Tsukuba Economics Working Papers
No. 2019-003

Accelerated Failure Time Models with Log-concave Errors

by

Ruixuan Liu and Zhengfei Yu

November 2019

UNIVERSITY OF TSUKUBA
Department of Economics
1-1-1 Tennodai
Tsukuba, Ibaraki 305-8571
JAPAN

Accelerated Failure Time Models with Log-concave Errors

Ruixuan Liu[†] and Zhengfei Yu[‡]

[†]*Department of Economics, Emory University, 201 Dowman Drive, Atlanta, GA, 30322*
E-mail: ruixuan.liu@emory.edu

[‡]*Faculty of Humanities and Social Sciences, University of Tsukuba, 1-1-1, Tennodai, Tsukuba, Ibaraki, Japan.*
E-mail: yu.zhengfei.gn@u.tsukuba.ac.jp

Summary

We study accelerated failure time (AFT) models in which the survivor function of the additive error term is log-concave. The log-concavity assumption covers large families of commonly-used distributions and also represents the aging or wear-out phenomenon of the baseline duration. For right-censored failure time data, we construct semi-parametric maximum likelihood estimates of the finite dimensional parameter and establish the large sample properties. The shape restriction is incorporated via a nonparametric maximum likelihood estimator (NPMLE) of the hazard function. Our approach guarantees the uniqueness of a global solution for the estimating equations and delivers semiparametric efficient estimates. Simulation studies and empirical applications demonstrate the usefulness of our method.

Keywords: *Accelerate Failure Time Models, NPMLE, Weighted Rank Estimation, Shape Restriction.*

1. INTRODUCTION

The accelerated failure time (AFT) model provides an attractive alternative to the popular proportional hazards model (Cox, 1972) for analyzing censored duration/failure time data. Let Y denote the logarithm of the duration T, C be the corresponding (log-transformed) censoring time, $\Delta = 1(Y \leq C)$, and $V = \min(Y, C)$. The model of interest is

$$Y_i = X_i'\beta_0 + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1.1}$$

where X_i stands for d-dimensional covariates and the independent error term ε_i has some unknown distribution F. We denote exponential transform of ε by T_0, which represents the baseline duration variable in the absence of the covariate effect. The AFT model directly examines the effect of covariates on the length of survival, in contrast to the proportional hazards model which focuses on the hazard rate. In many applications, it is easier to visualize the concept that a treatment intervention or exposure to certain environment increases or decreases the length of survival itself by some amount, as compared to the notion that the hazard rate is changed.[1]

In the presence of censoring and unknown error distribution, weighted log-rank estimators (Prentice, 1978) have been proposed to estimate the unknown regression coefficient β_0. The large sample properties of parameter estimates are established by Tsiatis (1990),

[1]This natural type of regression relationship led Sir David Cox himself (Reid, 1994) to acknowledge "the physical or substantive basis for...proportional hazards models...is one of its weakness...accelerated life models are in many ways more appealing because of their quite direct physical interpretation."

Ritov (1990), Lai and Ying (1991), and Ying (1993). Semi-parametric efficient estimators are first discussed in Lai and Ying (1991) and subsequently studied by Zeng and Lin (2007) and Ding and Nan (2011) based on kernel and sieve methods, respectively. Despite the aforementioned theoretical advance, the practical applications of these estimators of the AFT model remain limited. A fundamental issue is whether the estimating equation has a global unique solution at the true parameter value, as commented by Kalbfleisch and Prentice (2002) [p.219] "when viewed as a function of β_0, these test statistics are step functions, and furthermore they are, in general, not monotone in β_0. This gives rise to the possibility of multiple solutions to estimating equations." For some special and inefficient weighting schemes such as the Gehan's weight (Fygenson and Ritov, 1994), the resulting estimating equation is monotone and thus delivers the global uniqueness and consistency properties. Jin et al. (2003) develop a feasible iterative algorithm that uses the rank estimator based on Gehan's weights as the pilot estimate. Instead of narrowing down the weighting scheme, Ying (1993) explore another path that utilizes shape restrictions on the error term to show the global properties of weighted rank estimators.

We study the AFT model under the restriction that the survivor function of the error term ε is log-concave. Throughout this paper, a function $g(x)$ is said to be *log-concave* if the function $\log(g(x))$ is strictly concave. It is well-known that the log-concavity of the survival function is equivalent to the increasingness of the corresponding hazard function[2]. The main advantage of such a shape restriction is that the increasing hazard function guarantees the uniqueness of global solution to the population-level estimating equations for various weighted rank estimators (Ying, 1993). To the best of our knowledge, this key insight has not yet been converted to an efficient estimation method that formally takes into account of the log-concave restriction. For that purpose, we propose semi-parametric maximum likelihood estimation methods that incoporate such a restriction for the AFT model. Specifically, for any given β, we estimate the hazard rate of $(Y - X'\beta)$ using nonparametric maximum likelihood estimator (NPMLE) and enforce the monotonicity restriction as in Huang and Wellner (1995). Note that the general NPMLE for the hazard function does not exist without such shape restrictions; see Zeng and Lin (2007). Thereafter, we obtain the estimate of β that solves the efficient score function (see Theorem 4.3 in Ding and Nan, 2011) by plugging in the NPMLE of the hazard rate and the kernel smoothed estimate for the derivative of the hazard rate. We establish the asymptotic distribution theory of the resulting estimators for the finite dimensional parameter β_0 and show that our estimators achieve the semi-parametric efficiency bound (Bickel et al., 1993). Note that the efficiency bound under our log-concave restriction remains unchanged as the one under the standard smoothness assumption. We establish this result by showing the tangent set under the shape restriction can be well approximated by the one under the smoothness restriction and vice versa. A formal proof can be found in Section S1.3 of the online supplement. In sum, the main appealing property of our approach is that it achieves the same semiparametric efficiency as the kernel or sieve based estimators (Lai and Ying, 1991; Ding and Nan, 2011) and at the same time guarantees the uniqueness of a global solution for the estimation equations (and the resulting computational convenience), which previously was only shown for the inefficient Gehan's weighting scheme. In principle, a shape-constrained sieve MLE can serve as an alterna-

[2]See Proposition C.1.c. in Chapter 4 of Marshall and Olkin (2007) and the proof therein. Throughout this paper, the increasingness (decreasingness) of a function refers to the *strictly* increasingness (decreasingness).

tive for incorporating the log-concave restriction. However, it usually requires imposing additional inequality restrictions on the coefficients of basis functions, which may further complicate the numerical optimization procedure; see Section 2.3.5 of Chen (2007). In contrast, our NPMLE-based approach is made possible by the underlying log-concave restriction and automatically satisfies it. In this sense, the shape restriction is a blessing instead of a burden for our estimation approach.

The framework of log-concave errors comprises numerous parametric duration models where the failure time T (conditional on covariates) can follow exponential, Weibull, Gamma, Log-normal, Log-logistic, and many other distributions (Bagnoli and Bergstrom, 2005; Marshall and Olkin, 2007). It is worth clarifying the relationship between the log-concavity restriction on ε and the shape of the baseline hazard function of $T_0 = \exp(\varepsilon)$. On one hand, if one believes that baseline duration T_0 exhibits some aging or wear-out phenomenon, as encoded by an increasing baseline hazard function, then its logarithm transform ε will also have an increasing hazard rate. This feature often arises from a non-stationary job search model where the increasing exit rate of unemployment status is caused by a declining reservation wage and/or an increasing search intensity (Burdett and Vishwanath, 1988; Van den Berg, 1990). On the other hand, the increasing hazard of ε itself does not necessarily restrict the shape of the hazard of T_0. Our Example 2.1 provides an encompassing parametric family in which the hazard shape of T_0 can be increasing, decreasing, or non-monotonic, under the log-concavity restriction on ε.

In establishing the large sample properties of our estimators, our main technical contribution is a formal analysis of the asymptotics of the NPMLE of the hazard rate and its kernel smoothed derivative, adapting the recent development from Groeneboom et al. (2010), Groeneboom and Hendrickx (2018), and Westling and Carone (2019). This is not a standard problem, as the NPMLE of the hazard rate is a piece-wise constant function with random jump locations determined by the data, and the kernel smoothed estimate of NPMLE is not a linear functional of the empirical measure (Groeneboom et al., 2010). Compared with the binary choice model in Groeneboom and Hendrickx (2018), the characterization of NPMLE is more complicated in our setting, because its min-max representation in equation (3.15) involves random denominators, whereas the one in Groeneboom and Hendrickx (2018) does not; see Westling and Carone (2019) for more discussions. Our proofs that combine empirical process theory and the characterization of the (smoothed) NPMLE for the monotone hazard are also of independent interest. The shape restricted estimation and inference constitute a rich and evolving literature in econometrics and statistics, as reviewed by Groeneboom and Jongbloed (2014) and Chetverikov et al. (2018). The rest of the paper is organized as follows: Section 2 discusses the log-concavity restriction. Section 3 proposes efficient semiparametric maximum likelihood estimation (SMLE) methods subject to the log-concave restriction on the error term. Section 4 derives the asymptotic properties of the parameter estimates, proving the consistency, asymptotic normality, and semi-parametric efficiency. Section 5 conducts simulation studies and applies the proposed methods to a real data set. The final section concludes. Main results are proved in Appendix. Other proofs, technical lemmas, and an additional real data example are relegated to the online supplement.

2. LOG-CONCAVITY OF ERROR TERMS IN AFT MODELS

We begin with a clarification about the log-concavity restriction used in this paper. Proposition 2.1 collects some important properties related to the log-concavity restric-

tion and states the relationship between the hazard functions of ε and T_0. Let $f(u)$ be the (absolutely continuous) density function of the error term ε in (1.1), $F(u)$ be the cumulative distribution function, $\bar{F}(u) = 1 - F(u)$ be the survivor function and $\lambda(u) = f(u)/\bar{F}(u)$ be the hazard function. Let $h(u)$ be the hazard function of the baseline duration $T_0 = \exp(\varepsilon)$. The log-concave restriction in this paper refers to the survivor function $\bar{F}(u)$ being log-concave.

PROPOSITION 2.1. *(i). If $f(u)$ is log-concave, then both $F(u)$ and $\bar{F}(u)$ are log-concave. (ii). $\bar{F}(u)$ is log-concave if and only if $\lambda(u)$ is increasing. (iii). If $h(u)$ is increasing, then $\lambda(u)$ is also increasing.*

As shown in part (i) of Proposition 2.1, the log-concavity restriction on $\bar{F}$ can be derived from the log-concave restriction on the density function f. For parametric models, it is often easier to directly verify the log-concavity of the density function. Example 2.1 below shows that a large family of commonly-used parametric duration models falls into our modeling framework. From the perspective of semiparametric estimation for the AFT model, the log-concave restriction yields nice properties for weighted rank estimation. Specifically, Lemma 4.1 shows that an increasing hazard λ is sufficient for the population level estimating equation to have a unique global solution.

EXAMPLE 2.1. *(Generalized F-distribution). Various frequently used continuous parametric failure time models are special cases of the model (1.1) in which the error term ε is distributed as the logarithm of the F-variate with the degrees of freedom equal to $(2m_1, 2m_2)$ on [p.38] of Kalbfleisch and Prentice (2002). The density function of ε writes:*

$$f(u) = (m_1/m_2)^{m_1} e^{m_1 u} \left(1 + (m_1/m_2)e^{u}\right)^{-(m_1+m_2)} / B(m_1, m_2), \tag{2.2}$$

where $B(m_1, m_2)$ is Beta function. The density function (2.2) reduces to the logistic density when $(m_1, m_2) = (1, 1)$; it reduces to the extreme value density (that leads to the Weibull model) when $(m_1, m_2) = (1, \infty)$; it generates the Gamma model when $m_2 = \infty$; it approaches to the standard normal density as $(m_1, m_2) \to (\infty, \infty)$. It is easy to show that the second order derivative of the logarithm of $f(u)$ is

$$(\log f(u))'' = -(m_1 + m_2)(m_2/m_1)e^{-u} / \left(1 + (m_2/m_1)e^{-u}\right)^2 < 0, \tag{2.3}$$

for any $m_1, m_2 > 0$. Hence, the density function (2.2) is log-concave and the model is nested by our setup.

Note that the converse of part (iii) of Proposition 2.1 is not true. In other words, our log-concave restriction on ε does not restrict the hazard function of T_0 to be monotone. One can easily see this by inspecting the following expression:

$$h(t) = \left[f(\log(t))/t\right] / \bar{F}(\log(t)) = \lambda(\log t)/t. \tag{2.4}$$

Therefore, our shape restriction can still be plausible in the cases where empirical evidence rejects a monotonic hazard function of T_0; see Christofides and McKenna (1996) for such an example[3]. Figure 1 plots hazard functions of the error term ε and the baseline duration T_0 for parametric models encompassed by Example 2.1. Despite an increasing hazard rate

[3]Christofides and McKenna (1996) model the baseline duration by a generalized Gamma distribution that is log-concave, in order to accommodate a non-monotonic hazard of the unemployment duration.

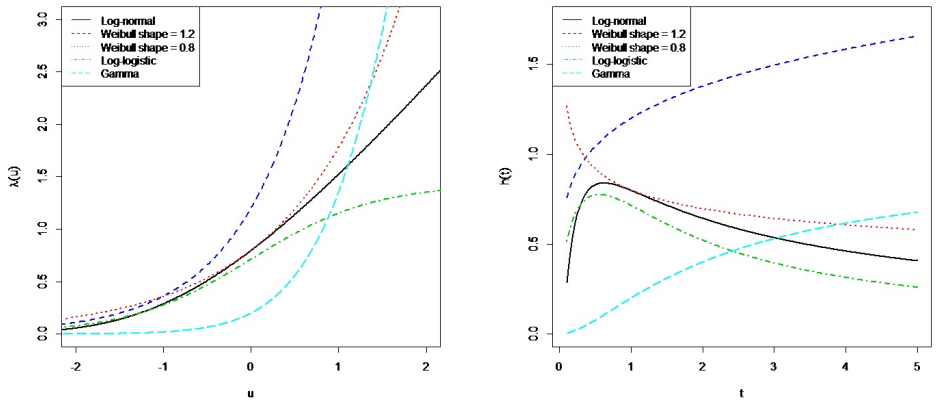

Figure 1. Hazard functions of the error term ε (left) and the baseline duration T_0 (right) for various parametric models.

for ε, the hazard rate of its exponential transform T_0 can be increasing, decreasing or non-monotonic. We observe the following: (i). The log-normal failure time model, in which ε has a standard normal distribution, yields a increasing hazard function for ε, but an inverse U-shaped hazard function for T_0. (ii). The Weibull model, in which the baseline hazard function takes the form $h(t) = \gamma t^{\gamma-1}$, has increasing hazard functions for both ε and T_0 when the shape parameter $\gamma > 1$; it implies a increasing hazard for ε and a decreasing hazard for T_0 when $0 < \gamma < 1$. (iii). The Log-logsitic model, in which ε has a logistic density $e^{u/\sigma}/\sigma(1+e^{u/\sigma})^2$, produces a increasing hazard function for ε and an inverse U-shaped function for T_0, when $0 < \sigma < 1$ ($\sigma = 0.7$ in the figure). (iv). The Gamma model, in which T_0 has a density $t^{k-1}e^{-t}/\Gamma(k)$, yields increasing hazard functions for both ε and T_0 when the shape parameter $k > 1$ ($k = 3$ in the figure).

Inspired by Abbring (2012) which studies durations defined by threshold-crossing rules, we provide a model that yields an increasing hazard function for T_0. Conditional on $X = x$, the duration T is defined as the survival time after a random number of shocks and the arrival of shocks is governed by a Poisson process $N_x(t)$, so that

$$\Pr\{T > t | X = x\} = \sum_{k=0}^{\infty} (\lambda_x t)^k e^{-\lambda_x t}/k! \overline{P}_k, \tag{2.5}$$

where $\lambda_x \equiv \exp(-x'\beta)$ is the failure rate of $N_x(t)$ and $\overline{P}_k$ is the probability of surviving after k shocks. The following two examples are from Esary and Marshall (1973) and spell out the interpretative content of the underlying failure mechanism based on different specifications of the sequence $\overline{P}_k$.

Example 2.2. *(Cumulative Shock Models) The failure is caused by cumulative shocks U_1, $U_1 + U_2$, $\cdots$ reaching some positive random threshold Z, as in Section 5 of Esary and Marshall (1973), with the following specification on $\bar{P}_k$ for $k \geq 1$:*

$$\overline{P}_k = \Pr\{U_1 + ... + U_k \leq Z\} = \int_0^{\infty} F_U^{(k)}(z) dG_Z(z),$$

where G_Z is the cumulative distribution function of Z, and $F_U^{(k)}$ stands for the k-th fold convolution of the function F_U, which is the distribution function of i.i.d. shocks U_k for $k \geq 1$.

