Vol. 37 (2017)
No. 4

数 学 杂 志
J. of Math. (PRC)

A MODIFIED BIC TUNING PARAMETER SELECTOR FOR SICA-PENALIZED COX REGRESSION MODELS WITH DIVERGING DIMENSIONALITY

SHI Yue-yong[1,3], JIAO Yu-ling[2], YAN Liang[1], CAO Yong-xiu[2]

(*1.School of Economics and Management, China University of Geosciences, Wuhan 430074, China*)

(*2.School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China*)

(*3.Center for Resources and Environmental Economic Research, China University of Geosciences, Wuhan 430074, China*)

Abstract: This paper proposes a modified BIC (Bayesian information criterion) tuning parameter selector for SICA-penalized Cox regression models with a diverging number of covariates. Under some regularity conditions, we prove the model selection consistency of the proposed method. Numerical results show that the proposed method performs better than the GCV (generalized cross-validation) criterion.

Keywords: Cox models; modified BIC; penalized likelihood; SICA penalty; smoothing quasi-Newton

2010 MR Subject Classification: 62N01; 62N02

Document code: A **Article ID:** 0255-7797(2017)04-0723-08

1 Introduction

The commonly used Cox model [4] for survival data assumes that the hazard function $h(t|\mathbf{z})$ for the failure time T associated with covariates $\mathbf{z} = (z_1, \cdots, z_d)^T$ takes the form

$$h(t|\mathbf{z}) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{z}), \tag{1.1}$$

where t is the time, $h_0(t)$ is an arbitrary unspecified baseline Hazard function and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T$ is an unknown vector of regression coefficients. In this paper, we consider the following so-called SICA-penalized log partial likelihood (SPPL) problem

$$\hat{\boldsymbol{\beta}} := \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^d}\{Q_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_{\lambda,\tau}(\beta_j)\}, \tag{1.2}$$

* **Received date:** 2016-11-10 **Accepted date:** 2016-12-20

Foundation item: Supported by National Natural Science Foundation of China (11501579); Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (CUGW150809).

Biography: Shi Yueyong (1984–), male, born at Luzhou, Sichuan, lecturer, major in biostatistics.

Corresponding author: Cao Yongxiu.

where

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \{\boldsymbol{\beta}^T \mathbf{z}_i - \log[\sum_{j=1}^{n} Y_j(\tilde{T}_i) \exp(\boldsymbol{\beta}^T \mathbf{z}_j)]\} \tag{1.3}$$

is the logarithm of the partial likelihood function, $\tilde{T}_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, $Y_i(t) = I(\tilde{T}_i \geq t)$, and T_i and C_i are the failure time and censoring time of subject $i\,(i = 1, \cdots, n)$, respectively; $p_{\lambda,\tau}(\beta_j) = \lambda(\tau+1)|\beta_j|/(|\beta_j| + \tau)$ is the SICA penalty function proposed by Lv and Fan [9], and λ and τ are two positive tuning (or regularization) parameters. In particular, λ is the sparsity tuning parameter obtaining sparse solutions and τ is the shape (or concavity) tuning parameter making SICA a bridge between L_0 ($\tau \to 0+$) and L_1 ($\tau \to \infty$), where L_0 and L_1 admit $p_\lambda(\beta_j) = \lambda I(|\beta_j| \neq 0)$ and $p_\lambda(\beta_j) = \lambda|\beta_j|$, respectively. $\hat{\boldsymbol{\beta}}$, which is dependent on λ and τ, i.e., $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda, \tau)$, is denoted as a SPPL estimator.

Although penalized likelihood methods can select variables and estimate coefficients simultaneously, their optimal properties heavily depend on an appropriate selection of the tuning parameters. Thus, an important issue in variable selection using penalized likelihood methods is the choice of tuning parameters. Some common used tuning parameter selection criteria are GCV [1, 6, 8, 13], AIC [17] and BIC [14, 15].

Shi et al. [12] proposed using the SPPL approach combined with a GCV tuning parameter selector for variable selection in Cox's proportional hazards model with diverging dimensionality. As shown in Wang et al. [14, 15] in the linear model case, it is known that GCV tends to over-fit the true model and BIC can identify the true model consistently. Thus, when the primary goal is variable selection and identification of the true model, BIC may be preferred over GCV. In this paper, in the context of right-censored data, we modify the classical BIC to select tuning parameters for (1.2) and prove its consistency when the number of regression coefficients tends to infinity. Simulation studies are given to illustrate the performance of the proposed approach.

