Journal of Machine Learning Research 21 (2020) 1-37 Submitted 10/19; Revised 9/20; Published 10/20

Functional Martingale Residual Process for High-Dimensional Cox Regression with Model Averaging

Baihua He HEBAIHUA@WHU.EDU.CN
Yanyan Liu LIUYY@WHU.EDU.CN
School of Mathematics and Statistics
Wuhan University
Wuhan 430072, China

Yuanshan Wu WU@ZUEL.EDU.CN
School of Statistics and Mathematics
Zhongnan University of Economics and Law
Wuhan 430073, China

Guosheng Yin GYIN@HKU.HK
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road, Hong Kong

Xingqiu Zhao XINGQIU.ZHAO@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong

Editor: Jie Peng

Abstract

Regularization methods for the Cox proportional hazards regression with high-dimensional survival data have been studied extensively in the literature. However, if the model is misspecified, this would result in misleading statistical inference and prediction. To enhance the prediction accuracy for the relative risk and the survival probability, we propose three model averaging approaches for the high-dimensional Cox proportional hazards regression. Based on the martingale residual process, we define the delete-one cross-validation (CV) process, and further propose three novel CV functionals, including the end-time CV, integrated CV, and supremum CV, to achieve more accurate prediction for the risk quantities of clinical interest. The optimal weights for candidate models, without the constraint of summing up to one, can be obtained by minimizing these functionals, respectively. The proposed model averaging approach can attain the lowest possible prediction loss asymptotically. Furthermore, we develop a greedy model averaging algorithm to overcome the computational obstacle when the dimension is high. The performances of the proposed model averaging procedures are evaluated via extensive simulation studies, demonstrating that our methods achieve superior prediction accuracy over the existing regularization methods. As an illustration, we apply the proposed methods to the mantle cell lymphoma study.

Keywords: Asymptotic Optimality, Censored Data, Cross Validation, Greedy Algorithm, Martingale Residual Process, Prediction, Survival Analysis

1. Introduction

High-dimensional survival data often arise in clinical studies where the number of predictors is large and sometimes even much larger than the sample size. Under the sparsity assumption that only a few among a large number of predictors are truly associated with survival times, regularization methods have been extended from linear regression to the Cox proportional hazards regression (Cox, 1972). These methods include, to name a few, the LASSO (Tibshirani, 1996; Huang et al., 2013), the adaptive LASSO (Zhang and Lu, 2007), the SCAD (Fan and Li, 2002), and the Dantzig selector (Antoniadis et al., 2010). The effectiveness of such regularization approaches relies heavily on the correctness of model specification, which is essential for drawing sound statistical conclusions. However, model misspecification often occurs in practice, especially for complex data such as high-dimensional survival data that are subject to right censoring. As a remedy, model averaging that combines the strength of a set of candidate models can mitigate the risk of misspecification, and thus enhance the prediction accuracy for statistical quantities of practical interest, such as the relative risk and the survival probability at a particular time. However, research in model averaging for survival data is very limited, let alone for high-dimensional survival data.

Most of the model averaging approaches are developed under standard settings where the sample size is larger than the number of predictors and the responses can be completely observed without censoring. In the Bayesian model averaging framework (Hoeting et al., 1999; Eklund and Karlsson, 2007), the posterior model probabilities are assigned to the candidate models. From the frequentist perspective, various strategies have been proposed to determine the weights for individual models, for example, the Mallows C_p model averaging (Hansen, 2007; Wan et al., 2010), optimal mean squared error averaging (Liang et al., 2011), optimal model averaging for linear mixed-effects models (Zhang et al., 2014), jackknife model averaging (Hansen and Racine, 2012), predictive likelihood model averaging (Ando and Tsay, 2010), and optimal model averaging for generalized linear models based on the Kullback–Leibler loss with a penalty term (Zhang et al., 2016). Nevertheless, all the aforementioned methods are developed for the situation where the sample size is larger than the number of predictors as well as under the constraint that the sum of weights of all candidate models is equal to one. Without imposing the summation constraint, Ando and Li (2014) proposed the delete-one cross-validation (CV) model averaging for high-dimensional linear regression and showed that it achieves the lowest possible prediction loss asymptotically. Recently, Ando and Li (2017) made a further extension to the high-dimensional generalized linear regression models.

As far as model averaging in survival analysis, Hjort and Claeskens (2006) extended the focused information criterion model averaging from linear regression to the Cox proportional hazards regression. Their approach aims to enhance the prediction accuracy for regression parameters of interest by considering the local root-n perturbation. However, it is difficult to apply their method to the high-dimensional survival data setting because regression parameter estimation itself is a nontrivial task. Based on the martingale residual process, we define the delete-one CV process, and further propose three novel model averaging approaches: the end-time CV, integrated CV, and supremum CV, for high-dimensional Cox regression to enhance the prediction accuracy for the relative risk and the survival

probability. The optimal weights for candidate models, without constraining the summation to be one, can be obtained by minimizing these functionals. Under certain conditions, we show that, based on the resulting optimal weights, all the three functionals of the CV process attain the lowest possible prediction loss asymptotically.

The number of candidate models is typically large when considering model averaging for high-dimensional data, which makes it challenging to find the optimal weights for candidate models. Ando and Li (2014) adopted the sure independent screening (SIS) procedure (Fan and Lv, 2008) to screen out the predictors that are less correlated with response marginally, so as to reduce the dimension of optimization. However, such SIS-based dimension reduction may remove some truly important predictors. Moreover, the choice of the cutting point for the number of predictors to be retained has not been explored. We develop a greedy model averaging algorithm to bypass these practical issues. Instead of ranking the predictors, we rank the candidate models, and the importance of candidate models is assessed in the greedy model averaging algorithm.

The rest of the article is organized as follows. We propose the model averaging procedures for the Cox regression with high-dimensional survival data in Section 2. We establish the asymptotic optimality of the proposed model averaging procedures in Section 3. To overcome the computational burden when the dimension is very high, we develop the greedy model averaging algorithm in Section 4. In Section 5, extensive simulation studies are conducted to demonstrate the superiority of the proposed methods in comparison with traditional regularization procedures for the high-dimensional Cox regression, which is further illustrated with the mantle cell lymphoma study in Section 6. Section 7 concludes with some remarks, and technical proofs are relegated to the Appendix.

2. Model Averaging Cox Regression

We first briefly review the partial likelihood estimation procedure for the conventional (low-dimensional) Cox proportional hazards regression model and then illustrate the strategy of preparing candidate Cox models in the high-dimensional case. We further propose three functionals of the CV process to produce the optimal weights for model averaging.

2.1 Partial Likelihood and Survival Prediction

Let T denote the failure time and C denote the censoring time. Let $\mathbf{Z}$ be a p-vector of predictors, $X = T \wedge C$ be the observed time, and $\Delta = I(T \leq C)$ be the censoring indicator, where $a \wedge b$ is the minimum of a and b, and $I(\cdot)$ is the indicator function. Assume that T and C are conditionally independent given $\mathbf{Z}$. The conditional hazard function associated with covariate $\mathbf{Z}$ is defined as

$$\lambda(t|\mathbf{Z}) = \lim_{h \to 0+} \mathbb{P}(t \leq T < t + h | T \geq t, \mathbf{Z})/h.$$

The Cox proportional hazards model (Cox, 1972) takes the form of

$$\lambda(t|\mathbf{Z}) = \lambda(t) \exp(\mathbf{Z}^\top \boldsymbol{\beta}), \tag{1}$$

where $\lambda(t)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}$ is a p-vector of unknown regression parameters. For $i = 1, \ldots, n$, let $(X_i, \Delta_i, \mathbf{Z}_i)$ be independent and identically distributed

copies of $(X, \Delta, \mathbf{Z})$. The log partial likelihood (Cox, 1975) based on the observed data can be written as

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_i^{\top}\boldsymbol{\beta} - \log \left\{ \sum_{j=1}^{n} Y_j(u) \exp(\mathbf{Z}_j^{\top}\boldsymbol{\beta}) \right\} \right] \mathrm{d}N_i(u),$$

where $N_i(t) = \Delta_i I(X_i \leq t)$ denotes the counting process, $Y_i(t) = I(X_i \geq t)$ the at-risk process and τ the end time of the study duration. By maximizing $l_n(\boldsymbol{\beta})$, we obtain a consistent and efficient estimator of $\boldsymbol{\beta}$, denoted as $\widehat{\boldsymbol{\beta}}$. Further, the Breslow estimator (Breslow, 1975) for the cumulative baseline hazard function, defined as $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$, is given by

$$\widehat{\Lambda}(t) = \int_0^t \frac{\sum_{i=1}^{n} \mathrm{d}N_i(u)}{\sum_{i=1}^{n} Y_i(u) \exp(\mathbf{Z}_i^{\top}\widehat{\boldsymbol{\beta}})}.$$

In practice, it is common to predict some risk quantity $Q(\boldsymbol{\beta}, \Lambda)$ for a specific patient at a particular time. For example, $Q(\boldsymbol{\beta}, \Lambda) = \exp(\mathbf{z}^{\top}\boldsymbol{\beta})$ represents the relative risk for patients with $\mathbf{Z} = \mathbf{z}$, and $Q(\boldsymbol{\beta}, \Lambda) = \exp\{-\Lambda(t_0)\exp(\mathbf{z}^{\top}\boldsymbol{\beta})\}$ is the survival probability for patients with $\mathbf{Z} = \mathbf{z}$ at time point $t_0 \in [0, \tau]$.

2.2 Candidate Cox Models

When the dimension p of predictors $\mathbf{Z}$ is larger than sample size n, the usual partial likelihood approach to predicting the risk indices of interest would not work. Although regularization methods such as the LASSO, MCP (Zhang, 2010), SCAD, and Dantzig selector can be applied, the fundamental assumption is that model (1) is correct under the sparsity setting. Obviously, correct specification of the true model is a nontrivial task and model diagnostic for high-dimensional regression itself is challenging. As a viable alternative, we propose a model averaging approach to predicting the risk quantities based on the high-dimensional survival data. To emphasize the dependence of the dimension p on sample size n, we rewrite p as p_n.

For simplicity, we use $[p_n]$ to denote the set $\{1, \ldots, p_n\}$. Let $\{\mathcal{A}_k : k = 1, \ldots, K_n\}$ be a family of sets with each element $\mathcal{A}_k$ being a nonempty subset of $[p_n]$, where K_n is some positive integer depending on n. Furthermore, the cardinality of $\mathcal{A}_k$, $|\mathcal{A}_k|$, is assumed to be much smaller than sample size n for $k = 1, \ldots, K_n$. For a p_n-dimensional vector $\mathbf{a} = (a_1, \ldots, a_{p_n})^{\top}$, let $\mathbf{a}_{(k)}$ denote the subvector of $\mathbf{a}$ indexed by set $\mathcal{A}_k$, or equivalently, $\mathbf{a}_{(k)} = (a_j : j \in \mathcal{A}_k)^{\top}$. Consequently, based on these index subsets $\mathcal{A}_k$'s, we can construct K_n candidate Cox models with the kth model given by

$$\lambda_k(t|\mathbf{Z}_{(k)}) = \lambda_{*k}(t) \exp\{\mathbf{Z}_{(k)}^{\top}\boldsymbol{\beta}_{*k}\}, \tag{2}$$

where $\mathbf{Z}_{(k)}$ is the subvector of $\mathbf{Z}$ consisting of coordinates indexed by set $\mathcal{A}_k$, $\lambda_{*k}(t)$ is an unknown baseline hazard function and $\boldsymbol{\beta}_{*k}$ is an $|\mathcal{A}_k|$-vector of unknown regression parameters. It is worth emphasizing that no model is imposed between survival time T and the predictors $\mathbf{Z}$ directly, while a set of candidate models is used to explore the large model space. Let $\widehat{\boldsymbol{\beta}}_{*k}$ denote the maximizer of the working log partial likelihood function,

$$l_{nk}(\boldsymbol{\beta}_{*k}) = \sum_{i=1}^{n} \int_0^{\tau} \left[\mathbf{Z}_{i(k)}^{\top}\boldsymbol{\beta}_{*k} - \log \left\{ \sum_{j=1}^{n} Y_j(u) \exp\{\mathbf{Z}_{j(k)}^{\top}\boldsymbol{\beta}_{*k}\} \right\} \right] \mathrm{d}N_i(u), \tag{3}$$

where $\mathbf{Z}_{i(k)} = (Z_{ij} : j \in \mathcal{A}_k)^\top$ and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$. Moreover, the estimator for $\Lambda_{*k}(t) = \int_0^t \lambda_{*k}(u)\mathrm{d}u$ is given by

$$\widehat{\Lambda}_{*k}(t) = \int_0^t \frac{\sum_{i=1}^n \mathrm{d}N_i(u)}{\sum_{i=1}^n Y_i(u)\exp\{\mathbf{Z}_{i(k)}^\top \widehat{\boldsymbol{\beta}}_{*k}\}}.$$

Based the kth candidate model, the prediction for the risk index can be written as $\widehat{Q}_k = Q(\widehat{\boldsymbol{\beta}}_{*k}, \widehat{\Lambda}_{*k})$, while it is critical to develop a sensible way to combine $\widehat{Q}_k, k = 1, \ldots, K_n$, for enhancing the overall prediction accuracy.

2.3 Optimal Weights

The σ-field, defined by

$$\mathcal{F}_t^i = \sigma\{N_i(u), I(X_i \le u, \Delta_i = 0), \mathbf{Z}_i : 0 \le u \le t\},$$

represents the history information for the ith subject up to time t. The conditional hazard function satisfies

$$\mathbb{E}\{\mathrm{d}N_i(t)|\mathcal{F}_{t-}^i\} = Y_i(t)\lambda(t|\mathbf{Z}_i)\mathrm{d}t, \tag{4}$$

where $\mathrm{d}N_i(t) = N_i\{(t + \mathrm{d}t)-\} - N_i(t-)$ is the increment of $N_i(\cdot)$ over the small interval $[t, t+\mathrm{d}t)$. Following the work of Lin et al. (2000) and the independent censoring assumption, (4) implies that

$$\mathbb{E}\{\mathrm{d}N_i(t)|\mathbf{Z}_i, C_i \ge t\} = \mathbb{E}\{\mathrm{d}N_i(t)|\mathbf{Z}_i\} = Y_i(t)\lambda(t|\mathbf{Z}_i)\mathrm{d}t.$$

Define the integrated intensity function

$$\mu_i(t) = \mathbb{E}\{N_i(t)|\mathbf{Z}_i\} = \int_0^t Y_i(u)\lambda(u|\mathbf{Z}_i)\mathrm{d}u, \tag{5}$$

which is unspecified as no model assumption is imposed on T_i and $\mathbf{Z}_i$. Let $\boldsymbol{\mu}(t) = (\mu_1(t), \ldots, \mu_n(t))^\top$, and we aim to approximate $\boldsymbol{\mu}(t)$ using the K_n candidate Cox models. Mimicking (5), we define the integrated intensity function associated with the kth working model in (2) as

$$\mu_{ik}(t, \boldsymbol{\beta}_{*k}, \Lambda_{*k}) = \int_0^t Y_i(u)\exp\{\mathbf{Z}_{i(k)}^\top \boldsymbol{\beta}_{*k}\}\mathrm{d}\Lambda_{*k}(u).$$

Denote $\widehat{\mu}_{ik}(t) = \mu_{ik}(t, \widehat{\boldsymbol{\beta}}_{*k}, \widehat{\Lambda}_{*k})$ and $\widehat{\boldsymbol{\mu}}_k(t) = (\widehat{\mu}_{1k}(t), \ldots, \widehat{\mu}_{nk}(t))^\top$. Let the K_n-vector weight $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{K_n})^\top$ be from the unit hypercube of $\mathbb{R}^{K_n}$,

$$\boldsymbol{\Omega}_n = \{\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{K_n})^\top \in [0,1]^{K_n} : 0 \le \omega_k \le 1, k = 1, \ldots, K_n\}.$$

The averaged intensity function is given by

$$\widehat{\boldsymbol{\mu}}(t) = \sum_{k=1}^{K_n} \omega_k \widehat{\boldsymbol{\mu}}_k(t),$$

which can be considered as an approximation for the true intensity function $\boldsymbol{\mu}(t)$ if the weights are chosen properly. To obtain the weights, we construct the quadratic loss process,

$$L_n(\boldsymbol{\omega}, t) = \|\boldsymbol{\mu}(t) - \widehat{\boldsymbol{\mu}}(t)\|^2, \tag{6}$$

where $\|\cdot\|$ denotes the Euclidean norm. However, minimization of the quadratic loss process is an infeasible task, as it involves the unknown intensity function $\boldsymbol{\mu}(t)$.

