

Statistica Sinica Preprint No: SS-2026-0289	
Title	A General Approach to Maximum Likelihood Estimation of Semiparametric Models with Massive Sample Sizes
Manuscript ID	SS-2026-0289
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0289
Complete List of Authors	Jianqiao Wang, Donglin Zeng and Danyu Lin
Corresponding Authors	Danyu Lin
E-mails	lin@bios.unc.edu
Notice: Accepted author version.	

A General Approach to Maximum Likelihood Estimation of Semiparametric Models with Massive Sample Sizes

Jianqiao Wang¹, Donglin Zeng² and Dan-Yu Lin¹

¹*University of North Carolina at Chapel Hill*, ²*University of Michigan*

Abstract: We propose a computationally efficient approach to performing maximum likelihood estimation of semiparametric models with big data containing potentially millions of observations. We calculate an initial estimator for the vector of finite-dimensional parameters of primary interest using a small fraction of the data and improve its efficiency by one-step estimation with semiparametric efficient score functions on the remaining data. We estimate the efficient scores by the numerical differences of the profile log-likelihood function, eliminating the need for an analytic form of these scores, which is often unavailable for semiparametric problems. We show that the resulting estimator is consistent and asymptotically normal with a covariance matrix that attains the semiparametric efficiency bound. Because the majority of its computing time is spent in calculating the initial estimator using a small subset of data, the proposed approach substantially reduces computational burden compared to the standard method of performing maximum likelihood estimation on the entire dataset. We evaluate the performance of the proposed methods through extensive simulation studies on fitting the Cox proportional hazards model to an interval-censored failure time and the partially linear model to a continuous outcome. Finally, we apply the proposed methods to interval-censored data from the UK Biobank.

Key words and phrases: Cox model, Efficient score, Interval-censored data, One-step estimation, Partially linear model, Profile likelihood.

1. Introduction

Semiparametric models, which are characterized by finite-dimensional parameters of primary interest and possibly infinite-dimensional nuisance parameters, are widely used in scientific studies. An important example is the Cox (1972) proportional hazards model, which includes a vector of regression parameters to quantify the effects of covariates on a possibly censored failure time and an arbitrary, infinite-dimensional baseline hazard function. A second example is the partially linear model, which uses a vector of regression parameters to capture the linear effects of some covariates on a continuous outcome and nonparametric functions for the nonlinear effects of other covariates. Thus far, semiparametric models have been mostly applied to small-scale studies.

Recent years have witnessed a proliferation of large-scale data systems, such as the UK Biobank (Bycroft et al., 2018), the Million Veteran Program (Gaziano et al., 2016), and the China Kadoorie Biobank (Chen et al., 2011). These databases, which involve hundreds of thousands to millions of participants, have significantly advanced our understanding of human health (Bi et al., 2020; Dey et al., 2022). However, the immense size of such data presents enormous computational challenges in fitting semiparametric models, as the number of parameters associated with the infinite-dimensional component increases with the sample size.

One approach to tackling these computational challenges is to randomly divide the entire dataset into several subsets, analyze each subset separately, and combine the resulting estimators. This so-called divide-and-conquer approach has been suggested for parametric models (Zhang et al., 2013; Neiswanger et al., 2014; Schifano et al., 2016; Lee et al., 2017) and some semiparametric problems such as the partially linear model for a continuous

response variable (Zhao et al., 2016) and the Cox proportional hazards model for a right-censored failure time (Wang et al., 2021, 2022), but it has not been investigated for general semiparametric models. The divide-and-conquer approach is computationally efficient under the distributed computing architecture but requires separate estimation of the nuisance parameter for each subset of data, which may reduce statistical efficiency or cause numerical instability.

In this paper, we propose a computationally efficient approach for fitting general semiparametric models to big data. Specifically, we divide the entire dataset into two subsets, with the first subset being much smaller than the second. We perform maximum profile likelihood estimation (Murphy and van der Vaart, 2000) on the first subset of data and then update the initial estimator by using one-step estimation with semiparametric efficient scores on the remaining data. These scores are estimated by the derivatives or numerical differences of the profile log-likelihood function, so the approach applies to a wide variety of semiparametric problems, even when an analytical expression of the score function is not available. The proposed approach requires only a small fraction of the computation time that is required to maximize the likelihood over the entire dataset. Using modern empirical process theory (van der Vaart and Wellner, 1996), we show that the proposed estimator is consistent and asymptotically normal with a covariance matrix that attains the semiparametric efficiency bound.

2. Methods

2.1 Maximum profile likelihood estimation

We consider a random sample of n observations $\mathbf{O}_1, \dots, \mathbf{O}_n$ with distribution $\mathbb{P}_{\boldsymbol{\theta}_0, \eta_0}$, where $\{\mathbb{P}_{\boldsymbol{\theta}, \eta}; \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p, \eta \in \mathcal{F}\}$ is a family of statistical models parameterized by $(\boldsymbol{\theta}, \eta)$, $\boldsymbol{\theta}$ is a p -vector of parameters of primary interest in the space Θ , η is a possibly infinite-dimensional nuisance parameter in the space $\mathcal{F}$, and $\boldsymbol{\theta}_0$ and η_0 are the true values of $\boldsymbol{\theta}$ and η . The standard method for estimating $\boldsymbol{\theta}$ is to maximize the likelihood with the entire dataset. Let $l(\mathbf{O}; \boldsymbol{\theta}, \eta)$ denote the log-likelihood function of a single observation $\mathbf{O}$. The profile log-likelihood function with $(\mathbf{O}_1, \dots, \mathbf{O}_n)$ is defined as

$$pl_n(\boldsymbol{\theta}) = \max_{\eta \in \mathcal{F}_n} \sum_{i=1}^n l(\mathbf{O}_i; \boldsymbol{\theta}, \eta),$$

where $\mathcal{F}_n$ is a sequence of approximation spaces to $\mathcal{F}$. The maximizer of $pl_n(\boldsymbol{\theta})$, denoted by $\hat{\boldsymbol{\theta}}_n$, is the maximum likelihood estimator of $\boldsymbol{\theta}$.

Under appropriate regularity conditions, $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mathbf{I}_0^{-1}$, where $\mathbf{I}_0$ is the efficient information matrix for $\boldsymbol{\theta}$ (Murphy and van der Vaart, 2000). The limiting covariance matrix of $\hat{\boldsymbol{\theta}}_n$ can be consistently estimated by $\hat{\mathbf{I}}_n^{-1}$, where $\hat{\mathbf{I}}_n$ is the negative second-order derivative of $pl_n(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}_n$. If the derivative does not have an analytical expression, we approximate the (j, k) th element of $\hat{\mathbf{I}}_n$ by

$$-\frac{pl_n(\hat{\boldsymbol{\theta}}_n) - pl_n(\hat{\boldsymbol{\theta}}_n + h_n \mathbf{e}_j) - pl_n(\hat{\boldsymbol{\theta}}_n + h_n \mathbf{e}_k) + pl_n(\hat{\boldsymbol{\theta}}_n + h_n \mathbf{e}_j + h_n \mathbf{e}_k)}{h_n^2},$$

where $\mathbf{e}_j$ is the j th canonical vector in $\mathbb{R}^p$, and h_n is a constant of order $n^{-1/2}$.

For some semiparametric models with certain data structures, such as the partially linear model for a continuous response and the Cox proportional hazards model for a right-censored

failure time, there are explicit expressions for the profile likelihood. For most semiparametric problems, however, the structure of the data or the nature of the underlying distribution does not allow the derivation of an explicit expression for the profile likelihood. The absence of an explicit formula makes it computationally challenging to estimate $\boldsymbol{\theta}$ and conduct statistical inference. This challenge is significantly amplified with big data.

2.2 One-step efficient score estimator

To motivate our approach, we first consider parametric models, under which η is finite-dimensional and $l(\mathbf{O}; \boldsymbol{\theta}, \eta)$ is a smooth function. We let $\dot{l}_{\boldsymbol{\theta}}(\mathbf{O}; \boldsymbol{\theta}, \eta)$, $\dot{l}_{\eta}(\mathbf{O}; \boldsymbol{\theta}, \eta)$, $\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\mathbf{O}; \boldsymbol{\theta}, \eta)$, $\ddot{l}_{\boldsymbol{\theta}\eta}(\mathbf{O}; \boldsymbol{\theta}, \eta)$ and $\ddot{l}_{\eta\eta}(\mathbf{O}; \boldsymbol{\theta}, \eta)$ denote the first- and second-order derivatives of $l(\mathbf{O}; \boldsymbol{\theta}, \eta)$.

We divide the entire dataset into two subsets, indexed by $\mathcal{J}_1$ and $\mathcal{J}_2$, with m and $n - m$ observations in $\mathcal{J}_1$ and $\mathcal{J}_2$, respectively.

Let $\mathbb{P}_m$ and $\mathbb{P}_n$ denote the empirical distributions of the observations in $\mathcal{J}_1$ and in the entire dataset, respectively. In addition, let E denote the expectation under the true underlying measure. Finally, we use the operator notation for the expectation, such that $Pf = \int fdP$ for every measurable function f and a probability measure P .

We maximize the likelihood with the data in $\mathcal{J}_1$. Let $\tilde{\boldsymbol{\theta}}_m$ and $\tilde{\eta}_m$ denote the resulting estimators for $\boldsymbol{\theta}$ and η , respectively. By the standard maximum likelihood estimation theory, the covariance matrix of $\tilde{\boldsymbol{\theta}}_m$ can be consistently estimated by $\tilde{\mathbf{I}}_m^{-1}$, where

$$\tilde{\mathbf{I}}_m = -m\mathbb{P}_m\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) + m\mathbb{P}_m\ddot{l}_{\boldsymbol{\theta}\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) \left\{ \mathbb{P}_m\ddot{l}_{\eta\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) \right\}^{-1} \mathbb{P}_m\ddot{l}_{\eta\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m),$$

provided that $-\mathbb{P}_m\ddot{l}_{\eta\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m)$ and $\tilde{\mathbf{I}}_m$ are positive definite. We incorporate the data from the second block through one-step estimation (van der Vaart, 1998, pages 71–75) to improve the

efficiency of $\tilde{\boldsymbol{\theta}}_m$. The consistency of $\tilde{\boldsymbol{\eta}}_m$ is sufficient to establish that the one-step estimator is asymptotically efficient.

