Lifetime Data Analysis
https://doi.org/10.1007/s10985-022-09568-2

Double bias correction for high-dimensional sparse additive hazards regression with covariate measurement errors

Xiaobo Wang[1] · Jiayu Huang[1] · Guosheng Yin[2] · Jian Huang[3] · Yuanshan Wu[4]

Received: 18 September 2021 / Accepted: 6 July 2022

Abstract

We propose an inferential procedure for additive hazards regression with high-dimensional survival data, where the covariates are prone to measurement errors. We develop a double bias correction method by first correcting the bias arising from measurement errors in covariates through an estimating function for the regression parameter. By adopting the convex relaxation technique, a regularized estimator for the regression parameter is obtained by elaborately designing a feasible loss based on the estimating function, which is solved via linear programming. Using the Neyman orthogonality, we propose an asymptotically unbiased estimator which further corrects the bias caused by the convex relaxation and regularization. We derive the convergence rate of the proposed estimator and establish the asymptotic normality for the low-dimensional parameter estimator and the linear combination thereof, accompanied with a consistent estimator for the variance. Numerical experiments are carried

✉ Yuanshan Wu
wu@zuel.edu.cn

Xiaobo Wang
wangxiaobo@whu.edu.cn

Jiayu Huang
jerry_huang@whu.edu.cn

Guosheng Yin
gyin@hku.hk

Jian Huang
jian-huang@uiowa.edu

1 School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China

2 Department of Statistics and Actuarial Science, University of Hong Kong, Pokfulam Road, Hong Kong, China

3 Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242-1419, U.S.A.

4 School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, Hubei 430073, China

Published online: 22 July 2022

out on both simulated and real datasets to demonstrate the promising performance of the proposed double bias correction method.

Keywords Bias correction · Confidence interval · Error-in-variable · Estimating equation · High dimensions · Survival analysis

1 Introduction

With rapid advancement of information technology, high-dimensional complex data can be collected at a fast pace and a low cost in many scientific disciplines and applications. Nevertheless, the measurement error is commonly encountered in the data collection process, which plays an ever more critical role and often imposes new challenges in statistical analysis. There is a vast amount of literature on tackling measurement error problems in low-dimensional settings (Carroll et al. 2006), which however remains a rather unexplored area in high-dimensional settings. The issue becomes more challenging when the outcome is subject to right censoring, a common scenario in survival analysis.

When all variables are fully observed and accurately measured, regularization is one of the most-widely used approaches to high-dimensional sparse regression models; see Bühlmann and van de Geer (2011) and Wainwright (2019) for an overview. Recently, substantial endeavors have been devoted to developing statistical inference, such as construction of confidence intervals and hypothesis testing, for high-dimensional sparse linear regression models. Using the sample splitting techniques, Wasserman and Roeder (2009) and Meinshausen et al. (2009) proposed significance testing procedures for high-dimensional regression coefficients. However, the sample splitting may lead to potential efficiency loss. Lockhart et al. (2014) developed a significance test for variables along the Lasso (Tibshirani 1996) solution path. Lee et al. (2016) proposed an exact post-selection inference procedure for the regression coefficients conditional on the model selected by Lasso. The resulting confidence interval may change with the selected model and thus is difficult to interpret. Another line of research in constructing confidence intervals takes the debiased approach based on Lasso, which exploits the idea of low-dimensional projection while considering the remaining regression coefficients as nuisance parameters (Javanmard and Montanari 2014; Zhang and Zhang 2014). By inverting the Karush–Kuhn–Tucker condition, van de Geer et al. (2014) extended the debiased method to the high-dimensional generalized linear models and proposed to construct confidence intervals using the de-sparsified Lasso estimator. Ning and Liu (2017) constructed confidence intervals for high-dimensional penalized M-estimators based on the decorrelated score statistic. For high-dimensional censored survival data, Fang et al. (2017) and Yu et al. (2021) proposed hypothesis testing and confidence interval procedures in the framework of the Cox proportional hazards model based on the decorrelated and debiased approaches, respectively.

Covariate measurement errors might arise as a consequence of device failures or measurement cost savings. Recent progress on dealing with measurement errors mainly focuses on the high-dimensional sparse linear models. By relaxing the upper bound restriction for the ℓ_∞-norm of the gradient in the Dantzig selector (Candès and Tao 2007), Rosenbaum and Tsybakov (2010) showed that the resulting estimator can correctly recover the sparsity pattern with high probability. Sørensen et al. (2018) made an extension to high-dimensional generalized linear models. As the loss function is no longer convex due to covariate measurement errors, Loh and Wainwright (2012) proposed a non-convex modification of the Lasso and developed a projected gradient descent algorithm. On the other hand, Datta and Zou (2017) advocated the convex approximation to the non-convex loss as convexity is essential for the Lasso method and they further obtained the error bounds of the proposed estimators. Belloni et al. (2017) proposed simultaneous confidence intervals for a subset of regression coefficients where the critical values were obtained using the multiplier bootstrap.

Mismeasured covariates may also arise in survival regression analysis. As a valuable complement of the Cox proportional hazards model, the additive hazards model possesses a distinct interpretation and an explicit solution from the martingale-based estimating equation (Lin and Ying 1994). In the low-dimensional settings, Huang and Wang (2000) proposed a nonparametric-correction approach for the Cox model, and Kulich and Lin (2000) developed an empirical moment plug-in approach for the additive hazards model when there exists a validation set. Yan and Yi (2016) also studied the measurement error effect on the structure of the additive hazard function, and proposed the regression calibration method to reduce it. In the high-dimensional settings with accurately measured covariates, Huang et al. (2013) and Lin and Lv (2013) established the error bounds for the Lasso estimators under the proportional and additive hazards models, respectively.

In this work, we address a more challenging problem of high-dimensional survival data with covariate measurement errors. Focusing on the additive hazards model, our main contributions can be summarized in two different facets. First, multiple layers of bias correction are needed due to multiple sources for bias, for which we propose novel strategies to correct the biases adaptive to different sources. In the first step, we develop a surrogate loss function by correcting the bias from covariate measurement errors. However, the convexity of such loss function cannot be guaranteed, thus leading to an unstable regularized solution. We propose to relax the strict zero-root constraint of the penalized estimating equation and expand the feasible region over a parameter space at the sacrifice of introducing additional bias. We formulate a convex optimization problem following the idea of the Dantzig selector, which thereby greatly facilitates the computation and theoretical analysis. In the second step, we propose the nearly Neyman-orthogonal estimating function of the individual coefficient of interest such that the impact from the regularization bias and estimation of nuisance parameters is negligible (Neyman 1959; Newey 1994; Chernozhukov et al. 2018). Second, by allowing the dimensionality to increase at an exponential rate with respect to the sample size, we establish the error bound for the one-step bias corrected estimator and construct the confidence interval and hypothesis testing for the individual coefficient based on the second-step bias corrected estimator.

The rest of this paper is organized as follows. Section 2 describes the first-step bias corrected estimator under the additive hazards model with covariate measurement errors and establishes its convergence rate under some regularity conditions. In Sect. 3, we construct the confidence intervals and hypothesis testing for the low-dimensional components of high-dimensional regression coefficients and demonstrate the connection from the second-step bias correction for regularization. We conduct simulation studies in Sect. 4, including parameter estimation, confidence intervals, and size and power analysis for hypothesis testing, to evaluate the finite-sample performances of our propose methods. A real dataset is analyzed in Sect. 5, and Sect. 6 concludes with some discussions. The proofs of main theorems are relegated to the Appendix while additional preliminary lemmas are collected in the online supplementary material.

2 Additive Hazards Regression

Let T denote the failure time, C denote the censoring time, and $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$ be a p-vector of covariates. Assume that T and C are conditionally independent given $\mathbf{Z}$. Let $X = \min(T, C)$ denote the observed survival time and $\Delta = I(T \leq C)$ denote the failure indicator, where $I(\cdot)$ is the indicator function. In contrast to the Cox proportional hazards model (Cox 1972), the additive hazards model specifies that the hazard function associated with covariate $\mathbf{Z}$ is the sum of, rather than the product of, the baseline hazard function and the regression function of covariates. Specifically, the hazard function takes the form of

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) + \mathbf{Z}^\top \boldsymbol{\beta}_0,$$

where $\lambda_0(t)$ is the unknown baseline hazard function and $\boldsymbol{\beta}_0$ is a p-vector of unknown regression parameters. For $i = 1, \ldots, n$, let $(X_i, \Delta_i, \mathbf{Z}_i)$ be the independent and identically distributed copies of $(X, \Delta, \mathbf{Z})$. Let $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ denote the counting process and $Y_i(t) = I(X_i \geq t)$ denote the at-risk process, and $\overline{\mathbf{Z}}(t) = \sum_{i=1}^n Y_i(t)\mathbf{Z}_i / \sum_{i=1}^n Y_i(t)$. Adopting the martingale estimating equation, Lin and Ying (1994) proposed a pseudo-score function,

$$\mathbf{G}_n^*(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i - \overline{\mathbf{Z}}(t)\}\{Y_i(t)\mathbf{Z}_i^\top \boldsymbol{\beta} \mathrm{d}t - \mathrm{d}N_i(t)\},$$

where τ is the end time of the study duration and $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^p$ with $\mathcal{B}$ being the parameter space. We rewrite $\mathbf{G}_n^*(\boldsymbol{\beta}) = \mathbf{B}_n^* \boldsymbol{\beta} - \mathbf{b}_n^*$, where

$$\mathbf{b}_n^* = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \{\mathbf{Z}_i - \overline{\mathbf{Z}}(t)\} \mathrm{d}N_i(t)$$

and

$$\mathbf{B}_n^* = \frac{1}{n} \sum_{i=1}^n \int_0^\tau Y_i(t)\{\mathbf{Z}_i - \overline{\mathbf{Z}}(t)\}^{\otimes 2} \mathrm{d}t$$

with $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\top$ for a vector $\mathbf{a}$. Clearly, the resulting estimator can be obtained by solving the zero root of $\mathbf{G}_n^*(\boldsymbol{\beta})$. However, neither is such an estimation procedure applicable in the high-dimensional settings where p is larger than n nor in the situation where covariates $\mathbf{Z}$ are prone to measurement errors. When both situations are present in high-dimensional complex survival data, the problem becomes even more challenging.

We assume the classic additive measurement error structure as follows,

$$\mathbf{W} = \mathbf{Z} + \mathbf{U},$$

where $\mathbf{U} = (U_1, \ldots, U_p)^\top$ is a p-variate symmetrically distributed random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}$. We do not directly observe $\mathbf{Z}$ but only observe its surrogate $\mathbf{W}$, because covariate $\mathbf{Z}$ may be measured with error. If some covariates are error-free, we can simply set the corresponding terms in $\mathbf{V}$ to be zero. To facilitate development of our double bias correction approach, we assume $\mathbf{V}$ to be known provisionally. We further make the typical surrogacy assumption that (T, C) and $\mathbf{W}$ are conditionally independent given covariate $\mathbf{Z}$, which is obviously satisfied when $\mathbf{U}$ is independent of $(T, C, \mathbf{Z})$, in conjunction with the random censoring mechanism. Assume that $\mathbf{W}_1, \ldots, \mathbf{W}_n$ are independent copies of $\mathbf{W}$, and let $\mathcal{O}_n = \{X_i, \Delta_i, \mathbf{W}_i\}_{i=1}^n$ denote the observed data and correspondingly $\mathcal{U}_n = \{X_i, \Delta_i, \mathbf{Z}_i\}_{i=1}^n$.

