Projection-Based and Cross-Validated Estimation in High-Dimensional Cox Model

Haixiang Zhang[1*], Jian Huang[2] and Liuquan Sun[3]

[1] *Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China*

[2] *Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA*

[3] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China*

Summary. We propose a projection-based cross-validation method for estimating a low-dimensional parameter in the presence of a high-dimensional nuisance parameter in the Cox regression model. We show that the proposed estimator is asymptotically normal, which enables us to conduct hypothesis test for the parameter of interest with high-dimensional nuisance parameters. Three decision rules are presented to avoid the influence of random splitting of samples. Simulation studies indicate that our method is more powerful than that of Fang et al. (2017, *JRSSB*) when the coefficients of predictors are high-dimensional and not very sparse. As an illustrative example, we apply our procedure to a breast cancer study.

Keywords: Cross-Validation; High-dimensional nuisance parameters; Hypothesis test; Oracle inequality; Sample splitting

1 Introduction

Statistical analysis of censored survival data with high-dimensional covariates is of great practical importance. For example, in cancer genetic studies, an important problem is to identify genetic elements that are potentially related to patient's survival from high-throughput and high-dimensional genomic data. A critical issue is how to estimate their effects on the survival and make statistical inference about their significance. This problem can be formulated as that of estimating treatment effects in the presence of a large number of nuisance parameters. Here we interpret a treatment effect parameter broadly as any low-dimensional parameter

*Corresponding author. Email: haixiang.zhang@tju.edu.cn

in the model. Therefore, it is interesting to propose an approach to statistical inference in high-dimensional Cox regression (Cox, 1972) because of its central role in the analysis of censored survival data and its wide applications (Fleming and Harrington,1991; Kalbfleisch and Prentice, 2002).

Several penalty-based variable selection approaches, including the lasso (Tibshirani, 1996) and the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001) methods, have been adapted to survival models. For example, Tibshirani (1997) and Fan and Li (2002) applied the lasso and SCAD methods to the partial likelihood for the Cox model. Zhang and Lu (2007) and Zou (2008) considered the weighted lasso for low-dimensional Cox model. Huang et al. (2013) and Kong and Nan (2014) derived error bounds for the lasso in sparse and high-dimensional Cox model.

However, penalized procedures only yield point estimates but do not provide inferential statements such as confidence interval and hypothesis testing about a parameter of interest. To deal with this problem, Zhang and Zhang (2014) proposed a regularized projection approach for constructing asymptotically normal estimators of low-dimensional parameters in high-dimensional linear models. van de Geer et al. (2014) extended the approach of Zhang and Zhang (2014) and proposed a novel method by "inverting" the Karush-Kuhn-Tucker conditions for the lasso to construct estimators of low-dimensional parameters in linear and generalized linear models. Javanmard and Montanari (2014) constructed confidence intervals and p-values for high-dimensional linear models based on a "de-biased" version of regularized M-estimators. Wasserman and Roeder (2009) and Meinshausen et al. (2009) constructed p-values for high-dimensional regression via sample-splitting based methods. However, these authors did not consider the statistical inference problem in the high-dimensional Cox model. In the context of survival analysis, Zhong et al. (2015) considered hypothesis testing for low-dimensional coefficients in the high-dimensional additive hazards model, but it is unclear how to extend their method to the Cox model. Another closely related work is Fang et al. (2017), who have proposed a method for hypothesis test and confidence interval construction for the high-dimensional Cox model based on projection of score functions. However, their method is conservative and suffers from inefficiency when the coefficients of predictors are high-dimensional and not very sparse (See page 24 of online supplementary materials of Fang et al., 2017).

In this paper, we propose a projection-based cross-validation approach to inference about a low-dimensional parameter of interest in the Cox model in the presence of a high-dimensional nuisance parameter. There are three important aspects of our proposed approach that are different from the above-mentioned methods. First, we use a weighted lasso estimator as the initial estimator. With this estimator, we only penalize the nuisance parameters, but not the parameter of interest. This is different from the methods of Zhang and Zhang (2014) and Fang et al. (2017) in which they used a fully penalized estimator as an initial estimator. Second, our method only needs to calculate the least favorable direction related to the scores of the selected nuisance parameters rather than the whole set of the nuisance parameters as in Fang et al. (2017). Third, our two-stage projection-based cross-validation technique is different from the sample splitting method in Meinshausen et al. (2009). Roughly speaking, we randomly split the sample into two halves, and obtain a weighted lasso estimator using the first half of the sample. Then we fit the Cox model using the variables selected based on the first half of the sample and use the second half of the data to estimate the parameter of interest; and vice versa. The proposed estimator is then the average of these two estimators. To avoid the influence of random splitting of samples, we further provide three decision rules for the hypothesis test of interest.

The remainder of this article is organized as follows. In Section 2 we describe the Cox model and propose a projection-based cross-validation estimator. In Section 3 we first state an oracle inequality for the weighted lasso in the high-dimensional Cox model. We then establish the asymptotic normality of the proposed estimator, which provides a theoretical basis for making statistical inference. In Section 4 we conduct simulation studies and demonstrate the proposed method on a breast cancer gene expression dataset. In Section 5 we give concluding remarks. All proofs are deferred to the Appendix.

2 Model and method

2.1 Model

Consider an n-dimensional counting process $N^{(n)}(t) = (N_1(t), \cdots, N_n(t))$, $t > 0$ on a time interval $[0, \tau]$ with $\tau > 0$, where $N_i(t)$ counts the number of observed events for the ith individual in the time interval $[0, t]$, $i = 1, \cdots, n$. Let $\mathcal{F}_t$ be the filtration representing all the information available up to time $t > 0$. Following Andersen and Gill (1982), we assume that for $\{\mathcal{F}_t, t \geq 0\}$, $N^{(n)}$ has a predictable compensator $\Lambda^{(n)} = (\Lambda_1, \cdots, \Lambda_n)$ with

$$d\Lambda_i(t) = Y_i(t) \exp\{\beta^{\mathrm{T}} X_i(t) + \eta^{\mathrm{T}} Z_i(t)\} d\Lambda_0(t), i = 1, \ldots, n, \tag{2.1}$$

where $\beta \in \mathbb{R}^d$ is a parameter vector of interest, $\eta \in \mathbb{R}^q$ is a vector of nuisance parameters, $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ is an unknown baseline cumulative hazard function, and $Y_i(t) \in \{0, 1\}$ is predictable. We assume the dimension d of the parameter vector of interest β is fixed and small, but the dimension q of the nuisance parameter η can be large or even larger than the sample size.

Denote $V_i(t) = (X_i(t)^T, Z_i(t)^{\mathrm{T}})^{\mathrm{T}}$ and let $\theta_0 = (\beta_0^T, \eta_0^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p$ be the true values of the regression coefficients, where p is possibly much bigger than n. Define $S_0 = \{j : \theta_{j0} \neq 0\}$ with its complement denoted by $S_0^c = \{j : \theta_{j0} = 0\}$. Let $d_0 = |S_0|$ be the cardinality of S_0 with $d_0 \ll n$.

To estimate the parameter θ in the fixed-dimensional settings with $p < n$, Cox (1975) proposed the partial likelihood method. The negative log-partial likelihood function for (2.1) is

$$\ell(\theta) = \frac{1}{n} \left[\int_0^\tau \log \Big[\sum_{i=1}^n Y_i(t) \exp\{\theta^{\mathrm{T}} V_i(t)\} \Big] d\bar{N}(t) - \sum_{i=1}^n \int_0^\tau \{\theta^{\mathrm{T}} V_i(t)\} dN_i(t) \right], \tag{2.2}$$

where $\bar{N} = \sum_{i=1}^n N_i$. The maximum partial likelihood estimator can be obtained by minimizing $\ell(\theta)$. However, in high-dimensional settings with $p \gg n$, the maximum partial likelihood estimator is not well defined. Thus statistical inference cannot be based on the partial likelihood directly.

For any given set $I \subset \{1, \cdots, n\}$ and $S \subset \{1, \cdots, p\}$, define

$$\Phi_k(t, \theta; I, S) \quad = \quad \frac{1}{|I|} \sum_{i \in I} V_{iS}^{\otimes k}(t) Y_i(t) \exp\{\theta_S^{\mathrm{T}} V_{iS}(t)\}, \quad \text{for} \quad k = 0, 1, 2; \tag{2.3}$$

$$\Sigma(\theta; I, S) = \frac{1}{|I|} \sum_{i\in I} \int_0^\tau \left[\frac{\Phi_2(t,\theta; I, S)}{\Phi_0(t,\theta; I, S)} - \left\{ \frac{\Phi_1(t,\theta; I, S)}{\Phi_0(t,\theta; I, S)} \right\}^{\otimes 2} \right] dN_i(t),$$

and

$$\ell(\theta; I, S) = \frac{1}{|I|} \left[\int_0^\tau \log \Big[\sum_{i\in I} Y_i(t) \exp\{\theta_S^{\mathrm{T}} V_{iS}(t)\} \Big] d\bar{N}(t; I) - \sum_{i\in I} \int_0^\tau \{\theta_S^{\mathrm{T}} V_{iS}(t)\} dN_i(t) \right], \quad (2.4)$$