The efficient score function for $\boldsymbol{\theta}$ (Bickel et al., 1993, pages 28–30) based on a single observation $\mathbf{O}$ is

$$l_0(\mathbf{O}; \boldsymbol{\theta}, \eta_0) = \dot{l}_{\boldsymbol{\theta}}(\mathbf{O}; \boldsymbol{\theta}, \eta_0) - E\ddot{l}_{\boldsymbol{\theta}\eta}(\boldsymbol{\theta}_0, \eta_0) \left\{ E\ddot{l}_{\eta\eta}(\boldsymbol{\theta}_0, \eta_0) \right\}^{-1} \dot{l}_{\eta}(\mathbf{O}; \boldsymbol{\theta}, \eta_0).$$

We obtain an empirical counterpart $\tilde{l}_0(\mathbf{O}; \boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}_m)$ by replacing η_0 , $E\ddot{l}_{\boldsymbol{\theta}\eta}(\boldsymbol{\theta}_0, \eta_0)$ and $E\ddot{l}_{\eta\eta}(\boldsymbol{\theta}_0, \eta_0)$ in $l_0(\mathbf{O}; \boldsymbol{\theta}, \eta_0)$ with $\tilde{\boldsymbol{\eta}}_m$, $\mathbb{P}_m\ddot{l}_{\boldsymbol{\theta}\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_m)$ and $\mathbb{P}_m\ddot{l}_{\eta\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_m)$, respectively. In addition, let $\tilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\eta}} \mathbb{P}_m l(\boldsymbol{\theta}, \boldsymbol{\eta})$ for any fixed $\boldsymbol{\theta}$, and let $\dot{\tilde{\boldsymbol{\eta}}}_{\tilde{\boldsymbol{\theta}}_m}$ denote the first-order derivative of $\tilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$. We have

$$\tilde{l}_0(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_m) = \dot{l}_{\boldsymbol{\theta}}(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m}) + \dot{\tilde{\boldsymbol{\eta}}}_{\tilde{\boldsymbol{\theta}}_m}^T \dot{l}_{\eta}(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m}). \quad (2.1)$$

This result can be verified by noting that $\tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m}$ solves the equation

$$\mathbb{P}_m\ddot{l}_{\boldsymbol{\theta}\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m}) + \dot{\tilde{\boldsymbol{\eta}}}_{\tilde{\boldsymbol{\theta}}_m}^T \mathbb{P}_m\ddot{l}_{\eta\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m}) = 0,$$

which can be derived by differentiating the equation $\mathbb{P}_m \dot{l}_{\eta}(\boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}}) = 0$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}_m$. Of note, the right side of (2.1) is the total derivative of $l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}_m$. The one-step estimator for $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}_f = \tilde{\boldsymbol{\theta}}_m + \left(n\tilde{I}_m/m \right)^{-1} \sum_{i \in \mathcal{J}_2} \tilde{l}_0(\mathbf{O}_i; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_m).$$

In the special case where no nuisance parameter is present, $\tilde{l}_0(\mathbf{O}_i; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_m)$ and $\tilde{I}_m$ become $\dot{l}_{\boldsymbol{\theta}}(\mathbf{O}_i; \tilde{\boldsymbol{\theta}}_m)$ and $-m\mathbb{P}_m\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_m)$, and this estimator reduces to the one-step estimator proposed by Jordan et al. (2019).

We provide a brief explanation on the efficiency of $\hat{\boldsymbol{\theta}}_f$ as $n^{1/2}/m \rightarrow 0$. Clearly, $\mathbb{P}_m \tilde{l}_0(\tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_m) = 0$. By the Taylor series expansion,

$$(\tilde{I}_m/m)n^{1/2}(\hat{\boldsymbol{\theta}}_f - \boldsymbol{\theta}_0)$$

$$\begin{aligned}
&= n^{1/2} \mathbb{P}_n \mathbf{l}_0(\boldsymbol{\theta}_0, \eta_0) \\
&+ n^{1/2} \left[\tilde{\mathbf{I}}_m / m + \mathbb{P}_n \ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}_m, \bar{\eta}_m) \right. \\
&\quad \left. - \mathbb{P}_m \ddot{l}_{\boldsymbol{\theta}\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) \left\{ \mathbb{P}_m \ddot{l}_{\eta\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) \right\}^{-1} \mathbb{P}_n \ddot{l}_{\eta\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}_m, \bar{\eta}_m) \right] (\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0) \\
&- n^{1/2} \left[\mathbb{P}_m \ddot{l}_{\boldsymbol{\theta}\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) \left\{ \mathbb{P}_m \ddot{l}_{\eta\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) \right\}^{-1} \right. \\
&\quad \left. - E \ddot{l}_{\boldsymbol{\theta}\eta}(\boldsymbol{\theta}_0, \eta_0) \left\{ E \ddot{l}_{\eta\eta}(\boldsymbol{\theta}_0, \eta_0) \right\}^{-1} \right] \mathbb{P}_n \dot{l}_{\eta}(\boldsymbol{\theta}_0, \eta_0) \\
&+ n^{1/2} \left[\mathbb{P}_n \ddot{l}_{\boldsymbol{\theta}\eta}(\bar{\boldsymbol{\theta}}_m, \bar{\eta}_m) \right. \\
&\quad \left. - \mathbb{P}_m \ddot{l}_{\boldsymbol{\theta}\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) \left\{ \mathbb{P}_m \ddot{l}_{\eta\eta}(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m) \right\}^{-1} \mathbb{P}_n \ddot{l}_{\eta\eta}(\bar{\boldsymbol{\theta}}_m, \bar{\eta}_m) \right] (\tilde{\eta}_m - \eta_0), \tag{2.2}
\end{aligned}$$

where $(\bar{\boldsymbol{\theta}}_m, \bar{\eta}_m)$ is on the line segment between $(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_m)$ and $(\boldsymbol{\theta}_0, \eta_0)$. By the standard maximum likelihood estimation theory, $\|\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\| + \|\tilde{\eta}_m - \eta_0\| = O_P(m^{-1/2})$, where $\|\cdot\|$ denotes the Euclidean norm. By the central limit theorem, the last three terms on the right side of (2.2) are $O_P(n^{1/2}/m) + o_P(1)$. The efficiency of $\hat{\boldsymbol{\theta}}_f$ then follows from Slutsky's theorem as $n^{1/2}/m \rightarrow 0$.

The condition that $\|\tilde{\eta}_m - \eta_0\| = O_P(m^{-1/2})$ is often violated in the semiparametric setting. As long as the last three terms on the right side of (2.2) are $o_P(1)$, however, the asymptotic efficiency of the one-step estimator as $n \rightarrow \infty$ is guaranteed. In the semiparametric setting, the efficient score function for $\boldsymbol{\theta}$ often lacks an explicit formula, which makes the computation of the one-step estimator intractable. In addition, the Taylor series expansion for the Euclidean parameters and the central limit theorem cannot be used when η is infinite-dimensional, which makes it challenging to establish the semiparametric efficiency of the one-step estimator.

Motivated by (2.1), we approximate the semiparametric efficient score function for $\boldsymbol{\theta}$ by the total derivative of $l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\eta}_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$, where $\tilde{\eta}_{\boldsymbol{\theta}} = \arg \max_{\eta \in \mathcal{F}_m} \mathbb{P}_m l(\boldsymbol{\theta}, \eta)$ for

any fixed $\boldsymbol{\theta}$. The final estimator for $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}_f = \tilde{\boldsymbol{\theta}}_m + \left(n\tilde{\mathbf{I}}_m/m\right)^{-1} \sum_{i \in \mathcal{J}_2} D_{\boldsymbol{\theta}} l(\mathbf{O}_i; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m}), \quad (2.3)$$

where $\tilde{\mathbf{I}}_m$ is the information matrix for the first subset of data, and $D_{\boldsymbol{\theta}} l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}})$ denotes the total derivative of $l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\boldsymbol{\eta}}_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$. When $D_{\boldsymbol{\theta}} l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m})$ does not have an explicit expression, we approximate it by the numerical difference $D_{\boldsymbol{\theta}, \delta_n} l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m})$. One possible choice is the symmetric difference quotient,

$$D_{\boldsymbol{\theta}, \delta_n} l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m}) = (2\delta_n)^{-1} \left\{ l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m + \delta_n \mathbf{e}_j, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m + \delta_n \mathbf{e}_j}) - l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m - \delta_n \mathbf{e}_j, \tilde{\boldsymbol{\eta}}_{\tilde{\boldsymbol{\theta}}_m - \delta_n \mathbf{e}_j}) \right\}_{j=1, \dots, p}. \quad (2.4)$$

We show in Section 3 that $\hat{\boldsymbol{\theta}}_f$ has the same asymptotic distribution as $\hat{\boldsymbol{\theta}}_n$. Thus, we estimate the covariance matrix of $\hat{\boldsymbol{\theta}}_f$ by $m\tilde{\mathbf{I}}_m^{-1}/n$.

Remark 1. The classic one-step estimator, as studied by Bickel et al. (1993) and van der Vaart (1998), requires the initial estimator to be within an $O_p(n^{-1/2})$ neighborhood of $\boldsymbol{\theta}_0$ and relies on the consistent estimation of efficient scores, which often lack explicit expressions for semiparametric problems. By contrast, our proposed one-step estimator only requires $\|\tilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\|$ to be $O_p(m^{-1/2})$ and does not require the analytic formula of the semiparametric efficient scores.

Remark 2. The efficient score function used in our approach shares some spirit with the decorrelated score developed by Ning and Liu (2017), and the use of a small initial block to achieve computational efficiency is conceptually related to the communication-efficient surrogate likelihood framework proposed by Jordan et al. (2019). However, Ning and Liu (2017) assumed a finite-dimensional nuisance parameter and did not address the computational bottlenecks associated with massive sample sizes. Furthermore, the work of Jordan

et al. (2019) is strictly restricted to parametric models. In contrast, our approach directly addresses the theoretical and computational challenges of fitting general semiparametric models with infinite-dimensional components to massive datasets. Additionally, Chernozhukov et al. (2018) developed a double/debiased machine learning framework for debiasing parameters via explicit Neyman-orthogonal scores and K -fold cross-fitting. Their framework requires estimating the nuisance parameter K separate times, each time utilizing a training set of $n(K-1)/K$ observations. This repeated, large-scale estimation is computationally expensive for massive datasets. More importantly, it relies on explicit analytical expressions for the score function, which are unavailable for complex semiparametric models, such as the Cox proportional hazards model with interval-censored data and the random-effects proportional intensity model with recurrent events. By contrast, our approach estimates the nuisance parameter only once on a small initial block and completely circumvents the need for explicit expressions of semiparametric efficient scores.

