Title	Confidence intervals for high-dimensional Cox models
Manuscript ID	SS-2018-0247
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0247
Complete List of Authors	Yi Yu Jelena Bradic and Richard J. Samworth
Corresponding Author	Richard J. Samworth
E-mail	r.samworth@statslab.cam.ac.uk

Statistica Sinica

CONFIDENCE INTERVALS FOR HIGH-DIMENSIONAL COX MODELS

Yi Yu[1], Jelena Bradic[2] and Richard J. Samworth[3]

[1] *School of Mathematics, University of Bristol*

[2] *Department of Mathematics, University of California at San Diego*

[3] *Statistical Laboratory, University of Cambridge*

Abstract: We provide theoretical justification for post-selection inference in high-dimensional Cox models, based on the celebrated debiased Lasso procedure (e.g. Zhang and Zhang, 2014; van de Geer et al., 2014). Our generic model setup allows time-dependent covariates and an unbounded time interval, which is unique among post-selection inference studies on high-dimensional survival analysis. In addition, we adopt a novel proof technique to replace the use of Rebolledo's central limit theorem as in the seminal work of Andersen and Gill (1982). Our theoretical results, which provide conditions under which our confidence intervals are asymptotically valid, are supported by extensive numerical experiments.

Key words and phrases: Survival analysis; High-dimension statistical inference; Debiased Lasso.

1. Introduction

Over the last 45 years, the Cox proportional hazards model (Cox, 1972) has become central to the analysis of censored survival data. It posits that the conditional hazard rate at time $t \in \mathcal{T}$ for the survival time $\tilde{T}$ of an individual given their p-variate covariate vector $\boldsymbol{Z}(t)$ can be expressed as

$$\lambda(t) := \lambda_0(t) \exp\big\{\boldsymbol{\beta}^{o\top}\boldsymbol{Z}(t)\big\}, \tag{1.1}$$

where $\boldsymbol{\beta}^o \in \mathbb{R}^p$ is an unknown vector of regression coefficients and $\lambda_0(\cdot)$ is an unknown baseline hazard function. With n individuals from a population, we assume that for each $i = 1, \ldots, n$ we observe a (possibly right-censored) survival time T_i, an indicator δ_i of whether or not failure is observed, and the corresponding covariate processes $\{\boldsymbol{Z}_i(t) : t \in \mathcal{T}\}$.

When $p < n$, the maximum partial likelihood estimator (MPLE) (Cox, 1975) may be used to estimate $\boldsymbol{\beta}^o$. In the classical setting where the dimension p is assumed to be fixed and the sample size n is allowed to diverge to infinity, and under a strong (and hard to check) condition on the weak convergence of the sample covariance processes, Andersen and Gill (1982) derived the asymptotic normality of the MPLE using counting process arguments and Rebolledo's martingale central limit theorem. This result may be used to provide asymptotically valid confidence intervals for components of $\boldsymbol{\beta}^o$ (or more generally, for linear combinations $\mathbf{c}^\top\boldsymbol{\beta}^o$, for some fixed $\mathbf{c} \in \mathbb{R}^p$).

Our interest in this paper lies in providing corresponding confidence intervals in the high-dimensional regime, where p may be much larger than n. The motivation for such methodology arises from many different application areas, but particularly in biomedicine, where Cox models are ubiquitous and data on each individual, which may arise in the form of combinations of genetic information, greyscale values for each pixel in a scan and many other types, are often plentiful. Our construction begins with the Lasso penalised partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ studied in Huang et al. (2013), which is used as an initial estimator and which is sparse. We then seek a sparse estimator of the inverse of negative Hessian matrix, which we will refer to as a *sparse precision matrix estimator*. In Zhang and Zhang (2014) and van de Geer et al. (2014), who consider similar problems in the linear and generalised linear model settings respectively, this sparse precision matrix estimator is constructed via nodewise Lasso regression (Meinshausen and Bühlmann, 2006). On the other hand, Javanmard and Montanari (2014) derived their precision matrix estimators by minimising the trace of the product of the sample covariance matrix and the precision matrix, and the covariates are assumed to be centred. However, in the Cox model setting, the counterpart of the design matrix is a mean-shifted design matrix, where the mean is based on a set of tilting weights, and this destroys the necessary independence structure. Instead, we adopt a modification of the CLIME estimator (Cai et al., 2011) as the sparse precision matrix estimator, which allows us to handle the mean subtraction. Adjusting $\widehat{\boldsymbol{\beta}}$ by the product of our sparse precision matrix estimator and the score vector yields a debiased estimator $\widehat{\boldsymbol{b}}$, and our main theoretical result (Theorem 1) provides conditions under which $\boldsymbol{c}^\top\widehat{\boldsymbol{b}}$ is asymptotically normally distributed around $\boldsymbol{c}^\top\boldsymbol{\beta}^o$. The desired confidence intervals can then be obtained straightforwardly. Further very recent applications of the debiasing idea, outside the regression problem context, can be found in e.g. Janková and van de Geer (2018).

The success of the debiased Lasso approach for high-dimensional post-selection inference means it has received a great deal of attention in recent years. However, this is the first attempt to provide theoretical justification for the debiased Lasso in the important area of survival analysis. In addition to this main contribution, we believe that our novel proof techniques can provide the survival analysis community with new tools that will be applicable in other related problems, and list three technical contributions below.

- We avoid the difficult assumption on the weak convergence of sample covariance processes inherent in the martingale central limit theorem approach (Bradic et al., 2011). This entails a completely different line of attack, which provides new insights even in the low-dimensional setting. In particular, we introduce a new finite-sample concentration inequality (Lemma S2), which controls the largest deviations of the weighted sample covariate process from its population analogue.

- We allow the upper limit t_+ of the time index set $\mathcal{T}$ to be infinite, and do not assume that each subject has a constant, positive probability of remaining in the at risk set at time t_+. This is in contrast to the work of, e.g., Fang et al. (2017), where the authors propose hypothesis tests based on decorrelated scores and decorrelated partial likelihood ratios. Since our concentration inequality mentioned above is only useful when sufficiently many individuals remain under study, this feature of the problem necessitates a novel truncation argument.

- Our theory handles settings where p may be much larger than n; in fact, we only assume that $p = o(\exp(n^a))$, for every $a > 0$; this is sometimes called the ultrahigh dimensional setting (e.g. Fan et al., 2009).

Our estimators and inference procedure are given in Section 2, and our theoretical arguments are presented in Section 3. Section 4 is devoted to extensive numerical studies of our methdology on both simulated and real data. These reveal in particular that valid p-values and confidence intervals for the noise variables can be obtained with a relatively small sample size, while a larger sample size is needed for good coverage of signal variables. Various auxiliary results and proofs are given in the Supplementary Material.

We conclude this introduction with some notation used throughout the paper. For any set S, let $|S|$ denote its cardinality. For a vector $\boldsymbol{v} = (v_1, \ldots, v_m)^\top \in \mathbb{R}^m$, let $\|\boldsymbol{v}\|_1$, $\|\boldsymbol{v}\|$ and $\|\boldsymbol{v}\|_\infty$ denote its ℓ_1, ℓ_2 and ℓ_∞ norms, respectively; we also write $\boldsymbol{v}^{\otimes 2} := \boldsymbol{v}\boldsymbol{v}^\top$. Given a set $J \subseteq \{1, \ldots, m\}$, we write $\boldsymbol{v}_J := (v_j)_{j \in J} \in \mathbb{R}^{|J|}$. For a matrix $\boldsymbol{A} = (A_{ij})_{i,j=1}^m \in \mathbb{R}^{m \times m}$, let $\|\boldsymbol{A}\|_\infty := \max_{i,j=1,\ldots,m} |A_{ij}|$ be the entrywise maximum absolute norm, and let $\|\boldsymbol{A}\|_{\mathrm{op},\infty} := \sup_{\boldsymbol{v} \neq 0}\bigl(\|\boldsymbol{A}\boldsymbol{v}\|_\infty / \|\boldsymbol{v}\|_\infty\bigr)$ and $\|\boldsymbol{A}\|_{\mathrm{op},1} := \sup_{\boldsymbol{v} \neq 0}\bigl(\|\boldsymbol{A}\boldsymbol{v}\|_1 / \|\boldsymbol{v}\|_1\bigr)$ denote its operator ℓ_∞ and operator ℓ_1 norms respectively. We recall in Lemma S1 in the Supplementary Material that $\|\boldsymbol{A}\|_{\mathrm{op},\infty}$ and $\|\boldsymbol{A}\|_{\mathrm{op},1}$ are, respectively, the maximum of the ℓ_1 norms of the rows of $\boldsymbol{A}$ and the maximum of the ℓ_1 norms of its columns. Given two real sequences (a_n) and (b_n), we write $a_n \asymp b_n$ to mean $0 < \liminf_{n \to \infty} |a_n / b_n| \leq \limsup_{n \to \infty} |a_n / b_n| < \infty$. Given a distribution function F, we write $\bar{F} := 1 - F$. All probabilities and expectations are taken under the true model with baseline hazard λ_0 and regression parameter $\boldsymbol{\beta}^o$, though we suppress this in our notation.

2. Methodology

Recall that $\mathcal{T} \subseteq [0, \infty)$ denotes our time index set. We assume that, for $i = 1, \ldots, n$, there exist independent triples $\big(\tilde{T}_i, U_i, \{\boldsymbol{Z}_i(t) : t \in \mathcal{T}\}\big)$, where $\tilde{T}_i$ is a non-negative random variable indicating failure time, U_i is a non-negative random variable indicating a censoring time, and $\{\boldsymbol{Z}_i(t) : t \in \mathcal{T}\}$ is a p-variate, predictable time-varying covariate process. We further assume that $\tilde{T}_i$ and U_i are conditionally independent given $\{\boldsymbol{Z}_i(t) : t \in \mathcal{T}\}$. Writing $T_i := \min(\tilde{T}_i, U_i)$ and $\delta_i := \mathbb{1}_{\{\tilde{T}_i \leq U_i\}}$, our observations are $\big\{\big(T_i, \delta_i, \{\boldsymbol{Z}_i(t) : t \in \mathcal{T}\}\big) : i = 1, \ldots, n\big\}$. We regard these observations as independent copies of a generic triple $\big(T, \delta, \{\boldsymbol{Z}(t) : t \in \mathcal{T}\}\big)$.

