Statistics and Computing (2026) 36:98
https://doi.org/10.1007/s11222-026-10849-1

ORIGINAL PAPER

Computationally efficient likelihood-based estimation and variable selection for the Cox model with incomplete covariates

Ngok Sang Kwok[1] · **Kin Yau Wong**[1,2]

Received: 29 July 2025 / Accepted: 3 February 2026

Abstract

Regression analysis with missing data is a long-standing and challenging problem, particularly when there are many missing variables with arbitrary missing patterns. Likelihood-based methods, although theoretically appealing, are often computationally inefficient or even infeasible when dealing with a large number of missing variables. In this paper, we consider the Cox regression model with incomplete covariates that are missing at random. We develop an expectation-maximization (EM) algorithm for nonparametric maximum likelihood estimation, employing a transformation technique in the E-step so that it involves only a one-dimensional integration. This innovation makes our methods computationally tractable even when the number of missing variables is large. In addition, for variable selection, we extend the proposed EM algorithm to accommodate a Lasso penalty in the likelihood. We demonstrate the feasibility and advantages of the proposed methods by large-scale simulation studies and apply the proposed methods to a cancer genomic study.

Keywords EM algorithm · Lasso · missing data · nonparametric maximum likelihood estimation · penalized regression · survival analysis

1 Introduction

In public health and medical studies, we often study the association between covariates, such as treatment received, demographic information, or other personal characteristics, and times to disease events or death. One complication that often arises in practice is that the covariates may not be available for all study subjects. For example, The Cancer Genome Atlas (TCGA) collected genomic and clinical data for many types of cancer, but for a substantial number of subjects, protein expressions were not measured. Another example is the North Staffordshire Osteoarthritis Project (Wilkie et al. 2019), where researchers investigated the association between mortality and symptomatic osteoarthritis. Covariates such as walking frequency, depression, and BMI, collected via questionnaires, had missing values for some subjects. Missing data pose significant theoretical and computational challenges for regression analysis.

There are multiple approaches for handling missing covariates in regression analyses, many of which have been applied to survival analysis. One is the likelihood approach, in which we incorporate a model for the missing covariates into the likelihood. An advantage of this approach is that maximum likelihood estimation (MLE) is efficient. Herring and Ibrahim (2001) considered the Cox model with incomplete covariates, which could be categorical or continuous, and developed an expectation-maximization (EM) algorithm (Dempster et al. 1977) for computation. Zhou et al. (2022) implemented the EM algorithm with two-stage data augmentation for the Cox model with interval-censored survival time. However, MLE may become computationally intensive or even infeasible when there are a large number of missing covariates.

Another approach is inverse-probability weighting (IPW) (Wooldridge, 2002; Wooldridge, 2007), where only subjects with complete observations are used and are weighted according to the probability of complete observation. Martinussen et al. (2016) considered a Cox model and used an IPW approach to handle missing data. Thiessen et al. (2022) proposed a two-step estimator for the Cox model

✉ Kin Yau Wong
kin-yau.wong@polyu.edu.hk

[1] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom , Hong Kong

[2] Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China

Published online: 01 March 2026

with incomplete covariates, where two inverse-probability weighted estimators are combined to form a more efficient estimator. IPW, however, is in general inefficient as it discards information contained in subjects with partial observations.

A third approach is (multiple) imputation, where the missing values are filled in based on the observed data, enabling standard analysis on the completed data (Rubin 2004). A widely used technique is multiple imputation by chained equations (MICE) (van Buuren and Groothuis-Oudshoorn 2011; Azur et al. 2011), also termed "fully conditional specification," where conditional models are specified for each incomplete variable. MICE iteratively regresses each incomplete variable on the remaining imputed dataset and draws new imputations based on the regression models. Bartlett et al. (2015) extended this framework to accommodate a nonlinear outcome model with interaction and polynomial terms. Deng and Lumley (2023) proposed multiple imputation through XGBoost (MIXGB), which uses XGBoost for capturing the distribution of the incomplete variables. Alternatively, matrix completion imputes missing entries by solving a rank-constrained optimization problem (Hastie et al. 2015). Although imputation is intuitive, estimators derived from imputed data lack guaranteed theoretical properties.

In addition to missing data, high-dimensional covariates present another challenge. When the number of available covariates is large, standard approaches that regress on all covariates, such as MLE or estimating equations based on inverse-probability weighting, may suffer from overfitting and difficulties in interpretation, or may even be infeasible. In such cases, we are often interested in selecting a subset of covariates that are associated with the outcome. Penalized regression methods, such as Lasso (Tibshirani 1996), are popular approaches to reduce overfitting and to perform variable selection.

Penalized regression or variable selection with missing data is highly challenging, with limited research in this area. For likelihood-based methods, Garcia et al. (2010) developed EM algorithms for the Cox model with Lasso, adaptive Lasso (Zou 2006), and the smoothly clipped absolute deviation (Fan and Li 2001) penalties. Sabbe et al. (2013) studied variable selection in logistic regression using a Lasso penalty via a stochastic EM algorithm. For IPW, Johnson et al. (2008) and Wolfson (2011) incorporated a penalty term into inverse-probability-weighted estimating equations for performing variable selection. For multiple imputation, Wood et al. (2008) investigated methods for combining variable selection results from multiply imputed datasets. Deng et al. (2016) extended the MICE approach to high-dimensional settings by fitting a penalized regression model for each missing covariate. Liang et al. (2024) developed an iterative imputation method based on matrix completion and a randomized Lasso method based on bootstrap. However, these approaches suffer the shortcomings of their unpenalized counterparts, such as computational or estimation inefficiency and a lack of theoretical justifications. They may also require computationally intensive tuning.

In this paper, we study the likelihood approach and develop a novel algorithm for MLE that overcomes the computational challenges of existing methods. In particular, we consider the Cox proportional hazards model with incomplete covariates and develop an EM algorithm for computation, which is computationally feasible even when the missing pattern is arbitrary and a large number of covariates are missing for each subject. The algorithm involves linear transformations of the missing covariates, applied separately to each subject according to their individual missing data pattern. By exploiting the fact that linear combinations of Gaussian random variables are also Gaussian, only one-dimensional numerical integrations are required in the E-step after the transformation. This is a major advance over existing likelihood-based methods, which require multi-dimensional numerical integration (that is, integration over a space with dimensionality equal to the number of missing covariates) and thus are feasible only for a small number of missing covariates.

Similar transformation techniques for the computation of expectations under multivariate Gaussian distribution have been employed in Bayesian analysis and in latent variable modeling, where parameters or latent variables need to be integrated out in the posterior distribution or the likelihood (Albert and Chib 1993; González et al. 2006). While our approach makes use of the same basic property about Gaussian variables, to the best of our knowledge, there are no existing methods that adapt these techniques to handle missing covariates and apply them within an EM framework.

Under the likelihood framework, we can naturally perform variable selection by incorporating a penalty term. In particular, we consider a Lasso penalty and develop an EM algorithm for computation. Employing the transformation technique, the E-step involves only a one-dimensional numerical integration, as in the unpenalized case. In the M-step, we perform quadratic approximation of the log-partial likelihood and adopt the coordinate-descent algorithm. Consequently, the estimation and variable selection procedures remain computationally feasible even with many missing covariates.

This paper is structured as follows. In Section 2, we describe the proposed model and formulate the EM algorithm for both the unpenalized and penalized cases. In Section 3, we perform large-scale simulation studies and compare the performance of the proposed methods with existing methods. In Section 4, we demonstrate the feasibility and advantages of the proposed method in a cancer genomics study. In Section 5, we present the computation times for the proposed and alternative methods on real and simulated datasets. We

provide some concluding remarks and possible extensions in Section 6.

2 Methods

2.1 Model and likelihood

Let T^* be an event time of interest and $\boldsymbol{X}$ be a p-vector of covariates. Assume that T^* given $\boldsymbol{X}$ follows the Cox proportional hazards model, with the hazard function given by

$$\lambda(t \mid \boldsymbol{X}) = \lambda(t) e^{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}},$$

where $\boldsymbol{\beta}$ is a vector of regression parameters, and $\lambda(\cdot)$ is a nonparametric baseline hazard function. We assume that $\boldsymbol{X}$ follows the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here, we for simplicity of presentation impose a parametric model on all components of $\boldsymbol{X}$, but the proposed methods can be easily generalized to the milder condition that a subset of components of $\boldsymbol{X}$, $\boldsymbol{X}_{\mathcal{S}}$, is multivariate normal given the remaining components, $\boldsymbol{X}_{-\mathcal{S}}$ provided that $\boldsymbol{X}_{-\mathcal{S}}$ is always observed. This extension is discussed in Section 2.4. Suppose that T^* may be subject to right-censoring. Let C be the censoring time, $Y = \min(T^*, C)$, and $\Delta = I(T^* \le C)$.

We allow components of $\boldsymbol{X}$ to be missing. Let $\boldsymbol{M} \equiv (M_1, \ldots, M_p)^{\mathrm{T}}$ denote a vector of missing indicators, where $M_j = 1$ if X_j is missing and $M_j = 0$ otherwise. Assume missing at random, such that $\boldsymbol{M}$ and $\boldsymbol{X}$ are independent given $\{X_j : P(M_j = 0) = 1\}$, Y, and Δ. Also, assume that T^* and C are independent given $\{X_j : P(M_j = 0) = 1\}$. For a sample of size n, the observed data consist of $\mathcal{O}_i \equiv \{Y_i, \Delta_i, \boldsymbol{M}_i, \boldsymbol{X}_{i,-\boldsymbol{M}_i}\}$ for $i = 1, \ldots, n$, where $\boldsymbol{X}_{i,-\boldsymbol{M}_i}$ denotes the subvector of $\boldsymbol{X}_i$ consisting of components that correspond to $M_{ij} = 0$. Let $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ and $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \Lambda, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the set of all unknown parameters. The (observed-data) likelihood $L_{\mathrm{obs}}(\boldsymbol{\theta})$ is proportional to

$$\prod_{i=1}^{n} \int \left\{\lambda(Y_i) e^{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}}\right\}^{\Delta_i} e^{-\Lambda(Y_i) e^{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}}} |\boldsymbol{\Sigma}|^{-1/2} \times e^{-\frac{1}{2}(\boldsymbol{X}_i-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})}\,\mathrm{d}\boldsymbol{X}_{i,\boldsymbol{M}_i},$$

where $\boldsymbol{X}_{i,\boldsymbol{M}_i}$ denotes the subvector of $\boldsymbol{X}_i$ consisting of the components that correspond to $M_{ij} = 1$.

