ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 5, *pp.* 1415–1437

Testing and confidence intervals for high dimensional proportional hazards models

Ethan X. Fang,

Pennsylvania State University, University Park, USA

Yang Ning

Cornell University, Ithaca, USA

and Han Liu

Princeton University, USA

[Received June 2015. Final revision October 2016]

Summary. The paper considers the problem of hypothesis testing and confidence intervals in high dimensional proportional hazards models. Motivated by a geometric projection principle, we propose a unified likelihood ratio inferential framework, including score, Wald and partial likelihood ratio statistics for hypothesis testing. Without assuming model selection consistency, we derive the asymptotic distributions of these test statistics, establish their semiparametric optimality and conduct power analysis under Pitman alternatives. We also develop new procedures to construct pointwise confidence intervals for the baseline hazard function and conditional hazard function. Simulation studies show that all tests proposed perform well in controlling type I errors. Moreover, the partial likelihood ratio test is empirically more powerful than the other tests. The methods proposed are illustrated by an example of a gene expression data set.

Keywords: Censored data; High dimensional inference; Proportional hazards model; Sparsity; Survival analysis

1. Introduction

The proportional hazards model (Cox, 1972) is one of the most important tools for analysing time-to-event data. It finds wide applications in epidemiology, medicine, economics and sociology (Kalbfleisch and Prentice, 2011). This model is semiparametric by treating the baseline hazard function as a nuisance parameter. To infer the finite dimensional parameter of interest, Cox (1972, 1975) proposed a partial likelihood approach which is invariant to the baseline hazard function. In low dimensional settings, Tsiatis (1981) and Andersen and Gill (1982) have established the consistency and asymptotic normality of the maximum partial likelihood estimator.

In high dimensional settings, when the number of covariates d is larger than the sample size n, the maximum partial likelihood estimation is an ill-posed problem. To solve this problem, we resort to regularized estimators (Tibshirani, 1996, 1997; Fan and Li, 2002; Antoniadis *et al.*, 2010). Other types of estimation procedures and their theoretical properties have been studied by Cai *et al.* (2005), Zhang and Lu (2007), Wang *et al.* (2009) and Zhao and Li (2012). In particular, under the ultrahigh dimensional regime that $d = o\{\exp(s^{-1}n)\}$, Bradic *et al.* (2011),

Address for correspondence: Han Liu, Department of Operations Research and Financial Engineering, Sherred Hall 224, Princeton University, Princeton, NJ 08544, USA.
E-mail: hanliu@princeton.edu

 1369–7412/17/791415

Huang *et al*. (2013) and Kong and Nan (2014) have established the oracle properties and error bounds of penalized maximum partial likelihood estimators, where s denotes the number of non-zero parameters in the Cox model. We note that Bradic *et al*. (2011) also established the limiting distribution of the oracle estimator. However, such an inferential result hinges on model selection consistency, which ignores model selection uncertainty, and thus may not be a practical inferential procedure in real applications.

Though significant progress has been made towards developing estimation theory, to the best of our knowledge, how to perform statistical inference (e.g. to test hypotheses or to construct confidence intervals) of high dimensional proportional hazard models remains an open problem. This paper aims to close this gap by developing valid inferential methods and theory for high dimensional proportional hazards models. In particular, we test hypotheses and construct confidence regions for a low dimensional component of a d-dimensional parameter vector. The main challenge for developing valid inferential methods is due to the presence of a high dimensional nuisance parameter, which makes the existing partial-likelihood-based inference (e.g. partial score test and partial likelihood ratio test) infeasible.

In this paper, we develop a unified inferential framework by extending the classical score, Wald and partial likelihood ratio tests to high dimensional proportional hazards models. To handle the high dimensional nuisance parameter, we construct a decorrelated score function by applying a high dimensional projection of the score function of the parameter of interest to the nuisance space. The solution of the decorrelated score function or its one-step approximation defines an estimator of the parameter of interest, which is parallel to the classical maximum partial likelihood estimator and can be used to construct a Wald test statistic. Towards the goal of performing likelihood-based inference, we further propose a new type of decorrelated partial likelihood function which is used to construct the likelihood ratio test. Theoretically, we establish the asymptotic distributions of score, Wald and partial likelihood ratio statistics under both the null and the Pitman alternatives. Empirically, we find that the partial likelihood ratio test is more powerful than the Wald and score tests, which shows the advantage of our likelihood ratio inference in finite samples. Following a similar idea, we also construct pointwise confidence intervals for the baseline hazard function and the conditional hazard function, and establish their asymptotic properties. In comparison with oracle inference in Bradic *et al*. (2011), our method does not require any type of irrepresentable condition or the minimal signal strength condition. The method proposed is still applicable even if the model selection is incorrect and thus is more practical in applications.

Various recent works (van de Geer *et al*., 2014; Belloni *et al*., 2016; Lockhart *et al*., 2014; Zhang and Zhang, 2014; Ning and Liu, 2016; Zhong *et al*., 2015) have considered high dimensional inference under the linear, generalized linear and additive hazard models. In what follows, we highlight the main differences. Lockhart *et al*. (2014) considered conditional inference given the event that a set of covariates is selected, whereas we consider unconditional inference; see Section 7 for further details. Zhang and Zhang (2014) proposed a novel low dimensional parameter method for inference in high dimensional linear models. However, their method strongly relies on the linear structure of the model. For instance, their method is motivated by the decomposition of a closed form expression of the univariate least squares estimator (i.e. equation (4) in Zhang and Zhang (2014)). A similar idea was used by Zhong *et al*. (2015) to study additive hazard models. However, it is unclear whether the low dimensional parameter method can be easily extended to the proportional hazards model, because of the model's non-linearity; see also the discussion in Zhong *et al*. (2015). The method in van de Geer *et al*. (2014) is based on inverting the Karush–Kuhn–Tucker condition of the lasso estimator in generalized linear models. In comparison with van de Geer *et al*. (2014), which focused only on the lasso estimator,

our methods and theory also apply to non-convex estimators such as smoothly clipped absolute deviation (SCAD) and the minimax concave penalty. In addition, our inference allows the inverse of the Fisher information matrix corresponding to the nuisance parameter to be non-sparse, which is weaker than the assumption in van de Geer *et al*. (2014).

After this work, Ning and Liu (2016) further extended the decorrelated score test to the general model with independently and identically distributed samples. The current paper is different in the following two aspects. First, as a methodological development, we build on the decorrelated score function and further propose a novel partial likelihood ratio test. The partial likelihood ratio test proposed retains the well-known Wilks phenomenon and is empirically more powerful. This agrees with the convention that, when all Wald, score and likelihood ratio tests are available, the likelihood ratio test is generally recommended. Second, owing to the presence of censored data, our technical development is quite different from Ning and Liu (2016). To handle time-dependent covariates, we need to use the counting process formulation, which is a unique challenge in survival analysis. This formulation

> 'permits a regression analysis of the intensity of a recurrent event allowing for complicated censoring patterns and time-dependent covariate'

(Andersen and Gill, 1982). More importantly, the log-partial-likelihood no longer has the sum of independently and identically distributed samples structure, which is different from the set-up in Ning and Liu (2016). To address this challenge, we

(a) develop refined concentration inequalities based on empirical process theory to control the approximation error and
(b) fully utilize the curvature structure of the partial likelihood function to obtain sharp theoretical results.

To be more specific, we illustrate the detailed technical challenges in the analysis of the proportional hazards model in remark 1 in Section 4.

The rest of this paper is organized as follows. In Section 2, we provide some background on the proportional hazards model. In Section 3, we propose methods for testing hypotheses and constructing confidence intervals for a single component of regression parameters. In Section 4, we provide theoretical analysis of the methods proposed. In the on-line supplementary materials, we extend the procedures to conduct inference on a multi-dimensional parameter of interest. Inferences on the baseline hazard and survival functions are studied in Section 5. In Section 6, we investigate the empirical performance of these methods. Section 7 contains a summary and discussions. Additional technical details, an extension to the multivariate failure time model and more extensive simulation studies are presented in the supplementary materials. The code that was used in the simulation can be downloaded from `http://www.personal.psu.edu/xxf13/Code/CoxHDInference.R`.

2. Background

We start with an introduction of the notation. Let $\mathbf{a} = (a_1, \ldots, a_d)^{\mathrm{T}} \in \mathbb{R}^d$ be a d-dimensional vector and $\mathbf{A} = (a_{jk}) \in \mathbb{R}^{d \times d}$ be a $d \times d$ matrix. Let $\mathrm{supp}(\mathbf{a}) = \{j : a_j \neq 0\}$. For $0 < q < \infty$, we define l_0, l_q and l_∞ vector norms as $\|\mathbf{a}\|_0 = \mathrm{card}\{\mathrm{supp}(\mathbf{a})\}$, $\|\mathbf{a}\|_q = (\Sigma_{j=1}^d \|\mathbf{a}_j\|^q)^{1/q}$ and $\|\mathbf{a}\|_\infty = \max_{1 \leqslant j \leqslant d} |a_j|$. We define the matrix l_∞-norm as the elementwise sup-norm that $\|\mathbf{A}\|_\infty = \max_{1 \leqslant j,k \leqslant d} |a_{jk}|$ and let $\|\mathbf{A}\|_0 = \Sigma_{1 \leqslant j,k \leqslant d} \mathbf{1}(a_{jk} \neq 0)$ and $\|\mathbf{A}\|_1 = \Sigma_{1 \leqslant j,k \leqslant d} |a_{jk}|$. Let $\mathbf{I}_d$ be the identity matrix in $\mathbb{R}^{d \times d}$. For a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a random variable Y, we denote X_n weakly converges to Y by $X_n \rightarrow^{\mathrm{d}} Y$. We denote $(n) = \{1, \ldots, n\}$.

2.1. Cox's proportional hazards model

We briefly review Cox's proportional hazards model. Let Q be the time to event, R be the censoring time and $\mathbf{X}(t) = (X_1(t), \ldots, X_d(t))^{\mathrm{T}}$ be the d-dimensional time-dependent covariates at time t. We consider the non-informative censoring setting that Q and R are conditionally independent given $\mathbf{X}(t)$. Let $W = \min\{Q, R\}$ and $\Delta = \mathbf{1}(Q \leqslant R)$ denote the observed survival time and censoring indicator, where $\mathbf{1}(\cdot)$ denotes the indicator function. Let τ be the end-of-study time. We observe n independent copies of $\{(\mathbf{X}(t), W, \Delta) : 0 \leqslant t \leqslant \tau\}$,

$$\{(\mathbf{X}_i(t), W_i, \Delta_i) : 0 \leqslant t \leqslant \tau\}_{i \in [n]}.$$

We denote $\lambda\{t|\mathbf{X}(t)\}$ as the conditional hazard rate function at time t given the covariates $\mathbf{X}(t)$. Under the proportional hazards model, we assume that

$$\lambda\{t|\mathbf{X}(t)\} = \lambda_0(t) \exp\{\mathbf{X}^{\mathrm{T}}(t)\boldsymbol{\beta}^*\},$$

where $\lambda_0(t)$ is an unknown baseline hazard rate function, and $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is an unknown parameter.

2.2. Penalized estimation

Following Andersen and Gill (1982), we introduce some counting process notation. For each i, let $N_i(t) := \mathbf{1}(W_i \leqslant t, \Delta_i = 1)$ be the counting process, and $Y_i(t) := \mathbf{1}(W_i \geqslant t)$ be the at-risk process for subject i. Assume that the process $Y_i(t)$ is left continuous with its right-hand limits satisfying $\mathbb{P}\{Y_i(t) = 1, 0 \leqslant t \leqslant \tau\} > C_\tau$ for some positive constant C_τ. The negative log-partial-likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{n}\left(\sum_{i=1}^{n} \int_0^\tau \mathbf{X}_i^{\mathrm{T}}(u)\boldsymbol{\beta}\, \mathrm{d}N_i(u) - \int_0^\tau \log\left[\sum_{i=1}^{n} Y_i(u) \exp\{\mathbf{X}_i^{\mathrm{T}}(u)\boldsymbol{\beta}\}\right] \mathrm{d}\bar{N}(u)\right),$$

where $\bar{N}(t) = \sum_{i=1}^n N_i(t)$.

When the dimension d is fixed and smaller than the sample size n, $\boldsymbol{\beta}^*$ can be estimated by the maximum partial likelihood estimator (Andersen and Gill, 1982). However, in high dimensional settings with $n < d$, the maximum partial likelihood estimator is not well defined. To solve this problem, Tibshirani (1997) and Fan and Li (2002) imposed the sparsity assumption and proposed the following penalized estimator:

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\arg\min}\{\mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta})\}, \tag{2.1}$$

where $\mathcal{P}_\lambda(\cdot)$ is a sparsity inducing penalty function, and λ is a tuning parameter. Bradic *et al.* (2011) and Huang *et al.* (2013) established the rates of convergence and oracle properties of the penalized maximum partial likelihood estimators $\hat{\boldsymbol{\beta}}$ by using SCAD and lasso penalties. For notational simplicity, we focus on the lasso estimator (Tibshirani, 1997) in this paper and indicate that similar properties hold for the SCAD and other non-convex penalized estimators. Existing works generally impose the following assumptions.

Assumption 1. The covariate is uniformly bounded:

$$\sup_{0 \leqslant t \leqslant \tau} \max_{1 \leqslant i \leqslant n} \max_{1 \leqslant j \leqslant d} |X_{ij}(t)| \leqslant C_X,$$

for some constant $C_X > 0$.

Assumption 2. For any set $\mathcal{S} \subset \{1, \ldots, d\}$ where $|\mathcal{S}| \asymp s$ and any vector $\mathbf{v}$ belonging to the cone $\mathcal{C}(\xi, \mathcal{S}) = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}_{\mathcal{S}^C}\|_1 \leqslant \xi \|\mathbf{v}_{\mathcal{S}}\|_1\}$, there is a constant $\lambda_{\min}$ such that

$$\kappa\{\xi, \mathcal{S}; \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)\} = \inf_{0 \neq \mathbf{v} \in \mathcal{C}(\xi, \mathcal{S})} \frac{s^{1/2}\{\mathbf{v}^{\mathrm{T}} \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)\mathbf{v}\}^{1/2}}{\|\mathbf{v}_{\mathcal{S}}\|_1} \geqslant \lambda_{\min} > 0.$$

Assumption 1 is the bounded covariate condition, which was imposed by both Bradic *et al.* (2011) and Huang *et al.* (2013) and holds in most real applications. Assumption 2 is known as the compatibility factor condition which was also used by Huang *et al.* (2013). This assumption essentially bounds the minimal eigenvalue of the Hessian matrix $\nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)$ from below for those directions within the cone $\mathcal{C}(\xi, \mathcal{S})$. In particular, the validity of this assumption has been verified in theorem 4.1 of Huang *et al.* (2013). Under these assumptions, Huang *et al.* (2013) derived the rate of convergence of the lasso estimator $\hat{\boldsymbol{\beta}}$ under the l_1-norm. More specifically, they proved that under assumptions 1 and 2, if $\|\boldsymbol{\beta}^*\|_0 = s$ and $\lambda \asymp \sqrt{\{n^{-1} \log(d)\}}$, it holds that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda), \tag{2.2}$$

which establishes the estimation consistency in the high dimensional regime.

For theoretical development, we introduce some additional notation. For a vector u, we denote by $\mathbf{u}^{\otimes 0} = 1$, $\mathbf{u}^{\otimes 1} = \mathbf{u}$ and $\mathbf{u}^{\otimes 2} = \mathbf{u}\mathbf{u}^{\mathrm{T}}$. Denote

$$\left.\begin{aligned} S^{(r)}(t, \boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^{\otimes r}(t) Y_i(t) \exp\{\mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}\} \qquad \text{for } r = 0, 1, 2, \\ \bar{\mathbf{Z}}(t, \boldsymbol{\beta}) &= S^{(1)}(t, \boldsymbol{\beta}) / S^{(0)}(t, \boldsymbol{\beta}), \\ \mathbf{V}_n(t, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{Y_i(t) \exp\{\mathbf{X}_i(t)^{\mathrm{T}}\boldsymbol{\beta}\}}{n S^{(0)}(t, \boldsymbol{\beta})} \{\mathbf{X}_i(t) - \bar{\mathbf{Z}}(t, \boldsymbol{\beta})\}^{\otimes 2} = \frac{S^{(2)}(t, \boldsymbol{\beta})}{S^{(0)}(t, \boldsymbol{\beta})} - \bar{\mathbf{Z}}(t, \boldsymbol{\beta})^{\otimes 2}. \end{aligned}\right\} \tag{2.3}$$

The gradient of $\mathcal{L}(\boldsymbol{\beta})$ is

$$\nabla \mathcal{L}(\boldsymbol{\beta}) = \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \{\mathbf{X}_i(u) - \bar{\mathbf{Z}}(u, \boldsymbol{\beta})\} \mathrm{d}N_i(u), \tag{2.4}$$

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

$$\nabla^2 \mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n} \int_0^{\tau} \mathbf{V}_n(u, \boldsymbol{\beta}) \mathrm{d}\bar{N}(u) = \frac{1}{n} \int_0^{\tau} \left\{ \frac{S^{(2)}(u, \boldsymbol{\beta})}{S^{(0)}(u, \boldsymbol{\beta})} - \bar{\mathbf{Z}}(u, \boldsymbol{\beta})^{\otimes 2} \right\} \mathrm{d}\bar{N}(u). \tag{2.5}$$

We denote the population versions of the above-defined quantities by

$$\left.\begin{aligned} \mathbf{s}^{(r)}(t, \boldsymbol{\beta}) &= \mathbb{E}[Y(t)\mathbf{X}(t)^{\otimes r} \exp\{\mathbf{X}(t)^{\mathrm{T}}\boldsymbol{\beta}\}] \qquad \text{for } r = 0, 1, 2, \\ \mathbf{e}(t, \boldsymbol{\beta}) &= \mathbf{s}^{(1)}(t, \boldsymbol{\beta}) / \mathbf{s}^{(0)}(t, \boldsymbol{\beta}), \end{aligned}\right\} \tag{2.6}$$

and

$$\left.\begin{aligned} \mathbf{H}(\boldsymbol{\beta}) &= \mathbb{E}\left[\int_0^{\tau} \left\{ \frac{\mathbf{s}^{(2)}(t, \boldsymbol{\beta})}{\mathbf{s}^{(0)}(t, \boldsymbol{\beta})} - \mathbf{e}(t, \boldsymbol{\beta})^{\otimes 2} \right\} \mathrm{d}N(t) \right], \\ \mathbf{H}^* &= \mathbf{H}(\boldsymbol{\beta}^*), \end{aligned}\right\} \tag{2.7}$$

where $\mathbf{H}^*$ is the Fisher information matrix based on the partial likelihood.

3. Hypothesis test and confidence interval

Whereas the estimation consistency has been established in high dimensions, it remains challenging to develop inferential procedures (e.g. valid confidence intervals and hypotheses testing)

for the high dimensional proportional hazards model. In this section, we propose three novel hypothesis testing procedures. The tests can be viewed as high dimensional counterparts of the conventional score, Wald and partial likelihood ratio tests. Hereafter, for notational simplicity, we partition the vector $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\alpha, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}$, where $\alpha = \beta_1 \in \mathbb{R}$ is the parameter of interest; $\boldsymbol{\theta} = (\beta_2, \ldots, \beta_d)^{\mathrm{T}} \in \mathbb{R}^{d-1}$ is the vector of nuisance parameters and we denote $\mathcal{L}(\boldsymbol{\beta})$ by $\mathcal{L}(\alpha, \boldsymbol{\theta})$. Let $\nabla^2_{\alpha\alpha}\mathcal{L}(\boldsymbol{\beta})$, $\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta})$ and $\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta})$ be the corresponding partitions of $\nabla^2\mathcal{L}(\boldsymbol{\beta})$. Let $\mathbf{H}^*_{\alpha\alpha}$, $\mathbf{H}^*_{\alpha\boldsymbol{\theta}}$ and $\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}$ be the corresponding partitions of $\mathbf{H}^*$, where $\mathbf{H}^*$ is defined in expression (2.7). For instance, $\mathbf{H}^*_{\boldsymbol{\theta}\alpha} = \mathbf{H}^*_{2:d,1} \in \mathbb{R}^{d-1}$ and $\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}) = \nabla^2_{2:d,2:d}\mathcal{L}(\boldsymbol{\beta}) \in \mathbb{R}^{(d-1)\times(d-1)}$. In this section, without loss of generality, we test the hypothesis H_0: $\alpha^* = 0$ *versus* H_1: $\alpha^* \neq 0$ for some univariate parameter of interest α. The extension to the multi-dimensional parameter of interest $\boldsymbol{\alpha} \in \mathbb{R}^{d_0}$, where d_0 is fixed, is provided in section G of the on-line supplementary materials.

3.1. Decorrelated score test

In the classical low dimensional setting, we exploit the profile partial score function

$$S(\alpha) = \nabla_\alpha \mathcal{L}(\alpha, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\alpha)} \tag{3.1}$$

to conduct tests, where $\hat{\boldsymbol{\theta}}(\alpha) = \arg\min_\theta \mathcal{L}(\alpha, \boldsymbol{\theta})$ is the maximum partial likelihood estimator for $\boldsymbol{\theta}$ with a fixed α. Under the null hypothesis that $\alpha^* = 0$, when d is fixed while $n \to \infty$ it holds that $\sqrt{n}S(0) \to \mathrm{d}N(0, H_{\alpha|\boldsymbol{\theta}})$, where $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$. If $nH^{-1}_{\alpha|\boldsymbol{\theta}}S^2(0)$ is larger than the $(1-\eta)$th quantile of a χ^2-distribution with 1 degree of freedom, we reject the null hypothesis. Classical asymptotic theory shows that this procedure controls the type I error with significance level η.

However, in high dimensions, the profile partial score function $S(\alpha)$ with $\hat{\boldsymbol{\theta}}(\alpha)$ replaced by a penalized estimator, say the corresponding components of $\hat{\boldsymbol{\beta}}$ in expression (2.1), does not yield a tractable limiting distribution owing to the existence of a large number of nuisance parameters. To address this problem, we construct a new score function for α that is asymptotically normal even in high dimensions. The key component is a high dimensional decorrelation method, aiming to handle the effect of the high dimensional nuisance vector.

More specifically, we propose a decorrelated score test for H_0: $\alpha^* = 0$. We first estimate $\boldsymbol{\theta}^*$ by $\hat{\boldsymbol{\theta}}$ using the l_1-penalized estimator $\hat{\boldsymbol{\beta}}$ in expression (2.1). Next, we calculate a linear combination of the partial score function $\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \hat{\boldsymbol{\theta}})$ to approximate $\nabla_\alpha \mathcal{L}(0, \hat{\boldsymbol{\theta}})$ best. The population version of the vector of coefficients in the best linear combination can be calculated as

$$\begin{aligned}\mathbf{w}^* &= \arg\min_{\mathbf{w}\in\mathbb{R}^{d-1}} \mathbb{E}\{\nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*) - \mathbf{w}^{\mathrm{T}}\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \boldsymbol{\theta}^*)\}^2 \\ &= \mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \boldsymbol{\theta}^*)^{\mathrm{T}}\}^{-1}\mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \boldsymbol{\theta}^*)\nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*)\} = \mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha},\end{aligned} \tag{3.2}$$

where the last equality holds by the second Bartlett identity (Tsiatis, 1981). In fact, $\mathbf{w}^{*\mathrm{T}}\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \boldsymbol{\theta}^*)$ can be interpreted as the projection of $\nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*)$ onto the linear span of the partial score function $\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \boldsymbol{\theta}^*)$. In high dimensions, we cannot directly estimate $\mathbf{w}^*$ by the corresponding sample version since the problem is ill posed. Motivated by the definition of $\mathbf{w}^*$ in equation (3.2), we estimate it by the lasso-type estimator

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^{d-1}}\{\tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\hat{\boldsymbol{\beta}})\mathbf{w} - \mathbf{w}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\hat{\boldsymbol{\beta}}) + \lambda'\|\mathbf{w}\|_1\}, \tag{3.3}$$

where λ' is a tuning parameter. Similarly, other non-convex penalty functions such as SCAD or

the minimax concave penalty can be applied. For simplicity, we focus on the lasso-type estimator in equation (3.3). Given $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{w}}$, we propose a decorrelated score function for α as

$$\hat{U}(\alpha, \hat{\boldsymbol{\theta}}) = \nabla_\alpha \mathcal{L}(\alpha, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{w}}^{\mathrm{T}} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\alpha, \hat{\boldsymbol{\theta}}). \tag{3.4}$$

The decorrelated score function in equation (3.4) can be rewritten as

$$\hat{U}(\alpha, \hat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau [X_{i1}(u) - \hat{\mathbf{w}}^{\mathrm{T}} \mathbf{X}_{i2:d}(u) - \{\bar{Z}_1(u, \alpha, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{w}}^{\mathrm{T}} \bar{\mathbf{Z}}_{2:d}(u, \alpha, \hat{\boldsymbol{\theta}})\}] \mathrm{d}N_i(u).$$

Recall that the standard score function is given by equation (2.4). By the definition of $\bar{\mathbf{Z}}(u, \boldsymbol{\beta})$, we find that $\hat{U}(\alpha, \hat{\boldsymbol{\theta}})$ has the same structure as $\nabla \mathcal{L}(\boldsymbol{\beta})$ with $\mathbf{X}_i(u)$ replaced by $X_{i1}(u) - \hat{\mathbf{w}}^{\mathrm{T}} \mathbf{X}_{i2:d}(u)$ and the risk set average $\bar{\mathbf{Z}}(u, \boldsymbol{\beta})$ of $\mathbf{X}_i(u)$ replaced by the risk set average of $X_{i1}(u) - \hat{\mathbf{w}}^{\mathrm{T}} \mathbf{X}_{i2:d}(u)$. Hence, the method proposed implicitly constructs a new covariate $\tilde{X}_i(u) := X_{i1}(u) - \hat{\mathbf{w}}^{\mathrm{T}} \mathbf{X}_{i2:d}(u)$, and the decorrelated score function $\hat{U}(\alpha, \hat{\boldsymbol{\theta}})$ can be interpreted as the integrated difference between the new covariate $\tilde{X}_i(u)$ and its risk set average. The covariate $\tilde{X}_i(u)$ is constructed, such that the (weighted) correlation between $\tilde{X}_i(u)$ and $\mathbf{X}_{i2:d}(u)$ is reduced, where the weight is introduced to account for the non-linearity of the Cox model. If the (weighted) correlation is sufficiently weak, we can perform the marginal analysis to infer the regression coefficient of $\tilde{X}_i(u)$. This also explains why our method is called the decorrelation method.

Geometrically, the decorrelated score function is approximately orthogonal to any component of the nuisance score function $\nabla_\theta \mathcal{L}(0, \boldsymbol{\theta}^*)$. This orthogonality property, which does not hold for the original score function $\nabla_\alpha \mathcal{L}(\alpha, \hat{\boldsymbol{\theta}})$, reduces the variability that is caused by the nuisance parameter estimation. A geometric illustration of the decorrelation-based methods is provided in Fig. 1, which also incorporates an illustration of the decorrelated Wald and partial likelihood ratio tests to be introduced in the following subsections. In contrast with the original score function $\nabla_\alpha \mathcal{L}(\alpha, \hat{\boldsymbol{\theta}})$, the proposed decorrelated score function $\hat{U}(\alpha, \hat{\boldsymbol{\theta}})$ yields test statistics with tractable limiting distributions. In the next section, we show that $\hat{U}(0, \hat{\boldsymbol{\theta}})$ converges weakly to $N(0, H_{\alpha|\boldsymbol{\theta}})$ under the null hypothesis, where $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}} \mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathbf{H}^*_{\boldsymbol{\theta}\alpha}$. This result holds in

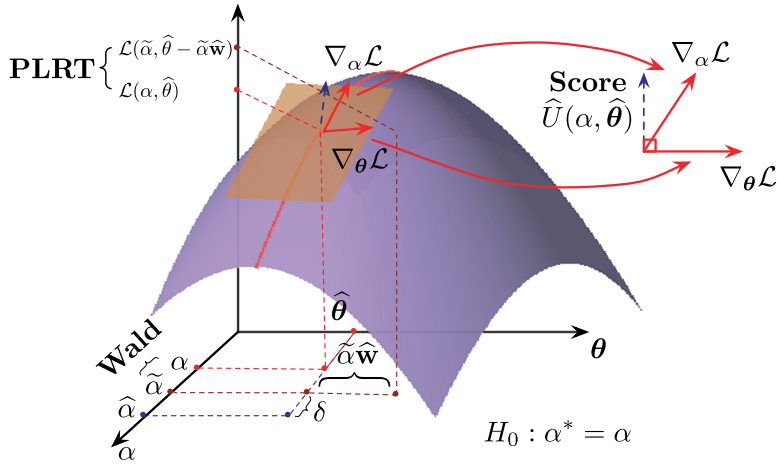

Fig. 1. Geometric illustration of the decorrelated score, Wald and partial likelihood ratio tests: the purple surface corresponds to the log-partial-likelihood function; the orange plane is the tangent plane of the surface at point $(\alpha, \hat{\boldsymbol{\theta}})$; the two red arrows in the orange plane represent $\nabla_\alpha \mathcal{L}$ and $\nabla_{\boldsymbol{\theta}} \mathcal{L}$; the decorrelated score function in blue is the projection of $\nabla_\alpha \mathcal{L}$ onto the space orthogonal to $\nabla_{\boldsymbol{\theta}} \mathcal{L}$; given the lasso estimator $\hat{\alpha}$, the decorrelated Wald estimator is $\tilde{\alpha} = \hat{\alpha} - \delta$, where $\delta = \{\partial \hat{U}(\hat{\alpha}, \hat{\boldsymbol{\theta}})/\partial \alpha\}^{-1} \hat{U}(\hat{\alpha}, \hat{\boldsymbol{\theta}})$; the decorrelated partial likelihood ratio test compares the log-partial-likelihood function values at $(\alpha, \hat{\boldsymbol{\theta}})$ and $(\tilde{\alpha}, \hat{\boldsymbol{\theta}} - \tilde{\alpha}\hat{\mathbf{w}})$

the high dimensional setting. We also point out that, in the low dimensional setting, it can be shown that the decorrelated score function $\hat{U}(\alpha, \hat{\boldsymbol{\theta}})$ is asymptotically equivalent to the profile partial score function $S(\alpha)$ in equation (3.1).

