Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Survival Analysis via Ordinary Differential Equations

Weijing Tang, Kevin He, Gongjun Xu & Ji Zhu

To cite this article: Weijing Tang, Kevin He, Gongjun Xu & Ji Zhu (2022): Survival Analysis via Ordinary Differential Equations, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2051519

To link to this article: https://doi.org/10.1080/01621459.2022.2051519

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–16: Theory and Methods
https://doi.org/10.1080/01621459.2022.2051519

Taylor & Francis
Taylor & Francis Group

Survival Analysis via Ordinary Differential Equations

Weijing Tang[a], Kevin He[b], Gongjun Xu[a], and Ji Zhu[a]

[a]Department of Statistics, University of Michigan, Ann Arbor, MI; [b]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI

ABSTRACT

This article introduces an Ordinary Differential Equation (ODE) notion for survival analysis. The ODE notion not only provides a unified modeling framework, but more importantly, also enables the development of a widely applicable, scalable, and easy-to-implement procedure for estimation and inference. Specifically, the ODE modeling framework unifies many existing survival models, such as the proportional hazards model, the linear transformation model, the accelerated failure time model, and the time-varying coefficient model as special cases. The generality of the proposed framework serves as the foundation of a widely applicable estimation procedure. As an illustrative example, we develop a sieve maximum likelihood estimator for a general semiparametric class of ODE models. In comparison to existing estimation methods, the proposed procedure has advantages in terms of computational scalability and numerical stability. Moreover, to address unique theoretical challenges induced by the ODE notion, we establish a new general sieve M-theorem for bundled parameters and show that the proposed sieve estimator is consistent and asymptotically normal, and achieves the semiparametric efficiency bound. The finite sample performance of the proposed estimator is examined in simulation studies and a real-world data example. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2020
Revised November 2021

KEYWORDS
Linear transformation model; Ordinary differential equation; Semiparametric efficiency; Sieve maximum likelihood estimator; Survival analysis; Time varying effects

1. Introduction

Survival analysis is an important branch of statistical modeling, where the primary outcome of interest is the time to a certain event. In practice, event times may not be observed due to a limited observation time window or missing follow-up during the study, which is referred to as censored data. Many statistical models have been developed to deal with censored data in the literature. For example, the Cox proportional hazard model is probably the most classical semiparametric model for handling censored data (Cox 1975), and it assumes that the covariates have a constant multiplicative effect on the hazard function. Although easy to interpret, the constant hazard ratio assumption is often considered as overly strong for real-world applications. As a result, many other semiparametric models have been proposed as attractive alternatives, such as accelerated failure time (AFT) models, transformation models, and additive hazards models. See Aalen (1980), Buckley and James (1979), Gray (1994), Bennett (1983), Cheng, Wei, and Ying (1995), Fine, Ying, and Wei (1998), and Chen, Jin, and Ying (2002) for a sample of references. Given different assumptions made in these semiparametric models, different estimation and inference procedures have also been developed accordingly, such as maximum partial likelihood based estimators (MPLE) (Zucker and Karr 1990; Gray 1994; Bagdonavicius and Nikulin 2001; Chen, Jin, and Ying 2002), least square and rank-based methods (Buckley and James 1979; Tsiatis 1990; Lai and Ying 1991; Jin et al. 2003; Jin, Lin, and Ying 2006), nonparametric maximum likelihood estimators (NPMLE) (Murphy, Rossini, and van der Vaart 1997; Zeng and Lin 2007b), and sieve maximum likelihood estimators (MLE) (Huang 1999; Shen and Wong 1994; Ding and Nan 2011; Zhao, Wu, and Yin 2017).

In this article, we introduce a novel Ordinary Differential Equation (ODE) notion and show that it provides a unified view of aforementioned survival models and, more importantly, facilitates the development of a scalable and easy-to-implement estimation and inference procedure, which can be applied to a wide range of ODE survival models. We note that the proposed approach is founded upon well-established numerical solvers and sensitivity analysis tools for ODEs, and it overcomes various practical limitations of existing estimation methods when applied to different survival models for large-scale studies.

Specifically, the proposed framework models the dynamic change of the cumulative hazard function through an ODE. Let T be the event time and X be covariates. Denote the conditional cumulative hazard function of T given $X = x$ as $\Lambda_x(t)$. Then $\Lambda_x(t)$ is characterized by the following ODE with a fixed initial value

$$\begin{cases} \Lambda_x'(t) = f(t, \Lambda_x(t), x) \\ \Lambda_x(t_0) = c(x) \end{cases}, \tag{1}$$

where the derivative is with respect to t, $f(\cdot)$ and $c(\cdot)$ are functions to be specified, and t_0 is a predefined initial time point. In particular, function $c(\cdot)$ determines the probability of an event occurring after t_0; for instance, $\Lambda_x(0) = 0$ corresponds to the

CONTACT Ji Zhu jizhu@umich.edu Department of Statistics, University of Michigan, Ann Arbor, MI.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

case when no event occurs before time 0. Further, function $f(\cdot)$ determines how covariates x affect the hazard function at time t given an individual's own cumulative hazard. Thus, different specifications of the function $f(\cdot)$ lead to different ODE models.

Next, we comment on both benefits of the ODE approach in terms of modeling and computation and new theoretical challenges induced by the ODE notion.

- First, the ODE modeling framework is general enough to unify many aforementioned existing survival models through different specifications of the function $f(\cdot)$, which serves as the foundation of a widely applicable estimation procedure that will be developed later. For example, the ODE (1) is equivalent to the Cox model when $f(\cdot)$ takes the form $\alpha(t)\exp(x^T\beta)$ for some function $\alpha(\cdot)$, and it is equivalent to the AFT model when $f(\cdot)$ takes the form $q(\Lambda_x(t))\exp(x^T\beta)$ for some function $q(\cdot)$. Similarly, we can obtain many more models such as the time-varying variants of the Cox model, the linear transformation model, and the additive hazards model to name a few (see Section 2 for details). We note that the ODE notion can provide new and sometimes more explicit interpretations in terms of the hazard by rewriting the existing models in the ODE form. In addition, the generality of the proposed framework offers an opportunity for designing more flexible model structures and model diagnostics.
- Second, and also more importantly, introducing the ODE notion facilitates the development of a general and easy-to-implement procedure for estimation and inference in large-scale survival analysis. In this article, we illustrate the proposed procedure by using a general class of ODE models as an example. In particular, this general class includes the most flexible linear transformation model, where both the transformation function and the error distribution are unspecified. Since the $f(\cdot)$ function for the general model contains both finite-dimensional and infinite-dimensional parameters, we propose a spline-based sieve MLE that directly maximizes the likelihood in a sieve space. We provide an easy-to-implement gradient-based optimization algorithm founded upon *local sensitivity analysis* tools for ODEs (Dickinson and Gelinas 1976), where numerical ODE solvers are used to compute the log-likelihood function and its gradients. Since efficient implementations of both ODE solvers and splines are available in many software, the resultant algorithm is easy to carry out in practice. It is worth noting that, in comparison to existing estimation methods, the proposed procedure has advantages in various aspects, such as scalability against MPLE for the time-varying Cox model, optimization-parameter efficiency against NPMLE, statistical efficiency and numerical stability against rank-based methods for the linear transformation model. We demonstrate these advantages through extensive simulation studies. For example, when the sample size is 8000, it takes the proposed ODE approach about 6 sec to estimate the semiparametric ODE-AFT model while the rank-based method needs 350 sec.
- Finally, we note that the ODE notion brings new challenges to asymptotic distributional theory. While many asymptotic distributional theories for M-estimation in semiparametric models have been developed (see Huang 1999; Shen 1997; Ai and Chen 2003; Wellner and Zhang 2007; Zhang, Hua, and Huang 2010; He, Xue, and Shi 2010; Ding and Nan 2011 for a sample of references), they cannot be directly applied to our setting. Among them, the proposed theory in Ding and Nan (2011) considers bundled parameters where the infinite-dimensional parameter is an unknown function of the finite-dimensional Euclidean parameter and has been applied to the AFT model, and recently, to the accelerated hazards model in Zhao, Wu, and Yin (2017). However, for the general class of ODE models, the estimation criterion is parameterized with more general bundled parameters where the nuisance parameter is an unknown function of not only finite-dimensional regression parameters of interest but also other infinite-dimensional nuisance parameters. To accommodate this different and challenging scenario induced by the ODE notion, we develop a new sieve M-theorem for more general bundled parameters. By applying it to the general class of ODE models along with ODE related methodologies (Walter 1998), we show consistency, asymptotic normality, and semiparametric efficiency for the estimated regression parameters. The proposed theory can also be extended to develop the asymptotic normality of estimators for other ODE models.

The rest of the article is organized as follows. We introduce the ODE framework and present a general class of ODE models as special cases in Section 2. We provide the estimation procedure in Section 3 and establish theoretical properties in Section 4. Simulation studies and a real-world data example are presented in Sections 5 and 6, respectively.

2. The ODE Framework

To characterize the conditional distribution of T given X, the conditional hazard function, denoted as $\lambda_x(t) = \Lambda_x'(t)$, provides a popular modeling target as it describes the instantaneous rate at which the event occurs given survival. In this article, we view the hazard function as the dynamic change of the cumulative hazard function and quantify them using an ODE.

In our ODE framework, the hazard function depends not only on the time and covariates but also on the cumulative hazard as shown in (1), where function $f(\cdot)$ specifies the dynamic change of $\Lambda_x(t)$ and covariates x serve as additional parameters in terms of the ODE. The initial value in (1) implies that, for an individual with covariates x, the probability for an event to occur after t_0 is controlled by $\exp(-c(x))$. For example, it is often the case that time 0 is defined prior to the occurrence of events, which implies that an event always occurs after time 0, that is, the survival function $S_x(0) = 1$, and it follows that $\Lambda_x(0) = 0$. We use this initial value in the ODE framework hereafter for simplicity, while the estimation method and the theoretical properties established later can be extended to the general case where $c(x)$ can be a function of covariates. Under certain smoothness conditions (Walter 1998, p. 108), the initial value problem (1) has exactly one solution, which uniquely characterizes the conditional distribution of the event time.

Next, we present a general class of ODE models as an instantiation of the ODE framework. Suppose there are two groups of

covariates denoted by $X \in \mathbf{R}^{d_1}$ and $Z \in \mathbf{R}^{d_2}$, respectively. We consider ODE models in the form of

$$\Lambda'_{x,z}(t) = \alpha(t)\exp(x^T\beta + z^T\boldsymbol{\eta}(t))q(\Lambda_{x,z}(t)), \tag{2}$$

where $\alpha(\cdot)$ and $q(\cdot)$ are two unknown positive functions, and given an individual's own cumulative hazard, both covariates x and z have multiplicative effects on the hazard, one with time-independent coefficients $\beta \in \mathbf{R}^{d_1}$ and the other with time-varying coefficients $\boldsymbol{\eta}(t) \in \mathbf{R}^{d_2}$. Here $\boldsymbol{\eta}(\cdot) = (\eta_1(\cdot), \ldots, \eta_{d_2}(\cdot))^T$.[1] We note that this general class of ODE models is a specific example; other examples beyond this class are included in Remark 2 to further illustrate the flexibility of the proposed ODE framework. In particular, this general class covers many existing models as special cases. As shown below, model (2) reduces to the time-varying Cox model when $q(\cdot) = 1$, to the linear transformation model when covariates z are not considered, and further reduces to the AFT model if $\alpha(\cdot) = 1$. In the following sections, we will also show that by rewriting many existing models under the format (1), the ODE framework brings them new interpretations in terms of the hazard function.

2.1. Cox Model and Time-Varying Cox Model

The Cox proportional hazard model assumes that the covariates have a multiplicative effect on the hazard function, that is, $\lambda_x(t) = \alpha(t)\exp(x^T\beta)$, where $\alpha(t)$ is a baseline hazard function and $\exp(x^T\beta)$ is the relative risk, and extensions of the Cox model allow for time-varying coefficients (Zucker and Karr 1990; Gray 1994). Here we write the Cox model with both time-independent and time-varying effects as a simple ODE, whose right-hand side does not depend on the cumulative function, that is

$$\Lambda'_{x,z}(t) = \alpha(t)\exp(x^T\beta + z^T\boldsymbol{\eta}(t)), \tag{3}$$

which allows covariates x to have time-independent effects and covariates z to have time-varying effects on the hazard function. The baseline hazard function $\alpha(t)$ and time-varying effects $\boldsymbol{\eta}(t)$ can be specified in a parametric model or left unspecified in a semiparametric model.

2.2. Accelerated Failure Time Model

The AFT model assumes that the log transformation of T is linearly correlated with covariates, that is, $\log T = -X^T\beta + \epsilon$. In the proposed ODE framework, the AFT model can be written as

$$\Lambda'_x(t) = q(\Lambda_x(t))\exp(x^T\beta), \tag{4}$$

where the function $q(\cdot)$ uniquely determines the distribution of the error ϵ in the following way. Let $H_q(u) = \int_0^{-\ln u} q^{-1}(v)dv$ and $G_q(u) = H_q^{-1}(u)$, then G_q is the survival function of $\delta = \exp(\epsilon)$ as shown in Bagdonavicius and Nikulin (2001). For example, if $q(t) = vk^{\frac{1}{v}}t^{1-\frac{1}{v}}$, then δ follows a Weibull distribution with $G_q(t) = \exp(-kt^v)$. When the error distribution is unknown (as in a semiparametric AFT model), we can leave the function $q(\cdot)$ unspecified.