Let F_T denote the distribution function of T, and let $t_+ := \inf\{t \geq 0 : F_T(t) = 1\}$ denote the upper limit of the support of T. If $t_+ < \infty$, we assume that $\mathcal{T} = [0, t_+]$; if $t_+ = \infty$, then we assume $\mathcal{T} = [0, \infty)$. In this sense, we assume that $\mathcal{T}$ covers the entire support of the distribution of T, so in particular, there are no individuals in the risk set at time t_+.

For $i = 1, \ldots, n$, define processes $\{N_i(t) : t \in \mathcal{T}\}$ and $\{Y_i(t) : t \in \mathcal{T}\}$ by $N_i(t) := \mathbb{1}_{\{T_i \leq t, \delta_i = 1\}}$ and $Y_i(t) := \mathbb{1}_{\{T_i \geq t\}}$. We regard these as independent copies of processes $\{N(t) : t \in \mathcal{T}\}$ and $\{Y(t) : t \in \mathcal{T}\}$ respectively. Let

$\bar{N}(t) := n^{-1}\sum_{i=1}^{n} N_i(t)$. The natural σ-field at time $t \in \mathcal{T}$ is therefore $\mathcal{F}_t := \sigma\big(\{(N_i(t), Y_i(t), \{\boldsymbol{Z}_i(s) : s \in [0,t]\}) : i = 1,\ldots,n\}\big)$. The Cox model (1.1) entails that $N_i(t)$ has predictable compensator

$$\Lambda_i(t, \boldsymbol{\beta}^o) := \int_0^t Y_i(s)\exp\{\boldsymbol{\beta}^{o\top}\boldsymbol{Z}_i(t)\}\lambda_0(s)\,ds$$

with respect to the filtration $(\mathcal{F}_t : t \in \mathcal{T})$.

Define the log-partial likelihood function, divided by n, at $\boldsymbol{\beta} \in \mathbb{R}^p$ by

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \ell_n(\boldsymbol{\beta})\\ &:= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\boldsymbol{\beta}^\top\boldsymbol{Z}_i(s)\,dN_i(s) - \int_{\mathcal{T}}\log\Bigg[\sum_{j=1}^{n}Y_j(s)\exp\{\boldsymbol{\beta}^\top\boldsymbol{Z}_j(s)\}\Bigg]\,d\bar{N}(s).\end{aligned}$$

Inspired by Zhang and Zhang (2014) and van de Geer et al. (2014), our main object of interest is the one-step type estimator

$$\widehat{\boldsymbol{b}} := \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\dot{\ell}(\widehat{\boldsymbol{\beta}}), \tag{2.1}$$

where $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1,\ldots,\hat{\beta}_p)^\top$ is an initial estimator of $\boldsymbol{\beta}^o$, where $\widehat{\boldsymbol{\Theta}} = (\hat{\Theta}_{ij})_{i,j=1}^{p}$ is a sparse precision matrix estimator that approximates the inverse of the negative Hessian $-\ddot{\ell}(\boldsymbol{\beta}^o)$ and where $\dot{\ell}(\widehat{\boldsymbol{\beta}})$ is the score function evaluated at the initial estimator. In the rest of this section, we will elucidate the definition and rationale for our choices of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Theta}}$. We remark that our proposals for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Theta}}$ will depend on certain tuning parameters, and this dependence is suppressed in our notation. However, in our theoretical results, we will give explicit conditions on these tuning parameters. We remark that a similar construction has also been proposed in a later submission Kong et al. (2018), which focuses on the utility of such a construction under model misspecification.

2.1 Initial estimator

Following Huang et al. (2013), for $\lambda > 0$, let

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\lambda) := \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \bigl\{ -\ell(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \bigr\}. \tag{2.2}$$

For $i = 1, \ldots, n$ and $t \in \mathcal{T}$, let $\tilde{w}_i(t, \boldsymbol{\beta}) := Y_i(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\}$ be the ith weight and let

$$w_i(s, \boldsymbol{\beta}) := \frac{\tilde{w}_i(s, \boldsymbol{\beta})}{\sum_{j=1}^n \tilde{w}_j(s, \boldsymbol{\beta})}$$

be the ith normalised weight, with the convention that $0/0 := 0$. The weighted average of the covariate processes is defined by

$$\bar{\boldsymbol{Z}}(s, \boldsymbol{\beta}) := \sum_{i=1}^n \boldsymbol{Z}_i(s) w_i(s, \boldsymbol{\beta}).$$

Then it follows from the subgradient conditions for optimality (Karush–Kuhn–Tucker conditions) that there exists $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_p)^\top$ such that

$$0 = -\dot{\ell}(\widehat{\boldsymbol{\beta}}) + \lambda \widehat{\boldsymbol{\tau}} = -\frac{1}{n} \sum_{i=1}^n \int_{\mathcal{T}} \bigl\{ \boldsymbol{Z}_i(s) - \bar{\boldsymbol{Z}}(s, \widehat{\boldsymbol{\beta}}) \bigr\}\, dN_i(s) + \lambda \widehat{\boldsymbol{\tau}},$$

where $\|\widehat{\boldsymbol{\tau}}\|_\infty \leq 1$ and $\hat{\tau}_j = \operatorname{sgn}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$.

2.2 The estimator of the precision matrix

For $\boldsymbol{\beta} \in \mathbb{R}^p$, we have

$$\ddot{\ell}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \int_{\mathcal{T}} \{\boldsymbol{Z}_i(s) - \bar{\boldsymbol{Z}}(s,\boldsymbol{\beta})\}^{\otimes 2} w_i(s,\boldsymbol{\beta})\, d\bar{N}(s),$$

but the presence of the weights in this integral makes it hard to analyse directly. As a first step towards obtaining a more tractable expression, we therefore rewrite this equation as

$$\ddot{\ell}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} \{\boldsymbol{Z}_i(s) - \bar{\boldsymbol{Z}}(s,\boldsymbol{\beta})\}^{\otimes 2} \tilde{w}_i(s,\boldsymbol{\beta})\, d\widehat{\Lambda}(s,\boldsymbol{\beta}),$$

where we define $\widehat{\Lambda}(t,\boldsymbol{\beta}) := n\int_0^t \{\sum_{j=1}^n \tilde{w}_j(s,\boldsymbol{\beta})\}^{-1}\, d\bar{N}(s)$ to be the Breslow estimator of $\int_0^t \lambda_0(s)\, ds$ (Breslow, 1972). Now recall from, e.g., Andersen et al. (1993, p. 66) that the process $\{N(t) : t \in \mathcal{T}\}$ has the Doob–Meyer decomposition

$$N(t) = M(t) + \int_0^t \tilde{w}(s,\boldsymbol{\beta}^o)\lambda_0(s)\, ds, \tag{2.3}$$

where $\{M(t) : t \in \mathcal{T}\}$ is a mean-zero martingale. This motivates us to define a population approximation to $-\ddot{\ell}(\boldsymbol{\beta}^o)$ by

$$\begin{aligned}\boldsymbol{\Sigma} &:= \mathbb{E}\int_{\mathcal{T}} \{\boldsymbol{Z}(s) - \boldsymbol{\mu}(s,\boldsymbol{\beta}^o)\}^{\otimes 2}\, dN(s)\\ &= \mathbb{E}\int_0^{t_+} \{\boldsymbol{Z}(s) - \boldsymbol{\mu}(s,\boldsymbol{\beta}^o)\}^{\otimes 2} \tilde{w}(s,\boldsymbol{\beta}^o)\lambda_0(s)\, ds,\end{aligned}$$

where, for $t \in \mathcal{T}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\boldsymbol{\mu}(t, \boldsymbol{\beta}) := \frac{\mathbb{E}\{\boldsymbol{Z}(t) Y(t) \exp(\boldsymbol{\beta}^\top \boldsymbol{Z}(t))\}}{\mathbb{E}\{Y(t) \exp(\boldsymbol{\beta}^\top \boldsymbol{Z}(t))\}}.$$

Our goal in this subsection is to define an estimator of $\boldsymbol{\Sigma}^{-1}$ whose properties we can analyse. To this end, observe that an oracle, with knowledge of $\boldsymbol{\beta}^o$, could estimate $\boldsymbol{\Sigma}$ by

$$\begin{aligned}\widehat{\mathcal{V}}(\boldsymbol{\beta}^o) &:= \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{T}} \big\{\boldsymbol{Z}_i(s) - \bar{\boldsymbol{Z}}(s, \boldsymbol{\beta}^o)\big\}^{\otimes 2} \, dN_i(s) \\ &= \frac{1}{n} \sum_{i=1}^n \delta_i \big\{\boldsymbol{Z}_i(T_i) - \bar{\boldsymbol{Z}}(T_i, \boldsymbol{\beta}^o)\big\}^{\otimes 2}.\end{aligned}$$

This suggests the genuine estimator

$$\widehat{\mathcal{V}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^n \delta_i \big\{\boldsymbol{Z}_i(T_i) - \bar{\boldsymbol{Z}}(T_i, \widehat{\boldsymbol{\beta}})\big\}^{\otimes 2}. \tag{2.4}$$

While both $-\ddot{\ell}(\widehat{\boldsymbol{\beta}})$ and $\widehat{\mathcal{V}}(\widehat{\boldsymbol{\beta}})$ can be considered as estimators of $\boldsymbol{\Sigma}$, it turns out that the latter is the much more convenient expression to study from a theoretical perspective.

As mentioned in the introduction, both Zhang and Zhang (2014) and van de Geer et al. (2014) employ nodewise regression to obtain a sparse precision matrix estimator $\widehat{\boldsymbol{\Theta}}$. In those cases, the design matrices consist of independent rows, which facilitate the adoption of Lasso-type methods; in the Cox model, however, we do not have the luxury of row independence since $\widehat{\mathcal{V}}$ defined in (2.4) involves $\bar{\boldsymbol{Z}}(T_i, \widehat{\boldsymbol{\beta}})$.

