Memorial Sloan-Kettering Cancer Center
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2017 *Paper* 36

Tuning Parameter Selection in Cox Proportional Hazards Model with a Diverging Number of Parameters

Andy Ni* Jianwen Cai†

*memorial sloan kettering cancer center, nia@mskcc.org

†University of North Carolina at Chapel Hill, cai@bios.unc.edu

Tuning Parameter Selection in Cox Proportional Hazards Model with a Diverging Number of Parameters

Andy Ni and Jianwen Cai

Abstract

Regularized variable selection is a powerful tool for identifying the true regression model from a large number of candidates by applying penalties to the objective functions. The penalty functions typically involve a tuning parameter that control the complexity of the selected model. The ability of the regularized variable selection methods to identify the true model critically depends on the correct choice of the tuning parameter. In this study we develop a consistent tuning parameter selection method for regularized Cox's proportional hazards model with a diverging number of parameters. The tuning parameter is selected by minimizing the generalized information criterion. We prove that, for any penalty that possesses the oracle property, the proposed tuning parameter selection method identifies the true model with probability approaching one as sample size increases. Its finite sample performance is evaluated by simulations. Its practical use is demonstrated in the Cancer Genome Atlas (TCGA) breast cancer data.

Memorial Sloan-Kettering Cancer Center
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2017 *Paper* 36

Tuning parameter selection in Cox model]{Tuning Parameter Selection in Cox Proportional Hazards Model with a Diverging Number of Parameters

Andy Ni*

*memorial sloan kettering cancer center, nia@mskcc.org

Tuning parameter selection in Cox model]{Tuning Parameter Selection in Cox Proportional Hazards Model with a Diverging Number of Parameters

Andy Ni

Abstract

Regularized variable selection is a powerful tool for identifying the true regression model from a large number of candidates by applying penalties to the objective functions. The penalty functions typically involve a tuning parameter that control the complexity of the selected model. The ability of the regularized variable selection methods to identify the true model critically depends on the correct choice of the tuning parameter. In this study we develop a consistent tuning parameter selection method for regularized Cox's proportional hazards model with a diverging number of parameters. The tuning parameter is selected by minimizing the generalized information criterion. We prove that, for any penalty that possesses the oracle property, the proposed tuning parameter selection method identifies the true model with probability approaching one as sample size increases. Its finite sample performance is evaluated by simulations. Its practical use is demonstrated in the Cancer Genome Atlas (TCGA) breast cancer data.

SCANDINAVIAN JOURNAL OF STATISTICS

Tuning Parameter Selection in Cox Proportional Hazards Model with a Diverging Number of Parameters

Ai Ni

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center

Jianwen Cai

Department of Biostatistics, University of North Carolina at Chapel Hill

SUMMARY: Regularized variable selection is a powerful tool for identifying the true regression model from a large number of candidates by applying penalties to the objective functions. The penalty functions typically involve a tuning parameter that control the complexity of the selected model. The ability of the regularized variable selection methods to identify the true model critically depends on the correct choice of the tuning parameter. In this study we develop a consistent tuning parameter selection method for regularized Cox's proportional hazards model with a diverging number of parameters. The tuning parameter is selected by minimizing the generalized information criterion. We prove that, for any penalty that possesses the oracle property, the proposed tuning parameter selection method identifies the true model with probability approaching one as sample size increases. Its finite sample performance is evaluated by simulations. Its practical use is demonstrated in the Cancer Genome Atlas (TCGA) breast cancer data.

KEY WORDS: Cox proportional hazards model; diverging number of parameter; generalized information criterion; TCGA data; tuning parameter selection; variable selection.

1. Introduction

In modern epidemiological and biomedical research, investigators are increasingly facing large-scale data with numerous variables. Investigators are often interested in identifying which of those variables are associated with the outcome of interest. Therefore, variable selection becomes an important task for large-scale data analysis. In order to avoid missing any potentially important variables and functional forms of them such as polynomials and interactions, it is desirable to include in the variable selection process as many candidate variables and their functions as the sample size allows. Regularized variable selection method is an effective and efficient tool to identifying important variables from a large number of candidates. In this method, a penalty is applied to the objective function to shrink the estimates of regression coefficients and achieve sparsity by estimating small coefficients as exactly zero. Many penalty functions have been proposed in the literature including Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006), and smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001), among others. It has been shown that certain penalty functions possess the so-called oracle property that they identify the true model with probability approaching one as sample size goes to infinity and estimate the nonzero parameters as efficient as if the true model is known with a proper choice of the tuning parameter (Fan & Li, 2001).

In variable selection literature, the number of parameters p is typically categorized into three scenarios according to its relationship with sample size n. In the first category, p is considered fixed as $n \to \infty$. In the next category, p is allowed to increase to infinity with n but at a slower rate. The relationship is commonly assumed to be $p = O(n^{\kappa})$ where $0 < \kappa < 1$. Models in this category are often said to have a diverging dimension. In the last category, p is assumed to increase to infinity at a faster rate than n such as $p = O(n^{\kappa})$ with $\kappa > 1$ or $\log(p) = O(n)$. Models in this category are called high-dimensional, and some researcher call them ultra high-dimensional when $\log(p) = O(n)$. In this paper we are concerned with the second category where p goes to infinity but at a slower rate than n. This scenario is useful in many practical situations. For example, in studies

that involve gene sequencing data, the number of observed single nucleotide polymorphisms and other gene alterations usually increases with the number of subjects under study. If each alteration is considered as a covariate, then it is necessary to allow the number of parameters in the model to increase with sample size. Many high-dimensional variable selection problems with $p \gg n$ can be reduced to problems with a diverging number of parameters by applying a pre-screening procedure (Fan & Lv, 2008; Fan et al., 2010b,a; Wang & Zhu, 2011).

Tuning parameter is an important component of any penalty function. It controls the complexity of the selected model. The oracle property of the penalty functions only ensures the existence of a tuning parameter that leads to the true model, but it does not provide a method to identify such tuning parameter consistently. Under the fixed-p scenario, Fan & Li (2001) used generalized cross-validation to choose the tuning parameter. This method has been shown to be analogous to the Akaike information criterion (Akaike, 1973) and overfit models with a positive probability asymptotically (Wang et al., 2007). The same authors proposed a Bayesian information criterion-based tuning parameter selection method and proved its model selection consistency in linear models. Zhang et al. (2010) further introduced a generalized information criterion for generalized linear models. Su et al. (2016) proposed an approximate information criterion for variable selection in Cox proportional hazards model. Under the diverging model dimension scenario, Wang et al. (2009) proposed a modified Bayesian information criterion for tuning parameter selection in linear models. Under the high-dimensional model setting, Wang & Zhu (2011) proposed a family of Bayesian information criteria for linear models. The authors proved that the generalized information criterion identifies the true model consistently if the penalty coefficient diverges to infinity with a rate slower than $n^{1/2}$. Fan & Tang (2013) extended this criterion to generalized linear models and established the divergence rate of its penalty coefficient for model selection consistency. More recently, Luo et al. (2015) tackled the problem from a Bayesian perspective and proposed the Extended Bayesian information criteria by modifying the prior distribution of the model space. The

authors established model selection consistency under high-dimensional ($p = O(n^{\kappa})$ with $\kappa > 1$) Cox proportional hazards model but with the requirement that the number of nonzero parameters is finite. In this paper, we extend the generalized information criterion to the Cox proportional hazards model with diverging numbers of candidate as well as nonzero parameters by establishing the required divergence rate of the penalty coefficient in the information criterion.

2. Generalized Information Criterion under Cox Proportional Hazards Model

Suppose there are n independent subjects. Let T and C be respectively the time to the outcome of interest and the censoring time. Let $X = \min(T, C)$ be the observed time and $\Delta = I(T \leqslant C)$ be the censoring indicator, where $I(\cdot)$ is an indicator function. Let $Z_i(t)$ be the $d_n \times 1$ possibly time-dependent covariate vector for subject i at time t. T and C are assumed to be independent conditional on Z. Let $\beta = (\beta_1, ..., \beta_{d_n})^T \in \mathscr{B} \subset \mathscr{R}^{d_n}$ be a vector of regression coefficients and $\beta_0 = (\beta_{01}, ..., \beta_{0d_n})^T$ be its true value. Assume there are k_n nonzero components of β_0 and $d_n - k_n$ zero components. We allow both d_n and k_n to increase to infinity with n but with a slower rate than n. Although the dimensions of β, β_0, and $Z_i(t)$ all depend on n, we omit n from the subscript for notational simplicity. Define for subject i the counting process $N_i(t) = I(X_i \leqslant t, \Delta_i = 1)$, and the at risk process $Y_i(t) = I(X_i \geqslant t)$. The log-partial likelihood under Cox proportional hazards model is

$$\ell_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left(\beta^T Z_i(t) - \log \left[\frac{1}{n} \sum\nolimits_{j=1}^{n} Y_j(t) \exp \left\{ \beta^T Z_j(t) \right\} \right] \right) dN_i(t), \tag{1}$$

where τ is the time at the end of study. This log-partial likelihood is slightly different from the conventional definition by including a $1/n$ term inside the logarithm. This leads to the same score function and estimate of β as the conventional definition but will facilitate the theoretical derivations in this paper. Let $P_\lambda(\cdot)$ be a penalty function with tuning parameter λ. We assume that the penalty function possesses the oracle property. The penalized maximum partial likelihood

estimator $\hat{\beta}_\lambda$ is the maximizer of the following objective function,

$$\ell_n(\beta) - n\sum_{j=1}^{d_n} P_\lambda(|\beta_j|). \tag{2}$$

Let α_λ be the model that is identified by the tuning parameter λ . Let α_0 be the true model. Let $|\alpha_\lambda|$ be the size of model α_λ. Then $|\alpha_0| = k_n$. We consider the generalized information criterion

$$\mathrm{GIC}(\lambda) = \frac{1}{n}\left\{-\ell_n(\hat{\beta}_\lambda) + a_n|\alpha_\lambda|\right\}, \tag{3}$$

where the penalty coefficient a_n is a positive sequence depending on n. When $a_n = 1$ the criterion becomes the AIC statistic. Wang et al. (2007) noted that, when d_n is small compared to n, the AIC statistic is approximately equal to the generalized cross-validation statistic (Craven & Wahba, 1979), which is frequently used for tuning parameter selection in Cox model (Tibshirani, 1997; Fan & Li, 2002; Cai et al., 2005; Zhang & Lu, 2007). When $a_n = \log(n)/2$, the criterion becomes the Bayesian information criterion (BIC). Although there is no direct use of BIC for Cox model selection, some modified forms of BIC have been proposed for Cox model selection in the literature (Volinsky & Raftery, 2000; Luo et al., 2015). The selected tuning parameter $\hat{\lambda}$ is the minimizer of (3). The oracle property guarantees the existence of at least one λ that gives rise to the true model α_0. The goal of this paper is to determine the characteristic of the sequence a_n so that the λ leading to the true model is identified with probability tending to one as sample size goes to infinity.

3. Notations and Regularity Conditions

In addition to the penalized estimator $\hat{\beta}_\lambda$, we also define the unpenalized maximum partial likelihood estimator $\hat{\beta}_{\alpha_\lambda}$ for model α_λ, which maximizes (1). Note that $\hat{\beta}_\lambda$ is a function of λ and $\hat{\beta}_{\alpha_\lambda}$ is a function of the model. For a given model α_λ, we define its true parameter $\beta^0_{\alpha_\lambda}$ as the minimizer of the Kullback-Leibler distance $D(\beta_{\alpha_\lambda}) = n^{-1}E\{\ell_n(\beta_0) - \ell_n(\beta_{\alpha_\lambda})\}$. The expectation is taken under the true model.

Let $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$ for a vector a. Define the following notations for each n:

$$S_n^{(k)}(\beta, t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) Z_i(t)^{\otimes k} e^{\beta^T Z_i(t)}, \quad k = 0, 1, 2,$$

$$s_n^{(k)}(\beta, t) = E\{S_n^{(k)}(\beta, t)\}, \quad k = 0, 1, 2,$$

$$I_n(\beta) = -\frac{1}{n} E\left\{\frac{\partial^2 \ell_n(\beta)}{\partial \beta^2}\right\}.$$

We require the following regularity conditions:

(A) $\int_0^\tau h_0(t)dt < \infty$, where $h_0(t)$ is the baseline hazard function.

(B) $E\{Y(\tau)\} > 0$.

(C) $|Z_{ij}(0)| + \int_0^\tau |dZ_{ij}(t)| < C_1 < \infty$ almost surely for some constant C_1 and $i = 1, ..., n$ and $j = 1, ..., d_n$. It implies that $K_n = \max_{1 \leqslant j \leqslant d_n, 1 \leqslant i \leqslant n} \|Z_{ij}(t)\|_\infty < \infty$, where $\|\cdot\|_\infty$ denotes the supremum norm.

(D) For any model α_λ, there exists a neighborhood $\mathscr{B}_{\alpha_\lambda}$ of $\beta^0_{\alpha_\lambda}$ such that for all $\beta_{\alpha_\lambda} \in \mathscr{B}_{\alpha_\lambda}$ and $t \in [0, \tau]$, $\partial s_n^{(0)}(\beta_{\alpha_\lambda}, t)/\partial \beta_{\alpha_\lambda} = s_n^{(1)}(\beta_{\alpha_\lambda}, t)$, and $\partial^2 s_n^{(0)}(\beta_{\alpha_\lambda}, t)/\partial \beta_{\alpha_\lambda} \partial \beta_{\alpha_\lambda}^T = s_n^{(2)}(\beta_{\alpha_\lambda}, t)$. The functions $s_n^{(k)}(\beta_{\alpha_\lambda}, t)$ $(k = 0, 1, 2)$ are continuous and bounded and $s_n^{(0)}(\beta_{\alpha_\lambda}, t)$ is bounded away from 0 on $\mathscr{B}_{\alpha_\lambda} \times [0, \tau]$.

(E) For any model α_λ, there exists a neighborhood $\mathscr{B}_{\alpha_\lambda}$ of $\beta^0_{\alpha_\lambda}$ such that for all $\beta_{\alpha_\lambda} \in \mathscr{B}_{\alpha_\lambda}$, there exists positive constant C_2 and C_3 such that

$$0 < C_2 < \text{eigen}_{\min}\{I_n(\beta_{\alpha_\lambda})\} \leqslant \text{eigen}_{\max}\{I_n(\beta_{\alpha_\lambda})\} < C_3 < \infty,$$

where $\text{eigen}_{\min}(\cdot)$ and $\text{eigen}_{\max}(\cdot)$ are the minimum and maximum eigenvalues of a matrix, respectively.

(F) $L_n = \sup_{\beta \in \mathscr{B}} \|\beta\|_1 < \infty$, where $\|\cdot\|_1$ denotes the L_1 norm. As a consequence of this condition and Condition (C), we can define $\exp\{|\beta^T Z_i(t)|\} \leqslant \exp(L_n K_n) = U_n < \infty$ for $i = 1, ..., n$ and $\beta \in \mathscr{B}$.

