A Single-Index Cox Model Driven by Lévy Subordinators

Ruixuan Liu*
Department of Economics
Emory University
Atlanta, Georgia 30322

First Version: September 2015
This Version: February 2016

Abstract

I propose a new duration model where the duration outcome is defined as the first time a Lévy subordinator—a stochastic process with stationary, independent and non-negative increments—crosses an exponential threshold, and the effect from observable covariate is acting multiplicatively on the latent process. Our specification is a natural variant of the mixed proportional hazard model from a stochastic process point of view. When the multiplicative effect is parameterized as a linear index, our model reduces to a single-index proportional hazard model, where the unknown link function captures the characteristic of the latent process. The large sample property of a sieve maximum partial likelihood estimator of the finite dimensional parameter is studied with right censored data.

Keywords: First Passage Time, Lévy Processes, Sieve Partial Likelihood, Semiparametric Efficiency, Single-Index Models

JEL Codes: C14; C24; C41

*Email: ruixuan.liu@emory.edu. This paper is based on the first chapter of my dissertation submitted to University of Washington. I would like to thank Professors Jaap Abbring, Yanqin Fan, Jianhua Z. Huang, Roger Klein, and Jon Wellner for helpful comments.
# 1 Introduction

Since the seminal work of Cox (1972), the proportional hazard model (PH) has unarguably been the focal point in the duration or survival analysis. PH is semiparametric in nature where the hazard density function of the duration outcome or survival time $T$ with covariate $X$ is specified as:

$$
\lambda_T(t|x) \equiv \lim_{\delta \to 0} \frac{1}{\delta} P\{t \leq T < t + \delta | T \geq t, X = x\}
$$

$$
= \lambda(t) \exp(x' \beta),
$$

(1.1)

depending on a nonparametric baseline hazard density function $\lambda(t)$ and a linear index $x' \beta$ via the exponential link function. Partial likelihood method (Cox, 1975) is used to estimate the regression coefficient $\beta$ without estimating the baseline hazard function given right censored survival data. With the estimator of $\beta$ in hand, the estimation of the baseline cumulative hazard function $\Lambda(t) = \int_0^t \lambda(s) ds$ readily follows from Breslow’s proposal (Breslow, 1972). This estimation procedure delivers root-$n$ consistent and asymptotic normal estimators for both the finite dimensional regression parameter and cumulative baseline hazard function, see Andersen and Gill (1982). More importantly, the estimators are semiparametrically efficient in the sense of Begun, Hall, Huang and Wellner (1983). Despite its theoretical elegance, PH comes with several restrictions. First of all, the covariates’ effect is linear in the conditional (log-)hazard function. Second, the plain Cox model does not incorporate any unobservable heterogeneity term. Last but not least, PH may not be applicable to structural durations triggered by stochastic processes, which naturally arises from optimal stopping time problems in economics. Indeed, developments have been made in the literature extending the scope of PH in all three directions. Maintaining the proportional hazard structure, the functional form $\exp(x' \beta)$ can be relaxed to a fully nonparametric version as in Tibshirani and Hastie (1987), Fan, Gijbels and King (1997), Chen and Zhou (2007) or semiparametric variants in Nielsen, Linton and Bickel (1998), Huang (1999), Fan, Lin and Zhou (2006). Introducing a multiplicative heterogeneity term (also known as the frailty term) in the conditional hazard density function, one arrives at Lancaster’s (1979) mixed proportional hazard model (MPH). Finally, the recent surge of process-based duration models of Abbring (2012) or Botosaru (2013) also borrow several insights from Cox’s original construction, while being able to capture economic agents’ optimizing behaviors in a dynamic setup with certain time-varying heterogeneity.

In this paper, I propose a new duration model that inherits prominent features from all three aforementioned branches. Here the duration outcome is defined as the first time a Lévy subordinator—a stochastic process with stationary, independent and non-negative increments—crosses an exponential threshold, and the effect from observable covariate is acting multiplicatively on the latent process. Such a model is of substantial interest because not only is it related to optimal stopping time models where agents optimally time their discrete actions, but it also applies to the scenario in which the termination of duration is caused by some gradual and irreversible accumulation of damage. In fact, I provide a unified framework of duration modelling, where the duration outcome is specified as the first passage time of a latent stochastic process crossing over a random threshold, including both traditional hazard-based models and recent process-based...
models. Thereafter, our model could be seen as a natural variant of the mixed proportional hazard model from a stochastic process point of view. Even though the model is constructed in a highly structural way, its empirical content is very tractable. When the multiplicative effect is parameterized as a linear index, our model exhibits a single-index structure via the conditional hazard function:

\[ \lambda_T(t|x) = \lambda(t) \exp \left[ \psi \left( x' \beta \right) \right] \]  

(1.2)

with an unknown smooth link function \( \psi(\cdot) \). This single-index model incorporate the nonlinear effect while alleviating the curse of dimension when multivariate covariates are considered. The restrictive parametric link function such as the identity map in PH is avoided, while greater estimation precision for finite dimensional parameter \( \beta \) is still possible, in contrast to fully nonparametric estimation. Notice this unknown link function is not just motivated from technical relaxation, but it is related to the characteristics of the latent stochastic process.

The contributions of our paper are mainly threefold. First, we present a unified perspective of modelling duration outcomes. The synthesis is that we could always view the duration or survival time as the first passage time of a certain stochastic process, bridging the conceptual gap between the existing hazard-based models and the process-based models. Second, our new construction of the structural duration model driven by Lévy subordinators could be seen as a natural variant of MPH where this latent process replaces the role of the static frailty term, embedding time-varying heterogeneity. The tractability associated with Lévy subordinators lead us to investigate econometric properties of the single-index PH model. The third contribution of this paper is a complete study of the large sample properties of the estimator for \( \beta \) from sieve partial likelihood maximization. Namely, the sieve maximum partial likelihood estimator for \( \beta \) is proved to be asymptotically normal with root-\( n \) convergence rate and it achieves the semiparametric efficiency bound. Furthermore, I present the efficient score function and information bound in explicit forms, and prove the bootstrap consistency with general exchangeable bootstrap weights.

We are now in a position to describe the agenda for this paper. Section 2 contains a brief survey of the related literature. Section 3 provides a unified framework of the duration modelling. Thereafter, I present the single-index Cox model driven by Lévy subordinators as well as its identifiability and estimation in Section 4, while Section 5 containing main theoretical results. A small scale Monte Carlo experiment is conducted in Section 6 examining the finite sample performance of the bootstrapped confidence intervals. We conclude in Section 7. Technical lemmas and proofs are collected in the Appendix. For a vector \( v \), we let \( v' \) be the transpose of it and \( v \otimes 2 = vv' \) in the sequel.

## 2 Literature Review

In this section, we first review the related literature extending the scope of Cox’s PH model in three categories. Then we discuss the existing development of single-index PH model.

Although the baseline hazard function is left unspecified in Cox model, the effect of covariates on the
(log-)hazard function is restricted to be linear. Obviously, any misspecification of the effect of covariates would lead to inconsistent estimation and misleading inference, which motivates consideration of flexible nonparametric or semiparametric methods. For a nonparametric smooth risk function \( \phi(x) \) measuring the covariates’ effect, Tibshirani and Hastie (1987) initiate the local partial likelihood for estimating the nonparametric Cox model. Subsequently, Fan, Gijbels and King (1997), Chen and Zhou (2007) have developed large sample theory for estimating the derivative or the difference of the unknown risk function. For various semiparametric siblings, we refer the readers to the semiparametric model with parametric baseline function in Nielsen, Linton and Bickel (1997), the partial linear and additive model in Huang (1999), the functional ANOVA model in Huang, Kooperberg, Stone, and Truong (2000), and the varying-coefficient model in Fan, Lin, and Zhou (2006).

The characterization of individual heterogeneity is a central issue in econometric duration analysis. Ignoring the unobservable heterogeneity would inevitably bias the estimation and interpretation of the effect from explanatory variables. Lancaster’s (1979) mixed proportional hazard (MPH) model has become a standard toolkit allowing researchers to incorporate both observed and unobserved variables into the conditional hazard function of the duration outcome. Relaxing the parametric assumption in Lancaster (1979), Elbers and Riddler (1982), Heckman and Singh (1984) have demonstrated the possibility of identifying MPH with a single spell of duration in a completely nonparametric way, further broadening its scope. When the heterogeneity follows the gamma distribution as in Lancaster (1979), Han and Hausman (1990) propose an estimator assuming the baseline hazard function is piece-wise constant. Horowitz (1999) comes up with the first nonparametric estimator allowing both baseline hazard and frailty’s distribution to be unspecified, see also Chen (2002). Without parameterizing the heterogeneity term’s distribution function, the semiparametric efficiency bound of MPH for the finite dimensional parameters in multiplicative effect and baseline hazard function could be singular as first realized by Hahn (1994), which excludes the possibility of obtaining root-\( n \) consistent estimators. Ridder and Woutersen (2003) present sufficient conditions which avoid this singularity. For this regular semiparametric model, Bearse, Canals-Cerda, and Rilstone (2007) have studied semiparametric efficient estimation. Hausman and Woutersen (2014) propose a different root-\( n \) consistent estimator for the finite dimensional parameters incorporating time-varying covariates and discretely observed durations.

Needless to say, the hazard based duration models play dominant roles in empirical applications, but economic theory in general does not lead to the above specification, which complicates the structural interpretation of various reduced-form estimates (van den Berg, 2001). A typical duration of interest to economists very often arises from solving an optimal stopping time problem, where optimizing economic agents make decisions about the time at which to switch from one state to another, such as in firms’ entry/exit choice (Dixit, 1989), worker’s unemployment decision (Mortensen and Pissarides, 1994), and real option type investment (McDonald and Siegels, 1986). In these examples, the structural duration is related to some threshold crossing behavior of a latent stochastic process (Abbring, 2012). Indeed what is usually
disregarded in the traditional approach to duration or survival analysis is that the particular duration is the end point of some process (Aalen, Borgan, and Gjessing, 2008; Lee and Whitmore, 2006). Parallel to the traditional hazard-based modeling strategy, many process-based models have been developed by making use of a specific parametric sub-class of Lévy processes. For example, the Brownian motion appears in Lancaster’s (1972) strike model and Whitmore’s (1979) job tenure model, while the gamma process has been utilized in Singpurwalla (1995), Lawless and Crowder (2004) to model the degradation process. Abbring (2012) has pioneered the study of identifiability of process-based duration models, without making any parametric assumption. Abbring (2012) specifies the duration formally as first passage time of a random threshold by the sample path of a latent stochastic process in his Mixed Hitting-Time model (MHT). Abbring (2012) has demonstrated that not only MHT incorporates the time-varying heterogeneity represented by the latent process, but also it arises more naturally from the optimal stopping time problem where the solution could be described by this threshold-crossing behavior. When the process is chosen to be the spectral negative Lévy process, the identifiability of all structural components could be shown, adapted from the strategies analyzing the identifiability of MPH. So far the identification of MHT is restricted to the univariate duration model, and the suggested estimation is fully parametric. Botosaru (2013) proposes another univariate duration model making use of the Lévy subordinator, which could be either viewed as a hazard-based model with certain random hazard density or a process-based model with the latent process being a double stochastic integral of the Lévy subordinator. Identification in Botosaru’s (2013) random hazard density model relates to solving a nonlinear integral equation. For the identified case, Botosaru (2013) presents a sieve type estimator and proves its consistency.

There is no doubt that the single-index structure is one of the most popular semiparametric modelling choice for conditional mean function (Ichimura, 1993), conditional distribution function (Klein and Spady, 1993; Chen, 2007), or conditional quantile function (Ichimura and Lee, 2010) and its extension to the Cox model has already been envisioned earlier (Ichimura, 1993; Nielsen, Linton and Bickel; Chen, 2007; Ding and Nan, 2011); However, the rigorous large sample theory for (1.2) has not been presented in a entirely satisfactory fashion when right censored data is available. Chen, Li, and Wang (1999) are among the first to consider the model in (1.2) based on sliced inverse regression technique which requires very strong assumptions on the covariates, such as symmetry. Wang (2004) suggests to iterate certain kernel based estimator while assuming some preliminary root-$n$ consistent estimator for $\beta$ is available beforehand, which is presumably taken to be the estimator in Chen, Li, and Wang (1999). Huang and Liu (2006) have proposed maximizing sieve partial likelihood with $\psi$ being replaced by a polynomial spline function, but the number of knots is held to be fixed. Consequently, they treat the link function exactly to be some spline function under consideration and derive parametric type asymptotics assuming one of the spline functions with the priorly determined number of knots is the true unknown link function.
3 A Unified Perspective of Durations Modelling

We provide a unified perspective on the structural duration model described by the threshold-crossing behavior of some latent process, including PH in Cox (1972), MPH in Lancaster (1979), MHT in Abbring (2012), and a random hazard density model (RHD) in Botosaru (2013).

Now we start with a generic duration model where the duration is related to the threshold crossing behavior of a latent stochastic process. Formally, the structural duration is defined as the first passage time of a latent process $\Lambda^*(t)$ crossing a random threshold $e$, with structural proportional effect $\phi(X)$ acting on the process:

$$ T \equiv \inf \{ t : \phi(X) \Lambda^*(t) \geq e \}. $$

(3.1)

In all models discussed in the paper, $X$, $\Lambda^*(t)$ and $e$ are assumed to be mutually independent. The rule specified in (3.1) resembles the classical binary choice model closely in a dynamic setup.\textsuperscript{1} But instead of making a discrete choice based on static comparisons, it stems from the optimal stopping time model where agent optimally chooses when to switch to another state at time $T$ with complete information, as in de Paula and Honore (2010). Namely, the economic agent is facing a utility maximization problem where the current state generates a payoff to the agent with $e$ and switching to another state gives $\phi(X) \Lambda^*(t)$, i.e., the duration in (3.1) is the optimal solution of

$$ \max_T \left\{ \int_0^T e \exp(-\rho s) \, ds + \int_T^{\infty} \phi(X) \Lambda^*(t) \exp(-\rho s) \, ds \right\} $$

with the discount rate equal to $\rho$.

