Journal of Multivariate Analysis 167 (2018) 79–96

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On the sign consistency of the Lasso for the high-dimensional Cox model

Shaogao Lv [a], Mengying You [a], Huazhen Lin [a,*], Heng Lian [b], Jian Huang [c,d]

[a] Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, 611130, China
[b] Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong
[c] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
[d] Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA, USA

ARTICLE INFO

Article history:
Received 14 January 2018
Available online 22 April 2018

AMS subject classifications:
62N02

Keywords:
Cox proportional
Empirical process
Hazard model
Lasso
Mutual coherence
Oracle property
Sparse recovery

ABSTRACT

In this paper we study the ℓ_1-penalized partial likelihood estimator for the sparse high-dimensional Cox proportional hazards model. In particular, we investigate how the ℓ_1-penalized partial likelihood estimation recovers the sparsity pattern and the conditions under which the sign support consistency is guaranteed. We establish sign recovery consistency and ℓ_∞-error bounds for the Lasso partial likelihood estimator under suitable and interpretable conditions, including mutual incoherence conditions. More importantly, we show that the conditions of the incoherence and bounds on the minimal non-zero coefficients are necessary, which provides significant and instructional implications for understanding the Lasso for the Cox model. Numerical studies are presented to illustrate the theoretical results.

1. Introduction

High-dimensional data, including high-throughput genomic data, are becoming increasingly available as data collection technology advances. Finding significant genetic factors for clinical outcomes based on high-throughput and high-dimensional genomic data, such as age of disease onset or survival time after treatment, is fundamental to clinical medicine. In view of the central role of the Cox model [5] in survival analysis, its widespread use and the proliferation of high-dimensional covariate data, it is important to study its properties under high-dimensional settings.

When the number p of features far exceeds the sample size, many standard approaches, such as likelihood and partial likelihood methods are not applicable. To handle high-dimensional problems, various regularization methods have been developed for sparse modeling and variable selection. In particular, the Lasso or the ℓ_1 regularization methods [4,23] have proven to be powerful for high-dimensional problems. The convexity of Lasso-type methods makes the implementation efficient and facilitates theoretical analysis.

Tibshirani [24] proposed using the Lasso for estimation and variable selection in the Cox model. Recently, several authors have studied the Lasso in the Cox model under sparse, high-dimensional settings. For example, Kong and Nan [14] considered prediction and ℓ_1 estimation error bounds for the Lasso in Cox models. Lemler [17] considered joint estimation of the baseline hazard function and regression coefficients in the Cox model, and established theoretical guarantees for the prediction performance and error bounds of the Lasso estimator under very weak conditions and some incoherent conditions,

* Corresponding author.
E-mail address: linhz@swufe.edu.cn (H. Lin).

https://doi.org/10.1016/j.jmva.2018.04.005
0047-259X/

respectively. Under a natural extension of the compatibility and the cone invertibility conditions of the Hessian matrix, Huang et al. [12] established ℓ_q estimation error bounds of the Lasso estimator for the Cox model with time-dependent covariates when $q \geq 1$.

The aforementioned studies mainly focused on the estimation properties of the Lasso in the Cox model, but there has not been any systematic analysis of the selection properties. Particularly, under sparsity conditions, what are the sufficient and necessary condition for the Lasso to be able to recover the positions of its non-zero entries, or to guarantee the sign consistency of estimator? This problem, also known as support recovery or model selection consistency, arises in a variety of contexts, including compressed sensing [8], sparse approximation [7] and structure estimation in graphical models [21]. The estimation of sparsity patterns is a central concern in high-dimensional analysis, and there has been considerable effort devoted to the linear regression model setting; see, e.g., [21,27–29].

In the context of high-dimensional survival analysis, Lin and Lv [18] considered the additive hazard model in a high-dimensional setting, and they investigated the sparse pattern recovery and estimation problems based on a class of general penalties, including the Lasso. There is an important difference between the additive hazards model and the Cox model. The former leads to a least squares-type loss function and hence the key quantity (the Hessian matrix) for statistical inference is independent of the estimated coefficients. By contrast, the Hessian matrix involved in the Cox model is a function of the regression coefficient, which is more involved and needs new technical tools.

For the high-dimensional Cox model, Bradic et al. [3] considered estimation as well as variable selection and oracle properties using general concave penalties. Although their work provided important insights into the properties of various regularization methods in the Cox model, they required that the sample Hessian matrix $\widehat{\Sigma}$ converges uniformly to its population version Σ with respect to the Frobenius norm $\|\cdot\|_2$ over a neighborhood $\mathcal{B}$ of the true parameter value in high-dimensional settings; see Condition 2 (i) of [3]. This is a strong assumption. First, requiring uniform convergence in a neighborhood $\mathcal{B}$ of the true parameter value is very demanding in high-dimensional settings. Second, even for the much simpler case of standard sample covariance matrix when the covariates are independent Gaussian random variables, Condition 2 (i) of [3] is highly nontrivial in $p > n$ settings [27].

Also, all of the above mentioned work only provided sufficient conditions to guarantee oracle properties. It is not clear whether these conditions are essential. Since most of the conditions for high-dimensional inference are often hard to verify, establishing necessary and sufficient conditions for statistical inference has significant and instructional implications for understanding the proposed methods.

The problem of sparse pattern recovery is difficult in the Cox model for counting processes data, due to the dynamic nature of the data, the complexity introduced by censoring, and the fact that the Hessian matrix of the partial likelihood is non-quadratic and depends on the model parameter. This paper aims at establishing a complete characterization for the selection consistency of the Lasso in the Cox model with counting process data and time-dependent covariates. We provide sufficient conditions for sign consistency of the Lasso penalized partial likelihood estimator. Two critical conditions to guarantee this are the *mutual incoherence condition* [29] and the *minimum value of the coefficient* condition. Furthermore, we show that the sign consistency will fail with probability at least 1/2 if either of the above two conditions does not hold. To the best of our knowledge, there is no work on this topic in the context of the Cox model for counting process data.

The rest of this paper is organized as follows. In Section 2, we review the Cox proportional hazard model and then introduce the penalized partial likelihood with the ℓ_1 penalty. The theoretical properties of the Lasso Cox estimator are studied in Section 3, while some numerical examples are presented to back up the obtained theory in Section 4. In Section 5 we give an outline of proof of theoretical guarantee of performance of the method. Most of the proof and technical details are relegated to a series of Appendices.

Here is some standard notation used throughout the paper. For vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, we use the usual inner product of $\mathbb{R}^p$, given by $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle = \boldsymbol{\alpha}^\top \boldsymbol{\beta}$, and $\|\boldsymbol{\alpha}\|_2 = \sqrt{\langle \boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle}$ is denoted to be its ℓ_2-norm. Similarly the ℓ_1-norm is given by $\|\boldsymbol{\alpha}\|_1 = |\alpha_1| + \cdots + |\alpha_p|$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top$. For some subset $A \subseteq \{1, \ldots, p\}$, we denote by $\|\boldsymbol{\alpha}_A\|_\infty = \max_{j \in A} |\alpha_j|$. For a vector $z \in \mathbb{R}^p$ and a subset $S \subseteq \{1, \ldots, p\}$, we use $z_S \in \mathbb{R}^S$ to denote the vector z restricted to S. Given sequences $f(n)$ and $g(n)$, the notion $f(n) = O_p\{g(n)\}$ means that there exists a constant c such that $f(n) \leq cg(n)$; Similarly, $f(n) = o_p\{g(n)\}$ means that $f(n)/g(n) \to 0$ as $n \to \infty$, and sometimes we write it $f(n) \ll g(n)$. For a matrix $M = (M_{ij})$, we define the spectral norm, given by $\|M\|_2 = \max_{\|x\|_2 = 1} \|Mx\|_2$. Also we write $\|M\|_\infty = \max_i \sum_j |M_{ij}|$.

2. The Cox model with Lasso

Suppose a random process of n individuals is chosen, i.e., consider an n-dimensional counting process $\mathbf{N}^{(n)}(t) = (N_1(t), \ldots, N_n(t))^\top$ with $t \geq 0$, where $N_i(t)$ is the number of observed events in the time interval $[0, t]$ for the ith individual. The sample paths of $N_1, \ldots, N_n$ are the standard Poisson processes adding size 1 at each jump. For $t \geq 0$, let $\mathcal{F}_t$ be the σ-filtration representing all the information available up to time t. In the study of the dependence of survival time τ on covariates $\mathbf{Z}_i(t) = (Z_{i1}(t), \ldots, Z_{ip}(t))^\top$, one assumes that $\mathbf{N}^{(n)}$ has predictable compensator $\boldsymbol{\Lambda}^{(n)} = (\Lambda_1, \ldots, \Lambda_n)^\top$ with respect to $\{\mathcal{F}_t : t \geq 0\}$ such that

$$d\Lambda_i(t) = Y_i(t) \exp\{\mathbf{Z}_i(t)^\top \boldsymbol{\beta}^0\} d\Lambda_0(t), \tag{1}$$

where $\boldsymbol{\beta}^0$ is the true regression coefficient associated with the p-dimensional covariates, and Λ_0 is an unknown baseline cumulative hazard function. For each $i \in \{1, \ldots, n\}$, $Y_i(t) \in \{0, 1\}$ is a predictable at risk indicator process that can be

constructed from data. In this specific setting, we can use the natural filtration of the processes $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), \mathbf{Z}_i(s) : s \le t, i \in \{1, \ldots, n\}\}$ or a larger one. In a high-dimensional setting, a standard assumption is that the vector $\boldsymbol{\beta}^0$ is sparse, in the sense that the cardinality $s_0 = |S(\boldsymbol{\beta}^0)|$ satisfies $s_0 \ll p$, where $S(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$.

Define the logarithm of the Cox partial likelihood at time t by

$$C(\boldsymbol{\beta}; t) = \sum_{i=1}^{n} \int_0^t \mathbf{Z}_i(s)^\top \boldsymbol{\beta} dN_i(s) - \int_0^t \ln \left\{ \sum_{i=1}^{n} Y_i(s) e^{\mathbf{Z}_i(s)^\top \boldsymbol{\beta}} \right\} d\bar{N}(s),$$

where $\bar{N} = N_1 + \cdots + N_n$. The Lasso program for the Cox model is to minimize a ℓ_1-penalized negative log-partial likelihood criterion, given by

$$\ell(\boldsymbol{\beta}, \lambda) = \ell(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\ell(\boldsymbol{\beta}) = -C(\boldsymbol{\beta}; \tau)/n$, λ is a penalty parameter and $0 < \tau < \infty$ is the end of the study. Since the minimization problem for (2) is a convex program, a global solution of the Lasso program (2) always exists, and we denote by $\hat{\boldsymbol{\beta}}$ any one of them, viz.

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda) \in \arg\min_{\boldsymbol{\beta}} \{\ell(\boldsymbol{\beta}, \lambda)\}. \tag{3}$$

2.1. Additional useful notation

For a vector $\mathbf{v}$, let $\mathbf{v}^{\otimes 0} = 1 \in \mathbb{R}$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^\top$. For any vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we define, for each $k \in \{0, 1, 2\}$,

$$S^{(k)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i(t)^{\times k} Y_i(t) e^{\mathbf{Z}_i(t)^\top \boldsymbol{\beta}}, \quad \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}) = \frac{S^{(1)}(t, \boldsymbol{\beta})}{S^{(0)}(t, \boldsymbol{\beta})},$$

$$V_n(t, \boldsymbol{\beta}) = \sum_{i=1}^{n} w_{ni}(t, \boldsymbol{\beta}) \left\{ \mathbf{Z}_i(t) - \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}) \right\}^{\otimes 2} = \frac{S^{(2)}(t, \boldsymbol{\beta})}{S^{(0)}(t, \boldsymbol{\beta})} - \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta})^{\otimes 2},$$

where $w_{ni}(t, \boldsymbol{\beta}) = Y_i(t) \exp\{\mathbf{Z}_i(t)^\top \boldsymbol{\beta}\} \big/ \left\{ n S^{(0)}(t, \boldsymbol{\beta}) \right\}$. Note that $S^{(0)}$ is a scalar, $S^{(1)}$ and $\bar{\mathbf{Z}}$ are p-dimensional vectors, and $S^{(2)}$ and V_n are $p \times p$ matrices. It has been shown as in Andersen and Gill [1] that the gradient of $\ell(\boldsymbol{\beta})$ is represented as

$$\dot{\ell}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \{\mathbf{Z}_i(s) - \bar{\mathbf{Z}}_n(s, \boldsymbol{\beta})\} dN_i(s), \tag{4}$$

and the Hessian matrix of $\ell(\boldsymbol{\beta})$ is

$$\ddot{\ell}(\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau V_n(s, \boldsymbol{\beta}) dN_i(s).$$

Since $S^{(k)}$ and V_n depend on a random sample, we need to define their population counterparts for theoretical analysis. For each $k \in \{0, 1, 2\}$, let

$$\mathbf{s}^{(k)}(t, \boldsymbol{\beta}) = \mathrm{E}\{\mathbf{Z}(t)^{\times k} Y(t) e^{\mathbf{Z}(t)^\top \boldsymbol{\beta}}\}, \quad \mathbf{e}(t, \boldsymbol{\beta}) = \mathbf{s}^{(1)}(t, \boldsymbol{\beta}) / \mathbf{s}^{(0)}(t, \boldsymbol{\beta}).$$

The population counterpart of $\ddot{\ell}(\boldsymbol{\beta})$ is given by

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \mathrm{E}\left[\int_0^\tau \left\{ \frac{\mathbf{s}^{(2)}(s, \boldsymbol{\beta})}{\mathbf{s}^{(0)}(s, \boldsymbol{\beta})} - \mathbf{e}(s, \boldsymbol{\beta})^{\otimes 2} \right\} dN(s) \right].$$

The matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\beta}^0)$ characterizes the covariance structure of (1) and will play a critical role in our high-dimensional analysis.

