SURVIVAL ANALYSIS WITH INCOMPLETE OBSERVATION OR INSUFFICIENT FOLLOW-UP

KWOK NGOK SANG

MPhil

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Applied Mathematics

Survival analysis with incomplete observation or insufficient follow-up

Kwok Ngok Sang

A thesis submitted in partial fulfilment of the requirements for the degree of
Master of Philosophy

May 2025

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Kwok Ngok Sang

Abstract

In the first part of the thesis, we address the issue of missing covariates in survival analysis. One approach to handle missing data is the likelihood-based approach, where the incomplete variables are modeled. Although likelihood-based approaches are theoretically appealing, they often become computationally inefficient or even infeasible when dealing with a large number of missing variables. We consider the Cox regression model with Gaussian covariates that are missing at random. We develop an expectation-maximization (EM) algorithm for nonparametric maximum likelihood estimation, utilizing a transformation technique in the E-step that involves only one-dimensional integration. This innovation enhances the scalability of our methods with respect to the dimensionality of the missing variables. We demonstrate the feasibility and advantages of the proposed methods over existing methods via large-scale simulation studies and apply the proposed methods to a cancer genomic study.

In the second part of the thesis, we address the issue of insufficient follow-up in survival analysis with a cure fraction. For some events of interest, not all subjects are susceptible; these non-susceptible subjects are referred to as being cured. When follow-up time is insufficient, the survival model may not be identifiable, and, in particular, the cure probability may not be consistently estimated. We focus on the promotion time cure model and develop a two-step approach for estimation. In the first step, we perform nonparametric maximum likelihood estimation. In the second step, we utilize extreme value theory and the tail behavior of the estimated hazard function to extrapolate the cure probability. We demonstrate the feasibility and advantages of the proposed methods using large-scale simulation studies.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Wong Kin Yau, for his unwavering guidance, insightful feedback, and constant encouragement throughout this research journey. His expertise and patience were invaluable in shaping this thesis. I extend my sincere thanks to Dr. Lee Chun Yin, for his constructive critiques and suggestions that strengthened this work. Finally, my deepest gratitude goes to my girlfriend, Kwok Yan Kiu, whose love, patience, and unwavering belief in me carried me through every challenge.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Survival analysis

Survival analysis is a statistical methodology for analyzing time-to-event data, where the outcome variable represents the time until the occurrence of a specific event. The primary objective of survival analysis is to characterize the distribution of the event time. Survival analysis has diverse applications across disciplines, including economics (duration of unemployment), engineering (time to mechanical failure), and medical research (time to tumor progression). One common issue in survival analysis is the presence of right censoring, where the event of interest is not observed but is known to occur after the end of the follow-up period.

Often, we can observe a set of explanatory variables, termed covariates, alongside the time-to-event data. Our interest lies in understanding the association between these covariates and the event time. For instance, we might investigate the effectiveness of a medical treatment on the survival of patients with a particular disease, or identify significant factors from a large pool of genetic covariates in a cancer genomic study. Widely used survival models include random survival forests, accelerated failure time models, and the Cox proportional hazards model. This thesis specifically focuses on the

Cox model, whose mathematical formulation is introduced below.

Let T be an event time of interest and $\boldsymbol{X}$ be a p-vector of covariates. The distribution of T given $\boldsymbol{X}$ is characterized by the hazard function $\lambda(t \mid \boldsymbol{X})$, which describes the instantaneous risk of experiencing the event at time t, conditional on survival up to that time. Under the proportional hazards assumption, the hazard function is given by

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

where $\lambda(\cdot)$ is a nonparametric baseline hazard function, and $\boldsymbol{\beta}$ is a p-vector of regression coefficients. Let $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ denote the cumulative baseline hazard. The corresponding survival function is:

$$P(T > t \mid \boldsymbol{X}) = \exp\left\{-\Lambda(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X})\right\}.$$

While the Cox model has been a powerful tool in the context of survival analysis, its application faces challenges such as missing data and insufficient follow-up. Missing data may arise from non-response in surveys or improper data collection. On the other hand, insufficient follow-up may result from early termination of the study and loss to follow-up. These issues hinder the estimation procedure and possibly introduce bias into the estimator. We will discuss these challenges separately in the following subsections.

1.2 Missing data

Statistical models, such as regression, decision trees, and neural networks are invaluable tools for analyzing real-world problems. However, these models typically require complete data, a requirement that is often impractical due to the prevalence of missing values in a large number of real-world datasets. Missing data presents a major challenge in data analysis. Traditional methods for handling missing data include inverse probability weighting, multiple imputation, and maximum likelihood estimation (MLE). While these methods have proven effective in empirical research, they often become computationally intractable as data dimensionality increases. Therefore, there is a pressing need for scalable methods that can effectively handle missing data.

Missing data is prevalent in biomedical studies. For example, there are multiple cancer datasets available in The Cancer Genome Atlas (TCGA) program, where the data are collected from different hospitals in the US. These datasets contain genomic and clinical data for cancer patients, where protein expressions were not measured or partially observed for a large fraction of subjects. Another example is the MIMIC3 database (Johnson et al., 2016), which collected data for patients who stayed in critical care units. Laboratory test data from this database contain missing values because certain tests are possibly skipped if the clinical provider believes those tests were not helpful for diagnosis or treatment of patients. Handling missing values is unavoidable when analyzing clinical data.

One naive approach for missing data is complete-case analysis, which discards subjects with missing values and analyze the complete-data. Although complete-case analysis is easy to implement and gives valid estimates in case of missing completely at random (MCAR), there are noticeable drawbacks that we should not overlook. For instance, complete-case analysis yields a biased estimator when the data are missing at random (MAR). An improved version of complete-case analysis is inverse probability weighting (Wooldridge, 2002; Wooldridge, 2007), which assigns weights to different subjects in the complete data, according to their probability of complete observation. These two methods are both inefficient because they do not incorporate partially observed subjects in the estimation procedure. One way to address this is to employ a unified framework, which uses partially observed subjects to update the inverse-probability weighted estimator (Thiessen et al., 2022).

Another straightforward approach is the missing indicator method (Cohen and Cohen, 1975). In this method, missing entries are imputed with some constant values, and an indicator vector, denoting whether the covariates were missing, is incorporated in the regression model. While this method is simple to implement and retains all available data, it has been criticized for introducing substantial bias (Jones, 1996). Despite these concerns, recent studies by Zhao and Ding (2024) and Zhao et al. (2024) demonstrate that, in randomized experiments with incomplete covariates, the missing indicator method is asymptotically more efficient than complete-case analysis. This suggests that the missing indicator method may offer practical advantages under certain condition.

Multiple imputation fills in missing values by simulating plausible numbers derived from distributions of the missing data conditioned on the observed data. After creating several completed datasets, we apply standard statistical analysis on each of them and eventually pool the estimates to obtain a final estimate based upon Rubin's rule (Rubin, 2004). There are a large variety of methods to generate imputations, one of the most popular methods is multiple imputation by chained equations (MICE) (van Buuren and Groothuis-Oudshoorn, 2011;Azur et al., 2011). MICE iteratively regresses each incomplete variable on the remaining imputed dataset and draws new imputations based on the regression models. Deng and Lumley (2024) considered using XGBoost, a supervised machine learning model, to generate imputations in a scalable manner. Imputation is intuitive, but it may be challenging to make inference on the resulting estimator.

For MLE, we maximize the likelihood, which includes both the outcome model and the covariates model. Generally, the observed-data likelihood does not have an analytical form, which causes difficulties in computation. One technique to facilitate computation is through the use of the expectation-maximization (EM) algorithm (Dempster et al., 1977), which maximizes the observed-data likelihood by iteratively performing E-step and M-step. Herring and Ibrahim (2001) developed an EM algorithm for the Cox model with partially observed covariates, which could be categorical or continuous. Zhou et al. (2022) implemented the EM algorithm with two-stage data augmentation for the Cox model with interval-censored survival time. Although MLE is efficient and has many theoretical guarantees, it could be computationally infeasible when both the dimension of covariates (p) and the number of missing entries grow.

In addition to missing data, high-dimensional covariates present another challenge. When the dimension of 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 , difficulty 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 in the presence of missing entries is highly challenging, and there is 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 inverse probability weighting, Johnson et al. (2008) and Wolfson (2011) incorporated a penalty term to 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. Additionally, they may require computationally intensive tuning.

1.3 Cure models

Conventional statistical methods in survival analysis typically assume that all subjects would eventually experience the event of interest, such as tumor progression or death in cancer studies. However, in many applications, it is plausible to postulate the presence of "cured subjects," who are long-term survivors not susceptible to the event even after a long-term follow-up (Peng and Yu, 2021). For example, Othus et al. (2012) showed a substantial proportion of subjects with multiple myeloma who remained recurrence-free for more than 15 years after receiving total therapy of tandem autotransplant. Similarly, in a large-scale European study of childhood cancer, Botta et al. (2022) reported that 81% of patients (in a sample of 135,847 patients) survived beyond five years since diagnosis. Another large-scale cancer study in France identified a noticeable cure fraction of skin melanoma, thyroid, and testis cancer patients (Romain et al., 2019).

When analyzing time-to-event data with a cure, we are often interested in the proportion of cure and how covariates — including personal characteristics, cancer type, and treatment received — affect the cure probability. A widely used framework for this purpose is the mixture cure model, which assumes that the population consists of two distinct groups: cured individuals and susceptible individuals. This model composes of two components: an incidence model that captures the cure probability and a latency model that captures the time to event for the susceptible subjects. Logistic or semiparametric single-index models have been considered for the incidence model (Farewell, 1982; Amico et al., 2019; Li et al., 2020; Lee et al., 2024), whereas parametric models, semiparametric proportional hazards model, transformation model, accelerated failure time model, and quantile regression model have been considered for the latency model (Sy and Taylor, 2000; Peng and Dear, 2000; Lu and Ying, 2004; Zhang and Peng, 2007; Lu, 2010; Wu and Yin, 2013, 2017a,b).

An alternative approach, the promotion time cure model, is motivated by the biological mechanisms underlying cancer progression (Yakovlev and Tsodikov, 1996; Tsodikov, 1998; Chen et al., 1999). This model postulates that the event time is governed by a random number of latent carcinogenic processes, each corresponding to a potentially active malignant cell. Cured individuals have no such carcinogenic cells, whereas uncured individuals harbor a random number of them and would experience the event when any of these cells become active and progress to malignancy. In particular, let $\boldsymbol{X}$ be a p-vector of covariates, N be the number of potentially active carcinogenic cells, and $Z_1, Z_2, \ldots$ be the times to activation of the carcinogenic cells. Assume that N follows a Poisson$(e^{\alpha+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}})$ distribution conditional on $\boldsymbol{X}$, where α and $\boldsymbol{\beta}$ are regression coefficients. Assume that Z_j's are independent and identically distributed with cumulative distribution function (CDF) F with $F(0) = 0$, which is typically assumed to be fully nonparametric. The time-to-event T is defined to be the minimum of these event times, $T = \min(Z_1, \ldots, Z_N)$, if $N > 0$, and $T = \infty$ if $N = 0$. The survival function of T is

$$\begin{aligned} P(T > t \mid \boldsymbol{X}) &= P(N = 0 \mid \boldsymbol{X}) + \sum_{j=1}^{\infty} P(Z_1 > t, \ldots, Z_j > t \mid N = j) P(N = j \mid \boldsymbol{X}) \\ &= \exp\{-\exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X})\} \left\{ 1 + \sum_{j=1}^{\infty} [1 - F(t)]^j \frac{\exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X})^j}{j!} \right\} \\ &= \exp\{-\exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}) F(t)\}. \end{aligned}$$

The probability of cure is given by

$$P(T = \infty \mid \boldsymbol{X}) = P(N = 0 \mid \boldsymbol{X}) = \exp\{-\exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X})\}.$$

The promotion time cure model offers several advantages over the mixture cure model. First, it aligns more closely with the biological mechanisms of cancer recurrence, whereas the two-part structure of the mixture cure model may be less biologically plausible. Second, the promotion time cure model retains a proportional hazards structure, which facilitates interpretation of the covariate effects. In contrast, in mixture cure models, especially those with the same covariates in both the incidence and latency models, the effect of a covariate on the event time is difficult to interpret.

Computation of the nonparametric maximum likelihood estimator (NPMLE) of $(\alpha, \boldsymbol{\beta}, F)$ involves solving a constrained optimization problem, for which various computational methods have been developed. Zeng et al. (2006) employed a Newton–Raphson method. Ma and Yin (2008) introduced a backfitting algorithm. Portier et al. (2017) utilized a profile likelihood approach. However, these methods are computationally inefficient due to the additional procedures required for handling a Lagrange multiplier. In contrast, Beyhum et al. (2022) noted the connection between the promotion time cure model and the Cox proportional hazards model and suggested that the NPMLE of $(\alpha, \boldsymbol{\beta}, F)$ can be derived from the NPMLE of the Cox model through a simple transformation. As a result, the NPMLE of the promotion time cure model can be computed as easily as that for the Cox model.

Despite its advantages, the promotion time cure model suffers an identifiability problem under insufficient follow-up. We say that the follow-up is insufficient if the right endpoint of the support of the censoring time distribution, denoted as τ_c, is smaller than that of the survival time for susceptible subjects (Maller and Zhou, 1994). Let $(\alpha_0, \boldsymbol{\beta}_0, F_0)$ be the true parameter values, where $F_0(\tau_c) < 1$. For any $(\alpha^*, \boldsymbol{\beta}^*, F^*) = (\alpha_0 + \log k, \boldsymbol{\beta}_0, k^{-1}F_0)$ such that $k \geq F_0(\tau_c)$, F^* is a proper CDF over $[0, \tau_c]$, and

$$\exp\{-\exp(\alpha_0 + \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})F_0(t)\} = \exp\{-\exp(\alpha^* + \boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})F^*(t)\} \text{ for all } t \in [0, \tau_c].$$

Therefore, even under an infinite sample and the survival function $P(T > t \mid \boldsymbol{X})$ is virtually observed over $t \in [0, \tau_c]$, the model parameters, and in particular the cure probability, cannot be consistently estimated from a likelihood approach.

A popular ad hoc solution for the identifiability issue is to impose a zero-tail constraint

(Sy and Taylor, 2000), which assumes that all subjects censored beyond the largest observed event time are cured. Although this restriction ensures the identifiability of the promotion time cure model, it introduces substantial bias in the estimation of α and F when $F(\tau_c)$ is far below 1. This bias arises from the potential misclassification of some susceptible subjects censored after the last observed event time as cured.

To mitigate the bias induced from the zero-tail constraint, we propose leveraging extreme value theory to refine the estimation of the cure probability. Extreme value theory provides a principled framework for extrapolating tail probabilities beyond observed censoring limits, thereby improving prediction accuracy. In the absence of covariates, Escobar-Bach and Van Keilegom (2019) and Escobar-Bach et al. (2022) employed this technique to extrapolate the Kaplan–Meier estimates and demonstrated a successful reduction of bias in cure probability estimation. Later, this technique is extended to the mixture cure model with covariates (Escobar-Bach and Van Keilegom, 2023).

In Chapter 3, we introduce a novel extrapolation approach for the promotion time cure model based on extreme value theory. Under some general regularity conditions, the tail behavior of a CDF can be characterized by one of three parametric forms, each corresponding to a domain of attraction of extreme value distribution. In the proposed approach, we first estimate F using NPMLE and then fit one of the parametric forms to the tail of the estimated F. The fitted tail structure can be used to estimate the cure probability. In contrast to Escobar-Bach and Van Keilegom (2019), Escobar-Bach et al. (2022), and Escobar-Bach and Van Keilegom (2023), the proposed method uses a segment of the NPMLE of F to fit the parametric tail structure, instead of using a few selected points of the function. Also, we develop an inferential procedure for the cure probability in addition to a point estimator.

1.4 Organization of thesis

This thesis is organized as follows. In Chapter 2, we consider the Cox proportional hazards model with covariates that are missing at random. We develop an EM algorithm for the NPMLE, employing a transformation technique in the E-step so that it involves only one-dimensional integration. In Chapter 3, we consider the promotion time cure model under insufficient follow-up. We develop a novel extrapolation method based on extreme value theory. In Chapter 4, we outline potential directions for future research, including possible extensions of the methodologies presented.

Chapter 2

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

2.1 Overview

In Chapter 2, we consider the Cox proportional hazards model with incomplete data. Section 2.2 describes the proposed model and formulates the EM algorithm for both the unpenalized and penalized cases. Section 2.3 reports large-scale simulation studies results and compares the performance of the proposed methods with existing methods. Section 2.4 demonstrates the feasibility and advantages of the proposed method in a cancer genomics study. Section 2.5 provides some concluding remarks and possible extensions.

2.2 Methods

2.2.1 Model and likelihood

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. Suppose that T may be subject to right-censoring. Let C be the censoring time, $Y = \min(T, C)$, and $\Delta = I(T \leq C)$.