To circumvent the difficulty, we adopt the delete-one CV approach in Hansen and Racine (2012) to obtain the optimal weights. Let $\widehat{M}_{ik}(t) = N_i(t) - \widehat{\mu}_{ik}(t)$ denote the pseudo martingale residual process associated with the kth candidate model, and then $\sum_{i=1}^n \widehat{M}_{ik}(t) = 0$ for any $t \in [0, \tau]$ and $\sum_{i=1}^n \mathbf{Z}_{i(k)}\widehat{M}_{ik}(\tau) = 0$. Despite the usual properties of residuals in linear regression, the nuisance function Λ_{*k} and the dependence on time t certainly require extra efforts.

Let $\widetilde{\boldsymbol{\beta}}_{*k}^{(-i)}$ denote the delete-one estimator for $\boldsymbol{\beta}_{*k}$ based on all the observations except for the ith subject $(X_i, \Delta_i, \mathbf{Z}_{i(k)})$. The corresponding delete-one estimator for $\Lambda_{*k}(t)$ is given by

$$\widetilde{\Lambda}_{*k}^{(-i)}(t) = \int_0^t \frac{\sum_{j\neq i}^n \mathrm{d}N_j(u)}{\sum_{j\neq i}^n Y_j(u)\exp\{\mathbf{Z}_{j(k)}^\top \widetilde{\boldsymbol{\beta}}_{*k}^{(-i)}\}}.$$

For ease of expression, let $\widetilde{\mu}_{ik}(t) = \mu_{ik}(t, \widetilde{\boldsymbol{\beta}}_{*k}^{(-i)}, \widetilde{\Lambda}_{*k}^{(-i)})$ and $\widetilde{\boldsymbol{\mu}}_k(t) = (\widetilde{\mu}_{1k}(t), \ldots, \widetilde{\mu}_{nk}(t))^\top$. The averaged delete-one intensity function estimator is then given by

$$\widetilde{\boldsymbol{\mu}}(t) = \sum_{k=1}^{K_n} \omega_k \widetilde{\boldsymbol{\mu}}_k(t),$$

and we further define the delete-one CV process as

$$\mathrm{CV}_n(\boldsymbol{\omega}, t) = \|\mathbf{N}(t) - \widetilde{\boldsymbol{\mu}}(t)\|^2$$

for $t \in [0, \tau]$. At the end time of the study duration, we propose the end-time cross-validation (ECV) criterion,

$$\mathrm{ECV}_n(\boldsymbol{\omega}) = \mathrm{CV}_n(\boldsymbol{\omega}, \tau),$$

and the corresponding optimal weight is defined as $\widehat{\boldsymbol{\omega}}_{\mathrm{E}} = \arg\min_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \mathrm{ECV}_n(\boldsymbol{\omega})$. Alternatively, similar to the Cramér–von Mises approach, we propose an integrated cross-validation (ICV) criterion by "smoothing" out time t as

$$\mathrm{ICV}_n(\boldsymbol{\omega}) = \int_0^\tau \mathrm{CV}_n(\boldsymbol{\omega}, t)\mathrm{d}t,$$

and the optimal weight is given by $\widehat{\boldsymbol{\omega}}_{\mathrm{I}} = \arg\min_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \mathrm{ICV}_n(\boldsymbol{\omega})$. In fact, both the ECV and ICV criteria can be considered as special cases of the general cross-validation (GCV) criterion which smooths out time t with some known non-negative weight function $\phi(t)$ as follows,

$$\mathrm{GCV}_n(\boldsymbol{\omega}) = \int_0^\tau \mathrm{CV}_n(\boldsymbol{\omega}, t)\phi(t)\mathrm{d}t.$$

By taking $\phi(t) = \delta_\tau(t)$ where $\delta_\tau(t) = I(t = \tau)$ is the Dirac measure, GCV reduces to ECV; if $\phi(t) = 1$, GCV reduces to ICV. Motivated by the Kolmogorov–Smirnov approach, we further propose the supremum cross-validation (SCV) criterion,

$$\mathrm{SCV}_n(\boldsymbol{\omega}) = \sup_{t\in[0,\tau]} \mathrm{CV}_n(\boldsymbol{\omega}, t),$$

and the corresponding optimal weight $\widehat{\boldsymbol{\omega}}_{\mathrm{S}}$ can be obtained by minimizing $\mathrm{SCV}_n(\boldsymbol{\omega})$.

In general, the three criteria can be viewed as the functionals of the delete-one CV process. The preservation of convexity facilitates both theoretical and numerical development. For ease of expression, let $\widehat{\boldsymbol{\omega}}$ be any of the optimal weights $\widehat{\boldsymbol{\omega}}_{\mathrm{E}}$, $\widehat{\boldsymbol{\omega}}_{\mathrm{S}}$, or $\widehat{\boldsymbol{\omega}}_{\mathrm{I}}$. To predict the risk index of interest, we combine estimators from individual models as $\widehat{\boldsymbol{\omega}}^\top \widehat{\mathbf{Q}}$, with $\widehat{\mathbf{Q}} = (\widehat{Q}_1, \ldots, \widehat{Q}_{K_n})^\top$.

3. Theory of Optimality

Based on the quadratic loss function $L_n(\boldsymbol{\omega}, t)$ in (6), we define the corresponding risk function,

$$R_n(\boldsymbol{\omega}, t) = \mathbb{E}\{L_n(\boldsymbol{\omega}, t) \mid \mathbf{Z}_1, \ldots, \mathbf{Z}_n\}.$$

We denote $a_n = \inf_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \inf_{t\in[c_0,\tau]} R_n(\boldsymbol{\omega}, t)$ for some $c_0 > 0$ small enough and $a_n^* = \inf_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} R_n(\boldsymbol{\omega}, t)$. Suppose that $\widehat{\boldsymbol{\beta}}_{*k}$ converges to $\boldsymbol{\beta}_{0k}$ in probability and $\widehat{\Lambda}_{*k}$ converges to Λ_{0k}. Denote $\boldsymbol{\mu}^0(t) = \sum_{k=1}^{K_n} \boldsymbol{\omega}_k \boldsymbol{\mu}_k^0(t)$, where $\boldsymbol{\mu}_k^0(t) = (\mu_{1k}^0(t), \ldots, \mu_{nk}^0(t))^\top$ and $\mu_{ik}^0(t) = \mu_{ik}(t, \boldsymbol{\beta}_{0k}, \Lambda_{0k})$. We impose the following conditions throughout the theoretical derivation.

C1 $\mathbf{Z}$ is mean-zero sub-Gaussian vector. There exists a universal constant $c_\dagger \geq 1$ such that

$$1/c_\dagger \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq c_\dagger,$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{Z}$ and $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denotes the minimum and maximum eigenvalues of matrix $\boldsymbol{\Sigma}$, respectively.

C2 The size of each candidate model, $|\mathcal{A}_k|$, is fixed, where $k = 1, \ldots, K_n$.

C3 The parametric space $\mathcal{B}_k$ of $\boldsymbol{\beta}_{*k}$ is compact and $\int_0^\tau \lambda_{*k}(t)\mathrm{d}t$ is bounded uniformly over k; $\inf_{\mathbf{Z}\in\mathcal{Z}} P(Y(\tau) = 1|\mathbf{Z})$ is bounded away from zero, where $\mathcal{Z}$ is the support of $\mathbf{Z}$.

C4 For each k, $\inf_{\boldsymbol{\beta}_k\in\mathcal{B}_k} \lambda_{\min}(\mathbf{I}_k(\boldsymbol{\beta}_k))$ is bounded away from zero, where $\mathbf{I}_k(\boldsymbol{\beta}_k)$ is the information matrix under the kth candidate model.

C5 For any $\epsilon > 0$, it holds

$$\mathbb{P}\left(\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} \|\boldsymbol{\mu}(t) - \boldsymbol{\mu}^0(t)\| \geq \epsilon\right) \leq 2\exp(-c_\ddagger n^{-1/2} K_n^{-2} \epsilon^2)$$

for some constant $c_\ddagger > 0$.

C6 (i) $a_n^{-1} K_n^2 \varrho_n^2 n^{-1} \log(nK_n) \to 0$ and (ii) $a_n^{-1} K_n^2 \varrho_n^2 (n \log n)^{1/2} \to 0$ in probability, where $\varrho_n = \exp\{(\log(nK_n))^{2/3}\}$.

C7 $n/a_n^* \to 0$ in probability.

Condition C1 is a common assumption in high-dimensional data analysis. Condition C2 requires that the size of each candidate model should not be divergent while conditions C3 and C4 are regular assumptions in survival analysis, which guarantee that all candidate models can yield feasible estimators. Even when the model is misspecified, it was shown by Lin and Wei (1989) that $\widehat{\boldsymbol{\beta}}_{*k} - \boldsymbol{\beta}_{0k} = O_p(n^{-1/2})$, and so is for $\widehat{\Lambda}_{*k}$. Condition C5 states the asymptotic order of difference between the true and model-based predictions, implicitly illustrating the effectiveness of candidate models. Condition C6 states the divergence rate of the minimum risk for the ECV and ICV criteria, which excludes the degenerate case of $t = 0$ due to sparse data, a common treatment in survival analysis. In particular, if $a_n = O_p(n^{\eta+1/2}\log n)$ for some constant $\eta > 0$, condition C6 is satisfied by choosing $K_n = O(n^{\delta})$, where $0 < \delta < \eta/2$, as there exists a constant $0 < \upsilon < \eta - 2\delta$ small enough such that $\varrho_n^2 \le n^{\upsilon}$. Condition C7 needs a higher divergence rate of the minimum risk for SCV as it considers the worst-case scenario, which holds if $\eta \ge 1/2$.

Theorem 1 *Under conditions C1–C6, it holds for the ECV criterion that*

$$\frac{L_n^{\mathrm{E}}(\widehat{\boldsymbol{\omega}}_{\mathrm{E}})}{\inf_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} L_n^{\mathrm{E}}(\boldsymbol{\omega})} \to 1$$

in probability as $n \to \infty$, where $L_n^{\mathrm{E}}(\boldsymbol{\omega}) = L_n(\boldsymbol{\omega}, \tau)$.

Theorem 2 *Under conditions C1–C6, it holds for the ICV criterion that*

$$\frac{L_n^{\mathrm{I}}(\widehat{\boldsymbol{\omega}}_{\mathrm{I}})}{\inf_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} L_n^{\mathrm{I}}(\boldsymbol{\omega})} \to 1$$

in probability as $n \to \infty$, where $L_n^{\mathrm{I}}(\boldsymbol{\omega}) = \int_0^{\tau} L_n(\boldsymbol{\omega}, t)\mathrm{d}t$.

Theorem 3 *Under conditions C1–C7, it holds for the SCV criterion that*

$$\frac{L_n^{\mathrm{S}}(\widehat{\boldsymbol{\omega}}_{\mathrm{S}})}{\inf_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} L_n^{\mathrm{S}}(\boldsymbol{\omega})} \to 1$$

in probability as $n \to \infty$, where $L_n^{\mathrm{S}}(\boldsymbol{\omega}) = \sup_{t\in[0,\tau]} L_n(\boldsymbol{\omega}, t)$.

Theorems 1–3 lay out the theoretical foundations of the proposed model averaging approaches for high-dimensional survival data. Although the smallest possible quadratic losses are infeasible to achieve because the underlying true intensity function $\boldsymbol{\mu}(t)$ is unknown, the functional delete-one CV criteria provide a viable solution. It is worth noting that Theorems 1 and 2 still hold under less stringent conditions as they consider a single time point τ or integration with respect to t. We delineate the proofs of theorems in a unified structure based on Lemmas 1–7 in the Appendix.

4. Computation

We formulate a general framework to construct the candidate models by partitioning the indices of predictors $[p_n]$, while in practice important predictors are usually grouped together by utilizing some prior information. The SIS methods for ultrahigh-dimensional survival data (Zhao and Li, 2012; Song et al., 2014; Wu and Yin, 2015) are effective tools to rank and group predictors based on their marginal dependence on survival times. As the screening method of Zhao and Li (2012) is particularly designed for the Cox proportional hazards model, we adopt it to rank the p_n predictors and then evenly partition them into K_n groups without overlapping; each group is formulated as a candidate model.

Intuitively, more candidate models should be considered as the dimension of predictors grows. When the number of candidate models K_n is moderate, the optimal weights can be readily obtained via the quadratic programming optimization with box constraints. However, such optimization turns out to be challenging for large K_n due to the computational burden of quadratic programming. To alleviate the high-dimensional optimization issue, we develop a greedy model averaging algorithm, which is described in detail as follows.

Algorithm 1 Greedy model averaging algorithm based on the ECV criterion

Initialize $\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(0)} \in \boldsymbol{\Omega}_n$

for $\ell = 1, 2, \ldots,$ do

$\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell)} = \arg\min_{\boldsymbol{\gamma} \in \boldsymbol{\Omega}_n} \langle \nabla \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell-1)}), \boldsymbol{\gamma} \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product

$\widehat{\alpha}_{\mathrm{E}}^{(\ell)} = \arg\min_{\alpha \in [0,1]} \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell-1)} + \alpha(\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell)} - \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell-1)}))$

$\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)} = \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell-1)} + \widehat{\alpha}_{\mathrm{E}}^{(\ell)}(\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell)} - \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell-1)})$

if $\|\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)} - \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell-1)}\|_\infty \le \kappa$, then break

end for

Output $\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}$

Denote the infinity norm by $\|\cdot\|_\infty$, and set $\kappa = 0.001$ and the initial value $\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(0)}$ as $\mathbf{0}$, $\mathbf{1}$, or $\mathbf{e}_1$, where $\mathbf{e}_1$ denotes the first vector of the canonical basis of $\mathbb{R}^{K_n}$. We choose the best coordinate according to the gradient $\nabla \mathrm{ECV}_n(\cdot)$, which is standard in the greedy-type algorithm. In particular, it selects candidate models by optimizing a linear function of gradient over the box constraints at each iteration, which thus overcomes the difficulty caused by high dimensionality. Dai et al. (2012) proposed a greedy algorithm for model aggregation, while they set $\widehat{\alpha}^{(\ell)} = 2/(\ell+1)$ and minimized the linear function of gradient over the canonical basis. Likewise, we can replace the ECV criterion by ICV or SCV in the algorithm to obtain $\widehat{\boldsymbol{\omega}}_{\mathrm{I}}^{(\ell)}$ or $\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}$ in the ℓth iteration.

Theorem 4 *(a) For the ECV criterion, let* $h_{\mathrm{E},n} = \max_{k\in[K_n]} \|\widetilde{\boldsymbol{\mu}}_k(\tau)\|^2$*, then*

$$\mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell+1)}) \le \min_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \mathrm{ECV}_n(\boldsymbol{\omega}) + \frac{16K_n^2 h_{\mathrm{E},n}}{\ell}.$$

(b) For the ICV criterion, let $h_{\mathrm{I},n} = \max_{k\in[K_n]} \int_0^\tau \|\widetilde{\boldsymbol{\mu}}_k(t)\|^2 \mathrm{d}t$ *and assume that the derivative with respect to* $\boldsymbol{\omega}$ *and the integral with respect to* t *are exchangeable, then*

$$\mathrm{ICV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{I}}^{(\ell+1)}) \le \min_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \mathrm{ICV}_n(\boldsymbol{\omega}) + \frac{16K_n^2 h_{\mathrm{I},n}}{\ell}.$$

(c) For the SCV criterion, let $h_{\mathrm{S},n} = \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \lambda_{\max}(\nabla^2 \mathrm{SCV}_n(\boldsymbol{\omega}))$*, then*

$$\mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell+1)}) \le \min_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \mathrm{SCV}_n(\boldsymbol{\omega}) + \frac{8K_n h_{\mathrm{S},n}}{\ell}.$$

If a sufficient number of iterations are carried out in the greedy model averaging algorithms, the resulting weights can reach the optimal solutions as the remainders approach zero as ℓ goes to infinity. The proof of Theorem 4 is also provided in the Appendix. For a unified exposition, Theorem 4 quantifies the approximation starting from the second-step iteration. Although not crucial in practice, the first-step approximation is declared in the proof.

