Open Access Library Journal
2020, Volume 7, e6824
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Polynomial-Based Smoothing Estimation for a Semiparametric Accelerated Failure Time Partial Linear Model

Wei Chen[1]*, Fengling Ren[2]

[1]School of Zhangjiagang, Jiangsu University of Science and Technology (JUST), Zhangjiagang, China
[2]School of Computer Science and Engineering, Xinjiang University of Finance and Economics, Urumqi, China
Email: *chenweixiyang@sina.com

How to cite this paper: Chen, W. and Ren, F.L. (2020) Polynomial-Based Smoothing Estimation for a Semiparametric Accelerated Failure Time Partial Linear Model. *Open Access Library Journal*, **7**: e6824.
https://doi.org/10.4236/oalib.1106824

Received: September 15, 2020
Accepted: October 16, 2020
Published: October 19, 2020

Abstract

The accelerated failure time partial linear model allows the functional form of the effect of covariates to be possibly nonlinear and unknown. We propose to approximate the nonparametric component by cubic *B*-splines and construct a Gehan estimating function similar to that under the AFT model. Due to its non-smoothness, which will lead to computational challenge in estimating standard error, we propose a polynomial-based smoothing Gehan estimating function and compute the estimate of the parameters involved using the limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm. Asymptotic properties of the resulting estimators are established. The proposed method presents a good performance in the simulation studies and is applied to two real data sets.

Subject Areas

Mathematical Statistics

Keywords

Accelerated Failure Time Model, Partial Linear Model, Polynomial-Based Smoothing, Rank Estimation, Quasi-Newton Methods

1. Introduction

In survival analysis, it is often of interest to explore the relationship between the failure time and a collection of covariates. For this purpose, a large number of semiparametric regression models and estimation methods have been developed. Among them, the Cox [1] proportional hazards (PH) model may be the most popular and widely-used statistical tool for analyzing survival data partly due to the efficient inference based on the partial likelihood and the availability of implementation in almost all existing softwares. However, the PH model requires that the hazard ratio is always constant over time between any two subjects with distinct covariates. In some situations, this assumption seems to be rather restrictive and hard to be met. See [2]. Thus, other alternatives to the PH model with non-proportional risks or more flexibility of modelling the covariates are desirable.

The AFT model assumes that the logarithm of survival time is linearly correlated to a vector of covariates of interest, which can be specified as

$$\log(T) = \beta^{\mathrm{T}} Z + \epsilon , \tag{1.1}$$

where T denotes the failure time; β is a p-vector of regression coefficients to be estimated; Z is a p-vector of covariates; ϵ is the random error with mean zero but unknown distribution; and the superscript "T" denotes the transpose of a column vector. In the presence of right censored data, several semiparametric estimates have been proposed, such as the least square estimator in [3] [4] and rank-based estimator in [5] [6] [7]. Nevertheless, these estimators have not been wildly used as it should be in practice due to lack of efficient and reliable computation algorithm to obtain the parameter and its standard error estimation. Especially, for the rank-based estimator, the computational challenge arises from two aspects: the estimating function used is non-smooth, *i.e.* a step function with respect to β; and the asymptotic slope matrix of the estimating function depends on the unknown hazard function of the error and its first derivative, which makes the direct estimation from the observed data impossible. Thus, most existing covariance matrix estimation methods rely on the bootstrap approach that is computation-demanding. The authors in [8] provided the first reliable and accurate estimating procedure via linear programming (LP) to obtain the Gehan estimator, a special case of the general weighted logrank estimator, which is implemented in R package "lss". However, the merit of their LP strategy is greatly discounted for large, even modest sample sizes in [9] [10].

In view of these limitations, some more computationally efficient procedures for the Gehan estimator are introduced by [9] [11] [12] [13]. Explicitly speaking, Brown and Wang [12] developed a pseudo-Bayesian approach in [11] to derive a smoothing version, which is asymptotically equivalent to the original discontinuous Gehan estimating function. They called this technique the induced smoothing (IS) method. A significant advantage is that the smoothed estimating function is differential so that one can estimate the regression parameters and the covariance matrix simultaneously with common numerical methods, avoiding the computationally extensive resampling. Later on, the theoretical justification for the IS method was provided by [14], and they extended the method to clustered failure time data. On the other hand, Hellner [13] considered to directly approximate the indicator function in the Gehan estimating function with a known distribution function and showed the resulting estimator converges in distribution to a normal random vector with mean zero and covariance matrix which can be straightforwardly estimable. A detailed review is available in [9], where they also suggested a polynomial-based smoothing method which has more well-behaved performance than the two counterparts mentioned above. This technique will also be applied to our problem depicted in this paper.

Although the AFT model is useful, the assumption that each covariate has a linear effect on the log survival time is not appropriate in some situations. For example, in many clinical trials and biomedical studies, one is primarily concern about identifying the effect of a treatment when a confounding factor of less interest exists. In such cases, it is reasonable and useful to treat the confounding factor as a nonparametric component without loss of the easy interpretation of the treatment effect in [15]. In the literature, one usually characterizes these covariate effects through a model referred to as the partial linear (PL) model, which can be written as