To test the null hypothesis $\alpha^* = 0$, we need to standardize $\hat{U}(0, \hat{\boldsymbol{\theta}})$ to construct the test statistic. We estimate $H_{\alpha|\boldsymbol{\theta}}$ by

$$\hat{H}_{\alpha|\boldsymbol{\theta}} = \nabla^2_{\alpha\alpha}\mathcal{L}(\hat{\alpha}, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\hat{\alpha}, \hat{\boldsymbol{\theta}}). \tag{3.5}$$

Hence, we define the decorrelated score test statistic as

$$\hat{S}_n = n\hat{H}^{-1}_{\alpha|\boldsymbol{\theta}}\hat{U}^2(0, \hat{\boldsymbol{\theta}}), \tag{3.6}$$

where $\hat{U}(0, \hat{\boldsymbol{\theta}})$ and $\hat{H}_{\alpha|\boldsymbol{\theta}}$ are defined in equations (3.4) and (3.5). In the next section, we show that, under the null hypothesis, $\hat{S}_n$ converges weakly to a χ^2-distribution with 1 degree of freedom. Given a significance level $\eta \in (0, 1)$, the score test $\psi_S(\eta)$ is

$$\psi_S(\eta) = \begin{cases} 0 & \text{if } \hat{S}_n \leqslant \chi^2_1(1-\eta), \\ 1 & \text{otherwise,} \end{cases} \tag{3.7}$$

where $\chi^2_1(1-\eta)$ denotes the $(1-\eta)$th quantile of a χ^2 random variable with 1 degree of freedom, and the null hypothesis $\alpha^* = 0$ is rejected if and only if $\psi_S(\eta) = 1$.

3.2. Confidence intervals and decorrelated Wald test

The score test proposed does not directly provide a confidence interval for α^*. In low dimensions, by looking at the limiting distribution of the maximum partial likelihood estimator, we can obtain a confidence interval for α^* (Andersen and Gill, 1982), which is equivalent to the classical Wald test. In this subsection, we extend the classical Wald test under the proportional hazards model to high dimensional settings.

The key idea of performing a Wald test is to derive a regular estimator for α^*. Our procedure is based on the decorrelated score function $\hat{U}(\alpha, \hat{\boldsymbol{\theta}})$ in equation (3.4). Since $\hat{U}(\alpha, \hat{\boldsymbol{\theta}})$ serves as an approximately unbiased estimating equation for α, the root of the equation $\hat{U}(\alpha, \hat{\boldsymbol{\theta}}) = 0$ with respect to α defines an estimator for α^*. However, searching for the root may be computationally intensive, especially when α is multi-dimensional as seen in the on-line supplementary materials, section G. To reduce the computational cost, we exploit a closed form estimator $\tilde{\alpha}$ obtained by linearizing $\hat{U}(\alpha, \hat{\boldsymbol{\theta}}) = 0$ at the initial estimator $\hat{\alpha}$. More specifically, letting $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\boldsymbol{\theta}}^{\mathrm{T}})^{\mathrm{T}}$ be the l_1-penalized estimator in expression (2.1), we adopt the following one-step estimator:

$$\tilde{\alpha} = \hat{\alpha} - \left\{ \frac{\partial \hat{U}(\hat{\alpha}, \hat{\boldsymbol{\theta}})}{\partial \alpha} \right\}^{-1} \hat{U}(\hat{\alpha}, \hat{\boldsymbol{\theta}}), \qquad \hat{U}(\hat{\alpha}, \hat{\boldsymbol{\theta}}) = \nabla_\alpha \mathcal{L}(\hat{\alpha}, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{w}}^{\mathrm{T}}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\hat{\alpha}, \hat{\boldsymbol{\theta}}). \tag{3.8}$$

In the next section, we prove that $\sqrt{n}(\tilde{\alpha} - \alpha^*)$ converges weakly to $N(0, H^{-1}_{\alpha|\boldsymbol{\theta}})$. Hence, let $Z_{1-\eta/2}$ be the $(1-\eta/2)$th quantile of $N(0, 1)$. We have that

$$[\tilde{\alpha} - n^{-1/2}Z_{1-\eta/2}\hat{H}^{-1/2}_{\alpha|\boldsymbol{\theta}}, \tilde{\alpha} + n^{-1/2}Z_{1-\eta/2}\hat{H}^{-1/2}_{\alpha|\boldsymbol{\theta}}]$$

is a $100(1-\eta)\%$ confidence interval for α^*. From the perspective of hypothesis testing, the decorrelated Wald test statistic for H_0: $\alpha^* = 0$ *versus* H_1: $\alpha^* \neq 0$ is

$$\hat{W}_n = n\hat{H}_{\alpha|\boldsymbol{\theta}}\tilde{\alpha}^2, \tag{3.9}$$

where $\tilde{\alpha}$ and $\hat{H}_{\alpha|\boldsymbol{\theta}}$ are defined in equations (3.8) and (3.5) respectively. Consequently, the decorrelated Wald test with significance level η is

$$\psi_{\mathrm{W}}(\eta) = \begin{cases} 0 & \text{if } \hat{W}_n \leqslant \chi_1^2(1-\eta), \\ 1 & \text{otherwise,} \end{cases} \tag{3.10}$$

and the null hypothesis $\alpha^* = 0$ is rejected if and only if $\psi_{\mathrm{W}}(\eta) = 1$.

3.3. Decorrelated partial likelihood ratio test

Under low dimensional settings, the likelihood ratio inference enjoys great success in the statistical literature. Under the proportional hazards model, the partial likelihood ratio test statistic is $\mathrm{PLRT} = 2n[\mathcal{L}\{0, \hat{\boldsymbol{\theta}}_{\mathrm{P}}(0)\} - \mathcal{L}(\hat{\alpha}_{\mathrm{P}}, \hat{\boldsymbol{\theta}}_{\mathrm{P}})]$, where $\hat{\boldsymbol{\theta}}_{\mathrm{P}}(0) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta})$ and $(\hat{\alpha}_{\mathrm{P}}, \hat{\boldsymbol{\theta}}_{\mathrm{P}}) = \arg\min_{\alpha, \boldsymbol{\theta}} \mathcal{L}(\alpha, \boldsymbol{\theta})$ are the maximum partial likelihood estimators under the null and alternative hypotheses. Hence, PLRT evaluates the validity of the null hypothesis by comparing the partial likelihood under H_0 with that under H_1. Similarly to the partial score test, the partial likelihood ratio test also fails in the high dimensional setting because of a large number of nuisance parameters. In this section, we propose a new type of the partial likelihood ratio test which is applicable in the high dimensional setting.

To handle the effect of high dimensional nuisance parameters, we define the (negative) decorrelated partial likelihood for α as $\mathcal{L}_{\mathrm{decor}}(\alpha) = \mathcal{L}(\alpha, \hat{\boldsymbol{\theta}} - \alpha\hat{\mathbf{w}})$. The intuition of this decorrelated partial likelihood is to approximate the likelihood of a submodel with the direction $(1, -\mathbf{w}^*)$, which also corresponds to the least favourable direction for estimating α. In the low dimensional setting, the decorrelated partial likelihood $\mathcal{L}_{\mathrm{decor}}(\alpha)$ is asymptotically equivalent to the profile partial likelihood $\mathcal{L}\{\alpha, \hat{\boldsymbol{\theta}}(\alpha)\}$. Hence, we view $\mathcal{L}_{\mathrm{decor}}(\alpha)$ as an extension of the classical profile partial likelihood to high dimensions. The decorrelated partial likelihood ratio test statistic is defined as

$$\hat{L}_n = 2n\{\mathcal{L}_{\mathrm{decor}}(0) - \mathcal{L}_{\mathrm{decor}}(\tilde{\alpha})\}, \qquad \mathcal{L}_{\mathrm{decor}}(\alpha) = \mathcal{L}(\alpha, \hat{\boldsymbol{\theta}} - \alpha\hat{\mathbf{w}}), \tag{3.11}$$

and $\tilde{\alpha}$ is given in expression (3.8). As discussed in the previous subsection, $\tilde{\alpha}$ is a one-step approximation of the global minimizer of $\mathcal{L}_{\mathrm{decor}}(\alpha)$. Hence, the log-likelihood ratio $\hat{L}_n$ evaluates the validity of the null hypothesis by comparing the decorrelated partial likelihood under H_0 with that under H_1.

In the next section, we show that $\hat{L}_n$ converges weakly to a χ^2-distribution with 1 degree of freedom. Therefore, a decorrelated partial likelihood ratio test with significance level η is

$$\psi_L(\eta) = \begin{cases} 0 & \text{if } \hat{L}_n \leqslant \chi_1^2(1-\eta), \\ 1 & \text{otherwise,} \end{cases} \tag{3.12}$$

and $\psi_L(\eta) = 1$ indicates a rejection of the null hypothesis.

4. Asymptotic properties

In this section, we derive the limiting distributions of the decorrelated test statistics under the null hypothesis. The limiting distributions of the test statistics under the Pitman alternative are shown in the on-line supplementary materials, section B. In our analysis, we assume the following conditions.

Assumption 3. The true hazard is uniformly bounded, i.e.

$$\sup_{t \in [0,\tau]} \max_{i \in [n]} |\mathbf{X}_i^{\mathrm{T}}(t)\boldsymbol{\beta}^*| = \mathcal{O}(1).$$

Assumption 4. It holds that $\|\mathbf{w}^*\|_0 = s' \asymp s$, and

$$\sup_{t\in[0,\tau]} \max_{i\in[n]} |\mathbf{X}_{i,2:d}^{\mathrm{T}}(t)\mathbf{w}^*| = \mathcal{O}(1).$$

Assumption 5. The maximum and minimum eigenvalues of the Fisher information matrix are bounded, $C_h \leqslant \Lambda_{\min}(\mathbf{H}^*) \leqslant \Lambda_{\max}(\mathbf{H}^*) \leqslant 1/C_h$ for some constant $C_h > 0$.

To connect these assumptions with existing literature, assumptions 3 and 4 extend assumption (iv) of theorem 3.3 in van de Geer *et al.* (2014) to the proportional hazards model. In particular, the sparsity of $\mathbf{w}^*$ is assumed in order to establish the consistency of $\hat{\mathbf{w}}$ to $\mathbf{w}^*$. By the block matrix inversion formula, the sparsity assumption of $\mathbf{w}^*$ is equivalent to the corresponding row or column of $\mathbf{H}^{*-1}$ being sparse. Assumption 5 is related to the Fisher information matrix, which is essential even in low dimensional settings.

To derive the asymptotic property of our test statistics, a crucial step in our analysis is the consistency of the estimator $\hat{\mathbf{w}}$ in equation (3.3). The following lemma provides a fast rate of convergence of $\hat{\mathbf{w}}$.

Lemma 1. Under assumptions 1–5, if $\lambda' \asymp \sqrt{\{n^{-1}\log(d)\}}$, we have

$$\|\hat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}[(s'+s)\sqrt{\{n^{-1}\log(d)\}}, \tag{4.1}$$

where $\mathbf{w}^*$ and $\hat{\mathbf{w}}$ are defined in equations (3.2) and (3.3).

Consequently, the following result characterizes the asymptotic normality of the decorrelated score function $\hat{U}(0, \hat{\theta})$ in equation (3.4) under the null hypothesis.

Theorem 1. Suppose that assumptions 1–5 hold. If $\lambda \asymp \sqrt{\{n^{-1}\log(d)\}}$, $\lambda' \asymp \sqrt{\{n^{-1}\log(d)\}}$ and $n^{-1/2}s\log(d) = o(1)$, under the null hypothesis $\alpha^* = 0$, the decorrelated score function $\hat{U}(0, \hat{\boldsymbol{\theta}})$ defined in equation (3.4) satisfies

$$\sqrt{n}\hat{U}(0, \hat{\boldsymbol{\theta}}) \xrightarrow{\mathrm{d}} Z, \qquad Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}), \tag{4.2}$$

and $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$.

We note that theorem 1 holds if (n, s, d) satisfies $n^{-1/2}s\log(d) = o(1)$, which agrees with the assumption in the existing work for the linear model and the generalized linear model; see van de Geer *et al.* (2014) and Ning and Liu (2016). In addition, our tuning parameters λ and λ' follow the conventional $\sqrt{\{n^{-1}\log(d)\}}$-rate for high dimensional estimation and inference.

Remark 1. In this remark, we emphasize the additional technical challenges in the analysis of the Cox model compared with the existing works on the linear model and the generalized linear model. First, since the log-partial likelihood does not have the independently and identically distributed samples structure, our theoretical results are built on the concentration inequalities for the score function and Hessian matrix via empirical process theory; see the technical lemmas in section E of the on-line supplementary materials. The main theoretical tools are a refinement of Talagrand's inequality (Massart (2007), equation (5.50)) and maximal inequalities for empirical processes. Second, it is technically more complicated to control the uncertainty of the score function and Hessian matrix evaluated at $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{w}}$. For instance, to attain the fast rate in lemma 1, we need to exploit the structure of the Hessian matrix fully and to separate the uncertainty of the Hessian matrix from the plug-in error of $\hat{\boldsymbol{\beta}}$ carefully, since $\hat{\mathbf{w}}$ defined in equation (3.3) depends on $\hat{\boldsymbol{\beta}}$. A direct analysis based on the local Lipschitz property of the Hessian matrix on $\boldsymbol{\beta}$ yields a weaker result $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}[s's\sqrt{\{n^{-1}\log(d)\}}]$, which is slower than the rate in

lemma 1. Thus, this slower rate leads to stronger assumptions than $n^{-1/2}s\log(d)=o(1)$ required in theorem 1. Such challenges do not appear in the context of the linear model and the generalized linear model.

As we have discussed, the limiting variance of the decorrelated score function can be estimated by $\hat{H}_{\alpha|\boldsymbol{\theta}}=\nabla^2_{\alpha\alpha}\mathcal{L}(\hat{\alpha},\hat{\boldsymbol{\theta}})-\hat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\hat{\alpha},\hat{\boldsymbol{\theta}})$. The next lemma shows the consistency of $\hat{H}_{\alpha|\boldsymbol{\theta}}$.

Lemma 2. Suppose that assumptions 1–5 hold. If $\lambda\asymp\sqrt{\{n^{-1}\log(d)\}}$ and $\lambda'\asymp\sqrt{\{n^{-1}\log(d)\}}$, we have

$$|H_{\alpha|\boldsymbol{\theta}}-\hat{H}_{\alpha|\boldsymbol{\theta}}|=\mathcal{O}_{\mathbb{P}}\left[s\sqrt{\left\{\frac{\log(d)}{n}\right\}}\right].$$

By theorem 1 and lemma 2, the following corollary shows that, under the null hypothesis, the type I error of the decorrelated score test $\psi_S(\eta)$ in expression (3.7) converges asymptotically to the significance level η. Let the associated p-value of the decorrelated score test be $P_S=2\{1-\Phi(\hat{S}_n)\}$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable and $\hat{S}_n$ is the score test statistic defined in equation (3.6). The distribution of P_S converges to a uniform distribution asymptotically.

Corollary 1. Suppose that assumptions 1–5 hold, $\lambda\asymp\sqrt{\{n^{-1}\log(d)\}}$, $\lambda'\asymp\sqrt{\{n^{-1}\log(d)\}}$ and $n^{-1/2}s\log(d)=o(1)$. The decorrelated score test and the p-value satisfy

$$\lim_{n\to\infty}\mathbb{P}\{\psi_S(\eta)=1|\alpha^*=0\}=\eta,\qquad\text{and}\qquad P_S\xrightarrow{\mathrm{d}}\mathrm{Unif}[0,1],\qquad\text{when }\alpha^*=0,$$

where Unif[0, 1] denotes a random variable uniformly distributed in [0, 1].

We then analyse the Wald test under the null hypothesis. We derive the limiting distribution of the one-step estimator $\tilde{\alpha}$ defined in expression (3.8) in the next theorem.

Theorem 2. Suppose that assumptions 1–5 hold, and $\lambda\asymp\sqrt{\{n^{-1}\log(d)\}}$, $\lambda'\asymp\sqrt{\{n^{-1}\log(d)\}}$ and $n^{-1/2}s\log(d)=o(1)$. When the null hypothesis $\alpha^*=0$ holds, the decorrelated estimator $\tilde{\alpha}$ satisfies

$$\sqrt{n}\tilde{\alpha}\xrightarrow{\mathrm{d}}Z,\qquad Z\sim N(0,H^{-1}_{\alpha|\boldsymbol{\theta}}).\tag{4.3}$$

Utilizing the asymptotic normality of $\tilde{\alpha}$, we can establish the limiting type I error of $\psi_{\mathrm{W}}(\eta)$ in equation (3.10), in the next corollary. It is straightforward to generalize the result to $\sqrt{n}(\tilde{\alpha}-\alpha^*)\to\mathrm{d}Z$, where $Z\sim N(0,H^{-1}_{\alpha|\boldsymbol{\theta}})$ for any α^*. This gives us a confidence interval of α^*.

Corollary 2. Under assumptions 1–5, suppose that $\lambda\asymp\sqrt{\{n^{-1}\log(d)\}}$, $\lambda'\asymp\sqrt{\{n^{-1}\log(d)\}}$ and $n^{-1/2}s\log(d)=o(1)$. The type I error of the decorrelated Wald test $\psi_{\mathrm{W}}(\eta)$ and its corresponding p-value $P_{\mathrm{W}}=2\{1-\Phi(\hat{W}_n)\}$ satisfy

$$\lim_{n\to\infty}\mathbb{P}\{\psi_W(\eta)=1|\alpha^*=0\}=\eta,\qquad\text{and}\qquad P_{\mathrm{W}}\xrightarrow{\mathrm{d}}\mathrm{Unif}[0,1]\qquad\text{when }\alpha^*=0.$$

In addition, an asymptotic $100(1-\eta)\%$ confidence interval of α^* is

$$\left(\tilde{\alpha}-\frac{\Phi^{-1}(1-\eta/2)}{\sqrt{(n\hat{H}_{\alpha|\boldsymbol{\theta}})}},\tilde{\alpha}+\frac{\Phi^{-1}(1-\eta/2)}{\sqrt{(n\hat{H}_{\alpha|\boldsymbol{\theta}})}}\right).$$

Finally, we present our main result on the limiting distribution of the decorrelated partial likelihood ratio test statistic $\hat{L}_n$ that was introduced in expression (3.11).

Theorem 3. Suppose that assumptions 1–5 hold, $\lambda \asymp \sqrt{\{n^{-1}\log(d)\}}$, $\lambda' \asymp \sqrt{\{n^{-1}\log(d)\}}$ and $n^{-1/2}s\log(d) = o(1)$. If the null hypothesis $\alpha^* = 0$ holds, the decorrelated likelihood ratio test statistic $\hat{L}_n$ in expression (3.11) satisfies

$$\hat{L}_n \xrightarrow{\mathrm{d}} Z_\chi, \qquad Z_\chi \sim \chi_1^2. \tag{4.4}$$

Theorem 3 illustrates the Wilks phenomenon of $\hat{L}_n$ and justifies the decorrelated partial likelihood ratio test $\psi_L(\eta)$ in equation (3.12). Also, let the p-value that is associated with the decorrelated partial likelihood ratio test be $P_L = 1 - F(\hat{L}_n)$, where $F(\cdot)$ is the cumulative distribution function of χ_1^2. Similarly to corollaries 1 and 2, we characterize the type I error of the test $\psi_L(\eta)$ in equation (3.12) and its corresponding p-value below.

Corollary 3. Suppose that assumptions 1–5 hold, $\lambda \asymp \sqrt{\{n^{-1}\log(d)\}}$, $\lambda' \asymp \sqrt{\{n^{-1}\log(d)\}}$ and $n^{-1/2}s\log(d) = o(1)$. The type I error of the decorrelated partial likelihood ratio test $\psi_L(\eta)$ with significance level η and its associated p-value P_L satisfy

$$\lim_{n\to\infty} \mathbb{P}\{\psi_L(\eta) = 1 | \alpha^* = 0\} = \eta, \quad \text{and} \quad P_L \xrightarrow{\mathrm{d}} \mathrm{Unif}[0,1] \qquad \text{when } \alpha^* = 0.$$

By corollaries 1–3, we see that the decorrelated score, Wald and partial likelihood ratio tests are asymptotically equivalent since, under the same assumptions, it holds that

$$\hat{S}_n = \hat{W}_n + o_{\mathbb{P}}(1) = \hat{L}_n + o_{\mathbb{P}}(1).$$

To summarize this section, corollaries 1–3 characterize the asymptotic distributions of the proposed decorrelated test statistics under the null hypothesis. It is known that $H_{\alpha|\theta}$ is the semiparametric information lower bound for inferring α. Theorem 2 shows that $\tilde{\alpha}$ achieves the semiparametric information bound, which indicates the semiparametric efficiency of $\tilde{\alpha}$. By asymptotic equivalence, all our test statistics are semiparametrically efficient (van der Vaart, 2000). Although the three tests are asymptotically equivalent, our numerical results suggest that the partial likelihood ratio test outperforms the remaining tests empirically.

5. Inference on the baseline hazard function

The baseline hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u$$

is treated as a nuisance function in the log-partial-likelihood method. In practice, inferences on the baseline hazard function can be of interest also. To the best of our knowledge, such problems remain unexplored in high dimensional settings. In this section, we aim to construct confidence intervals for the baseline hazard function and the survival function. In addition, we extend the procedure to conduct inference on the conditional hazard function in the on-line supplementary materials, section D.

We consider the following Breslow-type estimator for the baseline hazard function. Given an l_1-penalized estimator $\hat{\boldsymbol{\beta}}$ derived from equation (2.1), the plug-in estimator for the baseline hazard function at time t is

$$\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) = \int_0^t \frac{\sum_{i=1}^n \mathrm{d}N_i(u)}{\sum_{i=1}^n Y_i(u)\exp\{\mathbf{X}_i^{\mathrm{T}}(u)\hat{\boldsymbol{\beta}}\}}. \tag{5.1}$$

Since the plug-in estimator $\hat{\boldsymbol{\beta}}$ does not have a tractable distribution, inference based on the estimator $\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}})$ is difficult. To handle this problem, we adopt a bias correction procedure to reduce the uncertainty that is caused by plugging $\hat{\boldsymbol{\beta}}$ into $\hat{\Lambda}_0(t, \boldsymbol{\beta})$. Specifically, we estimate $\Lambda_0(t)$ by the sample version of $\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - (\nabla\Lambda_0(t, \boldsymbol{\beta}^*))^{\mathrm{T}}\mathbf{H}^{*-1}\nabla\mathcal{L}(\hat{\boldsymbol{\beta}})$, where

$$\Lambda_0(t, \boldsymbol{\beta}) = \mathbb{E}\left\{ \int_0^t \frac{\mathrm{d}N_i(u)}{S^{(0)}(u, \boldsymbol{\beta})} \right\},$$

and the gradient $\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$ is taken with respect to the corresponding $\boldsymbol{\beta}$-component, and $\mathbf{H}^*$ is the Fisher information matrix defined in equation (2.7). We propose to estimate $\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$ by the Dantzig-type estimator

$$\hat{\mathbf{u}}(t) = \arg\min \|\mathbf{u}(t)\|_1, \qquad \text{subject to } \|\nabla\hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - \nabla^2\mathcal{L}(\hat{\boldsymbol{\beta}})\mathbf{u}(t)\|_\infty \leqslant \delta, \tag{5.2}$$

where δ is a tuning parameter. It can be shown that the estimator $\hat{\mathbf{u}}(t)$ converges to $\mathbf{u}^*(t) = \mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$ under the following regularity assumption.

Assumption 6. It holds that $\|\mathbf{u}^*(t)\|_0 = s' \asymp s$ for all $0 \leqslant t \leqslant \tau$, and

$$\sup_{t\in[0,\tau]} \max_{i\in[n]} |\mathbf{X}_i^{\mathrm{T}}(t)\mathbf{u}^*(t)| = \mathcal{O}(1).$$

Assumption 6 plays the same role as assumption 4 in the previous section. Hence, the decorrelated baseline hazard function estimator at time t is

$$\tilde{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) = \hat{\Lambda}_0(t, \hat{\boldsymbol{\beta}}) - \hat{\mathbf{u}}(t)^{\mathrm{T}}\nabla\mathcal{L}(\hat{\boldsymbol{\beta}}), \tag{5.3}$$

where $\hat{\mathbf{u}}(t)$ is defined in expression (5.2). On the basis of estimator (5.3), the survival function $S_0(t) = \exp\{-\Lambda_0(t)\}$ is estimated by $\tilde{S}(t, \hat{\boldsymbol{\beta}}) = \exp\{-\tilde{\Lambda}_0(t, \hat{\boldsymbol{\beta}})\}$. The main theorem of this section characterizes the pointwise asymptotic normality of $\tilde{\Lambda}_0(t, \hat{\boldsymbol{\beta}})$ and $\tilde{S}(t, \hat{\boldsymbol{\beta}})$ as follows.

Theorem 4. Suppose that assumptions 1–3, 5 and 6 hold, $\lambda \asymp \sqrt{\{n^{-1}\log(d)\}}$, $\delta \asymp s' \times \sqrt{\{n^{-1}\log(d)\}}$ and $n^{-1/2}s^2\log(d) = o(1)$. We have that, for any $t \in [0, \tau]$, the decorrelated baseline hazard function estimator $\tilde{\Lambda}_0(t, \hat{\boldsymbol{\beta}})$ in equation (5.3) satisfies

$$\sqrt{n}\{\Lambda_0(t) - \tilde{\Lambda}_0(t, \hat{\boldsymbol{\beta}})\} \xrightarrow{\mathrm{d}} Z, \qquad Z \sim N\{0, \sigma_1^2(t) + \sigma_2^2(t)\},$$

where

$$\begin{aligned} \sigma_1^2(t) &= \int_0^t \frac{\lambda_0(u)\,\mathrm{d}u}{\mathbb{E}[\exp\{\mathbf{X}^{\mathrm{T}}(u)\boldsymbol{\beta}^*\}Y(u)]}, \\ \sigma_2^2(t) &= \nabla\Lambda_0(t, \boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*). \end{aligned} \tag{5.4}$$

The estimated survival function $\tilde{S}(t, \hat{\boldsymbol{\beta}})$ satisfies

$$\sqrt{n}\{\tilde{S}(t, \hat{\boldsymbol{\beta}}) - S_0(t)\} \xrightarrow{\mathrm{d}} Z', \qquad Z' \sim N\left[0, \frac{\sigma_1^2(t) + \sigma_2^2(t)}{\exp\{2\Lambda_0(t)\}}\right].$$

As a final remark, the assumption $n^{-1/2}s^2\log(d) = o(1)$ in theorem 4 is stronger than those in theorems 1 and 3. The main reason is that the estimand $\Lambda_0(t, \boldsymbol{\beta})$ is a non-linear function of $\boldsymbol{\beta}$, which requires stronger technical assumptions to construct confidence intervals.

6. Numerical results

This section reports numerical results of our proposed methods by using both simulated and real data.