An outline for this paper is as follows. In Section 2, we first describe the Modified BIC method for SPPL and then give theoretical results and corresponding proofs. The finite sample performance of the proposed method through simulation studies are demonstrated in Section 3. We conclude the paper with Section 4.

2 Modified BIC (MBIC) for SPPL

2.1 Methodology

The ordinary BIC procedure is implemented by minimizing

$$\mathrm{BIC} = \mathrm{BIC}(\hat{\boldsymbol{\beta}}) = -2l_n(\hat{\boldsymbol{\beta}}) + \log(n)\widehat{\mathrm{DF}}, \tag{2.1}$$

where $\widehat{\mathrm{DF}}$ is an estimator of the degrees of freedom corresponding to $\hat{\boldsymbol{\beta}}$. Motivated by [17], we take $\widehat{\mathrm{DF}} = \|\hat{\boldsymbol{\beta}}\|_0 = |\{j : \hat{\beta}_j \neq 0\}| \triangleq \hat{d}$. In order to account for a diverging number of parameters, enlightened by [5], we propose a modified BIC (MBIC) minimizing

$$\mathrm{MBIC}_{k_n} = \mathrm{MBIC}_{k_n}(\hat{\boldsymbol{\beta}}) = -2l_n(\hat{\boldsymbol{\beta}}) + k_n\hat{d}, \tag{2.2}$$

where k_n is a positive number that depends on the sample size n with $k_n > \log(n)$.

2.2 Theoretical Results

Without loss of generality, we write the true parameter vector as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$, where $\boldsymbol{\beta}_{10}$ consists of all s nonzero components and $\boldsymbol{\beta}_{20}$ consists of the remaining zero components. Correspondingly, we write the maximizer of (1.2) as $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$. Define $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$ and $\widehat{\mathcal{A}} = \{j : \hat{\beta}_j \neq 0\}$. Hereafter, sometimes we use d_n, s_n, λ_n and τ_n rather than d, s, λ and τ to emphasize their dependence on n. The regularity conditions (C1)–(C7) in [12] are assumed in the following theoretical results.

Theorem 1 (Existence of SPPL estimator) Under conditions (C1)–(C7) in [12], with probability tending to one, there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q_n(\boldsymbol{\beta})$, defined in (1.2), such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(\sqrt{d_n/n})$, where $\|\cdot\|_2$ is the L_2 norm on the Euclidean space.

Theorem 2 (Oracle property) Under conditions (C1)–(C7) in [12], with probability tending to 1, the $\sqrt{n/d_n}$-consistent local maximizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ in Theorem 1 must be such that

(i) (Sparsity) $\hat{\boldsymbol{\beta}}_2 = 0$;

(ii) (Asymptotic normality) For any nonzero constant $s_n \times 1$ vector c_n with $c_n^T c_n = 1$,

$$\sqrt{n} c_n^T \Gamma_{11}^{-\frac{1}{2}} A_{11} \{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\} \to N(0, 1) \tag{2.3}$$

in distribution, where A_{11} and Γ_{11} consist of the first s_n columns and rows of $A(\boldsymbol{\beta}_{10}, \mathbf{0})$ and $\Gamma(\boldsymbol{\beta}_{10}, \mathbf{0})$ respectively, and $A(\boldsymbol{\beta})$ and $\Gamma(\boldsymbol{\beta})$ are defined in Appendix of [12].

Regularity conditions and detailed proofs for Theorem 1 and Theorem 2 can be found in Appendix of [12]. We now present the main result on the selection consistency of the MBIC under conditions (C1)–(C7) in [12] and an extra condition

(C8) $\rho_n \sqrt{n/(d_n k_n)} \to \infty$ and $d_n/k_n \to 0$ as $n \to \infty$, where $\rho_n = \min_{j \in \mathcal{A}} |\beta_{j0}|$.