Remark 3. Note that the summation in the one-step estimator (2.3) is taken over only the second block rather than the entire dataset. Because $\tilde{\theta}_m$ and $\tilde{\eta}_{\tilde{\theta}_m}$ are the maximum likelihood estimators for the first block, the summation of the efficient score functions over the first block is zero.

2.3 Examples

In this subsection, we present three examples to show how the proposed approach can be applied to various semiparametric problems.

Example (Partially linear model). The partially linear model assumes that the outcome Y ,

p -dimensional covariates $\mathbf{X}$, and a one-dimensional covariate Z follow the relationship

$$Y = \boldsymbol{\theta}^T \mathbf{X} + g(Z) + \epsilon,$$

where $\boldsymbol{\theta}$ is a p -vector of unknown regression parameters, $g(\cdot)$ is an unknown smooth function, and ϵ is a random error with $E(\epsilon \mid \mathbf{X}, Z) = 0$ and $E(\epsilon^2 \mid \mathbf{X}, Z) = \sigma^2(\mathbf{X}, Z)$. For simplicity, we assume that ϵ is zero-mean normal with an unknown constant variance σ^2 and is independent of $\mathbf{X}$ and Z . The log-likelihood for a single observation $\mathbf{O} = (Y, \mathbf{X}, Z)$ takes the form

$$l(\mathbf{O}; \boldsymbol{\theta}, g, \sigma^2) = -(2\sigma^2)^{-1} \{Y - \boldsymbol{\theta}^T \mathbf{X} - g(Z)\}^2 - \frac{1}{2} \log \sigma^2.$$

In this case, the nuisance parameters consist of $g(\cdot)$ and σ^2 .

We adopt the method of sieves (Geman and Hwang, 1982; Shen and Wong, 1994) to estimate the unknown function g . Specifically, we consider a polynomial splines sieve space and approximate $g(Z)$ by $\boldsymbol{\omega}^T \mathbf{p}(Z)$, where $\mathbf{p}(Z) = \{p_1(Z), \dots, p_{r+k_m}(Z)\}^T$ denotes the basis of the space of polynomial splines of order r with k_m knots, and $\boldsymbol{\omega} = (\omega_1, \dots, \omega_{r+k_m})^T$ is a vector of unknown parameters. We replace $g(Z)$ with $\boldsymbol{\omega}^T \mathbf{p}(Z)$ in the log-likelihood function with $\mathcal{J}_1$ to obtain

$$l_m(\boldsymbol{\theta}, \boldsymbol{\omega}, \sigma^2) = -(2\sigma^2)^{-1} \sum_{i \in \mathcal{J}_1} \{Y_i - \boldsymbol{\theta}^T \mathbf{X}_i - \boldsymbol{\omega}^T \mathbf{p}(Z_i)\}^2 - (m/2) \log \sigma^2.$$

It can be shown that for a fixed $\boldsymbol{\theta}$, the function $\boldsymbol{\omega}, \sigma^2 \mapsto l_m(\boldsymbol{\theta}, \boldsymbol{\omega}, \sigma^2)$ is maximal for $\tilde{\boldsymbol{\omega}}_{\boldsymbol{\theta}} = \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) \mathbf{p}(Z_i)^T \right\}^{-1} \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) (Y_i - \boldsymbol{\theta}^T \mathbf{X}_i)$ and $\tilde{\sigma}_{\boldsymbol{\theta}}^2 = m^{-1} \sum_{i \in \mathcal{J}_1} \{Y_i - \boldsymbol{\theta}^T \mathbf{X}_i - \tilde{\boldsymbol{\omega}}_{\boldsymbol{\theta}}^T \mathbf{p}(Z_i)\}^2$. The profile log-likelihood for $\boldsymbol{\theta}$ then takes the form

$$pl_m(\boldsymbol{\theta}) = -\frac{m}{2} \log \left[\sum_{i \in \mathcal{J}_1} \{Y_i - \boldsymbol{\theta}^T \mathbf{X}_i - \tilde{\boldsymbol{\omega}}_{\boldsymbol{\theta}}^T \mathbf{p}(Z_i)\}^2 \right] + \frac{m}{2} (\log m - 1).$$

It can be further shown that the sieve maximum likelihood estimator for $\boldsymbol{\theta}$ is

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_m = & \left[\sum_{i \in \mathcal{J}_1} \mathbf{X}_i \mathbf{X}_i^T - \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{X}_i \mathbf{p}(Z_i)^T \right\} \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) \mathbf{p}(Z_i)^T \right\}^{-1} \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) \mathbf{X}_i^T \right\} \right]^{-1} \\ & \times \left[\sum_{i \in \mathcal{J}_1} Y_i \mathbf{X}_i - \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{X}_i \mathbf{p}(Z_i)^T \right\} \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) \mathbf{p}(Z_i)^T \right\}^{-1} \left\{ \sum_{i \in \mathcal{J}_1} Y_i \mathbf{p}(Z_i) \right\} \right]. \end{aligned}$$

By taking the derivatives of the profile log-likelihood, we conclude that the covariance matrix of $\tilde{\boldsymbol{\theta}}_m$ can be estimated by $\tilde{\mathbf{I}}_m^{-1}$, where

$$\tilde{\mathbf{I}}_m = \tilde{\sigma}_{\tilde{\boldsymbol{\theta}}_m}^{-2} \left[\sum_{i \in \mathcal{J}_1} \mathbf{X}_i \mathbf{X}_i^T - \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{X}_i \mathbf{p}(Z_i)^T \right\} \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) \mathbf{p}(Z_i)^T \right\}^{-1} \left\{ \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) \mathbf{X}_i^T \right\} \right].$$

Next, we derive the expression of the proposed estimator for $\boldsymbol{\theta}$. Write $\tilde{g}_{\boldsymbol{\theta}}(\cdot) = \tilde{\boldsymbol{\omega}}_{\boldsymbol{\theta}}^T \mathbf{p}(\cdot)$ and $\dot{\tilde{g}}_{\boldsymbol{\theta}}(\cdot) = \dot{\tilde{\boldsymbol{\omega}}}_{\boldsymbol{\theta}}^T \mathbf{p}(\cdot)$. By differentiating $l(\mathbf{O}; \boldsymbol{\theta}, \tilde{g}_{\boldsymbol{\theta}}, \tilde{\sigma}_{\boldsymbol{\theta}}^2)$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}_m$ and noting that $\dot{\tilde{\sigma}}_{\tilde{\boldsymbol{\theta}}_m}^2 = 0$, we have

$$D_{\boldsymbol{\theta}} l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{g}_{\tilde{\boldsymbol{\theta}}_m}, \tilde{\sigma}_{\tilde{\boldsymbol{\theta}}_m}^2) = \tilde{\sigma}_{\tilde{\boldsymbol{\theta}}_m}^{-2} \left\{ Y - \tilde{\boldsymbol{\theta}}_m^T \mathbf{X} - \tilde{g}_{\tilde{\boldsymbol{\theta}}_m}(Z) \right\} \left\{ \mathbf{X} + \dot{\tilde{g}}_{\tilde{\boldsymbol{\theta}}_m}(Z) \right\}.$$

This mirrors the efficient score function for $\boldsymbol{\theta}$, which takes the form $\mathbf{l}_0(\mathbf{O}; \boldsymbol{\theta}, g_0, \sigma_0^2) = \sigma_0^{-2} \{Y - \boldsymbol{\theta}^T \mathbf{X} - g_0(Z)\} \{\mathbf{X} - E(\mathbf{X} \mid Z)\}$ (Bickel et al., 1993, pages 110–111), where σ_0^2 and $g_0(\cdot)$ are the true values of σ^2 and $g(\cdot)$, respectively. Clearly, $\dot{\tilde{\boldsymbol{\omega}}}_{\tilde{\boldsymbol{\theta}}_m} = -\left\{ \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) \mathbf{p}(Z_i)^T \right\}^{-1} \sum_{i \in \mathcal{J}_1} \mathbf{p}(Z_i) \mathbf{X}_i^T$, and $\dot{\tilde{g}}_{\tilde{\boldsymbol{\theta}}_m}(Z)$ is the sieve least-squares estimator of $-E(\mathbf{X} \mid Z)$. Thus, $D_{\boldsymbol{\theta}} l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{g}_{\tilde{\boldsymbol{\theta}}_m}, \tilde{\sigma}_{\tilde{\boldsymbol{\theta}}_m}^2)$ is an empirical counterpart of $\mathbf{l}_0(\mathbf{O}; \boldsymbol{\theta}, g_0, \sigma_0^2)$, with $\boldsymbol{\theta}$, $g_0(Z)$, $-E(\mathbf{X} \mid Z)$, and σ_0^2 replaced by $\tilde{\boldsymbol{\theta}}_m$, $\tilde{g}_{\tilde{\boldsymbol{\theta}}_m}(Z)$, $\dot{\tilde{g}}_{\tilde{\boldsymbol{\theta}}_m}(Z)$, and $\tilde{\sigma}_{\tilde{\boldsymbol{\theta}}_m}^2$, respectively.

The proposed estimator is then given by

$$\hat{\boldsymbol{\theta}}_f = \tilde{\boldsymbol{\theta}}_m + \left(n \tilde{\mathbf{I}}_m / m \right)^{-1} \sum_{i \in \mathcal{J}_2} (\tilde{\sigma}_{\tilde{\boldsymbol{\theta}}_m}^2)^{-1} \left\{ Y_i - \tilde{\boldsymbol{\theta}}_m^T \mathbf{X}_i - \tilde{g}_{\tilde{\boldsymbol{\theta}}_m}(Z_i) \right\} \left\{ \mathbf{X}_i + \dot{\tilde{g}}_{\tilde{\boldsymbol{\theta}}_m}(Z_i) \right\}.$$

Example (Cox proportional hazards model with right-censored data). The Cox (1972) proportional hazards model specifies that the hazard function of a failure time T conditional on

a p -vector of potentially time-dependent covariates $\mathbf{Z}$ takes the form

$$\lambda(t; \mathbf{Z}) = \lambda(t)e^{\boldsymbol{\theta}^T \mathbf{Z}(t)},$$

where $\boldsymbol{\theta}$ is a p -vector of unknown regression parameters, and $\lambda(t)$ is an arbitrary baseline hazard function. When T is subject to right censoring by C , we observe $\mathbf{O} = (Y, \Delta, \mathbf{Z})$, where $Y = \min(T, C)$, $\Delta = I(T \leq C)$, and $I(\cdot)$ is the indicator function. It is assumed that C is independent of T conditional on $\mathbf{Z}$. The log-likelihood for a single observation is

$$l(\mathbf{O}; \boldsymbol{\theta}, \Lambda) = \Delta\{\boldsymbol{\theta}^T \mathbf{Z}(Y) + \log \lambda(Y)\} - \int_0^Y e^{\boldsymbol{\theta}^T \mathbf{Z}(t)} d\Lambda(t),$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$, which is an infinite-dimensional nuisance parameter.