In spite of the fact that PH or MPH is constructed from a regression modeling on the hazard rate, it is interesting to ask whether it could be put into the same framework where the duration is specified as the first passage time of a latent process. Indeed this is possible. This alternative point of view further reveals the restrictions on the latent process and threshold imposed by MPH, and it provides the link of our specification to MPH.

Example 3.1 (PH) It turns out that the duration in Cox model could be equivalently defined as in (3.1)

$$ T \equiv \inf \{ t : \exp(X') \Lambda(t) \geq e \}. $$

(3.2)

by setting $\Lambda^*(t)$ equal to the cumulative baseline hazard function $\Lambda(t)$. Furthermore, the multiplicative effect is parameterized as $\exp(X' \beta)$, and the threshold follows unit exponential distribution, i.e., $e \sim \text{Exp}(1)$. The assertion is easy to see as

$$ \Pr \{ T > t | X = x \} = \Pr \{ e > \Lambda(t) e^{x' \beta} | X = x \} = \exp \left( - \Lambda(t) e^{x' \beta} \right). $$

\textsuperscript{1}In discrete-time setting, Heckman and Navarro (2007) have constructed a general mixture duration model based on a latent process crossing thresholds.

\textsuperscript{2}Because only the comparison plays a role from the agent’s perspective, it is not restrictive to let one side to become time invariant.
This alternative view is favored by Singpurwalla (2006) who has advocated this interpretation of the cumulative function \( \Lambda(t) \) is more sensible than merely treating it as the primitive of a hazard density.

**Example 3.2 (MPH)** The duration in a nonparametric MPH could be also defined as in (3.1) by

\[
T \equiv \inf \{ t : \phi(X) \zeta \Lambda(t) \geq e \},
\]

(3.3)

where the latent process is equal to \( \Lambda^*(t) = \zeta \Lambda(t) \) with the cumulative baseline hazard function \( \Lambda(t) \). Again, the threshold follows unit exponential distribution, i.e. \( e \sim \text{Exp}(1) \).

In those two classical hazard-based models, the driving stochastic process is either deterministic (as in PH) or its randomness is pinned down at time zero (the frailty term \( \zeta \) is static in MPH). In MPH given the realization of \( \zeta \), we will just have a deterministic and nondecreasing trend \( \Lambda(t) \) approaching the threshold from below. The individual heterogeneity does not evolve over time along the entire spell of duration, hence it is certainly desirable to construct a model with time-varying heterogeneity. Apparently, one has to make some assumptions on the underlying class for \( \Lambda^*(t) \), in order to maintain a tractable structure. In the sequel, we shall restrict our attention to general Lévy processes.

Lévy processes constitute a very rich and attractive class of stochastic processes, including the commonly encountered Brownian motion, gamma process and stable process as special cases. They have attracted considerable attention due to the flexibility for a wide variety of modeling issues in finance, insurance and engineering, see Kyprianou (2006). In the study of duration models, parallel to the traditional hazard-based modeling, many process-based models have been developed by making use of a specific parametric sub-class of Lévy processes. For example, Brownian motion appears in Lancaster’s (1972) strike model and Whitmore’s (1979) job tenure model, while gamma process has been utilized in Singpurwalla (1995), Lawless and Crowder (2004) to model the degradation process. Abbring (2012) has pioneered the study of identifiability of process based duration models, without making any parametric assumptions. Before reviewing his contribution, it is worthwhile to introduce the formal definition of a univariate Lévy process and two important subclasses. To avoid digressions, we refer the readers to Sato (1999) and Kyprianou (2006) for more detailed discussions.

**Definition 3.3** A univariate Lévy process \( L(t) \) is a right continuous stochastic process with left limits such that for every \( t \) and \( r \geq 0 \), the increment \( L(t + r) - L(t) \) is independent of \( \{L(s) : 0 \leq s \leq t\} \) and has the same distribution as \( L(r) \).

**Definition 3.4** A univariate Lévy subordinator \( L(t) \) is a Lévy process with almost surely nondecreasing sample path, i.e. for \( t \geq s \) one has \( L(t) \geq L(s) \).

**Definition 3.5** A univariate spectral negative Lévy process \( L(t) \) is a Lévy process with no positive jumps.

The most remarkable property of a Lévy process is that even it is purely defined in terms of descriptive features about the sample path, it admits a very concrete analytic characterization via the well-known
Lévy-Khintchine representation, see Sato (1999). Specifically, its Laplace transform could be expressed as

\[ E \{ \exp [-zL(t)] \} = \exp [-t\Phi(z)], \]

where \( \Phi(\cdot) \) is the so-called Lévy-Laplace exponent function, which is nonparametric and time-invariant. This elegant analytic characterization plays a crucial role in identification problems of those duration models driven by Lévy processes, including Abbring (2012) and Botosaru (2013).

**Example 3.6 (MHT)** Motivated from the optimal stopping time problems, Abbring (2012) starts with a model where the structural duration is formally defined to be

\[ T \equiv \inf \left\{ t : \phi(X) \tilde{L}(t) \geq e \right\}, \tag{3.4} \]

where \( e \) is an arbitrary random threshold and \( \tilde{L}(t) \) is a spectral negative Lévy process. The empirical content of MHT is revealed through the conditional Laplace transform \( \mathcal{L}_T(s|X) \) of the structural duration \( T \):

\[ \mathcal{L}_T(s|X) \equiv E \{ \exp (-sT) | X \} = \mathcal{L}_e \left( \hat{\Lambda}(s) \phi(X) \right), \tag{3.5} \]

where \( \mathcal{L}_e \) is the Laplace transform of \( e \) and \( \hat{\Lambda}(s) \) is the largest root satisfying \( \Phi(\hat{\Lambda}(s)) = s \). Despite that there are quite different functions (i.e. \( \mathcal{L}_T(s|X) \) and \( \hat{\Lambda}(s) \)) appearing in (3.5), the mathematical structure is almost identical to the one in MPH. Assuming regular variations of certain functions, Abbring (2012) has shown one could identify the triple \( (\mathcal{L}_e, \hat{\Lambda}, \phi) \) from \( \mathcal{L}_T(s|X) \) nonparametrically.

**Example 3.7 (RHD)** Given our discussion on linking the hazard-based models to their threshold-crossing behavior over the unit exponential threshold, it is clear now that the duration in Botosaru’s (2013) random hazard density model could be expressed as

\[ T \equiv \inf \left\{ t : \phi(X) \Lambda^*(t) \geq e \right\}, \]

where the latent crossing process is a double stochastic integral of the Lévy subordinator, as \( \Lambda^*(t) = \int_0^t \int_0^s f(u) dL(u) ds \) with a transformation function \( f(u) \). When the process \( L(u) \) is taken to be a Lévy subordinator, the conditional survival function could be written as

\[ S_T(t|x) = \exp \left[ - \int_0^t \Phi(\phi(x) f(u)(t-u)) du \right]. \]

The nonparametric identification of \( (\Phi, f, \phi) \) is related to solving a nonlinear Volterra integral equation of the first kind with unknown kernel, see Botosaru (2013).

**4 A New Duration Model**

Inspired by Abbring (2012) and Botosaru (2013), I construct a new duration model under the framework of (3.1), replacing the product of the cumulative baseline hazard function and the frailty term \( \zeta \Lambda(t) \) in MPH all together with a latent stochastic process \( L(\Lambda(t)) \). Here \( L(\cdot) \) is a Lévy subordinator and \( \Lambda(t) \) is
a deterministically increasing time-change function. This time-change function could be seen as a suitable transformation from certain abstract time scale that the process evolves to the rate of economic transactions (Kyprianou, 2006), which simply turns out to be the cumulative baseline hazard function. Hence the duration in our model becomes

\[ T \equiv \inf \{ t : \phi(X) L(\Lambda(t)) \geq e \} , \tag{3.6} \]

where the random threshold is still assumed to be unit exponentially distributed. Notice the latent Lévy subordinator \( L(\cdot) \) has replaced the static frailty term \( \zeta \) and its sample path property inherits the key feature of a monotonic driving trend in MPH. Another close connection to MPH is that the distribution of frailty term is often taken to be infinitely divisible, see Hougaard (1986). Correspondingly in our present model, it’s well known the Lévy subordinator has non-negative infinitely divisible distribution (Sato, 1999).

The conditional survival function of \( T \) in (3.6) could be derived in a straightforward way by the Lévy-Khintchine representation:

\[ S_T(t|x) = \exp \left( -\Lambda(t) \Phi(\phi(x)) \right) , \tag{3.7} \]

which leads to the conditional hazard density function being

\[ \lambda_T(t|x) = \lambda(t) \Phi(\phi(x)) , \tag{3.8} \]

where \( \lambda(t) \) is the derivative of \( \Lambda(t) \).

Our new model complements the traditional hazard-based models and the recent study in Abbring (2012) and Botosaru (2013), with the great advantage that the proportional hazard structure is maintained as in (3.8). The target here is not to argue the current model (3.6) is more general than any other existing proposal, but mainly to offer an interesting and somewhat surprising alternative which delivers very tractable econometric or statistical properties.

Now we present some examples with the latent Lévy subordinator restricted to be chosen from specific parametric sub-classes, which make the function \( \Phi(\cdot) \) parameterized.

**Example 4.1** When \( L(\cdot) \) is the gamma process, our model could be seen as a natural variant of MPH with gamma frailty term as originally in Lancaster (1979). One nice property of the gamma frailty model is that the conditional distribution of frailty term \( \zeta \), given survival until any time (i.e. conditional on \( Y^* > t \)), is also gamma with the original shape parameter, see van den Berg (2001). Abbring and van den Berg (2007) provide another justification, showing that the distribution of frailty term among survivors would always converge to a gamma distribution upon suitable normalization when \( t \to \infty \). In Figure 1, we plot a typical realization of the first passage time or structural duration with gamma process in our model (1.2). Its conditional survival function is

\[ S_T(t|x) = \exp \left( -\Lambda(t) \left( 1 + \frac{\phi(x)}{\nu} \right)^\rho \right) , \]

where \((\nu, \rho)\) stand for the scale and shape parameter in the gamma distribution.
Example 4.2 When \( L(\cdot) \) is the stable process, our model could be seen as a natural variant of MPH with stable frailty term as originally in Hougaard (1986). One advantage of the stable frailty model under MPH is that one still gets a proportional hazard model when integrating out \( \zeta \) (see Section 5.2 in van den Berg, 2001), which is not the case for the gamma frailty model. In Figure 2, we plot a typical realization of the first passage time or structural duration with stable process in our model (1.2). Its conditional survival function is

\[
S_T(t|\mathbf{x}) = \exp\left[-\Lambda(t) \nu \phi(x)^\rho\right],
\]

where \((\nu, \rho)\) represent the stable index and power parameter in the stable distribution. Compared with gamma process, the sample path of stable process is more irregular and exhibits larger jumping magnitude.

Example 4.3 When we take the Levy subordinator to be a compound Poisson process, i.e., \( L(t) = \sum_{i=1}^{N(t)} \xi_i \) where the arrival of shocks is governed by a (homogeneous) Poisson process \( N(t) \) with hazard rate \( \lambda \) and individual i.i.d. shocks \( \{\xi_i\} \) are assumed to non-negative. In the absence of covariate \( \mathbf{X} \), our model boils down to the random shock model studied by Esary, Marshall and Proschan (1973). Under this circumstance, the marginal survival function of \( T \) could be determined explicitly by model primitives without referring to the Lévy-Khintchine representation. The marginal survival function of \( T \) is equal to

\[
\Pr\{T > t\} = \sum_{k=0}^{\infty} \frac{(\lambda t)^k \exp(-\lambda t)}{k!} \mathcal{P}_k,
\]

where \( \mathcal{P}_k \) is the probability of surviving after \( k \) shocks:

\[
\mathcal{P}_k = \Pr\{\xi_1 + \ldots + \xi_k \leq e\}.
\]

It is well known that the compound Poisson process has a piece-wise constant sample path, and it is the only process whose Lévy measure is finite (see Sato, 1999).

4.1 Identification of the Single-Index Model

Given the proportional hazard structure in our model (3.7), disentangling the effects from covariates and latent stochastic processes boils down to study the composition of two function. Once the multiplicative effect \( \phi(x) \) is parameterized to be \( \exp(\mathbf{x}' \beta) \) as in the Cox regression, we actually arrive at a single-index proportional hazard model with unknown link function

\[
\psi(\cdot) = \log \circ \Phi \circ \exp(\cdot), \tag{3.9}
\]
as in (1.2). The identification of all model primitives in (3.7) could be achieved without invoking the identification-at-limit strategy, in contrast with Elbers and Ridder (1982), Heckman and Singer (1984), or Abbring (2012). It is well-known the finite dimensional parameter \( \beta \) is only identified up to scale in the single-index model (Ichimura, 1993), so we adopt the following convention by fixing the first element to be 1 (see Ichimura and Lee, 2010). We partition the covariate vector \( X = \left( X_1, X_2 \right) \) with a univariate component \( X_1 \), and stack \( \beta \) with one 1 as \( \tilde{\beta} = (1, \beta')' \). Moreover, we denote \( X_0 = X_1 + X_2' \beta_0 \) with the true parameter \( \beta_0 \). The regularity conditions below required for identification are fairly weak.