6.1. Simulated data

We first investigate empirical performances of the decorrelated score, Wald and partial likelihood ratio tests on the parametric component $\boldsymbol{\beta}^*$ as proposed in Section 3. In our simulation, we let $\beta_1^* = 0$ and randomly select s out of the next $d-1$ components to be non-zero, where the rest of the entries are set to be 0. To estimate $\boldsymbol{\beta}^*$ and $\mathbf{w}^*$, we choose the tuning parameters λ by tenfold cross-validation and set $\lambda' = \sqrt{\{n^{-1}\log(d)\}}$. We find that our simulation results are insensitive to the choice of λ'. We conduct decorrelated score, Wald and partial likelihood ratio tests for β_1 which is set to be 0 under the null hypothesis H_0: $\beta_1^* = 0$ *versus* the alternative H_a: $\beta_1^* \neq 0$, where we set the level of significance to be $\eta = 0.05$. In each setting, we simulate $n = 150$ independent samples from a multivariate Gaussian distribution $\mathbf{X}_i \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ for $d = 100, 200, 500$, where $\boldsymbol{\Sigma}$ is a Toeplitz matrix with $\Sigma_{jk} = \rho^{|j-k|}$ and $\rho = 0.25$, 0.4, 0.6, 0.75. The cardinality of the active set s is either 2 or 3, and the regression coefficients in the active set are either all 1s (Dirac) or drawn randomly from the uniform distribution Unif[0, 2]. We set the baseline hazard rate function to be the identity. Thus, the ith survival time follows an exponential distribution with mean $\exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)$. The ith censoring time is independently generated from an exponential

Table 1. Average type I error of the decorrelated tests with $\eta = 5\%$ where $(n, s) = (150, 2)$

Method	*d*	*Results (%) for* $\rho = 0.25$		*Results (%) for* $\rho = 0.4$		*Results (%) for* $\rho = 0.6$		*Results (%) for* $\rho = 0.75$	
		Dirac	*Unif [0,2]*	*Dirac*	*Unif [0,2]*	*Dirac*	*Unif [0,2]*	*Dirac*	*Unif [0,2]*
Score	100	5.2	5.1	4.8	4.7	5.3	5.2	5.1	5.0
	200	5.3	4.9	4.7	5.0	5.2	5.3	5.1	4.8
	500	6.2	6.3	5.9	5.2	4.5	4.8	4.2	3.8
Wald	100	5.5	5.2	5.1	5.3	4.9	4.7	5.1	5.3
	200	5.5	5.3	5.4	5.1	4.8	4.7	4.4	4.5
	500	6.5	6.2	5.7	5.6	5.7	4.4	4.6	3.7
PLRT	100	5.3	5.1	5.3	4.9	5.1	4.8	4.7	4.8
	200	5.6	5.7	5.4	5.3	4.7	5.5	4.8	4.6
	500	6.2	6.5	6.2	5.6	4.8	4.4	4.0	3.8

Table 2. Average type I error of the decorrelated tests with $\eta = 5\%$ where $(n, s) = (150, 3)$

Method	*d*	*Results (%) for* $\rho = 0.25$		*Results (%) for* $\rho = 0.4$		*Results (%) for* $\rho = 0.6$		*Results (%) for* $\rho = 0.75$	
		Dirac	*Unif [0,2]*	*Dirac*	*Unif [0,2]*	*Dirac*	*Unif [0,2]*	*Dirac*	*Unif [0,2]*
Score	100	5.4	5.3	4.9	5.4	5.2	4.7	5.4	5.3
	200	5.3	5.1	4.8	5.4	5.3	5.7	4.6	4.4
	500	6.2	6.3	6.0	4.5	4.7	4.8	3.5	3.7
Wald	100	5.3	5.1	5.5	5.2	4.9	5.0	5.2	4.8
	200	4.9	4.8	4.7	5.2	5.3	5.5	4.3	4.5
	500	6.6	6.7	6.2	6.1	5.3	4.8	4.1	3.8
PLRT	100	5.2	5.2	5.1	5.4	5.2	5.3	5.0	4.7
	200	5.3	5.4	5.2	4.8	5.3	5.5	5.3	4.5
	500	6.7	6.5	5.9	5.4	4.7	4.4	3.9	3.6

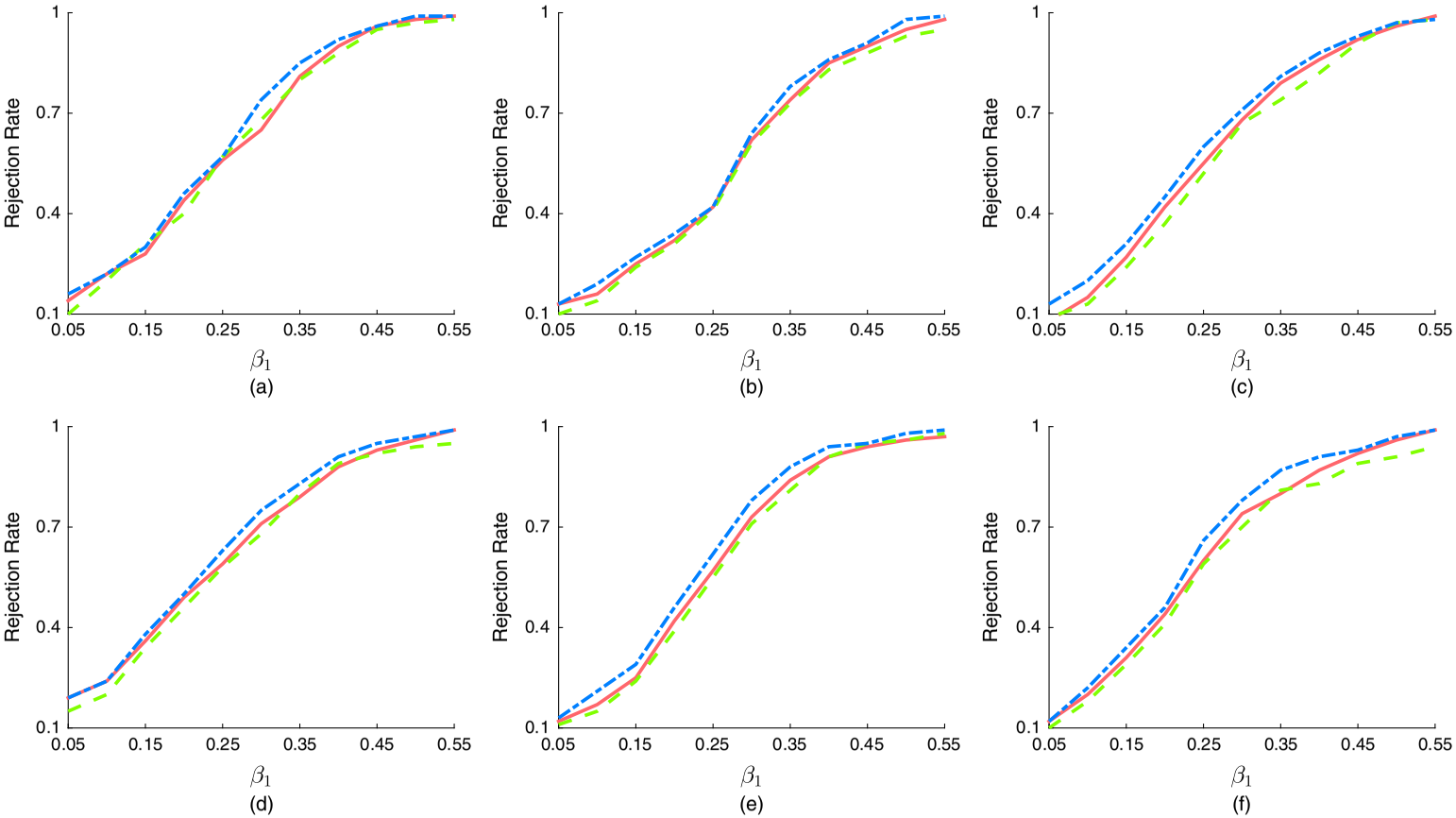

Fig. 2. Empirical rejection rates of the decorrelated score, Wald and partial likelihood ratio tests on simulated data with different active set sizes and dimensionality (——, score; - - -, Wald; – – –, partial likelihood ratio): (a) $(s, d) = (2, 100)$; (b) $(s, d) = (2, 200)$; (c) $(s, d) = (2, 500)$; (d) $(s, d) = (3, 100)$; (e) $(s, d) = (3, 200)$; (f) $(s, d) = (3, 500)$

distribution with mean $U \exp(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^*)$, where $U \sim \mathrm{Unif}[1, 3]$. As discussed in Fan and Li (2002), this censoring scheme results in about 30% censored samples.

The simulation is repeated 1000 times. The empirical type I errors of the decorrelated score, Wald and partial likelihood ratio tests are summarized in Tables 1 and 2. We see that the empirical type I errors of all three tests are close to the desired 5% level of significance, which supports our theoretical results. This observation holds for the whole range of ρ, s and d specified in the data-generating procedures. In addition, as expected, the empirical type I errors further deviate from the level of significance as d increases for all three tests, illustrating the effects of dimensionality d on finite sample performance.

We also investigate the empirical power of the tests proposed. Instead of setting $\beta_1 = 0$, we generate the data with $\beta_1 = 0.05, 0.1, 0.15, \ldots, 0.55$, following the same simulation scheme as introduced above. We plot the rejection rates of the three decorrelated tests for testing $H_0 : \beta_1 = 0$ with significance level 0.05 and $\rho = 0.25$ in Fig. 2. We see that, when $d = 100$, the three tests share similar power. However, for larger d (e.g. $d = 500$), the decorrelated partial likelihood ratio test is the most powerful test. In addition, the Wald test is less effective for problems with higher dimensionality. On the basis of our simulation results, we recommend the use of the decorrelated partial likelihood ratio test for inference in high dimensional problems.

In the on-line supplementary materials, we conduct more thorough simulation studies to examine the empirical performance of our proposed methods. Specifically, we consider the simulation scenarios with non-sparse $\boldsymbol{\beta}^*$ (i.e. $s = 10$), different data-generating procedures for the covariate $\mathbf{X}_i$ and some high censoring settings. In addition, we carefully examine the bias, standard deviation, estimated standard deviation and empirical coverage probability for both zero components of $\boldsymbol{\beta}^*$ and non-zero components. Finally, we investigate the empirical performance of the proposed methods for inferring the baseline hazard function $\Lambda_0(t)$ under a variety of simulation scenarios. All these numerical results illustrate that the methods proposed work well in practice. For brevity, we refer the readers to the on-line supplementary materials for the detailed results.

6.2. Analysing a gene expression data set

We apply the proposed testing procedures to analyse a genomic data set, which is collected from a diffuse large B-cell lymphoma study analysed by Alizadeh *et al.* (2000). The data set can be downloaded from `http://llmpp.nih.gov/lymphoma/data.shtml`. One of the goals in this study is to investigate how different genes in B-cell malignancies are associated with the survival time. The expression values for over 13412 genes in B-cell malignancies are measured by microarray experiments. The data set contains 40 patients with diffuse large B-cell lymphoma who are recruited and followed until death or the end of the study. A small proportion (about 5%) of the gene expression values were not well measured and were treated as missing values by Alizadeh *et al.* (2000). For simplicity, we impute the missing values of each gene by the median of the observed values of the same gene. The average survival time is 43.9 months and the censoring rate is 55%.

We apply the proposed score, Wald and partial likelihood ratio tests to the data. The same strategy for choosing the tuning parameters as that in the simulation studies is adopted. We repeatedly apply the testing procedures for all parameters. To control the familywise error rate due to the multiple testing, the p-values are adjusted by Bonferroni's method. To be more conservative, we report only the genes with adjusted p-values that are less than 0.05 by all of the three methods in Table 3. Many of the genes which are significant in the hypothesis tests are biologically related to lymphoma. For instance, the relationship between lymphoma and genes FLT3 (Meierhoff *et al.*, 1995), CDC10 (Di Gaetano *et al.*, 2003), CHN2 (Nishiu *et al.*,

Table 3. Genes with adjusted *p*-values less than 0.05 by using score, Wald and partial likelihood ratio tests for the large B-cell lymphoma gene expression data set

Gene	*Results for the following tests:*		
	Score	*Wald*	*PLRT*
SP1	5.38×10^{-6}	2.17×10^{-5}	6.53×10^{-6}
PTMAP1	4.21×10^{-5}	6.35×10^{-5}	4.13×10^{-5}
Emv11	6.13×10^{-5}	1.81×10^{-4}	4.49×10^{-5}
CDC10	1.57×10^{-4}	4.91×10^{-4}	4.72×10^{-4}
NR2E3	2.41×10^{-3}	4.51×10^{-3}	2.65×10^{-3}
FLT3	1.75×10^{-3}	3.72×10^{-4}	2.52×10^{-4}
GPD2	3.85×10^{-3}	4.49×10^{-3}	5.66×10^{-4}
TAP2	4.39×10^{-3}	1.63×10^{-2}	6.97×10^{-3}
CHN2	5.65×10^{-3}	3.25×10^{-3}	7.15×10^{-4}
CD137	4.51×10^{-2}	2.91×10^{-3}	1.05×10^{-4}

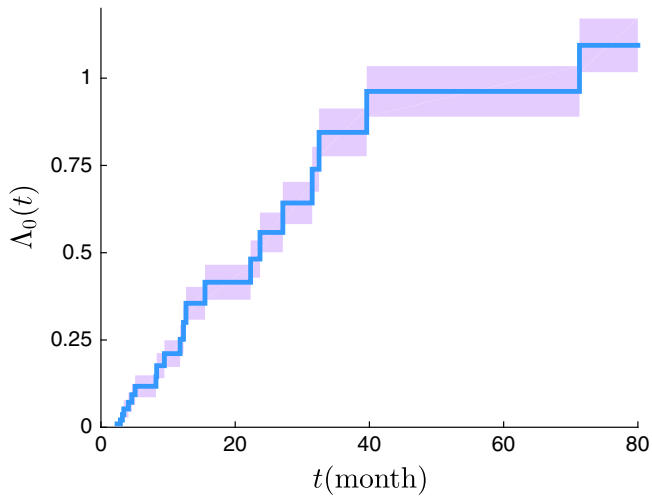

Fig. 3. Estimation and 95% confidence interval of the baseline hazard function

2002), Emv11 (Hiai *et al.*, 2003), CD137 (Alizadeh *et al.*, 2011) and TAP2 (Nielsen *et al.*, 2015) have been experimentally confirmed. This provides evidence that our methods can be used to discover findings in scientific applications involving high dimensional covariates. We further plot the estimated baseline hazard function and its 95% confidence interval in Fig. 3 for illustration.

7. Discussion

In this paper, we focus on Cox's proportional hazards model for univariate survival data. In practice, many biomedical studies involve multiple survival outcomes. For instance, in the Framingham Heart Study (Dawber, 1980), both time to coronary heart disease and time to

cerebrovascular accident are observed. We further extend the proposed procedures to analyse jointly multivariate survival data in the high dimensional setting in section H of the on-line supplementary materials.

The methods proposed involve two tuning parameters λ and λ'. The presence of multiple tuning parameters in inferential procedures has been encountered in many recent works (van de Geer *et al.*, 2014; Zhang and Zhang, 2014; Ning and Liu, 2016). Theoretically, we prove asymptotic normality of the test statistics when $\lambda \asymp \lambda' \asymp \sqrt{\{n^{-1}\log(d)\}}$. Empirically, our numerical results suggest that cross-validation can be used to determine the value of λ. Together with the choice of $\lambda' = \sqrt{\{n^{-1}\log(d)\}}$, we observe satisfactory type I errors in our numerical studies.

We comment that post-selection conditional inference (Lockhart *et al.*, 2014) and the proposed unconditional inference address different inferential problems. To be specific, consider the linear regression model $Y_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i + \epsilon_i$ $(i = 1, \ldots, n)$. Post-selection conditional inference aims to construct a 95% confidence interval of $\boldsymbol{\beta}^{\mathrm{P}}$, where

$$\boldsymbol{\beta}^{\mathrm{P}} = \arg\min_{\mathbf{b}^{\mathrm{P}}} \mathbb{E}(Y_i - \mathbf{X}_{iM}^{\mathrm{T}}\mathbf{b}^{\mathrm{P}})^2,$$

and $\mathbf{X}_{iM}$ denotes the components of $\mathbf{X}_i$ in the set $M \subset \{1, \ldots, d\}$. However, it is important to note that in general $\boldsymbol{\beta}^{\mathrm{P}} \neq \boldsymbol{\beta}^*_M$, where $\boldsymbol{\beta}^*_M$ are the components of the true value $\boldsymbol{\beta}^*$ in set M. In contrast, our unconditional inference constructs confidence intervals for the unknown value β^*_j for $1 \leqslant j \leqslant d$.

Whether conditional or unconditional inference is more appropriate depends on the context. For instance, in our real data applications, the goal is to study how genes in B-cell malignancies are associated with the survival time with all other genes adjusted. That means we are interested in constructing confidence intervals (or testing hypotheses) for the unknown true value β^*_j for all $1 \leqslant j \leqslant d$. However, the conditional inference on $\boldsymbol{\beta}^{\mathrm{P}}$ does not directly address this scientific question. Thus, our unconditional inference can be more appropriate in our real data application.

In practice, the method proposed has the following two added values, compared with the standard variable selection method. First, the p-values that are produced by our procedures provide an indication on how likely a covariate is associated with the survival time with all other covariates adjusted. The covariate with a smaller p-value means that it is statistically more significant in the joint Cox model. In contrast, the standard variable selection method (e.g. the lasso estimator) does not provide p-values. One cannot know how likely a covariate is associated with the survival time on the basis of the magnitude of the point estimator, because some covariates may have large coefficients as well as large standard deviations. Second, as seen in our real data analysis, the p-values can be adjusted by the Bonferroni method to control the familywise error rate at any given level (e.g. 0.05). This is a commonly used procedure in the analysis of genomic data, because it provides the explicit confidence level (e.g. 0.05) on quantifying the probability of false discoveries. However, such a measure of uncertainty is not provided by the standard variable selection method.

Acknowledgements

We thank the Joint Editor, Associate Editor and the three referees for their helpful comments, which significantly improved the paper. We also thank Professor Bradic for providing very helpful comments. This research is partially supported by National Science Foundation career grants DMS 1454377, IIS1408910 and IIS1332109, and National Institutes of Health grants R01MH102339, R01GM083084 and R01HG06841.

Appendix A: Proof of main theorems

In this appendix, we aim to prove our main result on the limiting distribution of the partial likelihood ratio test in theorem 3. Since this theorem is built on theorem 1, we first provide the proof of theorem 1.

A.1. Proof of theorem 1

We first provide a key lemma which characterizes the asymptotic normality of $\nabla\mathcal{L}(\boldsymbol{\beta}^*)$. This lemma is essential in our later proofs to derive the asymptotic distributions of the test statistics.

Lemma 3. Under assumptions 1, 4 and 5 for any vector $\mathbf{v}\in\mathbb{R}^d$, if $\|\mathbf{v}\|_0\leqslant s'$, $\|\mathbf{v}\|_2=1$,

$$\sup_{t\in[0,\tau]}\max_{i\in[n]}|\mathbf{X}_i^{\mathrm{T}}(t)\mathbf{v}|$$

is bounded and $\sqrt{\{s'\log(d/n)\}}=o(1)$, it holds that

$$\frac{\sqrt{n}\mathbf{v}^{\mathrm{T}}\nabla\mathcal{L}(\boldsymbol{\beta}^*)}{\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{H}^*\mathbf{v})}}\xrightarrow{\mathrm{d}}N(0,1),$$

where $\mathbf{H}^*$ is defined in equation (2.7).

Proof. Let $M_i(t)=N_i(t)-\int_0^t Y_i(u)\lambda_0(u)\,\mathrm{d}u$. By the definition of $\nabla\mathcal{L}(\boldsymbol{\beta}^*)$ in equation (2.4), we have

$$\begin{aligned}\nabla\mathcal{L}(\boldsymbol{\beta}^*)&=-\frac{1}{n}\sum_{i=1}^n\int_0^\tau\{\mathbf{X}_i(u)-\bar{\mathbf{Z}}(u,\boldsymbol{\beta}^*)\}\,\mathrm{d}M_i(u)\\&=-\frac{1}{n}\sum_{i=1}^n\int_0^\tau\{\mathbf{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\}\,\mathrm{d}M_i(u)-\frac{1}{n}\sum_{i=1}^n\int_0^\tau\{\mathbf{e}(u,\boldsymbol{\beta}^*)-\bar{\mathbf{Z}}(u,\boldsymbol{\beta}^*)\}\,\mathrm{d}M_i(u).\end{aligned}$$

Thus, by the identity $\mathbf{H}^*=\sqrt{n}\,\mathrm{var}\{\nabla\mathcal{L}(\boldsymbol{\beta}^*)\}$, we have

$$\begin{aligned}\frac{\sqrt{n}\mathbf{v}^{\mathrm{T}}\nabla\mathcal{L}(\boldsymbol{\beta}^*)}{\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{H}^*\mathbf{v})}}=&-\underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{v}^{\mathrm{T}}}{\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{H}^*\mathbf{v})}}\sum_{i=1}^n\int_0^\tau\{\mathbf{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\}\,\mathrm{d}M_i(u)}_{S}\\&-\underbrace{\frac{1}{\sqrt{n}}\frac{\mathbf{v}^{\mathrm{T}}}{\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{H}^*\mathbf{v})}}\sum_{i=1}^n\int_0^\tau\{\mathbf{e}(u,\boldsymbol{\beta}^*)-\bar{\mathbf{Z}}(u,\boldsymbol{\beta}^*)\}\,\mathrm{d}M_i(u)}_{E}.\end{aligned}$$

For the first term S, denote by

$$\xi_i=\frac{\mathbf{v}^{\mathrm{T}}}{\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{H}^*\mathbf{v})}}\int_0^\tau\{\mathbf{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\}\,\mathrm{d}M_i(u).$$

We have $\mathbb{E}(\xi_i)=0$, and $\mathrm{var}(n^{-1/2}S)=1$. Thus S is a sum of n independent random variables with mean 0. To obtain the asymptotic distribution of $n^{-1/2}S$, we verify the Lyapunov condition. Indeed,

$$\frac{1}{n^{3/2}}\sum_{i=1}^n\mathbb{E}\left|\frac{\mathbf{v}^{\mathrm{T}}}{\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{H}^*\mathbf{v})}}\int_0^\tau\{\mathbf{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\}\,\mathrm{d}M_i(u)\right|^3\leqslant\frac{C}{C_h^{3/2}n^{3/2}}\sum_{i=1}^n\sup_{u\in[0,\tau]}|\mathbf{v}^{\mathrm{T}}\{\mathbf{X}_i(u)-\mathbf{e}(u,\boldsymbol{\beta}^*)\}|^3=\mathcal{O}(n^{-1/2}),$$

where the inequality follows by assumption 5 for some constant C. Thus, the Lyapunov condition holds. Applying the Lindeberg–Feller central limit theorem, we have $n^{-1/2}S\to^{\mathrm{d}}N(0,1)$.

Next, we prove that the second term $E=o_{\mathbb{P}}(1)$. Since

$$\begin{aligned}E&=\frac{1}{\sqrt{n}}\frac{\mathbf{v}^{\mathrm{T}}}{\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{H}^*\mathbf{v})}}\sum_{i=1}^n\int_0^\tau[\{\mathbf{e}(u,\boldsymbol{\beta}^*)-\bar{\mathbf{Z}}(u,\boldsymbol{\beta}^*)\}\mathbf{1}\,\mathrm{d}M_i(u)]\\&\leqslant\frac{1}{\sqrt{n}}\frac{s'^{1/2}}{\lambda_{\min}}\sup_{u\in[0,\tau]}\|\mathbf{e}(u,\boldsymbol{\beta}^*)-\bar{\mathbf{Z}}(u,\boldsymbol{\beta}^*)\|_\infty\int_0^\tau\left|\sum_{i=1}^n\mathbf{1}\,\mathrm{d}M_i(u)\right|.\end{aligned}$$

By lemma E.1 in the on-line supplementary materials, it holds that $\sup_{u\in[0,\tau]}\|\mathbf{e}(u,\boldsymbol{\beta}^*)-\bar{\mathbf{Z}}(u,\boldsymbol{\beta}^*)\|_\infty=\mathcal{O}_{\mathbb{P}}[\sqrt{\{n^{-1}\log(d)\}}]$. It holds that, for some constant $C>0$,

$$E \leqslant \frac{C}{\sqrt{n}} \frac{1}{\lambda_{\min}} \sqrt{\left\{ \frac{s' \log(d)}{n} \right\}} \int_0^\tau \left| \sum_{i=1}^n \mathbf{1} \, \mathrm{d}M_i(u) \right|.$$

It remains to bound the term $\int_0^\tau |\Sigma_{i=1}^n \mathbf{1} \, \mathrm{d}M_i(u)|$. By theorem 2.11.9 and example 2.11.16 of van der Vaart and Wellner (1996), $\tilde{G}(t) := n^{-1/2} \Sigma_{i=1}^n M_i(t)$ converges weakly to a tight Gaussian process $G(t)$. Furthermore, by the strong embedding theorem of Shorack and Wellner (2009), there is another probability space such that $(S^{*(0)}(\boldsymbol{\beta}, t), S^{*(1)}(\boldsymbol{\beta}, t), \tilde{G}^*(t))$ converges almost surely to $(s^{*(0)}(\boldsymbol{\beta}, t), \mathbf{s}^{*(1)}(\boldsymbol{\beta}, t), G^*(t))$, where '$*$' indicates the existences in a new probability space. This implies that $\int_0^\tau |\mathrm{d}G(t)| = \int_0^\tau |\mathrm{d}G^*(t)| + o_{\mathbb{P}}(1)$. We have, by our assumption that $\sqrt{\{s' \log(d/n)\}} = o_{\mathbb{P}}(1)$, the term E satisfies that

$$E = \mathcal{O}_{\mathbb{P}} \left[\sqrt{\left\{ \frac{s' \log(d)}{n} \right\}} \right] = o_{\mathbb{P}}(1).$$

Together with the result that $n^{-1/2} S \to \mathrm{d}N(0, 1)$, we conclude the proof. □

Before proving theorem 1, we need some technical lemmas to characterize several concentration results. We present them here and defer the proofs to section F of the on-line supplementary materials.

Lemma 4. Under assumptions 1–5 there is a positive constant C, such that, with probability at least $1 - \mathcal{O}(d^{-3})$,

$$\|\nabla \mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leqslant C \sqrt{\left\{ \frac{\log(d)}{n} \right\}}.$$

Lemma 5. Under assumptions 1–5, let $\hat{\boldsymbol{\beta}}$ be the estimator for $\boldsymbol{\beta}^*$ estimated by expression (2.1) satisfying the result in expression (2.2) that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$ with $\lambda \asymp \mathcal{O}[\sqrt{\{n^{-1} \log(d)\}}]$. Then, for any $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + u(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ with $u \in [0, 1]$, we have $\|\nabla^2 \mathcal{L}(\tilde{\boldsymbol{\beta}})\|_\infty = \mathcal{O}_{\mathbb{P}}(1)$,

$$\|\nabla^2 \mathcal{L}(\tilde{\boldsymbol{\beta}}) - \mathbf{H}^*\|_\infty = \mathcal{O}_{\mathbb{P}} \left[s \sqrt{\left\{ \frac{\log(d)}{n} \right\}} \right]$$

and

$$\|\{\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(\tilde{\boldsymbol{\beta}}) - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\} \mathbf{w}^*\|_\infty = \mathcal{O}_{\mathbb{P}} \left[s \sqrt{\left\{ \frac{\log(d)}{n} \right\}} \right].$$

A.1.1. Proof of theorem 1

To derive the asymptotic distribution of $\sqrt{n} \hat{U}(0, \hat{\boldsymbol{\theta}})$, we start with decomposing $\hat{U}(0, \hat{\boldsymbol{\theta}})$ into several terms:

$$\begin{aligned} \hat{U}(0, \hat{\boldsymbol{\theta}}) &= \nabla_\alpha \mathcal{L}(0, \hat{\boldsymbol{\theta}}) - \hat{\mathbf{w}}^{\mathrm{T}} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \hat{\boldsymbol{\theta}}) \\ &= \nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*) + \nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \{\hat{\mathbf{w}}^{\mathrm{T}} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*) + \hat{\mathbf{w}}^{\mathrm{T}} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\} \\ &= \underbrace{\nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*) - \mathbf{w}^{*\mathrm{T}} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)}_{S} + \underbrace{(\mathbf{w}^* - \hat{\mathbf{w}})^{\mathrm{T}} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)}_{E_1} + \underbrace{\{\nabla^2_{\alpha\boldsymbol{\theta}} \mathcal{L}(0, \bar{\boldsymbol{\theta}}) - \hat{\mathbf{w}}^{\mathrm{T}} \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(0, \tilde{\boldsymbol{\theta}})\}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_2}, \end{aligned} \tag{A.1}$$

where the second equality holds by the mean value theorem for some $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $u, u' \in [0, 1]$.

We consider the terms S, E_1 and E_2 separately. For the first term S, by lemma 3, taking $\mathbf{v} = (1, -\mathbf{w}^{*\mathrm{T}})^{\mathrm{T}}$, we have

$$\sqrt{n} S \xrightarrow{\mathrm{d}} Z, \qquad Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}). \tag{A.2}$$

For the term E_1, we have,

$$E_1 \leqslant \|\hat{\mathbf{w}} - \mathbf{w}^*\|_1 \|\nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}[s' \lambda' \sqrt{\{n^{-1} \log(d)\}}], \tag{A.3}$$

where $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s' \lambda')$ by lemma 1, and $\|\nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}[\sqrt{\{n^{-1} \log(d)\}}]$ by lemma 4.

For the term E_2, we have

$$E_2 = \underbrace{\{\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(0,\bar{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\tilde{\boldsymbol{\theta}})\}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_{21}} + \underbrace{(\mathbf{w}^* - \hat{\mathbf{w}})^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_{22}}. \tag{A.4}$$

Considering the terms E_{21} and E_{22} separately, first, we have

$$E_{21} = \nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(0,\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \mathbf{w}^{*\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \mathcal{O}_{\mathbb{P}}\{n^{-1}s\log(d)\}, \tag{A.5}$$

where the last equality holds by the proof of lemma 2 in the on-line supplementary materials.

For the second term E_{22} in equation (A.4), we have, by the Cauchy–Schwarz inequality,

$$|E_{22}| \leqslant \tfrac{1}{2}(\hat{\mathbf{w}} - \mathbf{w}^*)^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\tilde{\boldsymbol{\theta}})(\hat{\mathbf{w}} - \mathbf{w}^*) + \tfrac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \mathcal{O}_{\mathbb{P}}\{n^{-1}s\log(d)\}, \tag{A.6}$$

where the last equality follows by expression (A.5) in the supplementary materials. Plugging expressions (A.5) and (A.6) into equation (A.4), we have $E_2 = \mathcal{O}_{\mathbb{P}}\{n^{-1}s\log(d)\}$. Combining with expression (A.3), we have

$$|E_1| + |E_2| = \mathcal{O}_{\mathbb{P}}\left\{\frac{s\log(d)}{n}\right\} = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right), \tag{A.7}$$

where the last equality holds by the assumption that $n^{-1/2}s\log(d) = o(1)$. Combining expressions (A.7), (A.2) and (A.1), our claim (4.2) holds as desired.