5. Simulation Studies

We evaluate the finite-sample performances of the proposed model averaging methods and make comparisons with various regularization methods via simulation studies, including the LASSO, MCP, SCAD, Elastic Net (EN) with ratio 0.5, Ridge, adaptive LASSO (ALASSO) approaches. We generate survival time T_i from the Cox proportional hazards model,

$$\lambda(t|\mathbf{Z}_i) = \lambda(t)\exp(\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}),$$

where the baseline hazard function is $\lambda(t) = (t-0.5)^2$ and the high-dimensional predictor $\mathbf{Z}_i = (Z_{i1},\ldots,Z_{ip_n})^\top$ follows a p_n-dimensional normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = (0.8^{|j-j'|})$ for $j, j' = 1,\ldots,p_n$. The first 15 elements of $\boldsymbol{\beta}$ are set to be 0.2 and the rest $\mathbf{0}$. The censoring time is $C_i = \widetilde{C}_i \wedge \tau$, where $\widetilde{C}_i$ is generated from an exponential distribution, Exp(0.12), and the study duration τ is chosen to yield a censoring rate of 20%. We consider sample size $n = 100$ and 200, coupled with the dimension of predictors $p_n = 1000$ and 2000.

The SIS method of Zhao and Li (2012) is adopted to rank the importance of each predictor and then every 10 or 20 predictors are grouped together to formulate a candidate Cox model. This leads to a total of $K_n = 100$ or 50 candidate models for $p_n = 1000$ and $K_n = 200$ or 100 for $p_n = 2000$. We evaluate the relative risk (RR) for a subject with predictors drawn from a p_n-dimensional normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, as well as the survival probability (SP) at time $t_0 = 2$. For each configuration, we replicate 100 simulations and present the mean squared errors (MSEs) of predictions for the RR and SP.

				Proposed methods											
				ECV				ICV				SCV			
Index	n	p_n	K_n	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$
RR	100	1000	50	0.027	0.033	0.035	0.036	0.030	0.036	0.037	0.036	0.139	0.184	0.182	0.174
			100	0.013	0.012	0.014	0.013	0.027	0.014	0.015	0.017	0.091	0.107	0.109	0.105
		2000	100	0.223	0.398	0.368	0.320	0.208	0.370	0.380	0.343	0.952	0.909	0.821	0.868
			200	0.138	0.216	0.186	0.187	0.095	0.199	0.227	0.213	0.898	0.726	0.715	0.786
	200	1000	50	0.013	0.012	0.010	0.011	0.013	0.011	0.011	0.012	0.016	0.013	0.014	0.014
			100	0.014	0.005	0.006	0.008	0.017	0.007	0.008	0.008	0.044	0.035	0.029	0.026
		2000	100	0.161	0.226	0.219	0.208	0.138	0.225	0.206	0.197	0.150	0.186	0.194	0.200
			200	0.110	0.144	0.152	0.145	0.098	0.157	0.151	0.151	0.425	0.294	0.324	0.330
SP	100	1000	50	0.017	0.025	0.024	0.022	0.016	0.025	0.022	0.021	0.022	0.025	0.026	0.027
			100	0.004	0.009	0.010	0.008	0.006	0.010	0.009	0.009	0.013	0.014	0.014	0.014
		2000	100	0.088	0.089	0.087	0.092	0.069	0.077	0.084	0.086	0.027	0.031	0.035	0.037
			200	0.051	0.053	0.051	0.053	0.037	0.045	0.050	0.049	0.019	0.028	0.030	0.032
	200	1000	50	0.014	0.015	0.013	0.014	0.015	0.014	0.014	0.014	0.005	0.009	0.009	0.010
			100	0.004	0.007	0.006	0.006	0.005	0.006	0.006	0.006	0.007	0.009	0.009	0.009
		2000	100	0.055	0.079	0.076	0.072	0.046	0.077	0.071	0.068	0.038	0.039	0.043	0.044
			200	0.041	0.049	0.052	0.050	0.032	0.051	0.048	0.047	0.024	0.032	0.035	0.036

			Regularized methods					
Index	n	p_n	LASSO	MCP	SCAD	EN	Ridge	ALASSO
RR	100	1000	0.118	0.514	0.433	0.169	0.193	0.165
		2000	0.934	4.205	5.542	2.140	2.001	1.084
	200	1000	0.053	0.286	0.230	0.134	0.132	0.059
		2000	0.434	4.274	3.688	0.311	0.423	0.469
SP	100	1000	0.043	0.069	0.059	0.037	0.049	0.047
		2000	0.031	0.065	0.050	0.065	0.057	0.047
	200	1000	0.022	0.048	0.043	0.037	0.047	0.020
		2000	0.032	0.072	0.046	0.033	0.042	0.030

Table 1: Mean squared errors over 100 simulations for predictions of the relative risk (RR) and the survival probability (SP) using the proposed model averaging methods and various regularization methods when the failure times are generated from the Cox model.

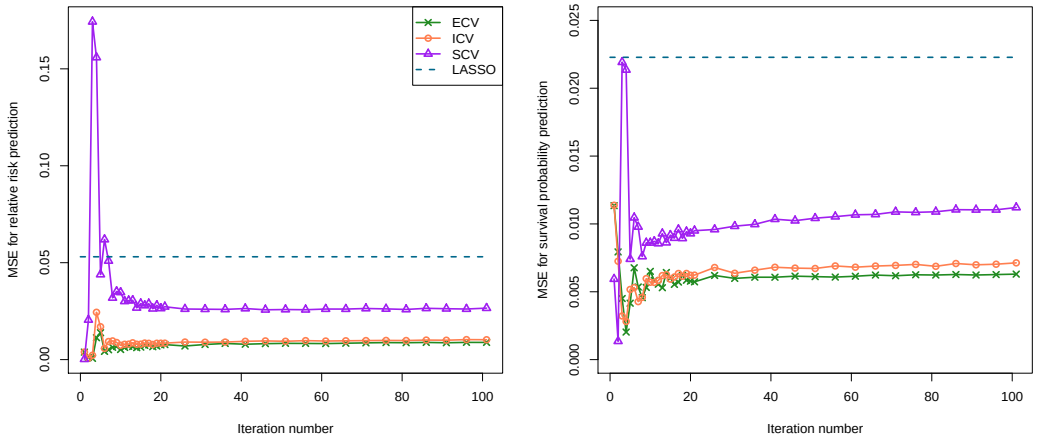

Figure 1: Mean squared errors for predictions of the relative risk (left) and survival probability (right) using the greedy model averaging algorithms and the LASSO method in the first 100 iterations with $n = 200$, $p_n = 1000$ and $K_n = 100$ when the failure times are generated from the Cox model.

Table 1 summarizes the simulation results for the proposed methods and various regularization approaches. In most cases, the proposed methods yield smaller MSEs for the predictions of the relative risk and survival probability than regularization ones. Our methods also benefit from the increasing number of candidate models as they could explore the model space more sufficiently. Figure 1 exhibits the MSEs in the first 100 iterations for the case with $n = 200$, $p_n = 1000$ and $K_n = 100$ where we also plot those of LASSO for comparison, demonstrating superior performances and stable convergence paths of the greedy model averaging algorithms.

We also consider an accelerated failure time (AFT) model,

$$\log T_i = \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i,$$

where the error ϵ_i follows the standard normal distribution. The remaining setups are kept the same as those under the Cox model. Obviously, the proportional hazards structure does not hold under the AFT model. Nevertheless, the proposed model averaging methods under the Cox model framework are still applied to the data arise from the AFT model, aiming to investigate the robustness of our approach. Simulation results in Table 2 show that the proposed methods deliver much smaller MSEs than the regularization methods. It indicates that the proposed Cox model averaging methods do not rely upon the correct specification of the underlying model as much as the regularization methods.

				Proposed methods											
				ECV				ICV				SCV			
Index	n	p_n	K_n	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$
RR	100	1000	50	0.138	0.102	0.117	0.135	0.136	0.121	0.168	0.158	0.433	0.651	0.633	0.658
			100	0.126	0.054	0.064	0.072	0.153	0.055	0.075	0.084	0.330	0.234	0.232	0.247
		2000	100	0.309	0.479	0.519	0.457	0.434	0.393	0.423	0.435	0.648	0.462	0.600	0.649
			200	0.121	0.149	0.156	0.163	0.095	0.143	0.145	0.167	0.293	0.214	0.240	0.258
	200	1000	50	0.055	0.079	0.089	0.092	0.063	0.105	0.108	0.115	0.198	0.334	0.320	0.325
			100	0.139	0.100	0.111	0.119	0.160	0.117	0.131	0.135	0.240	0.264	0.256	0.260
		2000	100	0.150	0.221	0.214	0.205	0.139	0.220	0.203	0.193	0.163	0.160	0.169	0.173
			200	0.084	0.148	0.143	0.135	0.077	0.137	0.139	0.135	0.131	0.164	0.206	0.212
SP	100	1000	50	0.028	0.025	0.027	0.028	0.029	0.025	0.029	0.029	0.042	0.051	0.050	0.052
			100	0.021	0.014	0.014	0.016	0.026	0.014	0.017	0.019	0.044	0.037	0.038	0.040
		2000	100	0.065	0.076	0.086	0.084	0.059	0.075	0.081	0.078	0.038	0.059	0.056	0.056
			200	0.037	0.045	0.050	0.051	0.030	0.040	0.044	0.049	0.018	0.037	0.035	0.036
	200	1000	50	0.018	0.022	0.024	0.024	0.020	0.027	0.027	0.028	0.037	0.051	0.049	0.050
			100	0.016	0.015	0.017	0.017	0.018	0.019	0.020	0.020	0.032	0.042	0.041	0.041
		2000	100	0.048	0.072	0.069	0.066	0.043	0.070	0.064	0.060	0.045	0.046	0.049	0.050
			200	0.029	0.044	0.044	0.041	0.024	0.043	0.040	0.038	0.017	0.030	0.032	0.032

			Regularized methods					
Index	n	p_n	LASSO	MCP	SCAD	EN	Ridge	ALASSO
RR	100	1000	3.038	4.793	2.401	1.758	4.356	4.041
		2000	1.003	6.679	3.782	1.130	2.102	1.182
	200	1000	2.817	9.890	13.643	2.088	4.562	5.353
		2000	0.462	5.702	13.604	0.403	0.540	0.852
SP	100	1000	0.199	0.156	0.166	0.211	0.095	0.220
		2000	0.042	0.063	0.053	0.067	0.055	0.053
	200	1000	0.228	0.169	0.174	0.225	0.138	0.231
		2000	0.033	0.103	0.052	0.035	0.058	0.043

Table 2: Mean squared errors over 100 simulations for predictions of the relative risk (RR) and the survival probability (SP) using the proposed model averaging methods and various regularization methods when the failure times are generated from an accelerated failure time model.

The delete-one CV procedure, which is also called the n-fold CV, is advocated for the proposed model averaging methods. Nevertheless, our methods can be readily coupled with general ν-fold CV with $\nu < n$. Essentially, we can use $100 \times (1 - 1/\nu)\%$ of the total samples to train the proposed methods to predict the remaining samples. We consider $\nu = 5$ and 10 to investigate the performances of the proposed methods. It can be seen from simulation results in Table 3 that the proposed methods perform equally well for different values of ν. We also plot the first 10 weights for $n = 200$ and $p_n = 1000$ in Figure 2, which shows that the first two models attain the largest weights. The drastically decreasing trend further demonstrates that the greedy model averaging algorithm can quickly identify the effective candidate models. Further, we compare the running time in minutes per simulation by taking an average over 100 simulations for each method as shown in Table 4, when the failure times are generated form the Cox model. Compared with the regularization methods, the proposed methods are time costly as they all involve CV procedure, especially for the ICV and SCV criteria where integrating and maximizing the functional CV process over $[0, \tau]$ consume substantial computational memory. However, the running time under the 5-fold ECV criterion is comparable to that of the regularization methods, indicating the proposed greedy model averaging algorithm may not be the cause of the burden of computation.

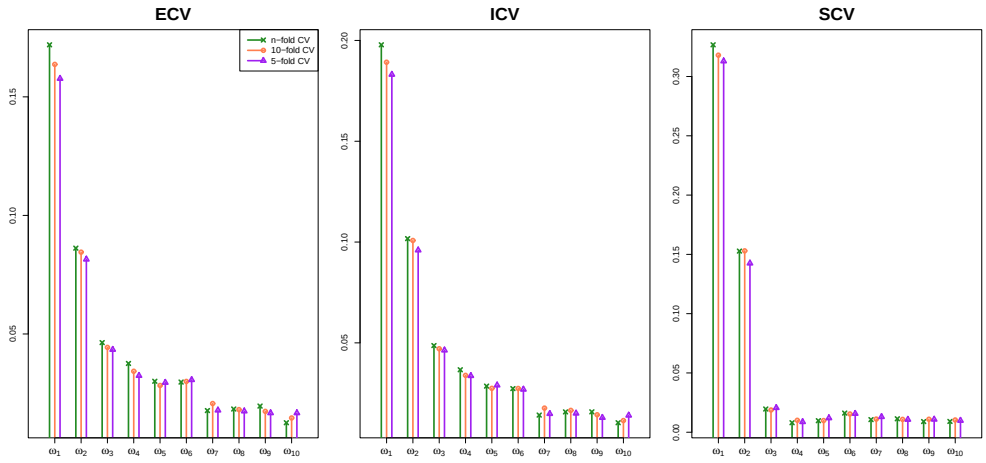

Figure 2: First 10 components of weights using the greedy model averaging algorithms in the last iteration with $n = 200$, $p_n = 1000$ and $K_n = 100$ when the failure times are generated from the Cox model.

				Proposed methods											
				ECV				ICV				SCV			
Index	n	p_n	ν	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$
RR	100	1000	5	0.009	0.026	0.019	0.022	0.013	0.024	0.019	0.023	0.163	0.111	0.108	0.098
			10	0.013	0.016	0.016	0.015	0.021	0.017	0.017	0.021	0.171	0.130	0.118	0.120
		2000	5	0.165	0.180	0.170	0.176	0.133	0.163	0.173	0.184	0.775	0.532	0.546	0.575
			10	0.148	0.170	0.155	0.178	0.106	0.172	0.177	0.192	1.003	0.670	0.683	0.683
	200	1000	5	0.011	0.005	0.005	0.006	0.012	0.006	0.006	0.008	0.041	0.029	0.023	0.022
			10	0.010	0.005	0.005	0.006	0.015	0.006	0.007	0.007	0.039	0.028	0.027	0.025
		2000	5	0.131	0.161	0.168	0.163	0.098	0.174	0.166	0.160	0.378	0.279	0.293	0.297
			10	0.113	0.151	0.154	0.149	0.101	0.156	0.159	0.152	0.452	0.272	0.294	0.298
SP	100	1000	5	0.005	0.011	0.011	0.011	0.005	0.013	0.011	0.011	0.016	0.014	0.013	0.013
			10	0.005	0.009	0.010	0.009	0.006	0.011	0.010	0.010	0.016	0.014	0.015	0.014
		2000	5	0.063	0.058	0.056	0.056	0.048	0.049	0.050	0.053	0.018	0.027	0.029	0.030
			10	0.057	0.055	0.051	0.054	0.041	0.046	0.047	0.051	0.020	0.028	0.030	0.032
	200	1000	5	0.004	0.007	0.006	0.006	0.005	0.007	0.006	0.007	0.007	0.008	0.008	0.008
			10	0.004	0.007	0.006	0.006	0.005	0.006	0.006	0.006	0.007	0.008	0.009	0.009
		2000	5	0.048	0.056	0.057	0.056	0.036	0.056	0.054	0.051	0.022	0.032	0.036	0.037
			10	0.044	0.052	0.054	0.052	0.034	0.052	0.051	0.049	0.023	0.031	0.034	0.035

			Regularized methods					
Index	n	p_n	LASSO	MCP	SCAD	EN	Ridge	ALASSO
RR	100	1000	0.118	0.514	0.433	0.169	0.193	0.165
		2000	0.934	4.205	5.542	2.140	2.001	1.084
	200	1000	0.053	0.286	0.230	0.134	0.132	0.059
		2000	0.434	4.274	3.688	0.311	0.423	0.469
SP	100	1000	0.043	0.069	0.059	0.037	0.049	0.047
		2000	0.031	0.065	0.050	0.065	0.057	0.047
	200	1000	0.022	0.048	0.043	0.037	0.047	0.020
		2000	0.032	0.072	0.046	0.033	0.042	0.030

Table 3: Mean squared errors over 100 simulations for prediction of the relative risk (RR) and the survival probability (SP) using the proposed model averaging methods and various regularization methods under the general ν-fold CV procedure with $K_n = p_n/10$ when the failure times are generated from the Cox model.