3. Primal–dual witness technique and statistical theory

In this section, we first construct a biased oracle estimator based on the primal–dual witness (PDW) technique, developed by Wainwright [27]. Then we give two useful lemmas to describe the solutions of the Lasso program (3). In Section 3.2, we introduce some technical conditions for building analysis framework, and then establish estimation error of the biased oracle estimator. Based on these results, we state the weak oracle property of the Lasso estimator in Section 3.3, and in Section 3.4 we show that the mutual coherent condition and the minimum value of the true nonzero coefficients are essential to guarantee the corrected sign recovery.

3.1. Primal–dual witness construction

We now outline the main steps of the PDW, which we will use to establish the support recovery property for the program (3). Define the active set $S = S(\boldsymbol{\beta}_0) = \{j : \beta_j^0 \neq 0\}$, and its supplementary $S^c = \{j : \beta_j^0 = 0\}$. Without loss of generality, we assume that the last $p - s_0$ components of $\boldsymbol{\beta}^0$ are 0, so we can write $\boldsymbol{\beta}^0 = ((\boldsymbol{\beta}_1^0)^\top, \mathbf{0}^\top)^\top$. A vector $z \in \mathbb{R}^p$ is a subgradient for the ℓ_1-norm evaluated at $\boldsymbol{\beta}$, denoted $z \in \partial\|\boldsymbol{\beta}\|_1$, if the elements satisfy the following relation: $z_j = \text{sign}(\beta_j)$ if $\beta_j \neq 0$, and $z_j \in [-1, 1]$ otherwise, where $\text{sign}(\beta_j) = 1, 0$, or -1 if $\beta_j > 0, = 0$ or < 0, respectively.

The key steps of the PDW argument are as follows.

Step 1: Define a biased oracle estimator as $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_S, \mathbf{0})$, where

$$\check{\boldsymbol{\beta}}_S = \min_{\boldsymbol{\beta}_S \in \mathbb{R}^S} \{\ell(\boldsymbol{\beta}_S, \mathbf{0}) + \lambda\|\boldsymbol{\beta}_S\|_1\}. \tag{5}$$

Here we impose the additional constraint such that $S(\check{\boldsymbol{\beta}}) \subseteq S(\boldsymbol{\beta}^0) = S$. The solution to this restricted convex program (5) is guaranteed to be unique under the invertibility condition on $\ddot{\ell}_{SS}(\boldsymbol{\beta}_S, \mathbf{0})$. Note that here $\check{\boldsymbol{\beta}}$ is just a theoretical construction assuming the true sparsity structure were known, but not a real estimator.

Step 2: We choose $\check{z}_S \in \mathbb{R}^S$ as an element of the subdifferential of the $\|\cdot\|_1$ norm evaluated at $\check{\boldsymbol{\beta}}_S$. By definition of subgradient, we see that $\|\check{z}_S\|_\infty \leq 1$.

Step 3: We solve for a vector $\check{z}_{S^c} \in \mathbb{R}^{p-s_0}$ to satisfy the zero subgradient condition

$$\dot{\ell}(\check{\boldsymbol{\beta}}) + \lambda\check{z} = 0, \tag{6}$$

where $\check{z} = (\check{z}_S, \check{z}_{S^c})$. Then we check whether or not the dual feasibility condition $|\check{z}_j| \leq 1$ for all $j \in S^c$ is satisfied. To ensure uniqueness, we need to check strict dual feasibility of $\check{z}_S$, i.e., $\max_{j\in S^c}|\check{z}_j| < 1$.

Step 4: We check whether the sign consistency condition $\check{z}_S = \text{sign}(\boldsymbol{\beta}_1^o)$ is satisfied.

Note that, in high dimensions, the Lasso program (3) is not guaranteed to be strictly convex, so there may be multiple solutions generated by (3). Note that the justification of the PDW argument mainly relies on some basic knowledge of convex optimization such as Kuhn–Tucker conditions, as shown in Lemma 5 in the Appendix.

3.2. Estimation error of the oracle estimator $\check{\boldsymbol{\beta}}$

In this section, we establish the error bound for the Lasso Cox estimator $\check{\boldsymbol{\beta}}$. To this end, we need the following assumptions.

Assumption 1. (1) $\{Y_i(t), \mathbf{Z}_i(t), N_i(t) : t \geq 0\}$ with $i \in \{1, \ldots, n\}$ are iid time-dependent sample from the underlying process $\{Y(t), \mathbf{Z}(t), N(t) : t \geq 0\}$; (2) $\Pr[\max_i\{N_i(\tau)\} \leq 1] = 1$; (3) no two components of the N_is jump at the same time.

Assumption 2. (1) There exists some constant K such that

$$\forall_{q\leq i<i'\leq n} \quad \sup_{t\in[0,\tau]} \max_{j\leq p}|Z_{ij}(t) - Z_{i'j}(t)| \leq K.$$

(2) $\Lambda_0(\tau) < \infty$. (3) $\Pr\{Y(\tau) = 1\} > 0$. (4) The sample paths $Z_1, \ldots, Z_p$ are of uniformly bounded variation.

Assumption 3. $\|\boldsymbol{\Sigma}_{SS}\|_2 = O_p(1)$ and $\|(\boldsymbol{\Sigma}_{SS})^{-1}\|_2 = O_p(1)$.

Assumption 1 and parts (2) and (3) of Assumption 2 are standard for survival models [1]; part (1) of Assumption 2 controls the behavior of the covariates and such condition for linear regression has been imposed on the deterministic Gram design. Actually, our analysis still holds under a relaxed condition such as sub-Gaussian ensembles, but this will complicate our proof; part (4) of Assumption 2 is a mild technique condition that will control entropy integrals involved empirical process, which has been imposed in [18]. Assumption 3 is required to avoid technical complexity in the dynamic Cox model and may not be necessary. This condition holds obviously when s_0 is fixed and has also been required by [3] under similar settings.

To estimate $\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_\infty$, we only need to consider the subvector in the first s_0 component, i.e., $\|\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0\|_\infty$, because $\check{\boldsymbol{\beta}}_{S^c} = \boldsymbol{\beta}_{S^c}^0 = \mathbf{0}$. Hence, the consistency of the estimator $\check{\boldsymbol{\beta}}$ can be obtained from the consistent result of Theorem 3.1 [12] over the restricted space $\mathbb{R}^S$.

Lemma 1 (*Estimation Error of $\check{\boldsymbol{\beta}}_S$*). *Suppose that Assumptions 1–3 hold over $\mathbb{R}^S$. Then in the event $\|\dot{\ell}(\boldsymbol{\beta}^0)_S\|_\infty = O_p(\lambda)$, we have*

$$\|\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0\|_2 = O_p\big(\sqrt{s_0}\lambda\big).$$

Moreover, let $\lambda = O_p\{\sqrt{\ln(p/\delta)/n}\}$ with a small $\delta \in (0, 1)$, then with probability at least $1 - \delta$,

$$\|\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0\|_2 = O_p\left\{\sqrt{s_0\ln(p/\delta)/n}\right\}.$$

The proof of Lemma 1 is given in Appendix A. By Lemma 1(a), we can see that, in order to prove $\check{\boldsymbol{\beta}}$ is an optimum of the Lasso program (3), we still need to show that $\check{z}$ is one of subgradients of the ℓ_1-norm at $\check{\boldsymbol{\beta}}$.

3.3. Weak oracle property

An estimator is said to have the weak oracle property if it is both estimation consistent and model selection consistent, proposed originally in [20]. By parts (a) and (b) of Lemma 5, to derive the weak oracle property of the Lasso estimator defined in (3), we shall provide a sufficient condition to ensure that $\hat{z} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$ and $\max_{j\in S(\hat{\boldsymbol{\beta}})^c}|\hat{z}_j| < 1$.

Assumption 4. There exists $\gamma \in (0, 1]$ such that $\|\boldsymbol{\Sigma}_{S^cS}(\boldsymbol{\Sigma}_{SS})^{-1}\|_\infty \le 1-\gamma$. (the mutual incoherence condition)

Assumption 4 is an analog of mutual incoherent conditions related to the covariance matrix $\boldsymbol{\Sigma}$, which have been considered in linear regression [29] and in additive hazards regression [18]. It should be pointed out that the restriction on the correlation structure in Assumption 4 is more complex, compared to linear or general additive models. In fact, the matrix $\boldsymbol{\Sigma}$ depends on the failure process and censoring mechanism, as well as the distribution of the covariates. Although Assumption 4 is stringent in some sense, we will show in Theorem 3 that such an incoherent condition is necessary to guarantee support recovery of the Cox model via the Lasso.

We now state some properties of the PDW method, which will be useful for deriving the oracle property of the Lass estimator.

Theorem 1. *Under Assumptions 1–3 as above.*

(a) *If Steps 1 through 3 of the PDW method succeed with strict dual feasibility in Step 3, then the Lasso (3) has a unique solution given by $\check{\boldsymbol{\beta}}$ with $S(\check{\boldsymbol{\beta}}) \subseteq S$.*
(b) *If Steps 1 through 4 of the PDW method succeed with strict dual feasibility in Step 3. When $|\check{\beta}_j| > 0$ is satisfied for $j \in S$, $\check{\boldsymbol{\beta}}$ is the unique solution of the Lasso (3), such that the corrected signed support holds, i.e., $\mathrm{sign}(\check{\boldsymbol{\beta}}) = \mathrm{sign}(\boldsymbol{\beta}^0)$.*
(c) *Conversely, if either Step 3 or 4 of the PDW method fails, then the Lasso fails to recover the corrected signed support.*

It is seen from Theorem 1, the main task in the PDW construction lies in verifying the dual feasibility condition in Step 3, and the sign consistency condition in Step 4.

Note also that part(b) of Theorem 1 holds only when $|\check{\beta}_j| > 0$ $(j \in S)$ is satisfied, which can be verified as long as the lower bound of $\{|\beta_j^0|, j \in S\}$ is not too small. This will be shown clearly in Theorem 2.

Combining Theorem 1 and technical lemmas presented in the Appendix, we obtain the weak oracle property of the proposed estimator.

Theorem 2. *Under Assumptions 1–4 and $\mathrm{E}\{\|\mathbf{Z}(t)^{\otimes 2}_{SS}\|_2\}$ and $\|(\boldsymbol{\Sigma}_{SS})^{-1}\|_\infty$ are bounded uniformly on $t \in [0, \tau]$. If $n \gg s_0^4 \ln(p)$ and let $\lambda = O_p\{\sqrt{\ln(p/\delta)/n}\}$, then with probability at least $1-\delta$, $\check{\boldsymbol{\beta}}$ is the unique solution of the Lasso program (3), i.e., $\hat{\boldsymbol{\beta}} = \check{\boldsymbol{\beta}}$, such that (a) $\hat{\boldsymbol{\beta}}_{S^c} = \mathbf{0}$ and (b) $\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0\|_\infty = O_p\{\sqrt{\ln(p/\delta)/n}\}$. If additionally, we have a lower bound of the form $\min_{j\in S}|\beta_j^0| \gg \sqrt{\ln(p/\delta)/n}$, then it is guaranteed that $\hat{\boldsymbol{\beta}}$ is sign-consistent for $\boldsymbol{\beta}^0$.*

The proof of Theorem 2 is relegated to Section 5.1. Due to the semiparametric structure and the censoring mechanism, we require constraints $n \gg s_0^4 \ln(p)$ to guarantee sparse recovery. This sufficient condition on p, s_0 and n implies that the Lasso Cox estimator can handle a non-polynomially growing dimension of covariates as high as $p = o_p\{\exp(n/s_0^4)\}$, and the dimension of the true sparse model growing as $s_0 = o_p(n^{1/4})$. Theorem 2 guarantees that $\check{\boldsymbol{\beta}}$ is the unique sign-consistent estimator for $\boldsymbol{\beta}^0$, as long as $\min_{j\in S}|\beta_j^0| \gg \sqrt{\ln(p)/n}$ are satisfied.

3.4. Necessary conditions for sign consistency

Now we turn to the results related to the failure of the sign recovery consistency, providing that either mutual incoherence condition or the constraint on minimum value of the true coefficient is violated.

Theorem 3. *Suppose that Assumptions 1–3 hold.*

(a) *As opposed to Assumption 4 (the mutual incoherence condition), we assume that, for some $\nu > 0$,*

$$\max_{j\in S^c}|e_j^\top \boldsymbol{\Sigma}_{S^cS}(\boldsymbol{\Sigma}_{SS})^{-1}\mathrm{sign}(\boldsymbol{\beta}_1^0)| = 1+\nu. \tag{7}$$

Then for any $\lambda > 0$ and sufficiently large n, the probability that sign support recovery fails is no less than $1/2$, i.e., $\Pr\{\mathrm{sign}(\hat{\boldsymbol{\beta}}) \ne \mathrm{sign}(\boldsymbol{\beta}^0)\} \ge 1/2$.
(b) *For each $j \in S$, we define the quantity $h(\lambda) = \lambda e_j^\top(\boldsymbol{\Sigma}_{SS})^{-1}\mathrm{sign}(\boldsymbol{\beta}_1^0)$. If there exists $j \in S$ with $\beta_j^0 \in (0, h(\lambda))$ for $h(\lambda) > 0$ or $\beta_j^0 \in (h(\lambda), 0)$ for $h(\lambda) < 0$, then $\Pr\{\mathrm{sign}(\hat{\boldsymbol{\beta}}) \ne \mathrm{sign}(\boldsymbol{\beta}^0)\} \ge 1/2$.*

Theorem 3(a) implies that, the mutual incoherence condition in Assumption 4 is essential to ensure the corrected sign recovery for the Lasso partial likelihood estimator. The same result has been established in linear regression models [27,29]. For sign consistency, Theorem 3(b) shows that the value $\min_{j\in S}|\beta_j^0|$ cannot decay to zero faster than the penalty parameter λ. Note that (7) is not a complete complementary event of Assumption 4, where $\gamma > 0$ is used to guarantee the uniqueness of the estimator $\hat{\boldsymbol{\beta}}$.