By treating $\lambda(\cdot)$ as piecewise constant between uncensored failure times (Kalbfleisch and Prentice, 2002), elementary calculus shows that for a fixed $\boldsymbol{\theta}$, the function $\Lambda \mapsto \sum_{i \in \mathcal{J}_1} l(\mathbf{O}_i; \boldsymbol{\theta}, \Lambda)$ is maximal for

$$\tilde{\Lambda}_{\boldsymbol{\theta}}(t) = \sum_{i \in \mathcal{J}_1} \frac{I(Y_i \leq t)\Delta_i}{\sum_{j \in \mathcal{J}_1} I(Y_j \geq Y_i)e^{\boldsymbol{\theta}^T \mathbf{Z}(Y_i)}}.$$

The profile log-likelihood for $\boldsymbol{\theta}$ then equals to the partial likelihood for $\boldsymbol{\theta}$ (Cox, 1975), and $\tilde{\boldsymbol{\theta}}_m$ is the maximum partial likelihood estimator based on $\mathcal{J}_1$.

By differentiating $l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\Lambda}_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}_m$, we have

$$D_{\boldsymbol{\theta}} l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m}) = \int_0^\infty \mathbf{Z}(t) - \tilde{\mathbf{E}}(t) d\tilde{M}(t),$$

where $\tilde{\mathbf{E}}(t) = \sum_{j \in \mathcal{J}_1} I(Y_j \geq t)e^{\tilde{\boldsymbol{\theta}}_m^T \mathbf{Z}(t)} \mathbf{Z}(t) / \sum_{j \in \mathcal{J}_1} I(Y_j \geq t)e^{\tilde{\boldsymbol{\theta}}_m^T \mathbf{Z}(t)}$ and $\tilde{M}(t) = \Delta I(Y \leq t) - \int_0^t I(Y \geq s)e^{\tilde{\boldsymbol{\theta}}_m^T \mathbf{Z}(s)} d\tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m}(s)$. This is an empirical counterpart of the efficient score function for $\boldsymbol{\theta}$, which takes the form

$$\mathbf{l}_0(\mathbf{O}; \boldsymbol{\theta}, \Lambda_0) = \int_0^\infty \mathbf{Z}(t) - \mathbf{E}\{\mathbf{Z}(t) \mid Y \geq t\} dM(t),$$

where $M(t) = \Delta I(Y \leq t) - \int_0^t I(Y \geq s) e^{\boldsymbol{\theta}^T \mathbf{Z}(s)} d\Lambda_0(t)$ and $\Lambda_0(\cdot)$ is the true value of $\Lambda(\cdot)$. The proposed estimator then becomes the Cox-OSSES estimator (Wang et al., 2024), which will not be considered further.

Example (Cox proportional hazards model with interval-censored data). We now consider the model of the previous example but with general interval-censored data instead of right-censored data. When T is known only to lie in a time interval induced by a sequence of monitoring times $U_1 < \cdots < U_K$, we observe $\mathbf{O} = (L, R, \mathbf{Z})$, where $L = \max\{U_k : U_k < T, k = 0, \dots, K\}$ and $R = \min\{U_k : U_k \geq T, k = 1, \dots, K+1\}$, with $U_0 = 0$ and $U_{K+1} = \infty$. Of note, $L = 0$ indicates that T is left-censored, and $R = \infty$ indicates that T is right-censored. The monitoring times and K are assumed to be independent of T conditional on $\mathbf{Z}$. The log-likelihood for a single observation is

$$l(\mathbf{O}; \boldsymbol{\theta}, \Lambda) = \log \left[\exp \left\{ - \int_0^L e^{\boldsymbol{\theta}^T \mathbf{Z}(t)} d\Lambda(t) \right\} - \exp \left\{ - \int_0^R e^{\boldsymbol{\theta}^T \mathbf{Z}(t)} d\Lambda(t) \right\} \right].$$

We adopt the nonparametric maximum likelihood approach to estimate $\boldsymbol{\theta}$ and Λ . Unlike the Cox model with right-censored data, $\tilde{\Lambda}_{\boldsymbol{\theta}} = \arg \max_{\Lambda} \sum_{i \in \mathcal{J}_1} l(\mathbf{O}_i; \boldsymbol{\theta}, \Lambda)$ does not have an analytical form. The expectation-maximization (EM) algorithm (Zeng et al., 2016) provides a stable method to estimate $\tilde{\Lambda}_{\boldsymbol{\theta}}$ and maximize the likelihood. Let $pl_m(\boldsymbol{\theta}) = \sum_{i \in \mathcal{J}_1} l(\mathbf{O}_i; \boldsymbol{\theta}, \tilde{\Lambda}_{\boldsymbol{\theta}})$, and let $\tilde{\boldsymbol{\theta}}_m$ denote the nonparametric maximum likelihood estimator of $\boldsymbol{\theta}$ based on $\mathcal{J}_1$. The covariance matrix of $\tilde{\boldsymbol{\theta}}_m$ is estimated by $\tilde{\mathbf{I}}_m^{-1}$, where the (j, k) th element of $\tilde{\mathbf{I}}_m$ is

$$-\frac{pl_m(\tilde{\boldsymbol{\theta}}_m) - pl_m(\tilde{\boldsymbol{\theta}}_m + h_m \mathbf{e}_j) - pl_m(\tilde{\boldsymbol{\theta}}_m + h_m \mathbf{e}_k) + pl_m(\tilde{\boldsymbol{\theta}}_m + h_m \mathbf{e}_j + h_m \mathbf{e}_k)}{h_m^2},$$

or

$$m \mathbb{P}_m \left[h_m^{-2} \left\{ l(\tilde{\boldsymbol{\theta}}_m + h_m \mathbf{e}_j, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m + h_m \mathbf{e}_j}) - l(\tilde{\boldsymbol{\theta}}_m, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m}) \right\} \left\{ l(\tilde{\boldsymbol{\theta}}_m + h_m \mathbf{e}_k, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m + h_m \mathbf{e}_k}) - l(\tilde{\boldsymbol{\theta}}_m, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m}) \right\} \right],$$

and h_m is a constant of order $m^{-1/2}$.

To calculate the proposed estimator, we approximate $D_{\boldsymbol{\theta}}l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\Lambda}_{\boldsymbol{\theta}})$ by a numerical difference. Specifically, we choose a constant δ_n of order $n^{-1/2}$ and compute $\tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m + \delta_n \mathbf{e}_j}(\cdot)$ and $\tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m - \delta_n \mathbf{e}_j}(\cdot)$ ($j = 1, \dots, p$) by the EM algorithm. We then calculate $l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m + \delta_n \mathbf{e}_j, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m + \delta_n \mathbf{e}_j})$ and $l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m - \delta_n \mathbf{e}_j, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m - \delta_n \mathbf{e}_j})$, and we approximate $D_{\boldsymbol{\theta}}l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m})$ by

$$D_{\boldsymbol{\theta}, \delta_n}l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m}) = (2\delta_n)^{-1} \left\{ l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m + \delta_n \mathbf{e}_j, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m + \delta_n \mathbf{e}_j}) - l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m - \delta_n \mathbf{e}_j, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m - \delta_n \mathbf{e}_j}) \right\}_{j=1, \dots, p}.$$

The proposed estimator can then be calculated by (2.3) upon replacing $D_{\boldsymbol{\theta}}l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m})$ with $D_{\boldsymbol{\theta}, \delta_n}l(\mathbf{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\Lambda}_{\tilde{\boldsymbol{\theta}}_m})$.

Example (Proportional hazards or intensity model with random effects for dependent failure times). To encompass various multivariate or dependent failure times, we consider the counting processes $N_{ikj}^*(\cdot)$ ($i = 1, \dots, n; k = 1, \dots, K; j = 1, \dots, n_{ik}$), where i pertains to a subject or cluster, k to the type of event and j to individuals within a cluster (Zeng and Lin, 2007). The specific choices of $K = n_{ik} = 1$, $n_{ik} = 1$, and $K = 1$ correspond to recurrent events, multiple types of events, and clustered failure times, respectively. We specify that the cumulative hazard or intensity function for $N_{ikj}^*(t)$ conditional on covariates $\mathbf{Z}_{ikj}(t)$ takes the form

$$\Lambda_k(t \mid \mathbf{Z}_{ikj}; \mathbf{b}_i) = \int_0^t R_{ikj}^*(s) \exp\{\boldsymbol{\theta}^T \mathbf{Z}_{ikj}(s) + \mathbf{b}_i^T \tilde{\mathbf{Z}}_{ikj}(s)\} d\Lambda_k(s),$$

where $R_{ikj}^*(\cdot)$ is the indicator process, $\boldsymbol{\theta}$ is a p -vector of unknown regression parameters, $\tilde{\mathbf{Z}}_{ikj}$ is a subset of $\mathbf{Z}_{ikj}$ plus the unit component, $\mathbf{b}_i$ ($i = 1, \dots, n$) are independent d -dimensional random vectors with multivariate density function $f(\mathbf{b}; \boldsymbol{\gamma})$ indexed by a set of parameters $\boldsymbol{\gamma}$, and $\Lambda_k(\cdot)$ ($k = 1, \dots, K$) are arbitrary increasing functions (Zeng and Lin, 2007). Let T_{ikj} denote the failure time. For survival data, $R_{ikj}^*(t) = I(T_{ikj} \geq t)$, where $I(\cdot)$ is the indicator function; and for recurrent events, $R_{ikj}^*(\cdot) = 1$.