As an alternative, we adapt the CLIME estimator of Cai et al. (2011), originally proposed in the context of precision matrix estimation. Let $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\Theta}}_1, \ldots, \widehat{\boldsymbol{\Theta}}_p)^\top$ be defined by

$$\widehat{\boldsymbol{\Theta}}_j \in \operatorname*{argmin}_{\boldsymbol{b}_j \in \mathbb{R}^p} \Big\{ \|\boldsymbol{b}_j\|_1 : \big\| \widehat{\mathcal{V}}(\widehat{\boldsymbol{\beta}})\boldsymbol{b}_j - \boldsymbol{e}_j \big\|_\infty \leq \lambda_n \Big\}, \tag{2.5}$$

where $\boldsymbol{e}_j^\top := (\mathbb{1}_{\{j=l\}})_{l=1}^p \in \mathbb{R}^p$ for $j = 1, \ldots, p$. The original proposal of Cai et al. (2011) symmetrised $\widehat{\boldsymbol{\Theta}}$ by taking both the (i,j)th and (j,i)th off-diagonal entries to be the corresponding entry of $\widehat{\boldsymbol{\Theta}}$ with smaller absolute value. In our theoretical analysis, it turned out to be convenient not to symmetrise in this way, and in practice, we found the the difference to be negligible; see Section 4.1.

For $j = 1, \ldots, p$, let $\dot{\ell}_j(\boldsymbol{\beta})$ denote the jth component of the score vector at $\boldsymbol{\beta}$, and let $\ddot{\ell}_j(\boldsymbol{\beta}) \in \mathbb{R}^p$ have lth component $\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_l \partial \beta_j}$. By a Taylor expansion, for each $j = 1, \ldots, p$, there exists $\widetilde{\boldsymbol{\beta}}_j$ lying on the line segment between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^o$, such that

$$\dot{\ell}_j(\widehat{\boldsymbol{\beta}}) = \dot{\ell}_j(\boldsymbol{\beta}^o) + \ddot{\ell}_j(\widetilde{\boldsymbol{\beta}}_j)^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o). \tag{2.6}$$

Now let $\boldsymbol{M}(\widetilde{\boldsymbol{\beta}}) \in \mathbb{R}^{p \times p}$ be the matrix with jth row $\ddot{\ell}_j(\widetilde{\boldsymbol{\beta}}_j)^\top$. It follows that with $\widehat{\boldsymbol{b}}$ defined as in (2.1), and for any $\mathbf{c} \in \mathbb{R}^p$ with $\|\mathbf{c}\|_1 = 1$, we can write

$$\begin{aligned} &\mathbf{c}^\top(\widehat{\boldsymbol{b}} - \boldsymbol{\beta}^o) = \mathbf{c}^\top\big\{\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\dot{\ell}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}^o\big\} \\ &= \mathbf{c}^\top\boldsymbol{\Sigma}^{-1}\dot{\ell}(\boldsymbol{\beta}^o) + \mathbf{c}^\top\big(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Sigma}^{-1}\big)\dot{\ell}(\boldsymbol{\beta}^o) + \mathbf{c}^\top\widehat{\boldsymbol{\Theta}}\big\{\dot{\ell}(\widehat{\boldsymbol{\beta}}) - \dot{\ell}(\boldsymbol{\beta}^o)\big\} + \mathbf{c}^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) \\ &= \mathbf{c}^\top\boldsymbol{\Sigma}^{-1}\dot{\ell}(\boldsymbol{\beta}^o) + \mathbf{c}^\top(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Sigma}^{-1})\dot{\ell}(\boldsymbol{\beta}^o) + \mathbf{c}^\top\{\widehat{\boldsymbol{\Theta}}\boldsymbol{M}(\widetilde{\boldsymbol{\beta}}) + \boldsymbol{I}\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o). \quad (2.7) \end{aligned}$$

In Section 3 below, we will provide conditions under which, when both sides of (2.7) are rescaled by $n^{1/2}$, the first, dominant term is asymptotically normal, and the second and third terms are asymptotically negligible. This is the main step in deriving asymptotically valid confidence intervals for $\mathbf{c}^\top\boldsymbol{\beta}^o$.

3. Theory

3.1 Assumptions and main result

Recall that our underlying processes are n independent copies of the triple $\big(\tilde{T}, U, \boldsymbol{\mathcal{Z}}\big)$, where $\boldsymbol{\mathcal{Z}} := \{\boldsymbol{Z}(t) : t \in \mathcal{T}\}$, and that we assume $\tilde{T}$ and U are conditionally independent given $\boldsymbol{\mathcal{Z}}$. Our observations are n independent copies of $\big(T, \delta, \{\boldsymbol{Z}(t) : t \in \mathcal{T}\}\big)$, and we assume that the conditional hazard function of $\tilde{T}$ at time t given $\boldsymbol{\mathcal{Z}}$ satisfies (1.1)[1] for some $\boldsymbol{\beta}^o \in \mathbb{R}^p$. We will

[1] In the terminology of, e.g., Kalbfleisch and Prentice (2002, Section 6.3), this means that all time-dependent covariates are *external*.

make use of the following assumptions:

(A1) **(a)** The process $\{\boldsymbol{Z}(t) : t \in \mathcal{T}\}$ is predictable and there exists a deterministic $K_Z > 0$ with $\sup_{t\in\mathcal{T}} \mathbb{P}\{\|\boldsymbol{Z}(t)\|_\infty \leq K_Z\} = 1$.

(b) The process $\{\boldsymbol{Z}(t) : t \in \mathcal{T}\}$ is uniformly Lipschitz in the sense that there exists a deterministic $L > 0$ such that

$$\mathbb{P}\left\{ \sup_{s,t\in\mathcal{T}, s\neq t} \frac{\|\boldsymbol{Z}(s) - \boldsymbol{Z}(t)\|_\infty}{|s-t|} \leq L \right\} = 1.$$

(A2) **(a)** The random variable T has a bounded density f_T with respect to Lebesgue measure.

(b) $\int_0^{t_+} t^\alpha f_T(t)\, dt < \infty$ for some $\alpha > 0$.

(A3) **(a)** $p = p_n = o(e^{n^a})$, for every $a > 0$.

(b) $d_o := |\{j : \beta_j^o \neq 0\}|$ satisfies $d_o = o\big(n^{1/2}/\log^{1/2}(np)\big)$.

(A4) **(a)** Writing $\mathcal{S} := \{j : \beta_j^o \neq 0\}$, $\mathcal{N} := \{j : \beta_j^o = 0\}$ and

$$\kappa := \inf_{\{\mathbf{v}\in\mathbb{R}^p\setminus\{0\}: \|\mathbf{v}_\mathcal{N}\|_1 \leq 2\|\mathbf{v}_\mathcal{S}\|_1\}} \frac{d_o^{1/2}\{\mathbf{v}^\top \ddot{\ell}(\boldsymbol{\beta}^o)\mathbf{v}\}^{1/2}}{\|\mathbf{v}_\mathcal{S}\|_1},$$

we have that $1/\kappa = O_p(1)$.

(b) $\max_{j=1,\ldots,p} \Sigma_{jj} = O(1)$ as $n \to \infty$.

(c) $\liminf_{n\to\infty} \|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op},1} > 0$ and writing $r_j := \sum_{i=1}^p \mathbb{1}_{\{(\boldsymbol{\Sigma}^{-1})_{ij}\neq 0\}}$ for

$j = 1, \ldots, p$, there exists $\delta_0 > 0$ such that

$$\|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op},1}^2 \max\left\{\frac{d_o^2 \log(np)}{n^{1/2}}, \, d_o n^{-(1/3-\delta_0)}\right\} \max_{j=1,\ldots,p} r_j = o\left(\frac{1}{\log^{1/2}(np)}\right).$$

Some discussion of these assumptions is in order. Condition **(A1)** concerns the boundedness and Lipschitz continuity of the covariate process. It is likely that the first of these conditions could be replaced with a tail condition, at the expense of further complicating the theoretical analysis. Indeed, in our simulations in Section 4, we explore settings in which $\|\boldsymbol{Z}(t)\|_\infty$ is unbounded. Condition **(A2)** consists of two mild and interpretable conditions on the distribution of the observed failure times. Condition **(A3)(a)** controls the rate of growth of the dimensionality as the sample size increases, and in particular allows super-polynomial growth; however, the sparsity assumption **(A3)(b)** ensures that the number of important variables (those with non-zero regression coefficient) is more tightly controlled. Condition **(A4)(a)** is a high-level condition on the so-called compatability factor of $\ddot{\ell}(\boldsymbol{\beta}^o)$; in the presence of our other assumptions, we will see in the discussion following Lemma 1 that this essentially amounts to a condition on the smallest eigenvalue of $\boldsymbol{\Sigma}$. The other parts of **(A4)** also imposes further conditions on $\boldsymbol{\Sigma}$, and, in the case of**(A4)(c)**, the way its properties interact with the sparsity level of $\boldsymbol{\beta}^o$.

The confidence intervals for the regression coefficients are constructed based on the results derived in the following theorem.

Theorem 1. *Assume **(A1)-(A4)** and let $\boldsymbol{c} \in \mathbb{R}^p$ be such that $\|\boldsymbol{c}\|_1 = 1$ and $\mathbf{c}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{c} \to \nu^2 \in (0,\infty)$. For $\widehat{\boldsymbol{\beta}}$ in (2.2), let $\lambda \asymp n^{-1/2}\log^{1/2}(np)$, and for $\widehat{\boldsymbol{\Theta}}$ in (2.5), let*

$$\lambda_n \asymp \left\{\max\left(\|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op},1}\frac{d_o\log(np)}{n^{1/2}}, \|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op},1}n^{-(1/3-\delta_0)}\right)\right\}.$$

Then for $\widehat{\boldsymbol{b}}$ defined in (2.7), we have

$$n^{1/2}\boldsymbol{c}^\top(\widehat{\boldsymbol{b}} - \boldsymbol{\beta}^o) \xrightarrow{d} \mathcal{N}(0,\nu^2)$$

as $n \to \infty$. Moreover,

$$n^{1/2}\boldsymbol{c}^\top(\widehat{\boldsymbol{b}} - \boldsymbol{\beta}^o)/(\boldsymbol{c}^\top\widehat{\boldsymbol{\Theta}}\boldsymbol{c})^{1/2} \xrightarrow{d} \mathcal{N}(0,1).$$

Remark: Theorem 1 can be extended to cover situations where one is interested in testing a hypothesis about a fixed-dimensional sub-vector of $\boldsymbol{\beta}^o$, such as $H_0 : \beta_1^o = \beta_2^o = \beta_3^o = 0$, by choosing an appropriate matrix $\boldsymbol{C}$ in place of the vector $\mathbf{c}$, but for simplicity of exposition, we state the result in terms of a single linear combination of the components of $\boldsymbol{\beta}^o$.