(G) $d_n^4/n \to 0$ and $k_n/d_n \to c \in [0, 1)$ as $n \to \infty$.

Condition (A) ensures finite baseline cumulative hazard. Condition (B) ensures non-empty risk

set by the end of the study. Condition (C) requires the stochastic process of each time-dependent covariate to have bounded total variation almost surely. Condition (D) essentially requires $\exp\{\beta_{\alpha_\lambda}^T Z_i(t)\}$ to be integrable under a diverging dimension so that integration and differentiation with respect to $S_n^{(k)}(\beta_{\alpha_\lambda}, t)$ $(k = 0, 1)$ can be interchanged, which is a standard condition for the proportional hazards model. Condition (E) ensures that the covariance matrices of the score function are positive definite and have uniformly bounded eigenvalues for all n. The same condition has been assumed in the variable selection literature (Peng & Fan, 2004; Cai et al., 2005; Cho & Qu, 2013). Condition (F) confines our investigation to the parameters with a finite total effect on the hazard function, which is very reasonable in practice. Condition (G) specifies the divergence rate of the number of candidate and nonzero parameters that is required to establish the theoretical results in this paper.

4. Asymptotic Properties of the Generalized Information Criterion

Let $\lambda_{\max}$ be the smallest λ that results in an empty model with no nonzero estimates. We partition the tuning parameter space $\Omega = [0, \lambda_{\max}]$ into the underfit, true, and overfit subspaces as follows,

$$\Omega_- = \{\lambda : \alpha_\lambda \not\supset \alpha_0\}, \quad \Omega_0 = \{\lambda : \alpha_\lambda = \alpha_0\}, \quad \Omega_+ = \{\lambda : \alpha_\lambda \supsetneq \alpha_0\},$$

where $a \supsetneq b$ means a contains b but is not equal to b. Since $\hat{\beta}_\lambda$ is the maximizer of potentially nonconcave objective function (2) due to nonconcave penalties, the asymptotic property of $\ell_n(\hat{\beta}_\lambda)$ is difficult to study. Instead, we work with the unpenalized version of the log-partial likelihood. Define

$$\mathrm{GIC}^*(\alpha_\lambda) = \frac{1}{n}\left\{-\ell_n(\hat{\beta}_{\alpha_\lambda}) + a_n|\alpha_\lambda|\right\}.$$

Note that $\mathrm{GIC}^*(\alpha_\lambda)$ is a function of the model whereas $\mathrm{GIC}(\lambda)$ is a function of the tuning parameter. We only present main results in this section, the proofs of which are outlined in the Web Appendix. There are two challenges in the proofs that are unique to the log-partial likelihood. First, the log-partial likelihood and its score function are summations of dependent terms. We introduce two intermediate quantities to tackle this difficulty. Second, the log-partial likelihood does not

possess the Lipschitz property (Kong & Nan, 2014) so certain asymptotic properties cannot be established uniformly for β. We instead establish the pointwise properties for any given β, which suffices our purpose as we are only concerned with the maximum partial likelihood estimator $\hat{\beta}_{\alpha_\lambda}$.

The following lemma states that, for any λ, the difference between $\mathrm{GIC}(\lambda)$ and $\mathrm{GIC}(\lambda_0)$ is no less than that between $\mathrm{GIC}^*(\alpha_\lambda)$ and $\mathrm{GIC}^*(\alpha_0)$ with probability tending to one.

LEMMA 1: *If the penalty function possesses the oracle property for the log-partial likelihood* (1)*, then for any* $\lambda \in \Omega$ *and* $\lambda_0 \in \Omega_0$*, as* $n \to \infty$*,*

$$\mathrm{pr}\left\{\mathrm{GIC}(\lambda) - \mathrm{GIC}(\lambda_0) \geqslant \mathrm{GIC}^*(\alpha_\lambda) - \mathrm{GIC}^*(\alpha_0)\right\} \to 1.$$

Lemma 1 allows us to study the asymptotic properties of $\mathrm{GIC}^*(\alpha_\lambda)$ instead of $\mathrm{GIC}(\lambda)$. Cai et al. (2005) established oracle property for SCAD penalty under Cox model with a growing number of parameters. Bradic et al. (2011) further proved the oracle property for the class of folded concave penalties including SCAD, minimax concave penalty (MCP), and Lasso under Cox model with non-polynomial dimensionality which includes diverging dimension as a special case.

The following theorem describes the uniform stochastic order of the difference between $\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})$ and $D(\beta^0_{\alpha_\lambda})$ over all possible model α_λ, the number of which increases combinatorially fast with sample size. All expectations are taken under the true model.

THEOREM 1: *Under Conditions (A) to (G), uniformly for all models,*

$$\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\} \right| = O_p[n^{1/2}\{\log(d_n)\}^{1/2}].$$

Based on Theorem 1, for all underfitted model $\alpha_\lambda \not\supset \alpha_0$ we have that,

$$\begin{aligned}
&\inf_{\alpha_\lambda \not\supset \alpha_0} \{\mathrm{GIC}^*(\alpha_\lambda) - \mathrm{GIC}^*(\alpha_0)\} \\
&= \inf_{\alpha_\lambda \not\supset \alpha_0} \frac{1}{n} \Big[\ell_n(\hat{\beta}_{\alpha_0}) - \ell_n(\hat{\beta}_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_0}) - \ell_n(\beta^0_{\alpha_\lambda})\} + E\{\ell_n(\beta^0_{\alpha_0}) - \ell_n(\beta^0_{\alpha_\lambda})\} \\
&\quad + a_n(|\alpha_\lambda| - |\alpha_0|) \Big] \\
&\geqslant -\frac{1}{n} \sup_{\alpha_\lambda \not\supset \alpha_0} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\} \right| - \frac{1}{n} \left| \ell_n(\hat{\beta}_{\alpha_0}) - E\{\ell_n(\beta^0_{\alpha_0})\} \right| + \inf_{\alpha_\lambda \not\supset \alpha_0} D(\beta^0_{\alpha_\lambda})
\end{aligned}$$

$$
\begin{aligned}
&+ \inf_{\alpha_\lambda \not\supset \alpha_0} \frac{1}{n} a_n(|\alpha_\lambda| - |\alpha_0|) \\
&\geqslant -\frac{1}{n} \sup_{\alpha_\lambda \not\supset \alpha_0} \left| \ell_n(\hat{\beta}_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\} \right| - \frac{1}{n} \left| \ell_n(\hat{\beta}_{\alpha_0}) - E\{\ell_n(\beta^0_{\alpha_0})\} \right| + \delta_n - \frac{1}{n} a_n k_n \\
&= -\frac{1}{n} O_p[\{d_n n \log(d_n)\}^{1/2}] - \frac{1}{n} O_p[\{d_n n \log(d_n)\}^{1/2}] + \delta_n - \frac{1}{n} a_n k_n \\
&= -O_p[\{d_n \log(d_n)\}^{1/2} n^{-1/2}] + \delta_n - \frac{1}{n} a_n k_n, \qquad (4)
\end{aligned}
$$

where $\delta_n = \inf_{\alpha_\lambda \not\supset \alpha_0} D(\beta^0_{\alpha_\lambda})$ defines the smallest Kullback–Leibler distance to the true model among all underfitted models. It can be deemed as the signal strength of the true model. Since δ_n is always positive, if δ_n and a_n satisfy the conditions $\delta_n n^{1/2} \{d_n \log(d_n)\}^{-1/2} \to \infty$ and $a_n = o(\delta_n n k_n^{-1})$, then (4) is positive with probability tending to one. Then by Lemma 1,

$$
\mathrm{pr}\left[\inf_{\lambda \in \Omega_-} \{\mathrm{GIC}(\lambda) - \mathrm{GIC}(\lambda_0)\} > 0 \right] \to 1
$$

as $n \to \infty$. This result suggests that as long as the signal strength of the true model does not decay to zero too fast and the sequence a_n does not go to infinity too fast, the generalized information criterion of all underfitted models is larger than that of the true model with probability tending to one.

For overfitted models, the Kullback–Leibler distance based method is no longer useful because $D(\beta^0_{\alpha_\lambda}) = 0$ for all $\alpha_\lambda \supsetneq \alpha_0$ so δ_n cannot be well defined. We instead study the asymptotic property of $\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})$ directly. If the dimension of the model is finite, it is known that 2 times the log-partial likelihood ratio converges to a χ^2 distribution with $|\alpha_\lambda| - |\alpha_0|$ degree of freedom. However, when the model dimension goes to infinity, we have to consider higher order terms in the linearization of the log-partial likelihood ratio statistic. Moreover, obtaining a uniform stochastic order of $\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})$ over all overfitted models is challenging since the number of such models increases to infinity combinatorially fast.

THEOREM 2: *Under Conditions (A) to (G), uniformly for all* $\alpha_\lambda \supsetneq \alpha_0$,

$$
\sup_{\alpha_\lambda \supsetneq \alpha_0} \frac{1}{|\alpha_\lambda| - |\alpha_0|} \left\{ \ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) \right\} = O_p\{\log(d_n)\}.
$$

As a consequence of Theorem 2, uniformly for all overfitted models we have that

$$\begin{aligned}
&\inf_{\alpha_\lambda \supsetneq \alpha_0} \frac{\mathrm{GIC}^*(\alpha_\lambda) - \mathrm{GIC}^*(\alpha_0)}{|\alpha_\lambda| - |\alpha_0|} \\
&= \inf_{\alpha_\lambda \supsetneq \alpha_0} \frac{1}{n(|\alpha_\lambda| - |\alpha_0|)} \left\{ \ell_n(\hat{\beta}_{\alpha_0}) - \ell_n(\hat{\beta}_{\alpha_\lambda}) + a_n(|\alpha_\lambda| - |\alpha_0|) \right\} \\
&= -O_p\{n^{-1}\log(d_n)\} + \frac{a_n}{n}. \qquad (5)
\end{aligned}$$

Therefore, when $a_n/\log(d_n) \to \infty$, (5) is positive with probability tending to one. Since $|\alpha_\lambda| - |\alpha_0|$ is positive for all overfitted models, it follows that $\inf_{\alpha_\lambda \supsetneq \alpha_0} \mathrm{GIC}^*(\alpha_\lambda) - \mathrm{GIC}^*(\alpha_0)$ is positive with probability approaching one when $a_n/\log(d_n) \to \infty$. By Lemma 1 it follows that

$$\mathrm{pr}\left[\inf_{\lambda \in \Omega_+} \{\mathrm{GIC}(\lambda) - \mathrm{GIC}(\lambda_0)\} > 0\right] \to 1$$

as $n \to \infty$. With Theorem 1 and 2, we arrive at the following theorem.

THEOREM 3: *Under Conditions (A) to (G), if* $\delta_n n^{1/2}\{d_n \log(d_n)\}^{-1/2} \to \infty$, $a_n = o(\delta_n n k_n^{-1})$, *and* $a_n/\log(d_n) \to \infty$, *then as* $n \to \infty$,

$$\mathrm{pr}\left\{\inf_{\lambda \in \Omega_- \cup \Omega_+} \mathrm{GIC}(\lambda) > \mathrm{GIC}(\lambda_0)\right\} \to 1.$$

Theorem 3 is a direct consequence of Theorem 1 and 2. It entails that, if the signal strength of the true model does not decrease to zero too fast and a_n diverges with sample size with a proper range of rates, then by minimizing the generalized information criterion we can identify the tuning parameter that leads to the true model with probability tending to one as sample size goes to infinity. From the three requirements specified in Theorem 3 we can see that the lower bound of the divergence rate of a_n is $\log(d_n)$. The upper bound depends on the signal strength δ_n. If δ_n satisfies the first requirement, then $a_n = O[\{n d_n \log(d_n)\}^{1/2}/k_n]$ always satisfies the second requirement. Hence, any a_n with a divergence rate between $\log(d_n)$ and $\{n d_n \log(d_n)\}^{1/2}/k_n$ gives model selection consistency as sample size goes to infinity. Notably, the AIC statistic where $a_n = 2$ does not satisfy the requirements listed in Theorem 3, hence its inconsistency in model selection. The BIC statistic where $a_n = \log(n)$ does satisfy the model selection consistency requirements. Moreover, there is a range of other consistent information criteria as long as their a_n satisfies the

requirements in Theorem 3. In the simulation study that follows, we will investigate the finite sample performance of AIC, BIC, and one other consistent information criterion.

5. Simulation Studies

We use the smoothly clipped absolute deviation penalty (Fan & Li, 2001) to demonstrate the finite sample performance of the proposed tuning parameter selection method. Cai et al. (2005) has established the oracle property of this penalty function in Cox model with a diverging number of parameters.

Independent failure times are generated from the exponential hazard model. We set the baseline hazard $h_0(t) = 2$ and the dimension of β to be $d_n = [10n_c^{1/5}]$ to reflect that it increases with the number of cases n_c and in turn with the sample size. We set d_n as a function of n_c rather than n because the former better represents the amount of information contained in the dataset. The first component of β is the smallest nonzero parameter in terms of the absolute value, denoted by $\beta_{\min}$, which is related to δ_n, the signal strength of the true model. As it is not possible to specify the required convergence rate of δ_n under finite sample size, we consider three different values of $\beta_{\min}$: 1.0, 0.34, and 0.18 corresponding to hazard ratio of 2.8, 1.4, and 1.2, respectively. The other nonzero parameters recycle from 0.6 and -0.8. There is one nonzero parameter for every two zero parameters. The pattern of β is $(\beta_{\min}, 0, 0, 0.6, 0, 0, -0.8, 0, 0, 0.6, 0, 0, -0.8, 0, 0, ...)$. We generate the design matrix Z as a mixture of correlated binary and continuous variables. First, a d_n-dimensional multivariate standard normal variable Z^* is generated with $\mathrm{corr}(Z_i^*, Z_j^*) = 0.5^{|i-j|}$. Then the first three components of Z^* are kept continuous, and the next three components are dichotomized at zero, and this pattern is repeated for the rest of Z^*. Thus half of the covariates become binary with parameter 0.5. Censoring times C_i are generated from a uniform distribution $\mathrm{U}(0, c)$ where c is adjusted to achieve desired censoring percentage.

Various sample sizes and censoring rates are considered for each of the two $\beta_{\min}$ values. Variable selection performance of the generalized information criterion is assessed for three choices of a_n: 1,

$\log(n)/2$, and $\log\{\log(d_n)\}\log(d_n)$. The first two choices correspond to AIC and BIC, respectively. The third one has a divergence rate between AIC and BIC. We also include as a comparison the extended BIC (EBIC) (Luo et al., 2015) where $a_n|\alpha_\lambda|$ in the proposed GIC is replaced by $\log(n)|\alpha_\lambda|/2 + \gamma \log\binom{d_n}{|\alpha_\lambda|}$. Following the authors, we set $\gamma = 1 - 1/\{4\log(d_n)/\log(n)\}$. Since the objective function (2) is not concave, we use local quadratic approximation algorithm to obtain the estimates and their standard errors (Fan & Li, 2001). As a benchmark, we include the hard threshold variable selection procedure, where the component of the unpenalized maximum partial likelihood estimator from the full model is selected if its p-value from the Wald test is less than 0.05. We also include the result from the oracle procedure where the correct subset of covariates is used to fit the model. For each setting 500 replications are conducted.