Suppose $\Omega \subsetneqq \mathbb{R}^2$. We define $\Omega_- = \{(\lambda, \tau) \in \Omega : \mathcal{A} \not\subset \widehat{\mathcal{A}}\}$, $\Omega_0 = \{(\lambda, \tau) \in \Omega : \mathcal{A} = \widehat{\mathcal{A}}\}$ and $\Omega_+ = \{(\lambda, \tau) \in \Omega : \mathcal{A} \subsetneqq \widehat{\mathcal{A}}\}$. In other words, Ω_0, Ω_- and Ω_+ are three subsets of Ω, in which the true, underfitted and overfitted models can be produced. It easily follows that $\Omega = \Omega_0 \cup \Omega_+ \cup \Omega_-$ (disjoint union) and $\mathcal{A} \neq \widehat{\mathcal{A}} \Leftrightarrow (\lambda, \tau) \in \Omega_- \cup \Omega_+$. Let $\hat{\boldsymbol{\beta}}(\lambda_n^*, \tau_n^*)$ be the local maxima of SPPL described in Theorem 1.

Theorem 3 Under conditions (C1)–(C8),

$$P[\inf_{(\lambda,\tau) \in \Omega_- \cup \Omega_+} \mathrm{MBIC}_{k_n}(\lambda, \tau) > \mathrm{MBIC}_{k_n}(\hat{\boldsymbol{\beta}}(\lambda_n^*, \tau_n^*))] \to 1. \tag{2.4}$$

Proof Since $k_n > \log(n)$, without loss of generality, we assume $k_n > 1$. We prove this theorem by considering two different cases, i.e., underfitting and overfitting.

Case 1 Underfitted model, i.e., $\mathcal{A} \not\subset \widehat{\mathcal{A}}$, which means $\exists j^* \in \mathcal{A}$, $j^* \notin \widehat{\mathcal{A}}$. Let $\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} l_n(\boldsymbol{\beta})$, namely, $\tilde{\boldsymbol{\beta}}$ is the ordinary maximum partial likelihood estimator (MLE). By

the second-order Taylor expansion of the log partial likelihood, we have

$$\begin{aligned}&\mathrm{MBIC}_{k_n}(\hat{\boldsymbol{\beta}}) - \mathrm{MBIC}_{k_n}(\tilde{\boldsymbol{\beta}})\\ =& -2l_n(\hat{\boldsymbol{\beta}}) + k_n\hat{d} + 2l_n(\tilde{\boldsymbol{\beta}}) - k_n d_n = -2[l_n(\hat{\boldsymbol{\beta}}) - l_n(\tilde{\boldsymbol{\beta}})] + k_n(\hat{d} - d_n)\\ =& -2[(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^T l_n'(\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^T l_n''(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})] + k_n(\hat{d} - d_n),\end{aligned}$$

where $\bar{\boldsymbol{\beta}}$ is between $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. Since $\tilde{\boldsymbol{\beta}}$ is MLE, we have $l_n'(\tilde{\boldsymbol{\beta}}) = 0$, and it follows that

$$\mathrm{MBIC}_{k_n}(\hat{\boldsymbol{\beta}}) - \mathrm{MBIC}_{k_n}(\tilde{\boldsymbol{\beta}}) = -(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^T l_n''(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + k_n(\hat{d} - d_n) \triangleq I_1 + I_2.$$

Noting that $-l_n''(\bar{\boldsymbol{\beta}})/n = A(\boldsymbol{\beta}_0) + o_p(1)$, where $A(\boldsymbol{\beta}_0)$ is defined in condition (C3), we have

$$I_1 = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^T\{nA(\boldsymbol{\beta}_0)[1 + o_p(1)]\}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) > nr[1 + o_p(1)]\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_2^2,$$

where $r = \lambda_{\min}\{A(\boldsymbol{\beta}_0)\}$. Since $j^* \notin \widehat{\mathcal{A}}$, we have $\hat{\beta}_{j^*} = 0$. Condition (C6) implies $\rho_n/\alpha_n \to \infty$. Together with $\rho_n = \min_{j\in\mathcal{A}}|\beta_{j0}|$ and $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(\alpha_n)$, we have

$$\begin{aligned}\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_2 &\geq |\hat{\beta}_{j^*} - \tilde{\beta}_{j^*}| = |\tilde{\beta}_{j^*}| = |\tilde{\beta}_{j^*} - \beta_{0j^*} + \beta_{0j^*}| \geq |\beta_{0j^*}| - |\tilde{\beta}_{j^*} - \beta_{0j^*}|\\ &\geq \rho_n - \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = \rho_n - O_p(\alpha_n) = \rho_n[1 - O_p(\alpha_n/\rho_n)] = \rho_n[1 + o_p(1)],\end{aligned}$$

