Forward regression for Cox models with high dimensional covariates

Hyokyoung G. Hong

Department of Statistics and Probability, Michigan State University, U.S.A.

Qi Zheng

Department of Bioinformatics and Biostatistics, University of Louisville, U.S.A

Yi Li

Department of Biostatistics, University of Michigan, U.S.A.

Abstract

Forward regression, a classical variable screening method, has been widely used for model building when the number of covariates is relatively low. However, forward regression is seldom used in high-dimensional settings because of the cumbersome computation and unknown theoretical properties. Some recent works have shown that forward regression, coupled with an extended Bayesian information criterion (EBIC)-based stopping rule, can consistently identify all relevant predictors in high-dimensional linear regression settings. However, the results are based on the sum of residual squares from linear models and it is unclear whether forward regression can be applied to more general regression settings, such as Cox proportional hazards models. We introduce a forward variable selection procedure for Cox models. It selects important variables sequentially according to the increment of partial likelihood, with an EBIC stopping rule. To our knowledge, this is the first study that investigates the partial likelihood-based forward regression in high-dimensional survival settings and establishes selection consistency results. We show that, if the dimension of the true model is finite, forward regression can discover all relevant predictors within a finite number of steps and their order of entry is determined by the size of the increment in partial likelihood. As partial likelihood is not a regular density-based likelihood, we develop some new theoretical results on partial likelihood and use these results to establish the desired sure screening properties. The practical utility of the proposed method is examined via extensive simulations and analysis of a subset of the Boston Lung Cancer Survival Cohort study, a hospital-based study for identifying biomarkers related to lung cancer patients' survival.

Keywords: Forward selection, partial likelihood, sure screening properties, extended Bayesian information criteria, high-dimensional predictors

1. Introduction

New biotechnologies have generated a vast amount of high-throughput data. In the Boston Lung Cancer Survival Cohort study, a hospital-based study for lung cancer patients, identifying high-throughput predictors such as molecular profiles that are associated with patients' survival is a major research goal for understanding disease progression processes and designing more effective gene therapies. When the number of covariates (p) is less than the sample size (n), the Cox proportional hazards model has been routinely used for modeling survival data in many practical settings. When $p > n$, penalized partial likelihood methods have been proposed by various authors [1, 2] and the oracle properties and statistical error bounds of estimation have been established [3]. However, when $p \gg n$, these methods often fail because of serious challenges in "computational expediency, statistical accuracy, and algorithmic stability" [4]. Recently, [5] established the oracle properties of the regularized partial likelihood estimates under a high-dimensional setting. However, the results require the estimates to be unique and global optimizers, which is, in general, difficult to verify, especially when the dimension of covariates is exceedingly high.

Forward regression has been widely used for model selection, but it has often been criticized for not achieving selection consistency as it fails to account for multiple comparisons in the model building process. Recently, some authors, e.g. [6], [7], [8], [9], [10] and [11], have revamped forward regression in the context of linear regression or varying-coefficient linear models. The advantages can be summarized as follows. First, these authors have shown that, with some proper stopping criteria, forward regression can achieve screening consistency even in high dimensional settings. Second, the variables are sequentially selected into the final model with the entry order determined by the size of the likelihood increment, which might reflect the relative importance of each selected variable. Third, the implementation is simple as no cross-validation for tuning parameters is needed. Finally, the method only needs assumptions on the original model and does not require restrictive faithfulness assumptions, in which the marginal models reflect the original model. However, to our knowledge, the aforementioned forward regression approaches are either based on the sum of residual squares from linear models [6, 11] or Lasso estimation [7]. It is unclear whether forward regression can be applied to more general regression settings, such as the Cox proportional hazards models.

On the other hand, there has been active research in developing high-dimensional screening tools for survival data. The works include the principled sure screening by [12], the feature aberration at survival times screening by [13] and the conditional screening by [14], the quantile adaptive sure independence screening by [15], the censored rank independence screening procedure by [16], and the integrated powered density screening by [17]; see [18] for an extensive review. However, the screening methods require a threshold to dictate how many variables to retain, for which unfortunately there are no clear rules. [12] did tie the threshold with false discoveries, but it still needs to prefix the number of false positives that users are willing to tolerate. Recently, [19] designed a model-free measure, namely, survival impact index, that sensibly captures the overall influence of a covariate on the survival outcome and can help guide selecting important variables. However, even this method, like the other screening methods, does not directly lead to a final model, for which extra modeling steps have to be implemented.

We introduce a new forward variable selection procedure for survival data based on partial likelihood. It selects important variables sequentially according to the increment of partial likelihood, with a stopping rule based on EBIC. We show that if the dimension of the true model is finite, within a finite number of steps forward regression can discover all relevant predictors, with the entry order determined by the size of the likelihood increment.

Our work is novel in the following aspects. It likely registers as the first attempt to thoroughly investigate the forward regression in high-dimensional survival settings, methodologically, theoretically and numerically. Technically this paper is also novel. First, our work represents technical advances and a more broadened scope compared to the existing forward regression [6, 7, 11]. This may be the first work that investigates the partial likelihood-based forward regression in survival models with high-dimensional predictors, and establishes rigorous selection consistency results when the extended Bayesian information criterion (EBIC) [20] is used. It improves the partial likelihood-based variable selection developed by [21] and [22] for survival data in low dimensional settings. Second, as partial likelihood is not a regular density-based likelihood, it fails to satisfy the requirements for theories of forward regression. We revisit partial likelihood and develop some new inequalities, based on which we establish the desired sure screening properties. The derived theoretical framework and techniques will facilitate the extension of the procedure to other general likelihood-based settings, such as generalized linear regression models. Finally, we note that forward selection starts with an empty model or some important variables identified *a priori* and then sequentially recruits variables given important variables identified in the previous steps. This may resemble the conditional screening approach [14], which incorporates prior knowledge into variable screening. However, our method is valid even in the absence of such information.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed forward regression procedure. In Section 3, we rigorously establish forward regression's screening consistency and asymptotic normality under some regularity conditions. We carry out simulation studies to assess the performance of the proposed method in Section 4, and apply the method in Section 5 to analyze a subset of the Boston Lung Cancer Survival Cohort study, our motivating study for identifying biomarkers related to lung cancer patients' survival. We conclude the paper with a natural extension of the proposal in Section 6. Technical proofs and all of the lemmas are presented in the appendix.

2. Partial likelihood-based forward regression

Suppose we have n independent subjects with p covariates, where $p \gg n$. For subject i, denote by X_{ij} the jth covariate for subject i, write $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top$, and let T_i and C_i be the underlying survival and censoring times.

We, however, only observe $Y_i = \min\{T_i, C_i\}$, and the event indicator $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. We assume random censoring such that C_i and T_i are independent given $\mathbf{X}_i$. We assume that $(Y_i, \delta_i, \mathbf{X}_i)$ are independently and identically distributed (iid). In particular, we assume that $(Y_i, T_i, X_{ij}), i = 1, \dots, n$, are iid copies of (Y, T, X_j), the random variables that underlie the observed survival time, true survival time and covariates.

To link T_i to $\mathbf{X}_i$, we consider the following Cox proportional hazards model:

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t)\exp(\boldsymbol{\beta}_0^\top \mathbf{X}_i), \tag{1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function and $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0p})^\top$ is the vector of regression coefficients. Without loss of generality, we assume that $\mathrm{E}(X_j) = 0,\ j = 1, \dots, p$. Denote the true model as $\mathcal{M} = \{j : \beta_{0j} \neq 0\}$. The overarching goal of variable screening is to estimate $\mathcal{M}$ and we let its estimate be $\hat{\mathcal{M}}$.

We introduce more notation. For an index set $S \subset \{1, \dots, p\}$ and a p-dimensional vector $\mathbf{A}$, we use $\mathbf{A}_S = \{A_j : j \in S\}$ to denote the subvector of $\mathbf{A}$ corresponding to S. For example, $\mathbf{X}_{iS}$ denotes the collection of covariates for the ith individual corresponding to S. We use $|S|$ to denote the cardinality of S and let S^c denote the complement of S.

Now we elaborate on the idea of forward regression under model (1). Initialize $S_0 = \emptyset$. We can also start with a set of given variables according to some prior knowledge, which is in the same spirit as conditional screening [14] but is followed by a sequential selection process. Specifically, we sequentially select the sets of covariates as follows:

$$S_0 \subset S_1 \subset S_2 \subset \cdots \subset S_k,$$

where $S_k \subset \{1, \dots, p\}$ is the index set of the selected covariates upon completion of the kth step, with $k \geq 0$. Then at the $(k+1)$th step, we need to choose a new candidate variable not in S_k and then decide whether we should stop at the kth step or we should include the new candidate in our selection and proceed to the next step. We emphasize that our selection criterion is based on the partial likelihood. The framework is much broader than that of the one based on the reduction in sum of squared residuals proposed by [6] and [11] and can be extended to more general regression settings.

Now, given S_k, we consider estimation of the extended Cox model by adding a new variable index to S_k. Specifically, for every $j \in S_k^c$, we denote by $S_{k,j} = S_k \cup \{j\}$, and fit a Cox model on the variables indexed by $S_{k,j}$. We then compute the increment of log partial likelihood for each $j \in S_k^c$:

$$\ell_{S_{k,j}}(\hat{\boldsymbol{\beta}}_{S_{k,j}}) - \ell_{S_k}(\hat{\boldsymbol{\beta}}_{S_k}).$$

Here, for a covariate set S, $\ell_S(\boldsymbol{\beta}_S)$ is the log partial likelihood function given $\mathbf{X}_S$:

$$\ell_S(\boldsymbol{\beta}_S) = \sum_{i=1}^{n} \int_0^\tau \left[\boldsymbol{\beta}_S^\top \mathbf{X}_{iS} - \ln\left\{ \sum_{l=1}^{n} \bar{Y}_l(t)\exp(\boldsymbol{\beta}_S^\top \mathbf{X}_{lS}) \right\} \right] dN_i(t), \tag{2}$$

and $\hat{\boldsymbol{\beta}}_S$ maximizes (2), where $N_i(t) = I(Y_i \leq t, \delta_i = 1)$ is the counting process, $\bar{Y}_i(t) = I(Y_i \geq t)$ is the at-risk process, and $\tau > 0$ is the study duration such that $P(Y \geq \tau) > 0$. Then, the candidate index is chosen as

$$j^* = \underset{j \notin S_k}{\arg\max}\, \ell_{S_{k,j}}(\hat{\boldsymbol{\beta}}_{S_{k,j}}) - \ell_{S_k}(\hat{\boldsymbol{\beta}}_{S_k}).$$

Upon completion of the $(k+1)$th step, update $S_{k+1} = S_k \cup \{j^*\}$.

We are now in a position to decide whether to stop at the kth step or to include variable j^* in our selection and proceed to the next step. In the survival setting, the effective sample size is the number of uncensored events, in which case [21] showed that replacement of the sample size with the number of uncensored events in the penalty term of EBIC gives a better approximation to the Bayes factor. Therefore, we propose the following as the modified EBIC criterion for ultrahigh-dimensional survival data:

$$\begin{aligned} \mathrm{EBIC}(S_{k+1}) &= -2\ell_{S_{k+1}}(\hat{\boldsymbol{\beta}}_{S_{k+1}}) + |S_{k+1}|(\ln d + 2\eta \ln p) \\ &= -2\ell_{S_{k+1}}(\hat{\boldsymbol{\beta}}_{S_{k+1}}) + (k+1)(\ln d + 2\eta \ln p), \end{aligned} \tag{3}$$

where $d = \sum \delta_i$ is the number of events and η is some positive constant.

We stop if $\mathrm{EBIC}(S_{k+1}) > \mathrm{EBIC}(S_k)$ and declare $\hat{\mathcal{M}} = S_k$; otherwise, we proceed to the next step.

3. Theoretical Properties

We first introduce more notation. Let $\rightarrow_p$ denote convergence in probability. Given iid samples $Z_1, \ldots, Z_n$, let $\mathbb{E}_n\{f(Z_i)\} := n^{-1}\sum_{i=1}^n f(Z_i)$ and $\mathbb{G}_n\{f(Z_i)\} := n^{-1/2}\sum_{i=1}^n (f(Z_i) - \mathrm{E}\{f(Z_i)\})$. For a column vector $\mathbf{v}$, let $\mathbf{v}^{\otimes 0} = 1, \mathbf{v}^{\otimes 1} = \mathbf{v}$, and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^\top$. We denote the l_q-norm of $\mathbf{v}$ by $\|\mathbf{v}\|_q$ for $q \geq 1$, and, in particular, denote its l_2-norm by $\|\mathbf{v}\|$. For any symmetric matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ represent the smallest and largest eigenvalues. Given an index set S and an index $j \in S$, we use $S \setminus j$ to denote the set $\{r : r \in S, r \neq j\}$.

Given an index set $S \subset \{1, \ldots, p\}$, for $k = 0, 1, 2$, define

$$R_S^{(k)}(\boldsymbol{\beta}_S, t) = \mathbb{E}_n\left\{\bar{Y}_i(t)\mathbf{X}_{iS}^{\otimes k}\exp\left(\boldsymbol{\beta}_S^\top \mathbf{X}_{iS}\right)\right\}, \qquad r_S^{(k)}(\boldsymbol{\beta}_S, t) = \mathrm{E}\left\{R_S^{(k)}(\boldsymbol{\beta}_S, t)\right\},$$
$$V_S^{(k)}(t) = \mathbb{E}_n\left\{\bar{Y}_i(t)\mathbf{X}_{iS}^{\otimes k}\lambda_0(t)\exp(\boldsymbol{\beta}_0^\top \mathbf{X}_i)\right\}, \qquad v_S^{(k)}(t) = \mathrm{E}\left\{V_S^{(k)}(t)\right\}.$$

In addition, we use $\boldsymbol{\beta}_S^*$ to denote the least false value, which is the unique root of

$$\int_0^\tau \left\{v_S^{(1)}(t) - \frac{r_S^{(1)}(\boldsymbol{\beta}_S, t)}{r_S^{(0)}(\boldsymbol{\beta}_S, t)} v_S^{(0)}(t)\right\} dt = 0. \tag{4}$$

We use $F_T(t; \mathbf{X}_\mathcal{M})$, $f_T(t; \mathbf{X}_\mathcal{M})$ and $S_T(t; \mathbf{X}_\mathcal{M})$ to denote the conditional cumulative distribution function (cdf), probability density function (pdf), and survival function of T given the true model $\mathbf{X}_\mathcal{M}$, respectively. Likewise, the conditional cdf, pdf, and survival function of C given $\mathbf{X}_{\mathcal{M}_\mathcal{A}}$ are denoted by $F_C(t; \mathbf{X}_{\mathcal{M}_\mathcal{A}})$, $f_C(t; \mathbf{X}_{\mathcal{M}_\mathcal{A}})$, and $S_C(t; \mathbf{X}_{\mathcal{M}_\mathcal{A}})$, respectively, where $\mathbf{X}_{\mathcal{M}_\mathcal{A}}$ is the collection of covariates that are related to the censoring time C. Let $\mathcal{M}_O = \mathcal{M} \cup \mathcal{M}_\mathcal{A}$.

3.1. Regularity conditions

We posit the regularity conditions, followed by explanations. The assumptions are, in general, mild, well justified and follow the same lines as suggested by the existing literature.

(A) The study has a finite duration τ such that $\omega := \Pr(Y \geq \tau) > 0$.

(B) The X_j are time-independent and bounded by a constant $K > 1$ with $\mathrm{E}(X_j) = 0$ and $\mathrm{E}(X_j^2) = 1$ for all $1 \leq j \leq p$.

(C) There exist two positive constants $0 < \kappa_{\min} < \kappa_{\max} < \infty$, such that

$$\kappa_{\min} < \lambda_{\min}\left\{\mathrm{E}\left(\mathbf{X}_S^{\otimes 2}\right)\right\} \leq \lambda_{\max}\left\{\mathrm{E}\left(\mathbf{X}_S^{\otimes 2}\right)\right\} < \kappa_{\max},$$

uniformly in $S \subset \{1, \ldots, p\}$ satisfying $|S| \leq \rho$ for some $\rho > |\mathcal{M}|$.

(D) $\sup_{|S| \leq \rho} \|\boldsymbol{\beta}_S^*\|_1 \leq L$, for some constant L.

(E) For some constant $0 < \alpha < 1/2$,

$$\inf_{j \in \mathcal{M}} \left|\int_0^\tau \mathrm{E}\{X_j f_T(t; \mathbf{X}_\mathcal{M}) S_C(t; \mathbf{X}_{\mathcal{M}_\mathcal{A}})\} dt\right| \geq K n^{-\alpha}.$$

(F) There exists a $\zeta > 0$, such that for all $0 < t$,

$$\kappa_{\min} \leq \inf_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|_\infty \leq \zeta, |S| \leq \rho} \lambda_{\min}\left[\int_0^\tau \left\{\frac{r_S^{(2)}(\boldsymbol{\beta}_S, t)}{r_S^{(0)}(\boldsymbol{\beta}_S, t)} - \frac{(r_S^{(1)}(\boldsymbol{\beta}_S, t))^{\otimes 2}}{\left(r_S^{(0)}(\boldsymbol{\beta}_S, t)\right)^2}\right\} v_S^{(0)}(t) dt\right]$$
$$\leq \sup_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|_\infty \leq \zeta, |S| \leq \rho} \lambda_{\max}\left\{\int_0^\tau \frac{r_S^{(2)}(\boldsymbol{\beta}_S, t)}{r_S^{(0)}(\boldsymbol{\beta}_S, t)} v_S^{(0)}(t) dt\right\} \leq \kappa_{\max}.$$

(G) $\mathrm{E}\left\{X_j S_T(t; \mathbf{X}_\mathcal{M}) f_C(t; \mathbf{X}_{\mathcal{M}_\mathcal{A}}) | \mathbf{X}_{\mathcal{M}_O \setminus j}\right\}$ and $\mathrm{E}\left\{X_j S_T(t; \mathbf{X}_\mathcal{M}) S_C(t; \mathbf{X}_{\mathcal{M}_\mathcal{A}}) | \mathbf{X}_{\mathcal{M}_O \setminus j}\right\}$ have the same sign across t, for any $j \in \mathcal{M}$.

Condition (A) is standard in survival models with censored data [see, e.g., 23]. Conditions (B) and (C) are commonly assumed in the literature for variable selection and screening [see, e.g., 6, 7, 11, 24]. The boundedness of $\mathbf{X}$ is adopted to simplify our theoretical development and can be relaxed to the Cramer condition as in [5]. Condition (D) replaces the Lipschitz assumption in [25] and has a similar flavor to the conditions in [12], and [23]. Condition (E) is introduced in [12], which is an adapted version of the conditions in [26] to survival data. Condition (F) is analogous to Condition 2 considered in [5] for regularized Cox models. The condition essentially requires that the concavity of the log partial likelihood is well bounded in a neighborhood of $\boldsymbol{\beta}_S^*$. We invoke Condition (G) in order to analyze the least false value $\boldsymbol{\beta}_S^*$. The condition often holds in practice; for example, Lemma 1 shows that it is satisfied if $\mathcal{M} \cap \mathcal{M}_{\mathcal{A}} = \emptyset$.

