(\zeta) = 0$ for $\zeta > m\lambda$.

(C3) If $\lambda_n \to 0$ as $n \to \infty$, then the penalty function satisfies

$$\liminf_{n\to\infty} \liminf_{\zeta\to 0+} \sqrt{n} p'_{\lambda_n}(\zeta) \to \infty.$$

Condition (C1) indicates that a smaller regularization parameter is needed if the sample size is large. Condition (C2) assures that the resulting penalized likelihood estimate is asymptotically unbiased (Fan and Li 2001). Both the SCAD and Zhang's (2007) minimax concave penalties (MCP) satisfy this condition. Condition (C3) is adapted from Fan and Li's (2001) equation (3.5), which is used to study the oracle property.

Remark 1. Both the SCAD penalty and L_q penalty for $0 < q \le 1$ are singular at the origin. Hence, it becomes challenging to maximize their corresponding penalized likelihood functions. Accordingly, Fan and Li (2001) proposed the LQA algorithm for finding the solution of the nonconcave penalized likelihood. In LQA, $p_\lambda(|\beta|)$ is locally approximated by a quadratic function $q_\lambda(|\beta|)$, whose first derivative is given by

$$q'_\lambda(|\beta_j|) = \{p'_\lambda(|\beta_j|)/|\beta_j|\}\beta_j.$$

The above equation is evaluated at the $(k+1)$-step iteration of the Newton–Raphson algorithm if $\beta_j^{(k)}$ is not very close to zero. Otherwise, the resulting parameter estimator of β_j is set to 0.

2.2 Generalized Information Criterion

Before introducing the GIC, we first define the candidate model, which is involved in variable selections.

Definition 1 (Candidate model). We define α, a subset of $\bar{\alpha} = \{1, \ldots, d\}$, as a candidate model, meaning that the corresponding predictors labelled by α are included in the model. Accordingly, $\bar{\alpha}$ is the full model. In addition, we denote the size of model α (i.e., the number of nonzero parameters in α) and the coefficients associated with the predictors in model α by d_α and $\boldsymbol{\beta}_\alpha$, respectively. Moreover, we denote the collection of all candidate models by $\mathcal{A}$. For a penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$ that minimizes the objective function (1), denote the model associated with $\hat{\boldsymbol{\beta}}_\lambda$ by α_λ.

In the normal linear regression model, $y_i = \mathbf{x}_i^T\boldsymbol{\beta}_\alpha + e_i$ and $e_i \sim \mathrm{N}(0, \sigma^2)$ for $i = 1, \ldots, n$, Nishii (1984) proposed the GIC for classical variable selections. It is

$$\mathrm{GIC}_{\kappa_n}(\alpha) = \log\hat{\sigma}_\alpha^2 + \frac{1}{n}\kappa_n d_\alpha,$$

where $\boldsymbol{\beta}_\alpha$ is the parameter of the candidate model α, $\hat{\sigma}_\alpha^2$ is the maximum likelihood estimator of σ^2, and κ_n is a positive number that controls the properties of variable selection. Note that Nishii's GIC is different from the GIC proposed by Konishi and Kitagawa (1996). When $\kappa_n = 2$, GIC becomes AIC, while $\kappa_n = \log(n)$ leads to GIC being BIC. Because GIC contains a broad range of selection criteria, this motivates us to propose the following GIC-type regularization parameter selector,

$$\mathrm{GIC}_{\kappa_n}(\lambda) = \frac{1}{n}\{G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda) + \kappa_n df_\lambda\}, \tag{2}$$

where $G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)$ measures the fitting of model α_λ, $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\hat{\boldsymbol{\beta}}_\lambda$ is the penalized parameter estimator obtained by maximizing Equation (1) with respect to $\boldsymbol{\beta}$, and df_λ is the degrees of

freedom of model α_λ. For any given κ_n, we select λ that minimizes $\mathrm{GIC}_{\kappa_n}(\lambda)$. We can see that, the larger κ_n is, the higher the penalty for models with more variables. Therefore, for some given data, the size of the selected model decreases as κ_n increases.

Remark 2. For any given model α (including the penalized model α_λ and the full model $\bar{\alpha}$), we are able to obtain the nonpenalized parameter estimator $\hat{\boldsymbol{\beta}}^*_\alpha$ by maximizing the log-likelihood function $\ell(\boldsymbol{\beta})$ in (1). Then, Equation (2) becomes

$$\mathrm{GIC}^*_{\kappa_n}(\alpha) = \frac{1}{n}\{G(\mathbf{y}, \hat{\boldsymbol{\beta}}^*_\alpha) + \kappa_n d_\alpha\}, \tag{3}$$

which can be used for classical variable selections. In addition, $\mathrm{GIC}^*_{\kappa_n}(\alpha)$ turns into Nishii's $\mathrm{GIC}_{\kappa_n}(\alpha)$ if we replace $G(\mathbf{y}, \hat{\boldsymbol{\beta}}^*_\alpha)$ in (3) with the -2log-likelihood function of the fitted normal regression model.

We next study the degrees of freedom used in the second term of GIC. In the selection of the regularization parameter, Fan and Li (2001, 2002) proposed that the degrees of freedom be the trace of the approximate linear projection matrix, that is,

$$df_L(\lambda) \triangleq \mathrm{tr}\big\{(\nabla^{\otimes 2}_\lambda Q^*(\hat{\boldsymbol{\beta}}_\lambda))^{-1}\nabla^{\otimes 2}_\lambda \ell(\hat{\boldsymbol{\beta}}_\lambda)\big\}, \tag{4}$$

where $Q^*(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^d q_\lambda(|\beta_j|)$, $[\nabla^{\otimes 2}_\lambda Q^*(\boldsymbol{\beta})]_{jj'} = \frac{\partial^2}{\partial\beta_j\beta_{j'}}Q^*(\boldsymbol{\beta})$, and $[\nabla^{\otimes 2}_\lambda \ell(\boldsymbol{\beta})]_{jj'} = \frac{\partial^2}{\partial\beta_j\beta_{j'}}\ell(\boldsymbol{\beta})$ for j, j' such that $\hat{\beta}_j \neq 0$ and $\hat{\beta}_{j'} \neq 0$. To understand the large sample property of $df_L(\lambda)$, we show its asymptotic behavior given below.

Proposition 1. Assume that the penalized likelihood estimator $\hat{\boldsymbol{\beta}}_\lambda$ is sparse (i.e., with probability tending to one, $\hat{\beta}_{\lambda j} = 0$ if the true value of β_j is 0) and consistent, where $\hat{\beta}_{\lambda j}$ is the jth component of $\hat{\boldsymbol{\beta}}_\lambda$. Under Conditions (C1) and (C2), we have

$$P\big\{df_L(\lambda) = d_{\alpha_\lambda}\big\} \to 1,$$

where d_{α_λ} is the size of model α_λ.

Proof. After algebraic simplifications,

$$df_L(\lambda) = \mathrm{tr}\big\{(\nabla^{\otimes 2}_\lambda \ell_{\alpha_\lambda}(\hat{\boldsymbol{\beta}}_\lambda) + n\Sigma_\lambda)^{-1}\nabla^{\otimes 2}_\lambda \ell(\hat{\boldsymbol{\beta}}_\lambda)\big\},$$

where $\Sigma_\lambda = \mathrm{diag}_{\hat{\beta}_{\lambda j}\neq 0}\{p'_\lambda(|\hat{\beta}_{\lambda j}|)/|\hat{\beta}_{\lambda j}|\}$. Because of the consistency and sparsity of $\hat{\boldsymbol{\beta}}_\lambda$, $\hat{\beta}_{\lambda j}$ converges to β_j with probability tending to 1 for all j such that $\hat{\beta}_{\lambda j} > 0$. Hence, those $\hat{\beta}_{\lambda j}$ are all bounded from 0. This result, together with Conditions (C1) and (C2), implies that $\Sigma_\lambda = \mathbf{0}$ with probability tending to 1. Subsequently, using the fact that $n^{-1}\nabla^{\otimes 2}_\lambda \ell(\hat{\boldsymbol{\beta}}_\lambda) = O_P(1)$, we complete the proof.

The above proposition suggests that the difference between $df_L(\lambda)$ and the size of the model, d_{α_λ}, is small. Because d_{α_λ} is simple to calculate, we use it as the degrees of freedom df_λ in (2). In linear regression models, Efron et al. (2004) and Zou, Hastie, and Tibshirani (2007) also suggested using d_{α_λ} as an estimator of the degrees of freedom for LASSO. Moreover, Zou, Hastie, and Tibshirani (2007) showed that d_{α_λ} is an asymptotically unbiased estimator. In this article, our asymptotical results are valid without regard to the use of $df_L(\lambda)$ or d_{α_λ} as the degrees of freedom. When the sample size is small, however, $df_L(\lambda)$ should be considered. In the following section, we explore the properties of GIC for the generalized linear models which have been widely used in various disciplines.

3. CONSISTENCY

In this section, we assume that the set of candidate models contains the unique true model, and that the number of parameters in the full model is finite. Under this assumption, we are able to study the asymptotic consistency of GIC by introducing the following definition and condition.

Definition 2 (Underfitted and overfitted models). We assume that there is a unique true model α_0 in $\mathcal{A}$, whose corresponding coefficients are nonzero. Therefore, any candidate model $\alpha \not\supset \alpha_0$, is referred to as an underfitted model, while any $\alpha \supset \alpha_0$ other than α_0 itself is referred to as an overfitted model.

Based on the above definitions, we partition the tuning parameter interval $[0, \lambda_{\max}]$ into the underfitted, true, and overfitted subsets, respectively,

$$\Omega_- = \{\lambda : \alpha_\lambda \not\supset \alpha_0\},$$
$$\Omega_0 = \{\lambda : \alpha_\lambda = \alpha_0\}, \qquad \text{and}$$
$$\Omega_+ = \{\lambda : \alpha_\lambda \supset \alpha_0 \text{ and } \alpha_\lambda \neq \alpha_0\}.$$

This partition allows us to assess the performance of regularization parameter selections.