EXAMPLE 2.3. *(Maximum Shock Models) The failure is caused by the maximum of independent shocks U_1, U_2, $\cdots$ reaching a fixed threshold z, as in Section 6 of Esary and Marshall (1973), with the following specification on $\bar{P}_k$ for $k \geq 1$:*

$$\overline{P}_k = \Pr\{\max(U_1, \cdots, U_k) \leq z\} = \Pi_{i=1}^k F_{U_i}(z),$$

where the distribution function F_{U_k} is not necessarily identical for $k \geq 1$.

We show in the following proposition that the construction via equation (2.5) generates an AFT model with some baseline duration T_0 that has an increasing hazard function under reasonable assumptions on $\overline{P}_k$ for both Examples 2.2 and 2.3. Therefore, the corresponding additive error term ε also has an increasing hazard rate, according to part (iii) of Proposition 2.1.

PROPOSITION 2.2. *The duration T defined by equation (2.5) generates an AFT model with a log-concave baseline duration T_0 if the following conditions hold: (i) For the cumulative shock models, both the shock U and threshold Z have log-concave densities; (ii) For the maximum shock models, $F_{U_k}(u)$ is decreasing in k for any u (so that subsequent shocks are stronger).*

3. SEMI-PARAMETRIC EFFICIENT ESTIMATION WITH SHAPE RESTRICTIONS

The data consists of *i.i.d.* observations of $\{V_i, X_i, \Delta_i; i = 1, \cdots, n\}$, where $V_i = \min(Y_i, C_i)$ and $\Delta_i = 1(Y_i \leq C_i)$. The error term ε_i in (1.1) is assumed to be independent of X_i and the right-censoring time C_i. Let the marginal distribution function of covariates X be $H(x)$ for x in the support $\mathcal{X}$. Following Ying (1993), we denote the conditional distribution and density functions of the censoring variable C as $G_x(\cdot)$ and $g_x(\cdot)$, respectively. For a given β, define $\epsilon_\beta = V - X'\beta$ and $\epsilon_0 = V - X'\beta_0$. We use the standard empirical process notations as follows. For a function f of a random vector $Z = (V, X, \Delta)$ that follows distribution P,

$$Pf = \int f(z)dP(z), \quad \mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(Z_i), \quad \mathbb{G}_n f = n^{1/2}\left(\mathbb{P}_n - P\right)f. \tag{3.6}$$

Function f can be replaced by a random function $z \mapsto \hat{f}_n(z; Z_1, \cdots, Z_n)$ (Nan and Wellner, 2013), so we also write $P\hat{f}_n$, $\mathbb{P}_n\hat{f}_n$, or $\mathbb{G}_n\hat{f}_n$. Furthermore, define

$$D_n^{(0)}(t,\beta) = \mathbb{P}_n\{1(\epsilon_\beta \geq t)\}, \quad D^{(0)}(t,\beta) = P\{1(\epsilon_\beta \geq t)\}, \tag{3.7}$$

$$D_n^{(1)}(t,\beta) = \mathbb{P}_n\{X1(\epsilon_\beta \geq t)\}, \quad D^{(1)}(t,\beta) = P\{X1(\epsilon_\beta \geq t)\}, \tag{3.8}$$

and

$$\hat{\eta}_n(t,\beta) = D_n^{(1)}(t,\beta)/D_n^{(0)}(t,\beta). \tag{3.9}$$

The probability limit of $\hat{\eta}_n(t,\beta)$ for any given β is

$$\eta_0(t,\beta) = D^{(1)}(t,\beta)/D^{(0)}(t,\beta) = E[X|V - X'\beta \geq t]. \tag{3.10}$$

In addition, we define

$$N_n(t,\beta) = \mathbb{P}_n\{1(\epsilon_\beta \le t, \Delta = 1)\}, \quad N(t,\beta) = P\{1(\epsilon_\beta \le t, \Delta = 1)\}. \tag{3.11}$$

The likelihood function (omitting terms irrelevant for our parameters of interest) is

$$L_n(\beta,\lambda) = n^{-1}\sum_{i=1}^{n}\Big\{\Delta_i \log\lambda(V_i - X_i'\beta) - \int_{-\infty}^{V_i}\lambda(t - X_i'\beta)dt\Big\}; \tag{3.12}$$

also see Zeng and Lin (2007) and Ding and Nan (2011). The unknown parameters consist of $\beta \in \mathbf{B}$ and $\lambda(t) \in \mathcal{A}$, where $\mathbf{B}$ is a compact set in $\mathbb{R}^d$ and $\mathcal{A}$ is the space for increasing and non-negative functions. We first derive $\hat\lambda_n(\cdot,\beta)$ that maximizes (3.12) for any fixed β. It is an estimator of the increasing function $\lambda_0(t,\beta) = (dN(t,\beta)/dt)\,/D^{(0)}(t,\beta)$ and can be obtained by applying NPMLE for the monotone hazard function (Huang and Wellner, 1995). Recall that $\epsilon_{i,\beta} = V_i - X_i'\beta$, and the corresponding indicator is $\Delta_{i,\beta}$ for $i = 1,\cdots,n$. Let $\epsilon_{(1),\beta} \le \cdots \le \epsilon_{(n),\beta}$ be the order statistics, and $\Delta_{(i),\beta}$ be the indicator associated with the i-th order statistic $\epsilon_{(i),\beta}$. Define the weights

$$w_{i,\beta} = n\{\epsilon_{(i+1),\beta} - \epsilon_{(i),\beta}\}\int_{u\ge\epsilon_{(i+1),\beta}} d\mathbb{P}_n(u), \text{ for } i = 1,...,n-1. \tag{3.13}$$

Assuming that the function $\lambda_0(t,\beta) \le M_\lambda$ for some constant M_λ and for all $\beta \in \mathbf{B}$,[4] the NPMLE $\hat\lambda_n(\cdot,\beta)$ can be characterized as the left derivative of the greatest convex minorant of the following "cumulative sum diagram" with points

$$P_0 = (0,0), \text{ and } P_i = \left(\sum_{j=1}^{i} w_{j,\beta}, \sum_{j=1}^{i} w_{j,\beta}\Delta_{(i),\beta}/w_{i,\beta}\right), \text{ for } i = 1,\cdots,n-1. \tag{3.14}$$

By the min-max characterization (Groeneboom and Jongbloed, 2014), the NPMLE is a step-wise constant function uniquely determined at $\epsilon_{i,\beta}$ for $i = 1,\cdots,n-1$ as

$$\hat\lambda_n(\epsilon_{(i),\beta},\beta) = \max_{1\le s\le i}\min_{i\le t\le n}\frac{\sum_{j=s}^{t}\Delta_{(j),\beta}}{\sum_{j=s}^{t}(n-j+1)(\epsilon_{(j),\beta} - \epsilon_{(j-1),\beta})}, \tag{3.15}$$

and $\hat\lambda_n(\epsilon_{(n),\beta},\beta) = M_\lambda$. In practice, it is enough to compute $\hat\lambda_n(\epsilon_{(i),\beta},\beta)$ at the first $n-1$ points as in (3.15) and thus there is no need to specify M_λ. The established pool-adjacent-violators algorithm (PAVA) can be used to compute NPMLE in a very efficient way; see Groeneboom and Jongbloed (2014). The probability limit of $\hat\lambda_n(\cdot,\beta)$ is $\lambda_0(t,\beta)$. We also denote $\lambda_0(t) \equiv \lambda_0(t,\beta_0)$. It is worthwhile emphasizing that for any given β, the likelihood function cannot be meaningfully maximized without the shape restriction on λ (Zeng and Lin, 2007).

Referring to the semi-parametric efficient estimator presented below, smoothing is inevitable, because the efficient score function involves the derivative of the hazard function. Thus, we adopt the smoothed maximum likelihood estimator in Groeneboom et al.

[4]This boundedness assumption is required for maximizing the likelihood in (3.12). Otherwise, $\hat\lambda_n(\epsilon_{(n),\beta},\beta)$ can be chosen arbitrarily large. After obtaining the solution under the upper bound M_λ, one can allow M_λ to grow arbitrarily large. This type of argument is common in isotonic optimization problems, see [p338-339] of Robertson et al. (1988) and [p38-39] of Groeneboom and Jongbloed (2014). In computation, one can set $\hat\lambda_n(\epsilon_{(n),\beta},\beta) = \hat\lambda_n(\epsilon_{(n-1),\beta},\beta)$.

(2010) and obtain

$$\dot{\hat{\lambda}}_n(t,\beta) = \int K_h(t-u)d\hat{\lambda}_n(u,\beta), \tag{3.16}$$

for a kernel density function $K(\cdot)$ and $K_h(\cdot) \equiv K(\cdot/h)/h$, with the bandwidth equal to h. It is straightforward to find that the probability limit of the kernel smoothed estimate is

$$\dot{\lambda}_0(t,\beta) = \partial\lambda_0(t,\beta)/\partial t. \tag{3.17}$$

To estimate β_0, we consider the following random map

$$\Psi_n(\beta,\eta,\rho) \equiv \frac{1}{n}\sum_{i=1}^{n} \Delta_i \rho(V_i - X_i'\beta,\beta)\{X_i - \eta(V_i - X_i'\beta,\beta)\}, \tag{3.18}$$

where $\beta \in \mathbf{B}$ is the d-dimensional Euclidean parameter of interest with an unknown true value of β_0, η and ρ are functions that can be viewed as infinite dimensional nuisance parameters. Intuitively speaking, η provides the correct centering term for the regressor X under the censoring, whereas the weight ρ influences the efficiency. In this paper, we consider $\Psi_n(\beta,\hat{\eta}_n,\hat{\rho}_n)$ which approximates the efficient score function and belongs to the general weighted rank estimating equation as in Tsiatis (1990), with an efficient weighting scheme. Specifically, these nuisance components are estimated by

$$\hat{\eta}_n(t,\beta) = D_n^{(1)}(t,\beta)/D_n^{(0)}(t,\beta), \quad \hat{\rho}_n(t,\beta) = \mathbb{I}\{t \geq \tau_U\}\dot{\hat{\lambda}}_n(t,\beta)/\hat{\lambda}_n(t,\beta). \tag{3.19}$$

Following Tsiatis (1990) (see also Lai and Ying, 1991; Ding and Nan, 2011), the trimming constant τ_U is introduced to avoid the instability of $D_n^{(0)}$ and $\hat{\lambda}_n$ near the right tail. The corresponding probability limit of $\hat{\rho}_n(t,\beta)$ is denoted by $\rho_0(t,\beta)$.

Our estimator $\hat{\beta}_n$ is the root of the estimating function $\Psi_n(\beta,\hat{\eta}_n(\cdot,\beta),\hat{\rho}_n(\cdot,\beta))$, which is discontinuous in β. Therefore, $\Psi_n(\beta,\hat{\eta}_n(\cdot,\beta),\hat{\rho}_n(\cdot,\beta)) = 0$ may not hold exactly. Following Groeneboom and Hendrickx (2018), we define $\hat{\beta}_n$ as a *zero-crossing* point such that each coordinate value of $\Psi_n(\beta,\hat{\eta}_n(\cdot,\beta),\hat{\rho}_n(\cdot,\beta))$ changes sign in the left and right neighborhoods of $\hat{\beta}_n$.

Definition 3.1. (Zero-crossing) *We say that β_* is a zero-crossing of a function $C : \mathbf{B} \mapsto \mathbb{R}$ if each open neighborhood of β_* contains points $\beta_1, \beta_2 \in \mathbf{B}$ such that $C(\beta_1)C(\beta_2) \leq 0$. Moreover, we say that a function $\tilde{C} : \mathbf{B} \mapsto \mathbb{R}^d$ crosses zero at point β_* if β_* is zero-crossing in each component $\tilde{C}_j$ for $j = 1, \cdots, d$.*

We propose another estimation procedure which uses a smoothed version of $\hat{\lambda}_n(t,\beta)$. It is well documented in the literature (Groeneboom et al., 2010) that smoothed NPMLE often performs better in finite samples. For the kernel density function $K(\cdot)$, we define $\mathbb{K}(t) = \int_{-\infty}^{t} K(u)du$ and $\mathbb{K}_{\tilde{h}}(u) = \mathbb{K}(u/\tilde{h})$. The smoothed estimator for $\lambda_0(t,\beta)$ is

$$\tilde{\lambda}_n(t,\beta) = \int \mathbb{K}_{\tilde{h}}(t-u)d\hat{\lambda}_n(u,\beta). \tag{3.20}$$

Therefore, the alternative estimating equation is

$$\tilde{\Psi}_n(\beta,\eta,\rho) \equiv \frac{1}{n}\sum_{i=1}^{n} \Delta_i \tilde{\rho}_n(V_i - X_i'\beta,\beta)\{X_i - \hat{\eta}_n(V_i - X_i'\beta,\beta)\}, \tag{3.21}$$

with

$$\tilde{\rho}_n(t,\beta) = \mathbb{I}\{t \geq \tau_U\}\dot{\hat{\lambda}}_n(t,\beta)/\tilde{\lambda}_n(t,\beta). \tag{3.22}$$

The corresponding solution is denoted by $\tilde{\beta}_n$.

Figure 2 illustrates typical shapes of estimating functions Ψ_n and $\tilde{\Psi}_n$ for a single parameter β using the unsmoothed and smoothed estimates of $\lambda_0(t,\beta)$. The sample size is 500 and is generated from a $N(2, 0.5^2)$ covariate, a standard normal error term, and a uniform censoring variable on $[0,4]$. The true parameter is $\beta_0 = 1$. The bandwidths are $5 \times n^{-1/5}$ for $\tilde{\lambda}_n(t,\beta)$ and $5 \times n^{-1/7}$ for $\dot{\hat{\lambda}}_n(t,\beta)$, respectively. Both methods deliver unique zero-crossing points, whereas $\tilde{\Psi}_n$ exhibits smoother variation.

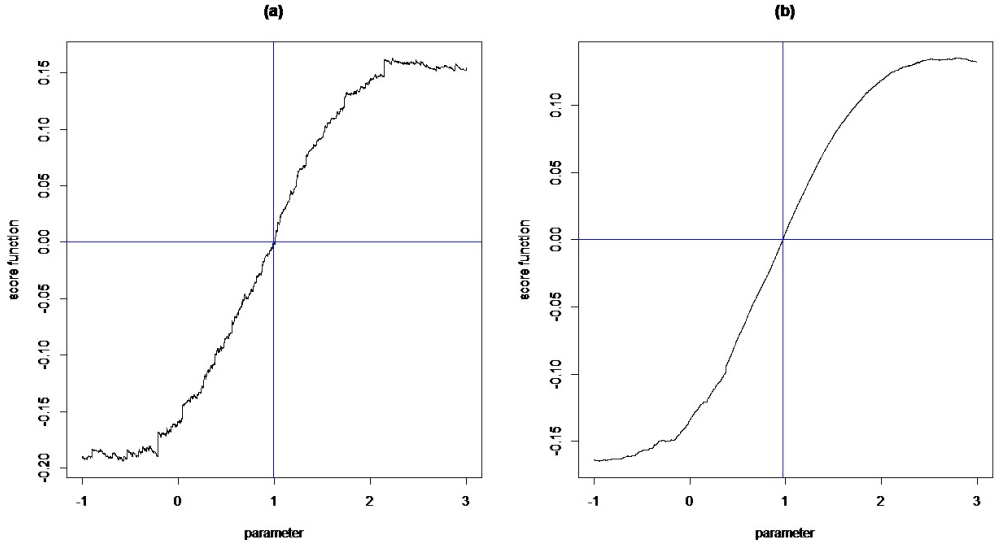

Figure 2. The estimating equations Ψ_n (left) and $\tilde{\Psi}_n$ (right) as functions of the single parameter β.

REMARK 3.1. *Lai and Ying (1991) are among the first who propose to use the kernel smoothing methods to estimate the underlying hazard and its derivative functions in the efficient score function. Their specific construction consists of (i) dividing the sample into two disjoint subsets and evaluating a preliminary consistent estimate $\hat{b}_j$ of β from the jth subsample ($j = 1, 2$), (ii) finding from the uncensored residuals in the jth subsample a smooth consistent estimate $\hat{\lambda}_j$ of the hazard function, (iii) smoothing the estimated hazard to obtain a smooth consistent estimate $\hat{\rho}_j$ of $\dot{\lambda}/\lambda$, and (iv) using $\hat{\rho}_1$ (respectively $\hat{\rho}_2$) as the weight function for the linear rank statistic of the second (respectively first) sample of residuals $Y_i - X_i'b$. The sum $S(b)$ of these two linear rank statistics is used to define the rank estimator as the minimizer of $\| S(b) \|$. However, there are practical difficulties in carrying out this procedure, as reviewed by Kim and Lai (2000). First, this efficient rank estimator is difficult to compute when the regressor X is multidimensional.*

Second, the kernel smoothing approach employed does not give good results unless the sample size is very large.