A.2. Proof of theorem 3

We have

$$\begin{aligned}\mathcal{L}(\tilde{\alpha}, \hat{\boldsymbol{\theta}} - \tilde{\alpha}\hat{\mathbf{w}}) - \mathcal{L}(0,\hat{\boldsymbol{\theta}}) &= \tilde{\alpha}\nabla_{\alpha}\mathcal{L}(0,\hat{\boldsymbol{\theta}}) - \tilde{\alpha}\hat{\mathbf{w}}^{\mathrm{T}}\nabla_{\boldsymbol{\theta}}\mathcal{L}(0,\hat{\boldsymbol{\theta}}) + \frac{\tilde{\alpha}^2}{2}\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha},\hat{\boldsymbol{\theta}}) \\ &\quad + \frac{\tilde{\alpha}^2}{2}\hat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\bar{\boldsymbol{\theta}})\hat{\mathbf{w}} - \tilde{\alpha}^2\hat{\mathbf{w}}^{\mathrm{T}}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\bar{\alpha}',\hat{\boldsymbol{\theta}}) \\ &= \underbrace{\tilde{\alpha}\hat{U}(0,\hat{\boldsymbol{\theta}})}_{T_1} + \underbrace{\frac{\tilde{\alpha}^2}{2}\{\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha},\hat{\boldsymbol{\theta}}) + \hat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\bar{\boldsymbol{\theta}})\hat{\mathbf{w}} - 2\hat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\bar{\alpha}',\bar{\boldsymbol{\theta}}')\}}_{T_2},\end{aligned} \tag{A.8}$$

where the first equality follows by the mean value theorem with $\bar{\alpha} = u_1\hat{\alpha}$, $\bar{\alpha}' = u_2\hat{\alpha}$, $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u_3(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $\bar{\boldsymbol{\theta}}' = \boldsymbol{\theta}^* + u_4(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ for some $0 \leqslant u_1, u_2, u_3, u_4 \leqslant 1$.

We first look at the term T_1. Under the null hypothesis $\alpha^* = 0$, $\sqrt{n}\hat{U}(0,\hat{\boldsymbol{\theta}}) \to \mathrm{d}Z$ and $\sqrt{n}\tilde{\alpha} = -H^{-1}_{\alpha|\boldsymbol{\theta}}\hat{U}(0,\hat{\boldsymbol{\theta}}) + o_{\mathbb{P}}(1)$ by theorems 1 and 2, where $Z \sim N(0, H_{\alpha|\boldsymbol{\theta}})$. We have

$$2nT_1 = -2\hat{U}^2(0,\hat{\boldsymbol{\theta}}) + o_{\mathbb{P}}(1) \xrightarrow{\mathrm{d}} -2Z^2H^{-1}_{\alpha|\boldsymbol{\theta}}. \tag{A.9}$$

Next, we look at the term T_2,

$$\begin{aligned}T_2 &= \underbrace{\frac{\tilde{\alpha}^2}{2}(\mathbf{H}^*_{\alpha\alpha} + \mathbf{H}_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - 2\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha})}_{T_{21}} \\ &\quad + \underbrace{\frac{\tilde{\alpha}^2}{2}[\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha},\hat{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\alpha} + \hat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\bar{\boldsymbol{\theta}})\hat{\mathbf{w}} - \mathbf{w}^*\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{w}^* - 2\{\tilde{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\bar{\alpha}',\bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{w}^*\}]}_{T_{22}}.\end{aligned} \tag{A.10}$$

It holds by theorem 2 that $\sqrt{n}\tilde{\alpha} \to^{\mathrm{d}} H^{-1}_{\alpha\boldsymbol{\theta}}Z$. Together with $H^*_{\alpha|\boldsymbol{\theta}} = \mathcal{O}(1)$, we have

$$2nT_{21} = n\tilde{\alpha}^2H_{\alpha|\boldsymbol{\theta}} \xrightarrow{\mathrm{d}} H^{-1}_{\alpha|\boldsymbol{\theta}}Z^2. \tag{A.11}$$

Considering the term T_{22}, we have

$$T_{22} = \frac{\tilde{\alpha}^2}{2}[\underbrace{\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \hat{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\alpha}}_{R_1} + \underbrace{\hat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}})\hat{\mathbf{w}} - \mathbf{w}^*\mathbf{H}^*_{\boldsymbol{\theta\theta}}\mathbf{w}^*}_{R_2}$$
$$- \underbrace{2\{\tilde{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{w}^{*\mathrm{T}}\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\}}_{R_3}]. \tag{A.12}$$

For the first term $|R_1|$, we have, by lemma 5, $|R_1| = |\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \hat{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\alpha}| = \mathcal{O}_{\mathbb{P}}(s\lambda)$. For the second term, we have

$$|R_2| = |\hat{\mathbf{w}}^{\mathrm{T}}\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}})\hat{\mathbf{w}} - \mathbf{w}^*\mathbf{H}^*_{\boldsymbol{\theta\theta}}\mathbf{w}^*| = \mathcal{O}_{\mathbb{P}}(s\lambda), \tag{A.13}$$

where the last equality follows by the same arguments as in the proof of lemma 1 in the supplementary materials. For the third term $|R_3|$, we can similarly show that

$$|R_3| \leqslant 2[|\{\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\}\hat{\mathbf{w}}| + |\mathbf{H}^*_{\alpha\boldsymbol{\theta}}(\hat{\mathbf{w}} - \mathbf{w}^*)|] = \mathcal{O}_{\mathbb{P}}[s\sqrt{\{n^{-1}\log(d)\}}], \tag{A.14}$$

where the last equality follows by lemma E.4 in the supplementary materials and lemma 5. Combining the results above, we have

$$T_{22} = \frac{\tilde{\alpha}^2}{2}\mathcal{O}_{\mathbb{P}}(s\lambda) = \mathcal{O}_{\mathbb{P}}\left\{\frac{s\log(d)}{n^{3/2}}\right\} = o_{\mathbb{P}}(n^{-1}), \tag{A.15}$$

where the second equality follows by theorem 2 that $\tilde{\alpha} = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ under the null hypothesis, and the last equality follows by the assumption that $n^{-1/2}s\log(d) = o(1)$. Combining equations (A.11) and (A.15) with equation (A.10) we have

$$2nT_2 \xrightarrow{\mathrm{d}} H^{-1}_{\alpha|\boldsymbol{\theta}}Z^2, \qquad Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}). \tag{A.16}$$

Plugging expressions (A.9) and (A.16) into equation (A.8), by theorem 1,

$$-2n\{\mathcal{L}(\tilde{\alpha}, \hat{\boldsymbol{\theta}} - \tilde{\alpha}\hat{\mathbf{w}}) - \mathcal{L}(0, \hat{\boldsymbol{\theta}})\} \xrightarrow{\mathrm{d}} Z^2_\chi, \qquad Z_\chi \sim \chi^2_1,$$

which concludes the proof.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, Jr, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. and Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alizadeh, A. A., Gentles, A. J., Alencar, A. J., Liu, C. L., Kohrt, H. E., Houot, R., Goldstein, M. J., Zhao, S., Natkunam, Y., Advani, R. H., Gascoyne, R. D., Briones, J., Tibshirani, R. J., Myklebust, J. H., Plevritis, S. K., Lossos, I. S. and Levy, R. (2011) Prediction of survival in diffuse large B-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. *Blood*, **118**, 1350–1358.

Andersen, P. K. and Gill, R. D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, **10**, 1100–1120.

Antoniadis, A., Fryzlewicz, P. and Letué, F. (2010) The Dantzig selector in Cox's proportional hazards model. *Scand. J. Statist.*, **37**, 531–552.

Belloni, A., Chernozhukov, V. and Wei, Y. (2016) Post-selection inference for generalized linear models with many controls. *J. Bus. Econ. Statist.*, to be published.

Bradic, J., Fan, J. and Jiang, J. (2011) Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.*, **39**, 3092–3120.

Cai, J., Fan, J., Li, R. and Zhou, H. (2005) Variable selection for multivariate failure time data. *Biometrika*, **92**, 303–316.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.

Dawber, T. R. (1980) *The Framingham Study: the Epidemiology of Atherosclerotic Disease*. Cambridge: Harvard University Press.

Di Gaetano, N., Cittera, E., Nota, R., Vecchi, A., Grieco, V., Scanziani, E., Botto, M., Introna, M. and Golay, J. (2003) Complement activation determines the therapeutic activity of rituximab in vivo. *J. Immunol.*, **171**, 1581–1587.

Fan, J. and Li, R. (2002) Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, **30**, 74–99.
van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–1202.
Hiai, H., Tsuruyama, T. and Yamada, Y. (2003) Pre-B lymphomas in SL/Kh mice: a multifactorial disease model. *Cancer Sci.*, **94**, 847–850.
Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013) Oracle inequalities for the Lasso in the Cox model. *Ann. Statist.*, **41**, 1142–1165.
Kalbfleisch, J. D. and Prentice, R. L. (2011) *The Statistical Analysis of Failure Time Data*. New York: Wiley.
Kong, S. and Nan, B. (2014) Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Statist. Sin.*, **24**, 25–42.
Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014) A significance test for the Lasso. *Ann. Statist.*, **42**, 413–468.
Massart, P. (2007) *Concentration Inequalities and Model Selection*. New York: Springer.
Meierhoff, G., Dehmel, U., Gruss, H., Rosnet, O., Birnbaum, D., Quentmeier, H., Dirks, W. and Drexler, H. (1995) Expression of FLT3 receptor and FLT3-ligand in human leukemia-lymphoma cell lines. *Leukemia*, **9**, 1368–1372.
Nielsen, K. R., Steffensen, R., Bendtsen, M. D., Rodrigo-Domingo, M., Baech, J., Haunstrup, T. M., Bergkvist, K. S., Schmitz, A., Boedker, J. S., Johansen, P., Dybkaeær, K., Boeøgsted, M. and Johnsen, H. E. (2015) Inherited inflammatory response genes are associated with B-cell non-Hodgkins lymphoma risk and survival. *PLOS ONE*, **10**, article e0139329.
Ning, Y. and Liu, H. (2016) A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.*, to be published.
Nishiu, M., Yanagawa, R., Nakatsuka, S.-I., Yao, M., Tsunoda, T., Nakamura, Y. and Aozasa, K. (2002) Microarray analysis of gene-expression profiles in diffuse large b-cell lymphoma: identification of genes related to disease progression. *Cancer Sci.*, **93**, 894–901.
Shorack, G. R. and Wellner, J. A. (2009) *Empirical Processes with Applications to Statistics*. Philadelphia: Society for Industrial and Applied Mathematics.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tibshirani, R. (1997) The Lasso method for variable selection in the Cox model. *Statist. Med.*, **16**, 385–395.
Tsiatis, A. A. (1981) A large sample study of Cox's regression model. *Ann. Statist.*, **9**, 93–108.
van der Vaart, A. W. (2000) *Asymptotic Statistics*. New York: Cambridge University Press.
van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer.
Wang, S., Nan, B., Zhu, N. and Zhu, J. (2009) Hierarchically penalized Cox regression with grouped variables. *Biometrika*, **96**, 307–322.
Zhang, H. H. and Lu, W. (2007) Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691–703.
Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B, **76**, 217–242.
Zhao, S. D. and Li, Y. (2012) Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multiv. Anal.*, **105**, 397–411.
Zhong, P.-S., Hu, T. and Li, J. (2015) Tests for coefficients in high-dimensional additive hazard models. *Scand. J. Statist.*, **42**, 649–664.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials to "Testing and confidence intervals for high dimensional proportional hazards models"'.

Supplementary Materials to "Testing and Confidence Intervals for High Dimensional Proportional Hazards Model"

Ethan X. Fang* Yang Ning† Han Liu‡

A Proofs in Section 4

In this section, we provide the detailed proofs in Section 4.

Proof of Lemma 4.4. Let $M(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\mathbf{w} - \mathbf{w}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) + \lambda'\|\mathbf{w}\|_1$. By the optimality of $\widehat{\mathbf{w}}$, we have $M(\widehat{\mathbf{w}}) \leq M(\mathbf{w}^*)$. Letting $\widehat{\boldsymbol{\Delta}} = \widehat{\mathbf{w}} - \mathbf{w}^*$, we have

$$\begin{aligned}\frac{1}{2}\widehat{\boldsymbol{\Delta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\Delta}} &\leq \widehat{\boldsymbol{\Delta}}^T(\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\mathbf{w}^*) + \lambda'\|\mathbf{w}^*\|_1 - \lambda'\|\widehat{\mathbf{w}}\|_1 \\ &= \underbrace{\widehat{\boldsymbol{\Delta}}^T(\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\boldsymbol{\beta}^*) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\mathbf{w}^*)}_{I_1} + \underbrace{\lambda'\|\mathbf{w}^*\|_1 - \lambda'\|\widehat{\mathbf{w}}\|_1}_{I_2} \\ &\quad + \underbrace{\widehat{\boldsymbol{\Delta}}^T\Big[(\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\boldsymbol{\beta}^*)) - (\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*))\mathbf{w}^*\Big]}_{I_3}. \qquad \text{(A.1)}\end{aligned}$$

where $\widehat{\boldsymbol{\Delta}} = \widehat{\mathbf{w}} - \mathbf{w}^*$. For the first term I_1, we have

$$|I_1| \leq \|\widehat{\boldsymbol{\Delta}}\|_1 \cdot \|\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\boldsymbol{\beta}^*) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\mathbf{w}^*\|_\infty \leq C\sqrt{\frac{\log d}{n}}\|\widehat{\boldsymbol{\Delta}}\|_1,$$

where the last step follows from Lemma E.3, and C is some positive constant.

For I_2, denoting by $\mathcal{S}$ the support of $\mathbf{w}^*$, by the triangle inequality, it is seen that

$$I_2 = \lambda'\|\mathbf{w}^*_{\mathcal{S}}\|_1 - \lambda'\|\widehat{\mathbf{w}}_{\mathcal{S}}\|_1 - \lambda'\|\widehat{\mathbf{w}}_{\mathcal{S}^c}\|_1 \leq \lambda'\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_1 - \lambda'\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_1,$$

*Department of Statistics, Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802, USA; e-mail: xxf13@psu.edu

†Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA; email: yn265@cornell.edu

‡Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu

where we use the fact that $\|\mathbf{w}^*_{\mathcal{S}^c}\|_1 = 0$.

Finally, we consider I_3. Let $a_i = \boldsymbol{\Delta}_1^T\{\boldsymbol{X}_i(t) - \overline{\boldsymbol{Z}}(t, \boldsymbol{\beta}^*)\}$, where $\boldsymbol{\Delta}_1 = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Let $w_i = Y_i(t)\exp\{\boldsymbol{\beta}^{*T}\boldsymbol{X}_i(t)\}$. It is not difficult to verify that

$$\nabla^2\mathcal{L}(\boldsymbol{\beta}^*) = \frac{1}{n}\int_0^\tau \frac{\sum_{i,j} w_i w_j\{\boldsymbol{X}_i(t) - \boldsymbol{X}_j(t)\}^{\otimes 2}}{\sum_{i,j} 2w_i w_j} d\overline{N}(t),$$

$$\nabla^2\mathcal{L}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\int_0^\tau \frac{\sum_{i,j} w_i w_j\{\boldsymbol{X}_i(t) - \boldsymbol{X}_j(t)\}^{\otimes 2}\exp(a_i + a_j)}{\sum_{i,j} 2w_i w_j\exp(a_i + a_j)} d\overline{N}(t).$$

For notational simplicity, let $\boldsymbol{X}_i = (X_{i\alpha}, \boldsymbol{X}_{i\boldsymbol{\theta}}^T)^T$, $h_{ij} = \widehat{\boldsymbol{\Delta}}^T\{\boldsymbol{X}_{i\boldsymbol{\theta}}(t) - \boldsymbol{X}_{j\boldsymbol{\theta}}(t)\}$, $g_{ij} = \boldsymbol{\Delta}_1^T\{\boldsymbol{X}_i(t) - \boldsymbol{X}_j(t)\}$ and $b_{ij} = X_{i\alpha} - X_{j\alpha}$. We now focus on $\widehat{\boldsymbol{\Delta}}^T\{\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\boldsymbol{\beta}^*)\}$, which is equal to

$$\begin{aligned}
&\frac{1}{n}\int_0^\tau \Big\{\frac{\sum_{i,j} w_i w_j h_{ij} b_{ij}\exp(a_i + a_j)}{\sum_{i,j} 2w_i w_j\exp(a_i + a_j)} - \frac{\sum_{i,j} w_i w_j h_{ij} b_{ij}}{\sum_{i,j} 2w_i w_j}\Big\} d\overline{N}(t) \\
&= \underbrace{\frac{1}{n}\int_0^\tau \frac{\sum_{i,j} w_i w_j h_{ij} b_{ij}[\exp(a_i + a_j) - 1]}{\sum_{i,j} 2w_i w_j} d\overline{N}(t)}_{I_{31}} \\
&+ \underbrace{\frac{1}{n}\int_0^\tau \Big[\sum_{i,j} w_i w_j h_{ij} b_{ij}\exp(a_i + a_j)\Big]\Big[\frac{1}{\sum_{i,j} 2w_i w_j\exp(a_i + a_j)} - \frac{1}{\sum_{i,j} 2w_i w_j}\Big] d\overline{N}(t)}_{I_{32}}.
\end{aligned}$$

For the term I_{31}, by Cauchy-Schwarz inequality,

$$\begin{aligned}
&\frac{\sum_{i,j} w_i w_j h_{ij} b_{ij}[\exp(a_i + a_j) - 1]}{\sum_{i,j} 2w_i w_j} \le \sqrt{\frac{\sum_{i,j} w_i w_j h_{ij}^2 b_{ij}^2}{\sum_{i,j} 2w_i w_j}}\sqrt{\frac{\sum_{i,j} w_i w_j[\exp(a_i + a_j) - 1]^2}{\sum_{i,j} 2w_i w_j}} \\
&\lesssim \sqrt{\frac{\sum_{i,j} w_i w_j h_{ij}^2 b_{ij}^2}{\sum_{i,j} 2w_i w_j}}\sqrt{\frac{\sum_{i,j} w_i w_j(a_i + a_j)^2}{\sum_{i,j} 2w_i w_j}} = \sqrt{\frac{\sum_{i,j} w_i w_j h_{ij}^2 b_{ij}^2}{\sum_{i,j} 2w_i w_j}}\sqrt{\frac{\sum_{i,j} w_i w_j g_{ij}^2}{\sum_{i,j} 2w_i w_j}},
\end{aligned}$$

where the second step follows from the proof of Lemma 3.2 in Huang et al. (2013) and $\exp(x) - 1 \lesssim x$ for $x = o(1)$, and the last step follows from $\sum_i a_i w_i = 0$. Thus, applying Cauchy-Schwarz inequality again, we obtain

$$|I_{31}| \lesssim \sqrt{\widehat{\boldsymbol{\Delta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\widehat{\boldsymbol{\Delta}}}\sqrt{\widehat{\boldsymbol{\Delta}}_1^T\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\widehat{\boldsymbol{\Delta}}_1} \lesssim \sqrt{\frac{s\log d}{n}}\sqrt{\widehat{\boldsymbol{\Delta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\widehat{\boldsymbol{\Delta}}},$$

where the last step follows from the proof of Theorem 3.1 in Huang et al. (2013). We now consider

the term I_{32}. By similar arguments, we can show that

$$
\begin{aligned}
|I_{32}| &\lesssim \frac{1}{n}\int_0^\tau \frac{\sum_{i,j} w_i w_j h_{ij} b_{ij} \exp(a_i + a_j)}{\sum_{i,j} 2w_i w_j \exp(a_i + a_j)} \sqrt{\frac{\sum_{i,j} 2w_i w_j g_{ij}^2}{\sum_{i,j} 2w_i w_j}} d\overline{N}(t) \\
&\leq \frac{1}{n}\int_0^\tau \sqrt{\frac{\sum_{i,j} w_i w_j h_{ij}^2 b_{ij}^2 \exp(a_i + a_j)}{\sum_{i,j} 2w_i w_j \exp(a_i + a_j)}} \sqrt{\frac{\sum_{i,j} w_i w_j \exp(a_i + a_j)}{\sum_{i,j} 2w_i w_j \exp(a_i + a_j)}} \sqrt{\frac{\sum_{i,j} 2w_i w_j g_{ij}^2}{\sum_{i,j} 2w_i w_j}} d\overline{N}(t) \\
&\lesssim \sqrt{\frac{1}{n}\int_0^\tau \frac{\sum_{i,j} w_i w_j h_{ij}^2 b_{ij}^2 \exp(a_i + a_j)}{\sum_{i,j} 2w_i w_j \exp(a_i + a_j)} d\overline{N}(t)} \sqrt{\frac{1}{n}\int_0^\tau \frac{\sum_{i,j} 2w_i w_j g_{ij}^2}{\sum_{i,j} 2w_i w_j} d\overline{N}(t)} \\
&\lesssim \sqrt{\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\Delta}}} \sqrt{\widehat{\boldsymbol{\Delta}}_1^T \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}_1} \lesssim \sqrt{\frac{s \log d}{n}} \sqrt{\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}},
\end{aligned}
$$

where in the last step we use the fact that

$$
\left| \widehat{\boldsymbol{\Delta}}^T \left\{ \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \right\} \widehat{\boldsymbol{\Delta}} \right| \lesssim s \sqrt{\frac{\log d}{n}} |\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}|,
$$

which further implies

$$
|\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}| \lesssim |\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\Delta}}| \lesssim |\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}|,
$$

given $s\sqrt{\log d/n} = o(1)$. Combining the bounds for I_{31} and I_{32}, we have that

$$
|I_3| \leq C'' \sqrt{\frac{s \log d}{n}} \sqrt{\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}}.
$$

Choosing $\lambda' = 2C\sqrt{\log d/n}$ in (A.1) and by the previous arguments, we obtain

$$
\begin{aligned}
\frac{1}{2} \widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}} &\leq C'' \sqrt{\frac{s \log d}{n}} \cdot \sqrt{\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}} \\
&\quad + 3C \sqrt{\frac{\log d}{n}} \cdot \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_1 - C \sqrt{\frac{\log d}{n}} \cdot \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_1. \qquad \text{(A.2)}
\end{aligned}
$$

If $[\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}]^{1/2} > 2C'' \sqrt{s \log d/n}$, (A.2) implies $\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_1 \leq 3\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_1$. By Lemma E.5, it holds that

$$
\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\Delta}} \geq \frac{1}{2} \kappa^2(1, s'; \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)) \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_2^2,
$$

which implies that

$$
\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{\Delta}} \geq \frac{1}{2} \kappa^2(1, s'; \nabla^2 \mathcal{L}(\boldsymbol{\beta}^*)) s'^{-1} \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_1^2.
$$

By plugging into (A.2), we have

$$
[\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}]^{1/2} \lesssim \sqrt{\frac{(s + s') \log d}{n}}. \qquad \text{(A.3)}
$$

If $6\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_1 \geq \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}^c}\|_1$, by the same argument, we have

$$
\|\widehat{\boldsymbol{\Delta}}\|_1 \leq 7\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}}\|_1 \lesssim \sqrt{s'} \cdot [-\widehat{\boldsymbol{\Delta}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\beta}^*) \widehat{\boldsymbol{\Delta}}]^{1/2} \lesssim (s + s') \sqrt{\frac{\log d}{n}}.
$$

On the other hand, if $6\|\widehat{\mathbf{\Delta}}_{\mathcal{S}}\|_1 \le \|\widehat{\mathbf{\Delta}}_{\mathcal{S}^c}\|_1$, (A.2) implies

$$0 \le \frac{1}{2}\widehat{\mathbf{\Delta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\widehat{\mathbf{\Delta}} \le C''\sqrt{\frac{s\log d}{n}}\cdot\sqrt{\widehat{\mathbf{\Delta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\widehat{\mathbf{\Delta}}} - \frac{C}{2}\sqrt{\frac{\log d}{n}}\cdot\|\widehat{\mathbf{\Delta}}_{\mathcal{S}^c}\|_1. \tag{A.4}$$

Combining (A.3) and (A.4),

$$\|\widehat{\mathbf{\Delta}}\|_1 \le \frac{7}{6}\|\widehat{\mathbf{\Delta}}_{\mathcal{S}^c}\|_1 \lesssim \sqrt{s}\cdot[\widehat{\mathbf{\Delta}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\widehat{\mathbf{\Delta}}]^{1/2} \lesssim (s+s')\sqrt{\frac{\log d}{n}}. \tag{A.5}$$

Thus, in both cases, we have $\|\widehat{\mathbf{\Delta}}\|_1 \lesssim (s+s')\sqrt{\log d/n}$. □

Proof of Lemma 4.7. By the definition of $H_{\alpha|\boldsymbol{\theta}}$ and $\widehat{H}_{\alpha|\boldsymbol{\theta}}$, we have

$$|H_{\alpha|\boldsymbol{\theta}} - \widehat{H}_{\alpha|\boldsymbol{\theta}}| \le \underbrace{|\mathbf{H}^*_{\alpha\alpha} - \nabla^2_{\alpha\alpha}\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})|}_{E_1} + \underbrace{|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})|}_{E_2}. \tag{A.6}$$

We consider the two terms separately. For the first term E_1, we have by Lemma I.3, $E_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$. For the second term E_2, we have,

$$\begin{aligned} E_2 &= |\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})| = |\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \widehat{\mathbf{w}}^T\mathbf{H}^*_{\boldsymbol{\theta}\alpha} + \widehat{\mathbf{w}}^T\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})| \\ &\le \underbrace{|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \widehat{\mathbf{w}}^T\mathbf{H}^*_{\boldsymbol{\theta}\alpha}|}_{E_{21}} + \underbrace{|\widehat{\mathbf{w}}^T\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})|}_{E_{22}}. \end{aligned}$$

For the term E_{21}, we have, by Hölder's inequality,

$$E_{21} \le \|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \widehat{\mathbf{w}}^T\|_1\|\mathbf{H}^*_{\boldsymbol{\theta}\alpha}\|_\infty = \mathcal{O}_{\mathbb{P}}(s'\lambda'), \tag{A.7}$$

where the last inequality holds by the fact that $\|\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \widehat{\mathbf{w}}^T\|_1 = \mathcal{O}_{\mathbb{P}}(s'\lambda')$, and $\|\mathbf{H}^*_{\boldsymbol{\theta}\alpha}\|_\infty = \mathcal{O}(1)$ by Assumption 4.3.

For the second term E_{22}, we have, by Hölder's inequality,

$$E_{22} \le \big|\widehat{\mathbf{w}}^T\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \mathbf{w}^{*T}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}\big| + \big|\mathbf{w}^{*T}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})\big| = \mathcal{O}_{\mathbb{P}}(s'\lambda') \tag{A.8}$$

where the last equality holds by the result $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s'\lambda')$ by (4.1) and by Lemma I.3 that $\|\mathbf{w}^{*T}\mathbf{H}^* - \widehat{\mathbf{w}}^T\nabla^2\mathcal{L}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})\|_\infty = \mathcal{O}_{\mathbb{P}}\big((s+s')\lambda\big)$.

Combining (A.7) and (A.8), we have, $E_2 \le E_{21} + E_{22} = \mathcal{O}_{\mathbb{P}}(s'\lambda')$. Together with the result that $E_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$, the claim holds as desired. □

Proof of Corollary 4.8. The claim follows immediately from Theorem 4.5 and Lemma 4.7. □

Proof of Theorem 4.9. Based on our construction of $\widetilde{\alpha}$ in (3.8), we have

$$\begin{aligned} \widetilde{\alpha} &= \widehat{\alpha} - \Big\{\frac{\partial\widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})}{\partial\alpha}\Big\}^{-1}\widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}}) = \widehat{\alpha} - H^{-1}_{\alpha|\boldsymbol{\theta}}\widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}}) + \underbrace{\widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})\Big[H^{-1}_{\alpha|\boldsymbol{\theta}} - \Big\{\frac{\partial\widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})}{\partial\alpha}\Big\}^{-1}\Big]}_{R_1} \\ &= \widehat{\alpha} - H^{-1}_{\alpha|\boldsymbol{\theta}}\Big\{\widehat{U}(0,\widehat{\boldsymbol{\theta}}) + \frac{(\widehat{\alpha}-0)\partial\widehat{U}(\bar{\alpha},\widehat{\boldsymbol{\theta}})}{\partial\alpha}\Big\} + R_1 \qquad (A.9) \\ &= \widehat{\alpha} - H^{-1}_{\alpha|\boldsymbol{\theta}}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) - \widehat{\alpha}H^{-1}_{\alpha|\boldsymbol{\theta}}H_{\alpha|\boldsymbol{\theta}} + \underbrace{\widehat{\alpha}H^{-1}_{\alpha|\boldsymbol{\theta}}\Big\{H_{\alpha|\boldsymbol{\theta}} - \frac{\partial\widehat{U}(\bar{\alpha},\widehat{\boldsymbol{\theta}})}{\partial\alpha}\Big\}}_{R_2} + R_1 = -H^{-1}_{\alpha|\boldsymbol{\theta}}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) + R_1 + R_2, \end{aligned}$$

where (A.9) holds by the mean value theorem for some $\bar{\alpha} = u\widehat{\alpha}$ and $u \in [0,1]$. For both terms R_1 and R_2, we have that they are of the order $\mathcal{O}(n^{-1}s\log d)$ by similar arguments as in the proof of Lemma 4.4. Consequently, it holds that,

$$\sqrt{n}\widetilde{\alpha} \xrightarrow{d} Z, \text{ where } Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}^{-1}),$$

which follows by Theorem 4.5 and our the assumption that $n^{-1/2}s\log d = o(1)$. The claim follows as desired. □

Proof of Corollary 4.10. The claim follows from the argument Theorem 4.9 by replacing 0 by α^*. □

Proof of Corollary 4.12. The claim follows from Theorem 4.11 directly. □

B Limiting Distributions under the Alternative

Statistical power under the alternative hypothesis is one of the most important criteria to compare different tests. Due to the root n convergence of the estimator $\widetilde{\alpha}$ in (3.8) as illustrated in Theorem 4.9, it is of interest to examine the corresponding tests under the alternative where α^* shrinks to the null in a suitable rate.