The ODE (4) provides a new and clear interpretation on how covariates affect the hazard for the AFT model. Specifically, it implies that given an individual's own cumulative hazard, covariates x have a multiplicative constant effect on the hazard function. Further, besides the direct effects of covariates, if $q(\cdot)$ is a monotonic increasing function, then an individual with a higher cumulative hazard at a particular time would have a higher "baseline" hazard. Note that although we can also present the hazard directly as a function of covariates and time, that is, $\lambda_x(t) = \lambda_\delta(t\exp(x^T\beta))\exp(x^T\beta)$, the covariate effects are entangled with the baseline hazard λ_δ in this representation, which is more difficult to interpret.

2.3. Linear Transformation Model

As an extension of the AFT model, the linear transformation model assumes that, after a monotonic increasing transformation $\varphi(\cdot)$, the event time T is linearly correlated with covariates, that is, $\varphi(T) = -X^T\beta + \epsilon$. In the proposed ODE framework, it can be written as

$$\Lambda'_x(t) = q(\Lambda_x(t))\exp(x^T\beta)\alpha(t), \tag{5}$$

where $q(\cdot)$ corresponds to the distribution of ϵ in the same way as in the AFT model, and $\alpha(\cdot)$ is uniquely determined by the equation $\varphi(t) = \log\int_0^t \alpha(s)ds$. In comparison to model (4), the hazard function at time t depends not only on the current cumulative hazard and covariates, but also on the current time t directly.

Different specifications of $\varphi(\cdot)$ and ϵ have been proposed in the literature for the linear transformation model. We consider the case where both the transformation and the error distribution are unknown. This specification is especially preferred when parametric assumptions on the transformation function or the error distribution cannot be properly justified. However, when both $q(\cdot)$ and $\alpha(\cdot)$ are unknown, they may not be identifiable. The equivalent linear regression representation, $\varphi(T) = -x^T\beta + \epsilon$, allows us to see the identifiability issue clearly. Note that, when no covariate is associated with survival, that is, $\beta = 0$, nonidentifiability issue arises because parameters (φ, ϵ) and $(f(\varphi), f(\epsilon))$ give the same event time distribution for any arbitrary function f. Therefore, we consider $\beta \neq 0$, in which case Horowitz (1996) showed that the model parameters are identifiable up to a scale and a location normalization under certain regularity conditions. Following that result, we have developed Proposition 1 that characterizes the identifiability of parameters in (5), while Proposition 2 provides necessary and sufficient degeneration conditions for AFT and Cox models. The proofs are given in the supplementary materials.

Proposition 1. Suppose at least one of the covariates in x is continuous and this covariate has a nonzero β coefficient, which without loss of generality is assumed to be positive. Let $(q(\cdot), \beta, \alpha(\cdot))$ specify the survival distribution through (5). Then for any other $(\tilde{q}(\cdot), \tilde{\beta}, \tilde{\alpha}(\cdot))$ that gives the same survival distribution, if and only if there exist positive constants c_1

[1] Throughout this article, we bold vectors only when each element is a function.

and c_2 such that $\tilde{\beta} = c_1\beta$, $\int_0^t \tilde{\alpha}(s)ds = c_2(\int_0^t \alpha(s)ds)^{c_1}$, and $\int_0^t \tilde{q}^{-1}(s)ds = c_2(\int_0^t q^{-1}(s)ds)^{c_1}$ for any $t > 0$.

Proposition 2. Suppose the conditions in Proposition 1 hold, then the linear transformation model in (5) coincides with the Cox model if and only if there exist positive constants c_1 and c_2 such that $q(u) = c_2u^{1-c_1}$, and it coincides with the AFT model if and only if there exist positive constants c_1 and c_2 such that $\alpha(t) = c_2t^{c_1-1}$ for $t > 0$.

Remark 1. Note that the original forms of the AFT model and the linear transformation model do not directly take time-varying coefficients. Existing works on the linear transformation model that consider varying coefficients choose to model them as a function of certain covariates rather than a function of time (Chen and Tong 2010; Qiu and Zhou 2015). In contrast, the equivalent ODE forms of the AFT model in (4) and the linear transformation model in (5) can naturally accommodate time-varying coefficients. For example, we can consider the generalization in (2), where given an individual's own cumulative hazard covariates z have time-varying multiplicative effects $\boldsymbol{\eta}(t)$ on the hazard. In particular, this generalization is equivalent to a covariate-dependent transformation model

$$\varphi_Z(T) = -X^T\beta + \epsilon,$$

where $\varphi_z(t) = \log\int_0^t \alpha(s)\exp(z^\top\boldsymbol{\eta}(s))ds$, that is, covariates z have multiplicative time-varying effect $\boldsymbol{\eta}(t)$ on the gradient of $\exp(\varphi_z(t))$.

Remark 2. The proposed ODE framework is general enough to cover other existing models as well. For example, both the additive hazard model (Aalen 1980; Mckeague and Sasieni 1994) and the additive-multiplicative hazard model (Lin and Ying 1995) can be viewed as a specific ODE model, that is, $\Lambda'_{x,z}(t) = r_1(x^T\beta)+\alpha(t)r_2(z^T\eta)$, where $r_1(\cdot)$ and $r_2(\cdot)$ are some known link functions. Subsequently, the generalized additive hazards model and the generalized additive-multiplicative hazards model (Bagdonavicius and Nikulin 2001) can be written as $\Lambda'_x(t) = q(\Lambda_x(t))(r_1(x) + \alpha(t)r_2(x))$. The generalized Sedyakin's model (Bagdonavicius and Nikulin 2001), which was proposed as an extension of the AFT model, can also be viewed as a special case of (1) with $\Lambda'_x(t) = f(\Lambda_x(t), x)$.

Remark 3. Further, the proposed ODE framework and the estimation method in Section 3 can also be extended to deal with time-varying covariates. Suppose the covariate is a stochastic process $X(t), t \geq 0$ and $T_{X(\cdot)}$ is the failure time under $X(\cdot)$. Denote the conditional survival, the hazard function, and the cumulative function by $S_{x(\cdot)}(t) = P\{T_{X(\cdot)} \geq t|X(s) = x(s), 0 \leq s \leq t\}$, $\lambda_{x(\cdot)}(t) = -\frac{S'_{x(\cdot)}(t)}{S_{x(\cdot)}(t)}$, and $\Lambda_{x(\cdot)}(t) = -\log(S_{x(\cdot)}(t))$, respectively. Then the ODE (1) can be extended to $\Lambda'_{x(\cdot)}(t) = f(t, \Lambda_{x(\cdot)}(t), x(t))$. This extension also covers many existing models as special cases. For example, the linear transformation model with time-varying covariates (Zeng and Lin 2006) can be written as $\Lambda'_{x(\cdot)}(t) = q(\Lambda_{x(\cdot)}(t))\exp(x(t)^T\beta)\alpha(t)$, and the Cox model with time-varying covariates can be viewed as a special case with $q(\cdot) \equiv 1$. For presentation simplicity, we focus on models in the form of (2) in this article.

2.4. Related Estimation Methods and their Limitations

The maximum partial likelihood estimator (MPLE) (Cox 1975) was first proposed for the Cox model, and the asymptotic property of MPLE was established by Andersen and Gill (1982) via the counting process martingale theory. For time-varying Cox models, many different estimation methods have been developed while relying on maximizing the partial likelihood (Zucker and Karr 1990; Gray 1994). However, evaluating the partial likelihood for an uncensored individual requires access to all other observations who were in its risk set. This prevents parallel computing for partial likelihood-based methods, which is a drawback when analyzing large-scale data.

For the linear transformation model, different specifications of the transformation and the error distribution along with different estimation methods have been proposed. For example, Cheng, Wei, and Ying (1995), Fine, Ying, and Wei (1998), Shen (1998), Chen, Jin, and Ying (2002), and Bagdonavicius and Nikulin (1999) have considered an unknown transformation with a known error distribution, which includes the Cox model and the proportional odds model (Bennett 1983) as special cases. The corresponding modified MPLE (Chen, Jin, and Ying 2002; Bagdonavicius and Nikulin 1999), sieve MLE (Shen 1998), and NPMLE (Murphy, Rossini, and van der Vaart 1997; Zeng and Lin 2007b) have also been developed. However, due to the large number of nuisance parameters, it is difficult to obtain NPMLE in practice, especially in large-scale applications. Alternatively, Cai, Tian, and Wei (2005) considered a parametric Box-Cox transformation with an unknown error distribution, which includes the semiparametric AFT model as a special case, and least square and rank-based methods have been proposed to estimate the regression parameters (Buckley and James 1979; Lai and Ying 1991; Tsiatis 1990; Jin et al. 2003; Jin, Lin, and Ying 2006). Nevertheless, they are not asymptotically efficient and may suffer additional numerical errors resulting from discrete objective functions. Subsequently, under the AFT model, Zeng and Lin (2007a) and Lin and Chen (2012) proposed efficient estimators based on a kernel-smoothed profile likelihood, and Ding and Nan (2011) developed an efficient sieve MLE. When both the transformation function and the error distribution are unknown, a partial rank-based method has been proposed (Song et al. 2006; Khan and Tamer 2007), and its computation is analogous to that of the partial likelihood, where the rank of an uncensored individual is determined by all other individuals in its risk set, and thus, the computational challenge for large-scale applications still remains.

As evident from the above discussion, many existing estimation methods suffer from important limitations in practice. In Section 3, we propose a scalable, easy-to-implement and efficient estimation method that can be applied to a wide range of models.

3. Maximum Likelihood Estimation

In this section, we propose a general estimation procedure that can be applied to a wide range of ODE models. Here we use the ODE model in (2) as an illustrative example, and the proposed

estimation method can also be applied to other models such as those mentioned in Remark 2.

We denote the event time as T, the censoring time as C. Let $Y = \min\{T, C\}$ and $\Delta = \mathbb{1}(T \le C)$, where $\mathbb{1}(\cdot)$ denotes the indicator function. Our data consist of n independent and identically distributed observations $\{Y_i, \Delta_i, X_i, Z_i\}$, $i = 1, \ldots, n$. Since $\alpha(\cdot)$ and $q(\cdot)$ in (2) are positive, we set $\gamma(\cdot) = \log\alpha(\cdot)$ and $g(\cdot) = \log q(\cdot)$. Under the conditional independence between T and C given covariates (X, Z), the log-likelihood function of the parameters $(\beta, \gamma(\cdot), \boldsymbol{\eta}(\cdot), g(\cdot))$ is given by

$$\begin{aligned} l_n(\beta, \gamma(\cdot), g(\cdot), \boldsymbol{\eta}(\cdot)) = \frac{1}{n}\sum_{i=1}^{n}[\Delta_i\{\gamma(Y_i) + X_i^T\beta + Z_i^T\boldsymbol{\eta}(Y_i) \\ + g(\Lambda_i(Y_i; \beta, \gamma, g, \boldsymbol{\eta}))\} \\ - \Lambda_i(Y_i; \beta, \gamma, \boldsymbol{\eta}, g)], \end{aligned} \tag{6}$$

where $\Lambda_i(t; \beta, \gamma, \boldsymbol{\eta}, g)$ denotes the solution of ODE (2) parameterized by $(\beta, \gamma, \boldsymbol{\eta}, g)$ given covariates $X = X_i$ and $Z = Z_i$. The log-likelihood function (6) includes both finite-dimensional parameter β and infinite-dimensional parameters $\gamma, \boldsymbol{\eta}, g$.

We propose a sieve MLE that maximizes the log-likelihood over a sequence of finite-dimensional parameter spaces that are dense in the original parameter space as the sample size increases. The sieve space can be chosen as linear spans of many types of basis functions with desired properties (Chen 2007). In particular, we construct the sieve space using polynomial splines due to their capacity in approximating complex functions and the simplicity of their construction. Under suitable smoothness conditions, $\gamma_0(\cdot)$, $\boldsymbol{\eta}_0(\cdot)$, and $g_0(\cdot)$, the true parameters associated with the data generating distribution, can be well approximated by some functions in the space of polynomial splines as defined in Schumaker (2007, p. 108, Definition 4.1). Further, there exists a group of spline bases such that functions in the space of polynomial splines can be written as linear combinations of the spline bases (Schumaker 2007, p. 117, Corollary 4.10). Different groups of spline bases may be used for the estimation of different parameters $(\gamma, \boldsymbol{\eta})$ and g because of their different domains.