where for any vector a, $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^{\mathrm{T}}$; a_S denotes the subvector of a with components whose indices are in S; $|I|$ denotes the cardinality of set I, and $\bar{N}(t; I) = \sum_{i\in I} N_i(t)$. Hereafter, for notational simplicity, we assume that $|I| = n/2$ if n is even and $|I| = (n+1)/2$ if n is odd. We partition the matrix $\Sigma(\theta; I, S)$ into

$$\Sigma(\theta; I, S) = \begin{bmatrix} \Sigma_{11}(\theta; I, S) & \Sigma_{12}(\theta; I, S) \\ \Sigma_{21}(\theta; I, S) & \Sigma_{22}(\theta; I, S) \end{bmatrix}, \quad (2.5)$$

where $\Sigma_{11}(\theta; I, S) \in \mathbb{R}^{d\times d}$, $\Sigma_{21}(\theta; I, S) \in \mathbb{R}^{(|S|-d)\times d}$, and $\Sigma_{22}(\theta; I, S) \in \mathbb{R}^{(|S|-d)\times(|S|-d)}$. Let $\Sigma_{\beta|\eta}(\theta; I, S) = \Sigma_{11}(\theta; I, S) - \Sigma_{12}(\theta; I, S)\Sigma_{22}^{-1}(\theta; I, S)\Sigma_{21}(\theta; I, S)$, we denote the population versions of the quantities in (2.3) as

$$\begin{aligned} \phi_k(t,\theta; S) &= E[V_S^{\otimes k}(t) Y(t) \exp\{\theta_S^{\mathrm{T}} V_S(t)\}], \quad \text{for} \quad k = 0, 1, 2; \\ \Sigma^*(\theta; S) &= E\left\{ \int_0^\tau \left[\frac{\phi_2(t,\theta; S)}{\phi_0(t,\theta; S)} - \left\{ \frac{\phi_1(t,\theta; S)}{\phi_0(t,\theta; S)} \right\}^{\otimes 2} \right] dN(t) \right\}. \end{aligned}$$

We partition the matrix $\Sigma^*(\theta; S)$ according to (2.5) as

$$\Sigma^*(\theta; S) = \begin{bmatrix} \Sigma_{11}^*(\theta; S) & \Sigma_{12}^*(\theta; S) \\ \Sigma_{21}^*(\theta; S) & \Sigma_{22}^*(\theta; S) \end{bmatrix}, \quad (2.6)$$

and let $\Sigma_{\beta|\eta}^*(\theta; S) = \Sigma_{11}^*(\theta; S) - \Sigma_{12}^*(\theta; S)\Sigma_{22}^{*-1}(\theta; S)\Sigma_{21}^*(\theta; S)$.

2.2 Projection-based cross-validation method.

In this section, we describe the proposed two-stage projection-based cross-validation approach to statistical inference for the high-dimensional Cox model. Our basic idea is to split the data randomly into two halves I_1 and I_2, and perform model selection using the first half of the data I_1. Then we fit the Cox model on the basis of the variables selected in the first stage, and calculate a projection-based estimator $\hat{\beta}_1$ using the second half of the data

I_2. We then switch the roles of I_1 and I_2 and use the same procedure to obtain an estimator $\hat{\beta}_2$. Below we describe the proposed method in details.

Stage 1. We split the data randomly into two halves I_1 and I_2. Using the first half of the data I_1, we obtain a weighted lasso estimator, which is defined as

$$\breve{\theta} = (\breve{\beta}, \breve{\eta}) = \underset{\beta,\eta}{\operatorname{argmin}} \left\{ \ell(\theta; I_1, S_p) + \lambda \sum_{j=1}^{q} w_j |\eta_j| \right\}, \tag{2.7}$$

where $\ell(\theta; I_1, S_p)$ is defined in (2.4), $S_p = \{1, \cdots, p\}$; $\lambda > 0$ is a tuning parameter, and $w_j \geq 0$ are weights for the nuisance parameters η_j, $j = 1, \cdots, q$. Let $S_1 = \{j : \breve{\theta}_j \neq 0\}$ be the index set of the nonzero estimated coefficients. Our goal is to make statistical inference about β, we only penalize η while β is not penalized. Thus, the estimator $\breve{\theta}$ can be referred as a "semi-penalized" estimator.

Stage 2. Consider a sub-model based on the variables selected in the first stage (S_1), using the second half of the data I_2,

$$d\Lambda_i(t) = Y_i(t) \exp\{\theta_{S_1}^{\mathrm{T}} V_{iS_1}(t)\} d\Lambda_0(t), \quad i \in I_2, \tag{2.8}$$

where $Y_i(t)$ and $\Lambda_0(t)$ are given in (2.1). The negative log-partial likelihood function based on (2.8) is

$$\ell(\theta; I_2, S_1) = \frac{1}{|I_2|} \left[\int_0^\tau \log \left[\sum_{i \in I_2} Y_i(t) \exp\{\theta_{S_1}^{\mathrm{T}} V_{iS_1}(t)\} \right] d\bar{N}(t; I_2) - \sum_{i \in I_2} \int_0^\tau \{\theta_{S_1}^{\mathrm{T}} V_{iS_1}(t)\} dN_i(t) \right] \tag{2.9}$$

where $\bar{N}(t; I_2) = \sum_{i \in I_2} N_i(t)$. Let $M_i(t) = N_i(t) - \int_0^t Y_i(u) \exp\{\theta_{S_1}^{\mathrm{T}} V_{iS_1}(u)\} d\Lambda_0(u)$ be the martingales with predictable variation processes $\langle M_i, M_i \rangle(t) = \int_0^t Y_i(u) \exp\{\theta_{S_1}^{\mathrm{T}} V_{iS_1}(u)\} d\Lambda_0(u)$, and $\langle M_i, M_j \rangle = 0$ for $i \neq j$. The gradient of $\ell(\theta; I_2, S_1)$ is

$$\dot{\ell}(\theta; I_2, S_1) = \frac{\partial \ell(\theta; I_2, S_1)}{\partial \theta_{S_1}} = -\frac{1}{|I_2|} \sum_{i \in I_2} \int_0^\tau \{V_{iS_1}(t) - \bar{V}(t, \theta; I_2, S_1)\} dN_i(t),$$

and the Hessian matrix of $\ell(\theta; I_2, S_1)$ is

$$\ddot{\ell}(\theta; I_2, S_1) = \Sigma(\theta; I_2, S_1) = \frac{1}{|I_2|} \int_0^\tau \left[\frac{\Phi_2(t, \theta; I_2, S_1)}{\Phi_0(t, \theta; I_2, S_1)} - \left\{ \frac{\Phi_1(t, \theta; I_2, S_1)}{\Phi_0(t, \theta; I_2, S_1)} \right\}^{\otimes 2} \right] d\bar{N}(t; I_2),$$

where $\bar{V}(t, \theta; I_2, S_1) = \Phi_1(t, \theta; I_2, S_1)/\Phi_0(t, \theta; I_2, S_1)$, and $\Phi_k(t, \theta; I, S)$ is defined in (2.3), k=0 and 1. For notational simplicity, we partition the gradient $\dot{\ell}(\theta; I_2, S_1)$ into $\dot{\ell}(\theta; I_2, S_1) =$

$(\dot{\ell}_\beta(\theta; I_2, S_1), \dot{\ell}_\eta(\theta; I_2, S_1)^{\mathrm{T}})^{\mathrm{T}}$, where $\dot{\ell}_\beta(\theta; I_2, S_1) \in \mathbb{R}^d$ is the score function for the low-dimensional parameter of interest β, and $\dot{\ell}_\eta(\theta; I_2, S_1) \in \mathbb{R}^{|S_1|-d}$ is the score function of the nuisance parameters.

To remove the effects of the nuisance parameters, we project $\dot{\ell}_\beta(\theta; I_2, S_1)$ onto the linear span of the partial score function $\dot{\ell}_\eta(\theta; I_2, S_1)$ and consider the projected partial score function for β,

$$U(\theta_0, h_0; I_2, S_1) = \dot{\ell}_\beta(\theta_0; I_2, S_1) - h_0^{\mathrm{T}}\dot{\ell}_\eta(\theta_0; I_2, S_1), \tag{2.10}$$

where $h_0 = \arg\min_h E\{\dot{\ell}_\beta(\theta_0; I_2, S_1) - h_0^{\mathrm{T}}\dot{\ell}_\eta(\theta_0; I_2, S_1)\}^{\otimes 2}$ with an explicit expression

$$\begin{aligned} h_0 &= E\{\dot{\ell}_\eta(\theta_0; I_2, S_1)\dot{\ell}_\eta^{\mathrm{T}}(\theta_0; I_2, S_1)\}^{-1}E\{\dot{\ell}_\eta(\theta_0; I_2, S_1)\dot{\ell}_\beta(\theta_0; I_2, S_1)\} \\ &= \Sigma_{22}^{*-1}(\theta_0; S_1)\Sigma_{21}^{*}(\theta_0; S_1). \end{aligned} \tag{2.11}$$

To better understand (2.10), we focus on the geometric interpretation for $U(\theta_0, h_0; I_2, S_1)$. The linear space $\mathcal{H}$ spanned by the score function $\dot{\ell}(\theta; I_2, S_1)$ is the closure of $\{a_\theta^{\mathrm{T}}\dot{\ell}_\beta(\theta; I_2, S_1) + b_\theta^{\mathrm{T}}\dot{\ell}_\eta(\theta; I_2, S_1) : a_\theta \in \mathbb{R}^d, b_\theta \in \mathbb{R}^{|S_1|-d}\}$. As indicated by the notation, a_θ and b_θ can depend on θ. By Small and McLeish (1994), the space $\mathcal{H}$ is a Hilbert space with an inner product given by $\langle g_1(\theta; I_2, S_1), g_2(\theta; I_2, S_1)\rangle = E\{g_1(\theta; I_2, S_1)g_2(\theta; I_2, S_1)\}$ for any $g_1 \in \mathcal{H}$ and $g_2 \in \mathcal{H}$. We further consider the linear space $\mathcal{H}_N$ spanned by the nuisance score functions $\{b_\theta^{\mathrm{T}}\dot{\ell}_\eta(\theta; I_2, S_1)\}$ with $b_\theta \in \mathbb{R}^{|S_1|-d}$, and its orthogonal complement $\mathcal{H}_N^{\perp} = \{g \in \mathcal{H}, \langle g, f\rangle = 0, \forall f \in \mathcal{H}_N\}$. Since $\dot{\ell}_\beta(\theta_0; I_2, S_1) \in \mathcal{H}$, and $\mathcal{H}_N$ is a closed space, the projection of $\dot{\ell}_\beta(\theta_0; I_2, S_1)$ to $\mathcal{H}_N$ is well defined and identical to $U(\theta_0, h_0; I_2, S_1)$.