			Proposed methods								
			ECV			ICV			SCV		
n	p_n	K_n	$\nu = 5$	$\nu = 10$	$\nu = n$	$\nu = 5$	$\nu = 10$	$\nu = n$	$\nu = 5$	$\nu = 10$	$\nu = n$
100	1000	50	0.034	0.063	0.058	0.466	0.868	0.943	0.171	0.318	0.290
		100	0.107	0.470	0.866	1.749	9.222	17.058	0.538	2.369	4.345
	2000	100	0.084	0.160	0.143	1.154	2.142	2.322	0.419	0.808	0.719
		200	0.272	1.189	2.102	4.393	23.014	42.017	1.363	5.878	10.591
200	1000	50	0.058	0.110	0.103	0.848	1.558	1.750	0.298	0.551	0.524
		100	0.190	1.740	3.202	3.188	34.898	62.454	0.956	8.865	16.033
	2000	100	0.118	0.223	0.220	1.702	3.110	3.667	0.590	1.136	1.100
		200	0.392	3.672	6.807	6.586	76.346	134.080	1.991	18.564	35.084

		Regularized methods					
n	p_n	LASSO	MCP	SCAD	EN	Ridge	ALASSO
100	1000	0.033	0.042	0.072	0.082	0.025	0.034
	2000	0.047	0.053	0.098	0.125	0.060	0.059
200	1000	0.078	0.165	0.247	0.250	0.054	0.047
	2000	0.117	0.198	0.331	0.373	0.132	0.104

Table 4: Averaged running time (in minutes) per simulation over 100 runs for prediction MSEs of the proposed methods and various regularization methods when the failure times are generated from the Cox model.

6. Application

As an illustration, we apply the proposed model averaging approaches to the mantle cell lymphoma (MCL) study, which was also analyzed by Rosenwald et al. (2003). The gene expression data set available from `http://llmpp.nih.gov/MCL/` contains expression values of 6312 cDNA elements after excluding genes with missing values. Based on the morphologic and immunophenotypic criteria, 92 patients with MCL were included in the study. During the 14 years' follow-up, 64 patients died of the MCL and the other 28 were censored, leading to a censoring rate of 30.4%. Our primary goal is to predict the relative risk and survival probability of patients with high-dimensional predictors of gene expressions.

We first apply the SIS method of Zhao and Li (2012) to rank the importance of genes, and then construct candidate models by grouping every 10 genes, leading to $K_n = 632$ candidate models. A new predictor, denoted by $\mathbf{Z}_0$, is taken to be the column-wise median of the 6312 gene expressions. Figure 3 shows the relative risk for the patient with predictor $\mathbf{Z}_0$ over the first 100 iterations in the greedy model averaging algorithm. It shows that the relative risks using the ECV and ICV criteria agree well with each other, while the SCV criterion, corresponding to the worst-case consideration, yields more serious relative risk. The regularization methods also tend to deliver higher relative risk. There is no much difference across the n-, 10- and 5-fold CV procedures. To predict the survival probability of the patient with predictor $\mathbf{Z}_0$, we confine the time duration of interest to 14 years and

equally partition the time axis by an interval of length 0.01. The survival probabilities at the grid points are predicted using the proposed model averaging methods and regularization methods. As shown in Figure 4, the ECV and ICV criteria result in higher predicted survival probabilities than the SCV criterion and regularization methods. The convergence paths of the predicted survival probabilities are stable for the first 100 iterations. Furthermore, we take averages over all the estimated survival curves of all subjects with predictors $\mathbf{Z}_i, i = 1, \ldots, n$, using the model averaging methods and various regularization methods, respectively. Figure 5 shows that the predicted survival curves using the model averaging method with the SCV criterion, and the LASSO, EN, and MCP regularized methods generally agree well with the Kaplan–Meier curve, while the model averaging methods with the ECV and ICV criteria tend to overestimate the survival probability.

The concordance index (C-index) is commonly used in survival analysis for assessment of prediction performance (Uno et al., 2011), which is given by

$$C_n(\widehat{\boldsymbol{\omega}}) = \frac{\sum_{i \neq j} \Delta_i I(X_i < X_j) I(\sum_{k=1}^{K_n} \widehat{\omega}_k \mathbf{Z}_{i(k)}^\top \widehat{\boldsymbol{\beta}}_{*k} > \sum_{k=1}^{K_n} \widehat{\omega}_k \mathbf{Z}_{j(k)}^\top \widehat{\boldsymbol{\beta}}_{*k})}{\sum_{i \neq j} \Delta_i I(X_i < X_j)}$$

in the framework of model averaging, where $\widehat{\boldsymbol{\omega}}$ represents $\widehat{\boldsymbol{\omega}}_{\mathrm{E}}$, $\widehat{\boldsymbol{\omega}}_{\mathrm{I}}$ or $\widehat{\boldsymbol{\omega}}_{\mathrm{S}}$. A higher C-index implies a better prediction performance. As shown in Table 5, the proposed model averaging methods deliver higher C-index values than various regularization methods. The SCV criterion exhibits the best concordance behavior among all the considered methods. The overall running time that each method consumes for analyzing the mantle cell lymphoma study is summarized in Table 6, from which we can draw conclusions similar to those of simulations.

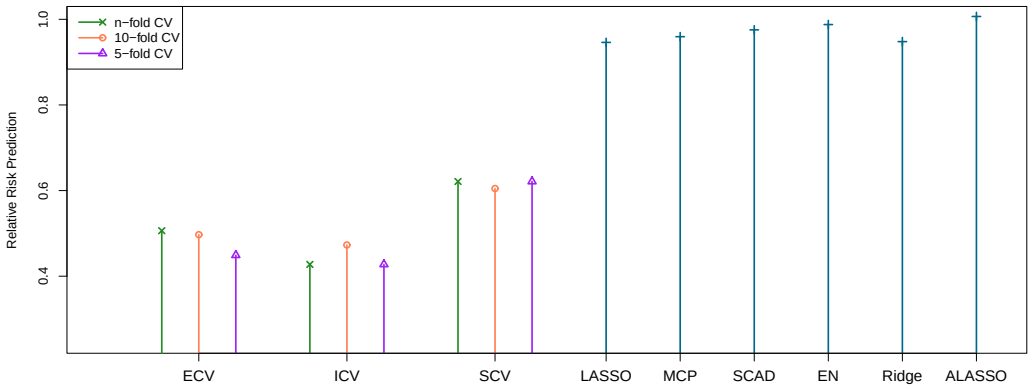

Figure 3: Prediction of the relative risk for a patient with column-wise medians of gene expressions in the mantle cell lymphoma study using the proposed methods with the first 100 convergence pathes under the greedy model averaging algorithms, and various regularization methods.

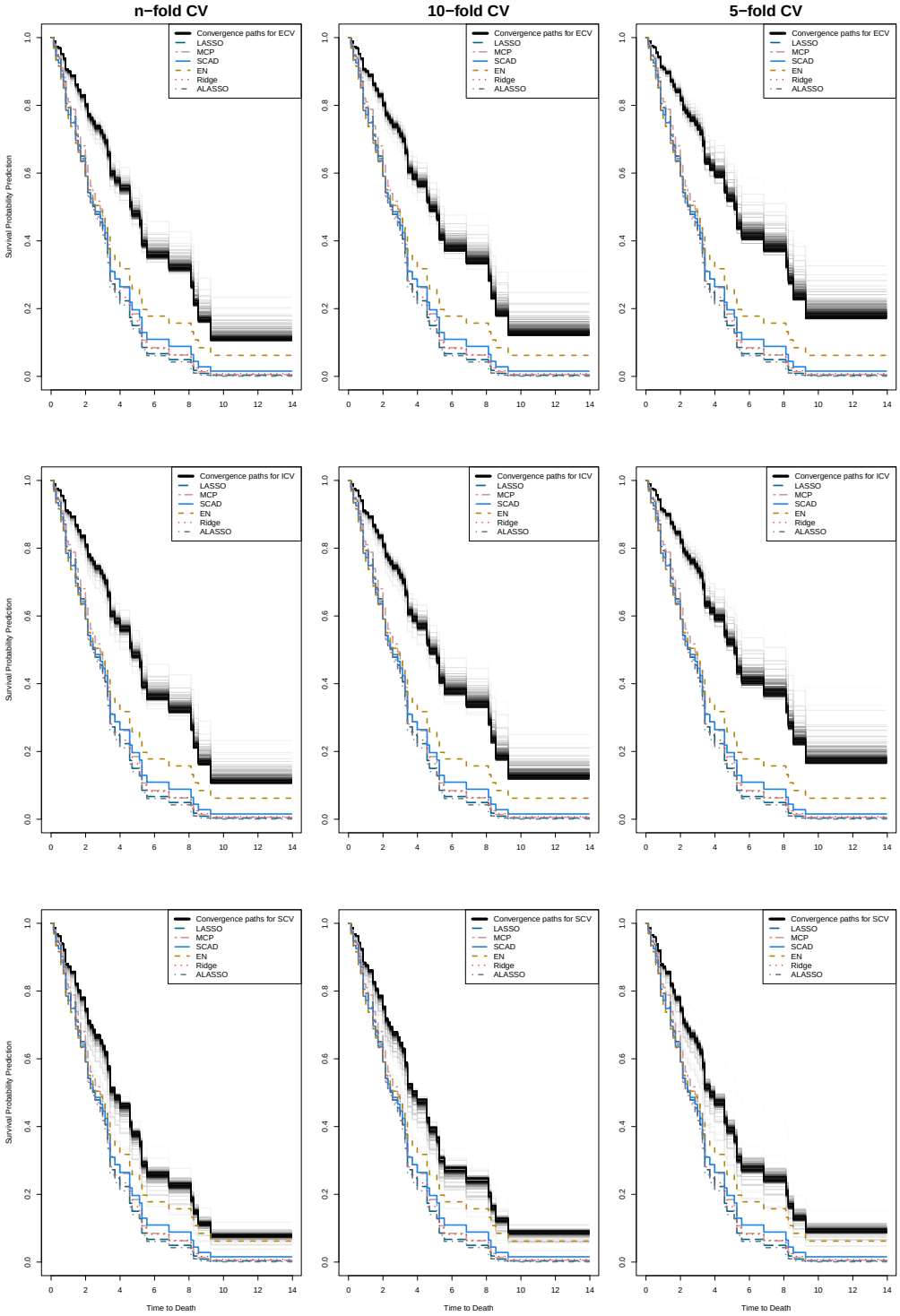

Figure 4: Prediction of the survival probability for a patient with column-wise medians of gene expressions in the mantle cell lymphoma study using the proposed methods with the first 100 iterations under the greedy model averaging algorithms, and the regularization methods.

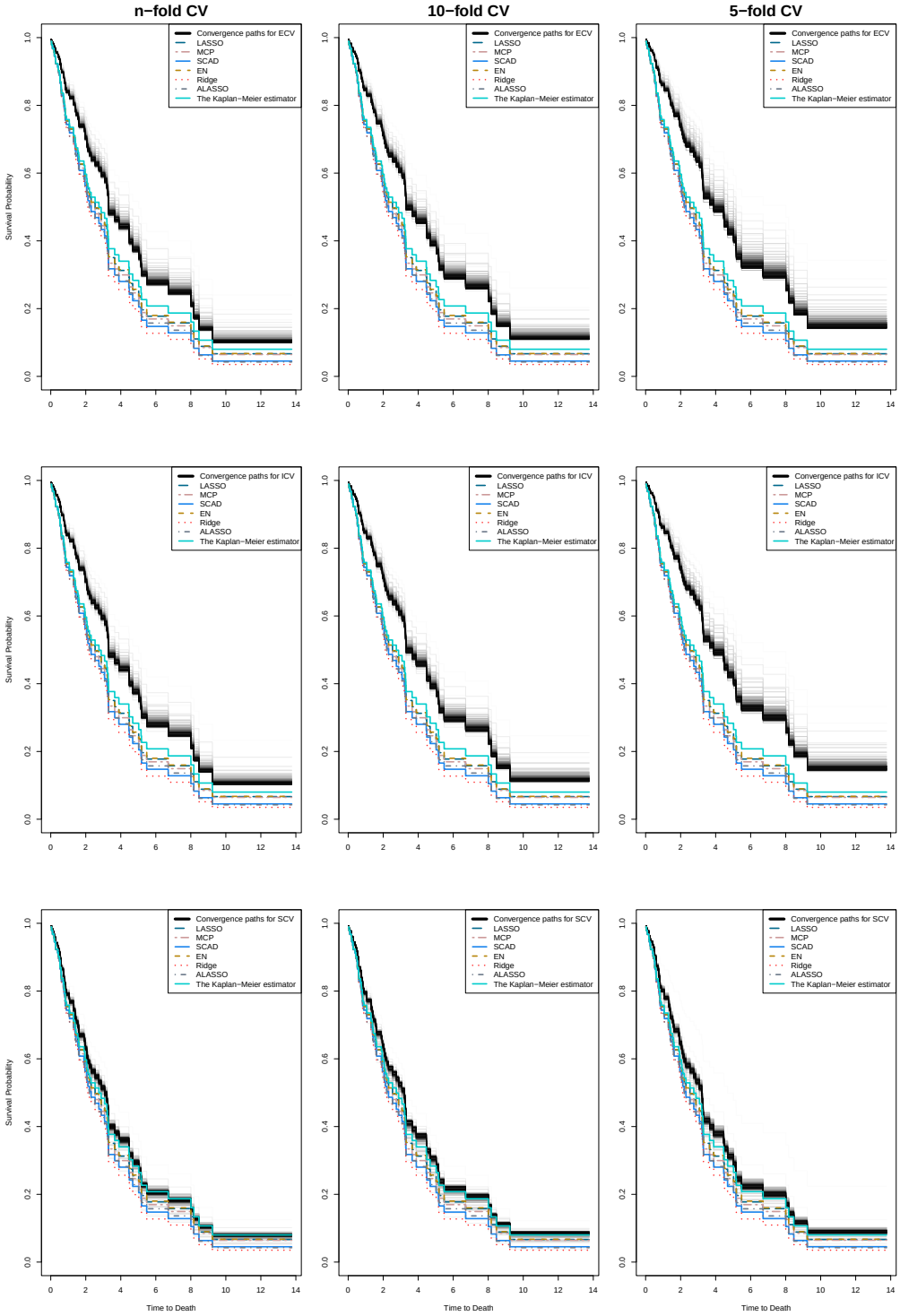

Figure 5: The Kaplan–Meier survival curve and predicted survival curve using the proposed methods with the first 100 iterations under the greedy model averaging algorithms, and various regularization methods for the mantle cell lymphoma study.