We develop the procedure for the first correction of the bias caused by simply replacing $\mathbf{Z}_i$ with $\mathbf{W}_i$ in $\mathbf{G}_n^*(\boldsymbol{\beta})$. For ease of exposition, let

$$\mathbf{b}_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\mathbf{W}_i - \overline{\mathbf{W}}(t)\}\mathrm{d}N_i(t)$$

and

$$\mathbf{B}_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{\mathbf{W}_i - \overline{\mathbf{W}}(t)\}^{\otimes 2}\mathrm{d}t,$$

where $\overline{\mathbf{W}}(t) = \sum_{i=1}^n Y_i(t)\mathbf{W}_i / \sum_{i=1}^n Y_i(t)$. By noting that

$$E(\mathbf{W}_i \mid \mathcal{U}_n) = \mathbf{Z}_i \quad \text{and} \quad E(\mathbf{W}_i\mathbf{W}_j^\top \mid \mathcal{U}_n) = \mathbf{Z}_i\mathbf{Z}_j^\top + I(i=j)\mathbf{V}$$

and following some algebra, we can derive that

$$E(\mathbf{B}_n\boldsymbol{\beta} - \mathbf{b}_n \mid \mathcal{U}_n) = \mathbf{G}_n^*(\boldsymbol{\beta}) + d_n\mathbf{V}\boldsymbol{\beta},$$

where

$$d_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\mathrm{d}t - \frac{\tau}{n}.$$

Inspired by the corrected score method proposed by Stefanski (1989), we immediately obtain the desirable corrected estimating function, depending only on $\mathcal{O}_n$ and $\boldsymbol{\beta}$, as follows,

$$\mathbf{G}_n(\boldsymbol{\beta}) = \mathbf{D}_n\boldsymbol{\beta} - \mathbf{b}_n, \quad \text{with } \mathbf{D}_n = \mathbf{B}_n - d_n\mathbf{V},$$

which satisfies $E\{\mathbf{G}_n(\boldsymbol{\beta}) \mid \mathcal{U}_n\} = \mathbf{G}_n^*(\boldsymbol{\beta})$. Ideally, following Johnson et al. (2008) and Lin and Lv (2013), the penalized estimating equation with the Lasso penalty is given by

$$\mathbf{G}_n(\boldsymbol{\beta}) + \delta \cdot \text{sign}(\boldsymbol{\beta}) = \mathbf{0}, \tag{2.1}$$

where sign$(\cdot)$ is the pointwise sign function and $\delta > 0$ is the tuning parameter. Due to the extra term $d_n\mathbf{V}$ arising from the measurement error, the matrix $\mathbf{D}_n = \mathbf{B}_n - d_n\mathbf{V}$ cannot be guaranteed to be non-negative definite. This prohibits a unique solution of (2.1), and the resulting estimator would be unstable. A more plausible way is to relax the strict zero-root constraint of (2.1) and expand the feasible region over the parameter space $\mathcal{B}$ at the sacrifice of introducing additional bias. In particular, we define the estimator $\widehat{\boldsymbol{\beta}}$ as the solution of the minimization problem,

$$\min_{\boldsymbol{\beta}\in\mathcal{B}} \{|\boldsymbol{\beta}|_1 : |\mathbf{G}_n(\boldsymbol{\beta})|_\infty \le \eta|\boldsymbol{\beta}|_1 + \delta\}, \tag{2.2}$$

where $|\cdot|_r$ denotes the ℓ_r-norm for a vector with $1 \le r \le \infty$ with $|\cdot|_\infty$ being the maximal norm. The term $\eta|\boldsymbol{\beta}|_1$ is added to further loosen the upper bound tailored by the tuning parameter $\eta > 0$ (Rosenbaum and Tsybakov 2010; Sørensen et al. 2018). We show in Lemma 3 that the true parameter $\boldsymbol{\beta}_0$ falls in the feasible region of (2.2) with probability tending to one. By introducing a slack vector $\mathbf{u} = (u_1, \ldots, u_p)^\top$, minimizing (2.2) is equivalent to solving the optimization problem,

$$\begin{aligned} \text{minimize} \quad & \mathbf{1}^\top\mathbf{u} \\ \text{subject to} \quad & -\mathbf{u} \le \boldsymbol{\beta} \le \mathbf{u} \\ & -\eta\mathbf{1}\mathbf{1}^\top\mathbf{u} + \mathbf{D}_n\boldsymbol{\beta} \le \delta\mathbf{1} + \mathbf{b}_n \\ & -\eta\mathbf{1}\mathbf{1}^\top\mathbf{u} - \mathbf{D}_n\boldsymbol{\beta} \le \delta\mathbf{1} - \mathbf{b}_n \\ & \mathbf{u} \ge \mathbf{0}, \end{aligned} \tag{2.3}$$

where $\mathbf{1}$ is a vector of 1 and $\mathbf{0}$ is a vector of 0. For notational simplicity, the length of a vector is omitted when there is no ambiguity. We write $\boldsymbol{\beta} \le \mathbf{u}$ if $\beta_j \le u_j$ for $j = 1, \ldots, p$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$. The resulting estimator, denoted by $\widehat{\boldsymbol{\beta}}$, can be obtained by solving (2.3) using linear programming. Because it is a crude estimator obtained by expanding the feasible region, multiple solutions to (2.3) may exist and they can all be considered as $\widehat{\boldsymbol{\beta}}$.

The rate of $\widehat{\boldsymbol{\beta}}$ converging to the true parameter $\boldsymbol{\beta}_0$ should first be established before accounting for the bias of $\widehat{\boldsymbol{\beta}}$ explicitly. Let $\mathcal{A}$ denote a set of indices and $\mathbf{v}$ denote a vector, and let $\mathbf{v}_{\mathcal{A}}$ be the sub-vector of $\mathbf{v}$ with indices corresponding to those in set $\mathcal{A}$. Denote the active set by $\mathcal{A}_0 = \{j : \beta_{0j} \ne 0\}$, where β_{0j} is the j-th element of the true regression parameter $\boldsymbol{\beta}_0$, for $j = 1, \ldots, p$. Let $s_0 = |\mathcal{A}_0|$, the cardinality of set $\mathcal{A}_0$. We first state some regularity conditions as follows.

C1 Assume that $\Lambda_0(\tau) = \int_0^\tau \lambda_0(t)\mathrm{d}t < \infty$ and $P\{Y(\tau) = 1\} \geq \tau_0 > 0$ for some constant τ_0, where $Y(t) = I(X \geq t)$ for $t \in [0, \tau]$.

C2 Each component of $\mathbf{Z}$ and $\mathbf{U}$ is sub-Gaussian.

C3 There exists a constant $\zeta_0 > 0$, such that

$$\min_{\mathbf{v}\in\mathbb{R}^p\setminus\{\mathbf{0}\}, |\mathbf{v}_{\mathcal{A}_0^c}|_1 \leq 3|\mathbf{v}_{\mathcal{A}_0}|_1} \frac{\mathbf{v}^\top \mathbf{B}^* \mathbf{v}}{\mathbf{v}^\top \mathbf{v}} \geq \zeta_0,$$

where

$$\mathbf{B}^* = E\left\{\int_0^\tau Y(t)\,(\mathbf{Z} - \bar{\mathbf{z}}(t))^{\otimes 2}\,\mathrm{d}t\right\}$$

with $\bar{\mathbf{z}}(t) = E\{Y(t)\mathbf{Z}\}/E\{Y(t)\}$.

Condition C1 is a standard assumption in survival analysis. Condition C2 bounds the tail probability for every component of the covariate and the measurement error. It implies that the event

$$\Omega_L = \left\{\max\left(\max_{1\leq j\leq p} |Z_j|,\ \max_{1\leq j\leq p} |U_j|\right) \leq L\right\}$$

can occur with high probability for sufficiently large L. Condition C3 is regarding the restricted eigenvalue condition, an inherent condition in high-dimensional regression models. We denote $c_n \asymp c_n^*$ if c_n and c_n^* are of the same order of magnitude. Let $L_{\max}^{c_1,c_2} = \max\{L^{c_1}, |\mathbf{V}|_{\max}^{c_2}\}$ for some constants c_1 and c_2, where $|\mathbf{V}|_{\max} = \max_{i,j} |V_{ij}|$ with $\mathbf{V} = (V_{ij})_{p\times p}$. The error bound for the proposed estimator $\widehat{\boldsymbol{\beta}}$ is given in the following theorem.

Theorem 1 *Assume* $\eta \asymp \delta \asymp L_{\max}^{2,1}\sqrt{\log p/n}$ *and* $s_0 L^2\sqrt{\log p/n} = o(1)$. *Under conditions* C1–C3, *we have*

$$|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|_r = O_P\left(s_0^{1/r} L_{\max}^{2,1}\sqrt{\frac{\log p}{n}}\right)$$

for $r = 1$ *and* 2.

The proof of Theorem 1 is delegated to the Appendix. When it holds $s_0^{1/r} L_{\max}^{2,1}\sqrt{\log p/n} = o(1)$, the ℓ_r-bound converges to zero in probability. To present our results in a more general framework, we explicitly emphasize the dependence of error bounds on L, wherein L can be chosen large enough to ensure that the event Ω_L happens with probability close to one. On the other hand, if $\mathbf{Z}$ is uniformly bounded by a constant as assumed in Bühlmann and van de Geer (2011), the extra effect on the error bound arises from the measurement error $\mathbf{U}$ and its covariance matrix $\mathbf{V}$. As the first step of bias correction, the estimator $\widehat{\boldsymbol{\beta}}$ is obtained via the corrected estimating function, while the convex relaxation in the penalized term unfortunately introduces additional bias. Thus, the second step of bias correction is necessary when we construct the confidence interval and hypothesis testing procedure.

3 Hypothesis Test and Confidence Interval

3.1 Inference for One Parameter

Suppose that we are interested in testing the k-th component of $\boldsymbol{\beta}_0$,

$$H_0 : \beta_{0k} = \theta_0 \quad \text{versus} \quad H_1 : \beta_{0k} \neq \theta_0, \tag{3.1}$$

where θ_0 is a prespecified constant. Obviously, the nuisance parameters $\boldsymbol{\beta}_{0(-k)} = \{\beta_{0j} : j = 1, \ldots, p, j \neq k\}$ should also be taken into account in the hypothesis testing procedure. Employing the semiparametric efficiency theory, we propose a new test procedure for (3.1) such that the effect from estimating the high-dimensional nuisance parameter becomes asymptotically negligible. We rewrite $\boldsymbol{\beta}_0 = (\beta_{0k}, \boldsymbol{\beta}_{0(-k)}^\top)^\top$, and let $G_{nk}(\boldsymbol{\beta})$ denote the k-th component of $\mathbf{G}_n(\boldsymbol{\beta})$ and $\mathbf{G}_{n(-k)}(\boldsymbol{\beta})$ denote the remaining components. For a matrix $\mathbf{A}$, let A_{jk} denote the (j, k)-th element of $\mathbf{A}$. Denote $\mathbf{A}_{j,-k}$ as the j-th row of $\mathbf{A}$ but excluding the k-th element A_{jk} and $\mathbf{A}_{-j,k}$ as the k-th column of $\mathbf{A}$ but excluding the j-th element A_{jk}. Let $\mathbf{A}_{-j,-k}$ denote the $(p-1) \times (p-1)$ submatrix of $\mathbf{A}$ by deleting its j-th row and k-th column.