To investigate the asymptotic properties of the regularization parameter selectors, we introduce the technical condition given below:

(C4) For any candidate model $\alpha \in \mathcal{A}$, there exists $c_\alpha > 0$ such that $\frac{1}{n}G(\mathbf{y}, \hat{\boldsymbol{\beta}}^*_\alpha) \xrightarrow{P} c_\alpha$. In addition, for any underfitted model $\alpha \not\supset \alpha_0$, $c_\alpha > c_{\alpha_0}$, where c_{α_0} is the limit of $\frac{1}{n}G(\mathbf{y}, \boldsymbol{\beta}_{\alpha_0})$ and $\boldsymbol{\beta}_{\alpha_0}$ is the parameter vector of the true model α_0.

The above condition assures that the underfitted model yields a larger measure of model fitting than that of the true model. We next explore the asymptotic consistency of GIC for the generalized linear models which have been used in various disciplines.

Consider the generalized linear model (GLIM)—see McCullagh and Nelder (1989)—whose conditional density function of y_i given x_i is

$$f_i(y_i; \theta_i, \phi) = \exp\big\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\big\}, \tag{5}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$ are suitably chosen functions, θ_i is the canonical parameter, $E(y_i|x_i) = \mu_i = b'(\theta_i)$, $g(\mu_i) = \theta_i$, g is a link function, and ϕ is a scale parameter. Throughout this paper, we assume that ϕ is known (such as in the logistic regression model and the Poisson log-linear model) or that it can be estimated by fitting the data with the full model (e.g., the normal linear model). In addition, we follow the classical regression approach to model θ_i by $\mathbf{x}_i^T\boldsymbol{\beta}$. Based on (5), the log-likelihood-based function in (1) is

$$\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\mu}; \mathbf{y}) = \ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log f_i(y_i; \theta_i, \phi)$$
$$= \sum_{i=1}^n \big[\{y_i\mathbf{x}_i^T\boldsymbol{\beta} - b(\mathbf{x}_i^T\boldsymbol{\beta})\}/a(\phi) + c(y_i, \phi)\big], \tag{6}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$. Then, the resulting scaled deviance of a penalized estimate $\hat{\boldsymbol{\beta}}_\lambda$ is

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}_\lambda) = 2\{\ell(\mathbf{y}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}_\lambda; \mathbf{y})\},$$

where $\hat{\boldsymbol{\mu}}_\lambda = (g^{-1}(\mathbf{x}_1^T\hat{\boldsymbol{\beta}}_\lambda), \ldots, g^{-1}(\mathbf{x}_n^T\hat{\boldsymbol{\beta}}_\lambda))^T$.

For model α_λ, we employ the scaled deviance $D(\mathbf{y}; \hat{\boldsymbol{\mu}}_\lambda)$ as the goodness-of-fit measure, $G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)$, in (2) so that the resulting GIC for GLIM is

$$\mathrm{GIC}_{\kappa_n}(\lambda) = \frac{1}{n}D(\mathbf{y}; \hat{\boldsymbol{\mu}}_\lambda) + \frac{1}{n}\kappa_n df_\lambda. \tag{7}$$

In addition, when we fit the data with the nonpenalized likelihood approach under model α, GIC becomes

$$\mathrm{GIC}^*_{\kappa_n}(\alpha) = \frac{1}{n}D(\mathbf{y}; \hat{\boldsymbol{\mu}}^*_\alpha) + \frac{1}{n}\kappa_n d_\alpha, \tag{8}$$

where $\hat{\boldsymbol{\mu}}^*_\alpha = (g^{-1}(\mathbf{x}_1^T\hat{\boldsymbol{\beta}}^*_\alpha), \ldots, g^{-1}(\mathbf{x}_n^T\hat{\boldsymbol{\beta}}^*_\alpha))^T$, and $\hat{\boldsymbol{\beta}}^*_\alpha$ is the nonpenalized maximum likelihood estimator of $\boldsymbol{\beta}$. Accordingly, GIC* can be used in classical variable selection [see equation (3.10) of McCullagh and Nelder 1989]. Next, we show the asymptotic performances of GIC.

Theorem 1. Suppose the density function of the generalized linear model satisfies Fan and Li's 2001 three regularity conditions (A)–(C) and that the technical condition (C4) holds.

(A) If there exists a positive constant M such that $\kappa_n < M$, then the tuning parameter $\hat{\lambda}$ selected by minimizing $\mathrm{GIC}_{\kappa_n}(\lambda)$ in (7) satisfies

$$P\{\hat{\lambda} \in \Omega_-\} \to 0 \qquad \text{and} \qquad P\{\hat{\lambda} \in \Omega_+\} \geq \pi,$$

where π is a nonzero probability.

(B) Suppose that Conditions (C1)–(C3) are satisfied. If $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$, then the tuning parameter $\hat{\lambda}$ selected by minimizing $\mathrm{GIC}_{\kappa_n}(\lambda)$ in (7) satisfies $P\{\alpha_{\hat{\lambda}} = \alpha_0\} \to 1$.

The proof is given in Appendix A.

Theorem 1 provides guidance on the choice of the regularization parameter. Theorem 1(A) implies that the GIC selector with bounded κ_n tends to overfit without regard to which penalty function is being used. Analogous to the classical variable selection criterion AIC, we refer to GIC with $\kappa_n = 2$ as the AIC selector, while we name GIC with $\kappa_n \to 2$ the AIC-type selector. In contrast, because $\kappa_n = \log(n)$ fulfills the conditions of Theorem 1(B), we call GIC with $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$ the BIC-type selector. Theorem 1(B) indicates that the BIC-type selector identifies the true model consistently. Thus, the nonconcave penalized likelihood of the generalized linear model with the BIC-type selector possesses the oracle property.

Remark 3. In linear regression models, Fan and Li (2001) applied the GCV selector given below to choose the regularization parameter,

$$\mathrm{GCV}^*(\lambda) = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2}{n\{1 - df_L(\lambda)/n\}^2}, \tag{9}$$

where $\|\cdot\|$ denotes the Euclidean norm, $X = (x_1^T, \ldots, x_n^T)^T$, and $df_L(\lambda)$ is defined in (4). To extend the application of GCV, we replace the residual sum of squares in (9) by $G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)$, and then choose λ that minimizes

$$\mathrm{GCV}(\lambda) = \frac{G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)}{n\{1 - df_\lambda/n\}^2}. \tag{10}$$

Using the Taylor expansion, we further have that

$$\mathrm{GCV}(\lambda) \approx \frac{1}{n}\left\{G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda) + df_\lambda\left[\frac{2G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)}{n}\right]\right\}.$$

Because $G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)/n$ is bounded via Condition (C4), GCV yields an overfitted model with a positive probability. Therefore, the penalized partial likelihood with the GCV selector does not possess the oracle property.

Remark 4. In linear regression models, Wang, Li, and Tsai (2007b) demonstrated that Fan and Li's (2001) GCV-selector for the SCAD penalized least squares procedure cannot select the tuning parameter consistently. They further proposed the following BIC tuning parameter selector:

$$\mathrm{BIC}^*(\lambda) = \log(\hat{\sigma}^2_\lambda) + \frac{1}{n}\log(n)\, df_L(\lambda),$$

where $\hat{\sigma}^2_\lambda = \sum_{i=1}^n (y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\lambda)^2/n$. Using the result that $\log(1 + t) \approx t$ for small t, BIC^*_λ is approximately equal to

$$\mathrm{BIC}^{**}(\lambda) = \frac{1}{n}D^{**}(\mathbf{y}; \hat{\boldsymbol{\mu}}_\lambda) + \frac{1}{n}\log(n)\, df_L(\lambda),$$

where $D^{**}(\mathbf{y}; \hat{\boldsymbol{\mu}}_\lambda) = \hat{\sigma}^2_\lambda/\hat{\sigma}^2_\alpha$ is the scaled deviance of the normal distribution, and $\hat{\sigma}^2_\alpha$ is the dispersion parameter estimator computed from the full model. It can be seen that BIC** is a BIC-type selector. Under the conditions in Theorem 1(B), the SCAD penalized least squares procedure with the BIC** selector possesses the oracle property, which is consistent with the findings in Wang, Li, and Tsai (2007b).

4. EFFICIENCY

Under the assumption that the true model is included in a family of candidate models, we established the consistency of BIC-type selectors. In practice, however, this assumption may not be valid, which motivates us to study the asymptotic efficiency of the AIC-type selector. In the literature, the L_2 norm has been commonly used to assess the efficiency of the classical AIC procedure (see Shibata 1981, 1984 and Li 1987) in linear regression models. Hence, we focus on the efficiency of linear regression model selections via L_2 norm.

Consider the following model:

$$y_i = \mu_i + \epsilon_i \quad \text{for } i = 1, \ldots, n,$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ is an unknown mean vector, and ϵ_i's are independent and identically distributed (iid) random errors with mean 0 and variance σ^2. Furthermore, we assume that $\mathbf{X}\boldsymbol{\beta}$ constitutes the nearest representation of the true mean vector $\boldsymbol{\mu}$, and hence the full model is not necessarily a correct model. Adapting the formulation of Li (1987), we allow d, the dimension of $\boldsymbol{\beta}$, to tend to infinity with n, but $d/n \to 0$.

For the given dataset $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$, we follow the formulation of (1) to define the penalized least squares function

$$Q^{LS}(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^d p_\lambda(|\beta_j|). \tag{11}$$

The resulting penalized estimator of $\boldsymbol{\mu}$ in model α_λ is $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$. In addition, the nonpenalized estimator of $\boldsymbol{\mu}$ in model α is $\hat{\boldsymbol{\mu}}^*_\alpha = \mathbf{X}\hat{\boldsymbol{\beta}}^*_\alpha$. It is noteworthy that $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\beta}}^*_\alpha$ have been defined in

Definition 1 and Remark 2, respectively. In practice, the tuning parameter λ is unknown and can be selected by minimizing

$$\mathrm{GIC}_{\kappa_n}^{LS}(\lambda) = \frac{1}{n}\left\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\lambda)^2 + \kappa_n\sigma^2 d_{\alpha_\lambda}\right\}, \tag{12}$$

where σ^2 is assumed to be known, and the case with the unknown σ^2 will be discussed later. When $\hat{\boldsymbol{\beta}}_\lambda$ in (12) is replaced by the least squares estimator $\hat{\boldsymbol{\beta}}_\alpha^*$, $\mathrm{GIC}_{\kappa_n}^{LS}$ with $\kappa_n = 2$ becomes the Mallows' C_p criterion that has been used for classical variable selections.