$$\log(T) = \beta^{\mathrm{T}} Z + h(U) + \epsilon, \tag{1.2}$$

where U is an univariate covariate such as the confounding factor, $h(.)$ is an unknown smooth function playing the role of the nonparametric component, and other notation are defined as above. In the linear regression setting, when the response variable T is completely observed, many researchers have studied the PL model, see [16]. Rather, to the best of our knowledge, there is little investigation for inference of model (1.2) with right censored data except several authors, where the model (1.2) is also called the semiparametric accelerated failure time partial linear model (AFT-PLM). Orbe in [17] adapted [18]'s method and proposed a penalized weighted least square method with unknown function h being approximated by the cubic splines. But the statistical properties of the resulting estimators are not well established (page 112 of [17]). Chen in [15] developed a strategy to eliminate the function $h(.)$ by a proper stratification, thus proposed an estimation method, which is a Gehan-type extension of the Wilcoxon-Mann-Whitney estimating function. They proved the acquired rank estimate to be consistent and asymptotically normal. However, duo to stratifications, the nonparametric component is not likely to be estimated appropriately. Recently, Zou in [19] incorporated the penalized spline into the Gehan-type estimating function occurred in the rank-based inference for the AFT model and obtained estimate of the regression coefficients and nonparametric component simultaneously. Nevertheless, all current methods are either computationally extensive, which is more severe for large sample sizes (>400), because they all rely on the bootstrap technique to estimate the covariance matrix, or fail to provide an estimate of the nonparametric component when the effect of U is also of interest. Specially, with regard to [19]'s procedure, the estimating function is non-smooth and may bring numerical difficulties when more covariates are incorporated into the model (1.2). In view of advances on dealing with non-smooth estimating function summarized above, it is possible to develop a smoothing estimation method for the semiparametric AFT-PLM. In this paper, we consider this issue. Once the smoothed estimating function is derived, under certain regularity conditions, the common inference techniques can be applied. Therefore, one can consistently estimate the covariance matrix by the plug-in rule without resorting to the time-consuming resampling method.

The rest of this paper is organized as follows. In Section 2, we introduce some notation and assumptions, and derive a smoothed Gehan estimating function through a polynomial-based smoothing method given that the nonparametric component *h* in (1.2) is parametrically approximated by the cubic *B*-splines. Under some regularity conditions, the resulting estimator is shown to be consistent and asymptotically normal. In particular, the asymptotic covariance matrix of estimators of the regression coefficients can be straightforwardly estimated, avoiding the substantial computation needed in the resampling approach. In addition, by virtue of the fact that the smoothed Gehan estimating function can be written as the gradient of a smooth convex loss function, we develop the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm to solve the numerical problem. The procedure is implemented in our simulation studies presented in Section 3. The results show our method performs quiet well, no matter for the estimation of the regression coefficients or the unknown function *h*. Section 4 presents the results of the application to two real data sets and, finally, some discussions in Section 5 conclude the paper.

2. Methods

2.1. Notation, Model and Assumptions

Suppose that there are *n* independent subjects under study. For the i^{th} subject, $i = 1, \cdots, n$, let T_i denote the failure time and Z_i be the *p*-vector of covariates. In addition, one auxiliary covariate U_i is also measured. Since the failure time is subject to right censoring, we will instead observe the i.i.d. vectors $(Y_i, \delta_i, Z_i, U_i)$ of (Y, δ, Z, U), where $Y = \min(T, C)$, $\delta = I(T \le c)$ is the censoring indicator taking values 1 if the failure time is observed and 0 otherwise. Here C denotes the right-censoring time. In this paper, we assume (T_i, Z_i, U_i) $(i = 1, \cdots, n)$ satisfy the AFT-PLM defined in (1.2). As in most cases, we also assume T and C are independent given (Z, U) and, (Z, U) and ϵ are independent. For technical reasons, we assume the covariates Z and U have bounded supports, and without loss of generality, we take the support of U as the unit interval [0, 1]. Furthermore, we also assume $\mathbb{E}[h(U)] = 0$, which is also required in the context of [20] [21].

2.2. Estimation Procedures

If the nonparametric component *h* is known or fully parametric, the statistical inference problem to be solved can be readily reduced to the usual AFT rank-based problem. Under suitable regularity conditions, it can be shown the resulting estimator is consistent and asymptotically normal. However, as argued in the introduction, the effect of U on the survival time is not certain and a misspecified form of *h* will lead to biased conclusions. To attain the flexibility of modelling and reliable results, it is more desirable not to impose an explicit form on *h*. However, just as to the incorporation of the unknown function *h*, the commonly-used rank-based inference approach can not directly be applied because doing so may suffer from the so-called "curse of dimensionality".

A simple but useful method is to approximate the unknown function *h* by a spline. In the survival literature, the use of splines is common in [22] and among others. More details on splines can be found in [23]. In this paper, we assume that the smooth function $h(u)$ can be expressed as a function of *B*-splines, *i.e.*

$$h(u) = \sum_{l=-\varrho}^{L} \gamma_l B_l(u), \tag{2.1}$$

where $B_l(u), l = -\varrho, \cdots, L$, are the *B*-spline basis functions of degree $\rho \geq 1$ associated a sequence of knots

$$t_{-\rho} = \cdots = t_{-1} = t_0 = 0 < t_1 < \cdots < t_L < 1 = t_{L+1} = \cdots = t_{L+\varrho+1}.$$

Let $B(u) = \{B_{-\varrho}(u), \cdots, B_L(u)\}^{\mathrm{T}}$ and $\gamma = (\gamma_{-\varrho}, \cdots, \gamma_L)^{\mathrm{T}}$. Then one can write the expression (3) as

$$h(u) = \gamma^{\mathrm{T}} B(u). \tag{2.2}$$

In our numerical studies and real data analysis, the cubic *B*-splines, *i.e.* $\varrho = 3$, are used in the basis expansion of $h(u)$. Generally 3 - 10 internal knots are adequate in practice in [22]. In our implementation, we choose the number of internal knots, $L = 3, 5, 7$, respectively. As demonstrated in the following simulation studies, our strategy is appropriate and the results obtained are not sensitive to the selection of different numbers of internal knots. In addition, once the number is given, we put the knots equally spaced between the smallest and largest values of U_i's.