REMARK 3.2. *Besides the guarantee of the unique solution, another advantage of our efficient weighted rank estimator over the sieve MLE in Ding and Nan (2011) stems from its computational convenience. In the former, only the regression parameter β needs to be solved from the estimating equations, whereas both β and the spline coefficients (the number of such coefficients also diverges with the sample size to eliminate the approximation bias) need to be solved in the latter. For example, the sieve approach using cubic splines and two internal knots has to solve six more parameters from the optimization procedure than our approach. Meanwhile, the pool-adjacent-violators algorithm (PAVA) used to compute the NPMLE of the hazard rate (Groeneboom and Jongbloed, 2014) is faster than the iterative Newton-Raphson method for the sieve MLE estimate.*

4. ASYMPTOTIC PROPERTIES

The following regularity conditions, adapted from Tsiatis (1990), Lai and Ying (1991), and Ying (1993), are imposed throughout. We let $\mathcal{T}$ to be the support of ε trimmed from the right tail at some given τ_U (Tsiatis, 1990).

ASSUMPTION 4.1. *The covariates X_i are uniformly bounded, i.e., $\max_{1\leq i\leq n}\|X_i\| \leq M_x$ for some finite constant M_x.*

ASSUMPTION 4.2. *The error term ε is independent of (X, C) and has a finite mean. The density f and its derivative $\dot{f}$ are bounded. Moreover, $\int_{-\infty}^{\infty}\left(\dot{f}(t)/f(t)\right)^2 f(t)dt < \infty$.*

ASSUMPTION 4.3. *The conditional density $g_x(t)$ of the censoring time C is uniformly bounded; that is, $\sup_{t\in\mathcal{C},x\in\mathcal{X}}|g_x(t)| \leq M_c$ for some finite constant M_c, where $\mathcal{X}$ is the support of covariates X and $\mathcal{C}$ is the support of censoring variable C.*

ASSUMPTION 4.4. *The hazard rate λ_0 of error term ε is increasing and is continuously third-order differentiable. Moreover, $\lambda_0(t,\beta)$ is uniformly bounded away from zero for any β in the parameter space and $t \in \mathcal{T}$. Its derivative $\dot{\lambda}_0(t,\beta)$ is also uniformly bounded away from 0 and ∞.*

ASSUMPTION 4.5. *For the finite positive constant τ_U, there exists a value of ξ such that $Pr(V - X'\beta_0 \geq \tau_U) \geq \xi > 0$.*

ASSUMPTION 4.6. *(i) The kernel function $K(u)$ is a kernel density function with compact support such that $\int K(u)du = 1$, $\int uK(u)du = 0$, and $\int u^2K(u)du < \infty$. Moreover, its derivative $k(u)$ is uniformly continuous over the support. The kernel function and its derivative can be written in the form of $\varphi(p(x))$, with some function $\varphi(\cdot)$ being of bounded variation and $p(x)$ a real polynomial on $\mathbb{R}$. (ii) The bandwidths satisfy $h \asymp n^{-1/7}$ and $\tilde{h} \asymp n^{-1/5}$.*

Assumptions 4.1 to 4.3 are from Ying (1993). The monotonicity restriction in Assumption 4.4 results from our log-concave restriction on the survival function of ε through Proposition 2.1 (ii). The class of hazard rate functions that are bounded away from zero

when t approaches to the left boundary is discussed extensively in Ridder and Woutersen (2003). With log-concave errors, the moment bound in Condition 4 of Ying (1993) is automatically satisfied. When the error term ε has a finite mean and an increasing hazard rate, it has a sub-exponential tail, cf. Theorem 4.1 in Barlow and Marshall (1964), which gives rise to finite moments of all orders. In accordance with Assumption 4.5, we trim the right tail at some large constant τ_U; see equation (3.1) of Tsiatis (1990). This is a standard practice in analyzing censored regression models (Lai and Ying, 1991; Ritov, 1990; Ding and Nan, 2011). In fact, the calculation of the semiparametric information bound takes this trimming parameter τ_U as a fixed constant; see Example 4 on [p.284] of Bickel et al. (1993). It is possible to allow τ_U to be infinity through more technical proofs (see Lemma 2 of Ying, 1993), but it will not be attempted here. In Assumption 4.6, we collect the standard requirements on kernel smoothing methods, as in Groeneboom et al. (2010). Under the restriction, we have the following two VC-type functional classes due to Nolan and Pollard (1987):

$$\mathcal{K}_1 = \left\{K\left(h^{-1}\left(x-\cdot\right)\right) : x \in \mathbb{R},\, h>0\right\} \quad \text{and} \quad \mathcal{K}_2 = \left\{k\left(h^{-1}\left(x-\cdot\right)\right) : x \in \mathbb{R},\, h>0\right\},$$

which are needed to establish the convergence rates for $\hat{\dot{\lambda}}_n$ and $\tilde{\lambda}_n$. The population level estimating equations can be written as

$$\Psi(\beta, \eta_0(\cdot,\beta), \rho_0(\cdot,\beta)) = P\left[\Delta\rho_0(V-X'\beta,\beta)\{X-\eta_0(V-X'\beta,\beta)\}\right]. \tag{4.23}$$

To show the existence of a solution of estimating equations (3.18) and (3.21) with probability tending to one and its consistency, we verify that the population level estimating function $\Psi(\beta, \eta_0(\cdot,\beta), \rho_0(\cdot,\beta))$ has a unique root under the shape restriction given by Assumption 4.4. This result essentially follows from the discussion in Section 5 of Ying (1993) and we restate it as Lemma 4.1 for completeness.

LEMMA 4.1. *If the hazard function λ of the error term ε is increasing, as assumed in Assumption 4.4, then $\beta_0 \in \mathbf{B}$ is the unique solution to the population level estimating functions $\Psi(\beta, \eta_0(\cdot,\beta), \rho_0(\cdot,\beta))$.*

In particular, Ying (1993) shows that the population criterion function has a unique zero-crossing point globally when the error term has an increasing hazard function and the weights do not have alternating signs (either all positive or all negative). In our case, the weighting function $\hat{\rho}_n(\cdot)$ defined in (3.19) is automatically non-negative, so is its probability limit. In sharp contrast, standard kernel or sieve type estimators for $\hat{\rho}_n(\cdot)$ without such shape constraint fail to deliver a unique global solution; see Kim and Lai (2000). Based on Lemma 4.1, the consistency of our estimators can be obtained from the uniform convergence of $\Psi_n(\beta, \hat{\eta}_n(\cdot,\beta), \hat{\rho}_n(\cdot,\beta))$ and the Glivenko-Cantelli theorem.

THEOREM 4.1. (CONSISTENCY) *Suppose that Assumptions 4.1 to 4.6 hold. Then, for all large n, a zero-crossing $\hat{\beta}_n$ for $\Psi_n\left(\hat{\beta}_n, \hat{\eta}_n(\cdot,\hat{\beta}_n), \hat{\rho}_n(\cdot,\hat{\beta}_n)\right)$ exists with probability tending to one and is a consistent estimator of β_0. The same conclusion holds for $\tilde{\beta}_n$, which is a zero-crossing point for $\tilde{\Psi}_n\left(\tilde{\beta}_n, \hat{\eta}_n(\cdot,\tilde{\beta}_n), \tilde{\rho}_n(\cdot,\tilde{\beta}_n)\right)$.*

Because our estimation procedure belongs to the general Z-estimation with bundled parameter and nonparametric nuisance components, we prove the root-n rate and asymptotic normality of $\hat{\beta}_n$ following the route in Nan and Wellner (2013). Note that under our

assumptions, $\rho_0(\epsilon_\beta, \beta)$ and $\eta_0(\epsilon_\beta, \beta)$ are both continuously differentiable with derivatives denoted by $\dot{\rho}_{0\beta}$ and $\dot{\rho}_{0\beta}$, respectively. As shown in Theorem 3.3 in Nan et al. (2009), this implies that $\Psi(\beta, \eta_0(\cdot, \beta), \rho_0(\cdot, \beta))$ is differentiable in β with the bounded derivative $\dot{\Psi}_\beta(\beta, \eta_0(\cdot, \beta), \rho_0(\cdot, \beta))$ in $\mathbf{B}$. Given the explicit form of $\Psi(\beta, \eta(\cdot, \beta), \rho(\cdot, \beta))$, the path derivatives (Ichimura and Lee, 2010) with respect to both η and ρ exist, and we denote them by $\dot{\Psi}_\eta(\beta, \eta(\cdot, \beta), \rho(\cdot, \beta))$ and $\dot{\Psi}_\rho(\beta, \eta(\cdot, \beta), \rho(\cdot, \beta))$, respectively.

THEOREM 4.2. (ASYMPTOTIC NORMALITY) *Suppose that Assumptions 4.1 to 4.6 hold. Then, we have the following linear representation*

$$-\dot{\Psi}_\beta(\beta_0, \eta_0(\cdot, \beta_0), \rho_0(\cdot, \beta_0))n^{1/2}(\hat{\beta}_n - \beta_0) = \mathbb{G}_n \int \rho_0(t, \beta_0)\{X - \eta_0(t, \beta_0)\}dM(t) + o_p(1),$$

where $M(t)$ is the martingale:

$$M(t) = \Delta 1(V - X'\beta_0 \leq t) - \int_{-\infty}^{t} 1(V - X'\beta_0 \geq s)\lambda_0(s)ds. \tag{4.24}$$

Therefore, we obtain

$$n^{1/2}(\hat{\beta}_n - \beta_0) \to \mathbb{N}(0, I^{-1}(\beta_0))$$

in distribution. The information matrix $I(\beta_0)$ is the semi-parametric efficient information matrix given by Lemma S1 in the online supplement. The same conclusion holds for $\tilde{\beta}_n$.

The calculation of semi-parametric information bound embedding shape restrictions is not trivial; see Tripathi (2000), and Kuchibhotla et al. (2017). We formally verify in the online supplement that for the increasing hazard rate, the information bound remains unchanged as in Bickel et al. (1993). In order to determine the bound, the tangent set and the projection of the (parametric) score function to the tangent set need to be calculated. The score function is not affected by the shape restriction. The crux in our proof is to show the tangent set remains unchanged by showing that scores for smooth sub-models lie in the set and by exhibiting a family of smooth sub-models with scores that can approximate any element of the set arbitrarily well are dense in the set.

REMARK 4.1. *The standard errors of $\hat{\beta}_n$ can be obtained based on the following result:*

$$\frac{1}{n}\sum_{i=1}^{n} \Delta_i \left[\hat{\rho}_n(V_i - X_i'\hat{\beta}_n, \hat{\beta}_n)\{X_i - \hat{\eta}_n(V_i - X_i'\hat{\beta}_n, \hat{\beta}_n)\}\right]^{\otimes 2} \to_p I(\beta_0). \tag{4.25}$$

In the proofs of Theorem 4.1 and Theorem 4.2, we have $\Psi_n\left(\hat{\beta}_n, \hat{\eta}_n(\cdot, \hat{\beta}_n), \hat{\rho}_n(\cdot, \hat{\beta}_n)\right) = \mathbb{P}_n\{l^*_{\beta_0}(V, \Delta, X)\} + o_p(1)$, where $l^*_{\beta_0}$ is the efficient score function such that $I(\beta_0) = \mathbf{E}[l^*_{\beta_0}(V_i, X_i, \Delta_i)^{\otimes 2}]$. The experssion of $l^*_{\beta_0}$ is given by Lemma S1 in the online supplement. An adaption to the result in equation (4.25) by the Glivenko-Cantelli property is straightforward; see Theorem 4.3 in Ding and Nan (2011).

REMARK 4.2. *Our asymptotic analysis also produces a uniform consistent estimator for the hazard function λ_0 with a cubic-root rate (modulo some logarithm term):*

$$\| \hat{\lambda}_n(t, \hat{\beta}_n) - \lambda_0(t) \|_2 = O_p(n^{-1/3}\log n). \tag{4.26}$$

Given the natural bound via the triangular inequality:

$$\| \hat{\lambda}_n(t, \hat{\beta}_n) - \lambda_0(t) \|_2 \leq \| \hat{\lambda}_n(t, \hat{\beta}_n) - \lambda_0(t, \hat{\beta}_n) \|_2 + \| \lambda_0(t, \hat{\beta}_n) - \lambda_0(t) \|_2,$$

(4.26) follows from Lemma S8 in the online supplement and the root-n consistency of $\hat{\beta}_n$.

REMARK 4.3. *The foundation to incorporate the time-dependent covariates in AFT models is laid down by Robins and Tsiatis (1992) and further developed by Lin and Ying (1995) and Zeng and Lin (2007). In this setup, it is more convenient to work with the duration* T *without making the logarithmic transformation. Recall in the setting with time independent covariates* X*, the essence of the AFT model is that there exists some independent baseline duration* T_0 *such that* $T_0 = Te^{-X'\beta_0}$*. With time-dependent covariates* $X(t)$*, the following specification generalizes the AFT model:*

$$T_{0,i} = \int_0^{T_i} \exp(X_i'(s)\beta_0)ds, \quad i = 1, \cdots, n, \tag{4.27}$$

where $T_{0,i}$ *is independent of the covariates. Lin and Ying (1995) have derived the efficient estimating function extending equation (3.18), whereas Zeng and Lin (2007) formally develop a semiparametric efficient estimator for the AFT model with time-dependent covariates. It is interesting to extend our methodology to this scenario in future work.*

5. NUMERICAL RESULTS

5.1. Monte Carlo Simulations

We conduct Monte Carlo simulations to evaluate the finite sample performances of our estimators. The logarithm of the duration $Y = \log T$ is generated from the model: $Y = 2 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where $\beta_1 = \beta_2 = 1$. The covariate X_1 is Bernoulli with a success probability 0.5, and the covariate X_2 is normal with mean zero and standard deviation 0.5. X_1 and X_2 are independent. We consider two types of log-concave errors: (I). The error term ε is the logarithm of a Weibull-distributed variable. The density of ε is $f(u) = \gamma e^{\gamma u} \exp(-e^{\gamma u})$, where $\gamma = 1.2$. In this case, the duration T conditional on (X_1, X_2) has a Weibull distribution and its hazard function is increasing. (II). The error term ε is the logarithm of a Gamma-distributed variable. The density of ε is $f(u) = \exp(ku - e^u)/\Gamma(k)$, where $k = 0.7$. In this case, the duration T conditional on (X_1, X_2) has a Gamma distribution and its hazard function is decreasing. Recall that $V = \min(Y, C)$ and $\Delta = 1\{Y \leq C\}$. The censoring time C is drawn from a uniform distribution supported on $[0, \tau]$, where τ is chosen to produce a 25% censoring rate. The data in each replication is an i.i.d. sample of $\{V_i, \Delta_i, X_{1i}, X_{2i}\}_{i=1}^n$ for $n = 500$ and 1000. We obtain semi-parametric maximum likelihood estimates of $(\beta_1, \beta_2)'$ using an unsmoothed hazard rate (SMLE-u) and a smoothed hazard rate (SMLE-s), by repetitively solving the estimating equations (3.18) and (3.21). We use the quartic kernel function $K(u) = \frac{15}{16}\left(1 - u^2\right)^2 1\{|u| \leq 1\}$, which satisfies Assumption 4.6. The bandwidths for smoothing the estimated hazard rate and its derivative are set to $h_c \times n^{-1/7}$ or $h_c \times n^{-1/7}$, respectively. The constant $h_c = 2$ and 5. Regarding the right-tail trimming factor τ_U in Assumption 4.5, we trim at the 98% quantile of $Y - \beta_1 X_1 - \beta_2 X_2$ for each (β_1, β_2). As comparison, we also include the Gehan-weight estimator (Jin et al., 2003), which is inefficient, and the sieve ML estimator (Ding and Nan, 2011, using cubic splines and two interior knots). In our simulations, SMLE methods are computationally much faster than the sieve estimator. On a 3.0 GHz Intel Core i5 processor and with 10 GB of RAM, SMLE methods on average take about 25 seconds (elapsed time) while the sieve estimator takes

about 300 seconds for each replication with sample size equal to 500. This is mainly because SMLE methods solve two-dimensional parameter β from the estimating equations, whereas the sieve estimator has to solve both β and additional six spline coefficients.

Table 1 presents the bias, empirical standard error (SE) and square root of the mean squared error (RMSE) for SMLE-u, SMLE-s, Genhan and sieve estimators. Results are based on 1000 simulated samples. We make the following observations when comparing these estimators. First, the performances of the proposed SMLE methods are stable across two sets of bandwidths. SMLE-s using the smoothed hazard rate yields smaller empirical standard errors than its unsmoothed version SMLE-u in all scenarios. Second, SMLE methods, especially SMLE-s, turn out to be more efficient (having smaller empirical standard error) than Gehan-weighted estimator in all scenarios. This substantiates the efficiency gain of SMLE methods over the rank estimator using inefficient weights. Third, the standard errors of SMLE-s are similar to those of the efficient sieve estimator in all scenarios. Fourth, the SMLE methods and the Gehan-weighted estimator yield smaller bias for β_1 than the sieve estimator. Bias performances are comparable for β_2. In sum, our SMLE methods achieve the same accuracy as the sieve ML estimator and enjoy two advantages: the guarantee of a unique global solution and faster computation.