and then we get

$$I_1 > nr\rho_n^2[1 + o_p(1)]. \tag{2.5}$$

Next we consider I_2. It easily follows that

$$I_2 = k_n(\hat{d} - d_n) > -k_n d_n. \tag{2.6}$$

By (2.5), (2.6)) and condition (C8), we have

$$\begin{aligned}\mathrm{MBIC}_{k_n}(\hat{\boldsymbol{\beta}}) - \mathrm{MBIC}_{k_n}(\tilde{\boldsymbol{\beta}}) &> nr\rho_n^2[1 + o_p(1)] - d_n k_n\\ &= d_n k_n\Big(\frac{nr\rho_n^2[1 + o_p(1)]}{d_n k_n} - 1\Big) \xrightarrow{p} \infty,\ n \to \infty,\end{aligned}$$

which yields

$$P[\inf_{(\lambda,\tau)\in\Omega_-} \mathrm{MBIC}_{k_n}(\lambda, \tau) > \mathrm{MBIC}_{k_n}(\tilde{\boldsymbol{\beta}})] \to 1. \tag{2.7}$$

Thus we deduce that the minimum MBIC can not be selected from the underfitted model.

Case 2 Overfitted model, i.e., $\mathcal{A} \subsetneqq \widehat{\mathcal{A}}$, which means $\forall j \in \mathcal{A}$, $j \in \widehat{\mathcal{A}}$, but $\exists j^* \in \widehat{\mathcal{A}}$, $j^* \notin \mathcal{A}$. In this case, we have $\hat{d} > s_n$. Define $\check{\boldsymbol{\beta}}$ a vector with the same length of $\hat{\boldsymbol{\beta}}$ by $\check{\boldsymbol{\beta}}_{\mathcal{A}^c} = 0$ and $\check{\boldsymbol{\beta}}_{\mathcal{A}} = \hat{\boldsymbol{\beta}}_{\mathcal{A}}$. According to Theorem 1 and Theorem 2, we have $\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(\alpha_n)$, where $\alpha_n = \sqrt{d_n/n}$. By the definition of MBIC, it follows that

$$\begin{aligned}&\mathrm{MBIC}_{k_n}(\hat{\boldsymbol{\beta}}) - \mathrm{MBIC}_{k_n}(\check{\boldsymbol{\beta}}) = -2l_n(\hat{\boldsymbol{\beta}}) + k_n\hat{d} - [-2l_n(\check{\boldsymbol{\beta}}) + k_n s_n]\\ =& -2l_n(\hat{\boldsymbol{\beta}}) + 2l_n(\check{\boldsymbol{\beta}}) + k_n(\hat{d} - s_n) = -2[l_n(\hat{\boldsymbol{\beta}}) - l_n(\boldsymbol{\beta}_0)] + 2[l_n(\check{\boldsymbol{\beta}}) - l_n(\boldsymbol{\beta}_0)] + k_n(\hat{d} - s_n)\\ \geq& -2(\hat{\boldsymbol{\beta}} - \check{\boldsymbol{\beta}})^T l_n'(\boldsymbol{\beta}_0) - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T l_n''(\bar{\boldsymbol{\beta}}_1)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T l_n''(\bar{\boldsymbol{\beta}}_2)(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + k_n\\ \triangleq& I_1 + I_2 + I_3 + I_4,\end{aligned} \tag{2.8}$$

where $\bar{\boldsymbol{\beta}}_1$ is between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and $\bar{\boldsymbol{\beta}}_2$ is between $\check{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. By using similar arguments as in Theorem 1, we can prove that the first three terms in (2.8) are all of the order $O_p(n\alpha_n^2) = O_p(d_n)$. Since $d_n/k_n \to 0$, we obtain

$$\mathrm{MBIC}_{k_n}(\hat{\boldsymbol{\beta}}) - \mathrm{MBIC}_{k_n}(\check{\boldsymbol{\beta}}) \geq O_p(d_n) + k_n = k_n[O_p(d_n/k_n) + 1] \xrightarrow{p} \infty,$$

which implies

$$P[\inf_{(\lambda,\tau)\in\Omega_+} \mathrm{MBIC}_{k_n}(\lambda,\tau) > \mathrm{MBIC}_{k_n}(\check{\boldsymbol{\beta}})] \to 1. \tag{2.9}$$

Thus we deduce that the minimum MBIC can not be selected from the overfitted model.

The results of Cases 1 and 2 complete the proof.