We adopt the nonparametric likelihood estimation (NPMLE) approach. Let $t_1 < \cdots < t_m$ be the ordered unique observed event times, where $m = \sum_{i=1}^n \Delta_i$. We set Λ to be a step function that jumps only at $t_1, \ldots, t_m$ and let the corresponding jump sizes be $\lambda_1, \ldots, \lambda_m$. In the likelihood, we replace $\lambda(Y_i)$ by the corresponding jump size. In the sequel, we use L_{obs} to denote this nonparametric version of the likelihood.

2.2 EM algorithm for unpenalized estimation

When the dimension of $\boldsymbol{X}$ is low and we are not interested in variable selection, we estimate $\boldsymbol{\theta}$ by the NPMLE $(\widehat{\boldsymbol{\beta}}, \widehat{\Lambda}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, which is the maximizer of L_{obs}. We adopt the EM algorithm to compute the NPMLE, with $\boldsymbol{X}_{i,\boldsymbol{M}_i}$ treated as missing data for $i = 1, \ldots, n$. The complete-data log-likelihood is

$$\log L_{\mathrm{com}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big\{ \Delta_i (\log \lambda_{l(i)} + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}) - \sum_{j:t_j \le Y_i} \lambda_j e^{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}} - \frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| \Big\} + \text{const},$$

where $l(i)$ is a mapping such that $t_j = Y_{l(i)}$ for $i \in \{i = 1, \ldots, n : \Delta_i = 1\}$, and "const" is a constant term. In the E-step, we evaluate the conditional expectation of $\log L_{\mathrm{com}}(\boldsymbol{\theta})$ given the observed data at the current parameter estimate. In the M-step, we maximize the expected complete-data log-likelihood. In particular, at the $(k+1)$th iteration, we update

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i) \tag{1}$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}) - \left(\boldsymbol{\mu}^{(k+1)}\right)\left(\boldsymbol{\mu}^{(k+1)}\right)^{\mathrm{T}}, \tag{2}$$

where $\widehat{\mathrm{E}}^{(k)}$ denotes conditional expectation given the observed data, evaluated at the parameter estimate at the kth iteration. After profiling out $\lambda_1, \ldots, \lambda_m$, $\boldsymbol{\beta}$ maximizes the following "complete-data log-partial likelihood"

$$G^{(k)}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left[\widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i)^{\mathrm{T}}\boldsymbol{\beta} - \log\left\{ \sum_{j:Y_j \ge Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}}) \right\} \right].$$

This expression is similar to the objective function for $\boldsymbol{\beta}$ in Zeng and Lin (2007). Note that

$$\frac{\partial G^{(k)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \Delta_i \left\{ \widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i) - \frac{\sum_{j:Y_j \ge Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}} \boldsymbol{X}_j)}{\sum_{j:Y_j \ge Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}})} \right\}$$

$$\frac{\partial^2 G^{(k)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = -\sum_{i=1}^{n} \Delta_i \left[\frac{\sum_{j:Y_j \ge Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}} \boldsymbol{X}_j \boldsymbol{X}_j^{\mathrm{T}})}{\sum_{j:Y_j \ge Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}})} \right.$$

$$-\left\{\frac{\sum_{j:Y_j\geq Y_i}\widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}}\boldsymbol{X}_j)}{\sum_{j:Y_j\geq Y_i}\widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}})}\right\}^{\otimes 2}\Bigg].$$

We update $\boldsymbol{\beta}$ by the one-step Newton method:

$$\boldsymbol{\beta}^{(k+1)}=\boldsymbol{\beta}^{(k)}-\left(\frac{\partial^2 G^{(k)}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}\right)^{-1}\left(\frac{\partial G^{(k)}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}\right), \tag{3}$$

where $\boldsymbol{\beta}^{(k)}$ denotes the estimate of $\boldsymbol{\beta}$ at the kth iteration. Finally, we update the baseline hazard function using the Breslow-like estimator:

$$\lambda_{l(i)}^{(k+1)}=\frac{1}{\sum_{j:Y_j\geq Y_i}\widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}^{(k+1)}})} \tag{4}$$

for i such that $\Delta_i=1$.

The major computational challenge of the EM algorithm is that the conditional distribution of $\boldsymbol{X}_i$ does not have a closed form, and direct numerical integration for the expectations is infeasible when the dimension of $\boldsymbol{X}_{i,\boldsymbol{M}_i}$ is moderately high. To avoid multi-dimensional numerical integrations, we propose a transformation approach under which the expectations can be computed using at most one-dimensional numerical integrations.

For notational simplicity, we omit the subject index i in the following. Note that all operations are performed on a per-subject basis, as each subject may have a different missing data pattern. For a p-vector $\boldsymbol{U}$, let $\boldsymbol{U}_{\mathrm{mis}}$ and $\boldsymbol{U}_{\mathrm{obs}}$ denote the subvectors of $\boldsymbol{U}$ consisting of components that correspond to $M_j=1$ and 0, respectively. The expectations that need to be computed in the E-step are in one of the following forms:

$$\widehat{\mathrm{E}}^{(k)}\left\{\exp(\boldsymbol{X}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{mis}}^{(k)})\right\} \tag{5}$$

$$\widehat{\mathrm{E}}^{(k)}\left\{\exp(\boldsymbol{X}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{mis}}^{(k)})\right\} \tag{6}$$

$$\widehat{\mathrm{E}}^{(k)}\{g(\boldsymbol{X}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{mis}}^{(k)})\boldsymbol{X}_{\mathrm{mis}}\} \tag{7}$$

$$\widehat{\mathrm{E}}^{(k)}\{g(\boldsymbol{X}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{mis}}^{(k)})\boldsymbol{X}_{\mathrm{mis}}\boldsymbol{X}_{\mathrm{mis}}^{\mathrm{T}}\} \tag{8}$$

$$\widehat{\mathrm{E}}^{(k)}\left\{\exp(\boldsymbol{X}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{mis}}^{(k+1)})\right\}, \tag{9}$$

where g is either the exponential function or the constant function $g(\cdot)=1$.

First, if $\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}=\boldsymbol{0}$, then the conditional distribution of $\boldsymbol{X}_{\mathrm{mis}}$ given $\mathcal{O}$ is a multivariate normal distribution that does not depend on (Y,Δ). The expectations (5)–(9) have simple closed-form expressions.

For $\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\neq\boldsymbol{0}$, we define an orthogonal matrix $\boldsymbol{\Psi}$ with the first row being $(\boldsymbol{\beta}_{\mathrm{mis}}^{(k)})^{\mathrm{T}}/\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|$ and let $\widetilde{\boldsymbol{X}}=\boldsymbol{\Psi}\boldsymbol{X}_{\mathrm{mis}}$, where $\|\cdot\|$ denotes the L_2-norm. Note that $\boldsymbol{\Psi}$ and quantities defined based on it depend on the iteration number k, but we suppress the index k for simplicity of presentation. We see that the first component of $\widetilde{\boldsymbol{X}}$ is $\widetilde{X}_1=\boldsymbol{X}_{\mathrm{mis}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}/\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|$. Let $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ denote the mean and variance of $\boldsymbol{X}$ given $\boldsymbol{X}_{\mathrm{obs}}$, respectively, such that

$$\begin{aligned}\boldsymbol{\eta}&=\boldsymbol{\Psi}\boldsymbol{\mu}_{\mathrm{mis}}^{(k)}+\boldsymbol{\Psi}\boldsymbol{\Sigma}_{\mathrm{mis,obs}}^{(k)}(\boldsymbol{\Sigma}_{\mathrm{obs,obs}}^{(k)})^{-1}(\boldsymbol{X}_{\mathrm{obs}}-\boldsymbol{\mu}_{\mathrm{obs}}^{(k)})\\ \boldsymbol{\nu}&=\boldsymbol{\Psi}\boldsymbol{\Sigma}_{\mathrm{mis,mis}}^{(k)}\boldsymbol{\Psi}^{\mathrm{T}}-\boldsymbol{\Psi}\boldsymbol{\Sigma}_{\mathrm{mis,obs}}^{(k)}(\boldsymbol{\Sigma}_{\mathrm{obs,obs}}^{(k)})^{-1}\boldsymbol{\Sigma}_{\mathrm{obs,mis}}^{(k)}\boldsymbol{\Psi}^{\mathrm{T}},\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathrm{mis,mis}}$ and $\boldsymbol{\Sigma}_{\mathrm{obs,obs}}$ are the covariance matrices of $\boldsymbol{X}_{\mathrm{mis}}$ and $\boldsymbol{X}_{\mathrm{obs}}$, respectively, $\boldsymbol{\Sigma}_{\mathrm{mis,obs}}$ is the covariance between $\boldsymbol{X}_{\mathrm{mis}}$ and $\boldsymbol{X}_{\mathrm{obs}}$, and $\boldsymbol{\Sigma}_{\mathrm{obs,mis}}=\boldsymbol{\Sigma}_{\mathrm{mis,obs}}^{\mathrm{T}}$. Let $\widetilde{\boldsymbol{X}}_{-1}$ denote the subvector of $\widetilde{\boldsymbol{X}}$ consisting of all except the first component. Although the conditional distribution of $\widetilde{\boldsymbol{X}}$ given the observed data does not have a simple form, $\widetilde{\boldsymbol{X}}_{-1}$ given the observed data and $\widetilde{X}_1$ (at $\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}$) follows the multivariate normal distribution:

$$\begin{aligned}&\widetilde{\boldsymbol{X}}_{-1}\mid(Y,\Delta,\widetilde{X}_1,\boldsymbol{X}_{\mathrm{obs}})\\ &\quad\sim\mathrm{N}\left(\boldsymbol{\eta}_{-1}+\boldsymbol{\nu}_{-1,1}\frac{\widetilde{X}_1-\boldsymbol{\eta}_1}{\boldsymbol{\nu}_{1,1}},\boldsymbol{\nu}_{-1,-1}-\frac{\boldsymbol{\nu}_{-1,1}^{\otimes 2}}{\boldsymbol{\nu}_{1,1}}\right)\\ &\quad\equiv\mathrm{N}(\boldsymbol{m}(\widetilde{X}_1),\boldsymbol{V}),\end{aligned}$$

where $\boldsymbol{\nu}_{1,1}$ is the upper left element of $\boldsymbol{\nu}$, $\boldsymbol{\nu}_{-1,1}$ is the first column of $\boldsymbol{\nu}$ with the first component removed, and $\boldsymbol{\nu}_{-1,-1}$ is the lower right submatrix of $\boldsymbol{\nu}$, with the first row and column of $\boldsymbol{\nu}$ removed. The conditional density of $\widetilde{X}_1$ given $\mathcal{O}$ is proportional to

$$\begin{aligned}&f(\widetilde{x}_1\qquad\qquad;\mathcal{O})\\ &\quad\equiv\exp\left\{\Delta\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|\widetilde{x}_1-\Lambda^{(k)}(Y)e^{\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|\widetilde{x}_1+\boldsymbol{X}_{\mathrm{obs}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{obs}}^{(k)}}\right.\\ &\qquad\left.-\tfrac{1}{2}\boldsymbol{\nu}_{1,1}^{-1}(\widetilde{x}_1-\boldsymbol{\eta}_1)^2\right\}.\end{aligned}$$

Therefore, conditional expectations of functions of $\boldsymbol{X}_{\mathrm{mis}}\equiv\boldsymbol{\Psi}^{\mathrm{T}}\widetilde{\boldsymbol{X}}$ can be computed by first further conditioning on $\widetilde{X}_1$, where the conditional expectations have closed-form expressions, and then taking the expectation over $\widetilde{X}_1$, which can be performed by numerical integration.

Specifically, the expectation (5) is equal to $\widehat{\mathrm{E}}^{(k)}\left\{\exp(\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|\widetilde{X}_1)\right\}$. The expectation (7) is equal to

$$\boldsymbol{\Psi}^{\mathrm{T}}\widehat{\mathrm{E}}^{(k)}\{g(\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|\widetilde{X}_1)\widetilde{\boldsymbol{X}}\}=\boldsymbol{\Psi}^{\mathrm{T}}\widehat{\mathrm{E}}^{(k)}\left\{g(\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|\widetilde{X}_1)\begin{pmatrix}\widetilde{X}_1\\ \boldsymbol{m}(\widetilde{X}_1)\end{pmatrix}\right\}.$$

The expectation (8) is equal to

$$\begin{aligned}&\boldsymbol{\Psi}^{\mathrm{T}}\widehat{\mathrm{E}}^{(k)}\{g(\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|\widetilde{X}_1)\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\}\boldsymbol{\Psi}\\ &=\boldsymbol{\Psi}^{\mathrm{T}}\widehat{\mathrm{E}}^{(k)}\left\{g(\|\boldsymbol{\beta}_{\mathrm{mis}}^{(k)}\|\widetilde{X}_1)\begin{pmatrix}\widetilde{X}_1^2 & \widetilde{X}_1\boldsymbol{m}(\widetilde{X}_1)^{\mathrm{T}}\\ \widetilde{X}_1\boldsymbol{m}(\widetilde{X}_1) & \boldsymbol{V}+\boldsymbol{m}(\widetilde{X}_1)\boldsymbol{m}(\widetilde{X}_1)^{\mathrm{T}}\end{pmatrix}\right\}\boldsymbol{\Psi}.\end{aligned}$$

Finally, to evaluate (9), for any vector $\boldsymbol{a}$ of an appropriate dimension, let

$$\phi(\widetilde{X}_1; \boldsymbol{a}) = \exp\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{m}(\widetilde{X}_1) + \frac{1}{2}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{a}\right\},$$

which is the conditional expectation of $\exp(\widetilde{\boldsymbol{X}}_{-1}^{\mathrm{T}}\boldsymbol{a})$ given the observed data and $\widetilde{X}_1$, evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$. We can write (9) as

$$\begin{aligned}&\widehat{\mathrm{E}}^{(k)}\left\{\exp(\widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{\beta}_{\mathrm{mis}}^{(k+1)})\right\}\\ &\quad= \widehat{\mathrm{E}}^{(k)}\left\{\exp((\boldsymbol{\Psi}\boldsymbol{\beta}_{\mathrm{mis}}^{(k+1)})_1\widetilde{X}_1)\phi(\widetilde{X}_1; (\boldsymbol{\Psi}\boldsymbol{\beta}_{\mathrm{mis}}^{(k+1)})_{-1})\right\}.\end{aligned}$$

Therefore, all expectations involved in the E-step can be computed using one-dimensional numerical integrations over the conditional distribution of $\widetilde{X}_1$. In particular, for any function h, we have

$$\widehat{\mathrm{E}}^{(k)}\{h(\widetilde{X}_1)\} = \frac{\int h(\widetilde{x}_1) f(\widetilde{x}_1; \mathcal{O})\,\mathrm{d}\widetilde{x}_1}{\int f(\widetilde{x}_1; \mathcal{O})\,\mathrm{d}\widetilde{x}_1}.$$

The integrations can be approximated using the adaptive Gauss–Hermite quadrature (Liu and Pierce 1994). The proposed algorithm is summarized in Algorithm 1.

Algorithm 1: NPMLE

Input : $\{\mathcal{O}_i\}_{i=1,2,\ldots,n}$.
1 Initialize $(\boldsymbol{\beta}^{(0)}, \Lambda^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)})$.
2 Calculate (5), (7), and (8) for $i = 1, 2, \ldots, n$ and in turn the gradient and Hessian of $G^{(k)}(\boldsymbol{\beta})$.
3 Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by (1) and (2), respectively.
4 Update $\boldsymbol{\beta}$ by (3).
5 Calculate (9) for $i = 1, 2, \ldots, n$.
6 Update Λ by (4).
7 Repeat Steps 2–6 until convergence.
Output: $(\widehat{\boldsymbol{\beta}}, \widehat{\Lambda}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$.

2.3 EM algorithm for penalized estimation

When the number of covariates is large, it is often desirable to select a subset of covariates that are associated with the survival time. We propose a penalization approach with the following penalized observed-data log-likelihood:

$$p\ell(\boldsymbol{\theta}) = \log L_{\mathrm{obs}}(\boldsymbol{\theta}) - n\gamma\|\boldsymbol{\beta}\|_1,$$

where $\gamma > 0$ is a tuning parameter. The penalized NPMLE is the maximizer of $p\ell(\boldsymbol{\theta})$. Note that we assume that the sample size is sufficiently larger than the number of covariates, so no penalty is imposed for the covariance matrix $\boldsymbol{\Sigma}$.

To compute the penalized NPMLE, we adopt the proposed EM algorithm for the unpenalized estimator with some modifications. The E-step for the penalized estimator is the same as the previous algorithm. In the M-step, the estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the same as before.

After profiling out the baseline hazard function, $\boldsymbol{\beta}$ maximizes the (expected) penalized complete-data log-partial likelihood $n^{-1}G^{(k)}(\boldsymbol{\beta}) - \gamma\|\boldsymbol{\beta}\|_1$. Clearly, there is no closed-form solution, and the objective function is not differentiable. To update $\boldsymbol{\beta}$, we first approximate the objective function using a second-order Taylor expansion:

$$n^{-1}G^{(k)}(\boldsymbol{\beta}) - \gamma\|\boldsymbol{\beta}\|_1 \approx \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\beta} + \boldsymbol{P}^{\mathrm{T}}\boldsymbol{\beta} - \gamma\|\boldsymbol{\beta}\|_1 + \text{const}, \tag{10}$$

where

$$\begin{aligned}\boldsymbol{Q} &= \frac{1}{n}\left(\frac{\partial^2 G^{(k)}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}\right)\\ \boldsymbol{P} &= \frac{1}{n}\left\{\left(\frac{\partial G^{(k)}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}\right) - \left(\frac{\partial^2 G^{(k)}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}\right)\boldsymbol{\beta}^{(k)}\right\}.\end{aligned}$$

Then, to maximize the right-hand side of (10), we adopt the coordinate-descent algorithm (Simon et al. 2011). For $j = 1, \ldots, p$, we update β_j with $\boldsymbol{\beta}_{-j}$ fixed at the current estimates by setting

$$\beta_j = -\frac{S(\boldsymbol{Q}_{j,-j}\boldsymbol{\beta}_{-j} + \boldsymbol{P}_j, \gamma)}{\boldsymbol{Q}_{j,j}}, \tag{11}$$

where $S(x, \gamma) = \mathrm{sgn}(x)(|x| - \gamma)_+$. We iterate over components of $\boldsymbol{\beta}$ until convergence. After updating $\boldsymbol{\beta}$, we update Λ using the same Breslow-like estimator as before. This completes a single M-step. We summarize the procedure in Algorithm 2.