This subsection investigates the power of the decorrelated score, Wald and partial likelihood ratio tests under a sequence of local alternatives, named as Pitman alternatives. In particular, denote by H_a^n: $\alpha^* = n^{-1/2}c$ the alternative hypothesis, where c is a nonzero constant. Under H_a^n, as n goes to infinity, α^* approaches to 0 as specified in the null hypothesis. We first derive the asymptotic distribution of the decorrelated score function $\widehat{U}(0,\widehat{\theta})$ in (3.4) under Pitman alternatives.

Theorem B.1. Suppose that Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp \sqrt{n^{-1}\log d}$ and $n^{-1/2}s\log d = o(1)$. Under the alternative H_a^n: $\alpha^* = n^{-1/2}c$, the decorrelated score function $\widehat{U}(0,\widehat{\boldsymbol{\theta}})$ in (3.4) satisfies

$$\sqrt{n}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z', \text{ where } Z' \sim N(-cH_{\alpha|\boldsymbol{\theta}}, H_{\alpha|\boldsymbol{\theta}}).$$

Proof of Theorem B.1. Under the alternative hypothesis that $\alpha^* = n^{-1/2}c$, we look at the decorrelated score function $\widehat{U}(0,\widehat{\boldsymbol{\theta}})$. By the same derivation as in (I.1) and mean value theorem, it holds that

$$\widehat{U}(0,\widehat{\boldsymbol{\theta}}) = \underbrace{\nabla_\alpha \mathcal{L}(0,\boldsymbol{\theta}^*) - \mathbf{w}^{*T}\nabla_{\boldsymbol{\theta}}\mathcal{L}(0,\boldsymbol{\theta}^*)}_{S} + \underbrace{(\mathbf{w}^* - \widehat{\mathbf{w}})^T\nabla_{\boldsymbol{\theta}}\mathcal{L}(0,\boldsymbol{\theta}^*)}_{E_1}$$
$$+ \underbrace{\{\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(0,\bar{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\widetilde{\boldsymbol{\theta}})\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_2},$$

where $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u_1(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + u_2(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ for some $0 \le u_1, u_2 \le 1$.

The proof of Theorem 4.5 cannot be directly applied to characterize the asymptotic distribution of the first term S. This is because the vector $(0, \boldsymbol{\theta}^{*T})^T \neq \boldsymbol{\beta}^*$ under the alternative hypothesis, and

thus Lemma I.1 cannot be applied. To derive the asymptotic distribution of S, we have

$$\begin{aligned} S &= \nabla_\alpha \mathcal{L}(0, \boldsymbol{\theta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\theta}} \mathcal{L}(0, \boldsymbol{\theta}^*) \\ &= \underbrace{\nabla_\alpha \mathcal{L}(\alpha^*, \boldsymbol{\theta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\alpha^*, \boldsymbol{\theta}^*)}_{S_1} + \underbrace{\alpha^* \mathbf{w}^{*T} \nabla^2_{\boldsymbol{\theta}\alpha} \mathcal{L}(\bar{\alpha}', \boldsymbol{\theta}^*) - \alpha^* \nabla^2_{\alpha\alpha} \mathcal{L}(\bar{\alpha}, \boldsymbol{\theta}^*)}_{R}, \end{aligned}$$

where the second equality holds by mean value theorem for some $\bar{\alpha} = v_1 \alpha^*, \bar{\alpha}' = v_2 \alpha^*$ and $0 \leq v_1, v_2 \leq 1$.

By Lemma I.1, taking $\mathbf{v} = (1, -\mathbf{w}^{*T})^T$, under the alternative hypothesis, it holds that the first term

$$S_1 \xrightarrow{d} Z, \text{ where } Z \sim N(0, H_{\alpha|\boldsymbol{\theta}}). \tag{B.1}$$

For the second term R, we have

$$R = -\alpha^* (\mathbf{H}^*_{\alpha\alpha} - \mathbf{w}^{*T} \mathbf{H}^*_{\boldsymbol{\theta}\alpha}) + \underbrace{\alpha^* \{\mathbf{H}^*_{\alpha\alpha} - \nabla^2_{\alpha\alpha} \mathcal{L}(\bar{\alpha}, \boldsymbol{\theta}^*)\}}_{R_1} + \underbrace{\alpha^* \mathbf{w}^{*T} \{\nabla^2_{\boldsymbol{\theta}\alpha} \mathcal{L}(\bar{\alpha}', \boldsymbol{\theta}^*) - \mathbf{H}^*_{\boldsymbol{\theta}\alpha}\}}_{R_2}. \tag{B.2}$$

For the term R_1, we have, under the alternative hypothesis $\alpha^* = n^{-1/2} c$,

$$|R_1| = \alpha^* |\mathbf{H}^*_{\alpha\alpha} - \nabla^2_{\alpha\alpha} \mathcal{L}(\bar{\alpha}, \boldsymbol{\theta}^*)| \leq c n^{-1/2} \|\mathbf{H}^* - \nabla^2 \mathcal{L}(\bar{\alpha}, \boldsymbol{\theta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(n^{-1} \sqrt{\log d}), \tag{B.3}$$

where the last equality holds by Lemma I.3.

Next, we look at the term R_2. It holds that

$$|R_2| \leq |\alpha^*| \big\{ \mathbf{w}^{*T} \big(\nabla^2_{\boldsymbol{\theta}\alpha} \mathcal{L}(\bar{\alpha}', \boldsymbol{\theta}^*) - \mathbf{H}^*_{\boldsymbol{\theta}\alpha} \big) \big\} = \mathcal{O}_{\mathbb{P}}(n^{-1} s \sqrt{\log d}), \tag{B.4}$$

where the equality holds by Lemma I.3.

Plugging (B.3) and (B.4) into (B.2), and by the identity $H_{\alpha|\boldsymbol{\theta}} = \mathbf{H}^*_{\alpha\alpha} - \mathbf{w}^{*T} \mathbf{H}^*_{\boldsymbol{\theta}\alpha}$, we have $R = -\alpha^* H_{\alpha|\boldsymbol{\theta}} + \mathcal{O}_{\mathbb{P}}(n^{-1} s \sqrt{\log d})$. Combining this result with (B.1), we have, $S \xrightarrow{d} Z - \alpha^* H_{\alpha|\boldsymbol{\theta}}$.

Meanwhile, by the similar argument as in the proof of Theorem 4.5, it is not difficult to get that the two latter terms $E_1 = o_{\mathbb{P}}(n^{-1/2})$ and $E_2 = o_{\mathbb{P}}(n^{-1/2})$.

To summarize, under the alternative hypothesis that $\alpha^* = c n^{-1/2}$, the decorrelated score function satisfies

$$\sqrt{n} \widehat{U}(0, \widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z', \text{ where } Z' \sim N(-c H_{\alpha|\boldsymbol{\theta}}, H_{\alpha|\boldsymbol{\theta}}),$$

which concludes the proof. □

By Theorem B.1, we have the power of the decorrelated score test under the alternative, H^n_a: $\alpha^* = n^{-1/2} c$, is defined as

$$\mathbb{P}\big(\psi_S(\eta) = 1 \big| \alpha^* = n^{-1/2} c\big) = \mathbb{P}\big(\widehat{S}_n > \chi^2_1(1-\eta) \big| \alpha^* = n^{-1/2} c\big),$$

where $\widehat{S}_n$ is defined in (3.6), and $\chi^2_1(1-\eta)$ is the $(1-\eta)$-th quantile of a chi-squared distribution with one degree of freedom. Denote by $NC_{\chi_1}(\xi)$ the noncentral chi-squared distribution with one degree of freedom and noncentrality parameter ξ. By Theorem B.1, it follows that $\widehat{S}_n \xrightarrow{d} NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}})$. The following corollary of Theorem B.1 provides the power of the decorrelated score test $\psi_S(\eta)$ in (3.7) at a significance level η.

Corollary B.2. Suppose Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp \sqrt{n^{-1}\log d}$ and $n^{-1/2}s\log d = o(1)$. Under Pitman alternative hypothesis H_a^n: $\alpha^* = n^{-1/2}c$, the power of the decorrelated score test $\psi_S(\eta)$ at a significance level η is

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_S(\eta) = 1 | \alpha^* = n^{-1/2}c\big) = \mathbb{P}\big(NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}}) > \chi_1^2(1-\eta)\big).$$

This corollary implies the intuitive fact that the decorrelated score test is asymptotically more powerful when the null and alternative hypotheses become further separated (i.e., $|c|$ increases).

The next theorem provides the limiting distribution of the decorrelated Wald statistic $\widetilde{\alpha}$ in (3.8) and partial likelihood ratio test statistic $\widehat{L}_n$ in (3.12) under Pitman alternative H_a^n: $\alpha^* = n^{-1/2}c$. We also obtain the asymptotic power of these two tests.

Theorem B.3. Assume that Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp \sqrt{n^{-1}\log d}$ and $n^{-1/2}s\log d = o(1)$. Suppose that the alternative H_a^n: $\alpha^* = n^{-1/2}c$ holds. We have:

(a) The one-step estimator $\widetilde{\alpha}$ in (3.8) satisfies

$$\sqrt{n}\widetilde{\alpha} \xrightarrow{d} Z', \text{ where } Z' \sim N(c, H_{\alpha|\boldsymbol{\theta}}^{-1}).$$

The decorrelated Wald test $\psi_W(\eta)$ in (3.10) with a significance level η has asymptotic power

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_W(\eta) = 1 \big| \alpha^* = n^{-1/2}c\big) = \mathbb{P}\big(NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}}) > \chi_1^2(1-\eta)\big).$$

(b) The decorrelated partial likelihood ratio statistic $\widehat{L}_n$ satisfies

$$\widehat{L}_n \xrightarrow{d} Z'_\chi, \text{ where } Z'_\chi \sim NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}}).$$

The power of the decorrelated partial likelihood ratio test at a significance level η satisfies

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_L(\eta) = 1 \big| \alpha^* = n^{-1/2}c\big) = \mathbb{P}\big(NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}}) > \chi_1^2(1-\eta)\big).$$

Proof of Theorem B.3. (a). By the definition of $\widetilde{\alpha}$ in (3.8), we have

$$\widetilde{\alpha} = -H_{\alpha|\boldsymbol{\theta}}^{-1}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) + \underbrace{\widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})\Big[H_{\alpha|\boldsymbol{\theta}}^{-1} - \Big\{\frac{\partial \widehat{U}(\widehat{\alpha},\widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\}^{-1}\Big]}_{R_1} + \underbrace{\widehat{\alpha} H_{\alpha|\boldsymbol{\theta}}^{-1}\Big\{H_{\alpha|\boldsymbol{\theta}} - \frac{\partial \widehat{U}(\bar{\alpha},\widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\}}_{R_2}.$$

By Theorem B.1, we have

$$-\sqrt{n}H_{\alpha|\boldsymbol{\theta}}^{-1}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z, \text{ where } Z \sim N(\alpha^*, H_{\alpha|\boldsymbol{\theta}}^{-1}).$$

In addition, by the similar argument as in Theorem 4.9, we have $R_1 = o_{\mathbb{P}}(n^{-1/2})$ and $R_2 = o_{\mathbb{P}}(n^{-1/2})$. Under the null hypothesis $\alpha^* = n^{-1/2}c$, we have

$$\sqrt{n}\widetilde{\alpha} \xrightarrow{d} Z', \text{ where } Z' \sim N(c, H_{\alpha|\boldsymbol{\theta}}^{-1}),$$

and our claim holds as desired.

(b). By the definition of the test statistic of the decorrelated partial likelihood ratio test (3.11), we have

$$\mathcal{L}(\widetilde{\alpha}, \widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}}) - \mathcal{L}(0, \widehat{\boldsymbol{\theta}}) = \underbrace{\widetilde{\alpha}\widehat{U}(0, \widehat{\boldsymbol{\theta}})}_{T_1} + \underbrace{\frac{\widetilde{\alpha}^2}{2}\Big\{\nabla^2_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \widehat{\boldsymbol{\theta}}) + \widehat{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}})\widehat{\mathbf{w}} - 2\widetilde{\mathbf{w}}^T\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}')\Big\}}_{T_2},$$

where the equality holds by mean value theorem with $\bar{\alpha} = \widehat{\alpha} + u_1(0 - \widehat{\alpha})$, $\bar{\alpha}' = \widehat{\alpha} + u_2(0 - \widehat{\alpha})$ and $\bar{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} + u_3(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}})$ for some $0 \le u_1, u_2, u_3 \le 1$.

By Theorem B.1 and B.3, we have $T_1 = -\big\{\widehat{U}(0, \widehat{\boldsymbol{\theta}})\big\}^2 H^{-1}_{\alpha|\boldsymbol{\theta}} + o_{\mathbb{P}}(n^{-1})$. In addition, by the similar argument as in Theorem 4.11, we have $T_2 = \frac{1}{2}\{\widehat{U}(0, \widehat{\boldsymbol{\theta}})\}^2 H^{-1}_{\alpha|\boldsymbol{\theta}} + o_{\mathbb{P}}(n^{-1})$.

Consequently, the test statistic $\widehat{L}_n$ in (3.11) satisfies

$$2n\big\{\mathcal{L}(0, \widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widetilde{\alpha}, \widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}})\big\} = n\widehat{U}(0, \widehat{\boldsymbol{\theta}})^2 H^{-1}_{\alpha|\boldsymbol{\theta}} + o_{\mathbb{P}}(1) \xrightarrow{d} Z'_\chi + o_{\mathbb{P}}(1),$$

where $Z'_\chi \sim NC_{\chi_1}(c^2 H_{\alpha|\boldsymbol{\theta}})$ by Theorem B.1. Our claim follows as desired. □

In summary, Corollary B.2 and Theorem B.3 imply that the decorrelated score, Wald and partial likelihood ratio tests have the same local asymptotic power. This observation coincides with the conventional asymptotic equivalence among these tests.

C Proofs in Section 5

In this section, we provide detailed proofs in Section 5.

Lemma C.1. Under Assumptions 2.1, 2.2, 4.2, 4.3 and 5.1, $\|\nabla\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) - \nabla\Lambda_0(t, \boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d})$.

Proof. By the definition of $\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$ in (5.1), we have,

$$\|\nabla\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) - \nabla\Lambda_0(t, \boldsymbol{\beta}^*)\|_\infty = \Big\|\frac{1}{n}\int_0^t \frac{S^{(1)}(u, \widehat{\boldsymbol{\beta}})d\overline{N}(u)}{\{S^{(0)}(u, \widehat{\boldsymbol{\beta}})\}^2} - \mathbb{E}\int_0^t \frac{\mathbf{s}^{(1)}(u, \boldsymbol{\beta}^*)dN(u)}{\{\mathbf{s}^{(0)}(u, \boldsymbol{\beta}^*)\}^2}\Big\|_\infty = \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big),$$

where the last inequality follows by the same argument in Lemma I.3. □

A corollary of Lemma C.1 and Lemma 4.4 follows immediately which characterizes the rate of convergence of $\widehat{\mathbf{u}}(t)$.

Corollary C.2. Under Assumptions 2.1, 2.2, 4.2, 4.3 and 5.1, if $\delta \asymp s'\sqrt{n^{-1}\log d}$ we have,

$$\|\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)\|_1 = \mathcal{O}_{\mathbb{P}}\Big(ss'\sqrt{\frac{\log d}{n}}\Big), \quad (\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t))^T\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)(\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)) = \mathcal{O}_{\mathbb{P}}\Big(ss'\frac{\log d}{n}\Big).$$

Now, we are ready to prove Theorem 5.2.

Proof of Theorem 5.2. We first decompose $\sqrt{n}\big\{\Lambda_0(t) - \widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})\big\}$ into two terms that

$$\sqrt{n}\big\{\Lambda_0(t) - \widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})\big\} = \sqrt{n}\underbrace{\big\{\Lambda_0(t) - \widehat{\Lambda}_0(t, \boldsymbol{\beta}^*)\big\}}_{I_1(t)} + \sqrt{n}\underbrace{\big\{\widehat{\Lambda}_0(t, \boldsymbol{\beta}^*) - \widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})\big\}}_{I_2(t)}.$$

Let $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_0(u)du$. For the first term $\sqrt{n}I_1(t)$, we have

$$\sqrt{n}I_1(t) = \int_0^t \frac{\sqrt{n}\sum_{i=1}^n dM_i(u)}{\sum_{i=1}^n Y_i(u)\exp\{\boldsymbol{X}_i^T(u)\boldsymbol{\beta}^*\}}.$$

Since $M_i(t)$ is a martingale, $\sqrt{n}I_1(t)$ becomes a sum of martingale residuals. By Andersen and Gill (1982), we have, as $n \to \infty$, $\sqrt{n}I_1(t) \xrightarrow{d} N\big(0, \sigma_1^2(t)\big)$, where

$$\sigma_1^2(t) = \int_0^t \frac{\lambda_0(u)du}{\mathbb{E}\big[\exp\{\mathbf{X}^T(u)\boldsymbol{\beta}^*\}Y(u)\big]}.$$

For the second term $I_2(t)$, we have, by mean value theorem, for some $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, $\widetilde{\boldsymbol{\beta}}' = \boldsymbol{\beta}^* + t'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ and $0 \le t, t' \le 1$,

$$\begin{aligned}
I_2(t) &= \widehat{\Lambda}_0(t, \boldsymbol{\beta}^*) - \widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) + \{\widehat{\mathbf{u}}(t)\}^T\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T\nabla\widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) + \{\widehat{\mathbf{u}}(t)\}^T\{\nabla\mathcal{L}(\boldsymbol{\beta}^*) + \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\} \\
&= \{\mathbf{u}^*(t)\}^T\nabla\mathcal{L}(\boldsymbol{\beta}^*) + \underbrace{(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T\nabla\widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) + \{\mathbf{u}^*(t)\}^T\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}_{R_1} \\
&\quad + \underbrace{\{\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)\}^T\{\nabla\mathcal{L}(\boldsymbol{\beta}^*) + \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}}_{R_2}.
\end{aligned}$$

Next, we consider the two terms R_1 and R_2. For the term R_1, we have

$$\begin{aligned}
R_1 &= (\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T\nabla\widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) + \{\mathbf{u}^*(t)\}^T\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&= (\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T\big[\mathbf{H}^*\mathbf{H}^{*-1}\nabla\widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) - \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)\big] \\
&= \underbrace{(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T\{\nabla\widehat{\Lambda}_0(t, \widetilde{\boldsymbol{\beta}}) - \nabla\Lambda_0(t, \boldsymbol{\beta}^*)\}}_{R_{11}} + \underbrace{(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T\big[\mathbf{H}^* - \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')\big]\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)}_{R_{12}}.
\end{aligned}$$

It holds that $|R_{11}| \le \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1\|\nabla\Lambda_0(t, \widetilde{\boldsymbol{\beta}}) - \nabla\widehat{\Lambda}_0(t, \boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(s^2n^{-1}\log d)$ by (2.2) and Lemma C.1, and $|R_{12}| \le \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1\|(\mathbf{H}^* - \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}'))\mathbf{u}^*(t)\|_\infty = \mathcal{O}_{\mathbb{P}}(s^2n^{-1}\log d)$. Summing them up, by triangle inequality, we have $|R_1| = \mathcal{O}_{\mathbb{P}}(s^2n^{-1}\log d)$.

For the term R_2, we have

$$\begin{aligned}
|R_2| &\le \|\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + |(\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t))^T\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \\
&\le \|\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t)\|_1\|\nabla\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + |(\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t))^T\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\mathbf{u}}(t) - \mathbf{u}^*(t))|^{1/2} \\
&\quad \times |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}')(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|^{1/2} \\
&= \mathcal{O}_{\mathbb{P}}(s'sn^{-1}\log d) + \mathcal{O}_{\mathbb{P}}(\sqrt{s'}sn^{-1}\log d),
\end{aligned}$$

where the last inequality holds by Lemma I.2 and I.3 and Corollary C.2.

Meanwhile, by Lemma I.1, taking $\mathbf{v} = \mathbf{u}^*(t)$, we have the term $\sqrt{n}\mathbf{u}^{*T}(t)\nabla\mathcal{L}(\boldsymbol{\beta}^*) \xrightarrow{d} N(0, \sigma_2^2(t))$, where $\sigma_2^2(t) = \nabla\Lambda_0(t, \boldsymbol{\beta}^*)^T\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$. Thus, we have,

$$\sqrt{n}I_2(t) \xrightarrow{d} Z, \text{ where } Z \sim N(0, \sigma_2^2(t)),$$

and $\sigma_2^2(t) = \nabla\Lambda_0(t, \boldsymbol{\beta}^*)^T\mathbf{H}^{*-1}\nabla\Lambda_0(t, \boldsymbol{\beta}^*)$.

Following the standard martingale theory, the covariance between $I_1(t)$ and $I_2(t)$ is 0. Our claim holds as desired. □

D Extension to Conditional Hazard Function Inference

In this section, we extend the procedure proposed in Section 5 to conduct conditional hazard function inference given the covariate. For ease of presentation, we assume that the covariates are fixed through time t. Given the i-th sample's covariate $\boldsymbol{X}_i$, the conditional hazard rate function and the cumulative conditional hazard function at time t are

$$\lambda_0(t, \boldsymbol{X}_i) = \lambda_0(t) \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}^*), \text{ and } \Lambda_0(t, \boldsymbol{X}_i) = \int_0^t \lambda_0(u, \boldsymbol{X}_i) du = \int_0^t \lambda_0(u) du \cdot \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}^*).$$

Similar to Section 5, we adopt a Breslow-type estimator for the conditional hazard function. Given the initial penalized estimator $\widehat{\boldsymbol{\beta}}$, we use the direct plug-in estimator for the conditional hazard function at time t as

$$\widehat{\Lambda}_0(t, \boldsymbol{X}_i) = \int_0^t \frac{dN_i(u) \cdot \exp\left(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}\right)}{\sum_{i'=1}^n \exp\left(\boldsymbol{X}_{i'}^T \widehat{\boldsymbol{\beta}}\right) Y_{i'}(u)}.$$

Due to the intractable distribution of $\widehat{\boldsymbol{\beta}}$, we cannot directly conduct inference based on $\widehat{\Lambda}_0(t, \boldsymbol{X}_i)$. Using the decorrelation approach, we propose to estimate the conditional hazard function by the sample version of $\widehat{\Lambda}_0(t, \boldsymbol{X}_i) - \big\{\nabla \Lambda_0(t, \boldsymbol{X}_i)\big\} \mathbf{H}^{*-1} \nabla \mathcal{L}(\widehat{\boldsymbol{\beta}})$, where the gradient $\nabla \Lambda_0(t, \boldsymbol{X}_i)$ is taken with respect to $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^*$. Similar to (5.2), we directly estimate the product $\mathbf{H}^{*-1} \nabla \Lambda_0(t, \boldsymbol{X}_i, \boldsymbol{\beta}^*)$ by the following Dantzig type estimator

$$\widehat{\mathbf{u}}(t) = \operatorname{argmin} \|\mathbf{u}(t)\|_1, \text{ subject to } \|\nabla \widehat{\Lambda}_0(t, \boldsymbol{X}_i) - \nabla^2 \mathcal{L}(\widehat{\boldsymbol{\beta}}) \mathbf{u}(t)\|_\infty \leq \delta, \tag{D.1}$$

where δ is a tuning parameter. By the following assumption, which is analogous to Assumption 5.1 $\widehat{\mathbf{u}}(t)$ converges to $\mathbf{u}^*(t) = \mathbf{H}^{*-1} \nabla \Lambda_0(t, \boldsymbol{X}_i)$ at a fast rate.

Assumption D.1. It holds that $\|\mathbf{u}^*(t)\|_0 = s' \asymp s$ for all $0 \leq t \leq \tau$.

Hence, the decorrelated conditional hazard function estimator at time t is

$$\widetilde{\Lambda}_0(t, \boldsymbol{X}_i) = \widehat{\Lambda}_0(t, \boldsymbol{X}_i) - \widehat{\mathbf{u}}(t)^T \nabla \mathcal{L}(\widehat{\boldsymbol{\beta}}), \quad \text{where } \widehat{\mathbf{u}}(t) \text{ is defined in (D.1).} \tag{D.2}$$

Consequently, the conditional survival function can be estimated by $\widetilde{S}(t, \boldsymbol{X}_i) = \exp\{-\widetilde{\Lambda}_0(t, \boldsymbol{X}_i)\}$. The next theorem characterizes the asymptotic normality of $\widetilde{\Lambda}_0(t, \boldsymbol{X}_i)$ and $\widetilde{S}(t, \boldsymbol{X}_i)$. The proof is analogous to the proof to Theorem 5.2, which we omit here to avoid repetitions.

Theorem D.2. Suppose Assumptions 2.1, 2.2, 4.1, 4.3 and D.1 hold, $\lambda \asymp \sqrt{n^{-1} \log d}$, $\delta \asymp s' \sqrt{n^{-1} \log d}$ and $n^{-1/2} s^2 \log d = o(1)$. We have that for any $t \in [0, \tau]$, the decorrelated conditional hazard function estimator $\widetilde{\Lambda}_0(t, \boldsymbol{X}_i)$ in (D.2) satisfies

$$\sqrt{n}\big\{\Lambda_0(t, \boldsymbol{X}_i) - \widetilde{\Lambda}_0(t, \boldsymbol{X}_i)\big\} \xrightarrow{d} Z, \text{ where } Z \sim N\big(0, \sigma_1^2(t) + \sigma_2^2(t)\big),$$

where

$$\sigma_1^2(t) = \int_0^t \frac{\lambda_0(u, \boldsymbol{X}_i) du \cdot \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)}{\mathbb{E}\big\{\exp(\boldsymbol{X}^T \boldsymbol{\beta}^*) Y(u)\big\}}, \quad \text{and } \sigma_2^2(t) = \nabla \Lambda_0(t, \boldsymbol{X}_i)^T \mathbf{H}^{*-1} \nabla \Lambda_0(t, \boldsymbol{X}_i). \tag{D.3}$$

The estimated survival function $\widetilde{S}(t, \boldsymbol{X}_i)$ satisfies

$$\sqrt{n}\big\{\widetilde{S}(t, \boldsymbol{X}_i) - S_0(t, \boldsymbol{X}_i)\big\} \xrightarrow{d} Z', \text{ where } Z' \sim N\left(0, \frac{\sigma_1^2(t) + \sigma_2^2(t)}{\exp\big\{2\Lambda_0(t, \boldsymbol{X}_i)\big\}}\right).$$

Note that, the limiting variance can be estimated by plug-in estimators that

$$\widehat{\sigma}_1^2(t) = \int_0^t \frac{d\widehat{\Lambda}_0(u, \boldsymbol{X}_i)}{n^{-1}\sum_{i'=1}^n \exp\left(\boldsymbol{X}_{i'}^T\widehat{\boldsymbol{\beta}}\right)Y_{i'}(u)} \quad \text{and} \quad \widehat{\sigma}_2^2(t) = \left\{\nabla\widehat{\Lambda}_0(t, \boldsymbol{X}_i)\right\}^T\widehat{\mathbf{u}}(t).$$

To conclude, based on above Theorem D.2, we can conduct valid inference and construct confidence intervals for the conditional hazard function and survival function.

E Technical Lemmas

In this section, we prove some concentration results of the sample gradient $\nabla\mathcal{L}(\boldsymbol{\beta}^*)$ and sample Hessian matrix $\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$. The mathematical tools we use are mainly from empirical process theory.

We start from introducing the following notations. Let $\|\cdot\|_{\mathbb{P},r}$ denote the $L_r(\mathbb{P})$-norm. For any given $\epsilon > 0$ and the function class $\mathcal{F}$, let $N_{[\,]}\big(\epsilon, \mathcal{F}, L_r(\mathbb{P})\big)$ and $N\big(\epsilon, \mathcal{F}, L_2(\mathbb{Q})\big)$ denote the bracketing number and the covering number, respectively. The quantifies $\log N_{[\,]}\big(\epsilon, \mathcal{F}, L_r(\mathbb{P})\big)$ and $\log N\big(\epsilon, \mathcal{F}, L_2(\mathbb{Q})\big)$ are called entropy with bracketing and entropy. In addition, let F be an envelope of $\mathcal{F}$ where $|f| \leq F$ for all $f \in \mathcal{F}$. The bracketing integral and uniform entropy integral are defined as

$$J_{[\,]}\big(\delta, \mathcal{F}, L_r(\mathbb{P})\big) = \int_0^\delta \sqrt{\log N_{[\,]}\big(\epsilon, \mathcal{F}, L_r(\mathbb{P})\big)}d\epsilon,$$

and

$$J\big(\delta, \mathcal{F}, L_2\big) = \int_0^\delta \sqrt{\log \sup_{\mathbb{Q}} N\big(\epsilon\|F\|_{\mathbb{Q},2}, \mathcal{F}, L_2(\mathbb{Q})\big)}d\epsilon,$$

respectively, where the supremum is taken over all probability measures $\mathbb{Q}$ with $\|F\|_{\mathbb{Q},2} > 0$. Denote the empirical process by $\mathbb{G}_n(f) = n^{1/2}(\mathbb{P}_n - \mathbb{P})(f)$, where $\mathbb{P}_n(f) = n^{-1}\sum_{i=1}^n f(X_i)$ and $\mathbb{P}(f) = \mathbb{E}(f(X_i))$. The following three Lemmas characterize the bounds for the expected maximal empirical processes and the concentration of the maximal empirical processes.