Since the log partial likelihood function in (2) is the sum of non-iid random variables, we consider its asymptotically equivalent version, which can be expressed as the sum of iid terms:

$$\tilde{\ell}_S(\boldsymbol{\beta}_S) = \sum_{i=1}^{n} \int_0^\tau \left\{\boldsymbol{\beta}_S^\top \mathbf{X}_{iS} - \ln r_S^{(0)}(\boldsymbol{\beta}_S, t)\right\} dN_i(t). \tag{5}$$

According to [23], the log partial likelihood function (2) can then be viewed as a "working" model of (5), and the corresponding loss function becomes

$$\gamma_S(\boldsymbol{\beta}_S; \mathbf{X}_i, Y_i, \delta_i) := -\int_0^\tau \left\{\boldsymbol{\beta}_S^\top \mathbf{X}_{iS} - \ln r_S^{(0)}(\boldsymbol{\beta}_S, t)\right\} dN_i(t)$$

with the expected loss $\Gamma_S(\boldsymbol{\beta}_S) = \mathrm{E}\left[\gamma_S(\boldsymbol{\beta}_S; \mathbf{X}_i, Y_i, \delta_i)\right]$.

To validate the replacement of the log partial likelihood (5), a commonly assumed condition in the literature is that there exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ such that for $k = 0, 1, 2$,

$$\sup_{t \in [0,\tau], \boldsymbol{\beta} \in \mathcal{B}} \left\| R_{\mathcal{M}}^{(k)}(\boldsymbol{\beta}, t) - r_{\mathcal{M}}^{(k)}(\boldsymbol{\beta}, t) \right\| \to_p 0.$$

See, for example, [5, 12]. Under Conditions (B) and (D), Lemma 3 shows that the above condition holds uniformly for all $S \subset \{1, \ldots, p\}$ satisfying $|S| \leq \rho$.

We note that the proportional hazards assumption is made only on the true (and sparse) model. At each step of the screening procedure, we treat the misspecified Cox proportional hazards model as a working model following [27] and [28]. Our theoretical derivations depend on the least false value, which helps us characterize the asymptotic behavior of our estimator at each step even without the proportional hazards assumption. Specifically, similar to [27] and [28], the proposed estimator will converge to the least false value at each step under the working model, and when the second derivative of the log partial likelihood is bounded in a neighborhood of the least false value, adding an active variable will increase the partial-likelihood, even if a mis-specified model is under consideration.

3.2. Main results

Theorem 1. *Under Conditions (A) – (G), if* $\rho^4 \ln p/n \to 0$, *then with probability of at least* $1 - 8\exp(-3\rho \ln p)$,

$$\min_{S:|S|<\rho, \mathcal{M} \not\subset S} \max_{j \in S^c} \left\{\ell_{S \cup \{j\}}(\hat{\boldsymbol{\beta}}_{S \cup \{j\}}) - \ell_S(\hat{\boldsymbol{\beta}}_S)\right\} \geq c_1 n^{1-2\alpha} - c_2 \sqrt{n\rho^2 \ln p},$$

for some $0 < \alpha < 1/2$.

Theorem 1 shows that if $\mathcal{M} \not\subset S_k$ and $|S_k| < \rho$, then the increment of the log likelihood at the $(k+1)$th step is at least $c_1 n^{1-2\alpha} - c_2 \sqrt{n\rho^2 \ln p}$. Since the maximum increment is bounded by $|\ell_{S_0}(0)|$, we naturally obtain an upper bound on the number of steps for the forward selection, which is stated in the following corollary.

Corollary 1. *Suppose the same conditions as in Theorem 1 hold. If* $\sqrt{\rho^2 \ln p/n} = o(n^{-2\alpha})$ *and*

$$M := 2E\left\{\int_0^\tau \ln\left(n r_{S_0}^{(0)}(0, t)\right) dN_i(t)\right\} n^{2\alpha} < \rho$$

for some $0 < \alpha < 1/2$, *then* $\mathcal{M} \subset S_k$, *for some* $k \leq M$, *with probability of at least* $1 - 11\exp(-3\rho \ln p)$.

Corollary 1 establishes the screening consistency of the forward selection procedure. However, the upper bound M is not sharp, as it is calculated based on the lower bound of the increment of the log likelihood. The following corollary establishes an upper bound of the number of steps by evaluating how likely a signal variable will be selected at each step.

Corollary 2. *Under the same conditions as in Corollary 1, if the* $\mathbf{X}_{\mathcal{M}^c}$ *are independent of* $\mathbf{X}_{\mathcal{M}}$*, then* $\mathcal{M} \subset S_{c_3|\mathcal{M}|}$*, for some* $c_3 > 1$*, with probability of at least* $1 - 8\exp(-2\rho \ln p)$.

The condition that the $\mathbf{X}_{\mathcal{M}^c}$ are independent of $\mathbf{X}_{\mathcal{M}}$ stems from the assumption imposed in [12], and is similar to the partial orthogonality assumption introduced in [29]. It ensures that selecting a noise variable would bring less of an increment of the log likelihood compared to choosing a signal variable. Thus, it is much more likely for our procedure to select a signal variable at each step.

The following corollary follows from Corollary 2. We expect the proposed forward procedure to stop early with $\mathcal{M} \subset S_k$ for some k.

Corollary 3. *Under the same conditions as in Corollary 2, if* $\mathcal{M} \not\subset S_{k-1}$ *and* $\mathcal{M} \subset S_k$*, then with probability going to 1,*

(i) (screening consistency) the procedure stops at the kth step and $\mathcal{M} \subset \hat{\mathcal{M}} = S_k$*,*

(ii) (false discovery rate control) $|\hat{\mathcal{M}} \cap \mathcal{M}^c|/|\hat{\mathcal{M}}| \leq c_3 - 1$.

By Corollary 3, our proposed forward selection procedure will stop at a final step, denoted by $\hat{k}$, which is at most $c_3|\mathcal{M}|$. The final model $\hat{\mathcal{M}}$ not only achieves screening consistency, but also has well-controlled false discoveries.

4. Numerical Studies

Simulations were conducted to compare the performance of the proposed forward regression (FR) with two partial likelihood based screening methods, the principled sure independence screening (PSIS) by [12], and the conditional screening (CS) by [14]. The size of the models selected by the PSIS and CS was initially set to be $[n/\ln n]$ as suggested by [26]. To further reduce false positives, we applied Lasso [30], SCAD [31], and MCP [32] penalties to further reduce the sizes of models selected by each method. In the tables, we used screening method+penalty to denote the corresponding procedure. Although FR could start from a null model, we tried different initial sets for FR, including active or inactive variables. Particularly, we chose $\{X_1\}$ and $\{X_{10}\}$ to represent the active and inactive initial sets, respectively. When computing the model size for both FR and CS, we included the given initial set.

In this paper, we considered η as a fixed constant, which is analogous to the constant "a" parameter in the SCAD penalty function [31]. This distinguishes this from the other screening approaches that typically require data-driven thresholding tuning parameters and may incur more of a computational burden for finding them. To further justify the use of a fixed η, we considered various values of η between 0 and 1, the theoretical range of η in EBIC [20]. The BIC is a special case of EBIC when $\eta = 0$. We explored using BIC as the stopping criterion, but it incurred too many false positives compared to EBIC. This may cause overfitting of the Cox proportional hazards model with unreliably estimated regression coefficients and spuriously detected associations [33]. Thus, we elected not to use BIC.

We next considered three different values of η, 0.5, $1 - \ln d/(3\ln p)$, and 1; see Tables 1-3. Essentially, a larger η gives more penalty to a complicated model, which may incur more false negatives, while a smaller η less penalizes the complexity of models and may lead to more false positives. Based on Tables 1-3, it seems that under all of the scenarios considered, the choice of $\eta = 1 - \ln d/(3\ln p)$ strikes a good balance between false negatives and false positives.

We considered $p = 1,000$ and varied sample size $n = 200$ and 400. The survival time was generated from a Cox model $\lambda(t|\mathbf{X}) = \lambda_0(t)\exp(\boldsymbol{\beta}^\top \mathbf{X})$ with a Weibull baseline hazard. Specifically, $\lambda_0(t) = \alpha\gamma t^{\gamma-1}$, with $\alpha = 1$ and $\gamma = 1.5$. We considered various models for $\mathbf{X}$ and different parameter configurations for $\boldsymbol{\beta}$ in the following four examples. The censoring time was independently generated from a uniform distribution over $[0, c)$. We varied c for each example in order to yield mild (around 25%) and heavy censoring proportions (around 50%). For each configuration, a total of 500 simulated datasets were generated.

Example 1: We chose $\boldsymbol{\beta} = (1, 0.5, -1, 0, 1, \mathbf{0}_{p-5})$ and generated $\mathbf{X}$ from a multivariate normal distribution where the mean was 0, the variance 1, and $\mathrm{cor}(X_j, X_{j'}) = 0.5^{|j-j'|}$.
Example 2: We set $\boldsymbol{\beta} = (\mathbf{1}_5^\top, -2.5, \mathbf{0}_{p-6}^\top)$ and generated $\mathbf{X}$ from a multivariate normal distribution with mean 0, variance 1, and $\mathrm{cor}(X_j, X_{j'}) = 0.5$. In this case, since $\mathrm{cov}(T, X_6) = 0$, X_6 has a lower marginal utility than all of the noise variables with $\mathrm{cov}(T, X_j) = 1.25$ for $j \in \mathcal{M}^c$.
Example 3: We let $\boldsymbol{\beta} = (1, -1, 1, -1, 1, -\nu + \nu^2 - \nu^3 + \nu^4 - \nu^5, \mathbf{0}_{p-6}^\top)$ with $\nu = 0.5$. We generated $\mathbf{X}$ from a multivariate normal distribution with mean 0, variance 1, and $\mathrm{cor}(X_j, X_{j'}) = 0.5^{|j-j'|}$. In this case, since $\mathrm{cov}(T, X_6) = 0$, X_6 is an active but hidden variable. Furthermore, the signals of the active variables are weak due to signal cancellation.

[Table 1 about here.]

[Table 2 about here.]

Tables 1–2 report the average of the estimated probabilities of including the true models (PIT), the average numbers of true positive (TP) and false positive (FP), and their standard deviations in parenthesis, under mild and heavy censoring. We use p_0 to denote the number of true signals. We have observed the competing performance of the proposed method as detailed below.

First, Example 1 was designed in such a way that all of the true signals have nonzero marginal correlations with the outcome and are detectable by marginal screening methods. In particular, the dependence among the active variables in Example 1 can further strengthen the marginal correlations between them and the outcome. Even under these situations, FR was found to perform better than the marginal screening methods, including the conditional screening method, with larger PIT, TP and smaller FP. When the sample size decreases to 200 or with more censored events, FR's performance was still decent with, for example, a PIT around 0.8. In contrast, the performances of PSIS and CS deteriorated quickly with smaller sample sizes or with more censoring.

Second, Examples 2 and 3 were designed so that X_6, though active, has a 0 marginal correlation with the outcome and, therefore, is not detectable by marginal screening methods. As a result, PSIS and CS failed in this challenging situation. In contrast, FR remained competent by being able to detect the hidden variable. It even outperformed the conditional screening that used the prior information.

Third, even using penalties such as Lasso, SCAD, and MCP to further reduce false positives, the screening methods still incurred many false positives. The final models selected by the different penalties vary a bit. However, compared to FR, all of these penalties still perform similarly. On the other hand, with the EBIC-based stopping rule, the proposed FR caused much fewer false positives without the help of Lasso.

Finally, the simulation results hint that the performance of FR is quite robust toward the choice of the initial set. Even if the "wrong" set was employed as the initial set, the TP is almost the same as the one that starts from the null set.

We further explored the robustness of the method to the violation of the independence censoring assumption. We considered Examples 1^*–3^*, which have the same setup as Examples 1–3, except that the underlying survival times and censoring times have a common latent variable b which was generated from the standard normal distribution and the censoring times C_i were covariate-dependent. That is,

$$\lambda(t|\mathbf{X}) = \exp(b + \boldsymbol{\beta}^\top \mathbf{X}),$$

and

$$\lambda_C(t \mid \mathbf{X}) = c \exp(b + \boldsymbol{\alpha}^\top \mathbf{X}),$$

where $\boldsymbol{\alpha} = (0.75, 0.75, \mathbf{0}_{p-2})$ and c was chosen to censor approximately 25% of the observations. The results are documented in Table 3.

[Table 3 about here.]

We found that, with dependent censoring, the performance of all of the methods deteriorated a bit across the board. However, our proposed method still outperformed all the other methods, hinting at the usability of the proposal under dependent censoring. More work is warranted.

5. Analysis of the Boston Lung Cancer Survival Cohort (BLCSC) study

Recent studies demonstrate that aberrant methylation may be the most common mechanism of inactivating cancer-related genes in lung cancer. It occurs in the smoking-damaged bronchial epithelium from cancer-free individuals, can be reversed in vitro by demethylating agents, and may be a useful biomarker for lung cancer risk assessment [34]. It is thus of substantial interest to identify the methylations that play an important role in the pathogenesis of lung cancer, which affects patients' overall survival.

The motivating data represented a subset of the Boston Lung Cancer Survival Cohort (BLCSC) and included 124 samples, each with 442,613 methylations. The median follow up time of the subjects was 6.2 years and during the follow up, 84 deaths were observed. Each methylation resides within a certain gene. Prior literature has suggested that the following genes are associated with the development of lung cancer: ROS1, RET, PIK3CA, NRAS, BRAF, ALK, AKT 1, VGLL2, MET, KRAS, EGFR, KDM4, ST3GAL3, and CDH13. We used the array annotations from the Bioconductor package FDb.InfiniumMethylation.hg19 (version 2.2.0) to identify methylations that lie within these genes. A total of 589 methylations were identified. The other available environmental exposure and demographic variables in the data included lifetime tobacco exposure (SMOK), computed by multiplying the number of packs of cigarettes smoked per day by the number of years the person has smoked until the beginning of the study; AGE, the age at diagnosis in years; and SEX (1=male; 0=female).

Our analytical goal was to explore what methylations and their interactions with demographic and environmental exposure variables might be related to patients' overall survival. Thus, the outcome was the time to death, while the dependent variables included the aforementioned variables, consisting of demographic information, environmental exposures, and methylations and their interactions, for a total of 2,359 variables. We applied FR to the dataset and identified cg04187088×SEX and cg14363146×SMOK.

To check the model adequacy for the final model obtained by FR, we plotted the Cox-Snell residuals based on the final model that includes two predictors, cg04187088×SEX and cg14363146×SMOK. Figure 1 shows that the final model fits the data reasonably well.

[Figure 1 about here.]

Furthermore, we conducted the score test for the scaled Schoenfeld residuals to test the proportional hazards assumption on each included predictor in the final model. We obtained the p-values of 0.542 and 0.670 for cg04187088×SEX and cg14363146×SMOK, respectively. It appears that the proportional hazards assumption is not rejected for either of them.

We also applied competing methods to the BLCSC dataset, including the PSIS, CS, and SII. For each competing method, we selected the top $[n/\ln(n)]$=25 genes, and compared them with the genes selected by FR.

In terms of computing time, PSIS, CS, SII, and FR took 2.47, 2.52, 1473.65, and 20.89 seconds, respectively. Due to the sequential nature of the proposed procedure, FR is understandably more computationally intensive than the marginal screening approaches such as PSIS and CS. However, FR appears to run faster than SII, a nonparametric approach.

Table 4 lists the overlapping genes across the four methods. It appears that two genes selected by FR did not overlap with any genes selected by PSIS, CS and SII.

[Table 4 about here.]

On the other hand, using the SMOK, AGE, and SEX as the initial set, FR further selected cg11704212, in addition to these identified interactions. Our Pubmed review did not detect any previous literature that discusses these two SNPs. This may indicate the ability of the proposed FR to identify some novel SNPs, which may not have been detected by the existing methods. We expect that future studies are warranted to confirm and study the functionality of these detected biomarkers.

To elucidate the identified effects, we further conducted Kaplan-Meier analysis. We first dichotomized methylations by using the median values and used "+" or "−" if a subject's methylation is higher or lower than its median, respectively. Figure 2 compared survival curves across various subgroups. Figure 2(a) clearly indicated that female patients with low cg04187088 had a higher survival probability than the other comparison groups, while Figure 2(b) revealed that heavy smokers with high cg14363146 had a much lower survival probability than the other comparison groups. Figure 2(c) also showed that patients with high cg11704212 had a larger survival probability than those with low cg11704212.

[Figure 2 about here.]

6. Concluding Remarks

This article proposes forward regression with partial likelihood for high-dimensional survival data, and has obtained computationally and theoretically useful results. We envision the established theoretical framework will facilitate the extension of the procedure to other general settings, such as generalized linear models.

To further improve the computational efficiency, we can consider a natural extension of the proposed forward selection in the spirit of boosting [35]. That is, at each step, we use the selected variables and the obtained coefficient estimates to construct an offset, and search for a variable that will maximize the partial likelihood with such an offset. The advantage is that we need to maximize the partial likelihood with respect to only one covariate at each step, which may enhance computational efficiency. Specifically, we denote the log partial likelihood with an offset term O and covariate j in the model by

$$\ell_{O,j}(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[O_i + \beta X_{ij} - \ln \left\{ \sum_{l=1}^{n} \bar{Y}_l(t) \exp(O_l + \beta_j X_{lj}) \right\} \right] dN_i(t),$$

Here O_i refers to the offset term O evaluated at the ith subject. We let $O^{(k)}$ be the offset term evaluated at kth step and S_k be the set of indices of covariates selected up to the kth step. For FR, we initialize $S_0 = \emptyset$ and set $O^{(0)} = 0$. For $j \in \{1, \ldots, p\}$, compute $\hat{\beta}_j^{(1)} = \arg\max_\beta \ell_{O^{(0)},j}(\beta)$. Then $j_1 = \arg\max_{j\in\{1,\ldots,p\}} \ell_{O^{(0)},j}(\hat{\beta}_j^{(1)})$. Now set $O^{(1)} = \hat{\beta}_{j_1}^{(1)} X_{j_1}$ and $S_1 = \{j_1\}$. Given $O^{(k)}$ and S_k, compute $\hat{\beta}_j^{(k+1)} = \arg\max_\beta \ell_{O^{(k)},j}(\beta)$ for $j \in S_k^c$. Then $j_{k+1} = \arg\max_{j\in S_k^c} \ell_{O^{(k)},j}(\hat{\beta}_j^{(k+1)})$. Now set $O^{(k+1)} = O^{(k)} + \hat{\beta}_{j_{k+1}}^{(k+1)} X_{j_{k+1}}$ and $S_{k+1} = S_k \cup \{j_{k+1}\}$. We note this proposal does not require re-estimation of the coefficients of the covariates selected in the previous steps, which expedites computation. We will explore this further.

We employed a modified EBIC to select the final models. Although it worked well in our simulations, it tends to be conservative in real data analysis and recruits too few variables, whereas BIC recruits too many variables. It would be interesting to investigate the optimal η in the EBIC penalty term to strike a balance between EBIC and BIC.

Appendix

Preliminary lemmas

We present some preliminary lemmas in this section. Given an index set $S \subset \{1, \ldots, p\}$, we use S_{+r} to denote $S \cup \{r\}$ for some $r \in S^c$.

Lemma 1. *Condition (G) is satisfied if* $\mathcal{M} \cap \mathcal{M}_{\mathcal{A}} = \emptyset$.

