Nonparametric regression models for right-censored data using Bernstein polynomials

Muhtarjan Osman *, Sujit K. Ghosh

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA

ARTICLE INFO

Article history:
Received 8 February 2011
Received in revised form 29 June 2011
Accepted 31 August 2011
Available online 18 September 2011

Keywords:
Bernstein polynomials
Censored data
Nonparametric regression
Nonproportional hazards
Sieve

ABSTRACT

In some applications of survival analysis with covariates, the commonly used semiparametric assumptions (e.g., proportional hazards) may turn out to be stringent and unrealistic, particularly when there is scientific background to believe that survival curves under different covariate combinations will cross during the study period. We present a new nonparametric regression model for the conditional hazard rate using a suitable sieve of Bernstein polynomials. The proposed nonparametric methodology has three key features: (i) the smoothest estimator of the conditional hazard rate is shown to be a unique solution of a strictly convex optimization problem for a wide range of applications; making it computationally attractive, (ii) the model is shown to encompass a proportional hazards structure, and (iii) large sample properties including consistency and convergence rates are established under a set of mild regularity conditions. Empirical results based on several simulated data scenarios indicate that the proposed model has reasonably robust performance compared to other semiparametric models particularly when such semiparametric modeling assumptions are violated. The proposed method is further illustrated on the gastric cancer data and the Veterans Administration lung cancer data.

1. Introduction

In regression analysis of survival data, the proportional hazard model (Cox, 1972) has become by far the most widely used method by researchers in many disciplines especially in the field of biostatistics. The most appealing features of the Cox model include the unspecified baseline hazard function and the straightforward interpretation for the effect of categorical covariates such as treatment assignment. In many applications the proportional hazard (PH hereafter) assumption is generally considered reasonable for the cases with a long follow-up time (Perperoglou et al., 2007). In some other situations, however, the validity of this simplification is certainly questionable. It is known that the violation of the PH assumption could lead to erroneous inference in some circumstances (see e.g., Schemper, 1992). In these cases, several alternatives such as proportional odds (PO hereafter) model (Bennett, 1983) and accelerated failure time (AFT hereafter) model (Kalbfleisch and Prentice, 1980) have been proposed. But these semiparametric models may also turn out to be stringent or even unrealistic in some cases. For instance, consider the case of crossing survival curves, in which there is scientific background to believe that survival curves under different covariate combinations will cross during the study period. For example, in the well known gastric cancer clinical trial (Stablein et al., 1981), patients receiving only chemotherapy may have higher survival rates initially but such rates decay much faster compared to the group of patients receiving chemotherapy and radiotherapy. None of the above models (PH, PO, AFT) can accommodate such a feature of the data. The cases with nonproportional hazards can also occur in non-medical applications. For example, Dolton and O’Neill (1996) reported a non-proportional effect of government official counseling on the hazard function of unemployment benefits sign-offs in the United Kingdom.

* Corresponding author.
E-mail addresses: mosman@ncsu.edu (M. Osman), sujit_ghosh@ncsu.edu (S.K. Ghosh).

© 2011 Elsevier B.V. All rights reserved.

doi:10.1016/j.csda.2011.08.019
In the Cox PH model, the conditional hazard function \( h(t|Z) \) is modeled as \( h(t|Z) = h_0(t) \exp(\beta^T Z) \) where \( h_0(\cdot) \) denotes the baseline hazard, \( Z \) represents the vector of covariates and the parameter of interest \( \beta \) is constant over time. Several semiparametric extensions of the Cox PH model have been proposed by various authors to relax the proportionality assumption. The most popular approach is to include a time-varying effect \( \beta(t) \) by replacing \( \beta \) in the Cox PH model. A challenging step in the time-varying coefficient model is the estimation of the effect function \( \beta(\cdot) \). Murphy and Sen (1991) assumed \( \beta(\cdot) \) is piecewise constant and proposed a histogram sieve estimator. Zucker and Karr (1990) used smoothing splines based on partial likelihood. The time-varying coefficient model has received extensive attention in literature recently. For further details on more recent approaches on this subject, we refer to the papers by Martinussen and Scheike (2002), Cai and Sun (2003) and Tian et al. (2005). Most of the methodologies involving the time-varying effect may turn out to be computationally intensive because the form of likelihood function is usually very complicated. Besides, Perperoglou et al. (2007) pointed out that when survival curves cross, over-emphasizing on the regression coefficient might not be appropriate. The reason is that the appealing feature of easy interpretation and estimation of the Cox PH model, which comes from the separation of time and covariate effects, will be lost under nonproportional hazards. Instead, the entire conditional hazard or survival curve will be more informative to medical researchers. Therefore, in this paper we take a different direction by directly modeling the conditional hazard function using Bernstein polynomials.