Specifically, we construct the proposed sieve estimator as follows. Let $\mathcal{B} \subset \mathbf{R}^{d_1}$ be the parameter space of β. Let $\{B_j^1, 1 \le j \le q_n^1\}$ and $\{B_j^2, 1 \le j \le q_n^2\}$ be two groups of spline bases that are used for the estimation of parameters $(\gamma, \boldsymbol{\eta})$ and g, respectively. Here the number of spline bases, q_n^i, should grow sublinearly in rate $O(n^{v_i})$ for some $v_i \in (0, 0.5)$, $i = 1, 2$ for convergence guarantee (see Section 4 for rigorous definitions). Overall, we wish to find $d_2 + 1$ members $(\gamma, \eta_1, \ldots, \eta_{d_2})$ from the space of polynomial splines associated with $\{B_j^1\}$, one member g from that associated with $\{B_j^2\}$, along with $\beta \in \mathcal{B}$ to maximize the log-likelihood function (6). Let $Z_{i0} = 1$, $Z_i = (Z_{i1}, \ldots, Z_{id_2})^T$. Then the objective function can be written as

$$\begin{aligned} l_n(\beta, a, b) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\Delta_i\{X_i^T\beta + \sum_{l=0}^{d_2}\sum_{j=1}^{q_n^1} a_j^l B_j^1(Y_i)Z_{il} \\ + \sum_{j=1}^{q_n^2} b_j B_j^2(\Lambda_i(Y_i; \beta, a, b))\} - \Lambda_i(Y_i; \beta, a, b)\Bigg], \end{aligned} \tag{7}$$

where $a = \left(a_j^l\right)_{j=1,\ldots,q_n^1, l=0,\ldots,d_2}$ and $b = \left(b_j\right)_{j=1,\ldots,q_n^2}$ are the coefficients of the spline bases, and $\Lambda_i(t; \beta, a, b)$ is the solution of

$$\begin{cases} \Lambda_i'(t) = \exp(X_i^T\beta + \sum_{l=0}^{d_2}\sum_{j=1}^{q_n^1} a_j^l B_j^1(t)Z_{il} \\ \qquad + \sum_{j=1}^{q_n^2} b_j B_j^2(\Lambda_i(t))), \\ \Lambda_i(0) = 0. \end{cases} \tag{8}$$

The proposed sieve estimators are given by $\hat{\beta}_n = \hat{\beta}$, $\hat{\boldsymbol{\eta}}_n(\cdot) = \left(\sum_{j=1}^{q_n^1}\hat{a}_j^1 B_j^1(\cdot), \ldots, \sum_{j=1}^{q_n^1}\hat{a}_j^{d_2} B_j^1(\cdot)\right)$, $\hat{\gamma}_n(\cdot) = \sum_{j=1}^{q_n^1}\hat{a}_j^0 B_j^1(\cdot)$, and $\hat{g}_n(\cdot) = \sum_{j=1}^{q_n^2}\hat{b}_j B_j^2(\cdot)$, where $(\hat{\beta}, \hat{a}, \hat{b})$ maximizes the objective function (7).

Note that the objective function (7) contains the solution of a parameterized ODE (i.e., (8)), and this is different from most traditional optimization problems. In particular, it is nontrivial to evaluate the objective function and its gradient with respect to parameters when there is no closed-form solution for the ODE. To address this optimization challenge, we develop a gradient-based optimization algorithm by taking advantage of local sensitivity analysis (Dickinson and Gelinas 1976; Petzold et al. 2006) and well-implemented ODE solvers. Specifically, we evaluate the objective function and its gradient as follows:

1. we numerically calculate $\Lambda_i(Y_i; \beta, a, b)$ by solving (8) given the current parameter estimates β, a, b and covariates X_i, Z_i, the initial value at $t_0 = 0$, and the evaluating time $t = Y_i$;
2. we evaluate the derivative of $\Lambda_i(Y_i; \beta, a, b)$ with respect to the parameters β, a, and b through solving another ODE which is derived by local sensitivity analysis, and calculate the gradient of the objective function by the chain rule.

We summarize the results of the local sensitivity analysis in the following, and provide detailed derivations in the supplementary materials. The local sensitivity analysis is a technique that studies the rate of change in the solution of an ODE system with respect to the parameters. There are two ways to obtain the sensitivity: forward sensitivity analysis and adjoint sensitivity analysis. Both of them require solving another ODE with some fixed initial value. For example, we consider to compute the gradient of $\Lambda(y; \theta)$ with respect to its parameter θ, where $\Lambda(t; \theta)$ is the solution of (8) and θ consists of parameters β, a, and b in our case. For presentation simplicity, we denote the right-hand side of (8) by the function $f(t, \Lambda; \theta)$, that is

$$f(t, \Lambda; \theta) = \exp(X^T\beta + \sum_{l=0}^{d_2}\sum_{j=1}^{q_n^1} a_j^l B_j^1(t)Z_j + \sum_{j=1}^{q_n^2} b_j B_j^2(\Lambda)),$$

and its partial derivative with respect to θ and Λ by f_θ' and f_Λ', respectively. In forward sensitivity analysis, it can be shown that the partial derivative of $\Lambda(y; \theta)$ with respect to θ is given by the solution of (9) at $t = y$, that is, $\Lambda_\theta'(y; \theta) = F_1(y)$ with F_1 satisfying

$$\begin{cases} F_1'(t) = f_\theta'(t, \Lambda; \theta) + f_\Lambda'(t, \Lambda; \theta)F_1, \\ F_1(0) = 0. \end{cases} \tag{9}$$

In the alternative adjoint sensitivity analysis, we can show that the partial derivative can also be obtained by evaluating the

solution of (10) at $t = 0$, that is, $\Lambda'_\theta(y;\theta) = F_2(0)$ with F_2 satisfying

$$\begin{cases} (\kappa(t); F_2'(t)) = (-\kappa \cdot f'_\Lambda(t, \Lambda; \theta); -\kappa \cdot f'_\theta(t, \Lambda; \theta)), \\ (\kappa(t); F_2(t))|_{t=y} = (1; \mathbf{0}). \end{cases} \tag{10}$$

Thus, after plugging the form of $f(t, \Lambda; \theta)$ into either (9) or (10), we can obtain the gradients through solving the corresponding ODE. In Remark 4, we compare the computational complexity of forward and adjoint sensitivity analyses and provide a general guidance on which sensitivity analysis to use when computing gradients under survival ODE models.

It is worth noting that the proposed estimation method can be easily implemented using existing computing packages. For example, the "Optimization Toolbox" in MATLAB contains "fminunc" for unconstrained optimization and "fmincon" for constrained optimization; both require initialization and the objective function. In our implementation, we also provide evaluation of the gradient for faster and more reliable computations. In particular, we compute both the objective function and the gradient by well-implemented ODE solvers in MATLAB. In addition, we construct the sieve space using B-splines for its numerical simplicity, whose implementation is available in the "Curve Fitting Toolbox."

Remark 4. In general, forward sensitivity analysis is computationally more efficient when the dimension of the ODE system is relatively large and the number of parameters is small, while adjoint sensitivity analysis is best suited in the complementary scenario. See Dickinson and Gelinas (1976) and Petzold et al. (2006) for more details. For a general ODE model such as (1) where the size of the ODE system is 1 and the number of parameters increases as the sample size n grows, we can use the adjoint sensitivity analysis along with parallel computing for n independent individuals. Alternatively, if the memory permits, we can combine ODEs for n individuals into a large ODE system with n dimensions, which is larger than the number of parameters, and then the forward sensitivity analysis is preferred.

Remark 5. Moreover, we introduce a computational trick for the general class of ODE models in (2) that can significantly accelerate the evaluation of the objective and gradients, where we need to solve ODEs for n independent individuals. Specifically, the trick transforms the problem of solving n different ODEs at their respective observed times into a problem of solving a single ODE at n different time points. More generally, this trick can be applied to any ODE model where the right-hand side is separable in the way that $f(t, \Lambda_x; \theta, x) = f_1(t; \theta, x) f_2(\Lambda_x; \theta)$ with two functions f_1 and f_2. We refer to the supplementary materials for more details about this computational trick.

Remark 6. The proposed sieve MLE can also be applied to many existing models. For example, for the time-varying Cox model where $q(\cdot) = 1$, we can remove the function $g(\cdot)$ from the objective function (6). For the semiparametric AFT model where Z is not considered and $\alpha(\cdot) = 1$, we can just keep parameters β and $g(\cdot)$ in (6). For the linear transformation model, if either $q(\cdot)$ or $\alpha(\cdot)$ is specified, we can replace the corresponding term in (6) with the specified finite-dimensional parametric form. Also note that in comparison to existing estimation methods in Section 2.4, the proposed estimation method allows parallel computing, which is especially important for large-scale applications. Specifically, since the log-likelihood of each individual only depends on its own observations, the evaluation for independent data points can be carried out simultaneously. Further, compared with the NPMLE where the number of optimization parameters is linear in n (Murphy, Rossini, and van der Vaart 1997; Zeng and Lin 2007b), the number of optimization parameters used in sieve MLE increases more slowly with the sample size.

Remark 7. The objective function (7) is convex with respect to β and a for the (time-varying) Cox model, where the parameter b is not included, and the global optimum can be achieved quickly. For the general case, the objective function is nonconvex and the optimization algorithm may converge to a local optimum. Nevertheless, based on our extensive simulation studies, the algorithm generally performs well with appropriately chosen initialization, such as initializing the algorithm with the estimates from the Cox model.

Remark 8. Note that different identifiability conditions are required for different survival models. Thus, we need to add corresponding constraints in the optimization algorithm.

- For the general ODE model (2) where both covariates X (with time-independent effects) and Z (with at least one nonzero time-varying effect) are considered, two groups of parameters $(\beta, \gamma, g, \boldsymbol{\eta})$ and $(\tilde{\beta}, \tilde{\gamma}, \tilde{g}, \tilde{\boldsymbol{\eta}})$ give the same survival distribution if and only if $\beta = \tilde{\beta}$, $\gamma = \tilde{\gamma} + c$, $g = \tilde{g} - c$, and $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$ for some constant c. To guarantee the identifiability, we can constrain either the value of $\gamma(\cdot)$ at a fixed time point t^* or the norm of $\gamma(\cdot)$, in which the former leads to a linear constraint on the coefficients of spline bases.
- For the linear transformation model where the time-varying effects are not considered and at least one component of X has a nonzero coefficient, parameters (β, γ, g) are identifiable up to two scaling factors as shown in Proposition 1. To guarantee identifiability, we can put constraints on β and γ. For β, we can either constrain the first element of β to be 1 (Song et al. 2006; Khan and Tamer 2007), which can be naturally achieved by arranging covariates if we know which covariate has a nonzero effect, or set $\|\beta\| = 1$. For γ, we can add a similar constraint as that for the general ODE model (2). Alternatively, we can put constraints on γ and g by setting $\int_0^{t^*} \exp(\gamma(s))ds = c_1$ and $\int_0^{t^*} \exp(-g(s))ds = c_2$, with some positive constants $c_1 \neq c_2 > 0$ and a fixed time point t^*. In our implementation, we choose to use two linear constraints, that is, set the first element of β to 1 and $\gamma(t^*) = 0$ for simplicity in optimization.

4. Theoretical Properties

In this section, we study the theoretical properties of the proposed sieve MLE. Although many works have investigated asymptotic distributional theories for M-estimation with bundled parameters (Ai and Chen 2003; Chen, Linton, and Van Keilegom 2003; Ding and Nan 2011), their results cannot be directly applied to our setting. In particular, the nuisance

parameters in existing works often take the form of an unknown function of only some finite-dimensional Euclidean parameters of interest. However, our work focuses on a more general scenario, where the nuisance parameter is an unknown function of not only the Euclidean parameters but also some other infinite-dimensional nuisance parameters. To deal with theoretical challenges due to the additional functional nuisance parameters, we develop a new sieve M-theorem for the asymptotic theory of a general family of semiparametric M-estimators. Moreover, we apply the proposed general theorem to establish the asymptotic normality and semiparametric efficiency of the proposed sieve MLE $\hat{\beta}_n$ when the convergence rate of the sieve estimator of the nuisance parameter can be slower than $\sqrt{n}$. We present regularity conditions and main theorems in this section and give all the proofs in the supplementary materials.

For the simplicity of notation, we focus on model (2) without covariates Z, that is, the linear transformation model (5), and the results can be similarly extended to the general case with additional regularity conditions on Z (see Remark 11). Recall that we have set $\gamma(\cdot) = \log\alpha(\cdot)$ and $g(\cdot) = \log q(\cdot)$ to ensure the positivity of $\alpha(\cdot)$ and $q(\cdot)$ in (5). Then we reformulate the ODE model as follows,

$$\begin{cases} \Lambda'(t) = \exp(x^T\beta + \gamma(t) + g(\Lambda(t))) \\ \Lambda(0) = 0 \end{cases}. \tag{11}$$

Note that the parameter β is identifiable when time-varying effects are considered, but in (11) it is identifiable only up to a scaling factor when both γ and g are unknown as shown in Proposition 1. To guarantee the identifiability, we constrain the first element of β to be 1 and $\gamma(t^*) = c$ with some constant c for simplicity in optimization. Specifically, denote $X = (X_{(1)}, X_{(-1)})$, $\beta = (1, \tilde{\beta}^T)^T$, $\bar{\gamma}(\cdot) = \gamma(\cdot) - \gamma(t^*)$ with $\bar{\gamma}(t^*) \equiv 0$, and $\bar{X}_{(1)} = X_{(1)} + \gamma(t^*)$, then we have $X^T\beta + \gamma(t) = \bar{X}_{(1)} + X_{(-1)}^T\tilde{\beta} + \bar{\gamma}(t)$. We substitute $\tilde{\beta}$, $\bar{\gamma}$, and $\bar{X}_{(1)}$ by β, γ, and $X_{(1)}$, respectively for notational simplicity hereafter, and the ODE (11) is then equivalent to

$$\begin{cases} \Lambda'(t) = \exp(x_{(1)} + x_{(-1)}^T\beta + \gamma(t) + g(\Lambda(t))) \\ \Lambda(0) = 0 \end{cases}, \tag{12}$$

with $\gamma(t^*) \equiv 0$. Before stating the regularity conditions, we first introduce some notations. We denote the solution of (12) by $\Lambda(t, x, \beta, \gamma, g)$ to explicitly indicate that the solution of (12) depends on covariates x and parameters (β, γ, g). We denote the true parameters associated with the data generating distribution by (β_0, γ_0, g_0) and simplify $\Lambda(t, x, \beta_0, \gamma_0, g_0)$ as $\Lambda_0(t, x)$. In addition, some commonly used notations in the empirical process literature will be used in this section as well. Let $Pf = \int f(x)Pr(dx)$, where Pr is a probability measure, and denote the empirical probability measure as $\mathbb{P}_n$.

Then we assume the following regularity conditions.

(C1) The true parameter β_0 is an interior point of a compact set $\mathcal{B} \subset \mathbf{R}^d$.