By virtue of the expansion of *h* defined in (2.2), the AFT-PLM in (1.2) can be rewritten as

$$\log(T) = \beta^{\mathrm{T}} Z + \gamma^{\mathrm{T}} B(U) + \epsilon = \theta^{\mathrm{T}} X + \epsilon, \tag{2.3}$$

where $X = \left(Z^{\mathrm{T}}, B(U)^{\mathrm{T}}\right)^{\mathrm{T}}, \theta = \left(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}}\right)^{\mathrm{T}}$. Consequently, by applying the weighted logrank estimation method in [5] [6] along with the Gehan weight function, we have the following estimating function

$$\Psi_G(\theta) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \left(X_i - X_j\right) I\left(e_i \leq e_j\right), \tag{2.4}$$

where $X_i = \left(Z_i^{\mathrm{T}}, B(U_i)^{\mathrm{T}}\right)^{\mathrm{T}}, e_i = e_i(\theta) = \log Y_i - \theta^{\mathrm{T}} X_i$, which is often referred to as the Gehan estimating function. On the other hand, the Gehan estimating function $\Psi_G(\theta)$ in (2.4) is the gradient of the following convex loss function

$$f_G(\theta) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i \left(e_j - e_i\right) I\left(e_i \leq e_j\right), \tag{2.5}$$

which is called the Gehan loss function in [9]. Naturally one can define the Gehan estimator of θ as the minimizer of the objective function $f_G(\theta)$, denoted by $\hat{\theta}=\left(\hat{\beta}^{\mathrm{T}},\hat{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}$, and the optimization problem can be solved by linear programming in [8] for small sample sizes. To derive the large sample properties of the proposed estimators, we assume the smooth function $h(.)$ is a spline with pre-specified knots. Doing so is due to mainly computational and theoretical consideration. The idea is also employed in issues investigated by [22] [24] among others. Under some regularity conditions *C*1 - *C*4 described in [6], it is shown that the resulting estimator $\hat{\theta}$ of θ is consistent and asymptotically normal with mean zero and an indirectly estimable covariance matrix. Thus to make inference for β, one has to resort to the resampling approach which is computationally intensive, especially for large sample sizes, even modest sample sizes.

Note that the challenge encountered in current background is also reflected in the rank-based inference problem for the AFT model with censored data. As reviewed in the introduction, these difficulties arise from the nonsmoothness of the Gehan estimating function. Building on the recent advances of the smoothed rank-based method, it enables us to develop an easily-implemented estimation method for both the regression coefficients β and the possibly nonlinear function *h*.

Define the following smoothing approximation to the Gehan loss function in (2.5),

$$f_{G,\varepsilon}(\theta)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i K_{\varepsilon}\left(e_i-e_j\right), \tag{2.6}$$

where K_{ε} is a sufficiently smooth real-valued function, having the form

$$K_{\varepsilon}(v)=\begin{cases}-v, & \text{if } v\le-\varepsilon;\\ -\dfrac{1}{16\varepsilon^3}(v-\varepsilon)^4-\dfrac{1}{4\varepsilon^2}(v-\varepsilon)^3, & \text{if } -\varepsilon<v\le\varepsilon;\\ 0, & \text{if } v>\varepsilon;\end{cases}$$

with sufficiently small but strictly positive ε. Clearly, $f_{G,\varepsilon}$ is identical to f_G in the entire line outside of the interval $(-\varepsilon,\varepsilon]$ and replaces f_G by a polynomial function in that interval. Through simple calculation, it can be seen that the function $f_{G,\varepsilon}$ has a continuous second order derivative for any $\varepsilon>0$, especially, $\lim_{\varepsilon\to0} f_{G,\varepsilon}(\theta)=f_G(\theta)$. Up to now, we define the estimator of θ as the minimizer of the smooth objective function $f_{G,\varepsilon}(\theta)$ in (2.6), *i.e.*

$$\tilde{\theta}=\left(\tilde{\beta}^{\mathrm{T}},\tilde{\gamma}^{\mathrm{T}}\right)^{\mathrm{T}}=\arg\min_{\theta} f_{G,\varepsilon}(\theta).$$

Then *h* can be estimated by $\tilde{h}(u)=\sum_{l=-\varrho}^{L}\tilde{\gamma}_l B_l(u)$. In fact, if the minimizer exists, it is also the solution to the estimating function $\Psi_{G,\varepsilon}(\theta)$, where

$$\Psi_{G,\varepsilon}(\theta)=\frac{\partial}{\partial\theta}f_{G,\varepsilon}(\theta)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\left(X_i-X_j\right)k_{\varepsilon}\left(e_i-e_j\right), \tag{2.7}$$

$$k_{\varepsilon}(v)=-\frac{\partial}{\partial v}K_{\varepsilon}(v)=\begin{cases}1, & \text{if } v\le-\varepsilon;\\ \dfrac{1}{4\varepsilon^3}(v-\varepsilon)^3+\dfrac{3}{4\varepsilon^2}(v-\varepsilon)^2, & \text{if } -\varepsilon<v\le\varepsilon;\\ 0, & \text{if } v>\varepsilon;\end{cases}$$

and $\Psi_{G,\varepsilon}(\theta)$ is actually the smoothed version of $\Psi_G(\theta)$ in (2.4).

Remark 1. In the asymptotical analysis, the tuning parameter ε should decrease as the sample size n increases. In this paper, we set ε to be 10^{-4} when used. This idea is common in statistical computing and adopted by [9] and among others.