To assess the performance of the tuning parameter selector, we adopt the approach of Shibata (1981) (also see Li 1987; Shao 1997), and define the average squared loss (or the L_2 loss) associated with the estimator $\hat{\boldsymbol{\beta}}_\lambda$ to be

$$L(\hat{\boldsymbol{\beta}}_\lambda) = \frac{1}{n}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_\lambda\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\lambda)^2. \tag{13}$$

Accordingly, the risk function is $R(\hat{\boldsymbol{\beta}}) = E[L(\hat{\boldsymbol{\beta}})]$.

Using the average squared loss measure, we further define the asymptotic loss efficiency.

Definition 3 (Asymptotically loss efficient). A tuning parameter selection procedure is said to be asymptotically loss efficient if

$$\frac{L(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})}{\inf_{\lambda\in[0,\lambda_{\max}]} L(\hat{\boldsymbol{\beta}}_\lambda)} \to 1, \tag{14}$$

in probability, where $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ is associated with the tuning parameter $\hat{\lambda}$ selected by this procedure. We also say $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ is asymptotically loss efficient if (14) holds.

Moreover, we introduce the following technical conditions for studying the asymptotic loss efficiency of the AIC-type selector in linear regression:

(C5) $(\frac{1}{n}\mathbf{X}^T\mathbf{X})^{-1}$ exists, and its largest eigenvalue is bounded by a constant number C.

(C6) $E\epsilon_1^{4q} < \infty$, for some positive integer q.

(C7) The risks of the least squares estimators $\hat{\boldsymbol{\beta}}_\alpha^*$ for all $\alpha \in \mathcal{A}$ satisfy

$$\sum_{\alpha\in\mathcal{A}}[nR(\hat{\boldsymbol{\beta}}_\alpha^*)]^{-q} \to 0.$$

(C8) Let $\mathbf{b} = (b_1, \ldots, b_d)^T$, where $b_j = p'_\lambda(|\hat{\beta}_{\lambda j}|)\,\mathrm{sgn}(\hat{\beta}_{\lambda j})$ for all j such that $|\hat{\beta}_{\lambda j}| > 0$, and $b_j = 0$ otherwise, and $\hat{\beta}_{\lambda j}$ is the j-th component of the penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$. In addition, let $\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^*$ be the least squares estimator of β obtained from model α_λ. Then, we assume that, in probability,

$$\sup_{\lambda\in[0,\lambda_{\max}]} \frac{\|\mathbf{b}\|^2}{R(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^*)} \to 0.$$

Condition (C5) has been commonly considered in the literature. Conditions (C6) and (C7) are adopted from conditions (A.2) and (A.3), respectively, in Li (1987). It can be shown that if the true model is approximated by a set of the candidate models (e.g., the true model is of infinite dimension), then Condition (C7) holds. Condition (C8) ensures that the difference between the penalized mean function estimator and the corresponding least squares mean function estimator is small in comparison with the risk of the least squares estimator (see Lemma 3 in Appendix B). Sufficient conditions for (C8) are also given in Appendix C. We next show the asymptotic efficiency of the AIC-type selector.

Theorem 2. Assume Conditions (C5)–(C8) hold. Then, the tuning parameter $\hat{\lambda}$ selected by minimizing $\mathrm{GIC}_{\kappa_n}^{LS}(\lambda)$ in (12) with $\kappa_n \to 2$ yields an asymptotically loss efficient estimator, $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$, in the sense of (14).

The proof is given in Appendix B.

Theorem 2 demonstrates that the AIC-type selector is asymptotically loss efficient. In addition, using the result that $\log(1+t) \approx t$ for small t, the AIC selector behaves similarly to the following AIC^* selector,

$$\mathrm{AIC}^*(\lambda) = \log(\hat{\sigma}_\lambda^2) + \frac{2\sigma^2 d_{\alpha_\lambda}}{n}.$$

Accordingly, both AIC and AIC^* selectors are asymptotically loss efficient.

Applying Lemma 4 and Equation (B.1) in Appendix B, we find that if $\sup_{\lambda\in[0,\lambda_{\max}]}|n^{-1}(\kappa_n - 2)\sigma^2 d_{\alpha_\lambda}/R(\hat{\boldsymbol{\beta}}_\lambda)| \to 0$ in probability, then $\mathrm{GIC}_{\kappa_n}^{LS}$ is asymptotically loss efficient. Note that it can be shown $n^{-1}\sigma^2 d_{\alpha_\lambda}/R(\hat{\boldsymbol{\beta}}_\lambda)$ is bounded by 1. As a result, $\kappa_n \to 2$ (including $\kappa_n = 2$) is critical in establishing the asymptotic loss efficiency. This finding is similar to the classical efficient criteria (see Shibata 1980; Shao 1997; and Yang 2005). It is also not surprising that the BIC-type selectors do not possess asymptotic loss efficiency, which is consistent with the finding of classical variable selections (see Li 1987 and Shao 1997).

In practice, σ^2 is often unknown. It is natural to replace the σ^2 in GIC_2^{LS} by its consistent estimator (see Shao 1997). The following corollary shows that the asymptotical property of GIC still holds.

Corollary 1. If the tuning parameter $\hat{\lambda}$ is selected by minimizing $\mathrm{GIC}_{\kappa_n}^{LS}(\lambda)$ with $\kappa_n \to 2$ and σ^2 being replaced by its consistent estimator $\tilde{\sigma}^2$, then the resulting procedure is also asymptotically loss efficient.

The proof is given in Appendix B.

Remark 5. As suggested by an anonymous reviewer, we study the asymptotic loss efficiency of the generalized linear model in (5). Following the spirit of the deviance measure commonly used in the GLIM, we adopt the Kullback–Leibler (KL) distance measure to define the KL loss of an estimate $\hat{\boldsymbol{\beta}}$ (either the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_\alpha^*$ or the nonconcave penalized estimate $\hat{\boldsymbol{\beta}}_\lambda$), as

$$L_{KL}(\hat{\boldsymbol{\beta}}) = \frac{2}{n}E_0\{\ell(\boldsymbol{\theta}_0) - \ell(\hat{\boldsymbol{\theta}})\}, \tag{15}$$

where $\ell(\cdot)$ is defined in (6), $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0n})^T$ is the true unknown canonical parameter, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)^T = \mathbf{X}\hat{\boldsymbol{\beta}}$ under canonical link function, and E_0 denotes the expectation under the true model (see, e.g., McQuarrie and Tsai 1998). It can be

shown that the KL loss is identical to the squared loss for normal linear regression models with known variance. To obtain the asymptotic loss efficiency for GLIM, however, it is necessary to employ the Taylor expansion to expand $b(\hat{\theta}_i)$ at the true value of θ_{0i} for $i = 1, \ldots, n$. Accordingly, we need to impose some strong assumptions to establish the asymptotic loss efficiency for the nonconcave penalized estimate under the KL loss if the tuning parameter $\hat{\lambda}$ is selected by minimizing GIC in (7) with $\kappa_n \to 2$. To avoid considerably increasing the length of the paper, we do not present detailed discussions, justifications, and proofs here, but they are given in the online supplemental material available in *JASA* website.

5. NUMERICAL STUDIES

In this section, we present three examples comprised of two Monte Carlo experiments and one empirical example. It is noteworthy that Example 1 considers a setting in which the true model is in a set of candidate models with logistic regressions. This setting allows us to examine the finite sample performance of the consistent selection criterion, which is expected to perform well via Theorem 1. In contrast, Example 2 considers a setting in which the true model is not included in a set of candidate models with Gaussian regressions. This setting enables us to assess the finite sample performance of the efficient selection criterion, which is expected to perform well via Theorem 2.

Example 1. Adapting the model setting from Tibshirani (1996) and Fan and Li (2001), we simulate the data from the logistic regression model, $y|\mathbf{x} \sim \text{Bernoulli}\{p(\mathbf{x}^T\boldsymbol{\beta})\}$, where

$$p(\mathbf{x}^T\boldsymbol{\beta}) = \mu(\mathbf{x}^T\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T\boldsymbol{\beta})}.$$

In addition, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0)^T$, $\mathbf{x}$ is a 12-dimensional random vector, the first nine components of $\mathbf{x}$ are multivariate normal with covariance matrix $\Sigma = (\sigma_{ij})$, in which $\sigma_{ij} = 0.5^{|i-j|}$, and the last three components of $\mathbf{x}$ are generated from an independent Bernoulli distribution with success probability 0.5. Moreover, we conduct 1000 realizations with the sample sizes $n = 200$ and 400.

To assess the finite sample performance of the proposed methods, we report the percentage of models correctly fitted, underfitted, and overfitted with 1, 2, 3, 4, 5 or more parameters by SCAD-AIC (i.e., $\kappa_n = 2$), SCAD-BIC [i.e., $\kappa_n = \log(n)$], SCAD-GCV, AIC, and BIC selectors as well as via the oracle procedure (i.e., the simulated data were fitted with the true model). Their corresponding standard errors can be calculated by $\sqrt{\hat{p}(1-\hat{p})/1000}$, where $\hat{p}$ is the observed proportion in 1000 simulations. Moreover, we report the average number of zero coefficients that were correctly (C) and incorrectly (I) identified by different methods. To compare model fittings, we further calculate the following model error for the new observation $(\mathbf{x}, y)$:

$$\text{ME}(\hat{\boldsymbol{\beta}}) = E_{\mathbf{x}}\{\mu(\mathbf{x}^T\boldsymbol{\beta}) - \mu(\mathbf{x}^T\hat{\boldsymbol{\beta}})\}^2,$$

where the expectation is taken with respect to the new observed covariate vector $\mathbf{x}$, and $\mu(\mathbf{x}^T\boldsymbol{\beta}) = E(y|\mathbf{x})$. Then, we report the median of the relative model error (MRME), where the relative model error is defined as $\text{RME} = \text{ME}/\text{ME}_{\text{full}}$, and ME_{full} is the model error calculated by fitting the data with the full model.