**Assumption (I1)** The support of \( X \) is a convex set with at least one interior point.

**Assumption (I2)** The support of \( X_0 \) contains a nonempty open interval.

**Assumption (I3)** Assume \( E[\psi(X_0)] = 0 \).

**Proposition 4.4** Let \( \phi(x) = \exp \left( x' \tilde{\beta} \right) \) in (3.7) and assume Assumptions (I1)-(I3), then the triple \((\beta, \Phi, \Lambda)\) is identifiable.

Our specification regarding in how the covariates enter into the model is only one of many possible choices. Indeed, one could consider a nonparametric multiplicative function as \( \phi(x) = \phi_1(x_1) \cdots \phi_d(x_d) \), and the problem boils down to the identification of the transformation model (Horowitz, 1999; Chiappori, Komunjer, and Kristensen, 2015) or the generalized accelerated failure time model (Ridder 1990; Abbring and Ridder, 2015). To be consistent with the asymptotic analysis in sequel, we shall focus on the single-index model.

### 4.2 Sieve Maximum Partial Likelihood Estimation

In practice, survival data is often right censored due to termination of the study or early withdrawal from the study, so we observe the random sample consisting of i.i.d data of \( Z_i = (V_i, \Delta_i, X_i) \) where \( V_i = \min(T_i, C_i) \) and \( \Delta_i = I[T_i \leq C_i] \). The logarithm of partial likelihood function (neglecting terms independent of parameters of interest) is

\[
n(\beta, \psi) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \left\{ \log \left( \sum_{k: V_k \geq V_i} \exp \left[ \psi \left( X_{k1} + X_{k2}' \beta \right) \right] \right) \right\}.
\]

Since the smooth link function \( \psi \) is unknown, it is natural to replace it by an approximating B-spline function\(^3\) \( \psi_n(\cdot) \) (Schumaker, 1981). Let \( T_{K_n} = \{ t_1, ..., t_{K_n} \} \) be a set of partition points of \([a, b]\) with \( K_n = O(n^v) \) and \( \max_j |t_j - t_{j-1}| = O(n^{-v}) \) for some constant \( v \in (0, 1/2) \). Let \( S_n(T_{K_n}, K_n, p) \) be the space of polynomial splines of order \( p \geq 1 \), then there exists a set of B-spline basis functions \( \{ B_j, 1 \leq j \leq q_n \} \) with \( q_n = K_n + p \) s.t. for any \( s \in S_n(T_{K_n}, K_n, p) \) we can write \( s(\cdot) = \sum_{j=1}^{q_n} \gamma_j B_j(\cdot) \).

Now with \( \psi_n(\cdot) = \sum_{j=1}^{q_n} \gamma_j B_j(\cdot) \) substituting \( \psi \), we are maximizing the following criterion function in terms of parameters \((\beta, \gamma)\):

\[
n(\beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \left\{ \sum_{j=1}^{q_n} \gamma_j B_j \left( X_{i1} + X_{i2}' \beta \right) - \log \left( \sum_{k: V_k \geq V_i} \exp \left[ \sum_{j=1}^{q_n} \gamma_j B_j \left( X_{k1} + X_{k2}' \beta \right) \right] \right) \right\}.
\]

\(^3\)Other sieve basis could potentially work as well, see those presented in Chen (2007).
A Newton-Raphson algorithm or any gradient-based search algorithm can be applied to solve for the score equations to get estimators \((\hat{\beta}, \hat{\gamma})\). Given the identification restriction of \(\psi\), we will center the estimator as follows. Let

\[
\hat{\psi}_n^\star (\cdot) = \sum_{j=1}^q \hat{\gamma}_j B_j (\cdot) \quad \text{and} \quad \bar{\psi}_n^\star (\cdot) = \frac{\sum_{i=1}^n \Delta_i \hat{\psi}_n^\star (\cdot)}{\sum_{i=1}^n \Delta_i}.
\]

The resulting estimator of \(\psi\) is a centered version and defined to be

\[
\hat{\psi}_n (\cdot) \equiv \hat{\psi}_n^\star (\cdot) - \bar{\psi}_n^\star (\cdot),
\]

so it satisfies \(\sum_{i=1}^n \Delta_i \hat{\psi}_n (X_{i1} + X_{i2} \hat{\beta}) = 0\), see a similar procedure in Huang (1999). We refer readers to Huang and Liu (2006) for detailed descriptions of the computation and simulation results.

## 5 Theoretical Results

It is clear that our construction results in a single-index Cox model where the unknown link function relates to the Lévy exponent function of the underlying Lévy subordinator. In fact, this link function defined by (3.9) is infinitely order differentiable, because the Lévy exponent function \(\Phi (\cdot)\) possesses derivatives of any finite order with alternating signs. Formally, \(\Phi (\cdot)\) is a Bernstein function, see Schilling, Song and Vondracek (2012). From now on, we focus on a more general class of models where the link function \(\psi (\cdot)\) is regularly smooth, hence it is only required to be differentiable up to a finite order to be specified later.

A close examination of the existing literature on semiparametric estimation theory reveals the novelty and technicality associated with (1.2). In contrast with Huang (1999), we are dealing with the bundled parameter \((\beta, \eta (\cdot, \beta))\) where \(\eta (x, \beta) = \psi (x \beta)\) in the present scenario. This terminology is used by Huang and Wellner (1997) referring to statistical models where the parameter of interest and nuisance parameter are bundled together. General theory for handling semiparametric estimation with bundled parameter has been developed by Ding and Nan (2011) for some criterion function which could be written as the sample average of \(i.i.d\). observations. In our estimation, the sample criterion function is chosen to be the logarithm of partial likelihood function, which is neither an average of \(i.i.d\). terms as in Ichimura and Lee (2010), Ding and Nan (2011), nor weakly dependent as in Chen (2007) due to the counting over the at-risk set, see Andersen and Gill (1982). Compared with Nan and Wellner (2013), we have to approximate the additional unknown function \(\psi (\cdot)\) by spline sieves. In fact, Ding and Nan (2011) lists the extension to the single-index Cox model as a challenging open problem. All that said, our large sample theory for (1.2) is indeed a successful marriage of the aforementioned papers.

Now we introduce some necessary notations. Let \(P_n\) be the empirical measure of \(Z_i \equiv (V_i, \Delta_i, X_i)\) and let \(P\) be the associated probability measure. Let \(P_{\Delta n}\) be the subprobability empirical measure of \(Z_i\) when \(\Delta_i = 1\), with its population version \(P_{\Delta n}\). The linear functional notation is convenient, as \(P_{\Delta n} f = \int f \Delta dP_{\Delta n} = n^{-1} \sum_{i=1}^n \Delta_i f (Z_i)\). Recall that we let the true parameters be denoted as \((\beta_o, \psi_o, \lambda_o)\), and \(X = (X_1, X_2)\) and \(X_0 = X_1 + X_2 \beta_o\). Furthermore, stack \(\beta\) with one 1 as \(\tilde{\beta} = (1, \beta')\). The following
standard counting process notations in Andersen and Gill (1982) would facilitate subsequent presentation, so \( N_i(t) = I[V_i \leq t, \Delta_i = 1] \), \( Y_i(t) = I[V_i > t] \) and
\[
M_i(t) = N_i(t) - \int_0^t Y_i(u) \exp \left[ \psi_o \left( X_{i1} + X_{i2}' \beta_o \right) \right] \lambda_o(u) \, du,
\]
are the corresponding counting process, at-risk process and associated martingale respectively. The following important observation by Sasieni (1992a) would appear repeatedly in the remaining section
\[
E[K(t)g(Z)] = E[g(Z)|V = t, \Delta = 1]
\]
where
\[
K(t) = \frac{Y(t) \exp \left[ \psi_o \left( X_{i1} + X_{i2}' \beta_o \right) \right]}{E \left\{ Y(t) \exp \left[ \psi_o \left( X_{i1} + X_{i2}' \beta_o \right) \right] \right\}}.
\]

The regularity assumptions are listed below.

**Assumption (A1)** We have i.i.d. observations of \( Z_i \equiv (V_i, \Delta_i, X_i) \). The duration outcome \( T \) and censoring time \( C \) are conditionally independent given covariates \( X \).

**Assumption (A2)** The finite dimensional parameter space \( \mathcal{B} \) is a compact subset of \( \mathbb{R}^d \) and the true parameter \( \beta_o \) is an interior point of \( \mathcal{B} \). The covariate \( X \) has bounded support and for any \( \beta \neq \beta_o \) we have
\[
\Pr \left\{ X_{i2}' \beta \neq X_{i2}' \beta_o \right\} > 0. \quad \text{EX}_2^{\otimes 2} \text{is a stricty positive definite matrix.}
\]

**Assumption (A3)** The truncation time \( \tau < \infty \) satisfies \( \Lambda_o(\tau) = \int_0^\tau \lambda_o(t) \, dt < \infty \), moreover \( \Pr (\Delta = 1|X) > 0 \) and \( \Pr (T > \tau|X) > 0 \) almost surely.

**Assumption (A4)** Let \( 0 < c_1 < c_2 < \infty \) be two constants. The joint sub-density \( f \left( t, x' \beta_o, \Delta = 1 \right) \) satisfies \( c_1 \leq f \left( t, x' \beta_o, \Delta = 1 \right) < c_2 \) for all \( (t, x' \beta_o) \in [0, \tau] \times [a, b] \).

**Assumption (A5)** For two arbitrary univariate functions \( f_1, f_2 \), we have
\[
f_1(V) + f_2(X_0) = 0, a.s.-P_{\Delta} \text{ if and only if } f_1(V) = f_2(X_0) = 0, a.s.-P_{\Delta}
\]

**Assumption (A6)** Let \( \Psi^p \) denote the collection of bounded functions \( \psi \) on \([a, b]\) with bounded derivatives \( \psi^{(j)}, j = 1, \ldots, k \) and the \( k \)-th derivative \( \psi^{(k)} \) satisfies the following Lipschitz condition:
\[
\left| \psi^{(k)}(s) - \psi^{(k)}(t) \right| \leq L |s - t|^{\alpha} \text{ for } \forall s, t \in [a, b]
\]
where \( k \) is a positive integer and \( \alpha \in (0, 1] \) s.t. \( p = k + \alpha \geq 3 \) and \( L < \infty \). The true unknown link function \( \psi_o(\cdot) \in \Psi^p \).

**Remark 5.1** Assumptions (A1)-(A3) are standard for deriving large sample properties of estimators in PH or its semiparametric variants, see Andersen and Gill (1982), Huang (1999), Ding and Nan (2011), Nan and Wellner (2013). The conditional independence assumption in (A1) ensures the censoring mechanism does not affect the identification. For the dependent censoring and partial identification analysis, we refer readers to Fan and Liu (2013). Assumptions (A4)-(A5) are needed to ensure the projections and information bound are well defined while one considers the projection onto the sumspace of tangent sets, see Sasieni
(1992b). Assumption (A6) imposes smoothness restriction of the unknown link function in order to utilize the nonparametric estimation. Notice that our postulated model driven by Lévy subordinators leads to a link function possessing derivatives of any order, hence it satisfies the Assumption (A6) automatically.

5.1 Semiparametric Information Bound

In this section, we calculate the semiparametric information bound for the estimation of $\beta$. We refer readers to Bickel, Klaassen, Ritov, and Wellner (1993) for an authoritative and book-length treatment of the semiparametric information bound for parameters in infinite dimensional models. In the standard single-index model via the conditional moment restrictions, the calculation of the efficient score function is done by projecting onto a single tangent space (Ichimura, 1993; Klein and Spady, 1993; Newey and Stoker, 1993). In contrast, we have two nuisance nonparametric components in the model. Thereafter, we have to consider the projection onto a sum-space of two non-orthogonal tangent spaces (see Bickel, Klaassen, Ritov, and Wellner, 1993). The iterated projection presented below could be viewed as a variant of the classical result of Frisch and Waugh (1933) to obtain the least-squares estimates for a sub-vector, which states the estimated regression parameter of interest is algebraically equal to run the least-squares regression of one set of residuals against the other. Recall the true parameters are denoted as $(\beta_0, \psi_0, \lambda_0)$, and $X = \begin{pmatrix} X_1, X_2 \end{pmatrix}'$ and $X_0 = X_1 + X_2 \beta_0$.