Remark 1 Theorem 3 implies that if $\hat{\boldsymbol{\beta}}(\lambda_n^*, \tau_n^*)$ is chosen to minimize MBIC with an appropriately chosen k_n, then $\hat{\boldsymbol{\beta}}(\lambda_n^*, \tau_n^*)$ is consistent for model selection.

3 Computation

3.1 Algorithm

We apply the smoothing quasi-Newton (SQN) method to optimize $Q_n(\boldsymbol{\beta})$ in (1.2). Since the SICA penalty function is singular at the origin, we first smooth the objective function by replacing $|\beta_j|$ with $\sqrt{\beta_j^2 + \varepsilon}$, where ε is a small positive quantity. It follows that $\sqrt{\beta_j^2 + \varepsilon} \to |\beta_j|$ when $\varepsilon \to 0$. Then we maximize

$$Q_n^{\varepsilon}(\boldsymbol{\beta}) = l_n(\beta) - n\sum_{j=1}^{d_n} p_{\lambda,\tau}(\sqrt{\beta_j^2 + \varepsilon}) \tag{3.1}$$

instead of maximizing $Q_n(\boldsymbol{\beta})$ by using the DFP quasi-Newton method with backtracking linear search algorithm procedure (e.g. [11]). In practice, taking $\varepsilon = 0.01$ gives good results. The pseudo-code for our algorithmic implementation can be found in [12]. More theoretical results about smoothing methods for nonsmooth and noconvex minimization can be found in [2, 3].

Remark 2 Like the LQA (local quadratic approximation) algorithm in [6], the sequence $\boldsymbol{\beta}^k$ obtained from SQN(DFP) may not be sparse for any fixed k and hence is not directly suitable for variable selection. In practice, we set $\beta_j^k = 0$ if $|\beta_j^k| < \varepsilon_0$ for some sufficiently small tolerance level ε_0, where β_j^k is the jth element of $\boldsymbol{\beta}^k$.

3.2 Covariance Estimation

Following [12], we estimate the covariance matrix (i.e., standard errors) for $\hat{\boldsymbol{\beta}}_1$ (the nonvanishing component of $\hat{\boldsymbol{\beta}}$) by using the sandwich formulae

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_1) = \{\nabla^2 l_n(\hat{\boldsymbol{\beta}}_1) - n\Sigma_{\lambda,\tau,\varepsilon}(\hat{\boldsymbol{\beta}}_1)\}^{-1}\widehat{\mathrm{cov}}\{\nabla l_n(\hat{\boldsymbol{\beta}}_1)\}\{\nabla^2 l_n(\hat{\boldsymbol{\beta}}_1) - n\Sigma_{\lambda,\tau,\varepsilon}(\hat{\boldsymbol{\beta}}_1)\}^{-1}, \tag{3.2}$$

where $\Sigma_{\lambda,\tau,\varepsilon}(\boldsymbol{\beta}) = \mathrm{diag}\{p'_{\lambda,\tau,\varepsilon}(|\beta_1|)/|\beta_1|, \cdots, p'_{\lambda,\tau,\varepsilon}(|\beta_d|)/|\beta_d|\}$ and $p_{\lambda,\tau,\varepsilon}(\beta_j) = p_{\lambda,\tau}(\sqrt{\beta_j^2 + \varepsilon})$. $\nabla^2 l_n(\hat{\boldsymbol{\beta}}_1)$ and $\Sigma_{\lambda,\tau,\varepsilon}(\hat{\boldsymbol{\beta}}_1)$ are the first $\hat{d} \times \hat{d}$ elements of $\nabla^2 l_n(\hat{\boldsymbol{\beta}})$ and $\Sigma_{\lambda,\tau,\varepsilon}(\hat{\boldsymbol{\beta}})$, respectively. For variables with $\hat{\beta}_j = 0$, the estimated standard errors are 0.