Table 1 shows that the MRME of SCAD-BIC is smaller than that of SCAD-AIC. As the sample size increases, the MRME of SCAD-BIC approaches that of the oracle estimator, whereas the MRME of SCAD-AIC remains at the same level. Hence, SCAD-BIC is superior to SCAD-AIC in terms of model error measures in this example. Moreover, Figure 1 presents the boxplots of RMEs, which lead to a similar conclusion.

In model identifications, SCAD-BIC has a higher chance than SCAD-AIC of correctly setting the nine true zero coefficients to zero, while SCAD-BIC is slightly more prone than SCAD-AIC to incorrectly set the three nonzero coefficients to zero when the sample size is small. In addition, SCAD-BIC has a much higher possibility of correctly identifying the true model

Table 1. Simulation results for the logistic regression model. MRME is the median relative model error; C denotes the average number of the nine true zero coefficients that were correctly identified as zero, while I denotes the average number of the three nonzero coefficients that were incorrectly identified as zero; the numbers in parentheses are standard errors; Under, Exact, and Overfitted represent the proportions of corresponding models being selected in 1000 Monte Carlo realizations, respectively

	MRME	Zeros		Under	Exact	Overfitted (%)				
Method	(%)	C	I	(%)	(%)	1	2	3	4	≥5
$n = 200$										
SCAD-AIC	58.03	7.4 (1.6)	0.0 (0.0)	0.2	30.2	23.2	19.6	13.6	7.5	5.9
SCAD-BIC	16.90	8.8 (0.5)	0.0 (0.1)	0.7	83.7	13.2	2.4	0.4	0.0	0.0
SCAD-GCV	81.91	5.8 (1.9)	0.0 (0.0)	0.0	7.8	12.0	18.1	18.8	17.6	25.7
AIC	57.05	7.4 (1.2)	0.0 (0.0)	0.1	19.9	31.5	26.7	14.8	4.3	2.8
BIC	18.27	8.8 (0.5)	0.0 (0.1)	0.8	80.4	17.0	2.3	0.1	0.0	0.0
Oracle	12.49	9.0 (0.0)	0.0 (0.0)	0.0	100.0	0.0	0.0	0.0	0.0	0.0
$n = 400$										
SCAD-AIC	64.45	7.4 (1.5)	0.0 (0.0)	0.0	29.4	21.9	21.0	15.1	8.5	4.1
SCAD-BIC	19.03	8.9 (0.4)	0.0 (0.0)	0.0	89.9	7.9	1.8	0.2	0.2	0.0
SCAD-GCV	82.13	6.0 (1.9)	0.0 (0.0)	0.0	9.2	14.7	18.6	19.8	15.9	21.8
AIC	63.17	7.4 (1.2)	0.0 (0.0)	0.0	21.7	29.2	27.3	14.6	5.5	1.7
BIC	20.46	8.8 (0.4)	0.0 (0.0)	0.0	86.3	12.1	1.4	0.2	0.0	0.0
Oracle	15.88	9.0 (0.0)	0.0 (0.0)	0.0	100.0	0.0	0.0	0.0	0.0	0.0

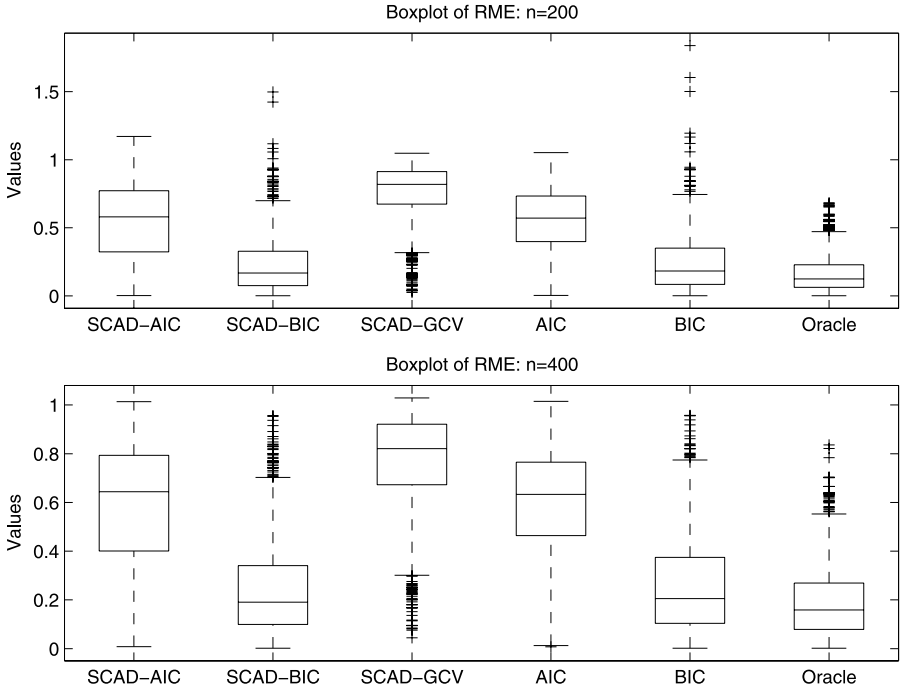

Figure 1. Boxplot of relative model error (RME) with $n = 200$ and $n = 400$.

than that of SCAD-AIC. Moreover, among the overfitted models, the SCAD-BIC method is likely to include only one irrelevant variable, whereas the SCAD-AIC method often includes two or more. When the sample size increases, the SCAD-BIC method yields a lesser degree of overfitting, while the SCAD-AIC method still overfits quite seriously. These results are consistent with the theoretical findings presented in Theorem 1.

It is not surprising that SCAD-GCV performs similar to SCAD-AIC. Furthermore, the classical AIC and BIC selection criteria behave as SCAD-AIC and SCAD-BIC, respectively, in terms of model errors and identifications. It is of interest to note that SCAD-AIC usually suggests a sparser model than AIC. In sum, SCAD-BIC performs the best.

Example 2. In practice, it is plausible that the full model fails to include some important explanatory variable. This motivates us to mimic such a situation and then assess various selection procedures. Accordingly, we consider a linear regression model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, where $\mathbf{x}_i$'s are iid multivariate normal random variables with dimension 13, the correlation between x_i and x_j is $0.5^{|i-j|}$, and ϵ_i's are iid $\mathrm{N}(0, \sigma^2)$ with $\sigma = 4$. In addition, we partition $\mathbf{x} = (\mathbf{x}_{\mathrm{full}}^T, \mathbf{x}_{\mathrm{exc}}^T)^T$, where $\mathbf{x}_{\mathrm{full}}$ contains $d = 12$ covariates of the full model and $\mathbf{x}_{\mathrm{exc}}$ is the covariate excluded from model fittings. Accordingly, we partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathrm{full}}^T, \boldsymbol{\beta}_{\mathrm{exc}}^T)^T$, where $\boldsymbol{\beta}_{\mathrm{full}}$ is a 12×1 vector and $\boldsymbol{\beta}_{\mathrm{exc}}$ is a scalar. To investigate the performance of the proposed methods under various parameter structures, we let $\boldsymbol{\beta}_{\mathrm{full}} = \boldsymbol{\beta}_0 + \gamma\boldsymbol{\delta}/\sqrt{n}$, where $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0)^T$, $\boldsymbol{\delta} = (0, 0, 1.5, 1.5, 1, 1, 0, 0, 0, 0, 0.5, 0.5)^T$, γ ranges from 0 to 10, and $\boldsymbol{\beta}_{\mathrm{exc}} = 0.2$. Because the candidate model is the subset of the full model that contains 12 covariates in $\mathbf{x}_{\mathrm{full}}$, the above model settings ensure that the true model is not included in the set of candidate models.

We simulate 1000 data sets with $n = 400$, 800, and 1600. To study the performance of selectors, we define the finite sample's loss efficiency,

$$\mathrm{LE}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}}) = \frac{L(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})}{\inf_\lambda L(\hat{\boldsymbol{\beta}}_\lambda)},$$

where $\hat{\lambda}$ is chosen by SCAD-GCV, SCAD-AIC, and SCAD-BIC. It is noteworthy that $\boldsymbol{\beta}_{\mathrm{full}}$ depends on the sample size and γ, a sensible approach to compare the loss efficiency of the selection criterion across different sample sizes is to compare the loss efficiency under the same value of $\boldsymbol{\beta}_{\mathrm{full}}$.

Because the performance of SCAD-GCV is very similar to that of SCAD-AIC, it is not reported here. Figure 2(a) and (b) depict the LEs of SCAD-AIC and SCAD-BIC, respectively, across various γ. Figure 2(a) clearly indicates that the loss efficiency of SCAD-AIC converges to 1 regardless of the value of γ, which corroborates the theoretical finding in Theorem 2. However, the loss efficiency of SCAD-BIC in Figure 2(b) does not show this tendency. Specifically, when γ is close to 0 (i.e., the model is nearly sparse), SCAD-BIC results in smaller loss due to its larger penalty function. As γ increases so that the full model contains the medium-sized coefficients, SCAD-BIC is likely to choose the model that is too sparse and hence increases the loss. When γ becomes large enough and the coefficient $\boldsymbol{\beta}_{\mathrm{exc}}$ is dominated by the remaining coefficients in $\boldsymbol{\beta}_{\mathrm{full}}$, the consistent property of SCAD-BIC tends to have the smaller loss. In sum, Figure 2 shows that SCAD-AIC is efficient, whereas SCAD-BIC is not. Finally, we examine the efficiencies of AIC and BIC, and find that the performances of AIC and BIC are similar to those of SCAD-AIC and SCAD-BIC, respectively, and hence are omitted.