First, note that the log-likelihood function for a sample of size one is

$$\Delta \psi \left( X_1\tilde{\beta} \right) + \Delta \lambda (V) - \exp \left[ \psi \left( X_1\tilde{\beta} \right) \right] \Lambda (V),$$

dropping the terms which do not involve parameters of interest. Consider a parametric smooth submodel \{\lambda_\chi (\cdot): \chi \in R\} and \{\psi_\gamma (\cdot): \gamma \in R\} that runs through the true model, i.e., $\lambda_0 (\cdot) = \lambda_\alpha (\cdot)$ and $\psi_0 (\cdot) = \psi_\gamma (\cdot)$. Moreover, we let

$$a (\cdot) = \frac{\partial}{\partial \chi} \log \lambda_\chi (\cdot) \quad \text{and} \quad h (\cdot) = \frac{\partial}{\partial \gamma} \log \psi_\gamma (\cdot)$$

represent possible directions that can approach the true model. Thereafter following Sasieni (1992ab), the score vector for regression parameter $\beta_0$ is

$$i_\beta (Z) = \Delta \psi \left( X_1\tilde{\beta} \right) X_2 - \exp \left[ \psi \left( X_1\tilde{\beta} \right) \right] \Lambda (V) \psi \left( X_1\tilde{\beta} \right) X_2$$

$$= \int \psi \left( X_1\tilde{\beta} \right) X_2 dM (t),$$

where $M (t)$ is the natural martingale in (3.10). Likewise, we have the following two score operators for the nonparametric components:

$$i_\psi h (Z) = \int h \left( X_1\tilde{\beta} \right) dM (t),$$

$$i_\lambda a (Z) = \int a (t) dM (t),$$

where $a (\cdot) = \frac{\partial}{\partial \chi} \log \lambda_\chi (\cdot)$ and $h (\cdot) = \frac{\partial}{\partial \gamma} \log \psi_\gamma (\cdot)$ are possible directions approaching $\lambda (\cdot)$ and $\psi (\cdot)$ respectively from some index sets. The natural Hilbert spaces where those functions $a (\cdot)$ and $h (\cdot)$ sit in
are
\[ L_{2,V} = \{ a : E \left[ \Delta a^2 (V) \right] < \infty \}, \]
\[ L_{2,X_0}^0 = \{ h : E \left[ \Delta h (X_0) \right] = 0, E \left[ \Delta h^2 (X_0) \right] < \infty \}. \]

Hence the two tangent sets are
\[ A_\lambda = \left\{ \lambda a : a \in L_{2,V} \right\}, \]
\[ H_\psi = \left\{ \psi h : h \in L_{2,X_0}^0 \right\}. \]

Under this circumstance, there are two nonparametric components whose tangent sets are not orthogonal. We shall rely on the techniques in Sasieni (1992ab) to find the efficient score function for the regression parameter \( \beta_\alpha \).

**Theorem 5.2** The efficient score for estimating \( \beta \) in the single-index Cox model is
\[
\tilde{i}_\beta^* (Z) = \int \left[ \psi (X_0) X_2 - a^* (t) - h^* (X_0) \right] dM (t)
\]
where \( a^* (\cdot) \) and \( h^* (\cdot) \) are the unique functions minimizing
\[
E \Delta \left\| \psi (X_0) X_2 - a (V) - h (X_0) \right\|^2.
\]
Here they take the following forms with
\[
a^* (t) = E \left[ \psi (X_0) X_2 - h^* (X_0) | V = t, \Delta = 1 \right]
\]
and \( h^* (\cdot) \) satisfies
\[
E \left\{ \psi (X_0) X_2 - h^* (X_0) - E \left[ \psi (X_0) X_2 - h^* (X_0) | V = t, \Delta = 1 \right] | X_0 = x_0, \Delta = 1 \right\} = 0, \ a.s. \ w.r.t \ P_{X_0}.
\]

Moreover, the semiparametric information bound for estimating \( \beta_\alpha \) is
\[
I^* (\beta_\alpha) = E \left[ \tilde{i}_\beta (Z)^2 \right].
\]

**Proof.** Let \( \Pi \) denote the projection operator. The efficient score function is of the orthogonal projection of \( \tilde{i}_\beta \) to the closure of the nuisance sum-space \( A_\lambda + H_\psi \), hence we need to find the least favorable direction \( (a^*, h^*) \) such that \( \tilde{i}_\beta - \tilde{i}_\lambda a^* - \tilde{i}_\psi h^* \) is orthogonal to the sumspace \( A_\lambda + H_\psi \). Assumptions (A3)-(A5) together with Proposition 1 in Sasieni (1992b) guarantee this projection is well-defined. Furthermore, Proposition 3 in Sasieni (1992b) points out an explicit way to calculate this orthogonal projection as
\[
\Pi \left[ \tilde{i}_\beta | (A_\lambda + H_\psi)^\perp \right] = \Pi \left[ \tilde{i}_\beta | (A_\lambda)^\perp \right] \Pi \left[ H_\psi | (A_\lambda)^\perp \right].
\]
Hence, we shall first eliminate the hazard function by projecting the scores $l_\beta$ and $l_\psi$ onto the tangent space for the hazard then and projecting the residual of the projection of $l_\beta$ onto the subspace generated by the residual score of the projection of $l_\psi$. By Theorem 1 of Sasieni (1992b), the two residual scores are

$$K_\beta = \int D_X (X_0, t) dM (t),$$

and

$$Kh = \int Dh (X_0, t) dM (t),$$

where

$$D_X (x_0, t) = \hat{\psi} (x_0) x_2 - E \left[ \hat{\psi} (x_0) x_2 | V = t, \Delta = 1 \right],$$

and

$$Dh (x_0, t) = h (x_0) - E \left[ h (x_0) | V = t, \Delta = 1 \right].$$

Thus, the least favorable direction $h^*$ minimizing $E \| K_\beta - Kh \|^2$ satisfies

$$E \left[ (K_\beta - Kh^*) Kh \right] = 0 \quad (3.13)$$

for all $h \in L_{2,W}$. By Lemma 1 in Sasieni (1992b), (3.13) is equivalent as

$$D^* (D_X - Dh^*) = 0 \text{ a.s.}$$

where $D^*$ is the adjoint of operator $D$. Now equations (ii) and (iii) from Lemma 3 in Sasieni (1992b) give us

$$D^* D_X = E \left\{ \hat{\psi} (X_0) X_2 - E \left[ \hat{\psi} (X_0) X_2 | V = t, \Delta = 1 \right] | x_0 = x_0, \Delta = 1 \right\}$$

$$D^* Dh^* = E \left\{ h (X_0) - E \left[ h (X_0) | V = t, \Delta = 1 \right] | X_0 = x_0, \Delta = 1 \right\}$$

Whence it is straightforward to see the efficient direction $h^* (\cdot)$ satisfies the following condition:

$$E \left\{ \hat{\psi} (X_0) X_2 - h^* (X_0) - E \left[ \hat{\psi} (X_0) X_2 - h^* (X_0) | V = t, \Delta = 1 \right] | X_0 = x_0, \Delta = 1 \right\}$$

$$= 0, \text{ a.s. } P_{X_0}.$$  

Moreover, the least favorable direction of the corresponding hazard function is

$$a^* (t) = E \left[ \hat{\psi} (X_0) X_2 - h^* (X_0) | V = t, \Delta = 1 \right].$$

\[\text{■}\]

5.2 Large Sample Properties

The following notations are standard in deriving large sample properties in Cox regression, see Andersen and Gill (1982), Huang (1999):

$$S_{0n} (t, \theta) = \frac{1}{n} \sum_{i=1}^{n} Y_i (t) \exp \left[ \psi \left( X_i^\prime \beta \right) \right],$$
\[ S_0 (t, \theta) = E \left[ Y (t) \exp \left( \psi \left( X' \beta \right) \right) \right], \]
\[ S_{1n} (t, \theta) [h] = \frac{1}{n} \sum_{i=1}^{n} Y_i (t) \exp \left( \psi \left( X_i' \beta \right) \right) h (X_i), \]
\[ S_1 (t, \theta) [h] = E \left[ Y (t) \exp \left( \psi \left( X' \beta \right) \right) h (X) \right], \]
with a real-valued function \( h (\cdot) \) depending on \( x \). Furthermore, for \( u = (t, x, \delta) \) define
\[ s_n (u, \theta) [h] = h (x) - \frac{S_{1n} (t, \theta) [h]}{S_0 (t, \theta) [h]} \]
\[ s (u, \theta) [h] = h (x) - \frac{S_1 (t, \theta) [h]}{S_0 (t, \theta) [h]} \]
The sample criterion function is
\[ M_n (\theta) = \frac{1}{n} \sum_{i=1}^{n} I \left[ V_i \leq \tau \right] \Delta, \left\{ \psi \left( X_i' \beta \right) - \log S_{0n} (V_i, \theta) \right\}. \]
In the sequel, we shall omit writing out \( I \left[ V_i \leq \tau \right] \) without confusion. The population criterion function is
\[ M (\theta) = P_\Delta \left\{ \psi \left( X' \beta \right) - \log S_0 (V, \theta) \right\}. \]
Notice \( \dot{\zeta} (x, \beta) = \frac{\partial}{\partial \beta} \psi \left( x' \beta \right) = \psi \left( x' \beta \right) x. \)
The first result concerns the consistency and rates of convergence of parameter \( \theta = (\beta, \zeta (\cdot, \beta)) \) in terms of the following metric
\[ d (\theta_1, \theta_2) = |\beta_1 - \beta_2| + \| \zeta_1 (\cdot, \beta_1) - \zeta_2 (\cdot, \beta_2) \| \]
where
\[ \| \zeta_1 (\cdot, \beta_1) - \zeta_2 (\cdot, \beta_2) \| = \int \left[ \psi_1 \left( x' \beta_1 \right) - \psi_2 \left( x' \beta_2 \right) \right]^2 dF_X (x). \]
Let \( \Psi^p_n = S_n (T_{K_n, K_n, p}) \), and denote
\[ \mathcal{H}^p_n = \left\{ \zeta (\cdot, \beta) : \zeta (x, \beta) = \psi \left( x' \beta \right), \psi \in \Psi^p_n, x \in \mathcal{X}, \beta \in \mathcal{B} \right\} \]
and \( \Theta^p_n = \mathcal{B} \times \mathcal{H}^p_n \). The proof of Theorem 5.3 is relegated to the Appendix.

**Theorem 5.3** Let \( K_n = O (n^\nu) \) where \( \nu \) satisfies the restriction \( \frac{1}{4p-3} < \nu < \frac{1}{2p} \). Then given Assumptions A1-A6, we have (i) \( d \left( \hat{\theta}_n, \theta_0 \right) \rightarrow 0 \) and its rate of convergence is
\[ d \left( \hat{\theta}_n, \theta_0 \right) = O_p \left( n^{-\min (p, (1 - \nu)/2)} \right). \]
Notice for \( \nu = \frac{1}{2p+1} \) we get the optimal convergence rate as \( O_p \left( n^{-\frac{p}{2p+1}} \right) \).

**Theorem 5.4** Given Assumptions (A1)-(A6), we have
\[ \sqrt{n} \left( \hat{\beta}_n - \beta_0 \right) = I^* (\beta_0)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{i}_\beta^* (Z_i) + o_p (1) \]
\[ \Rightarrow N \left( 0, I^* (\beta_0)^{-1} \right), \]
with the efficient score \( \hat{i}_\beta^* \) and information \( I^* (\beta_0) \) in Theorem 5.2.
The proof of Theorem 5.4 is adapted from Theorem 2.1 in Ding and Nan (2011). For earlier results without bundled parameter issue, we refer the readers to Huang (1999) for a similar exposition.

**Proof.** We will prove the following claims separately in Appendix. First of all, two estimating equations hold:

\begin{equation}
\begin{aligned}
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) \left[ \hat{\zeta}_n (\cdot, \beta_o) \right] \right\} &= o_p \left( n^{-1/2} \right), \\
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) [h^*] \right\} &= o_p \left( n^{-1/2} \right),
\end{aligned}
\end{equation}

Therefore

\begin{equation}
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) [\bar{m}_o] \right\} = o_p \left( n^{-1/2} \right),
\end{equation}

for the least favorable direction \( \bar{m}_o = \zeta_o (\cdot, \beta_o) - h^* \). Moreover, the stochastic equicontinuity holds for the following processes

\begin{equation*}
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) \left[ \hat{\zeta}_n (\cdot, \beta_o) \right] - s_n \left( \theta_o \right) \left[ \hat{\zeta}_o (\cdot, \beta_o) \right] \right\} \\
- P_{\Delta} \left\{ s \left( \hat{\theta}_n \right) \left[ \hat{\zeta}_o (\cdot, \beta_o) \right] - s \left( \theta_o \right) \left[ \hat{\zeta}_o (\cdot, \beta_o) \right] \right\} = o_p \left( n^{-1/2} \right),
\end{equation*}

\begin{equation*}
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) [h^*] - s_n \left( \theta_o \right) [h^*] \right\} \\
- P_{\Delta} \left\{ s \left( \hat{\theta}_n \right) [h^*] - s \left( \theta_o \right) [h^*] \right\} = o_p \left( n^{-1/2} \right),
\end{equation*}

leading to

\begin{equation*}
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) [\bar{m}_o] - s_n \left( \theta_o \right) [\bar{m}_o] \right\} \\
- P_{\Delta} \left\{ s \left( \hat{\theta}_n \right) [\bar{m}_o] - s \left( \theta_o \right) [\bar{m}_o] \right\} = o_p \left( n^{-1/2} \right).
\end{equation*}

Considering (3.15), one arrives at

\begin{equation*}
- P_{\Delta} \left\{ s \left( \hat{\theta}_n \right) [\bar{m}_o] - s \left( \theta_o \right) [\bar{m}_o] \right\} = P_{\Delta n} \left\{ s_n \left( \theta_o \right) [\bar{m}_o] \right\} + o_p \left( n^{-1/2} \right).
\end{equation*}