Remark 2. As an alternative, we propose to complete the computation by quasi-Newton methods, which avoid calculating the Hessian matrix. This advantage is more apparent when the dimension of parameter to be estimated is high or there exists an ill-posed problem. Explicitly, we recommend the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method in [25]. And we terminate the iterative step when the relative tolerance is smaller than 10^{-4} in our implementation. As we will see, this procedure works well.

2.3. Inference

Under the conditions aforementioned and assumptions A1-A3 in [14], in line of arguments in the Appendix of [9] [13], it can be shown that the smoothed estimating function $\Psi_{G,\varepsilon}(\theta)$ defined in (2.7) is asymptotically equivalent to the non-smooth one $\Psi_G(\theta)$ in (2.4). Furthermore, applying the arguments in the Appendix of [14], we can show that $\tilde{\theta}$ is consistent, and $n^{1/2}(\tilde{\theta}-\theta_0)$ converges in distribution to a normal vector with mean zero and covariance matrix $A_\varepsilon^{-1}(\theta_0)D_\varepsilon(\theta_0)A_\varepsilon^{-1}(\theta_0)$, where θ_0 is the true value of θ,

$$A_\varepsilon(\theta_0)=\lim\nolimits_{n\to\infty}\mathbb{E}\left\{\frac{\partial}{\partial\theta}\Psi_{G,\varepsilon}(\theta)\right\}\Big|_{\theta=\theta_0},$$

$$D_\varepsilon(\theta_0)=\lim\nolimits_{n\to\infty}\operatorname{Var}\left\{n^{1/2}\Psi_{G,\varepsilon}(\theta)\right\}\Big|_{\theta=\theta_0},$$

which can be consistently estimated by

$$\tilde{A}_\varepsilon(\tilde{\theta})=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\left(X_i-X_j\right)^{\otimes 2}k_{1,\varepsilon}\left(e_i-e_j\right)\Big|_{\theta=\tilde{\theta}},$$

$$\tilde{D}_\varepsilon(\tilde{\theta})=\frac{1}{n^3}\sum_{i=1}^{n}\delta_i\left[\sum_{j=1}^{n}\left(X_i-X_j\right)I\left(e_i\le e_j\right)\right]^{\otimes 2}\Bigg|_{\theta=\tilde{\theta}},$$

respectively, where $k_{1,\varepsilon}(v)=\frac{\partial}{\partial v}k_\varepsilon(v)$ and $\otimes^2$ denotes aa^{T} for a vector *a*. With regard to the second matrix, we obtain it by virtue of the asymptotic equivalence between $n^{1/2}\Psi_{G,\varepsilon}(\theta)$ and $n^{1/2}\Psi_G(\theta)$. Of course, we can also replace it with one derived in terms of of the smoothed estimating function $\Psi_{G,\varepsilon}(\theta)$. Compared to the non-smoothed one, $\tilde{D}_\varepsilon(\tilde{\theta})$ is computationally convenient. Therefore, one can make inference for θ via the estimated covariance matrix $\tilde{A}_\varepsilon^{-1}(\tilde{\theta})\tilde{D}_\varepsilon(\tilde{\theta})\tilde{A}_\varepsilon^{-1}(\tilde{\theta})$. It is implemented in our simulation studies and real data analysis.

3. Numerical Studies

To assess the performance of our estimation method, we conduct an extensive simulation study under various scenarios. We independently generated the survival time T_i from the following model

$$\log(T_i) = \beta_{10}Z_{1i} + \beta_{20}Z_{2i} + h_0(U_i) + \epsilon_i,\ i = 1,\cdots,n,$$

where $(\beta_{10},\beta_{20}) = (-0.3, 0.3)$, $Z_{1i} \sim \text{Uniform}(-3,3)$, $Z_{2i} \sim \text{Binomial}(1,0.5)$, $U_i \sim \text{Uniform}(0,1)$, and $\epsilon_i \sim \text{Normal}(0,1)$. In all simulations, the covariates (Z_{1i}, Z_{2i}, U_i) and error ϵ_i are independently generated. The censoring times C_i are generated independently from an exponential distribution with means varying to yield the censoring rates about 15%, 30%, 50%, respectively. Two different functions for the nonparametric component $h_0(.)$ are considered. For the first one, called Case I, $h_0(U_i) = \sin(2\pi U_i)$, which has one peak and one valley in the domain [0, 1], respectively. For the second one,

$h_0(U_i) = \log(1+U_i^2) - 0.2639$, denoted as Case II. For spline approximation to the nonparametric component, we select the number of internal knots as $L = 3,5,7$, respectively, and locate them equally-spaced between the smallest and largest observations of U_i's. We set ε defined in K_ε to be 10^{-4}. The sample sizes $n = 200, 500, 1000$, which correspond to the small, middle, and large levels respectively as used in [9], are considered in each scenario with 1000 runs of simulations. All simulations are implemented using the software Matlab, and the initial values are generated from a standard multivariate normal distribution. For $\tilde{\beta}$, we record its empirical bias (Bias), sample standard deviation (SD), standard error estimation (SEE), and empirical coverage probability of 95% confidence interval. For the nonparametric part, we use the mean estimated integrated square error (IMSE), where

$$\text{IMSE} = (1/ngrid)\sum_{i=1}^{ngrid}\left(\tilde{h}(u_i) - h_0(u_i)\right)^2,$$

at the fixed grid points $\{u_i\}$ between zero and one with step 0.01, $ngrid$ is the number of these grid points.