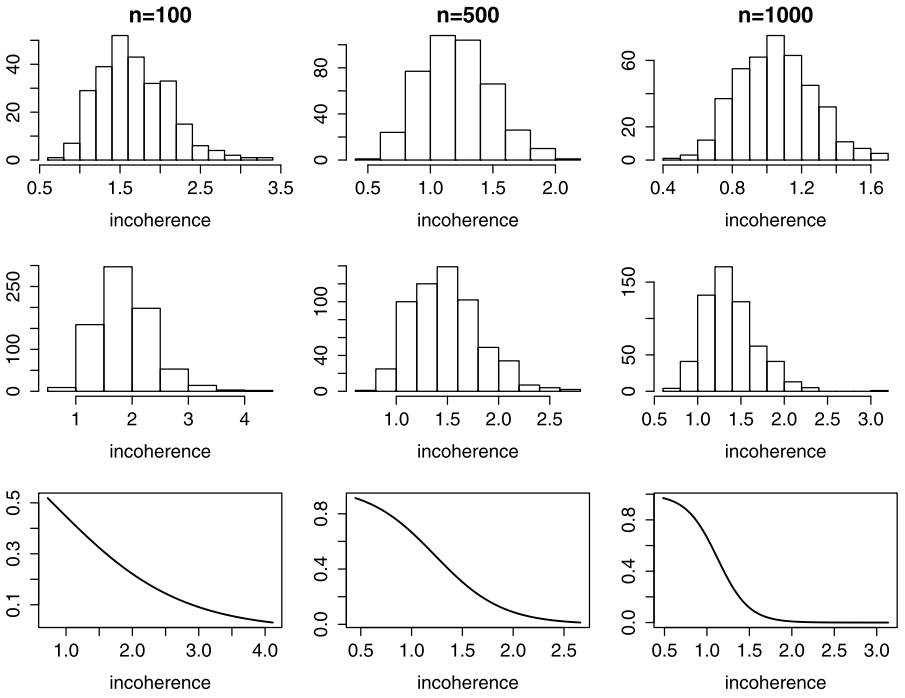

Fig. 1. Example 1—relating incoherence measure to sign consistency.

4. Numerical results

4.1. Simulations

We conduct some simulations to demonstrate the model selection consistency of the Lasso estimator, as well as to investigate whether the consistency is related to incoherence.

In the first example, we generate data from a Cox model with hazard function given by

$$\lambda(t|\mathbf{Z}_i) = \exp\big(\mathbf{Z}_i^\top \boldsymbol{\beta}^0 - 4\big),$$

where $\mathbf{Z}_i$ is generated by first generating p-dimensional random vectors from a multivariate Gaussian distribution with covariance matrix generated from a Wishart distribution with $p \times p$ identity matrix as the scale matrix and p degrees of freedom, and then standardizing the predictors to have mean zero and variance 1. We set $\boldsymbol{\beta}^0 = (3.5, 3, 2.5, -2, -1.5, 0, \ldots, 0)^\top$. The censoring times are independently generated from an exponential distribution with mean 1000 such that the censoring rate is roughly around 20%.

We take $n \in \{100, 500, 1000\}$ and $p = 20$ and generate 1000 datasets in each setting. A solution path for each dataset is obtained using the R package `glmnet` and we examine the path to see if there is a solution that satisfies sign consistency. We also compute the incoherence measure $\eta = \|\boldsymbol{\Sigma}_{S^cS}(\boldsymbol{\Sigma}_{SS})^{-1}\text{sign}(\boldsymbol{\beta}_1^o)\|_\infty$, with $\boldsymbol{\Sigma}$ estimated from generated data.

In Fig. 1, we show the histogram of values of η when the dataset admits a sign consistent solution (first row) and the histogram of value of η when the dataset does not admit a sign consistent solution (second row). It is seen that the incoherence values are on average smaller when we can have a sign consistent solution. Based on the 1000 values of incoherence in each scenario, we also fit a logistic regression model using incoherence value as the predictor to predict whether we have a sign consistent solution (1 for yes and 0 for no) and the predicted probability of having a sign consistent solution given incoherence value is shown in the third row of Fig. 1. It is seen that the probability increases as incoherence value decreases, and the increase is faster with larger sample sizes.

In the second example, we vary p and s (the size of nonzero components of the coefficients). The set up is the same as before except that we now simply set $\boldsymbol{\beta}^0 = (1, \ldots, 1, 0, \ldots, 0)$ (s ones followed by $p - s$ zeros). We fix $n = 300$, take $p \in \{20, 50, 100, 500\}$, $s \in \{p/10, p/2\}$, and generate 1000 datasets for each pair of values (p, s). Table 1 presents the simulation results. The first three rows in the table shows some summary statistics for the incoherence η, including its mean, standard deviation, and the proportion of times $\eta < 1$, and the next row, "path consistency", reports the proportion of times there is a sign consistent solution on the solution path. The results show that the incoherence condition is more likely to satisfy with small dimension p and small s (very sparse model). Similarly, sign consistency on the solution path is obtained more likely in these situations. In this example, sign consistency on the solution path is never obtained for $p = 500$.

Table 1
Simulation results in Example 2.

		$p = 20$		$p = 50$		$p = 100$		$p = 500$	
		$s = 2$	$s = 10$	$s = 5$	$s = 25$	$s = 10$	$s = 50$	$s = 50$	$s = 250$
Incoherence	Mean	0.689	1.874	0.900	2.476	1.180	3.378	3.662	25.756
	std	0.157	0.675	0.168	0.571	0.183	0.624	0.434	8.384
	prop < 1	0.965	0.044	0.737	0	0.154	0	0	0
	Path consistent	0.972	0.062	0.836	0	0.415	0	0	0
BIC	Consistent	0.029	0	0.002	0	0	0	0	0
	# NZ	4.756	12.776	13.166	33.932	24.612	71.423	158.319	180.144
	# NZT	1.500	7.492	3.750	18.752	7.500	37.496	37.474	108.789
	MSE	0.561	2.706	1.660	6.982	4.270	32.528	22.305	205.893
GCV	Consistent	0	0	0	0	0	0	0	0
	# NZ	9.877	12.573	26.411	32.332	55.829	66.058	157.658	231.207
	# NZT	1.500	7.488	3.750	18.749	7.500	37.498	37.324	132.385
	MSE	0.513	2.917	1.294	7.830	2.668	16.795	22.961	249.226
GIC	Consistent	0.025	0	0.003	0	0	0	0	0
	# NZ	5.041	12.860	11.373	33.407	17.528	69.739	52.759	70.450
	# NZT	1.500	7.492	3.750	18.750	7.500	37.497	28.234	49.125
	MSE	0.555	2.700	1.767	6.951	5.068	25.791	48.308	248.189

Table 2
Results for the real data. In each cell, the first number is incoherence and the second number is the number of genes selected. When no genes are selected, the incoherence value cannot be calculated.

Method	$p = 20$	$p = 50$	$p = 100$	$p = 200$	$p = 500$
BIC	1.165, 8	1.850, 13	NA, 0	NA, 0	NA, 0
GCV	1.170, 8	1.792, 14	1.339, 23	1.469,43	2.143, 68
GIC	1.166, 8	NA, 0	NA, 0	NA, 0	NA, 0

Tuning parameter selection is an important and challenging problem in practice. BIC and GCV are commonly used to select a final estimator on the solution path as detailed in [24]. Recent work including Fan and Tang [9] propose the GIC method for tuning parameter selection with strong theoretical guarantees. The next four rows in Table 1 shows the proportion of times the estimator is sign consistent, the average number of estimated nonzero coefficients (NZ), the average number of estimated nonzero coefficients that are also truly nonzero in the generating model (NZT), and the mean squared error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2$ (MSE) using the tuning parameter selected by BIC. The last eight rows in Table 1 shows those measures when using the tuning parameter selected by GCV and GIC. These tuning parameter selectors almost never select the consistent solution even when there is one on the path. In terms of model selection, there is no clear winner between BIC, GCV and GIC. When $p > n$ ($p = 500$), the results show that GIC tends to select larger values of tuning parameter λ than those for other regularization methods to enforce the model sparsity, which is consistent with that in [9].

4.2. Empirical application

We use a published dataset of DLBCL (https://llmpp.nih.gov/DLBCL/) by Rosenwald et al. [22] to illustrate the Cox model in relating microarray gene expression data to censored survival phenotypes. This dataset includes a total of 240 patients with DLBCL, including 138 patient deaths during the follow-ups with a median death time of 2.8 years.

Expression measurements on 7399 genes are available for analysis. We first reduce the number of genes by performing an initial screening based on marginal Cox models fitted separately for each single gene as done in [13]. We consider dimensions 20, 50, 100, 200, 500 and use BIC, GCV and GIC to select the tuning parameter. In each case, after we obtained the estimator $\hat{\beta}$ and the estimated support set $\hat{S}$, we use these to calculate the incoherence. The incoherence values and the number of genes selected are reported in Table 2. All incoherences are larger than 1 with higher dimension resulting in larger incoherence as in our simulations. This indicates that model selection consistency may be in doubt for the data, especially if the dimension is high.

5. Proofs of the main theorems

A key step in the proof of Theorem 2 is to verify the strict dual feasibility condition In the PDW procedure. To this end, We first derive two sufficient conditions to guarantee the strict dual feasibility condition. Then, we shall prove that these two conditions are satisfied under our Assumptions 1–4. To prove Theorem 3, we need an application of Martingale Central Limit Theorem.

5.1. Proof of Theorem 2

In the proof, we first derive conditions that allow us to establish the strict dual feasibility conditions required so as to apply Theorem 1. By the zero-subgradient condition (6), we rewrite it as

$$\dot{\ell}(\check{\boldsymbol{\beta}}) - \dot{\ell}(\boldsymbol{\beta}^0) + \dot{\ell}(\boldsymbol{\beta}^0) + \lambda\check{z} = 0.$$

Let $\widehat{Q} = \int_0^1 \ddot{\ell}\{\boldsymbol{\beta}^0 + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\}d\theta$. Then one gets $\widehat{Q}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \dot{\ell}(\boldsymbol{\beta}^0) + \lambda\check{z} = 0$. Since $S(\check{\boldsymbol{\beta}}) \subseteq S$, the above equality can be rewritten in a block form as

$$\begin{bmatrix} \widehat{Q}_{SS} & \widehat{Q}_{SS^c} \\ \widehat{Q}_{S^cS} & \widehat{Q}_{S^cS^c} \end{bmatrix} \begin{bmatrix} \check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0 \\ 0 \end{bmatrix} + \begin{bmatrix} \dot{\ell}(\boldsymbol{\beta}^0)_S \\ \dot{\ell}(\boldsymbol{\beta}^0)_{S^c} \end{bmatrix} + \lambda \begin{bmatrix} \check{z}_S \\ \check{z}_{S^c} \end{bmatrix} = 0. \tag{8}$$

By computing the top block of Eq. (8), we have

$$\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0 = -(\widehat{Q}_{SS})^{-1}\{\dot{\ell}(\boldsymbol{\beta}^0)_S + \lambda\check{z}_S\}. \tag{9}$$

Furthermore, a simple algebraic manipulations for the bottom of (8) yields

$$\check{z}_{S^c} = 1/\lambda\Big[\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\{\dot{\ell}(\boldsymbol{\beta}^0)_S + \lambda\check{z}_S\} - \dot{\ell}(\boldsymbol{\beta}^0)_{S^c}\Big]. \tag{10}$$

Based on Eq. (10), we have the following result.

Proposition 1. *Under the PDW construction of (6), the strict dual feasibility holds, provided that λ is chosen to satisfy the following inequalities*

$$\|\dot{\ell}(\boldsymbol{\beta}^0)\|_\infty \le \frac{\gamma}{8+2\gamma}\lambda, \tag{11}$$

and

$$\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\|_\infty \le 1 - \gamma/2. \tag{12}$$

Proof. Since $\|\check{z}_S\|_\infty \le 1$ by definition, Eq. (10) is applied to yield that

$$\begin{aligned} \|\check{z}_{S^c}\|_\infty &\le 1/\lambda\|\dot{\ell}(\boldsymbol{\beta}^0)\|_\infty + 1/\lambda\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^0)_S\big\|_\infty + \big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty \\ &\le 1/\lambda\|\dot{\ell}(\boldsymbol{\beta}^0)\|_\infty + \big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty\{1 + 1/\lambda\|\dot{\ell}(\boldsymbol{\beta}^0)_S\|_\infty\} \\ &\le 1 - \gamma/4, \end{aligned}$$

provided that (11) and (12) hold simultaneously. □

By Proposition 1, in order to verify the strict dual feasibility condition, we need to give an appropriate bound for $\|\dot{\ell}(\boldsymbol{\beta}^0)\|_\infty$ and provide sufficient conditions to guarantee that (12) holds.