We allow components of $\boldsymbol{X}$ to be missing. Let $\boldsymbol{R} \equiv (R_1, \dots, R_p)^{\mathrm{T}}$ denote a vector of missing indicators, where $R_j = 1$ if X_j is missing and $R_j = 0$ otherwise. Assume missing at random, such that $\boldsymbol{R}$ and $\boldsymbol{X}$ are independent given $\{X_j : P(R_j = 0) = 1\}$, Y, and Δ. Also, assume that T and C are independent given $\{X_j : P(R_j = 0) = 1\}$. For a sample of size n, the observed data consist of $\mathcal{O}_i \equiv \{Y_i, \Delta_i, \boldsymbol{R}_i, \boldsymbol{X}_{i,-\boldsymbol{R}_i}\}$ for $i = 1, \dots, n$, where $\boldsymbol{X}_{i,-\boldsymbol{R}_i}$ denote the subvector of $\boldsymbol{X}_i$ consisting of components that correspond to $R_{ij} = 0$. 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 is

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

where $\boldsymbol{X}_{i,\boldsymbol{R}_i}$ denote the subvector of $\boldsymbol{X}_i$ consisting of the components that correspond to $R_{ij} = 1$.

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 $t_1, \dots, t_m$ and let the corresponding jump sizes be $\lambda_1, \dots, \lambda_m$. 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.

2.2.2 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 $\boldsymbol{X}_{i,\boldsymbol{R}_i}$ treated as missing data for $i = 1, \ldots, n$. The complete-data log-likelihood is

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

where $j(i)$ is such that $Y_{j(i)} = t_j$ for $i \in \{i = 1, \ldots, n : \Delta_i = 1\}$. In the E-step, we evaluate the conditional expectation of $\log L_{\text{com}}(\boldsymbol{\theta})$ 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

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i) \tag{2.1}$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i \boldsymbol{X}_i^{\text{T}}) - \left(\boldsymbol{\mu}^{(k+1)}\right) \left(\boldsymbol{\mu}^{(k+1)}\right)^{\text{T}}, \tag{2.2}$$

where $\widehat{\mathrm{E}}^{(k)}$ denote 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"

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

Note that

$$
\begin{aligned}
\frac{\partial Q^{(k)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \Delta_i \left\{ \widehat{\mathrm{E}}^{(k)}(\boldsymbol{X}_i) - \frac{\sum_{j:Y_j \geq Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}} \boldsymbol{X}_j)}{\sum_{j:Y_j \geq Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}})} \right\} \\
\frac{\partial^2 Q^{(k)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} &= -\sum_{i=1}^{n} \Delta_i \left[\frac{\sum_{j:Y_j \geq Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}} \boldsymbol{X}_j \boldsymbol{X}_j^{\mathrm{T}})}{\sum_{j:Y_j \geq Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}})} - \left\{ \frac{\sum_{j:Y_j \geq Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}} \boldsymbol{X}_j)}{\sum_{j:Y_j \geq Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}})} \right\}^{\otimes 2} \right].
\end{aligned}
$$

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

$$
\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left(\left. \frac{\partial^2 Q^{(k)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} \right)^{-1} \left(\left. \frac{\partial Q^{(k)}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} \right), \tag{2.3}
$$

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

$$
\lambda_{j(i)}^{(k+1)} = \frac{1}{\sum_{j:Y_j \geq Y_i} \widehat{\mathrm{E}}^{(k)}(e^{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{\beta}^{(k+1)}})} \tag{2.4}
$$

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 $\boldsymbol{X}_{i,\boldsymbol{R}_i}$ 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. Let $\boldsymbol{\beta}_{\boldsymbol{R}_i}$ and $\boldsymbol{\beta}_{-\boldsymbol{R}_i}$ denote the subvector of $\boldsymbol{\beta}$ consisting of components that correspond to $R_{ij} = 1$ and 0 respectively. The same method is used to denote subvectors of $\boldsymbol{\mu}$. Let $\boldsymbol{\Sigma}_{\mathcal{A}_{i1},\mathcal{A}_{i2}}$ denote the submatrix of $\boldsymbol{\Sigma}$ with rows indexed by $\mathcal{A}_{i1}$ and columns indexed by $\mathcal{A}_{i2}$, and $\mathcal{A}_{ij}$ is either $\boldsymbol{R}_i$ or $-\boldsymbol{R}_i$ for $j = 1, 2$. The expectations that need to be computed in the E-step are in one of the following forms:

$$
\mathrm{E}\big\{ \exp(\boldsymbol{X}_{i,\boldsymbol{R}_i}^{\mathrm{T}} \boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)}) | \mathcal{O}_i \big\} \tag{2.5}
$$

$$\mathrm{E}\{g(\boldsymbol{X}_{i,\boldsymbol{R}_i}^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)})\boldsymbol{X}_{i,\boldsymbol{R}_i}|\mathcal{O}_i\} \tag{2.6}$$

$$\mathrm{E}\{g(\boldsymbol{X}_{i,\boldsymbol{R}_i}^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)})\boldsymbol{X}_{i,\boldsymbol{R}_i}\boldsymbol{X}_{i,\boldsymbol{R}_i}^{\mathrm{T}}|\mathcal{O}_i\} \tag{2.7}$$

$$\mathrm{E}\big\{\exp(\boldsymbol{X}_{i,\boldsymbol{R}_i}^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k+1)})|\mathcal{O}_i\big\}, \tag{2.8}$$

where g is either the exponential function or the constant function $g(\cdot) = 1$. Note that the expectations are evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$.

First, if $\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)} = \mathbf{0}$, then the conditional distribution of $\boldsymbol{X}_{i,\boldsymbol{R}_i}$ given $\mathcal{O}_i$ is a multivariate normal distribution that does not depend on (Y_i, Δ_i). The expectations (2.5)–(2.8) have simple closed-form expressions.

For $\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)} \neq \mathbf{0}$, we define an orthogonal matrix $\boldsymbol{\Psi}_i$ with the first row being $(\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)})^{\mathrm{T}}/\|\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)}\|$ and let $\widetilde{\boldsymbol{X}}_i = \boldsymbol{\Psi}_i\boldsymbol{X}_{i,\boldsymbol{R}_i}$, where $\|\cdot\|$ denote the L_2-norm. Note that the first component of $\widetilde{\boldsymbol{X}}_i$ is $\widetilde{X}_{i1} = \boldsymbol{X}_{i,\boldsymbol{R}_i}^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)}/\|\boldsymbol{\beta}_{\boldsymbol{R}_i}^{(k)}\|$. Let $\boldsymbol{\eta}_i$, and $\boldsymbol{\nu}_i$ denote the mean and variance of $\widetilde{\boldsymbol{X}}_i$ given $\boldsymbol{X}_{-\boldsymbol{R}_i}$, where

$$\begin{aligned}
\boldsymbol{\eta}_i &= \boldsymbol{\Psi}_i\boldsymbol{\mu}_{\boldsymbol{R}_i} + \boldsymbol{\Psi}_i\boldsymbol{\Sigma}_{\boldsymbol{R}_i,-\boldsymbol{R}_i}\boldsymbol{\Sigma}_{-\boldsymbol{R}_i,-\boldsymbol{R}_i}^{-1}(\boldsymbol{X}_{i,-\boldsymbol{R}_i} - \boldsymbol{\mu}_{-\boldsymbol{R}_i})\\
\boldsymbol{\nu}_i &= \boldsymbol{\Psi}_i\boldsymbol{\Sigma}_{\boldsymbol{R}_i,\boldsymbol{R}_i}\boldsymbol{\Psi}_i^{\mathrm{T}} - \boldsymbol{\Psi}_i\boldsymbol{\Sigma}_{\boldsymbol{R}_i,-\boldsymbol{R}_i}\boldsymbol{\Sigma}_{-\boldsymbol{R}_i,-\boldsymbol{R}_i}^{-1}\boldsymbol{\Sigma}_{-\boldsymbol{R}_i,\boldsymbol{R}_i}\boldsymbol{\Psi}_i^{\mathrm{T}}.
\end{aligned}$$

Let $\widetilde{\boldsymbol{X}}_{i,-1}$ denote the subvector of $\widetilde{\boldsymbol{X}}_i$ consisting of all but the first component. Although the conditional distribution of $\widetilde{\boldsymbol{X}}_i$ given the observed data does not have a simple form, $\widetilde{\boldsymbol{X}}_{i,-1}$ given the observed data and $\widetilde{X}_{i1}$ (at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$) follows the multivariate normal distribution:

$$\begin{aligned}
\widetilde{\boldsymbol{X}}_{i,-1} \mid (Y_i, \Delta_i, \widetilde{X}_{i1}, \boldsymbol{X}_{i,-\boldsymbol{R}_i}) \sim \mathrm{N}\left((\boldsymbol{\eta}_i)_{-1} + (\boldsymbol{\nu}_i)_{-1,1}\frac{\widetilde{X}_{i1} - (\boldsymbol{\eta}_i)_1}{(\boldsymbol{\nu}_i)_{1,1}}, (\boldsymbol{\nu}_i)_{-1,-1} - \frac{(\boldsymbol{\nu}_i)_{-1,1}^{\otimes 2}}{(\boldsymbol{\nu}_i)_{1,1}}\right)\\
\equiv \mathrm{N}(\boldsymbol{m}_i(\widetilde{X}_{i1}), \boldsymbol{V}_i),
\end{aligned}$$

where $(\boldsymbol{\nu}_i)_{1,1}$ is the upper left element of $\boldsymbol{\nu}_i$, $(\boldsymbol{\nu}_i)_{-1,1}$ is the first column of $\boldsymbol{\nu}_i$ with the first component removed, and $(\boldsymbol{\nu}_i)_{-1,-1}$ is the lower right submatrix of $\boldsymbol{\nu}_i$, with the first row and column of $\boldsymbol{\nu}_i$ removed. The conditional density of $\widetilde{X}_{i1}$ given $\mathcal{O}_i$ is proportional to

$$f(\widetilde{x}_{i1};\mathcal{O}_i) \equiv \exp\left\{\Delta_i\|\boldsymbol{\beta}^{(k)}_{\boldsymbol{R}_i}\|\widetilde{x}_{i1} - \Lambda^{(k)}(Y_i)e^{\|\boldsymbol{\beta}^{(k)}_{\boldsymbol{R}_i}\|\widetilde{x}_{i1}+\boldsymbol{X}^{\mathrm{T}}_{i,-\boldsymbol{R}_i}\boldsymbol{\beta}^{(k)}_{-\boldsymbol{R}_i}} - \frac{1}{2(\boldsymbol{\nu}_i)_{1,1}}(\widetilde{x}_{i1}-(\boldsymbol{\eta}_i)_1)^2\right\}.$$

Therefore, conditional expectations of functions of $\boldsymbol{X}_{i,\boldsymbol{R}_i} \equiv \boldsymbol{\Psi}_i^{\mathrm{T}}\widetilde{\boldsymbol{X}}_i$ can be computed by first further conditioning on $\widetilde{X}_{i1}$, where the conditional expectations have closed-form expressions, and then taking the expectation over $\widetilde{X}_{i1}$, which can be performed by numerical integration.

Specifically, the expectation (2.5) is equal to $\mathrm{E}\big\{\exp(\|\boldsymbol{\beta}^{(k)}_{\boldsymbol{R}_i}\|\widetilde{X}_{i1})|\mathcal{O}_i\big\}$. The expectation (2.6) is equal to

$$\boldsymbol{\Psi}_i^{\mathrm{T}}\mathrm{E}\big\{g(\|\boldsymbol{\beta}^{(k)}_{\boldsymbol{R}_i}\|\widetilde{X}_{i1})\widetilde{\boldsymbol{X}}_i|\mathcal{O}_i\big\} = \boldsymbol{\Psi}_i^{\mathrm{T}}\mathrm{E}\left\{g(\|\boldsymbol{\beta}^{(k)}_{\boldsymbol{R}_i}\|\widetilde{X}_{i1})\begin{pmatrix}\widetilde{X}_{i1}\\ \boldsymbol{m}_i(\widetilde{X}_{i1})\end{pmatrix}\middle|\mathcal{O}_i\right\}.$$

The expectation (2.7) is equal to

$$\begin{aligned}&\boldsymbol{\Psi}_i^{\mathrm{T}}\mathrm{E}\big\{g(\|\boldsymbol{\beta}^{(k)}_{\boldsymbol{R}_i}\|\widetilde{X}_{i1})\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}|\mathcal{O}_i\big\}\boldsymbol{\Psi}_i\\ &=\boldsymbol{\Psi}_i^{\mathrm{T}}\mathrm{E}\left\{g(\|\boldsymbol{\beta}^{(k)}_{\boldsymbol{R}_i}\|\widetilde{X}_{i1})\begin{pmatrix}\widetilde{X}_{i1}^2 & \widetilde{X}_{i1}\boldsymbol{m}_i(\widetilde{X}_{i1})^{\mathrm{T}}\\ \widetilde{X}_{i1}\boldsymbol{m}_i(\widetilde{X}_{i1}) & \boldsymbol{V}_i+\boldsymbol{m}_i(\widetilde{X}_{i1})\boldsymbol{m}_i(\widetilde{X}_{i1})^{\mathrm{T}}\end{pmatrix}\middle|\mathcal{O}_i\right\}\boldsymbol{\Psi}_i.\end{aligned}$$

Finally, to evaluate (2.8), let

$$\phi_i(\widetilde{X}_{i1};\boldsymbol{a}) = \mathrm{E}\big\{\exp(\widetilde{\boldsymbol{X}}^{\mathrm{T}}_{i,-1}\boldsymbol{a}) \mid \widetilde{X}_{i1},\mathcal{O}_i\big\} = \exp\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{m}_i(\widetilde{X}_{i1})+\frac{1}{2}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{V}_i\boldsymbol{a}\right\}$$

for any vector $\boldsymbol{a}$ of an appropriate dimension. We can write (2.8) as

$$\mathrm{E}\big\{\exp(\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\Psi}_i\boldsymbol{\beta}^{(k+1)}_{\boldsymbol{R}_i})|\mathcal{O}_i\big\} = \mathrm{E}\big\{\exp((\boldsymbol{\Psi}_i\boldsymbol{\beta}^{(k+1)}_{\boldsymbol{R}_i})_1\widetilde{X}_{i1})\phi_i(\widetilde{X}_{i1};(\boldsymbol{\Psi}_i\boldsymbol{\beta}^{(k+1)}_{\boldsymbol{R}_i})_{-1}) \mid \mathcal{O}_i\big\}.$$

Therefore, all expectations involved in the E-step can be computed using one-dimensional numerical integrations over the conditional distribution of $\widetilde{X}_{i1}$. In particular, for any function h, we have

$$\mathrm{E}\{h(\widetilde{X}_{i1}) \mid \mathcal{O}_i\} = \frac{\int h(\widetilde{x}_{i1}) f(\widetilde{x}_{i1}; \mathcal{O}_i)\, \mathrm{d}\widetilde{x}_{i1}}{\int f(\widetilde{x}_{i1}; \mathcal{O}_i)\, \mathrm{d}\widetilde{x}_{i1}}.$$

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 (2.5), (2.6), and (2.7) for $i = 1, 2, \ldots, n$ and in turn the gradient and
    Hessian of $Q^{(k)}(\boldsymbol{\beta})$.
3 Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by (2.1) and (2.2), respectively.
4 Update $\boldsymbol{\beta}$ by (2.3).
5 Calculate (2.8) for $i = 1, 2, \ldots, n$.
6 Update $\Lambda$ by (2.4).
7 Repeat Steps 2–6 until convergence.
  Output : $(\widehat{\boldsymbol{\beta}}, \widehat{\Lambda}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$.
```

2.2.3 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 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 $n^{-1}Q^{(k)}(\boldsymbol{\beta}) - \gamma\|\boldsymbol{\beta}\|_1$. 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:

$$n^{-1}Q^{(k)}(\boldsymbol{\beta}) - \gamma\|\boldsymbol{\beta}\|_1 \approx -\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{P}^{\mathrm{T}}\boldsymbol{\beta} - \gamma\|\boldsymbol{\beta}\|_1 + \text{const}, \tag{2.9}$$

where

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

Then, to maximize the right-hand side of (2.9), 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

$$\beta_j = -\frac{S(\boldsymbol{A}_{j,-j}\boldsymbol{\beta}_{-j} + \boldsymbol{P}_j, \gamma)}{\boldsymbol{A}_{j,j}}, \tag{2.10}$$

where $S(x,\gamma) = \operatorname{sgn}(x)(|x| - \gamma)_+$. 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. Let $\boldsymbol{\beta}_\gamma$ be the LASSO estimator corresponding to γ and $\mathfrak{B}_\gamma$ denote the active set of $\boldsymbol{\beta}_\gamma$. Since the LASSO estimator is biased, we refit the model using NPMLE, with only the coefficients in the active set $\mathfrak{B}_\gamma$ allowed to be nonzero. We summarize the procedure in Algorithm 2.