Without loss of generality, we assume that X_r is the last element of $\mathbf{X}_{S_{+r}}$, with β_r^* being the corresponding least false coefficient under the model S_{+r}.

Lemma 2. *Given* $S \subset \{1, \ldots, p\}$ *satisfying* $|S| < \rho$ *and* $r \in S^c$, *under Conditions (B) and (G),*

(i) if $r \in \mathcal{M}^c$ *and the* $\mathbf{X}_{\mathcal{M}^c}$ *are independent of* $\mathbf{X}_{\mathcal{M}}$, *then* $\boldsymbol{\beta}_{S_{+r}}^* = (\boldsymbol{\beta}_S^{*\top}, 0)^\top$, *i.e. the least false coefficient for* X_r *under the model* S_{+r} *is 0.*

(ii) if $r \in \mathcal{M}$, *then* $\boldsymbol{\beta}_{S_{+r}}^* \neq (\boldsymbol{\beta}_S^{*\top}, 0)^\top$.

(iii) if $r \in \mathcal{M}$ *and Conditions (E) and (F) are satisfied,* $\left\|\boldsymbol{\beta}_{S_{+r}}^* - (\boldsymbol{\beta}_S^{*\top}, 0)^\top\right\| \geq K\kappa_{\max}^{-1} n^{-\alpha}$.

Lemma 2 quantifies the difference between $\boldsymbol{\beta}_{S_{+r}}^*$ and $(\boldsymbol{\beta}_S^{*\top}, 0)^\top$ when a noise variable or a signal variable is selected into the model.

Lemma 3. *Under Conditions (B) and (D), if $\rho^3 \ln p/n \to 0$, then for each $S \subset \{1, \dots, p\}$ satisfying $|S| \le \rho$, we can find a neighborhood $\mathcal{B}_S^0(c)$ of $\boldsymbol{\beta}_S^*$, for some constant c, such that*

$$\sup_{|S|\le\rho, t\in[0,\tau], \boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)} \left\|R_S^{(k)}(\boldsymbol{\beta}_S, t) - r_S^{(k)}(\boldsymbol{\beta}_S, t)\right\| \to_p 0, \qquad k = 0, 1, 2.$$

Define

$$Z_S(\boldsymbol{\beta}_S) := \left|[\mathbb{E}_n\{\gamma_S(\boldsymbol{\beta}_S; \mathbf{X}_i, Y_i, \delta_i)\} - \Gamma_S(\boldsymbol{\beta}_S)] - [\mathbb{E}_n\{\gamma_S(\boldsymbol{\beta}_S^*; \mathbf{X}_i, Y_i, \delta_i)\} - \Gamma_S(\boldsymbol{\beta}_S^*)]\right|.$$

Lemma 4. *Under the same conditions as in Lemma 3,*

$$\Pr\left\{\sup_{|S|\le\rho, \boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)} Z_S(\boldsymbol{\beta}_S) \ge 2\frac{c}{K}a_n + 3\frac{c\sqrt{\rho}}{K\sqrt{s}}a_n + 6c\tau\sqrt{\rho\ln p/n} + 2K\sqrt{\rho}\zeta_n\right\}$$
$$\le 3\exp(-6\rho\ln p),$$

where $a_n = \sqrt{2K^2\ln(2p)/n} + K\ln(2p)/n$ and $\zeta_n = c/(Kn^2)$.

Define $D_S(\boldsymbol{\beta}_S) := n^{-1}\left|\left\{\ell_S(\boldsymbol{\beta}_S) - \tilde{\ell}_S(\boldsymbol{\beta}_S)\right\} - \left\{\ell_S(\boldsymbol{\beta}_S^*) - \tilde{\ell}_S\left(\boldsymbol{\beta}_S^*\right)\right\}\right|$.

Lemma 5. *Under Conditions (A), (B), and (D), we have*

$$\Pr\left\{\sup_{|S|\le\rho, \boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)} D_S(\boldsymbol{\beta}_S) \ge A_{10}\sqrt{\rho^2\ln p/n}\right\} \le 2\exp(-3\rho\ln p),$$

for some constant A_{10} that does not depend on n.

Lemma 6. *Under Conditions (B), (D), and (F), given an index set S satisfying $|S| \le \rho$, then for any $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(c)$,*

$$\frac{1}{2}\kappa_{\min}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2 \le \Gamma_S(\boldsymbol{\beta}_S) - \Gamma_S(\boldsymbol{\beta}_S^*) \le \frac{1}{2}\kappa_{\max}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2.$$

Lemma 7. *Under Conditions (A) – (F), if $\rho^4\ln p/n \to 0$, there exist two constants A_{11} and A_{12} that do not depend on n such that*

(i)

$$\Pr\left\{\sup_{|S|\le\rho}\left\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\right\| \le A_{11}\left(\rho^2\ln p/n\right)^{1/4}\right\} \ge 1 - 5\exp(-3\rho\ln p) \;\; \textit{and}$$

(ii)

$$\Pr\left\{\sup_{|S|\le\rho} n^{-1}\left|\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_S(\boldsymbol{\beta}_S^*)\right| \le A_{12}\sqrt{\rho^2\ln p/n}\right\} \ge 1 - 5\exp(-3\rho\ln p).$$

Lemma 8. *Under Conditions (B) and (D), there exists some constant A_{14}, which does not depend on n, such that*

$$\Pr\left[\sup_{|S|<\rho, r\in S^c}\left|n^{-1}\left\{\ell_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) - \ell_S(\boldsymbol{\beta}_S^*)\right\} - \left\{-\Gamma_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) + \Gamma_S(\boldsymbol{\beta}_S^*)\right\}\right|\right.$$
$$\left.\ge A_{14}\sqrt{\rho\ln p/n}\right] \le 3\exp\left(-3\rho\ln p\right).$$

Proofs of main theoretical results in Section 3.2

In the following, we provide the proofs for the theoretical results in Section 3.2.

Proof of Theorem 1: We prove the theorem for a generic index set S satisfying $S \subset \{1, \ldots, p\}$, $\mathcal{M} \not\subset S$ and $|S| < \rho$.

The change of log likelihood by adding a variable X_r, where $r \in S^c$, can be decomposed as

$$\begin{aligned}
&n^{-1}\left\{\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_S(\hat{\boldsymbol{\beta}}_S)\right\} \\
&= n^{-1}\left[\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_S(\hat{\boldsymbol{\beta}}_S) - \left\{\ell_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) - \ell_S(\boldsymbol{\beta}^*_S)\right\}\right] \\
&\quad + \left[n^{-1}\left\{\ell_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) - \ell_S(\boldsymbol{\beta}^*_S)\right\} - \left\{-\Gamma_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) + \Gamma_S(\boldsymbol{\beta}^*_S)\right\}\right] - \left\{\Gamma_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) - \Gamma_S(\boldsymbol{\beta}^*_S)\right\}.
\end{aligned}$$

We restrict our attention $\Omega_5^c \cap \Omega_6^c$, where

$$\begin{aligned}
\Omega_5 &:= \left\{\sup_{|S|\le\rho} n^{-1}\left|\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_S(\boldsymbol{\beta}^*_S)\right| > A_{12}\sqrt{\rho^2 \ln p/n}\right\} \text{ and} \\
\Omega_6 &:= \left\{\sup_{|S|<\rho, r\in S^c} \left|n^{-1}\left\{\ell_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) - \ell_S(\boldsymbol{\beta}^*_S)\right\} - \left\{-\Gamma_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) + \Gamma_S(\boldsymbol{\beta}^*_S)\right\}\right|\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left.\ge A_{14}\sqrt{\rho \ln p/n}\right\}.
\end{aligned}$$

According to Lemmas 7 and 8, $\Omega_5^c \cap \Omega_6^c$ holds with probability of at least $1 - 8\exp(-3\rho \ln p)$.

If $r \in \mathcal{M}$, then by Lemma 2 (*iii*), $\left\|\boldsymbol{\beta}^*_{S_{+r}} - (\boldsymbol{\beta}^{*\top}_S, 0)^\top\right\| \ge K\kappa_{\max}^{-1} n^{-\alpha}$. For any $\boldsymbol{\beta}_{S_{+r}}$ such that $\|\boldsymbol{\beta}_{S_{+r}} - \boldsymbol{\beta}^*_{S_{+r}}\| = K\kappa_{\max}^{-1} n^{-\alpha}$, noting that $\boldsymbol{\beta}^*_{S_{+r}}$ is the solution to (4) under model S_{+r}, we apply Taylor's expansion to obtain

$$\begin{aligned}
\Gamma_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}) - \Gamma_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) &= \frac{1}{2}\left(\boldsymbol{\beta}^*_{S_{+r}} - \boldsymbol{\beta}_{S_{+r}}\right)^\top \\
&\times \left[\int_0^\tau \frac{r^{(2)}_{S_{+r}}(\tilde{\boldsymbol{\beta}}_{S_{+r}}, u)}{r^{(0)}_{S_{+r}}(\tilde{\boldsymbol{\beta}}_{S_{+r}}, u)} - \frac{\left\{r^{(1)}_{S_{+r}}(\tilde{\boldsymbol{\beta}}_{S_{+r}}, u)\right\}^{\otimes 2}}{\left\{s^{(0)}(\tilde{\boldsymbol{\beta}}_{S_{+r}}, u)\right\}^2} v^{(0)}(u)du\right]\left(\boldsymbol{\beta}^*_{S_{+r}} - \boldsymbol{\beta}_{S_{+r}}\right) \\
&\ge K^2\kappa_{\min}\kappa_{\max}^{-2} n^{-2\alpha} =: c_1 n^{-2\alpha},
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{S_{+r}}$ is between $\boldsymbol{\beta}_{S_{+r}}$ and $\boldsymbol{\beta}^*_{S_{+r}}$, the last inequality follows from Condition (F) and $c_1 := K^2\kappa_{\min}\kappa_{\max}^{-2}$. By the convexity of $\Gamma_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}})$, we have $\Gamma_{S_{+r}}\{(\boldsymbol{\beta}^{*\top}_S, 0)^\top\} - \Gamma_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) \ge c_1 n^{-2\alpha}$. Thus,

$$\begin{aligned}
&n^{-1}\left\{\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_S(\hat{\boldsymbol{\beta}}_S)\right\} \\
&\ge -\left\{n^{-1}\left|\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}})\right| + n^{-1}\left|\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_S(\boldsymbol{\beta}^*_S)\right|\right\} \\
&\quad - \left|n^{-1}\left\{\ell_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) - \ell_S(\boldsymbol{\beta}^*_S)\right\} - \left\{-\Gamma_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}}) + \Gamma_S(\boldsymbol{\beta}^*_S)\right\}\right| + \left\{\Gamma_S(\boldsymbol{\beta}^*_S) - \Gamma_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}})\right\} \\
&\ge -2A_{12}\sqrt{\rho^2 \ln p/n} - A_{14}\sqrt{\rho \ln p/n} + c_1 n^{-2\alpha}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\sup_{|S|<\rho, r\in S^c} n^{-1}\left\{\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_S(\hat{\boldsymbol{\beta}}_S)\right\} \\
&\ge c_1 n^{-2\alpha} - 2A_{12}\sqrt{\rho^2 \ln p/n} - A_{14}\sqrt{\rho \ln p/n} \ge c_1 n^{-2\alpha} - c_2\sqrt{\rho^2 \ln p/n},
\end{aligned}$$

for some constant c_2 that does not depend on n. Then, we obtain that

$$\max_{r\in S^c}\left\{\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_S(\hat{\boldsymbol{\beta}}_S)\right\} \ge c_1 n^{1-2\alpha} - c_2\sqrt{n\rho^2 \ln p}.$$

Withdrawing the restriction on $\Omega_5^c \cap \Omega_6^c$ completes the proof of Theorem 1. □

Proof of Corollary 1: As shown in Theorem 1, for any S such that $|S| < \rho, \mathcal{M} \not\subset S$, with probability of at least $1 - 8\exp(-3\rho \ln p)$,

$$\max_{r \in S^c} \ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \geq c_1 n^{1-2\alpha} - c_2\sqrt{n\rho^2 \ln p}.$$

If $\sqrt{n\rho^2 \ln p} = o(n^{1-2\alpha})$, then $\ln d + 2\eta \ln p = o(n^{1-2\alpha})$, and consequently,

$$\begin{aligned}
&\mathrm{EBIC}(S_{+r}) - \mathrm{EBIC}(S) \\
&= -2\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) + (|S|+1)(\ln d + 2\eta \ln p) - \left\{-2\ell_S(\hat{\boldsymbol{\beta}}_S) + |S|(\ln d + 2\eta \ln p)\right\} \\
&\leq c_2\sqrt{n\rho^2 \ln p} - c_1 n^{1-2\alpha} + (\ln d + 2\eta \ln p) < 0.
\end{aligned}$$

Therefore, our forward selection does not stop when $\mathcal{M} \not\subset S$ and $|S| < \rho$ with probability of at least $1 - 8\exp(-3\rho \ln p)$.

Noting that $S_0 = \emptyset$, if $\mathcal{M} \not\subset S_k$, then

$$\begin{aligned}
&\sum_{i=1}^{n}\left[\int_0^\tau \ln\left\{nR_{S_0}^{(0)}(0,t)\right\} dN_i(t)\right] - 0 \\
&\geq \{\ell_{S_1}(\hat{\boldsymbol{\beta}}_{S_1}) - \ell_{S_0}(\hat{\boldsymbol{\beta}}_{S_0})\} + \{\ell_{S_2}(\hat{\boldsymbol{\beta}}_{S_2}) - \ell_{S_1}(\hat{\boldsymbol{\beta}}_{S_1})\} + \cdots + \{\ell_{S_k}(\hat{\boldsymbol{\beta}}_{S_k}) - \ell_{S_{k-1}}(\hat{\boldsymbol{\beta}}_{S_{k-1}})\} \\
&\geq k c_1 n^{1-2\alpha}/2,
\end{aligned}$$

when n is sufficiently large such that $2c_2\sqrt{\rho^2 \ln p/n} \leq c_1 n^{-2\alpha}/2$.

As shown in the proof of Lemma 8,

$$\mathbb{E}_n\left\{\int_0^\tau \ln R_{S_0}^{(0)}(0,t) dN_i(t)\right\} \to \mathrm{E}\left\{\int_0^\tau \ln r_{S_0}^{(0)}(0,t) dN_i(t)\right\},$$

with probability of at least $1 - 3\exp(-3\rho \ln p)$. If $\mathcal{M} \not\subset S_M$, then

$$\frac{\mathrm{E}\left[\int_0^\tau \ln\left\{n r_{S_0}^{(0)}(0,t)\right\} dN_i(t)\right]}{c_1 n^{-2\alpha}/2} > M,$$

which contradicts the definition of M. Hence, we have some k such that $\mathcal{M} \subset S_k$. This completes the proof of Corollary 1. □

Proof of Corollary 2: By Lemma 2 (i), If $r \in \mathcal{M}^c$ and $\mathbf{X}_{\mathcal{M}^c}$ are independent of $\mathbf{X}_{\mathcal{M}}$, then $\boldsymbol{\beta}^*_{S_{+r}} = (\boldsymbol{\beta}_S^{*\top}, 0)^\top$. Thus, under $\Omega_5^c \cap \Omega_6^c$,

$$n^{-1}\left\{\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_S(\hat{\boldsymbol{\beta}}_S)\right\} \leq n^{-1}\left|\ell_{S_{+r}}(\hat{\boldsymbol{\beta}}_{S_{+r}}) - \ell_{S_{+r}}(\boldsymbol{\beta}^*_{S_{+r}})\right| + n^{-1}\left|\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_S(\boldsymbol{\beta}^*_S)\right| \leq 2A_{12}\sqrt{\rho^2 \ln p/n}.$$

If $\sqrt{n\rho^2 \ln p} = o(n^{1-2\alpha})$, we have for any S such that $|S| < \rho$ and $\mathcal{M} \not\subset S$,

$$\arg\min_{r \in S^c} \mathrm{EBIC}(S_{+r}) - \mathrm{EBIC}(S) \in \mathcal{M},$$

when n is sufficiently large.

Withdrawing the restriction on $\Omega_5^c \cap \Omega_6^c$, we obtain that, at each step, the probability of selecting a noise variable is at most $8\exp(-3\rho \ln p)$.

Since $\mathcal{M} \not\subset S_k$ implies that at more than $k - |\mathcal{M}|$ steps, a noise variable is selected, then for $k = c_3|\mathcal{M}|$,

$$\begin{aligned}
P(\mathcal{M} \not\subset S_k) &\leq \sum_{j=k-|\mathcal{M}|}^{k}\binom{k}{j}\left\{8\exp(-3\rho \ln p)\right\}^j \leq |\mathcal{M}|k^{|\mathcal{M}|}\left\{8\exp(-3\rho \ln p)\right\}^{k-|\mathcal{M}|} \\
&\leq 8\exp(-3\rho \ln p + \ln|\mathcal{M}| + |\mathcal{M}|\ln k) \leq 8\exp(-2\rho \ln p).
\end{aligned}$$

Therefore, $\mathcal{M} \subset S_{c_3|\mathcal{M}|}$ with probability of at least $1 - 8\exp(-2\rho \ln p)$. This completes the proof of Corollary 2. □

Proof of Corollary 3: By Corollary 2, we know that $\mathcal{M} \subset S_k$, for some $k \leq c_3|\mathcal{M}|$. Thus, both $S_k \subset S_{k+1} \in \mathscr{A}_0$, where

$$\mathscr{A}_0 := \left\{S : \mathcal{M} \subset S, |S| \leq c_3|\mathcal{M}|\right\}.$$

(i): It can be shown that $\text{EBIC}(S_{k+1}) < \text{EBIC}(S_k)$ if and only if $2\ell_{S_{k+1}}(\hat{\boldsymbol{\beta}}_{S_{k+1}}) - 2\ell_S(\hat{\boldsymbol{\beta}}_{S_k}) \geq \ln d + 2\eta \ln p$. Following the same arguments used to show Equation (14) in [36], we can show that with probability tending to 1,

$$2\ell_{S_{k+1}}(\hat{\boldsymbol{\beta}}_{S_{k+1}}) - 2\ell_S(\hat{\boldsymbol{\beta}}_{S_k}) < \ln d + 2\eta \ln p,$$

for all $\eta > 0$. Therefore, with probability tending to 1, the procedure stops at the kth step and $\mathcal{M} \subset \hat{\mathcal{M}} = S_k$.
(ii): From Corollary 2 and Part (i), we have $|\hat{\mathcal{M}} \cap \mathcal{M}^c| = |\hat{\mathcal{M}}| - |\mathcal{M}|$ as well as $|\hat{\mathcal{M}}| \geq |\mathcal{M}|$ with probability going to 1. Hence, the stated result follows. □

Proofs of Lemmas

Proof of Lemma 1: We first show that $r \in \mathcal{M}$, $\mathrm{E}\left[X_r S_T(t; \mathbf{X}_{\mathcal{M}}) f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_O \backslash r}\right]$ has the same sign across $t > 0$. Let $S_0(t) = \exp\left(-\int_0^t \lambda_0(u) du\right)$.