In what follows, we need an initial consistent estimator $\tilde{\theta}$ for estimating h_0. First, we obtain a weighted lasso estimator

$$\tilde{\theta} = \arg\min_\theta \left\{\ell(\theta; I_2, S_p) + \lambda\sum_{j=1}^{p} w_j|\theta_j|\right\}, \tag{2.12}$$

where $S_p = \{1, \cdots, p\}$, and $\ell(\theta; I, S)$ is defined in (2.9); $\lambda > 0$ is the penalty parameter, and w_i is a weight. In view of (2.11), we can estimate h_0 by its sample version and plug-in the weighted lasso estimator $\tilde{\theta}$ for θ. The resulting estimator has an explicit expression:

$$\tilde{h} = \Sigma_{22}^{-1}(\tilde{\theta}; I_2, S_1)\Sigma_{21}(\tilde{\theta}; I_2, S_1), \tag{2.13}$$

where Σ and $\tilde{\theta}$ are defined in (2.5) and (2.12), respectively. We construct an estimated projected partial score function

$$U(\beta, \tilde{\eta}, \tilde{h}; I_2, S_1) = \dot{\ell}_\beta(\beta, \tilde{\eta}; I_2, S_1) - \tilde{h}^{\mathrm{T}} \dot{\ell}_\eta(\beta, \tilde{\eta}; I_2, S_1),$$

where $\tilde{\eta}$ and $\tilde{h}$ are defined in (2.12) and (2.13), respectively. Note that $U(\beta, \tilde{\eta}, \tilde{h}; I_2, S_1)$ can be regarded as an approximately unbiased estimating function for β. We define an estimator $\hat{\beta}_1$ as the solution to $U(\beta, \tilde{\eta}, \tilde{h}; I_2, S_1) = 0$, which can be solved by the Newton-Raphson algorithm. In practice, we use the weighted lasso estimator $\tilde{\beta}$ in (2.12) as the initial value to start the algorithm.

Similarly, we first select variables using the second half of the data I_2 and denote the active set as $S_2 = \{j : \breve{\theta}_j \neq 0\}$. We then consider the sub-model based on the variables whose indices are in S_2,

$$d\Lambda_i(t) = Y_i(t) \exp\{\theta_{S_2}^{\mathrm{T}} V_{iS_2}(t)\} d\Lambda_0(t), \quad i \in I_1. \tag{2.14}$$

Based on (2.14), we obtain a projected partial score estimator $\hat{\beta}_2$ parallel to the estimation procedure for $\hat{\beta}_1$. The two-stage projection-based cross-validation (TPCV) estimator of β is defined as

$$\hat{\beta} = \frac{\hat{\beta}_1 + \hat{\beta}_2}{2}. \tag{2.15}$$

We use a diagram to illustrate the above two-stage estimation procedure in Figure 1. There are three attractive features of our method. First, it has effectively handled the uncertainty due to variable selection via cross-validation, because we use one half of data to do model selection, and fit the selected variables using another half of the data. In addition, the martingale theory is still applicable in deriving the theoretical properties, since the selection of active variables in Stage 1 is independent of the samples used in Stage 2. Second, the TPCV estimator $\hat{\beta}$ makes use of all the information in the data by using cross validation twice. Third, the estimated projection vector $\tilde{h}$ has an explicit expression and its dimension is much smaller than p. Therefore, our method is easy to implement for practical applications.

3 Theoretical results

3.1 Non-asymptotic oracle inequality

For the two-stage projection-based cross-validated estimation procedure, we adopt the weighted lasso to select active variables. Similar to Fang et al. (2017), we need to prove that the weighted lasso estimator $\tilde{\theta}$ has the convergence rate $\|\tilde{\theta}-\theta_0\|_1 = O_P(\lambda d_0)$, which ensures estimation consistency under some regularity conditions. In addition, the non-asymptotic oracle inequality for the weighted lasso has its independent interest. For example, the convergence rate for penalty-based estimator plays an important role in establishing distributional results for confidence interval and hypothesis testing in high-dimensional models (Zhang and Zhang, 2014; Fang et al., 2017; Ning and Liu, 2017; Neykov et al., 2018). Huang et al. (2013) and Kong and Nan (2014) considered oracle inequalities for the lasso in the high-dimensional Cox model. Zhang et al. (2017) studied oracle inequalities for weighted lasso estimator in the high-dimensional additive hazards model. Below, we present some general convergence results for weighted lasso estimator in the high-dimensional Cox model (2.1), which are suitable for the estimator given by (2.12) in Stage 2. Let $w \in \mathbb{R}^p$ be a (possibly estimated) weight vector with nonnegative elements w_j, $1 \le j \le p$ and $W = \mathrm{diag}\{w\}$. For any vector $a \in \mathbb{R}^p$ and matrix $A \in \mathbb{R}^{p\times p}$, we define $\|a\|_1 = \sum_{i=1}^p |a_i|$, $\|a\|_\infty = \max_{1\le i\le p} |a_i|$, and $\|A\|_\infty = \max_{1\le i,j\le p} |a_{ij}|$. The weighted L_1 loss function is

$$Q(\theta) = \ell(\theta) + \lambda \|W\theta\|_1,$$

where $\lambda \ge 0$ is a penalty parameter, and $\ell(\cdot)$ is defined in (2.2). The weighted lasso estimator is given by

$$\tilde{\theta} = \arg\min_{\theta} Q(\theta). \tag{3.1}$$

Note that if the variables in $S \subset \{1, \cdots, p\}$ are of primary interest, it is not necessary to penalize θ_S, which leads to "semi-penalized" estimators with $w_j = 0$ for $j \in S$ and $w_j \neq 0$ for $j \in S^c$. In what follows, it is sufficient to require $\min\{w_{S^c}\} > 0$. A vector $\tilde{\theta}$ is a global minimizer of (3.1) if and only if it satisfies the Karush-Kuhn-Tucker (KKT) conditions

$$\begin{cases} \dot{\ell}_j(\tilde{\theta}) = -\lambda w_j \mathrm{sgn}(\tilde{\theta}_j), & \text{if } \tilde{\theta}_j \neq 0, \\ |\dot{\ell}_j(\tilde{\theta})| \le \lambda w_j, & \text{if } \tilde{\theta}_j = 0. \end{cases} \tag{3.2}$$

Theorem 1 *Let $\tilde{\theta}$ be the weighted lasso estimator defined in (3.1), and $\tilde{R} = \tilde{\theta} - \theta_0$. Then the following inequality holds:*

$$(\lambda - z_0)\|W_{S^c}\tilde{R}_{S^c}\|_1 \le D(\tilde{\theta}, \theta_0) + (\lambda - z_0)\|W_{S^c}\tilde{R}_{S^c}\|_1 \le (\lambda\|w_S\|_\infty + z_0)\|\tilde{R}_S\|_1,$$

where $z_0 = \max\{\|\dot{\ell}(\theta_0)_S\|_\infty,\ \|W_{S^c}^{-1}\dot{\ell}(\theta_0)_{S^c}\|_\infty\}$, and $D(\tilde{\theta}, \theta) = (\tilde{\theta} - \theta)^{\mathrm{T}}\{\dot{\ell}(\tilde{\theta}) - \dot{\ell}(\theta)\}$ is the Bregman divergence. Furthermore, for any $\xi > \|w_S\|_\infty$, we have $\|W_{S^c}\tilde{R}_{S^c}\|_1 \le \xi\|\tilde{R}_S\|_1$ in the event $\{z_0 \le (\xi - \|w_S\|_\infty)/(\xi + 1)\lambda\}$, where W_{S^c} denotes the submatrix of W with components in S^c.