5.2. Real Data Example

We re-visit the 4-month panel of the Current Population Survey (CPS) from September to December 1993, previously studied by Romeo (1999). We analyze the unemployment duration using the AFT model and focus on the sample with a positive employment duration of less than a year. The sample size is 399 and the censoring rate is 42%. We regress the natural logarithm of the unemployment duration (in weeks) on three covariates: race, gender, and a re-entrant indicator that indicates whether the individual is a labor force re-entrant. Table 2 reports coefficient estimates and standard errors. All methods suggest that on average, white, female, and re-entrants to the labor market have shorter unemployment durations. The negative sign of the coefficient on the re-entrant indicator suggests that people not in the labor force are more likely to return when they predict that there is a good chance of finding a job (Romeo, 1999). The magnitudes of estimates do not vary much across different methods. The proposed SMLE methods give smaller standard errors than other methods.

Figure 3 plots the estimated hazard of error term ε and the baseline hazard for the unemployment duration T_0. The left panel displays an increasing hazard function of ε, as a result of our log-concavity restriction. The right panel shows that the estimated baseline unemployment hazard function roughly takes a U-shape. The initial negative duration dependence can be a result of employers perceiving unemployment duration as a signal about the potential productivity of the worker, because the unemployed may lose valuable skills. On the other hand, the later positive duration dependence can be explained by falling reservation wage and/or increasing search intensity (Burdett and Vishwanath, 1988; Van den Berg, 1990). Such U-shaped duration dependence was also documented in empirical literature (Christofides and McKenna, 1996; McCall, 1996; Addison and Portugal, 2003).

Table 1. Finite sample performances of the SMLEs, sieve estimator, and Gehan-weight estimator.

				SMLE-u		SMLE-s		Gehan	Sieve
	n			$h_c = 2$	$h_c = 5$	$h_c = 2$	$h_c = 5$		
(I)	500	β_1	Bias	-.0007	-.0017	-.0007	-.0010	-.0041	-.0213
Weibull			SE	.0960	.0944	.0927	.0920	.1021	.0959
			RMSE	.0960	.0944	.0927	.0920	.1021	.0982
		β_2	Bias	.0075	.0055	.0053	.0050	.0056	.0006
			SE	.0970	.0954	.0931	.0936	.1043	.0914
			RMSE	.0973	.0956	.0932	.0937	.1044	.0914
	1000	β_1	Bias	.0021	.0035	.0018	.0030	.0016	-.0243
			SE	.0703	.0704	.0691	.0698	.0762	.0693
			RMSE	.0703	.0705	.0691	.0698	.0762	.0734
		β_2	Bias	.0049	.0030	.0041	.0030	.0067	.0050
			SE	.0696	.0698	.0689	.0691	.0752	.0672
			RMSE	.0698	.0698	.0690	.0692	.0755	.0674
(II)	500	β_1	Bias	-.0051	-.0020	-.0020	-.0010	.0036	.0181
Gamma			SE	.1421	.1368	.1368	.1337	.1551	.1366
			RMSE	.1422	.1368	.1369	.1338	.1551	.1378
		β_2	Bias	.0032	.0042	.0042	.0040	-.0014	.0133
			SE	.1428	.1418	.1396	.1395	.1577	.1388
			RMSE	.1429	.1419	.1396	.1396	.1577	.1394
	1000	β_1	Bias	.0017	.0012	.0012	.0012	.0014	-.0253
			SE	.0973	.0994	.0966	.0984	.1098	.1065
			RMSE	0974	.0994	.0966	.0984	.1098	.1094
		β_2	Bias	-.0003	-.0009	-.0004	-.0018	.0014	.0031
			SE	.0994	.1014	.0973	.0999	.1138	.1053
			RMSE	.0994	.1014	.0973	.0999	.1138	.1054

Note: SMLE-u and SMLE-s use the unsmoothed and smoothed hazard rates, respectively. Scenario (I): The error term is the logarithm of a Weibull-distributed variate; Scenario (II): The error term is the logarithm of a Gamma-distributed variate. The bandwidths for smoothing the hazard rate are $h_c \times n^{-1/5}$ (only used for SMLE-s) and $h_c \times n^{-1/7}$ for its derivative (used for SMLE-u and SMLE-s).

Table 2. AFT model estimates for unemployment duration.

Covariates	SMLE-u		SMLE-s		Gehan		Sieve	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Race	-.296	.099	-.211	.098	-.280	.185	-.366	.147
Gender	-.281	.098	-.164	.093	-.154	.174	-.259	.136
Re-entrant	-.274	.096	-.313	.096	-.259	.188	-.280	.148

Note: The bandwidths for smoothing the estimated hazard rate is $1 \times n^{-1/5}$ and for smoothing its derivative is $1 \times n^{-1/7}$.

6. CONCLUSION

In this paper, we study the AFT model under the constraint that the error term has a log-concave survivor function. Exploiting this shape restriction, we proposed two semi-

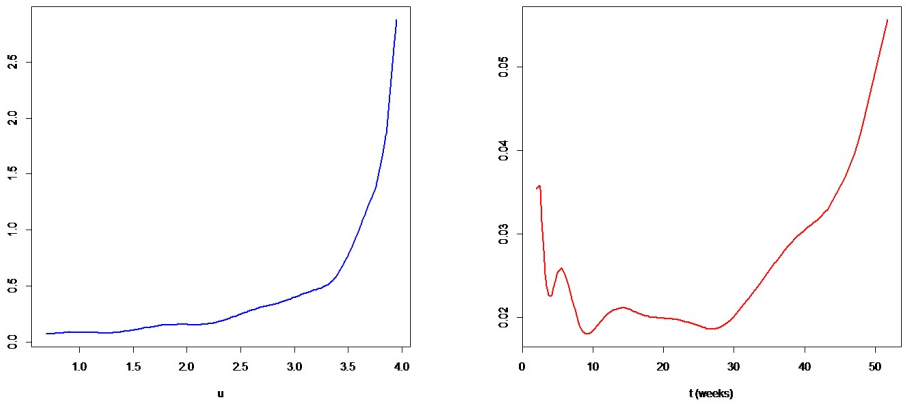

Figure 3. Estimated hazard functions of the error term ε (left) and the baseline duration T_0 (right).

parametric MLE estimators for the model coefficients. SMLE-u directly uses the NPMLE of the hazard function, which is piece-wise constant, and SMLE-s delivers smoothed estimates by resorting to the kernel smoothed version of the NPMLE. Both estimators obtain the semiparametric efficiency bound asymptotically. Simulation exercises show that SMLE-s has better finite sample performances than SMLE-u, even though the former introduces an additional bandwidth. Therefore, we recommend the SMLE-s to applied researchers. Our work makes the AFT model a more viable tool for analyzing duration data.

ACKNOWLEDGMENT

We thank two anonymous referees for their valuable comments. Ruixuan Liu acknowledges the financial support from the Program to Enhance Research and Scholarship (PERS) at Emory University. Zhengfei Yu acknowledges the support of JSPS KAKENHI Grant Number 19K13666.

REFERENCES

Abbring, J. H. (2012). Mixed hitting-time models. *Econometrica 80*(2), 783–819.

Addison, J. T. and P. Portugal (2003). Unemployment duration competing and defective risks. *Journal of Human Resources 38*(1), 156–191.

Bagnoli, M. and T. Bergstrom (2005). Log-concave probability and its applications. *Economic theory 26*(2), 445–469.

Barlow, R. and A. Marshall (1964). Bounds for distributions with monotone hazard rate I. *The Annals of Mathematical Statistics 35*(3), 1234–1257.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press.

Burdett, K. and T. Vishwanath (1988). Declining reservation wages and learning. *The Review of Economic Studies 55*(4), 656–665.

Chen, h. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6*, 5549–5632.

Chetverikov, D., A. Santos, and A. M. Shaikh (2018). The econometrics of shape restrictions. *Annual Review of Economics 10*, 31–63.

Christofides, L. N. and C. J. McKenna (1996). Unemployment insurance and job duration in Canada. *Journal of Labor Economics 14*, 286–312.

Cox, D. (1972). Regression models and life-tables. *Journal of Royal Statistical Society, Series B. 24*, 187–220.

Ding, Y. and B. Nan (2011). A sieve m-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *The Annals of Statistics 39*, 3032–3061.

Esary, J. and A. Marshall (1973). Shock models and wear processes. *The Annals of Probability 1*, 627–649.

Fygenson, M. and Y. A. Ritov (1994). Monotone estimating equations for censored data. *The Annals of Statistics 22*, 732–746.

Groeneboom, P. and K. Hendrickx (2018). Current status linear regression. *The Annals of Statistics 46*(4), 1415–1444.

Groeneboom, P. and G. Jongbloed (2014). *Nonparametric estimation under shape constraints*. Cambridge University Press.

Groeneboom, P., G. Jongbloed, and B. I. Witte (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *The Annals of Statistics 38*, 352–387.

Huang, J. and J. A. Wellner (1995). Estimation of a monotone density or monotone hazard under random censoring. *Scandinavian Journal of Statistics*, 3–33.

Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics 159*, 252–266.

Jin, Z., D. Lin, L. Wei, and Z. Ying (2003). Rank-based inference for the accelerated failure time model. *Biometrika 90*(2), 341–353.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The statistical analysis of failure time data*. John Wiley & Sons.

Kim, C. and T. Lai (2000). Efficient score estimation and adaptive m-estimators in censored and truncated regression models. *Statistica Sinica 10*, 731–749.

Kuchibhotla, A. K., R. K. Patra, and B. Sen (2017). Efficient estimation in convex single index models. *working paper*.

Lai, T. L. and Z. Ying (1991). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics 19*, 531–556.

Lin, D. Y. and Z. Ying (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. *Journal of Statistical Planning and Inference 44*, 47–63.

Marshall, A. W. and I. Olkin (2007). *Life distributions*. Springer.

McCall, B. P. (1996). Unemployment insurance rules, joblessness, and part-time work. *Econometrica*, 647–682.

Nan, B., J. D. Kalbfleisch, and M. Yu (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *The Annals of Statistics 37*, 2351–2376.

Nan, B. and J. A. Wellner (2013). A general semiparametric Z-estimation approach for case-cohort studies. *Statistica Sinica 23*, 1155–1180.

Nolan, D. and D. Pollard (1987). U-processes: Rates of convergence. *The Annals of Statistics 15*, 780–799.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika 65*, 167–179.

Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science 9*, 439–455.
Ridder, G. and T. M. Woutersen (2003). The singularity of the information matrix of the mixed proportional hazard model. *Econometrica 71*(5), 1579–1589.
Ritov, Y. (1990). Estimating in a linear regression model with censored data. *The Annals of Statistics 18*, 303–328.
Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order restricted statistical inference*. John Wiley & Sons.
Robins, J. and A. A. Tsiatis (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika 79*, 311–319.
Romeo, C. (1999). Conducting inference in semiparametric duration models under inequality restrictions on the shape of the hazard implied by job search theory. *Journal of Applied Econometrics 14*, 587–605.
Tripathi, G. (2000). Local semiparametric efficiency bounds under shape restrictions. *Econometric Theory 16*, 729–739.
Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics 18*, 354–372.
Van den Berg, G. J. (1990). Nonstationarity in job search theory. *The Review of Economic Studies 57*(2), 255–277.
Westling, T. and M. Carone (2019). A unified study of nonparametric inference for monotone functions. *Annals of Statistics*, forthcoming.
Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics 21*, 76–99.
Zeng, D. and D. Lin (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association 102*, 1387–1396.

APPENDIX: PROOFS OF MAIN RESULTS

The Appendix proves the asymptotic results in Section 4. The cited technical lemmas are collected in the online supplement. We first introduce some notations. Let $|\beta|$ denote the Euclidean norm for a vector β. For functional nuisance parameters ρ and η, we consider the semi-norm: $\| \rho \| \equiv \sup_{\beta \in \mathbf{B}} \| \rho(t,\beta) \|_2$ and $\| \eta \| \equiv \sup_{\beta \in \mathbf{B}} \| \eta(t,\beta) \|_2$ with $\| \cdot \|_2$ denoting the L_2 norm for the underlying function. Let $\mathcal{A}$ be the class of monotone functions with values in $[0, M]$ and $\mathcal{C}$ be the class of functions of bounded variation with values in $[0, M]$. We further define functional classes $\mathcal{F}_0$ and $\mathcal{F}_1$ as follows:

$$\mathcal{F}_0 \equiv \Big\{(y,x) \mapsto \mathbb{I}\{y - x'\beta \geq s\} : \beta \in \mathbf{B}, s \in \mathcal{T}\Big\},$$
$$\mathcal{F}_1 \equiv \Big\{(y,x) \mapsto x\mathbb{I}\{y - x'\beta \geq s\} : \beta \in \mathbf{B}, s \in \mathcal{T}\Big\}.$$

The corresponding convex hull of $\mathcal{F}_0$ and $\mathcal{F}_1$ are denoted as $\overline{conv}\mathcal{F}_0$ and $\overline{conv}\mathcal{F}_1$.

Proof of Theorem 4.1: We first show the following uniform convergence result

$$\| \Psi_n(\beta, \hat{\eta}_n, \hat{\rho}_n) - \Psi(\beta, \eta_0, \rho_0) \| \to_p 0. \tag{A.1}$$

Observe that

$$\| \Psi_n(\beta, \hat{\eta}_n, \hat{\rho}_n) - \Psi(\beta, \eta_0, \rho_0) \| \leq \| (\mathbb{P}_n - P)[\Delta \hat{\rho}_n(\epsilon_\beta, \beta)\{X - \hat{\eta}_n(\epsilon_\beta, \beta)\}] \| \tag{A.2}$$
$$+ \| P[\Delta(\hat{\rho}_n(\epsilon_\beta, \beta) - \rho_0(\epsilon_\beta, \beta))X] \| + \| P[\Delta(\hat{\rho}_n(\epsilon_\beta, \beta)\hat{\eta}_n(\epsilon_\beta, \beta) - \rho_0(\epsilon_\beta, \beta)\eta_0(\epsilon_\beta, \beta))] \| .$$

Note that for any β, the NPMLE estimator $\hat{\lambda}_n(u, \beta)$ is monotonically increasing and

$\dot{\hat{\lambda}}_n(u,\beta)$ is of bounded variation (see the proof of Lemma S6). Moreover, we prove in Lemma S6 the following functional class

$$\mathcal{G} \equiv \left\{(v,x,\delta) \mapsto \delta \frac{\dot{\lambda}(v-x'\beta,\beta)}{\lambda(v-x'\beta,\beta)}\left(x-\eta(v-x'\beta,\beta)\right) : \beta \in \mathbf{B}, \eta \in \bar{\mathcal{F}}, \lambda \in \mathcal{A}, \dot{\lambda} \in \mathcal{C}\right\}$$

is Glivenko-Cantelli, given that $\hat{\eta}_n \in \bar{\mathcal{F}} \equiv \overline{conv}\mathcal{F}_1/\overline{conv}\mathcal{F}_0$, $\hat{\lambda}_n \in \mathcal{A}$, and $\dot{\hat{\lambda}}_n \in \mathcal{C}$. Therefore, the first term on the right hand side of (A.2) converges to zero in outer probability. By Lemmas S8 and S12, we have $\parallel \hat{\rho}_n(\epsilon_\beta,\beta) - \rho_0(\epsilon_\beta,\beta) \parallel \to_p 0$. Hence the second term on the right hand side of (A.2) is bounded by

$$\parallel \hat{\rho}_n(\epsilon_\beta,\beta) - \rho_0(\epsilon_\beta,\beta) \parallel P|\Delta X| \to_p 0.$$

Lemma S7 implies that $\parallel \hat{\eta}_n(\epsilon_\beta,\beta) - \eta_0(\epsilon_\beta,\beta) \parallel \to_p 0$. Then the third term on the right hand side of (A.2) is bounded by

$$\begin{aligned}&\parallel \hat{\rho}_n(\epsilon_\beta,\beta)\hat{\eta}_n(\epsilon_\beta,\beta) - \rho_0(\epsilon_\beta,\beta)\eta_0(\epsilon_\beta,\beta) \parallel P|\Delta| \leq \parallel \hat{\rho}_n(\epsilon_\beta,\beta) - \rho_0(\epsilon_\beta,\beta) \parallel\parallel \hat{\eta}_n(\epsilon_\beta,\beta) \parallel P|\Delta| \\ &+ \parallel \hat{\eta}_n(\epsilon_\beta,\beta) - \eta_0(\epsilon_\beta,\beta) \parallel\parallel \rho_0(\epsilon_\beta,\beta) \parallel P|\Delta| \to_p 0.\end{aligned}$$

We now prove the consistency of $\hat{\beta}_n$. By Lemma 4.1, β_0 is the unique solution of $\Psi(\beta,\eta_0,\rho_0) = 0$. This implies that for any $\varepsilon > 0$, there is a $\delta > 0$ such that

$$\Pr\left[\mid \hat{\beta}_n - \beta_0 \mid > \varepsilon\right] \leq \Pr\left[\left|\Psi_n\left(\hat{\beta}_n, \eta_0(\cdot,\hat{\beta}_n), \rho_0(\cdot,\hat{\beta}_n)\right)\right| > \delta\right].$$

Observe that

$$\begin{aligned}&\left|\Psi_n\left(\hat{\beta}_n, \eta_0(\cdot,\hat{\beta}_n), \rho_0(\cdot,\hat{\beta}_n)\right)\right| \leq \left|\Psi_n\left(\hat{\beta}_n, \hat{\eta}_n(\cdot,\hat{\beta}_n), \hat{\rho}_n(\cdot,\hat{\beta}_n)\right)\right| \\ &+ \left|\Psi_n\left(\hat{\beta}_n, \eta_0(\cdot,\hat{\beta}_n), \rho_0(\cdot,\hat{\beta}_n)\right) - \Psi_n\left(\hat{\beta}_n, \hat{\eta}_n(\cdot,\hat{\beta}_n), \hat{\rho}_n(\cdot,\hat{\beta}_n)\right)\right| = o_p(1),\end{aligned}$$

where the last equality follows from the definition of $\hat{\beta}_n$ and (A.1). Hence $\hat{\beta}_n \to_p \beta_0$.