$$\begin{aligned}
&\mathrm{E}\left\{X_r S_T(t; \mathbf{X}_{\mathcal{M}}) f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_O \backslash r}\right\} = \mathrm{E}\left\{X_r S_T(t; \mathbf{X}_{\mathcal{M}}) | \mathbf{X}_{\mathcal{M} \backslash r}\right\} \mathrm{E}\left\{f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}\right\} \\
&= \mathrm{E}\left\{X_r S_0(t)^{\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)} \middle| \mathbf{X}_{\mathcal{M} \backslash r}\right\} \mathrm{E}\left\{f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}\right\} \\
&= \mathrm{E}\left\{X_r S_0(t)^{\exp\left(\boldsymbol{\beta}_{0\mathcal{M} \backslash r}^\top \mathbf{X}_{\mathcal{M} \backslash r}\right) \exp(\beta_{0r} X_r)} \middle| \mathbf{X}_{\mathcal{M} \backslash r}\right\} \mathrm{E}\left\{f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}\right\} \\
&= \frac{1}{\beta_{0r}} \mathrm{E}\left[\beta_r X_r \left\{S_0(t)^{\exp\left(\boldsymbol{\beta}_{0\mathcal{M} \backslash r}^\top \mathbf{X}_{\mathcal{M} \backslash r}\right)}\right\}^{\exp(\beta_{0r} X_r)} \middle| \mathbf{X}_{\mathcal{M} \backslash r}\right] \mathrm{E}\left\{f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}\right\} \\
&= -\frac{1}{\beta_{0r}} \mathrm{E}\left[-U \left\{S_0(t)^{\exp\left(\boldsymbol{\beta}_{0\mathcal{M} \backslash r}^\top \mathbf{X}_{\mathcal{M} \backslash r}\right)}\right\}^{\exp(U)} \middle| \mathbf{X}_{\mathcal{M} \backslash r}\right] \mathrm{E}\left\{f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}\right\},
\end{aligned}$$

where $U = \beta_{0r} X_r$. Noting that $S_0(t) \leq 1$ and $\exp\left(\boldsymbol{\beta}_{0\mathcal{M} \backslash r}^\top \mathbf{X}_{\mathcal{M} \backslash r}\right) > 0$, we have $S_0(t)^{\exp\left(\boldsymbol{\beta}_{0\mathcal{M} \backslash r}^\top \mathbf{X}_{\mathcal{M} \backslash r}\right)} \leq 1$. Therefore, given $\mathbf{X}_{\mathcal{M}/r}$,

$$-\left\{S_0(t)^{\exp\left(\boldsymbol{\beta}_{0\mathcal{M} \backslash r}^\top \mathbf{X}_{\mathcal{M} \backslash r}\right)}\right\}^{\exp(U)} \text{ is monotone increasing with respect to } U.$$

Then by [37],

$$\mathrm{E}\left[-U \left\{S_0(t)^{\exp\left(\boldsymbol{\beta}_{0\mathcal{M} \backslash r}^\top \mathbf{X}_{\mathcal{M} \backslash r}\right)}\right\}^{\exp(U)} \middle| \mathbf{X}_{\mathcal{M} \backslash r}\right] \geq 0,$$

for all $t > 0$, and hence

$$\mathrm{E}\left\{X_r S_T(t; \mathbf{X}_{\mathcal{M}})\right\} \mathrm{E}\left\{f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}\right\} = -\frac{1}{\beta_{0r}} \mathrm{E}\left[-U \left\{S_0(t)^{\exp\left(\boldsymbol{\beta}_{\mathcal{M} \backslash r}^\top \mathbf{X}_{\mathcal{M} \backslash r}\right)}\right\}^{\exp(U)} \middle| \mathbf{X}_{\mathcal{M} \backslash r}\right] \mathrm{E}\left\{f_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}\right\}$$

has the same sign as $-1/\beta_{0r}$, for all $t > 0$.

Next, by the same argument used above, $\mathrm{E}\left\{X_r S_T(t; \mathbf{X}_{\mathcal{M}}) S_C(t; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}}) | \mathbf{X}_{\mathcal{M}_O \backslash r}\right\}$ has the same sign as $-1/\beta_{0r}$ across $t > 0$. This completes the proof of Lemma 1. □

Proof of Lemma 2: We first note that $\boldsymbol{\beta}_S^*$ is the root of (4). Let $\mathbf{U}_S(\boldsymbol{\beta}_S) = \int_0^\tau \left\{v_S^{(1)}(t) - r_S^{(1)}(\boldsymbol{\beta}_S, t)v_S^{(0)}(t)/r_S^{(0)}(\boldsymbol{\beta}_S, t)\right\} dt$.

$$\begin{aligned}\mathbf{U}_S(\boldsymbol{\beta}_S^*) = \int_0^\tau & \mathrm{E}\left\{1\,(Y \geq u)\,\mathbf{X}_S\,\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} \\ & - \frac{\mathrm{E}\left\{1(Y \geq u)\mathbf{X}_S \exp\left(\boldsymbol{\beta}_S^\top \mathbf{X}_S\right)\right\}}{\mathrm{E}\left\{1(Y \geq u)\exp\left(\boldsymbol{\beta}_S^\top \mathbf{X}_S\right)\right\}}\mathrm{E}\left\{1(Y \geq u)\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} du = 0. \end{aligned} \tag{6}$$

Part (i): Let $S_{\mathcal{M}} = S \cap \mathcal{M}$ and $S_{\mathcal{M}^c} = S \cap \mathcal{M}^c$. By the condition that the $\mathbf{X}_{\mathcal{M}^c}$ is independent of $\mathbf{X}_{\mathcal{M}}$, the $\mathbf{X}_{S_{\mathcal{M}^c}}$ is independent of $\mathbf{X}_{S_{\mathcal{M}}}$. Thus, by Condition (B),

$$\begin{aligned} & \int_0^\tau \mathrm{E}\left\{1(Y \geq u)\mathbf{X}_{S_{\mathcal{M}^c}}\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} \\ & \qquad - \frac{\mathrm{E}\left\{1(Y \geq u)\mathbf{X}_{S_{\mathcal{M}^c}} \exp\left(\boldsymbol{\beta}_{S_{\mathcal{M}}}^{*\top} \mathbf{X}_{S_{\mathcal{M}}}\right)\right\}}{\mathrm{E}\left\{1\,(Y \geq u)\exp\left(\boldsymbol{\beta}_{S_{\mathcal{M}}}^{*\top} \mathbf{X}_{S_{\mathcal{M}}}\right)\right\}}\mathrm{E}\left\{1\,(Y \geq u)\,\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^{*T} \mathbf{X}_{\mathcal{M}}\right)\right\} du \\ = & \int_0^\tau \mathrm{E}\left(\mathbf{X}_{S_{\mathcal{M}^c}}\right)\mathrm{E}\left\{1\,(Y \geq u)\,\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} \\ & \qquad - \frac{\mathrm{E}\left(\mathbf{X}_{S_{\mathcal{M}^c}}\right)\mathrm{E}\left\{1\,(Y \geq u)\exp\left(\boldsymbol{\beta}_{S_{\mathcal{M}}}^{*\top} \mathbf{X}_{S_{\mathcal{M}}}\right)\right\}}{\mathrm{E}\left\{1\,(Y \geq u)\exp\left(\boldsymbol{\beta}_{S_{\mathcal{M}}}^{*\top} \mathbf{X}_{S_{\mathcal{M}}}\right)\right\}}\mathrm{E}\left\{1\,(Y \geq u)\,\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} du = 0, \end{aligned}$$

where the last equality follows from Condition (B). Combining the result and (6), it can be shown that

$$\begin{aligned} & \int_0^\tau \mathrm{E}\left\{1\,(Y \geq u)\,\mathbf{X}_S\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} \\ & \qquad - \frac{\mathrm{E}\left\{1\,(Y \geq u)\,\mathbf{X}_S \exp\left(\boldsymbol{\beta}_{S_{\mathcal{M}}}^{*\top} \mathbf{X}_{S_{\mathcal{M}}}\right)\right\}}{\mathrm{E}\left\{1\,(Y \geq u)\exp\left(\boldsymbol{\beta}_{S_{\mathcal{M}}}^{*\top} \mathbf{X}_{S_{\mathcal{M}}}\right)\right\}}\mathrm{E}\left\{1(Y \geq u)\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} du = 0. \end{aligned}$$

Therefore, $\boldsymbol{\beta}_S^* = (\boldsymbol{\beta}_{S_{\mathcal{M}}}^{*\top}, \mathbf{0}^\top)^\top$.

If $r \in \mathcal{M}^c$, then by the same arguments, we can show that $\boldsymbol{\beta}_{S_{+r}}^* = (\boldsymbol{\beta}_S^{*\top}, 0)^\top$.

Part (ii): Suppose $\boldsymbol{\beta}_{S_{+r}}^* = (\boldsymbol{\beta}_S^{*\top}, 0)^\top$. By the martingale property, $\boldsymbol{\beta}_{S_{+r}}^*$ is also the solution to the following equation,

$$\begin{aligned} & \int_0^\infty \mathrm{E}\left\{1(Y \geq u)X_r\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} du \\ & \qquad = \int_0^\tau \frac{\mathrm{E}\left\{1(Y \geq u)X_r \exp\left(\boldsymbol{\beta}_S^{*\top} \mathbf{X}_S\right)\right\}}{\mathrm{E}\left\{1(Y \geq u)\exp\left(\boldsymbol{\beta}_S^{*\top} \mathbf{X}_S\right)\right\}}\mathrm{E}\left\{1(Y \geq u)\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} du. \end{aligned}$$

On the one hand, it is straightforward to see that

$$\begin{aligned} & \int_0^\infty \mathrm{E}\left\{1(Y \geq u)X_r\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right)\right\} du \\ = & \int_0^\infty \mathrm{E}\left\{X_r S_T(u; \mathbf{X}_{\mathcal{M}})\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top \mathbf{X}_{\mathcal{M}}\right) S_C(u; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}})\right\} du \\ = & \int_0^\infty \mathrm{E}\left\{X_r f_T(u; \mathbf{X}_{\mathcal{M}})S_C(u; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}})\right\} du = \mathrm{E}\left\{X_r \int_0^\infty F_T(u; \mathbf{X}_{\mathcal{M}})f_C(u; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}})du\right\} \\ = & -\int_0^\infty \mathrm{E}\left\{X_r S_T(u; \mathbf{X}_{\mathcal{M}})f_C(u; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}})\right\} du = -\int_0^\infty \mathrm{E}\left[\mathrm{E}\left\{X_r S_T(u; \mathbf{X}_{\mathcal{M}})f_C(u; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}})|\mathbf{X}_{\mathcal{M}_O\backslash r}\right\}\right] du, \end{aligned}$$

which has the opposite sign as $\mathrm{E}\{X_r S_T(u;\mathbf{X}_{\mathcal{M}}) f_C(u;\mathbf{X}_{\mathcal{M}_{\mathcal{A}}})|\mathbf{X}_{\mathcal{M}_O\backslash r}\}$. On the other hand,

$$\begin{aligned}&\mathrm{E}\left[1(Y\geq u)X_r\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)\right]=\mathrm{E}\left[X_r S_T(u;\mathbf{X}_{\mathcal{M}})S_C(u;\mathbf{X}_{\mathcal{M}_{\mathcal{A}}})\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)\right]\\ &=\mathrm{E}\left[\mathrm{E}\left\{X_r S_T(u;\mathbf{X}_{\mathcal{M}})S_C(u;\mathbf{X}_{\mathcal{M}_{\mathcal{A}}})\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)|\mathbf{X}_{\mathcal{M}_O\backslash r}\right\}\right]\\ &=\mathrm{E}\Big[\mathrm{E}\{X_r S_T(u;\mathbf{X}_{\mathcal{M}})S_C(u;\mathbf{X}_{\mathcal{M}_{\mathcal{A}}})|\mathbf{X}_{\mathcal{M}_O\backslash r}\}\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)\Big],\end{aligned}$$

which has the same sign as $\mathrm{E}\,[X_r S_T(u;\mathbf{X}_{\mathcal{M}}) f_C(u;\mathbf{X}_{\mathcal{M}_{\mathcal{A}}})|\mathbf{X}_{\mathcal{M}_O\backslash r}]$ by Condition (G) and the fact that $\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)>0$.

Thus, $\int_0^\tau \mathrm{E}\left\{1(Y\geq u)X_r\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top\mathbf{X}_{\mathcal{M}}\right)\right\}du$ and

$$\int_0^\tau \frac{\mathrm{E}\left\{1(Y\geq u)X_r\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)\right\}}{\mathrm{E}\left\{1(Y\geq u)\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)\right\}}\mathrm{E}\left\{1(Y\geq u)\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top\mathbf{X}_{\mathcal{M}}\right)\right\}du$$

have the opposite signs. Contradiction! Therefore, $\boldsymbol{\beta}^*_{S_{+r}}\neq(\boldsymbol{\beta}_S^{*\top},0)^\top$.

Part (iii): Without loss of generality, we assume that X_r is the last element of $\mathbf{X}_{S_{+r}}$. Let $\mathbf{e}_r$ be a vector of length $(|S|+1)$ with the rth element 1 and all other elements 0. By definition, $v_S^{(0)}(u)=v_{S_{+r}}^{(0)}(u)$. Then by the mean value theorem,

$$\begin{aligned}&\int_0^\tau \mathrm{E}\left\{1(Y\geq u)X_r\lambda_0(u)\exp\left(\boldsymbol{\beta}_{\mathcal{M}}^{*\top}\mathbf{X}_{\mathcal{M}}\right)\right\}du-\int_0^\tau\frac{\mathrm{E}\left\{1(Y\geq u)X_r\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)\right\}}{r_S^{(0)}(\boldsymbol{\beta}_S^*,u)}v_S^{(0)}(u)du\\ &\quad-\int_0^\tau \mathrm{E}\left\{1(Y\geq u)X_r\lambda_0(u)\exp\left(\boldsymbol{\beta}_{\mathcal{M}}^{*\top}\mathbf{X}_{\mathcal{M}}\right)\right\}du+\int_0^\tau\frac{\mathrm{E}\left\{1(Y\geq u)X_r\exp\left(\boldsymbol{\beta}_{S_{+r}}^{*\top}\mathbf{X}_{S_{+r}}\right)\right\}}{r_{S_{+r}}^{(0)}(\boldsymbol{\beta}_S^*,u)}v_{S_{+r}}^{(0)}(u)du\\ &=\int_0^\tau\left[\frac{\mathrm{E}\left\{1(Y\geq u)X_r\mathbf{X}_{S_{+r}}^\top\exp\left(\tilde{\boldsymbol{\beta}}_{S_{+r}}^\top\mathbf{X}_{S_{+r}}\right)\right\}r_{S_{+r}}^{(0)}(\boldsymbol{\beta}^*_{S_{+r}},u)}{\left\{r_{S_{+r}}^{(0)}(\boldsymbol{\beta}^*_{S_{+r}},u)\right\}^2}\right.\\ &\qquad\left.-\frac{\mathrm{E}\left\{1\{Y\geq u\}X_r\exp\left(\boldsymbol{\beta}_{S_{+r}}^{*\top}\mathbf{X}_{S_{+r}}\right)\right\}\mathrm{E}\left\{1(Y\geq u)\mathbf{X}_{S_{+r}}^\top\exp\left(\boldsymbol{\beta}_{S_{+r}}^{*\top}\mathbf{X}_{S_{+r}}\right)\right\}}{\left\{r_{S_{+r}}^{(0)}(\boldsymbol{\beta}^*_{S_{+r}},u)\right\}^2}\right]\\ &\qquad\times\left\{\boldsymbol{\beta}^*_{S_{+r}}-(\boldsymbol{\beta}_S^{*\top},0)^\top\right\}\times\mathrm{E}\left\{1(Y\geq u)\lambda_0(u)\exp\left(\boldsymbol{\beta}_{\mathcal{M}}^{*\top}\mathbf{X}_{\mathcal{M}}\right)\right\}du\\ &=\int_0^\tau\mathbf{e}_r^\top\left[\frac{r_{S_{+r}}^{(2)}(\tilde{\boldsymbol{\beta}}_{S_{+r}},u)}{r_{S_{+r}}^{(0)}(\tilde{\boldsymbol{\beta}}_{S_{+r}},u)}-\frac{\left\{r_{S_{+r}}^{(1)}(\tilde{\boldsymbol{\beta}}_{S_{+r}},u)\right\}^{\otimes 2}}{\left\{r_{S_{+r}}^{(0)}(\tilde{\boldsymbol{\beta}}_{S_{+r}},u)\right\}^2}\right]\left(\boldsymbol{\beta}^*_{S_{+r}}-\boldsymbol{\beta}^*_S\right)v_{S_{+r}}^{(0)}(u)du,\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{S_{+r}}$ is between $\boldsymbol{\beta}^*_{S_{+r}}$ and $(\boldsymbol{\beta}_S^{*\top},0)^\top$. Thus,

$$\begin{aligned}&\left|\int_0^\tau \mathrm{E}\left\{1(Y\geq u)X_r\lambda_0(u)\exp\left(\boldsymbol{\beta}_{\mathcal{M}}^{*\top}\mathbf{X}_{\mathcal{M}}\right)\right\}du\right|\\ &\leq\left|\int_0^\tau \mathrm{E}\left\{1(Y\geq u)X_r\lambda_0(u)\exp\left(\boldsymbol{\beta}_{\mathcal{M}}^{*\top}\mathbf{X}_{\mathcal{M}}\right)\right\}du-\int_0^\tau\frac{\mathrm{E}\left\{1(Y\geq u)X_r\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)\right\}}{r_S^{(0)}(\boldsymbol{\beta}_S^*,u)}v_S^{(0)}(u)du\right|\\ &\leq\kappa_{\max}\|\mathbf{e}_r\|\left\|\boldsymbol{\beta}^*_{S_{+r}}-(\boldsymbol{\beta}_S^{*\top},0)^\top\right\|,\end{aligned}$$

where the first inequality follows from the proof of part (ii) that $\int_0^\tau \mathrm{E}\left\{1(Y\geq u)X_r\lambda_0(u)\exp\left(\boldsymbol{\beta}_{0\mathcal{M}}^\top\mathbf{X}_{\mathcal{M}}\right)\right\}du$ and

$$\int_0^\tau\frac{\mathrm{E}\left\{1(Y\geq u)X_r\exp\left(\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S\right)\right\}}{r_S^{(0)}(\boldsymbol{\beta}_S^*,u)}v_S^{(0)}(u)du$$

have the opposite signs, and the second inequality follows from Condition (F). Then, by Condition (E)

$$\left\|\boldsymbol{\beta}^*_{S_{+r}} - (\boldsymbol{\beta}^{*\top}_S, 0)^\top\right\| \geq \left|\int_0^\tau \mathrm{E}\left\{X_r f_T(u; \mathbf{X}_{\mathcal{M}}) S_C(u; \mathbf{X}_{\mathcal{M}_{\mathcal{A}}})\right\} du\right| \geq K \kappa_{\max}^{-1} n^{-\alpha}.$$

This completes the proof of Lemma 2. □

Proof of Lemma 3: We only prove $k = 1$, as $k = 2, 3$ can be proved similarly. Given an index set S of size $|S| = s \leq \rho$, let $\mathcal{B}^0_S(c) = \{\boldsymbol{\beta}_S : \|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leq c/(K\sqrt{s})\}$. By Conditions (B) and (D), it can be shown that for any $\boldsymbol{\pi} \in \mathbb{R}^s$ satisfying $\|\boldsymbol{\pi}\| = 1$,

$$0 \leq \|\boldsymbol{\pi}^\top \mathbf{X}\| \leq \sqrt{s} K, \tag{7}$$

$$\exp(\boldsymbol{\beta}_S^\top \mathbf{X}_S) \leq \exp((\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S)^\top \mathbf{X}_S) \exp(\boldsymbol{\beta}^{*\top}_S \mathbf{X}_S) \leq \exp(c + KL). \tag{8}$$

Define $h(\boldsymbol{\beta}_S, \boldsymbol{\pi}, t) = \left(\sqrt{s} K \exp(c + KL)\right)^{-1} \bar{Y}(t) \boldsymbol{\pi}^\top \mathbf{X}_S \exp\left(\boldsymbol{\beta}_S^\top \mathbf{X}_S\right)$. Then by (B) and (C), $h(\boldsymbol{\beta}_S, \boldsymbol{\pi}, u)$ is bounded between -1 and 1 uniformly over $\mathcal{B}^0_S(c)$, $\|\boldsymbol{\pi}\| = 1$, and $u \in [0, \tau]$. Define the function class

$$\mathcal{H}_S := \left\{h(\boldsymbol{\beta}_S, \boldsymbol{\pi}, u) : \boldsymbol{\beta}_S \in \mathcal{B}^0_S(c), \|\boldsymbol{\pi}\| = 1, u \in [0, \tau]\right\}.$$

Following the arguments used for Lemma 11 in [38] and Lemma 2.6.17 in [39], we can show that there exists some universal constant A_1 such that the class of functions $\mathcal{H}_S$ has a VC index bounded by $A_1 s$ (we refer the definitions of VC index page 85 in [39]). By Theorem 2.6.7 in [39], for any probability measure Q, there exists some universal constant A_2, such that the covering number $\sup_Q N[\epsilon \|\mathcal{H}_S\|_{Q,2}, \mathcal{H}_S, L_2(Q)]$ is bounded by $(A_2/\epsilon)^{2A_1 s}$ for any $\epsilon > 0$ (we refer the definition of covering numbers to page 83 in [39]).