In contrast to various extensions of the Cox PH model to accommodate nonproportionality, our method is completely nonparametric and computationally much simpler to implement. Although nonparametric models suffer from the curse of dimensionality, as Spierdijk (2008) pointed out they can serve as starting points for building parametric or semiparametric models in high dimensions. Fully nonparametric hazard regression models have been studied by many authors, most of which focused on the kernel method and smoothing splines (see e.g., Li and Doss, 1995; Gray, 1996; Spierdijk, 2008, among others). As pointed out by Tenbusch (1994, 1997), the estimators based on Bernstein polynomials can be regarded as kernel-based estimators with spatial adaptive polynomials as weight or kernel functions. So instead of using the same kernel function, the kernel function in the Bernstein polynomial based estimator adapts itself to the positions of the knots specified. It is also this property makes the estimators based on Bernstein polynomials enjoy better boundary behavior than the usual kernel-based estimators. Another important methodology in this line of research is referred to as HARE (Hazard Regression) in Kooperberg et al. (1995). HARE is a regression model based on linear splines and their tensor products for the conditional log-hazard function. In the HARE model, linear splines are used rather than quadratic or cubic splines in order to avoid numerical integration in the log-likelihood (and in its gradient and Hessian matrix). This simplification is essentially due to the model selection step involving stepwise addition and stepwise deletion incorporated in HARE. However, in some situations where the conditional log-hazard function takes complex form the linear splines and their tensor products may not capture the overall dependence of the event time on other covariates.

Bernstein polynomials have been considered in a wide range of statistical problems based on completely observed data. The most common application is density estimation, which dates back to the work of Vitale (1975). Some of the most recent work on this topic includes Petrone (1999), Babu et al. (2002) and Choudhuri et al. (2004) among many others. Bernstein polynomials have also been applied in the regression setting by Tenbusch (1997) and Chang et al. (2007). In the context of survival analysis with censored data, Chang et al. (2005) proposed using Bernstein polynomials for hazard rate estimation in a Bayesian framework for a homogeneous population, i.e., without any covariates.

In this paper, we consider nonparametric hazard regression based on Bernstein polynomials for right-censored data. As we will demonstrate later, Bernstein polynomials have several advantages in this particular setting. Monotonicity of the cumulative hazard function can be modeled naturally via Bernstein polynomials. In addition, Bernstein polynomials have nice differentiability properties such that the log-likelihood, its gradient, and Hessian matrix all take relatively easy forms, making our method very easy to implement as compared to other computationally intensive methods such as those based on the time-varying coefficient models. To obtain a smooth estimator for the conditional hazard function in a full nonparametric setting, we use a sieve maximum likelihood estimator (Grenander, 1981; Geman and Hwang, 1982). The proposed nonparametric regression model in this paper is shown to encompass a proportional hazards structure. The rest of the paper proceeds as follows. In Section 2, we describe the model for categorical and continuous covariates. We show that the sieve maximum likelihood estimate is consistent and the corresponding rate of convergence is derived in Section 3. In Section 4, the proposed method is demonstrated through simulated data as well as real data from two well known cancer studies. Finally, we conclude with some discussions in Section 5.