The proof of $\tilde{\beta}_n \to_p \beta_0$ is almost the same, with the only deviation of verifying $\parallel \tilde{\rho}_n(\epsilon_\beta,\beta) - \rho_0(\epsilon_\beta,\beta) \parallel \to_p 0$, where $\tilde{\rho}_n(\epsilon_\beta,\beta)$ is defined in (3.22). This immediately follows from Lemma S12. The proof of $\hat{\beta}_n$'s existence follows closely from Theorem 4.1 in Groeneboom and Hendrickx (2018) and is provided in S1.2 of the online supplement. □

Proof of Theorem 4.2: We present a detailed proof for $\hat{\beta}_n$ and highlight the necessary changes for $\tilde{\beta}_n$. First define $\Psi_n(\hat{\beta}_n, \hat{\eta}_n(\cdot,\hat{\beta}_n), \hat{\rho}_n(\cdot,\hat{\beta}_n)) = 0$ following the proof of Theorem 4.1 in Groeneboom and Hendrickx (2018): this can be done by taking convex combinations of Ψ_n evaluated at the left and right limit of $\hat{\beta}_n$, given the zero-crossing nature of $\hat{\beta}_n$. We then verify that the functional estimates $\hat{\eta}$ and $\hat{\rho}$ converge fast enough as

$$\parallel \hat{\eta}_n - \eta_0 \parallel = O_p(n^{-1/2}), \quad \parallel \hat{\rho}_n - \rho_0 \parallel = O_p(\log^2 n \times n^{-2/7}), \tag{A.3}$$

by using Lemmas S7 and S12. Note that both rates are faster than the critical rate $n^{-1/4}$ in Nan and Wellner (2013). By Lemma S6, the following stochastic equicontinuity result holds:

$$\mathbb{G}_n\left[\Delta\Big(\hat{\rho}_n(\epsilon_{\hat{\beta}_n},\hat{\beta}_n)\{X - \hat{\eta}_n(\epsilon_{\hat{\beta}_n},\hat{\beta}_n)\} - \rho_0(\epsilon_0,\beta_0)\{X - \eta_0(\epsilon_0,\beta_0)\}\Big)\right] = o_p(1). \tag{A.4}$$

Since $\Psi(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0)) = 0$, we get

$$\begin{aligned}
&n^{1/2}\left[\Psi(\hat\beta_n, \hat\eta_n(\cdot,\hat\beta_n), \hat\rho_n(\cdot,\hat\beta_n) - \Psi(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0))\right] \\
&= n^{1/2}(\Psi - \Psi_n)(\hat\beta_n, \hat\eta_n(\cdot,\hat\beta_n), \hat\rho_n(\cdot,\hat\beta_n)) + o_p(1) \\
&= -n^{1/2}(\Psi_n - \Psi)(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0)) + o_p(1),
\end{aligned}$$

where we make use of (A.4) in the second equality. Then we take the second-order Taylor expansion of the left hand side of the first equality, which leads to

$$\begin{aligned}
&-\dot\Psi_\beta(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0))n^{1/2}(\hat\beta_n - \beta_0) \\
&= n^{1/2}\left\{(\Psi_n - \Psi)(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0)) + \dot\Psi_\eta(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0))[(\hat\eta_n - \eta_0)(\cdot,\beta_0)]\right\} + o_p(1).
\end{aligned} \tag{A.5}$$

The second order terms are negligible due to (A.3). The first order effect of estimation for ρ_0 is canceled:

$$\dot\Psi_\rho(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0))[(\hat\rho_n - \rho_0)(\cdot,\beta_0)] = 0. \tag{A.6}$$

To prove this assertion, note that for any function $g(\epsilon_0, \beta_0)$,

$$P\left\{\int 1(\epsilon_0 \geq s)[X - \eta_0(\epsilon_0,\beta_0)]\dot\rho_{0\beta}(\epsilon_0,\beta_0)g(\epsilon_0,\beta_0)d\Lambda_0(s)\right\} = 0, \tag{A.7}$$

because the conditional expectation of the term inside the brackets is zero by first conditioning on $1\{V - X'\beta_0 \geq t\}$. Moreover, the following local martingale has a zero mean:

$$P\left\{\int [X - \eta_0(\epsilon_0,\beta_0)]\dot\rho_{0\beta}(\epsilon_0,\beta_0)g(\epsilon_0,\beta_0)dM(s)\right\} = 0, \tag{A.8}$$

where $M(t)$ is the martingale associated with the counting process:

$$M(t) = \Delta 1(V - X'\beta_0 \leq t) - \int_{-\infty}^{t} 1(V - X'\beta_0 \geq s)d\Lambda_0(s). \tag{A.9}$$

Therefore, the functional derivative with respect to ρ is equal to zero by summing up (A.7) and (A.8) and letting $g(\cdot,\beta_0) = [(\hat\rho_n - \rho_0)(\cdot,\beta_0)]$:

$$P\{[X - \eta_0(\epsilon_0,\beta_0)]\dot\rho_{0\beta}(\epsilon_0,\beta_0)\Delta[(\hat\rho_n - \rho_0)(\cdot,\beta_0)]\} = 0.$$

Further simplifying (A.5) leads to

$$\begin{aligned}
&n^{1/2}\left\{(\Psi_n - \Psi)(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0)) + \dot\Psi_\eta(\beta_0, \eta_0(\cdot,\beta_0), \rho_0(\cdot,\beta_0))[(\hat\eta_n - \eta_0)(\cdot,\beta_0)]\right\} \\
&\quad = \mathbb{G}_n \int \rho_0(t,\beta_0)\{X - \eta_0(t,\beta_0)\}dM(t) + o_p(1),
\end{aligned}$$

along the same line of Nan et al. (2009). The asymptotic covariance matrix of $\mathbb{G}_n \int \rho_0(t,\beta_0)\{X - \eta_0(t,\beta_0)\}dM(t)$ coincides with the efficient information matrix given by Lemma S1. Regarding the asymptotic analysis of $\tilde\beta_n$, the only change occurs to the convergence rate of the smoothed hazard function as

$$\| \tilde\lambda_n(t,\beta) - \lambda_0(t,\beta) \| = O_p(n^{-2/5}\log^2 n). \tag{A.10}$$

Nevertheless, this does not alter the rate of $\tilde\rho_n$ as the convergence rate of $\dot{\tilde\lambda}_n$ remains the slower and determining one. □

Online supplement of "Accelerated Failure Time Models with Log-concave Errors"

RUIXUAN LIU[†] AND ZHENGFEI YU[‡]

[†]*Department of Economics, Emory University, 201 Dowman Drive, Atlanta, GA, 30322*
E-mail: ruixuan.liu@emory.edu

[‡]*Faculty of Humanities and Social Sciences, University of Tsukuba, 1-1-1, Tennodai, Tsukuba, Ibaraki, Japan.*
E-mail: yu.zhengfei.gn@u.tsukuba.ac.jp

S1. PROOFS AND TECHNICAL LEMMAS

We first introduce some notations. Let $|\beta|$ denote the Euclidean norm for a vector β. For the functional nuisance parameters ρ and η, we consider the following semi-norm: $\| \rho \| \equiv \sup_{\beta \in \mathbf{B}} \| \rho(t,\beta) \|_2$ and $\| \eta \| \equiv \sup_{\beta \in \mathbf{B}} \| \eta(t,\beta) \|_2$ with $\| \cdot \|_2$ denoting the L_2 norm for the underlying function. We denote a large positive constant by M whose value might change line by line. We add subscripts for M if there is potential confusion. For two sequences a_n, b_n, $a_n \lesssim b_n$, if $a_n \leq b_n$ for some large M independent of n. Moreover, we denote $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ simultaneously. For any set E, $linE$ denotes the linear subspace generated by E and $\overline{E}$ stands for the L_2 closure of E.

S1.1. Log-concavity

Proof of Proposition 2.1: (i). See Theorems 1 and 3 of Bagnoli and Bergstrom (2005).
(ii). See Corollary 2 of Bagnoli and Bergstrom (2005).
(iii). Let $f_{T_0}(t)$ and $\bar{F}_{T_0}(t)$ be the density and the survival function of the baseline duration T_0, respectively. Since $\varepsilon = \log(T_0)$, one immediately obtains

$$\lambda(u) = \frac{f(u)}{\bar{F}(u)} = \frac{f_{T_0}(\exp(u))\exp(u)}{\bar{F}_{T_0}(\exp(u))} = h(\exp(u))\exp(u),$$

which is increasing in u as both $h(\cdot)$ and $\exp(u)$ are increasing functions. □

Proof of Proposition 2.2: It is straightforward to observe that the duration defined by the shock process can also be written as $T = T_0 \exp(X'\beta_0)$, where the baseline duration T_0 is defined as the survival time after a random number of shocks and the arrival of shocks is governed by a homogeneous Poisson process $N(t)$:

$$\Pr\{T_0 > t\} = \sum_{k=0}^{\infty} \frac{(t)^k e^{-t}}{k!} \overline{P}_k.$$

The conditional distribution of T is

$$\Pr\{T > t | X = x\} = \Pr\left\{T_0 > e^{-x'\beta} t | X = x\right\} = \sum_{k=0}^{\infty} \frac{\left(e^{-x'\beta} t\right)^k e^{-e^{-x'\beta} t}}{k!} \overline{P}_k.$$

The covariate effect $\exp(x'\beta)$ enters the model by changing the arrival rate of the null Poisson process $N(t)$. This feature is in accordance with the influence of covariates

on the conditional hazard function of the duration T in AFT models, as its role is to accelerate or decelerate the time to failure. According to Theorem 3.1 in Esary and Marshall (1973), T_0 has an increasing hazard function if $\overline{P}_k/\overline{P}_{k-1}$ is decreasing in k. The requirement stated in the proposition imply such restriction on the sequence $\bar{P}_k$ for Examples 2.2 and 2.3; see Theorem 5.5 and Theorem 6.1 in Esary and Marshall (1973). □

S1.2. Global Unique Solution

Proof of Lemma 4.1: For notation simplicity, we use $\Psi(\beta_0, \eta_0, \rho_0)$ to denote $\Psi(\beta, \eta_0(\cdot, \beta), \rho_0(\cdot, \beta))$ whenever no confusion is caused.

It suffices to show that $\Psi(\beta_0, \eta_0, \rho_0) = 0$ and $(\beta - \beta_0)'\Psi(\beta, \eta_0, \rho_0) > 0$ for all $\beta \in \mathbf{B} \setminus \{\beta_0\}$. Let $\bar{G}_x(\cdot) \equiv 1 - G_{C|X}(\cdot|X = x)$, where $G_{C|X}$ is the CDF of C conditional on X, and let $H(\cdot)$ be the CDF of X. Using equation (5.4) of Ying (1993), $\Psi(\beta, \eta_0, \rho_0)$ can be written as

$$\int \frac{\dot{\lambda}(t,\beta)}{\lambda(t,\beta)} \left[\int \bar{G}_x(t + \beta' x) f(t + (\beta - \beta_0)' x) dH(x) \right] q_0(t, \beta) dt,$$

where

$$\begin{aligned} q_0(t,\beta) \equiv & \frac{\int x \bar{G}_x(t + \beta' x) f(t + (\beta - \beta_0)' x) dH(x)}{\int \bar{G}_x(t + \beta' x) f(t + (\beta - \beta_0)' x) dH(x)} \\ & - \frac{\int x \bar{G}_x(t + \beta' x) \bar{F}(t + (\beta - \beta_0)' x) dH(x)}{\int \bar{G}_x(t + \beta' x) \bar{F}(t + (\beta - \beta_0)' x) dH(x)}. \end{aligned}$$

Obviously $q_0(t, \beta_0) = 0$. Hence $\Psi(\beta_0, \eta_0, \rho_0) = 0$. Then following equation (5.5) of Ying (1993), we obtain

$$(\beta - \beta_0)'\Psi(\beta, \eta_0, \rho_0) = \int \frac{\dot{\lambda}(t,\beta)}{\lambda(t,\beta)} \left[\int \bar{G}_x(t + \beta' x) f(t + (\beta - \beta_0)' x) dH(x) \right] q(t, \beta) dt,$$

where

$$\begin{aligned} q(t,\beta) \equiv & \frac{\int (\beta - \beta_0)' x \bar{G}_x(t + \beta' x) f(t + (\beta - \beta_0)' x) dH(x)}{\int \bar{G}_x(t + \beta' x) f(t + (\beta - \beta_0)' x) dH(x)} \\ & - \frac{\int (\beta - \beta_0)' x \bar{G}_x(t + \beta' x) \bar{F}(t + (\beta - \beta_0)' x) dH(x)}{\int \bar{G}_x(t + \beta' x) \bar{F}(t + (\beta - \beta_0)' x) dH(x)}. \end{aligned}$$

For $x \in \mathbb{R}^d$, let $R_1, R_2, ..., R_d$ be a sequence of orthogonal vectors satisfying $R_1 = \beta - \beta_0$. Define $\tilde{x}_\beta \equiv [\tilde{x}_{\beta,1}, \tilde{x}_{\beta,2}, ..., \tilde{x}_{\beta,d}]' \equiv R'x$, where $R \equiv [R_1, R_2, ..., R_d]$. For any $\beta \neq \beta_0$, we re-write $q(t, \beta)$ using $\tilde{x}_{\beta,1} = (\beta - \beta_0)'x$ and $x = (R')^{-1}\tilde{x}_\beta$.

$$q(t,\beta) = \int \tilde{x}_{\beta,1} q_1(t, \tilde{x}_{\beta,1}) d\tilde{H}(\tilde{x}_{\beta,1}) - \int \tilde{x}_{\beta,1} q_2(t, \tilde{x}_{\beta,1}) d\tilde{H}(\tilde{x}_{\beta,1}),$$

where

$$q_1(t,\tilde{x}_{\beta,1}) \equiv \frac{f(t+\tilde{x}_{\beta,1})}{\int f(t+\tilde{x}_{\beta,1})d\tilde{H}(\tilde{x}_{\beta,1})},$$
$$q_2(t,\tilde{x}_{\beta,1}) \equiv \frac{\bar{F}(t+\tilde{x}_{\beta,1})}{\int \bar{F}(t+\tilde{x}_{\beta,1})d\tilde{H}(\tilde{x}_{\beta,1})},$$
$$d\tilde{H}(\tilde{x}_{\beta,1}) \equiv \int_{\tilde{x}_{\beta,2},\ldots,\tilde{x}_{\beta,p}} \bar{G}_x(t+\beta'(R')^{-1}\tilde{x}_\beta)dH((R')^{-1}\tilde{x}_\beta).$$

Because the log-concavity of $F(\cdot)$ implies an increasing hazard function $f(\cdot)/\bar{F}(\cdot)$, the ratio of two density functions (with respect to $d\tilde{H}(\tilde{x}_{\beta,1})$) $q_1(t,\tilde{x}_{\beta,1})/q_2(t,\tilde{x}_{\beta,1})$ is increasing in $\tilde{x}_{\beta,1}$. Since the monotone likelihood ratio implies first order stochastic dominance (see Chapter 2 of Marshall and Olkin, 2007), it follows that $q(t,\beta) > 0$ for all $\beta \neq \beta_0$. Again by the log-concavity of $F(\cdot)$, $\dot{\lambda}(t,\beta) > 0$. Hence we conclude that $(\beta-\beta_0)'\Psi(\beta,\eta_0,\rho_0) > 0$ for all $\beta \in \mathbf{B} \setminus \{\beta_0\}$. □