The performance of the variable selection procedure is evaluated by the average number of zero parameters correctly estimated as zero (true negative number), the average number of nonzero parameters erroneously estimated as zero (false negative number), the average number of correctly identified parameters (both zero and nonzero), and the rate of identifying the true model. In addition, we define model error of a variable selection procedure as $\text{ME}(\hat{\mu}) = E\{E(T \mid z) - \hat{\mu}(z)\}^2$. Under the proportional hazard model with constant baseline hazard h_0, it can be shown that $\text{ME}(\hat{\mu}) = h_0^{-2} E\{\exp(-\hat{\beta}^T z) - \exp(-\beta_0^T z)\}^2$ and is estimated by $h_0^{-2} m^{-1} \sum_{i=1}^{m} \{\exp(-\hat{\beta}_i^T z_i) - \exp(-\beta_0^T z_i)\}^2$, where m is the number of simulation replications. The relative model error for a particular model is defined as the ratio of its model error to that of the unpenalized full model. We use the median and the median absolute deviation of the estimated relative model error to compare the performance of different variable selection procedures.

Table 1 summarizes the variable selection performance of different generalized information criteria under sample sizes 1500, 2500, and 5000 and censoring rates of 80% and 90%. Overall, the criterion with $a_n = \log\{\log(d_n)\}\log(d_n)$ gives the best performance in terms of rate of identifying the true model and the median relative model error in various settings. The performance of the

EBIC is remarkably close to that of $a_n = \log\{\log(d_n)\}\log(d_n)$ with the latter outperforming the former slightly but consistently across all scenarios. The only scenarios where the performance of $a_n = \log(n)/2$ is similar to or slightly better than that of $a_n = \log\{\log(d_n)\}\log(d_n)$ are when both of them have very high rate of identifying the true model or the signal strength is strong ($\beta_{\min} = 1.0$). Based on the average number of correctly and incorrectly identified zero parameters, the criterion with $a_n = 1$ tends to select more parameters into the final model than does the criterion with $a_n = \log\{\log(d_n)\}\log(d_n)$, whereas the criterion with $a_n = \log(n)/2$ tends to select less parameters than does the criterion with $a_n = \log\{\log(d_n)\}\log(d_n)$. This is consistent with the fact that $\log\{\log(d_n)\}\log(d_n)$ lies between 1 and $\log(n)/2$. We also evaluate the rates of identifying the true model and average percentages of correctly identified parameters for different generalized information criteria under wider range of sample sizes and censoring rates. The results are summarized in Figure 1 and 2. It is apparent that the generalized information criterion with $a_n = \log\{\log(d_n)\}\log(d_n)$ offers the best overall performance in variable selection under most sample sizes and censoring rates. The only scenarios where the choices of $a_n = 1$ or $\log(n)/2$ outperform $a_n = \log\{\log(d_n)\}\log(d_n)$ are those where the latter's performance is already very satisfactory.

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

6. Real Data Applications

The Cancer Genome Atlas (TCGA) Research Network is a large collection of publically available genomic sequence and mRNA expression data from tumor samples of various types of cancer (http://cancergenome.nih.gov). The availability of matched overall survival data makes it possible to conduct analysis to identify gene alterations and expressions that are potentially prognostic of

the overall survival. In this analysis we use the breast invasive carcinoma dataset that contains the mRNA expression data from 816 cancer patients (Ciriello et al., 2015). There are 119 death events, corresponding to a 85.4% censoring rate. The mRNA expression was measured by RNAseq technique and was standardized into z-scores for each gene by subtracting the mean and divided by the standard deviation of the mRNA expression of that gene in the normal samples in the TCGA database. Each gene was further categorized as significantly altered if the absolute value of its z-score is larger than 1.96 and not altered otherwise. We consider the 468 genes that constitute the IMPACT gene panel developed and routinely used at the Memorial Sloan Kettering Cancer Center (http://cmo.mskcc.org/cmo/resources/gene-lists). Since the number of genes is more than half of the sample size and exceeds the number of deaths, the Condition (G) imposed in this paper is likely to be violated. To overcome this difficulty, we pre-screen the candidate genes by only including those with an alteration frequency greater than 5% and a univariate log-rank test p value less than 0.05. These steps result in 35 genes that enter the subsequent SCAD-penalized variable selection procedure. The idea of pre-screening followed by penalized regression has been thoroughly studies in the literature (Fan et al., 2010a). The alteration frequency of the 35 genes range from 5% to 41%. We again use the three choices of a_n, EBIC, and the hard threshold method to select the genes. The chosen tuning parameters λs are: 0.41 for $a_n = 1$, 0.73 for $a_n = \log(n)/2$, 0.66 for $a_n = \log\{\log(d_n)\}\log(d_n)$, and 0.66 for EBIC. The identified genes are summarized in Table 2. Only genes that are selected by at least one method are listed.

The criterion with $a_n = \log\{\log(d_n)\}\log(d_n)$ identifies two genes: MLH1 and KRAS. The MLH1 gene mutation has been reported to be associated with over ten-fold increase in the incidence ratio of breast cancer (Scott et al., 2001). The KRAS gene amplification and mutation are well known to be present in a number of cancers including breast cancer, lung cancer, and endrometrial cancer (Kim et al., 2015; Pereira et al., 2013; Birkeland et al., 2012). Therefore, the identification of MLH1 and KRAS gene mutations makes biological sense. The EBIC method

identified the same two genes as $a_n = \log\{\log(d_n)\}\log(d_n)$, which is expected given the similar results of these two methods in the simulation studies. The criterion with $a_n = \log(n)/2$ misses the important KRAS gene mutation. On the other hand, the criterion with $a_n = 1$ identifies ten genes and the hard threshold method identifies five, many of which do not have previous literature to support their association with the overall survival.

[Table 2 about here.]

7. Discussion

The theorems developed in this paper specify theoretical range of the divergence rate of the sequence a_n for model selection consistency. Any rate within the range leads to selection consistency. Therefore, the choices of a_n is not unique. In real-data applications with finite sample sizes, different choice of a_n may yield different results. Our simulation studies numerically demonstrate that the choice of $a_n = \log\{\log(d_n)\}\log(d_n)$ offers an overall superior variable selection performance over wide ranges of sample sizes and censoring rates. Admittedly, there likely to be other situations where other choices of a_n may offer better performance. The main goal of this paper is to establish the theoretical requirement on a_n for selection consistency. It is not our intention to provide the best choices of a_n for all possible finite sample scenarios. In practice, we suggest practitioners to use a few different a_n choices as a sensitivity analysis to assess how robust the selected model is to the variation of a_n.

Although in this paper the model selection consistency of the generalized information criterion is investigated in the context of regularized variable selection in Cox's proportional hazards model with a diverging number of parameters, the conclusions of our study have a much broader application. In fact, the generalized information criterion developed in this paper can be used to identify the true model from any set of candidate Cox's regression models as long as the true model is contained in the set. Therefore, it can be equally applied to the best subset selection or

stepwise model selection procedures. In the context of regularized variable selection, the solution path corresponding to a sequence of tuning parameter forms the set of candidate models. The oracle property of the penalty function ensures that the solution path contains the true model.

A natural research question is to study the properties of the generalized information criterion when the set of candidate models does not contain the true model. This happens when the solution path of a regularized variable selection procedure fails to capture the true model or some covariates with nonzero true effects are not included in the initial family of candidate covariates. In these cases it is not clear if the proposed generalized information criterion can consistently identify all nonzero parameters. Another potential research direction is to evaluate a variable selection procedure by certain loss function of the estimated parameters rather than model selection consistency. Zhang et al. (2010) investigated the squared loss of the penalized estimator in linear models with fixed number of parameters and found that the Akaike but not the Bayesian information criterion is asymptotically loss efficient in that it identifies the model whose squared loss converges to the infimum of the squared loss of all possible models. It would be interesting to establish similar results for Cox's proportional hazards model with a diverging number of parameters.

Another future research direction is to apply the theoretical framework used in this paper to the variable selection method recently proposed by Su et al. (2016) under Cox model with a fixed model size. In their approach, the authors essentially approximate $|\alpha_\lambda|$ in our GIC with the "unit dent function" $\sum_{j=1}^{d_n} \tanh(n_c\gamma_j^2)$ and set a_n in our GIC to $\log(n_c)$, which lies in the range of divergence rate identified in our paper for selection consistency. Although the authors showed under a particular finite sample setting that the parameter estimation is robust to the choice of a_n, it would still be interesting to extend our theoretical framework to their approach to identify a theoretical range of divergence rate of a_n that ensures selection consistency under Cox model with a diverging number of parameters.

Acknowledgements

This work was partially supported by National Institutes of Health grants (P01 CA 142538, R01 ES 021900).

Supporting information

Additional information for this article is available online including proofs of the theorems and lemmas.

REFERENCES

AKAIKE, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255–265.

BIRKELAND, E., WIK, E., MJØS, S., HOIVIK, E., TROVIK, J., WERNER, H. M. J., KUSONMANO, K., PETERSEN, K., RÆDER, M. B., HOLST, F. et al. (2012). KRAS gene amplification and overexpression but not mutation associates with aggressive and metastatic endometrial cancer. *British Journal of Cancer* **107**, 1997–2004.

BRADIC, J., FAN, J. & JIANG, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39**, 3092.

CAI, J., FAN, J., LI, R. & ZHOU, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303–316.

CHO, H. & QU, A. (2013). Model selection for correlated data with diverging number of parameters. *Statist. Sinica* **23**, 901–927.

CIRIELLO, G., GATZA, M. L., BECK, A. H., WILKERSON, M. D., RHIE, S. K., PASTORE, A. & PEROU, C. M. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519.

CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions: Estimating the

correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377–403.

FAN, J., FENG, Y., WU, Y. et al. (2010a). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*. Institute of Mathematical Statistics, pp. 70–86.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

FAN, J. & LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74–99.

FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70**, 849–911.

FAN, J., SONG, R. et al. (2010b). Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist.* **38**, 3567–3604.

FAN, Y. & TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75**, 531–552.

KIM, R.-K., SUH, Y., YOO, K.-C., CUI, Y.-H., KIM, H., KIM, M.-J., KIM, I. G. & LEE, S.-J. (2015). Activation of KRAS promotes the mesenchymal features of basal-type breast cancer. *Experimental & Molecular Medicine* **47**, e137.

KONG, S. & NAN, B. (2014). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via lasso. *Statist. Sinica* **24**, 25–42.

LUO, S., XU, J. & CHEN, Z. (2015). Extended Bayesian information criterion in the Cox model with a high-dimensional feature space. *Ann. Inst. Statist. Math.* **67**, 287–311.

PENG, H. & FAN, J. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–961.

PEREIRA, C. B. L., LEAL, M. F., DE SOUZA, C. R. T., MONTENEGRO, R. C., REY, J. A.,

CARVALHO, A. A., ASSUMPÇÃO, P. P., KHAYAT, A. S., PINTO, G. R., DEMACHKI, S. et al. (2013). Prognostic and predictive significance of MYC and KRAS alterations in breast cancer from women treated with neoadjuvant chemotherapy. *PloS One* **8**, e60576.

SCOTT, R. J., MCPHILLIPS, M., MELDRUM, C. J., FITZGERALD, P. E., ADAMS, K., SPIGELMAN, A. D., DU SART, D., TUCKER, K., KIRK, J. & SERVICE, H. F. C. (2001). Hereditary nonpolyposis colorectal cancer in 95 families: differences and similarities between mutation-positive and mutation-negative kindreds. *The American Journal of Human Genetics* **68**, 118–127.

SU, X., WIJAYASINGHE, C. S., FAN, J. & ZHANG, Y. (2016). Sparse estimation of Cox proportional hazards models via approximated information criteria. *Biometrics* **72**, 751–759.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58**, 267–288.

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395.

VOLINSKY, C. T. & RAFTERY, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56**, 256–262.

WANG, H., LI, B. & LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71**, 671–683.

WANG, H., LI, R. & TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

WANG, T. & ZHU, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *J. Multivariate Anal.* **102**, 1141 – 1151.

ZHANG, H. H. & LU, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94**, 691–703.

ZHANG, Y., LI, R. & TSAI, C.-L. (2010). Regularization parameter selections via generalized

information criterion. *J. Amer. Statist. Assoc.* **105**, 312–323.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

Ai Ni, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center.
Email: nia@mskcc.org

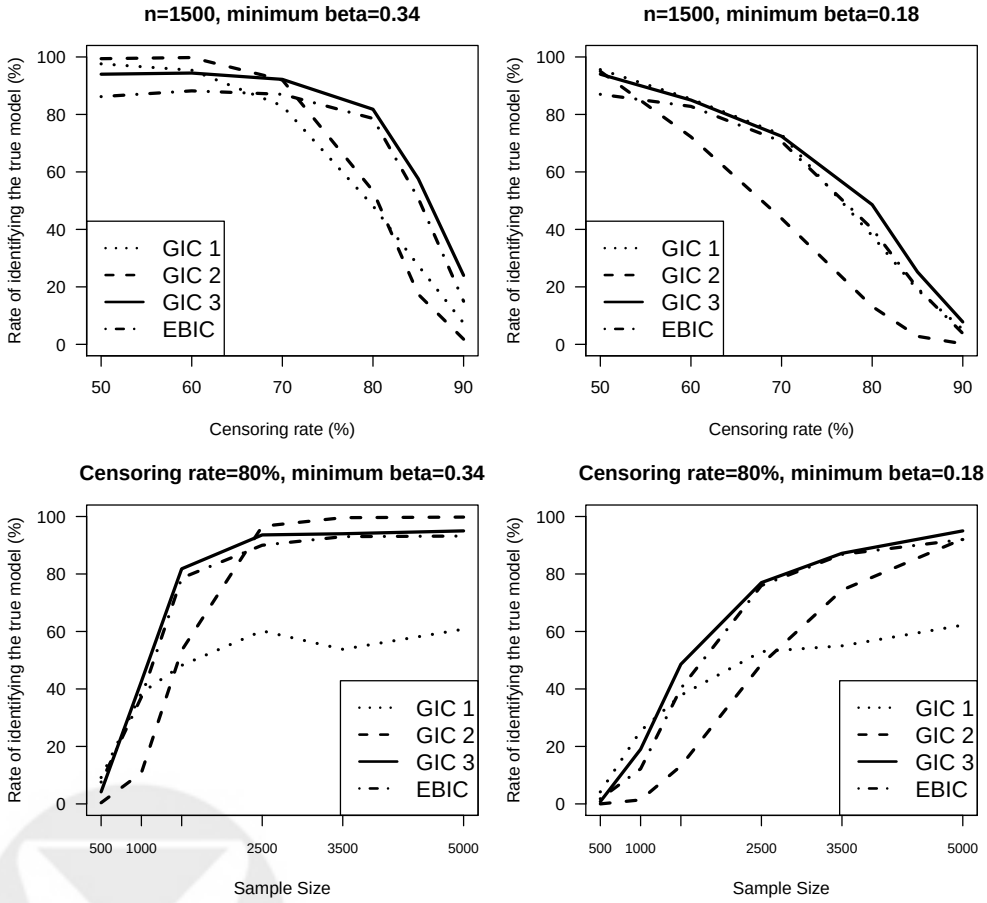

Figure 1. Rate of identifying the true model (RITM) of different choices of a_n in the generalized information criterion. GIC 1: $a_n = 1$; GIC 2: $a_n = \log(n)/2$; GIC 3: $a_n = \log\{\log(d_n)\}\log(d_n)$; EBIC: extended BIC.