For any measurable function f, let $\mathbb{P}_n f$ denote the expectation under the empirical probability measure $\mathbb{P}_n$. Denote

$$\mathbf{g}(\boldsymbol{\beta}_0) = -\int_0^\tau \{\mathbf{W} - \bar{\mathbf{z}}(t)\} \mathrm{d}M(t) + \int_0^\tau Y(t)[\{\mathbf{W} - \bar{\mathbf{z}}(t)\}\mathbf{U}^\top - \mathbf{V}]\mathrm{d}t\boldsymbol{\beta}_0,$$

where

$$M(t) = N(t) - \int_0^t Y(s)\{\mathrm{d}\Lambda_0(s) + \mathbf{Z}^\top \boldsymbol{\beta}_0 \mathrm{d}s\} \tag{3.2}$$

is the martingale process. The empirical version $\mathbb{P}_n\mathbf{g}(\boldsymbol{\beta}_0)$ can be viewed as an approximation of $\mathbf{G}_n(\boldsymbol{\beta}_0)$. The linearity of $\mathbf{g}(\boldsymbol{\beta}_0) = (g_1(\boldsymbol{\beta}_0), \ldots, g_p(\boldsymbol{\beta}_0))^\top$ implies that the Neyman orthogonal estimating function takes the form of

$$g_k^{\mathrm{o}}(\beta_{0k} \mid \boldsymbol{\beta}_{0(-k)}) = g_k(\boldsymbol{\beta}_0) - \mathbf{g}_{-k}^\top(\boldsymbol{\beta}_0)\boldsymbol{\gamma}_{0k},$$

where the projection direction $\boldsymbol{\gamma}_{0k}$ satisfies that

$$\partial_t E\{g_k^{\mathrm{o}}(\beta_{0k} \mid \boldsymbol{\beta}_{0(-k)} + t(\boldsymbol{\beta}_{-k}^* - \boldsymbol{\beta}_{0(-k)}))\}\big|_{t=0} = 0$$

along the submodel $t \mapsto \boldsymbol{\beta}_{0(-k)} + t(\boldsymbol{\beta}_{-k}^* - \boldsymbol{\beta}_{0(-k)})$ for any $\boldsymbol{\beta}_{-k}^*$. As a result, $\boldsymbol{\gamma}_{0k} = \mathbf{B}_{-k,-k}^{*-1}\mathbf{B}_{-k,k}^*$. At the population level, the Neyman orthogonal estimating function $g_k^{\mathrm{o}}(\beta_{0k} \mid \boldsymbol{\beta}_{0(-k)})$ is given by

$$g_k^{\mathrm{o}}(\beta_{0k} \mid \boldsymbol{\beta}_{0(-k)}) = g_k(\boldsymbol{\beta}_0) - \mathbf{g}_{-k}^\top(\boldsymbol{\beta}_0)\mathbf{B}_{-k,-k}^{*-1}\mathbf{B}_{-k,k}^*. \tag{3.3}$$

As $\mathbf{B}^*$ and $\mathbf{D}_n$ are close enough following Lemma 4, the estimator for the projection direction $\boldsymbol{\gamma}_{0k}$ can be obtained in the way of relaxation, which is defined as

$$\widehat{\boldsymbol{\gamma}}_k \in \arg\min_{\boldsymbol{\gamma}\in\mathcal{T}_k}\{|\boldsymbol{\gamma}|_1 : |(\mathbf{D}_n)_{-k,-k}\boldsymbol{\gamma} - (\mathbf{D}_n)_{-k,k}|_\infty \leq \eta_k|\boldsymbol{\gamma}|_1 + \delta_k\}, \tag{3.4}$$

where η_k and δ_k are the tuning parameters and $\mathcal{T}_k \subset \mathbb{R}^{p-1}$ is the parameter space. The minimization of (3.4) can be carried out using the linear programming procedure as (2.3). Mimicking (3.3), the Neyman orthogonal estimating function at the sample level is

$$G^{\mathrm{o}}_{nk}(\beta_k \mid \boldsymbol{\beta}_{-k}) = G_{nk}(\boldsymbol{\beta}) - \mathbf{G}^\top_{n(-k)}(\boldsymbol{\beta})\widehat{\boldsymbol{\gamma}}_k, \tag{3.5}$$

which is expected to be the desirable testing statistic for (3.1) by substituting $\widehat{\boldsymbol{\beta}}_{-k}$ for $\boldsymbol{\beta}_{-k}$. As shown in Lemma 7, the distance between $G^{\mathrm{o}}_{nk}(\beta_{0k} \mid \widehat{\boldsymbol{\beta}}_{-k})$ and $\mathbb{P}_n g^{\mathrm{o}}_k(\beta_{0k} \mid \boldsymbol{\beta}_{0(-k)})$ is negligible. Denote $\mathcal{A}_k = \{j : \gamma_{0k(j)} \neq 0\}$, where $\gamma_{0k(j)}$, $j = 1, \ldots, p-1$, is the j-th component of $\boldsymbol{\gamma}_{0k}$. Let $s_k = |\mathcal{A}_k|$. We need additional conditions in parallel with condition C3 to derive the convergence rate of $\widehat{\boldsymbol{\gamma}}_k$.

C4 There exists a constant $\zeta_{0k} > 0$, such that

$$\min_{\mathbf{v}\in\mathbb{R}^{p-1}\backslash\{\mathbf{0}\}, |\mathbf{v}_{\mathcal{A}_k^c}|_1 \leq 3|\mathbf{v}_{\mathcal{A}_k}|_1} \frac{\mathbf{v}^\top \mathbf{B}^*_{-k,-k}\mathbf{v}}{\mathbf{v}^\top\mathbf{v}} \geq \zeta_{0k}.$$

Theorem 2 *Assume $\eta_k \asymp \delta_k \asymp L^{2,1}_{\max}\sqrt{\log p/n}$ and $s_k L^2\sqrt{\log p/n} = o(1)$. Under conditions C1, C2 and C4, we have*

$$|\widehat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_{0k}|_r = O_P\left(s_k^{1/r} L^{2,1}_{\max}\sqrt{\frac{\log p}{n}}\right)$$

for $r = 1$ and 2.

The proof of Theorem 2 is provided in the Appendix. For notational simplicity, let $\Pi_{k,\theta_0}(\boldsymbol{\beta})$ denote the same vector as $\boldsymbol{\beta}$ except for the k-th component replaced by θ_0. Based on Theorems 1 and 2, we construct the test statistic under H_0,

$$G^{\mathrm{o}}_{nk}(\theta_0 \mid \widehat{\boldsymbol{\beta}}_{-k}) = G_{nk}(\Pi_{k,\theta_0}(\widehat{\boldsymbol{\beta}})) - \mathbf{G}^\top_{n(-k)}(\Pi_{k,\theta_0}(\widehat{\boldsymbol{\beta}}))\widehat{\boldsymbol{\gamma}}_k. \tag{3.6}$$

To derive the limiting distribution of (3.6), we need more conditions as follows.

C5 Assume $\log p = o(n^\alpha)$ for some $\alpha \in (0, 1/2)$.
C6 Assume

- (i) $n^{\alpha-1/2}L^{2,1}_{\max} = o(1)$;
- (ii) $\max(s_0, s_k)L^{4,2}_{\max}\log p/\sqrt{n} = o(1)$;
- (iii) $\max(s_0, s_k)L^{6,3}_{\max}\sqrt{\log p/n} = o(1)$.

Condition C5 indicates that the dimension p can attain an exponential rate of sample size n. Conditions C6-(ii) and C6-(iii) imply the assumptions $s_0L^2\sqrt{\log p/n} = o(1)$ and $s_kL^2\sqrt{\log p/n} = o(1)$ in Theorems 1 and 2 respectively. In the situation with no measurement error, i.e., $\mathbf{V} = \mathbf{0}$, when $\mathbf{Z}$ is uniformly bounded by some constant, C6-(i) automatically holds, C6-(ii) reduces to $\max(s_0, s_k)\log p/\sqrt{n} = o(1)$, and C6-(iii) to $\max(s_0, s_k)\sqrt{\log p/n} = o(1)$; clearly C6-(ii) is more stringent than C6-(iii). In fact, C6-(iii) is typically employed to show the consistency whereas C6-(ii) for the asymptotic distribution; e.g., see Ning and Liu (2017).

We denote the covariance matrix by $\mathbf{\Sigma}(\boldsymbol{\beta}_0) = E\{\mathbf{g}(\boldsymbol{\beta}_0)^{\otimes 2}\}$. Let $\Gamma_{k,\theta_0}(\boldsymbol{\beta})$ denote the vector obtained by inserting a scalar θ_0 between the $(k-1)$-th and k-th components of the vector $\boldsymbol{\beta}$. The limiting distribution of (3.6) is given in the following theorem.

Theorem 3 *Assume $\eta \asymp \delta \asymp \eta_k \asymp \delta_k \asymp L_{\max}^{2,1}\sqrt{\log p/n}$. Under conditions C1–C5, C6-(i) and C6-(ii), we have*

$$\sqrt{n}G_{nk}^{\mathrm{o}}(\theta_0 \mid \widehat{\boldsymbol{\beta}}_{-k}) \xrightarrow{\mathcal{D}} N(0, \sigma_k^2),$$

where $\sigma_k^2 = \boldsymbol{\varphi}_k^\top \mathbf{\Sigma}(\boldsymbol{\beta}_0)\boldsymbol{\varphi}_k$ with $\boldsymbol{\varphi}_k = \Gamma_{k,1}(-\boldsymbol{\gamma}_{0k})$.

To obtain a consistent estimator for σ_k^2, we consider the sample counterpart,

$$\begin{aligned}\widehat{\mathbf{g}}_i(\widehat{\boldsymbol{\beta}}) = &-\int_0^\tau \{\mathbf{W}_i - \overline{\mathbf{W}}(t)\}\left\{\mathrm{d}N_i(t) - Y_i(t)\mathrm{d}\widehat{\Lambda}_0(t)\right\} \\ &+ \int_0^\tau Y_i(t)\left[\{\mathbf{W}_i - \overline{\mathbf{W}}(t)\}\mathbf{W}_i^\top - \mathbf{V}\right]\mathrm{d}t\widehat{\boldsymbol{\beta}},\end{aligned}$$

where

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^n \int_0^t \frac{\mathrm{d}N_i(s) - Y_i(s)\mathbf{W}_i^\top\widehat{\boldsymbol{\beta}}\mathrm{d}s}{\sum_{j=1}^n Y_j(s)}$$

is the Breslow estimator for $\Lambda_0(t)$. Define an empirical estimator for $\mathbf{\Sigma}(\boldsymbol{\beta}_0)$ as

$$\widehat{\mathbf{\Sigma}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^n \{\widehat{\mathbf{g}}_i(\widehat{\boldsymbol{\beta}})\}^{\otimes 2}.$$

Theorem 4 *Assume $\eta \asymp \delta \asymp \eta_k \asymp \delta_k \asymp L_{\max}^{2,1}\sqrt{\log p/n}$ and $s_0 L_{\max}^{2,1}\sqrt{\log p/n} = o(1)$. Under conditions C1–C6, we have*

$$\widehat{\sigma}_k^2 = \sigma_k^2 + O_P\left(\max(s_0, s_k)L_{\max}^{6,3}\sqrt{\frac{\log p}{n}}\right),$$

where $\widehat{\sigma}_k^2 = \widehat{\boldsymbol{\varphi}}_k^\top \widehat{\mathbf{\Sigma}}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\varphi}}_k$ with $\widehat{\boldsymbol{\varphi}}_k = \Gamma_{k,1}(-\widehat{\boldsymbol{\gamma}}_k)$.

Theorem 5 *Assume $\eta \asymp \delta \asymp \eta_k \asymp \delta_k \asymp L_{\max}^{2,1}\sqrt{\log p/n}$. Under conditions C1–C6, we have*

$$\sqrt{n}G_{nk}^{\mathrm{o}}(\theta_0 \mid \widehat{\boldsymbol{\beta}}_{-k})/\widehat{\sigma}_k \xrightarrow{\mathcal{D}} N(0, 1).$$

The proofs of Theorems 3 and 4 are outlined in the Appendix while that of Theorem 5 follows directly from the former and is thus omitted. The null hypothesis H_0 is rejected at the significant level α if

$$|\sqrt{n}G^{\mathrm{o}}_{nk}(\theta_0 \mid \widehat{\boldsymbol{\beta}}_{-k})/\widehat{\sigma}_k| > z_{1-\alpha/2},$$

where z_α is the α-th lower quantile of the standard normal distribution. The asymptotic confidence interval of β_{0k} can be derived by reversely solving from the rejection region,

$$\frac{\widehat{\boldsymbol{\varphi}}_k^\top \mathbf{b}_n - (\mathbf{D}_n\widehat{\boldsymbol{\varphi}}_k)^\top_{-k}\widehat{\boldsymbol{\beta}}_{-k} \pm z_{1-\alpha/2}\widehat{\sigma}_k/\sqrt{n}}{\mathbf{e}_k^\top \mathbf{D}_n\widehat{\boldsymbol{\varphi}}_k}, \tag{3.7}$$

where $\mathbf{e}_k$ is the k-th vector of the canonical basis of $\mathbb{R}^p$. It is evident from the center-shifted confidence interval (3.7) for β_{0k} that we can define

$$\begin{aligned}\widetilde{\beta}_k &= \widehat{\beta}_k - \left(\widehat{\beta}_k - \frac{\widehat{\boldsymbol{\varphi}}_k^\top \mathbf{b}_n - (\mathbf{D}_n\widehat{\boldsymbol{\varphi}}_k)^\top_{-k}\widehat{\boldsymbol{\beta}}_{-k}}{\mathbf{e}_k^\top \mathbf{D}_n\widehat{\boldsymbol{\varphi}}_k}\right)\\ &= \widehat{\beta}_k - (\mathbf{e}_k^\top \mathbf{D}_n\widehat{\boldsymbol{\varphi}}_k)^{-1}\widehat{\boldsymbol{\varphi}}_k^\top(\mathbf{D}_n\widehat{\boldsymbol{\beta}} - \mathbf{b}_n),\end{aligned} \tag{3.8}$$

which can be viewed as the second-step bias correction. We further conclude that $\sqrt{n}(\widetilde{\beta}_k - \beta_{0k})$ converges in distribution to a zero-mean normal distribution with the standard error consistently estimated by $\widehat{\sigma}_k(\mathbf{e}_k^\top \mathbf{D}_n\widehat{\boldsymbol{\varphi}}_k)^{-1}$. However, this convergence property is not possessed by the first-step bias correction estimator $\widehat{\beta}_k$. We expand the double bias corrected estimator (3.8) by running k from 1 to p, leading to

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Xi}}^{-1}\widehat{\boldsymbol{\Phi}}\left(\mathbf{D}_n\widehat{\boldsymbol{\beta}} - \mathbf{b}_n\right), \tag{3.9}$$

where $\widehat{\boldsymbol{\Xi}} = \mathrm{diag}(\mathbf{e}_1^\top \mathbf{D}_n\widehat{\boldsymbol{\varphi}}_1, \ldots, \mathbf{e}_p^\top \mathbf{D}_n\widehat{\boldsymbol{\varphi}}_p)$ and $\widehat{\boldsymbol{\Phi}} = (\widehat{\boldsymbol{\varphi}}_1, \ldots, \widehat{\boldsymbol{\varphi}}_p)^\top$. This is in spirit similar to the debiased Lasso proposed by Zhang and Zhang (2014).