Example 3 (Mammographic mass data). Mammography is the most effective method for screening for the presence of

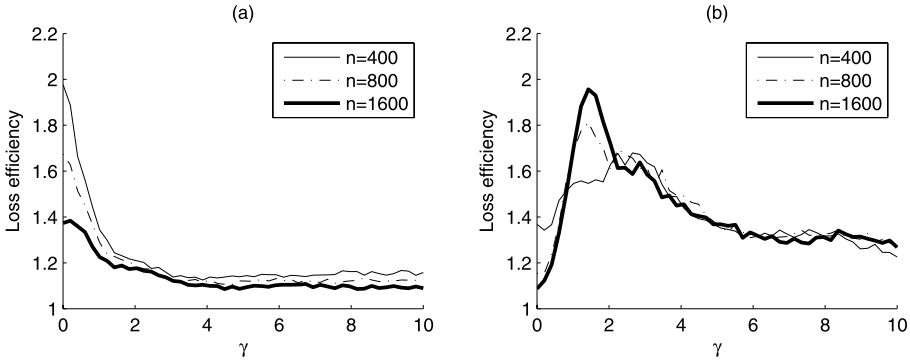

Figure 2. (a) The loss efficiency of SCAD-AIC; (b) The loss efficiency of SCAD-BIC.

breast cancer. However, mammogram interpretations yield approximately a 70% rate of unnecessary biopsies with benign outcomes. Hence, several computer-aided diagnosis (CAD) systems have been proposed to assist physicians in predicting breast biopsy outcomes from the findings of BI-RADS (Breast Imaging Reporting And Data System). To examine the capability of two novel CAD approaches, Elter, Schulz-Wendtland, and Wittenberg (2007) recently analyzed a dataset containing 516 benign and 445 malignant instances. We downloaded this data set from UCI Machine Learning Repository, and excluded 131 missing values and 15 coding errors. As a result, we considered a total of 815 cases.

To study the severity (benign or malignant), Elter, Schulz-Wendtland, and Wittenberg (2007) considered the following three covariates: x_1—*birads*, BI-RADS assessment assigned in a double-review process by physicians (definitely benign = 1 to highly suggestive of malignancy = 5); x_2—*age*, patient's age in years; x_3—*mass density*, high = 1, iso = 2, low = 3, fat-containing = 4. In addition, we employ the dummy variables, x_4 to x_6, to represent the four *mass shapes*: round = 1, oval = 2, lobular = 3, irregular = 4, where "irregular" is used as the baseline. Analogously, we apply the dummy variables, x_7 to x_{10}, to represent the five *mass margins*: circumscribed = 1, microlobulated = 2, obscured = 3, ill-defined = 4, spiculated = 5), where "spiculated" is used as the baseline.

To predict the value of the binary response, $y = 0$ (benign) or $y = 1$ (malignant), on the basis of the 10 explanatory variables, we fit the data with the following logistic regression model:

$$\log \frac{p(\mathbf{x}^T\boldsymbol{\beta})}{1 - p(\mathbf{x}^T\boldsymbol{\beta})} = \beta_0 + \sum_{j=1}^{10} x_j\beta_j,$$

where $\mathbf{x} = (x_1, \ldots, x_{10})^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})^T$, and $p(\mathbf{x}^T\boldsymbol{\beta})$ is the probability of the case being classified as malignant. As a result, the tuning parameters selected by SCAD-AIC and SCAD-BIC are 0.0332 and 0.1512, respectively, and SCAD-GCV yields the same model as SCAD-AIC. Because the true model is unknown, we follow an anonymous referee's suggestion to include the models selected by the delete-1 cross validation (delete-1 CV) and the 5-fold cross validation (5-fold CV). The detailed procedures for fitting those models can be obtained from the first author.

Because the model selected by the delete-1 CV is the same as SCAD-AIC, Table 2 only presents the nonpenalized maximum likelihood estimates from the full model as well as the SCAD-AIC, SCAD-BIC, and 5-fold CV estimates, together with their standard errors. It indicates that the nonpenalized maximum likelihood approach fits five spurious variables (x_3, x_6, and x_8 to x_{10}), while SCAD-AIC and 5-fold CV include two variables (x_6 and x_9) and one variable (x_9), respectively, with insignificant effects at a level of 0.05. In contrast, all variables (x_1, x_2, x_4, x_5, and x_7) selected by SCAD-BIC are significant. Because the sample size $n = 815$ is large, these findings are consistent with Theorem 1 and the simulation results. In addition, the p-value of the deviance test for assessing the SCAD-BIC model

Table 2. Estimates for mammographic mass data with standard deviations in parentheses

	MLE	SCAD-AIC	SCAD-BIC	5-Fold CV
β_0	−11.04 (1.48)	−11.17 (1.13)	−11.16 (1.07)	−11.49 (1.14)
BIRADS (x_1)	2.18 (0.23)	2.19 (0.23)	2.25 (0.23)	2.23 (0.23)
Age (x_2)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.05 (0.01)
Density (x_3)	−0.04 (0.29)	0 (–)	0 (–)	0 (–)
sRound (x_4)	−0.98 (0.37)	−0.99 (0.37)	−0.80 (0.34)	−0.81 (0.35)
sOval (x_5)	−1.21 (0.32)	−1.22 (0.32)	−1.07 (0.30)	−1.07 (0.30)
sLobular (x_6)	−0.53 (0.35)	−0.54 (0.34)	0 (–)	0 (–)
mCircum (x_7)	−1.05 (0.42)	−0.98 (0.32)	−1.01 (0.30)	−1.07 (0.31)
mMicro (x_8)	−0.03 (0.65)	0 (–)	0 (–)	0 (–)
mObscured (x_9)	−0.48 (0.39)	−0.42 (0.30)	0 (–)	−0.47 (0.30)
mIlldef (x_{10})	−0.09 (0.33)	0 (–)	0 (–)	0 (–)

against the full model is 0.41, and as a result, there is no evidence of lack of fit in the SCAD-BIC model.

Based on the model selected by SCAD-BIC, we conclude that a higher BI-RADS assessment or a greater age results in a higher chance for malignancy. In addition, the oval or round mass shape yields lower odds for malignance than does the irregular mass shape, while the odds for malignance designated by the lobular mass shape is not significantly different from that of the irregular mass shape. Moreover, the odds for malignance indicated by the microlobulated, obscured, and ill-defined mass margins are not significantly different from that of the spiculated mass margin. However, the circumscribed mass margin leads to lesser odds for malignance than that of the four other types of mass margins.

6. DISCUSSION

In the context of variable selection, we propose the generalized information criterion to choose regularization parameters for nonconcave penalized likelihood functions. Furthermore, we study the theoretical properties of GIC. If we believe that the true model is contained in a set of candidate models with the GLIM structure, then the BIC-type selector identifies the true model with probability tending to 1, while the GIC selectors with bounded κ_n tends to overfit with positive probability. However, if the true model is approximated by a family of candidate models, the AIC-type selector is asymptotically loss efficient, whereas the BIC-type selector is not, in general. Simulation studies support the finite sample performance of the selection criteria.

Although we obtain the theoretical properties of GIC for GLIM, the application of GIC is not limited. For example, we could employ GIC to select the regularization parameter in Cox proportional hazard and quasi-likelihood models. We believe that these efforts would further enhance the usefulness of GIC in real data analysis.

APPENDIX A: PROOF OF THEOREM 1

Before proving the theorem, we show the following two lemmas. Then, Theorems 1(A) and 1(B) follow from Lemmas 1 and 2, respectively. For the sake of convenience, we denote $D(\mathbf{y}; \hat{\boldsymbol{\mu}}_\lambda)$ in (7) and $D(\mathbf{y}; \hat{\boldsymbol{\mu}}^*_\alpha)$ in (8) by $D(\lambda)$ and $D^*(\alpha)$, respectively.

Lemma 1. Suppose the density function of the generalized linear model satisfies Fan and Li's (2001) three regularity conditions (A)–(C). Assume that there exists a positive constant M such that $\kappa_n < M$. Then under Condition (C4), we have

$$P\Big\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\lambda) > \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha})\Big\} \to 1 \tag{A.1}$$

and

$$\liminf_{n\to\infty} P\Big\{\inf_{\lambda\in\Omega_0} \mathrm{GIC}_{\kappa_n}(\lambda) > \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha})\Big\} \geq \pi, \tag{A.2}$$

as $n \to \infty$.

Proof. For a given λ, the nonpenalized maximum likelihood estimator, $\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}$, maximizes $\ell(\boldsymbol{\beta})$ under model α_λ. This implies $D(\lambda) \geq D^*(\alpha_\lambda)$ for any given λ, which leads to

$$\mathrm{GIC}_{\kappa_n}(\lambda) = D(\lambda)/n + \kappa_n df_\lambda/n > D^*(\alpha_\lambda)/n. \tag{A.3}$$

Then, we obtain that

$$\mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha}) > \frac{D^*(\alpha_\lambda)}{n} - \frac{D^*(\bar{\alpha})}{n} - \frac{\kappa_n d_{\bar{\alpha}}}{n},$$

holds true for any $\lambda \in \Omega_- = \{\lambda : \alpha_\lambda \not\supset \alpha_0\}$. This together with Condition (C4) and $\kappa_n d_{\bar{\alpha}}/n = o_P(1)$ obtained from the fact of $\kappa_n < M$, we have

$$\begin{aligned} &P\Big\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha}) > 0\Big\} \\ &\quad\geq P\Big\{\min_{\alpha\not\supset\alpha_0} \frac{D^*(\alpha)}{n} - \frac{D^*(\bar{\alpha})}{n} - \frac{\kappa_n d_{\bar{\alpha}}}{n} > 0\Big\} \\ &\quad= P\Big\{\min_{\alpha\not\supset\alpha_0} c_\alpha - c_{\bar{\alpha}} + o_P(1) > 0\Big\} \to 1, \end{aligned} \tag{A.4}$$

as $n \to \infty$. The last step uses the fact that the number of underfitting models is finite, hence $\min_{\alpha\not\supset\alpha_0} c_\alpha$ is strictly greater than $c_{\bar{\alpha}}$ (i.e., c_{α_0}) under Condition (C4). The above equation yields (A.1) immediately.