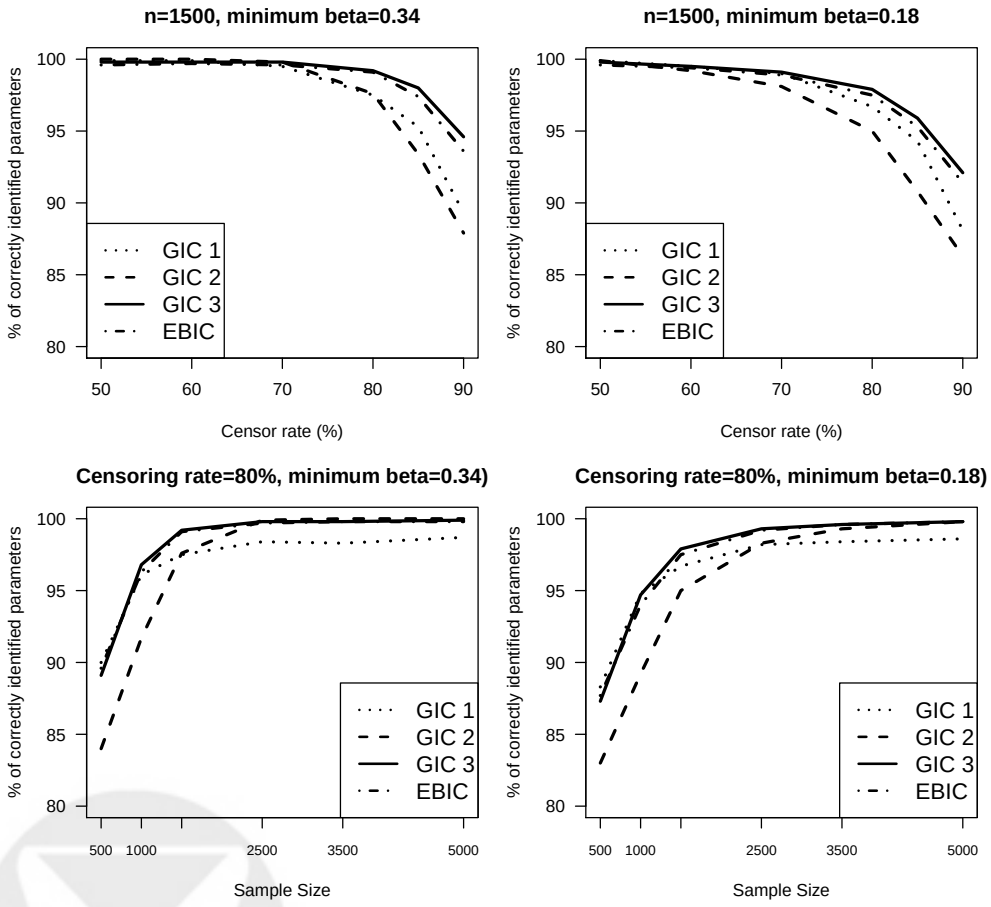

Figure 2. Average percentages of correctly identified parameters (both zero and nonzero) for different choices of a_n in the generalized information criterion. GIC 1: $a_n = 1$; GIC 2: $a_n = \log(n)/2$; GIC 3: $a_n = \log\{\log(d_n)\}\log(d_n)$; EBIC: extended BIC.

Table 1

Model selection performance of different choices of a_n in the generalized information criteria.

	80% Censored					90% Censored				
Method	RME median (MAD)	TN	FN	C	RITM (%)	RME median (MAD)	TN	FN	C	RITM (%)
	$n = 1500$, $\beta_{\min} = 1.0$, $d_n = 31$ for 80% censored, $d_n = 27$ for 90% censored									
HT	0.60 (0.19)	18.8	0.1	29.8	28.8	0.61 (0.23)	17.0	0.4	25.6	24.6
GIC 1	0.41 (0.15)	19.3	0.0	30.3	51.0	0.50 (0.13)	15.7	0.2	24.5	10.2
GIC 2	0.36 (0.15)	19.8	0.1	30.7	78.0	0.46 (0.29)	17.8	0.8	26.0	39.6
GIC 3	0.36 (0.16)	19.9	0.1	30.7	79.0	0.48 (0.31)	17.8	0.8	26.0	38.6
EBIC	0.36 (0.16)	19.8	0.1	30.7	78.8	0.50 (0.34)	17.8	0.9	26.0	37.8
Oracle	0.32 (0.14)	20.0	0.0	31.0	100.0	0.29 (0.14)	18.0	0.0	27.0	100.0
	$n = 1500$, $\beta_{\min} = 0.34$, $d_n = 31$ for 80% censored, $d_n = 27$ for 90% censored									
HT	0.72 (0.17)	18.8	0.1	29.7	29.6	0.81 (0.25)	16.9	0.6	25.2	16.6
GIC 1	0.46 (0.18)	19.3	0.0	30.2	48.0	0.71 (0.21)	15.3	0.2	24.1	8.0
GIC 2	0.56 (0.39)	20.0	0.7	30.3	52.8	3.64 (2.48)	18.0	3.3	23.7	2.0
GIC 3	0.36 (0.18)	19.9	0.1	30.8	80.6	0.86 (0.57)	17.8	1.2	25.5	24.4
EBIC	0.38 (0.17)	19.9	0.1	30.7	78.6	1.04 (0.91)	17.8	1.5	25.3	15.0
Oracle	0.33 (0.14)	20.0	0.0	31.0	100.0	0.29 (0.14)	18.0	0.0	27.0	100.0
	$n = 2500$, $\beta_{\min} = 0.34$, $d_n = 34$ for 80% censored, $d_n = 30$ for 90% censored									
HT	0.71 (0.15)	20.7	0.0	32.7	31.2	0.70 (0.19)	18.9	0.1	28.8	31.2
GIC 1	0.47 (0.18)	21.5	0.0	33.5	60.0	0.63 (0.18)	17.6	0.0	27.5	9.8
GIC 2	0.36 (0.16)	22.0	0.0	34.0	96.8	1.61 (1.25)	20.0	1.6	28.4	20.6
GIC 3	0.37 (0.16)	21.9	0.0	33.9	93.8	0.44 (0.25)	19.9	0.3	29.6	67.2
EBIC	0.38 (0.16)	21.9	0.0	33.9	90.0	0.44 (0.25)	19.8	0.5	29.3	52.6
Oracle	0.36 (0.15)	22.0	0.0	34.0	100.0	0.31 (0.13)	20.0	0.0	30.0	100.0
	$n = 2500$, $\beta_{\min} = 0.18$, $d_n = 34$ for 80% censored, $d_n = 30$ for 90% censored									
HT	0.71 (0.15)	20.7	0.1	32.7	26.0	0.69 (0.19)	18.9	0.4	28.5	21.8
GIC 1	0.49 (0.17)	21.5	0.1	33.4	54.6	0.66 (0.18)	17.6	0.2	27.4	10.6
GIC 2	0.45 (0.21)	22.0	0.6	33.4	47.6	2.27 (1.74)	20.0	2.5	27.5	3.0
GIC 3	0.40 (0.17)	21.9	0.2	33.8	77.2	0.50 (0.28)	19.9	0.8	29.1	35.8
EBIC	0.40 (0.17)	21.9	0.2	33.7	76.0	0.45 (0.27)	19.8	1.1	28.8	20.0
Oracle	0.36 (0.15)	22.0	0.0	34.0	100.0	0.32 (0.14)	20.0	0.0	30.0	100.0
	$n = 5000$, $\beta_{\min} = 0.18$, $d_n = 39$ for 80% censored, $d_n = 34$ for 90% censored									
HT	0.71 (0.18)	24.6	0.0	37.5	24.6	0.67 (0.16)	20.7	0.1	32.6	27.4
GIC 1	0.44 (0.16)	25.5	0.0	38.5	59.2	0.66 (0.18)	19.6	0.0	31.6	9.0
GIC 2	0.37 (0.15)	26.0	0.1	38.9	91.6	0.47 (0.20)	22.0	0.6	33.4	43.8
GIC 3	0.37 (0.15)	25.9	0.0	38.9	93.2	0.40 (0.17)	21.9	0.2	33.8	79.2
EBIC	0.36 (0.14)	25.9	0.0	38.9	92.0	0.41 (0.19)	21.9	0.3	33.6	64.8
Oracle	0.35 (0.14)	26.0	0.0	39.0	100.0	0.37 (0.16)	22.0	0.0	34.0	100.0

RME: estimated relative model error; MAD: median absolute deviation; TN: true negative number (average number of zero parameters correctly identified as zero); FN: false negative number (average number of nonzero parameters incorrectly identified as zero); C: average number of correctly identified parameters (both zero and nonzero); RITM: rate of identifying true model; HT: hard threshold; GIC 1: $a_n = 1$; GIC 2: $a_n = \log(n)/2$; GIC 3: $a_n = \log\{\log(d_n)\}\log(d_n)$; EBIC: extended BIC.

Table 2
Selected genes and estimated coefficients in the breast cancer TCGA data.

Variable	HT $\hat{\beta}$ (ŝe)	GIC 1 $\hat{\beta}$ (ŝe)	GIC 2 $\hat{\beta}$ (ŝe)	GIC 3 $\hat{\beta}$ (ŝe)	EBIC $\hat{\beta}$ (ŝe)
AKT1	0.72 (0.32)	0.54 (0.28)	0 (–)	0 (–)	0 (–)
APC	0 (–)	0.79 (0.35)	0 (–)	0 (–)	0 (–)
BCOR	0 (–)	0.64 (0.32)	0 (–)	0 (–)	0 (–)
CSF3R	0 (–)	0.81 (0.35)	0 (–)	0 (–)	0 (–)
ELF3	−1.22 (0.51)	−1.02 (0.44)	0 (–)	0 (–)	0 (–)
KRAS	0.93 (0.36)	0.81 (0.32)	0 (–)	1.03 (0.28)	1.03 (0.28)
MLH1	0.91 (0.29)	0.96 (0.24)	1.09 (0.23)	1.05 (0.23)	1.05 (0.23)
MPL	0 (–)	1.11 (0.31)	0 (–)	0 (–)	0 (–)
PPP2R1A	0.97 (0.41)	1.01 (0.39)	0 (–)	0 (–)	0 (–)
SDHC	0 (–)	0.38 (0.19)	0 (–)	0 (–)	0 (–)

HT: hard threshold; GIC 1: $a_n = 1$; GIC 2: $a_n = \log(n)/2$; GIC 3: $a_n = \log\{\log(d_n)\}\log(d_n)$; EBIC: extended BIC.

SCANDINAVIAN JOURNAL OF STATISTICS

Supporting Materials for "Tuning Parameter Selection in Cox Proportional Hazards Model with a Diverging Number of Parameters"

Ai Ni

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center

Jianwen Cai

Department of Biostatistics, University of North Carolina at Chapel Hill

Proof of lemma 1. We first consider $\hat{\beta}_{\lambda_0}$, the penalized estimate under the true model. By definition, $\hat{\beta}_{\lambda_0}$ solves the equations

$$\frac{\partial \ell_n(\beta)}{\partial \beta_j} - nP'_{\lambda_0}(|\beta_j|)\text{sgn}(\beta_j) = 0, \qquad j = 1, ..., k_n,$$

where β_j is the jth component of β. Since $\hat{\beta}_{\lambda_0}$ possesses the oracle property, it must follow that $\hat{\beta}_{\lambda_0 j}$ converges to β_{0j} in probability and $\text{pr}\{P'_\lambda(|\hat{\beta}_{\lambda_0 j}|) = 0\} \to 1$. As a result, with probability tending to one, $\hat{\beta}_{\lambda_0}$ solves the equations

$$\frac{\partial \ell_n(\beta)}{\partial \beta_j} = 0, \qquad j = 1, ..., k_n,$$

which are the same equations that the unpenalized estimate $\hat{\beta}_{\alpha_0}$ solves by definition. This implies that $\hat{\beta}_{\lambda_0} = \hat{\beta}_{\alpha_0}$ with probability tending to one. It follows that

$$\text{pr}\{\text{GIC}(\lambda_0) = \text{GIC}^*(\alpha_0)\} \to 1. \tag{1}$$

On the other hand, for any $\lambda \in \Omega$ and any model α_λ, by the definition of $\hat{\beta}_{\alpha_\lambda}$ we have

$$\text{GIC}(\lambda) \geqslant \text{GIC}^*(\alpha_\lambda). \tag{2}$$

Lemma 1 follows from (1) and (2).

The log-partial likelihood function under Cox proportional hazards model can be written as $\ell_n(\beta) = \sum_{i=1}^n \Delta_i \left(\beta^T Z_i(t_i) - \log\left[n^{-1}\sum_{j=1}^n Y_j(t_i)\exp\{\beta^T Z_j(t_i)\}\right]\right)$. Since the log-partial likelihood is a sum of dependent random variables, we introduce the following intermediate function to facilitate the theoretical derivation:

$$\bar{\ell}_n(\beta) = \sum_{i=1}^n \left[\beta^T Z_i(t_i) - \log\{s_n^{(0)}(\beta, t_i)\}\right]\Delta_i,$$

where $s_n^{(0)}(\beta, t)$ is defined in Section 3 of the main text. Define $\text{supp}(\beta)$ as the support of β consisting of indices of its nonzero components. Define set $\mathscr{B}_{\alpha_\lambda} = \{\beta \in \mathscr{B} : \text{ supp}(\beta) = \alpha_\lambda\} \cup \{\beta^0_{\alpha_\lambda}\}$. Then for any $\beta \in \mathscr{B}_{\alpha_\lambda}$ we define $N_{\alpha_\lambda} = \|\beta - \beta^0_{\alpha_\lambda}\|$ and

$$Z_{\alpha_\lambda}(\beta) = \frac{1}{n}\left|\ell_n(\beta) - \ell_n(\beta^0_{\alpha_\lambda}) - E\{\ell_n(\beta) - \ell_n(\beta^0_{\alpha_\lambda})\}\right|.$$

LEMMA 2: *Under Conditions (A) to (G), uniformly for all model* α_λ,

$$\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda| N_{\alpha_\lambda}} Z_{\alpha_\lambda}(\beta) = O_p\left[\left\{\frac{\log(d_n)}{n}\right\}^{1/2}\right].$$

Proof. We first restate a theorem from van de Geer (2008) that will be used in our proofs.