The tuning parameters offer us advantages of solving estimators by expanding the feasible region. We develop the cross-validation (CV) procedure by designing proper loss functions for selecting the tuning parameters. We first divide the dataset into J pieces equally. Let $\widehat{\boldsymbol{\beta}}_{[-j]}(\eta, \delta)$ denote the estimator using the same procedure of obtaining $\widehat{\boldsymbol{\beta}}$ when the j-th sub-dataset is removed. Motivated by the linear form of the corrected estimating function $\mathbf{G}_n(\boldsymbol{\beta})$, we define the CV criterion

$$\mathrm{CV}(\eta, \delta) = \frac{1}{J}\sum_{j=1}^{J} L_{[j]}(\widehat{\boldsymbol{\beta}}_{[-j]}(\eta, \delta)),$$

where $L_{[j]}(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{D}_{n[j]}\boldsymbol{\beta}/2 - \mathbf{b}_{n[j]}\boldsymbol{\beta}$ with $\mathbf{D}_{n[j]}$ and $\mathbf{b}_{n[j]}$ being calculated using the j-th sub-dataset. The tuning parameters can be selected by minimizing $\mathrm{CV}(\eta, \delta)$. Similar procedures can be developed for selecting the tuning parameters η_k and δ_k. As it is time-consuming to find the global minimizer of the CV criterion function, we alternatively search for the sub-optimum along the diagonal line $\eta = \delta$ or $\eta_k = \delta_k$, which is adopted in the numerical studies.

3.2 Inference for Multiple Parameters

The inference procedure for a single parameter can be extended to test multiple parameters simultaneously. For a matrix $\mathbf{A} = (a_{ij})_{p\times p}$, let $\mathcal{I}_1$ and $\mathcal{I}_2$ denote two subsets of $\{1, \ldots, p\}$, and let $\mathbf{A}_{\mathcal{I}_1,\mathcal{I}_2} = (a_{ij})_{(i,j)\in\mathcal{I}_1\times\mathcal{I}_2}$ be the correponding submatrix of $\mathbf{A}$. We consider the hypothesis test,

$$H_0 : \boldsymbol{\beta}_{0\mathcal{I}} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_{0\mathcal{I}} \neq \boldsymbol{\theta}_0,$$

where $\mathcal{I} \subset \{1, \ldots, p\}$ with $|\mathcal{I}|$ being fixed, $\boldsymbol{\beta}_{0\mathcal{I}} = \{\beta_{0j} : j \in \mathcal{I}\}$, and $\boldsymbol{\theta}_0$ is a prespecified $|\mathcal{I}|$-dimensional vector. Rewrite $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{I}}^\top, \boldsymbol{\beta}_{\mathcal{I}^c}^\top)^\top$, where $\mathcal{I}^c$ is the complement set of $\mathcal{I}$, i.e., $\mathcal{I}^c = \{1, \ldots, p\} - \mathcal{I}$. The Neyman orthogonal function can be derived as

$$\mathbf{G}_{n\mathcal{I}}^{\mathrm{o}}(\boldsymbol{\beta}_{\mathcal{I}} \mid \boldsymbol{\beta}_{\mathcal{I}^c}) = \mathbf{G}_{n\mathcal{I}}(\boldsymbol{\beta}) - \boldsymbol{\gamma}_{0\mathcal{I}}^\top \mathbf{G}_{n\mathcal{I}^c}(\boldsymbol{\beta}),$$

where $\mathbf{G}_{n\mathcal{I}}(\boldsymbol{\beta}) = \{G_{nk}(\boldsymbol{\beta}) : k \in \mathcal{I}\}$ and $\mathbf{G}_{n\mathcal{I}^c}(\boldsymbol{\beta}) = \{G_{nk}(\boldsymbol{\beta}) : k \in \mathcal{I}^c\}$. The true value of the direction of projection matrix $\boldsymbol{\gamma}_{0\mathcal{I}}$ satisfies the Neyman orthogonal equation,

$$\partial_r E\left\{\mathbf{G}_{n\mathcal{I}}^{\mathrm{o}}(\boldsymbol{\beta}_{0\mathcal{I}} \mid \boldsymbol{\beta}_{0\mathcal{I}^c} + r(\boldsymbol{\beta}_{\mathcal{I}^c} - \boldsymbol{\beta}_{0\mathcal{I}^c}))\right\} = \mathbf{0}$$

for any $\boldsymbol{\beta}_{\mathcal{I}^c} \in \mathbb{R}^{p-|\mathcal{I}|}$, which yields that

$$\boldsymbol{\gamma}_{0\mathcal{I}} = (\mathbf{D}_n)_{\mathcal{I}^c\mathcal{I}^c}^{-1}(\mathbf{D}_n)_{\mathcal{I}^c\mathcal{I}}.$$

Set $\boldsymbol{\gamma}_{0\mathcal{I}} = \left(\boldsymbol{\gamma}_{0(\cdot,j)} : j \in \mathcal{I}\right)$, where $\boldsymbol{\gamma}_{0(\cdot,j)}$ is the j-th column of matrix $\boldsymbol{\gamma}_{0\mathcal{I}}$. By mimicking the argument adopted in the test procedure for the single parameter, we propose the estimator of $\boldsymbol{\gamma}_{0(\cdot,j)}$ as

$$\widehat{\boldsymbol{\gamma}}_{\cdot,j} \in \arg\min_{\boldsymbol{\gamma}\in\widetilde{\mathcal{T}}_j}\{|\boldsymbol{\gamma}|_1 : |(\mathbf{D}_n)_{\mathcal{I}^c,\mathcal{I}^c}\boldsymbol{\gamma} - (\mathbf{D}_n)_{\mathcal{I}^c,j}|_\infty \leq \eta_{\mathcal{I},j}|\boldsymbol{\gamma}|_1 + \delta_{\mathcal{I},j}\},$$

where $\eta_{\mathcal{I},j}$ and $\delta_{\mathcal{I},j}$ are the tuning parameters for every $j \in \mathcal{I}$ and $\widetilde{\mathcal{T}}_j \subset \mathbb{R}^{p-|\mathcal{I}|}$ is the parameter space. As a result, the orthogonal score function is estimated as

$$\widehat{\mathbf{G}}_{n\mathcal{I}}^{\mathrm{o}}(\boldsymbol{\theta}_0 \mid \widehat{\boldsymbol{\beta}}_{\mathcal{I}^c}) = \mathbf{G}_{n\mathcal{I}}((\boldsymbol{\theta}_0^\top, \widehat{\boldsymbol{\beta}}_{\mathcal{I}^c}^\top)^\top) - \widehat{\boldsymbol{\gamma}}_{\mathcal{I}}^\top \mathbf{G}_{n\mathcal{I}^c}((\boldsymbol{\theta}_0^\top, \widehat{\boldsymbol{\beta}}_{\mathcal{I}^c}^\top)^\top),$$

where $\widehat{\boldsymbol{\gamma}}_{\mathcal{I}} = \left(\widehat{\boldsymbol{\gamma}}_{\cdot,j} : j \in \mathcal{I}\right)$. Furthermore, under the null hypothesis it holds

$$\sqrt{n}\widehat{\mathbf{G}}_{n\mathcal{I}}^{\mathrm{o}}(\boldsymbol{\theta}_0 \mid \widehat{\boldsymbol{\beta}}_{\mathcal{I}^c}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Psi}_{\mathcal{I}}\boldsymbol{\Sigma}(\boldsymbol{\beta}_{\mathbf{0}})\boldsymbol{\Psi}_{\mathcal{I}}^\top)$$

as $n \to \infty$, where $\boldsymbol{\Psi}_{\mathcal{I}} = (\mathbf{I}_{|\mathcal{I}|}, -\boldsymbol{\gamma}_{0\mathcal{I}}^{\top})$ and $\mathbf{I}_{|\mathcal{I}|}$ is the identity matrix of dimension $|\mathcal{I}|$. Under the null hypothesis, the asymptotic variance matrix can be consistently estimated by

$$\widehat{\boldsymbol{\Psi}}_{\mathcal{I}} \widehat{\boldsymbol{\Sigma}}((\boldsymbol{\theta}_0^{\top}, \widehat{\boldsymbol{\beta}}_{\mathcal{I}^c}^{\top})^{\top}) \widehat{\boldsymbol{\Psi}}_{\mathcal{I}}^{\top}$$

where $\widehat{\boldsymbol{\Psi}}_{\mathcal{I}} = (\mathbf{I}_{|\mathcal{I}|}, -\widehat{\boldsymbol{\gamma}}_{\mathcal{I}}^{\top})$. As a consequence, the test statistic is constructed as

$$\widehat{\chi}_n^2 = n \widehat{\mathbf{G}}_{n\mathcal{I}}^{\mathrm{o}}(\boldsymbol{\theta}_0 \mid \widehat{\boldsymbol{\beta}}_{\mathcal{I}^c})^{\top} \{\widehat{\boldsymbol{\Psi}}_{\mathcal{I}} \widehat{\boldsymbol{\Sigma}}((\boldsymbol{\theta}_0^{\top}, \widehat{\boldsymbol{\beta}}_{\mathcal{I}^c}^{\top})^{\top}) \widehat{\boldsymbol{\Psi}}_{\mathcal{I}}^{\top}\}^{-1} \widehat{\mathbf{G}}_{n\mathcal{I}}^{\mathrm{o}}(\boldsymbol{\theta}_0 \mid \widehat{\boldsymbol{\beta}}_{\mathcal{I}^c}),$$

which asymptotically follows the Chi-squared distribution with the degree of freedom $|\mathcal{I}|$ under the null hypothesis.