```
Algorithm 2: Penalized NPMLE
   Input   : $\{\mathcal{O}_i\}_{i=1,2,\dots,n}$ and $\gamma$.
1  Initialize $(\boldsymbol{\beta}^{(0)}, \Lambda^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)})$.
2  Calculate (2.5), (2.6), and (2.7) for $i = 1, 2, \dots, n$ and in turn the gradient and
     Hessian of $Q^{(k)}(\boldsymbol{\beta})$.
3  Update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by (2.1) and (2.2), respectively.
4  Iteratively update each component of $\boldsymbol{\beta}$ through (2.10) until convergence.
5  Calculate (2.8) for $i = 1, 2, \dots, n$.
6  Update $\Lambda$ through (2.4).
7  Repeat Steps 2–6 until convergence.
8  Refit NPMLE over the active set.
   Output : $(\widehat{\boldsymbol{\beta}}_\gamma^{\text{refit}}, \widehat{\Lambda}_\gamma^{\text{refit}}, \widehat{\boldsymbol{\mu}}_\gamma^{\text{refit}}, \widehat{\boldsymbol{\Sigma}}_\gamma^{\text{refit}})$.
```

We specify a grid of tuning parameters and calculate the penalized NPMLE corresponding to each γ. To search for the optimal tuning parameter γ^*, we choose the Bayesian information criterion (BIC) as our model selection criterion:

$$\text{BIC}(\gamma) = -2 \log \widehat{L}_{\text{obs}}(\mathfrak{B}_\gamma) + \log(n)|\mathfrak{B}_\gamma|,$$

where $\widehat{L}_{\text{obs}}(\mathfrak{B}_\gamma)$ is the maximum value of observed likelihood for the active set $\mathfrak{B}_\gamma$.

2.3 Simulation studies

2.3.1 Unpenalized estimation

In this subsection, we evaluate the empirical performance of the proposed unpenalized methods and two existing methods, namely complete-case analysis and single imputation.

We set $p = 4$ and generated $\boldsymbol{X}$ from a multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} \equiv (0.5^{|i-j|})_{i,j=1,\dots,p}$. We set $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)^{\mathrm{T}}$ and $\Lambda(t) = 0.04t^{5/4}$. We set the censoring time as $\min\{C^*, 50\}$, where $C^* \sim \text{Exp}(0.03)$; the censoring rate is approximately 34%.

We considered a sample size of $n = 500$ or 1000. For each subject, either all covariates are observed or the first two covariates are missing. We considered two missing mechanisms, namely MCAR and MAR. For MCAR, subjects with missing data were randomly assigned. For MAR, we mimicked a case-cohort study, where a subcohort consisting of 30% of the whole sample was set to have observed covariates. Then, we randomly selected subjects outside the subcohort with $\Delta = 1$ to have observed covariates. If all subjects with $\Delta = 1$ were selected and the missing proportion was still higher than the desired level, then we randomly selected subjects with $\Delta = 0$ to yield the desired missing proportion. We considered missing proportions (p_M) of 20% and 40%.

We considered the proposed NPMLE, complete-case analysis under the standard Cox regression, and single imputation. For single imputation, we estimated $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using MLE based on the fully observed $\boldsymbol{X}_i$'s, imputed the missing entries by their estimated conditional means given the partially observed $\boldsymbol{X}_i$'s, and then fitted the standard Cox model on the imputed data. For the proposed method, we used bootstrap to obtain standard error estimators and confidence intervals for the regression parameters, with 500 bootstrap replicates. We considered 500 simulation replicates.

The results for MCAR are shown in Table 2.1. Both the NPMLE and complete-case analysis yield unbiased estimation, whereas single imputation yields noticeably larger bias than the other two methods, especially under a missing proportion of 40%. This is because the imputed values are necessarily not as associated with the event time as the actual values, so the estimators are biased towards zero. Under all settings, single imputation yields the smallest standard errors, followed by NPMLE. This is because single imputation trades some bias for efficiency, and the NPMLE uses more subjects than complete-case analysis. For NPMLE, the standard errors of $\widehat{\beta}_3$ and $\widehat{\beta}_4$ tend to be smaller than those of $\widehat{\beta}_1$ and $\widehat{\beta}_2$, because X_3 and X_4 have more observations than X_1 and X_2 and thus have a larger "effective sample size."

In Figure 2.1, we present the average value of Λ over the replicates for different methods under MCAR. In this case, both the NPMLE and complete-case analysis are

		NPMLE				Complete Case		Single Imputation	
Setting	Parameter	Bias	SE	SEE	CP	Bias	SE	Bias	SE
$n = 500$	β_1	-0.0022	0.0771	0.0764	0.93	-0.0017	0.0782	-0.0206	0.0752
$p_M = 20\%$	β_2	0.0059	0.0856	0.0854	0.93	0.0064	0.0857	-0.0130	0.0816
	β_3	0.0061	0.0772	0.0801	0.94	0.0051	0.0824	-0.0303	0.0772
	β_4	0.0051	0.0701	0.0724	0.95	0.0062	0.0752	-0.0241	0.0698
$n = 500$	β_1	-0.0009	0.0901	0.0886	0.93	0.0014	0.0914	-0.0356	0.0843
$p_M = 40\%$	β_2	0.0061	0.0965	0.0995	0.94	0.0076	0.0982	-0.0290	0.0896
	β_3	0.0082	0.0825	0.0859	0.95	0.0086	0.0949	-0.0586	0.0836
	β_4	0.0049	0.0773	0.0776	0.93	0.0078	0.0860	-0.0486	0.0750
$n = 1000$	β_1	0.0055	0.0547	0.0534	0.94	0.0057	0.0553	-0.0148	0.0532
$p_M = 20\%$	β_2	-0.0006	0.0587	0.0592	0.95	-0.0005	0.0585	-0.0210	0.0551
	β_3	0.0066	0.0529	0.0558	0.95	0.0075	0.0580	-0.0305	0.0525
	β_4	0.0007	0.0514	0.0502	0.92	0.0007	0.0548	-0.0287	0.0513
$n = 1000$	β_1	0.0056	0.0629	0.0615	0.93	0.0062	0.0636	-0.0304	0.0588
$p_M = 40\%$	β_2	0.0000	0.0689	0.0681	0.94	0.0007	0.0695	-0.0361	0.0633
	β_3	0.0081	0.0554	0.0596	0.95	0.0092	0.0660	-0.0594	0.0548
	β_4	0.0005	0.0547	0.0536	0.92	0.0006	0.0625	-0.0530	0.0528

Note: "Bias" is the empirical bias; "SE" is the empirical standard error; "SEE" is the average standard error estimate; "CP" is the empirical coverage probability of a 95% confidence interval.

Table 2.1: Results for unpenalized estimators of $\boldsymbol{\beta}$ under MCAR

unbiased for Λ, whereas single imputation yields a biased estimator.

The simulation results for MAR are shown in Table 2.2. Under MAR, the NPMLE is unbiased, whereas complete-case analysis and single imputation are biased. Similar to the MCAR setting, single imputation yields the smallest standard errors overall. For the NPMLE, the standard errors of $\widehat{\beta}_3$ and $\widehat{\beta}_4$ tend to be smaller than those of $\widehat{\beta}_1$ and $\widehat{\beta}_2$. Note that the NPMLE yields the smallest mean squared error among all three methods in all settings, under MCAR and MAR.

In Figure 2.2, we present the average value of Λ over the replicates for different methods under MAR. The NPMLE yields unbiased estimation, whereas both complete-case analysis and single imputation are biased.

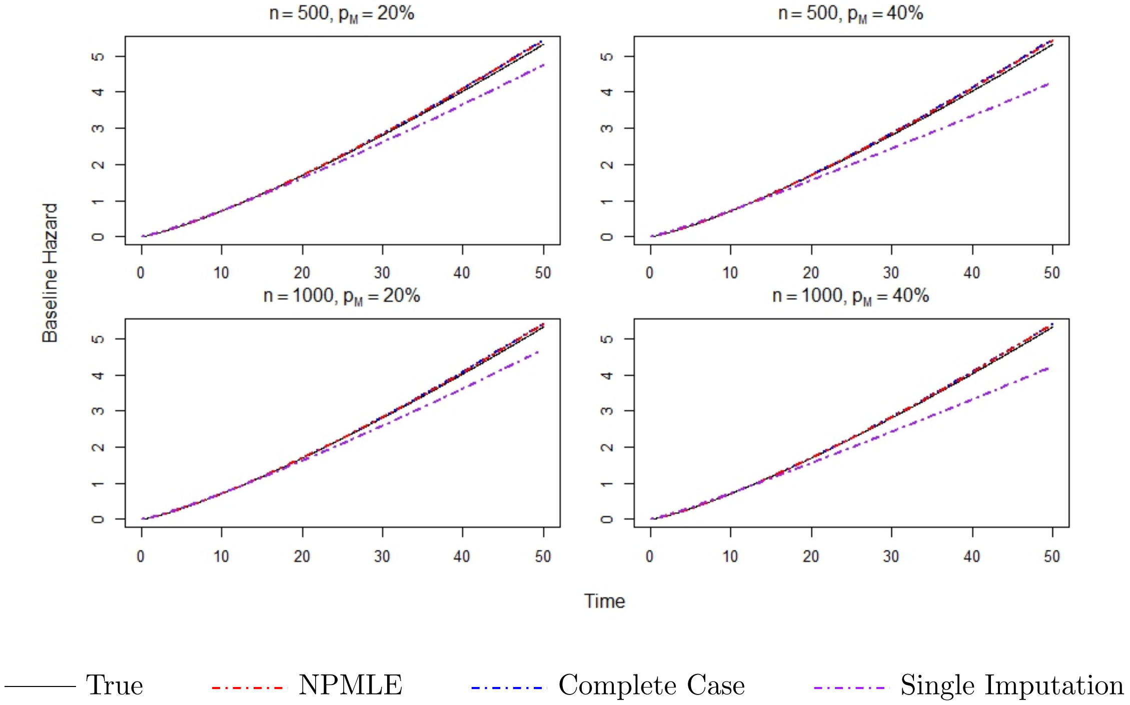

Figure 2.1: Results for unpenalized estimators of Λ under MCAR.

2.3.2 Penalized estimation

In this subsection, we compare the performance of penalized methods. We considered a sample size of $n = 500$ or 1000 and a number of covariates of $p = 100$. We draw the covariates from the multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\Sigma}_1 = (0.2^{|i-j|})_{i,j=1,\ldots,50}$ and $\boldsymbol{\Sigma}_2 = (0.5^{|i-j|})_{i,j=1,\ldots,50}$. For the survival model, we set $\boldsymbol{\beta} = (\underbrace{0.25, \ldots, 0.25}_{4}, \underbrace{0, \ldots, 0}_{92}, \underbrace{0.25, \ldots, 0.25}_{4})^{\mathrm{T}}$ and $\Lambda(t) = 0.04t^{5/4}$. We set the censoring time to be $\min\{C^*, 50\}$, where $C^* \sim \text{Exp}(0.035)$; the censoring rate is approximately 34%. For each subject, either all covariates are observed or only the covariates with even indices (i.e., $X_2, X_4, \ldots, X_{100}$) are observed. We considered missing mechanisms of MCAR and MAR, generated in the same way as the unpenalized case.

We considered the penalized NPMLE, complete-case analysis, and single imputation. For single imputation, we imputed the missing values in the same way as for the

		NPMLE				Complete Case		Single Imputation	
Setting	Parameter	Bias	SE	SEE	CP	Bias	SE	Bias	SE
$n = 500$	β_1	−0.0012	0.0773	0.0751	0.93	−0.0524	0.0709	−0.0510	0.0710
$p_M = 20\%$	β_2	0.0070	0.0848	0.0839	0.93	−0.0455	0.0778	−0.0654	0.0792
	β_3	0.0046	0.0761	0.0791	0.94	−0.0486	0.0753	0.0181	0.0770
	β_4	0.0046	0.0691	0.0707	0.94	−0.0484	0.0682	−0.0157	0.0719
$n = 500$	β_1	−0.0017	0.0906	0.0867	0.92	−0.0541	0.0838	−0.0678	0.0794
$p_M = 40\%$	β_2	0.0071	0.0943	0.0973	0.94	−0.0464	0.0871	−0.0877	0.0865
	β_3	0.0078	0.0824	0.0848	0.94	−0.0458	0.0896	-0.0079	0.0837
	β_4	0.0046	0.0726	0.0758	0.95	−0.0458	0.0763	−0.0412	0.0735
$n = 1000$	β_1	0.0043	0.0517	0.0522	0.94	−0.0479	0.0484	−0.0467	0.0490
$p_M = 20\%$	β_2	−0.0027	0.0591	0.0582	0.95	−0.0550	0.0532	−0.0757	0.0534
	β_3	0.0076	0.0523	0.0549	0.96	−0.0450	0.0504	0.0215	0.0550
	β_4	0.0009	0.0505	0.0491	0.92	−0.0517	0.0495	−0.0195	0.0520
$n = 1000$	β_1	0.0051	0.0621	0.0600	0.92	−0.0481	0.0572	−0.0629	0.0549
$p_M = 40\%$	β_2	−0.0016	0.0702	0.0670	0.93	−0.0559	0.0636	−0.0962	0.0628
	β_3	0.0074	0.0569	0.0587	0.95	−0.0476	0.0603	−0.0066	0.0577
	β_4	0.0009	0.0523	0.0524	0.94	−0.0511	0.0568	−0.0457	0.0512

Note: See Note to Table 2.1.

Table 2.2: Results for unpenalized estimators of $\boldsymbol{\beta}$ under MAR

unpenalized case. For complete-case analysis and single imputation, we obtained the active sets using maximum penalized partial likelihood estimation with a LASSO penalty using the observed or completed data and then refitted the model over the active sets. BIC was used in choosing the best model for all three methods. We report the true positive rate (TPR), false discovery rate (FDR), and mean squared error (MSE) for each method. These statistics, based on 500 simulation replicates, are summarized in Table 2.3.

The penalized NPMLE has the highest TPR and FDR among all three methods. This implies that the penalized NPMLE tends to select more covariates, both relevant and irrelevant ones. In terms of MSE, the penalized NPMLE is uniformly better than complete-case analysis, especially for MCAR. Single imputation could yield smaller MSE

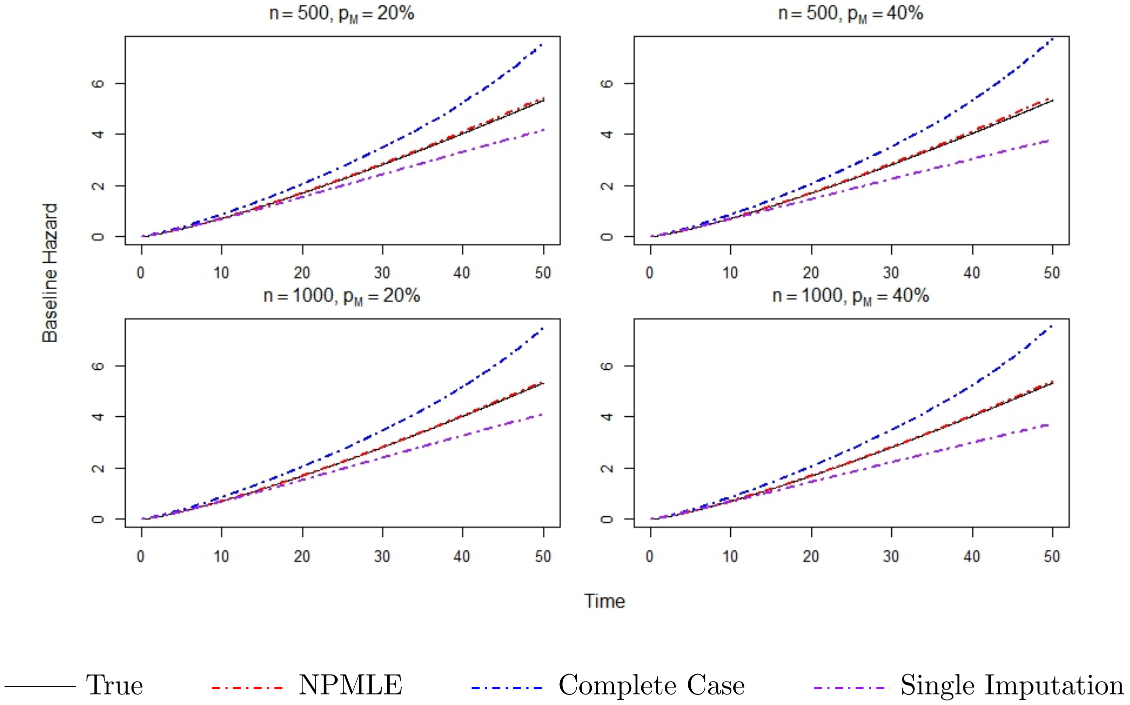

Figure 2.2: Results for unpenalized estimators of Λ under MAR.

than the penalized NPMLE under MCAR, probably because single imputation is biased towards zero, as demonstrated in the simulation studies for the unpenalized case. By contrast, single imputation always yields higher MSE than the penalized NPMLE under MAR.

2.3.3 Unpenalized and penalized estimation under a misspecified covariate distribution

To evaluate the sensitivity of the proposed methods to the normality assumption, we conduct simulation studies with a misspecified covariate distribution. In particular, we generated the multivariate normal random vector as described above. Then, we transformed each component of the normal vector by $F_5^{-1} \circ \Phi$ and set the transformed value as a covariate, where Φ and F_5 are the cumulative distribution functions of the standard normal distribution and the t distribution with five degrees of freedom,

			$p_M = 20\%$			$p_M = 40\%$		
n	Pattern	Method	TPR	FDR	MSE	TPR	FDR	MSE
500	MCAR	Penalized NPMLE	0.9335	0.1012	0.1013	0.8647	0.1233	0.1517
		Complete Case	0.9160	0.1012	0.1219	0.8472	0.1231	0.1934
		Single Imputation	0.9322	0.0917	0.0978	0.8502	0.0889	0.1420
500	MAR	Penalized NPMLE	0.9165	0.1092	0.1126	0.8417	0.1283	0.1680
		Complete Case	0.8985	0.1092	0.1126	0.8112	0.1283	0.1785
		Single Imputation	0.8867	0.0810	0.1207	0.7800	0.0902	0.1805
1000	MCAR	Penalized NPMLE	0.9948	0.0814	0.0367	0.9820	0.0921	0.0498
		Complete Case	0.9903	0.0810	0.0432	0.9738	0.0917	0.0677
		Single Imputation	0.9948	0.0760	0.0352	0.9785	0.0681	0.0472
1000	MAR	Penalized NPMLE	0.9918	0.0770	0.0373	0.9725	0.0926	0.0532
		Complete Case	0.9845	0.0768	0.0439	0.9605	0.0926	0.0672
		Single Imputation	0.9793	0.0538	0.0452	0.9355	0.0654	0.0730

Note: “TPR” is the true positive rate; “FDR” is the false discovery rate; “MSE” is the mean squared error.

Table 2.3: Results for penalized estimators of $\boldsymbol{\beta}$ under MCAR and MAR

respectively. As a result, each covariate marginally follows a t distribution. The event and censoring times were then generated in the same way as the above. We considered both the unpenalized and penalized estimators. The results are shown in Tables 2.4, 2.5, and 2.6.

For the unpenalized methods, the results are similar in pattern as those under a correctly-specified covariate distribution. In particular, the NPMLE and complete-case analysis are unbiased under MCAR, while single imputation tends to be biased. However, the standard errors of the NPMLE and single imputation are of similar level, and single imputation does not have noticeable smaller standard errors than the other methods. In terms of MSE, the NPMLE dominates the other two methods under all settings.

For the penalized methods, similar to the results in Table 2.3, the penalized NPMLE has overall the highest TPR and FDR. The penalized NPMLE always has a smaller MSE than complete-case analysis. Under MCAR, the penalized NPMLE and single imputation have similar values of MSE. Under MAR, single imputation has the largest MSE among all three methods.

2.4 Real data analysis

We analyzed a dataset of kidney renal clear cell carcinoma (KIRC) from TCGA. The dataset, released in November 2015, was downloaded through the RTCGA package (Kosinski et al., 2016) in R. In the study, times to new tumor events and death were collected, which were potentially subject to right censoring. Also, omic variables including gene expressions, measured by RNA sequencing, and protein expressions, measured by reverse-phase protein array, were collected for some or most subjects. In this section, we focus on the association between time to death since initial diagnosis and the omic variables.