Now we take the following expansion as in Lemma 5.4 of Huang (1999):

\begin{equation*}
P_{\Delta} \left\{ s \left( \hat{\theta}_n \right) [\bar{m}_o] - s \left( \theta_o \right) [\bar{m}_o] \right\} \\
= - P_{\Delta} \left\{ s \left( \theta_o \right) [\bar{m}_o] s \left( \theta_o \right) [\bar{m}_o] \right\} \left( \hat{\beta}_n - \beta_o \right) \\
- P_{\Delta} \left\{ s \left( \theta_o \right) [\bar{m}_o] s \left( \theta_o \right) \left[ \hat{\psi}_n - \psi_o \right] \right\} \\
= - P_{\Delta} \left\{ s \left( \theta_o \right) [\bar{m}_o] \right\} \hat{\beta}_n - \beta_o + o_p \left( n^{-1/2} \right),
\end{equation*}

where the final equality above follows from the fact that \( \bar{m}_o \) is the efficient direction. Thus combining all the results above, we arrive at

\begin{equation*}
\sqrt{n} P_{\Delta} \left\{ s \left( \theta_o \right) [\bar{m}_o] \right\} \hat{\beta}_n - \beta_o = \sqrt{n} P_{\Delta n} \left\{ s_n \left( \theta_o \right) [\bar{m}_o] \right\} + o_p \left( 1 \right).
\end{equation*}
Thus, we get
\[
\sqrt{n}P_{\Delta n}\{s_n (\cdot, \theta_o) | \hat{m}_o]\}
\]
\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T \left[ \psi \left( X_i' \beta \right) X_i - h^* \left( X_i' \beta \right) \right] dM_i (t)
\]
\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T \left[ \psi \left( X_i' \beta \right) X_i - h^* \left( X_i' \beta \right) \right] dM_i (t)
\]
\[
+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T \left[ S_1 (t, \theta_o) \left[ \tilde{c}_o (\cdot, \beta_o) - h^* \right] \right] dM_i (t)
\]
the second term is negligible by Lenglart’s inequality as in Anderson and Gill (1982):
\[
\frac{1}{n} \sum_{i=1}^{n} \int_0^T \left[ S_1 (t, \theta_o) \left[ \tilde{c}_o (\cdot, \beta_o) - h^* \right] \right] dM_i (t) + o_p (1)
\]
Hence
\[
\sqrt{n}P_{\Delta n}\{s_n (\cdot, \theta_o) | \hat{m}_o]\}
\]
\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^T \left[ \psi \left( X_i' \beta \right) X_i - h^* \left( X_i' \beta \right) \right] dM_i (t) + o_p (1)
\]
also notice that
\[
\frac{S_1 (t, \theta_o) \left[ \tilde{c}_o (\cdot, \beta_o) - h^* \right]}{S_0 (t, \theta_o)} = \alpha^* (t)
\]
In the end, the desired conclusion follows:
\[
\sqrt{n}P_{\Delta n}\{s_n (\cdot, \theta_o) | \hat{m}_o]\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I^*_\beta (V_i, X_i, \Delta_i) + o_p (1).
\]

5.3 Bootstrap Consistency

It is almost universal among semiparametric models that the construction of a valid confidence set of finite dimensional parameters requires plugging in additional nonparametric estimates, which complicates the inference problems in practice. Cheng and Huang (2010) prove the bootstrap consistency for a generic semiparametric M-estimator with exchangeable weights. They have treated the standard Cox regression as one demonstrating example, but made one simplifying assumption that \( s (\cdot, \theta) \) directly appears in the estimating equation instead of \( s_n (\cdot, \theta) \), pertaining to their \( i.i.d. \) assumption on the individual criterion function. In this section, we propose a novel bootstrap procedure consistent with the partial likelihood structure inspired by Nan and Wellner (2013). We establish the bootstrap consistency with general exchangeable bootstrap weights \( (W_{ni})_{i=1}^{n} \). Seeking for more general weighting scheme other than Efron’s multinomial weights is particularly important, as the later procedure often gives too many dies applied to censored data. Define the weighted empirical process as \( P^*_n f = \frac{1}{n} \sum_{i=1}^{n} W_{ni} f (X_i) \), we assume the following requirements on the bootstrap weight.
adapted from Cheng and Huang (2010). We have the product probability space underlying probability space and source of randomness. The following set of definitions and notations are conditional on observations $P_{in}$ with exchange weights independent from the sample observations, i.e., for the joint randomness from observed data and bootstrap weights. Furthermore, the bootstrap weights are

**Theorem 5.6**

Suppose Assumptions (A1)-(A6) and (W1)-(W5) hold. For the bootstrapped estimator $(\hat{\beta}_n^*, \hat{\gamma}_n^*)$ defined by maximizing

$$I_n^*(\beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} W_{ni} \Delta_i \left\{ \sum_{j=1}^{q} \gamma_j B_j \left( X_{i1} + X_{i2}^j \beta \right) - \log \left( \sum_{k:V_i \geq V_j} Q_{nk} \exp \left( \sum_{j=1}^{q} \gamma_j B_j \left( X_{ki} + X_{k2}^j \beta \right) \right) \right) \right\},$$

So the bootstrapped estimates $(\hat{\beta}_n^*, \hat{\gamma}_n^*)$ are defined by maximizing

$$I_n^*(\beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} W_{ni} \Delta_i \left\{ \sum_{j=1}^{q} \gamma_j B_j \left( X_{i1} + X_{i2}^j \beta \right) - \log \left( \sum_{k:V_i \geq V_j} Q_{nk} \exp \left( \sum_{j=1}^{q} \gamma_j B_j \left( X_{ki} + X_{k2}^j \beta \right) \right) \right) \right\}.$$

Before stating the theoretical results for the bootstrapped quantities, we should be explicit about the underlying probability space and source of randomness. The following set of definitions and notations are adapted from Cheng and Huang (2010). We have the product probability space

$$(\mathcal{Z}^\infty \times \mathcal{W}, \mathcal{A}^\infty \times \otimes, P_{ZW})$$

for the joint randomness from observed data and bootstrap weights. Furthermore, the bootstrap weights are independent from the sample observations, i.e., $P_{ZW} = P_Z \times P_W$.

**Definition 5.5** For a real-valued random variable $\Delta_n$, we define (i) $\Delta_n = o_P(1)$ in $P_Z$-probability if for any $\varepsilon, \delta > 0$

$$P_Z\{P_{W|Z}(|\Delta_n| > \varepsilon) > \delta\} \rightarrow 0.$$  

(ii) We define $\Delta_n = O_P(1)$ in $P_Z$-probability if for any $\delta > 0$ there exists a $M$ s.t.

$$P_Z\{P_{W|Z}(|\Delta_n| > M) > \delta\} \rightarrow 0.$$

**Theorem 5.6** Suppose Assumptions (A1)-(A6) and (W1)-(W5) hold. For the bootstrapped estimator $(\hat{\beta}_n^*, \hat{\gamma}_n^*)$ with exchange weights $W_n = (W_{n1}, \ldots, W_{nn})$, we have that

$$\|\hat{\beta}_n^* - \hat{\beta}_n\| = O_P\left(\frac{1}{n^{1/2}}\right)$$

in $P_Z$-probability. Furthermore,

$$\sup_{x \in \mathbb{R}^d} P_{W_n|Z} \left\{ \sqrt{n} \left( \hat{\beta}_n^* - \hat{\beta}_n \right) \leq x \right\} - P \left\{ N \left( 0, I^* (\beta_0)^{-1} \right) \leq x \right\} \rightarrow 0, \quad \text{conditional on observations } Z_n \text{ almost surely.}$$
A direct consequence of the last theorem is the validity of the bootstrap percentile type confidence interval, i.e.,

$$\lim_{n \to \infty} P_{ZW} (\beta_o \in BC(\alpha)) = 1 - \alpha.$$  

where the bootstrapped confidence set is defined to be

$$BC(\alpha) = [c_n^*(\alpha/2), c_n^*(1 - \alpha/2)],$$

with the $\alpha$-th quantile of bootstrap distribution $c_n^*(\alpha) = \inf \{\epsilon: P_{W_n}|Z_n \left( (\hat{\beta}_n^* - \beta_n) \leq \epsilon \right) \geq \alpha \},$ which could be easily simulated\(^4\).

6 Monte Carlo Results

As for the mean square error of the sieve maximum partial likelihood estimation in finite sample, we refer readers to the simulation results in Section 4 of Huang and Liu (2006). In this section we examine the finite sample coverage properties of our bootstrap percentile confidence intervals. We consider two models generated by a gamma process ($v = 1/2$ and $p = 1$) and a stable process ($v = 1/2$ and $p = 1$) respectively. We would also examine a single index model with link function $\psi(\cdot) = \sin(\cdot)$ as in the simulation study of Huang and Liu (2006), even though this is not driven by any Lévy subordinator. Referring to their reduced-form conditional hazard functions, we have

Model 1: $\Lambda_T(t|X) = \Lambda(t) \log \left(1 + \frac{\exp(X_1 + 2X_2 - 0.5X_3)}{2}\right),$  
Model 2: $\Lambda_T(t|X) = \Lambda(t) \frac{\exp(X_1 + 2X_2 - 0.5X_3)}{2},$  
Model 3: $\Lambda_T(t|X) = \Lambda(t) \exp[\sin(X_1 + 2X_2 - 0.5X_3)].$

In all three models, $\tilde{\beta}_o = (\beta_1, \beta_2, \beta_3) = (1, 2, -0.5)$ consistent with the scale normalization we choose earlier. We generate $X_2, X_3$ following independent normal distribution whose marginal distributions have mean equal to 1 and variance equal to 4 as in Huang and Liu (2006), and let $X_1 \sim U[-1, 1].$ The censoring variable $C$ is simulated independently from the rest of the data, following the unit exponential distribution. We also set the baseline cumulative hazard function as the identity map as in Huang and Liu (2006), i.e., $\Lambda(t) = t.$

Regarding a few tuning parameters we have to pick in the simulation, we choose a third order squally spaced B-spline function with number of knots equal to $2n^{1/7},$ which is certainly ad-hoc but consistent with the optimal rate of a third order differentiable function. Furthermore, we let the bootstrap replication number to be 200 and we adopt the Bayesian bootstrap (see Cheng and Huang, 2010) in which case all data points would be assigned with positive weights. Since the partial likelihood function needs to be maximized for each generated bootstrap weight, we only report the size properties via 1,000 simulations, tabulated below for $(\beta_2, \beta_3).$

\(^4\)The hybrid type bootstrap confidence interval in Cheng and Huang (2010) is valid as well. We only investigate the finite performance of percentile type version in simulation for space restrictions.
<table>
<thead>
<tr>
<th>Sample Size n</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 500</td>
<td>0.929</td>
<td>0.966</td>
<td>0.967</td>
</tr>
<tr>
<td>n = 1000</td>
<td>0.944</td>
<td>0.952</td>
<td>0.959</td>
</tr>
</tbody>
</table>

Table 1: Size Properties of Bootstrapped CIs

The above table confirms the nice coverage properties of our bootstrapped confidence intervals for a moderate sample size, hence we recommend the bootstrap inference for future empirical applications.

7 Conclusion

In this paper, we present a systematic study of duration models characterized by threshold-crossing behavior of latent Lévy subordinators. With its distributional property and nondecreasing sample path, the Lévy subordinator could be seen as the natural variant of the static heterogeneity terms in MPH from a process point of view, and it resembles the notion of usage and wear-out effect (Singpurwalla, 1995; Aalen, Borgan, and Gjessing, 2008) closely. Hence, besides those optimal stopping time problems in economics, our model also applies to the scenario concerning the life/failure time in biology, medical science, and engineering, where the traditional hazard-based models dominate. Furthermore, it is transparent how the structural parameters of theoretical models can be related to reduced-form parameters, and all model primitives could be point identified under semiparametric specifications. In spite of the sophisticated construction, our models exhibit very tractable mathematical structures, inducing flexible semiparametric estimation procedures with rigorously established large sample properties. In a companion paper, Liu (2015) considers a competing risks model where the dependence is generated via coupling two Lévy subordinators, and obtains semiparametric identification and estimation.

Numerous extensions are possible. First, one could incorporate endogenous and time-varying covariates by modeling additional stochastic processes explicitly. Renault, van den Heijden, and Werker (2014) present a structural model for transaction time (duration) and asset prices (associated marks) driven by multivariate Brownian motion. In their model, successive passage times of one latent Gaussian component relative to random boundaries define durations and the other correlated components generate the marks. It is both interesting and challenging to search for other processes which lead to tractable structure. Second, it is desirable to relax the parametric assumption on the threshold variable and to consider a more general Lévy process, i.e., the spectral positive Lévy process (Sato, 1999). Last but not least, it is worthwhile to study the empirical content of stochastic game models driven by a Lévy process, see Chapter 11 in Kyprianou (2006).

8 Appendix

We will first prove the identification result. After introducing several lemmas instrumental to the main results, we prove the claimed consistency and rates of convergence. We let $c$ denote a generic finite positive
constant, whose value may change from line to line. We shall introduce additional subscripts if more than one appear in the display. Moreover, $A \preceq B$ means there exists a positive constant $C$ which does not depend on $n$ s.t. $A \leq CB$.

The following lemma is a modification of Theorem 1 in Lin and Kulasekera (2007) with the additional univariate $X_1$. Recall we have partitioned $X$ as $X = \left( X_1, X_2 \right)'$.

**Lemma A.1** The support $\mathcal{X}$ of $m(\cdot)$ is a bounded convex set with at least one interior point, and $m(\cdot)$ is a nonconstant continuous function on $\mathcal{X}$. If

$$m(x) = \psi \left( x' \beta \right) = \varphi \left( x' \alpha \right),$$

for all $x \in \mathcal{X}$ for some continuous functions $\psi$ and $\varphi$, then $\beta = \alpha$ and $\psi \equiv \varphi$ on $\left\{ x_1 + x_2 \beta | x \in \mathcal{X} \right\}$.