2. Conditional hazard model using Bernstein polynomials

First, we consider the one-sample right-censored data with no covariates. Suppose an experiment or a clinical trial consists of \( n \) subjects, \( T_i \) denotes the continuously distributed time to certain event of interest for the subject \( i \) where \( i = 1, 2, \ldots, n \). The event time \( T_i \) is subject to random right-censoring \( C_i \) and hence for each subject we observe \((X_i, \Delta_i)\), where \( X_i = \min(T_i, C_i) \), \( \Delta_i = I(T_i \leq C_i) \), and \( I(\cdot) \) denotes the indicator function that takes the value 1 when the event \( A \) is true, otherwise \( I(A) = 0 \). We also assume that \( T_i \) is statistically independent of \( C_i \) for each \( i = 1, 2, \ldots, n \). For any \( t \geq 0 \), the cumulative hazard function is given by \( H(t) = -\log S(t) \) and the hazard function \( h(t) = \dot{H}(t) \), where \( S(t) = \Pr(T_i > t) \) is the survival function and \( H(t) \) denotes the derivative of \( H(t) \). Further, following standard practice (e.g. Tian et al., 2005) we assume that there exists a \( \tau < \infty \) such that \( \tau = \inf\{t : S(t) = 0\} \).
Now we define the model for the hazard function as
\[
    h_{m}(t, \gamma) = \sum_{k=1}^{m} \gamma_k g_{m,k}(t) = \gamma^T g_m(t), \quad 0 \leq t < \infty, \tag{1}
\]
where the coefficients \(\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)^T\) satisfy \(\gamma_k \geq 0\) for all \(k\)'s and \(g_m(t) = (g_{m,0}(t), g_{m,1}(t), \ldots, g_{m,m}(t))^T\) with the base functions satisfying \(g_{m,k}(\cdot) \geq 0\) and \(\int_0^\infty g_{m,k}(u)du < \infty\) for all \(k \leq m\). Therefore, we have the corresponding model for the cumulative hazard function
\[
    H_{m}(t, \gamma) = \sum_{k=1}^{m} \gamma_k G_{m,k}(t) = \gamma^T G_m(t), \quad 0 \leq t < \infty, \tag{2}
\]
where \(G_m(t) = (G_{m,0}(t), G_{m,1}(t), \ldots, G_{m,m}(t))^T\) with \(G_{m,k}(t) = \int_0^t g_{m,k}(u)du\). Clearly, the monotonicity of \(H_m(\cdot)\) is enforced by the restriction that \(\gamma_k \geq 0\) and \(g_{m,k}(\cdot) \geq 0\) for \(k = 1, 2, \ldots, m\). Although the above model for \(H_m(\cdot)\) does not satisfy the requirement that \(H_m(\tau) = \infty\), a simple tail adjustment can be made to satisfy the requirement (see Section 2.4 for further details).

In this paper, we construct the base functions \(g_{m,k}(\cdot)\) and \(G_{m,k}(\cdot)\) using Bernstein polynomials, which according to a result by Carnicer and Peña (1993) has the best shape-preserving property among all approximation polynomials. For a continuous function such as a cumulative hazard \(H(\cdot)\) on \((0, \tau]\), the approximating Bernstein polynomial of order \(m\) is given by
\[
    B(t; m, H) = \sum_{k=0}^{m} H\left(\frac{k}{m}\right)\left(\frac{m}{k}\right)(t/\tau)^k(1-t/\tau)^{m-k}.
\]
By the Weierstrass theorem, \(B(\cdot; m, H) \rightarrow H(\cdot)\) uniformly over \((0, \tau]\) as \(m \rightarrow \infty\) (Lorentz, 1953). Also, the derivative of the Bernstein polynomial for \(H\) can be written as
\[
    \dot{B}(t; m, H) = \frac{\partial B(t; m, H)}{\partial t} = \sum_{k=1}^{m} \left\{ H\left(\frac{k-1}{m}\right) - H\left(\frac{k}{m}\right) \right\} \frac{f_{\beta}(t/\tau; k, m - k + 1)}{\tau},
\]
where \(f_{\beta}(\cdot; k, m - k + 1)\) is the probability density function of the Beta distribution with shape parameters \(k\) and \(m - k + 1\). It is also well known that \(\dot{B}(\cdot; m, H) \rightarrow h(\cdot)\) uniformly on \((0, \tau]\) as \(m \rightarrow \infty\) (Lorentz, 1953).

Now we let the base function \(g_{m,k}(t) = f_{\beta}(t/\tau; k, m - k + 1)/\tau\) in \((1)\), so
\[
    G_{m,k}(t) = \int_0^t g_{m,k}(u)du = \int_0^t f_{\beta}(u/\tau; k, m - k + 1)d(u/\tau),
\]
i.e., \(G_{m,k}(t)\) becomes the cumulative distribution function of the Beta distribution with shape parameters \(k\) and \(m - k + 1\) evaluated at \(t/\tau\). Consequently, the log-likelihood function of \(\gamma\) corresponding to the models \((1)\) and \((2)\) can be written as
\[
    l(\gamma) = \sum_{i=1}^{n} \left\{ \Delta_i \log(h_m(X_i, \gamma)) - H_m(X_i, \gamma) \right\} = \sum_{i=1}^{n} \left\{ \Delta_i \log(U_i^T \gamma) - V_i^T \gamma \right\}, \tag{3}
\]
where \(\gamma \in \mathcal{C}_m = [0, \infty)^m, U_i = g_m(X_i), \) and \(V_i = G_m(X_i)\).