The dataset contains 20,531 gene expressions and 217 protein expressions. There are 530 subjects with both survival data and gene expression measurements. Among these subjects, 475 have measurements in protein expressions. Following Zhao et al. (2015), we filtered out gene expressions with 0 median absolute deviation, resulting in the removal of 2865 genes. Then, following The Cancer Genome Atlas Research Network (2013), we selected the top 1500 gene expressions with the largest maximum absolute deviation. We then performed the $\log(1+x)$-transformation on the gene expressions. In addition, we removed 5 protein expressions that were missing for over 90% of the subjects. After the above preliminary processing, we performed supervised screening by fitting a separate Cox model for time to death against each gene or protein expression and selected the top 150 covariates with the smallest p-values; here, complete-case analysis was used in the presence of missing values. This resulted in 16 protein expressions and 134 gene expressions selected for downstream analyses. The missing proportion for protein expressions is 10%, and the censoring rate is 58%.

We performed the penalized NPMLE, complete-case analysis, and single imputation approaches on the processed data. Note that covariates were standardized during variable selection, and models were refit on the original unstandardized scale. The analysis results are presented in Table 2.7. The penalized NPMLE, complete-case analysis, and single imputation selected 8, 4, and 5 covariates, respectively. This is consistent with the findings in the simulation studies that the penalized NPMLE tends to select the largest number of features.

We evaluate the prediction performance of the three methods as follows. We randomly split the data into training and testing sets with a 7:3 ratio of sample sizes and performed the three estimation procedures on the training data. Then, to facilitate evaluation of the fitted models, we imputed the missing values in the testing data by single imputation, where the whole data set was used to estimate the imputation model. We calculated the concordance index (C-index) (Harrell et al., 1982) between the event time and the estimated $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$ on the (imputed) testing data. The above procedure was repeated 100 times. Note that we imputed the testing data in the exact same way as in the single imputation method in the simulation studies. The average C-index values over the 100 splits for penalized NPMLE, complete-case analysis, and single imputation are 0.660, 0.656, and 0.658, respectively. The C-index values are similar due to the small missing proportion, with the proposed method having a slight advantage.

2.5 Discussion

In this chapter, 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.

As for likelihood-based methods in general, we need to impose modeling assumptions on the missing covariates (except when only a few covariates are involved, in which case a fully nonparametric model can be fitted). The proposed methods depend crucially on the Gaussian assumptions on the covariates; without these assumptions, the transformation approach to reduce the dimension of numerical integration is not applicable. One possible approach to relax the Gaussian assumptions is to assume that the observed covariates are transformed values of underlying Gaussian variables, that is, $X_j = g_j(Z_j)$ for some transformation function g_j and Gaussian variable Z_j; this also allows X_j's to be discrete. We then assume that the dependence between the outcome and covariates is mediated through $\boldsymbol{Z} \equiv (Z_1, \ldots, Z_p)^{\mathrm{T}}$, such that $\lambda(t \mid \boldsymbol{X}, \boldsymbol{Z}) = \lambda(t)e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}}$. In this scenario, the proposed transformation approach can still be adopted.

The proposed transformation technique can also be applied to random effect 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 modelled 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.

Due to its flexibility, multiple imputation is a popular approach for handling missing data. The proposed methods have advantages over multiple imputation approaches, especially when variable selection is desirable, in two key respects. 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. Second, multiple imputation is not amenable to simultaneous variable selection and estimation. One advantage of penalization methods is that they perform variable selection and estimation simultaneously: penalization shrinks the estimators towards zero, and some estimators are shrunk to exactly zero, thereby eliminating the corresponding covariates. However, even though penalized estimation can be performed on each imputed dataset to yield a sparse estimator, the final estimator that combines results from all imputed datasets is generally not sparse, as a variable would be retained even if it is selected in just one of the imputed datasets. By contrast, because the proposed method imposes a penalty on a single likelihood, it performs simultaneous variable selection and estimation.

In the proposed methods, we fit an unstructured covariance matrix for the covariates and estimate it by unpenalized MLE. As a result, we cannot accommodate a high-dimensional setting with $p > n$, as the variance estimator would not be positive definite. To accommodate high-dimensional data, one can consider shrinkage estimators for covariance estimation (Ledoit and Wolf, 2004; Warton, 2008). 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.

2.6 Appendix: Additional simulation results

		NPMLE				Complete Case		Single Imputation	
Setting	Parameter	Bias	SE	SEE	CP	Bias	SE	Bias	SE
$n = 500$	β_1	−0.0020	0.0637	0.0632	0.93	−0.0005	0.0648	−0.0266	0.0615
$p_M = 20\%$	β_2	0.0032	0.0710	0.0703	0.94	0.0053	0.0711	−0.0221	0.0668
	β_3	0.0070	0.0653	0.0669	0.95	0.0051	0.0686	−0.0456	0.0695
	β_4	0.0072	0.0595	0.0605	0.94	0.0072	0.0636	−0.0323	0.0596
$n = 500$	β_1	−0.0026	0.0745	0.0732	0.94	0.0023	0.0765	−0.0464	0.0681
$p_M = 40\%$	β_2	0.0014	0.0780	0.0817	0.95	0.0066	0.0789	−0.0432	0.0704
	β_3	0.0107	0.0710	0.0727	0.94	0.0087	0.0798	−0.0815	0.0777
	β_4	0.0077	0.0657	0.0658	0.92	0.0094	0.0732	−0.0624	0.0640
$n = 1000$	β_1	0.0041	0.0448	0.0439	0.92	0.0057	0.0454	−0.0225	0.0429
$p_M = 20\%$	β_2	−0.0004	0.0487	0.0483	0.95	0.0012	0.0487	−0.0269	0.0447
	β_3	0.0054	0.0452	0.0461	0.94	0.0046	0.0489	−0.0485	0.0488
	β_4	0.0019	0.0418	0.0417	0.94	0.0012	0.0443	−0.0381	0.0452
$n = 1000$	β_1	0.0030	0.0513	0.0505	0.93	0.0068	0.0524	−0.0424	0.0467
$p_M = 40\%$	β_2	−0.0023	0.0557	0.0554	0.93	0.0020	0.0568	−0.0476	0.0501
	β_3	0.0082	0.0479	0.0501	0.93	0.0059	0.0551	−0.0863	0.0520
	β_4	0.0026	0.0455	0.0453	0.94	0.0019	0.0507	−0.0685	0.0483

Note: See Note to Table 1.

Table 2.4: Results for unpenalized estimators of $\boldsymbol{\beta}$ with a misspecified distribution under MCAR

		NPMLE				Complete Case		Single Imputation	
Setting	Parameter	Bias	SE	SEE	CP	Bias	SE	Bias	SE
$n = 500$ $p_M = 20\%$	β_1	-0.0011	0.0625	0.0610	0.93	-0.0445	0.0587	-0.0480	0.0581
	β_2	0.0083	0.0681	0.0680	0.94	-0.0371	0.0636	-0.0590	0.0642
	β_3	0.0044	0.0643	0.0652	0.94	-0.0395	0.0632	0.0027	0.0684
	β_4	0.0052	0.0564	0.0583	0.94	-0.0385	0.0549	-0.0219	0.0629
$n = 500$ $p_M = 40\%$	β_1	0.0002	0.0759	0.0703	0.92	-0.0452	0.0711	-0.0690	0.0664
	β_2	0.0050	0.0773	0.0783	0.94	-0.0422	0.0733	-0.0935	0.0717
	β_3	0.0090	0.0680	0.0706	0.94	-0.0386	0.0714	-0.0327	0.0741
	β_4	0.0048	0.0610	0.0634	0.94	-0.0428	0.0655	-0.0543	0.0657
$n = 1000$ $p_M = 20\%$	β_1	0.0047	0.0423	0.0421	0.93	-0.0394	0.0403	-0.0431	0.0403
	β_2	0.0009	0.0481	0.0467	0.95	-0.0439	0.0445	-0.0663	0.0442
	β_3	0.0049	0.0432	0.0448	0.94	-0.0382	0.0422	0.0039	0.0481
	β_4	0.0004	0.0407	0.0404	0.93	-0.0425	0.0411	-0.0283	0.0435
$n = 1000$ $p_M = 40\%$	β_1	0.0044	0.0497	0.0483	0.93	-0.0426	0.0465	-0.0667	0.0445
	β_2	-0.0011	0.0561	0.0536	0.94	-0.0489	0.0518	-0.0995	0.0502
	β_3	0.0065	0.0464	0.0484	0.95	-0.0411	0.0494	-0.0338	0.0530
	β_4	0.0007	0.0427	0.0437	0.95	-0.0459	0.0451	-0.0607	0.0456

Note: See Note to Table 1.

Table 2.5: Results for unpenalized estimators of $\boldsymbol{\beta}$ with a misspecified distribution under MAR

			$p_M = 20\%$			$p_M = 40\%$		
n	Pattern	Method	TPR	FDR	MSE	TPR	FDR	MSE
500	MCAR	Penalized NPMLE	0.9848	0.1034	0.0535	0.9465	0.1238	0.0823
		Complete Case	0.9773	0.1034	0.0645	0.9430	0.1238	0.1094
		Single Imputation	0.9830	0.0877	0.0503	0.9300	0.0953	0.0833
500	MAR	Penalized NPMLE	0.9790	0.1077	0.0558	0.9285	0.1285	0.0939
		Complete Case	0.9715	0.1075	0.0582	0.9255	0.1282	0.0967
		Single Imputation	0.9603	0.0841	0.0660	0.8692	0.0943	0.1175
1000	MCAR	Penalized NPMLE	0.9988	0.0756	0.0214	0.9965	0.0958	0.0267
		Complete Case	0.9988	0.0752	0.0242	0.9948	0.0954	0.0379
		Single Imputation	0.9988	0.0675	0.0208	0.9960	0.0635	0.0265
1000	MAR	Penalized NPMLE	0.9985	0.0852	0.0216	0.9953	0.0976	0.0281
		Complete Case	0.9988	0.0848	0.0254	0.9935	0.0972	0.0362
		Single Imputation	0.9950	0.0670	0.0274	0.9760	0.0739	0.0443

Note: See Note to Table 3.

Table 2.6: Results for penalized estimators of $\boldsymbol{\beta}$ with a misspecified distribution under MCAR and MAR

Variable	Penalized NPMLE	Complete Case	Single Imputation
Protein – MAPK_pT202_Y204	−0.3288	−0.3209	−0.3266
Gene – CDCA3 (83461)	0.0797	0.1487	0.1172
Gene – SHOX2 (6474)	0.1089	0.1608	0.1432
Gene – LOC286467 (286467)	0.0883	0.1642	0.1338
Gene – DNASE1L3 (1776)	−0.0976	·	−0.1261
Gene – BRD9 (65980)	0.1717	·	·
Gene – PHF21A (51317)	0.4149	·	·
Gene – CARS (833)	0.1477	·	·

Note: For gene expressions, the Entrez IDs are given in the parentheses.

Table 2.7: Regression parameter estimates for the KIRC data

Chapter 3

Cure models with a parametric tail constraint under insufficient follow-up

3.1 Overview

In Chapter 3, we consider the promotion cure model under insufficient follow-up. Section 3.2 introduces the proposed estimation and inference procedures. Section 3.3 presents some preliminary theoretical results. Section 3.4 reports simulation results of the proposed method and the NPMLE based on the zero-tail constraint under different scenarios. Concluding remarks are given in Section 3.5.

3.2 Methodology

3.2.1 Nonparametric maximum likelihood estimation

Let C be a random censoring time, $Y = \min(T, C)$, and $\Delta = I(T \leq C)$. For a sample of size n, the observed data consist of $(Y_i, \Delta_i, \boldsymbol{X}_i)_{i=1,\ldots,n}$. Under the zero-tail constraint, the likelihood is

$$\prod_{i=1}^{n}\{\exp(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X_i})f(Y_i)\}^{\Delta_i}\exp\{-\exp(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)[\xi_i F(Y_i)+(1-\xi_i)F(\infty)]\},$$

where $\xi_i \equiv I(Y_i \leq t_{(n)})$ is an indicator for uncured subjects, $t_{(n)} \equiv \max\{Y_i : \Delta_i = 1\}$ is the largest observed event time, and f is the derivative of F.

We first estimate the model parameters using NPMLE, treating F as a nondecreasing step function that jumps at Y_i for which $\Delta_i = 1$. In this formulation, $F(t) = \sum_{i:Y_i\leq t} F\{Y_i\}$, where $F\{Y_i\}$ is the jump size of F at Y_i. The NPMLE $(\widehat{\alpha}_n, \widehat{\boldsymbol{\beta}}_n, \widehat{F}_n)$ maximizes

$$L(\alpha,\boldsymbol{\beta},F)=\prod_{i=1}^{n}\{\exp(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)F\{Y_i\}\}^{\Delta_i}\exp\{-\exp(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)[\xi_i F(Y_i)+(1-\xi_i)F(\infty)]\}, \tag{3.1}$$

subject to the constraint $F(\infty) = \sum_{i=1}^{n} F\{Y_i\} = 1$.

As demonstrated by Beyhum et al. (2022), the Cox proportional hazards model is closely related to the promotion time cure model. Consider the Cox model with the survival function of T conditional on $\boldsymbol{X}$ given by

$$P(T > t \mid \boldsymbol{X}) = \exp\big\{-\Lambda(t)e^{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{X}}\big\},$$

where $\boldsymbol{\eta}$ is a vector of regression coefficients, and Λ is a nonnegative, nondecreasing, nonparametric function. There is a one-to-one relationship between the parameter sets $(\alpha, \boldsymbol{\beta}, F)$ and $(\boldsymbol{\eta}, \Lambda)$, specifically that $\boldsymbol{\beta} = \boldsymbol{\eta}$, $\alpha = \lim_{t\to\infty} \log \Lambda(t)$, and $F = \exp(-\alpha)\Lambda$. Because the regression coefficients of $\boldsymbol{X}$ are the same in the two models, we use $\boldsymbol{\beta}$ instead of $\boldsymbol{\eta}$ in the Cox model formulation in the sequel. The likelihood presented in (3.1) can be reparameterized using $(\boldsymbol{\beta}, \Lambda)$. The NPMLE of the Cox model, where Λ is treated as a step function that jumps at the observed event times, is defined as

$$(\widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n) = \underset{(\boldsymbol{\beta}, \Lambda)}{\arg\max} \prod_{i=1}^{n} \{\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i)\Lambda\{Y_i\}\}^{\Delta_i} \exp\{-\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i)\Lambda(Y_i)\},$$

where $\Lambda\{Y_i\}$ is the jump size of Λ at Y_i. The NPMLE for the promotion time cure model is then obtained via the transformation

$$(\widehat{\alpha}_n, \widehat{\boldsymbol{\beta}}_n, \widehat{F}_n) = (\log \widehat{\Lambda}_n(\infty), \widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n / \widehat{\Lambda}_n(\infty)).$$