By Theorem 1, in the event $\{z_0 \le (\xi - \|w_S\|_\infty)/(\xi + 1)\lambda\}$, for any $\xi > \|w_S\|_\infty$, the estimation error $\tilde{\theta} - \theta_0$ belongs to the cone

$$G(\xi, S) = \{b \in \mathbb{R}^p : \|W_{S^c}b_{S^c}\|_1 \le \xi\|b_S\|_1\}. \tag{3.3}$$

To control estimation error of the weighted lasso in the Cox model, for the cone in (3.3) and the Hessian matrix $\ddot{\ell}(\theta_0)$, we use a compatibility factor as Huang et al. (2013),

$$\kappa(\xi, S) = \inf_{0 \ne b \in G(\xi, S)} \frac{d_0^{1/2}\{b^{\mathrm{T}}\ddot{\ell}(\theta_0)b\}^{1/2}}{\|b_S\|_1}.$$

In fact, the $\kappa(\xi, S)$ is a direct extension of the compatibility factor in linear models (van de Geer, 2007; van de Geer and Bühlmann, 2009; Huang and Zhang, 2012) by taking the Hessian of the log-partial likelihood at the true θ_0.

We make the following assumptions:

(C.1) $\int_0^\tau \lambda_0(t)dt < \infty$.

(C.2) The covariates are uniformly bounded: $\sup_{0 \le t \le \tau} \max_{1 \le i \le n} \max_{1 \le j \le p} |V_{ij}(t)| = O(1)$, where $V_{ij}(t)$ is the jth component of $V_i(t)$.

(C.3) The compatibility factor $\kappa(\xi, S)$ is strictly bounded away from zero.

Condition (C.1) has been similarly used by Andersen and Gill (1982) and Bradic et al. (2011) in their analysis of the partial likelihood estimator in the Cox model. Condition (C.2) was required by Huang et al. (2013) and Fang et al. (2017) in deriving the error bounds for the lasso in the Cox model, which is reasonable in most practical situations. Condition (C.3) was provided by Huang et al. (2013) under some regular assumptions.

The following result provides an upper bound of the estimation error for the weighted lasso estimator. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if $c \le a_n/b_n \le c'$ for some $c, c' > 0$.

Theorem 2 *Assume that Conditions (C.1) – (C.3) hold, and* $\lambda \asymp \sqrt{\{n^{-1}\log(p)\}}$. *Let* $\tilde{\theta}$ *be the weighted lasso estimator defined in (3.1),* K *is some positive constant and* $\rho = K\lambda d_0(1+\|w_S\|_\infty)(\xi+\min\{w_{S^c}\})^2/[4\min\{w_{S^c}\}\kappa^2(\xi,S)(\xi+1)]$ *with* $\rho \le 1/e$. *Then for* $\xi > |w_S|_\infty$, *in the event* $\{z_0 \le (\xi - \|w_S\|_\infty)/(\xi+1)\lambda\}$,

$$\|\tilde{\theta}-\theta_0\|_1 \le \frac{e^{\delta}\lambda d_0(1+\|w_S\|_\infty)(\xi+\min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}\kappa^2(\xi,S)(\xi+1)}, \tag{3.4}$$

where $\delta \le 1$ *is the smaller solution of* $\delta e^{-\delta} = \rho$.

By Huang et al. (2013) the term $\kappa(\xi, S)$ in (3.4) can be directly treated as a positive constant. Moreover, since the oracle inequality in Theorem 2 holds only within the event $\{z_0 \le (\xi - \|w_S\|_\infty)/(\xi+1)\lambda\}$, it is necessary to derive a probabilistic upper bound of z_0. It follows from Lemma 3.3 of Huang et al. (2013) that $P\{z_0 > Kx\} \le 2pe^{-nx^2/2}$. In order to better interpret the upper bound of the estimation error in (3.4), the conclusion can be simplified to the case that the convergence rate for the weighted lasso estimator is of order $O_P(\lambda d_0)$, which is used to establish the asymptotic properties in Theorem 3. Moreover, for the estimation error $\|\tilde{\theta}-\theta_0\|_1$ to be small with high probability, we need to ensure that $\lambda d_0 \to 0$ as $n \to \infty$. This requires the condition $p = \exp\{o(n/d_0^2)\}$. For bounded d_0, the dimension p can be as high as $e^{o(n)}$, which is in line with the lasso estimator of Huang et al. (2013).

3.2 Asymptotic normality

Let

$$\begin{aligned}\hat{\Sigma}_1 &= [\Sigma_{11}(\tilde{\theta}; I_2, S_1) - \Sigma_{12}(\tilde{\theta}; I_2, S_1)\Sigma_{22}^{-1}(\tilde{\theta}; I_2, S_1)\Sigma_{21}(\tilde{\theta}; I_2, S_1)]^{-1},\\ &= \Sigma_{\beta|\eta}^{-1}(\tilde{\theta}; I_2, S_1)\end{aligned} \tag{3.5}$$

and

$$\hat{\Sigma}_2 = [\Sigma_{11}(\tilde{\theta}; I_1, S_2) - \Sigma_{12}(\tilde{\theta}; I_1, S_2)\Sigma_{22}^{-1}(\tilde{\theta}; I_1, S_2)\Sigma_{21}(\tilde{\theta}; I_1, S_2)]^{-1},$$
$$= \Sigma_{\beta|\eta}^{-1}(\tilde{\theta}; I_1, S_2) \tag{3.6}$$

where Σ_{ij} is defined in (2.5) and $\tilde{\theta}$ is the weighted lasso estimate in Stage 2. The following theorem establishes the asymptotic normality of $\hat{\beta}$.

Theorem 3 *Suppose that Conditions (C.1)–(C.3) hold, $\lambda \asymp \sqrt{\{n^{-1}\log(p)\}}$ and $n^{-1/2}d_0\log(p) = o(1)$. Then as $n \to \infty$ we have*

$$\sqrt{n}\hat{\Sigma}^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, I_d),$$

where I_d is a $d \times d$ identity matrix, $\hat{\Sigma} = (\hat{\Sigma}_1 + \hat{\Sigma}_2)/2$ with $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ being defined in (3.5) and (3.6), respectively.

The conditions $\lambda \asymp \sqrt{\{n^{-1}\log(p)\}}$ and $n^{-1/2}d_0\log(p) = o(1)$ are also required in Fang et al. (2017) to ensure the asymptotic properties of their estimators. As an application, Theorem 3 provides a theoretical basis for conducting hypothesis test for a one-dimensional parameter $\beta_0 \in \mathbb{R}$ in the high-dimensional Cox model. Consider

$$H_0 : \beta_0 = 0 \quad \text{versus} \quad H_A : \beta_0 \neq 0, \tag{3.7}$$

we use the Wald statistic $T^w = \sqrt{n}\hat{\Sigma}^{-1/2}(\hat{\beta} - \beta_0)$, which is asymptotically distributed as $N(0, 1)$ under H_0. We reject H_0 if the p-value $P_w < 0.05$, where

$$P_w = 2\left\{1 - \Phi\left(\sqrt{n}\hat{\Sigma}^{-1/2}|\hat{\beta}|\right)\right\}, \tag{3.8}$$

and $\Phi(x)$ is the cumulative distribution function of $N(0, 1)$. To remove the potential influence of random splitting of samples, we repeat our proposed TPCV procedure B times. Denote the resulting p-values in (3.8) as $P_w^{(1)}, \cdots, P_w^{(B)}$. For the hypothesis test (3.7), we propose the following three decision rules:

- TPCV^1: Reject H_0 if $B^{-1}\sum_{b=1}^{B} P_w^{(b)} < 0.05$.
- TPCV^2: Reject H_0 if the median of $P_w^{(1)}, \cdots, P_w^{(B)}$ is smaller than 0.05.

• TPCV[3]: Reject H_0 if $B^{-1}\sum_{b=1}^{B} I(P_w^{(b)} < 0.05) > 0.5$, where $I(\cdot)$ is an indicator function. Of note, the TPCV[1] is coming from the mean of B p-values, and it may be affected by potential outliers. The TPCV[2] is based on the median, so it has the property of robustness. The TPCV[3] is from the idea of "majority voting", and it also owns the robustness. The performances of these decision rules will be evaluated via numerical simulation.

4 Numerical studies

In this section, we conduct simulation studies to evaluate the finite-sample performance of the proposed method. We also illustrate the application of the proposed method on a breast cancer gene expression dataset.

4.1 Simulation studies

We generate failure times $(T_1, \cdots, T_n)$ from the Cox model with an exponential hazards function $\exp(\theta_0^{\mathrm{T}} V_i)$, where $\theta_0 = (\beta_0, \eta_0^{\mathrm{T}})^{\mathrm{T}}$, and $V_i = (V_{i1}, \cdots, V_{ip})^{\mathrm{T}}$, $i = 1, \cdots, n$. First, we assume that the parameter of interest β_0 is one-dimensional and the nuisance parameter vector η_0 is chosen as follows:

- Case I: $\eta_0 = (\underbrace{1, \cdots, 1}_{10\ times}, 0, \cdots, 0)^{\mathrm{T}}$,
- Case II: $\eta_0 = (\underbrace{1, \cdots, 1}_{15\ times}, 0, \cdots, 0)^{\mathrm{T}}$,

where the dimension p =500 and 1000, respectively. The covariates $V_{ij} = \min(Z_{ij}, 10^3)$, and $Z_i = (Z_{i1}, \cdots, Z_{ip})'$ are generated from multivariate normal distribution with mean zero and covariance matrix $\Sigma_Z = (0.15^{|i-j|})$. The censoring times C_i are generated from the uniform distribution on [0, 5], which leads to about 40% censoring rate. The results presented below are based on 200 replications with sample size $n = 300$.