**Proof.** It suffices to show that $\beta = \alpha$. Recall that $\tilde{\beta} = \left( 1, \beta' \right)'$, $\tilde{\alpha} = \left( 1, \alpha' \right)'$ and the following two composition functions:

$$\tilde{\psi} \left( \cdot \right) = \psi \left( \| \tilde{\beta} \| \times \cdot \right), \text{ and } \tilde{\varphi} \left( \cdot \right) = \varphi \left( \| \tilde{\beta} \| \times \cdot \right),$$

where $\| \cdot \|$ is the standard Euclidean distance. If $\psi \left( x' \tilde{\beta} \right) = \varphi \left( x' \tilde{\alpha} \right)$, one gets

$$\tilde{\psi} \left( x' \tilde{\beta} \right) = \tilde{\varphi} \left( x' \tilde{\alpha} \right), \text{ for all } x \in \mathcal{X};$$

with $\tilde{\beta} = \beta / \| \tilde{\beta} \|$ and $\tilde{\alpha} = \alpha / \| \tilde{\alpha} \|$. Now we are ready to apply the arguments in Lin and Kulasekera (2007), as both $\tilde{\beta}$ and $\tilde{\alpha}$ have norm equal to 1 and their first coordinates are positive.

Suppose $\beta \neq \alpha$, then obviously $\tilde{\beta} \neq \tilde{\alpha}$ holds as well. By the stated assumptions, there exists a sphere $B(\bar{x}, r) \subset \mathcal{X}$ for some $\bar{x}$ s.t. $m(\cdot)$ is nonconstant on it. For all $t \in (-r, r)$ we have $\bar{x} + t \tilde{\beta} \in \mathcal{X}$ as $\tilde{\beta} \tilde{\beta} = 1$. So

$$\tilde{\psi} \left( x' \beta + t \right) = \tilde{\psi} \left( \left( \beta \left( \bar{x} + t \beta \right) \right) \right) = \tilde{\varphi} \left( \left( \alpha' \left( \bar{x} + t \alpha \right) \right) \right) \text{.}$$

(A.1)

Because the first non-zero components of $\tilde{\beta}$ and $\tilde{\alpha}$ are positive, we get $\tilde{\beta} \neq -\tilde{\alpha}$, hence $|\tilde{\beta} \tilde{\alpha}| < 1$. We could carry out the following iteration based on (A.1):

$$\tilde{\psi} \left( x' \beta + t \right) = \tilde{\varphi} \left( \left( \alpha' \left( \bar{x} + t \beta \right) \right) \right) = \tilde{\psi} \left( \left( \beta' \left( \bar{x} + t \beta \right) \right) ^2 \right) \text{.}$$

where we have used the continuity of $\tilde{\psi}$. This indicates $m(\cdot)$ is a constant function on $B(\bar{x}, r)$, leading to a contradiction. Thereafter we get $\tilde{\beta} = \tilde{\alpha}$, $\| \tilde{\beta} \| = \| \tilde{\alpha} \|$, resulting in $\beta = \alpha$ as desired.

**Proof.** (of Identification) The moment restriction in (B.3) allows us to identify $\Lambda(t)$ from the conditional cumulative hazard function $\Lambda_T(t|x)$ first, because $\log \Lambda(t) = E \left[ \log \Lambda_T(t|X) \right]$. For the single index structure, we shall apply the previous lemma to $m(x) = \psi \left( x' \beta \right)$, where $\psi(\cdot) = \log \Phi \circ \exp(\cdot)$ with the Lévy-Laplace

22
exponent function $\Phi$. $\Phi$ belongs to the class of Bernstein function whose derivative is completely monotone (see Schilling, Song and Vondracek, 2012), thus $\psi(\cdot)$ is real analytic. Given Assumption (I1), we get the identification of $\beta$ as $\psi(\cdot)$ is indeed a nonconstant continuous function. Finally, the variation of $X_0$ along an open interval leads to identification of the analytic function $\psi(\cdot)$ by Proposition 1 in Abbring and van den Berg (2003).

Now we present several preparatory lemmas. The first two essentially concern the smoothness of the model, and would be used in determining the rate of convergence.

**Lemma A.2** Under Assumptions (A1)-(A6), for a small $\varepsilon > 0$ we have

$$\sup_{d(\theta, \theta_0) \leq \varepsilon, \theta \in \Theta_e} \text{Var}[m(\cdot, \theta) - m(\cdot, \theta_0)] \leq C\varepsilon^2.$$  

**Proof.** By definition of the individual criterion function, we have

$$E[m(\cdot, \theta) - m(\cdot, \theta_0)]^2$$

$$= E \left[ \psi \left( X'_i \overline{\beta} \right) - \psi_0 \left( X'_i \overline{\beta}_0 \right) - \left[ \log S_0 (V_i, \theta) - \log S_0 (V_i, \theta_0) \right] \right]^2$$

$$\leq E \left[ \psi \left( X'_i \overline{\beta} \right) - \psi_0 \left( X'_i \overline{\beta}_0 \right) \right]^2 + E \left[ S_0 (V_i, \theta) - S_0 (V_i, \theta_0) \right]^2$$

$$\leq |\beta_0 - \beta|^2 + ||\zeta_0 (\cdot, \beta_0) - \zeta (\cdot, \beta)||^2.$$

**Lemma A.3** Under Assumptions (A1)-(A6), for a small $\varepsilon > 0$ we have

$$\inf_{d(\theta, \theta_0) \geq \varepsilon, \theta \in \Theta_e} [M(\theta_0) - M(\theta)] \geq C\varepsilon^2.$$  

**Proof.** Here we write $M(\theta) = M(\beta, \zeta (\cdot, \beta))$ to highlight the nature of the bundled parameter case. For any small $\varepsilon > 0$, and the point $\theta$ s.t. $d(\theta, \theta_0) \geq \varepsilon$, we consider the quadratic expansion for

$$M(\beta_0, \zeta_0 (\cdot, \beta_0)) - M(\beta, \zeta (\cdot, \beta))$$

$$= (\beta_0 - \beta)' M_{\beta \beta} (\beta_0 - \beta) + 2 (\beta_0 - \beta)' M_{\beta \psi} [\psi_0 - \psi]$$

$$+ M_{\psi \psi} [\psi_0 - \psi, \psi_0 - \psi] + o \left( d^2 (\theta, \theta_0) \right),$$

with those second order derivatives calculated following the rules in Ding and Nan (2011). A straightforward computation reveals that

$$M_{\beta \beta} = P \Delta \left\{ \left[ \psi^2 \left( X'_i \overline{\beta} \right) - \frac{P \left[ Y(t) e^\psi \left( X'_i \overline{\beta} \right) \left( \frac{\psi^2 \left( X'_i \overline{\beta} \right) + \psi \left( X'_i \overline{\beta} \right)}{P \left[ Y(t) e^\psi \left( X'_i \overline{\beta} \right) \right]} \right] \right] X_2^{\otimes 2} \right\}$$

$$+ P \Delta \left\{ \left[ \frac{P \left[ Y(t) e^\psi \left( X'_i \overline{\beta} \right) \psi \left( X'_i \overline{\beta} \right) X \right]}{P \left[ Y(t) e^\psi \left( X'_i \overline{\beta} \right) \right]} \right] \right\}^{\otimes 2}$$

$$= P \Delta \left\{ E \left[ K(t) \psi^2 \left( X'_i \overline{\beta} \right) X_2^{\otimes 2} \right] - E \left[ K(t) \left( \psi \left( X'_i \overline{\beta} \right) X_2 \right) \right] \right\}^{\otimes 2},$$
where we have used the definition of $K(t)$ in (3.12) and the fact (3.11), which leads to

\[
P_\Delta \left\{ \psi^2 \left( X' \right) - \frac{P \left[ Y(t) e^{\psi(X')} \psi^2 \left( X' \right) \right]}{P \left[ Y(t) e^{\psi(X')} \right]} \right\} X^\otimes 2 = 0.
\]

Similar calculations give us the other two derivatives:

\[
M[h] = P_n E h K(t) h X_0 e X_0 e X_2 i + E h K(t) h X_0 e X_2 i E h K(t) h X_0 e X_2 i,
\]

\[
M[h_1, h_2] = P_n E h_1 K(t) h_2 h X_0 e X_0 e h_1 K(t) h_2 h X_0 e X_2 i + E h_1 K(t) h_2 h X_0 e X_2 i E h_1 K(t) h_2 h X_0 e X_2 i,
\]

also see Lemma A.4 in Huang (1999).

After completing squares, one arrives at

\[
M(\beta_0, \zeta_0 (\cdot, \beta_0)) - M(\beta, \zeta (\cdot, \beta)) = P_\Delta \left\{ E K(t) [g^*(X) - E K(t) g^*(X)]^2 \right\}
\]

where

\[
g^*(X) = \psi \left( X' \right) X' (\beta_0 - \beta) + \psi_0 \left( X' \beta_0 \right) - \psi \left( X' \beta_0 \right).
\]

Now by the positive definiteness of $X^\otimes 2$ and

\[
0 < c_1 \leq E [K(t) |X|] \leq c_2 < \infty,
\]

we obtain

\[
\inf_{d(\theta, \theta_0) \geq \varepsilon} [M(\beta_0, \zeta_0 (\cdot, \beta_0)) - M(\beta, \zeta (\cdot, \beta))] \geq c d^2 \theta, \theta_0).
\]

The following lemma characterizes the order of approximation bias in sieve estimation, see Schumaker (1981). The additional term $n^{-(1-\nu)/2}$ comes from recentering the estimator, see Lemma A.5 in Huang (1999).

**Lemma A.4** For any $\psi_o \in \Psi^p$, there exists a function $\psi_{on} \in \Psi^p_n$ with $P_\Delta n \psi_{on} = 0$ s.t.

\[
\|\psi_{on} - \psi_o\|_{\infty} = O \left( n^{-p \nu} + n^{-(1-\nu)/2} \right).
\]

The next lemma is Lemma 4.2 in Kong and Nan (2014), and it (via Borel-Cantelli argument) ensures that the frequently appearing denominator term $S_{0n}(t, \theta)$ is uniformly (for $t \in [0, \tau]$ and $\theta \in \Theta^n_\tau$) bounded away from zero almost surely.

**Lemma A.5** For the truncation point $\tau$, and any $\eta > 0$, we have

\[
\Pr \left( \frac{1}{n} \sum_{i=1}^{n} 1 \left( V_i \geq \tau \right) \leq \frac{\eta}{2} \right) \leq 2 e^{-n \eta^2/2}.
\]
Now we introduce some standard terminology in empirical process theory (van der Vaart and Wellner, 1996). $\| \cdot \|_{\infty}$ is the usual $L_{\infty}$-norm for a function $f$ with $\| f \|_{\infty} < \infty$. The bracketing number $N_{\parallel}(\epsilon, \mathcal{F}, \| \cdot \|_{\infty})$ for subclass $\mathcal{F}$ is defined to be the minimum of $m$ such that $\exists f_1, f_1', \ldots, f_m, f_m'$ for $\forall f \in \mathcal{F}$, $f_i \leq f \leq f_i'$ for some $i$, and $\| f_i' - f_i \|_{\infty} \leq \epsilon$. What’s more the corresponding bracketing entropy integral is $J_{\parallel}(\epsilon, \mathcal{F}, \| \cdot \|_{\infty}) = \int_0^\epsilon \sqrt{1 + \log N_{\parallel}(\epsilon, \mathcal{F}, \| \cdot \|_{\infty})}\,d\epsilon$. The following lemma which is a restatement of Lemma 3.4.2 in van der Vaart and Wellner (1996) based on the supremum norm is useful to bound the normalized empirical process $G_n = \sqrt{n}(P_n - P)$ and $\| G_n \|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \| G_n(f) \|$. Let $\mathcal{F}$ be a uniformly bounded class of measurable functions such that $P f^2 \leq \delta^2$ and $\| f \|_{\infty} \leq M_0$, then

$$E_{\parallel} \| G_n \|_{\mathcal{F}} \lesssim J_{\parallel}(\delta, \mathcal{F}, \| \cdot \|_{\infty}) \left(1 + \frac{J_{\parallel}(\delta, \mathcal{F}, \| \cdot \|_{\infty}) M_0}{\delta^2 \sqrt{n}}\right).$$

(A.3)

The next lemma collects various entropy bounds for several functional classes involved in the current paper.

**Lemma A.6** Referring to the functional classes under consideration:

\begin{align*}
\mathcal{F}_1 & = \left\{ f_1 = \psi \left( x' \beta \right) - \log S_0 (t, \theta) : \theta \in \Theta_n^p, d(\theta, \theta_o) \leq \delta \right\}, \\
\mathcal{F}_2 & = \left\{ f_2 = \exp \left[ \psi \left( x' \beta \right) \right] I [t \leq \tau] : \theta \in \Theta_n^p, d(\theta, \theta_o) \leq \delta \right\}, \\
\mathcal{F}_3 & = \left\{ f_3 = \exp \left[ \psi \left( x' \beta \right) \right] h \left( x' \beta \right) I [t \leq \tau] : \theta \in \Theta_n^p, d(\theta, \theta_o) \leq \delta, \| h \|_2 \leq \delta \right\}, \\
\mathcal{F}_4 & = \left\{ f_4 = \exp \left[ \psi \left( x' \beta \right) \right] \psi \left( x' \beta \right) x I [t \leq \tau] : \theta \in \Theta_n^p, d(\theta, \theta_o) \leq \delta, \| \psi \|_2 \leq \delta \right\},
\end{align*}

we get

$$N_{\parallel}(\epsilon, \mathcal{F}, \| \cdot \|_{\infty}) \lesssim (\epsilon / \delta)^{c_{\mathcal{F}}} + d,$$

or in terms of the entropy integral:

$$J_{\parallel}(\delta, \mathcal{F}, \| \cdot \|_{\infty}) \lesssim \sqrt{\delta} + \delta \sqrt{\log (1/\delta)},$$

for the dimension of covariate $d$ and some finite positive constant $c$ which may change with $j = 1, ..., 4$.