To bound $\|\dot{\ell}(\boldsymbol{\beta}^0)\|_\infty$, we introduce additional notation on martingales. Since $\boldsymbol{\Lambda}^{(n)}$ is the predictable compensator of $\mathbf{N}^{(n)}$ and $N_i(t)$ are independent processes for all $i \in \{1, \ldots, n\}$, we have, for all $t > 0$,

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\exp\{\mathbf{Z}_i(s)^\top\boldsymbol{\beta}^0\}d\Lambda_0(s),$$

are local martingales on $[0, \tau]$ with predictable variation/covariation processes

$$\langle M_i, M_i\rangle(t) = \int_0^t Y_i(s)\exp\{\mathbf{Z}_i(s)^\top\boldsymbol{\beta}^0\}d\Lambda_0(s)$$

with $\langle M_i, M_j\rangle(t) = 0$ when $i \ne j$.

The following lemma is provided by [12], and another similar lemma can be found in [6].

Lemma 2. *(i) Let $f_n(t) = \sum_{i=1}^n \int_0^t a_i(s)dM_i(s)/n$ with $[-1, 1]$-valued predictable processes $a_i(s)$. Then, for all $c > 0$,*

$$\Pr\left\{\max_{t\in[0,\tau]}|f_n(t)| > cx,\ \sum_{i=1}^n\int_0^\tau Y_i(t)dN_i(t) \le c^2n\right\} \le 2\exp(-nx^2/2).$$

(ii) Suppose that $\sup_{t\ge0}\max_{i\le n, j\le p}|Z_{ij}(t) - \bar{Z}_{nj}(t, \boldsymbol{\beta}^0)| \le K$, where $\bar{Z}_{nj}(t, \boldsymbol{\beta}^0)$ are the components of $\bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0)$. Then, for all $c > 0$,

$$\Pr\left\{\|\dot{\ell}(\boldsymbol{\beta}^0)\|_\infty > cKx,\ \sum_{i=1}^n\int_0^\tau Y_i(t)dN_i(t) \le c^2n\right\} \le 2p\exp(-nx^2/2).$$

If additionally $\Pr\{\max_{i\le n}N_i(\tau) \le 1\} = 1$, we can take $c = 1$ and hence $\Pr(\|\dot{\ell}(\boldsymbol{\beta}^0)\|_\infty > Kx) \le 2p\exp(-nx^2/2)$.

It is shown in Lemma 2, with high probability we can take $\lambda = c_2\{\sqrt{\ln(p)/n}\}$ with a suitable constant c_2, so that Eq. (11) is satisfied.

However, to verify (12) in Proposition 1 under Assumption 4, we need to measure how close between $\widehat{Q}$ and $\mathbf{\Sigma}$. To be precise, since $\widehat{Q}$ is an empirical counterpart of the matrix $\mathbf{\Sigma}$ (note also depends on $\check{\boldsymbol{\beta}}$), a key step to verify (12) is to show that $\left(\widehat{Q}_{SS}\right)^{-1}$ and $\widehat{Q}_{S^cS}$ are close to $\left(\mathbf{\Sigma}_{SS}\right)^{-1}$ and $\mathbf{\Sigma}_{S^cS}$, respectively. These intermediate results are provided by the following proposition.

Proposition 2 (*Concentration of Empirical Matrices*). *Suppose that Assumptions 1–3 hold and* $\mathrm{E}\{\|\mathbf{Z}(t)_{SS}^{\otimes 2}\|_2\}$ *is bounded uniformly. If* $s_0^3 \ll n$, *with probability at least* $1-\delta$, *then*

$$\|(\widehat{Q}_{SS})^{-1} - (\mathbf{\Sigma}_{SS})^{-1}\|_2 = s_0 O_p\left\{\sqrt{s_0/n} + \sqrt{\ln(4s_0^2/\delta)/n}\right\}. \tag{13}$$

If additionally Assumption 4 also holds, and $n \gg s_0^3 \ln(p)$, *then, with the same probability as above, one has*

$$\left\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\right\|_\infty \le 1-\gamma/2.$$

Thus, under Assumptions 1–4, Eq. (12) is verified by Proposition 2. Until now, two conditions involved in Proposition 1 are both verified, so the strict dual feasibility of Step 3 in the PDW procedure holds. Then by part (a) of Theorem 1, we conclude that $\check{\boldsymbol{\beta}}$ is the unique solution of the Lasso program (3) with $S(\check{\boldsymbol{\beta}}) \subseteq S$. Next we shall obtain the inequality in Theorem 2 based on Eq. (9). In fact, by (13) and Assumption 3 we can treat $\left(\widehat{Q}_{SS}\right)^{-1}$ as some universal constant. Precisely, by (11) in Proposition 1, one gets

$$\begin{aligned}
\|\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0\|_\infty &\le \|(\widehat{Q}_{SS})^{-1}\|_\infty \|\dot{\ell}(\boldsymbol{\beta}^0)_S + \lambda \check{z}_S\|_\infty \\
&\le \left\{\|(\widehat{Q}_{SS})^{-1} - \left(\mathbf{\Sigma}_{SS}\right)^{-1}\|_\infty + \|\left(\mathbf{\Sigma}_{SS}\right)^{-1}\|_\infty\right\}\frac{(8+3\gamma)}{8+2\gamma}\lambda \\
&\le \left\{\sqrt{s_0}\|(\widehat{Q}_{SS})^{-1} - \left(\mathbf{\Sigma}_{SS}\right)^{-1}\|_2 + \|\left(\mathbf{\Sigma}_{SS}\right)^{-1}\|_\infty\right\}\frac{(8+3\gamma)}{8+2\gamma}\lambda.
\end{aligned}$$

With the choice of $\lambda = c_2\{\sqrt{\ln(p/\delta)/n}\}$ as before, the desired inequality in Theorem 2 follows from Proposition 2 immediately as long as $s_0^4 \ll n$ is satisfied.

In addition, if $\min_{j\in S}|\beta_j^0| \gg \sqrt{\ln(p/\delta)/n}$, without loss of generality, we assume that $\beta_j^0 > 0$. In this case, by part(b) of Theorem 2, we have that $\check{\beta}_j \ge \beta_j^0 - O_p\{\sqrt{\ln(p/\delta)/n}\} > 0$ for any $j \in S$. A similar argument also holds for $\beta_j^0 < 0$. As a result, we conclude that $\operatorname{sign}(\check{\boldsymbol{\beta}}_S) = \operatorname{sign}(\boldsymbol{\beta}_1^0)$. Since we have proved that $|\check{\beta}_j| > 0$ for all $j \in S$, it follows that $\check{z}_S = \operatorname{sign}(\check{\boldsymbol{\beta}}_S) = \operatorname{sign}(\boldsymbol{\beta}_1^0)$ by the definition of subgradient. So Step 4 in the PDW is verified. In sum, all the conditions of part(b) in Theorem 1 are satisfied, and based on this we obtain the second part of Theorem 2. □

5.2. Proof of Theorem 3

By part (c) of Theorem 1, it suffices to show that the dual feasibility check in Step 3, or the sign consistency check in Step 4 of the PDW must fail with probability at least 1/2. It may be assumed that $\check{z}_S = \operatorname{sign}(\boldsymbol{\beta}_1^0)$, otherwise, the sign consistency condition fails. Then, it remains to show that under this condition, the dual feasibility condition in Step 3 fails with probability at least 1/2. From (10), we recall that

$$\check{z}_{S^c} = \frac{1}{\lambda}\left[\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\left\{\dot{\ell}(\boldsymbol{\beta}^0)_S + \lambda \operatorname{sign}(\boldsymbol{\beta}_1^0)\right\} - \dot{\ell}(\boldsymbol{\beta}^0)_{S^c}\right]. \tag{14}$$

Let $j \in S^c$ be the index corresponding to the maximum which is achieved in the violating condition (7). Note that within the proof of Proposition 2, we have shown that $\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}$ converges to $\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}$ in $\|\cdot\|_\infty$-norm. Then, the violation of condition (7) implies that

$$\left|e_j^\top \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1^0)\right| \ge 1 + \nu/2$$

with high probability tending to 1. Without loss of generality, we may assume that $e_j^\top \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1^0) \ge 1 + \nu/2$. By (14), we have

$$\check{z}_j = e_j^\top \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\operatorname{sign}(\boldsymbol{\beta}_1^0) - 1/\lambda e_j^\top\left\{\dot{\ell}(\boldsymbol{\beta}^0)_{S^c} - \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^0)_S\right\}.$$

Hence, to prove $\check{z}_j > 1$, it suffices to show that $\Pr(W_j \le \nu/4) \ge 1/2$, where

$$W_j = 1/\lambda\left\{\dot{\ell}(\boldsymbol{\beta}^0)_j - e_j^\top \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^0)_S\right\}.$$

For this purpose, from the proof of Theorem 3.2 in [1], an application of the Martingale Central Limit Theorem yields that $\dot{\ell}(\boldsymbol{\beta}^0)_j$ and $\dot{\ell}(\boldsymbol{\beta}^0)_S$ are both asymptotically $\mathcal{N}(0,1)$. Note that

$$e_j^\top \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^0)_S = e_j^\top\left\{\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1} - \mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\right\}\dot{\ell}(\boldsymbol{\beta}^0)_S + e_j^\top\left\{\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\right\}\dot{\ell}(\boldsymbol{\beta}^0)_S.$$

It follows from the proof of Proposition 2 that $\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\|_\infty = o_p(1)$. Then the first term in the above equality converges to zero in probability, and this in turn means that $e_j^\top \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\dot{\ell}(\boldsymbol{\beta}^0)_S$ is also an asymptotically Gaussian variable with zero mean. As a result, W_j is also an asymptotically Gaussian variable with zero mean, so $\Pr\big(W_j \le \nu/4\big) > \Pr\big(W_j \le 0\big) = 1/2$ as $n \to \infty$. This completes the proof of part (a) in Theorem 3.

To prove part (b) of Theorem 3, we first recall from (9) that

$$\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0 = -(\widehat{Q}_{SS})^{-1}\big\{\dot{\ell}(\boldsymbol{\beta}^0)_S + \lambda \check{z}_S\big\}.$$

To recover the correct signed support, we must have $\check{z}_S = \operatorname{sign}(\boldsymbol{\beta}_1^0)$. Besides, by Proposition 2, we have shown that $\|\big(\widehat{Q}_{SS}\big)^{-1} - \big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\|_\infty = o_p(1)$. Thus, for $j \in S$ specified in part (b) of Theorem 3, $\check{\beta}_j$ can be expressed by

$$\begin{aligned}\check{\beta}_j &= e_j^\top\big\{\big(\boldsymbol{\Sigma}_{SS}\big)^{-1} - \big(\widehat{Q}_{SS}\big)^{-1}\big\}\big\{\dot{\ell}(\boldsymbol{\beta}^0)_S + \lambda\operatorname{sign}(\boldsymbol{\beta}_1^0)\big\} + \beta_j^0 - e_j^\top\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\big\{\dot{\ell}(\boldsymbol{\beta}^0)_S + \lambda\operatorname{sign}(\boldsymbol{\beta}_1^0)\big\}\\ &= o_p(1) + \big\{\beta_j^0 - h(\lambda)\big\} - e_j^\top\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\dot{\ell}(\boldsymbol{\beta}^0)_S.\end{aligned}$$

Without loss of generality, we only consider the situation $\beta_j^0 \in \big(0, h(\lambda)\big)$. As mentioned above, $e_j^\top\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\dot{\ell}(\boldsymbol{\beta}^0)_S$ is an asymptotically standard Gaussian with zero mean, which yields

$$\Pr(\check{\beta}_j < 0) > \Pr\big\{e_j^\top\big(\boldsymbol{\Sigma}_{SS}\big)^{-1}\dot{\ell}(\boldsymbol{\beta}^0)_S \ge 0\big\} = 1/2$$

as $n \to \infty$, where we used the fact that $\beta_j^0 - h(\lambda) < 0$. As a result, Step 4 of the PDW fails, i.e., $\check{z}_S \ne \operatorname{sign}(\boldsymbol{\beta}_1^0)$. This together with part (c) of Theorem 1 concludes the proof. □

6. Concluding remarks

In this paper, we studied variable selection and estimation of the ℓ_1-penalized partial likelihood estimation for the sparse high-dimensional Cox model with counting process data and time-dependent covariates. In particular, we provided sufficient and necessary conditions for sign consistency of the Lasso penalized partial likelihood estimator. We showed that the two conditions (mutual incoherence condition, the minimum value of the coefficient) are essential to guarantee sign consistency in terms of theory and empirical evidence. These results support that the Lasso is a useful model identification method for the high-dimensional Cox model. In addition, the rate of convergence $\sqrt{s_0\ln(p)/n}$ is actually optimal with an appropriate choice of the tuning parameter λ and under the aforementioned conditions in Theorem 2. This rate is, apart from logarithmic factor in p and n, identical to what could be achieved if the true sparse model would be known.

Although the current work focused exclusively on the Cox model, the empirical process techniques and the construction techniques (e.g., the primal–dual witness technique) used here are fairly general and can be easily adapted to other survival models. To be precise, a key step to establishing a high-dimensional theory for the Cox model under similar conditions is to characterize the concentration of the empirical information matrix around its population counterpart at the true coefficients. Note that this is quite different from the classical linear regression, since analyzing the Cox model in high dimensions involves a non-quadratic loss function and martingale processes. Besides, the primal–dual witness technique allows us to develop the convex-analytic conditions that characterize the optima of the L_1-regularized program (3). We then specify the construction technique to describe the proof procedure of our main results, and provide some basic lemmas so as to show how it characterizes the success (or failure) of the Lasso in recovering the correct support set.