Next we show that the model selected by the GIC selector with bounded κ_n is overfitted with probability bounded away from zero. For any $\lambda \in \Omega_0$, $\alpha_\lambda = \alpha_0$. Then subtracting $\mathrm{GIC}^*_{\kappa_n}(\bar{\alpha})$ from both sides of (A.3) and subsequently taking $\inf_{\lambda\in\Omega_0}$ over $\mathrm{GIC}_{\kappa_n}(\lambda)$, we have

$$\begin{aligned} &\inf_{\lambda\in\Omega_0} \mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha}) \\ &\qquad> D^*(\alpha_0)/n - [D^*(\bar{\alpha})/n + \kappa_n d_{\bar{\alpha}}/n]. \end{aligned} \tag{A.5}$$

Note that the right-hand side of the above equation does not involve λ.

$D^*(\alpha_0) - D^*(\bar{\alpha}) = -2[\ell(\hat{\boldsymbol{\beta}}^*_{\alpha_0}) - \ell(\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}})]$, where $\hat{\boldsymbol{\beta}}^*_{\alpha_0}$ and $\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}}$ are the nonpenalized maximum likelihood estimators computed via the true and full models, respectively. Under regularity conditions (A)–(C) in Fan and Li (2001), it can be shown that $\hat{\boldsymbol{\beta}}^*_{\alpha_0}$ and $\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}}$ are consistent and asymptotically normal. Hence, the likelihood ratio test statistic $-2[\ell(\hat{\boldsymbol{\beta}}^*_{\alpha_0}) - \ell(\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}})] \xrightarrow{\mathcal{L}} \chi^2_{d_{\bar{\alpha}}-d_{\alpha_0}}$. This result, together with $\kappa_n < M$ and (A.5), yields

$$\begin{aligned} &P\Big\{\inf_{\lambda\in\Omega_0} \mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha}) > 0\Big\} \\ &\quad\geq P\Big\{-\frac{2}{n}\big[\ell\big(\hat{\boldsymbol{\beta}}^*_{\alpha_0}\big) - \ell(\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}})\big] - \kappa_n d_{\bar{\alpha}}/n > 0\Big\} \\ &\quad\geq P\big\{-2\big[\ell\big(\hat{\boldsymbol{\beta}}^*_{\alpha_0}\big) - \ell(\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}})\big] > d_{\bar{\alpha}} M\big\} \\ &\quad\to P\big\{\chi^2_{d_{\bar{\alpha}}-d_{\alpha_0}} \geq d_{\bar{\alpha}} M\big\} \triangleq \pi. \end{aligned}$$

This implies (A.2), and we complete the proof of Lemma 1.

To prove Theorem 1(B), we present the following lemma.

Lemma 2. Suppose the density function of the generalized linear model satisfies Fan and Li's (2001) three regularity conditions (A)–(C). Assume Conditions (C1)–(C4) hold, and let $\lambda_n = \kappa_n/\sqrt{n}$. If κ_n satisfies $\kappa_n \to \infty$ and $\lambda_n \to 0$ as $n \to \infty$, then

$$P\big\{\mathrm{GIC}_{\kappa_n}(\lambda_n) = \mathrm{GIC}^*_{\kappa_n}(\alpha_0)\big\} \to 1 \tag{A.6}$$

and

$$P\Big\{\inf_{\lambda\in\Omega_-\cup\Omega_+} \mathrm{GIC}_{\kappa_n}(\lambda) > \mathrm{GIC}_{\kappa_n}(\lambda_n)\Big\} \to 1. \tag{A.7}$$

Proof. Without loss of generality, we assume that the first d_{α_0} coefficients of $\boldsymbol{\beta}_{\alpha_0}$ in the true model are nonzero and the rest are zeros. Note that the density function of the generalized linear model satisfies Fan and Li's (2001) three regularity conditions (A)–(C). These conditions, together with Condition (C3) and the assumptions of κ_n stated in Lemma 2, allow us to apply Fan and Li's theorems 1 and 2 to show that, with probability tending to 1, the last $d - d_{\alpha_0}$ components of $\hat{\boldsymbol{\beta}}_{\lambda_n}$ are zeros and the first d_{α_0} components of $\hat{\boldsymbol{\beta}}_{\lambda_n}$ satisfy the normal equations

$$\frac{\partial}{\partial\beta_j}\ell\big(\hat{\boldsymbol{\beta}}_{\lambda_n}\big) + b_{\lambda_n j} = 0 \quad \text{for } j = 1, \ldots, d_{\alpha_0}, \tag{A.8}$$

where $b_{\lambda_n j} = p'_{\lambda_n}(|\hat{\beta}_{\lambda_n j}|)\operatorname{sgn}(\hat{\beta}_{\lambda_n j})$, and $\hat{\beta}_{\lambda_n j}$ is the jth component of $\hat{\boldsymbol{\beta}}_{\lambda_n}$.

Using the oracle property, we have $|\hat{\beta}_{\lambda_n j}| \to |\beta_j| \geq \min_{1\leq j\leq d_{\alpha_0}} |\beta_j|$. Then, under Conditions (C1) and (C2), there exists a constant m so that $p'_{\lambda_n}(|\hat{\beta}_{\lambda_n j}|) = 0$ for $\min_{1\leq j\leq d_{\alpha_0}} |\beta_j| > m\lambda_n$ as n gets large. Accordingly, $P(b_{\lambda_n j} = 0) \to 1$ for $j = 1, \ldots, d_{\alpha_0}$. This together with (A.8) implies that, with probability tending to 1, the first d_{α_0} components of $\hat{\boldsymbol{\beta}}_{\lambda_n}$ solve the normal equations

$$\frac{\partial}{\partial \beta_j}\ell\big(\hat{\boldsymbol{\beta}}_{\lambda_n}\big) = 0, \qquad j = 1, \ldots, d_{\alpha_0},$$

and the remaining $d - d_{\alpha_0}$ components are zeros. This is exactly the same as the normal equation in solving the nonpenalized maximum likelihood estimator $\hat{\boldsymbol{\beta}}^*_{\alpha_0}$. As a result, $\hat{\boldsymbol{\beta}}^*_{\alpha_0} = \hat{\boldsymbol{\beta}}_{\lambda_n}$ with probability tending to 1. It follows that

$$P\{D(\lambda_n) = D^*(\alpha_0)\} = P\big\{\ell\big(\hat{\boldsymbol{\beta}}_{\lambda_n}\big) = \ell\big(\hat{\boldsymbol{\beta}}^*_{\alpha_0}\big)\big\} \to 1.$$

Moreover, using the result from Proposition 1, we have $P\{df_{\lambda_n} = d_{\alpha_0}\} \to 1$. Consequently,

$$\begin{aligned} &P\big\{\mathrm{GIC}_{\kappa_n}(\lambda_n) = \mathrm{GIC}^*_{\kappa_n}(\alpha_0)\big\} \\ &\quad = P\left\{\frac{1}{n}(D(\lambda) - D^*(\alpha_0)) + \frac{\kappa_n}{n}\big(df_{\lambda_n} - d_{\alpha_0}\big) = 0\right\} \\ &\quad \to 1. \end{aligned}$$

The proof of (A.6) is complete.

We next show that, for any λ which cannot identify the true model, the resulting $\mathrm{GIC}_{\kappa_n}(\lambda)$ is consistently larger than $\mathrm{GIC}_{\kappa_n}(\lambda_n)$. To this end, we consider two cases, underfitting and overfitting.

Case 1: Underfitted model (i.e., $\lambda \in \Omega_-$ so that $\alpha_\lambda \not\supset \alpha_0$). Applying (A.3) and (A.6), we obtain that, with probability tending to 1,

$$\mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}_{\kappa_n}(\lambda_n) > \frac{1}{n}D^*(\alpha_\lambda) - \frac{1}{n}D^*(\alpha_0) - \frac{\kappa_n}{n}d_{\alpha_0}.$$

Subsequently, take $\inf_{\lambda\in\Omega_-}$ of both sides of the above equation, which yields

$$\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}_{\kappa_n}(\lambda_n) > \min_{\alpha\not\supset\alpha_0} \frac{1}{n}D^*(\alpha) - \frac{1}{n}D^*(\alpha_0) - \frac{\kappa_n}{n}d_{\alpha_0}. \tag{A.9}$$

This, in conjunction with Condition (C4) and the number of underfitted models being finite, leads to

$$\begin{aligned} &P\left\{\min_{\alpha\not\supset\alpha_0} \frac{1}{n}D^*(\alpha) - \frac{1}{n}D^*(\alpha_0) - \frac{\kappa_n}{n}d_{\alpha_0} > 0\right\} \\ &\quad = P\left\{\min_{\alpha\not\supset\alpha_0} c_\alpha - c_{\alpha_0} + o_P(1) > 0\right\} \\ &\quad \to 1, \end{aligned}$$

as $n \to \infty$. As a result,

$$P\left\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\lambda) > \mathrm{GIC}_{\kappa_n}(\lambda_n)\right\} \to 1.$$

Case 2: Overfitted model (i.e., $\lambda \in \Omega_+$ so that $\alpha_\lambda \supsetneq \alpha_0$). According to Lemma 1, with probability tending to 1, $D(\lambda_n) = D^*(\alpha_0)$. In addition, Proposition 1 indicates that $df_\lambda - df_{\lambda_n} > \tau + o_P(1)$ for some $\tau > 0$ and $\lambda \in \Omega_+$. Moreover, as noticed in the proof of Lemma 1, $D(\lambda) \geq D^*(\alpha_\lambda)$. Therefore, with probability tending to 1,

$$\begin{aligned} &n\big(\mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}_{\kappa_n}(\lambda_n)\big) \\ &\quad = D(\lambda) - D(\lambda_n) + \big(df_\lambda - df_{\lambda_n}\big)\kappa_n \\ &\quad \geq D^*(\alpha_\lambda) - D^*(\alpha_0) + (\tau + o_P(1))\kappa_n. \end{aligned}$$

Because $D^*(\alpha_0) - D^*(\alpha)$ follows a χ^2 distribution with $d_\alpha - d_{\alpha_0}$ degrees of freedom asymptotically for any $\alpha \supsetneq \alpha_0$, it is of order $O_P(1)$. Accordingly,

$$\begin{aligned} &\inf_{\lambda\in\Omega_+} n\big(\mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}_{\kappa_n}(\lambda_n)\big) \\ &\quad \geq \min_{\alpha\supsetneq\alpha_0} \{D^*(\alpha_0) - D^*(\alpha)\} + (\tau + o_P(1))\kappa_n \\ &\quad \approx \tau\kappa_n. \end{aligned} \tag{A.10}$$

Using the fact that κ_n goes to infinity as $n \to \infty$, the right-hand side of Equation (A.10) goes to positive infinity, which guarantees that the left-hand side of Equation (A.10) is positive as $n \to \infty$. Hence, we finally have

$$P\left\{\inf_{\lambda\in\Omega_+} \mathrm{GIC}_{\kappa_n}(\lambda) > \mathrm{GIC}_{\kappa_n}(\lambda_n)\right\} \to 1.$$

The results of Cases 1 and 2 complete the proof of (A.7).