**Proof.** Due to similarity, we will only prove the assertion for the class $\mathcal{F}_4$ which is the most involved one. The construction here is analogous to Ding and Nan (2011). According to Schumaker (1981), for $\psi \in \Psi_n$ with $\| \psi - \psi_o \|_{\infty} \leq \delta$ and $\tilde{\psi} \in \Psi_n^{p-1}$ with $\| \tilde{\psi} - \dot{\psi}_o \|_{\infty} \leq \delta$ we get pairs of bracket $\left[ \psi_i^L, \psi_i^U \right]$ where $i = 1, ..., (\epsilon / \delta)^{c_{\mathcal{F}}}$ and $\left[ \tilde{\psi}_j^L, \tilde{\psi}_j^U \right]$ where $j = 1, ..., (\epsilon / \delta)^{c_{\mathcal{F}}}$. Since $\mathcal{B}$ is a compact subset of $\mathcal{R}^d$, it can be covered by $C \epsilon^{-d}$ balls with radius $\epsilon$. Given the boundedness of the support $\mathcal{X}$, for any $\beta \in \mathcal{B}$, there exists a $\beta_l$ satisfying $x' \beta \in \left[ x' \beta_l - c_0 \epsilon, x' \beta_l + c_0 \epsilon \right]$, for another constant $c_0$. Apparently the number of $l$ needed is of order $(\epsilon / \delta)^d$.

Now let $c_1^L, c_2^L$ be two constants that $\psi_i^L \left( x' \beta_l + c_1^L \epsilon \right)$ and $\psi_i^U \left( x' \beta_l + c_2^U \epsilon \right)$ are minimum and maximum values of $\psi_i^L$ and $\psi_i^U$ respectively in the interval $\left[ x' \beta_l - c_0 \epsilon, x' \beta_l + c_0 \epsilon \right]$. Similarly we define $c_3^L, c_4^U$ which make $\tilde{\psi}_j^L \left( x' \beta_l + c_3^L \epsilon \right)$ and $\tilde{\psi}_j^U \left( x' \beta_l + c_4^U \epsilon \right)$ to attain the corresponding minimum and maximum values in the same interval. Consider the set of brackets as:

$$\left\{ \left[ f_{ijl}, f_{ijl}' \right] : i = 1, ..., (\epsilon / \delta)^{c_{\mathcal{F}}}; j = 1, ..., (\epsilon / \delta)^{c_{\mathcal{F}}}, l = 1, ..., d (\epsilon / \delta)^d \right\}$$
where
\[
 f'_{ijt} = \exp \left[ \psi_i' \left( x' \beta_s + c_i \epsilon \right) \right] \psi_j' \left( x' \beta_t + c_j \epsilon \right) \, xI \, [t \leq \tau],
\]
\[
 f''_{ijt} = \exp \left[ \psi_i'' \left( x' \beta_s + c_i \epsilon \right) \right] \psi_j'' \left( x' \beta_t + c_j \epsilon \right) \, xI \, [t \leq \tau].
\]

By construction, the lower and upper bound functions are ordered and serve as a valid bracketing pair for the function \( f \in \mathcal{F}_4 \), and the overall number needed is of order \((\delta/\epsilon)^q n^{d+1}\) as claimed. Furthermore, it holds
\[
 \| f'_{ijt} - f''_{ijt} \| \leq C \left( \| \psi' - \psi'' \|_\infty + \| \psi'' - \psi'' \|_\infty + \| \beta_i - \beta_s \| \right) \lesssim \epsilon,
\]
utilizing the boundedness of \( \mathcal{X} \). ■

**Lemma A.7** Given the entropy bounds in Lemma A.6, we get

(i) for \( \Theta_\delta = \{ \theta : \| \psi - \psi_n \| \leq \delta, \| \psi \|_\infty \leq c \} \)

\[
 \sup_{t, \theta \in \Theta_\delta} \left| \frac{S_0(t, \theta)}{S_0(t, \theta)_{on}} - \frac{S_0(t, \theta)_{on}}{S_0(t, \theta)_{on}} \right| = n^{-1/2} O_p \left( \sqrt{q_n \delta + \delta \sqrt{\log(1/\delta)}} \right), \tag{A.4}
\]

(ii) for any \( g \) such that \( \| g \|_2 = o \left( q_n^{-1/2} \right) \)

\[
 \sup_t \left| \frac{S_{1t}(t, \theta_{on}) [g]}{S_{0t}(t, \theta_{on})} - \frac{S_{1t}(t, \theta_{on}) [g]}{S_{0t}(t, \theta_{on})} \right| = o_p \left( n^{-1/2} \right), \tag{A.5}
\]

(iii) for any \( \zeta (\cdot, \cdot) \in \mathcal{H}_n \) and \( g = \zeta (\cdot, \cdot) \) or \( g = h^* \).

\[
 \sup_t \left| \frac{S_{1t}(t, \theta_{on}) [g]}{S_{0t}(t, \theta_{on})} - \frac{S_{1t}(t, \theta_{on}) [g]}{S_{0t}(t, \theta_{on})} + \frac{S_{1t}(t, \theta_{on}) [g]}{S_{0t}(t, \theta_{on})} \right| = o_p \left( n^{-1/2} \right). \tag{A.6}
\]

**Proof.** Since
\[
 \frac{S_0(t, \theta)}{S_0(t, \theta)_{on}} - \frac{S_0(t, \theta)_{on}}{S_0(t, \theta)_{on}} = \frac{S_0(t, \theta)_{on}}{S_0(t, \theta)_{on}} \frac{S_0(t, \theta) - S_0(t, \theta)_{on}}{S_0(t, \theta) - S_0(t, \theta)_{on}},
\]
and the denominator is bounded away from zero with probability 1 by Lemma A.5, it is sufficient to consider the numerator on the r.h.s. of the equality above. This numerator could be expressed as
\[
 S_0(t, \theta) [S_0(t, \theta) - S_0(t, \theta)_{on} - S_0(t, \theta)_{on} + S_0(t, \theta)_{on}] \]
\[
 - \left[ S_0(t, \theta) - S_0(t, \theta_{on}) \right] [S_0(t, \theta) - S_0(t, \theta_{on})].
\]

It is clear that
\[
 S_0(t, \theta) - S_0(t, \theta_{on}) = S_0(t, \theta) + S_0(t, \theta_{on})
\]
\[
 = (P_n - P) \left\{ Y(t) \left[ \exp \left( \psi \left( x' \beta \right) \right) - \exp \psi_{on} \left( x' \beta \right) \right] \right\}
\]
\[
 = n^{-1/2} O_p \left( \sqrt{q_n \delta + \delta \sqrt{\log(1/\delta)}} \right).
\]

Moreover, \( S_0(t, \theta_{on}) - S_0(t, \theta_{on}) = n^{-1/2} \sqrt{q_n} \) and it holds
\[
 |S_0(t, \theta) - S_0(t, \theta_{on})| \leq E \left( t \left| \exp \left( \psi \left( x' \beta \right) \right) - \exp \psi_{on} \left( x' \beta \right) \right| \right)
\]
\[
 \lesssim \| \psi - \psi_n \|.
\]
Therefore, we get

\[ |S_{0n} (t, \theta_{on}) S_0 (t, \theta) - S_{0n} (t, \theta) S_0 (t, \theta_{on})| = n^{-1/2} O_p \left( \sqrt{q_n} \delta + \delta \sqrt{\log (1/\delta)} \right). \]

For the second conclusion,

\[
\frac{S_{1n} (t, \theta_{on}) [g]}{S_{0n} (t, \theta_{on})} - \frac{S_1 (t, \theta_{on}) [g]}{S_0 (t, \theta_{on})} = \frac{S_0 (t, \theta_{on}) \{ S_{1n} (t, \theta_{on}) [g] - S_1 (t, \theta_{on}) [g] \} - S_1 (t, \theta_{on}) [g] \{ S_{0n} (t, \theta_{on}) [g] - S_0 (t, \theta_{on}) [g] \}}{S_{0n} (t, \theta_{on}) S_0 (t, \theta_{on})}.
\]

Again, the conclusion follows from the fact that those two terms in the curly brackets are of \( o_p (n^{-1/2}) \) and the denominator is bounded away from zero with probability 1.

After carrying out the straightforward algebra as in Lemma A.7 of Huang (1999), the third conclusion follows from Donsker properties of \( S_{1n} (t, \theta) - S_1 (t, \theta) \) and \( S_{0n} (t, \theta) - S_0 (t, \theta) \).

**Proof. (of Consistency)** We shall prove the consistency result that \( d (\theta, \theta_o) \rightarrow_p 0 \), by verifying those conditions leading to Theorem 3.1 in Chen (2007). Condition 3.2 and 3.4 regarding the approximation and compactness properties are automatically satisfied by the B-spline sieves we use.

Condition 3.1 (ii) is ensured by our Lemma A.3 while letting \( g (\varepsilon) = \varepsilon^2 \) and \( \delta (k) = C \) in Chen (2007)

\[
\inf_{d (\theta, \theta_o) \geq \varepsilon, \theta \in \Theta_n^p} [M (\theta) - M (\theta)] \geq C \varepsilon^2.
\]

Condition 3.3 regarding the upper semicontinuity property is satisfied because for small enough \( \varepsilon > 0 \) and any \( \theta \) s.t. \( d (\theta, \theta_o) \leq \varepsilon \) we have

\[ |M (\theta) - M (\theta)| \leq C \varepsilon^2; \]

just by the upper bound of \( E [K (t) \mid X] \) in our proof of Lemma A.3.

Finally, because the functional class \( \mathcal{F}_1 \) has a finite bracketing number for any \( \delta \) as shown in Lemma A.6, it satisfies the uniform law of large number as in Theorem 2.4.1 in van der Vaart and Wellner (1996). Similarly, we have the uniform convergence of \( |S_{0n} (t, \theta) - S_0 (t, \theta)| \) to zero when \( \theta \) varies over \( \Theta_n^p \). Thus, Condition 3.5 in Chen (2007) is checked by the uniform convergence of \( M_n (\theta) \) to \( M (\theta) \).

**Proof. (of Convergence Rate)** Given the consistency result, we shall apply Theorem 3.2.5 in van der Vaart and Wellner (1996) to deduce the rate of convergence. We shall prove that

\[
E \sup_{d (\theta, \theta_o) \leq \delta} |M_n (\theta) - M_n (\theta_o) - (M (\theta) - M (\theta_o))| = n^{-1/2} \delta \left( q_n^{1/2} + \log^{1/2} \left( \frac{1}{\delta} \right) \right),
\]

and

\[
\inf_{d (\theta, \theta_o) \geq \varepsilon} [M (\theta) - M (\theta_o)] \geq c \delta^2 (\theta, \theta_o).
\]

The later separation condition has already be verified in our Lemma A.3. So the convergence rate \( r_n \) satisfies

\[
r_n^2 \left( r_n^{-1} q_n^{1/2} + r_n^{-1} \log^{1/2} (r_n) \right) = O \left( n^{1/2} \right),
\]

27
Hence it remains to bound the term
\[
M_n (\theta) - M_n (\theta_n) - (M (\theta) - M (\theta_n))
\]
\[
= (P_{\Delta n} - P_{\Delta}) [m (\cdot, \theta) - m (\cdot, \theta_n)] - P_{\Delta n} \left[ \log \frac{S_{0n} (\cdot, \theta)}{S_{0n} (\cdot, \theta_n)} - \log \frac{S_0 (\cdot, \theta)}{S_0 (\cdot, \theta_n)} \right]
\]
\[
\equiv L_{1n} (\theta) + L_{2n} (\theta).
\]
For the first term \(L_{1n} (\theta)\), we get
\[
E \sup_{d(\theta, \theta_n) \leq \delta} |L_{1n} (\theta)| \lesssim n^{-1/2} q_n^{1/2} \delta,
\]
applying our Lemma A.6 to the functional class \(\mathcal{F}_1\). While for the second term, we have
\[
\sup_{d(\theta, \theta_n) \leq \delta} |L_{2n} (\theta)| \leq 2 \sup_{d(\theta, \theta_n) \leq \delta, t \in [0, T]} \left| \log \frac{S_{0n} (\cdot, \theta)}{S_{0n} (\cdot, \theta_n)} - \log \frac{S_0 (\cdot, \theta)}{S_0 (\cdot, \theta_n)} \right|
\]
\[
\lesssim \sup_{d(\theta, \theta_n) \leq \delta, t \in [0, T]} \left| \frac{S_{0n} (\cdot, \theta) - S_0 (\cdot, \theta)}{S_{0n} (\cdot, \theta_n) - S_0 (\cdot, \theta_n)} \right|
\]
\[
\lesssim n^{-1/2} \delta \left( q_n^{1/2} + \log^{1/2} (1/\delta) \right),
\]
where we have made use of (A.4) Lemma A.7. Moreover for the approximation error, it follows upon (A.2)
\[
|M (\theta_o) - M (\theta)| \lesssim \|\psi_{on} - \psi_o\|_2 = O \left( n^{-p_o} \right),
\]
where \(\zeta_{on} (x, \beta_o) = \psi_{on} (x', \beta_o)\). In sum, the rate of convergence is \(O_p \left( \max \{ n^{-(1-p)/2}, n^{-p_o}, n^{-p_o} \} \right) = O_p \left( n^{-\min(p_o, (1-\nu)/2)} \right) \) as claimed. □

**Lemma A.8** Under Assumptions (A1)-(A6), we have the following two estimating equations hold approximately:
\[
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) [g] \right\} = o_p \left( n^{-1/2} \right),
\]
with \(g = \hat{\zeta}_o (\cdot, \beta_o)\) or \(g = h^*\).