It follows immediately from Theorem 1 that for any $q \in (0,1)$, an asymptotic $(1-q)$-level confidence interval for $\boldsymbol{c}^\top\boldsymbol{\beta}^o$ is given by

$$\big[\boldsymbol{c}^\top\widehat{\boldsymbol{b}} - z_{q/2}n^{-1/2}(\boldsymbol{c}^\top\widehat{\Theta}\boldsymbol{c})^{1/2}, \boldsymbol{c}^\top\widehat{\boldsymbol{b}} + z_{q/2}n^{-1/2}(\boldsymbol{c}^\top\widehat{\Theta}\boldsymbol{c})^{1/2}\big],$$

where z_q is the $(1-q)$th quantile of the standard normal distribution. In particular, for each $j = 1, \ldots, p$, an asymptotic $(1-q)$-level confidence interval for β_j^o is provided by

$$[\hat{b}_j - z_{q/2}n^{-1/2}(\widehat{\Theta}_{jj})^{1/2}, \hat{b}_j + z_{q/2}n^{-1/2}(\widehat{\Theta}_{jj})^{1/2}]. \tag{3.1}$$

3.2 Proof of Theorem 1

The proof of Theorem 1 contains three main steps: a) to provide properties of the initial estimator $\widehat{\boldsymbol{\beta}}$; b) to show the asymptotic normality of the first term in (2.7); c) to show that the remainder terms in (2.7) are negligible. These steps are tackled via the intermediate results in the following three subsections (though proofs are deferred to the supplementary material). The final subsection completes the proof.

To highlight a couple of features in the proof, in step b), the first term in (2.7) is split into two by subtracting and adding the population quantity $\boldsymbol{\mu}(s, \boldsymbol{\beta}^o)$ in the integrand of the expression for the score function $\dot{\ell}(\boldsymbol{\beta}^o)$ at $\boldsymbol{\beta}^o$. This allows us to apply the Lindeberg–Feller central limit theorem to the first (dominant) term to obtain its limiting distribution. The remainder term is a normalised sum of mean-zero, exchangeable random variables whose variances are controlled by weighted integrals over $\mathcal{T}$ of $\|\bar{\boldsymbol{Z}}(\cdot,\boldsymbol{\beta}^o)-\boldsymbol{\mu}(\cdot,\boldsymbol{\beta}^o)\|_\infty^2$. We expect this term to be small when the at-risk set size is reasonably large, but since we allow this at-risk set to be empty at t_+, we adopt adopt a novel truncation technique by setting $t_* := F_T^{-1}(1-n^{-1/2})$ and treating the time intervals from 0 to t_* and from t_* to t_+ separately. For the former interval, we develop a new finite-sample concentration inequality (Lemma S2) to control $\sup_{t\in[0,t_*)}\|\bar{\boldsymbol{Z}}(\cdot,\boldsymbol{\beta}^o)-\boldsymbol{\mu}(\cdot,\boldsymbol{\beta}^o)\|_\infty$, while for the latter, we exploit the boundedness of the process $\bar{\boldsymbol{Z}}(\cdot,\boldsymbol{\beta}^o)$ together with Jensen's inequality to argue that the weighted integral over this region is also asymptotically negligible.

For step c), we derive a special form of martingale concentration inequality by utilising the decoupling techniques developed in de la Peña (1999), as well as concentration inequalities for sub-gamma random variables.

3.2.1 The initial estimator

The following lemma gives the required properties for the score function at $\boldsymbol{\beta}^o$ and the initial estimator. The first result is proved in Lemma 3.3 of Huang et al. (2013), while the second combines Theorem 3.2 and Theorem 4.1 of the same paper.

Lemma 1. *(i) Assume **(A1)(a)**. Then for each $x > 0$,*

$$\mathbb{P}\{\|\dot{\ell}(\boldsymbol{\beta}^o)\|_\infty > x\} \leq 2pe^{-nx^2/(8K_Z^2)}.$$

*(ii) Assume **(A1)(a)**, **(A3)(b)** and **(A4)(a)**, and take*

$$\lambda \asymp n^{-1/2}\log^{1/2}(np)$$

in (2.2). Then

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\left(\frac{d_o \log^{1/2}(np)}{n^{1/2}}\right).$$

Remark: More generally, if we take a sequence (a_n) diverging to infinity arbitrarily slowly, and set $\lambda \asymp n^{-1/2}\log^{1/2}(a_n p)$ in (2.2), then under the conditions of Lemma 1(ii), we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\big(\frac{d_o \log^{1/2}(a_n p)}{n^{1/2}}\big)$. In fact, if we further assume that $p = p_n \to \infty$ as $n \to \infty$, then we may take $\lambda = An^{-1/2}\log^{1/2} p$ in (2.2), and for sufficiently large $A > 0$, conclude that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\big(\frac{d_o \log^{1/2} p}{n^{1/2}}\big)$.

We now discuss **(A4)(a)** in greater depth. For arbitrary finite $t^* \in \mathcal{T}$ and $M > 0$, let $C_1 := 1 + \Lambda_0(t^*)$, and let $C_2 := 2\Lambda_0(t^*)/r_*$, where $r_* := \mathbb{E}\big[Y(t^*)\min\{M, e^{\boldsymbol{\beta}^{o\top}\boldsymbol{Z}(t^*)}\}\big]$. Further, let

$$\boldsymbol{\Sigma}(t^*; M) := \mathbb{E}\int_0^{t^*}\big\{\boldsymbol{Z}(s) - \boldsymbol{\mu}(s, \boldsymbol{\beta}^o; M)\big\}^{\otimes 2} Y(s)\min\{M, e^{\boldsymbol{\beta}^{o\top}\boldsymbol{Z}(t^*)}\}\lambda_0(s)\,ds,$$

where

$$\boldsymbol{\mu}(t, \boldsymbol{\beta}^o; M) := \frac{\mathbb{E}\big[\boldsymbol{Z}(t) Y(t) \min\{M, e^{\boldsymbol{\beta}^{o\top}\boldsymbol{Z}(t)}\}\big]}{\mathbb{E}\big[Y(t) \min\{M, e^{\boldsymbol{\beta}^{o\top}\boldsymbol{Z}(t)}\}\big]}.$$

Write ρ^* for the smallest eigenvalue of $\boldsymbol{\Sigma}(t^*; M)$, and let

$$t_{n,p,\epsilon} := \max\left\{\frac{4}{3n}\log\left(\frac{2.221p(p+1)}{\epsilon}\right), \frac{2}{n^{1/2}}\log^{1/2}\left(\frac{2.221p(p+1)}{\epsilon}\right)\right\}.$$

Then the proof of Huang et al. (2013, Theorem 4.1) gives that for each $\epsilon \in (0, 1/3)$,

$$\mathbb{P}\left[\kappa < \rho^* - 36 d_o K_Z^2\left\{\frac{2^{1/2}C_1}{n^{1/2}}\log^{1/2}\left(\frac{p(p+1)}{\epsilon}\right) + C_2 t_{n,p,\epsilon}^2\right\}\right] \leq 3\epsilon + e^{-nr_*^2/(8M^2)}.$$

Since t^* and M are considered as fixed, it is natural to assume that both $\limsup_{n\to\infty}\max(C_1, C_2) < \infty$, and $\liminf_{n\to\infty}\min(\rho^*, r_*) > 0$. In that case, under **(A3)(b)**, we have $\mathbb{P}(\kappa < \liminf_{n\to\infty}\rho^*/2) \leq 4\epsilon$ for sufficiently large n, so **(A4)(a)** holds.

3.2.2 The dominant and remainder terms

We will describe the limiting behaviour of the dominant term in Proposition 1, and the limiting behaviour of the remainder terms in Propositions 2 and 3. All the proofs can be found in the Supplementary Material.

After rescaling by $n^{1/2}$, the leading term in (2.7) is

$$n^{1/2}\boldsymbol{c}^\top\boldsymbol{\Sigma}^{-1}\dot{\ell}(\boldsymbol{\beta}^o) = \frac{1}{n^{1/2}}\sum_{i=1}^{n}\int_{\mathcal{T}}\boldsymbol{c}^\top\boldsymbol{\Sigma}^{-1}\big\{\boldsymbol{Z}_i(s) - \bar{\boldsymbol{Z}}(s, \boldsymbol{\beta}^o)\big\}\,dN_i(s).$$

We will prove that its limiting distribution is Gaussian.

Proposition 1. *Assume* ***(A1)***, ***(A2)***, ***(A3)(a)*** *and* ***(A4)(c)***, *and let* $\boldsymbol{c} \in \mathbb{R}^p$ *be such that* $\|\boldsymbol{c}\|_1 = 1$ *and* $\boldsymbol{c}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{c} \to \nu^2 \in (0, \infty)$. *Then*

$$n^{1/2} \boldsymbol{c}^\top \boldsymbol{\Sigma}^{-1} \dot{\ell}(\boldsymbol{\beta}^o) \xrightarrow{d} \mathcal{N}(0, \nu^2),$$

as $n \to \infty$.

The two remainder terms in (2.7) are controlled in Propositions 2 and 3 below respectively.