Theorem A.1 in van de Geer (2008) (Bousquet concentration theorem):

Let $X_1, ..., X_n$ *be independent random variables in space* $\mathscr{X}$ *and let* Γ *be a class of real-valued functions on* $\mathscr{X}$ *satisfying for some positive constants* η_n *and* τ_n

$$\|\gamma\|_\infty \leqslant \eta_n \quad \textit{and} \quad \frac{1}{n}\sum_{i=1}^n \mathrm{var}\{\gamma(X_i)\} \leqslant \tau_n^2 \quad \forall \gamma \in \Gamma.$$

Define $Z = \sup_{\gamma\in\Gamma} \left|n^{-1}\sum_{i=1}^n \{\gamma(X_i) - E\gamma(X_i)\}\right|$. *Then for any* $\varepsilon > 0$,

$$\mathrm{pr}\left[Z \geqslant EZ + \varepsilon\left\{2(\tau_n^2 + 2\eta_n EZ)\right\}^{1/2} + \frac{2\varepsilon^2\eta_n}{3}\right] \leqslant \exp(-n\varepsilon^2).$$

We begin by introducing the following two intermediate quantities:

$$Q_{\alpha_\lambda}(\beta) = \frac{1}{n}\left|\bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda}) - E\{\bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda})\}\right|,$$
$$R_{\alpha_\lambda}(\beta) = \frac{1}{n}\left|\ell_n(\beta) - \ell_n(\beta^0_{\alpha_\lambda}) - \{\bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda})\}\right|.$$

It is easy to see that $Z_{\alpha_\lambda}(\beta) \leqslant Q_{\alpha_\lambda}(\beta) + R_{\alpha_\lambda}(\beta) + E\{R_{\alpha_\lambda}(\beta)\}$.

We will study the tail probabilities of the above two quantities separately.

To use Theorem A.1 in van de Geer (2008) to establish a probability bound for $Q_{\alpha_\lambda}(\beta)$, we first derive a bound for $E\{Q_{\alpha_\lambda}(\beta)\}$. Let $\epsilon_1, ..., \epsilon_n$ be a Rademacher sequence, independent of the random variables $\bar{\ell}_1(\beta) - \bar{\ell}_1(\beta^0_{\alpha_\lambda}), ..., \bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda})$. By symmetrization theorem presented in Lemma 2.3.1 of van der Vaart & Wellner (1996) with $\mathcal{F}$ being a class of only the identity function, we have

$$E\{Q_{\alpha_\lambda}(\beta)\} = \frac{1}{n}E\left|\bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda}) - E\{\bar{\ell}_n(\beta) - \bar{\ell}_n(\beta^0_{\alpha_\lambda})\}\right| \leqslant \frac{2}{n}E\left|\sum_{i=1}^n \epsilon_i\{\bar{\ell}_i(\beta) - \bar{\ell}_i(\beta^0_{\alpha_\lambda})\}\right|$$
$$= \frac{2}{n}E\left|\sum_{i=1}^n \epsilon_i\left(\left[\beta^T Z_i(t_i) - \log\left\{s_n^{(0)}(\beta, t_i)\right\}\right]\Delta_i - \left[(\beta^0_{\alpha_\lambda})^T Z_i(t_i) - \log\left\{s_n^{(0)}(\beta^0_{\alpha_\lambda}, t_i)\right\}\right]\Delta_i\right)\right|$$

$$\leqslant \frac{2}{n}E\left|\sum_{i=1}^{n}\epsilon_i\left\{\beta^T Z_i(t_i)-(\beta^0_{\alpha_\lambda})^T Z_i(t_i)\right\}\Delta_i\right|+\frac{2}{n}E\left|\sum_{i=1}^{n}\epsilon_i\left\{\log s_n^{(0)}(\beta,t_i)-\log s_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)\right\}\Delta_i\right|$$

$$= I_1 + I_2.$$

We first consider I_1. By Cauchy–Schwarz inequality and $E(\epsilon)=0$,

$$I_1=\frac{2}{n}E\left|\sum_{i=1}^{n}\epsilon_i\left\{\sum_{j=1}^{|\alpha_\lambda|}(\beta_j-\beta^0_{\alpha_\lambda j})Z_{ij}(t_i)\right\}\Delta_i\right|=\frac{2}{n}E\left|\sum_{j=1}^{|\alpha_\lambda|}\left\{(\beta_j-\beta^0_{\alpha_\lambda j})\sum_{i=1}^{n}\epsilon_i Z_{ij}(t_i)\Delta_i\right\}\right|$$

$$\leqslant \frac{2}{n}\|\beta-\beta^0_{\alpha_\lambda}\|E\left[\sum_{j=1}^{|\alpha_\lambda|}\left\{\sum_{i=1}^{n}\epsilon_i Z_{ij}(t_i)\Delta_i\right\}^2\right]^{1/2}\leqslant \frac{2}{n}\|\beta-\beta^0_{\alpha_\lambda}\|\left[\sum_{j=1}^{|\alpha_\lambda|}E\left\{\sum_{i=1}^{n}\epsilon_i Z_{ij}(t_i)\Delta_i\right\}^2\right]^{1/2}$$

$$\leqslant \frac{2}{n}\|\beta-\beta^0_{\alpha_\lambda}\|\left(\sum_{j=1}^{|\alpha_\lambda|}\left[\sum_{i=1}^{n}E\left\{\epsilon_i Z_{ij}(t_i)\Delta_i\right\}^2+\sum_{i=1}^{n}\sum_{k=1}^{n}E\left\{\epsilon_i Z_{ij}(t_i)\Delta_i\epsilon_k Z_{kj}(t_k)\Delta_k\right\}\right]\right)^{1/2}$$

$$\leqslant 2N_{\alpha_\lambda}|\alpha_\lambda|^{1/2}n^{-1/2}K_n.$$

Next we consider I_2. Due to its lack of Lipschitz property, we cannot study its properties uniformly for β as in van de Geer (2008). We instead study its pointwise property for any given β by mean value theorem. For some $\beta^*_{\alpha_\lambda}$ that lies between $\beta^0_{\alpha_\lambda}$ and β,

$$I_2=\frac{2}{n}E\left|\sum_{i=1}^{n}\epsilon_i\Delta_i\sum_{j=1}^{|\alpha_\lambda|}(\beta_j-\beta^0_{\alpha_\lambda j})\frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t_i)}\right|=\frac{2}{n}E\left|\sum_{j=1}^{|\alpha_\lambda|}(\beta_j-\beta^0_{\alpha_\lambda j})\sum_{i=1}^{n}\epsilon_i\Delta_i\frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t_i)}\right|,$$

where $s_{nj}^{(1)}(\beta,t)$ denotes the j-th component of $s_n^{(1)}(\beta,t)$, which is defined in Section 3 of the main text. By the definition of $s_{nj}^{(1)}(\beta,t)$ we have that

$$s_{nj}^{(1)}(\beta,t)=E\left[Y(t)Z_j(t)\exp\{\beta^T Z(t)\}\right]\leqslant K_n E\left[Y(t)\exp\{\beta^T Z(t)\}\right]=K_n s_n^{(0)}(\beta,t).$$

By Cauchy–Schwarz inequality and $E(\epsilon)=0$,

$$I_2\leqslant \frac{2}{n}\|\beta-\beta^0_{\alpha_\lambda}\|E\left[\sum_{j=1}^{|\alpha_\lambda|}\left\{\sum_{i=1}^{n}\epsilon_i\Delta_i\frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t_i)}\right\}^2\right]^{1/2}$$

$$\leqslant \frac{2}{n}\|\beta-\beta^0_{\alpha_\lambda}\|\left[\sum_{j=1}^{|\alpha_\lambda|}E\left\{\sum_{i=1}^{n}\epsilon_i\Delta_i\frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t_i)}\right\}^2\right]^{1/2}$$

$$\leqslant \frac{2}{n}\|\beta-\beta^0_{\alpha_\lambda}\|\left(\sum_{j=1}^{|\alpha_\lambda|}\left[\sum_{i=1}^{n}E\left\{\epsilon_i\Delta_i\frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t_i)}\right\}^2\right.\right.$$

$$+\sum_{i=1}^{n}\sum_{k=1}^{n} E\left\{\epsilon_i\Delta_i\frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t_i)}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t_i)}\epsilon_k\Delta_k\frac{s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t_k)}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t_k)}\right\}\Bigg]\Bigg)^{1/2}$$

$$\leqslant 2N_{\alpha_\lambda}|\alpha_\lambda|^{1/2}n^{-1/2}K_n.$$

It follows that $E\{Q_{\alpha_\lambda}(\beta)\} \leqslant I_1 + I_2 \leqslant 4N_{\alpha_\lambda}|\alpha_\lambda|^{1/2}n^{-1/2}K_n$.

Now we check the two conditions for Theorem A.1 in van de Geer (2008). By Cauchy-Schwarz inequality and mean value theorem, for all i we have

$$\left|\bar{\ell}_i(\beta)-\bar{\ell}_i(\beta^0_{\alpha_\lambda})\right| \leqslant \left|\beta^T Z_i(t_i)-(\beta^0_{\alpha_\lambda})^T Z_i(t_i)\right|\Delta_i + \left|\log s_n^{(0)}(\beta,t_i)-\log s_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)\}\right|\Delta_i$$

$$\leqslant |\alpha_\lambda|^{1/2}\|\beta-\beta^0_{\alpha_\lambda}\|K_n + \|\beta-\beta^0_{\alpha_\lambda}\|\frac{\left\{\sum_{j=1}^{|\alpha_\lambda|}K_n^2 s_n^{(0)}(\beta^*_{\alpha_\lambda},t_i)^2\right\}^{1/2}}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t_i)}$$

$$\leqslant 2|\alpha_\lambda|^{1/2}N_{\alpha_\lambda}K_n.$$

Thus $\|\bar{\ell}_i(\beta)-\bar{\ell}_i(\beta^0_{\alpha_\lambda})\|_\infty \leqslant 2|\alpha_\lambda|^{1/2}N_{\alpha_\lambda}K_n$ and $\mathrm{var}\{\bar{\ell}_i(\beta)-\bar{\ell}_i(\beta^0_{\alpha_\lambda})\} \leqslant E\{\bar{\ell}_i(\beta)-\bar{\ell}_i(\beta^0_{\alpha_\lambda})\}^2 \leqslant 4|\alpha_\lambda|N^2_{\alpha_\lambda}K_n^2$. Let $\eta_n = 2|\alpha_\lambda|^{1/2}N_{\alpha_\lambda}K_n$ and $\tau_n^2 = 4|\alpha_\lambda|N^2_{\alpha_\lambda}K_n^2$. Then by Theorem A.1 in van de Geer (2008) with $X_i = \bar{\ell}_n(\beta)-\bar{\ell}_n(\beta^0_{\alpha_\lambda})$, γ being the identity function, and $\Gamma = \{\gamma\}$, for any $\varepsilon > 0$,

$$\mathrm{pr}\left[Q_{\alpha_\lambda}(\beta) \geqslant \frac{4N_{\alpha_\lambda}|\alpha_\lambda|^{1/2}K_n}{n^{1/2}} + \varepsilon\left\{2(4|\alpha_\lambda|N^2_{\alpha_\lambda}K_n^2 + \frac{16|\alpha_\lambda|N^2_{\alpha_\lambda}K_n^2}{n^{1/2}}\right\}^{1/2} + \frac{4\varepsilon^2|\alpha_\lambda|^{1/2}N_{\alpha_\lambda}K_n}{3}\right]$$

$$= \mathrm{pr}\left[Q_{\alpha_\lambda}(\beta) \geqslant 2|\alpha_\lambda|^{1/2}N_{\alpha_\lambda}K_n\left\{\frac{2}{n^{1/2}} + \varepsilon(2+\frac{8}{n^{1/2}})^{1/2} + \frac{2\varepsilon^2}{3}\right\}\right] \leqslant \exp(-n\varepsilon^2). \quad (3)$$

Next we consider $R_{\alpha_\lambda}(\beta)$. By mean value theorem, for some $\beta^*_{\alpha_\lambda}$ that lies between $\beta^0_{\alpha_\lambda}$ and β we have that

$$R_{\alpha_\lambda}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Bigg|\Bigg(\log\left[\frac{1}{n}\sum_{j=1}^{n}\frac{Y_j(t_i)\exp\{\beta^T Z_j(t_i)\}}{s_n^{(0)}(\beta,t_i)}\right]$$

$$-\log\left[\frac{1}{n}\sum_{j=1}^{n}\frac{Y_j(t_i)\exp\{(\beta^0_{\alpha_\lambda})^T Z_j(t_i)\}}{s_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)}\right]\Bigg)\Delta_i\Bigg|$$

$$\leqslant \sup_{0\leqslant t\leqslant\tau}\left|\log\left\{\frac{S_n^{(0)}(\beta,t)}{s_n^{(0)}(\beta,t)}\right\} - \log\left\{\frac{S_n^{(0)}(\beta^0_{\alpha_\lambda},t)}{s_n^{(0)}(\beta^0_{\alpha_\lambda},t)}\right\}\right|$$

$$= \sup_{0\leqslant t\leqslant\tau}\left|(\beta-\beta^0_{\alpha_\lambda})^T\left\{\frac{S_n^{(1)}(\beta^*_{\alpha_\lambda},t)}{S_n^{(0)}(\beta^*_{\alpha_\lambda},t)} - \frac{s_n^{(1)}(\beta^*_{\alpha_\lambda},t)}{s_n^{(0)}(\beta^*_{\alpha_\lambda},t)}\right\}\right.$$

$$
\begin{aligned}
&\leqslant \sup_{0\leqslant t\leqslant\tau} \|\beta-\beta^0_{\alpha_\lambda}\| \left[\sum_{j=1}^{|\alpha_\lambda|}\left\{\frac{S^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)}{S^{(0)}_n(\beta^*_{\alpha_\lambda},t)}-\frac{s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)}{s^{(0)}_n(\beta^*_{\alpha_\lambda},t)}\right\}^2\right]^{1/2}\\
&= \sup_{0\leqslant t\leqslant\tau} \|\beta-\beta^0_{\alpha_\lambda}\| \left\{\sum_{j=1}^{|\alpha_\lambda|}\left(\frac{1}{S^{(0)}_n(\beta^*_{\alpha_\lambda},t)}\left[S^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)-s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)\right.\right.\right.\\
&\quad \left.\left.\left.+\frac{s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)}{s^{(0)}_n(\beta^*_{\alpha_\lambda},t)}\{s^{(0)}_n(\beta^*_{\alpha_\lambda},t)-S^{(0)}_n(\beta^*_{\alpha_\lambda},t)\}\right]\right)^2\right\}^{1/2}\\
&\leqslant \sup_{0\leqslant t\leqslant\tau} \|\beta-\beta^0_{\alpha_\lambda}\| \left\{\sum_{j=1}^{|\alpha_\lambda|}\left(\frac{1}{S^{(0)}_n(\beta^*_{\alpha_\lambda},t)}\left[\max_{1\leqslant j\leqslant|\alpha_\lambda|}\left|S^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)-s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)\right|\right.\right.\right.\\
&\quad \left.\left.\left.+K_n\left|S^{(0)}_n(\beta^*_{\alpha_\lambda},t)-s^{(0)}_n(\beta^*_{\alpha_\lambda},t)\right|\right]\right)^2\right\}^{1/2}\\
&= \sup_{0\leqslant t\leqslant\tau} \|\beta-\beta^0_{\alpha_\lambda}\||\alpha_\lambda|^{1/2}\frac{1}{S^{(0)}_n(\beta^*_{\alpha_\lambda},t)}\left\{\max_{1\leqslant j\leqslant|\alpha_\lambda|}\left|S^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)-s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)\right|\right.\\
&\quad \left.+K_n\left|S^{(0)}_n(\beta^*_{\alpha_\lambda},t)-s^{(0)}_n(\beta^*_{\alpha_\lambda},t)\right|\right\}\\
&\leqslant \|\beta-\beta^0_{\alpha_\lambda}\||\alpha_\lambda|^{1/2}\sup_{0\leqslant t\leqslant\tau}\frac{1}{S^{(0)}_n(\beta^*_{\alpha_\lambda},t)}\sup_{0\leqslant t\leqslant\tau}\left\{\max_{1\leqslant j\leqslant|\alpha_\lambda|}\left|S^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)-s^{(1)}_{nj}(\beta^*_{\alpha_\lambda},t)\right|\right.\\
&\quad \left.+K_n\left|S^{(0)}_n(\beta^*_{\alpha_\lambda},t)-s^{(0)}_n(\beta^*_{\alpha_\lambda},t)\right|\right\}. \qquad (4)
\end{aligned}
$$