Algorithm 2: Penalized NPMLE

Input : $\{\mathcal{O}_i\}_{i=1,2,\ldots,n}$ and γ.
1 Standardize the covariates. Initialize $(\boldsymbol{\beta}^{(0)}, \Lambda^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)})$.
2 Calculate (5), (7), and (8) for $i = 1, 2, \ldots, n$ and in turn the gradient and Hessian of $G^{(k)}(\boldsymbol{\beta})$.
3 Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by (1) and (2), respectively.
4 Iteratively update each component of $\boldsymbol{\beta}$ through (11) until convergence.
5 Calculate (9) for $i = 1, 2, \ldots, n$.
6 Update Λ through (4).
7 Repeat Steps 2–6 until convergence.
8 Rescale the NPMLE.
Output: $(\widehat{\boldsymbol{\beta}}_\gamma, \widehat{\Lambda}_\gamma, \widehat{\boldsymbol{\mu}}_\gamma, \widehat{\boldsymbol{\Sigma}}_\gamma)$.

To choose the tuning parameter, we use the corrected Akaike information criterion (AICc) (Sugiura 1978; Hurvich

and Tsai 1989) as our selection criterion:

$$\mathrm{AICc}(\gamma) = -2\log L_{\mathrm{obs}}(\widehat{\boldsymbol{\theta}}_\gamma) + 2k_\gamma + \frac{2k_\gamma(k_\gamma+1)}{n-k_\gamma-1},$$

where k_γ represents the number of nonzero components in $\widehat{\boldsymbol{\beta}}_\gamma$.

2.4 Relaxation of the normality assumption on complete covariates

In the previous subsections, we restrict all covariates to be jointly normal, but this assumption may not hold in practice. In this subsection, we relax this restriction: rather than assuming that the entire covariate vector is normal, we assume that the missing covariates follow a normal distribution conditional on the observed covariates.

To facilitate presentation, we redefine $\boldsymbol{X}$ as the set of potentially missing covariates (that is, covariates with $P(M_j = 1) > 0$) with dimension p, and we let $\boldsymbol{Z}$ be the fully observed covariates with dimension q. Here, we do not make distributional assumptions on $\boldsymbol{Z}$. The survival model becomes:

$$\lambda(t \mid \boldsymbol{X}, \boldsymbol{Z}) = \lambda(t)e^{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_2},$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are p- and q-dimensional vectors of coefficients, respectively, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$. We consider a linear model for $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}^* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Z}^* = (1, \boldsymbol{Z}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{A}$ is $p \times (q+1)$ matrix of regression coefficients, and $\boldsymbol{\varepsilon}$ is p-vector of Gaussian noise with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Thus, the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$ is $\mathrm{N}(\boldsymbol{A}\boldsymbol{Z}^*, \boldsymbol{\Sigma})$. The contribution of a subject to the complete-data likelihood is given by

$$\left\{\lambda_{l(i)}e^{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_2}\right\}^{\Delta_i} e^{-\Lambda(Y_i)e^{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_2}} \times |\Sigma|^{-1/2} e^{-\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{A}\boldsymbol{Z}_i^*)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{A}\boldsymbol{Z}_i^*)}$$

The proposed EM algorithm can be adopted with slight modifications. In particular, at the $(k+1)$th iteration, we update

$$\boldsymbol{A}^{(k+1)} = \left(\sum_{i=1}^{n}\widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i)\boldsymbol{Z}_i^{*\mathrm{T}}\right)\left(\sum_{i=1}^{n}\boldsymbol{Z}_i^*\boldsymbol{Z}_i^{*\mathrm{T}}\right)^{-1} \tag{12}$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}) - \frac{1}{n}\left(\sum_{i=1}^{n}\widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i)\boldsymbol{Z}_i^{*\mathrm{T}}\right)\left(\sum_{i=1}^{n}\boldsymbol{Z}_i^*\boldsymbol{Z}_i^{*\mathrm{T}}\right)^{-1} \times \left(\sum_{i=1}^{n}\boldsymbol{Z}_i^*\widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i)^{\mathrm{T}}\right). \tag{13}$$

Let $\boldsymbol{X}_{\mathrm{mis}}$ and $\boldsymbol{X}_{\mathrm{obs}}$ denote the missing and observed components of $\boldsymbol{X}$, respectively. Similarly, let $\boldsymbol{\beta}_{\mathrm{mis}}$ denote the subvector of $\boldsymbol{\beta}$ that corresponds to $\boldsymbol{X}_{\mathrm{mis}}$. In the E-step, expectations (5), (7), (8), and (9) are required. Note that the transformation technique remains applicable under this formulation. In particular, we first define $\boldsymbol{\Psi}$ in the same way as in Section 2.2. The mean and variance of $\widetilde{\boldsymbol{X}} = \boldsymbol{\Psi}\boldsymbol{X}_{\mathrm{mis}}$ given $(\boldsymbol{X}_{\mathrm{obs}}, \boldsymbol{Z})$ are

$$\boldsymbol{\eta} = \boldsymbol{\Psi}(\boldsymbol{A}^{(k)}\boldsymbol{Z}^*)_{\mathrm{mis}} + \boldsymbol{\Psi}\boldsymbol{\Sigma}^{(k)}_{\mathrm{mis,obs}}(\boldsymbol{\Sigma}^{(k)}_{\mathrm{obs,obs}})^{-1}(\boldsymbol{X}_{\mathrm{obs}} - (\boldsymbol{A}^{(k)}\boldsymbol{Z}^*)_{\mathrm{obs}})$$

$$\boldsymbol{\nu} = \boldsymbol{\Psi}\boldsymbol{\Sigma}^{(k)}_{\mathrm{mis,mis}}\boldsymbol{\Psi}^{\mathrm{T}} - \boldsymbol{\Psi}\boldsymbol{\Sigma}^{(k)}_{\mathrm{mis,obs}}(\boldsymbol{\Sigma}^{(k)}_{\mathrm{obs,obs}})^{-1}\boldsymbol{\Sigma}^{(k)}_{\mathrm{obs,mis}}\boldsymbol{\Psi}^{\mathrm{T}}.$$

The remaining procedures are the same as described in Section 2.2. Algorithm 3 summarizes the computation procedure of the NPMLE under this relaxed assumption.

Algorithm 3: NPMLE under relaxed assumption

Input : $\{\mathcal{O}_i\}_{i=1,2,\ldots,n}$.
1 Initialize $(\boldsymbol{\beta}^{(0)}, \Lambda^{(0)}, \boldsymbol{A}^{(0)}, \boldsymbol{\Sigma}^{(0)})$.
2 Calculate (5), (7), and (8) for $i = 1, 2, \ldots, n$ and in turn the gradient and Hessian of $G^{(k)}(\boldsymbol{\beta})$.
3 Update $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ by (12) and (13), respectively.
4 Update $\boldsymbol{\beta}$ by (3).
5 Calculate (9) for $i = 1, 2, \ldots, n$.
6 Update Λ by (4).
7 Repeat Steps 2–6 until convergence.
Output: $(\widehat{\boldsymbol{\beta}}, \widehat{\Lambda}, \widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\Sigma}})$.

3 Simulation studies

3.1 Unpenalized estimation

In this subsection, we evaluate the empirical performance of the unpenalized NPMLE and some competing methods. We considered $p = 5$ and generated $\boldsymbol{X}$ from $\mathrm{N}(\boldsymbol{0}, (0.5^{|i-j|})_{i,j=1,\ldots,p})$. We set $\boldsymbol{\beta} = (0.3, 0.3, 0.3, 0.3, 0.3)^{\mathrm{T}}$ and $\Lambda(t) = 0.1t^2$. We generated the censoring time from $\mathrm{Unif}(0, 5)$, resulting in a censoring rate of approximately 55%. We considered sample sizes of $n = 300$ and 1000.

We considered missing completely at random (MCAR) and missing at random (MAR) mechanisms. The covariate vector $\boldsymbol{X}$ was partitioned into four blocks: $\{X_1, X_2\}$, $\{X_3\}$, $\{X_4\}$, and $\{X_5\}$. Under MCAR, subjects were randomly selected to have incomplete covariates, and one block of variables was then randomly chosen to be missing for

each selected subject. Under MAR, we mimicked a case-cohort study: a subcohort comprising 10% of the sample had all covariates observed. Outside this subcohort, we randomly selected subjects with $\Delta = 1$ to have complete covariates. If necessary, we randomly selected subjects with $\Delta = 0$ to reach the target missing proportion. Similarly, one block was randomly chosen to be missing for each subject selected to have incomplete covariates. We considered missing proportions of 50% and 75%.

We compare NPMLE with the complete-case analysis (CCA), IPW, single imputation (SI), MICE, and MIXGB. For CCA, we discarded subjects with any missing values. For IPW, we maximized the weighted Cox nonparametric log-likelihood, using inverse (true) propensity scores as weights. For SI, we estimated $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using MLE based on the fully observed $\boldsymbol{X}_i$'s and then imputed missing entries by their estimated conditional mean given the partially observed $\boldsymbol{X}_i$'s. MICE and MIXGB are two multiple imputation approaches. We generated $m = 5$ imputed datasets using the `mice` and `mixgb` packages in R (van Buuren and Groothuis-Oudshoorn 2011; Deng and Lumley 2023), where we used the default hyperparameter values and included the event indicators and the Nelson–Aalen estimates in the imputation models. For `mixgb`, we selected the optimal number of boosting rounds using the built-in cross-validation function. Final estimates were pooled using Rubin (2004)'s rule. In this and subsequent simulation studies, we considered 500 replicates unless otherwise specified.