**Remark 1.** Since \(l(\gamma)\) as defined in \((3)\) is a strictly concave function and \(l(\gamma) \rightarrow -\infty\) as \(\gamma \rightarrow \partial \mathcal{C}_m\), where \(\partial \mathcal{C}_m\) denotes the boundary of \(\mathcal{C}_m\), the existence and uniqueness of the maximum likelihood estimator \(\hat{\gamma} = \arg \max_{\gamma \in \mathcal{C}_m} l(\gamma)\) follow immediately.

### 2.1. Categorical covariate

For simplicity, first we consider a dichotomous covariate, say \(Z \in \{0, 1\}\). For example, in most medical applications \(Z\) denotes the assignment of the treatment group: 0 for the placebo and 1 for the active treatment. Essentially, the model for the hazard function and the cumulative hazard function can be easily extended by using different sets of parameter vectors for different groups as it is in \((1)\) and \((2)\), namely,
\[
    h_m(t, \gamma|Z) = ((1 - Z)\gamma_0^T + Z\gamma_1^T)g_m(t) \quad \text{and} \quad H_m(t, \gamma|Z) = ((1 - Z)\gamma_0^T + Z\gamma_1^T)G_m(t), \tag{4}
\]
where $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T = (y_{01}, y_{02}, \ldots, y_{0m}, y_{11}, y_{12}, \ldots, y_{1m})^T$. $g_{m,k}(\cdot)$ and $G_{m,k}(\cdot)$ are the same as they were defined in the previous section. Accordingly, the log-likelihood function becomes

$$
  l(\mathbf{y}) = \sum_{i=1}^{n} \left\{ \Delta_i \log(h_m(X_i, \mathbf{y}|Z_i)) - H_m(X_i, \mathbf{y}|Z_i) \right\}
$$

$$
  = \sum_{i=1}^{n} \left\{ \Delta_i \log(U_i^T \mathbf{y}) - V_i^T \mathbf{y} \right\},
$$

(5)

where

$$
  U_i = \left[ \begin{array}{c} (1 - Z_i)g_m(X_i) \\ Z_i g_m(X_i) \end{array} \right] \quad \text{and} \quad V_i = \left[ \begin{array}{c} (1 - Z_i)G_m(X_i) \\ Z_i G_m(X_i) \end{array} \right].
$$

The log-likelihood function in (5) is of the same form as in the case of one-sample data with no covariates (see Eq. (3)), so the existence and uniqueness of the maximum likelihood estimator still hold in the presence of a binary covariate.

The model described in (4) can be regarded as modeling the discretized hazard function using 1-way ANOVA. As a result, it can be further extended to the cases when there are multiple categorical covariates and each has more than 2 levels. Generally, multiple categorical covariates can be summarized in terms of one categorical covariate taking multiple levels that index all the possible combinations of different levels in multiple categorical covariates. Suppose $Z \in \{1, 2, \ldots, J\}$ summarizes all categorical covariates, then the model in (4) can be expressed as

$$
  h_m(t, \mathbf{y}|Z) = \left\{ \sum_{j=1}^{J} I(Z = j) \mathbf{y}_j^T \right\} g_m(t) \quad \text{and}
$$

$$
  H_m(t, \mathbf{y}|Z) = \left\{ \sum_{j=1}^{J} I(Z = j) \mathbf{y}_j^T \right\} G_m(t),
$$

(6)

where $\mathbf{y}_j = (y_{1j}, y_{2j}, \ldots, y_{mj})^T$ and $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_J^T)^T$. Again, the corresponding log-likelihood function can be written in the form as in (5) and hence the existence and uniqueness of maximum likelihood estimate $\hat{\mathbf{y}}$ follow by the Remark 1 stated earlier.