	Proposed methods											
	ECV				ICV				SCV			
ν	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$	$\ell=5$	$\ell=10$	$\ell=15$	$\ell=20$
5	0.904	0.906	0.906	0.907	0.916	0.919	0.918	0.919	0.925	0.925	0.925	0.924
10	0.911	0.912	0.911	0.911	0.906	0.918	0.916	0.912	0.923	0.929	0.928	0.928
n	0.911	0.911	0.910	0.911	0.918	0.921	0.920	0.920	0.927	0.923	0.922	0.922

	Regularized methods					
	LASSO	MCP	SCAD	EN	Ridge	ALASSO
	0.890	0.800	0.833	0.903	0.765	0.868

Table 5: The C-indices of the proposed model averaging methods and various regularization methods in the mantle cell lymphoma study.

Proposed methods								
ECV			ICV			SCV		
$\nu=5$	$\nu=10$	$\nu=n$	$\nu=5$	$\nu=10$	$\nu=n$	$\nu=5$	$\nu=10$	$\nu=n$
0.501	0.790	5.313	13.832	27.794	246.236	48.136	74.137	512.597

Regularized methods					
LASSO	MCP	SCAD	EN	Ridge	ALASSO
0.118	0.109	0.162	0.227	0.208	0.159

Table 6: The overall running time (in minutes) under the proposed model averaging methods and various regularization methods in the mantle cell lymphoma study.

7. Remarks

As an effective tool for prediction, the model averaging methods have been extensively studied in linear regression, which however remain to be explored in high-dimensional survival analysis. Utilizing the martingale residuals, we propose three functionals of the CV process to conduct the Cox model averaging for high-dimensional survival data. The optimality of the model averaging methods is established using empirical process theory. We also develop the greedy model averaging algorithm to carry out the high-dimensional optimization. Numerical studies show that the proposed methods in conjunction with the greedy algorithms generally deliver superior performances over the regularization methods.

A fundamental feature of the proposed model averaging approach is that it allows the model weight to vary freely between zero and one without the usual constraint of summing up to one. We rank the candidate models based on marginal utility that quantifies associations between predictors and the outcome marginally, so that the higher-ranked models capture more information than the lower-ranked ones. Borrowing information from all working models can substantially enhance the prediction performance. Relaxation of

the constraint makes the theoretical derivation more challenging while the theoretical development for the constraint case can be considered as a special case. Furthermore, the greedy algorithm facilitates the implementation of the proposed model averaging approach in practice. The simplicity and generality of the greedy algorithm can also incorporate the constraint case, with slight modification by selecting one candidate model at each iteration.

The general ν-fold CV procedure for the proposed model averaging approaches is also recommended in practice for reducing the computational burden. The grid-search method for selecting ν using the block bootstrap approach suggested by Liu et al. (2019) can be also applicable. When the true model happens to be contained as a candidate model, the infimum risk converges to infinity at a lower rate, making conditions C6 and C7 no longer hold. The proposed methods thus require in theory that all candidate models are misspecified, which nevertheless is a common assumption in the literature of model averaging (Zhu et al, 2019).

Splitting the predictors into small sets to construct a series of candidate models can be considered as a way of dimension reduction to break the curse of high dimensionality. We adopt the strategy by fixing the size of each candidate models as in condition C2 while increasing K_n to accommodate the growth of p_n. As the set of candidate models is increased, the possible quadratic loss is decreased. In this sense, the larger value K_n, the better. In practice, if the computational cost is not concerned, we suggest to increase K_n as large as possible such that the set of candidate models is increased gradually. On the other hand, when the size of each candidate model is divergent with n, it imposes practical and theoretical obstacles for the model averaging approach. The usage of regularization method for each candidate model may result in tremendous computational cost and unstable numerical results as there are a large number of tuning parameters to tune. Theoretical derivation for evaluating the convergence rate between parameter estimation and its delete-one counterpart under model misspecification has its own importance even for a single model with divergent size. Instead of making effort on covering the true model using candidate models with large size, we employ the refined pieces of candidate models with small size to sufficiently explore the model space, which greatly facilitates the theoretical establishment and practical implementation.

Acknowledgments

We thank the action editor and reviewers for their many constructive comments that strengthened the work immensely. This research was supported in part by the National Natural Science Foundation of China (Projects 11971362, 11671311, 12071483, 11771366) and the Research Grants Council of Hong Kong (Projects 17307218, 15301218, 15303319). The corresponding author is Yuanshan Wu.

Appendix. Theoretical Proofs

We first introduce several lemmas as follows.

Lemma 1 *Under conditions C1–C4, for $k = 1, \ldots, K_n$, it holds that*

$$\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)} - \widehat{\boldsymbol{\beta}}_{*k} = O_p(n^{-1}),$$
$$\sup_{t\in[0,\tau]} \left|\widetilde{\Lambda}_{*k}^{(-s)}(t) - \widehat{\Lambda}_{*k}(t)\right| = O_p(n^{-1}),$$

where $s = 1, \ldots, n$.

Proof Following (3), the log partial likelihood function for the kth working model can also be rewritten as

$$l_{nk}(\boldsymbol{\beta}_{*k}) = \sum_{i=1}^{n} \Delta_i \mathbf{Z}_{i(k)}^{\top}\boldsymbol{\beta}_{*k} - \sum_{i=1}^{n} \Delta_i \log S_{nk}(X_i, \boldsymbol{\beta}_{*k}),$$

where $S_{nk}(t, \boldsymbol{\beta}_{*k}) = \sum_{j=1}^{n} Y_j(t)\exp(\mathbf{Z}_{j(k)}^{\top}\boldsymbol{\beta}_{*k})$. Without loss of generality, assume that the n observations are rearranged according to the order, $X_1 < X_2 < \cdots < X_n$. When the sth observation is deleted, the log partial likelihood function based on the remaining $n-1$ observations can be written as

$$\begin{aligned} l_{nk}^{(-s)}(\boldsymbol{\beta}_{*k}) &= \sum_{i=1, i\neq s}^{n} \Delta_i \mathbf{Z}_{i(k)}^{\top}\boldsymbol{\beta}_{*k} - \sum_{i=1}^{s-1} \Delta_i \log\Big\{S_{nk}(X_i, \boldsymbol{\beta}_{*k}) - \exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\Big\} \\ &\quad - \sum_{i=s+1}^{n} \Delta_i \log S_{nk}(X_i, \boldsymbol{\beta}_{*k}). \end{aligned}$$

Denote $l_{nk}^{(s)}(\boldsymbol{\beta}_{*k}) = l_{nk}(\boldsymbol{\beta}_{*k}) - l_{nk}^{(-s)}(\boldsymbol{\beta}_{*k})$. Straightforward calculation yields that

$$\begin{aligned} \dot{l}_{nk}^{(s)}(\boldsymbol{\beta}_{*k}) &= \frac{\partial l_{nk}^{(s)}(\boldsymbol{\beta}_{*k})}{\partial \boldsymbol{\beta}_{*k}} \\ &= \Delta_s \mathbf{Z}_{s(k)} - \Delta_s \frac{\dot{S}_{nk}(X_s, \boldsymbol{\beta}_{*k})}{S_{nk}(X_s, \boldsymbol{\beta}_{*k})} \\ &\quad + \sum_{i=1}^{s-1} \Delta_i \frac{\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\big\{\dot{S}_{nk}(X_i, \boldsymbol{\beta}_{*k}) - S_{nk}(X_i, \boldsymbol{\beta}_{*k})\mathbf{Z}_{s(k)}\big\}}{S_{nk}(X_i, \boldsymbol{\beta}_{*k})\big\{S_{nk}(X_i, \boldsymbol{\beta}_{*k}) - \exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\big\}}, \end{aligned}$$

where $\dot{S}_{nk}(t, \boldsymbol{\beta}_{*k}) = \sum_{j=1}^{n} Y_j(t)\exp(\mathbf{Z}_{j(k)}^{\top}\boldsymbol{\beta}_{*k})\mathbf{Z}_{j(k)}$. It follows from conditions C1–C3 and Anderson and Gill (1982) that $\sup_{\boldsymbol{\beta}_k\in\mathcal{B}_k} \|\dot{l}_{nk}^{(s)}(\boldsymbol{\beta}_k)\| = O_p(1)$. For simplicity, we further denote $\ddot{S}_{nk}(t, \boldsymbol{\beta}_{*k}) = \sum_{j=1}^{n} Y_j(t)\exp(\mathbf{Z}_{j(k)}^{\top}\boldsymbol{\beta}_{*k})\mathbf{Z}_{j(k)}^{\otimes 2}$, where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\top}$ for a vector $\mathbf{a}$. We have

$$\begin{aligned} \ddot{l}_{nk}^{(s)}(\boldsymbol{\beta}_{*k}) &= \frac{\partial^2 l_{nk}^{(s)}(\boldsymbol{\beta}_{*k})}{\partial \boldsymbol{\beta}_{*k}\partial \boldsymbol{\beta}_{*k}^{\top}} \\ &= -\Delta_s \frac{S_{nk}(X_s, \boldsymbol{\beta}_{*k})\ddot{S}_{nk}(X_s, \boldsymbol{\beta}_{*k}) - \dot{S}_{nk}^{\otimes 2}(X_s, \boldsymbol{\beta}_{*k})}{S_{nk}^2(X_s, \boldsymbol{\beta}_{*k})} \\ &\quad + \sum_{i=1}^{s-1} \frac{\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\Delta_i \ddot{S}_{nk}(X_i, \boldsymbol{\beta}_{*k})}{S_{nk}(X_i, \boldsymbol{\beta}_{*k})\big\{S_{nk}(X_i, \boldsymbol{\beta}_{*k}) - \exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\big\}} \end{aligned}$$

$$
\begin{aligned}
&-\sum_{i=1}^{s-1}\frac{\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\mathbf{Z}_{s(k)}^{\otimes 2}\Delta_i}{S_{nk}(X_i,\boldsymbol{\beta}_{*k})-\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})}\\
&-\sum_{i=1}^{s-1}\frac{\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\Delta_i\dot{S}_{nk}^{\otimes 2}(X_i,\boldsymbol{\beta}_{*k})}{S_{nk}^{2}(X_i,\boldsymbol{\beta}_{*k})\{S_{nk}(X_i,\boldsymbol{\beta}_{*k})-\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\}}\\
&+\sum_{i=1}^{s-1}\frac{2\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\Delta_i\dot{S}_{nk}(X_i,\boldsymbol{\beta}_{*k})\mathbf{Z}_{s(k)}^{\top}}{\{S_{nk}(X_i,\boldsymbol{\beta}_{*k})-\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\}^{2}}\\
&-\sum_{i=1}^{s-1}\frac{\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\Delta_i\dot{S}_{nk}^{\otimes 2}(X_i,\boldsymbol{\beta}_{*k})}{S_{nk}(X_i,\boldsymbol{\beta}_{*k})\{S_{nk}(X_i,\boldsymbol{\beta}_{*k})-\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\}^{2}}\\
&-\sum_{i=1}^{s-1}\frac{\exp(2\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\mathbf{Z}_{s(k)}^{\otimes 2}\Delta_i}{\{S_{nk}(X_i,\boldsymbol{\beta}_{*k})-\exp(\mathbf{Z}_{s(k)}^{\top}\boldsymbol{\beta}_{*k})\}^{2}}.
\end{aligned}
$$

Likewise, we have $\sup_{\boldsymbol{\beta}_k\in\mathcal{B}_k}\|\ddot{l}_{nk}^{(s)}(\boldsymbol{\beta}_k)\|=O_p(1)$. We take the first-order Taylor approximation for $\dot{l}_{nk}^{(-s)}(\boldsymbol{\beta}_{*k})$ around $\widehat{\boldsymbol{\beta}}_{*k}$ as follows,

$$
\dot{l}_{nk}^{(-s)}(\boldsymbol{\beta}_{*k})=\dot{l}_{nk}^{(-s)}(\widehat{\boldsymbol{\beta}}_{*k})+\ddot{l}_{nk}^{(-s)}(\widehat{\boldsymbol{\beta}}_{*k})(\boldsymbol{\beta}_{*k}-\widehat{\boldsymbol{\beta}}_{*k})+o_p(\|\boldsymbol{\beta}_{*k}-\widehat{\boldsymbol{\beta}}_{*k}\|). \tag{7}
$$

By setting $\boldsymbol{\beta}_{*k}=\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)}$ in (7) and observing that $\dot{l}_{nk}^{(-s)}(\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})=\dot{l}_{nk}(\widehat{\boldsymbol{\beta}}_{*k})=0$ and $l_{nk}^{(-s)}=l_{nk}-l_{nk}^{(s)}$, we have

$$
\dot{l}_{nk}^{(s)}(\widehat{\boldsymbol{\beta}}_{*k})=\{\ddot{l}_{nk}(\widehat{\boldsymbol{\beta}}_{*k})-\ddot{l}_{nk}^{(s)}(\widehat{\boldsymbol{\beta}}_{*k})\}(\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)}-\widehat{\boldsymbol{\beta}}_{*k})+o_p(\|\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)}-\widehat{\boldsymbol{\beta}}_{*k}\|). \tag{8}
$$

It follows from Lin and Wei (1989) that $n^{-1}\ddot{l}_{nk}(\widehat{\boldsymbol{\beta}}_{*k})$ converges to $\mathbf{I}(\boldsymbol{\beta}_{0k})$, the information matrix under the kth candidate model. Consequently, (8) boils down to

$$
\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)}-\widehat{\boldsymbol{\beta}}_{*k}=\left\{\mathbf{I}(\boldsymbol{\beta}_{0k})-O_p\left(n^{-1}\right)+o_p(1)\right\}^{-1}O_p\left(n^{-1}\right),
$$

which implies that

$$
\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)}-\widehat{\boldsymbol{\beta}}_{*k}=O_p\left(n^{-1}\right)
$$

under condition C4. Furthermore, following Lin and Wei (1989), we have

$$
\begin{aligned}
&\sup_{t\in[0,\tau]}\left|\widetilde{\Lambda}_{*k}^{(-s)}(t)-\widehat{\Lambda}_{*k}(t)\right|\\
=\;&\sup_{t\in[0,\tau]}\left|\int_0^t\frac{\sum_{j\neq s}^{n}\mathrm{d}N_j(u)}{\sum_{j\neq s}^{n}Y_j(u)\exp(\mathbf{Z}_{j(k)}^{\top}\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})}-\int_0^t\frac{\sum_{j=1}^{n}\mathrm{d}N_j(u)}{\sum_{j\neq s}^{n}Y_j(u)\exp(\mathbf{Z}_{j(k)}^{\top}\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})}\right|\\
&+\sup_{t\in[0,\tau]}\left|\int_0^t\frac{\sum_{j=1}^{n}\mathrm{d}N_j(u)}{\sum_{j\neq s}^{n}Y_j(u)\exp(\mathbf{Z}_{j(k)}^{\top}\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})}-\int_0^t\frac{\sum_{j=1}^{n}\mathrm{d}N_j(u)}{\sum_{j=1}^{n}Y_j(u)\exp(\mathbf{Z}_{j(k)}^{\top}\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})}\right|\\
&+\sup_{t\in[0,\tau]}\left|\int_0^t\frac{\sum_{j=1}^{n}\mathrm{d}N_j(u)}{\sum_{j=1}^{n}Y_j(u)\exp(\mathbf{Z}_{j(k)}^{\top}\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})}-\int_0^t\frac{\sum_{j=1}^{n}\mathrm{d}N_j(u)}{\sum_{j=1}^{n}Y_j(u)\exp(\mathbf{Z}_{j(k)}^{\top}\widehat{\boldsymbol{\beta}}_{*k})}\right|
\end{aligned}
$$

$$
\begin{aligned}
= & \sup_{t\in[0,\tau]}\left|\int_0^t \frac{\mathrm{d}N_s(u)}{\sum_{j\neq s}^n Y_j(u)\exp(\mathbf{Z}_{j(k)}^\top \widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})}\right| \\
& + \sup_{t\in[0,\tau]}\left|\int_0^t \frac{Y_s(u)\exp(\mathbf{Z}_{s(k)}^\top \widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})\sum_{j=1}^n \mathrm{d}N_j(u)}{S_{nk}(u,\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})\{S_{nk}(u,\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)}) - Y_s(u)\exp(\mathbf{Z}_{s(k)}^\top \widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})\}}\right| \\
& + \sup_{t\in[0,\tau]}\left|\int_0^t \frac{\dot{S}_{nk}^\top(u,\widehat{\boldsymbol{\beta}}_{*k}^\dagger)(\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)} - \widehat{\boldsymbol{\beta}}_{*k})}{S_{nk}^2(u,\widehat{\boldsymbol{\beta}}_{*k}^\dagger)}\sum_{j=1}^n \mathrm{d}N_j(u)\right| \\
= & \ O_p(n^{-1}),
\end{aligned}
$$

where $\widehat{\boldsymbol{\beta}}_{*k}^\dagger$ lies between $\widehat{\boldsymbol{\beta}}_{*k}$ and $\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)}$. Lemma 1 is thus shown. ■