Performance of the methods is evaluated using two metrics: the empirical mean squared error (MSE) of $\widehat{\boldsymbol{\beta}}$, which quantifies estimation accuracy, and the concordance index (C-index) (Harrell et al. 1996) between the event time and $\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$, which measures capacity to correctly rank individuals by their risk. The C-index is calculated on a fully observed validation dataset of 1000 subjects without censoring, generated from the same population.

Table 1 presents the results for unpenalized estimation under MCAR and MAR. Across all scenarios, all methods yield similar C-index values. By contrast, NPMLE, SI, and MICE consistently achieve lower MSEs compared to CCA, IPW, and MIXGB, particularly at higher missing proportions. Figure 1 shows the empirical mean of $\widehat{\Lambda}$ for unpenalized estimation under missing proportion of 50%. Under MCAR, $\widehat{\Lambda}$ are unbiased for all methods. Under MAR, only NPMLE, SI, MICE, and MIXGB yield unbiased estimates.

3.2 Penalized estimation

In this subsection, we compare the performance of penalized methods under a sparse setup. We considered $p = 100$ and set nonzero $\beta_j = 0.5$ for $j = 10, 20, \ldots, 100$. The remaining configuration for data generation was the same as that described in Section 3.1. The censoring rate is approximately 55%. We considered both MCAR and MAR mechanisms. The covariate vector $\boldsymbol{X}$ was partitioned into 20 blocks: $\{X_1, X_2, X_3, X_4, X_5\}$, $\{X_6, X_7, X_8, X_9, X_{10}\}$, …, $\{X_{96}, X_{97}, X_{98}, X_{99}, X_{100}\}$. We inserted missing data into the simulated dataset according to the procedure described in Section 3.1.

We considered NPMLE, CCA, IPW, SI, MICE, and MIXGB approaches, each with a Lasso penalty. We used the R package `glmnet` (Friedman et al. 2010; Simon et al. 2011; Tay et al. 2023) to perform Lasso estimation for all methods except NPMLE, with the optimal tuning parameter value selected via 10-fold cross-validation. For CCA, IPW, and SI, missing data were handled as previously described. For multiple imputation, each simulated dataset was imputed $m = 20$ times, as this is required for more stable variable selection. To ensure convergence of the chained equations, the number of iterations for MICE was set to 20. For both multiple imputation methods, Lasso estimates were obtained using the stacked method (Du et al. 2022).

In addition to MSE and C-index, we also report the true positive rate (TPR), which is the proportion of truly relevant variables that are identified, and false discovery rate (FDR), which is the proportion of selected variables that are irrelevant.

Table 2 presents the results for penalized estimation under MCAR and MAR. Overall, the performance of NPMLE, SI, MICE, and MIXGB are similar. NPMLE consistently achieves the lowest or near-lowest MSE and near-highest C-index among all methods. In term of variable selection, NPMLE maintains a high TPR, but it tends to have higher FDR than the other methods. Figure 2 shows the empirical mean of $\widehat{\Lambda}$ for penalized estimation under 50% missing. CCA and IPW give biased estimates under MAR, whereas NPMLE as well as the imputation methods give virtually unbiased estimates in all cases.

3.3 Unpenalized estimation under a misspecified outcome model

We evaluate the performance of the proposed methods under a misspecified outcome model. We set $p = 5$ and generated the covariates and the censoring time in the same manner as described in Section 3.1. The true outcome model was set to be

$$\lambda(t \mid \boldsymbol{X}) = 0.2te^{0.3X_1+0.3X_2+0.3X_3+0.3X_4+0.3X_5-0.5X_1X_2-0.5X_1X_3+0.3X_4^2}.$$

For each method, we fitted a misspecified Cox model with only linear terms in the regression equation, without quadratic terms. The implementation details of the estimation methods and the missing data mechanisms are the same as those described in Section 3.1. In this setting, MSE cannot be evaluated, and we only report the C-index values.

The simulation results are presented in Table S1 in the Supplementary Materials. The performance of NPMLE, SI,

Table 1 Simulation results for unpenalized estimation

Sample size	Missing proportion	Statistic	NPMLE	CCA	IPW	SI	MICE	MIXGB
MCAR								
300	50%	MSE	0.0810	0.1551	0.1551	0.0786	0.0761	0.0895
		C-index	0.7203	0.7162	0.7162	0.7202	0.7202	0.7201
	75%	MSE	0.0911	0.3845	0.3845	0.0890	0.0845	0.1028
		C-index	0.7197	0.7060	0.7060	0.7193	0.7194	0.7194
1000	50%	MSE	0.0215	0.0377	0.0377	0.0208	0.0204	0.0257
		C-index	0.7240	0.7229	0.7229	0.7240	0.7240	0.7238
	75%	MSE	0.0253	0.0831	0.0831	0.0251	0.0242	0.0330
		C-index	0.7237	0.7200	0.7200	0.7236	0.7236	0.7233
MAR								
300	50%	MSE	0.0765	0.1109	0.2108	0.0769	0.0759	0.0821
		C-index	0.7206	0.7157	0.7112	0.7205	0.7204	0.7202
	75%	MSE	0.0867	0.2013	0.3536	0.0843	0.0798	0.0957
		C-index	0.7200	0.7069	0.7048	0.7198	0.7198	0.7197
1000	50%	MSE	0.0205	0.0682	0.0756	0.0205	0.0209	0.0239
		C-index	0.7241	0.7224	0.7203	0.7240	0.7240	0.7239
	75%	MSE	0.0242	0.0718	0.1002	0.0240	0.0236	0.0316
		C-index	0.7238	0.7200	0.7189	0.7238	0.7238	0.7235

MICE, and MIXGB is similar, and these methods consistently yield higher C-index values than CCA and IPW. The results indicate that NPMLE and other imputation methods are robust to misspecification of the outcome model to a certain extent.

3.4 Unpenalized and penalized estimation under a misspecified covariate distribution

To assess the sensitivity of the proposed methods to the normality assumption, we conduct simulation studies using misspecified covariate distributions. Specifically, we first generated a multivariate normal random vector as previously described. Each component of this vector was then transformed using $F^{-1} \circ \Phi$, where Φ and F denote the cumulative distribution functions of the standard normal distribution and a specific target distribution, respectively. This transformation ensures that each covariate marginally follows the target distribution. We considered target distributions of the Student's t distribution with 5 degrees of freedom, the exponential distribution with a rate of 1, and the Gaussian mixture distribution given by $0.5\,\mathrm{N}(-2, 1) + 0.5\,\mathrm{N}(2, 1)$.

We simulated the performance of both unpenalized and penalized estimation. The data generation models, missing data mechanisms and implementation details for unpenalized and penalized estimation follow those described in Sections 3.1 and 3.2, respectively. We considered a sample size of 1000 and a missing proportion of 50%. We also considered both MCAR and MAR.

Model performance is assessed using MSE and C-index, with TPR and FDR additionally evaluated for penalized estimation. Tables S2 and S3 in the Supplementary Materials present the results for the unpenalized and penalized estimation respectively. The performance of NPMLE is similar to SI, MICE, and MIXGB under the student's t and the exponential distribution. NPMLE remains stable even when the covariate distribution deviates substantially from normality, demonstrating its robustness to distributional misspecification.

4 Real data analysis

We analyzed the TCGA kidney renal clear cell carcinoma (KIRC) dataset which was released in November 2015 and accessed through the `RTCGA` package in R (Kosinski 2025). This study considered outcomes such as time to new tumor events or death, which were potentially subject to right censoring. The omics variables included 20,531 RNA-sequencing gene expressions (available for 532 subjects) and 217 reverse-phase protein array protein expressions (available for 475 of the 532 subjects). We specifically investigated associations between these omics variables and time to death from initial diagnosis.

We applied a $\log_2(1 + x)$-transformation to the gene expression. To select high-variance genes while minimizing redundancy, we adopted the following procedure: First, gene expressions with zero median absolute deviation were