Proof of the existence of zero-crossing points in Theorem 4.1 The uniform convergence in (A.1) of the Appendix leads to

$$\Psi_n(\beta,\hat{\eta}_n(\beta),\hat{\rho}_n(\beta)) = \dot{\Psi}_{\beta_0}(\beta-\beta_0) + r_n(\beta), \tag{S.1}$$

where $r_n(\beta) = o_p(1) + o(\beta-\beta_0)$. We now define for $h > 0$,the function

$$\Psi_{n,h}(\beta,\hat{\eta}_n(\beta),\hat{\rho}_n(\beta)) = \dot{\Psi}_{\beta_0}(\beta-\beta_0) + \tilde{r}_{n,h}(\beta), \tag{S.2}$$

with

$$\tilde{r}_{n,h}(\beta) = h^{-d}\int K_h(u_1-\beta_1)\cdots K_h(u_1-\beta_1)r_n(u_1,\cdots,u_d)du_1\cdots du_d, \tag{S.3}$$

where K is standard kernel density function supported on $[-1,1]$ and $\beta' = (\beta_1,\cdots,\beta_d)'$. Note that $\lim_{h\to 0}\tilde{r}_{n,h}(\beta) = r_n(\beta)$. We reparameterize by defining $\gamma = \dot{\Psi}_{\beta_0}\beta$ and $\gamma_0 = \dot{\Psi}_{\beta_0}\beta_0$. This gives

$$\Psi_{n,h}(\beta,\hat{\eta}_n(\beta),\hat{\rho}_n(\beta)) = \gamma - \gamma_0 + \tilde{r}_{n,h}(\dot{\Psi}_{\beta_0}^{-1}\gamma). \tag{S.4}$$

Given the result in (S.1), the mapping $\gamma \mapsto \gamma_0 - \tilde{r}_n(\dot{\Psi}_{\beta_0}^{-1}\gamma)$ maps, for each $\delta > 0$, the ball $B_\delta(\gamma_0) = \{\gamma : |\gamma-\gamma_0| \leq \delta\}$ into $B_{\delta/2}(\gamma_0) = \{\gamma : |\gamma-\gamma_0| \leq \delta/2\}$ with probability approaching to 1. Therefore by Brouwer's fixed point theorem (Groeneboom and Hendrickx (2018)), the mapping $\gamma \mapsto \gamma_0 - \tilde{r}_{n,h}(\dot{\Psi}_{\beta_0}^{-1}\gamma)$ has a fixed point which we denote by $\gamma_{n,h}$. Let $\beta_{n,h} \equiv \dot{\Psi}_{\beta_0}^{-1}\gamma_{n,h}$, then we have

$$\Psi_{n,h}(\beta_{n,h},\hat{\eta}_n(\beta_{n,h}),\hat{\rho}_n(\beta_{n,h})) = 0 \tag{S.5}$$

By compactness of $\mathbf{B}$, the sequence $(\beta_{n,1/k})_{k=1}^{\infty}$ must have a subsequence $(\beta_{n,1/k_l})$ with a limit point $\bar{\beta}_n$ as $l \to \infty$. Finally, we prove as in Groeneboom and Hendrickx (2018) that $\Psi_n(\beta,\hat{\eta}_n(\beta),\hat{\rho}_n(\beta))$ has a crossing of zero at $\bar{\beta}_n$ by contradiction.

Suppose that the j-th component Ψ_n^j of Ψ_n does not have a crossing of zero at $\bar{\beta}_n$. Then there must be an open ball $B_\delta(\bar{\beta}_n) = \{\beta : |\beta-\bar{\beta}_n| < \delta\}$ of $\bar{\beta}_n$ such that Ψ_n^j has a constant sign in $B_\delta(\bar{\beta}_n)$, say $\Psi_n^j(\beta,\hat{\eta}_n(\beta),\hat{\rho}_n(\beta)) \geq c > 0$ for all $\beta \in B_\delta(\bar{\beta}_n)$ and some constant $c > 0$. Arguing as Groeneboom and Hendrickx (2018), the j-th component of $\Psi_{n,h}^j$ of $\Psi_{n,h}$ satisfies

$$\Psi_{n,h}^j(\beta,\hat{\eta}_n(\beta),\hat{\rho}_n(\beta)) \geq \frac{c}{2}, \tag{S.6}$$

for sufficiently small h and all $\beta \in B_\delta(\bar{\beta}_n)$, which contradicting (S.5), since $\beta_{n,h}$ for $h = 1/k_l$ belongs to $B_\delta(\bar{\beta}_n)$ for large k_l. □

S1.3. Semi-parametric Efficiency

The calculation of semi-parametric information bound embedding shape restrictions is not a trivial task. Adapting the argument in Kuchibhotla et al. (2017), we formally verify that for the increasing hazard rate as assumed in our paper, the information bound remains unchanged as in Bickel et al. (1993). In order to determine the bound, the tangent set and the projection of the (parametric) score function to the tangent set need to be calculated. The score function is not affected by the shape restriction. The crux in our proof is to show the tangent set remains unchanged by showing that scores for smooth submodels lie in the set and by exhibiting a family of smooth submodels with scores that can approximate any element of the set arbitrarily well are dense in the set. Note that the calculation of the semiparametric information bound takes this trimming parameter τ_U as fixed; see Example 4 on page 284 of Bickel et al. (1993). Thus, the range of the (stochastic) integral $l^*_{\beta_0}$ in Lemma S1 is taken to be $\mathcal{T}$ throughout. We suppress $\mathcal{T}$ without confusion for notational simplicity.

Lemma S1. (Semiparametric efficiency bound) *In model (**??**) with right censorship, the efficient score function for estimating β_0 is*

$$l^*_{\beta_0}(V_i, X_i, \Delta_i) = \int (X_i - \mathbf{E}[X_i | Y_i - \beta_0' X_i \geq t]) \left(\frac{\dot{\lambda}_0(t)}{\lambda_0(t)} \right) dM(t), \tag{S.7}$$

where

$$M(t) = \Delta_i 1\{V_i - \beta_0' X_i \leq t\} - \int_{-\infty}^{t} 1\{V_i - \beta_0' X_i \geq s\} d\Lambda_0(s), \tag{S.8}$$

and $\Lambda_0(\cdot)$ is the cumulative hazard function of the error term ε.

Proof of Lemma S1: The complication in our setting is that the true hazard rate $\lambda_0(t)$ is increasing. Recall that we assume $\lambda_0 \geq c > 0$ over its support and its derivative is bounded and strictly positive, say $\dot{\lambda}_0 \geq c > 0$. We consider the following parametric sub-family that contains the true model:

$$\beta_{h_1} = \beta_0 + h_1 \alpha, \quad \log \lambda_{h_2}(t) = \log \lambda_0(t) + h_2 \theta(t), \tag{S.9}$$

where h_1 and h_2 are both scalars converging to zero and α is a $d-$dimensional vector as in Lai and Ying (1991). The local perturbation function $\theta(t)$ is a univariate function that is uniformly bounded and has uniformly bounded first-order derivative $\dot{\theta}(t)$. These requirements are made to guarantee the monotonicity of $\lambda_{h_2}(t)$ for small enough h_2. Therefore, the score functions are

$$l_{\beta_0} = X \int \frac{\dot{\lambda}_0(t)}{\lambda_0(t)} dM(t), \tag{S.10}$$

$$l_{\lambda_0}[\theta] = \int \theta(t) dM(t), \tag{S.11}$$

where $M(t)$ is the natural martingale in our Theorem 3.2. The efficient score function is $l^*_{\beta_0} = l_{\beta_0} - l_{\lambda_0}[\theta^*]$ for which $l_{\beta_0} - l_{\lambda_0}[\theta^*]$ is orthogonal to the nuisance tangent space denoted by Λ_S:

$$\Lambda_S \equiv \overline{lin}\{l_{\lambda_0}[\theta] : \lambda_{h_2}(t) \quad \text{is increasing for small enough} \quad h_2 \quad \text{and} \quad \theta \in \mathbb{L}_2(P)\} \tag{S.12}$$

Obviously the L_2 closure of the linear span of $l_{\lambda_0}[\theta]$ where the function θ has bounded first-order derivative (Bickel et al., 1993) is a subset of Λ_S. As shown in Section 4 of Kuchibhotla et al. (2017), we get

$$\overline{lin}\{l_{\lambda_0}[\theta] : \theta \quad \text{has a bounded first-order derivative}\} = \overline{lin}\{l_{\lambda_0}[\theta] : \theta \in \mathbb{L}_2(P)\}, \tag{S.13}$$

leading to the conclusion that Λ_S is also equal to $\overline{lin}\{l_{\lambda_0}[\theta] : \theta \in \mathbb{L}_2(P)\}$. The linear span generated by those nuisance functional components agree with each other and the monotonicity of the nonparametric component does not change the efficient score calculation; also see Tripathi (2000) for a similar setting of where the functional nuisance parameter is an increasing function in the partial linear model. Now the efficient score function is the one obtained by Ritov and Wellner (1988):

$$l^*_{\beta_0}(V_i, X_i, \Delta_i) = \int (X_i - \mathbf{E}[X_i | Y_i - \beta_0' X_i \geq t]) \left(\frac{\dot{\lambda}_0(t)}{\lambda_0(t)} \right) dM(t), \tag{S.14}$$

and the resulting information $I(\beta_0)$ immediately follows from Lemma S1:

$$\begin{aligned} I(\beta_0) &= \mathbf{E}[l^*_{\beta_0}(V_i, X_i, \Delta_i)^{\otimes 2}] \\ &= \mathbf{E}\left[X_i X_i' J(C_i - \beta_0' X_i)\right] \\ &\quad - \int \mathbf{E}[X_i | C_i - \beta_0' X_i \geq t] \mathbf{E}[X_i | C_i - \beta_0' X_i \geq t]' \mathbf{E}[1\{C_i - \beta_0' X_i \geq t\}] dJ(t), \end{aligned} \tag{S.15}$$

where

$$J(t) = \int_{-\infty}^{t} \left(\frac{\dot{\lambda}_0(s)}{\lambda_0(s)} \right)^2 dF_0(s). \tag{S.16}$$

□

S1.4. Empirical Processes

We first restate some necessary definitions and Theorem 2.4.1 in Van Der Vaart and Wellner (1996) that will be used repeatedly in the sequel. Let $\mathcal{F}$ be the class of functions and $L_2(Q)$ be the L_2-norm defined by a probability measure Q . For any probability measure Q, let $N(\varepsilon, \mathcal{F}, L_2(Q))$ be the minimal number of balls of radius ε needed to cover the class $\mathcal{F}$. The entropy integral $J(\delta, \mathcal{F})$ is defined as

$$J(\delta, \mathcal{F}) \equiv \sup_Q \int_0^\delta \sqrt{1 + \log N(\varepsilon, \mathcal{F}, L_2(Q))} d\varepsilon.$$

An envelope function of a functional class $\mathcal{F}$ is a function F such that $|f(x)| \leq F(x)$ for all x and $f \in \mathcal{F}$.

LEMMA S2. (THEOREM 2.14.1 IN VAN DER VAART AND WELLNER (1996)) *Let P_0 be the distribution of the underlying observation and let $\mathcal{F}$ be a P_0-measurable class with an envelope function F. We have*

$$E \sup_{f \in \mathcal{F}} |\mathbb{G}_n f| \lesssim J(1, \mathcal{F}) \parallel F \parallel_{P_0,2} \tag{S.17}$$

LEMMA S3. (THEOREM 2.7.5 IN VAN DER VAART AND WELLNER (1996)) *Let $\mathcal{A}$ be the class of monotone functions with values in $[0, M]$, then for all $\delta > 0$,*

$$J(\delta, \mathcal{A}) \lesssim \sqrt{\delta}. \tag{S.18}$$

Let $\mathcal{C}$ be the class of functions of bounded variation with values in $[0, M]$, then for all $\delta > 0$,

$$J(\delta, \mathcal{C}) \lesssim \sqrt{\delta}. \tag{S.19}$$

LEMMA S4. (ENTROPY FOR CONVEX HULL OF VC CLASSES) *Let $\mathcal{F}$ be a VC-class with VC index equal to v, and denote $\overline{conv}\mathcal{F}$ be the convex hull of $\mathcal{F}$, then*

$$J(\delta, \overline{conv}\mathcal{F}) \lesssim \delta^{\frac{2}{v+2}}. \tag{S.20}$$

In our context, the following two VC classes play important roles in analyzing $\hat{\eta}_n$:

$$\mathcal{F}_0 \equiv \Big\{(y, x) \mapsto \{1(y - x'\beta \geq s) : \beta \in \mathbf{B}, s \in \mathcal{T}\Big\}, \tag{S.21}$$

and

$$\mathcal{F}_1 \equiv \Big\{(y, x) \mapsto x\{1(y - x'\beta \geq s) : \beta \in \mathbf{B}, s \in \mathcal{T}\Big\}. \tag{S.22}$$

We denote the corresponding convex hull of two classes as $\overline{conv}\mathcal{F}_0$ and $\overline{conv}\mathcal{F}_1$, respectively. Note that $D_n^{(k)}(s, \beta)$ $(k = 0, 1)$ are in $\overline{conv}\mathcal{F}_0$ and $\overline{conv}\mathcal{F}_1$, see Nan and Wellner (2013).

The next lemma provides the entropy bound for an important functional class in our remaining proofs; see also Lemma 10.1 in Groeneboom and Hendrickx (2018).

LEMMA S5. *Consider the following function class:*

$$\mathcal{F}^2_{K\delta,0} = \{\lambda(x'\beta) : \sup|\lambda| \leq K_1, |\beta - \beta_0| \leq K_2\}, \tag{S.23}$$

where the function $\lambda(\cdot)$ belongs to the class of monotone functions, then the following entropy bound holds:

$$H_{[]}\left(\epsilon, \mathcal{F}^2_{K\delta,0}, \|\cdot\|_2\right) \leq \frac{MK_1}{\epsilon}, \tag{S.24}$$

for some finite constant M.

Proof of Lemma S5: For any small ϵ_β, the compact neighborhood of β_0 can be covered by N_β neighborhoods with diameters no larger than ϵ_β, where $N_\beta \leq M\epsilon_\beta^{-q}$. Thus, for any β, we can find $i \in \{1, \cdots, N_\beta\}$ such that $|\beta - \beta_i| \leq \epsilon$. For the monotone function λ, we can find brackets $[\lambda_j^L, \lambda_j^U]$ with size ϵ covering the class of monotone functions with range restricted to $[-K_1, K_1]$. Moreover, the number of brackets N_λ is bounded by $\exp(K_1\epsilon^{-1})$ up to some finite constant.

Consider any function $f(x)$ in $\mathcal{F}^2_{K\delta,0}$, one has

$$f(x) \equiv \lambda(x'\beta) = \lambda(x'\beta_i + x'(\beta - \beta_i)), \tag{S.25}$$

which leads to

$$\lambda(x'\beta_i - M\epsilon_\beta) \leq f(x) \leq \lambda(x'\beta_i + M\epsilon_\beta), \tag{S.26}$$

given that the covariates X have compact support. Therefore, we can cover the element in $\mathcal{F}^2_{K\delta,0}$ by

$$\lambda_j^L(x'\beta_i - M\epsilon_\beta) \leq f \leq \lambda_j^U(x'\beta_i + M\epsilon_\beta), \tag{S.27}$$

for a pair $[\lambda_j^L, \lambda_j^U]$ that covers λ.

Now we verify the size of new bracket $[\lambda_j^L(x'\beta_i - M\epsilon_\beta), \lambda_j^U(x'\beta_i + M\epsilon_\beta)]$ is less than ϵ up to some finite constant with a proper choice of ϵ_β. We start with the following decomposition:

$$\begin{aligned}\| \lambda_j^U(x'\beta_i + M\epsilon_\beta) - \lambda_j^L(x'\beta_i - M\epsilon_\beta) \|_2 \leq & \| \lambda_j^U(x'\beta_i + M\epsilon_\beta) - \lambda(x'\beta_i + M\epsilon_\beta) \|_2 \\ & + \| \lambda(x'\beta_i + M\epsilon_\beta) - \lambda(x'\beta_i - M\epsilon_\beta) \|_2 \\ & + \| \lambda(x'\beta_i - M\epsilon_\beta) - \lambda_j^L(x'\beta_i - M\epsilon_\beta) \|_2 .\end{aligned}$$

Apparently, the first and third terms are bounded up by ϵ by the construction of $[\lambda_j^L, \lambda_j^U]$. Considering the second term, one get

$$\| \lambda(x'\beta_i + M\epsilon_\beta) - \lambda(x'\beta_i - M\epsilon_\beta) \|_2^2 \leq M \int_{-2M}^{2M} (\lambda(t) - \lambda(t - 2M\epsilon_\beta))^2 \, dt$$

by the change of variable. Now given the monotonicity of λ and the fact that it is bounded in absolute value by K, we have

$$\begin{aligned}\int_{-2M}^{2M} (\lambda(t) - \lambda(t - 2M\epsilon_\beta))^2 \, dt \leq & M \int_{-2M}^{2M} (\lambda(t - 2M\epsilon_\beta) - \lambda(t)) \, dt \\ = & M \left[\int_{-2M-2\epsilon_\beta M}^{-2M} \lambda(t - 2M\epsilon) dt - \int_{2M-2\epsilon_\beta M}^{2M} \lambda(t) dt \right] \\ \lesssim & \epsilon_\beta.\end{aligned}$$

Then we take $\epsilon_\beta = \epsilon^2$, we get $\| \lambda(x'\beta_i + M\epsilon_\beta) - \lambda(x'\beta_i - M\epsilon_\beta) \|_2 \lesssim \epsilon$. Thus, the overall bracketing entropy number is bounded by:

$$H_{[]}\left(\epsilon, \mathcal{F}^2_{K\delta,0}, \|\cdot\|_2\right) \leq \log N_\beta + \log N_\lambda \leq 2q \log(\epsilon^{-1}) + \frac{MK_1}{\epsilon} \lesssim \frac{MK_1}{\epsilon}.$$

□

Now we obtain the entropy bounds for the key functional class in our context and prove the asymptotic characterizations for terms in showing consistency and asymptotic normality. Note that the uniform convergence result in Lemma S8 implies that the estimated hazard rate function is uniformly bounded away from zero, given that the true hazard rate function is uniformly bounded away from zero. Therefore, we can restrict our attention to bounded monotone functions for the classes involving $1/\lambda(\cdot, \beta)$.