We first bound $\sup_{0\leqslant t\leqslant\tau}\{S^{(0)}_n(\beta^*_{\alpha_\lambda},t)\}^{-1}$. By Condition (F) we have

$$\inf_{\beta,Z(t)} S^{(0)}_n(\beta,t)=\frac{1}{n}\sum_{i=1}^n Y_i(t)\exp\{-\sup_{\beta,Z_i(t)}\beta^TZ_i(t)\}=U_n^{-1}\frac{1}{n}\sum_{i=1}^n Y_i(t).$$

Since $Y(t)$ is a non-increasing function of t, we have that

$$\inf_{0\leqslant t\leqslant\tau} S^{(0)}_n(\beta^*_{\alpha_\lambda},t)\geqslant U_n^{-1}\frac{1}{n}\sum_{i=1}^n Y_i(\tau),$$

and therefore

$$\sup_{0\leqslant t\leqslant\tau}\frac{1}{S^{(0)}_n(\beta^*_{\alpha_\lambda},t)}\leqslant U_n\left\{\frac{1}{n}\sum_{i=1}^n Y_i(\tau)\right\}^{-1}.$$

Define $\mu=E\{Y(\tau)\}$. By Lemma 2 in Kong & Nan (2014),

$$\mathrm{pr}\left\{\frac{1}{n}\sum_{i=1}^n Y_i(\tau)\leqslant\frac{\mu}{2}\right\}=\mathrm{pr}\left[\left\{\frac{1}{n}\sum_{i=1}^n Y_i(\tau)\right\}^{-1}\geqslant\frac{2}{\mu}\right]\leqslant 2\exp\left(-\frac{n\mu^2}{2}\right).$$

Therefore,

$$\mathrm{pr}\left\{\sup_{0\leqslant t\leqslant\tau}\frac{1}{S_n^{(0)}(\beta^*_{\alpha_\lambda},t)}\geqslant\frac{2U_n}{\mu}\right\}\leqslant 2\exp\left(-\frac{n\mu^2}{2}\right).$$

By a modification of Lemma 3 and 4 in Kong & Nan (2014) we have for any positive constant ε,

$$\mathrm{pr}\left\{\sup_{0\leqslant t\leqslant\tau}\left|S_n^{(0)}(\beta^*_{\alpha_\lambda},t)-s_n^{(0)}(\beta^*_{\alpha_\lambda},t)\right|\geqslant U_n\varepsilon\right\}\leqslant\frac{1}{5}W^2\exp(-n\varepsilon^2), \tag{5}$$

$$\mathrm{pr}\left\{\sup_{0\leqslant t\leqslant\tau}\max_{1\leqslant j\leqslant|\alpha_\lambda|}\left|S_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)-s_{nj}^{(1)}(\beta^*_{\alpha_\lambda},t)\right|\geqslant U_nK_n\varepsilon\right\}\leqslant\frac{1}{5}|\alpha_\lambda|W^2\exp(-n\varepsilon^2),$$

where W is a constant determined by the bracketing number of the class of functions indexed by t, $\mathscr{F}=\{Y(t)\exp\{\beta^TZ(t)\}U_n^{-1}:t\in[0,\tau],\exp\{\beta^TZ(t)\}\leqslant U_n\}$. Applying these results to (4) we have

$$\mathrm{pr}\left\{R_{\alpha_\lambda}(\beta)\geqslant\frac{2N_{\alpha_\lambda}|\alpha_\lambda|^{1/2}U_n^2K_n\varepsilon}{\mu}\right\}\leqslant 2\exp\left(-\frac{n\mu^2}{2}\right)+\frac{1}{5}(|\alpha_\lambda|+1)W^2\exp(-n\varepsilon^2). \tag{6}$$

Since $Z_{\alpha_\lambda}(\beta)\leqslant Q_{\alpha_\lambda}(\beta)+R_{\alpha_\lambda}(\beta)+E\{R_{\alpha_\lambda}(\beta)\}$, by (3) and (6) we have that

$$\begin{aligned}&\mathrm{pr}\left[Z_{\alpha_\lambda}(\beta)\geqslant 2N_{\alpha_\lambda}K_n|\alpha_\lambda|^{1/2}\left\{\frac{2}{n^{1/2}}+\varepsilon(2+\frac{8}{n^{1/2}})^{1/2}+\frac{2\varepsilon^2}{3}+\frac{U_n^2\varepsilon}{\mu}\right\}+E\{R_{\alpha_\lambda}(\beta)\}\right]\\&\leqslant 2\exp\left(-\frac{n\mu^2}{2}\right)+\left\{\frac{1}{5}(|\alpha_\lambda|+1)W^2+1\right\}\exp(-n\varepsilon^2).\end{aligned} \tag{7}$$

To establish the stochastic order of random sequences, we use the following result: for any random sequence X_n, a_n, b_n and any diverging constant sequence γ_n, $\mathrm{pr}(X_n\geqslant a_n+b_n\gamma_n)=o(1)$ implies that $X_n=O_p(a_n+b_n)$. Let $\varepsilon=n^{-1/2}\gamma_n$, where γ_n is any diverging sequence. Then (7) becomes

$$\begin{aligned}&\mathrm{pr}\left[Z_{\alpha_\lambda}(\beta)\geqslant 2N_{\alpha_\lambda}K_n|\alpha_\lambda|^{1/2}\left\{\frac{2}{n^{1/2}}+\frac{\gamma_n}{n^{1/2}}(2+\frac{8}{n^{1/2}})^{1/2}+\frac{2\gamma_n^2}{3n}+\frac{U_n^2\gamma_n}{n^{1/2}\mu}\right\}+E\{R_{\alpha_\lambda}(\beta)\}\right]\\&\leqslant 2\exp\left(-\frac{n\mu^2}{2}\right)+\left\{\frac{1}{5}(|\alpha_\lambda|+1)W^2+1\right\}\exp(-\gamma_n^2).\end{aligned} \tag{8}$$

Using the same method on (6) we get

$$\mathrm{pr}\left\{\frac{R_{\alpha_\lambda}(\beta)\mu n^{1/2}}{2N_{\alpha_\lambda}|\alpha_\lambda|^{1/2}U_n^2K_n}\geqslant\gamma_n\right\}\leqslant 2\exp\left(-\frac{n\mu^2}{2}\right)+\frac{1}{5}(|\alpha_\lambda|+1)W^2\exp(-\gamma_n^2).$$

From this tail inequality we can verify that $E\{R_{\alpha_\lambda}(\beta)\mu n^{1/2}N_{\alpha_\lambda}^{-1}|\alpha_\lambda|^{-1/2}U_n^{-2}K_n^{-1}/2\}<\infty$. Therefore, $E\{R_{\alpha_\lambda}(\beta)\}=N_{\alpha_\lambda}|\alpha_\lambda|^{1/2}O(n^{-1/2})$. Then from (8) it follows that $Z_{\alpha_\lambda}(\beta)N_{\alpha_\lambda}^{-1}|\alpha_\lambda|^{-1/2}=O_p(n^{-1/2})$.

Now we derive the probability bound for the supremum of $Z_{\alpha_\lambda}(\beta)$ over all possible models. Let $\varepsilon = \{|\alpha_\lambda| \log(d_n)\}^{1/2} n^{-1/2} \gamma_n$ in (7), where γ_n is any diverging sequence. Then,

$$\begin{aligned}
&\mathrm{pr}\left(Z_{\alpha_\lambda}(\beta) \geqslant \frac{2N_{\alpha_\lambda} K_n |\alpha_\lambda|}{n^{1/2}} \left[2|\alpha_\lambda|^{-1/2} + \gamma_n \left\{ \log(d_n)(2 + 8n^{-1/2}) \right\}^{1/2} + \frac{2|\alpha_\lambda|^{1/2} \log(d_n) \gamma_n^2}{3n^{1/2}} \right.\right. \\
&\quad \left.\left. + \frac{U_n^2 \{\log(d_n)\}^{1/2} \gamma_n}{\mu} + \frac{E\{R_{\alpha_\lambda}(\beta)\} n^{1/2}}{N_{\alpha_\lambda} |\alpha_\lambda|^{1/2}} \right] \right) \\
&\leqslant 2 \exp\left(-\frac{n\mu^2}{2} \right) + \left\{ \frac{1}{5}(|\alpha_\lambda| + 1) W^2 + 1 \right\} \exp\{ -|\alpha_\lambda| \log(d_n) \gamma_n^2 \}.
\end{aligned}$$

We use the fact that

$$\binom{d_n}{k} \leqslant (d_n e / k)^k, \quad 0 < k \leqslant d_n, \tag{9}$$

where e is the Euler's number, in the following derivation.

$$\begin{aligned}
&\mathrm{pr}\left(\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda| N_{\alpha_\lambda}} Z_{\alpha_\lambda}(\beta) \geqslant \frac{2K_n}{n^{1/2}} \left[2|\alpha_\lambda|^{-1/2} + \gamma_n \left\{ \log(d_n)(2 + 8n^{-1/2}) \right\}^{1/2} + \frac{2|\alpha_\lambda|^{1/2} \log(d_n) \gamma_n^2}{3n^{1/2}} \right.\right. \\
&\quad \left.\left. + \frac{U_n^2 \{\log(d_n)\}^{1/2} \gamma_n}{\mu} + \frac{E\{R_{\alpha_\lambda}(\beta)\} n^{1/2}}{N_{\alpha_\lambda} |\alpha_\lambda|^{1/2}} \right] \right) \\
&\leqslant \sum_{k=1}^{d_n} \sum_{|\alpha_\lambda| = k} \mathrm{pr}\left(Z_{\alpha_\lambda}(\beta) \geqslant \frac{2N_{\alpha_\lambda} K_n |\alpha_\lambda|}{n^{1/2}} \left[2|\alpha_\lambda|^{-1/2} + \gamma_n \left\{ \log(d_n)(2 + 8n^{-1/2}) \right\}^{1/2} \right.\right. \\
&\quad \left.\left. + \frac{2|\alpha_\lambda|^{1/2} \log(d_n) \gamma_n^2}{3n^{1/2}} + \frac{U_n^2 \{\log(d_n)\}^{1/2} \gamma_n}{\mu} + \frac{E\{R_{\alpha_\lambda}(\beta)\} n^{1/2}}{N_{\alpha_\lambda} |\alpha_\lambda|^{1/2}} \right] \right) \\
&\leqslant \sum_{k=1}^{d_n} \binom{d_n}{k} \left[2 \exp\left(-\frac{n\mu^2}{2} \right) + \left\{ \frac{1}{5}(k + 1) W^2 + 1 \right\} \exp\{ -k \log(d_n) \gamma_n^2 \} \right] \\
&\leqslant \sum_{k=1}^{d_n} \left(\frac{d_n e}{k} \right)^k \left[2 \exp\left(-\frac{n\mu^2}{2} \right) + \left\{ \frac{1}{5}(k + 1) W^2 + 1 \right\} \exp\{ -k \log(d_n) \gamma_n^2 \} \right] \\
&= \sum_{k=1}^{d_n} \left(\frac{e}{k} \right)^k \left[2 d_n^k \exp\left(-\frac{n\mu^2}{2} \right) + \left\{ \frac{1}{5}(k + 1) W^2 + 1 \right\} d_n^{(1 - \gamma_n^2) k} \right].
\end{aligned} \tag{10}$$

By Condition (G), $\{(d_n + 1) \log(d_n)/n\} = o(1)$. Thus $d_n^{d_n + 1} = o\{\exp(n)\}$ and the first term in the square brackets in (10) is $o(d_n^{-1})$. Since γ_n diverges to infinity, the second term in the square brackets in (10) is also $o(d_n^{-1})$. Moreover, $(e/k)^k < 1$ for all $k \geqslant 3$. Therefore, it is easy to see that (10) goes to 0 as $n \to \infty$. It follows that

$$\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda| N_{\alpha_\lambda}} Z_{\alpha_\lambda}(\beta) = O_p\left[\frac{2K_n}{n^{1/2}} \left(2|\alpha_\lambda|^{-1/2} + \gamma_n \left\{ \log(d_n)(2 + 8n^{-1/2}) \right\}^{1/2} + \frac{2|\alpha_\lambda|^{1/2} \log(d_n) \gamma_n^2}{3n^{1/2}} \right.\right.$$

$$+\frac{U_n^2\{\log(d_n)\}^{1/2}\gamma_n}{\mu}+\frac{E\{R_{\alpha_\lambda}(\beta)\}n^{1/2}}{N_{\alpha_\lambda}|\alpha_\lambda|^{1/2}}\Big)\Big]=O_p\left[\left\{\frac{\log(d_n)}{n}\right\}^{1/2}\right].$$