Lemma E.1. Under Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3, there exist some constant $C > 0$, such that, for $r = 0, 1, 2$, with probability at least $1 - \mathcal{O}(d^{-3})$,

$$\sup_{t\in[0,\tau]} \|\mathbf{s}^{(r)}(t, \boldsymbol{\beta}^*) - S^{(r)}(t, \boldsymbol{\beta}^*)\|_\infty \leq C\sqrt{\frac{\log d}{n}},$$

where $\mathbf{s}^{(r)}(t, \boldsymbol{\beta}^*)$ and $S^{(r)}(t, \boldsymbol{\beta}^*)$ are defined in (2.6) and (2.3).

Proof. We will only prove the case for $r = 1$, and the cases for $r = 0$ and 2 follow by the similar argument. For $j = 1, ..., d$, let

$$E_j = \sup_{t\in[0,\tau]} |S_j^{(1)}(t, \boldsymbol{\beta}^*) - s_j^{(1)}(t, \boldsymbol{\beta}^*)|,$$

where $S_j^{(1)}(t, \boldsymbol{\beta}^*)$ and $s_j^{(1)}(t, \boldsymbol{\beta}^*)$ denote the j-th component of $S^{(1)}(t, \boldsymbol{\beta}^*)$ and $s^{(1)}(t, \boldsymbol{\beta}^*)$, respectively. We will prove a concentration result of E_j.

First, we show the class of functions $\{X_j(t)Y(t)\exp\left(\boldsymbol{X}^T(t)\boldsymbol{\beta}^*\right) : t \in [0, \tau]\}$ has bounded uniform entropy integral. By Lemma 9.10 of Kosorok (2007), the class $\mathcal{F} = \{X_j(t) : t \in [0, \tau]\}$ is

a VC-hull class associated with a VC class of index 2. By Corollary 2.6.12 of van der Vaart and Wellner (1996), the entropy of the class $\mathcal{F}$ satisfies $\log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(\mathbb{Q})) \leq C'(1/\epsilon)$ for some constant $C' > 0$, and hence $\mathcal{F}$ has the uniform entropy integral $J(1, \mathcal{F}, L_2) \leq \int_0^1 \sqrt{K(1/\epsilon)}d\epsilon < \infty$. By the same argument, we have that $\big\{\exp\{\boldsymbol{X}(t)^T\boldsymbol{\beta}^*\} : t \in [0,\tau]\big\}$ also has a uniform entropy integral. Meanwhile, by example 19.16 of van der Vaart and Wellner (1996), $\{Y(t) : t \in [0,\tau]\}$ is a VC class and hence has bounded uniform entropy integral. Thus, by Theorem 9.15 of Kosorok (2007), we have $\big\{X_j(t)Y(t)\exp\{\boldsymbol{X}(t)^T\boldsymbol{\beta}^*\} : t \in [0,\tau]\big\}$ has bounded uniform entropy integral.

Next, taking the envelop F as $\sup_{t\in[0,\tau]}|X_j(t)Y(t)\exp\{\boldsymbol{X}^T(t)\boldsymbol{\beta}^*\}|$, by Lemma 19.38 of van der Vaart (2000),

$$\mathbb{E}(E_j) \leq C_1 n^{-1/2} J(1, \mathcal{F}, L_2)\|F\|_{\mathbb{P},2} \leq Cn^{-1/2},$$

for some positive constants C_1 and C. By the Talagrand's inequality (Massart, 2007, Equation (5.50)), we have, for any $\Delta > 0$,

$$\mathbb{P}\big(E_j \geq Cn^{-1/2}(1+\Delta)\big) \leq \mathbb{P}\big(E_j \geq \mathbb{E}(E_j) + n^{-1/2}C\Delta\big) \leq \exp(-C_2\Delta^2 L^{-2}),$$

for some positive constant C_2 and L, and the desired result follows by taking $\Delta = \sqrt{n^{-1}\log d}$ a union bound over $j = 1, ..., d$. □

Lemma E.2. Suppose the Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold, and $\lambda \asymp \sqrt{n^{-1}\log d}$. We have, for $r = 0, 1, 2$ and $t \in [0,\tau]$,

$$\|S^{(r)}(t,\widehat{\boldsymbol{\beta}}) - S^{(r)}(t,\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big).$$

Proof. Similar to the previous Lemma, we only prove the case for $r = 1$, and the other two cases follow by the similar argument. For the case $r = 1$, we have

$$\begin{aligned}
\|S^{(1)}(t,\widehat{\boldsymbol{\beta}}) - S^{(1)}(t,\boldsymbol{\beta}^*)\|_\infty &= \Big\|\frac{1}{n}\sum_{i=1}^n Y_i(t)\big[\exp\{\boldsymbol{X}_i^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*\}\big]\boldsymbol{X}_i(t)\Big\|_\infty \\
&\leq \max_i \big\{Y_i(t)\|\boldsymbol{X}_i(t)\|_\infty\big|\exp\{\boldsymbol{X}_i^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*\}\big|\big\} \\
&\leq C_X \cdot \max_i \big|\exp\{\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*\}\big[\exp\{\boldsymbol{X}_i^T(t)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\} - 1\big]\big| \qquad \text{(E.1)} \\
&\leq C_X \cdot C_1 \cdot \max_i \|\boldsymbol{X}_i(t)\|_\infty \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \qquad \text{(E.2)} \\
&= \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big),
\end{aligned}$$

where (E.1) holds by the Assumption 2.1 for some constant $C_X > 0$; (E.2) holds by Assumption 4.1 that $\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^* = \mathcal{O}(1)$ and $\exp(|x|) \leq 1 + 2|x|$ for any $|x|$ sufficiently small, and the last equality holds by (2.2). Our claim holds as desired. □

Lemma E.3. Under Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3, for any $1 \leq j, k \leq d$, there exists a positive constant C, such that with probability at least $1 - \mathcal{O}(d^{-1})$,

$$\max_{j,k=1,...,d}\big|\nabla^2_{jk}\mathcal{L}(\boldsymbol{\beta}^*) - \mathbf{H}^*_{jk}\big| \leq C\sqrt{\frac{\log d}{n}}. \qquad \text{(E.3)}$$

and

$$\max_{j=1,\ldots,d-1} \big|\big\{\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\mathbf{w}^*\big\}_j - \nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\boldsymbol{\beta}^*)\big\}_j\big| \le C\sqrt{\frac{\log d}{n}}.$$

Proof. We prove the first claim, and the second claim follows by similar arguments. By the definitions of $\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$ and $\mathbf{H}^*$ in (2.5) and (2.7), we have

$$\begin{aligned}
\nabla^2\mathcal{L}(\boldsymbol{\beta}^*) - \mathbf{H}^* = & \underbrace{\frac{1}{n}\int_0^\tau \Big\{\frac{S^{(2)}(t,\boldsymbol{\beta}^*)}{S^{(0)}(t,\boldsymbol{\beta}^*)} - \frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)}\Big\} d\overline{N}(t)}_{T_1} \\
& + \underbrace{\frac{1}{n}\int_0^\tau \frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)} d\overline{N}(t) - \mathbb{E}\Big[\int_0^\tau \frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)} dN(t)\Big]}_{T_2} \\
& + \underbrace{\frac{1}{n}\int_0^\tau \Big\{\mathbf{e}(t,\boldsymbol{\beta}^*)^{\otimes 2} - \overline{\boldsymbol{Z}}(t,\boldsymbol{\beta}^*)^{\otimes 2}\Big\} d\overline{N}(t)}_{T_3} \\
& + \underbrace{\mathbb{E}\Big[\int_0^\tau \mathbf{e}(t,\boldsymbol{\beta}^*)^{\otimes 2} dN(t)\Big] - \frac{1}{n}\int_0^\tau \mathbf{e}(t,\boldsymbol{\beta}^*)^{\otimes 2} d\overline{N}(t)}_{T_4}.
\end{aligned}$$

For the term T_1, we have, with probability at least $1 - \mathcal{O}(d^{-1})$,

$$\|T_1\|_\infty \le \sup_{t\in[0,\tau]} \Big\|\frac{S^{(2)}(t,\boldsymbol{\beta}^*)}{S^{(0)}(t,\boldsymbol{\beta}^*)} - \frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)}\Big\|_\infty \cdot \frac{1}{n}\int_0^\tau d\overline{N}(t) \le C_1\sqrt{\frac{\log d}{n}},$$

where the last inequality follows by Lemma E.1. Next, by Assumption 2.1, we have

$$\Big\|\frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}^*)}{\mathbf{s}^{(0)}(t,\boldsymbol{\beta}^*)}\Big\|_\infty < \infty.$$

Consequently, T_2 becomes an i.i.d. sum of mean 0 bounded random variables. Hoeffding's inequality gives that with probability at least $1 - \mathcal{O}(d^{-1})$, $\|T_2\|_\infty \le C_2\sqrt{n^{-1}\log d}$. Meanwhile, the terms T_3 and T_4 can be bounded similarly. Our claim holds as desired. □

Lemma E.4. Under Assumptions 2.1, 2.2 4.1, 4.2 and 4.3, it holds that

$$\|\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^{*T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty = \mathcal{O}_{\mathbb{P}}\Big(s\sqrt{\frac{\log d}{n}}\Big),$$

and

$$\|\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \widehat{\mathbf{w}}^{T}\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty = \mathcal{O}_{\mathbb{P}}\Big((s+s')\sqrt{\frac{\log d}{n}}\Big).$$

Proof. We prove the first claim, and the second claim follows by similar arguments. By triangle inequality, we have

$$\begin{aligned}
& \|\nabla^2_{\alpha\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^{*T}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\|_\infty \\
\le & \underbrace{\|\mathbf{H}^*_{\alpha\boldsymbol{\theta}} - \mathbf{w}^{*T}\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\|_\infty}_{E_1} + + \underbrace{\|\nabla^2_{\boldsymbol{\theta}\alpha}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \mathbf{H}^*_{\boldsymbol{\theta}\alpha}\|_\infty}_{E_2} + \underbrace{\|\mathbf{w}^{*T}\big\{\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}} - \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\big\}\|_\infty}_{E_3}.
\end{aligned}$$

It is seen that $E_1 = 0$ by the definition of $\mathbf{w}^* = \mathbf{H}^{*-1}_{\boldsymbol{\theta\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$ in (3.2). In addition, $E_2 = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d})$ by Lemma I.3. For the term E_3, we have

$$E_3 \le \underbrace{\|\mathbf{w}^{*T}\{\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\}\|_\infty}_{E_{31}} + \underbrace{\|\mathbf{w}^{*T}\{\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(\boldsymbol{\beta}^*) - \mathbf{H}^*_{\boldsymbol{\theta\theta}}\}\|_\infty}_{E_{32}}.$$

For the term E_{31}, by the definition of $\nabla^2\mathcal{L}(\cdot)$ in (2.5), we have

$$\begin{aligned}\mathbf{w}^{*T}\{\nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}}) - \nabla^2_{\boldsymbol{\theta\theta}}\mathcal{L}(\boldsymbol{\beta}^*)\} =& \underbrace{\mathbf{w}^{*T}\Big\{\frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{S^{(2)}(t,\widehat{\boldsymbol{\beta}})}{S^{(0)}(t,\widehat{\boldsymbol{\beta}})} - \frac{S^{(2)}(t,\boldsymbol{\beta}^*)}{S^{(0)}(t,\boldsymbol{\beta}^*)}dN_i(t)\Big\}_{\boldsymbol{\theta\theta}}}_{T_1}\\ &+ \underbrace{\mathbf{w}^{*T}\Big\{\frac{1}{n}\sum_{i=1}^n\int_0^\tau \overline{\boldsymbol{Z}}(t,\widehat{\boldsymbol{\beta}})^{\otimes 2} - \overline{\boldsymbol{Z}}(t,\boldsymbol{\beta}^*)^{\otimes 2}\Big\}_{\boldsymbol{\theta\theta}}}_{T_2}.\end{aligned}$$

For the term T_1, we have

$$T_1 = \frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{S^{(0)}(t,\boldsymbol{\beta}^*)\mathbf{w}^{*T}S^{(2)}_{\boldsymbol{\theta\theta}}(t,\widehat{\boldsymbol{\beta}}) - S^{(0)}(t,\widehat{\boldsymbol{\beta}})\mathbf{w}^{*T}S^{(2)}_{\boldsymbol{\theta\theta}}(t,\boldsymbol{\beta}^*)}{S^{(0)}(t,\widehat{\boldsymbol{\beta}})S^{(0)}(t,\boldsymbol{\beta}^*)}$$

For ease of notation, in the rest of the proof, let $\widehat{S}^{(r)}(t) := S^{(r)}(t,\widehat{\boldsymbol{\beta}})$ and $S^{*(r)}(t) := S^{(r)}(t,\boldsymbol{\beta}^*)$ for $r = 0,1,2$. We have, for the k-th component of T_1,

$$\begin{aligned}T_{1,k} =& \frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{S^{*(0)}(t)\frac{1}{n}\sum_{i'=1}^n y_{i'}(t)\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\}\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}dN_i(t)\\ &- \frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{\widehat{S}^{(0)}(t)\sum_{i'=1}^n y_{i'}(t)\exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}dN_i(t).\end{aligned}$$

Consequently, it holds that

$$\begin{aligned}&|T_{1,k}|\\ \le& \Big|\frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{\{S^{*(0)}(t) - \widehat{S}^{(0)}(t)\}\frac{1}{n}\sum_{i'=1}^n Y_{i'}(t)\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\}\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}dN_i(t)\Big|\\ &+ \Big|\frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{\widehat{S}^{(0)}(t)\frac{1}{n}\sum_{i'=1}^n Y_{i'}(t)\big[\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\big]\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}dN_i(t)\Big|\\ \le& \sup_{t\in[0,\tau]}\Big|\frac{1}{n}\sum_{i=1}^n\frac{\{S^{*(0)}(t) - \widehat{S}^{(0)}(t)\}\big[\frac{1}{n}\sum_{i'=1}^n Y_{i'}(t)\exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)\big]}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}\Big|\cdot\tau\\ &+ \Big|\frac{1}{n}\sum_{i=1}^n\frac{\{S^{*(0)}(t) - \widehat{S}^{(0)}(t)\}\Big[\frac{1}{n}\sum_{i'=1}^n Y_{i'}(t)\big[\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\big]\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)\Big]}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}\Big|\cdot\tau\\ &+ \frac{1}{n}\sum_{i=1}^n\frac{\widehat{S}^{(0)}(t)\frac{1}{n}\sum_{i'=1}^n Y_{i'}(t)\big[\exp\{\boldsymbol{X}_{i'}^T(t)\widehat{\boldsymbol{\beta}}\} - \exp\{\boldsymbol{X}_{i'}^T(t)\boldsymbol{\beta}^*\}\big]\mathbf{w}^{*T}\boldsymbol{X}_{i',\boldsymbol{\theta}}(t)X_{i',k}(t)}{\widehat{S}^{(0)}(t)S^{*(0)}(t)}\cdot\tau\\ =& \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d}),\end{aligned}$$

where the last equality holds by Assumptions 2.1 and 4.1 that $\boldsymbol{X}_i^T(t)\boldsymbol{\beta}^*$ is bounded, $S^{*(0)}(t)$ is bounded away from 0, and by Lemma E.2 that $|\widehat{S}^{(r)}(t) - S^{*(r)}(t)| = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d})$.

The term T_2 can be bounded by the similar argument, and our claim holds as desired. □

Lemma E.5. Under Assumptions 2.1 and 2.2, and if $n^{-1/2}s^3\log d = o(1)$, the RE condition holds for the sample Hessian matrix $\nabla^2\mathcal{L}(\widehat{\boldsymbol{\beta}})$. Specifically, for the vectors in the cone $\mathcal{C} = \{\mathbf{v}|\|\mathbf{v}_{\mathcal{S}}\|_1 \leq \xi\|\mathbf{v}_{\mathcal{S}^C}\|_1\}$, we have

$$\frac{\mathbf{v}^T\nabla^2\mathcal{L}(\widehat{\boldsymbol{\beta}})\mathbf{v}}{\|\mathbf{v}\|_2^2} \geq \frac{1}{2}\kappa^2\big(\xi, |\mathcal{S}|; \nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\big), \text{ for all } \mathbf{v} \in \mathcal{C}.$$

Proof. By Lemma 3.2 of Huang et al. (2013), we have $\exp(-2\xi_{\mathbf{b}})\nabla^2\mathcal{L}(\boldsymbol{\beta}) \preceq \nabla^2\mathcal{L}(\boldsymbol{\beta}+\mathbf{b})$, where $\xi_{\mathbf{b}} = \max_{u\geq 0}\max_{i,i',k,k'}|\mathbf{b}^T\{\boldsymbol{X}_{ik}(u) - \boldsymbol{X}_{i'k'}(u)\}|$. Let $\mathbf{b} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. By Assumption 2.1 that $\|\{\boldsymbol{X}_{ik}(u) - \boldsymbol{X}_{i'k'}(u)\}\|_\infty \leq C_X$, we have $\xi_{\mathbf{b}} = \mathcal{O}_{\mathbb{P}}(s\sqrt{n^{-1}\log d})$ by (2.2), we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s\lambda)$. By the scaling assumption that $n^{-1/2}s^3\log d = o(1)$, we have $\xi_{\mathbf{b}} \leq \frac{1}{2}\log 2$. Consequently, $\exp(-2\xi_{\mathbf{b}}) \geq 1/2$. We have $\nabla^2\mathcal{L}(\widehat{\boldsymbol{\beta}}) \succeq \frac{1}{2}\cdot\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)$. Since the cone $\mathcal{C}$ is a subset of $\mathbb{R}^d$, our claim follows as desired. □

F Proof of Lemmas in Appendix I

Proof of Lemma I.2. By definition, we have, for all $j = 1, ..., d$,

$$\begin{aligned}\nabla_j\mathcal{L}(\boldsymbol{\beta}^*) &= -\frac{1}{n}\sum_{i=1}^n\int_0^\tau \big\{X_{ij}(u,\boldsymbol{\beta}^*) - \overline{Z}_j(u,\boldsymbol{\beta}^*)\big\}dM_i(u)\\ &= \frac{1}{n}\sum_{i=1}^n\int_0^\tau \overline{Z}_j(u,\boldsymbol{\beta}^*)dM_i(u) - \frac{1}{n}\sum_{i=1}^n\int_0^\tau X_{ij}(u,\boldsymbol{\beta}^*)dM_i(u). \end{aligned} \tag{F.1}$$

For the first term, we have for all $t \in [0,\tau]$,

$$\overline{Z}_j(t,\boldsymbol{\beta}^*) - e_j(t,\boldsymbol{\beta}^*) = \frac{S_j^{(1)}(t,\boldsymbol{\beta}^*) - s_j^{(1)}(t,\boldsymbol{\beta}^*)}{S^{(0)}(t,\boldsymbol{\beta}^*)} - \frac{s_j^{(1)}(t,\boldsymbol{\beta}^*)\big\{S^{(0)}(t,\boldsymbol{\beta}^*) - s^{(0)}(t,\boldsymbol{\beta}^*)\big\}}{S^{(0)}(t,\boldsymbol{\beta}^*)s^{(0)}(t,\boldsymbol{\beta}^*)}. \tag{F.2}$$

By Assumption 2.1 and the fact that $\mathbb{P}(y(\tau) > 0) > 0$, we have that $\sup_{t\in[0,\tau]}|\overline{Z}_j(t,\boldsymbol{\beta}^*) - e_j(t)| \leq C_1$ for some constant $C_1 > 0$. In addition,

$$\frac{1}{n}\sum_{i=1}^n\int_0^\tau \overline{Z}_j(u,\boldsymbol{\beta}^*)dM_i(u) \leq \sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^n\int_0^\tau f(u)dM_i(u),$$

where $\mathcal{F}_j$ denotes the class of functions $f : [0,\tau] \to \mathbb{R}$ which have uniformly bounded variation and satisfy $\sup_{t\in[0,\tau]}|f(t) - e_j(t)| \leq \delta_1$ for some δ_1. By constructing ℓ_∞ balls centered at piecewise constant functions on a regular grid, one can show that the covering number of the class $\mathcal{F}_j$ satisfies $N(\epsilon, \mathcal{F}_j, \ell_\infty) \leq (C_2\epsilon^{-1})^{C_3\epsilon^{-1}}$ for some positive constants C_2, C_3. Let $\mathcal{G}_j = \{\int_0^\infty f(t)dM(t) : f \in \mathcal{F}_j\}$. Note that for any two f_1, $f_2 \in \mathcal{F}_j$,

$$\Big|\int_0^\tau f_1(t) - f_2(t)dM(t)\Big| \leq \sup_{u\in[0,\tau]}|f_1(u) - f_2(u)|\int_0^\tau |dM(t)|.$$

By Theorem 2.7.11 of van der Vaart and Wellner (1996), the bracketing number of the class $\mathcal{G}_j$ satisfies $N_{[\,]}\big(2\epsilon\|F\|_{\mathbb{P},2}, \mathcal{G}_j, \ell_2(\mathbb{P})\big) \leq N(\epsilon, \mathcal{F}_j, \|\cdot\|_\infty) \leq (C_2\epsilon^{-1})^{C_3\epsilon^{-1}}$, where $F = \int_0^\tau |dM(t)|$. Hence, $\mathcal{G}_j$ has bounded bracketing integral. An application of Corollary 19.35 of van der Vaart (2000) yields that

$$\mathbb{E}\bigg(\sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^n\int_0^\tau f(u)dM_i(u)\bigg) \leq n^{-1/2}C_4$$

for some constant $C_4 > 0$. Then, by Talagrand's inequality, for any $c > 0$

$$\mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^n\int_0^\tau \overline{Z}_j(u,\boldsymbol{\beta}^*)dM_i(u) > c\bigg) \leq \mathbb{P}\bigg(\sup_{f\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^n\int_0^\tau f(u)dM_i(u) > c\bigg) \leq \exp\bigg(-\frac{nc^2}{C_5}\bigg),$$

for some constant C_5. Following by the union bound, we have with probability at least $1-\mathcal{O}(d^{-3})$,

$$\bigg\|\frac{1}{n}\sum_{i=1}^n\int_0^\tau \overline{Z}_j(u,\boldsymbol{\beta}^*)dM_i(u)\bigg\|_\infty \leq C\sqrt{\frac{\log d}{n}}.$$

Note that the second term of (F.1) is a sum of i.i.d. mean-zero bounded random variables. Following by the Hoeffding inequality and the union bound, we have with probability at least $1-\mathcal{O}(d^{-3})$,

$$\bigg\|\frac{1}{n}\sum_{i=1}^n\int_0^\infty X_{ij}(u,\boldsymbol{\beta}^*)dM_i(u)\bigg\|_\infty \leq C\sqrt{\frac{\log d}{n}},$$

for some constant C. The claim follows as desired. □

Proof of Lemma I.3. Let $\xi = \max_{u\geq 0}\max_{i,i'}|\boldsymbol{\Delta}^T\{\boldsymbol{X}_i(u)-\boldsymbol{X}_{i'}(u)\}|$, where $\boldsymbol{\Delta} = \widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*$. By Lemma 3.2 of Huang et al. (2013), it holds that,

$$\exp(-2\xi)\nabla^2\mathcal{L}(\boldsymbol{\beta}^*) \preceq \nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}}) \preceq \exp(2\xi)\nabla^2\mathcal{L}(\boldsymbol{\beta}^*), \tag{F.3}$$

where $\mathbf{A} \preceq \mathbf{B}$ means that the matrix $\mathbf{B}-\mathbf{A}$ is a positive semidefinite matrix.

Note that the diagonal elements of a positive semidefinite matrix can only be nonnegative. In addition, for a positive semidefinite matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$, it is easy to see that $\|\mathbf{A}\|_\infty = \max_j\{a_{jj}\}_{j=1}^d$. We have,

$$\exp(-2\xi)\|\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq \|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})\|_\infty \leq \exp(2\xi)\|\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty.$$

By (2.2) that $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 = \mathcal{O}_\mathbb{P}(s\lambda)$, which implies that $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1 = \mathcal{O}(s\lambda)$ as $\widetilde{\boldsymbol{\beta}}$ is on the line segment connecting $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$. Hence, $\xi = \mathcal{O}_\mathbb{P}(s\lambda)$. By triangle inequality,

$$\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})-\mathbf{H}^*\|_\infty \leq \underbrace{\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})-\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty}_{E_1} + \underbrace{\|\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)-\mathbf{H}^*\|_\infty}_{E_2}.$$

We consider the two terms separately, for the first term E_1, we have, by (F.3) and taking the Taylor's expansion of $\exp(2\xi)$,

$$\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})-\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty \leq 2\,\|\xi\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty + o_\mathbb{P}(\xi).$$

Since $\xi = \mathcal{O}_\mathbb{P}(s\lambda)$, and by Assumption 4.3, we have,

$$\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})-\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}_\mathbb{P}(s\lambda),$$

and $E_1 = \mathcal{O}_\mathbb{P}(s\sqrt{n^{-1}\log d})$ as $\lambda \asymp \sqrt{n^{-1}\log d}$. In addition, $E_2 = \mathcal{O}_\mathbb{P}(\sqrt{n^{-1}\log d})$ by Lemma E.3. It further implies that $\|\nabla^2\mathcal{L}(\widetilde{\boldsymbol{\beta}})\|_\infty = \mathcal{O}_\mathbb{P}(1)$. The proof of the last result is similar and is omitted. □

G Extension to Multi-dimensional Parameter

In this section, we extend our procedures and asymptotic results to the case when the parameter of interest is of dimension $d_0 > 1$, i.e., $\boldsymbol{\alpha} = (\beta_1, ..., \beta_{d_0})^T \in \mathbb{R}^{d_0}$. We assume that d_0 is a fixed constant and does not increase with s, n or d. The null hypothesis of interest is H_0: $\boldsymbol{\alpha}^* = \mathbf{0}$.

We first consider the decorrelated score test. Similar to the univariate case, we first estimate the nuisance parameter $\boldsymbol{\theta}^*$ by $\widehat{\boldsymbol{\theta}}$ using the Lasso estimator. Then, we define a matrix

$$\mathbf{w}^* = \mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*)^T\}^{-1}\mathbb{E}\big\{\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*)\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*)^T\big\} = \mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\alpha}} \in \mathbb{R}^{(d-d_0)\times d_0}.$$

Similar to (3.3), we estimate $\mathbf{w}^*$ by $\widehat{\mathbf{w}} = (\widehat{\mathbf{w}}_1, ..., \widehat{\mathbf{w}}_{d_0})$, where

$$\widehat{\mathbf{w}}_j = \underset{\mathbf{w}\in\mathbb{R}^{d-d_0}}{\operatorname{argmin}} \Big\{\frac{1}{2}\mathbf{w}^T\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(\widehat{\boldsymbol{\beta}})\mathbf{w} - \mathbf{w}^T\nabla^2_{\boldsymbol{\theta}\alpha_j}\mathcal{L}(\widehat{\boldsymbol{\beta}}) + \lambda'\|\mathbf{w}\|_1\Big\}, \tag{G.1}$$

where λ' is a tuning parameter. To guarantee that $\widehat{\mathbf{w}}$ converges to $\mathbf{w}^*$ at a fast rate of convergence, we impose the following sparsity assumption on $\mathbf{w}^*$ to replace Assumption 4.2.

Assumption G.1. It holds that $\|\mathbf{w}^*\|_0 = s' \asymp s$, and $\sup_{t\in[0,\tau]}\max_{i\in[n]}\|\boldsymbol{X}^T_{i,(d_0+1):d}(t)\mathbf{w}^*\|_\infty = \mathcal{O}(1)$.

We then define a d_0 dimensional decorrelated score function for $\boldsymbol{\alpha}$ as

$$\widehat{U}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}}). \tag{G.2}$$

The next theorem characterizes the asymptotic distribution of $\widehat{U}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}})$. The proof is similar to the proof of Theorem 4.5. We provide the proof for completeness and omit proofs for the decorrelated Wald and partial likelihood ratio tests to avoid repetitions.

Theorem G.2. Suppose that Assumptions 2.1, 2.2, 4.1, 4.3 and G.1 hold, $\lambda \asymp \sqrt{n^{-1}\log d}$, $\lambda' \asymp \sqrt{n^{-1}\log d}$ and $n^{-1/2}s\log d = o(1)$. Under the null hypothesis that $\boldsymbol{\alpha}^* = \mathbf{0}$, the decorrelated score function $\widehat{U}(\mathbf{0}, \widehat{\boldsymbol{\theta}})$ in (G.2) satisfies

$$\sqrt{n}\widehat{U}(\mathbf{0}, \widehat{\boldsymbol{\theta}}) \xrightarrow{d} \boldsymbol{Z}, \text{ where } \boldsymbol{Z} \sim N(\mathbf{0}, \mathbf{H}_{\boldsymbol{\alpha}|\boldsymbol{\theta}}),$$

and $\mathbf{H}_{\boldsymbol{\alpha}|\boldsymbol{\theta}} = \mathbf{H}^*_{\boldsymbol{\alpha}\boldsymbol{\alpha}} - \mathbf{H}^*_{\boldsymbol{\alpha}\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\alpha}} \in \mathbb{R}^{d_0\times d_0}$.