3.2.2 Extrapolation-based cure probability estimator

Under insufficient follow-up, the NPMLE $\widehat{\alpha}_n$ could be severely biased. We propose an extrapolation approach based on extreme value theory to improve estimation. We first provide a brief overview of the relevant concepts in extreme value theory. A CDF F is said to belong to the domain of attraction of an extreme value distribution G if there exist normalizing constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\lim_{n\to\infty} F^n(a_n t + b_n) = G(t)$$

at every continuity point t of G. According to extreme value theory, G must take the form of one of the three classical extreme value distributions: Fréchet, Gumbel, and Weibull. Let $\tau_0 \equiv \sup\{t : F(t) < 1\}$ denote the right endpoint of the support of F. According to Theorem 1.2.6 of de Haan and Ferreira (2006), F belongs to a domain of attraction if and only if there exist a shape parameter $\gamma \in \mathbb{R}$, a positive continuous function h, and a positive function ψ, such that for all $t \in (\omega, \tau_0)$ with $\omega < \tau_0$,

$$1 - F(t) = \psi(t) \exp\left\{ -\int_{\omega}^{t} \frac{\mathrm{d}s}{h(s)} \right\}, \tag{3.2}$$

$\lim_{t\to\tau_0} \psi(t) = \psi \in (0, \infty)$, and the auxiliary function h satisfies

$$\begin{cases} \lim_{t\to\tau_0} h(t)/t = \gamma & \text{if } \gamma > 0, \\ \lim_{t\to\tau_0} h(t)/(\tau_0 - t) = -\gamma & \text{if } \gamma < 0, \\ \lim_{t\to\tau_0} h'(t) = 0 & \text{if } \gamma = 0. \end{cases} \tag{3.3}$$

The shape parameter γ, known as the extreme value index, determines the class of limiting distributions G, the tail behavior of F, and the nature of the right endpoint τ_0. The three domains of attraction are summarized in Table 3.1.

Domain of attraction	**Sign of γ**	**Tail behavior of F**	**Right endpoint (τ_0)**
Fréchet	$\gamma > 0$	heavy-tailed	infinite
Weibull	$\gamma < 0$	light-tailed	finite
Gumbel	$\gamma = 0$	moderate-tailed	finite or infinite

Table 3.1: Domains of attraction in extreme value theory

When F belongs to a domain of attraction, the aforementioned theorem implies that the tail of survival function $1-F$ can be approximated by some parametric functions. This serves as the foundation for our extrapolation method. In this chapter, we restrict our attention to distributions with $\tau_0 = \infty$, and therefore, the Weibull domain of attraction is not considered.

We now derive parametric tail approximations for F belonging to the Fréchet and Gumbel domains. As $t \to \tau_0 = \infty$, the function $\psi(t)$ approaches $\psi \in (0, \infty)$. For the Frćhet domain of attraction, (3.3) implies that for large t, we can approximate $h(t)$ as

$$h(t) \approx \gamma t.$$

Substituting this into (3.2), we obtain

$$1 - F(t) \approx \psi \exp\left\{-\int_\omega^t \frac{\mathrm{d}s}{\gamma s}\right\} = \psi \left(\frac{t}{\omega}\right)^{-1/\gamma}.$$

In this case, the survival function $1-F$ tends to zero under a power law with exponent $-\gamma^{-1}$. For the Gumbel domain of attraction, the tail behavior of $h(\cdot)$ is ambiguous due to the weaker condition $h'(t) \to 0$. In this case, we impose a strengthened von Mises condition (see Condition VM3' in Falk and Marohn (1993)), which serves as a sufficient condition for $h(t) \to \eta \in (0, \infty)$. Under this refined condition, we can approximate $h(t)$ for large t as

$$h(t) \approx \eta,$$

yielding the tail approximation

$$1 - F(t) \approx \psi \exp\left\{ -\int_{\omega}^{t} \frac{\mathrm{d}s}{\eta} \right\} = \psi \exp\left(-\frac{t-\omega}{\eta} \right).$$

Here, the survival function $1-F$ exhibits exponential decrease with rate parameter η^{-1}. By reparameterizing γ^{-1} and η^{-1} as μ, we obtain a unified parametric tail estimator $F_\gamma(\cdot\,; \psi, \mu)$, whose form depends on the domain of attraction:

$$F_\gamma(t; \psi, \mu) = \begin{cases} 1 - \psi(t/\omega)^{-\mu} & \text{if } \gamma > 0 \text{ (Fréchet)}, \\ 1 - \psi \exp\{-\mu(t-\omega)\} & \text{if } \gamma = 0 \text{ (Gumbel)}. \end{cases}$$

Given prior knowledge of the sign of the extreme value index γ — specifically, the domain of attraction to which F belongs — our goal is to estimate the parameter of the corresponding tail approximation. In Section 3.3, we demonstrate that the Breslow estimator within a promotion time cure model framework consistently estimates $\exp(\alpha_0)F_0(\cdot)$. Let $\boldsymbol{\theta}$ denote the set of parameters (α, ψ, μ) and define $L_\gamma(\cdot\,; \boldsymbol{\theta}) = \exp(\alpha)F_\gamma(\cdot\,; \psi, \mu)$. We propose finding the best fitting parametric tail $L_\gamma(\cdot\,; \boldsymbol{\theta})$ that minimizes the discrepancy between the parametric model and the Breslow estimator over an interval (ω, τ_c), where $\tau_c = \max\{Y_i\}$ represents the last follow-up time. In particular, we define a regularized estimator:

$$\widehat{\boldsymbol{\theta}}_{n,\lambda_n} = \arg\min_{\boldsymbol{\theta}} \int_{\omega}^{\tau_c} \left\{ L_\gamma(t;\boldsymbol{\theta}) - \widehat{\Lambda}_n(t) \right\}^2 \mathrm{d}t + \lambda_n\{\alpha^2 + (\log\psi)^2 + (\log\mu)^2\}, \tag{3.4}$$

where $\widehat{\Lambda}_n(\cdot)$ is the Breslow estimator of the baseline cumulative hazard, and $\lambda_n \geq 0$ is a tuning parameter. We penalize α, $\log\psi$, and $\log\mu$ to prevent overfitting, a common issue observed in unregularized settings. To obtain the minimizer, we employ an adaptive Newton algorithm (Mishchenko, 2023).

To select the tuning parameter λ_n, we adopt K-fold cross-validation, with

$$\lambda_n^* = \arg\min_{\lambda} \sum_{j=1}^{K} \left[\int_{\omega}^{\tau_c} \left\{ L_\gamma(t;\widehat{\boldsymbol{\theta}}_{n,\lambda}) - \widehat{\Lambda}_n^{(j)}(t) \right\}^2 \mathrm{d}t + \left\{ \exp(\widehat{\alpha}_{n,\lambda}^{(-j)}) - \exp(\widehat{\alpha}_n^{(-j)}) \right\}^2 \right],$$

where $\widehat{\boldsymbol{\theta}}_{n,\lambda}^{(-j)}$ are estimates obtained by excluding the j-th fold, $\widehat{\alpha}_n^{(-j)}$ is the NPMLE of α calculated under zero-tail constraint excluding the j-th fold, and $\widehat{\Lambda}_n^{(j)}$ is the Breslow estimator computed on the j-th fold. The cross-validation criterion comprises two components. The first integral quantifies the goodness-of-fit between the parametric tail estimator and the Breslow estimator on the validation fold. The second term addresses a practical identifiability issue observed in simulations, where multiple parameter sets yielded nearly identical goodness-of-fit scores. By penalizing deviations of $\widehat{\alpha}_{n,\lambda}^{(-j)}$ from $\widehat{\alpha}_n^{(-j)}$, we stabilize the solution, favoring parameter sets that align with the NPMLE.

The estimation procedure can be summarized as follows:

1. Compute the Breslow estimate $\widehat{\Lambda}_n(\cdot)$ and initialize a grid of tuning parameters.

2. Perform K-fold cross-validation to select the optimal tuning parameter λ_n^*. Estimate the parameter $\widehat{\boldsymbol{\theta}}_{n,\lambda_n^*}$ by minimizing the objective function in (3.4).

3. Calculate the final estimator as $\widetilde{\alpha}_n = \max(\widehat{\alpha}_n, \widehat{\alpha}_{n,\lambda_n^*})$ and $\widetilde{F}_n = \exp(-\widetilde{\alpha}_n)\widehat{\Lambda}_n$.

We call $\widetilde{\alpha}_n$ the extrapolation-based cure probability estimator. Note that $\exp\{-\exp(\widetilde{\alpha}_n)\}$

estimates the baseline cure probability, that is, the cure probability of a subject with $\boldsymbol{X} = \mathbf{0}$.

3.2.3 Inference for the cure probability

From NPMLE theory, $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\Lambda}_n$ are consistent and asymptotically normal, and we can use their asymptotic distributions to perform inference for $\boldsymbol{\beta}$ and Λ. However, $\widetilde{\alpha}_n$ is based on an approximation from extreme value theory, so inference for α is more difficult. In this subsection, we develop an inference procedure by assuming that the parametric tail structure holds exactly, in which case the extrapolation-based cure probability estimator is consistent. The asymptotic distribution of the estimator can be derived from the fact that it is a smooth transformation of the NPMLE of Λ.

By the delta method and for fixed λ, the estimator $\widehat{\boldsymbol{\theta}}_{n,\lambda}$ is approximately normally distributed with variance

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}_{n,\lambda}) = n^{-1}(\boldsymbol{A} + 2\lambda \boldsymbol{W})^{-1}\boldsymbol{B}(\boldsymbol{A} + 2\lambda \boldsymbol{W})^{-1},$$

where $\boldsymbol{W} = \mathrm{diag}(1, (1-\log\psi)/\psi^2, (1-\log\mu)/\mu^2)$ is a 3×3 diagonal matrix, and

$$\begin{aligned}\boldsymbol{A} =& 2\int_{\omega}^{\tau_c} \nabla L_\gamma(t; \widehat{\boldsymbol{\theta}}_{n,\lambda}) \nabla L_\gamma(t; \widehat{\boldsymbol{\theta}}_{n,\lambda})^{\mathrm{T}}\,\mathrm{d}t,\\ \boldsymbol{B} =& 4\int_{\omega}^{\tau_c}\int_{\omega}^{\tau_c} \widehat{\sigma}_n(s,t) \nabla L(s; \widehat{\boldsymbol{\theta}}_{n,\lambda}) \nabla L_\gamma(t; \widehat{\boldsymbol{\theta}}_{n,\lambda})^{\mathrm{T}}\,\mathrm{d}s\,\mathrm{d}t.\end{aligned}$$

Here, $\widehat{\sigma}_n$ is an estimator of the covariance process of the limiting distribution of $\{\sqrt{n}(\widehat{\Lambda}_n - \Lambda_0)(t) : t \in [0, \tau_c]\}$, given by

$$\begin{aligned}\widehat{\sigma}_n(s,t) = &\left\{\int_0^s Z(u; \widehat{\boldsymbol{\beta}}_n)\,\mathrm{d}\widehat{\Lambda}_n(u)\right\}\left\{\mathcal{I}_n(\widehat{\boldsymbol{\beta}}_n)\right\}^{-1}\left\{\int_0^t Z(u; \widehat{\boldsymbol{\beta}}_n)\,\mathrm{d}\widehat{\Lambda}_n(u)\right\}\\ &+ \int_0^{\min(s,t)} \frac{\mathrm{d}\widehat{\Lambda}_n(u)}{n^{-1}\sum_{j=1}^n I(Y_j \geq u)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}} \boldsymbol{X}_j)},\end{aligned}$$

where $\mathcal{I}_n(\widehat{\boldsymbol{\beta}}_n)^{-1}$ is a consistent estimator of the asymptotic variance of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$, and

$$Z(u;\widehat{\boldsymbol{\beta}}_n) = \frac{\sum_{j=1}^n I(Y_j \geq u)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X}_j)\boldsymbol{X}_j}{\sum_{j=1}^n I(Y_j \geq u)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X}_j)}.$$

Let $\widehat{v}_n$ be the upper-left element of $\widehat{\mathrm{Var}(\widehat{\boldsymbol{\theta}}_{n,\lambda_n^*})}$. We can construct a $(1-c)$ confidence interval for α as $\widetilde{\alpha}_n \pm z_{c/2}\widehat{v}_n^{1/2}$, where z_c is the c-upper quantile of the standard normal distribution.

3.3 Asymptotic properties

In this section, we establish the asymptotic properties of the NPMLE under a theoretical framework where the follow-up time diverges to infinity as the sample size increases. This serves as a preliminary result for investigating the asymptotic efficiency of the extrapolation cure probability estimator relative to the NPMLE in the future. We assume the following conditions:

(C.1) The covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$ have a bounded support.

(C.2) The true regression coefficients $(\alpha_0, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}$ belongs to a compact set $\mathcal{A} \times \mathcal{B}$, and the true CDF F_0 belongs to a space of absolutely continuous functions from $[0,\infty)$ to $[0,1]$.

(C.3) The censoring times are derived from a triangular array of random variables. In particular, there exists a diverging, nondecreasing sequence $\{\tau_n\}$ and i.i.d. random variables $\widetilde{C}_1, \widetilde{C}_2, \ldots$ with $P(\widetilde{C}_i = \infty \mid \boldsymbol{X}) > \delta$, such that under a sample of size n, the censoring time of the ith subject is $C_{ni} = \min(\widetilde{C}_i, \tau_n)$, where δ is a positive constant.

In the theoretical development, to emphasize that the distribution of the censoring time depends on the sample size, we denote the censoring time, the observed time, and event indicator of the ith subject by C_{ni}, $Y_{ni} \equiv \min(T_i, C_{ni})$ and $\Delta_{ni} \equiv I(T_i \leq C_{ni})$, respectively. Also, we define the generic random variables $\widetilde{Y} = \min(T, \widetilde{C})$ and $\widetilde{\Delta} = I(T \leq \widetilde{C})$ and use $\widetilde{Y}_i$ and $\widetilde{\Delta}_i$ to denote the realizations of $\widetilde{Y}$ and $\widetilde{\Delta}$ for the ith subject, respectively. Note that $Y_{ni} = \min(\widetilde{Y}_i, \tau_n)$ and $\Delta_{ni} = \widetilde{\Delta}_i I(\widetilde{Y}_i \leq \tau_n)$.

Let $\mathbb{P}_n$ and $\mathbf{P}$ denote the empirical measure and the true probability measure respectively. Remark that these measures are defined on the random vectors $(\widetilde{Y}, \widetilde{\Delta}, \boldsymbol{X})$. In particular, for any measurable function $Q(\widetilde{Y}, \widetilde{\Delta}, \boldsymbol{X})$, the empirical measure and true measures are given by:

$$\begin{aligned}\mathbb{P}_n[Q(\widetilde{Y}, \widetilde{\Delta}, \boldsymbol{X})] =& \frac{1}{n}\sum_{i=1}^{n} Q(\widetilde{Y}_i, \widetilde{\Delta}_i, \boldsymbol{X}_i),\\ \mathbf{P}[Q(\widetilde{Y}, \widetilde{\Delta}, \boldsymbol{X})] =& \mathrm{E}[Q(\widetilde{Y}, \widetilde{\Delta}, \boldsymbol{X})].\end{aligned}$$

Before presenting the main theoretical results, we first introduce a key lemma:

Lemma 1. Let $Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)$ be an arbitrary measurable and bounded function that satisfies $Q(\tau_n, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda) = Q(\infty, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)$ and that $\{Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda) : \boldsymbol{\beta} \in \mathcal{B}, \Lambda \in \mathcal{L}\}$ is a Glivenko–Cantelli class, where