Acknowledgments

The research was supported partially by National Natural Science Foundation of China (Grant No. 11571282, 11528102), and Fundamental Research Funds for the Central Universities of China (Grant Nos. JBK120509 and JBK140507), as well as KLAS-130026507.

Appendix A. Useful lemmas about empirical processes

In this paper, a main ingredient from the theoretical point of view is that the randomness of the problem should be taken care of. For example, the covariate $\mathbf{Z}_i(t)$ is a random function of t, and classical results from random matrix theory are invalid for our time-dependent data. In this situation, we need to consider the behavior of the empirical process. This paper adopts the notation of Rademacher complexity to characterize the functional capacity. Let us recall Rademacher random variables, which are independent $\{-1, 1\}$-valued random variables with probability $1/2$ of taking either value. Let $X_1,\ldots,X_n$ be iid random sample from the distribution ρ and $\sigma_1,\ldots,\sigma_n$ be iid Rademacher random variables. We define the empirical Rademacher average on the function space $\mathcal{G}$ by setting, for all $g \in \mathcal{G}$,

$$\widehat{R}_n(g) = \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i),$$

and the population Rademacher average $R_n(\mathcal{G})$ is given by

$$R_n(\mathcal{G}) = \sup_{g\in\mathcal{G}} \mathrm{E}_{\sigma,\rho}\left\{\widehat{R}_n(g)\right\},$$

where $\mathrm{E}_{\sigma,\rho}$ means taking expectation with respect to all the random variables (i.e., the data and the Rademacher variables). Now we give some basic properties of the Rademacher average.

Properties. *Given any function class $\mathcal{G}$ and constants $a, b \in \mathbb{R}$, denote the function class $\{h : h(x) = ag(x) + b\}$ by $a\mathcal{G} + b$. Then*

$$R_n(a\mathcal{G} + b) = |a| R_n(\mathcal{G}). \tag{15}$$

Furthermore, the Rademacher average can be bounded by the so-called Dudley's entropy integral, namely, there exists some constant c_0 such that

$$R_n(\mathcal{G}) \le c_0/\sqrt{n} \int_0^{\|\mathcal{G}\|_\infty} \sqrt{\ln N(\mathcal{G}, \varepsilon, d_n)}\, d\varepsilon, \tag{16}$$

where $N(\mathcal{G}, \varepsilon, d_n)$ is the empirical covering number of $\mathcal{G}$ with the radius ε. Here the metric is defined as

$$\left\{d_n(f, g)\right\}^2 = \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|^2$$

associated with available sample points $x_1, \ldots, x_n$. In the literature on empirical processes, it is known that using this Rademacher average can remove the unnecessary $\ln n$ factor of the VC bound, as well as refined constants.

Let $P_n(g) = \{g(X_1) + \cdots + g(X_n)\}/n$ and $P(g) = \mathrm{E}_\rho\left\{g(X)\right\}$. We now state the fundamental result involving Rademacher averages from Chapter 4 in [16].

Lemma 3. *If $\mathcal{G} \subseteq \{g : X \to [c, c+1]\}$ for any given constant c. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sup_{g\in\mathcal{G}} |P(g) - P_n(g)| \le 2R_n(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Note that by (15), we can extend the result of Lemma 3 to any bounded function space.

A major challenge in our proof is to characterize the concentration of the large matrix/vector given, for each $k \in \{0, 1, 2\}$, by

$$S^{(\kappa)}(t, \boldsymbol{\beta}^0) = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i(t)^{\times k} Y_i(t) e^{\mathbf{Z}_i(t)^\top \boldsymbol{\beta}^0}.$$

Note that each entry of the stochastic matrix is not a sum of independent terms, since it may vary with time t. Hence the functional complexity in the Law of Large Number has to be considered. To this end, we rely on Lemma 3 concerning the Rademacher complexity for empirical processes as our primary mathematical tools.

Lemma 4. *Assume that Assumption 2 holds. For all $\ell, k \in \{1, \ldots, p\}$, there exists some universal constant $c_0 > 0$, such that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sup_{t\in[0,\tau]} \left|S^{(0)}(t, \boldsymbol{\beta}^0) - s^{(0)}(t, \boldsymbol{\beta}^0)\right| \le c_0 s_0 e^{K\|\boldsymbol{\beta}_1^0\|_1} \sqrt{\ln(2/\delta)/n}, \tag{17}$$

$$\sup_{t\in[0,\tau]} \left|S_\ell^{(1)}(t, \boldsymbol{\beta}^0) - s_\ell^{(1)}(t, \boldsymbol{\beta}^0)\right| \le c_0 s_0 e^{K\|\boldsymbol{\beta}_1^0\|_1} \sqrt{\ln(2/\delta)/n},$$

$$\sup_{t\in[0,\tau]} \left|S_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0) - s_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)\right| \le c_0 s_0 e^{K\|\boldsymbol{\beta}_1^0\|_1} \sqrt{\ln(2/\delta)/n}, \tag{18}$$

where $S_\ell^{(1)}$ is the jth element of $S^{(1)}$ and $S_{\ell k}^{(2)}$ is the (ℓ, k)th entry of $S^{(2)}$, and the same is valid to $s_\ell^{(2)}$ and $s_{\ell k}^{(2)}$.

Proof. We only prove (18), and the other two inequalities follow similarly. For any given ℓ, k, we write $g(X_i) = w_i(t) Z_{i\ell}(t) Z_{ik}(t)$, where $w_i(t) = Y_i(t) e^{\mathbf{Z}_i(t)^\top \boldsymbol{\beta}^0}$, $t \in [0, \tau]$. That is, $S_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0) = P_n(g)$ and $\mathrm{E}\left\{S_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)\right\} = s_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0) = P(g)$. Since $\max_{i,\ell} |Z_{i\ell}(t)| \le K$ for all $t \in [0, \tau]$, it follows that $w_i(t) \le e^{K\|\boldsymbol{\beta}_1^0\|_1}$. Let $\tilde{g}(X_i) = g(X_i)/(2K^2 e^{K\|\boldsymbol{\beta}_1^0\|_1})$, then $\tilde{g}(X_i) \in [-1/2, 1/2]$. Based on Lemma 3 with $c = -1/2$, the following inequality holds with probability at least $1 - \delta$:

$$\forall_{\tilde{g}\in\mathcal{G}} \quad |P(\tilde{g}) - P_n(\tilde{g})| \le 2R_n(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2n}}, \tag{19}$$

where $\mathcal{G} = \{w(t) Z_\ell(t) Z_k(t)/(2K^2 e^{K\|\boldsymbol{\beta}_1^0\|_1}) : t \in [0, \tau]\}$ as the hypotheses set. By (16), it suffices to show that the class of functions $\mathcal{G}$ has bounded uniform entropy integral. Since a function of bounded variation can be expressed as the difference

of two increasing functions, it follows from Lemma 9.10 of [15] that $\mathcal{Z}_\ell = \{Z_\ell(t)/K : t \in [0,\tau]\}$ is a VC-hull class associated with a VC class of index 2. Then, by Corollary 2.6.12 of [26], the entropy of $\mathcal{Z}_\ell$ satisfies $\ln N(\mathcal{Z}_\ell, \varepsilon, d_n) = O_p(1/\varepsilon)$. Also, by Example 19.16 of [25], $\mathcal{Y} = \{Y(t) : t \in [0,\tau]\}$ is a VC class and hence has bounded uniform entropy integral. Thus, by Theorem 9.15 of [15], $\mathcal{Y}\mathcal{Z}_\ell\mathcal{Z}_k$ has bounded uniform entropy integral.

It remains to consider the set $\mathcal{H} = \{\exp(\mathbf{Z}(t)^\top\boldsymbol{\beta}^0 - K\|\boldsymbol{\beta}_1^0\|_1),\ t \in [0,\tau]\}$. For any functions $f(t_m) = \exp\{\mathbf{Z}(t_m)^\top\boldsymbol{\beta}^0 - K\|\boldsymbol{\beta}_1^0\|_1\}$, with $t_m \in [0,\tau]$ for $m \in \{1,2\}$, it is easy to check that

$$|f(t_1) - f(t_2)| \le \|\boldsymbol{\beta}_1^0\|_\infty \sum_{j\in S} |Z_j(t_1) - Z_j(t_2)|,$$

and it follows that

$$\ln N(\mathcal{H}, \varepsilon, d_n) \le s_0 \max_{l\in S}\{\ln N(\mathcal{Z}_\ell, \varepsilon/(\|\boldsymbol{\beta}_1^0\|_\infty s_0), d_n)\} = s_0^2 O_p(1/\varepsilon).$$

Then, applying Theorem 9.15 of [15] again, we conclude that the uniform entropy integral of $\mathcal{Y}\mathcal{Z}_\ell\mathcal{Z}_k\mathcal{H}$ is bounded by the order of s_0. Consequently, our desired result follows immediately from (19) and (16). □

We note that Lemma 4 is similar to Lemma C1 in Fang et al. [10], and the main idea for completing the proof is also similar to each other. By contrast, their rates are much sharper with respect to s_0, while it is required that p is diverging and this excludes the fixed dimensional case.

Appendix B. Proof of Theorem 1

Using some basic properties on convex optimization, we next consider the properties of the Lasso estimators defined in (3).

Lemma 5. *(a) A vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is optimal if and only if there exists a subgradient vector $\hat{z} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$ such that*

$$\frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i(s) - \bar{\mathbf{Z}}_n(s,\hat{\boldsymbol{\beta}})\} dN_i(s) - \lambda\hat{z} = 0. \tag{20}$$

(b) Suppose that the subgradient vector satisfies the strict dual feasibility condition $|\hat{z}_j| < 1$ for all $j \notin S(\hat{\boldsymbol{\beta}})$, then any optimal solution $\tilde{\boldsymbol{\beta}}$ to the Lasso program (3) satisfies $\tilde{\beta}_j = 0$ for all $j \notin S(\hat{\boldsymbol{\beta}})$.

(c) Under the conditions of part (b). If the $|S(\hat{\boldsymbol{\beta}})| \times |S(\hat{\boldsymbol{\beta}})|$ matrix $\ddot{\ell}_{S(\hat{\boldsymbol{\beta}})S(\hat{\boldsymbol{\beta}})}(\boldsymbol{\beta})$ is invertible, then $\hat{\boldsymbol{\beta}}$ is the unique optimal solution of the Lasso program.

The above lemma shows that the solution of (3) is unique to the support recovery, provided that the strict dual feasibility condition in (20) holds and the restricted Hessian matrix over $\mathbb{R}^{S(\hat{\boldsymbol{\beta}})}$ is invertible.

Proof of Lemma 5. Equivalently, the convex program (3) can be reformulated as the ℓ_1-constrained minimization, i.e., $\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\ell(\boldsymbol{\beta})$, subject to $\|\boldsymbol{\beta}\|_1 \le C$, where the penalty parameter λ and constraint level C are in one-to-one correspondence via Lagrangian duality. So by Weierstraß's theorem, the minimum is always achieved. Furthermore, by a standard condition for optimality in a convex program on the open set $\mathbb{R}^p$, a point $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ minimizing (3) if and only if there exists a subgradient $\hat{z} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$ such that $\dot{\ell}(\hat{\boldsymbol{\beta}}) + \lambda\hat{z} = 0$. Thus part (a) is derived from (4).

To prove part (b), by standard duality theory [2], given the subgradient $\hat{z} \in \partial\|\hat{\boldsymbol{\beta}}\|_1$, any optimal solution $\tilde{\boldsymbol{\beta}}$ to the group Lasso must satisfy the complementary slackness condition $\hat{z}^\top\tilde{\boldsymbol{\beta}} = \|\tilde{\boldsymbol{\beta}}\|_1$. Since $|\hat{z}_k| \le 1$ for all k, $|\hat{z}_j| < 1$ (for some j) implies $\tilde{\beta}_j = 0$. This establishes Lemma 5(b).

Finally, since the invertibility of $\ddot{\ell}_{S(\hat{\boldsymbol{\beta}})S(\hat{\boldsymbol{\beta}})}(\boldsymbol{\beta})$ implies that $\ddot{\ell}_{S(\hat{\boldsymbol{\beta}})S(\hat{\boldsymbol{\beta}})}(\boldsymbol{\beta})$ is strictly positive definite, then when restricted to vectors of the form $(\boldsymbol{\beta}_{S(\hat{\boldsymbol{\beta}})}, \mathbf{0})$, the group-Lasso program (3) is strictly convex, and so its optimum is uniquely determined, which ultimately yields part (c). □

Lemma 6. *(a) If Assumption 1(3) holds and $\lambda \ge \ln(n)/n$ is satisfied, all the optimum of the restricted program (5) stay within a bounded domain of $\mathbb{R}^S$.*

(b) If Assumptions 1–3 hold, then for any bounded vector $\mathbf{b}_S \in \mathbb{R}^S$, $\ddot{\ell}_{SS}\{\boldsymbol{\beta}^0 + (\mathbf{b}_S, \mathbf{0})\}$ is invertible.

The part(a) of Lemma 6 tells us that it is enough to consider all the possible solutions of (5) within a bounded domain, not the whole space $\mathbb{R}^S$. Furthermore, under additional conditions, part (b) of Lemma 6 implies that the solution of (5) is unique. This serves the proof of part (a) of Theorem 1.