2.2. Continuous covariate

Next we consider the case when $Z$ is a continuous covariate. As natural extensions of (1) and (2), the models for the hazard function and the cumulative hazard function can be written as

$$
  h_m(t, \mathbf{y}|Z) = \sum_{k=1}^{m} \gamma_k(Z)g_{m,k}(t) = \mathbf{y}(Z)^T g_m(t) \quad \text{and}
$$

$$
  H_m(t, \mathbf{y}|Z) = \sum_{k=1}^{m} \gamma_k(Z)G_{m,k}(t) = \mathbf{y}(Z)^T G_m(t).
$$

(7)

One possible approach of conditioning on a continuous covariate is to discretize it into several categories, then it could be fitted into the framework described in the previous section. In order to capture the overall dependence between $Z$ and the response (via the conditional hazard or cumulative hazard functions), it may require a relatively large number of discretized categories. As a consequence, it may lead to small number of samples within each category. Instead of modeling $\gamma_k(Z)$ as a linear function of $Z$, we propose approximating the function $\gamma_k(\cdot)$ with another $\tilde{m}$-th order Bernstein polynomial. To implement this, a transformation $\tilde{Z} = a(Z)$ is required to map the continuous covariate to the unit interval $(0, 1)$. Then the second fold of Bernstein polynomial can be written as $\gamma_k(Z) = \sum_{j=1}^{m} \gamma_{kj} \tilde{Z}^j(1 - \tilde{Z})^{\tilde{m}-j}$. If we let $\tilde{m} = m - 1$, the expression can be rewritten as

$$
  \gamma_k(Z) = \sum_{j=1}^{m} \gamma_{kj} \tilde{g}_{m,j}(\tilde{Z}) = \mathbf{y}_k^T \tilde{g}_m(\tilde{Z}),
$$

where $\tilde{g}_{m,j}(\tilde{Z}) = f_{\beta}(\tilde{Z}, j, m - j + 1)/m$. Note that even though $\tilde{g}_{m,j}(\cdot)$ and $g_{m,j}(\cdot)$ take very similar forms, they are not exactly identical because the function $g_{m,j}(\cdot)$ involves the cut-off time point $\tau$. We can now express the models as

$$
  h_m(t, \mathbf{y}|Z) = \sum_{k=1}^{m} \sum_{j=1}^{m} \gamma_{kj} \tilde{g}_{m,j}(\tilde{Z})g_{m,k}(t) \quad \text{and}
$$

$$
  H_m(t, \mathbf{y}|Z) = \sum_{k=1}^{m} \sum_{j=1}^{m} \gamma_{kj} \tilde{g}_{m,j}(\tilde{Z})G_{m,k}(t),
$$

(8)
where the unknown parameter $\gamma = (\gamma_1^T, \gamma_2^T, \ldots, \gamma_m^T)$ with $\gamma_k = (\gamma_{k1}, \gamma_{k2}, \ldots, \gamma_{km})^T$ for $k = 1, 2, \ldots, m$. Notice that it is not necessary to use $m = m - 1$, but we find it convenient to implement our method with one tuning parameter $m$.

Then the log-likelihood function becomes

$$l(\gamma) = \sum_{i=1}^{n} \Delta_i \log(\text{in}(X_i, \gamma | Z_i)) - \text{Homo}(X_i, \gamma | Z_i)$$

$$= \sum_{i=1}^{n} \{\Delta_i \log(U_i^T \gamma) - V_i^T \gamma\},$$

(9)

where $U_i = \text{vec}[\text{gin}(\tilde{Z}) \text{gin}(X_i)]$, $V_i = \text{vec}[\text{gin}(\tilde{Z}) \text{gin}(X_i)]$, and $\text{vec}[\cdot]$ denotes the vectorization by column operator applied to a matrix. Similarly, we impose the restriction $\gamma_{kj} > 0$ for $1 \leq k, j \leq m$ for $m = 1, 2, \ldots$. Because the above form is the same as in (3), it immediately guarantees the existence and uniqueness of maximum likelihood estimate for a given $m$.

As a special case of our model in (8), when the time effect and the covariate effect can be factored as $\gamma_{kj} = \gamma_k \tilde{Z}_j$ in (8), the model simplifies to

$$h_m(t, \gamma | Z) = \left\{ \sum_{k=1}^{m} \gamma_k \text{gin}_m(k)(t) \right\} \left\{ \sum_{j=1}^{m} \tilde{Z}_j \text{gin}_m(j)(\tilde{Z}) \right\} = h_0(t) \exp[\mu(\gamma(Z))] \quad \text{and}$$

$$H_m(t, \gamma | Z) = \left\{ \sum_{k=1}^{m} \gamma_k \text{gin}_m(k)(t) \right\} \left\{ \sum_{j=1}^{m} \tilde{Z}_j \text{gin}_m(j)(\tilde{Z}) \right\} = H_0(t) \exp[\mu(\gamma(Z))].$$

which is a proportional hazard model with the baseline hazard approximated by Bernstein polynomials and the covariate effect $\mu(Z) = \log \left[ \sum_{j=1}^{m} \tilde{Z}_j \text{gin}_m(j)(\tilde{Z}) \right]$. Notice that the Cox proportional hazard model is a special case of this model when $\mu(\cdot)$ is assumed linear in $Z$.