Lemma 2 *Under conditions C1–C2, it holds that,*

$$\max_{i\in[n]}\max_{k\in[K_n]}\|\mathbf{Z}_{i(k)}\| = O_p\left((\log(nK_n))^{1/2}\right), \tag{9}$$

$$\max_{i\in[n]}\max_{k\in[K_n]}\sup_{\boldsymbol{\beta}_k\in\mathcal{B}_k}\exp(\mathbf{Z}_{i(k)}^\top\boldsymbol{\beta}_k) = O_p\left(\varrho_n\right). \tag{10}$$

Proof Under condition C1, Hsu et al. (2012) showed that there exists $\sigma > 0$ such that for all $\boldsymbol{\alpha}$,

$$\mathbb{E}(\exp(\boldsymbol{\alpha}^\top\mathbf{Z})) \leq \exp(\|\boldsymbol{\alpha}\|^2\sigma^2/2).$$

Following Theorem 2.1 and Remark 2.1 in Hsu et al. (2012), we have for any $t > 0$,

$$\mathbb{P}\left(\|\mathbf{Z}\|^2 > \sigma^2\{p_n + 2(p_n t)^{1/2} + 2t\}\right) \leq \exp(-t).$$

Some calculations yield that

$$\mathbb{P}\left(\max_{i\in[n]}\|\mathbf{Z}_i\| > 2\sigma(\log n \vee p_n)^{1/2}\right) \to 0,$$

where $a \vee b = \max(a, b)$. Furthermore, we have

$$\mathbb{P}\left(\max_{i\in[n]}\max_{k\in[K_n]}\|\mathbf{Z}_{i(k)}\| > 2\sigma\{\log(nK_n) \vee \max_k|\mathcal{A}_k|\}^{1/2}\right) \to 0,$$

under condition C2,

$$\mathbb{P}\left(\max_{i\in[n]}\max_{k\in[K_n]}\|\mathbf{Z}_{i(k)}\| > 2\sigma(\log(nK_n))^{1/2}\right) \to 0,$$

which completes the proof of (9).

It follows from condition C3 that there exists a constant c_* such that $\sup_k \sup_{\boldsymbol{\beta}_k \in \mathcal{B}_k} \|\boldsymbol{\beta}_k\| \leq c_*$. The Cauchy–Schwarz inequality furhter implies that $\sup_{\boldsymbol{\beta}_k \in \mathcal{B}_k} \exp(\mathbf{Z}_{i(k)}^\top \boldsymbol{\beta}_k) \leq \exp\{c_* \|\mathbf{Z}_{i(k)}\|\}$. Hence, for any $t > 0$, it holds that

$$\mathbb{P}\left(\max_{i\in[n]} \max_{k\in[K_n]} \sup_{\boldsymbol{\beta}_k \in \mathcal{B}_k} \exp(\mathbf{Z}_{i(k)}^\top \boldsymbol{\beta}_k) > t \right) \leq \mathbb{P}\left(\max_{i\in[n]} \max_{k\in[K_n]} \|\mathbf{Z}_{i(k)}\| > \log t / c_* \right).$$

Set $\log t / c_* = 2\sigma(\log(nK_n))^{1/2}$, then $t = \exp\{2c_*\sigma(\log(nK_n))^{1/2}\}$, which indicates

$$\max_{i\in[n]} \max_{k\in[K_n]} \sup_{\boldsymbol{\beta}_k \in \mathcal{B}_k} \exp(\mathbf{Z}_{i(k)}^\top \boldsymbol{\beta}_k) = O_p\left(\exp\{2c_*\sigma(\log(nK_n))^{1/2}\} \right).$$

For any constant $c > 0$, it holds $c \leq (\log(nK_n))^{1/6}$ if n is sufficiently large. Then we have

$$\exp\{c(\log(nK_n))^{1/2}\} = (nK_n)^{c(\log(nK_n))^{-1/2}} \leq \varrho_n.$$

Consequently, (10) follows immediately as $c_*\sigma$ can be bounded uniformly under conditions C1 and C3. ■

Lemma 3 *Assume that all the items related to t are continuous with respect to $t \in [0, \tau]$. Under conditions C1–C3, for $k = 1, \ldots, K_n$, it holds that,*

$$\sup_{t\in[0,\tau]} \left| \boldsymbol{\mu}(t)^\top \{\mathbf{N}(t) - \boldsymbol{\mu}(t)\} \right| = O_p\left((n \log n)^{1/2} \right), \tag{11}$$

$$\sup_{t\in[0,\tau]} \left| \boldsymbol{\mu}_k^0(t)^\top \{\mathbf{N}(t) - \boldsymbol{\mu}(t)\} \right| = O_p\left(\varrho_n (n \log n)^{1/2} \right), \tag{12}$$

and

$$\sup_{t\in[0,\tau]} \left| \mathbf{N}(t)^\top \{\mathbf{N}(t) - \boldsymbol{\mu}(t)\} - \mathbb{E}[\mathbf{N}(t)^\top \{\mathbf{N}(t) - \boldsymbol{\mu}(t)\}] \right| = O_p\left((n \log n)^{1/2} \right). \tag{13}$$

Proof For simplicity, denote $G_n^*(t) = n^{-1} \sum_{i=1}^n N_i(t)\mu_i(t)$ and $G_n(t) = n^{-1} \sum_{i=1}^n \mu_i^2(t)$. Choose $0 = t_1 < \cdots < t_{d_n} = \tau$ such that $G_n(t_{j+1}) - G_n(t_j) = (n^{-1} \log n)^{1/2}$; thus $d_n \leq (n/\log n)^{1/2}$. It further holds that

$$\sup_{t\in[0,\tau]} |G_n^*(t) - G_n(t)| \leq 2 \max_{1\leq j\leq d_n} |G_n^*(t_j) - G_n(t_j)| + 2(n^{-1} \log n)^{1/2}$$

for sufficiently large n. On the other hand, note that, for $i = 1, \ldots, n$, $N_i(t_j)\mu_i(t_j) - \mu_i^2(t_j)$ are independent, $\mathbb{E}\left\{N_i(t_j)\mu_i(t_j) - \mu_i^2(t_j)\right\} = 0$, and $\mathbb{P}(\sup_{t\in[0,\tau]} |N_i(t)\mu_i(t) - \mu_i^2(t)| \leq 1) =$

1. Following the Hoeffding inequality (Hoeffding, 1963), we have

$$\begin{aligned}
& \mathbb{P}\left(\sup_{t\in[0,\tau]}|G_n^*(t)-G_n(t)|\geq(2^{3/2}+2)(n^{-1}\log n)^{1/2}\right)\\
\leq\ & \mathbb{P}\left(\max_{1\leq j\leq d_n}|G_n^*(t_j)-G_n(t_j)|\geq(2n^{-1}\log n)^{1/2}\right)\\
\leq\ & \sum_{j=1}^{d_n}\mathbb{P}\left(|G_n^*(t_j)-G_n(t_j)|\geq(2\log n/n)^{1/2}\right)\\
\leq\ & 2\sum_{j=1}^{d_n}\exp\left\{-2n(2n^{-1}\log n)/4\right\}\\
\leq\ & 2n^{-1}d_n\leq 2(n\log n)^{-1/2}\to 0,
\end{aligned}$$

which completes the proof of (11). Furthermore, (12) and (13) follow by similar arguments and Lemma 2. ■

Lemma 4 *Under conditions C1–C3 and C5–C6, it holds that*

$$\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\left|\frac{L_n(\boldsymbol{\omega},t)}{R_n(\boldsymbol{\omega},t)}-1\right|\to 0$$

in probability.

Proof Under conditions C2 and C5, it holds

$$\begin{aligned}
& \mathbb{P}\left(\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2\geq\epsilon\right)\\
=\ & \mathbb{P}\left(\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|\geq\epsilon^{1/2}\right)\\
\leq\ & 2\exp\{-c_\ddagger n^{-1/2}K_n^{-2}\epsilon\}.
\end{aligned}$$

Setting $\epsilon=(n\log n)^{1/2}K_n^2$, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2\geq(n\log n)^{1/2}K_n^2\right)\leq 2\exp(-c_\ddagger(\log n)^{1/2})\to 0.$$

Consequently, we further have

$$\mathbb{E}\left(\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2\right)\leq\int_0^\infty 2\exp\{-c_\ddagger n^{-1/2}K_n^{-2}\epsilon\}\mathrm{d}\epsilon=2c_\ddagger^{-1}n^{1/2}K_n^2.$$

It follows from the Markov inequality that

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{E}\left(\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2 \mid \mathcal{Z}_n\right)\geq (n\log n)^{1/2}K_n^2\right\}\\
\leq\ & \frac{\mathbb{E}\left(\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2\right)}{(n\log n)^{1/2}K_n^2}\to 0
\end{aligned}$$

as $n\to\infty$, where $\mathcal{Z}_n=(\mathbf{Z}_1,\ldots,\mathbf{Z}_n)$. We conclude that

$$\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2 = O_p\left((n\log n)^{1/2}K_n^2\right), \tag{14}$$

$$\mathbb{E}\left(\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2|\mathcal{Z}_n\right) = O_p\left((n\log n)^{1/2}K_n^2\right). \tag{15}$$

Noting that $\widehat{\boldsymbol{\beta}}_{*k}-\boldsymbol{\beta}_{0k}=O_p(n^{-1/2})$, $\widehat{\Lambda}_{*k}(t)-\Lambda_{0k}(t)=O_p(n^{-1/2})$, and $\Lambda_{0k}(\tau)$ is bounded, under conditions C1–C3, for any constant c_1, we have

$$\mathbb{P}\left(\max_{i\in[n]}\max_{k\in[K_n]}\sup_{t\in[0,\tau]}\widehat{\boldsymbol{\mu}}_{ik}(t)\geq c_1\varrho_n\right)\leq\mathbb{P}\left(\max_{i\in[n]}\max_{k\in[K_n]}\sup_{t\in[0,\tau]}\boldsymbol{\mu}^0_{ik}(t)\geq c_1\varrho_n/2\right)\to 0,$$

which implies that

$$\max_{i\in[n]}\max_{k\in[K_n]}\sup_{t\in[0,\tau]}\widehat{\mu}_{ik}(t)=O_p\left(\varrho_n\right).$$

Under conditions C1–C3 and using Lemma 12.6 of Kosorok (2008), for every individual, we have

$$\left\|\widehat{\mu}_{ik}(t)\{\widehat{\mu}_{ij}(t)-\mu^0_{ij}(t)\}\right\|_\infty\leq c_1\varrho_n\|\exp\{\mathbf{Z}^\top_{i(j)}\widehat{\boldsymbol{\beta}}_{*j}\}\widehat{\Lambda}_{*j}(t)-\exp\{\mathbf{Z}^\top_{i(j)}\boldsymbol{\beta}_{0j}\}\Lambda_{0j}(t)\|_\infty$$

on the event $\mathcal{E}_{\varrho_n}=\{\max_{i\in[n]}\max_{k\in[K_n]}\sup_{t\in[0,\tau]}\widehat{\mu}_{ik}(t)\leq c_1\varrho_n\}$. Furthermore, it follows from conditions C1–C3 and Goldberg and Kosorok (2012) that

$$\mathbb{P}\left(n^{1/2}\max_{i\in[n]}\max_{k,j\in[K_n]}\sup_{t\in[0,\tau]}\left|\widehat{\mu}_{ik}(t)\{\widehat{\mu}_{ij}(t)-\mu^0_{ij}(t)\}\right|>\varrho_n\epsilon\mid\mathcal{E}_{\varrho_n}\right)\leq 2\exp\{-c_2(\varrho_n)^{-2}\epsilon^2\},$$

where c_2 is a universal constant that depends $\mathcal{Z}$ and the bound of the parametric space. As a result,

$$\begin{aligned}
&\mathbb{P}\left(n^{1/2}\max_{i\in[n]}\max_{k,j\in[K_n]}\sup_{t\in[0,\tau]}\left|\widehat{\mu}_{ik}(t)\{\widehat{\mu}_{ij}(t)-\mu^0_{ij}(t)\}\right|>\varrho_n\epsilon\right)\\
\leq\ & 2\exp\{-c_2(\varrho_n)^{-2}\epsilon^2\}+\mathbb{P}(\mathcal{E}^c_{\varrho_n})\to 0.
\end{aligned}$$

Setting $\epsilon=(\log n)^{1/2}\varrho_n$ and based on some simple calculations, we obtain

$$\max_{k,j\in[K_n]}\sup_{t\in[0,\tau]}\left|\widehat{\boldsymbol{\mu}}_k(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}^0_j(t)\}\right| = O_p\left(\varrho_n^2(n\log n)^{1/2}\right), \tag{16}$$

$$\mathbb{E}\left(\max_{k,j\in[K_n]}\sup_{t\in[0,\tau]}\left|\widehat{\boldsymbol{\mu}}_k(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}^0_j(t)\}\right|\mid\mathcal{Z}_n\right) = O_p\left(\varrho_n^2(n\log n)^{1/2}\right). \tag{17}$$

Similarly, we can conclude

$$\max_{k,j\in[K_n]}\sup_{t\in[0,\tau]}\left|\{\boldsymbol{\mu}_k^0(t)\}^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| = O_p\left(\varrho_n^2(n\log n)^{1/2}\right), \quad (18)$$

$$\max_{j\in[K_n]}\sup_{t\in[0,\tau]}\left|\boldsymbol{\mu}(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| = O_p\left(\varrho_n(n\log n)^{1/2}\right), \quad (19)$$

$$\mathbb{E}\left(\max_{k,j\in[K_n]}\sup_{t\in[0,\tau]}\left|\{\boldsymbol{\mu}_k^0(t)\}^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| \mid \mathcal{Z}_n\right) = O_p\left(\varrho_n^2(n\log n)^{1/2}\right), \quad (20)$$

and

$$\mathbb{E}\left(\max_{j\in[K_n]}\sup_{t\in[0,\tau]}\left|\boldsymbol{\mu}(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right|\Big| \mathcal{Z}_n\right) = O_p\left(\varrho_n^2(n\log n)^{1/2}\right). \quad (21)$$