(C2) The density of X is bounded below by a constant $c > 0$ over its domain $\mathcal{X}$, which is a compact subset of $\mathbf{R}^{d+1}$, and $P(X_{(-1)}X_{(-1)}^T)$ is nonsingular.

(C3) There exists a truncation time $\tau < \infty$ such that, for some positive constant δ_0, $Pr(Y > \tau|X) \geq \delta_0$ almost surely with respect to the probability measure of X. Then there is a constant $\mu = \sup_{x\in\mathcal{X}} \Lambda_0(\tau, x) \leq -\log\delta_0$ such that $\Lambda_0(\tau, X) = -\log Pr(T > \tau|X) \leq \mu$ almost surely with respect to the probability measure of X.

(C4) Let $S^p([a, b])$ be the collection of bounded functions f on $[a, b]$ with bounded derivatives $f^{(j)}$, $j = 1, \ldots, k$, where the kth derivative $f^{(k)}$ satisfies the m-Hölder continuity condition:

$$|f^{(k)}(s) - f^{(k)}(t)| \leq L|s - t|^m \quad \text{for } s, t \in [a, b],$$

where k is a positive integer and $m \in (0, 1]$ with $p = m+k$, and $L < \infty$ is a constant. The true function $\gamma_0(\cdot)$ belongs to $\Gamma^{p_1} = \{\gamma \in S^{p_1}([0, \tau]) : \gamma(t^*) = 0\}$ with $p_1 \geq 2$ and the true function $g_0(\cdot)$ belongs to $S^{p_2}([0, \mu + \delta_1]) = \mathcal{G}^{p_2}$ with some positive constant δ_1 and $p_2 \geq 3$.

(C5) Denote $R(t) = \int_0^t \exp(\gamma_0(s))ds$, $V = X_{(1)} + X_{(-1)}^T\beta_0$, and $U = e^V R(Y)$. There exists $\eta_1 \in (0, 1)$ such that for all $u \in \mathbf{R}^d$ with $\|u\| = 1$,

$$u^T \mathrm{var}(X_{(-1)} \mid U, V)u \geq \eta_1 u^T P(X_{(-1)}X_{(-1)}^T \mid U, V)u \quad \text{almost surely.}$$

(C6) Let $\psi(t, x, \beta, \gamma, g) = x_{(1)} + x_{(-1)}^T\beta + \gamma(t) + g(\Lambda(t, x, \beta, \gamma, g))$ and denote its functional derivatives with respect to $\gamma(\cdot)$ and $g(\cdot)$ along the direction $v(\cdot)$ and $w(\cdot)$ at the true parameter by $\psi'_{0\gamma}(t, x)[v]$ and $\psi'_{0g}(t, x)[w]$, respectively, whose rigorous definitions are given by (S19)–(S20) in the supplementary materials. For any $v(\cdot) \in \Gamma^{p_1}$ and $w(\cdot) \in \mathcal{G}^{p_2}$, there exists $\eta_2 \in (0, 1)$ such that

$$\begin{aligned} &(P\{\psi'_{0\gamma}(Y, X)[v]\psi'_{0g}(Y, X)[w] \mid \Delta = 1\})^2 \\ &\leq \eta_2 P\{(\psi'_{0\gamma}(Y, X)[v])^2 \mid \Delta = 1\} \\ &\quad P\{(\psi'_{0g}(Y, X)[w])^2 \mid \Delta = 1\} \end{aligned}$$

almost surely.

Conditions (C1)–(C3) are common regularity assumptions in survival analysis. Condition (C4) requires $p_2 \geq 3$ to control the error rates of the spline approximation for the true function g_0 and its first and second derivatives. Moreover, together with $p_1 \geq 2$, (C4) will also be used to verify the assumptions (A4)–(A6) for the general M-theorem (Theorem 3) when we apply it to derive the asymptotic normality of the proposed sieve MLE (Theorem 2). A similar condition to (C5) was imposed by Wellner and Zhang (2007) for the panel count data, by Ding and Nan (2011) for the linear transformation model with a known transformation, and by Zhao, Wu, and Yin (2017) for the accelerated hazards model. When the transformation function is known, condition (C5) is equivalent to the assumption C7 in Ding and Nan (2011) and can be verified in many applications as shown in Wellner and Zhang (2007). For the general case where both the transformation function and the error distribution are unspecified, condition (C6) is assumed to avoid strong collinearity between $\psi'_{0\gamma}(Y, X)[v]$ and $\psi'_{0g}(Y, X)[w]$.

Note that the parameter $g(\cdot)$ takes $\Lambda(t, x, \beta, \gamma, g)$ as its argument in (12), which involves the other parameters β and $\gamma(\cdot)$. Thus, β, $\gamma(\cdot)$ and $g(\cdot)$ are bundled parameters. For any $g(\cdot) \in \mathcal{G}^{p_2}$, we directly consider the composite function $g(\Lambda(t, x, \beta, \gamma, g))$ as a function from $\mathcal{T} \times \mathcal{X} \times \mathcal{B} \times \Gamma^{p_1}$ to $\mathbf{R}$. And we define the collection of functions

$$\mathcal{H}^{p_2} = \{\zeta(\cdot,\beta,\gamma) : \zeta(t,x,\beta,\gamma) = g(\Lambda(t,x,\beta,\gamma,g)), t \in [0,\tau], x \in \mathcal{X}, \beta \in \mathcal{B}, \gamma \in \Gamma^{p_1}, g \in \mathcal{G}^{p_2} \text{ such that } \sup_{t\in[0,\tau],x\in\mathcal{X}} |\Lambda(t,x,\beta,\gamma,g)| \le \mu + \delta_1\},$$

with δ_1 given in condition (C4). For any $\zeta(\cdot,\beta,\gamma) \in \mathcal{H}^{p_2}$, we define its norm as

$$\|\zeta(\cdot,\beta,\gamma)\|_2 = \left[\int_{\mathcal{X}} \int_0^\tau [\zeta(t,x,\beta,\gamma)]^2 d\Lambda_0(t,x) dF_X(x)\right]^{1/2},$$

where $F_X(x)$ is the cumulative distribution function of X. Denote the parameter $\theta = (\beta, \gamma(\cdot), \zeta(\cdot,\beta,\gamma))$ and the true parameter $\theta_0 = (\beta_0, \gamma_0(\cdot), \zeta_0(\cdot,\beta_0,\gamma_0))$ with $\zeta_0(t,x,\beta_0,\gamma_0) = g_0(\Lambda(t,x,\beta_0,\gamma_0,g_0))$. Denote the parameter space by $\Theta = \mathcal{B} \times \Gamma^{p_1} \times \mathcal{H}^{p_2}$. For any θ_1 and θ_2 in Θ, we define the distance

$$d(\theta_1,\theta_2) = \left(\|\beta_1 - \beta_2\|^2 + \|\gamma_1 - \gamma_2\|_2^2 + \|\zeta_1(\cdot,\beta_1,\gamma_1) - \zeta_2(\cdot,\beta_2,\gamma_2)\|_2^2\right)^{1/2},$$

where $\|\cdot\|$ is the Euclidean norm and $\|\gamma\|_2 = (\int_0^\tau (\gamma(t))^2 dt)^{1/2}$ is the L_2 norm.

Next, we construct the sieve space as follows. Let $0 = t_0 < t_1 < \cdots < t_{K_n^1} < t_{K_n^1+1} = \tau$ be a partition of $[0,\tau]$ with $K_n^1 = O(n^{\nu_1})$ and $\max_{1\le j\le K_n^1+1} |t_j - t_{j-1}| = O(n^{-\nu_1})$ for some $\nu_1 \in (0, 0.5)$. Let $T_{K_n^1} = \{t_1, \ldots, t_{K_n^1}\}$ denote the set of partition points and $S_n(T_{K_n^1}, K_n^1, p_1)$ be the space of polynomial splines of order p_1 as defined in Schumaker (2007, p. 108, Definition 4.1). Similarly, let $T_{K_n^2}$ be a set of partition points of $[0,\mu]$ with $K_n^2 = O(n^{\nu_2})$ and $\max_{1\le j\le K_n^2+1} |t_j - t_{j-1}| = O(n^{-\nu_2})$ for some $\nu_2 \in (0, 0.5)$, and $S_n(T_{K_n^2}, K_n^2, p_2)$ be the space of polynomial splines of order p_2. According to Schumaker (2007, p. 117, Corollary 4.10), there exist two sets of B-spline bases $\{B_j^1, 1 \le j \le q_n^1\}$ with $q_n^1 = K_n^1 + p_1$ and $\{B_j^2, 1 \le j \le q_n^2\}$ with $q_n^2 = K_n^2 + p_2$ such that for any $s_1 \in S_n(T_{K_n^1}, K_n^1, p_1)$ and $s_2 \in S_n(T_{K_n^2}, K_n^2, p_2)$, we can write $s_1(t) = \sum_{j=1}^{q_n^1} a_j B_j^1(t)$ and $s_2(t) = \sum_{j=1}^{q_n^2} b_j B_j^2(t)$. Let $\Gamma_n^{p_1} = \{\gamma \in S_n(T_{K_n^1}, K_n^1, p_1) : \gamma(0) = 0\}$, $\mathcal{G}_n^{p_2} = S_n(T_{K_n^2}, K_n^2, p_2)$, and

$$\mathcal{H}_n^{p_2} = \{\zeta(\cdot,\beta,\gamma) : \zeta(t,x,\beta,\gamma) = g(\Lambda(t,x,\beta,\gamma,g)), g \in \mathcal{G}_n^{p_2}, t \in [0,\tau], x \in \mathcal{X}, \beta \in \mathcal{B}, \gamma \in \Gamma_n^{p_1}\}.$$

Let $\Theta_n = \mathcal{B} \times \Gamma_n^{p_1} \times \mathcal{H}_n^{p_2}$ be the sieve space. It is not difficult to see that $\Theta_n \subset \Theta_{n+1} \subset \cdots \subset \Theta$. We consider the sieve estimator $\hat\theta_n = (\hat\beta_n, \hat\gamma_n(\cdot), \hat\zeta_n(\cdot,\hat\beta_n,\hat\gamma_n))$, where $\hat\zeta_n(t,x,\hat\beta_n,\hat\gamma_n) = \hat g_n(\Lambda(t,x,\hat\beta_n,\hat\gamma_n,\hat g_n))$, that maximizes the log-likelihood (6) (without covariates Z and parameter $\boldsymbol{\eta}$) over the sieve space Θ_n. The consistency and convergence rate of the sieve MLE $\hat\theta_n$ are then established in the following theorem.

Theorem 1 (Convergence rate of $\hat\theta_n$). Let ν_1 and ν_2 satisfy the restrictions $\max\{\frac{1}{2(2+p_1)}, \frac{1}{2p_1} - \frac{\nu_2}{p_1}\} < \nu_1 < \frac{1}{2p_1}$, $\max\{\frac{1}{2(1+p_2)}, \frac{1}{2(p_2-1)} - \frac{2\nu_1}{p_2-1}\} < \nu_2 < \frac{1}{2p_2}$, and $2\min\{2\nu_1, \nu_2\} > \max\{\nu_1, \nu_2\}$. Suppose conditions (C1)–(C6) hold, then we have

$$d(\hat\theta_n, \theta_0) = O_p(n^{-\min\{p_1\nu_1, p_2\nu_2, \frac{1-\max\{\nu_1,\nu_2\}}{2}\}}).$$

Theorem 1 gives the convergence rate of the proposed estimator $\hat\theta_n$ to the true parameter θ_0, and its proof is provided in the supplementary materials by verifying the conditions in Shen and Wong (1994, Theorem 1). Note the subscripts 1 and 2 correspond to the space of the spline approximation for two infinite-dimensional parameters γ and g, respectively. The restrictions on ν_1 and ν_2 are feasible for p_1 and p_2 not far away from each other. For example, if $p_1 = p_2 = p$ and $\nu_1 = \nu_2 = \nu$, the restriction on ν is equivalent to $\frac{1}{2(1+p)} < \nu < \frac{1}{2p}$, and the convergence rate becomes $d(\hat\theta_n, \theta_0) = O_p(n^{-\min\{p\nu, \frac{1-\nu}{2}\}})$, which is the same as the case when there is only one infinite-dimensional parameter (Ding and Nan 2011; Zhao, Wu, and Yin 2017). Further, if $\nu = \frac{1}{1+2p}$, we have $d(\hat\theta_n, \theta_0) = O_p(n^{-\frac{p}{1+2p}})$, which achieves the optimal convergence rate in the nonparametric regression setting.

Although the convergence rate for the nuisance parameter is slower than the typical rate $n^{1/2}$, we will show that the sieve MLE of the regression parameter, that is, $\hat\beta_n$, is still asymptotically normal and achieves the semiparametric efficiency bound. First, we introduce two additional regularity conditions which are stated below.

(C7) There exist $\mathbf{v}^* = (v_1^*, \ldots, v_d^*)^T$ and $\mathbf{w}^* = (w_1^*, \ldots, w_d^*)^T$, where $v_j^* \in \Gamma^2$ and $w_j^* \in \mathcal{G}^2$ for $j = 1, \ldots, d$, such that $P\{\Delta \mathbf{A}^*(U,X)\psi'_{0\gamma}(Y,X)[v]\} = 0$ and $P\{\Delta \mathbf{A}^*(U,X)\psi'_{0g}(Y,X)[w]\} = 0$ hold for any $v \in \Gamma^{p_1}$ and $w \in \mathcal{G}^{p_2}$. Here U and V are defined the same as in condition (C5) and

$$\begin{aligned}\mathbf{A}^*(t,X) = &-\left(g_0'(\tilde\Lambda_0(t))\exp(g_0(\tilde\Lambda_0(t)))t + 1\right) X_{(-1)} \\ &+ g_0'(\tilde\Lambda_0(t))\exp(g_0(\tilde\Lambda_0(t))) \\ &\int_0^t \mathbf{v}^*(R^{-1}(se^{-V}))ds + \mathbf{v}^*(R^{-1}(te^{-V})) \\ &+ g_0'(\tilde\Lambda_0(t))\exp(g_0(\tilde\Lambda_0(t))) \\ &\int_0^{\tilde\Lambda_0(t)} \exp(-g_0(s))\mathbf{w}^*(s)ds + \mathbf{w}^*(\tilde\Lambda_0(t)),\end{aligned}$$

where $\tilde\Lambda_0(t)$ is the solution of $\tilde\Lambda_0'(t) = \exp(g_0(\tilde\Lambda_0))$ with $\tilde\Lambda_0(0) = 0$.