We denote C_{ikj} the censoring time, which is assumed to be independent of $N_{ikj}^*(\cdot)$ and $\mathbf{b}_i$ conditional on $\mathbf{Z}_{ikj}(\cdot)$ and be non-informative about $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and $\Lambda_k(k = 1, \dots, K)$. The observed data consist of $\{N_{ikj}(t), R_{ikj}(t), \mathbf{Z}_{ikj}(t); t \in [0, \tau]\}$, where $N_{ikj}(t) = N_{ikj}^*\{\min(t, C_{ikj})\}$, $R_{ikj}(t) = I(C_{ikj} \geq t)R_{ikj}^*(t)$, and τ is the duration of the study. The log-likelihood for $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and $\Lambda_k(k = 1, \dots, K)$ with a single cluster or subject i is

$$\begin{aligned} & l(\mathbf{O}_i; \boldsymbol{\theta}, \boldsymbol{\gamma}, \Lambda_1, \dots, \Lambda_K) \\ &= \log \int_{\mathbf{b}} \prod_{k=1}^K \prod_{j=1}^{n_{ik}} \prod_{t \leq \tau} \left[R_{ikj}(t) \lambda_k(t) \exp\{\boldsymbol{\theta}^\top \mathbf{Z}_{ikj}(t) + \mathbf{b}^\top \tilde{\mathbf{Z}}_{ikj}(t)\} \right]^{dN_{ikj}(t)} \\ & \quad \exp \left[- \int_0^\tau R_{ikj}(t) \exp\{\boldsymbol{\theta}^\top \mathbf{Z}_{ikj}(t) + \mathbf{b}^\top \tilde{\mathbf{Z}}_{ikj}(t)\} d\Lambda_k(t) \right] f(\mathbf{b}; \boldsymbol{\gamma}) d\mathbf{b}, \end{aligned}$$

where $\lambda_k(t) = \Lambda'_k(t)(k = 1, \dots, K)$ and $\mathbf{O}_i = [\{N_{ikj}(t), R_{ikj}(t), \mathbf{Z}_{ikj}(t)\}; t \in [0, \tau], k = 1, \dots, K, j = 1, \dots, n_{ik}]$.

We adopt the nonparametric maximum likelihood approach for inference. Since the estimates $\tilde{\boldsymbol{\gamma}}, \tilde{\Lambda}_{1,\boldsymbol{\theta}}, \dots, \tilde{\Lambda}_{K,\boldsymbol{\theta}}$ lack closed-form expressions, we utilize the EM algorithm (Zeng and Lin, 2007) to maximize the likelihood. The subsequent steps for variance estimation and the construction of the proposed estimator follow the same approach as in the interval-censored example. Specifically, we approximate $D_{\boldsymbol{\theta}} l(\mathbf{O}_i; \tilde{\boldsymbol{\theta}}_m, \tilde{\boldsymbol{\gamma}}_{\tilde{\boldsymbol{\theta}}_m}, \tilde{\Lambda}_{1,\tilde{\boldsymbol{\theta}}_m}, \dots, \tilde{\Lambda}_{K,\tilde{\boldsymbol{\theta}}_m})$ by the symmetric difference quotient with perturbation δ_n , and obtain the final estimator via (2.3).

3. Asymptotic Properties and Time Complexities

3.1 Asymptotic properties

We assume that $\mathcal{F}$ is a linear normed space with norm $\|\cdot\|_{\mathcal{F}}$ and that for any $f \in \mathcal{F}$, there exists $f_n \in \mathcal{F}_n$ such that $\|f_n - f\|_{\mathcal{F}} \rightarrow 0$ as $n \rightarrow \infty$. Let $\mathcal{H}$ be a normed space with norm $\|\cdot\|_{\mathcal{H}}$. We say that a map $\Phi : \mathcal{F} \mapsto \mathcal{H}$ is Fréchet-differentiable at $\eta \in \mathcal{F}$

with respect to norm $\|\cdot\|_{\mathcal{F}}$ if there exists a bounded linear map $\dot{\Phi}[\cdot] : \mathcal{F} \mapsto \mathcal{H}$ such that $\|\Phi(\eta + h) - \Phi(\eta) - \dot{\Phi}[h]\|_{\mathcal{H}}/\|h\|_{\mathcal{F}} \rightarrow 0$ as $\|h\|_{\mathcal{F}} \rightarrow 0$.

We impose the following regularity conditions:

Condition 1. There exists a neighborhood V of $(\boldsymbol{\theta}_0, \eta_0)$ such that $\boldsymbol{\theta} \mapsto l(\boldsymbol{O}; \boldsymbol{\theta}, \eta)$ is thrice differentiable and $\eta \mapsto l(\boldsymbol{O}; \boldsymbol{\theta}, \eta)$ is thrice Fréchet-differentiable. We denote the first- and second-order derivatives by $\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{O}; \boldsymbol{\theta}, \eta)$, $\dot{l}_{\eta}(\boldsymbol{O}; \boldsymbol{\theta}, \eta)[\cdot]$, $\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{O}; \boldsymbol{\theta}, \eta)$, $\ddot{l}_{\boldsymbol{\theta}\eta}(\boldsymbol{O}; \boldsymbol{\theta}, \eta)[\cdot]$ and $\ddot{l}_{\eta\eta}(\boldsymbol{O}; \boldsymbol{\theta}, \eta)[\cdot, \cdot]$.

Condition 2. The least favorable direction h_0 , defined as the function $h \in \mathcal{F}$ that minimizes $E\|\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{O}; \boldsymbol{\theta}_0, \eta_0) - \dot{l}_{\eta}(\boldsymbol{O}; \boldsymbol{\theta}_0, \eta_0)[h]\|^2$, exists and is unique.

Condition 3. There exists a neighborhood of $\boldsymbol{\theta}_0$ such that $\boldsymbol{\theta} \mapsto \tilde{\eta}_{\boldsymbol{\theta}}$ is thrice differentiable, and the derivatives are bounded in the neighborhood of $\boldsymbol{\theta}_0$. In addition, there exist positive numbers α, β, γ with $\alpha > 1/4$, $\alpha + \beta > 1/2$ and $\tilde{h}_0 \in \mathcal{F}$ such that as $\boldsymbol{\theta} \rightarrow_P \boldsymbol{\theta}_0$,

$$\|\tilde{\eta}_{\boldsymbol{\theta}} - \eta_0\|_{\mathcal{F}} = O_P(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|) + O_P(m^{-\alpha}),$$

$$\|\dot{\tilde{\eta}}_{\boldsymbol{\theta}} - \tilde{h}_0\|_{\mathcal{F}} = O_P(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|) + O_P(m^{-\beta}),$$

and

$$\|m^{-1}\tilde{\mathbf{I}}_m - \mathbf{I}_0\|_2 = O_P(m^{-\gamma}),$$

where $\dot{\tilde{\eta}}_{\boldsymbol{\theta}}$ denotes the first-order derivative of $\tilde{\eta}_{\boldsymbol{\theta}}$, and $\|\cdot\|_2$ is the $L_2(\mathbb{P}_{\boldsymbol{\theta}_0, \eta_0})$ norm.

Condition 4. The function $\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}_m}) + \dot{l}_{\eta}(\boldsymbol{O}; \tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}_m})[\dot{\tilde{\eta}}_{\tilde{\boldsymbol{\theta}}_m}]$ belongs to some Donsker class with square-integrable envelope functions, and for any fixed $h \in \mathcal{F}$, $\ddot{l}_{\eta\eta}(\boldsymbol{O}; \boldsymbol{\theta}_0, \tilde{\eta}_{\boldsymbol{\theta}_0})[\dot{\tilde{\eta}}_{\boldsymbol{\theta}_0}, h]$ and $\ddot{l}_{\boldsymbol{\theta}\eta}(\boldsymbol{O}; \boldsymbol{\theta}_0, \tilde{\eta}_{\boldsymbol{\theta}_0})[h]$ belong to some Glivenko-Cantelli classes with an envelope function bounded in $L_1(\mathbb{P}_{\boldsymbol{\theta}_0, \eta_0})$.

Condition 5. The size of the first block satisfies that $n/m^r \rightarrow 0$ as $n, m \rightarrow \infty$, where $r = \min(2, 4\alpha, 1 + 2\alpha, 1 + 4\beta, 2\alpha + 2\beta, 1 + 2\gamma)$.

Condition 6. The numerical difference $D_{\theta, \delta_n} l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\eta}_{\boldsymbol{\theta}})$ satisfies

$$\mathbb{P}_n D_{\theta, \delta_n} l(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}_m}) = \mathbb{P}_n D_{\boldsymbol{\theta}} l(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}_m}) + o_P(n^{-1/2})$$

and

$$\mathbb{P}_m D_{\theta, \delta_n} l(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}_m}) = \mathbb{P}_m D_{\boldsymbol{\theta}} l(\tilde{\boldsymbol{\theta}}_m, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}_m}) + o_P(n^{-1/2}).$$

Remark 4. Condition 1 requires the smoothness of the density function. Condition 2 ensures the existence and uniqueness of the least favorable direction. Condition 3 specifies the smoothness and convergence rate of the nuisance parameter estimator, which holds for most semiparametric problems. Condition 4 pertains to empirical process conditions, which are needed in establishing the asymptotic properties of the proposed estimator. Conditions 3 and 4 need to be verified for each specific semiparametric problem using empirical process theory and functional analysis. Examples are provided in Section S.2 of the Supplementary Materials. Condition 5 implies that the size of the first block can be $O(n^{1/r+\xi})$ for any $\xi > 0$, which is much smaller than n since $r > 1$ by Condition 3. Specifically, $r = 2$ for parametric models and the Cox proportional hazards model with right-censored data. Condition 6 ensures that the numerical difference does not introduce bias in the final one-step estimator. We can verify that the symmetric difference quotient in (2.4) with $\delta_n = O(n^{-1/2})$ satisfies this condition by showing that the third-order total derivative of $l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\eta}_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}_m$ belongs to some Glivenko-Cantelli class.

We state below the asymptotic properties of the proposed estimator.

Theorem 1. *Under Conditions 1–6, $n^{1/2}(\hat{\boldsymbol{\theta}}_f - \boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector with covariance matrix $\mathbf{I}_0^{-1}$.*

The theorem is proved in Section S.1 of the Supplementary Materials. The proof is based on the expansion of the score function for $\boldsymbol{\theta}$ and η and relies on modern empirical process theory (van der Vaart and Wellner, 1996). Theorem 1 implies that the proposed estimator is as efficient as the standard maximum likelihood estimator with the entire dataset and that the limiting covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}}_f - \boldsymbol{\theta}_0)$ attains the semiparametric efficiency bound. Thus, we estimate the covariance matrix of $\hat{\boldsymbol{\theta}}_f$ by $m\tilde{\mathbf{I}}_m^{-1}/n$.