2.3. Categorical and continuous covariates

When both a categorical covariate $Z_1$ and a continuous covariate $Z_2$ are present, generally we can stratify the data based on the levels of the categorical covariate $Z_1$ and fit a separate model with the continuous covariate $Z_2$ at each stratum. This stratified approach is suitable for large sample sizes where there are enough data at each stratum. But for moderate or small datasets, some subgroup may result in only a few samples so that corresponding $Z_2$’s (or transformed version $\tilde{Z}_2$) may not adequately cover the entire unit interval $(0, 1)$. Alternatively, if we are willing to accept the assumption that there is no interaction between the categorical covariate $Z_1$ and the continuous covariate $Z_2$, we can write a partially linear coefficient model as

$$h_m(t, \gamma | Z_1, Z_2) = \gamma(Z_1, Z_2)^T \text{gin}_m(t) = [\phi(Z_1) + \psi(Z_2)^T] \text{gin}_m(t) \quad \text{and}$$

$$H_m(t, \gamma | Z_1, Z_2) = [\phi(Z_1) + \psi(Z_2)^T] \text{gin}_m(t),$$

(10)

where $\gamma(Z_1, Z_2) = (\gamma_{11}(Z_1, Z_2), \gamma_{12}(Z_1, Z_2), \ldots, \gamma_{m1}(Z_1, Z_2)^T)$ with $\gamma_{kj}(Z_1, Z_2) = \phi_k(Z_1) + \psi_k(Z_2)$, $\phi_k(Z_1)$ is just an ANOVA model with mean effects at the levels of $Z_1$, and $\psi_k(Z_2)$ is a smooth function. Suppose $Z_1$ has $J$ levels, then we can write $\phi(Z_1)^T \text{gin}_m(t) = U_1^T \mu$ and $\psi(Z_1)^T \text{gin}_m(t) = V_1^T \mu$, where $\mu = (\mu_1^T, \mu_2^T, \ldots, \mu_J^T)^T$ with $\mu_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jm})^T$.

$$U_1 = \begin{bmatrix}
I(Z_1 = 1) \text{gin}_m(t) \\
I(Z_1 = 2) \text{gin}_m(t) \\
\vdots \\
I(Z_1 = J) \text{gin}_m(t)
\end{bmatrix} \quad \text{and} \quad V_1 = \begin{bmatrix}
I(Z_1 = 1) \text{gin}_m(t) \\
I(Z_1 = 2) \text{gin}_m(t) \\
\vdots \\
I(Z_1 = J) \text{gin}_m(t)
\end{bmatrix}. $$

Similar to the single continuous covariate case, we can write $\psi(Z_2)^T \text{gin}_m(t) = U_2^T w$ and $\psi(Z_2)^T \text{gin}_m(t) = V_2^T w$, where $w = (w_1^T, w_2^T, \ldots, w_m^T)^T$ with $w_k = (w_{k1}, w_{k2}, \ldots, w_{km})^T$, $U_2 = \text{vec} [\text{gin}_m(Z_2) \text{gin}_m(t)]$ and $V_2 = \text{vec} [\text{gin}_m(Z_2) \text{gin}_m(t)]$. Therefore, the partially linear coefficient model can be written in a unified form as

$$h_m(t, \gamma | Z_1, Z_2) = U^T \gamma \quad \text{and} \quad H_m(t, \gamma | Z_1, Z_2) = V^T \gamma,$$

where $\gamma = (\mu^T, w^T)^T$, $U = [U_1^T, U_2^T]^T$, and $V = [V_1^T, V_2^T]^T$. The log-likelihood function again takes the form as in (3) and hence the Remark 1 applies to this case as well.