Table 1 summarizes the results for case I with 5 internal knots. It is seen that the proposed estimates $\tilde{\beta}$ are quiet accurate. Moreover, the coverage probabilities are close to the nominal level 0.95. Remarkably, the results reported here are superior to those summarized in Table 1 in [19] based on the P-spline Gehan estimating function, where no smoothing is employed. It is worthnoting that the sample standard deviation is comparable to the standard error estimation obtained by the sandwich estimator based on the smoothed estimating function, even for small sample size ($n = 200$), implying that using our procedure to estimate the covariance matrix for making statistical inference is appropriate while little computational effort is involved. For a fixed censoring rate, when the sample size increases, both the biases and standard error estimates decrease; for a fixed sample size, with the increment of the censoring rate, the biases and standard error estimates will increase. The same tendency is also reflected in the IMSE. **Figure 1** shows the mean, median of estimation of the function h_0 from 1000 simulations with sample size 500, censoring rate 30%, and 5 internal equally-spaced internal

Table 1. Case I Sin curve model: results of parameters with 5 internal knots and cubic B-splines.

n	CR	$\tilde{\beta}_1$				$\tilde{\beta}_2$				$\tilde{h}(u)$
		Bias	SD	SEE	CP	Bias	SD	SEE	CP	IMSE
200	15%	0.0008	0.0172	0.0173	0.950	0.0013	0.0553	0.0543	0.950	0.1189
	30%	−0.0005	0.0181	0.0182	0.956	−0.0058	0.0642	0.0631	0.949	0.1197
	50%	−0.0014	0.0221	0.0220	0.955	0.0016	0.0720	0.0716	0.959	0.1093
500	15%	0.0000	0.0101	0.0101	0.947	0.0000	0.0352	0.0348	0.951	0.1172
	30%	0.0000	0.0110	0.0110	0.955	0.0002	0.0386	0.0391	0.950	0.1175
	50%	0.0000	0.0132	0.0136	0.951	0.0010	0.0449	0.0491	0.946	0.1191
1000	15%	−0.0003	0.0074	0.0074	0.942	−0.0000	0.0251	0.0251	0.941	0.1128
	30%	−0.0004	0.0080	0.0080	0.951	−0.0001	0.0268	0.0268	0.952	0.1129
	50%	−0.0006	0.0092	0.0092	0.945	−0.0001	0.0320	0.0319	0.941	0.1137

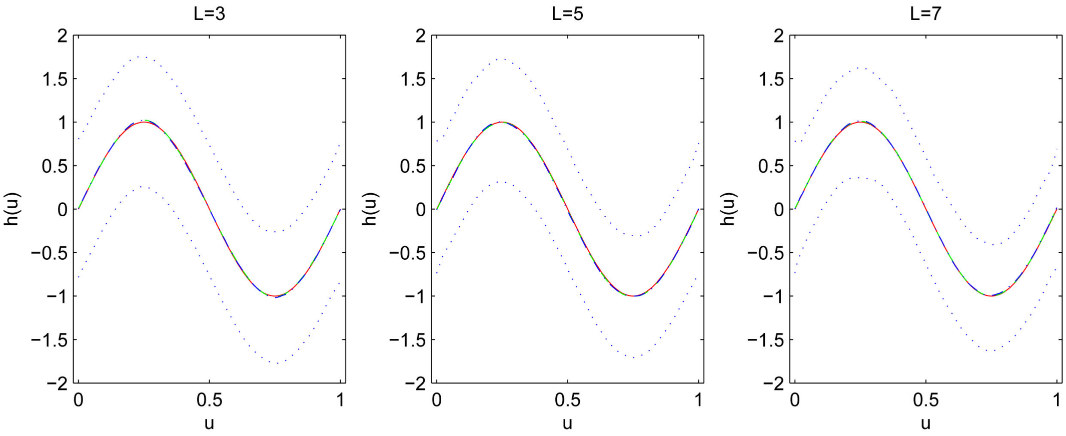

Figure 1. The mean (dotted), median (dash-dotted) of estimated function $\tilde{h}(u)$ and 95% pointwise Monte Carlo intervals (dotted) for case I model with number of internal knots $L = 3,5,7$, respectively.

knots, which are both close to the true curve $\sin(2\pi u)$, and its 95% pointwise Monte Carlo intervals, which are constructed using the 2.5% and 97.5% sample quantiles of the estimated functions. Similar phenomena are also occurred in cases with remaining censoring rates and numbers of internal knots but not shown. Therefore, even a few number of knots are determined, the regression coefficient estimates and the estimated curve perform well enough.

Table 2 and Figure 2 represent the results under the same setting as that in Table 1 when we estimated the parameters in the Case II model, and shows similar results.

In addition, as we have particularly stressed many times, for only a data set with sample size 1000, the proposed L-BFGS algorithm just need about 20 seconds to fulfill a complete inference, however, from our experiments, to compute the estimates of parameters and their standard errors through optimizing the

Table 2. Case II log curve model: results of parameters with 3, 5, 7 internal knots and cubic B-splines.

n	CR	$\tilde{\beta}_1$				$\tilde{\beta}_2$				$\tilde{h}(u)$
		Bias	SD	SEE	CP	Bias	SD	SEE	CP	IMSE
200	15%	−0.0012	0.0160	0.0159	0.948	0.0023	0.0572	0.0566	0.950	0.1160
	30%	0.0006	0.0184	0.0185	0.954	0.0032	0.0586	0.0583	0.951	0.1190
	50%	−0.0004	0.0216	0.0218	0.943	−0.0058	0.0721	0.0722	0.954	0.1236
500	15%	0.0002	0.0101	0.0104	0.958	0.0002	0.0344	0.0336	0.943	0.1113
	30%	−0.0005	0.0114	0.0111	0.950	−0.0018	0.0377	0.0362	0.959	0.1239
	50%	−0.0000	0.0129	0.0135	0.948	−0.0021	0.0449	0.0434	0.954	0.1199
1000	15%	0.0004	0.0072	0.0072	0.956	−0.0009	0.0253	0.0253	0.949	0.1087
	30%	−0.0003	0.0077	0.0077	0.954	−0.0009	0.0264	0.0264	0.946	0.1131
	50%	0.0003	0.0092	0.0092	0.952	−0.0021	0.0315	0.0313	0.950	0.1149