Proof of Lemma 6. To prove part (a) of Lemma 6, by minimization definition of $\check{\boldsymbol{\beta}}_S$, we have that

$$\ell\{(\check{\boldsymbol{\beta}}_S, \mathbf{0})\} + \lambda\|\check{\boldsymbol{\beta}}_S\|_1 \le \ell(\mathbf{0}) = \frac{1}{n}\int_0^\tau \ln\left\{\sum_{i=1}^n Y_i(s)\right\} d\bar{N}(s), \tag{21}$$

where $\bar{N}(s) = N_1(s) + \cdots + N_n(s)$. Since no two counting processes N_i jump at the same time, we have $|d\bar{N}(s)| = |dN_1(s) + \cdots + dN_n(s)| \le 1$, where $dN_i(s) = N_i(t) - N_i(t^-)$ denotes the jump of N_i at time t. By (21), it follows that $\|\check{\boldsymbol{\beta}}_S\|_1 \le \tau \ln(n)/(n\lambda) < \infty$, provided that $\lambda \ge \ln(n)/n$. On the other hand, under Assumption 2 and $\Pr[\max_i\{N_i(\tau)\} \le 1] = 1$, by Lemma 7, we have $\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\|_2 = o_p(1)$ provided that $s_0^4 \ll n$. Then, repeating the process as that from (36) to (37), this together with Assumption 3 implies that $\ddot{\ell}(\boldsymbol{\beta}^0)_{SS}$ is invertible with $\|\{\ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\}^{-1}\|_2 = O_p(1)$. Since $\ddot{\ell}(\boldsymbol{\beta}^0)_{SS}$ is symmetric, we see that $\ddot{\ell}(\boldsymbol{\beta}^0)_{SS}$ is strictly positive definite. Besides, over the restricted set $\mathbb{R}^S$, from Lemma 3.2 of [12], we see that

$$\ddot{\ell}\{\boldsymbol{\beta}^0 + (\mathbf{b}_S, \mathbf{0})\}_{SS} - e^{-2\eta_b}\ddot{\ell}(\boldsymbol{\beta}^0)_{SS} \text{ is nonnegative-definite,}$$

where $\eta_b = \max_t \max_{ij} |\mathbf{b}_S^\top \mathbf{Z}_{iS}(t) - \mathbf{b}_S^\top \mathbf{Z}_{jS}(t)|$. Since $\mathbf{b}_S$ and $\mathbf{Z}_S$ are both bounded by assumption, $e^{-2\eta_b} > 0$ holds. This further implies that $\ddot{\ell}\{\boldsymbol{\beta}^0 + (\mathbf{b}_S, \mathbf{0})\}_{SS}$ is also strictly positive definite. Note that $\ddot{\ell}\{\boldsymbol{\beta}^0 + (\mathbf{b}_S, \mathbf{0})\}_{SS} = \ddot{\ell}|_S(\boldsymbol{\beta}_1^0 + \mathbf{b}_S)$, and then part (b) of Lemma 6 is proved. □

Based on the conclusions derived in Lemmas 5 and 6, we can complete the proof of Theorem 1.

Proof of Theorem 1. Since $\check{\boldsymbol{\beta}}_S$ is an interior point over $\mathbb{R}^S$, it must be a zero-gradient point for the restricted program (5), hence $(\dot{\ell})|_S\{(\check{\boldsymbol{\beta}}_S, \mathbf{0})\} + \lambda \check{z}_S = 0$. By the chain rule, this implies that $\{\dot{\ell}(\check{\boldsymbol{\beta}})\}_S + \lambda \check{z}_S = 0$, where $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_S, 0_{S^c})$. Besides, since $\max_{j\in S^c} |\check{z}_j| \le 1$ by assumption, we treat $\check{z}$ as one extended subgradient of $\check{\boldsymbol{\beta}}$. Then, $\check{\boldsymbol{\beta}}$ is an optimal point of the group-Lasso scheme (3) based on part (a) of Lemma 5. Furthermore, since $S(\check{\boldsymbol{\beta}}) \subseteq S$ be definition, part (b) of Lemma 6 implies that $\ddot{\ell}_{S(\check{\boldsymbol{\beta}})S(\check{\boldsymbol{\beta}})}(\check{\boldsymbol{\beta}})$ is also invertible. Thus, by parts (b) and (c) of Lemma 5, we conclude that $\check{\boldsymbol{\beta}}$ is the unique solution of (3) satisfying $S(\check{\boldsymbol{\beta}}) \subseteq S$.

If additionally, the sign consistency condition in Step 4 is satisfied. Then since $\check{z}_S$ was chosen as an element of the subdifferential $\partial\|\check{\boldsymbol{\beta}}_S\|_1$ in Step 2, we must have $\operatorname{sign}(\check{\boldsymbol{\beta}}_S) = \operatorname{sign}(\boldsymbol{\beta}_1^o)$, provided that $\check{\beta}_j \ne 0$ for all $j \in S$. Then this implies Lemma 1(b) by noting that $S(\check{\boldsymbol{\beta}}) = S(\boldsymbol{\beta}^0)$.

To prove part (c), it suffices to prove the following equivalent assertion: if there exists a group-Lasso solution $\hat{\boldsymbol{\beta}}$ with $S(\hat{\boldsymbol{\beta}}) = S$ and $\operatorname{sign}(\hat{\boldsymbol{\beta}}_S) = \operatorname{sign}(\boldsymbol{\beta}_1^o)$, then the PDW method succeeds in producing a dual feasible vector $\check{z}$ with $\check{z}_S = \operatorname{sign}(\boldsymbol{\beta}_1^o)$. First, by (4), it is easy to verify that $\dot{\ell}(\hat{\boldsymbol{\beta}})_S = (\dot{\ell})|_S(\hat{\boldsymbol{\beta}}_S)$. So $\hat{\boldsymbol{\beta}}_S$ is an optimal point to the restricted program (5). Also note that $\ddot{\ell}_{SS}\{(\boldsymbol{\beta}_S, \mathbf{0})\}$ is invertible as proved in part (b) of Lemma 6, the vector $\hat{\boldsymbol{\beta}}_S$ must be the unique solution to (5). Note that $\operatorname{sign}(\hat{\boldsymbol{\beta}}_S) = \operatorname{sign}(\boldsymbol{\beta}_1^o)$ by condition, and we conclude that $\check{z}_S = \operatorname{sign}(\hat{\boldsymbol{\beta}}_S)$ is only subgradient that can be chosen in Step 2. Since $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S, \mathbf{0})$ is an optimal Lasso solution by assumption, then there must exist a dual feasible vector $\check{z}_{S^c}$ such that $\{\operatorname{sign}(\hat{\boldsymbol{\beta}}_S), \check{z}_{S^c}\}$ satisfies the zero subgradient condition (6). □

Appendix C. Proof for the concentration of Hessian matrix

Lemma 7. *Under Assumptions 2 and 1(2), the following inequality holds with probability at least* $1 - \delta$,

$$\left\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\right\|_2 = O_p\left\{s_0^2\sqrt{\ln(s_0^2/\delta)/n}\right\}.$$

Proof. Let $L_{\ell k}$ and $\boldsymbol{\Sigma}_{\ell k}$ be the (ℓ, k)th entries of the matrices $\ddot{\ell}(\boldsymbol{\beta}^0)$ and $\boldsymbol{\Sigma}$, respectively. First of all, we write

$$\begin{aligned} L_{\ell k} - \boldsymbol{\Sigma}_{\ell k} &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left\{\frac{S_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)}{S^{(0)}(t, \boldsymbol{\beta}^0)}\right\} dN_i(t) - \mathrm{E}\int_0^\tau \left\{\frac{s_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)}{s^{(0)}(t, \boldsymbol{\beta}^0)}\right\} dN(t) \\ &\quad - \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left[\frac{S_\ell^{(1)}(t, \boldsymbol{\beta}^0) S_k^{(1)}(t, \boldsymbol{\beta}^0)}{\{S^{(0)}(t, \boldsymbol{\beta}^0)\}^2}\right] dN_i(t) + \mathrm{E}\int_0^\tau \left[\frac{s_\ell^{(1)}(t, \boldsymbol{\beta}^0) s_k^{(1)}(t, \boldsymbol{\beta}^0)}{\{s^{(0)}(t, \boldsymbol{\beta}^0)\}^2}\right] dN(t) \\ &\equiv T_1 - T_2. \end{aligned}$$

To bound term T_1, note that

$$\frac{S_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)}{S^{(0)}(t, \boldsymbol{\beta}^0)} - \frac{s_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)}{s^{(0)}(t, \boldsymbol{\beta}^0)} = \frac{S_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0) - s_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)}{S^{(0)}(t, \boldsymbol{\beta}^0)} - \frac{s_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)\{S^{(0)}(t, \boldsymbol{\beta}^0) - s^{(0)}(t, \boldsymbol{\beta}^0)\}}{S^{(0)}(t, \boldsymbol{\beta}^0) s^{(0)}(t, \boldsymbol{\beta}^0)}. \tag{22}$$

Since $s^{(0)}(t, \boldsymbol{\beta}^0)/e^{K\|\boldsymbol{\beta}_1^0\|_1}$ is bounded away from zero by Assumption 2(3), $S^{(0)}(t, \boldsymbol{\beta}^0)/e^{K\|\boldsymbol{\beta}_1^0\|_1}$ is also bounded away from zero by (17). Then, from (17) and (18) in Lemma 4, we have

$$\sup_t \left|\frac{S_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)}{S^{(0)}(t, \boldsymbol{\beta}^0)} - \frac{s_{\ell k}^{(2)}(t, \boldsymbol{\beta}^0)}{s^{(0)}(t, \boldsymbol{\beta}^0)}\right| \le c_0 s_0 \sqrt{\ln(4/\delta)/n} \tag{23}$$

with probability at least $1-\delta/2$. Write

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{\frac{S_{\ell k}^{(2)}(t,\boldsymbol{\beta}^0)}{S^{(0)}(t,\boldsymbol{\beta}^0)} - \frac{s_{\ell k}^{(2)}(t,\boldsymbol{\beta}^0)}{s^{(0)}(t,\boldsymbol{\beta}^0)}\right\} dN_i(t) + (P_n - P)\int_0^\tau \left\{\frac{s_{\ell k}^{(2)}(t,\boldsymbol{\beta}^0)}{s^{(0)}(t,\boldsymbol{\beta}^0)}\right\} dN(t) \equiv T_{11} + T_{12},$$

Since $\Pr\big[\max_i\{N_i(\tau)\} \le 1\big] = 1$, it follows from (23) that $|T_{11}| = O_p\{s_0\sqrt{\ln(2/\delta)/n}\}$. Besides, note that the term T_{22} is an iid and bounded sum. Thus an application of Hoeffding inequality [11] yields that, with probability at least $1-\delta/2$,

$$|T_{12}| = O_p\left\{\sqrt{\ln(4/\delta)/n}\right\}.$$

Putting the bounds for T_{11} and T_{12} together, with probability at least $1-\delta$, we have $|T_1| = O_p\{s_0\sqrt{\ln(2/\delta)/n}\}$. Similarly, we can rewrite T_2 as

$$\begin{aligned} T_2 = {} & \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\frac{S_\ell^{(1)}(t,\boldsymbol{\beta}^0)S_k^{(1)}(t,\boldsymbol{\beta}^0)}{\{S^{(0)}(t,\boldsymbol{\beta}^0)\}^2} - \frac{s_i^{(1)}(t,\boldsymbol{\beta}^0)s_j^{(1)}(t,\boldsymbol{\beta}^0)}{\{s^{(0)}(t,\boldsymbol{\beta}^0)\}^2}\right] dN_i(t) \\ & + (P_n - P)\int_0^\tau \left[\frac{s_\ell^{(1)}(t,\boldsymbol{\beta}^0)s_k^{(1)}(t,\boldsymbol{\beta}^0)}{\{s^{(0)}(t,\boldsymbol{\beta}^0)\}^2}\right] dN(t) \equiv T_{21} + T_{22}. \end{aligned}$$

To bound term T_{21}, note that

$$\begin{aligned} & \frac{S_\ell^{(1)}(t,\boldsymbol{\beta}^0)S_k^{(1)}(t,\boldsymbol{\beta}^0)}{\{S^{(0)}(t,\boldsymbol{\beta}^0)\}^2} - \frac{s_\ell^{(1)}(t,\boldsymbol{\beta}^0)s_k^{(1)}(t,\boldsymbol{\beta}^0)}{\{s^{(0)}(t,\boldsymbol{\beta}^0)\}^2} = \frac{S_k^{(1)}(t,\boldsymbol{\beta}^0)}{\{S^{(0)}(t,\boldsymbol{\beta}^0)\}^2}\left\{S_\ell^{(1)}(t,\boldsymbol{\beta}^0) - s_\ell^{(1)}(t,\boldsymbol{\beta}^0)\right\} \\ & + \frac{s_\ell^{(1)}(t,\boldsymbol{\beta}^0)}{\{S^{(0)}(t,\boldsymbol{\beta}^0)\}^2}\left\{S_k^{(1)}(t,\boldsymbol{\beta}^0) - s_k^{(1)}(t,\boldsymbol{\beta}^0)\right\} - \frac{s_\ell^{(1)}(t,\boldsymbol{\beta}^0)s_k^{(1)}(t,\boldsymbol{\beta}^0)}{\{S^{(0)}(t,\boldsymbol{\beta}^0)s^{(0)}(t,\boldsymbol{\beta}^0)\}^2}\left[\left\{S^{(0)}(t,\boldsymbol{\beta}^0)\right\}^2 - \left\{s^{(0)}(t,\boldsymbol{\beta}^0)\right\}^2\right]. \end{aligned}$$

By the same arguments as in the proof of (22), it follows that

$$\left|\frac{S_\ell^{(1)}(t,\boldsymbol{\beta}^0)S_k^{(1)}(t,\boldsymbol{\beta}^0)}{\{S^{(0)}(t,\boldsymbol{\beta}^0)\}^2} - \frac{s_\ell^{(1)}(t,\boldsymbol{\beta}^0)s_k^{(1)}(t,\boldsymbol{\beta}^0)}{\{s^{(0)}(t,\boldsymbol{\beta}^0)\}^2}\right| \le c_0 s_0\sqrt{\ln(4/\delta)/n}$$

with probability at least $1-\delta/2$. This further implies that $|T_{21}| = O_p\{s_0\sqrt{\ln(4/\delta)/n}\}$. Also, note that the term T_{22} is an iid and bounded sum, an application of Hoeffding inequality yields that, with probability at least $1-\delta/2$, $|T_{22}| = O_p\{\sqrt{\ln(4/\delta)/n}\}$. Then, combining the bounds for T_{21} and T_{22} yields that

$$T_2 = O_p\left\{s_0\sqrt{\ln(4/\delta)/n}\right\}.$$

Finally, putting the bounds for T_1 and T_2 together tells us that, with probability at least $1-\delta$,

$$\left\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\right\|_{\max} = O_p\left\{s_0\sqrt{\ln(s_0^2/\delta)/n}\right\}, \tag{24}$$

where $\|\cdot\|_{\max}$ is denoted to be the elementwise norm for a matrix. Note that $\|M\|_2 \le s_0\|M\|_{\max}$ for any $M \in \mathbb{R}^{s_0\times s_0}$, the desired inequality follows from (24) immediately. □

We here introduce the following useful lemma on matrices [19].