Thus, we have shown that for a broad range of applications including categorical and continuous predictors, the resulting log-likelihood function can be written in the simple form (3) as in the case of a one-sample problem. This attractive feature of our method makes our nonparametric approach very easy to implement requiring a simple numerical iterative method to estimate the conditional hazard function without making any stringent restrictive assumptions.
2.4. Computational details

As discussed in the previous sections, the log-likelihood function in all cases can be written in a general form (3) as 
\[ l(\mathbf{y}) = \sum_{i=1}^{n} (\Delta_i \log(U_i^T \mathbf{y}) - V_i^T \mathbf{y}), \]
where \( U_i \) and \( V_i \) are some vectors depending only on data. Consequently, the gradient and the Hessian matrix of the log-likelihood function can be conveniently expressed as
\[
\frac{\partial l}{\partial \mathbf{y}} = \sum_{i=1}^{n} \left( \frac{\Delta_i U_i}{U_i^T \mathbf{y}} - V_i \right) \quad \text{and} \quad \frac{\partial^2 l}{\partial \mathbf{y} \partial \mathbf{y}^T} = - \sum_{i=1}^{n} \frac{\Delta_i U_i U_i^T}{(U_i^T \mathbf{y})^2}.
\]

Although the gradient of the log-likelihood function has a trackable form, it still seems difficult if not impossible to solve the likelihood equation analytically. However, since the gradient vector and the Hessian matrix are in simple closed forms, one can easily implement Newton–Raphson method or its various extensions to compute the maximum likelihood estimate \( \hat{\mathbf{y}} \). For example, the maximum can be obtained iteratively through 
\[
\hat{\mathbf{y}}^{(r+1)} = \hat{\mathbf{y}}^{(r)} - \left[ \frac{\partial^2 l}{\partial \mathbf{y} \partial \mathbf{y}^T} \right]^{-1} \frac{\partial l}{\partial \mathbf{y}} \bigg|_{\hat{\mathbf{y}}^{(r)}},
\]
where superscript \( r \) denotes the value at the \( r \)-th iteration starting with say \( \hat{\mathbf{y}}^{(0)} = 1 \). However, for the implementation in this paper we use a modification of the quasi-Newton method given by Byrd et al. (1995), which is available as an option in the R function “optim”. Finally, the variance–covariance matrix of \( \hat{\mathbf{y}} \) can be approximated by inverting the observed information matrix: 
\[
\hat{V}(\hat{\mathbf{y}}) = \left[ \sum_{i=1}^{n} \frac{\Delta_i U_i U_i^T}{(U_i^T \mathbf{y})^2} \right]^{-1}.
\]

In the previous section we mentioned that a suitable transformation \( a(\cdot) \) is needed to map the continuous covariate into the unit interval \((0, 1)\). Instead of a non-linear transformation we choose \( a(\cdot) \) such that \( \hat{Z} = a(Z) \) is approximately linear in \( Z \) on the observed range of \( Z \)‘s. Although many other transformations are possible, throughout this paper we use 
\[
\hat{Z} = \frac{Z - Z_{(1)} + \epsilon}{Z_{(n)} - Z_{(1)} + 2\epsilon},
\]
where we let \( \epsilon = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2} \). \( Z_{(1)} = \min\{Z_1, Z_2, \ldots, Z_n\} \), and \( Z_{(n)} = \max\{Z_1, Z_2, \ldots, Z_n\} \).

For the value of the cut-off time point, we let \( \hat{\tau} = X_{(n)} = \max\{X_1, X_2, \ldots, X_n\} \) which denotes the largest observed value. As described in the previous sections, we model the hazard function and the cumulative hazard function over \((0, \hat{\tau})\) using Bernstein polynomials. When \( \tau > \hat{\tau} \), since we do not have any information from data about the distribution of the event time, we simply let the hazard rate stay constant after the cut-off time point \( t = \hat{\tau} \). The overall hazard function over \( t \in [0, \infty) \) can be written as
\[
h_m(t, \mathbf{y}|Z) = \begin{cases} h_m(t, \mathbf{y}|Z), & \text{if } 0 \leq t < \hat{\tau} \\ \frac{m\gamma_m(Z)}{\tau}, & \text{if } t \geq \hat{\tau}. \end{cases}
\]
Correspondingly, the cumulative hazard function over \( t \in [0, \infty) \) becomes
\[
H_m(t, \mathbf{y}|Z) = \begin{cases} h_m(t, \mathbf{y}|Z), & \text{if } 0 \leq t < \hat{\tau} \\ \sum_{k=1}^{m} \gamma_k(Z) + \frac{m\gamma_m(Z)}{\tau} (t - \tau), & \text{if } t \geq \hat{\tau}. \end{cases}
\]

Now we have \( \int_0^{\infty} h_m^*(t, \mathbf{y}|Z) dt = \infty \) for any \( \mathbf{y} \) and \( Z \), so the tail adjustment suffices that the overall hazard \( h_m^*(\cdot) \) is a legitimate hazard function. On the other hand, since we let \( \hat{\tau} = X_{(n)} \), all the data entering into the likelihood are modeled by Bernstein polynomials, hence the properties of the model in the previous and the following sections remain unaffected.