**Proof.** We only prove the claim for \(\hat{\zeta}_o (\cdot, \beta_o)\), as the second one follows from Lemma 5.2 of Huang (1999) upon notational change. By the first order condition for the regression parameter, we already get
\[
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) \left[ \hat{\zeta}_n (\cdot, \beta_n) \right] \right\} = o_p \left( n^{-1/2} \right).
\]
Hence it remains to bound the term
\[
P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) \left[ \hat{\zeta}_n (\cdot, \beta_n) - \hat{\zeta}_o (\cdot, \beta_o) \right] \right\}
\]
\[
\equiv I_{1n} + I_{2n} + I_{3n},
\]
where
\[
I_{1n} = (P_{\Delta n} - P_{\Delta}) \left\{ \hat{\zeta}_n (\cdot, \beta_n) - \hat{\zeta}_o (\cdot, \beta_o) - \frac{S_1 \left( t, \hat{\theta}_n \right) \left[ \hat{\zeta}_n (\cdot, \beta_n) - \hat{\zeta}_o (\cdot, \beta_o) \right]}{S_0 \left( t, \hat{\theta}_n \right)} \right\},
\]
\[
I_{2n} = \left( P_{\Delta n} - P_{\Delta} \right) \left\{ \hat{\zeta}_n (\cdot, \beta_n) - \hat{\zeta}_o (\cdot, \beta_o) \right\}
\]
\[
I_{3n} = \left( P_{\Delta n} - P_{\Delta} \right) \left\{ \hat{\zeta}_n (\cdot, \beta_n) - \hat{\zeta}_o (\cdot, \beta_o) \right\}.
\]
\[ I_{2n} = P_{\Delta n} \left\{ \frac{S_1 \left( t, \hat{\theta}_n \right) \left[ \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \hat{\xi}_n \left( \cdot, \beta_0 \right) \right]}{S_0 \left( t, \hat{\theta}_n \right)} - \frac{S_{1n} \left( t, \hat{\theta}_n \right) \left[ \xi_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right]}{S_{0n} \left( t, \hat{\theta}_n \right)} \right\}, \]

\[ I_{3n} = P_{\Delta} \left\{ \frac{\hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right)}{S_{0} \left( t, \hat{\theta}_n \right)} - \frac{S_1 \left( t, \hat{\theta}_n \right) \left[ \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \hat{\xi}_n \left( \cdot, \beta_0 \right) \right]}{S_0 \left( t, \hat{\theta}_n \right)} \right\}. \]

Referring to the convergence rate results, we have \( \left\| \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right\|_2 = O_p \left( n^{-\min(p, (1-\nu)/2)} \right) \), thus by Lemma 7 in Stone (1986) we get

\[ \left\| \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right\|_\infty = \frac{q_{1/2}^{1/2} O_p \left( n^{-\min(p, (1-\nu)/2)} \right)}{O_p \left( n^{-\min(\nu, (1-\nu)/2)} \right)} = O_p \left( n^{-\min(p-\frac{1}{2}, \nu, (1-\nu)/2)} \right). \]

Now resorting to Cor. 6.21 in Schumaker (1981) we have

\[ \left\| \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right\|_\infty = O_p \left( n^{-\min(p-\frac{1}{2}, (1-\nu)/2)} \right). \]

(A.7)

Thus set \( \delta_n = n^{-\min(p-\frac{1}{2}, \nu, (1-\nu)/2)} \) in the maximal inequality in Lemma 3.4.2 of van der Vaart and Wellner (1996), one arrives at

\[ I_{1n} = O_p \left( n^{-1/2} q_{1/2}^{1/2} \delta_n \right) = n^{-1/2} O_p \left( n^{-\min(p-\nu, (1-\nu)/2)} \right) = o_p \left( n^{-1/2} \right). \]

While \( I_{2n} = o_p \left( n^{-1/2} \right) \) is confirmed by (A.5) in Lemma A.7, as \( \left\| \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right\|_\infty = o_p \left( q_{1/2}^{1/2} \right) \) given our assumptions on \( p \) and \( \nu \).

When it comes to the third term, recall that for any function \( g \)

\[ P_{\Delta} \left\{ g - \frac{S_1 \left( t, \theta_0 \right)}{S_0 \left( t, \theta_0 \right)} \left[ g \right] \right\} = E \left\{ \Delta g - E \left[ \Delta g | V = t, \Delta = 1 \right] \right\} = 0, \]

thus we get the desired conclusion

\[ I_{3n} \]

\[ = P_{\Delta} \left\{ \frac{S_1 \left( t, \theta_0 \right) \left[ \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right]}{S_0 \left( t, \theta_0 \right)} - \frac{S_1 \left( t, \hat{\theta}_n \right) \left[ \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right]}{S_0 \left( t, \hat{\theta}_n \right)} \right\} \]

\[ \leq \left\| \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right\|_\infty \left\| \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) - \xi_n \left( \cdot, \beta_0 \right) \right\|_2 \]

\[ = o_p \left( n^{-\min(p, (1-\nu)/2)} \right) \times O_p \left( n^{-\min(p-\nu, (1-\nu)/2)} \right) = o_p \left( n^{-1/2} \right). \]

Lemma A.9 Under Assumptions (A1)-(A6), we have the stochastic equicontinuity condition hold for the process:

\[ P_{\Delta n} \left\{ s_n \left( \hat{\theta}_n \right) \left[ g \right] - s_n \left( \theta_0 \right) \left[ g \right] \right\} - P_{\Delta} \left\{ s \left( \hat{\theta}_n \right) \left[ g \right] - s \left( \theta_0 \right) \left[ g \right] \right\} = o_p \left( n^{-1/2} \right), \]

with \( g = \hat{\xi}_n \left( \cdot, \hat{\beta}_n \right) \) or \( g = h^* \).
Proof. Following Huang (1999), we decompose the term on LHS into two parts:

\[ J_{1n} = (P_{\Delta n} - P_{\Delta}) \left\{ s \left( \tilde{\theta}_n \right) [g] - s (\cdot, \theta_o) [g] \right\}, \]

and

\[ J_{2n} = P_{\Delta n} \left\{ s_n \left( \tilde{\theta}_n \right) [g] - s_n (\cdot, \theta_o) [g] \right\} - P_{\Delta n} \left\{ s \left( \tilde{\theta}_n \right) [g] - s (\cdot, \theta_o) [g] \right\}. \]

\( J_{2n} = o_p \left( n^{1/2} \right) \) is proved by (A.6) in our Lemma A.7.

The conclusion of \( J_{1n} = o_p \left( n^{1/2} \right) \) would follow from the \( P \)–Donsker property of \( s (\cdot, \theta) [g] \) indexed by \( \theta \), hence the implied stochastic equicontinuity result. This assertion follows from the \( P \)–Donsker properties of \( \mathcal{F}_2 \) and \( \mathcal{F}_3 \) (or \( \mathcal{F}_4 \)) and Donsker permanence property of the convex hull when taking expectations over the class \( \mathcal{F}_2 \) and \( \mathcal{F}_3 \) (or \( \mathcal{F}_4 \)), appearing in the ratio term of \( s (\cdot, \theta) [g] \), see Theorems 2.10.2 and 2.10.3 in van der Vaart and Wellner (1996).

There are two key ingredients in our proof for the convergence rate and asymptotic normality. They are \( P \)–Donsker property and the maximal inequality (A.3), once the complexity of certain functional class is pinned down. When we move to the bootstrap world, the first property is guaranteed by the multiplier inequality in Lemma 3.6.7 of Van der Vaart and Wellner (1996) with general exchangeable weights. As for the second part, lemma 1 in Cheng and Huang (2010) would be important to characterize the convergence rate for bootstrapped processes. Let

\[ \mathcal{F}_\delta = \{ f (\cdot, \theta) - f (\cdot, \theta_o) : d (\theta, \theta_o) \leq \delta \}, \]

Let \( F_n (Z) \) be the envelope function of this class \( \mathcal{F}_{\delta_n} \). If we have

\[ \lim_{\lambda \to -\infty} \lim_{n \to \infty} \sup_{t \geq \lambda} P_Z (F_n (Z) > t) = 0, \tag{A.8} \]

for each sequence \( \delta_n \to 0 \). Then we have

\[ \sup_{d (\theta, \theta_o) \leq \delta} |G^*_n (f (\cdot, \theta) - f (\cdot, \theta_o))| = O_{P_{ZW}} (\varphi^*_n (\delta)), \]

where

\[ \varphi^*_n (\delta) \geq J_{\|\|} (\delta, \mathcal{F}_\delta, \| \cdot \|_\infty) \left[ 1 + \frac{J_{\|\|} (\delta, \mathcal{F}_\delta, \| \cdot \|_\infty)}{\delta^2 \sqrt{n}} \right]. \]

One sufficient condition to (A.8) is ensured by the uniform boundedness of the underlying class. Given our Assumptions (A1)-(A6), we do have the uniform boundedness for all those functional classes in Lemma A.6.

Proof. (of Bootstrap Consistency) Repeating the argument in our proof of Theorem 5.2, one could get

\[ d \left( \tilde{\bar{\theta}}^*_n, \theta_o \right) = o_{P_{ZW}} (1). \]

Then given Theorem 3 in Cheng and Huang, we have

\[ d \left( \tilde{\bar{\theta}}^*_n, \theta_o \right) = O_{P_{W}} \left( \frac{\min \{ p, (1 - \nu) / 2 \}}{n} \right), \]

in \( P_Z \)–probability. Following the proof of our Theorem 5.3, one arrives at

\[ \sqrt{n} P_{\Delta n} \left\{ s_n (\cdot, \theta_o) [\tilde{m}_o] \right\} \tilde{\beta}^*_n = \sqrt{n} P_{\Delta n} \left\{ s_n (\cdot, \theta_o) [\tilde{m}_o] \right\} + o_{P_{W}} (1), \]

30
in $P_Z$-probability. Notice that

\[ P_{\Delta n}^* \{ s_n (t, \theta_o) | \bar{m}_o \} = \frac{1}{n} \sum_{i=1}^{n} W_{ni} \Delta_i \left( \bar{m}_o (X_i) - \frac{S_{1n} (V_i, \theta_o) | \bar{m}_o \} }{S_{0n}^* (V_i, \theta_o)} \right), \] (A.9)

where

\[ S_{1n}^* (t, \theta) [h] = \frac{1}{n} \sum_{i=1}^{n} W_{ni} Y_i (t) \exp \left[ \psi \left( X_i' \beta \right) \right] h (X_i) \quad \text{and} \quad S_{0n}^* (t, \theta) [h] = \frac{1}{n} \sum_{i=1}^{n} W_{ni} Y_i (t) \exp \left[ \psi \left( X_i' \beta \right) \right]. \]

Because $\bar{m}_o$ and $\theta_o$ is fixed at the true value in (A.9), the convergence rate of $S_{1n}^* (t, \theta_o) | \bar{m}_o \} and $S_{0n}^* (t, \theta_o)$ to their corresponding limit $S_1 (t, \theta_o) | \bar{m}_o \}$ and $S_0 (t, \theta_o)$ would be with root-$n$ for $t \in [0, \tau]$. Hence we could proceed as in the Proposition 3 of Nan and Wellner (2013) by taking their

\[ p \in P \text{ and } i \in P \] 5.3, we obtain

Finally, combining the above equation with the linear representation for $\hat{\beta}_n - \beta_o$ in the proof of Theorem 5.3, we obtain

\[ \sqrt{n} P_{\Delta} \{ s (t, \theta_o) | \bar{m}_o \} \| \|^2 (\hat{\beta}_n - \beta_o) \]

\[ = \sqrt{n} (P_n^* - P_n) \left[ \Delta (\bar{m}_o (X) - \frac{S_1 (V, \theta_o) | \bar{m}_o \} }{S_0 (V, \theta_o)} \right) - \int s (t, \theta_o) | \bar{m}_o \} I (Y \geq t) e^{\psi (X_i' \hat{\beta}_o)} d\Lambda_o (t) \right] + o_{P_W} \] (1)

in $P_Z$-probability. The part associated with the operation $P$ in Nan and Wellner (2013) cancels out in our case, because of the independence of weights with the observation and the fact that $\frac{1}{n} \sum_{i=1}^{n} W_{ni} = 1$. Finally, combining the above equation with the linear representation for $\hat{\beta}_n - \beta_o$ in the proof of Theorem

\[ \sqrt{n} P_{\Delta} \{ s (t, \theta_o) | \bar{m}_o \} \| \|^2 (\hat{\beta}_n - \beta_o) \]

\[ = \sqrt{n} (P_n^* - P_n) \left[ \Delta (\bar{m}_o (X) - \frac{S_1 (V, \theta_o) | \bar{m}_o \} }{S_0 (V, \theta_o)} \right) - \int s (t, \theta_o) | \bar{m}_o \} I (Y \geq t) e^{\psi (X_i' \hat{\beta}_o)} d\Lambda_o (t) \right] + o_{P_W} \] (1)

in $P_Z$-probability. The final display of the bootstrapped version converges weakly to the same limit as its asymptotic counterpart conditionally for almost surely observations by Theorem 3.16.3 in van der Vaart and Wellner (1996).

References


Stable

Threshold

Y2

Duration
Poisson

Threshold

Y3

Duration