Thus, by Theorem 1.1 in [40], there exists some constant A_3 that depends on A_2 only, such that for all $\epsilon > 0$,

$$\Pr\left[\sup_{\boldsymbol{\beta}_S \in \mathcal{B}^0_S(c), \|\boldsymbol{\pi}\|=1, u \in [0,\tau]} \left|\sqrt{n} \mathbb{G}_n\{h(\boldsymbol{\beta}_S, \boldsymbol{\pi}, u)\}\right| \geq \sqrt{n}\epsilon\right] \leq \frac{A_3}{\epsilon} \left(\frac{A_3 \epsilon^2}{A_1 s}\right)^{A_1 s} \exp\left(-2\epsilon^2\right).$$

By choosing $\epsilon = A_4 \sqrt{\rho \ln p}$ for some universal constant A_4, we obtain that

$$\Pr\left[\sup_{\boldsymbol{\beta}_S \in \mathcal{B}^0_S(c), \|\boldsymbol{\pi}\|=1, u \in [0,\tau]} \left|\mathbb{G}_n\{h(\boldsymbol{\beta}_S, \boldsymbol{\pi}, u)\}\right| \geq A_4 \sqrt{\rho \ln p}\right] \leq \exp\left(-5\rho \ln p\right).$$

Consequently,

$$\Pr\left[\sup_{|S|=s, \boldsymbol{\beta}_S \in \mathcal{B}^0_S(c), u \in [0,\tau]} \left\|R^{(1)}_S(\boldsymbol{\beta}_S, u) - r^{(1)}_S(\boldsymbol{\beta}_S, u)\right\| \geq A_4 \sqrt{s} K \exp(c + KL) \sqrt{\rho \ln p/n}\right] \leq \left(\frac{ep}{s}\right)^s \exp\left(-5\rho \ln p\right),$$

where the inequality follows from the combinatoric inequality $\binom{p}{s} \leq (ep/s)^s$.

Let $A_5 := A_4 K \exp(c + KL)$. Then,

$$\Pr\left\{\sup_{|S| \leq \rho, \boldsymbol{\beta}_S \in \mathcal{B}^0_S(c), u \in [0,\tau]} \left\|R^{(1)}_S(\boldsymbol{\beta}_S, u) - r^{(1)}_S(\boldsymbol{\beta}_S, u)\right\| \geq A_5 \sqrt{\rho^2 \ln p/n}\right\}$$
$$\leq \sum_{s=1}^{\rho} \left(\frac{ep}{s}\right)^s \exp\left(-5\rho \ln p\right) \leq \exp\left(-3\rho \ln p\right),$$

when n is sufficiently large. Thus, if $\rho^2 \ln p/n \to 0$,

$$\sup_{|S| \leq \rho, \boldsymbol{\beta}_S \in \mathcal{B}^0_S(c), u \in [0,\tau]} \left\|R^{(1)}_S(\boldsymbol{\beta}_S, u) - r^{(1)}_S(\boldsymbol{\beta}_S, u)\right\| \to_p 0.$$

Similarly, we can show that

$$\Pr\left\{\sup_{|S|\le\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c),u\in[0,\tau]}\left\|R_S^{(0)}(\boldsymbol{\beta}_S,u)-r_S^{(0)}(\boldsymbol{\beta}_S,u)\right\|\ge A_6\sqrt{\rho\ln p/n}\right\}\le\exp\left(-3\rho\ln p\right)\ \text{and}$$

$$\Pr\left\{\sup_{|S|\le\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c),u\in[0,\tau]}\left\|R_S^{(2)}(\boldsymbol{\beta}_S,u)-r_S^{(2)}(\boldsymbol{\beta}_S,u)\right\|\ge A_7\sqrt{\rho^3\ln p/n}\right\}\le\exp\left(-3\rho\ln p\right),$$

for some constants A_6 and A_7 that do not depend on n. This completes the proof of Lemma 3. □

Proof of Lemma 4: Given an index set S such that $|S|<\rho$, it is easy to see that

$$\begin{aligned}Z_S(\boldsymbol{\beta}_S)&\le n^{-1/2}\left|\mathbb{G}_n\left[\int_0^\tau\left\{\boldsymbol{\beta}_S^\top\mathbf{X}_{iS}-\boldsymbol{\beta}_S^{*\top}\mathbf{X}_{iS}\right\}dN_i(t)\right]\right|\\&\quad+n^{-1/2}\left|\mathbb{G}_n\left[\int_0^\tau\left\{\ln r_S^{(0)}(\boldsymbol{\beta}_S,t)-\ln r_S^{(0)}(\boldsymbol{\beta}_S^*,t)\right\}dN_i(t)\right]\right|\\&=:I+II.\end{aligned}$$

For the item I, by Conditions (B) and (C), it can be shown that $\left|\boldsymbol{\beta}_S^\top\mathbf{X}_{iS}-\boldsymbol{\beta}_S^{*\top}\mathbf{X}_{iS}\right|\le c/\left(K\sqrt{s}\right)K\sqrt{s}\le c$, and $\operatorname{var}\left\{\left(\boldsymbol{\beta}_S^\top\mathbf{X}_{iS}-\boldsymbol{\beta}_S^{*T}\mathbf{X}_{iS}\right)N_i(\tau)\right\}\le\kappa_{\max}c^2/(K^2s)$, for any $\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)$. Let $\{\epsilon_i\}_{i=1}^n$ be a Rademacher sequence. Then we have

$$\begin{aligned}\mathrm{E}\left[\sup_{\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)}I\right]&\le2\mathrm{E}\left[\sup_{\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)}\left|\mathbb{E}_n\left\{\epsilon_i\left(\boldsymbol{\beta}_S^\top\mathbf{X}_{iS}-\boldsymbol{\beta}_S^{*\top}\mathbf{X}_{iS}\right)N_t(\tau)\right\}\right|\right]\\&\le2\mathrm{E}\left[\sup_{\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\|_1\max_{1\le j\le p}\left|\mathbb{E}_n\left\{\epsilon_iX_{ij}N_i(\tau)\right\}\right|\right]\le2\frac{c}{K\sqrt{s}}\sqrt{s}\mathrm{E}\left[\max_{1\le j\le p}\left|\mathbb{E}_n\left\{\epsilon_iX_{ij}N_i(\tau)\right\}\right|\right]\le2a_nc/K,\end{aligned}$$

where the first inequality follows from Lemma 2.3.1 in [39], the second inequality is trivial, and the third inequality follows from Condition (B) and Lemma A.1 in [25]. Applying Bousquet's concentration theorem [41] yields that for any $r>0$,

$$\Pr\left\{\sup_{\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)}I\ge2\frac{c}{K}a_n+ra_n\sqrt{2\left(\frac{c^2}{K^2s}+2\frac{c^2}{K}a_n\right)+\frac{2r^2a_n^2c}{3}}\right\}\le\exp\left(-nr^2a_n^2\right).$$

Choose $r=2\sqrt{\rho}$. As $\rho\sqrt{\ln p/n}\to0$, we obtain that, when n is sufficiently large,

$$\Pr\left(\sup_{\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)}I\ge2\frac{c}{K}a_n+3\frac{c\sqrt{\rho}}{K\sqrt{s}}a_n\right)\le\exp\left[-4n\rho\left\{\sqrt{2K^2\ln(2p)/n}+K\ln(2p)/n\right\}^2\right]\le\exp\left(-8K^2\rho\ln p\right).\tag{9}$$

For the item II, let $\mathcal{R}_s(c)$ denote a ball with dimensionality s and radius $c/(K\sqrt{s})$. Then $\mathcal{B}_S^0(c)=\mathcal{R}_s(c)+\boldsymbol{\beta}_S^*$. Let $C_s:=\{C(\boldsymbol{\xi}_l)\}$ be a collection of cubes that cover the ball $R_s(c)$, where $C(\boldsymbol{\xi}_l)$ is a cube containing $\boldsymbol{\xi}_l$ with sides of length $c/(K\sqrt{s}n^2)$. Then $|C_s|\le(4n^2)^s$ and $\|\boldsymbol{\xi}_l\|\le C/(K\sqrt{s})$. For any $\boldsymbol{\xi}\in C(\boldsymbol{\xi}_l)$, $\|\boldsymbol{\xi}-\boldsymbol{\xi}_l\|\le c/(Kn^2)=:\zeta_n$.

Let $T_S(\boldsymbol{\xi}) := \mathbb{E}_n \int_0^\tau \ln r_S^{(0)}(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}, t) dN_i(t)$. By the mean value theorem,

$$
\begin{aligned}
& \left| \int_0^\tau \ln r_S^{(0)}(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}_l, t) dN_i(t) - \int_0^\tau \ln r_S^{(0)}(\boldsymbol{\beta}_S^*, t) dN_i(t) \right| \\
= & \left| \int_0^\tau \frac{1}{r_S^{(0)}(\boldsymbol{\beta}_S^* + \tilde{\boldsymbol{\xi}}_l, t)} \mathrm{E}\left[\bar{Y}(t) \exp\left\{ \left(\boldsymbol{\beta}_S^* + \tilde{\boldsymbol{\xi}}_l \right)^\top \mathbf{X}_S \right\} \boldsymbol{\xi}_l^\top \mathbf{X}_S \right] dN_i(t) \right| \\
\le & \left| \int_0^\tau \frac{1}{r_S^{(0)}(\boldsymbol{\beta}_S^* + \tilde{\boldsymbol{\xi}}_l, t)} \mathrm{E}\left[\bar{Y}(t) \exp\left\{ \left(\boldsymbol{\beta}_S^* + \tilde{\boldsymbol{\xi}}_l \right)^\top \mathbf{X}_S \right\} \right] \|\boldsymbol{\xi}_l\|_1 \|\mathbf{X}\|_\infty dN_i(t) \right| \\
\le & \left| \int_0^\tau \frac{1}{r_S^{(0)}(\boldsymbol{\beta}_S^* + \tilde{\boldsymbol{\xi}}_l, t)} r_S^{(0)}(\boldsymbol{\beta}_S^* + \tilde{\boldsymbol{\xi}}_l, t) dN_i(t) \right| \|\boldsymbol{\xi}_l\|_1 \|\mathbf{X}\|_\infty \le \tau c/(K\sqrt{s})\sqrt{s}K = c\tau,
\end{aligned}
$$

where $\tilde{\boldsymbol{\xi}}_l$ is between $\boldsymbol{\xi}_l$ and 0. Applying Bernstein's inequality yields that for any $r > 0$,

$$
\Pr\left[n \left| T_S(\boldsymbol{\xi}_l) - T_S(0) - \mathrm{E}\left\{ T_S(\boldsymbol{\xi}_l) - T_S(0) \right\} \right| > r \right] \le 2 \exp\left(-\frac{1}{2} \frac{r^2}{nc^2\tau^2 + 2c\tau r/3} \right),
$$

and consequently, by choosing $r = 6c\tau\sqrt{n\rho \ln p}$,

$$
\begin{aligned}
& \Pr\left[\max_{1 \le l \le (4n^2)^s} \left| T_S(\boldsymbol{\xi}_l) - T_S(0) - \mathrm{E}\left\{ T_S(\boldsymbol{\xi}_l) - T_S(0) \right\} \right| > 6c\tau\sqrt{n\rho \ln p}/n \right] \\
\le & \; 2(4n^2)^s \exp\left(-\frac{1}{2} \frac{36c^2\tau^2 n\rho \ln p}{nc^2\tau^2 + 12c^2\tau^2\sqrt{n\rho \ln p}} \right) \le 2\left(4n^2\right)^s \exp\left(-12\rho \ln p\right) \\
\le & \; 2\exp(-12\rho \ln p + s\ln 4 + 2s\ln n) \le 2\exp(-10\rho \ln p), \qquad (10)
\end{aligned}
$$

where n is sufficiently large.

Given any $\boldsymbol{\xi} \in C(\boldsymbol{\xi}_l)$, we can similarly show that

$$
\left| \int_0^\tau \left\{ \ln r_S^{(0)}(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}, t) - \ln r_S^{(0)}(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}_l, t) \right\} dN_i(t) \right| \le K\sqrt{s}\zeta_n.
$$

Therefore,

$$
\left| T_S(\boldsymbol{\xi}) - T_S(\boldsymbol{\xi}_l) - \mathrm{E}\left\{ T_S(\boldsymbol{\xi}) - T_S(\boldsymbol{\xi}_l) \right\} \right| \le 2K\sqrt{s}\zeta_n \le 2K\sqrt{\rho}\zeta_n. \qquad (11)
$$

Combining (10) and (11) implies that

$$
\Pr\left(\sup_{\boldsymbol{\beta}_S \in \mathscr{B}_S^0(c)} II \ge 6c\tau\sqrt{\rho \ln p/n} + 2K\sqrt{\rho}\zeta_n \right) \le 2\exp(-10\rho \ln p). \qquad (12)
$$

By (9) and (12), we have

$$
\Pr\left(\sup_{\boldsymbol{\beta}_S \in \mathscr{B}_S^0(c)} Z_S(\boldsymbol{\beta}_S) \ge 2\frac{c}{K}a_n + 3\frac{c\sqrt{\rho}}{K\sqrt{s}}a_n + 6c\tau\sqrt{\rho \ln p/n} + 2K\sqrt{\rho}\zeta_n \right) \le \exp\left(-8K^2\rho \ln p\right) + 2\exp(-10\rho \ln p).
$$

By the combinatoric inequality $\binom{p}{s} \le (ep/s)^s$, we obtain that

$$
\begin{aligned}
& \Pr\left(\sup_{|S| \le \rho, \boldsymbol{\beta}_S \in \mathscr{B}_S^0(c)} Z_S(\boldsymbol{\beta}_S) \ge 2\frac{c}{K}a_n + 3\frac{c\sqrt{\rho}}{K\sqrt{s}}a_n + 6c\tau\sqrt{\rho \ln p/n} + 2K\sqrt{\rho}\zeta_n \right) \\
\le & \sum_{s=1}^{\rho} (ep/s)^s \left\{ \exp\left(-8K^2\rho \ln p\right) + 2\exp(-10\rho \ln p) \right\} \le 3\exp(-6\rho \ln p).
\end{aligned}
$$

This completes the proof of Lemma 4. □

Proof of Lemma 5: We first define the following events:

$$\Omega_1 := \left\{ \sup_{|S|\le\rho, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(c), u\in[0,\tau]} \left| R_S^{(0)}(\boldsymbol{\beta}_S, u) - r_S^{(0)}(\boldsymbol{\beta}_S, u) \right| \ge A_6 \sqrt{\rho \ln p/n} \right\} \text{ and}$$

$$\Omega_2 := \left\{ \sup_{|S|\le\rho, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(c), u\in[0,\tau]} \left\| R_S^{(1)}(\boldsymbol{\beta}_S, u) - r_S^{(1)}(\boldsymbol{\beta}_S, u) \right\| \ge A_5 \sqrt{\rho^2 \ln p/n} \right\}.$$

By Lemma 3, we obtain that $\Pr(\Omega_1) \le \exp(-3\rho \ln p)$ and $\Pr(\Omega_2) \le \exp(-3\rho \ln p)$. In the rest of the proof, we restrict our attention to $\Omega_1^c \cap \Omega_2^c$. By the arguments in [23], we can show that

$$\begin{aligned}
D_S(\boldsymbol{\beta}_S) &\le \mathbb{E}_n \left| \int_0^\tau \left[\ln\left\{ \frac{R_S^{(0)}(\boldsymbol{\beta}_S, t)}{r_S^{(0)}(\boldsymbol{\beta}_S, t)} \right\} - \ln\left\{ \frac{R_S^{(0)}(\boldsymbol{\beta}_S^*, t)}{r_S^{(0)}(\boldsymbol{\beta}_S^*, t)} \right\} \right] dN_i(t) \right| \le \sup_{0\le t\le\tau} \left| \ln\left\{ \frac{R_S^{(0)}(\boldsymbol{\beta}_S, t)}{r_S^{(0)}(\boldsymbol{\beta}_S, t)} \right\} - \ln\left\{ \frac{R_S^{(0)}(\boldsymbol{\beta}_S^*, t)}{r_S^{(0)}(\boldsymbol{\beta}_S^*, t)} \right\} \right| \\
&\le \sup_{0\le t\le\tau} \left| \frac{r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t)}{R_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t)} (\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^\top \left[\frac{R_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t)}{r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t)} - \frac{R_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t) r_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t)}{\left\{ r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t) \right\}^2} \right] \right\| \\
&= \sup_{0\le t\le\tau} \left| (\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^\top \frac{R_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t) - r_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t)}{R_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t)} \right| + \sup_{0\le t\le\tau} \left| \{\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^\top r_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t) \left\{ \frac{1}{R_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t)} - \frac{1}{r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t)} \right\} \right| \\
&=: I + II.
\end{aligned}$$

We first consider I. By Condition (B),

$$\begin{aligned}
r_S^{(0)}(\boldsymbol{\beta}_S, t) &= \mathrm{E}\left\{ 1(Y \ge t) \exp\left(\boldsymbol{\beta}_S^\top \mathbf{X}_S \right) \right] \ge \mathrm{E}\left\{ 1(Y \ge t) \exp\left(- \|\boldsymbol{\beta}_S\|_1 \|\mathbf{X}_S\|_\infty \right) \right\} \\
&\ge \mathrm{E}\left[1(Y \ge t) \exp\{-(c + KL)\} \right] \ge \omega \exp\{-(c + KL)\},
\end{aligned}$$

for any $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(c)$. We obtain that

$$\inf_{|S|\le\rho, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(c), t\in[0,\tau]} r_S^{(0)}(\boldsymbol{\beta}_S, t) \ge \omega \exp\{-(c + KL)\}.$$

Then,

$$\begin{aligned}
I &\le \sup_{0\le t\le\tau} \left| \frac{1}{R_S^{(0)}(\tilde{\boldsymbol{\beta}}_S, t)} \right| \left| (\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^\top \left\{ R_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t) - r_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t) \right\} \right| \\
&\le \frac{\left\| \boldsymbol{\beta}_S - \boldsymbol{\beta}_S^* \right\|}{\inf_{|S|\le\rho, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(c), t\in[0,\tau]} r_S^{(0)}(\boldsymbol{\beta}_S, t) - A_6 \sqrt{\rho \ln p/n}} \sup_{|S|\le\rho, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(c), t\in[0,\tau]} \left\| R_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t) - r_S^{(1)}(\tilde{\boldsymbol{\beta}}_S, t) \right\| \\
&\le \frac{1}{\omega \exp\{-(c + KL)\} - A_6 \sqrt{\rho \ln p/n}} \times \frac{c}{K\sqrt{s}} \times A_5 \sqrt{\rho^2 \ln p/n} \le A_8 \sqrt{\rho^2 \ln p/n},
\end{aligned}$$

for some constant A_8 that does not depend on n.