Proposition 2. *Assume conditions* ***(A1)***, ***(A2)(a)***, ***(A3)(b)***, ***(A4)(a)*** *and* ***(A4)(c)***. *For* $\widehat{\boldsymbol{\beta}}$ *in* (2.2), *let* $\lambda \asymp n^{-1/2} \log^{1/2}(np)$, *and for* $\widehat{\boldsymbol{\Theta}}$ *in* (2.5), *let*

$$\lambda_n \asymp \max\left(\|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op},1} \frac{d_o \log(np)}{n^{1/2}}, \|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op},1} n^{-(1/3-\delta_0)} \right).$$

Then for $\boldsymbol{c} \in \mathbb{R}^p$ *with* $\|\boldsymbol{c}\|_1 = 1$, *we have*

$$\boldsymbol{c}^\top \left(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Sigma}^{-1} \right) \dot{\ell}(\boldsymbol{\beta}^o) = o_p(n^{-1/2}).$$

Recall the definition of the matrix $\boldsymbol{M}(\widetilde{\boldsymbol{\beta}})$, which is defined just after (2.6), and which appears in (2.7).

Proposition 3. *Assume* ***(A1)***, ***(A2)(a)***, ***(A3)(b)*** *and* ***(A4)***. *For* $\widehat{\boldsymbol{\beta}}$ *in* (2.2), *let* $\lambda \asymp n^{-1/2} \log^{1/2}(np)$, *and for* $\widehat{\boldsymbol{\Theta}}$ *in* (2.5), *let*

$$\lambda_n \asymp \max\left(\|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op},1} \frac{d_o \log(np)}{n^{1/2}}, \|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op},1} n^{-(1/3-\delta_0)} \right).$$

Then for $\boldsymbol{c} \in \mathbb{R}^p$ *with* $\|\boldsymbol{c}\|_1 = 1$*, we have*

$$\boldsymbol{c}^\top(\widehat{\boldsymbol{\Theta}}\boldsymbol{M}(\widetilde{\boldsymbol{\beta}}) + \boldsymbol{I})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) = o_p(n^{-1/2}).$$

3.2.3 Completion of the proof

We now wrap up all the results in the previous three subsections.

Proof of Theorem 1. From (2.7), Proposition 1, Proposition 2 and 3, we deduce from Slutsky's theorem that under the stated assumptions, the first claim follows. To prove the second claim, note that

$$\left|\boldsymbol{c}^\top\widehat{\boldsymbol{\Theta}}\boldsymbol{c} - \boldsymbol{c}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{c}\right| \leq \left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Sigma}^{-1}\right\|_\infty = o_p(1),$$

where the final claim follows from (S2.5), Lemma **S3** and **(A4)(c)**. Another application of Slutsky's theorem therefore yields the second claim. □

4. Numerical experiments

In this section, we investigate the numerical performance of our proposed method. We begin by discussing various practical implementation issues in Section 4.1; in Sections 4.2 and 4.3, we present analyses of simulated data and real data, respectively.

4.1 Practical issues

4.1.1 Software

Recall that the debiased estimator $\widehat{\boldsymbol{b}}$ is obtained from a Lasso estimator $\widehat{\boldsymbol{\beta}}$ of the vector of true regression coefficients $\boldsymbol{\beta}^o = (\beta_1^o, \ldots, \beta_p^o)^\top$, as well as a CLIME-type estimator $\widehat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Sigma}^{-1}$, the population version of the inverse of the negative Hessian matrix. We use the R (R Core Team, 2017) package GLMNET (Friedman et al., 2010; Simon et al., 2011) to compute $\widehat{\boldsymbol{\beta}}$; and adapt the CLIME (Cai et al., 2012) and FLARE (Li et al., 2014) packages to obtain $\widehat{\boldsymbol{\Theta}}$. The CLIME package is more accurate, but is slow to compute for high-dimensional data; the FLARE algorithm computes only an approximate solution, but is faster. For simplicity, we will refer to the modified CLIME and FLARE algorithms as the CLIME and FLARE packages, respectively. In fact we also conducted analysis based on unmodified CLIME and FLARE (with `sym = ‘or’`) packages, and the differences were negligible.

4.1.2 Tuning parameters

Our theoretical results provide conditions on the tuning parameters λ and λ_n under which our confidence intervals are asymptotically valid; however, in practice, the unknown population quantities and the unspecified constants mean that these conditions do not provide a practical algorithm for choosing these tuning parameters. Therefore, to choose λ, we use the default 10-fold cross-validation algorithm implemented in the GLMNET package, with a grid of 100 different tuning parameters, equally spaced on the log scale. When using the CLIME and FLARE packages to compute $\widehat{\boldsymbol{\Theta}}$, the default 10-fold cross-validation algorithms were used to compute λ_n, with $\mathrm{tr}\big(\mathrm{diag}\big((\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\Theta}} - \boldsymbol{I})^2\big)\big)$ as the cross-validation criterion.

4.1.3 Covariates

Assumption **(A1)**(i) asks that the covariate process $\boldsymbol{\mathcal{Z}}$ should be bounded. However, in our numerical results, we generate the covariate processes from a multivariate Gaussian distribution, due to the convenience of generating different correlation structures. A simulation setting based on uniformly distributed covariates can be found in the Supplementary Material. We also focus for simplicity on time-independent covariates.

An important observation is that even if $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ has identity covariance matrix, this does not necessarily mean that $\boldsymbol{\Sigma} = (\Sigma_{ij})$ is the identity matrix. We can illustrate this in the case where $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}^Z)$, as follows: suppose that $(\boldsymbol{\Sigma}^Z)_{ij} = 0$ whenever $\beta_i^o \neq 0$ and $\beta_j^o = 0$. Then

- for any i, j with $\beta_i^o \neq 0$ and $\beta_j^o = 0$, we have $\Sigma_{ij} = 0$;

- for any i, j with $\beta_i^o = 0$ and $\beta_j^o = 0$, we have

$$\Sigma_{ij} = \mathbb{E}(Z_i Z_j)\mathbb{E}\int_0^{t_+} Y(s)\exp\Big(\sum_{l:\beta_l^o \neq 0} \beta_l^o Z_l\Big)\lambda_0(s)\,ds;$$

- for any i, j with $\beta_i^o \neq 0$ and $\beta_j^o = 0$, we have

$$\Sigma_{ij} = \mathbb{E}\int_0^{t_+} c_i(s)c_j(s)Y(s)\exp\Big(\sum_{l:\beta_l^o \neq 0} \beta_l^o Z_l\Big)\lambda_0(s)\,ds,$$

  where

$$c_i(s) := Z_i - \frac{\mathbb{E}\Big\{Z_i Y(s)\exp\Big(\sum_{l:\beta_l^o \neq 0}\beta_l^o Z_l\Big)\Big\}}{\mathbb{E}\Big\{Y(s)\exp\Big(\sum_{l:\beta_l^o \neq 0}\beta_l^o Z_l\Big)\Big\}}.$$

In order to satisfy the sparse precision matrix conditions, we consider the following two choices of $\mathbf{\Sigma}^Z$ in our simulations in Section 4.2.

a. $\mathbf{\Sigma}_a^Z = \boldsymbol{I}$;

b. $\mathbf{\Sigma}_b^Z = (\Sigma_b^Z)_{ij}$ with

$$(\Sigma_b^Z)_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0.5, & \text{if } i \neq j,\ \beta_i^o \neq 0, \beta_j^o \neq 0, \\ 0, & \text{if } i \neq j, \beta_i^o\beta_j^o = 0, |\beta_i^o| + |\beta_j^o| > 0, \\ 0.5^{|i-j|}, & \text{if } i \neq j, \beta_i^o = 0, \beta_j^o = 0. \end{cases}$$

4.1.4 A simple preliminary example

To illustrate several of the features that arise in more complicated settings, we consider the following two scenarios: let $n = 1000$; $p = 10$; $\boldsymbol{Z} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{I})$;

$\beta_1^o = \cdots = \beta_d^o = 1$, and $\beta_{d+1}^o = \cdots = \beta_p^o = 0$ for $d = 1, 3$; $\lambda_0(t) = 1$, for all $t > 0$; $U_i = 3$ when $d_o = 1$ and $U_i = 5$ when $d_o = 3$. In these settings, the average censoring rate is around 15%. In the top-left blocks of Tables 1 and 2, we report the average initial estimator error $\hat{\beta}_j - \beta_j^o$ for each index $j = 1, \ldots, p$, the average debiased estimator error $\hat{b}_j - \beta_j^o$, the average of empirical coverage (EC) of the 95% confidence intervals, their average widths, and the average p-values, based on 400 repetitions. Standard errors for all quantities are given in brackets.

Here, the results are quite encouraging: the biases of the estimates $\hat{\beta}_j$ of the signal variables are substantially corrected by the debiased estimator $\hat{b}_j$, the coverage probabilities are satisfactory (certainly in the $d_o = 1$ case) and the p-values for the noise variables appear to be approximately uniformly distributed (notice that, under uniformity, the standard errors should be close to $1/(400 \times 12)^{1/2} \approx 0.014$). Of course, this is a setting in which the usual inference for the maximum partial likelihood estimate (MPLE) is also valid, as illustrated in the bottom-right blocks of Tables 1 and 2 (for ease of exposition, the MPLE estimators are collected in the $\hat{b}_j - \beta_j^o$ columns). The MPLE was computed using the package SURVIVAL (Therneau, 2015).