$$\mathcal{L} = \{f : [0, \infty) \to \mathbb{R}, f(0) = 0, f \text{ is monotone increasing and bounded}\}.$$

Let $R_{ni} = I(\tau_n > Y_{ni} \geq y) + I(Y_{ni} \geq \tau_n)$. Under conditions (C.1)–(C.3), we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(Y_{ni} \geq \tau_n) Q(Y_{ni}, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \mathbf{P}[I(\widetilde{Y} = \infty) Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)\right| \xrightarrow{a.s.} 0, \quad (3.5)$$

$$\sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^{n} R_{ni}(y) Q(Y_{ni}, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \mathbf{P}[I(\widetilde{Y} \geq y) Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)\right| \xrightarrow{a.s.} 0, \quad (3.6)$$

$$\sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^{n} I(Y_{ni} \leq y)\Delta_{ni} Q(Y_{ni}, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \mathbf{P}[I(\widetilde{Y} \leq y)\widetilde{\Delta} Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)\right| \xrightarrow{a.s.} 0, \quad (3.7)$$

$$\sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^{n}\Delta_{ni}Q(Y_{ni},\boldsymbol{X}_i;\boldsymbol{\beta},\Lambda)-\mathbf{P}[\widetilde{\Delta}I(\widetilde{Y}<\infty)Q(\widetilde{Y},\boldsymbol{X};\boldsymbol{\beta},\Lambda)\right|\xrightarrow{a.s.}0. \tag{3.8}$$

Remark. (3.7) and (3.8) do not require the condition $Q(\tau_n,\boldsymbol{X};\boldsymbol{\beta},\Lambda)=Q(\infty,\boldsymbol{X};\boldsymbol{\beta},\Lambda)$.

The proof of the theoretical results are given in the appendix. We have the following theorem about the consistency of the NPMLE.

Theorem 1. Under conditions (C.1)–(C.3), the NPMLE of the promotion time cure model obtained from (3.1) is strongly consistent:

$$\|(\widehat{\alpha}_n,\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}})-(\alpha_0,\boldsymbol{\beta}_0^{\mathrm{T}})\|_1\xrightarrow{a.s.}0 \qquad \text{and} \qquad \sup_{t\in[0,\infty)}|\widehat{F}_n(t)-F_0(t)|\xrightarrow{a.s.}0.$$

Within the current theoretical framework, we have established the consistency of the NPMLE. Here, we present a preliminary outline of the proof for the extrapolation cure probability estimator, noting that the full theoretical derivation remains a work in progress. Intuitively, under insufficient follow-up, NPMLE $\widehat{\alpha}_n$ incurs a bias of $-\log F_0(\tau_c)$. We aim to demonstrate that the bias of the proposed estimator depends on the goodness-of-fit of the parametric tail, thereby potentially exhibiting a smaller bias compared to the NPMLE. Although the theoretical advantages of the proposed estimator require further characterization, simulation studies in Section 3.4 demonstrate that our approach outperforms NPMLE especially when $F_0(\tau_c)\ll 1$.

3.4 Simulation studies

3.4.1 Setup

In this section, we examine the finite sample performance of the proposed estimator. We set $p=5$ and simulate the covariate vector $\boldsymbol{X}$ from a multivariate normal distribution, with mean zero and an AR(1) covariance matrix $(0.5^{|i-j|})_{i,j=1,\ldots,5}$. The regression coefficients are set to $\boldsymbol{\beta}=(0.5,0.5,0.5,0.5,0.5)^{\mathrm{T}}$, and the baseline cure probability $\pi(\alpha)=\exp(-\exp(\alpha))$ (for a subject with $\boldsymbol{X}=\mathbf{0}$) is set to 0.25 or 0.50. We set the sample size to be $n = 1000$.

The distribution function F belongs to either the Fréchet or Gumbel domain of attraction. For the Fréchet domain of attraction, we consider the Fréchet distribution, half-Cauchy distribution, and generalized Pareto distribution. For the Gumbel domain of attraction, we consider the Gumbel distribution, Gamma distribution, and exponential distribution. Parameter configurations for these distributions are summarized in Table 3.2.

Domain of attraction	**Distribution**	**CDF**	**Support**	**Shape (a)**	**Scale (b)**
Fréchet	Fréchet	$\exp(-t^{-a})$	$t \in (0, \infty)$	1	$\cdot$
Fréchet	Cauchy	$\frac{2}{\pi} \arctan\left(\frac{t}{b}\right)$	$t \in (0, \infty)$	$\cdot$	1
Fréchet	Pareto	$1 - \left(\frac{b}{t}\right)^a$	$t \in (b, \infty)$	1	1
Gumbel	Gumbel	$\exp(-\exp(-\frac{t}{b}))$	$t \in \mathbb{R}$	$\cdot$	1
Gumbel	Gamma	$\Gamma(a)^{-1}\gamma(a, \frac{t}{b})$	$t \in (0, \infty)$	2	1
Gumbel	Exponential	$1 - \exp(-\frac{t}{b})$	$t \in (0, \infty)$	$\cdot$	1

Table 3.2: Distributions and parameter configurations.

Let q_{25} and q_{95} denote the 25th and 95th percentiles of the event time distribution, respectively. We set the last follow-up time to $\tau_c = r(q_{95} - q_{25}) + q_{25}$ for some $r > 0$, which is a linear interpolation between the 25th and 95th percentiles. We consider $r = 0.1, 0.2, \ldots, 1.5$. The censoring time is a mixture of $\text{Unif}[0, \tau_c]$ and a point mass at τ_c, with mixing probabilities 0.95 and 0.05, respectively.

We set ω to be the 90th empirical percentile of the observed event times. The estimator $\widetilde{\alpha}_n = \min(\widehat{\alpha}_{n,\lambda_n^*}, \widehat{\alpha}_n)$ is obtained following the procedure described in Section 3.2, where the optimal tuning parameter λ_n^* is determined through 5-fold cross-validation. We compare the proposed estimator with the NPMLE $\widehat{\alpha}_n$. For both estimators, we report the empirical mean, mean-squared error (MSE), and coverage probability (CP) of the 95% confidence intervals for the intercept α across all simulation scenarios. The results are presented in Figures 3.1–3.4. For each configuration, we simulate 1000 replicates.

3.4.2 Results

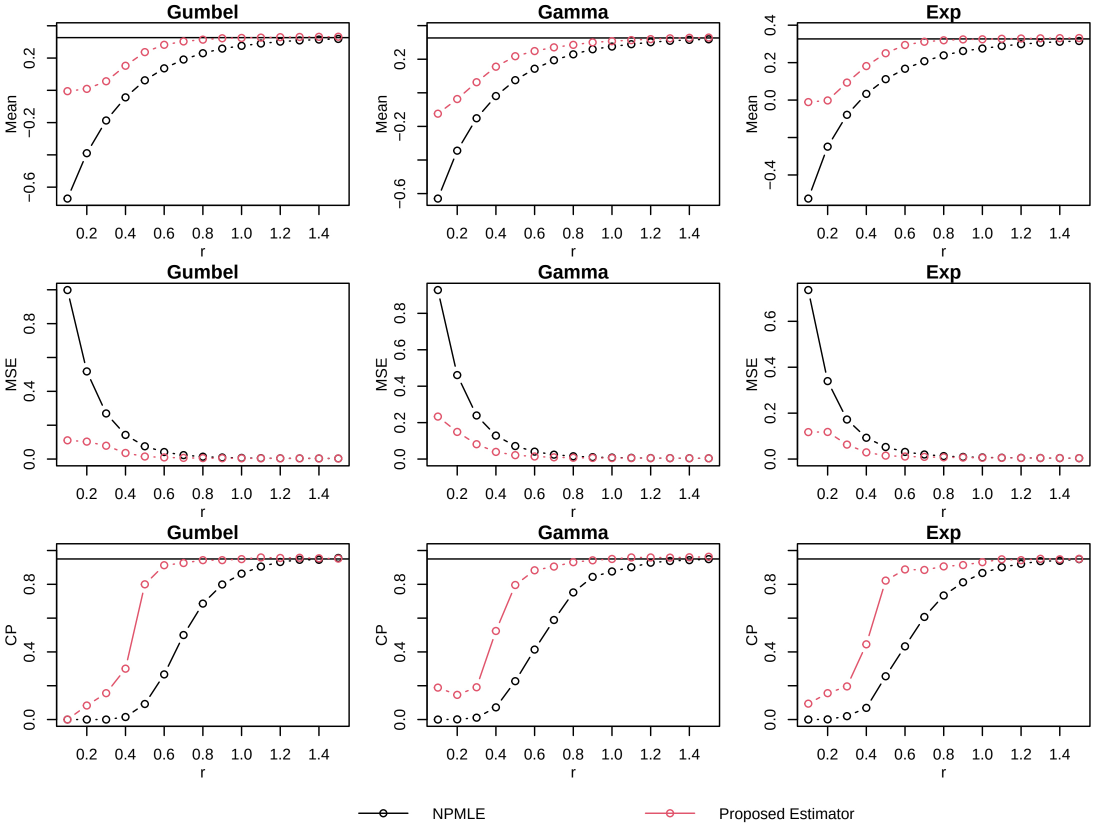

Figure 3.1: Statistic of α under the Gumbel domain and $\pi(\alpha) = 0.25$

Figures 3.1 and 3.2 illustrate the empirical performance for the proposed estimator and NPMLE under the Gumbel domain at baseline cure probabilities of 0.25 and 0.50, respectively. We draw three observations from the results. First, for sufficiently long follow-up ($r \geq 1$), the proposed estimator demonstrates comparable efficiency to the NPMLE across all distributions (Gumbel, Gamma, and Exponential distribution), as evidenced by overlapping empirical means, MSEs, and CPs. This alignment suggests that the proposed estimator is as efficient as the NPMLE when the follow-up is almost sufficient.

Second, for insufficient follow-up ($r < 1$), the proposed estimator uniformly outper-

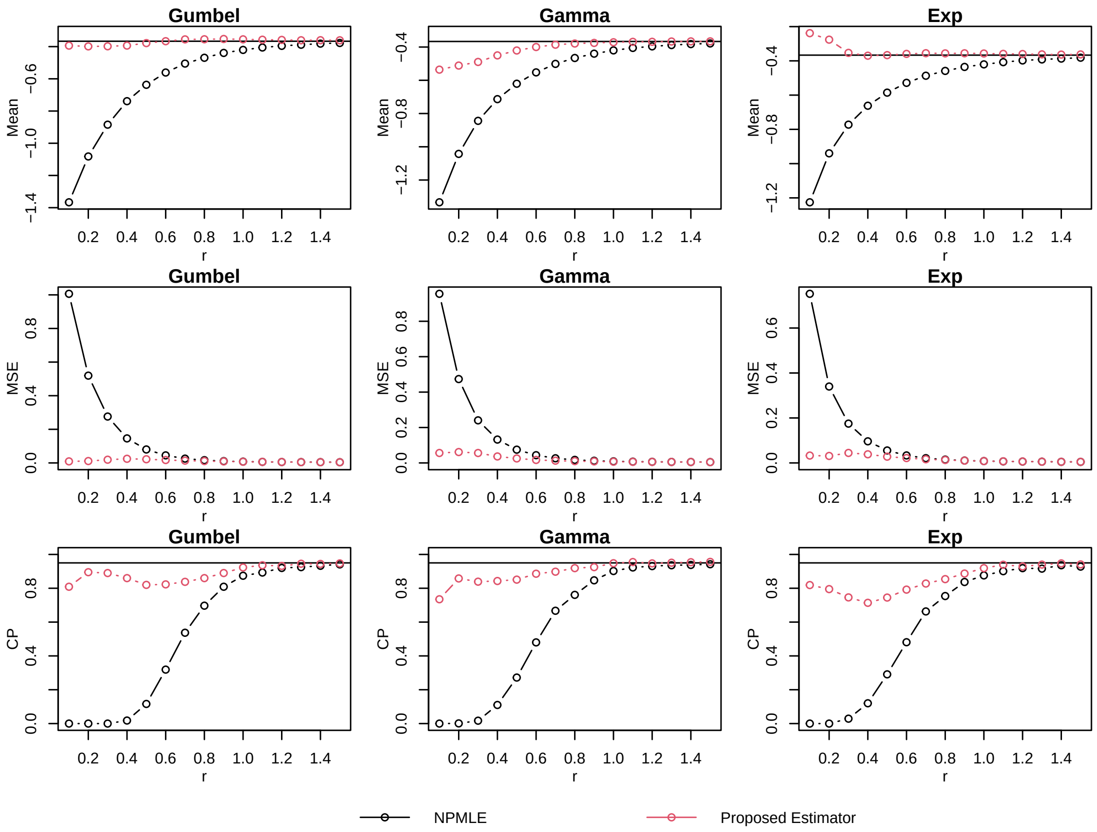

Figure 3.2: Statistic of α under the Gumbel domain and $\pi(\alpha) = 0.50$

forms the NPMLE. Specifically, it exhibits consistently smaller bias and lower MSE across all distributions. For instance, the proposed estimator maintains an MSE below 0.4 for $r < 1$, whereas the NPMLE exceeds 0.2 for $r < 0.4$ under the Gumbel and Gamma distribution. Similarly, the proposed estimator achieves near-nominal CPs for $r > 0.4$, whereas the NPMLE achieves that only for $r \geq 1$.

Third, the proposed estimator shows enhanced robustness at higher baseline cure probabilities. At $\pi(\alpha) = 0.50$, it substantially reduces MSE to below 0.1 for all r and has a near-zero bias. Furthermore, it achieves near-nominal CPs for $r > 0.4$. This is in contrast with $\pi(\alpha) = 0.25$, where the MSE reduction is less substantial and longer follow-up is required for achieving near-nominal CPs, suggesting that the estimator leverages increased cure-rate information more effectively at higher $\pi(\alpha)$.

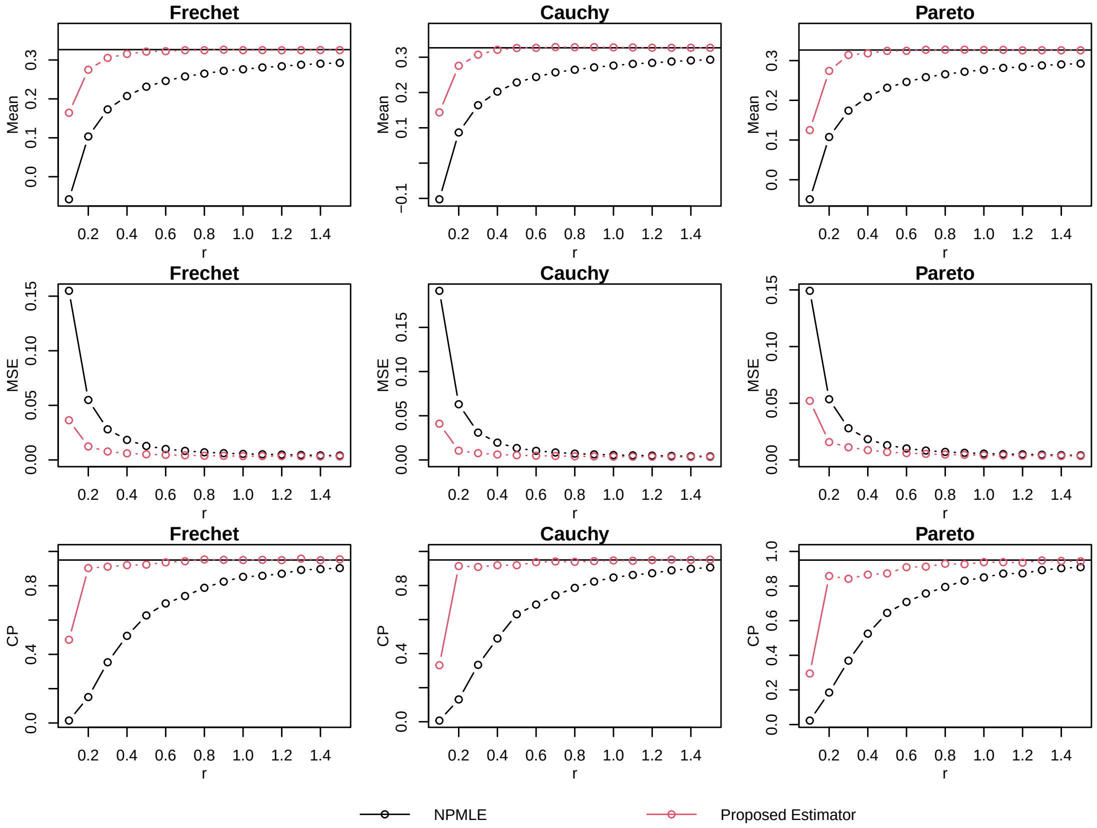

Figure 3.3: Statistic of α under the Fréchet domain and $\pi(\alpha) = 0.25$

Figures 3.3 and 3.4 illustrate the empirical performance for the proposed estimator and NPMLE under the Fréchet domain at baseline cure probabilities of 0.25 and 0.50, respectively. The proposed estimator under the Fréchet domain presents patterns that resemble the Gumbel case. For instance, the empirical performance of both estimators converges across all distributions (Fréchet, Cauchy, and Pareto distribution) as the follow-up becomes more sufficient and the proposed estimator has superior performance than the NPMLE for $r < 1$.

It is notable that the proposed estimator demonstrates distinct advantages specifically in the Fréchet domain. First, the estimator achieves exceptional stability across all distributions (Fréchet, Cauchy, Pareto distribution), with MSE values consistently below 0.05 for $r < 1$. This precision is coupled with near-zero bias and rapid convergence to

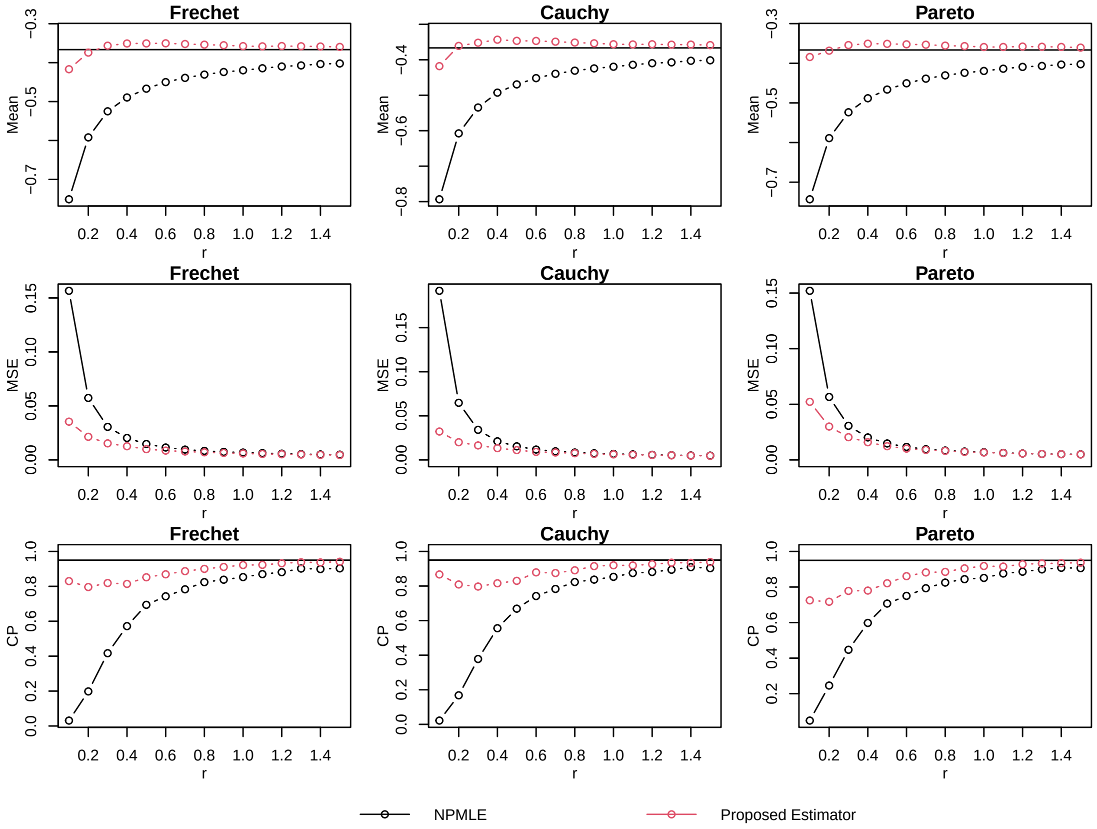

Figure 3.4: Statistic of α under the Fréchet domain and $\pi(\alpha) = 0.50$

nominal coverage probability. For instance, when $r = 0.2$ and $\pi(\alpha) = 0.25, 0.50$, the proposed estimator already achieves a performance that is similar to the performance as if we have sufficient follow-up.

We also conduct simulation for both the Fréchet and Gumbel domains under a nearly sufficient follow-up condition ($r = 10$). Table 3.3 summarizes the empirical performance of the NPMLE and the proposed estimator. Consistent with our earlier findings, the proposed estimator demonstrates efficiency comparable to the NPMLE when the follow-up is sufficient, as evidenced by their nearly identical MSEs. Additionally, both estimators achieve the nominal CP of 0.95.

			Distribution					
$\pi(\alpha)$	**Estimator**	**Statistic**	**Fréchet**	**Cauchy**	**Pareto**	**Gumbel**	**Gamma**	**Exp**
0.25	$\widehat{\alpha}_n$	MSE	0.0024	0.0024	0.0024	0.0024	0.0025	0.0025
		CP	0.944	0.945	0.946	0.938	0.941	0.940
	$\widetilde{\alpha}_n$	MSE	0.0025	0.0025	0.0025	0.0024	0.0025	0.0025
		CP	0.951	0.947	0.948	0.937	0.941	0.940
0.50	$\widehat{\alpha}_n$	MSE	0.0032	0.0031	0.0032	0.0030	0.0031	0.0031
		CP	0.941	0.940	0.946	0.947	0.946	0.944
	$\widetilde{\alpha}_n$	MSE	0.0032	0.0032	0.0032	0.0030	0.0031	0.0031
		CP	0.948	0.950	0.951	0.947	0.947	0.944

Table 3.3: Statistic of α under almost sufficient follow-up ($r = 10$)

3.5 Discussion

Our proposed method presupposes that the domain of attraction of F is known. However, this assumption may be impractical. To address this issue, graphical methods such as probability plots can be used to determine the domain of attraction. The fundamental concept behind constructing a probability plot is to apply distribution-specific transformations to both a random variable Z and its CDF F such that the resulting plot of the transformed Z against the transformed F forms a straight line with a slope of 1. Castillo et al. (1989) and Bhati and Ravi (2018) proposed the Gumbel and Fréchet probability plots, respectively. Their work illustrated the use of these graphical tools to identify the domain of attraction.

Insufficient follow-up complicates the diagnostic procedure. These graphical methods require a consistent estimator of F across the entire support $(0, \tau_0)$; however, this may not be possible under insufficient follow-up. Therefore, we propose an ad hoc procedure for identifying the domain of attraction, based on the methodology developed by Castillo et al. (1989):

1. Hypothesize that the CDF F belongs to the Gumbel domain of attraction. Estimate $\widetilde{F}_n$ using the proposed extrapolation method over an extrapolation interval (ω, τ_c).

2. Construct the Gumbel probability plot using the data points $\{(\log(-\log(\widetilde{F}_n^*(Y_i))), Y_i) : \Delta_i = 1\}$, where $\widetilde{F}_n^* = \frac{n}{n+1}\widetilde{F}_n$.

3. Classify F as belonging to the Gumbel domain of attraction if the tail of the Gumbel probability plot appears linear with a slope of 1. Classify F as belonging to the Fréchet domain of attraction if otherwise.

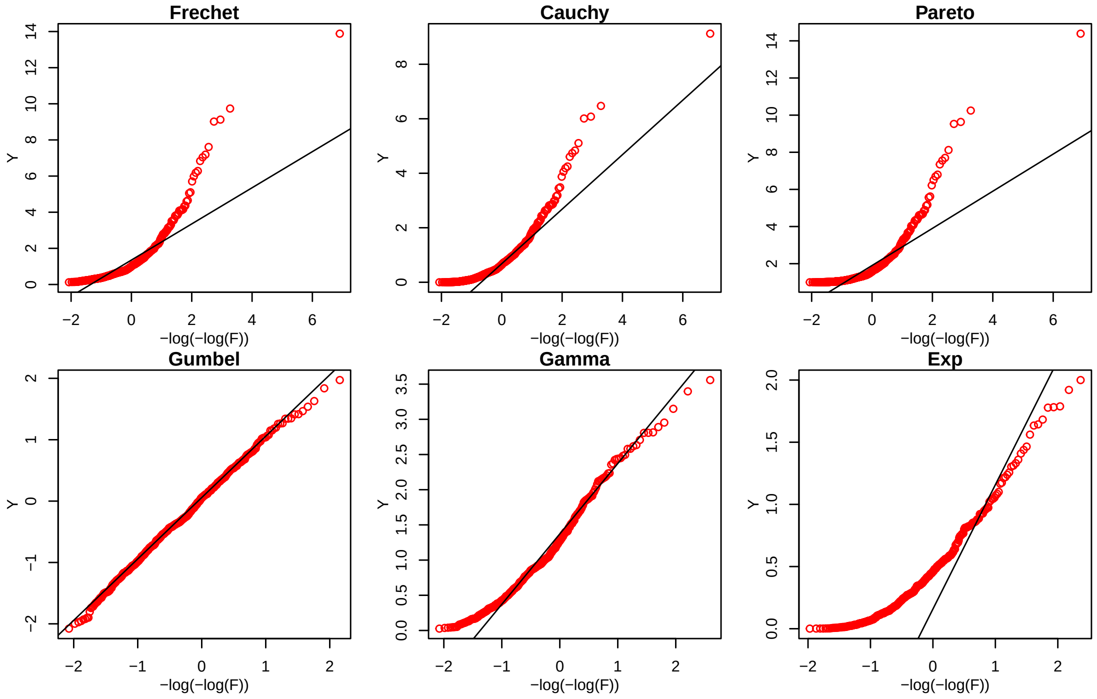

Figure 3.5: Gumbel probability plot at $\pi(\alpha) = 0.25$ and $r = 0.8$

Figure 3.5 presents the Gumbel probability plots under the setup described in Section 3.4, based on a single simulation. The straight line in the plot passes through the point $(-\log(-\log(\widetilde{F}_n^*(\widehat{q}_{90}))), \widehat{q}_{90})$ with a slope of 1, where $\widehat{q}_{90}$ denotes the 90-th empirical percentile of $\{Y_i : \Delta_i = 1\}$. In the Gumbel domain of attraction, the tails of the probability plot appear to be linear. In contrast, the tails in the Fréchet domain of attraction appear to be convex. To extend our work, we may develop a more comprehensive classification procedure and evaluate its accuracy through simulation in future work.

3.6 Appendix: Proof

Proof of Lemma 1.

Proof of (3.5). The expression can be bounded by

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n} I(Y_{ni} \geq \tau_n)Q(Y_{ni}, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \mathbb{P}_n[I(\widetilde{Y} = \infty)Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)]\right| \\ &\quad + \left|\mathbb{P}_n[I(\widetilde{Y} = \infty)Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)] - \mathbf{P}[I(\widetilde{Y} = \infty)Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)]\right|.\end{aligned}$$

Note that $\{I(\widetilde{Y} = \infty)\}$ is a Glivenko-Cantelli class. Under the algebraic operation of multiplication, the class $\{I(\widetilde{Y} = \infty)Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda) : \boldsymbol{\beta} \in \mathcal{B}, \Lambda \in \mathcal{L}\}$ is also Glivenko-Cantelli. Thus, the second term converges to 0 with probability 1. For the first term, we rewrite it as

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(\min(\widetilde{Y}_i, \tau_n) \geq \tau_n)Q(\min(\widetilde{Y}_i, \tau_n), \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{Y}_i = \infty)Q(\widetilde{Y}_i, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda)\right|.$$