We use the R package `glmnet` (Simon et al., 2011) to compute the weighted lasso estimator. The tuning parameter λ is determined by 10-folds cross-validation. For comparison, we consider the decorrelated score (DS) method in Eq. (3.8) of Fang et al. (2017). The DS method was implemented with R codes at *http://www.personal.psu.edu/xxf13/Code/CoxHDInference.R*.

As suggested by a reviewer, we also consider the two-stage projection-based (TP) method using the whole sample, i.e., $I_1 = I_2 = I$ in Stages 1 and 2 of our method. In Tables 1 and 2, we report the estimated bias (Bias) given by the sample mean of the estimates minus the true value, the sample mean of the estimated standard errors (ESE), the sample standard error (SSE) of the estimates, and the empirical coverage probability of the 95% confidence interval (CP). Tables 1 and 2 indicate that the proposed TPCV estimator is unbiased, and its ESE is close to SSE. The DS method leads to a biased estimator, especially for larger parameters ($\beta = 0.5$). Moreover, the ESE and SSE do not agree well for TP method, which uses the same dataset twice in Stages 1 and 2. Hence, the overall performance of TPCV is better than those of the DS and TP methods.

Tables 3 and 4 present the sizes and powers on testing $H_0: \beta_0 = 0$ vs. $H_A: \beta_0 \neq 0$ under Cases I and II, respectively. We consider the performances of our methods (TPCV^1, TPCV^2 and TPCV^3), the DS and TP methods. Due to the computation burden, we set the times of splitting as $B = 50$ (the conclusions are similar for a larger B). It can be seen from the tables that TPCV^2 and TPCV^3 have outstanding advantages over TPCV^1. One possible explanation is that TPCV^1 is based on the mean of p-values, which could be affected by potential outliers. The TPCV^1 still performs slightly better than the DS method. In brief, the proposed TPCV^2 and TPCV^3 methods are more powerful than the DS method when the coefficients of predictors are high-dimensional and not very sparse. Moreover, the TP method has an inflated type I error, which could leads to higher false positive rate than the prespecified nominal level.

We conduct the second simulation to assess the performance of the proposed estimation method for a two-dimensional vector $\beta_0 = (\beta_{10}, \beta_{20})'$. The data are generated as in the first simulation, except that $\beta_0 = (0.15, 0.30)'$. In Table 5, we report the Bias, ESE, SSE and CP for the estimates of β_{10} and β_{20}, respectively. It can be seen that the proposed method works well for estimating multiple parameters of interest.

4.2 Breast cancer gene expression data

Breast cancer is one of the most commonly diagnosed malignancy for women. Biomedical studies indicate that genomic measurements may have independent predictive power for breast cancer prognosis (van't Veer et al., 2002; Cheang et al., 2008). We apply the proposed method to a publicly available breast cancer gene expression data set (van't Veer et al., 2002). The data set consists of 295 tumor samples of breast cancer patients with expression measurements for 4919 genes. Among these patients, 79 died during the follow-up time and the remaining 216 observations are censored. We define the event time as the time from diagnosis to death. We first use the marginal Cox model to select top 500 genes, which are used as the covariates $V_i = (V_{i1}, \cdots, V_{ip})'$ in model (2.1) with $p = 500$.

We first take V_{i1} as the covariate of interest, and the remaining covariates $V_{i2}, \cdots, V_{ip}$ are regared as confounding variables. We apply the TPCV method to make inference on the first parameter of interest in the Cox model. We repeat this process for the other covariates and conduct inference about each coefficient using the proposed method. In Table 6, we report the estimated coefficient (Est), the corresponding standard error (SE), the 95% confidence interval (CI) on five genes with Bonferroni adjusted p-value $P_{adj} < 0.05$. Among these genes, the NM_001168 was shown to be biologically related to breast cancer (Goeman, 2010), which supports the effectiveness of our proposed approach.

5 Concluding remarks

We have considered the problem of statistical inference about a low-dimensional parameter of interest in the Cox model when the number of nuisance parameters is possibly greater than the sample size. A two-stage projection-based cross-validated estimation approach was proposed. Simulations and a gene-expression data example from a breast cancer study were used to illustrate the proposed method. Of note, we actually do not know beforehand which is the parameter of interest in many practical applications. For example, in the breast cancer gene expression data example, we are interested in finding the genes that are related to cancer in clinical research (van't Veer et al., 2002). In this setting, we need to conduct statistical inference about all the regression coefficients in the model. We can apply the proposed method to each coefficient in turn. This approach was also adopted in the real data analysis of Fang et al. (2017).

There are several questions that are of interest to be considered in the future. First, the weighted lasso estimator in our proposed method can be replaced with the SCAD or the minimax concave penalty estimator (Zhang, 2010). This usually involves a high dimensional nonconvex optimization problem and is more difficult to implement. The theoretical and computational aspects of using a concave penalty deserve further study. Second, a theoretical analysis of the test in (3.7) is desirable, such as its local asymptotic power behavior. Third, the proposed method can be extended to other survival models, such as the additive hazards model (Lin and Ying, 1994) and the accelerated failure time model (Huang et al., 2006).

Acknowledgement

The authors would like to thank the Editor, Professor Håkon K. Gjessing, an Associate Editor and two anonymous reviewers for their constructive and insightful comments that greatly improved the manuscript. We are grateful to Ethan X. Fang and Han Liu for sharing with us the code in the simulation. The work of Huang is supported in part by the NSF grant DMS-1916199. The work of Sun is supported in part by the National Natural Science Foundation of China (Grant Nos. 11771431, 11690015 and 11926341) and Key Laboratory of RCSDS, CAS (No. 2008DP173182).

6 Appendix

Proof of Theorem 1. Because $\ell(\theta)$ is a convex function, it follows that $D(\tilde{\theta}, \theta_0) = \tilde{R}^{\mathrm{T}}\{\dot{\ell}(\theta_0 + \tilde{R}) - \dot{\ell}(\theta_0))\} \geq 0$, and the first inequality holds. Note that $\tilde{R}_j = \tilde{\theta}_j$ for $j \in S^c$. By the KKT condition (3.2), we have

$$
\begin{aligned}
& \tilde{R}^{\mathrm{T}}\{\dot{\ell}(\theta_0 + \tilde{R}) - \dot{\ell}(\theta_0)\} \\
= \; & \sum_{j \in S^c} \tilde{R}_j \dot{\ell}_j(\theta_0 + \tilde{R}) + \sum_{j \in S} \tilde{R}_j \dot{\ell}_j(\theta_0 + \tilde{R}) + \tilde{R}^{\mathrm{T}}(-\dot{\ell}(\theta_0)) \\
\leq \; & \sum_{j \in S^c} \tilde{\theta}_j\big(-\lambda w_j \mathrm{sgn}(\tilde{\theta}_j)\big) + \sum_{j \in S} |\tilde{R}_j| \lambda w_j + \tilde{R}_{S^c}^{\mathrm{T}}(-\dot{\ell}_{S^c}(\theta_0)) + \tilde{R}_S^{\mathrm{T}}(-\dot{\ell}_S(\theta_0)) \\
= \; & -\lambda \|W_{S^c}\tilde{R}_{S^c}\|_1 + \lambda \|W_S \tilde{R}_S\|_1 + (W_{S^c}\tilde{R}_{S^c})^{\mathrm{T}}\big(-W_{S^c}^{-1}\dot{\ell}_{S^c}(\theta_0)\big) + \tilde{R}_S^{\mathrm{T}}(-\dot{\ell}_S(\theta_0)) \\
\leq \; & (z_0 - \lambda)\|W_{S^c}\tilde{R}_{S^c}\|_1 + (z_0 + \lambda\|w_S\|_\infty)\|\tilde{R}_S\|_1.
\end{aligned}
$$

Due to $\tilde{R}_j = \tilde{\theta}_j - \theta_{0j} = 0$ when $j \in S^c$ and $\tilde{\theta}_j = 0$, the first inequality above shows that $\dot{\ell}_j(\theta_0 + \tilde{R}) = -\lambda w_j \text{sgn}(\tilde{\theta}_j)$ only in the set $S^c \cap \{j : \tilde{\theta}_j \neq 0\}$. This completes the proof. $\square$

Proof of Theorem 2. Let $\tilde{R} = \tilde{\theta} - \theta_0 \neq 0$ and $b = \tilde{R}/\|\tilde{R}\|_1$. It follows from the convexity of $\ell(\beta_0 + xb)$ (as a function of x) and Theorem 1 that in the event $\{z_0 \leq (\xi - \|w_S\|_\infty)/(\xi+1)\lambda\}$,

$$b^{\mathrm{T}}\{\dot{\ell}(\theta_0 + xb) - \dot{\ell}(\theta_0)\} + \frac{\lambda(1 + \|w_S\|_\infty)}{\xi + 1}\|W_{S^c} b_{S^c}\|_1 \leq \frac{\xi\lambda(1 + \|w_S\|_\infty)}{\xi + 1}\|b_S\|_1, \tag{6.1}$$

where $x \in [0, \|\tilde{R}\|_1]$ and $b \in G(\xi, S)$. For any nonnegative x satisfying (6.1), due to $\delta_{xb} = \max_{0 \leq s \leq \tau} \max_{i,j} |xb^{\mathrm{T}} V_i(s) - xb^{\mathrm{T}} V_j(s)| \leq Kx\|b\|_1 = Kx$ and Lemma 3.2 in Huang, et al. (2013),

$$xb^{\mathrm{T}}\{\dot{\ell}(\theta_0 + xb) - \dot{\ell}(\theta_0)\} \geq x^2 \exp(-\delta_{xb}) b^{\mathrm{T}} \ddot{\ell}(\theta_0) b \geq x^2 \exp(-Kx) b^{\mathrm{T}} \ddot{\ell}(\theta_0) b. \tag{6.2}$$