Lemma 8. *Let $A,\ B \in \mathbb{R}^p$ be invertible. For any matrix norm $\|\cdot\|$, there holds*

$$\|A^{-1} - B^{-1}\| \le \frac{\|A^{-1}\|^2\cdot\|A-B\|}{1 - \|A^{-1}\|\cdot\|A-B\|}.$$

Note that, the upper bound of $\|A^{-1}-B^{-1}\|$ provided in Lemma 8 does not involve B^{-1}, which may give a better constant in some cases.

Proof of Proposition 2. First of all, by the triangle inequality, we have that

$$\|\widehat{\mathbb{Q}}_{SS} - \boldsymbol{\Sigma}_{SS}\|_2 \le \|\widehat{\mathbb{Q}}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\|_2 + \|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\|_2. \tag{25}$$

Since the second term of (25) has been studied in Lemma 7, it suffices to bound the first one. Recalling that $\ddot{\ell}(\boldsymbol{\beta}) = 1/n\int_0^\tau V_n(s,\boldsymbol{\beta})d\bar{N}(s)$, we have

$$\begin{aligned} \widehat{\mathbb{Q}} - \ddot{\ell}(\boldsymbol{\beta}^0) &= \int_0^1 \left[\ddot{\ell}\left\{\boldsymbol{\beta}^0 + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\right\} - \ddot{\ell}(\boldsymbol{\beta}^0)\right] d\theta \\ &= 1/n\int_0^1\int_0^\tau \left[V_n\{s,\boldsymbol{\beta}^0 + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\} - V_n(s,\boldsymbol{\beta}^0)\right] d\bar{N}(s)d\theta. \end{aligned} \tag{26}$$

Then, for any unit vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$,

$$\mathbf{a}^\top V_n(t, \boldsymbol{\beta})\mathbf{b} = \sum_{i=1}^n w_{in}(t, \boldsymbol{\beta})\mathbf{a}^\top \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta})\} \times \mathbf{b}^\top \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta})\}.$$

Following the above formulation, for any $\boldsymbol{\delta} \in \mathbb{R}^p$, we have

$$\mathbf{a}^\top \{V_n(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta}) - V_n(t, \boldsymbol{\beta}^0)\}\mathbf{b} \equiv I_1 - I_2 - I_3 + I_4, \tag{27}$$

where

$$\begin{aligned}
I_1 &= \sum_{i=1}^n \{w_{in}(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta}) - w_{in}(t, \boldsymbol{\beta}^0)\}\mathbf{a}^\top \mathbf{Z}_i(t)\mathbf{b}^\top \mathbf{Z}_i(t),\\
I_2 &= \sum_{i=1}^n w_{in}(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta})\mathbf{a}^\top \mathbf{Z}_i(t)\mathbf{b}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta}) - \sum_{i=1}^n w_{in}(t, \boldsymbol{\beta}^0)\mathbf{a}^\top \mathbf{Z}_i(t)\mathbf{b}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0),\\
I_3 &= \sum_{i=1}^n w_{in}(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta})\mathbf{b}^\top \mathbf{Z}_i(t)\mathbf{a}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta}) - \sum_{i=1}^n w_{in}(t, \boldsymbol{\beta}^0)\mathbf{b}^\top \mathbf{Z}_i(t)\mathbf{a}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0),\\
I_4 &= \sum_{i=1}^n w_{in}(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta})\mathbf{a}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta})\mathbf{b}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0 + \boldsymbol{\delta}) - \sum_{i=1}^n w_{in}(t, \boldsymbol{\beta}^0)\mathbf{a}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0)\mathbf{b}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0).
\end{aligned}$$

Now we consider I_1. To simplify expression, let $\delta_i = \delta_i(t) = \exp\{\mathbf{Z}_i(t)^\top \boldsymbol{\delta}\}$ and $w_i = w_i(t) = Y_i(t)\exp\{\mathbf{Z}_i(t)^\top \boldsymbol{\beta}^0 - K\|\boldsymbol{\beta}_1^0\|_1\}$. Since $\mathbf{Z}_i(t)$ is uniformly bounded by K, $\max_{i,t}[\{\mathbf{Z}_i(t)\}^\top \boldsymbol{\beta}^0] \le K\|\boldsymbol{\beta}_1^0\|_1$, which guarantees that $0 < w_i \le 1$ for all i and t. In this case, I_1 can be rewritten as

$$I_1 = \sum_{i=1}^n \left\{ \frac{w_i \sum_{k\neq i} w_k(\delta_i - \delta_k)}{\sum_{k,\ell=1} w_k\delta_k w_\ell} \right\} \mathbf{a}^\top \mathbf{Z}_i(t)\mathbf{b}^\top \mathbf{Z}_i(t).$$

Similarly, $I_2 - I_4$ can be rewritten as the following formulas,

$$\begin{aligned}
I_2 &= \sum_{i=1}^n \left\{ \frac{w_i \sum_{k\neq i} w_k(\delta_i - \delta_k)}{\sum_{k,\ell=1} w_k\delta_k w_\ell} \right\} \mathbf{a}^\top \mathbf{Z}_i(t)\mathbf{b}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0) + \left\{ \frac{\sum_{i=1}^n w_i \mathbf{a}^\top \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k} \right\} \times \left\{ \frac{\sum_{i=1}^n w_i\delta_i \mathbf{b}^\top \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k\delta_k} - \frac{\sum_{i=1}^n w_i \mathbf{b}^\top \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k} \right\},\\
I_3 &= \sum_{i=1}^n \left\{ \frac{w_i \sum_{k\neq i} w_k(\delta_i - \delta_k)}{\sum_{k,\ell=1} w_k\delta_k w_\ell} \right\} \mathbf{b}^\top \mathbf{Z}_i(t)\mathbf{a}^\top \bar{\mathbf{Z}}_n(t, \boldsymbol{\beta}^0) + \left\{ \frac{\sum_{i=1}^n w_i \mathbf{b}^\top \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k} \right\} \times \left\{ \frac{\sum_{i=1}^n w_i\delta_i \mathbf{a}^\top \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k\delta_k} - \frac{\sum_{i=1}^n w_i \mathbf{a}^\top \mathbf{Z}_i(t)}{\sum_{k=1}^n w_k} \right\},\\
I_4 &= \frac{\left\{\sum_{i=1}^n w_i\delta_i \mathbf{a}^\top \mathbf{Z}_i(t)\right\}\left\{\sum_{i=1}^n w_i\delta_i \mathbf{b}^\top \mathbf{Z}_i(t)\right\}}{\left(\sum_{k=1}^n w_k\delta_k\right)^2} - \frac{\left\{\sum_{i=1}^n w_i \mathbf{a}^\top \mathbf{Z}_i(t)\right\}\left\{\sum_{i=1}^n w_i \mathbf{b}^\top \mathbf{Z}_i(t)\right\}}{\left(\sum_{k=1}^n w_k\right)^2}.
\end{aligned}$$

Denote $\gamma_\delta = \gamma_\delta(t) = \max_{i,j}\{\mathbf{Z}_i(t) - \mathbf{Z}_j(t)\}^\top \boldsymbol{\delta} > 0$, and a direct computation yields that

$$|I_1| \le \exp(\gamma_\delta)\gamma_\delta \sum_{i=1}^n \left\{ w_i \left|\mathbf{a}^\top \mathbf{Z}_i(t)^{\otimes 2}\mathbf{b}\right| \right\} \Big/ \sum_{i=1}^n w_i. \tag{28}$$

Following Assumption 2(1), we have that $\gamma_\delta \le K\|\boldsymbol{\delta}\|_1$. Thus, it remains for bounding I_1 to estimate $\sum_{i=1}^n\{w_i\,|\mathbf{a}^\top \mathbf{Z}_i(t)^{\otimes 2}\mathbf{b}\,|\}/\sum_{i=1} w_i$, which can be solved by estimating its numerator and denominator, respectively.

Note that taking a supremum over unit vectors $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^S$, by definition we have

$$\sup_{\|\mathbf{a}\|_2=1, \|\mathbf{b}\|_2=1} \left\{ \frac{1}{n}\sum_{i=1}^n w_i \left|\mathbf{a}^\top \mathbf{Z}_i(t)^{\otimes 2}\mathbf{b}\right| \right\} = \left\| \frac{1}{n}\sum_{i=1}^n w_i \{\mathbf{Z}_i(t)^{\otimes 2}\}_{SS} \right\|_2,$$

which (28) further implies that

$$\tilde{I}_1(t) = \sup_{\|\mathbf{a}\|_2=1, \|\mathbf{b}\|_2=1} \{I_1\} \le K\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1 \times \left\| \frac{1}{n}\sum_{i=1}^n w_i \{\mathbf{Z}_i(t)^{\otimes 2}\}_{SS} \right\|_2 \Big/ \left(\sum_{i=1}^n w_i/n \right). \tag{29}$$

Note that by the triangle inequality, one gets

$$\left\| \frac{1}{n}\sum_{i=1}^n w_i \{\mathbf{Z}_i(t)^{\otimes 2}\}_{SS} \right\|_2 \le \left\| \frac{1}{n}\sum_{i=1}^n w_i \{\mathbf{Z}_i(t)^{\otimes 2}\}_{SS} - \mathrm{E}\left[w\{\mathbf{Z}(t)^{\otimes 2}\}_{SS} \right] \right\|_2 + \left\| \mathrm{E}\left[w\{\mathbf{Z}(t)^{\otimes 2}\}_{SS} \right] \right\|_2.$$

and by definition we have

$$\Big\|\frac{1}{n}\sum_{i=1}^{n} w_i\big\{\mathbf{Z}_i(t)^{\otimes 2}\big\}_{SS} - \mathrm{E}[w\big\{\mathbf{Z}(t)^{\otimes 2}\big\}_{SS}]\Big\|_2 \le \sqrt{s_0}\Big\|\frac{1}{n}\sum_{i=1}^{n} w_i\big\{\mathbf{Z}_i(t)^{\otimes 2}\big\}_{SS} - \mathrm{E}[w\big\{\mathbf{Z}(t)^{\otimes 2}\big\}_{SS}]\Big\|_{\max},$$

where $\|\cdot\|_{\max}$ is defined as that in Lemma 7. Moreover, by the same arguments as that for the proof of Lemma 4, with probability at least $1-\delta$, the following two inequalities both hold

$$\Big|\frac{1}{n}\sum_{i=1}^{n} w_i - \mathrm{E}(w)\Big| = O_p\Big\{\sqrt{\ln(4/\delta)/n}\Big\}, \tag{30}$$

$$\Big\|\frac{1}{n}\sum_{i=1}^{n} w_i\big\{\mathbf{Z}_i(t)^{\otimes 2}\big\}_{SS} - \mathrm{E}\Big[w\big\{\mathbf{Z}(t)^{\otimes 2}\big\}_{SS}\Big]\Big\|_2 = O_p\Big\{s_0^{1/2}\sqrt{\ln(s_0^2/\delta)/n}\Big\} \tag{31}$$

for all $t \in [0,\tau]$. Thus, since $\mathrm{E}(w) > 0$ by Assumption 2(3), from (30) we see that $|w_1+\cdots+W+n|/n$ is bounded away from zero with high probability. Thus, plugging (30) and (31) into (29), we obtain that

$$\sup_t\{\tilde{I}_1(t)\} = O_p\Big\{\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1\Big\},$$

since $\mathrm{E}[\|\{\mathbf{Z}(t)^{\otimes 2}\}_{SS}\|_2]$ is bounded uniformly by assumption. Besides, a similar argument shows that, for all $t \in [0,\tau]$,

$$\tilde{I}_2(t) = \sup_{\|\mathbf{a}\|_2=1,\|\mathbf{b}\|_2=1} I_2(t) = O_p\Big\{\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1\Big\},$$

$$\tilde{I}_3(t) = \sup_{\|\mathbf{a}\|_2=1,\|\mathbf{b}\|_2=1} I_3(t) = O_p\Big\{\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1\Big\},$$

$$\tilde{I}_4(t) = \sup_{\|\mathbf{a}\|_2=1,\|\mathbf{b}\|_2=1} I_4(t) = O_p\Big\{\exp(K\|\boldsymbol{\delta}\|_1)\|\boldsymbol{\delta}\|_1\Big\}.$$

Then, setting $\boldsymbol{\delta} = \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$, and combining with (26), (27) and all the derived bounds of $\tilde{I}_1 - \tilde{I}_4$, we have that

$$\|[V_n\{s,\boldsymbol{\beta}^0 + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\} - V_n(s,\boldsymbol{\beta}^0)]_{SS}\|_2 = O_p\{\exp(K\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1)\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1\}, \tag{32}$$

for all $s \in [0,\tau]$ and $\theta \in [0,1]$. Also, taking a supremum over unit vectors $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^S$, we see from (26) that

$$\|\widehat{Q}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\|_2 \le 1/n\int_0^1\int_0^\tau \Big\|\big[V_n\{s,\boldsymbol{\beta}^0 + \theta(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\} - V_n(s,\boldsymbol{\beta}^0)\big]_{SS}\Big\|_2 d\bar{N}(s)d\theta. \tag{33}$$

Since $N_i(\tau) \le 1$ for all $i \le n$, it follows from (33), (32) and the result of Lemma 1 that

$$\|\widehat{Q}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\|_2 = s_0 O_p\Big\{\sqrt{\ln(p/\delta)/n}\Big\}, \tag{34}$$

which converges to zero as $n \to \infty$ and $s_0^2 \ll n$.