Closer inspection, however, reveals that the situation is not perhaps ideal as it seems at first sight. First, while the bias correction works very well for the noise variables, it slightly under-corrects for the signal variables. Second, the widths of the confidence intervals are slightly smaller than those for the MPLE, which is an efficient estimator. These issues both arise from our choice of precision matrix estimator $\widehat{\boldsymbol{\Theta}}$, which aims to provide a good approximation to $\boldsymbol{\Sigma}^{-1}$ in different matrix norms. To attempt to address this, we therefore consider widening the intervals by replacing the diagonal entries of $\widehat{\boldsymbol{\Theta}}$ in (3.1) with the diagonal entries of $\widetilde{\boldsymbol{\Theta}}$, where $\widetilde{\boldsymbol{\Theta}} = (\widetilde{\Theta}_{ij}) \in \mathbb{R}^{p\times p}$ is given by

$$\widetilde{\Theta}_{ij} = \begin{cases} \widehat{\Theta}_{ij} & \text{if } i \neq j; \\ \max\{1/\widehat{\mathcal{V}}(\widehat{\boldsymbol{\beta}})_{jj}, \widehat{\Theta}_{jj}\} & \text{if } i = j. \end{cases} \tag{4.1}$$

The rationale behind our definition of $\widetilde{\boldsymbol{\Theta}}$ is that in an extreme case, when $\widehat{\mathcal{V}}(\widehat{\boldsymbol{\beta}})$ is a diagonal matrix, $\widehat{\boldsymbol{\Theta}}$ is still a biased estimator of $\boldsymbol{\Sigma}^{-1}$. Since our precision matrix estimators are also potentially sensitive to tuning parameter choice, and the default choice tends to over-penalise, we further consider alternative options to the 10-fold cross-validation choice λ_{CV} in the other blocks of Tables 1 and 2:

(1) Top-right: $\widehat{\boldsymbol{\Theta}}, 0.1\lambda_{\mathrm{CV}}$ – confidence interval constructed based on (3.1) with $0.1\lambda_{\mathrm{CV}}$ used in $\widehat{\boldsymbol{\Theta}}$, which is provided by the CLIME package;

(2) Middle-left: $\widetilde{\boldsymbol{\Theta}}$ – confidence interval replaces $\widehat{\boldsymbol{\Theta}}$ in (3.1) with $\widetilde{\boldsymbol{\Theta}}$, com-

puted using (4.1) with λ_{CV} in the CLIME package;

(3) Middle-right: $\widehat{\boldsymbol{\Theta}}$, FLARE – confidence interval based on (3.1), and $\widehat{\boldsymbol{\Theta}}$ is computed using the FLARE package;

(4) Bottom-left: Merge – confidence interval constructed based on (3.1), the tuning parameter for the sparse precision matrix is provided by the FLARE package using cross-validation, and $\widehat{\boldsymbol{\Theta}}$ is optimised by the CLIME package using the previously mentioned tuning parameter.

β_j^o	$\hat{\beta}_j - \beta_j^o$	$\hat{b}_j - \beta_j^o$	EC	Width	p-value	$\hat{b}_j - \beta_j^o$	EC	Width	p-value
		$\widehat{\boldsymbol{\Theta}}$, λ_{CV}				$\widehat{\boldsymbol{\Theta}}$, $0.1\lambda_{\text{CV}}$			
1	-.051(.002)	-.003(.002)	.937(.012)	.157(.000)	.000(.000)	.000(.002)	.940(.011)	.159(.000)	.000(.000)
0	.000(.001)	.000(.001)	.966(.009)	.123(.000)	.539(.014)	.000(.001)	.961(.009)	.125(.000)	.532(.014)
0	.000(.001)	.001(.001)	.952(.010)	.123(.000)	.520(.014)	.000(.001)	.952(.010)	.125(.000)	.510(.014)
0	.001(.001)	.001(.001)	.955(.010)	.124(.000)	.522(.014)	.001(.002)	.952(.010)	.125(.000)	.518(.014)
0	-.001(.001)	-.001(.001)	.943(.011)	.123(.000)	.532(.014)	-.002(.002)	.943(.011)	.125(.000)	.528(.014)
0	.001(.001)	.001(.001)	.943(.011)	.123(.000)	.514(.014)	.000(.002)	.947(.011)	.125(.000)	.509(.014)
0	-.002(.001)	-.003(.001)	.955(.010)	.123(.000)	.539(.014)	-.003(.001)	.950(.010)	.125(.000)	.529(.014)
0	.000(.001)	-.001(.001)	.934(.012)	.123(.000)	.532(.014)	-.001(.002)	.933(.012)	.125(.000)	.517(.014)
0	.000(.001)	-.001(.002)	.930(.012)	.123(.000)	.523(.014)	-.001(.002)	.929(.012)	.125(.000)	.513(.014)
0	.000(.001)	-.001(.001)	.957(.010)	.123(.000)	.520(.014)	-.001(.002)	.959(.010)	.125(.000)	.514(.014)
		$\widetilde{\boldsymbol{\Theta}}$				$\widehat{\boldsymbol{\Theta}}$, FLARE			
1	-.051(.002)	-.003(.002)	.961(.009)	.171(.000)	.000(.000)	-.007(.002)	.925(.013)	.153(.000)	.000(.000)
0	.000(.001)	.000(.001)	.972(.008)	.134(.000)	.564(.013)	.000(.001)	.965(.009)	.121(.000)	.544(.014)
0	.000(.001)	.000(.001)	.975(.008)	.134(.000)	.547(.013)	.001(.001)	.953(.010)	.121(.000)	.529(.014)
0	.001(.001)	.001(.001)	.968(.008)	.134(.000)	.548(.014)	.001(.001)	.948(.011)	.121(.000)	.526(.014)
0	-.001(.001)	-.002(.001)	.970(.008)	.134(.000)	.559(.014)	-.002(.001)	.948(.011)	.121(.000)	.539(.014)
0	.001(.001)	.001(.001)	.959(.010)	.134(.000)	.543(.014)	.001(.001)	.948(.011)	.121(.000)	.522(.014)
0	-.002(.001)	-.003(.001)	.968(.008)	.134(.000)	.560(.014)	-.003(.001)	.955(.010)	.121(.000)	.541(.014)
0	.000(.001)	-.001(.001)	.968(.008)	.134(.000)	.551(.013)	-.001(.001)	.937(.012)	.121(.000)	.534(.014)
0	.000(.001)	-.001(.002)	.952(.010)	.134(.000)	.544(.014)	-.001(.002)	.930(.012)	.121(.000)	.524(.015)
0	.000(.001)	-.001(.001)	.968(.008)	.134(.000)	.547(.014)	-.001(.001)	.958(.010)	.121(.000)	.527(.014)
		Merge				MPLE			
1	-.051(.002)	-.004(.002)	.938(.012)	.157(.000)	.000(.000)	.007(.002)	.950(.011)	.173(.000)	.000(.000)
0	.000(.001)	.000(.001)	.964(.009)	.123(.000)	.541(.014)	.000(.002)	.967(.009)	.136(.000)	.519(.014)
0	.000(.001)	.001(.001)	.955(.010)	.123(.000)	.525(.014)	.001(.002)	.953(.010)	.135(.000)	.497(.014)
0	.001(.001)	.001(.002)	.950(.011)	.123(.000)	.522(.014)	.001(.002)	.957(.010)	.136(.000)	.497(.014)
0	-.001(.001)	-.002(.001)	.948(.011)	.123(.000)	.536(.014)	-.002(.002)	.950(.011)	.136(.000)	.510(.014)
0	.001(.001)	.000(.001)	.948(.011)	.123(.000)	.518(.014)	.000(.002)	.950(.011)	.135(.000)	.493(.014)
0	-.002(.001)	-.003(.001)	.953(.010)	.123(.000)	.539(.014)	-.003(.002)	.962(.009)	.135(.000)	.516(.014)
0	.000(.001)	-.001(.001)	.941(.012)	.123(.000)	.529(.014)	.000(.002)	.945(.011)	.136(.000)	.504(.014)
0	.000(.001)	-.001(.002)	.927(.013)	.123(.000)	.518(.015)	-.001(.002)	.924(.013)	.135(.000)	.494(.015)
0	.000(.001)	-.001(.001)	.957(.010)	.123(.000)	.526(.014)	-.001(.002)	.960(.010)	.135(.000)	.501(.014)

Table 1: Simple preliminary example, $d_o = 1$.