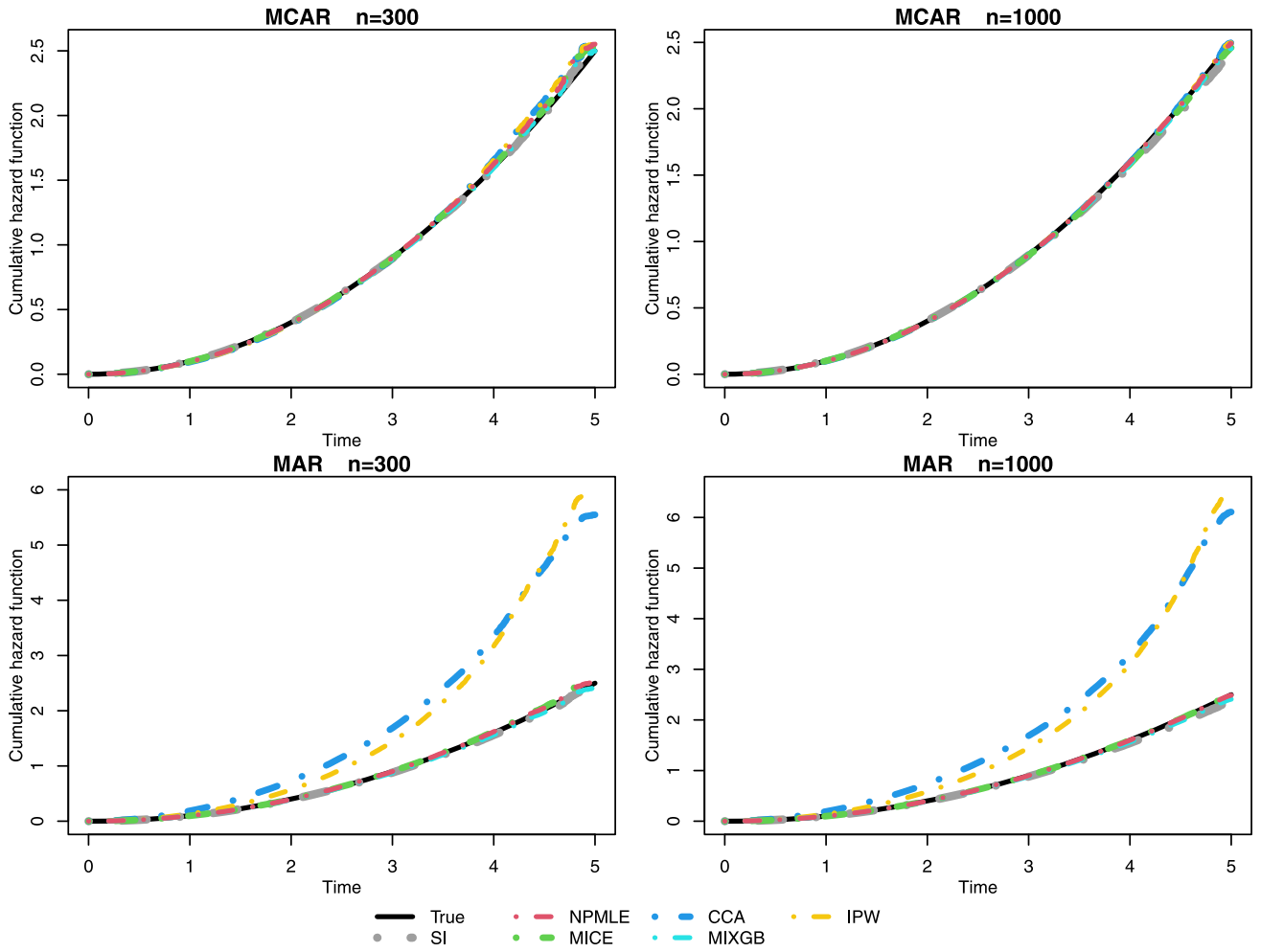

Fig. 1 Estimates of cumulative baseline hazard function for unpenalized estimation under 50% missing in the simulation studies.

filtered. Subsequently, hierarchical clustering (using 1 − |correlation| as the distance metric and complete linkage) partitioned the remaining genes into 2500 clusters; within each cluster, the gene exhibiting the highest median absolute deviation was retained (see Section 10.14.2 of Duda et al. (2000)). For protein expressions, we removed 5 proteins missing in 90% of the subjects. Finally, we performed supervised screening by fitting Cox models marginally on each retained gene and protein expression against time to death. Using complete-case analysis to handle missing values, we selected the top 150 omic variables (136 gene and 14 protein expressions) with the smallest p-values for downstream analysis.

We applied the NPMLE, CCA, IPW, SI, MICE, and MIXGB approaches with a Lasso penalty to the processed data. The implementation details for these methods, except IPW, follow those outlined in Section 3.2. For IPW, propensity scores were estimated using a logistic regression model with a Lasso penalty, incorporating the fully observed covariates, event indicators, and Nelson–Aalen estimates.

Since the true model is unknown, we are not able to use MSE, TPR, and FDR to evaluate the performance of different methods. We focused on the C-index, which was obtained through the following procedure:

1. Generate 20 imputed datasets using the `mixgb` package.
2. Randomly assign the subjects into a training set and a testing set with a 7:3 ratio. For subjects in the training set, remove all imputed entries. Note that this yields 20 testing sets, which consist of the same set of subjects but different imputed values, and a single training set with missing values.
3. Perform the six estimation procedures on the training set and obtain parameter estimates $\widehat{\boldsymbol{\beta}}$. For each approach, estimate the C-index on the 20 testing sets.
4. Compute the average of the C-index values across the 20 testing sets.
5. Repeat Steps 1–4 for 500 times.

Table 2 Simulation results for penalized estimation

Sample size	Missing proportion	Statistic	NPMLE	CCA	IPW	SI	MICE	MIXGB
MCAR								
300	50%	MSE	0.5778	1.1689	1.1689	0.6202	0.5514	0.5364
		C-index	0.7856	0.7426	0.7426	0.7877	0.7889	0.7901
		TPR	0.9984	0.8884	0.8884	0.9964	0.9988	0.9988
		FDR	0.6761	0.5802	0.5802	0.6269	0.6536	0.6332
	75%	MSE	0.5831	2.0922	2.0922	0.8468	0.5813	0.5610
		C-index	0.7852	0.6309	0.6309	0.7793	0.7876	0.7893
		TPR	0.9994	0.4002	0.4002	0.9926	0.9986	0.9990
		FDR	0.6670	0.4741	0.4741	0.6260	0.6586	0.6294
1000	50%	MSE	0.1475	0.2778	0.2778	0.1630	0.1645	0.1551
		C-index	0.8084	0.8039	0.8039	0.8094	0.8091	0.8095
		TPR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		FDR	0.7027	0.6560	0.6560	0.6814	0.6992	0.6850
	75%	MSE	0.1526	0.6031	0.6031	0.1980	0.1800	0.1659
		C-index	0.8083	0.7858	0.7858	0.8088	0.8088	0.8092
		TPR	1.0000	0.9956	0.9956	1.0000	1.0000	1.0000
		FDR	0.7030	0.6215	0.6215	0.6792	0.7020	0.6811
MAR								
300	50%	MSE	0.5361	1.6571	1.6337	0.5789	0.5254	0.5292
		C-index	0.7875	0.7105	0.6813	0.7893	0.7903	0.7907
		TPR	0.9992	0.7416	0.6400	0.9976	0.9990	0.9986
		FDR	0.6702	0.5499	0.5677	0.6329	0.6469	0.6326
	75%	MSE	0.5865	2.2096	2.1762	0.8167	0.5786	0.5613
		C-index	0.7857	0.6101	0.6067	0.7809	0.7882	0.7895
		TPR	0.9982	0.3202	0.3140	0.9918	0.9978	0.9982
		FDR	0.6622	0.4337	0.4521	0.6203	0.6532	0.6273
1000	50%	MSE	0.1422	0.6544	0.4513	0.1496	0.1493	0.1493
		C-index	0.8086	0.8018	0.7929	0.8095	0.8093	0.8095
		TPR	1.0000	1.0000	0.9996	1.0000	1.0000	1.0000
		FDR	0.7056	0.6769	0.7084	0.6871	0.6973	0.6856
	75%	MSE	0.1534	0.8612	0.6981	0.1890	0.1738	0.1641
		C-index	0.8083	0.7843	0.7778	0.8090	0.8088	0.8093
		TPR	1.0000	0.9948	0.9868	1.0000	1.0000	1.0000
		FDR	0.7046	0.6310	0.6569	0.6859	0.7090	0.6847

The above procedure yields 500 C-index values. In addition, we considered scenarios with extra missing data, where in Step 2 we removed protein expression values from randomly selected subjects to yield missing proportions of 30% and 50%. The distributions of the 500 C-index values are presented in Figures S1–3 in the Supplementary Materials for the original data and for 30% and 50% missingness, respectively.

Overall, the C-index values of NPMLE consistently have the highest or near-highest median and the tightest spread, whereas SI tends to have the lowest median. SI, MICE, and MIXGB have relatively more outliners in the lower tail under heavier missingness. Table 3 presents the average C-index values of the six approaches. The average C-index values of the six methods are similar, with the NPMLE having a slight advantage for all missing proportions.

5 Computation time

In this section, we compare the computation time for methods considered in Sections 3.1 and 3.2. All computations were conducted using R 4.5.1 on a 64-bit Windows 11 PC with an Intel i7-14700 CPU, an Nvidia GeForce RTX4060 Ti GPU,