LEMMA 3: *Under Conditions (A) to (G), uniformly for all model* α_λ*,*

$$\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|}\big\|\hat{\beta}_{\alpha_\lambda}-\beta^0_{\alpha_\lambda}\big\|=O_p\left[\left\{\frac{\log(d_n)}{n}\right\}^{1/2}\right].$$

Proof. Denote $\|\hat{\beta}_{\alpha_\lambda}-\beta^0_{\alpha_\lambda}\|=N_{\alpha_\lambda}$. Since $\hat{\beta}_{\alpha_\lambda}$ maximizes $\ell_n(\beta_{\alpha_\lambda})$, $\ell_n(\beta^0_{\alpha_\lambda})\leqslant\ell_n(\hat{\beta}_{\alpha_\lambda})$. Since $\beta^0_{\alpha_\lambda}$ minimizes the Kullback-Leibler distance, $E\{\ell_n(\beta^0_{\alpha_\lambda})\}\geqslant E\{\ell_n(\hat{\beta}_{\alpha_\lambda})\}$ and $\partial E\{\ell_n(\beta^0_{\alpha_\lambda})\}/\partial\beta=0$, where the expectation is taken under the true model. It follows that,

$$0\leqslant E\{\ell_n(\beta^0_{\alpha_\lambda})-\ell_n(\hat{\beta}_{\alpha_\lambda})\}\leqslant\ell_n(\hat{\beta}_{\alpha_\lambda})-E\{\ell_n(\hat{\beta}_{\alpha_\lambda})-\big[\ell_n(\beta^0_{\alpha_\lambda})-E\{\ell_n(\beta^0_{\alpha_\lambda})\}\big]\leqslant nZ_{\alpha_\lambda}(\hat{\beta}_{\alpha_\lambda}).\tag{11}$$

By Taylor expansion, for some $\beta^*_{\alpha_\lambda}$ that lies between $\hat{\beta}_{\alpha_\lambda}$ and $\beta^0_{\alpha_\lambda}$ we have that

$$E\{\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\beta^0_{\alpha_\lambda})\}=-\frac{n}{2}(\hat{\beta}_{\alpha_\lambda}-\beta^0_{\alpha_\lambda})^TI_n(\beta^*_{\alpha_\lambda})(\hat{\beta}_{\alpha_\lambda}-\beta^0_{\alpha_\lambda})\leqslant-\frac{n}{2}N^2_{\alpha_\lambda}C_3.\tag{12}$$

The last inequality in (12) hold by spectral decomposition on $I_n(\beta^*_{\alpha_\lambda})$ and Condition (E). By (11) and (12) we have that $N_{\alpha_\lambda}\leqslant 2Z_{\alpha_\lambda}(\hat{\beta}_{\alpha_\lambda})N^{-1}_{\alpha_\lambda}C_3^{-1}$. In the proof of Lemma 2 we have shown that $Z_{\alpha_\lambda}(\beta)N^{-1}_{\alpha_\lambda}=O_p(|\alpha_\lambda|^{1/2}n^{-1/2})$. It follows that $N_{\alpha_\lambda}=O_p(|\alpha_\lambda|^{1/2}n^{-1/2})$. Furthermore, by dividing both sides of the inequality $N_{\alpha_\lambda}\leqslant 2Z_{\alpha_\lambda}(\hat{\beta}_{\alpha_\lambda})\}N^{-1}_{\alpha_\lambda}C_3^{-1}$ by $|\alpha_\lambda|$ and taking supremum we arrive at

$$\sup_{\alpha_\lambda}\frac{N_{\alpha_\lambda}}{|\alpha_\lambda|}=\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|}\big\|\hat{\beta}_{\alpha_\lambda}-\beta^0_{\alpha_\lambda}\big\|=\sup_{\alpha_\lambda}\frac{2C_3^{-1}}{|\alpha_\lambda|N_{\alpha_\lambda}}Z_{\alpha_\lambda}(\hat{\beta}_{\alpha_\lambda})\}=O_p\left[\left\{\frac{\log(d_n)}{n}\right\}^{1/2}\right].$$

The last equality holds by Lemma 2.

LEMMA 4: *Under Conditions (A) to (G), uniformly for all model* α_λ*,*

$$\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|^{3/2}}\Big|\ell_n(\hat{\beta}_{\alpha_\lambda})-\ell_n(\beta^0_{\alpha_\lambda})\Big|=O_p[\{\log(d_n)\}^{1/2}].$$

Proof. By the definition of $\hat{\beta}_{\alpha_\lambda}$ and $\beta^0_{\alpha_\lambda}$, we have that $\ell_n(\beta^0_{\alpha_\lambda})\leqslant\ell_n(\hat{\beta}_{\alpha_\lambda})$ and $E\{\ell_n(\beta^0_{\alpha_\lambda})\}\geqslant$

$E\{\ell_n(\hat{\beta}_{\alpha_\lambda})\}$ for any model α_λ. Thus,

$$\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda}) \leqslant \ell_n(\hat{\beta}_{\alpha_\lambda}) - E\{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \left[\ell_n(\beta^0_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\}\right] \leqslant nZ_{\alpha_\lambda}(\hat{\beta}_{\alpha_\lambda}).$$

Define $N_{\alpha_\lambda} = \|\hat{\beta}_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\|$. By Lemma 2 we have

$$\sup_{\alpha_\lambda} \frac{\left|\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda})\right|}{|\alpha_\lambda| N_{\alpha_\lambda}} \leqslant \sup_{\alpha_\lambda} \frac{nZ_{\alpha_\lambda}(\hat{\beta}_{\alpha_\lambda})}{|\alpha_\lambda| N_{\alpha_\lambda}} = O[\{n\log(d_n)\}^{1/2}].$$

In the proof of Lemma 3 we have established that $N_{\alpha_\lambda} = O_p(|\alpha_\lambda|^{1/2} n^{-1/2})$ for any α_λ. It follows that $\sup_{\alpha_\lambda} |\alpha_\lambda|^{-3/2} |\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda})| = O_p[\{\log(d_n)\}^{1/2}]$.

LEMMA 5: *Under Conditions (A) to (G), uniformly for all model α_λ,*

$$\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}} \left|\ell_n(\beta^0_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\}\right| = O_p\left[\{n\log(d_n)\}^{1/2}\right].$$

Proof. Since $\ell_n(\beta^0_{\alpha_\lambda})$ is a sum of dependent random variables, we decompose the quantity in the statement of the lemma as follows,

$$\begin{aligned}
&\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}} \left|\ell_n(\beta^0_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\}\right| \\
&\leqslant \sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}} \left\{\left|\ell_n(\beta^0_{\alpha_\lambda}) - \bar{\ell}_n(\beta^0_{\alpha_\lambda})\right| + \left|\bar{\ell}_n(\beta^0_{\alpha_\lambda}) - E\{\bar{\ell}_n(\beta^0_{\alpha_\lambda})\}\right| + E\left|\ell_n(\beta^0_{\alpha_\lambda}) - \bar{\ell}_n(\beta^0_{\alpha_\lambda})\right|\right\} \\
&= \sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}} \{I_1 + I_2 + E(I_1)\}.
\end{aligned}$$

We first consider I_1. By mean value theorem,

$$I_1 \leqslant n \sup_{0\leqslant t\leqslant\tau} \left|\log\{S^{(0)}_n(\beta^0_{\alpha_\lambda}, t)\} - \log\{s^{(0)}_n(\beta^0_{\alpha_\lambda}, t)\}\right| = \sup_{0\leqslant t\leqslant\tau} \left|\frac{n}{S^*_n}\{S^{(0)}_n(\beta^0_{\alpha_\lambda}, t) - s^{(0)}_n(\beta^0_{\alpha_\lambda}, t)\}\right|, \tag{13}$$

where S^*_n lies between $S^{(0)}_n(\beta^0_{\alpha_\lambda}, t)$ and $s^{(0)}_n(\beta^0_{\alpha_\lambda}, t)$. It follows from (5) that $S^{(0)}_n(\beta^0_{\alpha_\lambda}, t)$ converges to $s^{(0)}_n(\beta^0_{\alpha_\lambda}, t)$ in probability uniformly on $t \in [0, \tau]$, and so does S^*_n. By Condition (D), $s^{(0)}_n(\beta^0_{\alpha_\lambda}, t)$ is uniformly bounded away from 0. Let C_5 be a constant satisfying $0 < C_5 < \inf_{0\leqslant t\leqslant\tau} s^{(0)}_n(\beta^0_{\alpha_\lambda}, t)$. Define the event $\mathscr{A}_n = \{S^*_n > C_5\}$. Denote $\mathscr{A}^c_n$ as the complement of $\mathscr{A}$. Consider

$$\begin{aligned}
&\mathrm{pr}\left[\sup_{0\leqslant t\leqslant\tau} \left|\log\{S^{(0)}_n(\beta^0_{\alpha_\lambda}, t)\} - \log\{s^{(0)}_n(\beta^0_{\alpha_\lambda}, t)\}\right| \geqslant \frac{U^2_n\varepsilon}{C_5}\right] \\
&\leqslant \mathrm{pr}\left[\sup_{0\leqslant t\leqslant\tau} \left|\frac{1}{S^*_n}\{S^{(0)}_n(\beta^0_{\alpha_\lambda}, t) - s^{(0)}_n(\beta^0_{\alpha_\lambda}, t)\}\right| \geqslant \frac{U^2_n\varepsilon}{C_5} \mid \mathscr{A}_n\right] + \mathrm{pr}(\mathscr{A}^c_n) = J_1 + J_2.
\end{aligned}$$

By (5) we have

$$J_1 \leqslant \mathrm{pr}\left[\sup_{0\leqslant t\leqslant\tau}\left|\frac{1}{C_5}\{S_n^{(0)}(\beta^0_{\alpha_\lambda},t)-s_n^{(0)}(\beta^0_{\alpha_\lambda},t)\}\right|\geqslant\frac{U_n^2\varepsilon}{C_5}\right]$$
$$=\mathrm{pr}\left\{\sup_{0\leqslant t\leqslant\tau}\left|S_n^{(0)}(\beta^0_{\alpha_\lambda},t)-s_n^{(0)}(\beta^0_{\alpha_\lambda},t)\right|\geqslant U_n^2\varepsilon\right\}\leqslant\frac{1}{5}W^2\exp(-n\varepsilon^2).$$

Further, we have that $J_2 = o(1)$ since S_n^* converges to $s_n^{(0)}(\beta^0_{\alpha_\lambda},t)$ in probability uniformly on $t\in[0,\tau]$. Therefore, by replacing ε with $n^{-1/2}\varepsilon$, from (13) we have that

$$\mathrm{pr}\left(I_1\geqslant\frac{n^{1/2}U_n^2\varepsilon}{C_5}\right)\leqslant\frac{1}{5}W^2\exp(-\varepsilon^2). \tag{14}$$

Next we consider I_2. For any i, $|\bar{\ell}_i(\beta^0_{\alpha_\lambda})|\leqslant|(\beta^0_{\alpha_\lambda})^TZ_i(t_i)-\log\{s_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)\}|\leqslant|(\beta^0_{\alpha_\lambda})^TZ_i(t_i)|+|\log\{s_n^{(0)}(\beta^0_{\alpha_\lambda},t_i)\}|\leqslant\log(U_n)+|\log(E[Y(t_i)\exp\{(\beta^0_{\alpha_\lambda})^TZ_i(t_i)\}])|\leqslant 2\log(U_n)$. It implies that $-2\log(U_n)\leqslant\bar{\ell}_i(\beta^0_{\alpha_\lambda})\leqslant 2\log(U_n)$ for all i. Thus, by Hoeffding's inequality (Hoeffding, 1963), for any $\varepsilon>0$,

$$\mathrm{pr}(I_2\geqslant n^{1/2}\varepsilon)\leqslant 2\exp\left[-\frac{2n\varepsilon^2}{\sum_{i=1}^n 4\{\log(U_n)\}^2}\right]=2\exp\left[-\frac{\varepsilon^2}{2\{\log(U_n)\}^2}\right]. \tag{15}$$

From (14) and (15) we get

$$\mathrm{pr}\left\{I_1+I_2+E(I_1)\geqslant\frac{n^{1/2}U_n\varepsilon}{C_5}+n^{1/2}\varepsilon+E(I_1)\right\}\leqslant\frac{1}{5}W^2\exp(-\varepsilon^2)+2\exp\left[-\frac{\varepsilon^2}{2\{\log(U_n)\}^2}\right].$$

Let $\varepsilon=\{\gamma_n|\alpha_\lambda|\log(d_n)\}^{1/2}$, where γ_n is any diverging sequence. Then,

$$\mathrm{pr}\left[I_1+I_2+E(I_1)\geqslant\{n\gamma_n|\alpha_\lambda|\log(d_n)\}^{1/2}\left(\frac{U_n}{C_5}+1\right)+E(I_1)\right]$$
$$\leqslant\frac{1}{5}W^2\exp\{-\gamma_n|\alpha_\lambda|\log(d_n)\}+2\exp\left[-\frac{\gamma_n|\alpha_\lambda|\log(d_n)}{2\{\log(U_n)\}^2}\right].$$

From (14) it can be verified that $E(I_1n^{-1/2}U_n^{-2}C_5)<\infty$. Therefore, $E(I_1)=O(n^{1/2})$. By using (9) we have that

$$\mathrm{pr}\left[\sup_{\alpha_\lambda}\frac{1}{|\alpha_\lambda|^{1/2}}\{I_1+I_2+E(I_1)\}\geqslant\{n\gamma_n\log(d_n)\}^{1/2}\left(\frac{U_n}{C_5}+1\right)+\frac{1}{|\alpha_\lambda|^{1/2}}E(I_1)\right]$$
$$\leqslant\sum_{k=1}^{d_n}\sum_{|\alpha_\lambda|=k}\mathrm{pr}\left[I_1+I_2+E(I_1)\geqslant\{n\gamma_n|\alpha_\lambda|\log(d_n)\}^{1/2}\left(\frac{U_n}{C_5}+1\right)+E(I_1)\right]$$
$$\leqslant\sum_{k=1}^{d_n}\left(\frac{e}{k}\right)^k\left[\frac{1}{5}W^2d_n^{k-k\gamma_n}+2d_n^{k-\frac{k\gamma_n}{2\{\log(U_n)\}^2}}\right]. \tag{16}$$

Since γ_n diverges to infinity, the two terms in the square brackets are both $o(d_n^{-1})$. Moreover,

$(e/k)^k < 1$ for all $k \geqslant 3$. Therefore, (16) goes to 0 as $n \to \infty$. Hence, $\sup_{\alpha_\lambda} |\alpha_\lambda|^{-1/2}\{I_1 + I_2 + E(I_1)\} = O_p[\{n\log(d_n)\}^{1/2}]$. Thus $\sup_{\alpha_\lambda} |\alpha_\lambda|^{-1/2}\left|\ell_n(\beta^0_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\}\right| = O_p\left[\{n\log(d_n)\}^{1/2}\right]$.