Since $\min(\widetilde{Y}_i, \tau_n) \geq \tau_n$ if and only if $\widetilde{Y}_i \geq \tau_n$ and given that $Q(\tau_n, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda) = Q(\infty, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)$, and $Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)$ is bounded by some constant $M > 0$, we can express this as

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n} I(\widetilde{Y}_i \geq \tau_n)Q(\infty, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{Y}_i = \infty)Q(\infty, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda)\right| \\ &\quad = \left|\frac{1}{n}\sum_{i=1}^{n} I(\tau_n \leq \widetilde{Y}_i < \infty)Q(\infty, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda)\right| \\ &\quad \leq M\left|\frac{1}{n}\sum_{i=1}^{n} I(\tau_n \leq \widetilde{Y}_i < \infty)\right|,\end{aligned}$$

Let $p_n = P(\tau_n \leq \widetilde{Y} < \infty)$. By the Hoeffding's inequality, for any $\varepsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(\tau_n \leq \widetilde{Y}_i < \infty) - p_n\right| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2).$$

Since $p_n \to 0$ and

$$\sum_{n=1}^{\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I(\tau_n \leq \widetilde{Y}_i < \infty) - p_n\right| \geq \epsilon\right) \leq \sum_{n=1}^{\infty} 2\exp(-2n\epsilon^2) = \frac{2\exp(-2\epsilon^2)}{1-\exp(-2\epsilon^2)} < \infty,$$

by the Borel-Cantelli Lemma, we have $n^{-1}\sum_{i=1}^{n} I(\tau_n \leq \widetilde{Y}_i < \infty)$ converges to 0 with probability 1. Hence, we finished the proof of (3.5).

Proof of (3.6). Notice that the supremum is bounded by

$$\sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^{n} I(Y_{ni} \geq \tau_n)Q(Y_{ni}, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \mathbf{P}[I(\widetilde{Y} = \infty)Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)\right|$$
$$+ \sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^{n} I(\tau_n > Y_{ni} \geq y)Q(Y_{ni}, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \mathbf{P}[I(\infty > \widetilde{Y} \geq y)Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)\right|.$$

By (3.5), the first term converges to 0 with probability 1. For the second term, the class $\{I(\infty > \widetilde{Y} \geq y)Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda) : y \in [0,\infty), \boldsymbol{\beta} \in \mathcal{B}, \Lambda \in \mathcal{L}\}$ is a Glivenko-Cantelli and therefore, it suffices to show that $\sup_{y\in[0,\infty)} |n^{-1}\sum_{i=1}^{n} I(\tau_n > Y_{ni} \geq y)Q(Y_{ni}, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \mathbb{P}_n[I(\infty > \widetilde{Y} \geq y)Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)]|$ converges to 0. We can show that this supremum is bounded above by

$$\sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^{n} I(\tau_n \leq \widetilde{Y}_i < \infty)I(\widetilde{Y}_i \geq y)M\right| \leq M\left|\frac{1}{n}\sum_{i=1}^{n} I(\tau_n \leq \widetilde{Y}_i < \infty)\right|.$$

Applying argument used in the proof (3.5), we complete the proof of (3.6).

Proof of (3.7). Since $\{I(\widetilde{Y} \leq y)\widetilde{\Delta}Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda) : y \in [0,\infty), \boldsymbol{\beta} \in \mathcal{B}, \Lambda \in \mathcal{L}\}$ is a Glivenko-Cantelli class, it suffices to show that the expression $\sup_{y\in[0,\infty)} |n^{-1}\sum_{i=1}^{n} I(Y_{ni} \leq y)\Delta_{ni}Q(Y_{ni}, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda) - \mathbb{P}_n[I(\widetilde{Y} \leq y)\widetilde{\Delta}Q(\widetilde{Y}, \boldsymbol{X}; \boldsymbol{\beta}, \Lambda)]|$ converges to 0. Note that we can write the expression as

$$\sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^{n} I(\min(\widetilde{Y}_i, \tau_n) \leq y)\widetilde{\Delta}_i I(\widetilde{Y}_i \leq \tau_n)Q(\min(\widetilde{Y}_i, \tau_n), \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda)\right.$$

$$
\begin{aligned}
&- \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{Y}_i \le y)\widetilde{\Delta}_i I(\widetilde{Y}_i < \infty)Q(\widetilde{Y}_i, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda)\Bigg| \\
&= \sup_{y\in[0,\infty)} \left| \frac{1}{n}\sum_{i=1}^{n} I(\widetilde{Y}_i \le y)\widetilde{\Delta}_i I(\tau_n < \widetilde{Y}_i < \infty)Q(\widetilde{Y}_i, \boldsymbol{X}_i; \boldsymbol{\beta}, \Lambda)\right| \\
&\le M \left| \frac{1}{n}\sum_{i=1}^{n} I(\tau_n < \widetilde{Y}_i < \infty)\right|.
\end{aligned}
$$

Applying argument used in the proof (3.5), we obtain the desired result.

Proof of (3.8). (3.8) is a special case of (3.7), and the argument is similar. Thus, we omit the proof.

Proof of Theorem 1. To begin, we establish that the NPMLE of the Cox model $(\widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n)$ is a consistent estimator for $(\boldsymbol{\beta}_0, \Lambda_0)$, where $\Lambda_0 = \exp(\alpha_0)F_0$. The proof is structured in three steps:

(a) The NPMLE $(\widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n)$ exists.

(b) $\widehat{\Lambda}_n(\infty)$ is uniformly bounded with probability 1.

(c) For any convergent subsequence $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}^*$ and $\widehat{\Lambda}_n \to \Lambda^*$, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\Lambda^* = \Lambda_0$.

Consequently, the strong consistency result follows from the continuous mapping theorem applied to the transformation $(\widehat{\alpha}_n, \widehat{\boldsymbol{\beta}}_n, \widehat{F}_n) = (\log \widehat{\Lambda}_n(\infty), \widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n/\widehat{\Lambda}_n(\infty))$.

Proof of (a): (Existence of NPMLE). The log-likelihood of the Cox model is

$$
\ell_n(\boldsymbol{\beta}, \Lambda) = \sum_{i=1}^{n} \Delta_{ni}\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i + \log \Lambda\{Y_{ni}\}\} - \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i)\Lambda(Y_{ni}),
$$

which is continuous in $\boldsymbol{\beta}$ and the jump sizes of Λ. Since $\boldsymbol{\beta}$ belongs to a compact set $\mathcal{B}$, it suffices to show that the jump sizes $\Lambda\{Y_{ni}\}$ must be finite. Since the log-likelihood remains finite when all jump sizes are finite and it diverges to negative infinity if any jump size becomes infinity, the maximum must be attained at finite jump sizes.

Proof of (b): (Uniformly boundedness). The jump size can be expressed as

$$\widehat{\Lambda}_n\{Y_{ni}\} = \frac{\Delta_{ni}}{\sum_{j=1}^{n} R_{nj}(Y_{ni}) \exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}} \boldsymbol{X}_j)},$$

where $R_{ni}(y) = I(\tau_n > Y_{ni} \geq y) + I(Y_{ni} \geq \tau_n)$. Here, we are essentially treating the subjects censored at τ_n as cured and hence, we use $R_{ni}(y)$ instead of the usual indicator function $I(Y_{ni} \geq y)$. For any fixed $\boldsymbol{\beta} \in \mathcal{B}$ and $y \in (0, \infty)$, we have

$$\frac{1}{n}\sum_{j=1}^{n} R_{nj}(y) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_j) \leq \frac{1}{n}\sum_{j=1}^{n} I(Y_{nj} \geq \tau_n) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_j).$$

Since $\{\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}) : \boldsymbol{\beta} \in \mathcal{B}\}$ is a bounded, Glivenko–Cantelli class, Lemma 1 implies that

$$\frac{1}{n}\sum_{j=1}^{n} I(Y_{nj} \geq \tau_n) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_j) \to \mathbf{P}[I(\widetilde{Y} = \infty) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X})].$$

Note that the expectation is bounded below by

$$\mathbf{P}[I(\widetilde{Y} = \infty) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X})] \geq \inf_{\boldsymbol{\beta} \in \mathcal{B}} \mathbf{P}[I(\widetilde{Y} = \infty) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X})] > 0,$$

where the positivity follows from the compactness of $\mathcal{B}$ and the assumption that the cure fraction is non-zero. Consequently, for sufficiently large n, $n^{-1}\sum_{j=1}^{n} R_{nj}(y) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_j)$ is bounded below by $2^{-1} \inf_{\boldsymbol{\beta} \in \mathcal{B}} \mathbf{P}[I(\widetilde{Y} = \infty) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X})]$. This implies for large n

$$\widehat{\Lambda}_n(\infty) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Delta_{ni}}{n^{-1}\sum_{j=1}^{n} R_{nj}(Y_{ni}) \exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}} \boldsymbol{X}_j)} \leq \frac{2}{\inf_{\boldsymbol{\beta} \in \mathcal{B}} \mathbf{P}[I(\widetilde{Y} = \infty) \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X})]} = B_0.$$

Hence, $\widehat{\Lambda}_n(\infty)$ is uniformly bounded.

Proof of (c): (Consistency). To establish consistency, we show that for every subsequence of $(\widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n)$, there is a further subsequence that converges to $(\boldsymbol{\beta}_0, \Lambda_0)$. For any subsequence of $(\widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n)$, by the Helly selection theorem, we can find a further converging subsequence. With an abuse of notation, we use the same subscript n to denote the subsequence. Suppose that $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}^*$ and $\widehat{\Lambda}_n \to \Lambda^*$. To complete the proof, we first construct a consistent estimator $\overline{\Lambda}_n$ by fixing $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Then we show that $n^{-1}\{\ell_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n) - \ell_n(\boldsymbol{\beta}_0, \overline{\Lambda}_n)\}$ converges to the negative Kullback–Leibler divergence. Since the likelihood function is identifiable, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\Lambda^* = \Lambda_0$.

Define $\overline{\Lambda}_n$ as a step function that jumps only at the observed event times, with the jump size at Y_{ni} equals

$$\overline{\Lambda}_n\{Y_{ni}\} = \frac{\Delta_{ni}}{\sum_{j=1}^n R_{nj}(Y_{ni})\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_j)}.$$

We claim that

$$\overline{\Lambda}_n(t) = \frac{1}{n}\sum_{i=1}^n \frac{I(Y_{ni} \le t)\Delta_{ni}}{n^{-1}\sum_{j=1}^n R_{nj}(Y_{ni})\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_j)} \to \mathrm{E}\left[\frac{I(\widetilde{Y} \le t)\widetilde{\Delta}}{\mathbf{P}[I(\widetilde{Y} \ge y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]_{y=\widetilde{Y}}}\right] = \Lambda_0(t).$$

To prove our claim, note that

$$\begin{aligned}
&\sup_{t\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^n \frac{I(Y_{ni} \le t)\Delta_{ni}}{n^{-1}\sum_{j=1}^n R_{nj}(Y_{ni})\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_j)} - \mathrm{E}\left[\frac{I(\widetilde{Y} \le t)\widetilde{\Delta}}{\mathbf{P}[I(\widetilde{Y} \ge y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]_{y=\widetilde{Y}}}\right]\right| \\
&\quad \le \sup_{t\in[0,\infty)}\left|\frac{1}{n^{-1}\sum_{j=1}^n R_{nj}(t)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_j)} - \frac{1}{\mathbf{P}[I(\widetilde{Y} \ge t)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]}\right| \\
&\qquad + \sup_{t\in[0,\infty)}\left|\frac{1}{n}\sum_{i=1}^n \frac{I(Y_{ni} \le t)\Delta_{ni}}{\mathbf{P}[I(\widetilde{Y} \ge y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]_{y=Y_{ni}}} - \mathbf{P}\left[\frac{I(\widetilde{Y} \le t)\widetilde{\Delta}}{\mathbf{P}[I(\widetilde{Y} \ge y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]_{y=\widetilde{Y}}}\right]\right|
\end{aligned}$$

The class $\{\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})\}$ is a bounded, Glivenko–Cantelli class. By Lemma 1, we obtain that $\sup_{t\in[0,\infty)} |n^{-1}\sum_{j=1}^n R_{nj}(t)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_j) - \mathbf{P}[I(\widetilde{Y} \ge t)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]|$ converges to 0. Additionally, the expectation $\mathbf{P}[I(\widetilde{Y} \ge t)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]$ is bounded below by $\inf_{\boldsymbol{\beta}\in\mathcal{B}} \mathbf{P}[I(\widetilde{Y} = \infty)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})] > 0$. Consequently, the first term converges to 0 as the transformation $1/x$ is Lipschitz on the domain $[c, \infty)$ for $c > 0$. Since $\{1/\mathbf{P}[I(\widetilde{Y} \ge y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]|_{y=\widetilde{Y}}\}$ is also a bounded, Glivenko–Cantelli class as it is Lipschitz transformation of a Glivenko–Cantelli class. By Theorem 1, the second term converges to 0.

From the construction of $\overline{\Lambda}_n$, we obtain that

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{n^{-1}\sum_{j=1}^n R_{nj}(y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_j)}{n^{-1}\sum_{j=1}^n R_{nj}(y)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X}_j)}\,\mathrm{d}\overline{\Lambda}_n(y),$$

where $\widehat{\Lambda}_n$ is absolutely continuous with respect to $\overline{\Lambda}_n$. We consider taking limit in both sides. Note that

$$\begin{aligned}
&\sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{j=1}^n R_{nj}(y)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X}_j) - \mathbf{P}[I(\widetilde{Y}\geq y)\exp(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})]\right| \\
&\quad\leq \sup_{y\in[0,\infty)}\left|\frac{1}{n}\sum_{j=1}^n R_{nj}(y)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X}_j) - \mathbb{P}_n[I(\widetilde{Y}\geq y)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X})]\right| \\