3.3 Tuning Parameter Selection

Numerical results suggest that the performance of SPPL estimator is robust to the choice of τ and $\tau = 0.01$ seems to give reasonable results in simulations, so we fix $\tau = 0.01$ and concentrate on tuning λ via

$$\hat{\lambda}^{\mathrm{MBIC}_{kn}} = \arg\min_{\lambda}\{\mathrm{MBIC}_{k_n}(\hat{\boldsymbol{\beta}}) = -2l_n(\hat{\boldsymbol{\beta}}) + k_n\hat{d}\}, \tag{3.3}$$

where we choose $k_n = 2\log(n)$ in the numerical experiments. We compare the performance of SPPL-MBIC with SPPL-GCV which solves

$$\hat{\lambda}^{\mathrm{GCV}} = \arg\min_{\lambda}\{\mathrm{GCV}(\hat{\boldsymbol{\beta}}) = \frac{-l_n(\hat{\boldsymbol{\beta}})}{n(1-\hat{d}/n)^2}\}. \tag{3.4}$$

In practice, we consider a range of values for $\lambda : \lambda_{\max} = \lambda_0 > \cdots > \lambda_G = 0$ for some positive number λ_0 and G, where λ_0 is an initial guess of λ, supposedly large, and G is the number of grid points (we take $G = 100$ in our numerical experiments).

3.4 Simulation Study

In this subsection, we illustrate the finite sample properties of SPPL-MBIC with a simulated example and compare it with the SPPL-GCV method. All simulations are conducted using MATLAB codes.

We simulated 100 data sets from the exponential hazards model

$$h(t|\mathbf{z}) = \exp(\boldsymbol{\beta}_0^T\mathbf{z}),$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^8$ with $\beta_{01} = 0.5$, $\beta_{02} = 1$, $\beta_{03} = -0.9$, and $\beta_{0j} = 0$, if $j \neq 1, 2, 3$. Thus $d = 8$ and $d_0 = 3$. The 8 covariates $\mathbf{z} = (z_1, \cdots, z_8)^T$ are marginally standard normal with pairwise correlations $\mathrm{corr}(z_j, z_k) = \rho^{|j-k|}$. We assume moderate correlation between the covariates by taking $\rho = 0.5$. Censoring times are generated from a uniform distribution $\mathrm{U}(0, r)$, where r is chosen to have approximately 25% censoring rate. Sample sizes $n = 150$ and 200 are considered.

To evaluate the model selection performance of both methods, for each estimate $\hat{\boldsymbol{\beta}}$, we record: the model size (MS), $|\widehat{\mathcal{A}}|$; the correct model (CM), $I\{\widehat{\mathcal{A}} = \mathcal{A}\}$; the false positive rate (FPR, the overfitting index), $|\widehat{\mathcal{A}}\backslash\mathcal{A}|/|\widehat{\mathcal{A}}|$; the false negative rate (FNR, the underfitting index), $|\mathcal{A}\backslash\widehat{\mathcal{A}}|/(d-|\widehat{\mathcal{A}}|)$; and the model error (ME), $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\Sigma(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$. Table 1 summarizes the average performance over 100 simulated datasets. With respect to parameter estimation, Table 2 presents the average of estimated nonzero coefficients (Mean), the average of estimated standard error (ESE) and the sample standard deviations (SSD).

Observing Table 1, both GCV and MBIC can work efficiently in all considered criteria, and the MBIC approach outperforms the GCV approach in terms of MS, CM, FNR and ME. In addition, all procedures have better performance in all metrics when the sample size increases from $n = 150$ to $n = 200$. From Table 2, we can see that Mean is close to its corresponding true value in all settings, and the proposed covariance estimation is shown to be reasonable in terms of ESE and SSD.

Table 1: Simulation results for model selection

n	Method	MS	CM	FPR	FNR	ME
150	MBIC	2.9400	88%	0.0075	0.0150	0.0555
	GCV	2.9100	85%	0.0075	0.0200	0.0602
200	MBIC	2.9700	95%	0.0025	0.0067	0.0408
	GCV	2.9200	90%	0.0025	0.0150	0.0498

Table 2: Simulation results for parameter estimation

		$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = -0.9$		
n	Method	Mean	ESE	SSD	Mean	ESE	SSD	Mean	ESE	SSD
150	MBIC	0.4753	0.1055	0.1838	1.0308	0.1490	0.1396	-0.8989	0.1362	0.1456
	GCV	0.4639	0.1022	0.2004	1.0349	0.1488	0.1445	-0.8972	0.1362	0.1432
200	MBIC	0.4794	0.0959	0.1488	0.9970	0.1258	0.1328	-0.8757	0.1136	0.1244
	GCV	0.4636	0.0912	0.1793	1.0002	0.1255	0.1335	-0.8734	0.1136	0.1256