Proof of Theorem 1. For all model α_λ we have that

$$\begin{aligned}&\sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}}\left|\ell_n(\hat\beta_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\}\right| \\ &\leqslant \sup_{\alpha_\lambda} \frac{d_n}{|\alpha_\lambda|^{3/2}}\left|\ell_n(\hat\beta_{\alpha_\lambda}) - \ell_n(\beta^0_{\alpha_\lambda})\right| + \sup_{\alpha_\lambda} \frac{1}{|\alpha_\lambda|^{1/2}}\left|\ell_n(\beta^0_{\alpha_\lambda}) - E\{\ell_n(\beta^0_{\alpha_\lambda})\}\right|. \qquad (17)\end{aligned}$$

By Lemma 4 and 5, (17) $= O_p[d_n\{\log(d_n)\}^{1/2}] + O_p[\{n\log(d_n)\}^{1/2}] = O_p[\{n\log(d_n)\}^{1/2}]$ under Condition (G).

Proof of Theorem 2. By Taylor expansion, for some β^* that lies between $\hat\beta_{\alpha_\lambda}$ and $\hat\beta_{\alpha_0}$,

$$\begin{aligned}\ell_n(\hat\beta_{\alpha_0}) - \ell_n(\hat\beta_{\alpha_\lambda}) &= (\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda})^T\ell_n'(\hat\beta_{\alpha_\lambda}) + \frac{1}{2}(\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda})^T\ell_n''(\hat\beta_{\alpha_\lambda})(\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}) \\ &+ \frac{1}{6}\sum_{i=1}^{n}\sum_{j,k,l=1}^{|\alpha_\lambda|}\ell'''_{ijkl}(\beta^*)(\hat\beta_{\alpha_0 j} - \hat\beta_{\alpha_\lambda j})(\hat\beta_{\alpha_0 k} - \hat\beta_{\alpha_\lambda k})(\hat\beta_{\alpha_0 l} - \hat\beta_{\alpha_\lambda l}) = I_1 + I_2 + I_3.\end{aligned}$$

Since $\hat\beta_{\alpha_\lambda}$ maximizes $\ell_n(\beta_{\alpha_\lambda})$, $I_1 = 0$. In the proof of Lemma 3 we have shown that $\|\hat\beta_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\| = O_p(|\alpha_\lambda|^{1/2}n^{-1/2})$ for any α_λ. Since $\alpha_\lambda \supsetneq \alpha_0$, $\beta^0_{\alpha_\lambda} = \beta^0_{\alpha_0}$. Therefore, $\|\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}\| \leqslant \|\hat\beta_{\alpha_0} - \beta^0_{\alpha_0}\| + \|\hat\beta_{\alpha_\lambda} - \beta^0_{\alpha_\lambda}\| = O_p(|\alpha_\lambda|^{1/2}n^{-1/2})$. We decompose I_2 as

$$\begin{aligned}I_2 &= \frac{1}{2}(\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda})^T\{\ell_n''(\hat\beta_{\alpha_\lambda}) + nI_n(\hat\beta_{\alpha_\lambda})\}(\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}) - \frac{1}{2}(\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda})^T nI_n(\hat\beta_{\alpha_\lambda})(\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}) \\ &= I_{21} - I_{22},\end{aligned}$$

where $I_n(\hat\beta_{\alpha_\lambda})$ is defined in Section 3 of the main text. It can be shown that for $\ell''_{njk}(\hat\beta_{\alpha_\lambda})$ and $I_{njk}(\hat\beta_{\alpha_\lambda})$, the (j,k)th component of $\ell_n''(\hat\beta_{\alpha_\lambda})$ and $I_n(\hat\beta_{\alpha_\lambda})$ respectively, we have that $\ell''_{njk}(\hat\beta_{\alpha_\lambda}) + nI_{njk}(\hat\beta_{\alpha_\lambda}) = O_p(n^{1/2})$. Thus, $I_{21} \leqslant \|\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}\|^2 O_p(n^{1/2}|\alpha_\lambda|) = \|\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}\|^2 o_p(n)$ under Condition (G). Furthermore, $I_{22} \geqslant n\|\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}\|^2 \text{eigen}_{\min}\{I_n(\hat\beta_{\alpha_\lambda})\}/2 \geqslant n\|\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}\|^2 C_3/2$ under Condition (E). It follows that $I_{21} = o_p(I_{22})$. It can be shown that $\ell'''_{ijkl}(\beta^*)$ is $O_p(1)$. Thus, $I_3 \leqslant O_p\{\|\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}\|^3|\alpha_\lambda|^{3/2}n\} = o_p(\|\hat\beta_{\alpha_0} - \hat\beta_{\alpha_\lambda}\|^2 n)$ under Condition (G). Thus, $I_3 = o_p(I_{22})$.

Let $R_1 = I_{21} + I_3 = o_p(I_{22}) = o_p(|\alpha_\lambda|)$, then

$$\ell_n(\hat{\beta}_{\alpha_0}) - \ell_n(\hat{\beta}_{\alpha_\lambda}) = \frac{1}{2}(\hat{\beta}_{\alpha_0} - \hat{\beta}_{\alpha_\lambda})^T n I_n(\hat{\beta}_{\alpha_\lambda})(\hat{\beta}_{\alpha_0} - \hat{\beta}_{\alpha_\lambda}) + R_1. \tag{18}$$

On the other hand, by Taylor expansion, for some β^{**} that lies between $\hat{\beta}_{\alpha_\lambda}$ and $\hat{\beta}_{\alpha_0}$,

$$\begin{aligned} 0 = \ell_n'(\hat{\beta}_{\alpha_\lambda}) &= \ell_n'(\hat{\beta}_{\alpha_0}) + \{\ell_n''(\hat{\beta}_{\alpha_0}) + nI_n(\hat{\beta}_{\alpha_0})\}(\hat{\beta}_{\alpha_\lambda} - \hat{\beta}_{\alpha_0}) - nI_n(\hat{\beta}_{\alpha_0})(\hat{\beta}_{\alpha_\lambda} - \hat{\beta}_{\alpha_0}) \\ &+ \frac{1}{2}\left(\sum_{i=1}^{n} \sum_{j,k=1}^{|\alpha_\lambda|} \ell'''_{ijk1}(\beta^{**})(\hat{\beta}_{\alpha_\lambda j} - \hat{\beta}_{\alpha_0 j})(\hat{\beta}_{\alpha_\lambda k} - \hat{\beta}_{\alpha_0 k}), ..., \right. \\ &\left. \sum_{i=1}^{n} \sum_{j,k=1}^{|\alpha_\lambda|} \ell'''_{ijk|\alpha_\lambda|}(\beta^{**})(\hat{\beta}_{\alpha_\lambda j} - \hat{\beta}_{\alpha_0 j})(\hat{\beta}_{\alpha_\lambda k} - \hat{\beta}_{\alpha_0 k}) \right)^T \\ &= J_1 + J_2 - J_3 + J_4. \end{aligned} \tag{19}$$

Denote the vector J_2 as $(\nu_1, ..., \nu_{|\alpha_\lambda|})^T$ and J_3 as $(\upsilon_1, ..., \upsilon_{|\alpha_\lambda|})^T$. Since we have shown that $I_{21} = o_p(I_{22})$, it follows that $\sum_{j=1}^{|\alpha_\lambda|}(\hat{\beta}_{\alpha_\lambda j} - \hat{\beta}_{\alpha_0 j})\nu_j = o_p\{\sum_{j=1}^{|\alpha_\lambda|}(\hat{\beta}_{\alpha_\lambda j} - \hat{\beta}_{\alpha_0 j})\upsilon_j\}$. Since $\ell_n''(\beta^0_{\alpha_\lambda}) + nI_n(\hat{\beta}_{\alpha_0})$ and $nI_n(\hat{\beta}_{\alpha_0})$ are both symmetric matrices, under Condition (E) we have that $\nu_j = o_p(\upsilon_j)$ for all j, and therefore $J_2 = o_p(J_3)$ component-wise. Since $I_3 = o_p(I_{22})$, similar argument gives that $J_4 = o_p(J_3)$ component-wise. Let $R_2 = J_2 + J_4 = o_p(J_3)$, then $J_1 - J_3 + R_2 = 0$ by (19). Using proof by contradiction, it is necessary that $R_2 = o_p(J_1) = o_p\{\ell_n'(\hat{\beta}_{\alpha_0})\}$ component-wise. By solving (19) we have that $\hat{\beta}_{\alpha_\lambda} - \hat{\beta}_{\alpha_0} = n^{-1}\{I_n(\hat{\beta}_{\alpha_0})\}^{-1}\{\ell_n'(\hat{\beta}_{\alpha_0}) + R_2\}$. Plug this result into (18) we get

$$\begin{aligned} &\ell_n(\hat{\beta}_{\alpha_0}) - \ell_n(\hat{\beta}_{\alpha_\lambda}) \\ &= -\frac{1}{2}\{\ell_n'(\hat{\beta}_{\alpha_0}) + R_2\}^T n^{-1}\{I_n(\hat{\beta}_{\alpha_0})\}^{-1} n I_n(\hat{\beta}_{\alpha_\lambda}) n^{-1}\{I_n(\hat{\beta}_{\alpha_0})\}^{-1}\{\ell_n'(\hat{\beta}_{\alpha_0}) + R_2\} + R_1. \end{aligned}$$

Since both $\hat{\beta}_{\alpha_\lambda}$ and $\hat{\beta}_{\alpha_0}$ converge to β_0 in probability, $\hat{\beta}_{\alpha_\lambda}$ also converges to $\hat{\beta}_{\alpha_0}$ in probability. Hence, $I_n(\hat{\beta}_{\alpha_\lambda}) = I_n(\hat{\beta}_{\alpha_0}) + o_p(1)$. Therefore,

$$\begin{aligned} &\ell_n(\hat{\beta}_{\alpha_0}) - \ell_n(\hat{\beta}_{\alpha_\lambda}) \\ &= -\frac{1}{2}\ell_n'(\hat{\beta}_{\alpha_0})^T n^{-1}\{I_n(\hat{\beta}_{\alpha_0})\}^{-1}\ell_n'(\hat{\beta}_{\alpha_0}) - \ell_n'(\hat{\beta}_{\alpha_0})^T n^{-1}\{I_n(\hat{\beta}_{\alpha_0})\}^{-1} R_2 \\ &- \frac{1}{2} R_2 n^{-1}\{I_n(\hat{\beta}_{\alpha_0})\}^{-1} R_2 + \frac{1}{2}\{\ell_n'(\hat{\beta}_{\alpha_0}) + R_2\}^T n^{-2}\{I_n(\hat{\beta}_{\alpha_0})\}^{-2}\{\ell_n'(\hat{\beta}_{\alpha_0}) + R_2\} o_p(1) + R_1 \end{aligned}$$

$= K_1 + K_2 + K_3 + K_4 + R_1.$

Since $R_2 = o_p\{\ell_n'(\hat{\beta}_{\alpha_0})\}$ component-wise, K_2 and K_3 are both $o_p(K_1)$. Also, $K_4 = o_p(K_1)$. Furthermore, by spectral decomposition and Condition (E),

$$K_1 \geqslant \|\ell_n'(\hat{\beta}_{\alpha_0})\|^2 n^{-1} \text{eigen}_{\min}[\{I_n(\hat{\beta}_{\alpha_0})\}^{-1}]/2 = O_p(|\alpha_\lambda|).$$

Thus, $R_1 = o_p(K_1)$ since $R_1 = o_p(|\alpha_\lambda|)$. For any $\alpha_\lambda \supsetneq \alpha_0$, $I_n(\hat{\beta}_{\alpha_0})$ is the covariance matrix of $n^{-1/2}\ell_n'(\hat{\beta}_{\alpha_0})$, it follows that $-2K_1$ converges to a Chi-square distribution with degree of freedom $|\alpha_\lambda| - |\alpha_0|$. Therefore, $2\{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})\}$ converges to a Chi-square distribution with degree of freedom $|\alpha_\lambda| - |\alpha_0|$ for any $\alpha_\lambda \supsetneq \alpha_0$. By the corollary of Lemma 1 in Laurent & Massart (2000), for $\varepsilon = \gamma_n \log(d_n)(|\alpha_\lambda| - |\alpha_0|)$ where γ_n is any diverging sequence,

$$\begin{aligned}
&\text{pr}\left[2\{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})\} \geqslant |\alpha_\lambda| - |\alpha_0| + 2\sqrt{(|\alpha_\lambda| - |\alpha_0|)^2\gamma_n \log(d_n)} + 2\gamma_n \log(d_n)(|\alpha_\lambda| - |\alpha_0|)\right] \\
&= \text{pr}\left(2\{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})\} \geqslant (|\alpha_\lambda| - |\alpha_0|)\left[1 + 2\left\{\gamma_n \log(d_n)\right\}^{1/2} + 2\gamma_n \log(d_n)\right]\right) \\
&\leqslant \exp\left\{-\gamma_n \log(d_n)(|\alpha_\lambda| - |\alpha_0|)\right\}.
\end{aligned}$$

Therefore, by using (9) we have that

$$\begin{aligned}
&\text{pr}\left[\sup_{\alpha_\lambda \supsetneq \alpha_0} \frac{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})}{|\alpha_\lambda| - |\alpha_0|} \geqslant \frac{1}{2} + \left\{\gamma_n \log(d_n)\right\}^{1/2} + \gamma_n \log(d_n)\right] \\
&\leqslant \sum_{k=|\alpha_0|+1}^{d_n} \sum_{|\alpha_\lambda|=k} \text{pr}\left(\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0}) \geqslant (|\alpha_\lambda| - |\alpha_0|)\left[\frac{1}{2} + \left\{\gamma_n \log(d_n)\right\}^{1/2} + \gamma_n \log(d_n)\right]\right) \\
&\leqslant \sum_{k=|\alpha_0|+1}^{d_n} \left(\frac{e}{k}\right)^k d_n^{\{k-(k-|\alpha_0|)\gamma_n\}}. \qquad (20)
\end{aligned}$$

Since γ_n diverges to infinity and $k = O(k - |\alpha_0|)$ under Condition (G), $d_n^{\{k-(k-|\alpha_0|)\gamma_n\}} = o(d_n^{-1})$. Moreover, $(e/k)^k < 1$ for all $k \geqslant 3$. Therefore, (20) goes to 0 as $n \to \infty$. Thus,

$$\sup_{\alpha_\lambda \supsetneq \alpha_0} \frac{1}{|\alpha_\lambda| - |\alpha_0|}\left\{\ell_n(\hat{\beta}_{\alpha_\lambda}) - \ell_n(\hat{\beta}_{\alpha_0})\right\} = O_p\left[\frac{1}{2} + \left\{\log(d_n)\right\}^{1/2} + \log(d_n)\right] = O_p\{\log(d_n)\}.$$

References

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

KONG, S. & NAN, B. (2014). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via lasso. *Statist. Sinica* **24**, 25–42.

LAURENT, B. & MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* , 1302–1338.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614–645.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.