Note that

$$\begin{aligned}
&L_n(\boldsymbol{\omega},t)-R_n(\boldsymbol{\omega},t)\\
=&\|\widehat{\boldsymbol{\mu}}(t)-\boldsymbol{\mu}^0(t)\|^2-\mathbb{E}\{\|\widehat{\boldsymbol{\mu}}(t)-\boldsymbol{\mu}^0(t)\|^2\mid\mathcal{Z}_n\}-2\{\widehat{\boldsymbol{\mu}}(t)-\boldsymbol{\mu}^0(t)\}^\top\{\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\}\\
&+2\mathbb{E}[\{\widehat{\boldsymbol{\mu}}(t)-\boldsymbol{\mu}^0(t)\}^\top\{\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\}\mid\mathcal{Z}_n]+\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2-\mathbb{E}\{\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2\mid\mathcal{Z}_n\}\\
\le&\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\sum_{k=1}^{K_n}\sum_{j=1}^{K_n}\omega_k\omega_j\left|\{\widehat{\boldsymbol{\mu}}_k(t)-\boldsymbol{\mu}_k^0(t)\}^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right|\\
&+\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}2\sum_{k=1}^{K_n}\sum_{j=1}^{K_n}\omega_k\omega_j\left|\{\boldsymbol{\mu}_k^0(t)\}^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right|\\
&+\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}2\sum_{j=1}^{K_n}\omega_j\left|\boldsymbol{\mu}(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right|+\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2\\
&+\mathbb{E}\left\{\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\sum_{k=1}^{K_n}\sum_{j=1}^{K_n}\omega_k\omega_j\left|\{\widehat{\boldsymbol{\mu}}_k(t)-\boldsymbol{\mu}_k^0(t)\}^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right|\mid\mathcal{Z}_n\right\}\\
&+2\mathbb{E}\left\{\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\sum_{k=1}^{K_n}\sum_{j=1}^{K_n}\omega_k\omega_j\left|\{\boldsymbol{\mu}_k^0(t)\}^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right|\mid\mathcal{Z}_n\right\}\\
&+2\mathbb{E}\left\{\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\sum_{j=1}^{K_n}\omega_j\left|\boldsymbol{\mu}(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right|\mid\mathcal{Z}_n\right\}\\
&+\mathbb{E}\left\{\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2\mid\mathcal{Z}_n\right\}
\end{aligned}$$

$$
\begin{aligned}
\leq & K_n^2 \max_{k,j\in[K_n]} \sup_{t\in[0,\tau]} \left|\widehat{\boldsymbol{\mu}}_k(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| + 3K_n^2 \max_{k,j\in[K_n]} \sup_{t\in[0,\tau]} \left|\{\boldsymbol{\mu}_k^0(t)\}^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| \\
& + 2K_n \max_{j\in[K_n]} \sup_{t\in[0,\tau]} \left|\boldsymbol{\mu}(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| + \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} \|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2 \\
& + K_n^2\mathbb{E}\left\{\max_{k,j\in[K_n]} \sup_{t\in[0,\tau]} \left|\widehat{\boldsymbol{\mu}}_k(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| \mid \mathcal{Z}_n\right\} \\
& + 3K_n^2\mathbb{E}\left\{\max_{k,j\in[K_n]} \sup_{t\in[0,\tau]} \left|\{\boldsymbol{\mu}_k^0(t)\}^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| \mid \mathcal{Z}_n\right\} \\
& + 2K_n\mathbb{E}\left\{\max_{j\in[K_n]} \sup_{t\in[0,\tau]} \left|\boldsymbol{\mu}(t)^\top\{\widehat{\boldsymbol{\mu}}_j(t)-\boldsymbol{\mu}_j^0(t)\}\right| \mid \mathcal{Z}_n\right\} + \mathbb{E}\left\{\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} \|\boldsymbol{\mu}(t)-\boldsymbol{\mu}^0(t)\|^2 \mid \mathcal{Z}_n\right\}.
\end{aligned}
$$

Therefore, it follows from (14) to (21) that

$$
\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} |L_n(\boldsymbol{\omega},t)-R_n(\boldsymbol{\omega},t)| = (n\log n)^{1/2}K_n^2\varrho_n^2 O_p(1).
$$

Thus, conditions C2 and C6 yield

$$
\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} \frac{|L_n(\boldsymbol{\omega},t)-R_n(\boldsymbol{\omega},t)|}{R_n(\boldsymbol{\omega},t)} \leq a_n^{-1}(n\log n)^{1/2}K_n^2\varrho_n^2 O_p(1) \to 0,
$$

which completes the proof of Lemma 4. ■

Lemma 5 *Under conditions C1–C6, it holds that*

$$
\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} \left|\frac{\widetilde{L}_n(\boldsymbol{\omega},t)}{L_n(\boldsymbol{\omega},t)}-1\right| \to 0 \tag{22}
$$

in probability, where $\widetilde{L}_n(\boldsymbol{\omega},t) = \|\boldsymbol{\mu}(t)-\widetilde{\boldsymbol{\mu}}(t)\|^2$.

Proof Following Lemmas 1 and 2, (10) and conditions C1–C4, we have

$$
\begin{aligned}
& \max_{k\in[K_n]} \sup_{t\in[0,\tau]} \|\widetilde{\boldsymbol{\mu}}_k(t)-\widehat{\boldsymbol{\mu}}_k(t)\|^2 \\
= & \sum_{s=1}^n \max_{k\in[K_n]} \sup_{t\in[0,\tau]} \left(\int_0^t Y_s(u)\exp(\mathbf{Z}_{s(k)}^\top\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})\mathrm{d}\widetilde{\Lambda}_{*k}^{(-s)}(u) - \int_0^t Y_s(u)\exp(\mathbf{Z}_{s(k)}^\top\widehat{\boldsymbol{\beta}}_{*k})\mathrm{d}\widehat{\Lambda}_{*k}(u)\right)^2 \\
\leq & 2\max_{k\in[K_n]} \sum_{s=1}^n \sup_{t\in[0,\tau]} \left(\int_0^t Y_s(u)\exp(\mathbf{Z}_{s(k)}^\top\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})\mathrm{d}\{\widetilde{\Lambda}_{*k}^{(-s)}(u)-\mathrm{d}\widehat{\Lambda}_{*k}(u)\}\right)^2 \\
& + 2\max_{k\in[K_n]} \sum_{s=1}^n \sup_{t\in[0,\tau]} \left(\int_0^t Y_s(u)\{\exp(\mathbf{Z}_{s(k)}^\top\widetilde{\boldsymbol{\beta}}_{*k}^{(-s)})-\exp(\mathbf{Z}_{s(k)}^\top\widehat{\boldsymbol{\beta}}_{*k})\}\mathrm{d}\widehat{\Lambda}_{*k}(u)\right)^2 \\
= & O_p\left(n^{-1}\log(nK_n)\varrho_n^2\right).
\end{aligned}
$$

As a result,

$$\begin{aligned}\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\|\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\|^2 &= \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\left\|\sum_{k=1}^{K_n}\omega_k\{\widetilde{\boldsymbol{\mu}}_k(t)-\widehat{\boldsymbol{\mu}}_k(t)\}\right\|^2\\ &\le K_n^2\max_{k\in[K_n]}\sup_{t\in[0,\tau]}\|\widetilde{\boldsymbol{\mu}}_k(t)-\widehat{\boldsymbol{\mu}}_k(t)\|^2\\ &\le n^{-1}K_n^2\log(nK_n)\varrho_n^2 O_p(1).\end{aligned}$$

Under conditions C2 and C6, and using Lemma 4, we have

$$\begin{aligned}\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\frac{\|\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\|^2}{L_n(\boldsymbol{\omega},t)} &= \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\frac{\|\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\|^2}{R_n(\boldsymbol{\omega},t)}\frac{R_n(\boldsymbol{\omega},t)}{L_n(\boldsymbol{\omega},t)}\\ &\le \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\frac{\|\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\|^2}{R_n(\boldsymbol{\omega},t)}\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\frac{R_n(\boldsymbol{\omega},t)}{L_n(\boldsymbol{\omega},t)}\\ &\le \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\frac{\|\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\|^2}{a_n}\\ &= a_n^{-1}n^{-1}K_n^2\log(nK_n)\varrho_n^2O_p(1)\to 0 \qquad (23)\end{aligned}$$

in probability. On the other hand, it follows from the Cauchy–Schwartz inequality that

$$\begin{aligned}|\widetilde{L}_n(\boldsymbol{\omega},t)-L_n(\boldsymbol{\omega},t)| &= |\|\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\|^2-2\langle\boldsymbol{\mu}(t)-\widehat{\boldsymbol{\mu}}(t),\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\rangle|\\ &\le \|\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\|^2+2\{L_n(\boldsymbol{\omega},t)\}^{1/2}\|\widetilde{\boldsymbol{\mu}}(t)-\widehat{\boldsymbol{\mu}}(t)\|. \qquad (24)\end{aligned}$$

Consequently, (22) follows directly from (23) and (24). ■

Lemma 6 *Under conditions C1–C4 and C6, it holds that*

$$\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\frac{|\langle\mathbf{N}(t)-\boldsymbol{\mu}(t),\boldsymbol{\mu}(t)-\widetilde{\boldsymbol{\mu}}(t)\rangle|}{R_n(\boldsymbol{\omega},t)}\to 0$$

in probability.

Proof Using Lemmas 1–3, (10) and conditions C1–C4, we have

$$\begin{aligned}&\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}|\langle\mathbf{N}(t)-\boldsymbol{\mu}(t),\boldsymbol{\mu}(t)-\widetilde{\boldsymbol{\mu}}(t)\rangle|\\ =\ &\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\left|\sum_{i=1}^{n}\left\{\{N_i(t)-\mu_i(t)\}\left(\mu_i(t)-\sum_{k=1}^{K_n}\omega_k\widetilde{\mu}_{ik}(t)\right)\right\}\right|\end{aligned}$$

$$
\begin{aligned}
\leq\ & \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\left|\sum_{i=1}^{n}\{N_i(t)-\mu_i(t)\}\mu_i(t)\right| + \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\sum_{k=1}^{K_n}\omega_k\left|\sum_{i=1}^{n}\{N_i(t)-\mu_i(t)\}\mu_{ik}^0(t)\right| \\
& + \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\left|\sum_{i=1}^{n}\left(\{N_i(t)-\mu_i(t)\}\sum_{k=1}^{K_n}\omega_k\{\mu_{ik}(t,\widehat{\boldsymbol{\beta}}_{*k},\Lambda_{0k})-\mu_{ik}^0(t)\}\right)\right| \\
& + \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\left|\sum_{i=1}^{n}\left(\{N_i(t)-\mu_i(t)\}\sum_{k=1}^{K_n}\omega_k\{\widehat{\mu}_{ik}(t)-\mu_{ik}(t,\widehat{\boldsymbol{\beta}}_{*k},\Lambda_{0k})\}\right)\right| \\
& + \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\left|\sum_{i=1}^{n}\left(\{N_i(t)-\mu_i(t)\}\sum_{k=1}^{K_n}\omega_k\{\mu_{ik}(t,\widetilde{\boldsymbol{\beta}}_{*k}^{(-i)},\widehat{\Lambda}_{*k})-\widehat{\mu}_{ik}(t)\}\right)\right| \\
& + \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\left|\sum_{i=1}^{n}\left(\{N_i(t)-\mu_i(t)\}\sum_{k=1}^{K_n}\omega_k\{\widetilde{\mu}_{ik}(t)-\mu_{ik}(t,\widetilde{\boldsymbol{\beta}}_{*k}^{(-i)},\widehat{\Lambda}_{*k})\}\right)\right| \\
=\ & O_p((n\log n)^{1/2}) + K_nO_p((n\log n)^{1/2}\varrho_n) + O_p(n)K_n\varrho_n(\log(nK_n))^{1/2}O_p(n^{-1/2}) \\
& + O_p(n)K_n\varrho_nO_p(n^{-1/2}) + O_p(n)K_n\varrho_n(\log(nK_n))^{1/2}O_p(n^{-1}) + O_p(n)K_n\varrho_nO_p(n^{-1}) \\
=\ & K_n(n\log(nK_n))^{1/2}\varrho_nO_p(1).
\end{aligned}
$$

It follows from conditions C2 and C6 that

$$
\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n}\sup_{t\in[0,\tau]}\frac{|\langle\mathbf{N}(t)-\boldsymbol{\mu}(t),\boldsymbol{\mu}(t)-\widetilde{\boldsymbol{\mu}}(t)\rangle|}{R_n(\boldsymbol{\omega},t)} \leq a_n^{-1}K_n(n\log(nK_n))^{1/2}\varrho_nO_p(1)\to 0
$$

in probability. Thus it completes the proof. ■

Lemma 7 *Under conditions C1–C6, it holds that*

$$
\mathrm{CV}_n(\boldsymbol{\omega},t)-\|\mathbf{N}(t)-\boldsymbol{\mu}(t)\|^2 = L_n(\boldsymbol{\omega},t)\{1+o_p(1)\},
$$

where $o_p(1)$ *is uniform in* $\boldsymbol{\omega}\in\boldsymbol{\Omega}_n$ *and* $t\in[0,\tau]$.

Proof We rewrite

$$
\begin{aligned}
& \mathrm{CV}_n(\boldsymbol{\omega},t) \\
=\ & \|\mathbf{N}(t)-\boldsymbol{\mu}(t)\|^2+\widetilde{L}_n(\boldsymbol{\omega},t)+2\langle\mathbf{N}(t)-\boldsymbol{\mu}(t),\boldsymbol{\mu}(t)-\widetilde{\boldsymbol{\mu}}(t)\rangle \\
=\ & \|\mathbf{N}(t)-\boldsymbol{\mu}(t)\|^2+L_n(\boldsymbol{\omega},t)\left(\frac{\widetilde{L}_n(\boldsymbol{\omega},t)}{L_n(\boldsymbol{\omega},t)}+\frac{2\langle\mathbf{N}(t)-\boldsymbol{\mu}(t),\boldsymbol{\mu}(t)-\widetilde{\boldsymbol{\mu}}(t)\rangle/R_n(\boldsymbol{\omega},t)}{L_n(\boldsymbol{\omega},t)/R_n(\boldsymbol{\omega},t)}\right).
\end{aligned}
$$

Consequently, Lemma 7 follows directly from Lemmas 4–6. ■

Proof of Theorem 1 Setting $t=\tau$ in Lemma 7, we have

$$
\mathrm{ECV}_n(\boldsymbol{\omega})-\|\mathbf{N}(\tau)-\boldsymbol{\mu}(\tau)\|^2 = L_n^{\mathrm{E}}(\boldsymbol{\omega})\{1+o_p(1)\},
$$

where $o_p(1)$ is uniform in $\boldsymbol{\omega} \in \boldsymbol{\Omega}_n$. Based on the definition of $\widehat{\boldsymbol{\omega}}_{\mathrm{E}}$, we obtain that

$$\frac{L_n^{\mathrm{E}}(\widehat{\boldsymbol{\omega}}_{\mathrm{E}})}{\inf_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} L_n^{\mathrm{E}}(\boldsymbol{\omega})} \to 1$$

in probability, which completes the proof of Theorem 1. ■

Proof of Theorem 2 As shown in the proof of Lemma 7 and utilizing the fact that the integral is a linear operator, we have

$$\mathrm{ICV}_n(\boldsymbol{\omega}) - \int_0^\tau \|\mathbf{N}(t) - \boldsymbol{\mu}(t)\|^2 \mathrm{d}t = L_n^{\mathrm{I}}(\boldsymbol{\omega})\{1 + o_p(1)\},$$

where $o_p(1)$ is uniform over $\boldsymbol{\omega} \in \boldsymbol{\Omega}_n$. Hence, Theorem 2 follows directly. ■

Proof of Theorem 3 Denote $R_n^{\mathrm{S}}(\boldsymbol{\omega}) = \sup_{t\in[0,\tau]} R_n(\boldsymbol{\omega}, t)$. Further calculation yields that

$$\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \left| \frac{L_n^{\mathrm{S}}(\boldsymbol{\omega})}{R_n^{\mathrm{S}}(\boldsymbol{\omega})} - 1 \right| \le \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} \left| \frac{L_n(\boldsymbol{\omega}, t)}{R_n(\boldsymbol{\omega}, t)} - 1 \right| \to 0$$

in probability using Lemma 4. Equations (11) and (13) in Lemma 3 imply that

$$\begin{aligned}
\sup_{t\in[0,\tau]} \|\mathbf{N}(t) - \boldsymbol{\mu}(t)\|^2 &\le \sup_{t\in[0,\tau]} \left| \sum_{i=1}^n (\{N_i(t) - \mu_i(t)\}N_i(t) - \mathbb{E}[\{N_i(t) - \mu_i(t)\}N_i(t)]) \right| \\
&\quad + \sup_{t\in[0,\tau]} \left| \sum_{i=1}^n \mathbb{E}[\{N_i(t) - \mu_i(t)\}N_i(t)] \right| \\
&\quad + \sup_{t\in[0,\tau]} \left| \sum_{i=1}^n \{N_i(t) - \mu_i(t)\}\mu_i(t) \right| \\
&= O_p\left((n\log n)^{1/2}\right) + O_p(n) + O_p\left((n\log n)^{1/2}\right) \\
&= O_p(n).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \sup_{t\in[0,\tau]} \frac{\|\mathbf{N}(t) - \boldsymbol{\mu}(t)\|^2}{L_n^{\mathrm{S}}(\boldsymbol{\omega})} &\le \frac{\sup_{t\in[0,\tau]} \|\mathbf{N}(t) - \boldsymbol{\mu}(t)\|^2}{\inf_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} R_n^{\mathrm{S}}(\boldsymbol{\omega})} \sup_{\boldsymbol{\omega}\in\boldsymbol{\Omega}_n} \frac{R_n^{\mathrm{S}}(\boldsymbol{\omega})}{L_n^{\mathrm{S}}(\boldsymbol{\omega})} \\
&\le O_p(n)\,(a_n^*)^{-1}\,(1 + o_p(1)) \to 0
\end{aligned}$$

in probability by conditions C6 and C7.