β_j^o	$\hat{\beta}_j - \beta_j^o$	$\hat{b}_j - \beta_j^o$	EC	Wid	pvals	$\hat{b}_j - \beta_j^o$	EC	Width	p-value
		$\widehat{\boldsymbol{\Theta}}$, λ_{CV}				$\widehat{\boldsymbol{\Theta}}$, $0.1\lambda_{\mathrm{CV}}$			
1	-.040(.002)	-.003(.002)	.917(.013)	.153(.000)	.000(.000)	.000(.002)	.919(.013)	.157(.000)	.000(.000)
1	-.041(.002)	-.005(.002)	.932(.012)	.153(.000)	.000(.000)	-.001(.002)	.940(.011)	.157(.000)	.000(.000)
1	-.042(.002)	-.006(.002)	.924(.012)	.153(.000)	.000(.000)	-.002(.002)	.930(.012)	.157(.000)	.000(.000)
0	.000(.001)	.000(.001)	.943(.011)	.123(.000)	.505(.013)	.000(.002)	.934(.011)	.125(.000)	.499(.013)
0	.000(.001)	.000(.001)	.941(.011)	.123(.000)	.505(.014)	.000(.002)	.938(.011)	.125(.000)	.499(.014)
0	.001(.001)	.001(.001)	.930(.012)	.123(.000)	.501(.014)	.001(.002)	.928(.012)	.125(.000)	.494(.014)
0	-.001(.001)	-.002(.001)	.932(.012)	.123(.000)	.520(.014)	-.001(.002)	.932(.012)	.125(.000)	.513(.014)
0	-.002(.001)	-.003(.002)	.936(.011)	.123(.000)	.510(.014)	-.003(.002)	.928(.012)	.125(.000)	.503(.014)
0	.000(.001)	.000(.002)	.928(.012)	.123(.000)	.497(.014)	.000(.002)	.928(.012)	.125(.000)	.491(.014)
0	.000(.001)	.000(.001)	.938(.011)	.123(.000)	.506(.014)	.000(.002)	.938(.011)	.125(.000)	.500(.014)
		$\widetilde{\boldsymbol{\Theta}}$				$\widehat{\boldsymbol{\Theta}}$, FLARE			
1	-.040(.002)	-.004(.002)	.919(.013)	.154(.000)	.000(.000)	-.008(.002)	.899(.015)	.148(.000)	.000(.000)
1	-.041(.002)	-.005(.002)	.932(.012)	.154(.000)	.000(.000)	-.009(.002)	.923(.013)	.149(.000)	.000(.000)
1	-.042(.002)	-.006(.002)	.923(.012)	.154(.000)	.000(.000)	-.010(.002)	.891(.015)	.148(.000)	.000(.000)
0	.000(.001)	.000(.001)	.962(.009)	.135(.000)	.535(.013)	-.001(.002)	.944(.011)	.121(.000)	.511(.014)
0	.000(.001)	.000(.001)	.957(.009)	.135(.000)	.534(.013)	-.001(.002)	.940(.012)	.121(.000)	.500(.015)
0	.001(.001)	.001(.001)	.962(.009)	.135(.000)	.530(.014)	.000(.002)	.923(.013)	.121(.000)	.511(.015)
0	-.001(.001)	-.001(.001)	.953(.010)	.135(.000)	.549(.013)	-.002(.002)	.935(.012)	.121(.000)	.524(.015)
0	-.002(.001)	-.003(.002)	.955(.010)	.135(.000)	.537(.014)	-.003(.002)	.937(.012)	.121(.000)	.514(.015)
0	.000(.001)	.000(.002)	.947(.010)	.135(.000)	.526(.014)	-.002(.002)	.935(.012)	.121(.000)	.497(.015)
0	.000(.001)	.000(.001)	.966(.008)	.135(.000)	.535(.014)	.001(.002)	.935(.012)	.121(.000)	.511(.015)
		Merge				MPLE			
1	-.040(.002)	-.004(.002)	.910(.014)	.153(.000)	.000(.000)	.006(.002)	.940(.012)	.172(.000)	.000(.000)
1	-.041(.002)	-.006(.002)	.928(.013)	.154(.000)	.000(.000)	.005(.002)	.958(.010)	.172(.000)	.000(.000)
1	-.042(.002)	-.006(.002)	.918(.014)	.154(.000)	.000(.000)	.005(.002)	.942(.012)	.171(.000)	.000(.000)
0	.000(.001)	-.002(.002)	.952(.011)	.123(.000)	.510(.015)	-.002(.002)	.955(.010)	.137(.000)	.495(.014)
0	.000(.001)	-.001(.002)	.942(.012)	.123(.000)	.497(.015)	-.001(.002)	.948(.011)	.136(.000)	.480(.014)
0	.001(.001)	.000(.002)	.925(.013)	.123(.000)	.506(.015)	.000(.002)	.945(.011)	.136(.000)	.493(.015)
0	-.001(.001)	-.002(.002)	.935(.012)	.123(.000)	.524(.015)	-.002(.002)	.940(.012)	.137(.000)	.511(.015)
0	-.002(.001)	-.003(.002)	.942(.012)	.123(.000)	.512(.015)	-.003(.002)	.950(.011)	.137(.000)	.500(.015)
0	.000(.001)	-.002(.002)	.935(.012)	.123(.000)	.499(.015)	-.002(.002)	.932(.013)	.136(.000)	.486(.014)
0	.000(.001)	.001(.002)	.932(.013)	.123(.000)	.509(.015)	.001(.002)	.948(.011)	.136(.000)	.493(.014)

Table 2: Simple preliminary example, $d_o = 3$.

Comparing the columns of $\hat{\beta}_j - \beta_j^o$ and $\hat{b}_j - \beta_j^o$, we can see that our proposed methods indeed correct the bias due to the shrinkage introduced by the Lasso estimators, but the biases for the signal variables are not fully corrected, and in terms of the signs of the errors, they all tend to be under-corrected, except the $\widehat{\mathbf{\Theta}}, 0.1\lambda_{\text{CV}}$ blocks. The differences between the $\widehat{\mathbf{\Theta}}, \lambda_{\text{CV}}$ and $\widehat{\mathbf{\Theta}}, 0.1\lambda_{\text{CV}}$ blocks show that the 10-fold cross-validation chosen tuning parameters still over-penalise the sparse precision matrix estimation and lead to under-correction of $\widehat{\boldsymbol{b}}$. From the EC and Width columns in the $\widehat{\mathbf{\Theta}}, \lambda_{\text{CV}}$ and $\widetilde{\mathbf{\Theta}}$ blocks, we can see that in some cases, using $\widetilde{\mathbf{\Theta}}$ indeed helps in terms of improving the coverages (naturally, the confidence intervals are a little wider). We can also see that the FLARE package does not produce identical solutions to the CLIME package even in this relatively simple context. It is worth noting that the $\widehat{\mathbf{\Theta}}$, FLARE and Merge blocks have the same initial estimators, the same tuning parameter grids for $\widehat{\mathbf{\Theta}}$ and the same cross-validation algorithms. Further investigation in the case $d_o = 1$ reveals that the FLARE package tends to choose slightly larger tuning parameters, which explains the better centering and coverage of the CLIME confidence intervals; see Table 3.

Packages	Mean	Median
CLIME	0.022	0.015
FLARE	0.026	0.025

Table 3: Selected tuning parameter comparisons.

4.2 Further simulated examples

In order to provide a deeper understanding of our proposed method, we consider the following 16 simulation settings described below, where CT is the censoring time and CR is censoring rate.

(1) $n = 1000$, $p = 10$, $\beta_j^o = 1$, $j = 1, 2, 3$, $\beta_j^o = 0$, $j = 4, \ldots, 10$, $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_a^Z)$, CT $= 5$, CR $\approx 15\%$;

(2) $n = 1000$, $p = 10$, $\beta_j^o = 1$, $j = 1, 2, 3$, $\beta_j^o = 0$, $j = 4, \ldots, 10$, $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_a^Z)$, CT $= 2$, CR $\approx 30\%$;

(3) $n = 1000$, $p = 10$, $(\beta_1^o, \beta_2^o, \beta_3^o) = (1.2, 1, 0.8)$, $\beta_j^o = 0$, $j = 4, \ldots, 10$, $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_a^Z)$, CT $= 5$, CR $\approx 15\%$;

(4) $n = 1000$, $p = 10$, $(\beta_1^o, \beta_2^o, \beta_3^o) = (1.2, 1, 0.8)$, $\beta_j^o = 0$, $j = 4, \ldots, 10$, $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_a^Z)$, CT $= 2$, CR $\approx 30\%$;

(5-8) As for (1)-(4), but with $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_b^Z)$, CT $= 10, 2.5, 10, 2.5$;

(9-10) As for (1)-(2), but with $p = 300$, $\beta_j^o = 1$, $j = 1, \ldots, 6$, $\beta_j^o = 0$, $j = 7, \ldots, 300$, $\text{CT} = 9, 2.5$;

(11-12) As for (3)-(4), but $p = 300$, $(\beta_1^o, \ldots, \beta_6^o) = (0.5, 0.7, 0.9, 1.1, 1.3, 1.5)$, $\beta_j^o = 0$, $j = 7, \ldots, 300$, $\text{CT} = 10, 3$;

(13-16) As for (9)-(12), but with $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_b^Z)$, $\text{CT} = 100, 7, 100, 7$.

In Table 4, we report averaged results for signal and noise variables separately, with $\widehat{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Theta}}$ chosen by 10-fold cross-validation. The simulations were run on a cluster, each node of which is a Intel(R) Xeon(R) CPU E5-2670 0@2.60GHz machine, with 16 CPUs. To conduct one repetition of a $(n, p) = (1000, 300)$ setting, it took on average 32 minutes. This explains why we limit our simulations to $p = 300$ even though our theory can handle $p \gg n$ settings.

				$\hat{\Theta}$			$\tilde{\Theta}$	
	$\hat{b}_j - \beta_j^o$	$\hat{\beta}_j - \beta_j^o$	EC	Width	p-values	EC	Width	p-values
(1) S	-.003(.001)	-.038(.001)	.933(.008)	.153(.000)	.000(.000)	.933(.008)	.154(.000)	.000(.000)
(1) N	-.001(.001)	-.001(.001)	.929(.008)	.123(.000)	.491(.010)	.956(.006)	.135(.000)	.521(.009)
(2) S	-.009(.001)	-.039(.001)	.907(.009)	.150(.000)	.000(.000)	.940(.008)	.165(.000)	.000(.000)
(2) N	-.002(.001)	-.001(.001)	.921(.008)	.123(.000)	.503(.010)	.957(.006)	.147(.000)	.556(.009)
(3) S	-.003(.001)	-.038(.001)	.940(.007)	.154(.000)	.000(.000)	.940(.007)	.155(.000)	.000(.000)
(3) N	-.002(.001)	-.001(.001)	.933(.008)	.123(.000)	.497(.010)	.951(.007)	.135(.000)	.527(.009)
(4) S	-.009(.001)	-.039(.001)	.883(.010)	.150(.000)	.000(.000)	.913(.009)	.166(.000)	.000(.000)
(4) N	-.002(.001)	-.001(.001)	.914(.009)	.123(.000)	.510(.010)	.957(.006)	.147(.000)	.565(.010)
(5) S	-.004(.002)	-.036(.002)	.937(.008)	.177(.000)	.000(.000)	.953(.006)	.194(.000)	.000(.000)
(5) N	.000(.001)	.000(.001)	.933(.008)	.152(.000)	.496(.009)	.937(.008)	.152(.000)	.496(.009)
(6) S	-.008(.002)	-.035(.002)	.887(.010)	.174(.000)	.000(.000)	.950(.007)	.211(.000)	.000(.000)
(6) N	.000(.001)	.000(.001)	.913(.009)	.151(.000)	.495(.01)	.921(.008)	.154(.000)	.508(.010)
(7) S	-.003(.002)	-.036(.002)	.930(.008)	.177(.000)	.000(.000)	.940(.007)	.194(.000)	.000(.000)
(7) N	.000(.001)	.000(.001)	.936(.008)	.152(.000)	.496(.009)	.936(.008)	.152(.000)	.494(.009)
(8) S	-.007(.002)	-.033(.002)	.903(.009)	.175(.000)	.000(.000)	.940(.007)	.212(.000)	.000(.000)
(8) N	-.001(.001)	.000(.001)	.917(.009)	.151(.000)	.496(.010)	.920(.009)	.154(.000)	.504(.010)
(9) S	-.169(.005)	-.264(.005)	.290(.026)	.242(.001)	.000(.000)	.322(.027)	.268(.001)	.000(.000)
(9) N	.000(.002)	.000(.001)	.984(.006)	.218(.001)	.625(.014)	.992(.004)	.251(.001)	.663(.014)
(10) S	-.078(.002)	-.155(.002)	.415(.016)	.138(.000)	.000(.000)	.495(.016)	.159(.000)	.000(.000)
(10) N	.000(.001)	.000(.000)	.976(.004)	.120(.000)	.609(.008)	.992(.002)	.149(.000)	.668(.007)
(11) S	-.063(.002)	-.150(.002)	.553(.016)	.143(.000)	.000(.000)	.612(.015)	.149(.000)	.000(.000)
(11) N	.000(.001)	.000(.000)	.977(.004)	.120(.000)	.586(.008)	.988(.003)	.136(.000)	.621(.008)
(12) S	-.081(.002)	-.154(.002)	.413(.016)	.141(.000)	.000(.000)	.485(.016)	.158(.000)	.000(.000)
(12) N	.000(.001)	.000(.000)	.976(.005)	.120(.000)	.608(.008)	.991(.002)	.147(.000)	.665(.007)
(13) S	-.034(.002)	-.122(.002)	.848(.011)	.178(.000)	.000(.000)	.895(.010)	.198(.000)	.000(.000)
(13) N	.000(.001)	.000(.000)	.985(.003)	.150(.000)	.593(.008)	.985(.003)	.150(.000)	.593(.008)
(14) S	-.052(.002)	-.126(.002)	.745(.014)	.177(.000)	.000(.000)	.852(.011)	.219(.000)	.000(.000)
(14) N	.000(.001)	.000(.000)	.988(.003)	.149(.000)	.624(.008)	.989(.003)	.151(.000)	.628(.008)
(15) S	-.028(.002)	-.122(.002)	.863(.011)	.180(.000)	.000(.000)	.897(.009)	.198(.000)	.000(.000)
(15) N	.000(.001)	.000(.000)	.985(.003)	.151(.000)	.593(.008)	.985(.003)	.151(.000)	.593(.008)
(16) S	-.046(.002)	-.126(.002)	.772(.013)	.178(.000)	.000(.000)	.845(.011)	.219(.000)	.000(.000)
(16) N	.000(.001)	.000(.000)	.987(.003)	.149(.000)	.624(.008)	.988(.003)	.151(.000)	.628(.008)