Proofs of Theorem 1

Lemma 1 implies that $\mathrm{GIC}_{\kappa_n}(\lambda)$, which produces the underfitted model, is consistently larger than $\mathrm{GIC}^*_{\kappa_n}(\bar{\alpha})$. Thus, the optimal model selected by minimizing the $\mathrm{GIC}_{\kappa_n}(\lambda)$ must contain all of the significant variables with probability tending to one. In addition, Lemma 1 indicates that there is a nonzero probability that the smallest value of $\mathrm{GIC}_{\kappa_n}(\lambda)$ for $\lambda \in \Omega_0$ is larger than that of the full model. As a result, there is a positive probability that any λ associated with the true model cannot be selected by $\mathrm{GIC}_{\kappa_n}(\lambda)$ as the regularization parameter. Theorem 1(A) follows.

Lemma 2 indicates that the model identified by λ_n converges to the true model as the sample size gets large. In addition, it shows λ's that fail to identify the true model cannot be selected by $\mathrm{GIC}_{\kappa_n}(\lambda)$ asymptotically. Theorem 1(B) follows.

APPENDIX B: PROOFS OF THEOREM 2 AND COROLLARY 1

Before proving the theorem, we establish the following two lemmas. Lemma 3 evaluates the difference between a penalized mean estimator $\hat{\boldsymbol{\mu}}_\lambda$ and its corresponding least squares mean estimator $\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}$, while Lemma 4 demonstrates that the losses of $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}$ are asymptotically equivalent.

Lemma 3. Under Condition (C5),

$$\big\|\hat{\boldsymbol{\mu}}_\lambda - \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}\big\|^2 \leq nC\|\mathbf{b}\|^2,$$

where C is the constant number in Condition (C5) and $\mathbf{b}$ is defined in Condition (C8).

Proof. Without loss of generality, we assume that the first d_{α_λ} components of $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}$ are nonzero, and denote them by $\hat{\boldsymbol{\beta}}^{(1)}_\lambda$ and $\hat{\boldsymbol{\beta}}^{*(1)}_{\alpha_\lambda}$, respectively. Thus, $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda = \mathbf{X}_{\alpha_\lambda}\hat{\boldsymbol{\beta}}^{(1)}_\lambda$ and $\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} = \mathbf{X}\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda} = \mathbf{X}_{\alpha_\lambda}\hat{\boldsymbol{\beta}}^{*(1)}_{\alpha_\lambda}$. From the proofs of theorems 1 and 2 in Fan and Li (2001), with probability tending to 1, we have that $\hat{\boldsymbol{\beta}}^{(1)}_\lambda$ is the solution of the following equation:

$$\frac{1}{n}\mathbf{X}^T_{\alpha_\lambda}\big(\mathbf{y} - \mathbf{X}_{\alpha_\lambda}\boldsymbol{\beta}^{(1)}_\lambda\big) + \mathbf{b}^{(1)} = \mathbf{0},$$

where $\mathbf{b}^{(1)}$ is the subvector of $\mathbf{b}$ that corresponds to $\hat{\boldsymbol{\beta}}^{(1)}_\lambda$. Accordingly,

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(1)}_\lambda &= \big(\mathbf{X}^T_{\alpha_\lambda}\mathbf{X}_{\alpha_\lambda}\big)^{-1}\mathbf{X}^T_{\alpha_\lambda}\mathbf{y} + \left(\frac{1}{n}\mathbf{X}^T_{\alpha_\lambda}\mathbf{X}_{\alpha_\lambda}\right)^{-1}\mathbf{b}^{(1)} \\ &= \hat{\boldsymbol{\beta}}^{*(1)}_{\alpha_\lambda} + \mathbf{V}_{\alpha_\lambda}\mathbf{b}^{(1)}, \end{aligned}$$

where $\mathbf{V}_{\alpha_\lambda} \triangleq (\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^T\mathbf{X}_{\alpha_\lambda})^{-1}$. In addition, the eigenvalues of $\mathbf{V}_{\alpha_\lambda}$ are bounded under Condition (C5). Hence,

$$\|\hat{\boldsymbol{\mu}}_\lambda - \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}\|^2 = \|\mathbf{X}_{\alpha_\lambda}(\hat{\boldsymbol{\beta}}^{(1)}_\lambda - \hat{\boldsymbol{\beta}}^{*(1)}_{\alpha_\lambda})\|^2 = n\mathbf{b}_1^T\mathbf{V}_{\alpha_\lambda}\mathbf{b}^{(1)} \le nC\|\mathbf{b}\|^2$$

for some positive constant number C. This completes the proof.

Lemma 4. If Conditions (C5)–(C8) hold, then

$$\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} - 1\right| \to 0,$$

in probability.

Proof. After algebraic simplification, we have

$$\begin{aligned} L(\hat{\boldsymbol{\beta}}_\lambda) - L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}) &= \frac{\|\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda\|^2}{n} + \frac{2(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda})^T(\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda)}{n} \\ &= I_1 + I_2. \end{aligned}$$

Under Conditions (C6) and (C7), Li (1987) showed that

$$\sup_{\alpha\in\mathcal{A}}\left|\frac{L(\hat{\boldsymbol{\beta}}^*_\alpha)}{R(\hat{\boldsymbol{\beta}}^*_\alpha)} - 1\right| \to 0.$$

This, together with Condition (C8) and Lemma 3, implies

$$\begin{aligned} &\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{I_1}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\right| \\ &= \sup_{\lambda\in[0,\lambda_{\max}]}\left\{\frac{\|\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda\|^2}{nR(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} - \frac{\|\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda\|^2}{nL(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\left[\frac{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} - 1\right]\right\} \\ &\to 0. \end{aligned}$$

Applying the Cauchy–Schwarz inequality, we next obtain

$$\begin{aligned} I_2 &\le \frac{2\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}\| \cdot \|\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda\|}{n} \\ &= 2\sqrt{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \cdot \frac{1}{\sqrt{n}}\|\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda\|. \end{aligned}$$

As a result, $\sup_{\lambda\in[0,\lambda_{\max}]}|\frac{I_2}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}| \to 0$, and Lemma 4 follows immediately.

Proof of Theorem 2

To show the asymptotic efficiency of the AIC-type selector, it suffices to demonstrate that minimizing $\mathrm{GIC}^{LS}_{\kappa_n}(\lambda)$ with $\kappa_n \to 2$ is the same as minimizing $L(\hat{\boldsymbol{\beta}}_\lambda)$ asymptotically. To this end, we need to prove that, in probability,

$$\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{\mathrm{GIC}^{LS}_{\kappa_n}(\lambda) - \|\boldsymbol{\epsilon}\|^2/n - L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right| \to 0. \tag{B.1}$$

Let the projection matrix corresponding to the model α be $\mathbf{H}_\alpha = \mathbf{X}_\alpha(\mathbf{X}_\alpha^T\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha^T$. Then,

$$\begin{aligned} \mathrm{GIC}^{LS}_{\kappa_n}(\lambda) &= \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2}{n} + \frac{\kappa_n\sigma^2 d_{\alpha_\lambda}}{n} \\ &= \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}\|^2}{n} + \frac{\|\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda\|^2}{n} + \frac{\kappa_n\sigma^2 d_{\alpha_\lambda}}{n} \\ &= \frac{\|\boldsymbol{\epsilon}\|^2}{n} + L(\hat{\boldsymbol{\beta}}_\lambda) + \big[L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}) - L(\hat{\boldsymbol{\beta}}_\lambda)\big] + \frac{1}{n}\|\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda\|^2 \\ &\quad + \frac{2}{n}\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbf{H}_{\alpha_\lambda})\boldsymbol{\mu} \\ &\quad + \frac{2}{n}\big(\sigma^2 d_{\alpha_\lambda} - \boldsymbol{\epsilon}^T\mathbf{H}_{\alpha_\lambda}\boldsymbol{\epsilon}\big) + \frac{1}{n}(\kappa_n - 2)\sigma^2 d_{\alpha_\lambda}. \end{aligned} \tag{B.2}$$

Let $J_1 = L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}) - L(\hat{\boldsymbol{\beta}}_\lambda)$, $J_2 = \|\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda\|^2/n$, $J_3 = 2\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbf{H}_{\alpha_\lambda})\boldsymbol{\mu}/n$, $J_4 = 2(\sigma^2 d_{\alpha_\lambda} - \boldsymbol{\epsilon}^T\mathbf{H}_{\alpha_\lambda}\boldsymbol{\epsilon})/n$ and $J_5 = (\kappa_n - 2)\sigma^2 d_{\alpha_\lambda}/n$. Using Lemma 3, Lemma 4, and similar arguments used in Li (1987), we obtain that, in probability,

$$\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{J_j}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right| \to 0 \quad \text{for } j = 1,\ldots,4.$$

Because $\kappa_n \to 2$ and the fact that $n^{-1}\sigma^2 d_{\alpha_\lambda}/R(\hat{\boldsymbol{\beta}}_\lambda)$ is bounded by 1, we can further show that $\sup_{\lambda\in[0,\lambda_{\max}]}|J_5/L(\hat{\boldsymbol{\beta}}_\lambda)| \to 0$ using Lemma 4 and Condition (C7). Accordingly, (B.1) holds, which implies that the difference between $\mathrm{GIC}^{LS}_{\kappa_n}(\lambda) - \frac{\|\boldsymbol{\epsilon}\|^2}{n}$ and $L(\hat{\boldsymbol{\beta}}_\lambda)$ is negligible in comparison to $L(\hat{\boldsymbol{\beta}}_\lambda)$. This completes the proof.