LEMMA S6. *The functional class $\mathcal{G}$ defined by*

$$\mathcal{G} \equiv \left\{ (y, x, \delta) \mapsto \delta\rho(y - x'\beta)[x - \eta(y - x'\beta)] : \beta \in \mathbf{B}, \eta(t, \beta) \in \bar{\mathcal{F}}, \rho \in \mathcal{H} \right\} \tag{S.28}$$

has bounded entropy integral, where $\bar{\mathcal{F}} \equiv \overline{conv}\mathcal{F}_1/\overline{conv}\mathcal{F}_0$ *and* $\mathcal{H} \equiv \mathcal{C}/\mathcal{A}$*. Therefore, we have the following Glivenko-Cantelli result*

$$\mathbb{P}_n\left[\Delta\hat{\rho}_n(\epsilon_{\hat{\beta}_n},\hat{\beta}_n)\{X-\hat{\eta}_n(\epsilon_{\hat{\beta}_n},\hat{\beta}_n)\}\right]-P\left[\rho_0(\epsilon_0,\beta_0)\{X-\eta_0(\epsilon_0,\beta_0)\}\right]=o_p(1),$$

and the stochastic equicontinuity as

$$\mathbb{G}_n\left[\Delta\Big(\hat{\rho}_n(\epsilon_{\hat{\beta}_n},\hat{\beta}_n)\{X-\hat{\eta}_n(\epsilon_{\hat{\beta}_n},\hat{\beta}_n)\}-\rho_0(\epsilon_0,\beta_0)\{X-\eta_0(\epsilon_0,\beta_0)\}\Big)\right]=o_p(1).$$

Proof of Lemma S6: We first verify that the uniform entropy integral $J(1,\mathcal{G})$ is bounded. We consider the following three subclasses:

$$\mathcal{G}_1 \equiv \Big\{(y,x,\delta)\mapsto 1/\lambda(y-x'\beta,\beta):\beta\in\mathbf{B},\lambda(t,\beta)\in\mathcal{A}\Big\}, \tag{S.29}$$

$$\mathcal{G}_2 \equiv \Big\{(y,x,\delta)\mapsto \dot{\lambda}(y-x'\beta,\beta):\beta\in\mathbf{B},\dot{\lambda}(t,\beta)\in\mathcal{C}\Big\}, \tag{S.30}$$

$$\mathcal{G}_3 \equiv \Big\{(y,x,\delta)\mapsto \delta[x-\eta(y-x'\beta)]:\beta\in\mathbf{B},\eta(t,\beta)\in\bar{\mathcal{F}}\Big\}. \tag{S.31}$$

Because the NPMLE estimator $\hat{\lambda}_n(t,\beta)$ is an increasing function for any given β, $\mathcal{G}_1$ is the class involving monotonically decreasing functions. Hence, the uniform entropy integral $J(1,\mathcal{G}_1)$ is bounded. A similar argument applies to $\mathcal{G}_2$ because the kernel estimator of its density function is $\dot{\hat{\lambda}}_n(t,\beta)$ is

$$\dot{\hat{\lambda}}_n(t,\beta)=\int K_h(t-u)d\hat{\lambda}_n(u,\beta),$$

for a continuous kernel function K and a function of bounded variation $\hat{\lambda}_n$ (since it is increasing). Therefore, $\dot{\hat{\lambda}}_n(t,\beta)$ is of bounded variation by Theorem I.5.c in Widder (1941). The composition of $\hat{\lambda}_n$ or $\dot{\hat{\lambda}}_n$ with the linear index $y-x'\beta$ still has the bounded entropy integral as shown in our Lemma S5. Moreover, $D_n^{(k)}(t,\beta)$ and $D^{(k)}(t,\beta)$ are in the convex hull of certain VC classes for $k=0,1$. Thus, the resulting uniform entropy integral is bounded. Thus, the conclusion for the whole functional class $\mathcal{G}=\mathcal{G}_1\cdot\mathcal{G}_2\cdot\mathcal{G}_3$ follows from Example 2.10.8 Van Der Vaart and Wellner (1996).
Thereafter, we show

$$\parallel P[\delta\{\rho[X-\eta]-\rho_0[X-\eta_0]\}]\parallel\leq\parallel P[\delta X(\rho-\rho_0)]\parallel+\parallel P[\delta(\rho\eta-\rho_0\eta_0)]\parallel\to 0,$$

as $|\beta-\beta_0|\to 0$, $\parallel\eta-\eta_0\parallel\to 0$, and $\parallel\rho-\rho_0\parallel\to 0$.
Hence, by (S.17) we get the desired stochastic equicontinuity as in (**??**) of Appendix A in the paper. □

We then present the linear expansion of $\hat{\eta}_n$ in Lemma S7, which has been established by Nan and Wellner (2013). We refer readers to [page 1172] of their paper for the proof.

LEMMA S7. *Suppose that Assumptions 4.1 to 4.6 hold. Then we have the following expansion*

$$n^{1/2}\left[\hat{\eta}_n(t,\beta)-\eta_0(t,\beta)\right]=D^{(0)}(t,\beta)^{-1}\mathbb{G}_n\left[1(\epsilon_\beta\geq t)\{X-\eta_0(t,\beta)\}\right]+o_p(1). \tag{S.32}$$

Therefore, $\parallel\hat{\eta}_n-\eta_0\parallel=O_p(n^{-1/2})$.

S1.5. Convergence Rate of Estimators of the Hazard Function and Its Derivative

This section is targeted to show the rate of convergence of $\hat{\rho}_n$ as $\| \hat{\rho}_n - \rho_0 \| = O_p(\log^2 n \times n^{-2/7})$. Given the ratio form of $\hat{\rho}_n$, the crux is to determine the rate of $\dot{\hat{\lambda}}_n$. We introduce additional notations. Let $\hat{\lambda}_n$ be the NPMLE and $\{\tau_1, ..., \tau_m\}$ with $m \leq n$ be its jump points. Denote intervals $J_i \equiv [\tau_i, \tau_{i+1})$ for $0 \leq i \leq m-1$. For $u \in J_i$, define

$$\hat{A}_n(u,\beta) = \begin{cases} \tau_i, & \text{if} \quad \forall t \in J_i : \lambda_0(t,\beta) > \hat{\lambda}_n(\tau_i,\beta) \\ s, & \text{if} \quad \exists s \in J_i : \lambda_0(s,\beta) = \hat{\lambda}_n(s,\beta) \\ \tau_{i+1}, & \text{if} \quad \forall t \in J_i : \lambda_0(t,\beta) < \hat{\lambda}_n(\tau_i,\beta). \end{cases}$$

We first state three lemmas regarding the NPMLE of the monotone hazard function. Their proofs are in analog with the ones in Groeneboom et al. (2010) and are hence omitted. Lemma S8 concerning the properties of the NPMLE for the increasing hazard rate $\lambda_0(t,\beta)$. The proof is in analog with Equations (2.1), (A.20), and (A.21) in Groeneboom et al. (2010).

LEMMA S8. *Suppose that Assumptions 4.1 to 4.6 hold. Then we have*

$$\begin{aligned} \Pr\left(\lim_{n\to\infty} \| \hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta) \| = 0\right) = 1, \\ \| \hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta) \| = O_p\left(\log n \times n^{-1/3}\right), \\ \| \hat{A}_n(u,\beta) - u \| = O_p\left(\log n \times n^{-1/3}\right). \end{aligned} \tag{S.33}$$

The following Lemma is in analogy with Lemma 4.1 of Groeneboom et al. (2010) .

LEMMA S9. *Suppose that Assumptions 4.1 to 4.6 hold. For any given t and β, the following representations hold:*

$$\int \mathbb{K}_{\tilde{h}}(t-u) d(\hat{\lambda}_n - \lambda_0)(u,\beta) = -\int \psi_{\tilde{h},t}(\hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta)) du, \tag{S.34}$$

and

$$\int K_h(t-u) d(\hat{\lambda}_n - \lambda_0)(u,\beta) = -\int \phi_{h,t}(\hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta)) du, \tag{S.35}$$

where

$$\psi_{\tilde{h},t}(u) = K_{\tilde{h}}(t-u)/p(v,x,\delta), \quad \textit{and} \quad \phi_{h,t}(u) = k_h(t-u)/p(v,x,\delta), \tag{S.36}$$

with $k_h(\cdot) \equiv K'(\cdot/h)/h^2$ and

$$p(v,x,\delta) = f^{\delta}(v - x'\beta_0)\bar{G}_x^{\delta}(v)\bar{F}^{1-\delta}(v - x'\beta_0) g_x^{1-\delta}(v) dH(x), \tag{S.37}$$

is the joint density function of (V, X, Δ).

Consider the piece-wise constant version of $\phi_{h,t}$ which is constant on the same intervals where the NPMLE $\hat{\lambda}_n(\cdot,\beta)$ remains constant. For $u \in J_i$, define

$$\bar{\phi}_{h,t}(u) = \phi_{h,t}(\hat{A}_n(u,\beta)) \quad \text{and} \quad \bar{\psi}_{\tilde{h},t}(u) = \psi_{\tilde{h},t}(\hat{A}_n(u,\beta)). \tag{S.38}$$

The next lemma spells out the difference between the functions $\psi_{\tilde{h},t}$, $\phi_{h,t}$ and their piecewise constant approximations $\bar{\psi}_{\tilde{h},t}$, $\bar{\phi}_{h,t}$. This lemma is in analogy with Lemma A.4 of Groeneboom et al. (2010).

LEMMA S10. *Suppose that Assumptions 4.1 to 4.6 hold. Then we have*

$$|\psi_{\tilde{h},t}(u) - \bar{\psi}_{\tilde{h},t}(u)| \lesssim \frac{|\hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta)| 1\{|t-u| \le \tilde{h}\}}{\tilde{h}^2}, \tag{S.39}$$

and

$$|\phi_{h,t}(u) - \bar{\phi}_{h,t}(u)| \lesssim \frac{|\hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta)| 1\{|t-u| \le h\}}{h^3}. \tag{S.40}$$

An immediate consequence of Lemma S10 is the characterization of some negligible terms in Lemma S11, which in turn will be used in the proof of Lemma S12. Recall that $\mathcal{T}$ is the support of ε trimmed from the right tail at come given τ_U as in Tsiatis (1990).

LEMMA S11. *Suppose that Assumptions Assumptions 4.1 to 4.6 hold. Then we have*

$$\int (\psi_{\tilde{h},t} - \bar{\psi}_{\tilde{h},t}(u))(\hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta)) dP = o_p(\log^2 n \times n^{-2/5}), \tag{S.41}$$

and

$$\int (\phi_{h,t} - \bar{\phi}_{h,t}(u))(\hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta)) dP = o_p(\log^2 n \times n^{-2/7}), \tag{S.42}$$

uniformly over $t \in \mathcal{T}, \beta \in \mathbf{B}$.

Proof of Lemma S11: We only prove the second claim to avoid repetition. We apply (S.40) in Lemma S10 and the Cauchy-Schwarz inequality to obtain:

$$\begin{aligned}
&\int (\phi_{h,t} - \bar{\phi}_{h,t}(u))(\hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta)) dP \\
&\lesssim h^{-3} \int_{u[\in t-h,t+h]} (\hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta))^2 dP \\
&\lesssim h^{-3+1/2} \parallel \hat{\lambda}_n(u,\beta) - \lambda_0(u,\beta) \parallel^2 = O_p(n^{5/14-2/3} \log^2 n) = o_p(\log^2 n \times n^{-2/7}),
\end{aligned}$$

which leads to the desired result. □

Now we state and prove the key lemma concerning the rate of convergence of $\dot{\hat{\lambda}}_n$ and $\tilde{\lambda}_n$. The convolution type estimation in (3.16) and (3.20) (of the paper) appear similar to kernel density estimation. However, the asymptotic analysis is more complicated (Groeneboom et al., 2010), because $\hat{\lambda}_n$ is not linear, in contrast with the empirical distribution function.

LEMMA S12. *Suppose that Assumptions 4.1 to 4.6 hold. Then, for $h \asymp n^{-1/7}$,*

$$\parallel \dot{\hat{\lambda}}_n(t - x'\beta, \beta) - \dot{\lambda}_0(t - x'\beta, \beta) \parallel = O_p\left(n^{-2/7} \log^2 n\right). \tag{S.43}$$

Moreover, for $\tilde{h} \asymp n^{-1/5}$,

$$\parallel \tilde{\lambda}_n(t - x'\beta, \beta) - \lambda_0(t - x'\beta, \beta) \parallel = O_p\left(n^{-2/5} \log^2 n\right). \tag{S.44}$$

Proof of Lemma S12: We only prove (S.43) which is in fact the harder one. The deterministic bias term can be easily dealt with based on standard kernel smoothing arguments:

$$\sup_{\beta,t}\left|\int K_h(t-u)d\lambda_0(u,\beta)-\lambda_0(t,\beta)\right|=O(h^2)=O(n^{-2/7}).$$

When it comes to the stochastic part, we rely on the representation in (S.42) to obtain

$$\begin{aligned}&\int K_h(t-(y-x'\beta))d(\hat{\lambda}_n-\lambda_0)(y-x'\beta,\beta)=\int\phi_{h,t}[\hat{\lambda}_n(y-x'\beta,\beta)-\lambda_0(y-x'\beta,\beta)]dP\\&=\int\bar{\phi}_{h,t}(\hat{\lambda}_n(y-x'\beta,\beta)-\lambda_0(y-x'\beta,\beta))dP\\&+\int(\phi_{h,t}-\bar{\phi}_{h,t})(\hat{\lambda}_n(y-x'\beta,\beta)-\lambda_0(y-x'\beta,\beta))dP\equiv I_{1n}+I_{2n}.\end{aligned}$$

The second term I_{2n} in the last equality is of a negligible order $O_p(\log n\times n^{-2/3})$ as shown in Lemma S11.