Proof. We first note that by similar arguments, we have the following multi-dimensional version of Lemma I.1, where we omit the details to avoid repetition.

Lemma G.3. Under Assumptions 2.1, 4.2 and 4.3, for any $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_{d_0}) \in \mathbb{R}^{d\times d_0}$, if $\|\mathbf{v}_j\|_0 \le s'$ for all $j \in [d_0]$ and $n^{-1/2}\sqrt{s'\log d} = o(1)$, it holds that

$$\big(\mathbf{v}^T\mathbf{H}^*\mathbf{v}\big)^{-1/2}\sqrt{n}\mathbf{v}^T\nabla\mathcal{L}(\boldsymbol{\beta}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_{d_0}), \quad \text{where } \mathbf{H}^* \text{ is defined in (2.7)}.$$

Let $\mathbf{w}^* = (\mathbf{w}_1^*, \mathbf{w}_2^*, ..., \mathbf{w}_{d_0}^*)$, $\widehat{\mathbf{w}} = (\widehat{\mathbf{w}}_1, \widehat{\mathbf{w}}_2^*, ..., \widehat{\mathbf{w}}_{d_0})$, where $\mathbf{w}_j^*, \widehat{\mathbf{w}}_j \in \mathbb{R}^{d-d_0}$ for $j \in [d_0]$, and let $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{d_0})^T \in \mathbb{R}^{d_0}$. By decomposing $\widehat{U}(\mathbf{0}, \widehat{\boldsymbol{\theta}})$ in (G.2), we have

$$
\begin{aligned}
&\widehat{U}(\mathbf{0}, \widehat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\mathbf{0}, \widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \widehat{\boldsymbol{\theta}}) \\
=&\nabla_{\boldsymbol{\alpha}} \mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*) - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*) + \begin{pmatrix} \nabla^2_{\alpha_1 \boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \bar{\boldsymbol{\theta}}_1)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ \vdots \\ \nabla^2_{\alpha_{d_0} \boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \bar{\boldsymbol{\theta}}_{d_0})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \end{pmatrix} - \begin{pmatrix} \widehat{\mathbf{w}}_1^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \widetilde{\boldsymbol{\theta}}_1)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ \vdots \\ \widehat{\mathbf{w}}_{d_0}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \widetilde{\boldsymbol{\theta}}_{d_0})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \end{pmatrix} \\
=&\underbrace{\nabla_{\alpha} \mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*) - \mathbf{w}^{*T} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*)}_{S} + \underbrace{(\mathbf{w}^* - \widehat{\mathbf{w}})^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*)}_{E_1} - \underbrace{\begin{pmatrix} \{\nabla^2_{\alpha_1 \boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \bar{\boldsymbol{\theta}}_1) - \widehat{\mathbf{w}}_1^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \widetilde{\boldsymbol{\theta}}_1)\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ \vdots \\ \{\nabla^2_{\alpha_{d_0} \boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \bar{\boldsymbol{\theta}}_1) - \widehat{\mathbf{w}}_{d_0}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}0, \widetilde{\boldsymbol{\theta}}_{d_0})\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \end{pmatrix}}_{E_2},
\end{aligned}
$$

where the second equality holds by applying mean-value theorem componentwisely, where each $\bar{\boldsymbol{\theta}}_j = \boldsymbol{\theta}^* + u_j(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, $\widetilde{\boldsymbol{\theta}}_j = \boldsymbol{\theta}^* + u_j'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $u_j, u_j' \in [0, 1]$ for all $j \in [d_0]$.

We consider the terms S, E_1 and E_2 separately. For the first term S, by Lemma G.3, we have

$$\sqrt{n} S \xrightarrow{d} \boldsymbol{Z}, \text{ where } \boldsymbol{Z} \sim N(\mathbf{0}, \mathbf{H}_{\alpha|\boldsymbol{\theta}}).$$

For the term E_1, we have

$$\|E_1\|_1 \leq \|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 \|\nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}(s' \lambda' \sqrt{n^{-1} \log d}),$$

where $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_1 = \mathcal{O}_{\mathbb{P}}(s'\lambda')$ holds by Lemma 4.4 that each $\|\widehat{\mathbf{w}}_j - \mathbf{w}_j^*\|_1 = \mathcal{O}_{\mathbb{P}}(s'\lambda')$ for all $j \in [d_0]$, and $\|\nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \boldsymbol{\theta}^*)\|_\infty = \mathcal{O}_{\mathbb{P}}\big(\sqrt{n^{-1} \log d}\big)$ by Lemma I.2.

For the term E_2, we have

$$\|E_2\|_1 = \sum_{j=1}^{d_0} \big|\{\nabla^2_{\alpha_j \boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \bar{\boldsymbol{\theta}}_j) - \widehat{\mathbf{w}}_1^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \widetilde{\boldsymbol{\theta}}_j)\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\big|. \tag{G.3}$$

For each j, we have that $\big|\{\nabla^2_{\alpha_j \boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \bar{\boldsymbol{\theta}}_j) - \widehat{\mathbf{w}}_1^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \mathcal{L}(\mathbf{0}, \widetilde{\boldsymbol{\theta}}_j)\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\big| = \mathcal{O}_{\mathbb{P}}(n^{-1} s \log d)$ by the same arguments in Theorem 4.5. Thus, we have that $\|E_2\|_1 = \mathcal{O}_{\mathbb{P}}(n^{-1} s \log d)$ as d_0 is a constant.

Combining the results above, our claim follows as desired. □

To standardize $\widehat{U}(\mathbf{0}, \widehat{\boldsymbol{\theta}})$, we estimate $\mathbf{H}_{\boldsymbol{\alpha}|\boldsymbol{\theta}}$ by

$$\widehat{\mathbf{H}}_{\boldsymbol{\alpha}|\boldsymbol{\theta}} = \nabla^2_{\boldsymbol{\alpha}\boldsymbol{\alpha}} \mathcal{L}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla^2_{\boldsymbol{\theta}\boldsymbol{\alpha}} \mathcal{L}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}). \tag{G.4}$$

Hence, the decorrelated score test statistic for multi-dimensional $\boldsymbol{\alpha}$ is defined as

$$\widehat{S}_n = n \widehat{U}^T(\mathbf{0}, \widehat{\boldsymbol{\theta}}) \widehat{\mathbf{H}}_{\boldsymbol{\alpha}|\boldsymbol{\theta}} \widehat{U}(\mathbf{0}, \widehat{\boldsymbol{\theta}}), \text{ where } \widehat{U}(\mathbf{0}, \widehat{\boldsymbol{\theta}}) \text{ and } \widehat{\mathbf{H}}_{\boldsymbol{\alpha}|\boldsymbol{\theta}} \text{ are defined in (G.2) and (G.4)},$$

and its associated test at significance level η is

$$\psi_S(\eta) = \begin{cases} 0 & \text{if } \widehat{S}_n \leq \chi^2_{d_0}(1 - \eta), \\ 1 & \text{otherwise,} \end{cases}$$

where $\chi^2_{d_0}(1-\eta)$ denotes the $(1-\eta)$-th quantile of the chi-squared distribution with d_0 degrees of freedom, and the null hypothesis $\boldsymbol{\alpha}^* = \mathbf{0}$ is rejected if and only if $\psi_S(\eta) = 1$. By Theorem G.2, it follows immediately that the test $\psi_S(\eta)$ satisfies

$$\lim_{n\to\infty} \mathbb{P}\big(\psi_S(\eta) = 1 | \boldsymbol{\alpha}^* = \mathbf{0}\big) = \eta.$$

Next, we extend the Wald test to the multi-dimensional case. Let $\left\{\frac{\partial \widehat{U}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\alpha}}\right\}^{-1} = \widehat{\boldsymbol{\Gamma}} = [\widehat{\boldsymbol{\Gamma}}_1, ..., \widehat{\boldsymbol{\Gamma}}_{d_0}] \in \mathbb{R}^{d_0 \times d_0}$, where each $\widehat{\boldsymbol{\Gamma}}_j \in \mathbb{R}^{d_0}$. Similar to (3.8), we construct an one-step estimator $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\alpha}_1, ..., \widetilde{\alpha}_{d_0}) \in \mathbb{R}^{d_0}$, where for each $j \in [d_0]$,

$$\widetilde{\alpha}_j = \widehat{\alpha}_j - \widehat{\boldsymbol{\Gamma}}_j^T \widehat{U}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}), \text{ and } \widehat{U}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}) = \nabla_\alpha \mathcal{L}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T \nabla_{\boldsymbol{\theta}} \mathcal{L}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}). \tag{G.5}$$

By similar arguments in Theorem 4.9, we have that $\sqrt{n}(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)$ converges weakly to $N(\mathbf{0}, \mathbf{H}^{-1}_{\boldsymbol{\alpha}|\boldsymbol{\theta}})$. Thus, the confidence region for $\boldsymbol{\alpha}$ can be obtained. Meanwhile, the deccorelated Wald test statistic and the associated test for H_0: $\boldsymbol{\alpha}^* = \mathbf{0}$ are

$$\widehat{W}_n = n\widetilde{\boldsymbol{\alpha}}^T \widehat{\mathbf{H}}_{\boldsymbol{\alpha}|\boldsymbol{\theta}} \widetilde{\boldsymbol{\alpha}}, \quad \text{and} \quad \psi_W(\eta) = \begin{cases} 0 & \text{if } \widehat{W}_n \le \chi^2_{d_0}(1-\eta), \\ 1 & \text{otherwise.} \end{cases}$$

In addition, based on the asymptotic distribution of $\widetilde{\boldsymbol{\alpha}}$, we can conduct inference on a linear combination of $\boldsymbol{\alpha}^*$. Specifically, consider the null hypothesis H_0: $\mathbf{v}^T\boldsymbol{\alpha}^* = 0$, for some $\mathbf{v} \in \mathbb{R}^{d_0}$. The decorrelated test statistic and the associated test are

$$\widehat{W}^L_n = n\big(\mathbf{v}^T \widehat{\mathbf{H}}_{\boldsymbol{\alpha}|\boldsymbol{\theta}} \mathbf{v}\big)^{-1} (\mathbf{v}^T \widetilde{\boldsymbol{\alpha}})^2, \quad \text{and} \quad \psi^L_W(\eta) = \begin{cases} 0 & \text{if } \widehat{W}^L_n \le \chi^2(1-\eta), \\ 1 & \text{otherwise.} \end{cases}$$

Finally, we extend the partial likelihood ratio test to the multi-dimensional case. Similar to Section 3.3, define the (negative) decorrelated partial likelihood for $\boldsymbol{\alpha}$ as $\mathcal{L}_{\text{decor}}(\boldsymbol{\alpha}) = \mathcal{L}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\theta}} - \widehat{\mathbf{w}}\boldsymbol{\alpha})$. Consequently, the decorrelated partial likelihood test statistic and the test are defined as

$$\widehat{L}_n = 2n\big\{\mathcal{L}_{\text{decor}}(\mathbf{0}) - \mathcal{L}_{\text{decor}}(\widetilde{\boldsymbol{\alpha}})\big\}, \quad \text{and} \quad \psi_L(\eta) = \begin{cases} 0 & \text{if } \widehat{L}_n \le \chi^2_{d_0}(1-\eta), \\ 1 & \text{otherwise.} \end{cases}$$

H Extensions to Multivariate Failure Time Data

In real applications, it is also of interest to study multivariate failure time outcomes. For example, Cai et al. (2005) consider the time to coronary heart disease and time to cerebrovascular accident. In their study, the primary sampling unit is the family. Using multivariate model, it takes the advantage to incorporate the assumption that the failure times for subjects within a family are likely to be correlated. In this section, we extend our method to conduct inference in the high dimensional multivariate failure time setting.

To be more specific about the model, assume there are n independent clusters (families). Each cluster i contains M_i subjects, and for each subject, there are K types of failure may occur. Thus, it is reasonable to assume that the number K is fixed that does not increase with dimensionality

d and sample size n. For example, $K = 2$ in the real example of Cai et al. (2005). Denote the d-dimensional covariates of the kth failure type of subject m in cluster i at time t by $\boldsymbol{X}_{ikm}(t)$. The marginal hazards model is taken as

$$\Lambda_{ikm}\{t|\boldsymbol{X}_{ikm}(t)\} = \Lambda_{0k}(t)\exp\{\boldsymbol{X}_{ikm}^T(t)\boldsymbol{\beta}\},$$

where the baseline hazard functions $\Lambda_{0k}(t)$'s are treated as nuisance parameters, and the model is known as mixed baseline hazards model. Using this model, our inference procedures are conducted based on the pseudo-partial likelihood approach, since the working model does not assume any correlation for the different failure times within each cluster. The log pseudo-partial likelihood loss function is

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{n}\Big[\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{m=1}^{M_i}\int_0^\tau \boldsymbol{X}_{ikm}^T(u)\boldsymbol{\beta}dN_{ikm}(u) - \sum_{k=1}^{K}\int_0^\tau \log\Big[\sum_{i=1}^{n}\sum_{m=1}^{M_i} Y_{ikm}(u)\exp\left\{\boldsymbol{X}_{ikm}^T(u)\boldsymbol{\beta}\right\}\Big]d\overline{N}_k(u)\Big],$$

where $Y_{ikm}(t)$ and $N_{ikm}(t)$ denote the at risk indicator and the number of observed failure event at time t of the kth type on subject m in cluster i, and $\overline{N}_k = \sum_{i=1}^n\sum_{m=1}^{M_i} N_{ikm}$ for each k. The penalized maximum pseudo likelihood estimator is

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}}\,\mathcal{L}(\boldsymbol{\beta}) + \mathcal{P}_\lambda(\boldsymbol{\beta}). \tag{H.1}$$

To connect the multivariate failure time model with Cox's proportional hazards model, first, we observe that we can drop the index m. This is by the fact that, for each (i,m) where $i \in \{1,...n\}$ and $m \in \{1,...,M_i\}$, we can map (i,m) to $i' = \sum_{j=1}^{i-1} M_j + m$, and we define $\sum_{j=1}^{0} M_j = 0$. It is not difficult to see the mapping is a bijection. After the mapping, the penalized estimator remains the same. Thus, without loss of generality, we assume $M_i = 1$ for all i, and we drop the index m. Next, we observe that the loss function $\mathcal{L}(\boldsymbol{\beta})$ is decomposable that

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\mathcal{L}^{(k)}(\boldsymbol{\beta}),$$

where

$$\mathcal{L}^{(k)}(\boldsymbol{\beta}) = -\frac{1}{n}\Big[\sum_{i=1}^{n}\int_0^t \boldsymbol{X}_{ik}^T(u)\boldsymbol{\beta}dN_{ik}(u) - \int_0^t \log\Big[\sum_{i=1}^{n} Y_{ik}(u)\exp\left\{\boldsymbol{X}_{ik}^T(u)\boldsymbol{\beta}\right\}\Big]d\overline{N}_k(u)\Big].$$

Thus, the loss function of multivariate failure time model can be decomposed into a sum of K loss functions of Cox's proportional hazards models. However, the extension of the inference of the Cox model to multivariate failure time model is not trivial since the loss function is derived from a pseudo-likelihood function.

First, we extend the estimation procedure to the multivariate failure time model in the high dimensional setting, where we take $\mathcal{P}_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$. It is not difficult to obtain that (2.2) holds

for the multivariate failure time model. An alternative approach is that we estimate $\boldsymbol{\beta}^*$ using each type k of failure time independently. Specifically, we construct the estimator $\widehat{\boldsymbol{\beta}}$ by

$$\widehat{\boldsymbol{\beta}} = K^{-1}\sum_{k=1}^{K}\widehat{\boldsymbol{\beta}}^{(k)}, \text{ where } \widehat{\boldsymbol{\beta}}^{(k)} = \underset{\boldsymbol{\beta}^{(k)}}{\operatorname{argmin}}\, \mathcal{L}^{(k)}(\boldsymbol{\beta}^{(k)}) + \lambda\|\boldsymbol{\beta}^{(k)}\|_1, \text{ for all } k.$$

Since for each $\widehat{\boldsymbol{\beta}}^{(k)}$, $\|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(\lambda s)$ by (2.2), it is readily seen that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_{\mathbb{P}}(\lambda s)$.

We extend the decorrelated score, Wald and partial likelihood ratio tests to the multivariate failure time model. We first introduce some notation. For $k = 1, ..., K$,

$$S_k^{(r)}(t,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{ik}^{\otimes r}(t)Y_{ik}(t)\exp\{\boldsymbol{X}_{ik}^T(t)\boldsymbol{\beta}\}, \text{ for } r = 0, 1, 2, \text{ and } \overline{\boldsymbol{Z}}_k(t,\boldsymbol{\beta}) = \frac{S_k^{(1)}(t,\boldsymbol{\beta})}{S_k^{(0)}(t,\boldsymbol{\beta})},$$

where their corresponding population versions are

$$\mathbf{s}_k^{(r)}(t,\boldsymbol{\beta}) = \mathbb{E}\big[Y_k(t)\boldsymbol{X}_{ik}(t)^{\otimes r}\exp\{\boldsymbol{X}_{ik}(t)\boldsymbol{\beta}\}\big], \text{ for } r = 0, 1, 2, \text{ and } \mathbf{e}_k(t,\boldsymbol{\beta}) = \mathbf{s}_k^{(1)}(t,\boldsymbol{\beta})/\mathbf{s}_k^{(0)}(t,\boldsymbol{\beta}).$$

Next, we derive the gradient and Hessian matrix at the point $\boldsymbol{\beta}$ of the loss function,

$$\nabla\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^\tau \big\{\boldsymbol{X}_{ik}(u) - \overline{\boldsymbol{Z}}_k(u,\boldsymbol{\beta})\big\}dN_{ik}(u),$$

and

$$\nabla^2\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{k=1}^{K}\int_0^\tau \Big\{\frac{S_k^{(2)}(u,\boldsymbol{\beta})}{S_k^{(0)}(u,\boldsymbol{\beta})} - \overline{\boldsymbol{Z}}_k(u,\boldsymbol{\beta})^{\otimes 2}\Big\}d\overline{N}_k(u).$$

The population version of the gradient and Hessian matrix are

$$\mathbf{g}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\mathbb{E}\Big[\int_0^\tau \big\{\boldsymbol{X}(u) - \mathbf{e}_k(u,\boldsymbol{\beta})\big\}d\overline{N}_k(u)\Big],$$

and

$$\mathbf{H}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\mathbb{E}\Big[\int_0^\tau \Big\{\frac{\mathbf{s}_k^{(2)}(u,\boldsymbol{\beta})}{\mathbf{s}_k^{(0)}(u,\boldsymbol{\beta})} - \mathbf{e}_k(u,\boldsymbol{\beta})^{\otimes 2}\Big\}d\overline{N}_k(u)\Big].$$

For notational simplicity, let $\mathbf{H}^* = \mathbf{H}(\boldsymbol{\beta}^*)$.

Note that, utilizing the decomposable structure, by the similar argument, the concentration results in Section E of Supplementary Materials hold for the empirical gradient and Hessian matrix. We estimate the decorrelation vector $\mathbf{w}^* = \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{*-1}\mathbf{H}_{\boldsymbol{\theta}\alpha}^*$ by the following estimator

$$\widehat{\mathbf{w}} = \operatorname{argmin}\|\mathbf{w}\|_1, \quad \text{subject to } \|\nabla_{\boldsymbol{\theta}\alpha}^2\mathcal{L}(0,\widehat{\boldsymbol{\theta}}) - \mathbf{w}^T\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}^2\mathcal{L}(0,\widehat{\boldsymbol{\theta}})\|_\infty \le \delta, \tag{H.2}$$

where δ is a tuning parameter.

We first introduce the decorrelated score test in multivariate failure time model. Suppose the null hypothesis is H_0: $\alpha^* = 0$, and the alternative hypothesis is H_a: $\alpha^* \neq 0$. The decorrelated score function is constructed similar to (3.4) that

$$\widehat{U}^M(0,\widehat{\boldsymbol{\theta}}) = \nabla_\alpha\mathcal{L}(0,\widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}}\mathcal{L}(0,\widehat{\boldsymbol{\theta}}). \tag{H.3}$$

The main technical difference between the multivariate failure time model and the univariate Cox's model is that, the loss function of Cox's model is a log profile likelihood function, and Bartlett's identity $\mathrm{Var}\{\nabla\mathcal{L}(\boldsymbol{\beta}^*)\} = \mathbb{E}(\nabla^2\mathcal{L}(\boldsymbol{\beta}^*))$ holds. In multivariate case, this identity does not hold. We need the following lemma which is analogous to Lemma I.1. We omit the proof details to avoid repetition.

Lemma H.1. For any vector $\mathbf{v} \in \mathbb{R}^d$, if $\|\mathbf{v}\|_0 \le s'$ and $\sqrt{s' \log d/n} = o(1)$, it holds that

$$\frac{\sqrt{n}\mathbf{v}^T\nabla\mathcal{L}(\boldsymbol{\beta}^*)}{\sqrt{\mathbf{v}^T\boldsymbol{\Omega}\mathbf{v}}} \xrightarrow{d} N(0,1). \quad \text{where } \Omega = \mathrm{Var}\{\sqrt{n}\nabla\mathcal{L}(\boldsymbol{\beta}^*)\} \in \mathbb{R}^{d\times d}.$$

By the similar argument as in Theorem 4.5, we derive the asymptotic normality of $\widehat{U}^M(0,\widehat{\boldsymbol{\theta}})$ in the next theorem.

Theorem H.2. Suppose that Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold. Let $\widehat{U}^M(0,\widehat{\boldsymbol{\theta}})$ be defined in (H.3). Under the null hypothesis that $\alpha^* = 0$ and if $\lambda \asymp \sqrt{n^{-1}\log d}$, $\delta \asymp s'\sqrt{n^{-1}\log d}$, $n^{-1/2}s^3\log d = o(1)$, we have

$$\sqrt{n}\widehat{U}^M(0,\widehat{\boldsymbol{\theta}}) \xrightarrow{d} Z, \text{ where } Z \sim N(0,\sigma^2) \text{ and } \sigma^2 = \Omega_{\alpha\alpha} - 2\mathbf{w}^{*T}\Omega_{\boldsymbol{\theta}\alpha} + \mathbf{w}^{*T}\Omega_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{w}^*.$$

Proof. By the definition of $\widehat{U}^M(0,\widehat{\boldsymbol{\theta}})$ and mean value theorem, we have, for some $z, z' \in [0,1]$, $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + z(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + z'(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$,

$$\begin{aligned}
\widehat{U}^M(0,\widehat{\boldsymbol{\theta}}) &= \nabla_\alpha\mathcal{L}(0,\widehat{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}}\mathcal{L}(0,\widehat{\boldsymbol{\theta}}) \\
&= \nabla_\alpha\mathcal{L}(0,\boldsymbol{\theta}^*) + \nabla_{\alpha\boldsymbol{\theta}}\mathcal{L}(0,\boldsymbol{\theta}^*) - \{\widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}}\mathcal{L}(0,\boldsymbol{\theta}^*) + \widehat{\mathbf{w}}\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\} \\
&= \underbrace{\nabla_\alpha\mathcal{L}(0,\boldsymbol{\theta}^*) - \mathbf{w}^{*T}\nabla_{\boldsymbol{\theta}}\mathcal{L}(0,\boldsymbol{\theta}^*)}_{S} + \underbrace{(\mathbf{w}^* - \widehat{\mathbf{w}})^T\nabla_{\boldsymbol{\theta}}\mathcal{L}(0,\boldsymbol{\theta}^*)}_{E_1} \\
&\quad + \underbrace{\{\nabla_{\alpha\boldsymbol{\theta}}\mathcal{L}(0,\bar{\boldsymbol{\theta}}) - \widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\widetilde{\boldsymbol{\theta}})\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}_{E_2}.
\end{aligned}$$

Using Lemma H.1, taking $\mathbf{b} = (1, -\mathbf{w}^{*T})^T$ and by the assumption that $\|\mathbf{w}^*\|_0 \le s'$, it holds that

$$\sqrt{n}S \xrightarrow{d} Z, \text{ where } Z \sim N(0,\sigma^2) \text{ and } \sigma^2 = \Omega_{\alpha\alpha} - 2\mathbf{w}^{*T}\Omega_{\boldsymbol{\theta}\alpha} + \mathbf{w}^{*T}\Omega_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{w}^*.$$

Following a similar proof as that in Theorem 4.5 and utilizing the separable structure in multivariate failure time model, we have $\sqrt{n}E_1 = o_{\mathbb{P}}(1)$ and $\sqrt{n}E_2 = o_{\mathbb{P}}(1)$. This concludes our proof. □

Remark H.3. Under the assumptions of H.2, using plug-in estimator $\widehat{\sigma}^2 = \widehat{\Omega}_{\alpha\alpha} - 2\widehat{\mathbf{w}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}\alpha} + \widehat{\mathbf{w}}^T\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}\boldsymbol{\theta}}\widehat{\mathbf{w}}$ converges to σ^2 at the rate of $\mathcal{O}_{\mathbb{P}}(s's\sqrt{n^{-1}\log d}) = o_{\mathbb{P}}(1)$.

Next, we extend the decorrelated Wald test to the multivariate failure time model, which constructs confidence intervals for α^*. We first estimate $\boldsymbol{\beta}^*$ by ℓ_1-penalized estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\alpha}, \widehat{\boldsymbol{\theta}})$. Let

$$\widetilde{\alpha}^M = \widehat{\alpha} - \left\{\frac{\partial\widehat{U}^M(\widehat{\alpha},\widehat{\boldsymbol{\theta}})}{\partial\alpha}\right\}^{-1}\widehat{U}^M(\widehat{\alpha},\widehat{\boldsymbol{\theta}}).$$

We derive the asymptotic normality of $\widetilde{\alpha}^M$ in the next theorem.

Theorem H.4. Suppose Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold. For $\lambda \asymp \sqrt{n^{-1}\log d}$, $\delta \asymp s'\sqrt{n^{-1}\log d}$ and $n^{-1/2}s^3 \log d = o(1)$, under the null hypothesis that $\alpha^* = 0$, we have

$$\sqrt{n}\widetilde{\alpha} \xrightarrow{d} Z, \quad \text{where } Z \sim N(0, \sigma^2/\gamma^4),$$

and $\sigma^2 = \Omega_{\alpha\alpha} - 2\mathbf{w}^{*T}\Omega_{\boldsymbol{\theta}\alpha} + \mathbf{w}^{*T}\Omega_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{w}^*$, $\gamma^2 = \mathbf{H}^*_{\alpha\alpha} - \mathbf{w}^{*T}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$.

Proof. By the definition of $\widetilde{\alpha}$, we have,

$$\begin{aligned}
\widetilde{\alpha} &= \widehat{\alpha} - \Big[\gamma^{-2} - \gamma^{-2} + \Big\{\frac{\partial \widehat{U}^M(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\}^{-1}\Big]\widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}) \\
&= \widehat{\alpha} - \gamma^{-2}\Big\{\widehat{U}^M(0, \widehat{\boldsymbol{\theta}}) + \frac{(\widehat{\alpha}-0)\partial \widehat{U}^M(\bar{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\} + \Big[\gamma^{-2} - \Big\{\frac{\partial \widehat{U}^M(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\}^{-1}\Big]\widehat{U}(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}), \text{ where} \\
&= \widehat{\alpha} - \gamma^{-2}\widehat{U}^M(0, \widehat{\boldsymbol{\theta}}) - \widehat{\alpha}\gamma^2\gamma^{-2} + \widehat{\alpha}\gamma^{-2}\Big\{\gamma^2 - \frac{\partial \widehat{U}^M(\bar{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\} + \widehat{U}^M(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})\Big[\gamma^{-2} - \Big\{\frac{\partial \widehat{U}^M(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}}{\partial \alpha}\Big\}^{-1}\Big], \\
&= \underbrace{-\gamma^{-2}\widehat{U}^M(0, \widehat{\boldsymbol{\theta}})}_{S} + \underbrace{\widehat{\alpha}\gamma^{-2}\Big\{\gamma^2 - \frac{\partial \widehat{U}^M(\bar{\alpha}, \widehat{\boldsymbol{\theta}})}{\partial \alpha}\Big\}}_{R_1} + \underbrace{\widehat{U}^M(\widehat{\alpha}, \widehat{\boldsymbol{\theta}})\Big[\gamma^{-2} - \Big\{\frac{\partial \widehat{U}^M(\widehat{\alpha}, \widehat{\boldsymbol{\theta}}}{\partial \alpha}\Big\}^{-1}\Big]}_{R_2},
\end{aligned}$$

where the second equality holds by mean value theorem for some $\bar{\alpha} = v\widehat{\alpha}$ and $v \in [0,1]$. For the first term above, we have $\sqrt{n}S \xrightarrow{d} Z$ where $Z \sim N(0, \sigma^2/\gamma^4)$ by Theorem H.2. In addition, $\sqrt{n}R_1 = o_{\mathbb{P}}(1)$ and $\sqrt{n}R_2 = o_{\mathbb{P}}(1)$ by the similar argument in Theorem 4.9. This concludes the proof. □

Finally, we extend the decorrelated partial likelihood ratio test to the multivariate failure time model. The test statistic is

$$2n\big\{\mathcal{L}(0, \widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widetilde{\alpha}, \widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}})\big\}.$$

Under the null hypothesis, the test statistic follows a weighted chi-squared distribution as shown in the following theorem.