Remark 5. A more accurate estimator of the covariance matrix can be obtained by using the profile likelihood approach with the entire data. However, its computation is more intensive than simply scaling $\tilde{\mathbf{I}}_m^{-1}$. Since Theorem 1 only requires the consistency of the covariance matrix estimator, it is sufficient to estimate the covariance matrix by $m\tilde{\mathbf{I}}_m^{-1}/n$.

We now apply Theorem 1 to the first and third examples introduced in Section 2.3.

Example (Partially linear model). We assume that $\boldsymbol{\theta}_0$ lies in the interior of a compact subset $B \subset \mathbb{R}^p$ and that both g_0 and $E(\mathbf{X} \mid Z)$ belong to a Hölder ball $\mathcal{G}$, where g_0 is the true value of g . We define $\mathcal{G} = \{g \in C^\rho[a, b] : \sup_z |g^{(j)}(z)| \leq L, j = 0, \dots, \rho, |g^{(\rho)}(z_1) - g^{(\rho)}(z_2)| \leq L|z_1 - z_2|^\kappa\}$, where $C^\rho[a, b]$ denotes the space of all ρ -times continuously differentiable real-valued functions on $[a, b]$, ρ is a nonnegative integer, and κ is a positive number. The smoothness of $\mathcal{G}$ is defined by $q = \rho + \kappa$, which is assumed to be greater than $1/2$. We further assume that σ_0^2 is positive and bounded and that both $E(\mathbf{X}\mathbf{X}^\top)$ and $E[\{\mathbf{X} - E(\mathbf{X} \mid Z)\}\{\mathbf{X} - E(\mathbf{X} \mid Z)\}^\top]$ are positive definite. These assumptions are standard for the partially linear model (Chen, 2007).

Under these assumptions and with $k_m = O\{m^{1/(2q+1)}\}$, α, β , and γ in Condition 3 are all $q/(2q+1)$, implying that r in Condition 5 is $4q/(2q+1)$. The validity of Conditions 1 and 2 follows from the definitions of differentiability and least favorable direction. Condition 4 can be verified by bounding the bracketing numbers of some function classes. Detailed proofs are provided in Section S.2.1 of the Supplementary Materials. It follows from Theorem 1 that the proposed estimator for the partially linear model is asymptotically normal and its limiting variance attains the semiparametric efficiency bound.

Example (Cox proportional hazards model with interval-censored data). We assume the regularity conditions imposed by Zeng et al. (2016). To verify Condition 1, we write the log-likelihood function for a single observation as

$$l(\mathbf{O}; \boldsymbol{\theta}, \Lambda) = \sum_{k=1}^K \Delta_k \log \left(\exp \left[- \int_0^{U_k} \exp\{\boldsymbol{\theta}^T \mathbf{Z}(t)\} d\Lambda(t) \right] \right. \\ \left. - \exp \left[- \int_0^{U_{k+1}} \exp\{\boldsymbol{\theta}^T \mathbf{Z}(t)\} d\Lambda(t) \right] \right),$$

where $\Delta_k = I(U_k < T \leq U_{k+1})$ ($k = 0, \dots, K$). Condition 1 then follows from the definition of differentiability. Conditions 2 and 4 have been verified by Zeng et al. (2016) and Xu et al. (2022).

We can show that the third-order total derivative of $l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\Lambda}_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}_m$ belongs to some Glivenko-Cantelli class. Thus, the symmetric difference quotient satisfies Condition 6. For Condition 3, Zeng et al. (2016) showed that $\alpha = 1/3$. Modern empirical process theory (van der Vaart and Wellner, 1996) can be used to further show that both β and γ in Condition 3 are $1/3$. Thus, the r in Condition 5 is $4/3$. Additional details about regularity conditions and the verification of Condition 3 are presented in Section S.2.2 of the Supplementary Materials. It follows from Theorem 1 that the proposed estimator for the

Cox proportional hazards model with interval-censored data is asymptotically normal and semiparametric efficient.

Example (Proportional hazards or intensity model with random effects for dependent failure times). We adopt the regularity conditions imposed by Zeng and Lin (2007). Condition 1 in our Theorem 1 follows from the definition of differentiability and the dominated convergence theorem. Conditions 2 and 4 hold by the results of Zeng and Lin (2010). Using arguments similar to those of Gao et al. (2019), we can establish that the third-order total derivative of $l(\mathbf{O}; \boldsymbol{\theta}, \tilde{\Lambda}_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}_m$ belongs to some Glivenko-Cantelli class, ensuring that the symmetric difference quotient satisfies Condition 6. Modern empirical process theory yields $\alpha = \beta = \gamma = 1/2$ for Condition 3, implying $r = 2$ in Condition 5. Additional details are provided in the Supplementary Material. By Theorem 1, the proposed estimator for the proportional hazards or intensity model with random effects is asymptotically normal and semiparametric efficient.

3.2 Time complexity

In this subsection, we evaluate the computational efficiency of the proposed method compared to the standard method of maximum likelihood estimation with the entire dataset and the divide-and-conquer method, as a function of sample size while holding p constant. As there is no existing divide-and-conquer method for general semiparametric problems, we use an ad-hoc divide-and-conquer approach for comparison. Specifically, we randomly divide the data into M blocks of equal sizes, perform maximum likelihood estimation on each block, and average the parameter and covariance matrix estimators. We assume that $m = O(n^{r_0})$ by Condition 5 for some $1/r < r_0 < 1$. We set $M = n/m$ to ensure the block size in the

divide-and-conquer method the same as the size of the first block in the proposed method. Detailed calculations are provided in Section S.3 of the Supplementary Materials.

For the partially linear model, the time complexities are $O\{n^{1+2/(1+2q)}\}$, $O\{n^{1+2r_0/(1+2q)}\}$ and $O\{n^{r_0+2r_0/(1+2q)} + n^{1+r_0/(1+2q)}\}$ for the standard, divide-and-conquer and proposed methods, respectively. For the Cox proportional hazards model with interval-censored data, we assume that the number of distinct monitoring times is proportional to the sample size. Let N_{iter} denote the number of iterations for the EM algorithm. The time complexities are $O(N_{iter}n^2)$, $O(N_{iter}n^{1+r_0})$ and $O(N_{iter}n^{2r_0} + n \log n)$ for the standard, divide-and-conquer and proposed methods, respectively. For both models, the time complexity of the proposed method is lower than that of the divide-and-conquer and standard methods since r_0 is less than 1. Thus, the computational advantages of the proposed estimator become more significant as n increases.

4. Simulation Studies

In this section, we present simulation studies comparing the proposed methods with the standard method and the divide-and-conquer method under the partially linear model, the Cox proportional hazards model with interval-censored data, and the random effects proportional hazards model for clustered failure times. Additional results for the random effects proportional intensity model with recurrent events and the random effects proportional hazards model with multiple event types are provided in Section S4 of the Supplementary Material. All simulations were conducted on Intel® Xeon® Gold 5220R @2.20GHz.

4.1 Partially linear model

We generated three covariates, $Z \sim \text{Un}(0, 1)$, $X_1 = \sin(Z) + \text{Un}(0, 1)$, and $X_2 = \cos(Z) + \text{Un}(0, 1)$, and we set $\theta_1 = 0.5, \theta_2 = -0.5, \sigma^2 = 1$, and $g(z) = 0.4f(z; 2, 3) + 0.6f(z; 3, 2)$, where $f(z; \alpha, \gamma)$ is the probability density function of $\text{Beta}(\alpha, \gamma)$. We considered a total sample size of 10^6 or 10^7 .

For the standard method, we performed sieve maximum likelihood estimation with the entire dataset. Specifically, we considered the B-spline basis of order 3 and used 10-fold cross-validation to select the number of knots from $\{5, 10, 20, 40, 80, 160, 320\}$. The knots were chosen according to the quantiles of the input Z . For the proposed method, we set $m = kn$, where $k = 0.01, 0.02, 0.05, 0.1$ or 0.2 . For the divide-and-conquer method, we divided the whole dataset into $M = 5, 10, 20, 50$ or 100 blocks of equal sizes and made the size of the first block in the proposed method the same as the block size in the divide-and-conquer method. For each method, we recorded the total CPU time.

Table 1 summarizes the simulation results based on 1000 replicates. The proposed estimator achieves the same statistical efficiency as the standard estimator for all choices of m . Under $n = 10^7$, the standard method took approximately 2.6 hours, and the total CPU time for the divide-and-conquer method was approximately 2 hours for all choices of M . By contrast, the proposed methods with $m = 0.01 \times n, 0.02 \times n, 0.05 \times n, 0.1 \times n$ and $0.2 \times n$ took only 1.1, 2.3, 6.1, 12.6, and 26 minutes, respectively.