(C8) Let $\boldsymbol{l}^*(\beta_0,\gamma_0,\zeta_0;W) = \int \mathbf{A}^*(t,X)dM(t)$, where $M(t) = \Delta \mathbb{1}(U \le t) - \int_0^t \mathbb{1}(U \ge s)d\tilde\Lambda_0(s)$ is the event counting process martingale. The information matrix $I(\beta_0) = P(\boldsymbol{l}^*(\beta_0,\gamma_0,\zeta_0;W)^{\otimes 2})$ is nonsingular. Here for a vector a, $a^{\otimes 2} = aa^T$.

The additional condition (C7) essentially requires the existence of the least favorable direction that is used to establish the semiparametric efficiency bound. The directions $\mathbf{v}^*$ and $\mathbf{w}^*$ may be found through the equations in (C7). We illustrate how to construct $\mathbf{v}^*$ and $\mathbf{w}^*$ for the Cox model and the linear transformation model with a known transformation, respectively, in Remark 10. Condition (C8) is a natural assumption that requires the information matrix to be invertible. The following theorem establishes the asymptotic normality and semiparametric efficiency of the sieve MLE $\hat\beta_n$ of the regression parameter for the general linear transformation model.

Theorem 2 (Asymptotic normality of $\hat{\beta}_n$). Suppose the conditions in Theorem 1 and (C7)–(C8) hold, then we have

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \sqrt{n} I^{-1}(\beta_0)\mathbb{P}_n \boldsymbol{l}^*(\beta_0, \gamma_0, \zeta_0; W) + o_p(1) \to_d N(0, I^{-1}(\beta_0))$$

with $I(\beta_0)$ given in condition (C8) and $\to_d$ denoting convergence in distribution.

Theorem 2 states that $\hat{\beta}_n$ is asymptotically normal with variance as the inverse of the information matrix. In practice, the information matrix can be approximated by the estimated information matrix of all parameters including the coefficients of spline bases.

We note that the existing sieve M-theorem for bundled parameters (Ding and Nan 2011; Zhao, Wu, and Yin 2017) cannot be directly applied to prove Theorem 2, because it does not allow the infinite-dimensional nuance parameter to be a function of other infinite-dimensional nuance parameters. Therefore, to study the asymptotic distribution of $\hat{\beta}_n$, we first establish a new general M-theorem for bundled parameters where the infinite-dimensional nuisance parameter is a function of not only the Euclidean parameter of interest but also other infinite-dimensional nuisance parameters. The established M-theorem under such a general scenario then enables us to prove Theorem 2 by verifying its assumptions for the linear transformation model. The details are provided in the supplementary materials. Since the new M-theorem can be useful for developing the asymptotic normality of sieve estimators for other ODE models, we state it below for readers of interest.

We first introduce the general setting and notation for the proposed sieve M-theorem. Let $m(\theta; W)$ be an objective function of unknown parameters $\theta = (\beta, \boldsymbol{\gamma}(\cdot), \zeta(\cdot, \beta, \boldsymbol{\gamma}))$ given a single observation W, where β is the finite-dimensional parameter of interest, $\boldsymbol{\gamma}(\cdot) = (\gamma_1(\cdot), \ldots, \gamma_{d_2}(\cdot))$ denotes infinite-dimensional nuisance parameters, and $\zeta(\cdot, \beta, \boldsymbol{\gamma})$ is another infinite-dimensional nuisance parameter that can be a function of β and $\boldsymbol{\gamma}$. Here "$\cdot$" represents some components of W. Given i.i.d. observations $\{W_i\}_{i=1}^n$, the sieve estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n(\cdot), \hat{\zeta}_n(\cdot, \hat{\beta}_n, \hat{\boldsymbol{\gamma}}_n))$ maximizes the objective function, $\mathbb{P}_n m(\theta; W)$, over certain sieve space. For example, $\hat{\theta}_n$ becomes the sieve MLE if m is the log-likelihood function. We denote the derivative of m with respect to β as m'_β, the functional derivative of m with respect to γ_j along the direction $v(\cdot)$ as $m'_{\gamma_j}[v]$ for $1 \le j \le d_2$, and the functional derivative of m with respect to ζ along the direction $h(\cdot)$ as $m'_\zeta[h]$, whose rigorous definitions are given in the supplementary materials. The following theorem then establishes the asymptotic normality of the sieve estimator, $\hat{\beta}_n$, under the above general setting.

Theorem 3 (A general M-theorem for bundled parameters). Under assumptions (A1)–(A6) in the supplementary materials, we have

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = A^{-1}\sqrt{n}\mathbb{P}_n \mathbf{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) + o_p(1) \to_d N(0, A^{-1}B(A^{-1})^T),$$

where

$$\begin{aligned}
&\mathbf{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) \\
&\quad = m'_\beta(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) \\
&\qquad - \sum_{j=1}^{d_2} m'_{\gamma_j}(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)[\mathbf{v}_j^*] \\
&\qquad - m'_\zeta(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)[\mathbf{h}^*(\cdot, \beta_0, \boldsymbol{\gamma}_0)], \\
&B = P\{\mathbf{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W) \\
&\qquad \mathbf{m}^*(\beta_0, \boldsymbol{\gamma}_0(\cdot), \zeta_0(\cdot, \beta_0, \boldsymbol{\gamma}_0); W)^T\},
\end{aligned}$$

with $\mathbf{v}_j^* = (v_{j1}^*, \ldots, v_{jd_1}^*)^T$, $\mathbf{h}^* = (h_1^*, \ldots, h_d^*)^T$ and A given in the assumption (A3).

Remark 9. The assumptions needed in Theorem 3 are similar to those in Ding and Nan (2011) (see the supplementary materials for details). However, our proposed theorem significantly differs from the main theorem in Ding and Nan (2011), because the latter considers $\zeta(\cdot, \beta)$ to be a function of only the finite-dimensional parameter β, while we consider a more general scenario of bundled parameters, where the nuisance parameter $\zeta(\cdot, \beta, \boldsymbol{\gamma})$ can be a function of both the finite-dimensional parameter β and other infinite-dimensional nuisance parameters $\boldsymbol{\gamma}$. The proposed theorem nontrivially extends the asymptotic distributional theories for M-estimation under this general scenario.

Remark 10. We note that to find the least favorable directions $\mathbf{v}^*$ and $\mathbf{w}^*$ required in (C7), we may solve the equations in (C7), which can be simplified to equations (S37) and (S39) provided in the supplementary materials. For illustration, we provide explicit constructions of the least favorable directions for the Cox model and for the linear transformation model with a known transformation, respectively. Specifically, for the Cox model, we have $g_0 \equiv 0$ and $\mathbf{v}^*$ can be derived as

$$\mathbf{v}^*(t) = \frac{P\{\mathbb{1}(Y \ge t)e^{X^T\beta_0}X\}}{P\{\mathbb{1}(Y \ge t)e^{X^T\beta_0}\}};$$

for the linear transformation model where γ_0 is known, $\mathbf{w}^*$ can be obtained as

$$\mathbf{w}^*(t) = \boldsymbol{\phi}(t) - g_0'(t)\int_0^t \boldsymbol{\phi}(s)ds,$$

where

$$\boldsymbol{\phi}(t) = \left(g_0'(t)\exp(g_0(t))\tilde{\Lambda}_0^{-1}(t) + 1\right)\frac{P\{\mathbb{1}(\Lambda_0(Y, X) \ge t)X\}}{P\{\mathbb{1}(\Lambda_0(Y, X) \ge t)\}}$$

with $\tilde{\Lambda}_0$ defined in (C7).

Given the above constructions of the least favorable directions, we can further simplify the nonsingularity condition of the information matrix in (C8). For the Cox model, the information matrix can be derived as $I(\beta_0) = \int_0^\infty P\left([-X + \boldsymbol{\mu}(t)]^{\otimes 2}\mathbb{1}(U \ge t)\right)dt$, where $\boldsymbol{\mu}(t) = P\{\mathbb{1}(U \ge t)e^{X^T\beta_0}X\}/P\{\mathbb{1}(U \ge t)e^{X^T\beta_0}\}$ with U defined in (C5). Respectively, for the linear transformation where γ_0 is known, the information matrix can be derived as $I(\beta_0) = \int_0^\infty m^2(t) \cdot \mathrm{var}(X|U \ge t) \cdot P(U \ge t) \cdot \exp(g_0(\tilde{\Lambda}_0(t)))dt$, where $m(t) = g_0'(\tilde{\Lambda}_0(t))\exp(g_0(\tilde{\Lambda}_0(t)))t + 1$. The nonsingularity condition requires the integral of a covariance matrix to be positive definite.

Remark 11. Moreover, for the general class of ODE models that include covariates Z with time-varying coefficients $\boldsymbol{\eta}(\cdot)$ in (2), we have further established the same convergence rate of the sieve estimator $-\hat{\theta}_n$ in Theorem 4 and the asymptotic normality of $\hat{\beta}_n$ in Theorem 5 in the supplementary materials. In particular, the conditions (C1)-(C8) have been revised to (C1′)–(C8′) with additional regularity conditions on covariates Z. We refer to the supplementary materials for the full list of conditions, rigorous statements of theorems, and their proofs.

5. Simulation Studies

In this section, we use simulation studies to show the finite sample performance of the sieve MLE under the time-varying Cox model and the general linear transformation model.

5.1. Time-Varying Cox model

We generate event times from the model

$$\Lambda'_x(t) = \alpha(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \eta(t)x_5),$$

where $(x_1, x_2, x_3, x_4, x_5)$ follows a multivariate normal distribution with mean 0 and autoregressive covariance truncated at ± 2, $\beta_1 = \beta_4 = 1$, and $\beta_2 = \beta_3 = -1$. Let $\eta(t) = \sin(\frac{3}{4}\pi t)$ be a time-varying coefficient for x_5 and the coefficients of all other covariates be time-independent. The baseline hazard $\alpha(t)$ is set to 0.5. The censoring times are generated from an independent uniform distribution $U(0, 3)$, which leads to a censoring rate around 50%. The sample size N varies from 1000 to 8000. We fit both the log-transformed baseline hazard function $\log\alpha(t)$ and time-varying coefficient $\eta(t)$ by cubic B-splines and set the number of knots $K_n = \lfloor N'^{\frac{1}{5}} \rfloor$, that is, the largest integer smaller than $N'^{\frac{1}{5}}$, where N' is the number of distinct observation time points. The interior knots are located at the K_n quantiles of the N' distinct observation time points. We compare the estimation accuracy and the computing time of the proposed sieve MLE with those of the partial likelihood-based estimator implemented in the "coxph" function in R with the "tt" argument set as the same cubic B-spline transformation of time.

Table 1 summarizes the estimates of regression coefficients β_1 and β_2 based on 1000 replications. The estimates of the other two regression coefficients β_3 and β_4 perform similarly, and the results are included in the supplementary materials. For the time-varying coefficient $\eta(t)$, we report the integrated mean square error (IMSE), which is the weighted sum of mean square error (MSE) of pointwise estimates over simulated time points from 0 to 2. As one can see, the mean and standard deviation of IMSE of the proposed sieve estimator decrease as the sample size increases. Remarkably, they are consistently smaller than those of the partial likelihood-based estimator. For time-independent coefficients, the proposed sieve estimator performs as well as the partial likelihood-based estimator. The mean of the standard error estimator, which is obtained by inverting the estimated information matrix of all parameters including the coefficients of spline bases, is approximate to the sample standard error, and the corresponding 95% confidence interval achieves a proper coverage proportion. From the left and middle panels of Figure 1, we can see that the means of the estimated $\alpha(t)$ and $\eta(t)$ are close to the true functions, and the 95% pointwise confidence bands cover the true functions well.

It is also worth noting that, in comparison to the partial likelihood-based estimation method whose relative computing time with respect to that with the smallest sample size increases quickly as the sample size grows, the proposed estimation method is computationally more efficient, especially when the sample size is large (see the right panel of Figure 1). When the number of knots increases with the sample size, the computation time of the proposed method grows at a rate slightly larger than the linear rate (but far below the quadratic rate).

5.2. Linear Transformation Model

We generate event times from the model $\Lambda'_x(t) = q(\Lambda_x(t))\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)\alpha(t)$. The covariates are independent normal with mean 0 and standard deviation 0.5 truncated at ± 2. We consider four different settings for $q(\cdot)$ and $\alpha(\cdot)$: (a) a constant $q(t) = 1$ and a monotonic increasing $\alpha(t) = t^3$, in which case the Cox model is correctly specified; (b) a monotonic decreasing $q(t) = e^{-t}$ and a constant $\alpha(t) = 2$; (c) a monotonic decreasing $q(t) = 2/(1 + t)$ and a constant $\alpha(t) = 1$; 4) an increasing $q(t) = \log(1 + t) + 2$ and an increasing $\alpha(t) = log(1 + t)$. In each setting, we generate the censoring time from an independent uniform distribution $U(0, c)$, where c is chosen to achieve approximately 25%–30% censoring rates. The sample size N varies from 1000 to 8000.