4 Concluding Remarks

Since the SICA penalty is modified from the transformed L_1 penalty $p_{\lambda,\tau}(\beta_j) = \lambda|\beta_j|/(|\beta_j|+\tau)$ proposed by Nikolova [10], it is straightforward to extend the SPPL-MBIC method to the penalty function

$$p_{\lambda,\tau}(\beta_j) = \frac{\lambda}{f(1)} f(\frac{|\beta_j|}{|\beta_j|+\tau}), \tag{4.1}$$

where λ (sparsity) and τ (concavity) are two positive tunning parameters, and f is an arbitrary function that satisfies the following two hypotheses

(H1) $f(x)$ is a continuous function w.r.t x, which has the first and second derivative in $[0, 1]$;

(H2) $f'(x) \geq 0$ on the interval $[0, 1]$ and $\lim\limits_{x\to 0} \frac{f(x)}{x} = 1$.

It is noteworthy that $p_{\lambda,\tau}(\beta_j)$ is the SELO penalty function proposed by Dicker et al. [5] when we take $f(x) = \log(x + 1)$.

References

[1] Cai J, Fan J, Li R, Zhou H. Variable selection for multivariate failure time data[J]. Biometrika, 2005, 92(2): 303–316.

[2] Chen X. Superlinear convergence of smoothing quasi-Newton methods for nonsmooth equations[J]. J. Comput. Appl. Math., 1997, 80(1): 105–126.

[3] Chen X. Smoothing methods for nonsmooth, nonconvex minimization[J]. Math. Prog., 2012, 134(1): 71–99.

[4] Cox D R. Regression models and life tables (with discussion)[J]. J. Royal Stat. Soc., 1972, 34(2): 187–220.

[5] Dicker L, Huang B, Lin X. Variable selection and estimation with the seamless-L_0 penalty[J]. Stat. Sinica, 2013, 23: 929–962.

[6] Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model[J]. Ann. Stat., 2002, 30(1): 74–99.

[7] Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters[J]. Ann. Stat., 2004, 32(3): 928–961.

[8] Huang J, Liu L, Liu Y, Zhao X. Group selection in the Cox model with a diverging number of covariates[J]. Stat. Sinica, 2014, 24: 1787–1810.

[9] Lv J, Fan Y. A unified approach to model selection and sparse recovery using regularized least squares[J]. Ann. Stat., 2009, 37(6A): 3498–3528.

[10] Nikolova M. Local strong homogeneity of a regularized estimator[J]. SIAM J. Appl. Math., 2000, 61(2): 633–658.

[11] Nocedal J, Wright S. Numerical optimization (2nd ed.)[M]. New York: Springer, 2006.

[12] Shi Y Y, Cao Y X, Jiao Y L, Liu Y Y. SICA for Cox's proportional hazards model with a diverging number of parameters[J]. Acta Math. Appl. Sinica, English Ser., 2014, 30(4): 887–902.

[13] Tibshirani R. The lasso method for variable selection in the Cox model[J]. Stat. Med., 1997, 16(4): 385–395.

[14] Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters[J]. J. Royal Stat. Soc., Ser. B (Stat. Meth.), 2009, 71(3): 671–683.

[15] Wang H, Li R, Tsai C L. Tuning parameter selectors for the smoothly clipped absolute deviation method[J]. Biometrika, 2007, 94(3): 553–568.

[16] Xu C. Applications of penalized likelihood methods for feature selection in statistical modeling[D]. Vancouver: Univ. British Columbia, 2012.

[17] Zou H, Hastie T, Tibshirani R. On the "degrees of freedom" of the lasso[J]. Ann. Stat., 2007, 35(5): 2173–2192.

发散维数SICA惩罚Cox回归模型的一种修正BIC调节参数选择器

石跃勇[1,3], 焦雨领[2], 严 良[1], 曹永秀[2]

(1.中国地质大学(武汉)经济管理学院, 湖北 武汉 430074)

(2.中南财经政法大学统计与数学学院, 湖北 武汉 430073)

(3.中国地质大学(武汉)资源环境经济研究中心, 湖北 武汉 430074)

摘要: 本文研究了发散维数SICA惩罚Cox回归模型的调节参数选择问题, 提出了一种修正的BIC调节参数选择器. 在一定的正则条件下, 证明了方法的模型选择相合性. 数值结果表明提出的方法表现要优于GCV准则.

关键词: Cox模型; 修正BIC; 惩罚似然; SICA惩罚; 光滑拟牛顿

MR(2010)**主题分类号**: 62N01; 62N02 **中图分类号**: O212.1