We now bound II. By Condition (B),

$$\mathrm{E}\left\{ 1(Y \ge t) X_j \exp\left(\boldsymbol{\beta}_S^\top \mathbf{X}_S \right) \right\} \le K \mathrm{E}\left\{ \exp\left(\boldsymbol{\beta}_S^\top \mathbf{X}_S \right) \right\} \le K \exp\left(\|\mathbf{X}_S\|_\infty \|\boldsymbol{\beta}_S\|_1 \right) \le K \exp(c + KL).$$

Thus,

$$
\begin{aligned}
II \le & \sup_{|S|\le\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c),t\in[0,\tau]} \left\| r_S^{(1)}(\boldsymbol{\beta}_S,t)\right\| \sup_{|S|\le\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c),t\in[0,\tau]} \left| \frac{1}{R_S^{(0)}(\tilde{\boldsymbol{\beta}}_S,t) r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S,t)} \right| \\
& \times \sup_{|S|\le\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c),t\in[0,\tau]} \left| R_S^{(0)}(\boldsymbol{\beta}_S,t) - r_S^{(0)}(\boldsymbol{\beta}_S,t)\right| \left\| \boldsymbol{\beta}_S - \boldsymbol{\beta}_S^* \right\| \\
\le & \frac{K\sqrt{\rho}\exp(c+KL)}{\omega\exp\{-(c+KL)\}\left[\omega\exp\{-(c+KL)\} - A_6\sqrt{\rho\ln p/n}\right]} \times \frac{A_6\sqrt{\rho\ln p/n}\times c}{K\sqrt{s}} \le A_9\sqrt{\rho^2\ln p/n},
\end{aligned}
$$

for some constant A_9 that is free of n.
Withdrawing the restriction to $\Omega_1^c \cap \Omega_2^c$, the above results indicate that

$$
\Pr\left(\sup_{|S|\le\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)} D_S(\boldsymbol{\beta}_S) \ge A_{10}\sqrt{\rho^2\ln p/n} \right) \le 2\exp(-3\rho\ln p),
$$

for some constant A_{10}. This completes the proof of Lemma 5. □

Proof of Lemma 6: Given any index set S satisfying $|S|\le\rho$ and $\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)$, by Taylor's expansion,

$$
\begin{aligned}
\Gamma_S(\boldsymbol{\beta}_S) &= \mathrm{E}\left[-\int_0^\tau \left\{\boldsymbol{\beta}_S^\top\mathbf{X}_S - \ln r_S^{(0)}(\boldsymbol{\beta}_S,t)\right\} dN(t)\right] \\
&= \mathrm{E}\left[-\int_0^\tau \left\{\boldsymbol{\beta}_S^{*\top}\mathbf{X}_S - \ln r_S^{(0)}(\boldsymbol{\beta}_S^*,t)\right\} dN(t)\right] + \mathrm{E}\left[-\int_0^\tau \left\{\mathbf{X}_S^\top - \frac{r_S^{(1)}(\boldsymbol{\beta}_S^*,t)}{r_S^{(0)}(\boldsymbol{\beta}_S^*,t)}\right\} dN(t)\right](\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) \\
&\quad + \frac{1}{2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^\top \mathrm{E}\left(\int_0^\tau \left[\frac{r_S^{(2)}(\tilde{\boldsymbol{\beta}}_S,t)}{r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S,t)} - \frac{\left\{r_S^{(1)}(\tilde{\boldsymbol{\beta}}_S,t)\right\}^{\otimes 2}}{\left\{r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S,t)\right\}^2}\right] dN(t)\right)(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*),
\end{aligned}
$$

where $\tilde{\boldsymbol{\beta}}_S$ is between $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_S^*$. Noting that

$$
\mathrm{E}\left[-\int_0^\tau \left\{\mathbf{X}_S^\top - \frac{r_S^{(1)}(\boldsymbol{\beta}_S^*,t)}{r_S^{(0)}(\boldsymbol{\beta}_S^*,t)}\right\} dN(t)\right] = 0,
$$

by Condition (F), we have $\Gamma_S(\boldsymbol{\beta}_S) - \Gamma_S(\boldsymbol{\beta}_S^*) \ge \kappa_{\min}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2/2$. Similarly,

$$
\begin{aligned}
\Gamma_S(\boldsymbol{\beta}_S) - \Gamma_S(\boldsymbol{\beta}_S^*) &= \frac{1}{2}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^\top \mathrm{E}\left(\int_0^\tau \left[\frac{r_S^{(2)}(\tilde{\boldsymbol{\beta}}_S,t)}{r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S,t)} - \frac{\left\{r_S^{(1)}(\tilde{\boldsymbol{\beta}}_S,t)\right\}^{\otimes 2}}{\left\{r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S,t)\right\}^2}\right] dN(t)\right)(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) \\
\le \frac{1}{2}&\left[\int_0^\tau (\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)^\top \left\{\frac{r_S^{(2)}(\tilde{\boldsymbol{\beta}}_S,t)}{r_S^{(0)}(\tilde{\boldsymbol{\beta}}_S,t)}\right\}(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*) dN(t)\right] \le \frac{1}{2}\kappa_{\max}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2,
\end{aligned}
$$

where the last inequality follows from Condition (F). This completes the proof of Lemma 6. □

Proof of Lemma 7: Let

$$
\Omega_3 := \left\{ \sup_{|S|\le\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)} Z_S(\boldsymbol{\beta}_S) \ge 2\frac{c}{K}a_n + 3\frac{c\sqrt{\rho}}{K\sqrt{s}}a_n + 6c\tau\sqrt{\rho\ln p/n} + 2K\sqrt{\rho}\zeta_n \right\} \text{ and}
$$

$$
\Omega_4 := \left\{ \sup_{|S|\le\rho,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(c)} D_S(\boldsymbol{\beta}_S) \ge A_{10}\sqrt{\rho^2\ln p/n} \right\}.
$$

We consider $\Omega_3^c \cap \Omega_4^c$, which holds with probability of at least $1 - 5\exp(-3\rho \ln p)$, by Lemmas 4 and 5. In the rest of the proof, we restrict our attention to $\Omega_3^c \cap \Omega_4^c$.

Given an index set S such that $|S| \le \rho$, for any $\boldsymbol{\beta}_S$ with $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| = A_{11}(\rho^2 \ln p/n)^{1/4}$ for some constant A_{11} defined later, we have $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(c)$ when n is sufficiently large such that $A_{11}(\rho^2 \ln p/n)^{1/4} \le c/(K\sqrt{\rho})$, since $\rho^4 \ln p/n \to 0$. Therefore, $\boldsymbol{\beta}_S \in \partial\mathcal{B}_S^0(c)$, where $\partial\mathcal{B}_S^0(c)$ denotes the boundary of $\mathcal{B}_S^0(c)$.

Noting that $n^{-1}\tilde{\ell}_S(\boldsymbol{\beta}_S) = -\mathbb{E}_n[\gamma_S(\boldsymbol{\beta}_S; \mathbf{X}_i, Y_i, \delta_i)]$,

$$\begin{aligned}
&n^{-1}\{-\ell_S(\boldsymbol{\beta}_S) + \ell_S(\boldsymbol{\beta}_S^*)\} = n^{-1}\{-\ell_S(\boldsymbol{\beta}_S) + \ell_S(\boldsymbol{\beta}_S^*)\} - n^{-1}\left\{-\tilde{\ell}_S(\boldsymbol{\beta}_S) + \tilde{\ell}_S(\boldsymbol{\beta}_S^*)\right\} \\
&\qquad + n^{-1}\left\{-\tilde{\ell}_S(\boldsymbol{\beta}_S) + \tilde{\ell}_S(\boldsymbol{\beta}_S^*)\right\} - \{\Gamma_S(\boldsymbol{\beta}_S) - \Gamma_S(\boldsymbol{\beta}_S^*)\} + \{\Gamma_S(\boldsymbol{\beta}_S) - \Gamma_S(\boldsymbol{\beta}_S^*)\} \\
&\ge \{\Gamma_S(\boldsymbol{\beta}_S) - \Gamma_S(\boldsymbol{\beta}_S^*)\} - n^{-1}\left|\{\ell_S(\boldsymbol{\beta}_S) - \ell_S(\boldsymbol{\beta}_S^*)\} - \left\{\tilde{\ell}_S(\boldsymbol{\beta}_S) - \tilde{\ell}_S(\boldsymbol{\beta}_S^*)\right\}\right| \\
&\quad - \left|n^{-1}\left\{-\tilde{\ell}_S(\boldsymbol{\beta}_S) + \tilde{\ell}_S(\boldsymbol{\beta}_S^*)\right\} - \{\Gamma_S(\boldsymbol{\beta}_S) - \Gamma_S(\boldsymbol{\beta}_S^*)\}\right| \\
&= \{\Gamma_S(\boldsymbol{\beta}_S) - \Gamma_S(\boldsymbol{\beta}_S^*)\} - |D_S(\boldsymbol{\beta}_S)| - |Z_S(\boldsymbol{\beta}_S)| \\
&\ge \frac{1}{2}\kappa_{\min}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2 - 2\frac{c}{K}a_n - 3\frac{c\sqrt{\rho}}{K\sqrt{s}}a_n - 6c\tau\sqrt{\rho \ln p/n} - 2K\sqrt{\rho}\zeta_n - A_{10}\sqrt{\rho^2 \ln p/n} \\
&\ge \frac{1}{2}\kappa_{\min}A_{11}^2\frac{\sqrt{\rho^2 \ln p}}{\sqrt{n}} - 2\frac{c}{K}\frac{\sqrt{2K^2\ln(2p)}}{\sqrt{n}} - 2\frac{c}{K}K\frac{\ln(2p)}{n} - 3\frac{c\sqrt{\rho}}{K\sqrt{s}}\frac{\sqrt{2K^2\ln(2p)}}{\sqrt{n}} \\
&\quad - 3\frac{c\sqrt{\rho}}{K\sqrt{s}}K\frac{\ln(2p)}{n} - 6c\tau\frac{\sqrt{\rho \ln p}}{\sqrt{n}} - 2K\sqrt{\rho}\frac{c}{Kn^2} - A_{10}\frac{\sqrt{\rho^2 \ln p}}{\sqrt{n}} > 0,
\end{aligned}$$

with some constant A_{11} satisfying $\kappa_{\min}A_{11}^2 > A_{10}$. Therefore,

$$\inf_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| = A_{11}\left(\rho^2 \ln p/n\right)^{1/4}} n^{-1}\{-\ell_S(\boldsymbol{\beta}_S) + \ell_S(\boldsymbol{\beta}_S^*)\} > 0.$$

By the concavity of ℓ_S, $\left\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\right\| \le A_{11}\left(\rho^2 \ln p/n\right)^{1/4}$. Withdrawing the restriction to $\Omega_3^c \cap \Omega_4^c$, we have

$$\Pr\left\{\sup_{|S| \le \rho}\left\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\right\| \le A_{11}\left(\rho^2 \ln p/n\right)^{1/4}\right\} \ge 1 - 5\exp(-3\rho \ln p).$$

On the other hand,

$$\begin{aligned}
\left|\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_S(\boldsymbol{\beta}_S^*)\right| &\le \frac{1}{2}\kappa_{\max}\frac{\sqrt{\rho^2 \ln p}}{\sqrt{n}} + 2\frac{c}{K}\frac{\sqrt{2K^2\ln(2p)}}{\sqrt{n}} + 2\frac{c}{K}K\frac{\ln(2p)}{n} + 3\frac{c\sqrt{\rho}}{K\sqrt{s}}\frac{\sqrt{2K^2\ln(2p)}}{\sqrt{n}} \\
&\quad + 3\frac{c\sqrt{\rho}}{K\sqrt{s}}K\frac{\ln(2p)}{n} + 6c\tau\frac{\sqrt{\rho \ln p}}{\sqrt{n}} + 2K\sqrt{\rho}\frac{c}{Kn^2} + A_{10}\frac{\sqrt{\rho^2 \ln p}}{\sqrt{n}} \le A_{12}\frac{\sqrt{\rho^2 \ln p}}{\sqrt{n}},
\end{aligned}$$

for some constant A_{12}. Similarly, we obtain that

$$\Pr\left\{\sup_{|S| \le \rho}\left|\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_S(\boldsymbol{\beta}_S^*)\right| \le A_{12}\sqrt{\rho^2 \ln p/n}\right\} \ge 1 - 5\exp(-3\rho \ln p).$$

This complete the proof of Lemma 7. □

Proof of Lemma 8: Given any $S \subset \{1, \ldots, p\}$ and r such that $|S| < \rho$, $r \in S^c$,

$$\begin{aligned}
&\left[n^{-1}\left\{\ell_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) - \ell_S(\boldsymbol{\beta}_S^*)\right\} - \left\{-\Gamma_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) + \Gamma_S(\boldsymbol{\beta}_S^*)\right\}\right] \\
&= \left[n^{-1}\left\{\ell_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) - \ell_S(\boldsymbol{\beta}_S^*)\right\} - n^{-1}\left\{\tilde{\ell}_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) - \tilde{\ell}_S(\boldsymbol{\beta}_S^*)\right\}\right] + \left[n^{-1}\left\{\tilde{\ell}_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) - \tilde{\ell}_S(\boldsymbol{\beta}_S^*)\right\} - \left\{-\Gamma_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) + \Gamma_S(\boldsymbol{\beta}_S^*)\right\}\right] \\
&=: I + II.
\end{aligned}$$

Noting that $n^{-1}\tilde{\ell}_S(\boldsymbol{\beta}_S) = -\mathbb{E}_n\{\gamma_S(\boldsymbol{\beta}_S;\mathbf{X}_i, Y_i, \delta_i)\}$, it is easy to check that

$$\begin{aligned}
|I| &\le \left|\mathbb{E}_n\left[\int_0^\tau \left\{\ln R_S^{(0)}(\boldsymbol{\beta}_S^*, u) - \ln r_S^{(0)}(\boldsymbol{\beta}_S^*, u)\right\} dN_i(u)\right]\right| \\
&\quad + \left|\mathbb{E}_n\left[\int_0^\tau \left\{\ln R_S^{(0)}(\boldsymbol{\beta}_{S_{+r}}^*, u) - \ln r_S^{(0)}(\boldsymbol{\beta}_{S_{+r}}^*, u)\right\} dN_i(u)\right]\right| =: I_1 + I_2. \\
|II| &= \left|\left[\mathbb{E}_n\left\{\gamma_S(\boldsymbol{\beta}_{S_{+r}}^*;\mathbf{X}_i, Y_i, \delta_i)\right\} - \Gamma_S(\boldsymbol{\beta}_{S_{+r}}^*)\right] - \left[\mathbb{E}_n\{\gamma_S(\boldsymbol{\beta}_S^*;\mathbf{X}_i, Y_i, \delta_i)\} - \Gamma_S(\boldsymbol{\beta}_S^*)\right]\right| \\
&\le \left|\mathbb{E}_n\left\{\gamma_S(\boldsymbol{\beta}_{S_{+r}}^*;\mathbf{X}_i, Y_i, \delta_i)\right\} - \Gamma_S(\boldsymbol{\beta}_{S_{+r}}^*)\right| + \left|\mathbb{E}_n\{\gamma_S(\boldsymbol{\beta}_S^*;\mathbf{X}_i, Y_i, \delta_i)\} - \Gamma_S(\boldsymbol{\beta}_S^*)\right| =: II_1 + II_2.
\end{aligned}$$

We restrict our attention to Ω_1^c, where Ω_1 is defined in Lemma 5. We consider I_1 first.

$$\begin{aligned}
|I_1| &= \left|\mathbb{E}_n\left[\int_0^\tau \left\{\ln R_S^{(0)}(\boldsymbol{\beta}_S^*, u) - \ln r_S^{(0)}(\boldsymbol{\beta}_S^*, u)\right\} dN_i(u)\right]\right| \le \sup_{0\le t\le\tau}\left|\ln\frac{R_S^{(0)}(\boldsymbol{\beta}_S^*, t)}{r_S^{(0)}(\boldsymbol{\beta}_S^*, t)}\right| \\
&= \sup_{0\le t\le\tau}\left|\frac{1}{\tilde{r}(t)}\left\{R_S^{(0)}(\boldsymbol{\beta}_S^*, t) - r_S^{(0)}(\boldsymbol{\beta}_S^*, t)\right\}\right| \le \sup_{0\le t\le\tau}\left|\frac{A_6\sqrt{\rho\ln p/n}}{r_S^{(0)}(\boldsymbol{\beta}_S^*, t) - A_6\sqrt{\rho\ln p/n}}\right| \\
&\le \frac{A_6\sqrt{\rho\ln p/n}}{\omega\exp\{-(c+KL)\} - A_6\sqrt{\rho\ln p/n}} \le A_{13}\sqrt{\rho\ln p/n},
\end{aligned}$$

where $\tilde{r}(t)$ is between $R_S^{(0)}(\boldsymbol{\beta}_S^*, t)$ and $r_S^{(0)}(\boldsymbol{\beta}_S^*, t)$. By the same argument, $I_2 \le A_{13}\sqrt{\rho\ln p/n}$ as well. Therefore, $|I| \le 2A_{13}\sqrt{\rho\ln p/n}$.