Given the fact that $\hat{\lambda}_n(y-x'\beta,\beta)$ is the greatest convex majorant of the cumulative sum diagram and the function $\bar{\phi}_{h,t}$ is constant on the same intervals where the NPMLE remains constant, we get

$$\int\bar{\phi}_{h,t}\left[\hat{\lambda}_n(v-x'\beta,\beta)-\frac{1\{\epsilon_\beta\le v-x'\beta,\delta=1\}}{D_n^{(0)}(v-x'\beta,\beta)}\right]d\mathbb{P}_n(v,x,\delta)=0,\tag{S.45}$$

see equation (11.81) on [page 348] of Groeneboom and Jongbloed (2014). Therefore, we insert (S.45) into I_{1n}:

$$\begin{aligned}I_{1n}&=\int\bar{\phi}_{h,t}(\hat{\lambda}_n(y-x'\beta,\beta)-\lambda_0(y-x'\beta,\beta))dP\\&\quad-\int\bar{\phi}_{h,t}\left[\hat{\lambda}_n(y-x'\beta,\beta)-\frac{1\{\epsilon_\beta\le y-x'\beta,\delta=1\}}{D_n^{(0)}(y-x'\beta,\beta)}\right]d\mathbb{P}_n\\&=-\int\bar{\phi}_{h,t}\left(\hat{\lambda}_n(y-x'\beta,\beta)-\frac{1\{\epsilon_\beta\le y-x'\beta,\delta=1\}}{D^{(0)}(y-x'\beta,\beta)}\right)d(\mathbb{P}_n-P)\\&\quad+\int\bar{\phi}_{h,t}1\{\epsilon_\beta\le y-x'\beta,\delta=1\}\left[\frac{1}{D_n^{(0)}(y-x'\beta,\beta)}-\frac{1}{D^{(0)}(y-x'\beta,\beta)}\right]d\mathbb{P}_n\\&\equiv J_{1n}+J_{2n}.\end{aligned}\tag{S.46}$$

In the above decomposition, we have made use of the fact that

$$\lambda_0(y-x'\beta,\beta)d(y-x'\beta)=\frac{dN(y-x'\beta,\beta)}{D^{(0)}(y-x'\beta,\beta)},$$

which leads to

$$\lambda_0(y-x'\beta,\beta)dP=\frac{1\{\epsilon_\beta\le y-x'\beta,\delta=1\}}{D^{(0)}(y-x'\beta,\beta)}dP.$$

The analysis of J_{2n} proceeds as follows

$$\begin{aligned} J_{2n} &= \int \bar{\phi}_{h,t} 1\{\epsilon_\beta \le y - x'\beta, \delta = 1\} \left[\frac{1}{D_n^{(0)}(y - x'\beta, \beta)} - \frac{1}{D^{(0)}(y - x'\beta, \beta)} \right] d(\mathbb{P}_n - P) \\ &\quad + \int \bar{\phi}_{h,t} 1\{\epsilon_\beta \le y - x'\beta, \delta = 1\} \left[\frac{1}{D_n^{(0)}(y - x'\beta, \beta)} - \frac{1}{D^{(0)}(y - x'\beta, \beta)} \right] dP \\ &\equiv J_{2n}^{(a)} + J_{2n}^{(b)}. \end{aligned} \tag{S.47}$$

We prove in Lemma S14 that

$$J_{2n}^{(a)} = o_p(n^{-2/7} \log^2 n), \text{ and } J_{2n}^{(b)} = O_p(n^{-1/2} \times h^{-1}) = O_p(n^{-5/14}).$$

Referring to the leading term J_{1n}, we get

$$\begin{aligned} -J_{1n} &= \int \bar{\phi}_{h,t} \left(\lambda_0(y - x'\beta, \beta) - \frac{1\{\epsilon_\beta \le y - x'\beta, \delta = 1\}}{D^{(0)}(y - x'\beta, \beta)} \right) d(\mathbb{P}_n - P) \\ &\quad + \int \bar{\phi}_{h,t} \left(\hat{\lambda}_n(y - x'\beta, \beta) - \lambda_0(y - x'\beta, \beta) \right) d(\mathbb{P}_n - P). \end{aligned}$$

The second term is denoted by R_n term and we prove that it is of smaller order in Lemma S13. Finally, we utilize Lemma A.7 in Groeneboom et al. (2010) to conclude:

$$\begin{aligned} &\int \bar{\phi}_{h,t} \left(\lambda_0(y - x'\beta, \beta) - \frac{1\{\epsilon_\beta \le y - x'\beta, \delta = 1\}}{D^{(0)}(y - x'\beta, \beta)} \right) d(\mathbb{P}_n - P) \\ &= \int \phi_{h,t} \left(\lambda_0(y - x'\beta, \beta) - \frac{1\{\epsilon_\beta \le y - x'\beta, \delta = 1\}}{D^{(0)}(y - x'\beta, \beta)} \right) d(\mathbb{P}_n - P) \\ &+ \int \left(\bar{\phi}_{h,t} - \phi_{h,t} \right) \left(\lambda_0(y - x'\beta, \beta) - \frac{1\{\epsilon_\beta \le y - x'\beta, \delta = 1\}}{D^{(0)}(y - x'\beta, \beta)} \right) d(\mathbb{P}_n - P) \\ &= O_p\left(n^{-2/7} \log^2 n \right). \end{aligned}$$

The proof of (S.44) is very similar with some notation-wise difference. When we take $\tilde{h} \asymp n^{-1/5}$, those smaller order terms as R_n, $J_{2n}^{(a)}$ and $J_{2n}^{(b)}$ can be shown of order $o_p(\log^2 n \times n^{-2/5})$. Meanwhile, the leading term as in J_{1n} is of order $O_p(\log^2 n \times n^{-2/5})$, following Lemma A.7 in Groeneboom et al. (2010), mutatis mutandis. □

Compared with the point-wise result in Groeneboom et al. (2010), The result of Lemma S12 contains extra $\log^2 n$ terms in the rate because we also need the uniform convergence over $t \in \mathcal{T}$ for the argument of the kernel functions. Nevertheless, the obtained rates are fast enough, i.e., of $o_p(n^{-1/4})$. The following two lemmas characterize a few smaller order terms in the proof of Lemma S12.

LEMMA S13. *Suppose that Assumptions 4.1 to 4.6 hold. Then we have the following characterization*

$$R_n \equiv \int \bar{\phi}_{h,t} \left(\hat{\lambda}_n(y - x'\beta, \beta) - \lambda_0(y - x'\beta, \beta) \right) d(\mathbb{P}_n - P) = o_p(\log^2 n \times n^{-2/7}),$$

uniformly over $\beta \in \mathbf{B}$ and t in the support $\mathcal{T}$.

Proof of Lemma S13: We first introduce some notations adapted from Groeneboom

et al. (2010). Define for

$$C_{t,n}(u) = \frac{k(n^{1/7}(t-u)/c)}{cp(y,x,\delta)} 1\{t - cn^{-1/7} \leq u \leq t + cn^{-1/7}\}, \tag{S.48}$$

the function

$$\xi_{A,B,t,n}(u) = C_{t,n}(A(u))B(u) \tag{S.49}$$

where $t \in \mathcal{T}$, A is an increasing function, and B is a function of bounded variation. And let

$$\mathcal{G}_{1,n} \equiv \{\xi_{A,B,t,n}(u) : A \in \mathcal{A}, B \in \mathcal{B}, t \in \mathcal{T}\}. \tag{S.50}$$

Given the result in (S8), for any small $\gamma > 0$ we can find finite constant term M_1 such that for all n sufficiently large:

$$\Pr\{\Upsilon_{1,n,M_1}\} \equiv \Pr\{\sup_{u,\beta} |\hat{\lambda}_n(u,\beta) - \lambda(u,\beta)| \leq M_1 n^{-1/3} \log n\} \geq 1 - \gamma/2.$$

Now for the vanishing sequence ν_n to be specified later, it is straightforward to arrive at

$$\begin{aligned}\Pr\{|n^{2/7}R_n| > \nu_n\} &= \Pr\{|n^{2/7}R_n| > \nu_n \cap \Upsilon_{1,n,M_1}\} + \Pr\{|n^{2/7}R_n| > \nu_n \cap \Upsilon^c_{1,n,M_1}\} \\ &\leq \nu_n^{-1} E|n^{2/7}R_n| 1\{\Upsilon_{1,n,M_1}\} + \gamma/2.\end{aligned}$$

Again by (S8), we have

$$\begin{aligned}E|n^{2/7}R_n| 1\{\Upsilon_{1,n,M_1}\} &\leq E \sup_{A\in\mathcal{A}, B\in\mathcal{B}_{M_1}} \left| n^{2/7-1/3} \log n \int \phi_{h,t}(A(u))B(u) d(\mathbb{P}_n - P) \right| \\ &\leq n^{4/7-5/6} \log n E \sup_{\xi\in\mathcal{G}_{1,n}} \left| \int \xi(u) d\mathbb{G}_n(u) \right|.\end{aligned}$$

The rest of the proof is to utilize Theorem 2.14.1 in Van Der Vaart and Wellner (1996) to bound the expectation in the last display. Following the construction in Groeneboom et al. (2010), we can select a minimal $n^{-1/7}\delta/(4M_1)$-net in $\mathcal{A}$ and a minimal $\delta/(2 \parallel C_n \parallel_\infty)$-net in $\mathcal{B}$. Referring to the part involving the kernel function $k(\cdot)$ indexed by t, we can obtain a minimal $n^{-1/7}\delta/(4M_1)$-net in $\mathcal{T}$ given the fact that the functional class is of VC-type Nolan and Pollard (1987). The number of functions in this net to cover $\mathcal{G}_{1,n}$ is bounded by $Mn^{1/7}/\delta \exp(n^{1/7}/\delta)$, so the entropy integral is bounded above by

$$J(1, \mathcal{G}_{1,n}) \lesssim n^{1/14} \log n. \tag{S.51}$$

Compared with Groeneboom et al. (2010), the uniformity over $t \in \mathcal{T}$ in the kernel function brings an extra $\log n$ term. The L_2-norm of the envelope function is of order $O_p(\sqrt{h}) = O_p(n^{-1/14})$ by standard arguments in kernel smoothing. Thus applying (S.17), we have

$$E|R_n| \leq n^{4/7-5/6} \log n \times E \sup_{\xi\in\mathcal{G}_{1,n}} \left| \int \xi(u) d\mathbb{G}_n(u) \right| \lesssim n^{4/7-5/6} \log^2 n,$$

which immediately leads to $R_n = o_p(\log^2 n \times n^{-2/7})$. □

LEMMA S14. *Suppose that Assumptions Assumptions 4.1 to 4.6 hold. Then* $J_{2n}^{(a)}$ *and* $J_{2n}^{(b)}$ *defined by (S.47) in Appendix B satisfy:*

$$J_{2n}^{(a)} = o_p(\log^2 n \times n^{-2/7}), \quad \textit{and} \quad J_{2n}^{(b)} = O_p(n^{-1/2} \times h^{-1}) = o_p(n^{-5/14}),$$

uniformly over $t \in \mathcal{T}, \beta \in \mathbf{B}$.

Proof of Lemma S14: After linearizing $1/D_n^{(0)}$, the first term $J_{2n}^{(a)}$ is determined by

$$R_n^{(a)} \equiv \int \bar{\phi}_{h,t} 1\{\epsilon_\beta \leq y - x'\beta, \delta = 1\} \frac{D^{(0)}(y - x'\beta, \beta) - D_n^{(0)}(y - x'\beta, \beta)}{D^{(0)2}(y - x'\beta, \beta)} d(\mathbb{P}_n - P), \tag{S.52}$$

modulo a second order term which is of $O_p(n^{-1}h^{-2})$. Note that both $D_n^{(0)}$ and $D^{(0)}$ are monotonically decreasing functions as defined in (**??**), hence their difference is a function of bounded variation. Now it is clear that $R_n^{(a)}$ can be dealt with in a completely analogous way as the term R_n in the previous lemma, which leads to $R_n^{(a)} = o_p(\log^2 n \times n^{-2/7})$. Referring to $J_{2n}^{(b)}$, we get

$$\begin{aligned}
&\int \bar{\phi}_{h,t} 1\{\epsilon_\beta \leq y - x'\beta, \delta = 1\} \left[\frac{1}{D_n^{(0)}(y - x'\beta, \beta)} - \frac{1}{D^{(0)}(y - x'\beta, \beta)} \right] dP \\
\leq &\| D_n^{(0)} - D^{(0)} \|_\infty \int \left| \frac{\bar{\phi}_{h,t} 1\{\epsilon_\beta \leq y - x'\beta, \delta = 1\}}{D_n^{(0)}(y - x'\beta, \beta) D^{(0)}(y - x'\beta, \beta)} \right| dP \\
\lesssim &\, n^{-1/2} \times h^{-1} = O_p(n^{-5/14}).
\end{aligned}$$

□

S2. AN ADDITIONAL EMPIRICAL APPLICATION

Methadone maintenance is a drug replacement therapy that uses methadone over a prolonged period of time as a treatment for heroin addicts. However, hospitals and clinics often have problems to retain the addicts. Barnett et al. (2000) finds that the median willing to pay of heroin addicts in Baltimore was much lower than the estimated cost of the methadone maintenance program. Meinhofer and Witman (2018) discover that the recent expansion of health insurance in US (Medicaid expansion) substantially increased admissions to medication-assistant treatment[1] for opioid use disorder in outpatient settings while hardly changed admissions to inpatient settings. Here we use the data of an Australian methadone program (Caplehorn and Bell, 1991) to investigate the retention time of heroin addicts in methadone maintenance. The data was collected from 238 addicts who were assigned to one of two public clinics based on their home address between February 1986 and August 1987. Both clinics serve adjacent areas of Sydney with similar socioeconomic characteristics. The censoring rate is 37%. In AFT model, we regress the days in treatment, censored at the end of the study, to the treatment variable is "Clinic" (taking value 1 or 2 to indicate two clinics that take different overall treatment polices (Kleinbaum and Klein, 2011), and two other variables "Prison" (that indicates prison record status) and "Dose" (that records the maximum methadone dose per day in mg). Estimates of parameters and standard errors are reported in Table S1. All methods suggest that Clinic 2 managed to retain addicts for a substantially longer period of time. The negative association between the prison record status and retention time is significant at 5% level for the proposed SMLE methods but insignificant for other methods. The

[1]The treatment uses methadone, buprenorphine, or naltrexone.

positive correlation between the maximum daily methadone and the retention time is highly significant for all methods.

Table S1. AFT model estimates for retention time in methadone maintenance.

Variables	SMLE-u		SMLE-s		Gehan		Sieve	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Clinic	.521	.121	.458	.124	.510	.203	.629	.178
Prison	-.219	.108	-.174	.065	-.304	.158	-.277	.153
Dose	.033	.004	.014	.003	.032	.007	.058	.005

Note: The bandwidths for smoothing the estimated hazard rate is $1 \times n^{-1/5}$ and for smoothing its derivative is $1 \times n^{-1/7}$.

Figure S1 displays the smoothed estimates of the hazard function of the error term ε and the hazard of T_0 in the retention of addicts data. By the log-concavity assumption on the error term, its estimated hazard function is increasing, as the left panel shows. The right panel exhibits that the overall trend of the estimated hazard function is increasing in the addicts' retention data. This suggest that the risk of patients dropping out of the methadone treatment keeps increasing over time, given the prison record status and maximum daily dose.

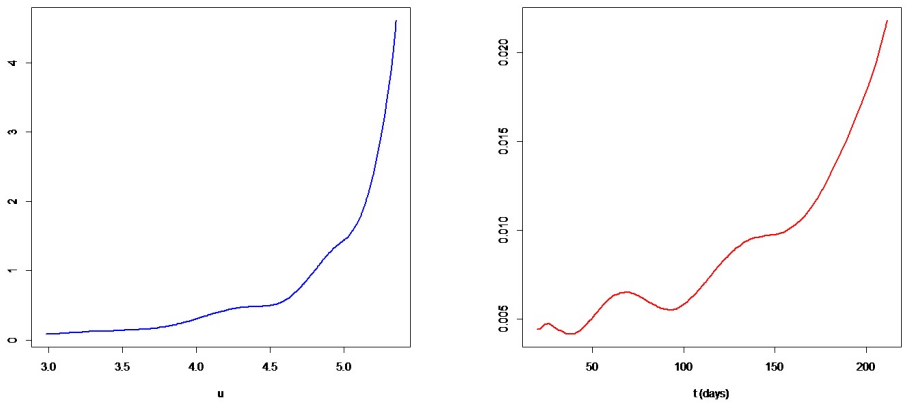

Figure S1. Estimated hazard functions of the error term ε (left) and the baseline duration T_0 (right).

REFERENCES

Bagnoli, M. and T. Bergstrom (2005). Log-concave probability and its applications. *Economic theory 26*(2), 445–469.

Barnett, P. G., S. S. Hui, et al. (2000). The cost-effectiveness of methadone maintenance. *Mount Sinai Journal of Medicine 67*(5-6), 365–374.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press.

Caplehorn, J. R. and J. Bell (1991). Methadone dosage and retention of patients maintenance treatment. *Medical Journal of Australia 154*(3), 195–199.

Esary, J. and A. Marshall (1973). Shock models and wear processes. *The Annals of Probability 1*, 627–649.

Groeneboom, P. and K. Hendrickx (2018). Current status linear regression. *The Annals of Statistics 46*(4), 1415–1444.

Groeneboom, P. and G. Jongbloed (2014). *Nonparametric estimation under shape constraints*. Cambridge University Press.

Groeneboom, P., G. Jongbloed, and B. I. Witte (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *The Annals of Statistics 38*, 352–387.

Kleinbaum, D. G. and M. Klein (2011). *Survival Analysis: A Self-Learning Text*. Springer.

Kuchibhotla, A. K., R. K. Patra, and B. Sen (2017). Efficient estimation in convex single index models. *working paper*.

Lai, T. L. and Z. Ying (1991). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics 19*, 531–556.

Marshall, A. W. and I. Olkin (2007). *Life distributions*. Springer.

Meinhofer, A. and A. E. Witman (2018). The role of health insurance on treatment for opioid use disorders: Evidence from the affordable care act medicaid expansion. *Journal of health economics 60*, 177–197.

Nan, B. and J. A. Wellner (2013). A general semiparametric Z-estimation approach for case-cohort studies. *Statistica Sinica 23*, 1155–1180.

Nolan, D. and D. Pollard (1987). U-processes: Rates of convergence. *The Annals of Statistics 15*, 780–799.

Ritov, Y. and J. A. Wellner (1988). Censoring, martingales, and the cox model. *Contemporary Mathematics 80*, 191–219.

Tripathi, G. (2000). Local semiparametric efficiency bounds under shape restrictions. *Econometric Theory 16*, 729–739.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics 18*, 354–372.

Van Der Vaart, A. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer.

Widder, D. V. (1941). *The Laplace transform*. Princeton University Press.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics 21*, 76–99.