Theorem H.5. Suppose Assumptions 2.1, 2.2, 4.1, 4.2 and 4.3 hold. If $\lambda \asymp \sqrt{n^{-1}\log d}$, $\delta \asymp s'\sqrt{n^{-1}\log d}$ and $n^{-1/2}s^3 \log d$, under the null hypothesis $\alpha^* = 0$, we have

$$2n\big\{\mathcal{L}(0, \widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widetilde{\alpha}, \widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}})\big\} \xrightarrow{d} \sigma^2\gamma^{-2}Z_\chi, \text{ where } Z_\chi \sim \chi^2_1,$$

and $\sigma^2 = \Omega_{\alpha\alpha} - 2\mathbf{w}^{*T}\Omega_{\boldsymbol{\theta}\alpha} + \mathbf{w}^{*T}\Omega_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{w}^*$, $\gamma^2 = \mathbf{H}^*_{\alpha\alpha} - \mathbf{w}^{*T}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}$.

Proof. We have, by mean value theorem, for some $\bar{\alpha} = v_1\widehat{\alpha}$, $\bar{\alpha}' = v_2\widehat{\alpha}$, $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + t_3(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $\bar{\boldsymbol{\theta}}' = \boldsymbol{\theta}^* + v_4(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $0 \le v_1, v_2, v_3, v_4 \le 1$,

$$\begin{aligned}
&\mathcal{L}(\widetilde{\alpha}, \widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}}) - \mathcal{L}(0, \widehat{\boldsymbol{\theta}}) \\
=&\widetilde{\alpha}\nabla_\alpha \mathcal{L}(0, \widehat{\boldsymbol{\theta}}) - \widetilde{\alpha}\widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}}\mathcal{L}(0, \widehat{\boldsymbol{\theta}}) + \frac{\widetilde{\alpha}^2}{2}\nabla_{\alpha\alpha}(\mathcal{L}(\bar{\alpha}, \widehat{\boldsymbol{\theta}}) + \widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}})\widehat{\mathbf{w}} - \widetilde{\alpha}^2\widehat{\mathbf{w}}^T\nabla_{\alpha\boldsymbol{\theta}}\mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}') \\
=&\underbrace{\widetilde{\alpha}\widehat{U}(0, \widehat{\boldsymbol{\theta}})}_{L} + \underbrace{\frac{\widetilde{\alpha}^2}{2}\Big\{\nabla_{\alpha\alpha}\mathcal{L}(\bar{\alpha}, \widehat{\boldsymbol{\theta}}) + \widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0, \bar{\boldsymbol{\theta}})\widehat{\mathbf{w}} - 2\widehat{\mathbf{w}}\nabla_{\alpha\boldsymbol{\theta}}\mathcal{L}(\bar{\alpha}', \bar{\boldsymbol{\theta}}')\Big\}}_{E}.
\end{aligned}$$

We first look at the term L. By Theorem H.2, we have $\widehat{U}(0,\widehat{\boldsymbol{\theta}}) = \widehat{U}(0,\widehat{\boldsymbol{\theta}}^*) + o_{\mathbb{P}}\big(n^{-1/2}\big)$, and by Theorem H.4 $\widetilde{\alpha} = -\gamma^{-2}\widehat{U}(0,\widehat{\boldsymbol{\theta}}) + o_{\mathbb{P}}\big(n^{-1/2}\big)$, we have

$$L = -\gamma^{-2}\widehat{U}^M(0,\widehat{\boldsymbol{\theta}})^2 + o_{\mathbb{P}}\big(n^{-1}\big).$$

Next, we look at the term E,

$$E = \underbrace{\frac{\widetilde{\alpha}^2}{2}\big(\mathbf{H}^*_{\alpha\alpha} + \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha} - 2\mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{H}^{*-1}_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{H}^*_{\boldsymbol{\theta}\alpha}\big)}_{E_1}$$
$$+ \underbrace{\frac{\widetilde{\alpha}^2}{2}\Big[\big\{\nabla_{\alpha\alpha}\mathcal{L}(\bar{\alpha},\widehat{\boldsymbol{\theta}}) - \mathbf{H}^*_{\alpha\alpha}\big\} + \big\{\widehat{\mathbf{w}}^T\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{L}(0,\bar{\boldsymbol{\theta}})\widehat{\mathbf{w}} - \mathbf{w}^*\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathbf{w}^*\big\} - 2\big\{\widetilde{\mathbf{w}}\nabla_{\alpha\boldsymbol{\theta}}\mathcal{L}(\bar{\alpha}',\bar{\boldsymbol{\theta}}') - \mathbf{H}^*_{\alpha\boldsymbol{\theta}}\mathbf{w}^*\big\}\Big]}_{E_2}.$$

By Theorem H.4, it holds that $2nE_1 \xrightarrow{d} \sigma^2\gamma^{-2}Z_\chi$. In addition, by the similar argument as in Theorem 4.11, we have $E_2 = o_{\mathbb{P}}\big(n^{-1}\big)$. Thus, we have

$$2n\big\{\mathcal{L}(0,\widehat{\boldsymbol{\theta}}) - \mathcal{L}(\widetilde{\alpha},\widehat{\boldsymbol{\theta}} - \widetilde{\alpha}\widehat{\mathbf{w}})\big\} \xrightarrow{d} \sigma^2\gamma^{-2}Z_\chi, \text{ where } Z_\chi \sim \chi^2_1,$$

which concludes our proof. □

I More Simulation Results

In this section, we provide more simulation results for the inference on the low dimensional parametric component α. Using the same data generating schemes in Section 6, we first provide more detailed simulation results about the decorrelated estimator $\widetilde{\alpha}$ defined in (3.8), where we provide the estimator's bias, standard deviation, estimated standard deviation and empirical coverage probability for $\beta_1 = 0$ and $\beta_2 = 1$ in Tables 1 and 2. We find that for both zero and nonzero coefficients the proposed estimator has very small bias. In addition, the estimated standard errors (ESE) are close to the empirical standard errors (SE) and the coverage probabilities are very close to 95%. This suggests that the theoretical results on the asymptotic normality of the proposed estimator work well in empirical studies.

Next, we investigate how our proposed methods work when $\boldsymbol{\beta}^*$ is not very sparse. Setting $s = 10$ and 20, using the same data generating scheme as in Section 6, we report the empirical Type I error in Tables 3 and 4. It is seen that when $d = 100$ and 200, the type I error of the proposed method is reasonably accurate. As d further increases to 500, the tests become conservative (i.e., type I error is smaller than the nominal level). This is mainly due to the fact that there exists significant estimation error of the initial Lasso estimator under the PH model for large s and d. This phenomenon is consistent with the existing literature such as the numerical results in Bradic et al. (2011).

We then consider a scheme where the censoring rate is higher. Using the same data generating scheme as in Section 6, except that the i-th censoring time is generated from an exponential distribution with mean $U \times \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)$, where $U \sim \text{Unif}[1,2]$. This censoring scheme results in about 50% censored samples. As seen in Tables 5 and 6, our methods work reasonably well for

moderately high dimensional setting (i.e., $d = 100$ and 200). As d further increases to 500, the tests tend to be more conservative (the empirical Type I error rate is slightly smaller than the nominal level), due to the high censoring rate.

Finally, we look at the case where the inverse of the Fisher information matrix is not sparse. Using the same data generating scheme as in Section 6, except that we let the covariance matrix X be $\Sigma_{jj} = 1$ and $\Sigma_{jk} = \rho$ if $j \neq k$. The results are reported in Tables 7 and 8. We find that, when $d = 100$ or 200, the Type I error of our methods is very close to the nominal level. When $d = 500$, our test becomes a little bit more conservative in the sense that the Type I error is smaller than the nominal level. Thus, under the setting with non-sparse Σ^{-1}, our methods still empirically work very well with moderately high dimensional covariates. For very high dimensional case (i.e., $d = 500$), the proposed test is less powerful, because the estimation of $\mathbf{w}^*$ seems less accurate.

J Simulation for the Inference on the Baseline Hazard Function on Simulated Data

In this section, we demonstrate the empirical performance of the decorrelated inference procedure on the baseline hazard function $\Lambda_0(t)$ proposed as in Section 5. We consider three scenarios with $\Lambda_0(t) = t$, $t^2/2$ and $t^3/3$. Note that when $\Lambda_0(t) = p^{-1}t^p$, the survival time follows a Weibull distribution with shape parameter p and scale parameter $\{p \exp(-\boldsymbol{X}_i^T \boldsymbol{\beta}^*)\}^{1/p}$, i.e., $W\big(p, \{p \exp(-\boldsymbol{X}_i^T \boldsymbol{\beta}^*)\}^{1/p}\big)$. We use the same data generating procedures for the covariate $\boldsymbol{X}_i$'s, parameter $\boldsymbol{\beta}^*$ and censoring time R as in the previous subsection.

In each simulation, we construct 95% confidence intervals for $\Lambda_0(t)$ at $t = 0.2$ using the procedures proposed in Section 5. The simulation is repeated 1,000 times. The results for the empirical coverage probabilities of $\Lambda_0(t)$ are summarized in Tables 9 and 10. It is seen that the coverage probabilities are all between 93% and 97%, which matches our theoretical results.

To further examine the performance of our method, we conduct additional simulation studies by plotting the 95% confidence intervals of $\Lambda_0(t)$ at $t = 0.05, 0.1, 0.15, ..., 0.5$, with $\Lambda_0(t) = t$ and $t^2/2$. The results are presented in Figures 1 and 2.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* 1100–1120.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.*, **39** 3092–3120.

Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika*, **92** 303–316.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle inequalities for the Lasso in the Cox model. *Ann. Statist.*, **41** 1142–1165.

Kosorok, M. R. (2007). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Massart, P. (2007). *Concentration inequalities and model selection*, vol. 6. Springer.

Table 1: Biases, standard errors (SE), estimated standard errors (ESE) and coverage probabilities (CP) for the Wald estimator for $\beta_1 = 0$ and $\beta_2 = 1$ with nominal coverage probability 95%, where $(s, n) = (2, 150)$.

		Bias		SE		ESE		CP	
ρ	d	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0.25	100	0.031	-0.019	0.184	0.203	0.231	0.217	94.7%	93.9%
	200	0.035	-0.049	0.237	0.231	0.312	0.272	94.6%	93.5%
	500	-0.038	-0.057	0.284	0.263	0.325	0.297	94.0%	92.9%
0.4	100	0.030	-0.042	0.245	0.184	0.262	0.209	94.8%	93.5%
	200	-0.029	-0.061	0.272	0.256	0.276	0.293	94.6%	93.0%
	500	-0.037	-0.066	0.330	0.312	0.381	0.365	93.8%	92.7%
0.6	100	-0.026	-0.043	0.264	0.198	0.287	0.234	94.9%	93.7%
	200	0.029	-0.051	0.294	0.269	0.327	0.294	94.5%	93.2%
	500	0.046	-0.065	0.385	0.310	0.415	0.348	95.6%	92.7%
0.75	100	0.023	-0.052	0.283	0.206	0.301	0.254	95.1%	93.6%
	200	-0.034	-0.057	0.342	0.257	0.352	0.298	95.3%	92.9%
	500	0.067	-0.070	0.411	0.325	0.398	0.402	96.2%	92.1%

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Table 2: Biases, standard errors (SE), estimated standard errors (ESE) and coverage probabilities (CP) for the Wald estimator for $\beta_1 = 0$ and $\beta_2 = 1$ with nominal coverage probability 95%, where $(s, n) = (3, 150)$.

		Bias		SE		ESE		CP	
ρ	d	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0.25	100	-0.025	-0.035	0.203	0.185	0.228	0.196	94.9%	93.4%
	200	-0.034	-0.049	0.248	0.233	0.312	0.261	95.2%	92.8%
	500	0.057	-0.065	0.362	0.357	0.349	0.382	93.2%	92.0%
0.4	100	0.024	-0.045	0.261	0.195	0.285	0.250	94.8%	93.7%
	200	-0.041	-0.065	0.284	0.254	0.307	0.273	95.4%	93.0%
	500	0.056	-0.081	0.322	0.343	0.424	0.362	93.9%	92.2%
0.6	100	0.028	-0.052	0.268	0.213	0.314	0.206	94.6%	93.5%
	200	0.034	-0.059	0.307	0.275	0.365	0.277	95.1%	92.7%
	500	-0.037	-0.083	0.379	0.324	0.431	0.356	94.2%	92.3%
0.75	100	-0.022	-0.047	0.274	0.212	0.325	0.265	95.0%	93.2%
	200	0.033	-0.062	0.325	0.271	0.388	0.307	94.4%	92.5%
	500	-0.052	-0.069	0.419	0.348	0.463	0.376	95.9%	91.7%

Table 3: Average Type I error under the non-sparse $\boldsymbol{\beta}^*$ setting with $\eta = 5\%$ and $(n, s) = (150, 10)$.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
Score	100	6.3%	6.6%	6.8%	6.7%	6.4%	6.2%	6.0%	5.7%
	200	6.4%	6.1%	5.5%	5.7%	3.8%	4.1%	3.6%	3.3%
	500	1.9%	2.0%	2.5%	2.6%	2.1%	1.5%	1.5%	0.9%
Wald	100	7.1%	6.8%	6.7%	7.2%	6.4%	6.2%	5.5%	5.2%
	200	7.0%	6.7%	5.7%	6.5%	4.1%	4.7%	3.8%	3.3%
	500	2.5%	1.9%	2.9%	2.6%	1.1%	1.7%	0.8%	1.5%
LRT	100	7.6%	7.0%	7.4%	7.2%	6.6%	6.1%	5.2%	5.3%
	200	7.3%	7.2%	7.1%	6.9%	3.9%	3.3%	3.5%	3.8%
	500	1.3%	1.1%	2.1%	2.0%	2.4%	1.8%	1.6%	0.9%

Table 4: Average Type I error of under the non-sparse $\boldsymbol{\beta}^*$ setting with $\eta = 5\%$ and $(n, s) = (150, 20)$.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
Score	100	6.2%	6.8%	6.0%	6.3%	7.2%	7.2%	6.7%	7.1%
	200	7.5%	7.4%	6.5%	6.4%	6.6%	6.8%	7.2%	6.9%
	500	1.7%	1.1%	1.6%	0.8%	1.7%	1.0%	1.9%	1.2%
Wald	100	6.8%	6.9%	6.6%	6.5%	7.1%	7.3%	7.0%	7.5%
	200	6.9%	6.2%	6.5%	6.8%	6.6%	6.3%	7.2%	7.3%
	500	1.6%	1.2%	1.1%	1.0%	1.4%	1.3%	1.6%	0.7%
LRT	100	6.5%	6.3%	7.3%	6.8%	7.2%	7.6%	7.4%	7.5%
	200	5.9%	6.1%	6.2%	6.9%	6.6%	7.1%	7.2%	7.0%
	500	1.7%	1.4%	1.2%	1.3%	0.7%	1.4%	1.5%	1.2%

Table 5: Average Type I error under high-censoring setting with $\eta = 5\%$ under high censoring scheme where $(s, n) = (2, 150)$.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
Score	100	5.0%	4.7%	4.9%	4.6%	4.5%	4.2%	4.4%	4.1%
	200	4.4%	4.5%	4.1%	3.7%	3.6%	3.3%	3.5%	3.0%
	500	2.8%	2.8%	2.2%	2.4%	2.1%	2.6%	1.8%	2.3%
Wald	100	5.1%	5.2%	4.9%	4.6%	4.2%	4.4%	4.1%	4.3%
	200	3.9%	4.3%	4.1%	3.6%	3.4%	3.6%	3.2%	3.0%
	500	2.7%	2.6%	2.5%	2.5%	2.3%	2.4%	2.4%	2.0%
LRT	100	4.8%	4.6%	4.8%	4.5%	4.5%	4.6%	4.4%	4.3%
	200	3.9%	4.3%	3.8%	4.0%	3.6%	3.5%	3.4%	3.1%
	500	3.0%	2.5%	2.7%	2.1%	2.5%	1.9%	2.3%	2.0%

Table 6: Average Type I error under high-censoring setting with $\eta = 5\%$ under high censoring scheme where $(s, n) = (3, 150)$.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
Score	100	5.2%	5.5%	4.6%	4.8%	4.3%	4.4%	4.7%	4.5%
	200	4.5%	4.4%	4.3%	4.3%	4.0%	3.6%	3.3%	3.2%
	500	2.9%	2.8%	2.6%	2.2%	2.1%	1.9%	1.6%	1.7%
Wald	100	5.4%	5.2%	4.9%	4.7%	4.2%	4.5%	4.3%	4.0%
	200	4.6%	4.5%	4.0%	4.1%	3.6%	3.3%	3.2%	2.9%
	500	2.6%	2.5%	2.3%	2.3%	1.9%	1.9%	1.7%	2.0%
LRT	100	5.2%	4.8%	4.7%	4.6%	4.2%	3.9%	4.1%	4.3%
	200	4.5%	4.4%	4.6%	4.4%	3.2%	3.5%	3.4%	3.1%
	500	2.9%	2.6%	2.4%	2.3%	2.0%	1.8%	1.8%	1.9%

Table 7: Average Type I error under non-sparse $\mathbf{w}^*$ setting with $\eta = 5\%$ where $(n, s) = (150, 2)$.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
Score	100	5.0%	5.3%	5.1%	5.3%	4.7%	4.5%	3.8%	3.5%
	200	5.6%	5.3%	4.6%	4.8%	3.7%	3.5%	3.6%	3.3%
	500	2.9%	2.1%	2.4%	2.1%	2.7%	2.5%	2.0%	1.7%
Wald	100	5.3%	5.2%	5.4%	4.6%	4.8%	5.6%	5.2%	5.5%
	200	4.7%	5.1%	4.3%	4.2%	3.8%	4.1%	3.8%	3.4%
	500	2.7%	2.2%	2.3%	1.8%	2.0%	1.7%	1.7%	1.6%
LRT	100	5.5%	5.4%	4.8%	5.3%	4.0%	4.4%	5.2%	5.7%
	200	4.6%	4.9%	3.5%	3.9%	3.2%	3.6%	3.4%	3.7%
	500	2.0%	2.6%	2.4%	2.5%	2.7%	2.2%	1.9%	1.6%

Table 8: Average Type I error under non-sparse $\mathbf{w}^*$ setting with $\eta = 5\%$ where $(n, s) = (150, 3)$.

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
Method	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
Score	100	4.5%	4.8%	5.1%	4.7%	4.0%	4.5%	4.2%	3.5%
	200	3.8%	4.2%	4.1%	4.3%	4.2%	4.0%	3.7%	3.8%
	500	2.0%	2.4%	2.3%	2.5%	1.6%	1.5%	0.7%	1.0%
Wald	100	4.5%	5.0%	4.8%	5.4%	3.7%	3.5%	4.4%	4.2%
	200	3.8%	3.9%	3.7%	4.2%	3.8%	3.5%	3.8%	3.7%
	500	2.3%	2.1%	2.4%	2.0%	1.9%	2.4%	0.9%	0.8%
LRT	100	5.6%	5.2%	4.4%	4.5%	3.7%	3.5%	3.8%	4.3%
	200	3.6%	3.9%	4.5%	4.2%	3.5%	3.9%	3.6%	3.5%
	500	2.9%	2.2%	2.4%	2.0%	0.7%	1.4%	0.9%	0.7%

Table 9: Empirical coverage probability of 95% confidence intervals for $\Lambda_0(t)$ at $t = 0.2$ with $(n, s) = (150, 2)$

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
$\Lambda_0(t)$	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
t	100	95.3%	95.1%	94.7%	95.1%	95.2%	94.6%	95.4%	94.9%
	200	95.5%	95.8%	95.7%	95.3%	94.6%	94.5%	94.4%	94.2%
	500	95.9%	96.2%	95.5%	94.8%	94.3%	94.1%	93.7%	93.5%
t^2	100	95.1%	95.3%	95.2%	95.0%	95.4%	94.7%	95.2%	95.3%
	200	95.5%	94.8%	95.4%	94.7%	94.6%	94.0%	94.4%	94.5%
	500	96.6%	96.7%	96.1%	95.4%	94.9%	94.3%	93.8%	93.6%
t^3	100	95.2%	95.0%	95.1%	95.3%	94.8%	95.1%	95.2%	94.7%
	200	95.4%	94.7%	94.6%	95.5%	95.2%	95.8%	94.6%	94.3%
	500	96.6%	95.9%	96.3%	95.9%	94.5%	94.7%	93.6%	93.4%

Table 10: Empirical coverage probability of 95% confidence intervals for $\Lambda_0(t)$ at $t = 0.2$ with $(n, s) = (150, 3)$

		$\rho = 0.25$		$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.75$	
$\Lambda_0(t)$	d	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$	Dirac	Unif$[0, 2]$
t	100	95.1%	94.8%	94.8%	95.2%	95.3%	95.1%	94.8%	95.4%
	200	95.6%	95.3%	95.4%	95.2%	94.7%	94.8%	94.2%	94.3%
	500	96.2%	95.9%	95.8%	96.1%	95.2%	94.3%	93.3%	93.6%
t^2	100	95.3%	94.7%	95.3%	94.9%	94.5%	95.3%	95.4%	95.2%
	200	94.7%	94.5%	95.4%	95.2%	94.1%	94.9%	94.3%	93.8%
	500	96.5%	96.2%	95.8%	96.0%	95.5%	95.1%	93.2%	93.7%
t^3	100	95.0%	95.2%	94.6%	94.8%	95.1%	95.4%	94.9%	95.5%
	200	95.3%	95.5%	95.2%	94.5%	94.3%	94.6%	93.8%	93.5%
	500	95.9%	96.3%	95.7%	96.0%	95.4%	94.7%	93.6%	93.1%

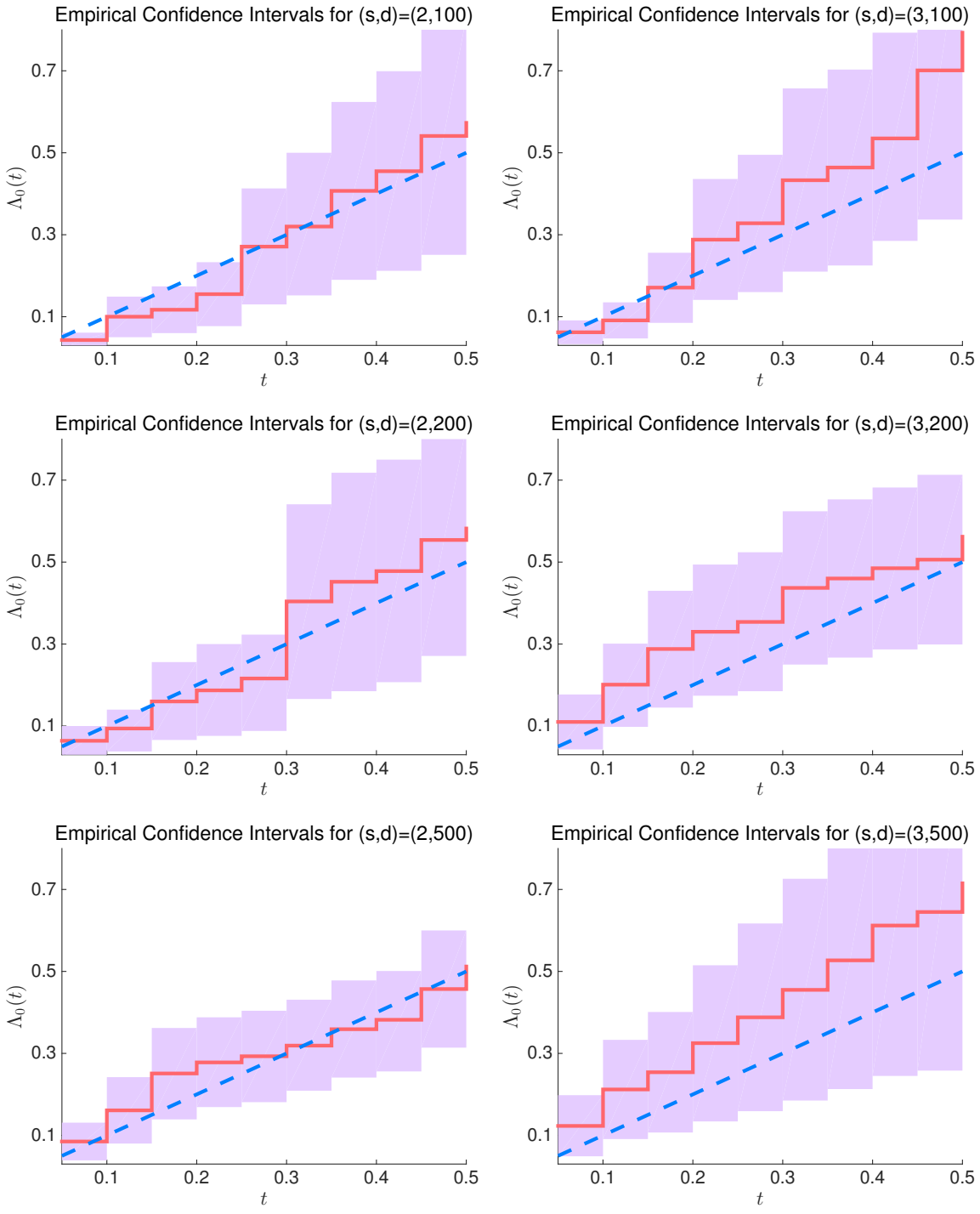

Figure 1: 95% confidence intervals for the baseline hazard function at $t = 0.05, 0.1, ..., 0.5$. The red solid line denotes the estimated baseline hazard function $\widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$, and blue dashed line denotes $\Lambda_0(t) = t$.

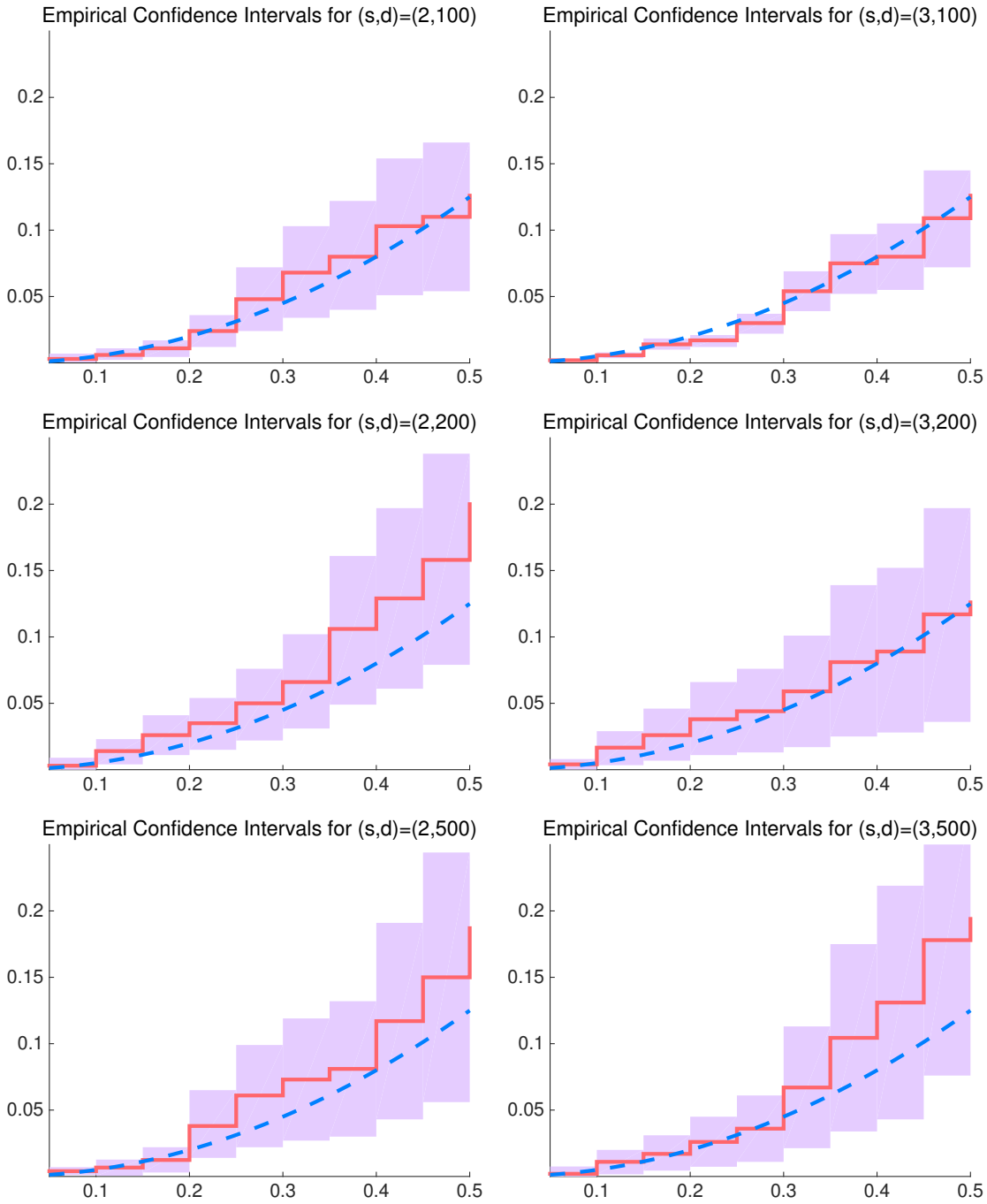

Figure 2: 95% confidence intervals for the baseline hazard function at $t = 0.05, 0.1, ..., 0.5$. The red solid line denotes the estimated baseline hazard function $\widetilde{\Lambda}_0(t, \widehat{\boldsymbol{\beta}})$, and the blue dashed line denotes $\Lambda_0(t) = t^2/2$.