By Conditions (B) and (D), $\left|\gamma_S(\boldsymbol{\beta}_S^*;\mathbf{X}_i, Y_i, \delta_i)\right| \le |\boldsymbol{\beta}_S^{*\top}\mathbf{X}_i| + \ln r_S^{(0)}(\boldsymbol{\beta}_S^*, t) \le 2KL$. Applying Bernstein's inequality yields

$$\begin{aligned}
&\Pr\left(\left|\mathbb{E}_n\{\gamma_S(\boldsymbol{\beta}_S^*;\mathbf{X}_i, Y_i, \delta_i)\} - \Gamma_S(\boldsymbol{\beta}_S^*)\right| > 6KL\sqrt{\rho\ln p/n}\right) \\
&\le 2\exp\left(-\frac{1}{2}\frac{36K^2L^2n\rho\ln p}{4nK^2L^2 + 12K^2L^2\sqrt{\rho n\ln p}/3}\right) \le 2\exp(-4\rho\ln p).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Pr\left(\sup_{|S|<\rho, r\in S^c}|II| > 12KL\sqrt{\rho\ln p/n}\right) \le \sum_{|S|<\rho}\sum_{r=1}^{p} 2\exp(-4\rho\ln p) \\
&\le p\sum_{s=1}^{\rho-1}(ep/s)^s 2\exp(-4\rho\ln p) \le 2\exp(-3\rho\ln p).
\end{aligned}$$

Withdrawing the restriction to Ω_1^c, we obtain that

$$\begin{aligned}
&\Pr\left[\sup_{|S|<\rho, r\in S^c}\left|n^{-1}\left\{\ell_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) - \ell_S(\boldsymbol{\beta}_S^*)\right\} - \left\{-\Gamma_{S_{+r}}(\boldsymbol{\beta}_{S_{+r}}^*) + \Gamma_S(\boldsymbol{\beta}_S^*)\right\}\right| \ge 2A_{13}\sqrt{\rho\ln p/n} + 12KL\sqrt{\rho\ln p/n}\right] \\
&\le 2\exp(-3\rho\ln p) + \exp(-3\rho\ln p) \le 3\exp(-3\rho\ln p).
\end{aligned}$$

Let $A_{14} = 2A_{13} + 12KL$ and we complete the proof of Lemma 8. □

References

[1] R. J. Tibshirani, The lasso method for variable selection in the Cox model, Statistics in Medicine 16 (4) (1997) 385–395.

[2] H. Zou, A note on path-based variable selection in the penalized proportional hazards model, Biometrika 95 (2008) 241–247.

[3] J. Huang, T. Sun, Z. Ying, Y. Yu, C.-H. Zhang, Oracle inequalities for the Lasso in the Cox model, The Annals of Statistics 41 (3) (2013) 1142–1165.

[4] J. Fan, R. Samworth, Y. Wu, Ultrahigh dimensional feature selection: Beyond the linear model, Journal of Machine Learning Research 10 (2009) 2013–2038.

[5] J. Bradic, J. Fan, J. Jiang, Regularization for Cox's proportional hazards model with NP-dimensionality, The Annals of Statistics 39 (6) (2011) 3092–3120. doi:10.1214/11-AOS911.

[6] H. Wang, Forward regression for ultra-high dimensional variable screening, Journal of the American Statistical Association 104 (488) (2009) 1512–1524.

[7] S. Luo, Z. Chen, Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space, Journal of the American Statistical Association 109 (507) (2014) 1229–1240.

[8] N. Hao, H. H. Zhang, Interaction screening for ultrahigh-dimensional data, Journal of the American Statistical Association 109 (507) (2014) 1285–1301.

[9] C.-K. Ing, T. L. Lai, A stepwise regression method and consistent model selection for high-dimensional sparse linear models, Statistica Sinica 21 (2011) 1473–1513.

[10] W. Zhong, T. Zhang, Y. Zhu, J. S. Liu, Correlation pursuit: forward stepwise variable selection for index models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74 (5) (2012) 849–870.

[11] M.-Y. Cheng, T. Honda, J.-T. Zhang, Forward variable selection for sparse ultra-high dimensional varying coefficient models, Journal of the American Statistical Association 111 (515) (2016) 1209–1221.

[12] S. D. Zhao, Y. Li, Principled sure independence screening for Cox models with ultra-high-dimensional covariates, Journal of Multivariate Analysis 105 (1) (2012) 397–411.

[13] A. Gorst-Rasmussen, T. Scheike, Independent screening for single-index hazard rate models with ultrahigh dimensional features, Journal of the Royal Statistical Society B 75 (2) (2013) 217–245.

[14] H. G. Hong, J. Kang, Y. Li, Conditional screening for ultra-high dimensional covariates with survival outcomes, Lifetime Data Analysis 24 (2018) 45–71.

[15] X. He, L. Wang, H. G. Hong, Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, The Annals of Statistics 41(1) (2013) 342–369.

[16] R. Song, W. Lu, S. Ma, X. J. Jeng, Censored rank independence screening for high-dimensional survival data, Biometrika 101 (4) (2014) 799–814.

[17] H. G. Hong, X. Chen, D. C. Christiani, Y. Li, Integrated powered density: Screening ultrahigh dimensional covariates with survival outcomes, Biometrics 74 (2) (2017) 421–429.

[18] H. G. Hong, Y. Li, Feature selection of ultrahigh-dimensional covariates with survival outcomes: a selective review, Appl Math Ser B 32 (2017) 379–396.

[19] J. Li, Q. Zheng, L. Peng, Z. Huang, Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes, Biometrics 72 (4) (2016) 1145–1154.

[20] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (2008) 759–771.

[21] C. T. Volinsky, A. E. Raftery, Bayesian information criterion for censored survival models, Biometrics 56 (2000) 256–262.

[22] R. Xu, F. Vaida, D. P. Harrington, Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models, Statistica Sinica 19 (2009) 819–842.

[23] S. Kong, B. Nan, Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso, Statistica Sinica 24 (1) (2014) 25–42.

[24] Q. Zheng, L. Peng, X. He, Globally adaptive quantile regression with ultra-high dimensional data, The Annals of Statistics 43 (5) (2015) 2225.

[25] S. A. van de Geer, High-dimensional generalized linear models and the lasso, The Annals of Statistics 36 (2) (2008) 614–645.

[26] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space (with discussion), Journal of Royal Statistical Society B 70 (5) (2008) 849–911.

[27] D. Y. Lin, L.-J. Wei, The robust inference for the Cox proportional hazards model, Journal of the American Statistical Association 84 (408) (1989) 1074–1078.

[28] J. Fine, Comparing nonnested cox models, Biometrika 89 (3) (2002) 635–648.

[29] J. Fan, R. Song, Sure independence screening in generalized linear models with NP-dimensionality, The Annals of Statistics 38 (2010) 3567–3604.

[30] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) 58 (1996) 267–288.

[31] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of the American Statistical Association 96 (456) (2001) 1348–1360.

[32] C. H. Zhang, Nearly unbiased variable selection under minimax concave penalty, The Annals of Statistics 38 (2) (2010) 894–942.

[33] E. Vittinghoff, C. E. McCulloch, Relaxing the rule of ten events per variable in logistic and Cox regression, American Journal of Epidemiology 165 (6) (2007) 710–718.

[34] S. Zöchbauer-Müller, J. D. Minna, A. F. Gazdar, Aberrant DNA methylation in lung cancer: biological and clinical implications, The Oncologist 7 (5) (2002) 451–457.

[35] J. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (5) (2001) 1189–1232.

[36] S. Luo, J. Xu, Z. Chen, Extended Bayesian information criterion in the Cox model with a high-dimensional feature space, Annals of the Institute of Statistical Mathematics 67 (2) (2015) 287–311.

[37] K. D. Schmidt, On the Covariance of Monotone Functions of a Random Variable, Professoren des Inst. für Math. Stochastik, 2003.

[38] A. Belloni, V. Chernozhukov, ℓ_1-penalized quantile regression in high-dimensional sparse models, The Annals of Statistics 39 (1) (2011) 82–130.

[39] A. W. van der Vaart, J. A. Wellner, Weak Convergence, Springer: New York, 1996.

[40] M. Talagrand, Sharper bounds for Gaussian and empirical processes, The Annals of Probability 22 (1) (1994) 28–76.

[41] O. Bousquet, A Bennett concentration inequality and its application to suprema of empirical processes, Comptes Rendus Mathematique 334 (6) (2002) 495–500.

List of Figures

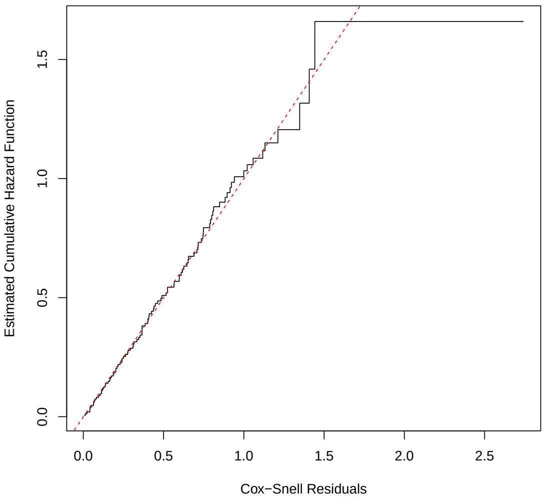
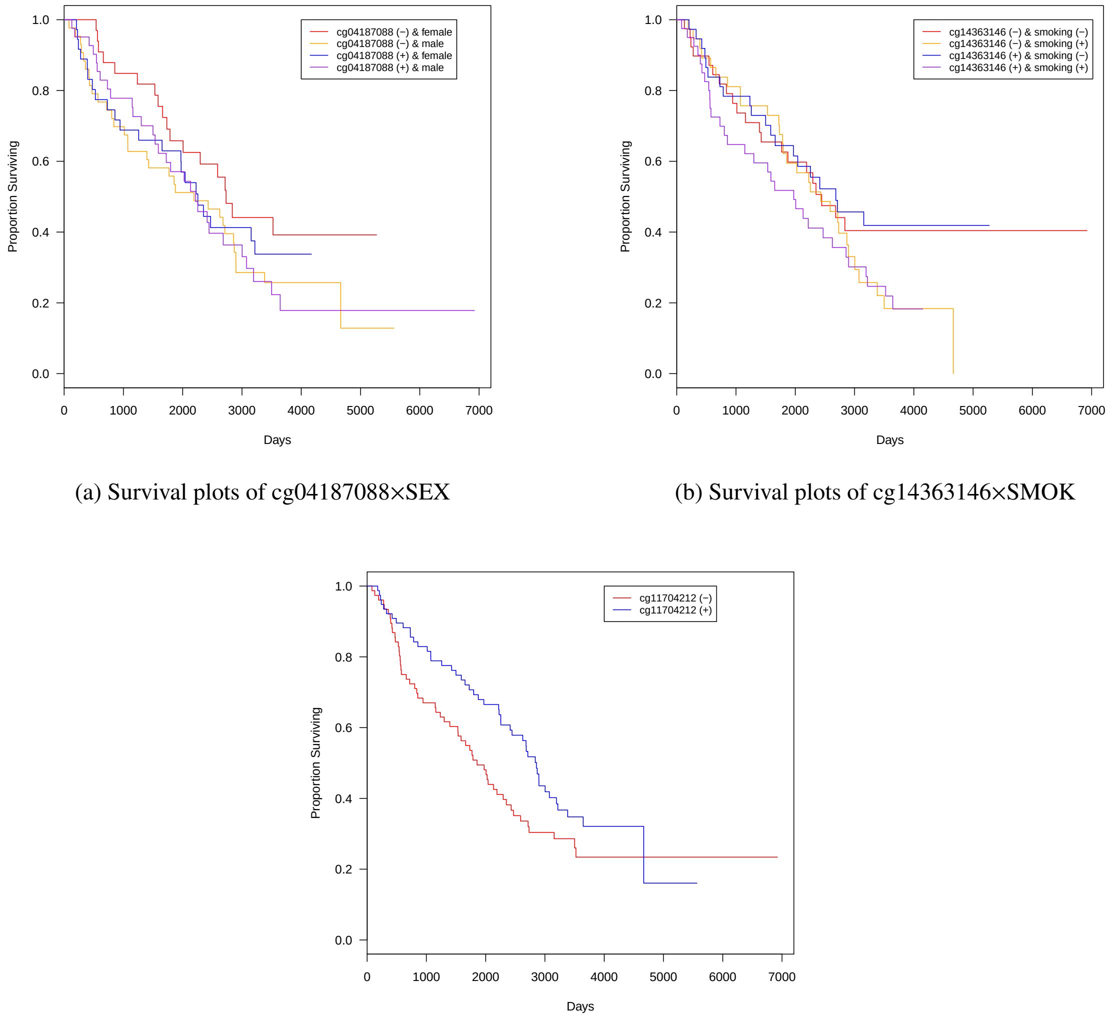


Figure 1: Cox-Snell residual plot.



(a) Survival plots of cg04187088×SEX

(b) Survival plots of cg14363146×SMOK

(c) Survival plots of the low and high level of cg11704212

Figure 2: Kaplan-Meier plots to illustrate the main and interaction effects identified by the proposed method.

List of Tables

Table 1: Comparisons of methods under mild censoring

Example	Method	$(n, p) = (200, 1000)$			$(n, p) = (400, 1000)$		
		PIT	TP	FP	PIT	TP	FP
1 ($p_0 = 4$)	FR	3.65 (0.52)	3.61 (0.49)	0.05 (0.21)	3.99 (0.18)	3.98 (0.14)	0.01 (0.12)
	FR+Lasso	3.65 (0.52)	3.02 (0.15)	0.63 (0.48)	3.99 (0.18)	3.01 (0.12)	0.98 (0.14)
	FR+MCP	3.65 (0.52)	3.02 (0.15)	0.63 (0.48)	3.99 (0.18)	3.01 (0.12)	0.98 (0.14)
	FR+SCAD	3.65 (0.52)	3.02 (0.15)	0.63 (0.48)	3.99 (0.18)	3.01 (0.12)	0.98 (0.14)
	FR$_{x_1}$	3.65 (0.52)	3.61 (0.49)	0.05 (0.21)	3.99 (0.18)	3.98 (0.14)	0.01 (0.12)
	FR$_{x_{10}}$	4.64 (0.52)	3.60 (0.49)	1.04 (0.20)	4.99 (0.19)	3.98 (0.15)	1.01 (0.12)
	FR(η_1)	5.04 (1.41)	3.89 (0.32)	1.16 (1.38)	4.53 (0.85)	4.00 (0.00)	0.53 (0.85)
	FR(η_2)	3.95 (0.62)	3.77 (0.42)	0.18 (0.45)	4.08 (0.33)	3.99 (0.09)	0.09 (0.31)
	PSIS	38.00 (0.00)	3.03 (0.55)	34.97 (0.55)	67.00 (0.00)	3.39 (0.49)	63.61 (0.49)
	PSIS+Lasso	4.44 (2.71)	2.43 (0.93)	2.01 (2.12)	6.82 (2.70)	3.13 (0.48)	3.69 (2.64)
	PSIS+MCP	4.05 (2.14)	2.47 (0.85)	1.58 (1.67)	4.81 (1.60)	3.08 (0.47)	1.74 (1.55)
	PSIS+SCAD	4.22 (2.55)	2.36 (0.89)	1.86 (1.98)	5.33 (1.96)	3.06 (0.53)	2.27 (1.85)
	CS	38.00 (0.00)	3.08 (0.48)	34.92 (0.48)	67.00 (0.00)	3.23 (0.42)	63.77 (0.42)
	CS+Lasso	4.63 (2.49)	2.51 (0.77)	2.12 (2.09)	6.50 (2.45)	2.95 (0.43)	3.54 (2.35)
	CS+MCP	4.00 (1.86)	2.44 (0.76)	1.55 (1.47)	4.78 (1.40)	2.92 (0.37)	1.86 (1.33)
	CS+SCAD	4.30 (2.23)	2.37 (0.75)	1.93 (1.81)	5.39 (1.82)	2.90 (0.44)	2.48 (1.67)
2 ($p_0 = 6$)	FR	0.91 (0.30)	5.89 (0.40)	1.34 (1.40)	1.00 (0.00)	6.00 (0.00)	1.01 (1.20)
	FR	6.16 (0.47)	6.00 (0.00)	0.16 (0.47)	6.02 (0.15)	6.00 (0.00)	0.02 (0.15)
	FR+Lasso	6.16 (0.47)	6.00 (0.00)	0.16 (0.47)	6.02 (0.15)	6.00 (0.00)	0.02 (0.15)
	FR+MCP	6.16 (0.47)	6.00 (0.00)	0.16 (0.47)	6.02 (0.15)	6.00 (0.00)	0.02 (0.15)
	FR+SCAD	6.08 (0.35)	5.91 (0.32)	0.16 (0.47)	6.02 (0.15)	6.00 (0.00)	0.02 (0.15)
	FR$_{x_1}$	6.08 (0.33)	6.00 (0.00)	0.08 (0.33)	6.02 (0.15)	6.00 (0.00)	0.02 (0.15)
	FR$_{x_{10}}$	7.13 (0.42)	6.00 (0.00)	1.13 (0.42)	7.01 (0.13)	6.00 (0.00)	1.01 (0.13)
	FR(η_1)	7.51 (1.78)	6.00 (0.00)	1.51 (1.78)	6.78 (1.16)	6.00 (0.00)	0.78 (1.16)
	FR(η_2)	6.34 (0.72)	6.00 (0.00)	0.34 (0.72)	6.11 (0.40)	6.00 (0.00)	0.11 (0.40)
	PSIS	38.00 (0.00)	3.42 (0.93)	34.58 (0.93)	67.00 (0.00)	4.83 (0.53)	62.17 (0.53)
	PSIS+Lasso	8.54 (4.97)	3.71 (0.97)	4.84 (4.66)	8.74 (5.33)	4.82 (0.54)	3.92 (5.21)
	PSIS+MCP	5.75 (2.39)	3.34 (1.01)	2.42 (1.97)	9.21 (3.05)	4.80 (0.61)	4.41 (2.77)
	PSIS+SCAD	6.73 (3.29)	3.51 (0.95)	3.21 (3.08)	8.80 (4.80)	4.77 (0.63)	4.03 (4.56)
	CS	38.00 (0.00)	4.45 (0.83)	33.55 (0.83)	67.00 (0.00)	5.66 (0.53)	61.34 (0.53)
	CS+Lasso	8.68 (4.49)	4.55 (0.93)	4.13 (4.07)	11.60 (4.50)	5.67 (0.53)	5.93 (4.22)
	CS+MCP	6.41 (2.38)	4.27 (1.09)	2.14 (1.97)	7.44 (2.20)	5.64 (0.63)	1.80 (2.46)
	CS+SCAD	6.68 (2.92)	4.34 (0.98)	2.34 (2.69)	6.83 (2.54)	5.63 (0.63)	1.20 (2.71)
3 ($p_0 = 6$)	FR	5.24 (0.95)	5.02 (1.03)	0.22 (0.49)	5.75 (0.45)	5.74 (0.44)	0.01 (0.12)
	FR+Lasso	5.24 (0.95)	5.19 (0.87)	0.05 (0.25)	5.75 (0.45)	5.74 (0.44)	0.01 (0.09)
	FR+MCP	5.24 (0.95)	5.19 (0.87)	0.05 (0.25)	5.75 (0.45)	5.74 (0.44)	0.01 (0.09)
	FR+SCAD	5.21 (0.94)	5.16 (0.87)	0.05 (0.25)	5.75 (0.45)	5.74 (0.44)	0.01 (0.09)
	FR$_{x_1}$	5.25 (0.85)	5.10 (0.84)	0.16 (0.40)	5.75 (0.45)	5.74 (0.44)	0.01 (0.12)
	FR$_{x_{10}}$	6.05 (1.28)	4.87 (1.28)	1.17 (0.41)	6.75 (0.45)	5.73 (0.44)	1.01 (0.12)
	FR(η_1)	6.84 (1.87)	5.38 (0.93)	1.45 (1.70)	6.43 (0.87)	5.93 (0.26)	0.50 (0.84)
	FR(η_2)	5.60 (1.09)	5.18 (1.00)	0.41 (0.73)	5.93 (0.48)	5.85 (0.36)	0.07 (0.33)
	PSIS	38.00 (0.00)	2.57 (0.72)	35.43 (0.72)	67.00 (0.00)	3.35 (0.58)	63.65 (0.58)
	PSIS+Lasso	3.43 (2.69)	2.06 (1.05)	1.37 (1.96)	6.18 (2.77)	3.22 (0.56)	2.96 (2.59)
	PSIS+MCP	3.04 (1.64)	2.21 (0.90)	0.83 (1.16)	3.51 (0.95)	3.11 (0.49)	0.40 (0.79)
	PSIS+SCAD	3.64 (2.24)	2.21 (0.99)	1.43 (1.61)	3.91 (1.18)	3.13 (0.49)	0.78 (1.04)
	CS	38.00 (0.00)	3.14 (0.68)	34.86 (0.68)	67.00 (0.00)	4.17 (0.62)	62.83 (0.62)
	CS+Lasso	4.74 (3.02)	2.67 (1.18)	2.07 (2.21)	7.97 (3.03)	4.05 (0.65)	3.93 (2.77)
	CS+MCP	3.76 (1.67)	2.84 (0.80)	0.93 (1.39)	4.31 (0.97)	3.94 (0.58)	0.38 (0.79)
	CS+SCAD	4.62 (2.16)	2.86 (0.95)	1.76 (1.69)	4.66 (1.08)	3.95 (0.58)	0.71 (0.98)

NOTE: FR, forward regression; PSIS, the principled sure independence screening; CS, the conditional screening with the given conditioning variable; PIT, estimated probability of including all true predictors in the selected predictors; TP, average number of true positives; FP, average number of false positives; p_0 denotes the number of true signals; numbers in the parentheses are standard deviations. We used $\eta_1 = .5$ and $\eta_2 = 1 - \ln d/(3 \ln p)$. When it is not noted, $\eta = 1$ was used. FR$_{S_0}$ denotes that FR was performed with the initial set of S_0. We considered two initial sets x_1 and x_{10}.