In setting (1), we compare the proposed sieve MLE for the ODE-Cox model, where the function $q(\cdot)$ is set to 1, with the partial-likelihood based estimator implemented using the R package *survival*. We fit $\log\alpha(\cdot)$ by cubic B-splines with $\lfloor N'^{\frac{1}{5}} \rfloor$ interior knots that are located at the quantiles of the distinct observation time points. In setting (2), we compare the proposed sieve MLE for the ODE-LT model, where the function $q(\cdot)$ is set to e^{-t}, with the NPMLE for the equivalent logarithmic transformation model considered in Zeng and Lin (2007b). We fit $\log\alpha(\cdot)$ by cubic B-splines with the same placement of interior knots. In setting (3), we compare the proposed sieve MLE for the ODE-AFT model, where the function α is set to 1, with the rank-based estimation approach implemented using the R package *aftgee*. We fit $\log q(t)$ by cubic B-splines with $\lfloor N^{\frac{1}{7}} \rfloor$ interior knots that are located at the quantiles of the estimated cumulative hazards under the Cox model. In setting (4) (as well as settings (1)–(3)), we fit the general linear transformation model (ODE-Flex) where both $q(\cdot)$ and $\alpha(\cdot)$ are unspecified, and compare the sieve MLE with the smoothed partial rank (SPR) method in Song et al. (2006). Both methods constrain $\beta_1 = 1$ for identifiability guarantee. For the sake of space, the results of the setting 4) are provided in the supplementary materials.

Tables 2 and 3 summarize the estimates of regression coefficients with the sample size $N = 4000$ based on 1000 replications. Full results for the other sample sizes are provided in the supplementary materials. Table 2 indicates that when any of the Cox model, the logarithmic transformation model, or the AFT model is correctly specified, the sieve estimator for the corresponding correctly specified ODE model achieves similar performance as the partial-likelihood based estimator for the Cox model, the NPMLE for the logarithmic transformation

Table 1. Simulation results under time-varying Cox model.

N	Method	$\beta_1 = 1$				$\beta_2 = -1$				IMSE($\eta(t)$)	
		Bias	SE	ESE	CP	Bias	SE	ESE	CP	Mean	SD
1000	ODE	0.008	0.070	0.070	0.958	−0.012	0.076	0.078	0.955	0.053	0.041
	Cox-MPLE	0.006	0.070	0.068	0.952	−0.010	0.075	0.075	0.950	0.0109	0.094
2000	ODE	0.004	0.048	0.048	0.958	−0.004	0.053	0.054	0.957	0.029	0.021
	Cox-MPLE	0.002	0.048	0.048	0.956	−0.003	0.053	0.053	0.959	0.053	0.041
4000	ODE	0.003	0.033	0.034	0.952	−0.003	0.038	0.038	0.938	0.016	0.011
	Cox-MPLE	0.003	0.033	0.034	0.950	−0.002	0.038	0.037	0.936	0.026	0.020
8000	ODE	0.000	0.024	0.024	0.962	−0.001	0.026	0.026	0.938	0.009	0.006
	Cox-MPLE	0.000	0.023	0.024	0.959	−0.001	0.026	0.026	0.936	0.013	0.009

Bias is the difference between the mean of estimates and the true value, SE is the sample standard error of the estimates, Mean is the mean of IMSE, and SD is the standard deviation of IMSE. ESE is the mean of the standard error estimators by inverting the estimated information matrix of all parameters, including the coefficients of spline bases, and CP is the corresponding coverage proportion of 95% confidence intervals.

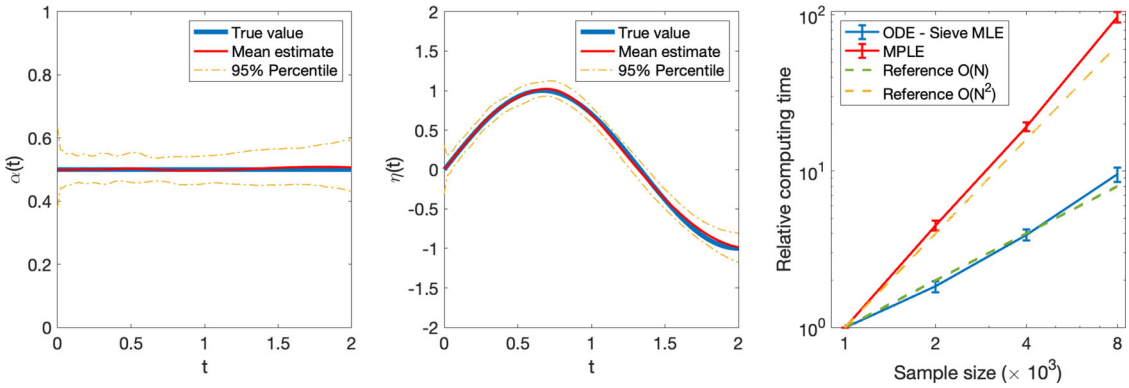

Figure 1. True $\alpha_0(t)$ and mean of $\hat{\alpha}(t)$ (left); true $\eta(t)$ and mean of $\hat{\eta}(t)$ (middle) with the sample size $N = 8000$; log-log plot of mean relative computation time (right) with respect to the sample size under the time-varying Cox model.

Table 2. Estimates of regression coefficients under correctly-specified ODE-Cox with $q(\cdot) \equiv 1$, ODE-LT with $q(t) = e^{-t}$, and ODE-AFT with $\alpha(\cdot) \equiv 1$.

	Method	$\beta_1 = 1$				$\beta_2 = 1$				$\beta_3 = 1$			
		Bias	SE	ESE	CP	Bias	SE	ESE	CP	Bias	SE	ESE	CP
(1)	MPLE	0.002	0.076	0.075	0.934	−0.003	0.075	0.075	0.941	−0.001	0.074	0.075	0.954
	ODE-Cox	0.003	0.076	0.076	0.936	−0.002	0.075	0.076	0.942	0.000	0.074	0.076	0.955
(2)	NPMLE	0.004	0.117	0.115	0.949	−0.001	0.114	0.115	0.951	0.003	0.113	0.115	0.960
	ODE-LT	0.005	0.117	0.115	0.950	−0.000	0.114	0.115	0.951	0.003	0.113	0.115	0.961
(3)	Rank-based	0.004	0.105	0.102	0.944	−0.001	0.102	0.102	0.950	0.002	0.100	0.103	0.954
	ODE-AFT	0.000	0.102	0.097	0.944	−0.005	0.100	0.097	0.944	−0.002	0.097	0.097	0.950

NOTE: Bias, SE, ESE and CP contain the same meanings as those in Table 1. Setting (1): the Cox model is correctly specified. Setting (2): the logarithmic transformation model is correctly specified. Setting (3): the AFT model is correctly specified.

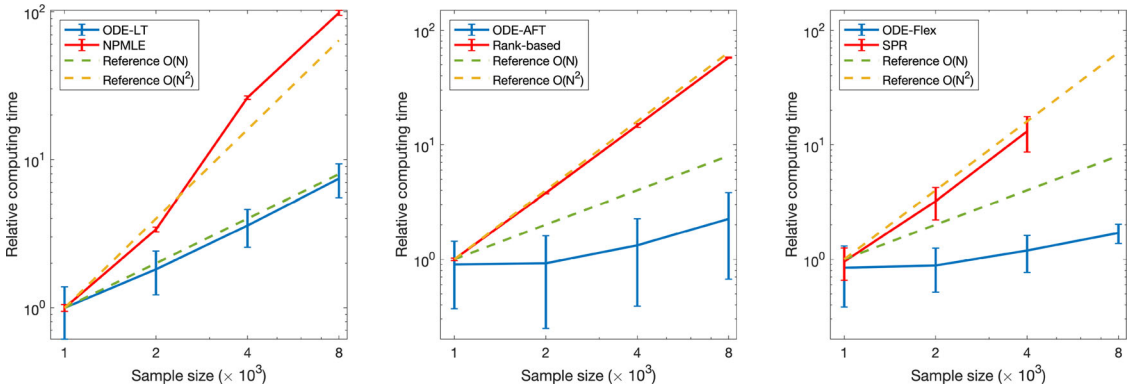

Figure 2. The log-log plots of mean relative computing time with respect to the sample size under the ODE-LT, the ODE-AFT model, and the ODE-Flex model are provided from left to right, respectively.

model, or the rank-based estimator for the AFT model. However, the relative computing time of the proposed ODE approach increases linearly as the sample size grows while that of the NPMLE for the logarithmic transformation model or the rank-based method for the AFT model increases in a quadratic rate as shown in Figure 2.

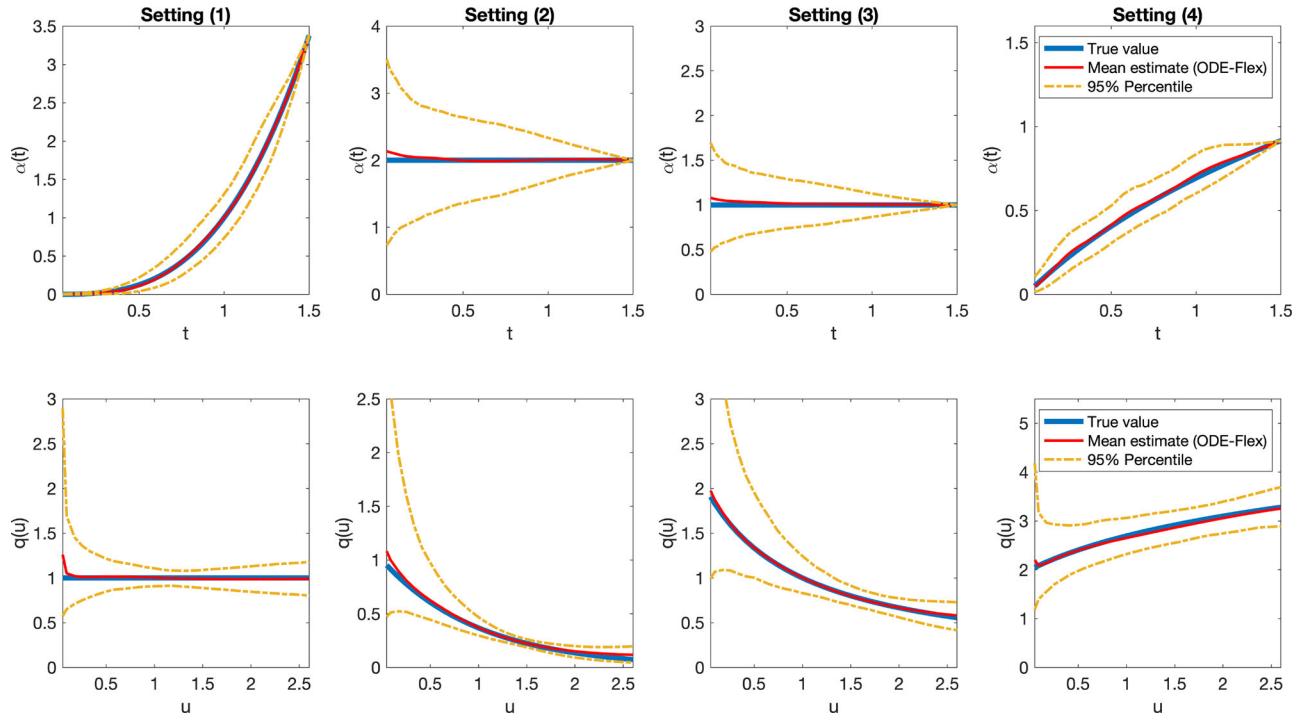

Figure 3. The solid blue curves are the true $q(\cdot)$ (upper row) and $\alpha(\cdot)$ (lower row). The solid red curves are the means of corresponding estimated $\hat{q}(\cdot)$ and $\hat{\alpha}(\cdot)$ under the general linear transformation model. The dashed yellow curves represent 95% pointwise confidence bands over 1000 replications. From left to right, the four columns correspond to settings (1)–(4), respectively.

Table 3. Estimates of regression coefficients under the general linear transformation model ODE-Flex with both $q(\cdot)$ and $\alpha(\cdot)$ unspecified.

	$\beta_2 = 1$				$\beta_3 = 1$			
Setting	Bias	SE	ESE	CP	Bias	SE	ESE	CP
(1)	0.008	0.106	0.107	0.947	0.012	0.104	0.107	0.959
(2)	−0.019	0.161	0.151	0.927	−0.016	0.159	0.151	0.938
(3)	−0.014	0.134	0.131	0.941	−0.012	0.131	0.132	0.945
(4)	0.001	0.092	0.090	0.939	0.005	0.091	0.090	0.954

NOTE: Bias, SE, ESE, and CP contain the same meanings as those in Table 1.

For the general linear transformation model, we find that the proposed ODE-Flex method has advantages against the existing SPR method in terms of estimation accuracy, numerical stability, and computational efficiency. We refer to the supplementary materials for detailed results and comparison with SPR. From Table 3, we can see that the bias of the ODE-Flex estimator is nearly negligible in all settings. The standard error estimators are close to the sample standard errors, and the corresponding 95% confidence intervals achieve a reasonable coverage proportion. When the Cox model, the logarithmic transformation model, or the AFT model is correctly specified, their estimators (in Table 2) achieve smaller standard errors than those for ODE-Flex (in Table 3), which is expected because both $q(\cdot)$ and $\alpha(\cdot)$ are unspecified in ODE-Flex. Figure 3 shows the mean of $\hat{\alpha}(\cdot)$ and $\hat{q}(\cdot)$, respectively. As one can see, the means of $\hat{\alpha}(\cdot)$ and $\hat{q}(\cdot)$ under the general linear transformation model are all close to the true functions. Moreover, the relative computing time of ODE-Flex increases in a much smaller rate than that of SPR as the sample size grows as shown in the right panel of Figure 2.