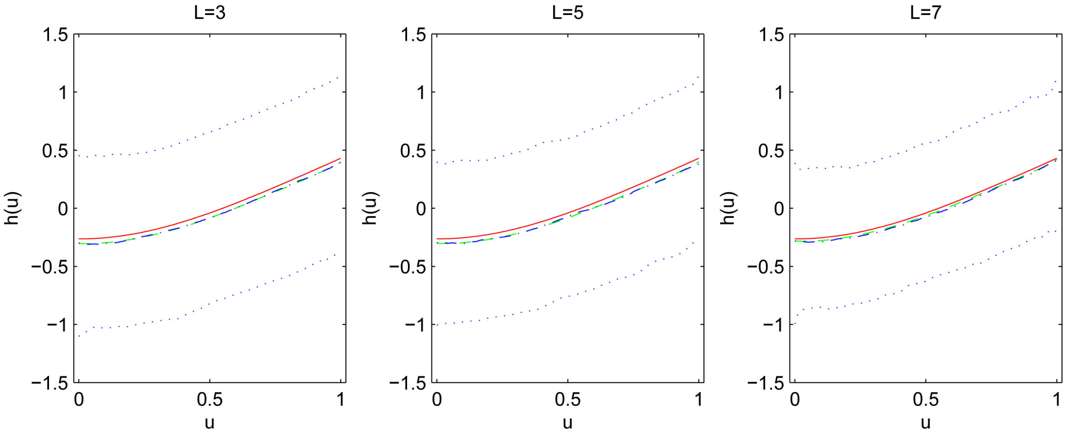

Figure 2. The mean (dotted), median (dash-dotted) of estimated function $\tilde{h}(u)$ and 95% pointwise Monte Carlo intervals (dotted) for case II model with number of internal knots $L = 3,5,7$, respectively.

non-smooth objective function as in (7) with 500 resamples requires about five hours. Clearly, our procedure proposed here is more efficient in computation for practitioners.

4. Application

4.1. Multiple Myeloma Data

The multiple myeloma data of set is the primary example in the PHREG procedure and available in the online SAS/STAT User's Guide, which is analyzed by [8] for the AFT model, and [15] for the AFT-PLM in (1.2). In the study, there are total 65 patients with 48 deaths and 17 survivals. We consider the data consisting of possibly censored survival times (T) and two independent covariates: logBUN, the logarithm of Blood Urea Nitrogen; and age. As pointed out by [15], age variable might be a confounding factor and it is treated as the nonparametric component in the following model,

$$\log(T) = \beta \times \log \mathrm{BUN} + h(\mathrm{age}) + \epsilon.$$

For the internal knots, we choose $L = 3$ or 5 as the number of the internal knots and locate them equally spaced between the smallest and largest observations of age variable. Applying our proposed method, for $L = 3$, $\tilde{\beta} = -1.4965$, its estimated standard error obtained from the sandwich estimator described in Subsection 2.3 is $\tilde{\sigma} = 0.0073$; for $L = 5$, $\tilde{\beta} = -1.5436$, $\tilde{\sigma} = 0.0109$. Note that the two slope estimates are both negative and the corresponding estimated standard error are rather small. That implies that Blood Urea Nitrogen is negatively related the log survival time. Similar conclusion is also attained by [15], where the estimate of β is −1.955 with standard error estimate 0.807. **Figure 3** displays the estimate of the nonparametric part. From it, it is visually justified to render the age variable has a nonlinear effect on the log survival time.

4.2. Nursing Home Usage Data

The data is from an experiment sponsored by National Center for Health Services Research in 1980-1982 designed to determine the effect of financial incentives on variation of patient care in nursing homes, involving 36 for profit nursing homes in San Diego, California. Full description of this data set is given in [26] and available from. The response variable T is measured in days and the total sample size is $n = 1601$. In the model we consider later, several covariates are incorporated, explicitly: treatment, sex, marital status, three health status indicators (HSI), and age,

$$\begin{aligned}\log T = {} & \beta_1 \times \mathrm{treatment} + \beta_2 \times \mathrm{sex} + \beta_3 \times \mathrm{maritalstatus} \\ & + \beta_4 \times \mathrm{HSI}_1 + \beta_5 \times \mathrm{HSI}_2 + \beta_6 \times \mathrm{HSI}_3 + h(\mathrm{age}) + \epsilon,\end{aligned}$$

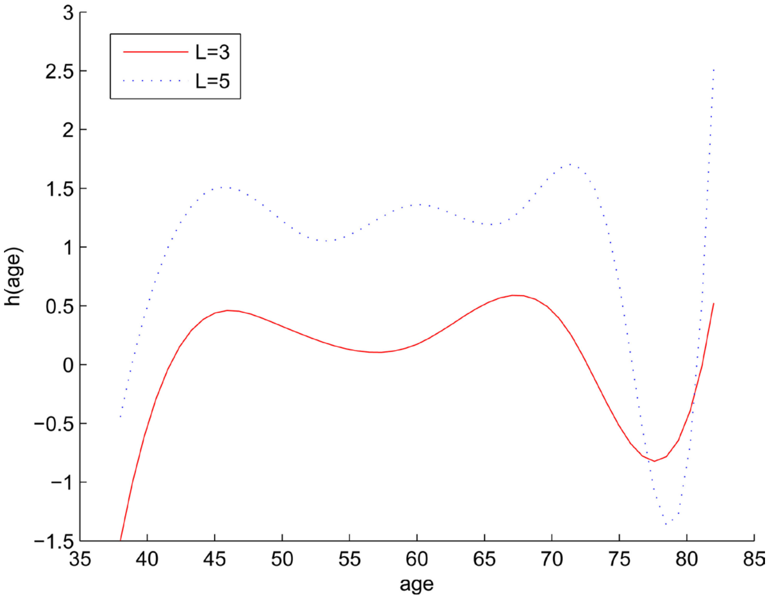

Figure 3. *L*, the number of internal knots used in the analysis of multiple myelome data.