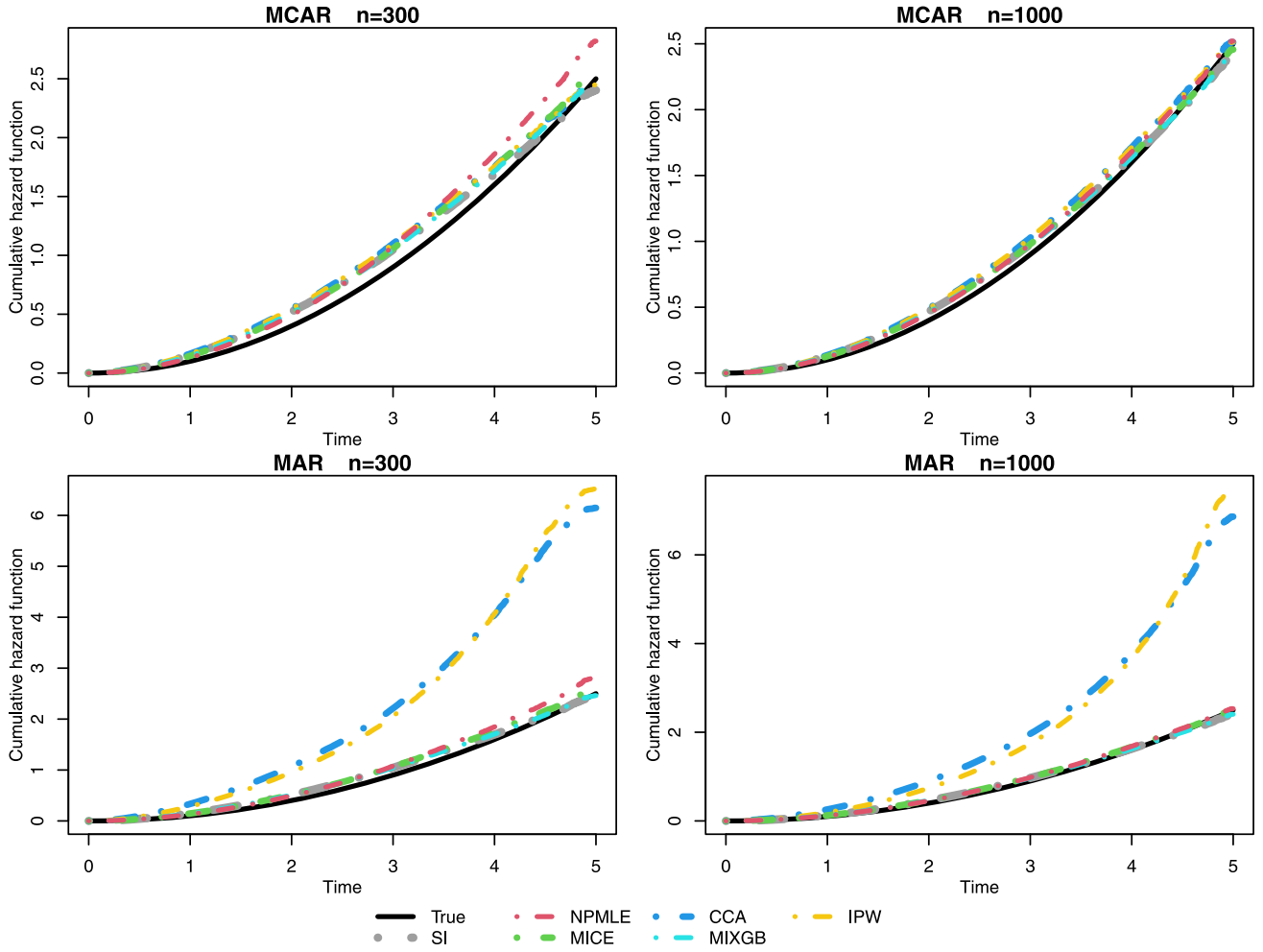

Fig. 2 Estimates of cumulative baseline hazard function for penalized estimation under 50% missing in the simulation studies.

Table 3 C-index values of different methods on the KIRC dataset

Missing proportion	NPMLE	CCA	IPW	SI	MICE	MIXGB
Original	0.7231	0.7183	0.7185	0.7154	0.7178	0.7189
30%	0.7192	0.7138	0.7142	0.7088	0.7157	0.7149
50%	0.7160	0.7078	0.7078	0.7062	0.7111	0.7115

and 64GB of RAM. They were performed using a single CPU core, without the use of GPU or parallel processing.

We first evaluate the computation time over simulated datasets. In particular, we set $n = 1000$ and $p = 20, 30, \ldots, 100$. We generated $\boldsymbol{X}$ and the censoring time from $\mathrm{N}(\mathbf{0}, (0.5^{|i-j|})_{i,j=1,\ldots p})$ and $\mathrm{Unif}(0, 5)$, respectively. We set $\Lambda(t) = 0.1t^2$ and $\boldsymbol{\beta}$ to be sparse, where $\beta_j = 0.5$ for $j = 1, 2, \ldots, 5$ and is zero elsewhere. We introduced missing data through the procedure described in Section 3.2, where we partitioned the covariates into blocks, with each block containing 5 consecutive covariates. We considered the MCAR mechanism and a missing proportion of 50%.

We incorporated a Lasso penalty to all methods. Note that we did not fully adopt the default hyperparameter values in the R packages `mice` and `mixgb`. In particular, we increased the number of imputations from the default of $m = 5$ to $m = 20$, as a larger m is needed for stable variable selection. Also, for MICE, the number of iterations was increased from 5 to 20 to ensure convergence of the chained equations. For each method, we used 100 logarithmically-spaced tuning parameter values ranging from $0.05\gamma_{\max}$ to $\gamma_{\max}$, where $\gamma_{\max}$ is the smallest tuning parameter value such that $\widehat{\boldsymbol{\beta}}_{\gamma_{\max}} = \mathbf{0}$. For NPMLE, we used 20 nodes for the Gauss–Hermite quadrature in the E-step, and the convergence criterion was $\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}\|_\infty < 10^{-4}$.

Table 4 Computation times (in seconds) on real and simulated datasets

Dataset	Dimension	Missing proportion	NPMLE	CCA	IPW	SI	MICE	MIXGB
KIRC	532×152	10.53%	52.38	0.40	1.66	0.45	131.03	21.20
Simulation	1000×102	50%	61.00	0.26	0.32	0.46	651.45	63.61

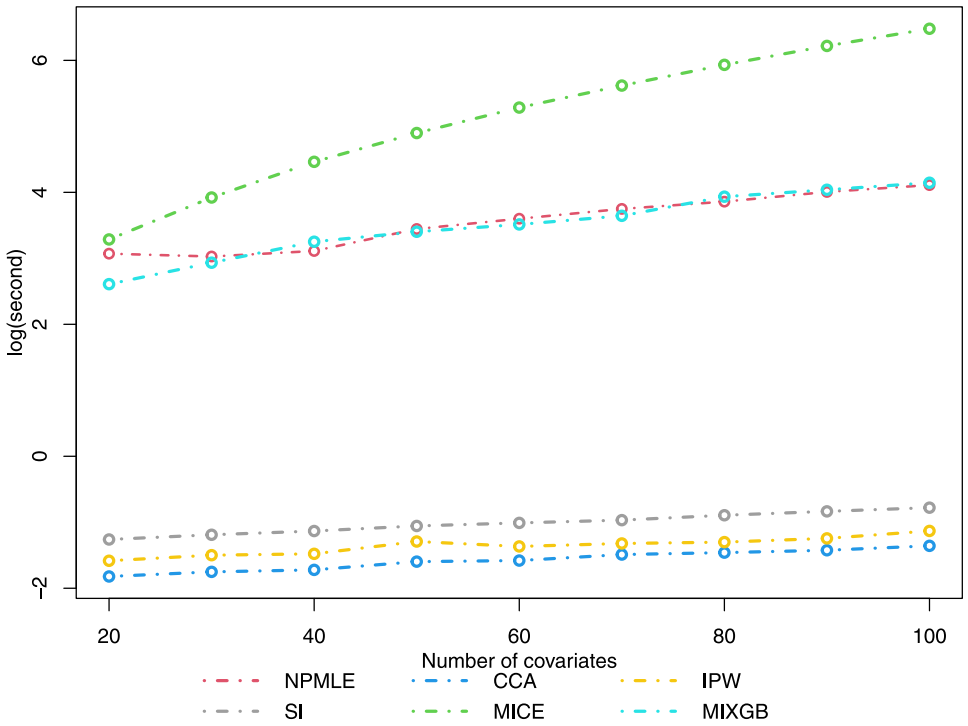

Fig. 3 Computation times on simulated datasets.

Table 4 and Figure 3 present the computation times for the six methods. We also include the computation times for the analysis of the KIRC dataset described in Section 4. The presented computation times are averages over 5 replicates, where a procedure is repeatedly applied to the same dataset 5 times. CCA, IPW, and SI are the fastest, with computations completed within 0.5 seconds in all simulation studies. These methods take only slightly longer time for the KIRC dataset. By contrast, NPMLE, MICE, and MIXGB are substantially slower. Overall, MICE is the slowest method for both the simulated and KIRC datasets. In the simulation studies, the computation times for NPMLE and MIXGB are similar. However, the computation of NPMLE is slower than MIXGB on the KIRC dataset.

Note that the computation time for NPMLE is particularly sensitive to the number of tuning parameter values considered, more so than for imputation methods. For each tuning parameter value, the proposed method repeatedly performs the E-step and M-step until convergence, with each M-step being comparable in complexity to running the entire coordinate descent algorithm on a complete dataset. As the number of tuning parameter values increases, the total computation time for NPMLE accumulates rapidly. In contrast, for a complete dataset, the coordinate descent algorithm for each tuning parameter value is relatively fast, so increasing the number of tuning parameter values does not substantially increase the computational burden. For imputation methods, the main computational cost arises from generating imputed datasets, making their overall computation time less sensitive to the number of tuning parameter values.

6 Discussion

In this paper, we propose a likelihood-based approach for (penalized) estimation of the Cox proportional hazards model, where covariates may be missing. We devise a novel EM algorithm that enables efficient computation under arbitrary missing patterns and a large number of missing covariates. Instead of performing multi-dimensional numerical integration over all dimensions of the missing covariates, we propose a linear transformation of the covariates so that the expectations of all but one component of the transformed variables have closed-form expressions. While the proposed methods demonstrate superior or competitive performance compared to some popular alternatives we considered, our primary objective is not to claim the overall superiority of the likelihood approach. Rather, we aim to illustrate that, even

with many missing covariates and arbitrary missingness patterns, the likelihood-based method remains a practical and viable option for researchers.

The likelihood approach offers several advantages over the popular multiple imputation approach. First, under multiple imputation, it is often difficult to explicitly define the estimator, as it is typically the limit of some iterative algorithm. This makes theoretical studies of the estimator very challenging. By contrast, the proposed estimator is the maximizer of the penalized likelihood, and existing techniques (such as Wang and Leng (2007)) can be applied to establish the theoretical properties. In fact, with a correctly specified model and a finite-dimensional setting, the likelihood approach is (semiparametrically) efficient under regularity conditions. Second, imputation method such as MICE may be computational intensive when the dimension of covariates is high or when the number of imputations is large. Sophisticated imputation models require substantial tuning, which may significantly slow down computation. By contrast, in the setting considered in Section 5, the proposed approach scales better with the number of covariates compared to MICE. Overall, we expect that the proposed methods have superior performance when the model is correctly specified and when the sample size is large.