Proof of Corollary 1

When σ^2 is unknown, the $\mathrm{GIC}^{LS}_{\kappa_n}(\lambda)$ in (B.2) becomes

$$\begin{aligned} \mathrm{GIC}^{LS}_{\kappa_n}(\lambda) &= \frac{\|\boldsymbol{\epsilon}\|^2}{n} + L(\hat{\boldsymbol{\beta}}_\lambda) \\ &\quad + J_1 + J_2 + J_3 + J_4 + J_5 + \frac{2(\tilde{\sigma}^2 - \sigma^2)d_{\alpha_\lambda}}{n}. \end{aligned}$$

Using Lemma 4 and Condition (C7), we have

$$\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{2(\tilde{\sigma}^2 - \sigma^2)d_{\alpha_\lambda}}{nL(\hat{\boldsymbol{\beta}}_\lambda)}\right| \to 0$$

in probability, which completes the proof.

APPENDIX C: SUFFICIENT CONDITIONS FOR (C8)

We first present three sufficient conditions for (C8).

(S1) There exists a constant M_1 such that $\lambda_{\max}$ satisfies $\sqrt{n}\lambda_{\max} < M_1$ for all n.

(S2) There exists a constant M_2 such that $p_\lambda(\theta)$ satisfies $p'_\lambda(\theta) \le M_2\lambda$ for any θ.

(S3) The average error of the full model $\bar{\alpha}$, that is, $\Delta_{\bar{\alpha}} \triangleq \|\boldsymbol{\mu} - \mathbf{H}_{\bar{\alpha}}\boldsymbol{\mu}\|^2/n$, satisfies $\frac{n\Delta_{\bar{\alpha}}}{d} \to \infty$, as $n \to \infty$.

We next provide motivations for these conditions. Assume that the true model is $\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}_0 = \sum_{j=1}^d x_{ij}\beta_{j0}$ for $i = 1,\ldots,n$, and then write $a_n = \max_{1\le j\le d}\{p'_{\lambda_n}(|\beta_{j0}|), \beta_{j0} \ne 0\}$. Under the condition $d = O(n^\nu)$ with $\nu < \frac{1}{4}$, Fan and Peng (2004) proved that there exists a local maximum of the penalized least squares function, so that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P\{\sqrt{d}(n^{-1/2} + a_n)\}$. Thus, Conditions (S1) and (S2) ensure that the penalized estimator is $\sqrt{n/d}$-consistent. As for Condition (S3), consider a case where the full model misses an important variable x_{exc} that is orthogonal to the rest of the variables. Accordingly, $\Delta_{\bar{\alpha}} = n^{-1}\sum_{i=1}^n x_{\mathrm{exc};i}^2\beta_{\mathrm{exc}}^2$, and hence $n\Delta_{\bar{\alpha}}/d$ is of the same order as $n\beta_{\mathrm{exc}}^2/d$. Consequently, (S3) holds if the coefficient β_{exc} satisfies $\sqrt{n/d}\beta_{\mathrm{exc}} \to \infty$, which is valid, for example, when β_{exc} is fixed.

In contrast to imposing Condition (S3) on the error of the full model, an alternative Condition (S3*) motivated by Fan and Peng (2004) is given below.

(S3*) The proportion of penalized coefficients satisfies $\sup_{\lambda\in[0,\lambda_{\max}]}\frac{1}{d_{\alpha_\lambda}}\sum_{j=1}^d I(0 < |\hat{\beta}_{\lambda j}| \le m\lambda) \to 0$, as $n \to \infty$, where m is defined in (C2).

Condition (S3*) means that the proportion of small or moderate size coefficients ($0 < \sqrt{n}|\hat{\beta}_{\lambda j}| \le mM_1$) is vanishing. This condition is satisfied under the identifiability assumption of Fan and Peng (2004) (i.e., $\min_{j:\beta_{0j}\ne 0}|\beta_{0j}|/\lambda \to \infty$, as $n \to \infty$). As a byproduct, a similar condition can also be established for the adaptive LASSO, which employs data-driven penalties with $b_j = \lambda/(\hat{\beta}^*_{\tilde{\alpha}j})^k$ for some $k > 0$ to cope with

the bias of LASSO. In sum, (S1) to (S3) (or S3*) are mild conditions, and two propositions given below demonstrate their sufficiency for Condition (C8).

Proposition 2. Under (S1) to (S3), Condition (C8) holds.

Proof. It is noteworthy that

$$R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}) = \Delta_{\alpha_\lambda} + \frac{d_{\alpha_\lambda}\sigma^2}{n} \geq \Delta_{\bar{\alpha}} + \frac{d_{\alpha_\lambda}\sigma^2}{n}. \tag{B.1}$$

On the right-hand side of (B.1), the first term dominates the second term via Condition (S3). From (S1), (S2), and (B.1), we have that

$$\frac{\|\mathbf{b}\|^2}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \leq \frac{M_2^2\lambda^2 d_{\alpha_\lambda}}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \leq M_2^2 \cdot n\lambda^2_{\max} \cdot \frac{1}{n\Delta_{\bar{\alpha}}/d},$$

which goes to zero independently of λ. This completes the proof.

Proposition 3. Assume that the penalty function satisfies Condition (C2). Under (S1) to (S3*), Condition (C8) holds.

Proof. Under Condition (C2), the components in $\mathbf{b}$ are zero except for those $\hat{\beta}_{\lambda j} \leq m\lambda$. Then, employing (S2) and (B.1), we obtain that

$$\frac{\|\mathbf{b}\|^2}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \leq \frac{\|\mathbf{b}\|^2}{n^{-1}d_{\alpha_\lambda}\sigma^2} \leq \frac{M_2^2}{\sigma^2} \cdot n\lambda^2_{\max} \cdot \frac{1}{d_{\alpha_\lambda}} \sum_{j=1}^{d} I(0 < |\hat{\beta}_{\lambda j}| \leq m\lambda).$$

On the right-hand side of the above equation, the first term is constant, the second term is bounded under (S1), and the third term goes to zero uniformly in λ under (S3*). This completes the proof.

SUPPLEMENTAL MATERIALS

Technical details for Remark 5: Detailed discussions, justifications, and proofs referenced in Remark 5. (suplv_rz_t.pdf)

[Received January 2008. Revised October 2009.]

REFERENCES

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723. [312]

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 377–403. [312]

Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [312,314]

Elter, M., Schulz-Wendtland, R., and Wittenberg, T. (2007), "The Prediction of Breast Cancer Biopsy Outcomes Using Two CAD Approaches That Both Emphasize an Intelligible Decision Process," *Medical Physics*, 34, 4164–4172. [319]

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. [312-315,317,320,321]

——— (2002), "Variable Selection for Cox's Proportional Hazards Model and Fratile Model," *The Annals of Statistics*, 30, 74–99. [312,314]

——— (2004), "New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis," *Journal of the American Statistical Association*, 99, 710–723. [312]

Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood With a Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961. [312,322]

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–148. [313]

Hunter, D., and Li, R. (2005), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617–1642. [312]

Konishi, S., and Kitagawa, G. (1996), "Generalised Information Criteria in Model Selection," *Biometrika*, 83, 875–890. [313]

Li, K.-C. (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975. [312,313,315,316,322]

Li, R., and Liang, H. (2008), "Variable Selection in Semiparametric Regression Modeling," *The Annals of Statistics*, 36, 261–286. [312]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman & Hall/CRC. [314,315]

McQuarrie, A. D. R., and Tsai, C.-L. (1998), *Regression and Time Series Model Selection* (1st ed.), Singapore: World Scientific Publishing Co., Pte. Ltd. [312,316]

Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics*, 12, 758–765. [312-314]

Park, M.-Y., and Hastie, T. (2007), "An L_1 Regularization-Path Algorithm for Generalized Linear Models," *Journal of the Royal Statistical Society, Ser. B*, 69, 659–677. [312]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 19 (2), 461–464. [312]

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7, 221–264. [312,313,316]

Shibata, R. (1980), "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *The Annals of Statistics*, 8, 147–164. [316]

——— (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45–54. [312,313,315,316]

——— (1984), "Approximation Efficiency of a Selection Procedure for the Number of Regression Variables," *Biometrika*, 71, 43–49. [312,315]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via LASSO," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [312,313,317]

Wang, H., and Leng, C. (2007), "Unified LASSO Estimation via Least Squares Approximation," *Journal of the American Statistical Association*, 102, 1039–1048. [312]

Wang, H., Li, G., and Tsai, C.-L. (2007a), "Regression Coefficient and Autoregressive Order Shrinkage and Selection via LASSO," *Journal of the Royal Statistical Society, Ser. B*, 69, 63–78. [312]

Wang, H., Li, R., and Tsai, C.-L. (2007b), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [312,315]

Yang, Y. (2005), "Can the Strengths of AIC and BIC Be Shared? A Conflict Between Model Identification and Regression Estimation," *Biometrika*, 92, 937–950. [316]

Yuan, M., and Lin, Y. (2007), "On the Non-Negative Garrotte Estimator," *Journal of the Royal Statistical Society, Ser. B*, 69, 143–161. [312]

Zhang, C.-H. (2007), "Penalized Linear Unbiased Selection," Technical Report 2007-003, Rutgers University, Dept. of Statistics. [313]

Zhang, H. H., and Lu, W. (2007), "Adaptive LASSO for Cox's Proportional Hazards Model," *Biometrika*, 94, 691–703. [312]

Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [312]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533. [312]

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the Degrees of Freedom of the LASSO," *The Annals of Statistics*, 35, 2173–2192. [314]