where HSI_1-HSI_3 are three binary health status indicators ranging from the best health to the worst health. When analyzing the data set, we discard ten observations where the observed T is zero, then utilize the remaining 1051 records to accomplish our analysis. Results of coefficient estimates are presented in Table 3. Figure 4 reports the nonparametric part. It seems that the claim that the age variable has a linear effect on the survival time is plausible. The drastic influction is possibly resulted from the fact that there is little observation of age available in the right tail. [9] analyzed this data using the AFT model and found that the age variable is not statistically significant, which agrees with our results.

5. Discussions

The accelerated failure time partial linear model (AFT-PLM) is a natural extension of the classic AFT model, which allows some covariate to relate to the log failure time in a nonlinear manner, and thus provides a more flexible and parsimonious way of modelling. In this paper, we employ the cubic B-splines to approximate the nonparametric smooth function in model (2), and doing so fascinates us to apply the efficient rank-based inference approach for the AFT model into our current situation. Explicitly, based on the recent achievements in dealing with non-smooth estimating function, we propose a polynomial-based smoothing Gehan estimating function and show that the resulting estimators are consistent and asymptotically normal under certain regularity conditions. Utilizing the smoothed version and the fact that it is the gradient of a smooth and convex loss function, to solve the solution to these equations, we develop the

Table 3. Analysis of nursing home usage data.

L	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_3$	$\tilde{\beta}_4$	$\tilde{\beta}_5$	$\tilde{\beta}_6$
3	0.0888	−0.6190	−0.2457	−0.6189	−0.8151	−1.5914
	(0.2743)	(0.2028)	(0.2800)	(0.3339)	(0.2537)	(0.2140)
5	0.0770	−0.6140	−0.2327	0.0341	−0.1552	−0.9387
	(0.2300)	(0.3163)	(0.2418)	(0.2463)	(0.2484)	(0.4121)

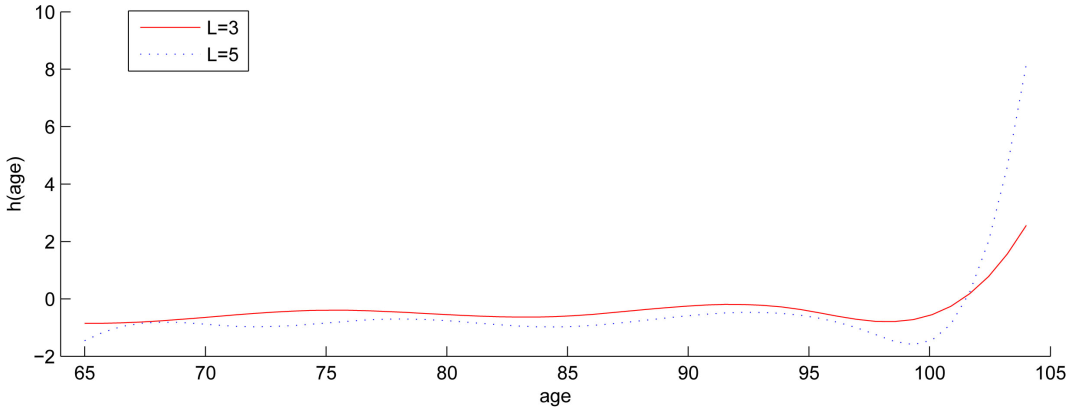

Figure 4. L, the number of internal knots used in the analysis of nursing home usage data.

L-BFGS method. In addition, the other primary advantage of our smoothing proposal is that one can estimate the standard error directly and efficiently through the sandwich-formed covariance matrix without resorting to computationally intensive resampling. As is seen in the simulation studies, our method performs well for estimation of both the regression coefficients and the nonparametric component.

Naturally, our method can be straightforwardly extended to cases where there are more than one covariate which has nonlinear effects. However, the number of the parameters to be estimated is increasing, especially when one chooses more internal knots to approximate those nonparametric components. At this time, it is necessary to incorporate a penalty term into the objective loss function defined in (2.6), and then proceed to make inference. To estimate the joint asymptotic covariance matrix, the sandwich estimator may encounter the unreliable numerical problem due to the high dimension; thus other efficient approaches should be further developed.

Another possible extension to the AFT-PLM in (1.2) is to consider the more general partial linear single index AFT model, which is specified as

$$\log(T) = \beta^{\mathrm{T}} Z + h\left(\alpha^{\mathrm{T}} U\right) + \epsilon ,$$

where U is a vector of nuisance covariates, and Z is a vector of covariates of primary interest, $h(.)$ is an unknown univariate smooth function that plays the role of a link function. This model is well-known due to the fact that it achieves dimension reduction purpose and avoids the "curse of dimensionality". Of course, how to adapt our proposed method to this model is interesting and will be investigated in future.

In this paper, we have assumed that the unknown function is a spline function with fixed number of knots in establishing the asymptotic properties. Through the simulation studies, we find that a few number of knots is enough and the bias caused by the spline approximation is small and doesn't affect the estimate of regression coefficients apparently. For cases without such assumptions, the number of knots should increase as the sample size increases, and developing asymptotic results in that setting is interesting but beyond the scope of this paper.