The (6.2) together with $\kappa(\xi, S)$ and (6.1) yields

$$\begin{aligned}
xe^{-Kx}\kappa^2(\xi, S)\|b_S\|_1^2/d_0 &\leq xe^{-Kx} b^{\mathrm{T}} \ddot{\ell}(\theta_0) b \\
&\leq \frac{\xi\lambda(1 + \|w_S\|_\infty)}{\xi + 1}\|b_S\|_1 - \frac{\lambda(1 + \|w_S\|_\infty)}{\xi + 1}\|W_{S^c} b_{S^c}\|_1 \\
&\leq \frac{\xi\lambda(1 + \|w_S\|_\infty)}{\xi + 1}\|b_S\|_1 - \frac{\lambda(1 + \|w_S\|_\infty)}{\xi + 1}\|b_{S^c}\|_1 \min\{w_{S^c}\} \\
&= \frac{\lambda(1 + \|w_S\|_\infty)(\xi + \min\{w_{S^c}\})}{\xi + 1}\|b_S\|_1 - \frac{\lambda \min\{w_{S^c}\}(1 + \|w_S\|_\infty)}{\xi + 1} \\
&\leq \frac{\lambda(1 + \|w_S\|_\infty)(\xi + \min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}(\xi + 1)}\|b_S\|_1^2.
\end{aligned}$$

For any nonnegative x satisfying (6.1), we have

$$Kx\exp(-Kx) \leq \frac{K\lambda d_0(1 + \|w_S\|_\infty)(\xi + \min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}\kappa^2(\xi, S)(\xi + 1)} = \rho. \tag{6.3}$$

Notice that $b^{\mathrm{T}}\{\dot{\ell}(\theta_0 + xb) - \dot{\ell}(\theta_0)\}$ is an increasing function of x. All nonnegative x satisfying (6.1) are a closed interval $[0, x^*]$ for some $x^* > 0$. By (6.3), we know that $Kx^* \leq \delta$, where δ is the smallest solution of $\delta e^{-\delta} = \rho$. Thus,

$$\|\tilde{R}\|_1 \leq x^* \leq \frac{\delta}{K} = \frac{e^\delta \lambda d_0(1 + \|w_S\|_\infty)(\xi + \min\{w_{S^c}\})^2}{4\min\{w_{S^c}\}\kappa^2(\xi, S)(\xi + 1)}.$$

This completes the proof. $\square$

Proof of Theorem 3. The proof consists of three steps.

Step 1: Based on the second half sample I_2 and the active variables in S_1, we fit a sub-model as

$$d\Lambda_i(t) = Y_i(t)\exp\{\theta_{S_1}^{\mathrm{T}} V_{iS_1}(t)\} d\Lambda_0(t), \quad i \in I_2, \tag{6.4}$$

where S_1 is the selected active index set using the first half sample I_1. The projected partial score function for β is

$$\begin{aligned} U(\theta_0, h_0; I_2, S_1) &= \dot{\ell}_\beta(\theta_0; I_2, S_1) - h_0^{\mathrm{T}} \dot{\ell}_\eta(\theta_0; I_2, S_1) \\ &= (1, -h_0^{\mathrm{T}})^{\mathrm{T}} \dot{\ell}(\theta_0; I_2, S_1), \end{aligned}$$

where

$$\dot{\ell}(\theta_0; I_2, S_1) = -\frac{1}{|I_2|} \sum_{i \in I_2} \int_0^\tau \{V_{iS_1}(t) - \bar{V}(t, \theta; I_2, S_1)\} dM_i(t),$$

and $M_i(t) = N_i(t) - \int_0^t Y_i(u)\exp\{\theta_{S_1}^{\mathrm{T}} V_{iS_1}(u)\} d\Lambda_0(u)$ are martingales with $\langle M_i, M_i\rangle(t) = \int_0^t Y_i(u)\exp\{\theta_{S_1}^{\mathrm{T}} V_{iS_1}(u)\} d\Lambda_0(u)$, and $\langle M_i, M_j\rangle = 0$ for $i \neq j$. The martingale theory is applicable to model (6.4), due to the selection of S_1 is independent of I_2. By Lemma G.3 of Fang et al. (2017), we can obtain

$$\sqrt{|I_2|} \cdot \{\mathbf{v}^{\mathrm{T}} \Sigma^*(\theta_0; S_1)\mathbf{v}\}^{-1/2} \mathbf{v}^{\mathrm{T}} \dot{\ell}(\theta_0; I_2, S_1) \xrightarrow{\mathcal{D}} N(0, I_d), \tag{6.5}$$

where $\mathbf{v}$ is similarly given in Lemma G.3 of Fang et al. (2017). Note that $|I_2| = n/2$, $U(\theta_0, h_0; I_2, S_1) = \mathbf{v}^{\mathrm{T}} \dot{\ell}(\theta_0; I_2, S_1)$ and $\mathbf{v}^{\mathrm{T}} \Sigma^*(\theta_0; S_1)\mathbf{v} = \Sigma^*_{\beta|\eta}(\theta_0; S_1)$. Then,

$$\sqrt{\frac{n}{2}} \cdot \{\Sigma^*_{\beta|\eta}(\theta_0; S_1)\}^{-1/2} U(\theta_0; I_2, S_1) \xrightarrow{\mathcal{D}} N(0, I_d). \tag{6.6}$$

Step 2. It follows from the mean value theorem that

$$U(\hat{\beta}_1, \tilde{\eta}, \tilde{h}; I_2, S_1) = U(\beta_0, \tilde{\eta}, \tilde{h}; I_2, S_1) + \dot{U}_\beta(\bar{\beta}, \tilde{\eta}, \tilde{h}; I_2, S_1)(\hat{\beta}_1 - \beta_0),$$

where $\bar{\beta}$ is on a line segment between $\hat{\beta}_1$ and β_0. Because $U(\hat{\beta}_1, \tilde{\eta}, \tilde{h}; I_2, S_1) = 0$,

$$\hat{\beta}_1 - \beta_0 = -\dot{U}_\beta^{-1}(\bar{\beta}, \tilde{\eta}, \tilde{h}; I_2, S_1) U(\beta_0, \tilde{\eta}, \tilde{h}; I_2, S_1) \tag{6.7}$$

$$= \underbrace{-\Sigma_{\beta|\eta}^{*-1}(\theta_0; S_1) U(\beta_0, \tilde{\eta}, \tilde{h}; I_2, S_1)}_{R_1} + \underbrace{U(\beta_0, \tilde{\eta}, \tilde{h}; I_2, S_1)[\Sigma_{\beta|\eta}^{*-1}(\theta_0; S_1) - \dot{U}_{\beta}^{-1}(\bar{\beta}, \tilde{\eta}, \tilde{h}; I_2, S_1)]}_{R_2}$$

To derive the asymptotic distribution of $\hat{\beta}_1$, we start with decomposing $U(\beta_0, \tilde{\eta}, \tilde{h}; I_2, S_1)$ as

$$\begin{aligned}
U(\beta_0, \tilde{\eta}, \tilde{h}; I_2, S_1) &= \dot{\ell}_{\beta}(\beta_0, \tilde{\eta}; I_2, S_1) - \tilde{h}^{\mathrm{T}} \dot{\ell}_{\eta}(\beta_0, \tilde{\eta}; I_2, S_1) \\
&= \dot{\ell}_{\beta}(\theta_0; I_2, S_1) + (\tilde{\eta} - \eta_0)^{\mathrm{T}} \ddot{\ell}_{\beta\eta}(\beta_0, \bar{\eta}; I_2, S_1) - \tilde{h}^{\mathrm{T}} \dot{\ell}_{\eta}(\theta_0; I_2, S_1) \\
&\quad - \tilde{h}^{\mathrm{T}} \ddot{\ell}_{\eta\eta}(\beta_0, \bar{\bar{\eta}}; I_2, S_1)(\tilde{\eta} - \eta_0) \\
&= \dot{\ell}_{\beta}(\theta_0; I_2, S_1) - h_0^{\mathrm{T}} \dot{\ell}_{\eta}(\theta_0; I_2, S_1) + \underbrace{(h_0 - \tilde{h})^{\mathrm{T}} \dot{\ell}_{\eta}(\theta_0; I_2, S_1)}_{E_1} \\
&\quad + \underbrace{(\tilde{\eta} - \eta_0)^{\mathrm{T}} \{\ddot{\ell}_{\beta\eta}(\beta_0, \bar{\eta}; I_2, S_1) - \ddot{\ell}_{\eta\eta}(\beta_0, \bar{\bar{\eta}}; I_2, S_1) \tilde{h}\}}_{E_2} \\
&= U(\theta_0; I_2, S_1) + E_1 + E_2, \qquad (6.8)
\end{aligned}$$

where $\bar{\eta} = \eta_0 + u(\tilde{\eta} - \eta_0)$ and $\bar{\bar{\eta}} = \eta_0 + u'(\tilde{\eta} - \eta_0)$ for some $u, u' \in [0, 1]$. By Lemmas 1 and 4 of Fang et al. (2017), together with $\|\tilde{\eta} - \eta_0\|_1 = O_p(\lambda d_0)$, we get

$$\|\tilde{h} - h_0\|_1 = O_P\left(d_0 \sqrt{\frac{\log(p)}{n}}\right) \quad \text{and} \quad \|\dot{\ell}_{\eta}(\theta_0; I_2, S_1)\|_{\infty} = O_P\left(\sqrt{\frac{\log(p)}{n}}\right).$$