Table 2: Comparisons of methods under heavy censoring

Example	Method	$(n,p) = (200, 1000)$			$(n,p) = (400, 1000)$		
		PIT	TP	FP	PIT	TP	FP
1 ($p_0 = 4$)	FR	3.33 (0.54)	3.29 (0.49)	0.04 (0.22)	3.86 (0.39)	3.85 (0.36)	0.01 (0.13)
	FR+Lasso	3.33 (0.54)	3.00 (0.16)	0.33 (0.48)	3.86 (0.39)	3.01 (0.10)	0.85 (0.36)
	FR+MCP	3.33 (0.54)	3.00 (0.16)	0.33 (0.48)	3.86 (0.39)	3.01 (0.10)	0.85 (0.36)
	FR+SCAD	3.33 (0.54)	3.00 (0.16)	0.33 (0.48)	3.86 (0.39)	3.01 (0.10)	0.85 (0.36)
	FR$_{x_1}$	3.33 (0.54)	3.29 (0.49)	0.04 (0.22)	3.86 (0.39)	3.85 (0.36)	0.01 (0.13)
	FR$_{x_{10}}$	4.35 (0.55)	3.29 (0.49)	1.05 (0.26)	4.85 (0.39)	3.84 (0.37)	1.01 (0.13)
	FR(η_1)	5.24 (1.98)	3.62 (0.49)	1.62 (1.88)	4.71 (1.04)	3.98 (0.15)	0.74 (1.02)
	FR(η_2)	3.64 (0.72)	3.45 (0.51)	0.19 (0.49)	4.02 (0.42)	3.93 (0.26)	0.09 (0.33)
	PSIS	38.00 (0.00)	2.93 (0.59)	35.07 (0.59)	67.00 (0.00)	3.37 (0.49)	63.63 (0.49)
	PSIS+Lasso	3.58 (2.57)	2.02 (0.92)	1.55 (1.98)	5.92 (2.46)	2.91 (0.58)	3.01 (2.26)
	PSIS+MCP	3.56 (2.23)	2.13 (0.87)	1.43 (1.74)	4.70 (1.60)	2.92 (0.53)	1.79 (1.49)
	PSIS+SCAD	3.66 (2.48)	2.01 (0.86)	1.65 (1.96)	5.22 (2.17)	2.85 (0.64)	2.37 (1.92)
	CS	38.00 (0.00)	2.90 (0.55)	35.10 (0.55)	67.00 (0.00)	3.27 (0.45)	63.73 (0.45)
	CS+Lasso	3.60 (2.29)	2.09 (0.83)	1.52 (1.81)	5.69 (2.51)	2.80 (0.56)	2.89 (2.30)
	CS+MCP	3.36 (1.91)	2.07 (0.79)	1.28 (1.50)	4.70 (1.66)	2.83 (0.47)	1.87 (1.51)
	CS+SCAD	3.67 (2.19)	2.04 (0.78)	1.63 (1.74)	5.21 (2.29)	2.74 (0.56)	2.48 (2.03)
2 ($p_0 = 6$)	FR	6.22 (0.65)	5.85 (0.61)	0.36 (0.74)	6.03 (0.18)	6.00 (0.00)	0.03 (0.18)
	FR+Lasso	6.21 (0.65)	5.95 (0.23)	0.26 (0.60)	6.03 (0.18)	6.00 (0.00)	0.03 (0.18)
	FR+MCP	6.21 (0.65)	5.95 (0.23)	0.26 (0.60)	6.03 (0.18)	6.00 (0.00)	0.03 (0.18)
	FR+SCAD	6.07 (0.5)	5.81 (0.45)	0.26 (0.60)	6.02 (0.17)	6.00 (0.06)	0.03 (0.18)
	FR$_{x_1}$	6.10 (0.45)	5.94 (0.33)	0.16 (0.46)	6.02 (0.17)	6.00 (0.00)	0.02 (0.17)
	FR$_{x_{10}}$	7.14 (0.58)	5.88 (0.49)	1.25 (0.60)	7.03 (0.17)	6.00 (0.00)	1.03 (0.17)
	FR(η_1)	8.36 (2.55)	5.97 (0.27)	2.39 (2.56)	6.98 (1.31)	6.00 (0.00)	0.98 (1.31)
	FR(η_2)	6.5 (0.90)	5.93 (0.44)	0.57 (0.95)	6.14 (0.43)	6.00 (0.00)	0.14 (0.43)
	PSIS	38.00 (0.00)	3.13 (0.94)	34.87 (0.94)	67.00 (0.00)	4.61 (0.67)	62.39 (0.67)
	PSIS+Lasso	8.17 (4.84)	3.38 (1.05)	4.80 (4.46)	8.98 (5.70)	4.62 (0.70)	4.36 (5.48)
	PSIS+MCP	5.34 (2.38)	2.97 (1.09)	2.37 (1.95)	7.82 (2.80)	4.47 (0.82)	3.35 (2.41)
	PSIS+SCAD	6.52 (3.31)	3.17 (1.06)	3.36 (3.02)	7.45 (4.02)	4.48 (0.77)	2.97 (3.73)
	CS	38.00 (0.00)	4.04 (0.87)	33.96 (0.87)	67.00 (0.00)	5.42 (0.64)	61.58 (0.64)
	CS+Lasso	8.41 (4.48)	4.13 (0.99)	4.29 (4.07)	10.64 (4.83)	5.41 (0.66)	5.23 (4.59)
	CS+MCP	5.81 (2.35)	3.68 (1.13)	2.12 (1.84)	7.53 (2.29)	5.33 (0.78)	2.20 (2.44)
	CS+SCAD	6.75 (3.19)	3.88 (1.07)	2.87 (2.81)	7.13 (3.07)	5.32 (0.79)	1.81 (3.17)
3 ($p_0 = 6$)	FR	4.45 (1.51)	4.08 (1.64)	0.37 (0.61)	5.49 (0.54)	5.46 (0.50)	0.03 (0.21)
	FR+Lasso	4.45 (1.51)	4.40 (1.44)	0.05 (0.24)	5.49 (0.54)	5.48 (0.50)	0.01 (0.16)
	FR+MCP	4.45 (1.51)	4.40 (1.44)	0.05 (0.24)	5.49 (0.54)	5.48 (0.50)	0.01 (0.16)
	FR+SCAD	4.44 (1.50)	4.39 (1.43)	0.05 (0.24)	5.49 (0.54)	5.48 (0.50)	0.01 (0.16)
	FR$_{x_1}$	4.57 (1.39)	4.30 (1.44)	0.26 (0.50)	5.49 (0.54)	5.46 (0.50)	0.03 (0.21)
	FR$_{x_{10}}$	4.97 (1.91)	3.72 (1.88)	1.24 (0.48)	6.48 (0.53)	5.45 (0.50)	1.03 (0.21)
	FR(η_1)	6.99 (2.59)	4.72 (1.40)	2.27 (2.29)	6.55 (1.18)	5.76 (0.43)	0.80 (1.09)
	FR(η_2)	4.93 (1.62)	4.35 (1.56)	0.58 (0.84)	5.74 (0.63)	5.60 (0.49)	0.14 (0.42)
	PSIS	38.00 (0.00)	2.32 (0.76)	35.68 (0.76)	67.00 (0.00)	3.20 (0.61)	63.8 (0.61)
	PSIS+Lasso	2.17 (2.01)	1.46 (1.01)	0.71 (1.28)	5.24 (2.50)	2.94 (0.73)	2.30 (2.17)
	PSIS+MCP	2.66 (1.86)	1.81 (0.98)	0.86 (1.25)	3.44 (1.16)	2.88 (0.54)	0.56 (1.02)
	PSIS+SCAD	2.78 (2.35)	1.66 (1.08)	1.11 (1.55)	4.23 (1.53)	2.92 (0.59)	1.31 (1.37)
	CS	38.00 (0.00)	2.87 (0.71)	35.13 (0.71)	67.00 (0.00)	3.88 (0.69)	63.12 (0.69)
	CS+Lasso	3.34 (2.81)	2.03 (1.17)	1.31 (1.92)	7.16 (3.05)	3.71 (0.75)	3.45 (2.71)
	CS+MCP	3.45 (1.88)	2.44 (0.98)	1.00 (1.34)	4.19 (1.15)	3.62 (0.64)	0.57 (0.92)
	CS+SCAD	4.13 (2.60)	2.42 (1.14)	1.71 (1.82)	4.80 (1.36)	3.66 (0.66)	1.14 (1.24)

Table 3: Comparisons of methods under covariate-dependent censoring

Example	Method	$(n, p) = (200, 1000)$			$(n, p) = (400, 1000)$		
		PIT	TP	FP	PIT	TP	FP
1^* ($p_0 = 4$)	FR	3.58 (0.51)	3.55 (0.50)	0.03 (0.17)	3.99 (0.15)	3.98 (0.13)	0.00 (0.06)
	FR+Lasso	3.58 (0.51)	3.01 (0.09)	0.57 (0.50)	3.99 (0.15)	3.00 (0.06)	0.98 (0.13)
	FR+MCP	3.58 (0.51)	3.01 (0.09)	0.57 (0.50)	3.99 (0.15)	3.00 (0.06)	0.98 (0.13)
	FR+SCAD	3.58 (0.51)	3.01 (0.09)	0.57 (0.50)	3.99 (0.15)	3.00 (0.06)	0.98 (0.13)
	FR_{x_1}	3.58 (0.51)	3.55 (0.50)	0.03 (0.17)	3.99 (0.15)	3.98 (0.13)	0.00 (0.06)
	$\text{FR}_{x_{10}}$	4.58 (0.51)	3.56 (0.50)	1.03 (0.17)	4.98 (0.14)	3.98 (0.13)	1.00 (0.04)
	FR(η_1)	5.05 (1.56)	3.84 (0.37)	1.21 (1.53)	4.56 (0.85)	4.00 (0.00)	0.56 (0.85)
	FR(η_2)	3.89 (0.64)	3.70 (0.46)	0.19 (0.48)	4.08 (0.32)	4.00 (0.06)	0.09 (0.31)
	PSIS	38.00 (0.00)	2.96 (0.51)	35.04 (0.51)	67.00 (0.00)	3.41 (0.50)	63.59 (0.50)
	PSIS+Lasso	4.39 (2.82)	2.29 (0.96)	2.11 (2.24)	6.58 (2.63)	3.06 (0.59)	3.53 (2.53)
	PSIS+MCP	4.00 (2.15)	2.37 (0.90)	1.63 (1.70)	4.81 (1.60)	3.08 (0.54)	1.73 (1.55)
	PSIS+SCAD	4.30 (2.54)	2.27 (0.91)	2.03 (2.01)	5.34 (1.91)	3.04 (0.59)	2.30 (1.83)
	CS	38.00 (0.00)	3.00 (0.49)	35.00 (0.49)	67.00 (0.00)	3.25 (0.43)	63.75 (0.43)
	CS+Lasso	4.50 (2.67)	2.36 (0.77)	2.14 (2.28)	6.50 (2.44)	2.92 (0.41)	3.58 (2.33)
	CS+MCP	4.01 (2.00)	2.38 (0.75)	1.63 (1.65)	4.72 (1.45)	2.91 (0.35)	1.81 (1.38)
	CS+SCAD	4.27 (2.39)	2.29 (0.76)	1.98 (1.96)	5.45 (1.87)	2.89 (0.39)	2.57 (1.77)
2^* ($p_0 = 6$)	FR	6.13 (0.39)	6.00 (0.04)	0.13 (0.39)	6.01 (0.09)	6.00 (0.00)	0.01 (0.09)
	FR+Lasso	6.12 (0.38)	5.99 (0.09)	0.13 (0.39)	6.01 (0.09)	6.00 (0.00)	0.01 (0.09)
	FR+MCP	6.12 (0.38)	5.99 (0.09)	0.13 (0.39)	6.01 (0.09)	6.00 (0.00)	0.01 (0.09)
	FR+SCAD	6.04 (0.25)	5.91 (0.30)	0.13 (0.39)	6.01 (0.09)	6.00 (0.00)	0.01 (0.09)
	FR_{x_1}	6.06 (0.26)	6.00 (0.04)	0.06 (0.25)	6.01 (0.09)	6.00 (0.00)	0.01 (0.09)
	$\text{FR}_{x_{10}}$	7.09 (0.32)	6.00 (0.04)	1.09 (0.31)	7.01 (0.11)	6.00 (0.00)	1.01 (0.11)
	FR(η_1)	7.55 (1.72)	6.00 (0.00)	1.55 (1.72)	6.68 (1.05)	6.00 (0.00)	0.68 (1.05)
	FR(η_2)	6.32 (0.64)	6.00 (0.00)	0.32 (0.64)	6.11 (0.39)	6.00 (0.00)	0.11 (0.39)
	PSIS	38.00 (0.00)	3.44 (0.92)	34.56 (0.92)	67.00 (0.00)	4.80 (0.59)	62.20 (0.59)
	PSIS+Lasso	8.55 (5.16)	3.63 (1.00)	4.92 (4.81)	9.14 (5.81)	4.79 (0.58)	4.34 (5.65)
	PSIS+MCP	5.50 (2.47)	3.25 (1.04)	2.24 (2.05)	9.01 (2.91)	4.76 (0.67)	4.25 (2.63)
	PSIS+SCAD	6.27 (3.19)	3.42 (0.99)	2.86 (2.98)	8.67 (4.77)	4.73 (0.68)	3.95 (4.50)
	CS	38.00 (0.00)	4.40 (0.89)	33.60 (0.89)	67.00 (0.00)	5.66 (0.53)	61.34 (0.53)
	CS+Lasso	8.74 (4.59)	4.47 (0.97)	4.27 (4.22)	11.42 (4.57)	5.66 (0.55)	5.76 (4.32)
	CS+MCP	6.19 (2.39)	4.12 (1.15)	2.06 (1.95)	7.30 (2.11)	5.62 (0.62)	1.67 (2.37)
	CS+SCAD	6.54 (2.98)	4.22 (1.08)	2.33 (2.76)	6.91 (2.81)	5.61 (0.66)	1.30 (2.98)
3^* ($p_0 = 6$)	FR	5.11 (1.02)	4.87 (1.13)	0.24 (0.51)	5.72 (0.47)	5.70 (0.46)	0.02 (0.13)
	FR+Lasso	5.11 (1.02)	5.06 (0.96)	0.04 (0.22)	5.72 (0.47)	5.71 (0.45)	0.01 (0.10)
	FR+MCP	5.11 (1.02)	5.06 (0.96)	0.04 (0.22)	5.72 (0.47)	5.71 (0.45)	0.01 (0.10)
	FR+SCAD	5.08 (1.01)	5.03 (0.94)	0.04 (0.22)	5.72 (0.47)	5.71 (0.45)	0.01 (0.10)
	FR_{x_1}	5.11 (0.98)	4.92 (1.00)	0.19 (0.46)	5.72 (0.47)	5.70 (0.46)	0.02 (0.13)
	$\text{FR}_{x_{10}}$	5.74 (1.49)	4.59 (1.50)	1.15 (0.38)	6.72 (0.49)	5.70 (0.46)	1.03 (0.17)
	FR(η_1)	7.11 (1.96)	5.38 (0.87)	1.73 (1.84)	6.60 (1.00)	5.91 (0.28)	0.69 (0.96)
	FR(η_2)	5.51 (1.09)	5.07 (1.02)	0.44 (0.73)	5.97 (0.54)	5.83 (0.37)	0.14 (0.39)
	PSIS	38.00 (0.00)	2.55 (0.72)	35.45 (0.72)	67.00 (0.00)	3.34 (0.63)	63.66 (0.63)
	PSIS+Lasso	3.08 (2.58)	1.85 (1.10)	1.23 (1.79)	5.93 (2.76)	3.19 (0.70)	2.74 (2.46)
	PSIS+MCP	3.10 (1.84)	2.10 (0.96)	1.00 (1.37)	3.48 (1.06)	3.05 (0.57)	0.43 (0.87)
	PSIS+SCAD	3.38 (2.40)	2.02 (1.11)	1.36 (1.67)	4.12 (1.48)	3.11 (0.58)	1.01 (1.37)
	CS	38.00 (0.00)	3.12 (0.70)	34.88 (0.70)	67.00 (0.00)	4.10 (0.68)	62.90 (0.68)
	CS+Lasso	4.47 (3.28)	2.53 (1.21)	1.94 (2.42)	7.88 (3.17)	3.96 (0.69)	3.92 (2.89)
	CS+MCP	3.97 (1.99)	2.77 (0.89)	1.20 (1.62)	4.35 (1.07)	3.86 (0.62)	0.50 (0.87)
	CS+SCAD	4.60 (2.57)	2.72 (1.05)	1.87 (1.97)	4.83 (1.28)	3.87 (0.63)	0.96 (1.13)

Table 4: The number of overlapped genes chosen by PSIS, CS, SII and FR

	PSIS	CS	SII	FR
PSIS	25	1	3	0
CS	1	25	0	0
SII	3	0	25	0
FR	0	0	0	2