Foundation Item

The Youth Foundation (ZJGQN004) of school of Zhangjiagang, JUST.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cox, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B* (*Methodological*), **34**, 187-220.

https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[2] Struthers, C.A. and Kalbfleisch, J.D. (1986) Misspecified Proportional Hazard Models. *Biometrika*, **73**, 363-369. https://doi.org/10.1093/biomet/73.2.363

[3] Buckley, J. and James, I. (1979) Linear Regression with Censored Data. *Biometrika*, **66**, 429-436. https://doi.org/10.1093/biomet/66.3.429

[4] Ritov, Y. (1990) Estimation in a Linear Regression Model with Censored Data. *The Annals of Statistics*, **18**, 303-328. https://doi.org/10.1214/aos/1176347502

[5] Tsiatis, A.A. (1990) Estimating Regression Parameters Using Linear Rank Tests for Censored Data. *The Annals of Statistics*, **18**, 354-372. https://doi.org/10.1214/aos/1176347504

[6] Ying, Z. (1993) A Large Sample Study of Rank Estimation for Censored Regression Data. *The Annals of Statistics*, **21**, 76-99. https://doi.org/10.1214/aos/1176349016

[7] Fygenson, M. and Ritov, Y. (1994) Monotone Estimating Equations for Censored Data. *The Annals of Statistics*, **22**, 732-746. https://doi.org/10.1214/aos/1176325493

[8] Jin, Z., Lin, D.Y. and Wei, L.J. (2003) Rank-Based Inference for the Accelerated Failure Time Model. *Biometrika*, **90**, 341-353. https://doi.org/10.1093/biomet/90.2.341

[9] Chung, M., Long, Q. and Johnson, B.A. (2013) A Tutorial on Rank-Based Coefficient Estimation for Censored Data in Small- and Large-Scale Problems. *Statistics and Computing*, **23**, 601-614. https://doi.org/10.1007/s11222-012-9333-9

[10] Chiou, S.H., Kang, S. and Yan, J. (2012) Fast Accelerated Failure Time Modeling for Case-Cohort Data. *Statistics and Computing*, **24**, 559-568.

[11] Brown, B.M. and Wang, Y.G. (2005) Standard Errors and Covariance Matrices for Smoothed Rank Estimators. *Biometrika*, **92**, 149-158. https://doi.org/10.1093/biomet/92.1.149

[12] Brown, B.M. and Wang, Y.G. (2007) Induced Smoothing for Rank Regression with Censored Survival Times. *Statistics in Medicine*, **26**, 828-836. https://doi.org/10.1002/sim.2576

[13] Heller, G. (2007) Smoothed Rank Regression with Censored Data. *Journal of the American Statistical Association*, **102**, 552-559. https://doi.org/10.1198/016214506000001257

[14] Johnson, L.M. and Strawderman, R.L. (2009) Induced Smoothing for the Semiparametric Accelerated Failure Time Model: Asymptotics and Extensions to Clustered data. *Biometrika*, **96**, 577-590. https://doi.org/10.1093/biomet/asp025

[15] Chen, K., Shen, J. and Ying, Z. (2005) Rank Estimation in Partial Linear Model with Censored Data. *Statistica Sinica*, **15**, 767-779.

[16] Hardle, W. and Liang, H. (2007) Partially Linear Models. Springer, Berlin Heidelberg.

[17] Orbe, J., Ferreira, E. and Nez-Anton, V. (2003) Censored Partial Regression. *Biostatistics*, **4**, 109-121. https://doi.org/10.1093/biostatistics/4.1.109

[18] Stute, W. (1993) Consistent Estimation under Random Censorship When Covariables Are Present. *Journal of Multivariate Analysis*, **45**, 89-103. https://doi.org/10.1006/jmva.1993.1028

[19] Zou, Y., Zhang, J. and Qin, G. (2011) A Semiparametric Accelerated Failure Time Partial Linear Model and Its Application to Breast Cancer. *Computational Statistics and Data Analysis*, **55**, 1479-1487. https://doi.org/10.1016/j.csda.2010.10.012

[20] Liu, X., Wang, L. and Liang, H. (2011) Estimation and Variable Selection for Semi-

parametric Additive Partial Linear Models (SS-09-140). *Statistica Sinica*, **21**, 1225. https://doi.org/10.5705/ss.2009.140

[21] Ma, S. and Kosorok, M.R. (2005) Penalized Log-Likelihood Estimation for Partly Linear Transformation Models with Current Status Data. *The Annals of Statistics*, **33**, 2256-2290. https://doi.org/10.1214/009053605000000444

[22] Huang, J.Z. and Liu, L. (2006) Polynomial Spline Estimation and Inference of Proportional Hazards Regression Models with Flexible Relative Risk Form. *Biometrics*, **62**, 793-802. https://doi.org/10.1111/j.1541-0420.2005.00519.x

[23] De Boor, C. (1978) A Practical Guide to Splines. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4612-6333-3

[24] Shang, S., Liu, M. and Zeleniuch-Jacquotte, A. (2013) Partially Linear Single Index Cox Regression Model in Nested Case-Control Studies. *Computational Statistics and Data Analysis*, **67**, 199-212. https://doi.org/10.1016/j.csda.2013.05.011

[25] Nocedal, J. and Wright, S.J. (2006) Numerical Optimization. Springer, New York.

[26] Morris, C.N., Norton, E.C. and Zhou, X.H. (1994) Parametric Duration Analysis of Nursing Home Usage. In: *Case Studies in Biometry*, John Wiley and Sons, New York, 231-248.