Note we have also considered other alternative knots placements (see the supplementary materials) and our numerical results suggest that knot selection does not appear critical for the proposed method.

6. Data Example

In this section, we apply the proposed method to a kidney post-transplantation mortality study. End-stage renal disease (ESRD) is one of the most deadly and costly diseases in the United States. From 2004 to 2016, ESRD incident cases increased from 345.6 to 373.4 per million people, with Medicare expenditures escalating from 18 to 35 billion dollars (Saran et al. 2017). Kidney transplantation is the renal replacement therapy for the majority of patients with ESRD. Successful kidney transplantation is associated with improved survival, improved quality of life, and health care cost savings when compared to dialysis. However, despite aggressive efforts to increase the number of donor kidneys, the demand far exceeds the supply of donor kidneys for transplantation and hence, the donor waiting list is very long. Currently about 130,000 patients are waiting for lifesaving organ transplants in the United States among whom 100,000 await kidney transplants and fewer than 15% of patients will receive transplants in their lifetime. To optimize the organ allocation, further research is essential to determine the risk factor associated with post-transplant mortality.

To better understand this problem, we considered the data obtained from the Organ Procurement and Transplantation Network (OPTN). There were 146,248 patients who received transplants between 1990 and 2008. Failure time (recorded in years) was defined as the time from transplantation to graft failure or death, whichever occurred first, where graft failure was considered to occur when the transplanted kidney ceased to function. Patient survival was censored at the end of study

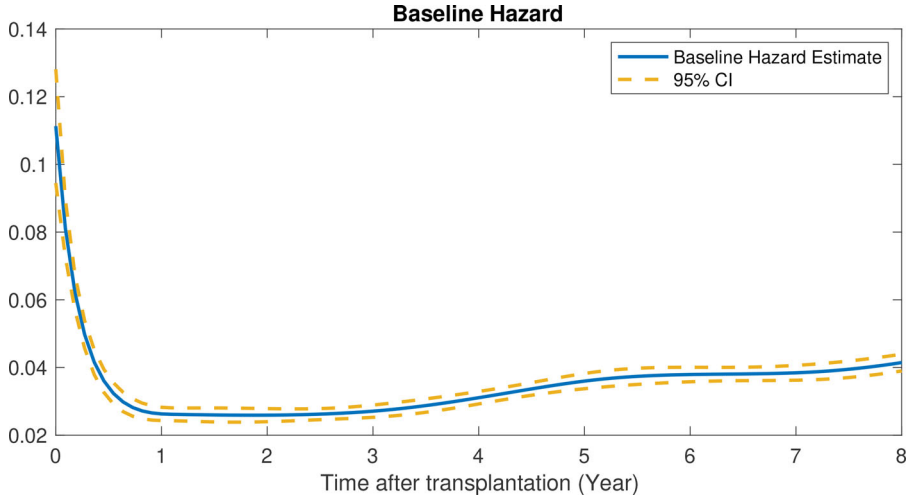

Figure 4. Estimated baseline hazard $\hat{\alpha}(t)$ using the proposed sieve MLE method for the kidney transplantation data.

Table 4. Summary of estimates for time-independent effects in kidney post-transplantation mortality study.

Variables	DCD	Polycystic	Diabetes	Hypertension
EST	−0.081	−0.511	0.333	−0.146
ESE	0.038	0.021	0.012	0.014
95% CI	[−0.156, −0.007]	[−0.553, −0.469]	[0.310, 0.357]	[−0.172, −0.119]
p-value	0.033	< 0.001	< 0.001	< 0.001

EST is the estimated time-independent effect, ESE is the estimated standard error by inverting the estimated information matrix of all parameters including the coefficients of spline basis, and CI is the confidence interval.

(2008). The median follow-up time was around 6 years and the censoring rate was 62%. Covariates included in this study were age at transplantation, race, gender, cold ischemic time, donation after cardiac death (DCD), BMI, expanded criteria donor (ECD), dialysis time, comorbidity conditions such as glomerulonephritis, polycystic kidney disease, diabetes, and hypertension. Detecting and accounting for time-varying effects are particularly important in the context of kidney transplantation, as nonproportional hazards have already been reported in the literature (Wolfe et al. 1999; He et al. 2017). Also, analyses with time-varying effects provide valuable clinical information that could be obscured otherwise.

However, existing statistical softwares become computationally infeasible when fitting a time-varying effects model on a dataset as large as what we have here. Thus, to estimate the potential time-varying effects, we fit the time-varying Cox model using the proposed sieve MLE, which is computationally scalable. Specifically, based on previous studies, DCD, Polycystic, Diabetes and Hypertension are modeled with time-independent effects, and the remaining variables are estimated with time-varying effects. The time-varying effects are all implemented by cubic B-splines with five interior knots, which is chosen based on the Bayesian information criterion. Figure 4 shows the estimated baseline hazard function. We can see that the post-transplant mortality is high in the short term after surgery, with a weakening association over time. Table 4 summarizes the estimated time-independent effects, and Figure 5 shows examples of fitted time-varying effects with 95% pointwise confidence intervals, where the standard error estimators were obtained by inverting the estimated information matrix of all parameters including time-independent coefficients and the coefficients of spline bases. As one can see, the effects of baseline age varied over time, resulting in an eventually strengthened association. Specifically, compared with the reference group (age at transplantation between 19 and 39), patients 40–49 years of age had a protective effect in the short term after transplantation. We can also see that the high cold ischemic time is a risk factor for mortality in the short run, with a weakening association over time. Thus, special care should be dedicated to improve the short-term outcome. As expected, longer waiting times on dialysis (greater than 5 years) negatively impact post-transplant survival, especially in the short run. Male gender was not significantly associated with mortality immediately after the renal transplantation but became a risk factor in the long run. As can be seen in Figure 5, underweight shows a protective effect in the short run, and then a slightly weakening association over time, which confirms the previous finding of Lafranca et al. (2015). The results regarding high BMI should be interpreted with caution. Although higher levels of BMI in the general population are typically associated with high mortality, in chronic kidney diseases, such as patients with kidney dialysis and kidney transplantation, higher BMI has been associated with better survival, which has been labeled as reverse epidemiology (Dekker et al. 2008; Kovesdy et al. 2010). Our results show that both overweight and obesity improved survival in the short term after kidney transplantation, but obesity became a risk factor after long-term exposure. One possible explanation is that BMI is a complex marker of visceral and nonvisceral adiposity and also of nutritional status including muscle mass (Kovesdy et al. 2010), and the improved short-term outcome associated with higher BMI may be related to differential benefits by one or more of these components. Our findings indicate a need to critically reassess the role of BMI in the risk stratification of kidney transplantation. A further assessment (such as subgroup analysis) of high BMI that differentiates between visceral adiposity, nonvisceral adiposity and higher muscle mass may improve risk stratification in kidney transplant recipients. In addition, our results show that graft survival for patients with Glomerulonephritis is better than patients with other primary diseases. Regarding racial disparities, the long-term survival outcomes for African

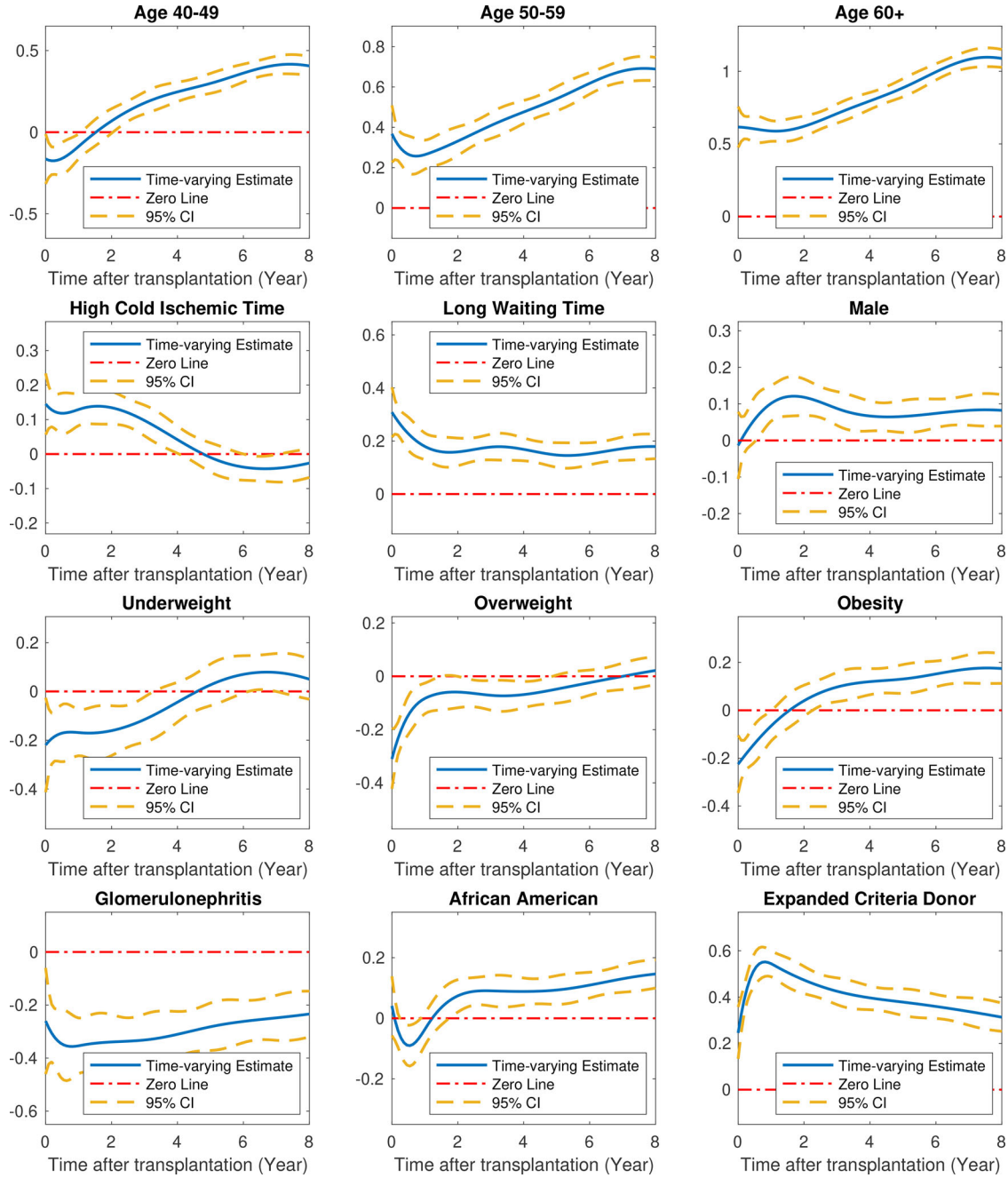

Figure 5. Estimated time-varying effects using the proposed sieve MLE method for the kidney transplantation data.

Americans continue to lag behind non-African Americans. Finally, as expected, the effect of expanded criteria donor (ECD) is not as good as optimal donor. When a suboptimal organ becomes available, patients and physicians must decide whether to accept the offer and special care must be dedicated to improve the survival benefit.

7. Discussion

In this article, we have proposed a novel ODE framework for survival analysis, which unifies the current literature, along with a general estimation procedure which is scalable and easy to implement. The ODE framework provides a new perspective for modeling censored data, which further allows us to use well-developed numerical solvers and local sensitivity analysis tools for ODEs in parameter estimation. Although we have only focused on one class of ODE models in this article, the ODE framework and the estimation method offer new opportunities for investigating more flexible model structures.

We note that a few recent works also use ODEs for survival analysis. Specifically, Tang et al. (2022) model the cumulative hazard as in the ODE (1) with the function $f(\cdot)$ being a neural network to improve feature representation. The method proposed in Tang et al. (2022) can be viewed as a neural-network-based extension of the general framework studied in this work, which demonstrates that the proposed ODE framework can be used to build flexible models. Groha, Schmon, and Gusev (2020) propose a neural-network-based ODE approach to model the Kolmogorov forward equation that characterizes the transition probabilities for multi-state survival analysis. Both the afore-

mentioned works focus on developing flexible models with powerful representation learning via neural networks to improve prediction performance. In this article, instead, we focus on estimation and inference for a general class of semiparametric ODE models, in which case the effects of certain covariates are often of interest. More importantly, we revisit the rich literature of survival analysis and provide a unified view of many existing survival models, which is the key insight that differentiates this work and the aforementioned ones. This unification merit serves as the foundation of the proposed widely applicable estimation procedure. We also establish the consistency and semiparametric efficiency of the proposed sieve estimator for a general class of semiparametric ODE models, with a new general sieve M-theorem.

The proposed general theory derives the asymptotic distribution of bundled parameters, where the nuisance parameter is a function of not only the regression parameters of interest but also other infinite-dimensional nuisance parameters. Though we have only illustrated the efficient estimation in the linear transformation model as an example to motivate such a theoretical development, the proposed general theory can be extended to other models.

In addition, an interesting application of the unified ODE framework is to check the model specification. In particular, the estimation and inference for a general ODE model can help test whether a nested model is appropriate for a dataset. For example, Proposition 2 implies that the function $q(\cdot)$ or $\alpha(\cdot)$ in the linear transformation model (5) should be a power function when it coincides with the Cox or the AFT model. Though we have established the consistency of the functional parameters $q(\cdot)$ and $\alpha(\cdot)$ in the nonparametric linear transformation model, it is worthwhile to further investigate their asymptotic distributional theory for model diagnostics as future work. As a preliminary study, we have explored a heuristic parametric approach for model diagnostics and provided its finite sample performance in the supplementary materials.