4.2 Cox proportional hazards model with interval-censored data

We generated two time-independent covariates, $Z_1 \sim \text{Ber}(0.5)$ and $Z_2 \sim \text{Un}(0, 1)$, set $\Lambda(t) = \log(1 + t/2)$ and $\boldsymbol{\theta} = (0.5, -0.5)^T$, and generated two monitoring times, $U_1 = \text{Un}(0, 9/4)$ and

Table 1: Summary statistics for the simulation study under the partially linear model

Method	$\theta_1 = 0.5$				$\theta_2 = -0.5$			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP
	$\times 10^{-4}$	$\times 10^{-3}$	$\times 10^{-3}$	$\times 100\%$	$\times 10^{-4}$	$\times 10^{-3}$	$\times 10^{-3}$	$\times 100\%$
$n = 10^6$ Standard	0.764	3.42	3.46	96	-1.391	3.42	3.46	95
Proposed								
$k = 0.01$	0.458	3.67	3.46	94	-1.301	3.65	3.46	94
$k = 0.02$	0.843	3.50	3.46	95	-1.254	3.48	3.46	95
$k = 0.05$	0.813	3.43	3.46	95	-1.396	3.43	3.46	95
$k = 0.1$	0.779	3.42	3.46	95	-1.391	3.43	3.46	95
$k = 0.2$	0.770	3.41	3.46	96	-1.407	3.43	3.46	95
Divide-and-conquer								
$M = 5$	0.759	3.42	3.46	96	-1.390	3.42	3.46	95
$M = 10$	0.757	3.42	3.46	95	-1.387	3.42	3.46	95
$M = 20$	0.759	3.42	3.46	95	-1.388	3.42	3.46	95
$M = 50$	0.744	3.42	3.46	96	-1.393	3.42	3.46	95
$M = 100$	0.715	3.43	3.46	96	-1.371	3.42	3.46	95
$n = 10^7$ Standard	-0.311	1.04	1.10	96	-0.384	1.10	1.10	95
Proposed								
$k = 0.01$	-0.298	1.05	1.10	97	-0.363	1.11	1.10	95
$k = 0.02$	-0.332	1.05	1.10	96	-0.385	1.10	1.10	95
$k = 0.05$	-0.313	1.05	1.10	97	-0.388	1.10	1.10	95
$k = 0.1$	-0.306	1.05	1.10	96	-0.383	1.10	1.10	95
$k = 0.2$	-0.312	1.05	1.10	96	-0.381	1.10	1.10	95
Divide-and-conquer								
$M = 5$	-0.311	1.04	1.10	97	-0.384	1.10	1.10	95
$M = 10$	-0.313	1.04	1.10	97	-0.384	1.10	1.10	95
$M = 20$	-0.310	1.04	1.10	97	-0.385	1.10	1.10	95
$M = 50$	-0.309	1.04	1.10	97	-0.385	1.10	1.10	95
$M = 100$	-0.314	1.04	1.10	96	-0.385	1.10	1.10	95

Bias, bias of the parameter estimator; SE, standard error of the parameter estimator; SEE, average of the standard error estimator; CP, coverage percentage of the 95% confidence interval.

$U_2 = \min\{0.1 + U_1 + \text{Exp}(2/3), 3\}$, where $\text{Exp}(\alpha)$ denotes the exponential distribution with mean $1/\alpha$. On average, 25–35% of the observations were left-censored, and 50–60% were right-censored. We set $n = 10^6$.

For the standard method, we performed nonparametric maximum likelihood estimation with the entire dataset using the EM algorithm of Zeng et al. (2016). For the proposed method, we set $m = kn$, where $k = 0.05, 0.1$ or 0.2 , and $\delta_n = 10 \times n^{-1/2}$ or $50 \times n^{-1/2}$. For the divide-and-conquer method, we divided the entire dataset into $M = 5, 10$ or 20 blocks of equal sizes to ensure that the block size is the same as the size of the first block in the proposed method. For variance estimation, we used the profile likelihood approach with $h_m = 5 \times m^{-1/2}$ for both the proposed and divide-and-conquer methods, and we used $h_n = 5 \times n^{-1/2}$ for the standard method. When computing time was of interest, we calculated the average total CPU time for each method based on 20 replicates, all performed on the same computer.

Table 2 summarizes the results based on 1000 replicates. The proposed estimator achieves the same statistical efficiency as the standard estimator. The parameter estimators are not affected by the choice of the perturbation constant, the variance estimators accurately reflect the true variabilities, and the confidence intervals have proper coverage probabilities. The standard method took approximately 2.2 hours, and the average total CPU time for the divide-and-conquer method with $M = 5, 10$ and 20 was 4.2, 6.5, and 9.1 hours. The divide-and-conquer method took more time than the standard method for all choices of M because estimating $\Lambda(t)$ with smaller blocks tended to require more iterations to converge. By contrast, the average total CPU time for the proposed method with $m = 0.05 \times n, 0.1 \times n$, and $0.2 \times n$ was only 0.5, 0.6, and 0.8 hours.

Table 2: Summary statistics for the simulation study under the Cox model with interval-censored data

		$\theta_1 = 0.5$				$\theta_2 = -0.5$			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP
Method		$\times 10^{-4}$	$\times 10^{-2}$	$\times 10^{-2}$	$\times 100\%$	$\times 10^{-4}$	$\times 10^{-2}$	$\times 10^{-2}$	$\times 100\%$
Standard		-0.01	2.59	2.62	96	0.47	4.69	4.47	95
Proposed									
$k = 0.05$	$\delta_n = 0.005$	0.86	2.64	2.63	95	4.99	5.04	4.53	94
	$\delta_n = 0.01$	0.63	2.63	2.63	95	4.02	4.98	4.53	94
$k = 0.1$	$\delta_n = 0.005$	0.55	2.61	2.63	95	1.91	4.87	4.53	95
	$\delta_n = 0.01$	0.57	2.60	2.63	96	1.61	4.84	4.53	95
$k = 0.2$	$\delta_n = 0.005$	0.37	2.60	2.63	96	0.75	4.72	4.52	95
	$\delta_n = 0.01$	0.43	2.59	2.63	96	0.35	4.71	4.52	95
Divide-and-conquer									
$M = 20$		1.68	2.59	2.63	96	-1.41	4.69	4.53	95
$M = 10$		0.89	2.59	2.63	96	-0.55	4.69	4.53	95
$M = 5$		0.48	2.59	2.63	96	-0.07	4.69	4.52	95

Bias, bias of the parameter estimator; SE, standard error of the parameter estimator; SEE, average of the standard error estimator; CP, coverage percentage of the 95% confidence interval.

4.3 Proportional hazards model for clustered failure times

We generated clustered failure times with cumulative hazard function $\Lambda_0(t) \exp(\theta_1 Z_{1ij} + \theta_2 Z_{2ij} + b_{i1} + b_{i2} Z_{1ij}) (j = 1, \dots, n_i; i = 1, \dots, n)$, where Z_{1ij} and Z_{2ij} are independent Bernoulli and uniform $[0, 1]$ variables, b_{i1} and b_{i2} are independently standard normal, $\theta_1 = -\theta_2 = 0.5$, and $\Lambda_0(t) = 0.5t$. We created censoring times from the uniform $[0, 2]$ distribution and set $\tau = 1$, producing approximately 64% censored observations. We sampled n_i from $(2, 3, 4, 5, 6)$ with probabilities $(0.2, 0.3, 0.3, 0.1, 0.1)$ and set $n = 10^6$.

For the standard method, we performed nonparametric maximum likelihood estimation with the entire dataset using the EM algorithm (Zeng and Lin, 2007). For the proposed method, we set $m = kn$, where $k = 0.05, 0.1$ or 0.2 , and $\delta_n = 0.5 \times n^{-1/2}$ or $5 \times n^{-1/2}$. For the divide-and-conquer method, we divided the whole dataset into $M = 5, 10$ or 20 blocks of equal size and made the size of the first block in the proposed method the same as the block size in the divide-and-conquer method. The convergence criterion for the EM algorithm was that the sum of absolute changes in parameter estimates between two successive iterations was less than 10^{-6} . We used 20 nodes for Gaussian quadratures. For variance estimation, we set $h_m = 5 \times m^{-1/2}$ for both the proposed and divide-and-conquer methods, and $h_n = 5 \times n^{-1/2}$ for the standard method. When computing time was of interest, we calculated the average total CPU time for each method based on 10 replicates, all performed on the same computer.

Table 3 summarizes the results based on 1000 replicates. The proposed estimator achieves the same statistical efficiency as the standard estimator. The parameter estimators are robust to the choice of the perturbation constant, the variance estimators accurately reflect the true variabilities, and the confidence intervals have proper coverage probabilities.

Table 3: Summary statistics for the simulation study for clustered failure times†

		$\theta_1 = 0.5$				$\theta_2 = -0.5$			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP
Method		$\times 10^{-3}$	$\times 10^{-3}$	$\times 10^{-3}$	$\times 100\%$	$\times 10^{-3}$	$\times 10^{-3}$	$\times 10^{-3}$	$\times 100\%$
Standard		0.15	2.86	2.80	95	0.09	3.69	3.77	95
Proposed	$k = 0.05 \ \delta_n = 0.0005$	0.08	2.86	2.81	95	0.13	3.69	3.78	94
	$\delta_n = 0.005$	0.10	2.86	2.81	95	0.10	3.69	3.77	95
	$k = 0.1 \ \delta_n = 0.0005$	0.11	2.86	2.80	95	0.11	3.69	3.78	95
	$\delta_n = 0.005$	0.13	2.86	2.80	95	0.10	3.69	3.78	95
	$k = 0.2 \ \delta_n = 0.0005$	0.13	2.86	2.80	95	0.10	3.69	3.77	95
	$\delta_n = 0.005$	0.14	2.86	2.80	95	0.09	3.69	3.77	95
Divide-and-conquer $M = 20$		0.15	2.86	2.81	95	0.09	3.69	3.78	95
$M = 10$		0.15	2.86	2.80	95	0.09	3.69	3.77	95
$M = 5$		0.15	2.86	2.80	95	0.09	3.69	3.77	95

†Bias, bias of the parameter estimator; SE, standard error of the parameter estimator; SEE, average of the standard error estimator; CP, coverage percentage of the 95% confidence interval.

The standard method took approximately 22 hours, and the average total CPU time for the divide-and-conquer method with $M = 5, 10$ and 20 was 25, 31, and 28 hours, respectively. By contrast, the average total CPU time for the proposed method with $k = 0.05, 0.1$ and 0.2 was only 2, 4, and 7 hours, respectively.

5. Application

We considered the UK Biobank, which is a cohort study of approximately half a million subjects (Bycroft et al., 2018). We related interval-censored age-at-onset of hypertension

to biological sex, race, birthplace, the top ten genetic principal components representing population substructures, and the polygenic risk score (PRS) for hypertension summarizing the estimated effects of genetic variants. Sex took value 1 for male and 0 for female, race took value 1 for white and 0 for non-white, birthplace took value 1 for England and 0 for others, and the PRS was obtained from Thompson et al. (2022).

The analysis dataset contained a total of 482507 subjects, with 13779 interval-censored, 258371 left-censored, and 210357 right-censored observations.

We randomly chose $k = 5\%$, 10% or 20% of the entire dataset as the first block and set $\delta_n = 0.5 \times n^{-1/2}$ for the proposed method, and we set $M = 5, 10$ or 20 for the divide-and-conquer method. For variance estimation, we used the profile likelihood approach with $h_m = m^{-1/2}$ for both the proposed and divide-and-conquer methods, and we used $h_n = n^{-1/2}$ for the standard method. The total CPU time for the standard method was approximately 32 hours. For the divide-and-conquer method, it was 44 hours with $M = 5$, 44 hours with $M = 10$ and 50 hours with $M = 20$. By contrast, the total CPU time for the proposed method with $k = 0.05, 0.1$ and 0.2 was only 1.9, 3.5 and 5.8 hours, respectively.