Table 4: Simulation settings (1)-(16). S and N rows are for results for signal and noise variables respectively.

It is reassuring to see that, in all cases, the confidence intervals for the noise variables have close to nominal coverage and the p-values for the noise variables appear to be uniformly distributed. Thus, our methodology is providing a reliable method for identifying signal variables, with uncertainty quantification. On the other hand, while the confidence intervals for the signal variables have good coverage when $p = 10$ (particularly with $\widetilde{\boldsymbol{\Theta}}$), it is much more challenging to ensure adequate coverage for the signal variables in the $p = 300$ case. Apparently, the sample size needs to be very large for the asymptotics to 'kick in', to the extent that we can think, for instance, that **(A4)(c)** is satisfied. The greater width of the intervals when using $\widetilde{\boldsymbol{\Theta}}$ yields improved coverage for the signal variables, but leads to some over-coverage for the noise variables.

One approach in high-dimensional settings, then, is to use our methodology as a screening method to identify signal variables (with false discovery guarantees), and then use the standard MPLE inference to obtain confidence intervals for the signal variables at a second stage. Further discussion can be found in the Supplementary Material.

4.3 Real data analysis

In this section, we apply our method to a diffuse large B-cell lymphoma (DLBCL) dataset, comprising survival times of 240 DLBCL patients and gene expression data from 7399 genes (Rosenwald et al., 2002). To reduce dimensionality, we computed the Lasso path, noting that the cross-validation algorithm picked the 16th largest value of λ on our grid of size 100. In total, 84 variables were selected at some stage in the first 25 λ values, and we therefore retained these 84 variables in our subsequent analysis.

In Figure 1, we plot the glmnet solution paths, with solid and black paths being the ones for those variables deemed to be significant according to our methodology, and dashed and grey paths for those variables deemed insignificant. The left and right panels correspond to the use of $\widehat{\boldsymbol{\Theta}}$ and $\widetilde{\boldsymbol{\Theta}}$ respectively, and the red vertical lines indicate the regularisation parameter values chosen by cross-validation. The only difference between the inferences drawn from the two precision matrix estimates is the confidence interval widths, so the selected variables when using $\widehat{\boldsymbol{\Theta}}$ are a proper subset of those obtained using $\widetilde{\boldsymbol{\Theta}}$.

It can be seen that some variables enter the model fairly early along the path, but appear not significant according to our methods. These variables are often omitted from the model at a later stage along the path, as other

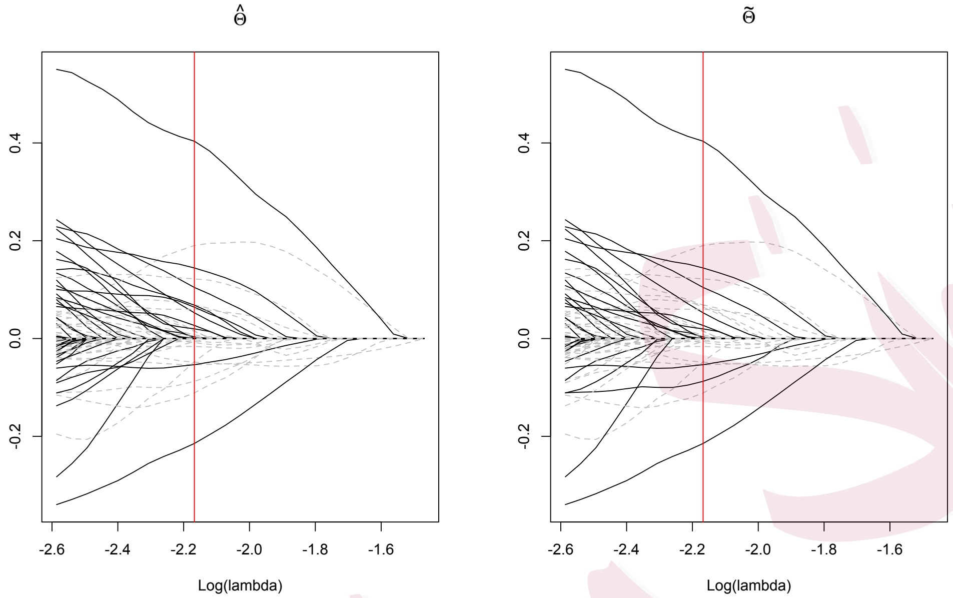

Figure 1: Solution paths

variables enter. This observation is demonstrated in Table 5, which presents the median life-spans of the corresponding variables, where the life-span is defined as the proportion of the locations on the solution paths for which a certain variable is chosen.

$\widehat{\boldsymbol{\Theta}}$			$\widetilde{\boldsymbol{\Theta}}$		
No.	Significant	Insignificant	No.	Significant	Insignificant
41	0.78	0.26	32	0.78	0.35

Table 5: Median life-spans for variables deemed significant and insignificant.

Acknowledgements: The third author is supported by Engineering and Physical Sciences Research Council fellowships EP/J017213/1 and EP/P031447/1 and grant RG81761 from the Leverhulme Trust. The first and third authors would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme 'Statistical Scalability' when work on this paper was undertaken. This work was supported by Engineering and Physical Sciences Research Council grant number EP/K032208/1. The second author is supported by the National Science Foundation grant number NSF-DMS/1712481.

References

Andersen, P. K., Borgan, Ø, Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes* Springer–Verlag, New York.

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

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

Breslow, N. E. (1972). Contribution to discussion of paper by D. R. Cox. *J. Roy. Statist. Soc., Ser. B*, **34**, 216–217.

Cai, T. T., Liu, W. and Luo, X. (2011). A Constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, **101**, 594–607.

Cai, T. T., Liu, W. and Luo, X. (2012). clime: Constrained L1-minimization for Inverse (covariance) Matrix Estimation. R package version 0.4.1. URL: https://CRAN.R-project.org/package=clime.

Cai, T. T., Liu, W. and Zhou, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.*, **44**, 455–88.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Stat. Soc., Ser. B*, **34**, 187–220.

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

de la Peña, V. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Probab.*, **27**, 537–564.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 2013–2038.

Fang, E. X., Yang, N. and Liu, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. Roy. Statist. Soc., Ser. B*, **79**, 1415–1437.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Soft.*, **33**, 1–22. URL http://www.jstatsoft.org/v33/i01/.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle inequalities for the Lasso in the Cox model. *Ann. Statist.*, **41**, 1142–65.

Janková, J., and van de Geer, S. (2018). De-biased sparse PCA: Inference and testing for eigenstructure of large covariance matrices. *ArXiv preprint*, arXiv:1801.10567.

Javanmard, A., and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, **15**, 2869–909.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data (2nd edition).* John Wiley & Sons, Inc., Hoboken, New Jersey.

Kong, S., Yu, Z., Zhang, X. and Cheng, G. (2018) High Dimensional Robust Inference for Cox Regression Models *ArXiv preprint*, arXiv:1811.00535.

Li, X., Zhao, T. Wang, L., Yuan, X. and Liu, H. (2014). flare: Family of Lasso Regression. R package version 1.5.0. https://CRAN.R-project.org/package=flare.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436–1462.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., Giltnane, J.M. and Hurt, E.M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, **346**, 1937–47.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Statist. Soft.*, **39**, 1–13. URL http://www.jstatsoft.org/v39/i05/.

Therneau T (2015). A Package for Survival Analysis in S. Version 2.38, URL: https://CRAN.R-project.org/package=survival.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–202.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Roy. Statist. Soc., Ser. B*, **76**, 217–42.

School of Mathematics, University of Bristol

E-mail: y.yu@bristol.ac.uk

Department of Mathematics, University of California at San Diego

E-mail: jbradic@ucsd.edu

Statistical Laboratory, University of Cambridge

E-mail: r.samworth@statslab.cam.ac.uk