Finally, we note that a few recent works have tried to address the computation burden of certain estimation methods for specific models on massive time-to-event data. In particular, Wang et al. (2019) proposed an efficient divide-and-conquer (DAC) algorithm for the sparse Cox model. Kawaguchi et al. (2020) developed an algorithm for reducing the computation cost of fitting the Fine-Gray (Fine and Gray 1999) proportional subdistributional hazards model by exploiting its special structure. Zuo et al. (2021) proposed a subsampling procedure to approximate the full-data estimator for the additive hazard model. Note that most of these methods are tailored for a specific model while our method can be applied more broadly. Further, our estimation procedure and these methods are not competitors. In contrast, some of the techniques used in these methods, such as DAC, can be naturally integrated into the proposed estimation procedure, which is an interesting future direction to be explored.

Supplementary Materials

The supplementary materials contain the detailed derivation of the local sensitivity analysis and the optimization algorithm, the proposed general M-theorem for bundled parameters (Theorem 3), the proofs of Theorems 1–3 and Propositions 1–2, the convergence rate and the asymptotic normality of the proposed sieve estimator for the general class of ODE models in the presence of covariates Z with time-varying coefficients, and additional simulation results.

Acknowledgments

The authors thank the editor, the associate editor, and four referees for their constructive comments and suggestions.

Funding

The research of Xu was supported by NSF SES-1846747. The research of Tang and Zhu was supported by NSF DMS-1821243.

References

Aalen, O. (1980), "A Model for Nonparametric Regression Analysis of Counting Processes," in *Mathematical Statistics and Probability Theory*, eds. P. Bauer, F. Konecny and W. Wertz, pp. 1–25, New York: Springer. [1,4]

Ai, C., and Chen, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843. [2,6]

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120. [4]

Bagdonavicius, V., and Nikulin, M. (2001), *Accelerated Life Models: Modeling and Statistical Analysis*, New York, NY: Chapman and Hall/CRC. [1,3,4]

Bagdonavicius, V. B., and Nikulin, M. S. (1999), "Generalized Proportional Hazards Model Based on Modified Partial Likelihood," *Lifetime Data Analysis*, 5, 329–350. [4]

Bennett, S. (1983), "Analysis of Survival Data by the Proportional Odds Model," *Statistics in Medicine*, 2, 273–277. [1,4]

Buckley, J., and James, I. (1979), "Linear Regression with Censored Data," *Biometrika*, 66, 429–436. [1,4]

Cai, T., Tian, L., and Wei, L. J. (2005), "Semiparametric Box-Cox Power Transformation Models for Censored Survival Observations," *Biometrika*, 92, 619–632. [4]

Chen, K., Jin, Z., and Ying, Z. (2002), "Semiparametric Analysis of Transformation Models with Censored Data," *Biometrika*, 89, 659–668. [1,4]

Chen, K., and Tong, X. (2010), "Varying Coefficient Transformation Models with Censored Data," *Biometrika*, 97, 969–976. [4]

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics* (Vol. 6B, 1st ed.), eds. J. Heckman and E. Leamer, pp. 5549–5632, Amsterdam: Elsevier. [5]

Chen, X., Linton, O., and Van Keilegom, I. (2003), "Estimation of Semiparametric Models When the Criterion Function is not Smooth," *Econometrica*, 71, 1591–1608. [6]

Cheng, S. C., Wei, L. J., and Ying, Z. (1995), "Analysis of Transformation Models with Censored Data," *Biometrika*, 82, 835–845. [1,4]

Cox, D. R. (1975), "Partial Likelihood," *Biometrika*, 62, 269–276. [1,4]

Dekker, F., Mutsert, R., Dijk, P., Zoccali, C., and Jager, K. (2008), "Survival Analysis: Time-Dependent Effects and Time-Varying Risk Factors," *Kidney International*, 74, 994–997. [13]

Dickinson, R. P., and Gelinas, R. J. (1976), "Sensitivity Analysis of Ordinary Differential Equation Systems—A Direct Method," *Journal of Computational Physics*, 21, 123–143. [2,5,6]

Ding, Y., and Nan, B. (2011), "A Sieve M-theorem for Bundled Parameters in Semiparametric Models, with Application to the Efficient Estimation in a Linear Model for Censored Data," *The Annals of Statistics*, 39, 3032–3061. [1,2,4,6,7,8,9]

Fine, J. P., and Gray, R. J. (1999), "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association*, 94, 496–509. [15]

Fine, J. P., Ying, Z., and Wei, L. J. (1998), "On the Linear Transformation Model for Censored Data," *Biometrika*, 85, 980–986. [1,4]

Gray, R. J. (1994), "Spline-based Tests in Survival Analysis," *Biometrics*, 50, 640–652. [1,3,4]

Groha, S., Schmon, S. M., and Gusev, A. (2020), "Neural Odes for Multi-State Survival Analysis," arXiv preprint arXiv:2006.04893 [14]

He, K., Yang, Y., Li, Y., Zhu, J., and Li, Y. (2017), "Modeling Time-Varying Effects with Large-Scale Survival Data: An Efficient Quasi-Newton Approach," *Journal of Computational and Graphical Statistics*, 26, 635–645. [13]

He, X., Xue, H., and Shi, N. (2010), "Sieve Maximum Likelihood Estimation for Doubly Semiparametric Zero-inflated Poisson Models," *Journal of Multivariate Analysis*, 101, 2026–2038. [2]

Horowitz, J. L. (1996), "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103–137. [3]

Huang, J. (1999), "Efficient Estimation of the Partly Linear Additive Cox Model," *The Annals of Statistics*, 27, 1536–1563. [1,2]

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003), "Rank-based Inference for the Accelerated Failure Time Model," *Biometrika*, 90, 341–353. [1,4]

Jin, Z., Lin, D. Y., and Ying, Z. (2006), "On Least-Squares Regression with Censored Data," *Biometrika*, 93, 147–161. [1,4]

Kawaguchi, E. S., Shen, J. I., Suchard, M. A., and Li, G. (2020), "Scalable Algorithms for Large Competing Risks Data," *Journal of Computational and Graphical Statistics*, 30, 685–693. [15]

Khan, S., and Tamer, E. (2007), "Partial Rank Estimation of Duration Models with General Forms of Censoring," *Journal of Econometrics*, 136, 251–280. [4,6]

Kovesdy, C., Czira, M., Rudas, A., Ujszaszi, A., Rosivall, L., Novak, M., Kalantar-Zadeh, K., Molnar, M., and Mucsi, I. (2010), "Survival Analysis: Time-Dependent Effects and Time-Varying Risk Factors," *American Journal of Transplantation*, 10, 2644–2651. [13]

Lafranca, J., IJermans, J., Betjes, M., and Frank, J. (2015), "Body Mass Index and Outcome in Renal Transplant Recipients: A Systematic Review and Meta-Analysis," *BMC Medicine*, 13, 111. [13]

Lai, T. L., and Ying, Z. (1991), "Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data," *The Annals of Statistics*, 19, 1370–1402. [1,4]

Lin, D. Y., and Ying, Z. (1995), "Semiparametric Analysis of General Additive-Multiplicative Hazard Models for Counting Processes," *The Annals of Statistics*, 23, 1712–1734. [4]

Lin, Y., and Chen, K. (2012), "Efficient Estimation of the Censored Linear Regression Model," *Biometrika*, 100, 525–530. [4]

Mckeague, I. W., and Sasieni, P. D. (1994), "A Partly Parametric Additive Risk Model," *Biometrika*, 81, 501–514. [4]

Murphy, S. A., Rossini, A. J., and van der Vaart, A. W. (1997), "Maximum Likelihood Estimation in the Proportional Odds Model," *Journal of the American Statistical Association*, 92, 968–976. [1,4,6]

Petzold, L., Li, S., Cao, Y., and Serban, R. (2006), "Sensitivity Analysis of Differential-Algebraic Equations and Partial Differential Equations," *Computers and Chemical Engineering*, 30, 1553–1559. [5,6]

Qiu, Z., and Zhou, Y. (2015), "Partially Linear Transformation Models with Varying Coefficients for Multivariate Failure Time Data," *Journal of Multivariate Analysis*, 142, 144–166. [4]

Saran, R., Robinson, B. M., Abbott, K. C., Agodoa, L. Y., Albertus, P., Ayanian, J. Z., Balkrishnan, R., Bragg-Gresham, J. L., Cao, J., Chen, J. L., Cope, E., Dharmarajan, S., Dietrich, X., Eckard, A., Eggers, P., Gaber, C. E., Gillen, D. L., Gipson, D. S., Gu, H., Hailpern, S. M., Hall, Y. N., Han, Y., He, K., Hébert, P., Helmuth, M. E., Herman, W. H., Heung, M., Hutton, D., Jacobsen, S. J., Ji, N., Jin, Y., Kalantar-Zadeh, K., Kapke, A., Katz, R., Kovesdy, C. P., Kurtz, V., Lavalee, D., Li, Y., Lu, Y., McCullough, K. P., Molnar, M. Z., Montez-Rath, M. E., Morgenstern, H., Mu, Q., Mukhopadhyay, P., Nallamothu, B. K., Nguyen, D. V., Norris, K. C., O'Hare, A. M., Obi, Y., Pearson, J., Pisoni, R. L., Plattner, B. W., Port, F. K., Potukuchi, P. K., Rao, P. S., Ratkowiak, K., Ravel, V. A., Ray, D., Rhee, C. M., Schaubel, D. E., Selewski, D. T., Shaw, S. F., Shi, J. M., Shieu, M., Sim, J. J., Song, P. X., Soohoo, M., Steffick, D. E., Streja, E., Tamura, M. K., Tentori, F., Tilea, A. M., Tong, L., Turf, M., Wang, D., Wang, M., Woodside, K. J., Wyncott, A., Xin, X., Zang, W., Zepel, L., Zhang, S., Zho, H., Hirth, R. A., and Shahinian, V. B. (2017), "US Renal Data System 2016 Annual Data Report: Epidemiology of Kidney Disease in the United States," *American Journal of Kidney Diseases*, 65, A7–A8. [12]

Schumaker, L. (2007), *Spline Functions: Basic Theory* (3rd ed.), Cambridge: Cambridge University Press, Cambridge Mathematical Library. [5,8]

Shen, X. (1997), "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591. [2]

——— (1998), "Propotional Odds Regression and Sieve Maximum Likelihood Estimation," *Biometrika*, 85, 165–177. [4]

Shen, X., and Wong, W. H. (1994), "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615. [1,8]

Song, X., Ma, S., Huang, J., and Zhou, X. (2006), "A Semiparametric Approach for the Nonparametric Transformation Survival Model with Multiple Covariates," *Biostatistics*, 8, 197–211. [4,6,10]

Tang, W., Ma, J., Mei, Q., and Zhu, J. (2022), "Soden: A Scalable Continuous-Time Survival Model Through Ordinary Differential Equation Networks," *Journal of Machine Learning Research*, 23, 1–29. [14]

Tsiatis, A. A. (1990), "Estimating Regression Parameters Using Linear Rank Tests for Censored Data," *The Annals of Statistics*, 18, 354–372. [1,4]

Walter, W. (1998), "First Order Systems. Equations of higher order," in *Ordinary Differential Equations*, pp. 105–157, New York: Springer. [2]

Wang, Y., Hong, C., Palmer, N., Di, Q., Schwartz, J., Kohane, I., and Cai, T. (2019), "A Fast Divide-and-Conquer Sparse Cox Regression," *Biostatistics*, 22, 381–401. [15]

Wellner, J. A., and Zhang, Y. (2007), "Two Likelihood-based Semiparametric Estimation Methods for Panel Count Data with Covariates," *The Annals of Statistics*, 35, 2106–2142. [2,7]

Wolfe, R., Ashbyv, V., Milfordv, E., Ojov, A., Ettengerv, R., Agodoav, L., Heldv, P., and Portv, F. (1999), "Comparison of Mortality in all Patients on Dialysis, Patients on Dialysis Awaiting Transplantation, and Recipients of a First Cadaveric Transplant," *The New England Journal of Medicine*, 341, 1725–1730. [13]

Zeng, D., and Lin, D. Y. (2006), "Efficient Estimation of Semiparametric Transformation Models for Counting Processes," *Biometrika*, 93, 627–640. [4]

——— (2007a), "Efficient Estimation for the Accelerated Failure Time Model," *Journal of the American Statistical Association*, 102, 1387–1396. [4]

——— (2007b), "Maximum Likelihood Estimation in Semiparametric Regression Models with Censored Data," *Journal of the Royal Statistical Society*, Series B, 69, 507–564. [1,4,6,10]

Zhang, Y., Hua, L., and Huang, J. (2010), "A Spline-based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data," *Scandinavian Journal of Statistics*, 37, 338–354. [2]

Zhao, X., Wu, Y., and Yin, G. (2017), "Sieve Maximum Likelihood Estimation for a General Class of Accelerated Hazards Models with Bundled Parameters," *Bernoulli*, 23, 3385–3411. [1,2,7,8,9]

Zucker, D. M., and Karr, A. F. (1990), "Nonparametric Survival Analysis with Time-Dependent Covariate Effects: A Penalized Partial Likelihood Approach," *The Annals of Statistics*, 18, 329–353. [1,3,4]

Zuo, L., Zhang, H., Wang, H., and Liu, L. (2021), "Sampling-based Estimation for Massive Survival Data with Additive Hazards Model," *Statistics in Medicine*, 40, 441–450. [15]