The estimation results are presented in Table S1 of the supplementary materials. The three methods yield highly similar estimates. For the standard method, the estimate of the regression coefficient for PRS is 0.3158 with a standard error of 2.159×10^{-3} . For the divide-and-conquer methods with $M = 5, 10$, and 20 , the coefficient estimates are 0.3158, 0.3160, and 0.3163, with corresponding standard errors of 2.159×10^{-3} , 2.162×10^{-3} , and 2.163×10^{-3} . For the proposed method with $k = 0.05, 0.1$, and 0.2 , the coefficient estimates are 0.3155, 0.3156, and 0.3156, with standard errors of 2.165×10^{-3} , 2.176×10^{-3} , and 2.160×10^{-3} , respectively. We tested the null hypothesis that PRS has no effect on the risk

of developing hypertension. By the proposed method with $k = 0.2$, the PRS has a p-value of less than 2.2×10^{-16} , leading to a strong rejection of the null hypothesis. These results suggest that individuals with higher PRS are more likely to develop hypertension compared to those with lower PRS.

6. Discussion

In this article, we presented a general framework for fitting semiparametric models to big data in a computationally and statistically efficient way. We proposed a one-step estimator with semiparametric efficient score functions to reduce the computation time in estimating the finite-dimensional parameters of interest. The proposed approach does not require an explicit formula of the semiparametric efficient score function or utilize distributed computing.

We introduced a novel way to estimate the semiparametric efficient score function when an analytical expression is unavailable. We established the asymptotic properties of the one-step estimator under mild regularity conditions. We verified the regularity conditions under the partially linear model for a continuous response variable and the Cox model for an interval-censored failure time through innovative use of modern empirical process theory. One important result of our work is that $\hat{\eta}_{\theta} \rightarrow_P -h_0$ as $\theta \rightarrow_P \theta_0$, which is presented in Section S.1 of the Supplementary Materials. This result provides a general approximation of the least favorable direction, which can be applied to a wide variety of semiparametric models.

The technical development of the present paper differs substantially from that of Wang et al. (2024). First, we established a rigorous asymptotic theory for general semiparametric models, for which efficient score functions may not have explicit expressions, whereas

Wang et al. (2024) relied heavily on the explicit form of efficient score functions. Second, we derived the technical conditions under which numerical derivatives of the profile likelihood correctly approximate efficient score functions, whereas Wang et al. (2024) avoided this technical challenge due to the explicit expression of efficient score functions. Finally, we verified that the technical conditions are satisfied for the partially linear model with a continuous response variable and the Cox model with an interval-censored failure time, for which nuisance parameters cannot be estimated in the parametric rate, unlike the case of Wang et al. (2024).

The proposed framework is highly general and readily covers more complex semiparametric problems, such as transformation models for counting processes and joint models for repeated measures and failure times (Zeng and Lin, 2007). For such problems, the standard full-data maximum likelihood estimation with massive sample sizes is computationally prohibitive, and the semiparametric efficient score functions lack analytical expressions. Our approach completely circumvents these challenges by one-step estimation and numerical differences of the profile log-likelihood.

The proposed approach can incorporate multiple blocks of data. Consider K blocks of data indexed by $\mathcal{J}_1, \dots, \mathcal{J}_K$, with sizes $n_1, \dots, n_K$, respectively. We obtain the maximum likelihood estimators of $\boldsymbol{\theta}$ and η using the first block of data, denoted by $\{\tilde{\boldsymbol{\theta}}^{(1)}, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}^{(1)}}^{(1)}\}$. For the remaining $(K - 1)$ blocks, we calculate

$$\tilde{\boldsymbol{\theta}}^{(k)} = \tilde{\boldsymbol{\theta}}^{(1)} + \left\{ n_k \tilde{\mathbf{I}}^{(1)} / n_1 \right\}^{-1} \sum_{i \in \mathcal{J}_k} D_{\boldsymbol{\theta}l} \left\{ \mathbf{O}_i; \tilde{\boldsymbol{\theta}}^{(1)}, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}^{(1)}}^{(1)} \right\}, \quad k = 2, \dots, K,$$

where $\tilde{\mathbf{I}}^{(1)}$ is the information matrix for $\tilde{\boldsymbol{\theta}}^{(1)}$. The proposed estimator is then defined as $\hat{\boldsymbol{\theta}}_f = \sum_{k=1}^K n_k \tilde{\boldsymbol{\theta}}^{(k)} / n$. Mathematically, $\hat{\boldsymbol{\theta}}_f$ is independent of K as long as n_1 remains constant, and thus Theorem 1 holds for the proposed estimator with multiple data blocks.

There are practical benefits for using more than 2 blocks. First, when the data are stored across multiple machines, the proposed approach can be implemented with distributed computing among the $(K - 1)$ machines, and only the estimates $\{\tilde{\boldsymbol{\theta}}^{(1)}, \tilde{\eta}_{\tilde{\boldsymbol{\theta}}^{(1)}}^{(1)}, \tilde{\mathbf{I}}^{(1)}\}$ from the initial analysis are communicated to the $(K - 1)$ machines. Thus, the proposed approach is highly efficient in both computation and communication. Second, data privacy is ensured when results from different sources are integrated because only estimates are communicated among the K blocks. Finally, the proposed procedure enables online learning by incorporating new data without revisiting previous data. By considering the new batch of data as the $(K + 1)$ th block, we can calculate $\tilde{\boldsymbol{\theta}}^{(K+1)}$ and update $\hat{\boldsymbol{\theta}}_f$.

The proposed approach can be extended to high-dimensional covariates. In particular, we may conduct variable selection using the first block of data. After debiasing the initial estimators, we carry out a one-step update on the coefficients of the selected covariates using the remaining data. The computation lies primarily in the first block analysis, which is always less intensive than directly analyzing the entire dataset. We expect that the asymptotic efficiency of the proposed estimators will not be affected by the uncertainty from variable selection.

In practice, big data are likely heterogeneous. We can account for heterogeneity by fitting a model stratified by subpopulations to the initial block of data and calculating the one-step estimator using the remaining data if the initial block contains subjects from all subpopulations. Alternatively, we can compute the one-step estimator separately for each subpopulation and then combine the estimators through meta-analysis (Lin and Zeng, 2010).

Supplementary Materials

The Supplementary Material contains technical details of the proposed methods and additional simulation studies.

References

- Bi, W., L. G. Fritsche, B. Mukherjee, S. Kim, and S. Lee (2020). A fast and accurate method for genome-wide time-to-event data analysis and its application to UK Biobank. *The American Journal of Human Genetics* 107(2), 222–233.
- Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.
- Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O’Connell, et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209.
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric model. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume VI, pp. 5549–5632. New York: Elsevier Science B.V.
- Chen, Z., J. Chen, R. Collins, Y. Guo, R. Peto, F. Wu, and L. Li (2011). China kadoorie niobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology* 40(6), 1652–1666.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1), C1–C68.
- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 34, 187–202.

- Cox, D. R. (1975). Partial likelihood. *Biometrika* 62(2), 269–276.
- Dey, R., W. Zhou, T. Kiiskinen, A. Havulinna, A. Elliott, J. Karjalainen, M. Kurki, A. Qin, FinnGen, S. Lee, et al. (2022). Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nature Communications* 13(1), 5437.
- Gao, F., D. Zeng, D. Couper, and D. Lin (2019). Semiparametric regression analysis of multiple right-and interval-censored events. *Journal of the American Statistical Association* 114(527), 1232–1240.
- Gaziano, J. M., J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, et al. (2016). Million veteran program: a mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology* 70, 214–223.
- Geman, S. and C.-R. Hwang (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics* 10, 401–414.
- Jordan, M. I., J. D. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* 114(526), 668–681.
- Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.
- Lee, J. D., Q. Liu, Y. Sun, and J. E. Taylor (2017). Communication-efficient sparse regression. *The American Journal of Human Genetics* 18(5), 1–30.
- Lin, D.-Y. and D. Zeng (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* 97(2), 321–332.
- Murphy, S. A. and A. W. van der Vaart (2000). On profile likelihood. *Journal of the American Statistical Association* 95(450), 449–465.
- Neiswanger, W., C. Wang, and E. Xing (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, Arlington, pp. pp. 623–632. AUAI Press.

- Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* 45(1), 158 – 195.
- Schifano, E. D., J. Wu, C. Wang, J. Yan, and M.-H. Chen (2016). Online updating of statistical inference in the big data setting. *Technometrics* 58(3), 393–403.
- Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics* (2), 580–615.
- Thompson, D. J., D. Wells, S. Selzam, I. Peneva, R. Moore, K. Sharp, W. A. Tarran, E. J. Beard, F. Riveros-Mckay, C. Giner-Delgado, et al. (2022). Uk biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. *MedRxiv*, 2022–06.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer.
- Wang, J., D. Zeng, and D. Y. Lin (2024). Fitting the Cox proportional hazards model to big data. *Biometrics* 80(1).
- Wang, W., S.-E. Lu, J. Q. Cheng, M. Xie, and J. B. Kostis (2022). Multivariate survival analysis in big data: a divide-and-combine approach. *Biometrics* 78(3), 852–866.
- Wang, Y., C. Hong, N. Palmer, Q. Di, J. Schwartz, I. Kohane, and T. Cai (2021, 09). A fast divide-and-conquer sparse Cox regression. *Biostatistics* 22(2), 381–401.
- Xu, Y., D. Zeng, and D. Y. Lin (2022, 11). Marginal proportional hazards models for multivariate interval-censored data. *Biometrika* 110(3), 815–830.
- Zeng, D. and D. Lin (2010). A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica* 20(2), 871.
- Zeng, D. and D. Y. Lin (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69(4), 507–564.

Zeng, D., L. Mao, and D. Y. Lin (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* 103, 253–271.

Zhang, Y., J. C. Duchi, and M. J. Wainwright (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research* 14(104), 3321–3363.

Zhao, T., G. Cheng, and H. Liu (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics* 44(4), 1400–1437.

Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill,
U.S.A

E-mail: jianqiao@live.unc.edu

Department of Biostatistics, School of Global Public Health, University of Michigan, U.S.A

E-mail: dzeng@umich.edu

Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill,
U.S.A

E-mail: lin@bios.unc.edu