4 Simulation Studies

We conduct simulation studies to investigate the finite-sample performance of the proposed method. For comparison, we consider the naive method that ignores measurement errors by treating $\mathbf{W}$ as error-free. As a benchmark, we further consider the oracle but practically infeasible method that presumes the underlying $\mathbf{Z}$ to be known. For both the naive and oracle methods, the decorrelated score in Ning and Liu (2017) is applied to obtain the one- and two-step estimators. We generate the failure time T from the additive hazards model,

$$\lambda(t|\mathbf{Z}) = 2t + 2 + \mathbf{Z}^{\top}\boldsymbol{\beta}_0,$$

where $\boldsymbol{\beta}_0 = (1, 1, -1, -1, \mathbf{0}^{\top})^{\top}$ with the first four regression coefficients nonzero and the remaining zero, and $\mathbf{Z}$ is generated from a p-variate normal distribution with mean zero and covariance matrix $(0.5^{|i-j|})_{i,j=1}^{p}$. The measurement error $\mathbf{U}$ is also generated from a zero-mean normal distribution with covariance matrix $\mathbf{V}$ and $\mathbf{W} = \mathbf{Z} + \mathbf{U}$. We set $\mathbf{V} = 0.25\mathbf{I}_p$ to guarantee the elementwise signal-to-noise ratio to be 0.8, where $\mathbf{I}_p$ is the identity matrix of size p. We take the censoring time $C = \min(\widetilde{C}, \tau)$, where $\widetilde{C}$ is generated from $U(0, \widetilde{\tau})$, and τ and $\widetilde{\tau}$ ($\tau < \widetilde{\tau}$) are chosen to yield a censoring rate of around 30%. We consider sample size $n = 200$, coupled with $p = 100$, 200, and 400, respectively. We test two hypotheses as follows: for the first signal covariate,

$$H_0 : \beta_{01} = 1 \quad \text{versus} \quad H_1 : \beta_{01} = \theta_1 \ (\theta_1 \neq 1), \tag{4.1}$$

and for the 34th non-signal one,

$$H_0 : \beta_{0(34)} = 0 \quad \text{versus} \quad H_1 : \beta_{0(34)} = \theta_{34} \ (\theta_{34} \neq 0). \tag{4.2}$$

For each configuration, we repeat 1000 simulations.

We apply the 10-fold CV procedure to select the tuning parameters along the diagonal line. In particular, we partition the prespecified interval equally on the log-scale

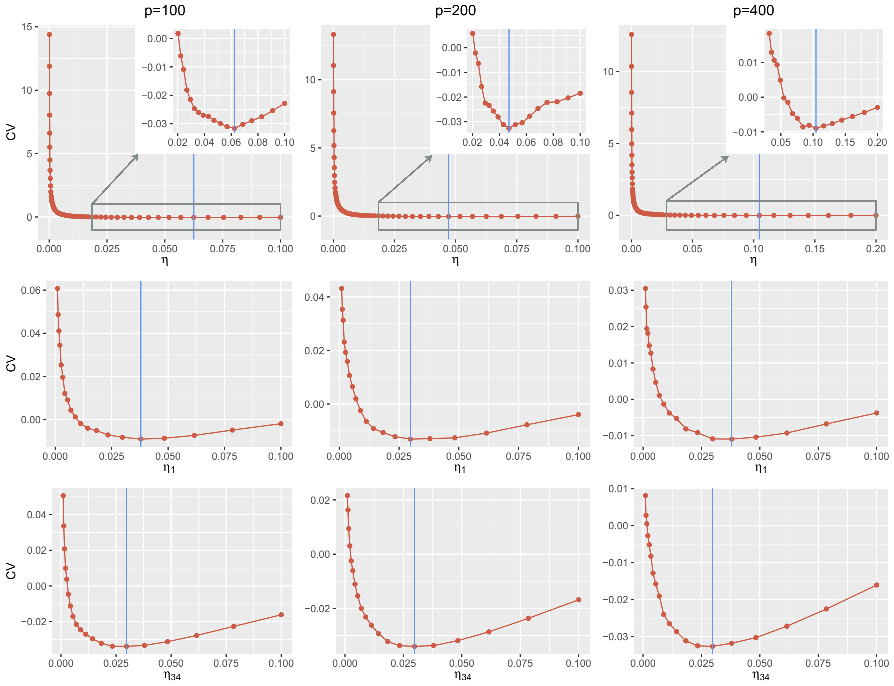

Fig. 1 Paths of the cross-validation (CV) procedure along the diagonal line where the minimum point is indicated by the vertical line

level and then identify which grid point corresponds to the minimum CV. As shown in Fig. 1, the proposed CV procedure can attain the optimum along the diagonal direction, demonstrating the practical feasibility of our strategy. We choose the tuning parameters for the naive and oracle methods by developing a similar 10-fold CV procedure. The proposed one-step and two-step bias corrected estimators, $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$, can be respectively obtained from (2.3) and (3.9), whereas the counterparts of the naive and oracle methods are obtained from (1.1) and (2.13) in Ning and Liu (2017). In Table 1, the columns labeled "$\widehat{\beta}_k$" and "$\widetilde{\beta}_k$" with sub-indices $k = 1$ and 34 are the averages of the one- and two-step bias corrected estimators, respectively; "SE" is the sample standard error of the two-step bias corrected estimators; "ESE" is the average of estimates of the corresponding standard errors; and "CP" is the coverage probability of 95% confidence intervals. For the signal covariate, the proposed estimator with double bias correction is indeed able to correct the bias, while the one-step bias correction estimator is still biased. Obviously, the naive method is seriously biased whereas the oracle method can effectively correct the bias of the one-step estimator. In contrast, for the non-signal covariate, the one-step bias corrected estimator is sufficient to reduce the bias under three approaches partially because the Lasso penalty in general leads to the shrunk estimator. Although the coverage probabilities of the proposed method are lower than the nominal level 95% for the signal covariate, the results are comparable

Table 1 Simulation results based on 1000 replications of the one-step ($\widehat{\beta}_k$) and two-step ($\widetilde{\beta}_k$) bias corrected estimators ($k = 1$ and 34) for the proposed, naive and oracle methods with sample size $n = 200$ and censoring rate 30%

Method	p	True value $\beta_{01} = 1$					
		$\widehat{\beta}_1$	SE	$\widetilde{\beta}_1$	SE	ESE	CP(%)
Proposed	100	0.228	0.210	0.905	0.357	0.304	90.8
	200	0.215	0.241	0.906	0.350	0.301	89.4
	400	0.092	0.143	0.931	0.358	0.314	88.7
Naive	100	0.273	0.255	0.658	0.249	0.229	62.5
	200	0.238	0.239	0.676	0.261	0.236	68.2
	400	0.133	0.193	0.599	0.226	0.230	57.1
Oracle	100	0.546	0.300	0.985	0.302	0.283	93.4
	200	0.467	0.315	0.955	0.311	0.275	92.2
	400	0.321	0.280	0.917	0.290	0.283	92.8
Method	p	True value $\beta_{0(34)} = 0$					
		$\widehat{\beta}_{34}$	SE	$\widetilde{\beta}_{34}$	SE	ESE	CP(%)
Proposed	100	0.000	0.024	−0.007	0.298	0.282	94.0
	200	0.002	0.042	−0.034	0.310	0.286	95.4
	400	0.001	0.011	0.008	0.326	0.293	95.9
Naive	100	0.001	0.031	0.010	0.206	0.216	96.7
	200	0.000	0.026	−0.013	0.199	0.220	97.2
	400	−0.002	0.030	0.007	0.225	0.214	94.6
Oracle	100	−0.001	0.052	−0.004	0.253	0.266	96.2
	200	−0.003	0.032	−0.003	0.237	0.256	97.0
	400	−0.001	0.018	0.016	0.259	0.262	95.8

to that of the oracle method. This situation can be improved by increasing the sample size or decreasing the censoring rate.

We corroborate the result of Theorem 5 with hypothesis tests of (4.1) and (4.2), accompanying with the oracle method. It can be seen from Table 2 that the sizes of the proposed and oracle testing procedures are similar; both are close to 5%. On the other hand, we explore the power of the two testing procedures by gradually enlarging the gap of the alternative hypothesis from the null. The power is enhanced accordingly as shown in Fig. 2, although the proposed method gains the power at a slightly lower rate than the oracle method when there is little discrimination between the null and alternative hypotheses. Besides the 1-st and 34-th coordinates, we randomly select the 3-rd, 4-th, 6-th, and 400-th ones to assess the normal approximation of Theorem 5 as well as the oracle method. The histograms of standardized testing statistics in Fig. 3 indicate the proposed and oracle testing statistics can almost equally and adequately approximate the normal distribution, yielding valid inferential procedures.

We further evaluate the finite-sample performance of the inference procedure for multiple parameters discussed in Sect. 3.2. We consider sample size $n = 200$, coupled

Table 2 Size and power of the proposed and oracle testing procedures (θ_k is the value of β_{0k} under the alternative hypothesis with $k = 1$ and 34 and the null values $\beta_{01} = 1$ and $\beta_{0(34)} = 0$) at the significant level of 5% based on 1000 simulation replications

Method	p	Size	Power of test (4.1)					
			$\theta_1 = 0$	$\theta_1 = 0.5$	$\theta_1 = 1.5$	$\theta_1 = 2$	$\theta_1 = 2.5$	$\theta_1 = 3$
Proposed	100	0.080	0.874	0.432	0.217	0.508	0.779	0.877
	200	0.065	0.864	0.401	0.200	0.511	0.718	0.880
	400	0.074	0.850	0.427	0.168	0.522	0.751	0.886
Oracle	100	0.066	0.959	0.531	0.282	0.668	0.911	0.990
	200	0.078	0.975	0.583	0.293	0.689	0.928	0.989
	400	0.072	0.965	0.530	0.235	0.643	0.916	0.952
	p	Size	Power of test (4.2)					
			$\theta_{34} = 0.5$	$\theta_{34} = 1$	$\theta_{34} = 1.5$	$\theta_{34} = 2$	$\theta_{34} = 2.5$	$\theta_{34} = 3$
Proposed	100	0.060	0.212	0.499	0.802	0.912	0.973	0.979
	200	0.046	0.165	0.470	0.759	0.917	0.978	0.989
	400	0.041	0.198	0.514	0.798	0.930	0.988	0.998
Oracle	100	0.036	0.235	0.721	0.960	1.000	0.995	1.000
	200	0.035	0.242	0.621	0.948	0.987	0.999	0.996
	400	0.042	0.228	0.679	0.890	0.989	1.000	1.000

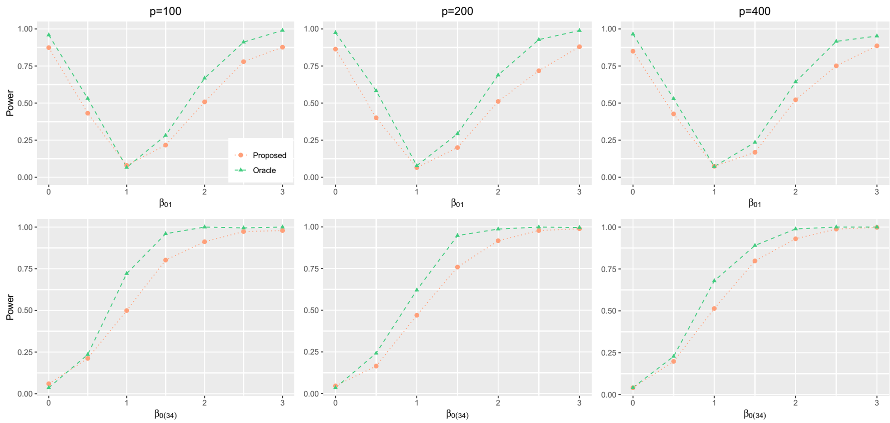

Fig. 2 Power curves of the proposed (dotted lines) and oracle (dashed lines) testing procedures for hypothesis tests in (4.1) and (4.2), respectively

with $p = 100, 200$, and 400, respectively. The remaining simulation setups are kept the same and the CV procedure for selecting the tuning parameters is developed likewise. Set $\mathcal{I} = \{1, 34\}$ and consider the hypothesis test,

$$H_0 : \boldsymbol{\beta}_{0\mathcal{I}} = (1, 0)^\top \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_{0\mathcal{I}} = \boldsymbol{\theta}_{\mathcal{I}} \quad (\boldsymbol{\theta}_{\mathcal{I}} \neq (1, 0)^\top). \tag{4.3}$$