Hence $E_1 = O_P\{n^{-1} d_0 \log(p)\}$. For the term E_2,

$$E_2 = \underbrace{(\tilde{\eta} - \eta_0)^{\mathrm{T}} \{\ddot{\ell}_{\beta\eta}(\beta_0, \bar{\eta}; I_2, S_1) - \ddot{\ell}_{\eta\eta}(\beta_0, \bar{\bar{\eta}}; I_2, S_1) h_0\}}_{E_{21}} + \underbrace{(h_0 - \tilde{h})^{\mathrm{T}} \ddot{\ell}_{\eta\eta}(\beta_0, \bar{\bar{\eta}}; I_2, S_1)(\tilde{\eta} - \eta_0)}_{E_{22}}.$$

From the Lemma E.4 in Fang et al. (2017) and $\|\tilde{\eta} - \eta_0\|_1 = O_p(\lambda d_0)$, we know that

$$E_{21} = (\tilde{\eta} - \eta_0)^{\mathrm{T}} \ddot{\ell}_{\beta\eta}(\beta_0, \bar{\eta}; I_2, S_1) - (\tilde{\eta} - \eta_0)^{\mathrm{T}} \ddot{\ell}_{\eta\eta}(\beta_0, \bar{\bar{\eta}}; I_2, S_1) h_0 = O_P\{n^{-1} d_0 \log(p)\}. \qquad (6.9)$$

By the Cauchy-Schwarz inequality and (A.6) of Fang et al. (2017),

$$\begin{aligned}
|E_{22}| &\leq \frac{1}{2}(h_0 - \tilde{h})^{\mathrm{T}} \ddot{\ell}_{\eta\eta}(\beta_0, \bar{\bar{\eta}}; I_2, S_1)(h_0 - \tilde{h}) + \frac{1}{2}(\tilde{\eta} - \eta_0)^{\mathrm{T}} \ddot{\ell}_{\eta\eta}(\beta_0, \bar{\bar{\eta}}; I_2, S_1)(\tilde{\eta} - \eta_0) \\
&= O_P\{n^{-1} d_0 \log(p)\}. \qquad (6.10)
\end{aligned}$$

It follows from (6.9) and (6.10) that $E_2 = O_P\{n^{-1} d_0 \log(p)\}$. From (6.8),

$$U(\beta_0, \tilde{\eta}, \tilde{h}; I_2, S_1) = U(\theta_0, h_0; I_2, S_1) + O_P\{n^{-1} d_0 \log(p)\}$$

$$= U(\theta_0, h_0; I_2, S_1) + o_P(n^{-1/2}), \tag{6.11}$$

where the last equality holds by the assumption that $n^{-1/2}d_0 \log(p) = o(1)$. Thus,

$$R_1 = -\Sigma^{*-1}_{\beta|\eta}(\theta_0; S_1)U(\theta_0, h_0; I_2, S_1) + o_P(n^{-1/2}).$$

By (6.7), (6.11) and Lemma 2 of Fang et al. (2017), together with the assumption $n^{-1/2}d_0 \log(p) = o(1)$, we can deduce that $R_2 = o_P(n^{-1/2})$. Note that

$$\hat{\beta}_1 - \beta_0 = -\Sigma^{*-1}_{\beta|\eta}(\theta_0; S_1)U(\theta_0, h_0; I_2, S_1) + o_P(n^{-1/2}). \tag{6.12}$$

Then,

$$\sqrt{\frac{n}{2}} \cdot \Sigma^{*1/2}_{\beta|\eta}(\theta_0; S_1)(\hat{\beta}_1 - \beta_0) = \sqrt{\frac{n}{2}} \cdot \{\Sigma^{*}_{\beta|\eta}(\theta_0; S_1)\}^{-1/2}U(\theta_0; I_2, S_1) + o_P(1).$$

Based on (6.6), together with the Slutsky's theorem, we obtain

$$\sqrt{\frac{n}{2}} \cdot \hat{\Sigma}_1^{-1/2}(\hat{\beta}_1 - \beta_0) \xrightarrow{\mathcal{D}} N(0, I_d), \tag{6.13}$$

where $\hat{\Sigma}_1$ is defined in (3.5).

Step 3. Based on the first half sample I_1 and the active variables with the index set S_2, we can fit a sub-model as

$$d\Lambda_i(t) = Y_i(t) \exp\{\theta^{\mathrm{T}}_{S_2} V_{iS_2}(t)\} d\Lambda_0(t), \quad i \in I_1.$$

Following similar arguments as in Steps 1 and 2, we have

$$\hat{\beta}_2 - \beta_0 = -\Sigma^{*-1}_{\beta|\eta}(\theta_0; S_2)U(\theta_0, h_0; I_1, S_2) + o_P(n^{-1/2}), \tag{6.14}$$

and

$$\sqrt{\frac{n}{2}} \cdot \hat{\Sigma}_2^{-1/2}(\hat{\beta}_2 - \beta_0) \xrightarrow{\mathcal{D}} N(0, I_d), \tag{6.15}$$

where $\hat{\Sigma}_2$ is defined in (3.6).

Note that the selections of S_1 and S_2 are determined by two independent datasets in I_1 and I_2, respectively. Then, $\Sigma^{*}_{\beta|\eta}(\theta_0; S_1)$ and $\Sigma^{*}_{\beta|\eta}(\theta_0; S_2)$ are independent. From (2.10), we know that $U(\theta_0, h_0; I_1, S_2)$ and $U(\theta_0, h_0; I_2, S_1)$ are formulated with two independent datasets in I_1 and I_2, respectively. Under mild conditions on the weights for the weighted Lasso to have the oracle property, for example, taking the weights to be the inverse of the an initial Lasso estimate (Zhang and Lu, 2007 and Huang and Zhang, 2012), we have $P(S_1 = S_0) \to 1$ and $P(S_2 = S_0) \to 1$, where S_1 and S_2 are given in Stage 1 of our method, and $S_0 = \{1, \cdots, d\} \cup \{j : \theta_{j0} \neq 0, j = d+1, \cdots, p\}$. The $U(\theta_0, h_0; I_1, S_2)$ and $U(\theta_0, h_0; I_2, S_1)$ are two asymptotically independent terms. Moreover, $\Sigma^*_{\beta|\eta}(\theta_0; S_1)$ is independent of $U(\theta_0, h_0; I_1, S_2)$, and $\Sigma^*_{\beta|\eta}(\theta_0; S_2)$ is independent of $U(\theta_0, h_0; I_2, S_1)$. In view of (6.12) and (6.14), $\hat{\beta}_1$ and $\hat{\beta}_2$ can be regarded as asymptotically independent. Thus, it follows from (2.15), (6.13) and (6.15) that

$$\sqrt{n}\hat{\Sigma}^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, I_d).$$

This completes the proof. $\square$

References

[1] Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, **10**, 1100-1120.

[2] Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-Dimensionality. *The Annals of Statistics*, **39**, 3092-3120.

[3] Cheang, M., van de Rijn, M. and Nielsen, T. (2008). Gene expression profiling of breast cancer. *Annual Review of Pathology: Mechanisms of Disease*, **3**, 67-97.

[4] Cox, D. R. (1972). Regression models and life-tables (with discussions). *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.

[5] Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

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

[7] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74-99.

[8] Fang, E., Ning, Y. and Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society, Series B*, **79**, 1415-1437.

[9] Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

[10] Goeman, J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, **52**, 70-84.

[11] Heikkinen, K., Rapakko, K., Karppinen, S. et al. (2006). RAD50 and NBS1 are breast cancer susceptibility genes associated with genomic instability. *Carcinogenesis*, **27**, 1593-1599.

[12] Huang, J., Ma, S. and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, **62**, 813-820.

[13] Huang, J. and Zhang, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *Journal of Machine Learning Research*, **13**, 1839-1864.

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

[15] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, **15**, 2869-2909.

[16] Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Second edition, John Wiley: New York.

[17] Kong, S. and Nan, B. (2014). Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Statistica Sinica*, **24**, 25-42.

[18] Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61-71.

[19] Ma, S., Huang, J., Wei, F., Xie, Y. and Fang, K. (2011). Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in Medicine*, **30**, 3361-3371.

[20] Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, **104**, 1671-1681.

[21] Neykov, M., Ning, Y., Liu, J. and Liu, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, **33**, 427-443.

[22] Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, **45**, 158-195.

[23] R Development Core Team (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0, http://www.R-project.org.

[24] Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, **39**, 1-13.

[25] Small, C. and McLeish, D. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. Wiley-Interscience.

[26] Sotiriou, C., Neo, S., McShane, L., Korn, E., Long, P., Jazaeri, A., Martiat, P., Fox, S., Harris, A. and Liu, E. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population based study. *Proceedings of the National Academy of Science*, **100**, 10393-10398.