&\qquad + \sup_{y\in[0,\infty)}\left|\mathbb{P}_n[I(\widetilde{Y}\geq y)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X})] - \mathbf{P}[I(\widetilde{Y}\geq y)\exp(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})]\right|.
\end{aligned}$$

Applying similar argument used in the proof of Theorem 1, we can show that the first term converges to 0. Since $\{I(\widetilde{Y}\geq y)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}) : y\in[0,\infty), \boldsymbol{\beta}\in\mathcal{B}\}$ is a Glivenko–Cantelli class and $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}^*$, by the monotone convergence theorem, the second term converges to 0. Hence, it holds that uniformly in y,

$$\frac{\widehat{\Lambda}_n\{y\}}{\overline{\Lambda}_n\{y\}} = \frac{n^{-1}\sum_{j=1}^n R_{nj}(y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_j)}{n^{-1}\sum_{j=1}^n R_{nj}(y)\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X}_j)} \to \frac{\mathbf{P}[I(\widetilde{Y}\geq y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]}{\mathbf{P}[I(\widetilde{Y}\geq y)\exp(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})]} = \frac{\lambda^*(y)}{\lambda_0(y)}.$$

From the aforementioned argument, we conclude that

$$\Lambda^*(t) = \int_0^t \frac{\mathbf{P}[I(\widetilde{Y}\geq y)\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})]}{\mathbf{P}[I(\widetilde{Y}\geq y)\exp(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})]}\,\mathrm{d}\Lambda_0(t).$$

The difference between the log-likelihood is

$$\begin{aligned}
n^{-1}\{\ell_n(\widehat{\boldsymbol{\beta}}_n,\widehat{\Lambda}_n) - \ell_n(\boldsymbol{\beta}_0,\overline{\Lambda}_n)\} &= \frac{1}{n}\sum_{i=1}^n \Delta_{ni}\log\left(\frac{\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X}_i)\widehat{\Lambda}_n\{Y_{ni}\}}{\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_i)\overline{\Lambda}_n\{Y_{ni}\}}\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \log\left(\frac{\exp\{-\exp(\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}\boldsymbol{X}_i)\widehat{\Lambda}_n(Y_{ni})\}}{\exp\{-\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_i)\overline{\Lambda}_n(Y_{ni})\}}\right)
\end{aligned}$$

$$\geq 0.$$

Take limit on both sides and applying the Glivenko-Cantelli theorem, Theorem 1, and the bounded convergence theorem, we have

$$\begin{aligned} n^{-1}\{\ell_n(\widehat{\boldsymbol{\beta}}_n, \widehat{\Lambda}_n) - \ell_n(\boldsymbol{\beta}_0, \overline{\Lambda}_n)\} \to & \mathrm{E}\left[\widetilde{\Delta} I(\widetilde{Y} < \infty) \log\left(\frac{\exp(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})\lambda^*(\widetilde{Y})}{\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})\lambda_0(\widetilde{Y})}\right)\right] \\ & + \mathrm{E}\left[\log\left(\frac{\exp\{-\exp(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X}_i)\Lambda^*(\widetilde{Y})\}}{\exp\{-\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X}_i)\Lambda_0(\widetilde{Y})\}}\right)\right], \end{aligned}$$

which is the negative Kullback–Leibler divergence. This implies, with probability 1,

$$\begin{aligned} & \{\exp(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})\lambda^*(\widetilde{Y})\}^{\widetilde{\Delta} I(\widetilde{Y}<\infty)} \exp\{-\exp(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})\Lambda^*(\widetilde{Y})\} \\ & \qquad = \{\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})\lambda_0(\widetilde{Y})\}^{\widetilde{\Delta} I(\widetilde{Y}<\infty)} \exp\{-\exp(\boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{X})\Lambda_0(\widetilde{Y})\}. \end{aligned}$$

By the identifiability of model, we obtain $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\Lambda^* = \Lambda_0$. By the continuous mapping theorem, we obtain the following with probability 1:

$$\begin{aligned} \widehat{\alpha}_n = \log\widehat{\Lambda}_n(\infty) &\to \log\Lambda_0(\infty) = \alpha_0 \\ \widehat{\boldsymbol{\beta}}_n &\to \boldsymbol{\beta}_0 \\ \widehat{F}_n = \widehat{\Lambda}_n/\widehat{\Lambda}_n(\infty) &\to \Lambda_0/\Lambda_0(\infty) = F_0. \end{aligned}$$

Chapter 4

Conclusion

4.1 Summary of the thesis

This thesis developed methodologies for survival analysis under two critical challenges: missing data and insufficient follow-up.

In Chapter 2, we handled missing covariates in the Cox proportional hazards model by proposing a novel transformation technique within the EM algorithm. By reducing the computational complexity of the E-step to one-dimensional integration, our method enables scalable estimation as the dimensionality of missing variables increases. We further extended this framework to penalized regression settings and validated the numerical performance of the NPMLE through simulations.

In Chapter 3, we tackled insufficient follow-up in survival data by developing an extrapolation-based cure probability estimator grounded in extreme value theory. Large-scale simulations demonstrated that our estimator outperforms the NPMLE in terms of efficiency when follow-up is severely insufficient. Additionally, we established preliminary asymptotic results.

4.2 Future research directions

For Chapter 2, we may consider shrinkage estimators for high-dimensional covariance estimation, which may help extend our method to high-dimensional settings. Also, we may relax the Gaussian assumption to accommodate categorical or mixed-type covariates via latent variable models. In addition, we may evaluate the performance of inverse probability weighting, multiple imputation, and our EM-based approach via simulations.

For Chapter 3, we may formalize the the superior efficiency of the proposed estimator over the NPMLE based on the preliminary results in Section 3.3. Also, we may investigate the empirical performance of the proposed domain-of-attraction classification rule under the setup described in Section 3.4. In addition, we may extend our extrapolation method to accommodate the Weibull domain of attraction.

References

Amico, M., Van Keilegom, I., and Legrand, C. (2019). The single-index/Cox mixture cure model. *Biometrics*, 75:452–462.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20:40–49.

Beyhum, J., El Ghouch, A., Portier, F., and Van Keilegom, I. (2022). On an extension of the promotion time cure model. *The Annals of Statistics*, 50:537–559.

Bhati, D. and Ravi, S. (2018). Diagnostic plots for identifying max domains of attraction under power normalization. *Journal of Applied Statistics*, 45:2394–2410.

Botta, L., Gatta, G., Capocaccia, R., Stiller, C., Cañete, A., Dal Maso, L., Innos, K., Mihor, A., Erdmann, F., Spix, C., Lacour, B., Rafael, M.-G., Murray, D., and Rossi, S. (2022). Long-term survival and cure fraction estimates for childhood cancer in europe (eurocare-6): results from a population-based study. *The Lancet Oncology*, 23:1525–1536.

Castillo, E., Galambos, J., and Sarabia, J. M. (1989). The selection of the domain of attraction of an extreme value distribution from a set of data. *In Extreme Value Theory. Lecture Notes in Statistics*, 51:181–190.

Chen, M. H., Ibrahim, J. G., and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94:909–919.

Cohen, J. and Cohen, P. (1975). *Applied Multiple Regression Correlation Analysis for the Behavioral Sciences*. John Wiley, New York.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–22.

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

Deng, Y. and Lumley, T. (2024). Multiple imputation through XGBoost. *Journal of Computational and Graphical Statistics*, 33:352–363.

Escobar-Bach, M., Maller, R., Van Keilegom, I., and Zhao, M. (2022). Estimation of the cure rate for distributions in the Gumbel maximum domain of attraction under insufficient follow-up. *Biometrika*, 109:243–256.

Escobar-Bach, M. and Van Keilegom, I. (2019). Non-parametric cure rate estimation under insufficient follow-up by using extremes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81:861–880.

Escobar-Bach, M. and Van Keilegom, I. (2023). Nonparametric estimation of conditional cure models for heavy-tailed distributions and under insufficient follow-up. *Computational Statistics & Data Analysis*, 183:107728.

Falk, M. and Marohn, F. (1993). von Mises conditions revisited. *The Annals of Probability*, 21:1310–1328.

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

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

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046.

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

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247:2543–2546.

Herring, A. H. and Ibrahim, J. G. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association*, 96:292–302.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:1–9.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103:672–680.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91:222–230.

Kosinski, M., Biecek, P., and Chodor, W. (2016). RTCGA: The Cancer Genome Atlas Data Integration. R package version 1.34.0. `https://rtcga.github.io/RTCGA/`.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.

Lee, C. Y., Wong, K. Y., and Bandyopadhyay, D. (2024). Partly linear single-index cure models with a nonparametric incidence link function. *Statistical Methods in Medical Research*, 33:498–514.

Li, P., Peng, Y., Jiang, P., and Dong, Q. (2020). A support vector machine based semiparametric mixture cure model. *Computational Statistics*, 35:931–945.

Liang, L., Zhuang, Y., and Yu, P. L. H. (2024). Variable selection for high-dimensional incomplete data. *Computational Statistics & Data Analysis*, 192:107877.

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

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*, 20:661–674.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, 91:331–343.

Ma, Y. and Yin, G. (2008). Cure rate model with mismeasured covariates under transformation. *Journal of the American Statistical Association*, 103:743–756.

Maller, R. A. and Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data. *Journal of the American Statistical Association*, 89:1499–1506.

Mishchenko, K. (2023). Regularized newton method with global convergence. *SIAM Journal on Optimization*, 33:1440–1462.

Othus, M., Barlogie, B., LeBlanc, M. L., and Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, 18:3731–3736.

Papageorgiou, G., Mauff, K., Tomer, A., and Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and its Application*, 6:223–240.

Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56:237–243.

Peng, Y. and Yu, B. (2021). *Cure Models: Methods, Applications, and Implementation.* Chapman and Hall/CRC, Boca Raton, Florida.

Portier, F., El Ghouch, A., and Van Keilegom, I. (2017). Efficiency and bootstrap in the promotion time cure model. *Bernoulli*, 23:3437–3468.

Romain, G., Boussari, O., Bossard, N., Remontet, L., Bouvier, A.-M., Mounier, M., Iwaz, J., Colonna, M., Jooste, V., and French Network of Cancer Registries (FRANCIM) (2019). Time-to-cure and cure proportion in solid cancers in france. a population based study. *Cancer Epidemiology*, 60:93–101.

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

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

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

Sun, J., Herazo-Maya, J. D., Molyneaux, P. L., Maher, T. M., Kaminski, N., and Zhao, H.

(2019). Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics*, 75:69–77.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56:227–236.

The Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499:43–49.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 58:267–288.

Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, 54:1508–1516.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–67.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102:1039–1048.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, 103:340–349.

Wolfson, J. (2011). EEBoost: A general method for prediction and variable selection based on estimating equations. *Journal of the American Statistical Association*, 106:296–305.

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

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

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

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

Wu, Y. and Yin, G. (2013). Cure rate quantile regression for censored data with a survival fraction. *Journal of the American Statistical Association*, 108:1517–1531.

Wu, Y. and Yin, G. (2017a). Cure rate quantile regression accommodating both finite and infinite survival times. *Canadian Journal of Statistics*, 45:29–43.

Wu, Y. and Yin, G. (2017b). Multiple imputation for cure rate quantile regression with censored data. *Biometrics*, 73:94–103.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications*. World Scientific, Hackensack, New Jersey.

Zeng, D., Yin, G., and Ibrahim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, 101:670–684.

Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine*, 26:3157–3171.

Zhao, A. and Ding, P. (2024). To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates. *Journal of the American Statistical Association*, 119:450–460.

Zhao, A., Ding, P., and Li, F. (2024). Covariate adjustment in randomized experiments with missing outcomes and covariates. *Biometrika*, 111:1413–1420.

Zhao, Q., Shi, X., Xie, Y., Huang, J., Shia, B., and Ma, S. (2015). Combining multidimensional genomic measurements for predicting cancer prognosis: Observations from TCGA. *Briefings in Bioinformatics*, 16:291–303.

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

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