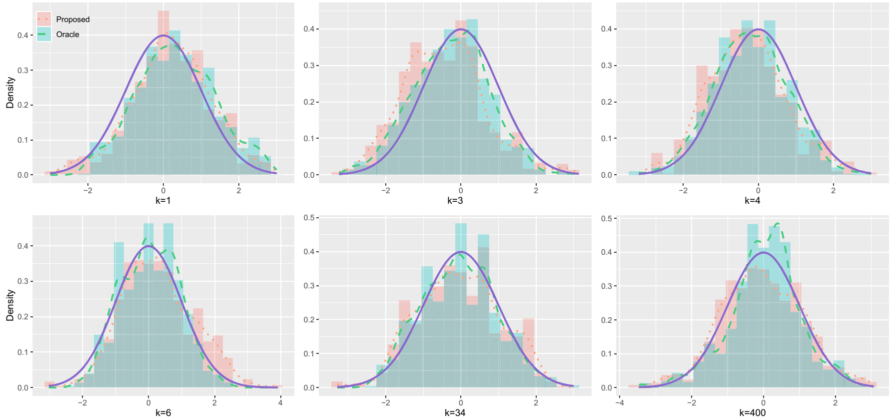

Fig. 3 The standard normal density curves (solid lines), histograms of standardized testing statistics, and the kernel smoothing curves for the proposed (dotted lines) and oracle (dashed lines) methods with $k = 1, 3, 4, 6, 34$ and 400, respectively

Table 3 Test size and power of the proposed testing procedure for multiple parameters ($\boldsymbol{\theta}_{\mathcal{I}}$ is the alternative value of $\boldsymbol{\beta}_{0\mathcal{I}}$ and the null value $\boldsymbol{\beta}_{0\mathcal{I}} = (1, 0)^\top$) at the significant level of 5% based on 1000 simulation replications

Method	p	Size	Power of test (4.3)				
			$\boldsymbol{\theta}_{\mathcal{I}} = (0, 0)^\top$	$\boldsymbol{\theta}_{\mathcal{I}} = (1, 1)^\top$	$\boldsymbol{\theta}_{\mathcal{I}} = (1, 2)^\top$	$\boldsymbol{\theta}_{\mathcal{I}} = (2, 2)^\top$	$\boldsymbol{\theta}_{\mathcal{I}} = (3, 3)^\top$
Proposed	100	0.075	0.796	0.575	0.941	0.912	0.972
	200	0.072	0.825	0.540	0.925	0.883	0.981
	400	0.065	0.833	0.578	0.951	0.859	0.969
Oracle	100	0.070	0.936	0.773	0.993	0.978	1.000
	200	0.084	0.960	0.772	1.000	0.972	1.000
	400	0.052	0.938	0.719	1.000	0.980	1.000

To explore the power of the proposed testing procedure, $\boldsymbol{\theta}_{\mathcal{I}}$ is set as $(0, 0)^\top$, $(1, 1)^\top$, $(1, 2)^\top$, $(2, 2)^\top$, and $(3, 3)^\top$, respectively. For each configuration, we repeat 1000 simulations. Table 3 shows that both the proposed and oracle inference procedures deliver promising and comparable performances.

5 Real Example

The China Health and Nutrition Survey (https://www.cpc.unc.edu/projects/china) is an ongoing open cohort and international collaborative project between the Carolina Population Center at the University of North Carolina at Chapel Hill and the National Institute for Nutrition and Health at the Chinese Center for Disease Control and Prevention. The survey aims to examine the effects of the social and economic transformation of Chinese society on the health and nutritional status of its population. As a subcohort study, we are primarily interested to investigate the physical fitness, living and studying behaviors as well as other confounding covariates on the health of adolescent girls, which is typically implied by the time to menarche. Early and delayed menarche is usually associated with some physical, psychological, or nutritional barrier. There are 63 eligible adolescent girls with age from 12 to 16 years old enrolled in the China Health and Nutrition Survey between 1993 and 2009. Their ages of menarche were recorded with a censoring rate of about 23.8%. There were 90 risk factors, including systolic blood pressure (SBP) and diastolic blood pressure (DBP), 25 biochemical items, 30 nutritional indexes, 13 urbanization indexes and other living and studying habits. The SBP and DBP were measured twice and 6 biochemical items were assessed by local and central laboratories in Beijing, China, respectively.

We apply our proposed method to quantify the effects of risk factors on the time to menarche. There were eight covariates prone to measurement errors and each was measured twice, and thus the corresponding coordinates of the covariance matrix of measurement errors can be estimated using the duplicated observations while the remaining coordinates are set as zero (Wang et al. 2012). The 10-fold CV procedure is adopted to select the tuning parameters along the diagonal line. We only report estimation results of the covariates that are shown to be significant under the proposed or naive method in Table 4. Among all mismeasured covariates, only our proposed method shows that albumin is significantly associated with the time to menarche. It also suggests that female adolescents with a higher level of low density lipoprotein cholesterol and a higher health score of the living community have a significantly higher risk of early menarche. Furthermore, there is a trend that a higher level of albumin, a higher market score of the living community and the habit of disliking for vegetables are associated with a higher risk of delayed menarche. The naive method shows that both Cholesterol and watching TV significantly affect the time to menarche.

6 Discussion

Both covariate measurement errors and regularization can result in estimation bias. A single-step bias correction due to covariate measurement errors or regularization is not adequate to completely remove the bias. In this article, we propose a double bias correction method for high-dimensional additive hazards model with covariate measurement errors. We establish the high-dimensional inferential procedure for the low-dimensional parameters by employing the empirical process theory and semiparametric efficient bound theory. As a byproduct, we obtain the convergence rate of the regularized one-step estimator that corrects the measurement errors in covariates, which has not been investigated before. Numerical results demonstrate the proposed method delivers reasonable finite-sample performances. However, our simulations indicate that it is a subtle issue of selecting the tuning parameters, for which we develop a CV procedure by searching for the suboptimum along the diagonal direction. Although such a strategy is plausible in practice, the result may not be globally optimal.

When the high-dimensional covariates are measured with error, Yan (2014) established the error bound for the Lasso estimator under the additive hazards model, where the non-convex surrogate loss is adopted for finding the local optima. We take

Table 4 Point estimation, confidence interval (CI) and hypothesis test for these covariates that are shown to be significant based on the proposed or naive method in the China Health and Nutrition Survey study

Covariate	Proposed method				
	$\widehat{\beta}$	$\widetilde{\beta}$	ESE	95% CI	p-value
Albumin	0.000	−0.022	0.011	(−0.043, −0.001)	0.041
Cholesterol	0.016	0.191	0.065	(0.064, 0.319)	0.006
Health	0.031	0.046	0.015	(0.017, 0.076)	0.009
Market	−0.012	−0.057	0.020	(−0.097, −0.018)	0.006
Vegetable	−0.027	−0.055	0.016	(−0.087, −0.024)	0.005
TV	0.000	−0.030	0.030	(−0.089, 0.029)	0.316
Covariate	Naive method				
	$\widehat{\beta}$	$\widetilde{\beta}$	ESE	95% CI	p-value
Albumin	0.000	0.065	0.065	(−0.063, 0.194)	0.318
Cholesterol	0.000	0.173	0.077	(0.022, 0.324)	0.025
Health	0.019	0.046	0.031	(−0.014, 0.107)	0.131
Market	0.000	−0.059	0.035	(−0.128, 0.009)	0.087
Vegetable	−0.023	−0.057	0.031	(−0.118, 0.004)	0.069
TV	0.000	−0.022	0.006	(−0.034, −0.010)	< 0.001

Cholesterol: Low density lipoprotein cholesterol, Health: Health score of the living community, Market: Modern market component score of the living community, Vegetable: Does the adolescent dislike vegetables? TV: Does the adolescent dislike watching TV?

a very different approach to establishing the error bound by following the idea of the Dantzig selector. Specifically, we investigate the tail probability of the corrected estimating function $\mathbf{G}_n(\boldsymbol{\beta}_0)$ through properly selecting x in Lemma 3, and then formulate the optimization problem of (2.2) by relaxing the strict zero-root constraint of (2.1) and expanding the feasible region. An additional term $\eta|\boldsymbol{\beta}|_1$ is introduced in (2.2) due to the measurement error. Furthermore, (2.2) is a convex optimization problem, thereby greatly facilitating the computation and theoretical analysis. As another main contribution, our work establishes the inference procedure for the components of high-dimensional regression coefficients, which is beyond the thesis of Yan (2014). In addition, the proposed inference procedure can test multiple components simultaneously using the Chi-squared test statistic based on the multivariate version of Theorem 5 (Mitra and Zhang 2016; Guo et al. 2021) as discussed in Sect. 3.2.

The measurement error $\mathbf{U}$ is typically assumed to be normally distributed in the literature. We relax the normal assumption by only imposing the second moment condition. However, the covariance matrix $\mathbf{V}$ of the measurement error $\mathbf{U}$ is assumed to be known. In fact, estimation of $\mathbf{V}$ based on (replicated) $\mathbf{W}$, even under a normal assumption, has its own right of interest and remains largely unexplored in the paradigm of high-dimensional settings (Donoho and Gavish 2014).

We have focused on the additive hazards model, while the issues under the Cox proportional hazards model are different and require new development. In the low-dimensional case and with normal measurement errors, it can only approximately correct the partial likelihood score of the Cox model because the exact correction does not exist (Nakamura 1992). The situation becomes much more challenging in the high-dimensional Cox model as the convergence rates of the approximately corrected score and oracle partial score need to be evaluated, which is the extra effort beyond our double bias correction method in the framework of additive hazards regression.

Supplementary Material

The online supplementary material contains Lemmas 1–10 and their technical proofs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10985-022-09568-2.

Acknowledgements The authors thank the Editor, the Associate Editor and two referees for their careful reviews of our manuscript and many insightful comments that have significantly improved the paper. Yin's research was partially supported by the Research Grants Council of Hong Kong (17308420), Huang's research was partially supported by the U.S. National Science Foundation (DMS-1916199), and Wu's research was partially supported by the National Natural Science Foundation of China (12071483).

Appendix: Proof of Theorems

Proof of Theorem 1 For ease of exposition, let c denote a generic positive constant that may be different from line to line. We first show that the true value $\boldsymbol{\beta}_0$ falls in the feasible region of the optimization problem (2.2), and then show that $\widehat{\boldsymbol{\beta}}$ satisfies the restricted eigenvalue condition as stated in Lemma 5. We finally conclude the proof by establishing variate concentration inequalities.

As $\eta|\boldsymbol{\beta}_0|_1 + \delta = O(a_0 L_{\max}^{2,1}\sqrt{\log p/n})$, the inequality $|\mathbf{G}_n(\boldsymbol{\beta}_0)|_\infty \le \eta|\boldsymbol{\beta}_0|_1 + \delta$ holds with probability tending to one by Lemma 3. For ease of exposition, our derivation is based on the event that $\boldsymbol{\beta}_0$ falls in the feasible region of the optimization problem (2.2).

Set $\widehat{\mathbf{h}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, and by the definition of $\widehat{\boldsymbol{\beta}}$, we have

$$\begin{aligned} |(\boldsymbol{\beta}_0)_{\mathcal{A}_0}|_1 + |(\boldsymbol{\beta}_0)_{\mathcal{A}_0^c}|_1 = |\boldsymbol{\beta}_0|_1 &\ge |\widehat{\boldsymbol{\beta}}|_1 = |\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}|_1 + |\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}|_1 \\ &= |\widehat{\mathbf{h}}_{\mathcal{A}_0} + (\boldsymbol{\beta}_0)_{\mathcal{A}_0}|_1 + |\widehat{\mathbf{h}}_{\mathcal{A}_0^c} + (\boldsymbol{\beta}_0)_{\mathcal{A}_0^c}|_1 \\ &\ge |(\boldsymbol{\beta}_0)_{\mathcal{A}_0}|_1 - |(\boldsymbol{\beta}_0)_{\mathcal{A}_0^c}|_1 - |\widehat{\mathbf{h}}_{\mathcal{A}_0}|_1 + |\widehat{\mathbf{h}}_{\mathcal{A}_0^c}|_1. \end{aligned} \tag{A.1}$$

Hence, (A.1) along with the definition of set $\mathcal{A}_0$ yields that $|\widehat{\mathbf{h}}_{\mathcal{A}_0}|_1 \ge |\widehat{\mathbf{h}}_{\mathcal{A}_0^c}|_1$ and thus

$$\widehat{\mathbf{h}}^\top \mathbf{B}_n^* \widehat{\mathbf{h}} \ge \zeta \widehat{\mathbf{h}}^\top \widehat{\mathbf{h}} \tag{A.2}$$

following Lemma 5.