Furthermore, by Lemma 7 we recall that

$$\|\boldsymbol{\Sigma}_{SS} - \ddot{\ell}(\boldsymbol{\beta}^0)_{SS}\|_2 = O_p\Big\{s_0^{3/2}\sqrt{\ln(s_0^2/\delta)/n}\Big\}, \tag{35}$$

with probability at least $1-\delta/2$. Therefore, combining (34), (35) with (25), we obtain that

$$\|\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\|_2 = O_p\Big\{s_0^{3/2}\sqrt{\ln(s_0^2/\delta)/n}\Big\}. \tag{36}$$

Since $\|(\boldsymbol{\Sigma}_{SS})^{-1}\|_2 = O_p(1)$ by Assumption 3, we still need to show that $\|(\widehat{Q}_{SS})^{-1}\|_2 = O_p(1)$ with high probability, so that applying Lemma 8 with the $\|\cdot\|_2$-norm yields our desired result. To this end, we decompose $(\widehat{Q}_{SS})^{-1}$ as

$$(\widehat{Q}_{SS})^{-1} = (\boldsymbol{\Sigma}_{SS})^{-1/2}\big\{I + (\boldsymbol{\Sigma}_{SS})^{-1/2}\big(\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\big)(\boldsymbol{\Sigma}_{SS})^{-1/2}\big\}^{-1}(\boldsymbol{\Sigma}_{SS})^{-1/2},$$

and let $\mathcal{A} = I + (\boldsymbol{\Sigma}_{SS})^{-1/2}\big(\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\big)(\boldsymbol{\Sigma}_{SS})^{-1/2}$. Then $(\widehat{Q}_{SS})^{-1} = (\boldsymbol{\Sigma}_{SS})^{-1/2}\mathcal{A}^{-1}(\boldsymbol{\Sigma}_{SS})^{-1/2}$. By the Bauer–Fike inequality, we have

$$|\lambda(\mathcal{A}) - 1| \le \|(\boldsymbol{\Sigma}_{SS})^{-1/2}\big(\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\big)(\boldsymbol{\Sigma}_{SS})^{-1/2}\|_2 \le \|(\boldsymbol{\Sigma}_{SS})^{-1/2}\|_2^2\|\widehat{Q}_{SS} - \boldsymbol{\Sigma}_{SS}\|_2.$$

Then by (36) and Assumption 3, $|\lambda(\mathcal{A}) - 1| = o_p(1)$. Hence $\lambda(\mathcal{A}^{-1}) = 1 + o_p(1)$. Since $\mathcal{A}$ is symmetrical, $\|\mathcal{A}^{-1}\|_2 = O_p(1)$. This together with Assumption 2 yields that

$$\|(\widehat{Q}_{SS})^{-1}\|_2 \le \|(\boldsymbol{\Sigma}_{SS})^{-1/2}\|_2\|\mathcal{A}^{-1}\|_2\|(\boldsymbol{\Sigma}_{SS})^{-1/2}\|_2 = O_p(1). \tag{37}$$

As a result, by (36) and (37), the first part of Proposition 2 follows easily from Lemma 8.

Next, we turn to the second term of Proposition 2, i.e., verify expression (12). Under Assumption 4, we first note that

$$\begin{aligned}\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty &\le \big\|\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big\|_\infty + \big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1} - \mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big\|_\infty \\ &\le 1-\gamma + \big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1} - \mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big\|_\infty. \end{aligned} \tag{38}$$

Then it suffices to show that $\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1} - \mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\|_\infty \le \gamma/2$. For this purpose, let $T = \widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1} - \mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}$, and we split T into two parts, viz. $T = T_1 + T_2$, where $T_1 = \big(\widehat{Q}_{S^cS} - \mathbf{\Sigma}_{S^cS}\big)(\widehat{Q}_{SS})^{-1}$, and $T_2 = \mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\big(\widehat{Q}_{SS} - \mathbf{\Sigma}_{SS}\big)(\widehat{Q}_{SS})^{-1}$.

Similarly as before, by the union bound, Lemma 4 and the result of Theorem 1, we have

$$\max_{j\in S^c}\big\|e_j^\top\big(\widehat{Q}_{S^cS} - \mathbf{\Sigma}_{S^cS}\big)\big\|_2 = O_p\Big\{s_0^{3/2}\sqrt{\ln(p/\delta)/n}\Big\}.$$

Since by (37), $\|(\widehat{Q}_{SS})^{-1}\|_2$ can be treated as positive constant. Then if $s_0^3 \ll n$, we have

$$\|T_1\|_\infty \le \max_{j\in S^c}\big\|e_j^\top\big(\widehat{Q}_{S^cS} - \mathbf{\Sigma}_{S^cS}\big)\big\|_2 \times \|(\widehat{Q}_{SS})^{-1}\|_2 = O_p\Big\{s_0^{3/2}\sqrt{\ln(p/\delta)/n}\Big\}. \tag{39}$$

To bound T_2, by Assumption 4 and (37), we have

$$\|T_2\|_\infty \le \|\mathbf{\Sigma}_{S^cS}(\mathbf{\Sigma}_{SS})^{-1}\|_\infty\|(\widehat{Q}_{SS})^{-1}\|_\infty\|\widehat{Q}_{SS} - \mathbf{\Sigma}_{SS}\|_\infty = \sqrt{s_0}O_p\big(\|\widehat{Q}_{SS} - \mathbf{\Sigma}_{SS}\|_2\big), \tag{40}$$

since $\|A\|_\infty \le \sqrt{s_0}\,\|A\|_2$ for any matrix $A \in \mathbb{R}^{S\times S}$. Then by bounds (36) and (40), we have

$$\|T_2\|_\infty = O_p\Big\{s_0^2\sqrt{\ln(s_0^2/\delta)/n}\Big\}. \tag{41}$$

So combined with inequalities (38), (39) and (41), we have

$$\big\|\widehat{Q}_{S^cS}(\widehat{Q}_{SS})^{-1}\big\|_\infty \le 1-\gamma + \|T_1\|_\infty + \|T_2\|_\infty \le 1-\gamma/2,$$

provided that $\{s_0^2\sqrt{\ln(p/\delta)/n}\} = o_p(1)$. □

Appendix D. Proof of Lemma 1

Since the estimation error established in Theorem 3.1 of [12] holds under the particular setting $p = s_0$, we can apply this result over $\mathbb{R}^S$ directly. To be precise, by Theorem 3.1 of [12], in the event $\|\dot{\ell}(\boldsymbol{\beta}_1^0)\|_\infty \le \lambda$, we have

$$\|\check{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_1^0\|_2 = O_p\Big\{\sqrt{s_0}\lambda\big/F_2(\xi, S)\Big\},$$

where $F_2(\xi, S)$ has been defined in Eq. (3.4) over there. Moreover, with high probability we have

$$F_2(\xi, S) \ge \lambda_{\min}\big\{\ddot{\ell}(\boldsymbol{\beta}_1^0)\big\} = \|\{\ddot{\ell}(\boldsymbol{\beta}_1^0)\}^{-1}\|_2 \ge 1/2\|(\mathbf{\Sigma}_{SS})^{-1}\|_2 = O_p(1),$$

where the first inequality follows from the definition of $F_2(\xi, S)$, and the second inequality follows from the results of Lemmas 7 and 8, and the last one holds from Assumption 3. In addition, if $\Pr\{\max_{i\le n}N_i(\tau) \le 1\} = 1$, as mentioned earlier, we have taken $\lambda = \sqrt{\ln(p/\delta)/n}$, which completes the proof of Lemma 1. □

References

[1] P.K. Andersen, R.D. Gill, Cox's regression model for counting processes: A large sample study, Ann. Statist. 10 (1982) 1100–1120.
[2] D.P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, MA, 1995.
[3] J. Bradic, J. Fan, J. Jiang, Regularization for Cox's proportional hazards model with NP-dimensionality, Ann. Statist. 39 (2011) 3092–3120.
[4] S. Chen, D.L. Donoho, M. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1998) 33–61.
[5] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B 34 (1972) 187–220.
[6] V.H. De La Peña, A general class of exponential inequalities for martingales and ratios, Ann. Probab. 27 (1999) 537–564.
[7] R.A. DeVore, G.G. Lorentz, Constructive Approximation, Springer, New York, 1993.
[8] D.L. Donoho, Compressing sensing, IEEE Trans. Inform. Theory 52 (2006) 1289–1306.
[9] Y. Fan, C.Y. Tang, Tuning parameter selection in high dimensional penalized likelihood, J. R. Stat. Soc. Ser. B 75 (2013) 531–552.
[10] E.X. Fang, Y. Ning, H. Liu, Testing and confidence intervals for high dimensional proportional hazards model, J. R. Stat. Soc. Ser. B 79 (2017) 1415–1437.
[11] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.
[12] J. Huang, T.N. Sun, Z.L. Ying, Y. Yu, C.H. Zhang, Oracle inequalities for the Lasso in the Cox model, Ann. Statist. 41 (2013) 1142–1165.
[13] T.K. Jenssen, W. Kuo, T. Stokke, E. Hovig, Associations between gene expressions in breast cancer and patient survival, Hum. Genet. 111 (2002) 411–420.
[14] S. Kong, B. Nan, Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso, Statist. Sin. 24 (2014) 25–42.
[15] M.R. Kosorok, Introduction To Empirical Processes and Semiparametric Inference, Springer, New York, 2008.
[16] M. Ledoux, M. Talagrand, Probability in Banach Spaces: Isoperimetry and Processes, Springer, Berlin, 2011.
[17] S. Lemler, Oracle inequalities for the Lasso for the conditional hazard rate in a high-dimensional setting, 2012. Available at arXiv:1206.5628.
[18] W. Lin, J. Lv, High dimensional sparse additive hazards regression, J. Amer. Statist. Assoc. 108 (2013) 247–264.
[19] P.L. Loh, M.J. Wainwright, Support recovery without incoherence: A case for nonconvex regularization, Ann. Statist. 45 (2017) 2455–2482.
[20] J. Lv, Y. Fan, A unified approach to model selection and sparse recovery using regularized least squares, Ann. Statist. 37 (2009) 3498–3528.

[21] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the Lasso, Ann. Statist. 34 (2006) 1436–1462.
[22] A. Rosenwald, G. Wright, W.C. Chan, J.M. Connors, E. Campo, R.I. Fisher, R.D. Gascoyne, H.K. Muller-Hermelink, E.B. Smeland, J.M. Giltnane, E.M. Hurt, H. Zhao, L. Averett, L. Yang, W.H. Wilson, E.S. Jaffe, R. Simon, R.D. Klausner, J. Powell, P.L. Duffey, D.L. Longo, T.C. Greiner, D.D. Weisenburger, W.G. Sanger, B.J. Dave, J.C. Lynch, J. Vose, J.O. Armitage, E. Montserrat, A. López-Guillermo, T.M. Grogan, T.P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, J. Delabie, H. Holte, P. Krajci, T. Stokke, L.M. Staudt, and the Lymphoma/Leukemia Molecular Profiling Project, The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, New Engl. J. Med. 346 (2002) 1937–1947.
[23] R.J. Tibshirani, Regression selection and shrinkage via the Lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[24] R.J. Tibshirani, The Lasso method for variable selection in the Cox model, Stat. Med. 16 (1997) 385–396.
[25] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, New York, 1998.
[26] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes: With Applications To Statistics, Springer, New York, 1996.
[27] M.J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso), IEEE Trans. Inform. Theory 55 (2009) 2183–2202.
[28] C.H. Zhang, J. Huang, The sparsity and bias of the Lasso selection in high-dimensional linear regression, Ann. Statist. 36 (2008) 1567–1594.
[29] P. Zhao, B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.