Nevertheless, the proposed methods have several limitations. A major limitation is that the transformation technique depends crucially on the Gaussian assumptions on the covariates. Specifically, we make use of the facts that linear transformations of Gaussian random vectors are Gaussian, and that the conditional distribution of any subvector of a Gaussian random vector, given the remaining components, is also Gaussian. A related issue is that the proposed methods do not accommodate interaction terms or nonlinear effects even if the individual covariates are Gaussian, because these polynomial or interaction terms are not Gaussian. In addition, we require that the variance estimator is positive definite at every iteration of the EM algorithm. This condition is violated when $p > n$ and may also fail when the number of missing entries is large relative to np.

In settings where these limitations apply, imputation methods may be appealing alternatives. First, imputation methods such as MIXGB and MissForest (Stekhoven and Bühlmann 2012) are fully nonparametric and impose no distributional assumptions on the covariates. They can handle continuous and categorical variables. These methods also naturally accommodate interaction and high-order terms, as such transformations can be applied after the imputation step. In addition, there are imputation approaches that can handle high-dimensional missing covariates through regularization.

The proposed methods can be extended in several directions. First, to accommodate high-dimensional settings, we can consider shrinkage estimators for covariance estimation (Ledoit and Wolf 2004; Warton 2008), which ensure the variance estimator is positive definite. Alternatively, we could impose structures on the covariance matrix to facilitate estimation. For example, we may fit a factor model for $\boldsymbol{X}$, such that $\boldsymbol{\Sigma}$ can be decomposed into a low-rank matrix plus a sparse or diagonal matrix (Fan et al. 2008). These approaches would require modifications to the M-step of the proposed algorithm, but the E-step remains the same.

Second, the current proposed methods cannot handle big data efficiently. To analyze a large dataset with, say, millions of subjects, we can use a divide-and-conquer strategy (Wang et al., 2021;Wang et al., 2022): partition the dataset into blocks, analyze each block using the proposed methods, and then aggregate the resulting estimators. By combining parallel processing technique with this strategy, we may obtain an overall estimator within a tractable amount of time.

Finally, the proposed transformation technique can also be applied to random effects models with Gaussian latent variables (Papageorgiou et al. 2019; Sun et al. 2019; Wong et al. 2022). In general, we can accommodate an outcome variable that follows a survival model or a generalized linear model that regresses on a linear combination of Gaussian latent variables. This outcome variable can also be jointly modeled with other Gaussian outcomes that regress linearly on the random effects. To compute the MLE, we can develop a similar EM algorithm, where in the E-step, we transform the latent variable vector such that the first component is the linear combination present in the survival or generalized linear model.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-026-10849-1.

Acknowledgements The authors gratefully acknowledge the Guang-Dong Basic and Applied Basic Research Foundation (Project No. 2021A1515110048) and the Hong Kong Research Grants Council under Grant 15303422.

Author Contributions K.Y.W. conceived and designed the study. K.Y.W. and N.S.K. developed the proposed methods. N.S.K. conducted the simulation studies and data analysis. K.Y.W. and N.S.K. written and reviewed the final manuscript.

Funding Open access funding provided by The Hong Kong Polytechnic University

Data Availability The data analyzed in the manuscript are publicly available from the R package RTCGA.

Declarations

Competing interests The authors declare no competing interests.

References

Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**, 669–679 (1993)

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J.: Multiple imputation by chained equations: What is it and how does it work? Int. J. Methods Psychiatr. Res. **20**, 40–49 (2011)

Bartlett, J.W., Seaman, S.R., White, I.R., Carpenter, J.R.: Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Stat. Methods Med. Res. **24**, 462–487 (2015)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B Methodol. **39**, 1–38 (1977)

Deng, Y., Chang, C., Ido, M.S., Long, Q.: Multiple imputation for general missing data patterns in the presence of high-dimensional data. Sci. Rep. **6**, 21689 (2016)

Deng, Y., Lumley, T.: Multiple imputation through XGBoost. J. Comput. Graph. Stat. **33**, 352–363 (2023)

Du, J., Boss, J., Han, P., Beesley, L.J., Kleinsasser, M., Goutman, S.A., Batterman, S., Feldman, E.L., Mukherjee, B.: Variable selection with multiply-imputed datasets: Choosing between stacked and grouped methods. J. Comput. Graph. Stat. **31**, 1063–1075 (2022)

Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons Inc, New York (2000)

Fan, J., Fan, Y., Lv, J.: High dimensional covariance matrix estimation using a factor model. J. Econom. **147**, 186–197 (2008)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)

Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**, 1–22 (2010)

Garcia, R.I., Ibrahim, J.G., Zhu, H.: Variable selection in the Cox regression model with covariates missing at random. Biometrics **66**, 97–104 (2010)

González, J., Tuerlinckx, F., De Boeck, P., Cools, R.: Numerical integration in logistic-normal models. Comput. Stat. Data Anal. **51**, 1535–1548 (2006)

Harrell, F.E., Lee, K.L., Mark, D.B.: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. **15**, 361–387 (1996)

Hastie, T., Mazumder, R., Lee, J.D., Zadeh, R.: Matrix completion and low-rank SVD via fast alternating least squares. J. Mach. Learn. Res. **16**, 3367–3402 (2015)

Herring, A.H., Ibrahim, J.G.: Likelihood-based methods for missing covariates in the Cox proportional hazards model. J. Am. Stat. Assoc. **96**, 292–302 (2001)

Hurvich, C.M., Tsai, C.-L.: Regression and time series model selection in small samples. Biometrika **76**, 297–307 (1989)

Johnson, B.A., Lin, D.Y., Zeng, D.: Penalized estimating functions and variable selection in semiparametric regression models. J. Am. Stat. Assoc. **103**, 672–680 (2008)

Kosinski, M. (2025). *RTCGA: The Cancer Genome Atlas Data Integration*. R package version 1.40.0

Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. **88**, 365–411 (2004)

Liang, L., Zhuang, Y., Yu, P.L.H.: Variable selection for high-dimensional incomplete data. Comput. Stat. Data Anal. **192**, 107877 (2024)

Liu, Q., Pierce, D.A.: A note on Gauss-Hermite quadrature. Biometrika **81**, 624–629 (1994)

Martinussen, T., Holst, K.K., Scheike, T.H.: Cox regression with missing covariate data using a modified partial likelihood method. Lifetime Data Anal. **22**, 570–588 (2016)

Papageorgiou, G., Mauff, K., Tomer, A., Rizopoulos, D.: An overview of joint modeling of time-to-event and longitudinal outcomes. Annu. Rev. Stat. Appl. **6**, 223–240 (2019)

Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons Inc, New York (2004)

Sabbe, N., Thas, O., Ottoy, J.-P.: EMLasso: Logistic lasso with missing data. Stat. Med. **32**, 3143–3157 (2013)

Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for Cox's proportional hazards model via coordinate descent. J. Stat. Softw. **39**, 1–13 (2011)

Stekhoven, D.J., Bühlmann, P.: MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics **28**, 112–118 (2012)

Sugiura, N.: Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Stat. - Theor. M. 7, 13–26 (1978)

Sun, J., Herazo-Maya, J.D., Molyneaux, P.L., Maher, T.M., Kaminski, N., Zhao, H.: Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. Biometrics **75**, 69–77 (2019)

Tay, J.K., Narasimhan, B., Hastie, T.: Elastic net regularization paths for all generalized linear models. J. Stat. Softw. **106**, 1–31 (2023)

Thiessen, D.L., Zhao, Y., Tu, D.: Unified estimation for Cox regression model with nonmonotone missing at random covariates. Stat. Med. **41**, 4781–4790 (2022)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B Stat. Methodol. **58**, 267–288 (1996)

van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in R. J. Stat. Softw. **45**, 1–67 (2011)

Wang, H., Leng, C.: Unified lasso estimation by least squares approximation. J. Am. Stat. Assoc. **102**, 1039–1048 (2007)

Wang, W., Lu, S.-E., Cheng, J.Q., Xie, M., Kostis, J.B.: Multivariate survival analysis in big data: A divide-and-combine approach. Biometrics **78**, 852–866 (2022)

Wang, Y., Hong, C., Palmer, N., Di, Q., Schwartz, J., Kohane, I., Cai, T.: A fast divide-and-conquer sparse Cox regression. Biostatistics **22**, 381–401 (2021)

Warton, D.I.: Penalized normal likelihood and ridge regularization of correlation and covariance matrices. J. Am. Stat. Assoc. **103**, 340–349 (2008)

Wilkie, R., Parmar, S.S., Blagojevic-Bucknall, M., Smith, D., Thomas, M.J., Seale, B.J., Mansell, G., Peat, G.: Reasons why osteoarthritis predicts mortality: Path analysis within a Cox proportional hazards model. RMD Open **5**, e001048 (2019)

Wolfson, J.: EEBoost: A general method for prediction and variable selection based on estimating equations. J. Am. Stat. Assoc. **106**, 296–305 (2011)

Wong, K.Y., Zeng, D., Lin, D.Y.: Semiparametric latent-class models for multivariate longitudinal and survival data. Ann. Stat. **50**, 487–510 (2022)

Wood, A.M., White, I.R., Royston, P.: How should variable selection be performed with multiply imputed data? Stat. Med. **27**, 3227–3246 (2008)

Wooldridge, J.M.: Inverse probability weighted M-estimators for sample selection, attrition, and stratification. Port. Econ. J. **1**, 117–139 (2002)

Wooldridge, J.M.: Inverse probability weighted estimation for general missing data problems. J. Econom. **141**, 1281–1301 (2007)

Zeng, D., Lin, D.Y.: Maximum likelihood estimation in semiparametric regression models with censored data. Journal of the Royal Statistical Society Series B (Statistical Methodology) **69**, 507–564 (2007)

Zhou, R., Li, H., Sun, J., Tang, N.: A new approach to estimation of the proportional hazards model based on interval-censored data with missing covariates. Lifetime Data Anal. **28**, 335–355 (2022)

Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)