Following Lemma 7, we have

$$\begin{aligned}
\mathrm{CV}_n(\boldsymbol{\omega}, t) &= L_n^{\mathrm{S}}(\boldsymbol{\omega}) \frac{L_n(\boldsymbol{\omega}, t)}{L_n^{\mathrm{S}}(\boldsymbol{\omega})} \left\{ \frac{\|\mathbf{N}(t) - \boldsymbol{\mu}(t)\|^2}{L_n(\boldsymbol{\omega}, t)} + 1 + o_p(1) \right\} \\
&= L_n^{\mathrm{S}}(\boldsymbol{\omega}) \left\{ \frac{\|\mathbf{N}(t) - \boldsymbol{\mu}(t)\|^2}{L_n^{\mathrm{S}}(\boldsymbol{\omega})} + \frac{L_n(\boldsymbol{\omega}, t)}{L_n^{\mathrm{S}}(\boldsymbol{\omega})} + o_p(1) \right\} \\
&= L_n(\boldsymbol{\omega}, t) + L_n^{\mathrm{S}}(\boldsymbol{\omega}) o_p(1),
\end{aligned}$$

which, by taking the supremum over $t \in [0, \tau]$ on both sides, implies that

$$\mathrm{SCV}_n(\boldsymbol{\omega}) = L_n^{\mathrm{S}}(\boldsymbol{\omega})\{1 + o_p(1)\}.$$

Theorem 3 follows directly. ■

Proof of Theorem 4 By the Taylor expansion, for fixed t we have

$$\mathrm{CV}_n(\boldsymbol{\omega}, t) = \mathrm{CV}_n(\boldsymbol{\omega}^\dagger, t) + (\boldsymbol{\omega} - \boldsymbol{\omega}^\dagger)^\top \nabla \mathrm{CV}_n(\boldsymbol{\omega}^\dagger, t) + \|\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\omega}}(t) - \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\omega}^\dagger}(t)\|^2 \tag{25}$$

for any $\boldsymbol{\omega}$ and $\boldsymbol{\omega}^\dagger \in \boldsymbol{\Omega}_n$, where $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\omega}}(t) = \sum_{k=1}^{K_n} \omega_k \widetilde{\boldsymbol{\mu}}_k(t)$ is adopted to emphasize its dependence on the weight $\boldsymbol{\omega}$. First, we derive the convergence of the greedy model averaging algorithm for the ECV criterion. Setting $t = \tau$, $\boldsymbol{\omega}^\dagger = \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}$ and $\boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)} + \alpha(\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)})$ in (25), it holds

$$\begin{aligned} & \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)} + \alpha(\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)})) - \alpha^2 \|\widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell+1)}}(\tau) - \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}}(\tau)\|^2 \\ = \; & \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}) + \alpha(\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)})^\top \nabla \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}). \end{aligned}$$

For any $\boldsymbol{\omega} \in \boldsymbol{\Omega}_n$, if we define $\Delta_{\mathrm{E}}^{(\ell)} = \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}) - \mathrm{ECV}_n(\boldsymbol{\omega})$, using the convexity of $\mathrm{ECV}_n(\cdot)$, we have

$$\begin{aligned} \Delta_{\mathrm{E}}^{(\ell)} &\le \langle \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)} - \boldsymbol{\omega}, \nabla \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}) \rangle \\ &\le \langle \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)} - \widehat{\boldsymbol{\gamma}}^{(\ell+1)}, \nabla \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}) \rangle. \end{aligned}$$

Consequently,

$$\begin{aligned} \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell+1)}) &\le \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)} + \alpha(\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)})) \\ &\le \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}) - \alpha \Delta_{\mathrm{E}}^{(\ell)} + \alpha^2 \|\widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell+1)}}(\tau) - \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}}(\tau)\|^2 \end{aligned} \tag{26}$$

for any $\alpha \in [0, 1]$. Immediately,

$$\begin{aligned} \Delta_{\mathrm{E}}^{(\ell+1)} &\le \Delta_{\mathrm{E}}^{(\ell)} + \min_{\alpha \in [0,1]} \left(-\alpha \Delta_{\mathrm{E}}^{(\ell)} + \alpha^2 \|\widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\gamma}}_{\mathrm{E}}^{(\ell+1)}}(\tau) - \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell)}}(\tau)\|^2 \right) \\ &\le \Delta_{\mathrm{E}}^{(\ell)} + \min_{\alpha \in [0,1]} (-\alpha \Delta_{\mathrm{E}}^{(\ell)} + 4\alpha^2 K_n^2 h_{\mathrm{E},n}). \end{aligned}$$

For $\ell = 0$, choosing $\alpha = 1$, then $\Delta_{\mathrm{E}}^{(1)} \le 4K_n^2 h_{\mathrm{E},n}$. Furthermore, $\Delta_{\mathrm{E}}^{(\ell+1)} \le \Delta_{\mathrm{E}}^{(\ell)}$ for any $\ell \ge 1$. It holds $\Delta_{\mathrm{E}}^{(\ell)} \le 4K_n^2 h_{\mathrm{E},n}$. For $\ell \ge 1$, choosing $\alpha = \Delta_{\mathrm{E}}^{(\ell)} / (8K_n^2 h_{\mathrm{E},n}) \in [0, 1/2]$, we have

$$\Delta_{\mathrm{E}}^{(\ell+1)} \le \Delta_{\mathrm{E}}^{(\ell)} - \frac{\{\Delta_{\mathrm{E}}^{(\ell)}\}^2}{16 K_n^2 h_{\mathrm{E},n}}.$$

This recursion implies that, for any $\ell \ge 1$ and $\boldsymbol{\omega} \in \boldsymbol{\Omega}_n$,

$$\Delta_{\mathrm{E}}^{(\ell+1)} = \mathrm{ECV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{E}}^{(\ell+1)}) - \mathrm{ECV}_n(\boldsymbol{\omega}) \le \frac{16 K_n^2 h_{\mathrm{E},n}}{\ell}.$$

Thus, we obtain the convergence of the greedy algorithm for the ECV criterion.

Assuming that the derivative with respect to $\boldsymbol{\omega}$ and the integral with respect to t are exchangeable for the ICV criterion and further noting that the integral operator is linear, we have

$$\begin{aligned}
\Delta_{\mathrm{I}}^{(\ell+1)} &\leq \Delta_{\mathrm{I}}^{(\ell)} + \min_{\alpha\in[0,1]} \left(-\alpha\Delta_{\mathrm{I}}^{(\ell)} + \alpha^2 \int_0^\tau \|\widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\gamma}}_{\mathrm{I}}^{(\ell+1)}}(t) - \widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\omega}}_{\mathrm{I}}^{(\ell)}}(t)\|^2 \mathrm{d}t \right) \\
&\leq \Delta_{\mathrm{I}}^{(\ell)} + \min_{\alpha\in[0,1]} \left(-\alpha\Delta_{\mathrm{I}}^{(\ell)} + \int_0^\tau 2\alpha^2 \|\widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\gamma}}_{\mathrm{I}}^{(\ell+1)}}(t)\|^2 \mathrm{d}t + \int_0^\tau 2\alpha^2 \|\widetilde{\boldsymbol{\mu}}_{\widehat{\boldsymbol{\omega}}_{\mathrm{I}}^{(\ell)}}(t)\|^2 \mathrm{d}t \right) \\
&\leq \Delta_{\mathrm{I}}^{(\ell)} + \min_{\alpha\in[0,1]} (-\alpha\Delta_{\mathrm{I}}^{(\ell)} + 4\alpha^2 K_n^2 h_{\mathrm{I},n})
\end{aligned}$$

using similar argument for (26). Therefore, the convergence of the greedy algorithm for the ICV criterion follows.

We finally consider the greedy algorithm for the SCV criterion. As $\mathrm{SCV}_n(\boldsymbol{\omega})$ is convex with respect to $\boldsymbol{\omega}$, we have

$$\mathrm{SCV}_n(\boldsymbol{\omega}) = \mathrm{SCV}_n(\boldsymbol{\omega}^\dagger) + (\boldsymbol{\omega} - \boldsymbol{\omega}^\dagger)^\top \nabla \mathrm{SCV}_n(\boldsymbol{\omega}^\dagger) + \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\omega}^\dagger)^\top \nabla^2 \mathrm{SCV}_n(\boldsymbol{\omega}^\dagger)(\boldsymbol{\omega} - \boldsymbol{\omega}^\dagger).$$

Setting $\boldsymbol{\omega}^\dagger = \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}$ and $\boldsymbol{\omega} = \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)} + \alpha(\widehat{\boldsymbol{\gamma}}_{\mathrm{S}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)})$ yields

$$\begin{aligned}
& \mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)} + \alpha(\widehat{\boldsymbol{\gamma}}_{\mathrm{S}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)})) - \frac{\alpha^2}{2}(\widehat{\boldsymbol{\gamma}}_{\mathrm{S}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)})^\top \nabla^2 \mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)})(\widehat{\boldsymbol{\gamma}}_{\mathrm{S}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}) \\
= \; & \mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}) + \alpha(\widehat{\boldsymbol{\gamma}}_{\mathrm{S}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)})^\top \nabla \mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}).
\end{aligned}$$

If we define $\Delta_{\mathrm{S}}^{(\ell)} = \mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}) - \mathrm{SCV}_n(\boldsymbol{\omega})$, by the convexity of $\mathrm{SCV}_n(\cdot)$,

$$\begin{aligned}
\Delta_{\mathrm{S}}^{(\ell)} &\leq \langle \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)} - \boldsymbol{\omega}, \nabla \mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}) \rangle \\
&\leq \langle \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)} - \widehat{\boldsymbol{\gamma}}_{\mathrm{S}}^{(\ell+1)}, \nabla \mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}) \rangle.
\end{aligned}$$

By the definition of $h_{\mathrm{S},n}$ and using similar argument to that of the ECV criterion, we have

$$\begin{aligned}
\Delta_{\mathrm{S}}^{(\ell+1)} &\leq \Delta_{\mathrm{S}}^{(\ell)} + \min_{\alpha\in[0,1]} \left(-\alpha\Delta_{\mathrm{S}}^{(\ell)} + \frac{\alpha^2}{2}(\widehat{\boldsymbol{\gamma}}_{\mathrm{S}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)})^\top \nabla^2 \mathrm{SCV}_n(\widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)})(\widehat{\boldsymbol{\gamma}}_{\mathrm{S}}^{(\ell+1)} - \widehat{\boldsymbol{\omega}}_{\mathrm{S}}^{(\ell)}) \right) \\
&\leq \Delta_{\mathrm{S}}^{(\ell)} + \min_{\alpha\in[0,1]} (-\alpha\Delta_{\mathrm{S}}^{(\ell)} + 2\alpha^2 K_n h_{\mathrm{S},n}) \\
&\leq \Delta_{\mathrm{S}}^{(\ell)} - \frac{\{\Delta_{\mathrm{S}}^{(\ell)}\}^2}{8K_n h_{\mathrm{S},n}},
\end{aligned}$$

from which the convergence of the greedy algorithm for the SCV criterion follows. ■

References

Tomohiro Ando and Ker-Chau Li. A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505): 254–265, 2014.

Tomohiro Ando and Ker-Chau Li. A weight-relaxed model averaging approach for high dimensional generalized linear models. *The Annals of Statistics*, 45(6): 2645–2679, 2017.

Tomohiro Ando and Ruey Tsay. Predictive likelihood for Bayesian model selection and averaging. *Internation Journal of Forecasting*, 26(4): 744–763, 2010.

Per K. Andersen and Richard D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4): 1100–1120, 1982.

Anestis Antoniadis, Piotr Fryzlewicz, and Frédérique Letué. The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics*, 37(4): 531–552, 2010.

Norman E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, 43(1): 45–57, 1975.

Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4): 477–505, 2007.

David R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34(2): 187–220, 1972.

David R. Cox. Partial likelihood. *Biometrika*, 62(2): 269–276, 1975.

Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3): 1878–1905, 2012.

Jana Eklund and Sune Karlsson. Forecast combination and model averaging using predictive measures. *Econometric Reviews*, 26(2): 329–363, 2007.

Jianqing Fan and Runze Li. Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1): 74–99, 2002.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B*, 70(5): 849–911, 2008.

Yair Goldberg and Michael R. Kosorok. An exponential bound for Cox regression. *Statistics and Probability Letters*, 82(7): 1267–1272, 2012.

Bruce E. Hansen. Least squares model averaging. *Econometrica*, 75(4): 1175–1189, 2007.

Bruce E. Hansen and Jeffrey S. Racine. Jackknife model averaging. *Journal of Econometrics*, 167(1): 38–46, 2012.

Nils L. Hjort and Gerda Claeskens. Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association*, 101(476): 1449–1464, 2006.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4): 382–401, 1999.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301): 13–30, 1963.

Daniel Hsu, Sham D. Kakade, and Tong Zhang. A tail inequality for quadratic forms of sub-Gaussian random vectors. *Electronic Communications in Probability*, 17(52): 1–6, 2012.

Jian Huang, Tingni Sun, Zhiliang Ying, Yi Yu, and Cun-Hui Zhang. Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics*, 41(3): 1142–1165, 2013.

Michael R. Kosorok. *Intoduction to Empirical Process and Semiparametric Inference*. New York: Springer, 2008.

Hua Liang, Guohua Zou, Alan T. K. Wan, and Xinyu Zhang. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, 106(495): 1053–1066, 2011.

Danyu Lin and Lee-Jen Wei. The robust inference for Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408): 1074–1078, 1989.

Danyu Lin, Lee-Jen Wei, and Zhiliang Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B*, 62(4): 711–730, 2000.

Guannan Liu, Wei Long, Xinyu Zhang, and Qi Li. Detecting financial data dependence structure by averaging mixture copulas. *Econometric Theory*, 35(4): 777–815, 2019.

Andreas Rosenwald, George Wright, Adrian Wiestner, and Wing C. Chan et al. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, 3(2): 185–197, 2003.

Rui Song, Wenbin Lu, Shuangge Ma and X. Jessie Jeng. Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101(4): 799–814, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1): 267–288, 1996.

Hajime Uno, Tianxi Cai, Michael Pencina, Ralph D'Agostion, and Lee-Jen Wei. On the C-statisics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10): 1105–1117, 2011.

Alan T. K. Wan, Xinyu Zhang, and Guohua Zou. Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156(2): 277–283, 2010.

Yuanshan Wu and Guosheng Yin. Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika*, 102(1): 65–76, 2015.

Cunhui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942, 2010.

Hao Helen Zhang and Wenbin Lu. Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94(3): 691–703, 2007.

Xinyu Zhang and Hua Liang. Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, 39(1): 174–200, 2011.

Xinyu Zhang, Guohua Zou, and Hua Liang. Model averaging and weight choice in linear mixed-effects models. *Biometrika*, 101(1): 205–218, 2014.

Xinyu Zhang, Dalei Yu, Guohua Zou, and Hua Liang. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111(516): 1775–1190, 2016.

Rong Zhu, Alan, T. K. Wan, Xinyu Zhang, and Guohua Zou. A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, 114(526): 882–892, 2019.

Sihai Dave Zhao and Yi Li. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1): 397–411, 2012.