Simple calculation entails that

$$\begin{aligned}\mathbf{B}_n^*\widehat{\mathbf{h}} &= \left(\mathbf{D}_n\widehat{\mathbf{h}} - \frac{\tau}{n}\mathbf{V}\widehat{\mathbf{h}}\right) - \frac{1}{n}\sum_{i=1}^n\int_0^\tau \{\mathbf{U}_i - \overline{\mathbf{U}}(t)\}\mathbf{Z}_i^\top\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t \\ &\quad -\frac{1}{n}\sum_{i=1}^n\int_0^\tau \mathbf{Z}_i\{\mathbf{U}_i - \overline{\mathbf{U}}(t)\}^\top\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t - \frac{1}{n}\sum_{i=1}^n\int_0^\tau [\mathbf{U}_i\{\mathbf{U}_i - \overline{\mathbf{U}}(t)\}^\top - \mathbf{V}]\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t \\ &\equiv (\mathbf{I}_1) - (\mathbf{I}_2) - (\mathbf{I}_3) - (\mathbf{I}_4),\end{aligned}$$

where $(\mathbf{I}_k),\ k = 1, \ldots, 4$, corresponds to each of the above four terms.

Considering term $(\mathbf{I}_1)$, by Lemma 3, we have

$$\left|\mathbf{G}_n(\boldsymbol{\beta}_0) - \frac{\tau}{n}\mathbf{V}\boldsymbol{\beta}_0\right|_\infty = O_P(a_0 L_{\max}^{2,1}\sqrt{\log p/n}).$$

Therefore,

$$\begin{aligned}|(\mathbf{I}_1)|_\infty &\le \left|\mathbf{G}_n(\boldsymbol{\beta}_0) - \frac{\tau}{n}\mathbf{V}\boldsymbol{\beta}_0\right|_\infty + \left|\mathbf{G}_n(\widehat{\boldsymbol{\beta}}) - \frac{\tau}{n}\mathbf{V}\widehat{\boldsymbol{\beta}}\right|_\infty \\ &\le O_P(a_0 L_{\max}^{2,1}\sqrt{\log p/n}) + \eta|\boldsymbol{\beta}_0|_1 + \delta + \left|\frac{\tau}{n}\mathbf{V}\widehat{\boldsymbol{\beta}}\right|_\infty \\ &= O_P(a_0 L_{\max}^{2,1}\sqrt{\log p/n}).\end{aligned} \tag{A.3}$$

Considering term $(\mathbf{I}_2)$, let U_{ij} denote the j-th component of $\mathbf{U}_i$ and Z_{ik} denote the k-th component of $\mathbf{Z}_i$. Noting that

$$|\widehat{\mathbf{h}}|_1 \le |\boldsymbol{\beta}_0|_1 + |\widehat{\boldsymbol{\beta}}|_1 \le 2|\boldsymbol{\beta}_0|_1 \le 2a_0$$

and using the Hoeffding inequality, we have

$$\begin{aligned}&P\left(\left|\frac{1}{n}\sum_{i=1}^n\int_0^\tau \mathbf{U}_i\mathbf{Z}_i^\top\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t\right|_\infty \ge cn^{-1/2}x \mid \Omega_L\right) \\ &\le P\left(\left|\frac{1}{n}\sum_{i=1}^n\int_0^\tau \mathbf{U}_i\mathbf{Z}_i^\top Y_i(t)\mathrm{d}t\right|_{\max} \left|\widehat{\mathbf{h}}\right|_1 \ge cn^{-1/2}x \mid \Omega_L\right) \\ &\le \sum_{j=1}^p\sum_{k=1}^p P\left(\left|\frac{1}{n}\sum_{i=1}^n\int_0^\tau U_{ij}Z_{ik}Y_i(t)\mathrm{d}t\right| \ge \frac{cn^{-1/2}x}{2a_0} \mid \Omega_L\right) \\ &\le 2p^2\exp\left(-\frac{cx^2}{a_0^2L^4}\right).\end{aligned} \tag{A.4}$$

Lemma 2 implies that

$$P\left(\sup_{t\in[0,\tau]}\left|\overline{\mathbf{U}}(t)\right|_\infty \geq cn^{-1/2}(L+x) \mid \Omega_L\right) \leq 3p\exp\left\{-c\min\left(n, \frac{x^2}{L^2}\right)\right\}. \tag{A.5}$$

Therefore, under conditions C1 and C2 and based on (A.4) and (A.5), we obtain

$$\begin{aligned} &P(|(\mathbf{I}_2)|_\infty \geq cn^{-1/2}(a_0L^2+x) \mid \Omega_L) \\ &\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n\int_0^\tau \mathbf{U}_i\mathbf{Z}_i^\top\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t\right|_\infty \geq \frac{cn^{-1/2}x}{2} \mid \Omega_L\right) \\ &\quad+P\left(\sup_{t\in[0,\tau]}\left|\overline{\mathbf{U}}(t)\right|_\infty\int_0^\tau\left|\frac{1}{n}\sum_{i=1}^n\mathbf{Z}_i^\top\widehat{\mathbf{h}}Y_i(t)\right|\mathrm{d}t \geq cn^{-1/2}\left(a_0L^2+\frac{x}{2}\right) \mid \Omega_L\right) \\ &\leq 5p^2\exp\left\{-c\min\left(n, \frac{x^2}{a_0^2L^4}\right)\right\}. \end{aligned} \tag{A.6}$$

Considering term ($\mathbf{I}_3$), based on symmetry with respect to the term ($\mathbf{I}_2$), we can calculate the tail bound of ($\mathbf{I}_3$) as

$$P(|(\mathbf{I}_3)|_\infty \geq cn^{-1/2}(a_0L^2+x) \mid \Omega_L) \leq 5p^2\exp\left\{-c\min\left(n, \frac{x^2}{a_0^2L^4}\right)\right\}. \tag{A.7}$$

Considering term ($\mathbf{I}_4$), we rewrite it as

$$(\mathbf{I}_4) = \frac{1}{n}\sum_{i=1}^n\int_0^\tau\left(\mathbf{U}_i^{\otimes 2}-\mathbf{V}\right)\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t + \frac{1}{n}\sum_{i=1}^n\int_0^\tau\mathbf{U}_i\overline{\mathbf{U}}(t)^\top\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t.$$

Under conditions C1 and C2, the tail bound for the first term of ($\mathbf{I}_4$) can be obtained by using similar arguments to (A.4), which is given by

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n\int_0^\tau\left(\mathbf{U}_i^{\otimes 2}-\mathbf{V}\right)\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t\right|_\infty \geq cn^{-1/2}x \mid \Omega_L\right) \leq 2p^2\exp\left(-\frac{cx^2}{a_0^2L_{\max}^{4,2}}\right).$$

Likewise, under conditions C1 and C2, we have

$$\begin{aligned} &P\left(\left|\frac{1}{n}\sum_{i=1}^n\int_0^\tau\mathbf{U}_i\overline{\mathbf{U}}(t)^\top\widehat{\mathbf{h}}Y_i(t)\mathrm{d}t\right|_\infty \geq cn^{-1/2}(a_0L^2+x) \mid \Omega_L\right) \\ &\leq 3p\exp\left\{-c\min\left(n, \frac{x^2}{a_0^2L^4}\right)\right\}. \end{aligned}$$

As a result,

$$P\left(|(\mathbf{I}_4)|_\infty \geq cn^{-1/2}(a_0L^2+x) \mid \Omega_L\right) \leq 5p^2\exp\left\{-c\min\left(n, \frac{x^2}{a_0^2L_{\max}^{4,2}}\right)\right\}. \tag{A.8}$$

Combining (A.3), (A.6), (A.7) and (A.8), we arrive at $|\mathbf{B}_n^*\widehat{\mathbf{h}}|_\infty = O_P(L_{\max}^{2,1}\sqrt{\log p/n})$. It follows from condition C3 and (A.2) that

$$|\widehat{\mathbf{h}}|_1^2 \leq 4|\widehat{\mathbf{h}}_{\mathcal{A}_0}|_1^2 \leq 4s_0|\widehat{\mathbf{h}}_{\mathcal{A}_0}|_2^2 \leq 4s_0|\widehat{\mathbf{h}}|_2^2 \leq 4s_0\zeta^{-1}\widehat{\mathbf{h}}^\top\mathbf{B}_n^*\widehat{\mathbf{h}} \leq 4s_0\zeta^{-1}|\widehat{\mathbf{h}}|_1|\mathbf{B}_n^*\widehat{\mathbf{h}}|_\infty,$$

which implies that

$$|\widehat{\mathbf{h}}|_1 = O_P(s_0L_{\max}^{2,1}\sqrt{\log p/n}),$$

and

$$|\widehat{\mathbf{h}}|_2 = O_P(s_0^{1/2}L_{\max}^{2,1}\sqrt{\log p/n}).$$

Thus, the proof is completed. □

Proof of Theorem 2 It follows the definition of $\boldsymbol{\gamma}_{0k}$ and Lemma 4 that

$$\begin{aligned}|(\mathbf{D}_n)_{-k,-k}\boldsymbol{\gamma}_{0k} - (\mathbf{D}_n)_{-k,k}|_\infty &\leq |\mathbf{B}^*_{-k,-k} - (\mathbf{D}_n)_{-k,-k}|_{\max}|\boldsymbol{\gamma}_{0k}|_1 + |\mathbf{B}^*_{-k,k} - (\mathbf{D}_n)_{-k,k}|_\infty \\ &= O_P(a_kL_{\max}^{2,1}\sqrt{\log p/n}),\end{aligned}$$

where we denote $a_k = |\boldsymbol{\gamma}_{0k}|_1$. As $\eta_k|\boldsymbol{\gamma}_{0k}|_1 + \delta_k = O(a_kL_{\max}^{2,1}\sqrt{\log p/n})$, the inequality $|(\mathbf{D}_n)_{-k,-k}\boldsymbol{\gamma}_{0k} - (\mathbf{D}_n)_{-k,k}|_\infty \leq \eta_k|\boldsymbol{\gamma}_{0k}|_1 + \delta_k$ holds with probability tending to one. Thus, we consider $\boldsymbol{\gamma}_{0k}$ falling in the feasible region of the optimization problem (3.4).

The definitions of $\widehat{\boldsymbol{\gamma}}_k$ and $\boldsymbol{\gamma}_{0k}$ imply

$$|\widehat{\boldsymbol{\gamma}}_k|_1 = |(\widehat{\boldsymbol{\gamma}}_k)_{\mathcal{A}_k}|_1 + |(\widehat{\boldsymbol{\gamma}}_k)_{\mathcal{A}_k^c}|_1 \leq |\boldsymbol{\gamma}_{0k}|_1 = |(\boldsymbol{\gamma}_{0k})_{\mathcal{A}_k}|_1.$$

We denote $\widehat{\mathbf{h}}_k = \boldsymbol{\gamma}_{0k} - \widehat{\boldsymbol{\gamma}}_k$ and then $|(\widehat{\boldsymbol{\gamma}}_k)_{\mathcal{A}_k}|_1 \geq |(\boldsymbol{\gamma}_{0k})_{\mathcal{A}_k}|_1 - |(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k}|_1$ using the triangular inequality. Immediately, we have $|(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k}|_1 \geq |(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k^c}|_1$, following which and Lemma 6, we further have

$$\zeta_k \leq \frac{\widehat{\mathbf{h}}_k^\top(\mathbf{B}_n^*)_{-k,-k}\widehat{\mathbf{h}}_k}{|\widehat{\mathbf{h}}_k|_2^2}. \tag{A.9}$$

The Cauchy–Schwarz inequality yields $|(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k}|_1 \leq \sqrt{s_k}|(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k}|_2$. As a result,

$$|\widehat{\mathbf{h}}_k|_1 = |(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k}|_1 + |(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k^c}|_1 \leq 2|(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k}|_1 \leq 2\sqrt{s_k}|(\widehat{\mathbf{h}}_k)_{\mathcal{A}_k}|_2 \leq 2\sqrt{s_k}|\widehat{\mathbf{h}}_k|_2.$$