[27] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267-288

[28] Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

[29] Tommiska, J., Seal, S., Renwick, A. et al. (2006). Evaluation of RAD50 in familial breast cancer predisposition. *International Journal of Cancer*, **118**, 2911-2916.

[30] van de Geer, S. (2007). On non-asymptotic bounds for estimation in generalized linear models with highly correlated design. *Lecture Notes-Monograph Series*, **55**, 121-134.

[31] van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, **3**, 1360-1392

[32] van de Geer, S., Bühlmann, P. and Ritov, Y. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**, 1166-1202.

[33] van't Veer, L., Dai, H., van de Vijver, M., He, Y., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530-536.

[34] Wasserman, L., and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics*, **37**, 2178-2201.

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

[36] Zhang, H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691-703.

[37] Zhang, H., Sun, L., Zhou, Y. and Huang, J. (2017). Oracle inequalities and selection consistency for weighted lasso in high-dimensional additive hazards model. *Statistica Sinica*, **27**, 1903-1920.

[38] Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series* B, **76**, 217-242.

[39] Zhong, P., Hu, T. and Li, J. (2015). Tests for coefficients in high-dimensional additive hazard models. *Scandinavian Journal of Statistics*, **42**, 649-664.

[40] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429.

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

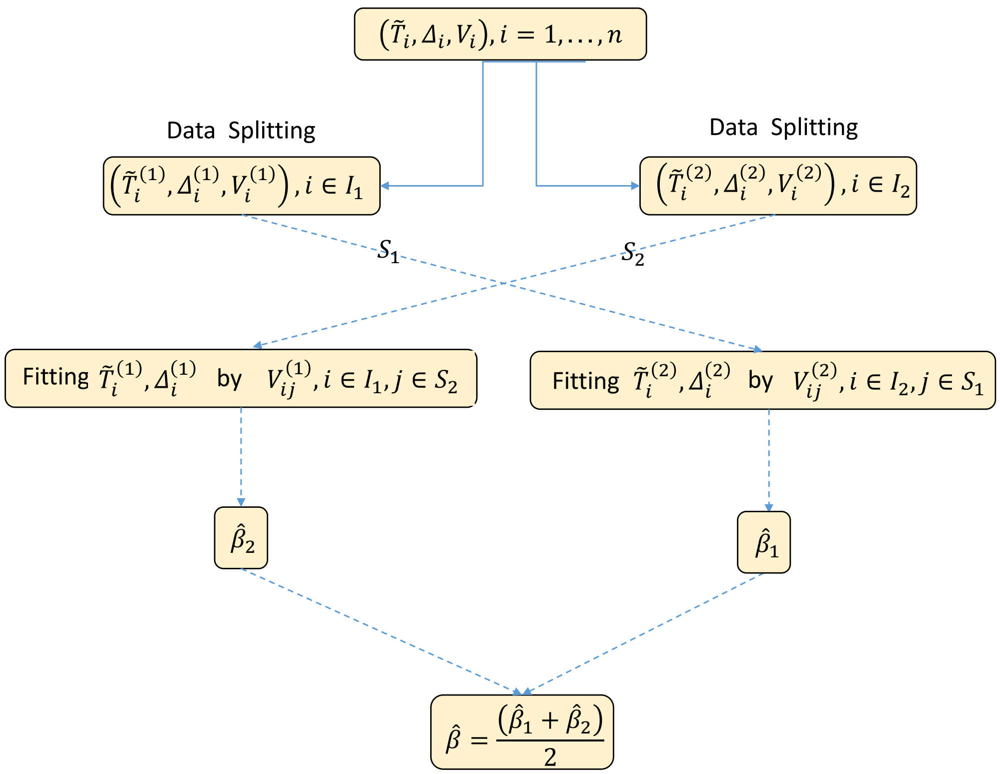

Figure 1. A scenario of two-stage projection-based cross-validation procedure.

Table 1. Estimation results on the parameter of interest β with Case I[†].

		$\beta = 0$				$\beta = 0.5$			
	Methods	Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
$p = 500$	TPCV	-0.0047	0.0866	0.0885	0.940	0.0232	0.0909	0.0930	0.940
	DS	0.0098	0.0749	0.0503	0.990	-0.1475	0.0753	0.0771	0.520
	TP	-0.0023	0.0799	0.0839	0.935	0.0201	0.0839	0.0862	0.950
$p = 1000$	TPCV	0.0031	0.0873	0.0994	0.925	0.0183	0.0924	0.1032	0.905
	DS	0.0169	0.0746	0.0482	0.990	-0.1727	0.0754	0.0736	0.360
	TP	0.0010	0.0803	0.0902	0.915	0.0229	0.0849	0.0980	0.895

† TPCV denotes our proposed method; "DS" denotes the decorrelated score method in Fang et al. (2017); "TP" denotes the two-stage projection-based method with the whole sample.

Table 2. Estimation results on the parameter of interest β with Case II[†].

	Methods	$\beta = 0$				$\beta = 0.5$			
		Bias	ESE	SSE	CP	Bias	ESE	SSE	CP
$p = 500$	TPCV	0.0028	0.0941	0.1035	0.945	0.0147	0.0989	0.0974	0.955
	DS	0.0134	0.0757	0.0492	0.990	-0.1898	0.0771	0.0719	0.335
	TP	0.0067	0.0835	0.0910	0.950	0.0189	0.0883	0.0947	0.945
$p = 1000$	TPCV	0.0065	0.0966	0.1059	0.905	0.0164	0.1009	0.1075	0.935
	DS	0.0135	0.0756	0.0414	1	-0.2057	0.0767	0.0678	0.200
	TP	0.0078	0.0842	0.0907	0.920	0.0303	0.0890	0.0988	0.935

† TPCV denotes our proposed method; "DS" denotes the decorrelated score method in Fang et al. (2017); "TP" denotes the two-stage projection-based method with the whole sample.

Table 3. Size/Power results with significance level $\alpha = 0.05$ (Case I)†.

	$p = 500$					$p = 1000$				
β	TPCV^1	TPCV^2	TPCV^3	DS	TP	TPCV^1	TPCV^2	TPCV^3	DS	TP
0	0.005	0.035	0.035	0	0.070	0.030	0.050	0.050	0.035	0.085
0.1	0.125	0.140	0.140	0.095	0.215	0.130	0.165	0.165	0.065	0.240
0.2	0.565	0.610	0.610	0.320	0.670	0.595	0.670	0.670	0.345	0.725
0.3	0.910	0.945	0.945	0.760	0.955	0.935	0.945	0.940	0.720	0.965
0.4	0.990	0.995	0.995	0.970	0.995	0.990	0.995	0.995	0.945	1
0.5	1	1	1	1	1	1	1	1	1	1

$\dagger$ TPCV$^{\mathrm{k}}$ denotes our proposed method, for $k = 1, 2, 3$; "DS" denotes the decorrelated score method in Fang et al. (2017); "TP" denotes the two-stage projection-based method with the whole sample.

Table 4. Size/Power results with significance level $\alpha = 0.05$ (Case II)†.

	$p = 500$					$p = 1000$				
β	TPCV^1	TPCV^2	TPCV^3	DS	TP	TPCV^1	TPCV^2	TPCV^3	DS	TP
0	0.025	0.040	0.035	0.020	0.060	0.025	0.055	0.055	0.015	0.085
0.1	0.115	0.215	0.215	0.075	0.305	0.115	0.190	0.190	0.060	0.300
0.2	0.565	0.685	0.670	0.350	0.735	0.515	0.625	0.620	0.255	0.685
0.3	0.825	0.890	0.890	0.650	0.925	0.855	0.920	0.920	0.550	0.945
0.4	0.985	0.990	0.990	0.930	0.990	0.980	1	1	0.870	0.990
0.5	1	1	1	1	1	1	1	1	0.985	1

† TPCV$^{\mathrm{k}}$ denotes our proposed method, for $k = 1, 2, 3$; "DS" denotes the decorrelated score method in Fang et al. (2017); "TP" denotes the two-stage projection-based method with the whole sample.

Table 5. Estimation results on the parameters of interest $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}}$.

		Bias		ESE		SSE		CP	
	p	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
Case I	p=500	0.0326	0.0312	0.0870	0.0891	0.0929	0.0892	0.925	0.935
	p=1000	0.0309	0.0412	0.0884	0.0903	0.0936	0.0972	0.920	0.925
Case II	p=500	0.0299	0.0204	0.0957	0.0962	0.1067	0.1009	0.930	0.930
	p=1000	0.0259	0.0176	0.0975	0.0987	0.1094	0.1062	0.925	0.935

Table 6. Summary of genes that are potentially related with breast cancer survival data†.

Gene identifier	Est	SE	CI	P_{adj}
Contig55111_RC	1.6400	0.4176	[0.8214, 2.4586]	0.0430
NM_006397	2.9614	0.7371	[1.5167, 4.4062]	0.0294
NM_006622	−2.1774	0.5040	[−3.1653, −1.1896]	0.0078
NM_016448	3.0523	0.7598	[1.5630, 4.5415]	0.0295
NM_001168	1.9191	0.3840	[1.1664, 2.6717]	0.0003

† "Est" denotes our TPCV-based estimator; "SE" denotes the corresponding standard error; "CI" denotes the 95% confidence interval; P_{adj} denotes the Bonferroni adjusted p-value.