Furthermore, combining the triangle inequality and Lemma 4, we obtain

$$\begin{aligned}|(\mathbf{B}_n^*)_{-k,-k}\widehat{\mathbf{h}}_k|_\infty &\le |(\mathbf{B}_n^*)_{-k,-k} - (\mathbf{D}_n)_{-k,-k}|_{\max}|\widehat{\boldsymbol{\gamma}}_k|_1 + |(\mathbf{D}_n)_{-k,k} - (\mathbf{B}^*)_{-k,k}|_\infty \\ &\quad + |(\mathbf{B}_n^*)_{-k,-k} - (\mathbf{B}^*)_{-k,-k}|_{\max}|\boldsymbol{\gamma}_{0k}|_1 + |(\mathbf{D}_n)_{-k,-k}\widehat{\boldsymbol{\gamma}}_k - (\mathbf{D}_n)_{-k,k}|_\infty \\ &= O_P(a_k L_{\max}^{2,1}\sqrt{\log p/n}).\end{aligned}$$

Consequently, there exists a constant $c > 0$ such that

$$\widehat{\mathbf{h}}_k^\top(\mathbf{B}_n^*)_{-k,-k}\widehat{\mathbf{h}}_k \le |(\mathbf{B}_n^*)_{-k,-k}\widehat{\mathbf{h}}_k|_\infty|\widehat{\mathbf{h}}_k|_1 \le c a_k L_{\max}^{2,1}\sqrt{\frac{s_k \log p}{n}}|\widehat{\mathbf{h}}_k|_2,$$

which along with (A.9) yields that

$$|\widehat{\mathbf{h}}_k|_2 = O_P\left(L_{\max}^{2,1}\sqrt{\frac{s_k \log p}{n}}\right)$$

and thus

$$|\widehat{\mathbf{h}}_k|_1 = O_P\left(s_k L_{\max}^{2,1}\sqrt{\frac{\log p}{n}}\right).$$

This completes the proof. □

Proof of Theorem 3 We rewrite

$$\begin{aligned}\sqrt{n}\widehat{\boldsymbol{\varphi}}_k^\top\mathbf{G}_n(\Pi_{k,\theta_0}(\widehat{\boldsymbol{\beta}})) &= \sqrt{n}\widehat{\boldsymbol{\varphi}}_k^\top\{\mathbf{G}_n(\Pi_{k,\theta_0}(\widehat{\boldsymbol{\beta}})) - \mathbf{G}_n(\boldsymbol{\beta}_0)\} \\ &\quad + \sqrt{n}(\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k)^\top\mathbf{G}_n(\boldsymbol{\beta}_0) + \sqrt{n}\boldsymbol{\varphi}_k^\top\mathbf{G}_n(\boldsymbol{\beta}_0).\end{aligned}$$

Using the definition of $\widehat{\boldsymbol{\gamma}}_k$ and Theorem 1 and following some basic algebraic calculations, we have

$$\begin{aligned}|\sqrt{n}\widehat{\boldsymbol{\varphi}}_k^\top\{\mathbf{G}_n(\Pi_{k,\theta_0}(\widehat{\boldsymbol{\beta}})) - \mathbf{G}_n(\boldsymbol{\beta}_0)\}| &= |\sqrt{n}\widehat{\boldsymbol{\varphi}}_k^\top\mathbf{D}_n(\Pi_{k,\theta_0}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_0)| \\ &\le \sqrt{n}|(\mathbf{D}_n)_{-k,-k}\widehat{\boldsymbol{\gamma}}_k - (\mathbf{D}_n)_{-k,k}|_\infty|\Pi_{k,\theta_0}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_0|_1 \\ &= O_P\left(s_0 L_{\max}^{4,2}\frac{\log p}{\sqrt{n}}\right).\end{aligned}$$

On the other hand, Theorem 2 and Lemma 3 imply that

$$\begin{aligned}\sqrt{n}|(\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k)^\top\mathbf{G}_n(\boldsymbol{\beta}_0)| &\le \sqrt{n}|\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k|_1|\mathbf{G}_n(\boldsymbol{\beta}_0)|_\infty \\ &= O_P\left(s_k L_{\max}^{4,2}\frac{\log p}{\sqrt{n}}\right).\end{aligned}$$

Therefore, Theorem 3 follows directly from the Slutsky theorem and Lemma 8. □

Proof of Theorem 4 Note that

$$
\begin{aligned}
&|\widehat{\sigma}_k^2 - \sigma_k^2| \\
&= |\widehat{\boldsymbol{\varphi}}_k^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\boldsymbol{\varphi}_k| \\
&\le |\widehat{\boldsymbol{\varphi}}_k^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k)| + |(\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\varphi}_k| + |\boldsymbol{\varphi}_k^\top \{\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}\boldsymbol{\varphi}_k| \\
&\le |\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})|_{\max}|\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k|_1|\widehat{\boldsymbol{\varphi}}_k|_1 + |\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})|_{\max}|\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k|_1|\boldsymbol{\varphi}_k|_1 \\
&\quad + |\boldsymbol{\varphi}_k^\top \{\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}\boldsymbol{\varphi}_k| \\
&\le \frac{4}{n}\sum_{i=1}^n |\mathbf{g}_i(\boldsymbol{\beta}_0)|_\infty^2 |\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k|_1|\boldsymbol{\varphi}_k|_1 + \frac{4}{n}\sum_{i=1}^n |\widehat{\mathbf{g}}_i(\widehat{\boldsymbol{\beta}}) - \mathbf{g}_i(\boldsymbol{\beta}_0)|_\infty^2 |\widehat{\boldsymbol{\varphi}}_k - \boldsymbol{\varphi}_k|_1|\boldsymbol{\varphi}_k|_1 \\
&\quad + |\boldsymbol{\varphi}_k^\top \{\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}\boldsymbol{\varphi}_k| \\
&= O_P\left(s_k L_{\max}^{6,3}\sqrt{\frac{\log p}{n}}\right) + O_P\left(s_0^2 s_k L_{\max}^{10,5}\left(\frac{\log p}{n}\right)^{3/2}\right) \\
&\quad + O_P\left(s_0 L_{\max}^{6,3}\sqrt{\frac{\log p}{n}}\right) \\
&= O_P\left(\max(s_0, s_k) L_{\max}^{6,3}\sqrt{\frac{\log p}{n}}\right)
\end{aligned}
$$

under condition $s_0 L_{\max}^{2,1}\sqrt{\log p/n} = o(1)$ and using Lemmas 9 and 10 in the online supplementary material. Thus, the proof is completed. □

References

Belloni A, Chernozhukov V, Kaul A (2017) Confidence bands for coefficients in high dimensional linear models with error-in-variables. *arXiv preprint*, arXiv:1703.00469. https://doi.org/10.48550/arXiv.1703.00469

Bühlmann P, van de Geer S (2011) Statistics for High-dimensional Data: Methods, Theory and Applications. Springer, New York

Candès EJ, Tao T (2007) The Dantzig selector: Statistical estimation when p is much larger than n. Ann Stat 35(6):2313–2351. https://doi.org/10.1214/009053606000001523

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement Error in Nonlinear Models: A Modern Perspective. Chapman and Hall Press, London

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey WK, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. Econom J 21:1–68. https://doi.org/10.1111/ectj.12097

Cox DR (1972) Regression models and life-tables (with discussion). J Roy Stat Soc B 34(2):187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Datta A, Zou H (2017) CoCoLasso for high-dimensional error-in-variables regression. Ann Stat 45(6):2400–2426. https://doi.org/10.1214/16-AOS1527

Donoho D, Gavish M (2014) Minimax risk of matrix denoising by singular value thresholding. Ann Stat 42(6):2413–2440. https://doi.org/10.1214/14-AOS1257

Fang EX, Ning Y, Liu H (2017) Testing and confidence intervals for high dimensional proportional hazards models. J Roy Stat Soc B 79(5):1415–1437. https://doi.org/10.1111/rssb.12224

Guo Z, Renaux C, Bühlmann P, Cai TT (2021) Group inference in high dimensions with applications to hierarchical testing. Electron J Stat 15(2):6633–6676. https://doi.org/10.1214/21-EJS1955

Huang J, Sun T, Ying Z, Yu Y, Zhang C-H (2013) Oracle inequalities for the Lasso in the Cox model. Ann Stat 41(3):1142–1165. https://doi.org/10.1214/13-AOS1098
Huang Y, Wang CY (2000) Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. J Am Stat Assoc 95(452):1209–1219. https://doi.org/10.1080/01621459.2000.10474321
Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 15(1):2869–2909. https://doi.org/10.5555/2627435.2697057
Johnson BA, Lin DY, Zeng D (2008) Penalized estimating functions and variable selection in semiparametric regression models. J Am Stat Assoc 103(482):672–680. https://doi.org/10.1198/016214508000000184
Kulich M, Lin DY (2000) Additive hazards regression with covariate measurement error. J Am Stat Assoc 95(449):238–248. https://doi.org/10.1080/01621459.2000.10473917
Lee JD, Sun DL, Sun Y, Taylor JE (2016) Exact post-selection inference with application to the Lasso. Ann Stat 44(3):907–927. https://doi.org/10.1214/15-AOS1371
Lin DY, Ying Z (1994) Semiparametric analysis of the additive risk model. Biometrika 81(1):61–71. https://doi.org/10.1093/biomet/81.1.61
Lin W, Lv J (2013) High-dimensional sparse additive hazards regression. J Am Stat Assoc 108(501):247–264. https://doi.org/10.1080/01621459.2012.746068
Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R (2014) A significance test for the Lasso. Ann Stat 42(2):413–468. https://doi.org/10.1214/13-AOS1175
Loh PL, Wainwright MJ (2012) High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. Ann Stat 40(3):1637–1664. https://doi.org/10.1214/12-AOS1018
Meinshausen N, Meier L, Bühlmann P (2009) P-values for high-dimensional regression. J Am Stat Assoc 104(488):1671–1681. https://doi.org/10.1198/jasa.2009.tm08647
Mitra R, Zhang C-H (2016) The benefit of group sparsity in group inference with de-biased scaled group lasso. Electron J Stat 10(2):1829–1873. https://doi.org/10.1214/16-EJS1120
Nakamura T (1992) Proportional hazards model with covariates subject to measurement error. Biom 48(3):829–838. https://doi.org/10.2307/2532348
Newey WK (1994) The asymptotic variance of semiparametric estimators. Econom 62(6):1349–1382. https://doi.org/10.2307/2951752
Neyman J (1959) Optimal asymptotic tests of composite statistical hypotheses. In: Grenander U (ed) Probability and Statistics. John Wiley & Sons, New York, pp 416–444
Ning Y, Liu H (2017) A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann Stat 45(1):158–195. https://doi.org/10.1214/16-AOS1448
Rosenbaum M, Tsybakov AB (2010) Sparse recovery under matrix uncertainty. Ann Stat 38(5):2620–2651. https://doi.org/10.1214/10-AOS793
Sørensen Ø, Hellton KH, Frigessi A, Thoresen M (2018) Covariate selection in high-dimensional generalized linear models with measurement error. J Comput Graph Stat 27(4):739–749. https://doi.org/10.1080/10618600.2018.1425626
Stefanski LA (1989) Unbiased estimation of a nonlinear function a normal mean with application to measurement error models. Commun Stat-Theory Methods 18(12):4335–4358. https://doi.org/10.1080/03610928908830159
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
van de Geer S, Bühlmann P, Ritov YA, Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Stat 42(3):1166–1202. https://doi.org/10.1214/14-AOS1221
Wainwright MJ (2019) High Dimensional Statistics: A Non-asymptotic Viewpoint. Cambridge University Press, New York
Wang H, Stefanski LA, Zhu Z (2012) Corrected-loss estimation for quantile regression with covariate measurement errors. Biom 99(2):405–421. https://doi.org/10.1093/biomet/ass005
Wasserman L, Roeder K (2009) High dimensional variable selection. Ann Stat 37(5A):2178–2201. https://doi.org/10.1214/08-AOS646
Yan Y (2014) Statistical Methods on Survival Data with Measurement Error. PhD Thesis, University of Waterloo. http://hdl.handle.net/10012/8699
Yan Y, Yi GY (2016) Analysis of error-prone survival data under additive hazards models: Measurement error effects and adjustments. Lifetime Data Anal 22:321–342. https://doi.org/10.1007/s10985-015-9340-1

Yu Y, Bradic J, Samworth RJ (2021) Confidence intervals for high-dimensional Cox models. Stat Sin 31:243–267. https://doi.org/10.5705/ss.202018.0247

Zhang C-H, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. J Roy Stat Soc B 76(1):217–242. https://doi.org/10.1111/rssb.12026

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.