As a side note, it is easy to verify that \( \hat{\tau}_n \to \hat{\tau} \) as \( n \to \infty \) where \( \tau = \inf\{t > 0 : S(t) = 0\} \).

3. Asymptotic properties

In this section we present the asymptotic properties of the sieve maximum likelihood estimator for the hazard rate of the event time: \( \hat{h}_m(t) = h_m(t, \mathbf{y} = \hat{\mathbf{y}}_{me}) \). For brevity, we only show the consistency and the rate of convergence for the one-sample case. Since the log-likelihood function takes similar forms as in (3) even in presence of categorical or continuous covariates, the asymptotic properties verified in the section can be readily extended to the case of regression model if we assume that the observations \( (X_i, \Delta_i, Z_i) \) are i.i.d. and \( T_i \) is independent of \( C_i \) given \( Z_i \) for each \( i = 1, 2, \ldots, n \).

In the context of right-censored data with no covariate, for each subject we observe i.i.d. copies of \( W = (X, \Delta) \) where \( X = \min(T, C) \) for the event (or failure) time \( T \) and the random censoring time \( C \) and \( \Delta = I(T \leq C) \). Throughout this article we assume that the distributions of \( T \) and \( C \) are dominated by Lebesgue measure. We further assume that the censoring is noninformative, that is, \( T \) and \( C \) are independent. In the presence of covariate \( Z \) we assume that \( T \) and \( C \) are conditionally independent given \( Z \). We let \( h_0(\cdot) \) denote the true hazard function of \( T \) and \( h_1(\cdot) \) denote the hazard function of \( C \).
Now suppose the data \( W_i = (X_i, \Delta_i) \overset{i.i.d.}{\sim} P_\theta \) and if we let \( \mu \) be the product measure of Lebesgue measure and counting measure on \([0, 1])\], we have the density of \( W_i\):

\[
p_\theta(w = (x, \delta)) = \frac{dP_\theta}{d\mu}(x, \delta) = \left\{ \frac{\partial(x)^5}{\exp \left( -\int_0^x \theta(u)du \right)} \right\} * \left\{ h_\gamma(x) \right\}^{1-\delta} \left\{ \int_0^x h_\gamma(u)du \right\},
\]

where the hazard rate of the censoring time \( h_\gamma(\cdot) \) can be regarded as a nuisance parameter and as it does not enter into the calculation of the maximum likelihood estimator we can ignore estimating this function. For brevity, let \( f(x, \delta) = \theta(x)^5 \exp \left\{ -\int_0^x \theta(u)du \right\} \) and \( g(x, \delta, h_\gamma) = h_\gamma(x)^{1-\delta} \exp \left\{ -\int_0^x h_\gamma(u)du \right\}. \) Finally, \( E_0[\cdot] \) and \( \text{VAR}_0[\cdot] \) denote the expectation and the variance operator with respect to the true density of \( W \): \( p_\theta(w) = f(h_0, x, \delta) * g(h_\gamma, x, \delta). \) We also assume that

(1) \( \tau = \inf \{ t > 0 : \int_0^t h_0(u)du = \infty \} < \infty. \)

We study the consistency and the rate of the convergence using the Hellinger distance as the metric of choice

\[
d(\theta_1, \theta_2) = \left\{ \int \left( p_{\theta_1}^{1/2} - p_{\theta_2}^{1/2} \right)^2 d\mu \right\}^{1/2},
\]

where \( \theta_1 = \theta_1(\cdot) \) and \( \theta_2 = \theta_2(\cdot) \) denote two hazard functions in the parameter space \( \Theta \) to be defined shortly. There are several reasons for using this metric. Although we are estimating the hazard function, a more natural and practical entity of interest to the medical researchers is the comparison of the survival curves, or ultimately the global behavior of the data. As pointed out by van de Geer (2000), the Hellinger distance is most convenient in this situation. Additionally, the Hellinger distance has the advantage that it is always well defined for arbitrary continuous density functions.

To study the asymptotic properties of \( \hat{h}_m(\cdot) \), the following conditions on the boundedness and smoothness of the true hazard function \( h_0 \) are assumed.

(II) \( h_0(\cdot) \) is continuous on \([0, \tau] \) and there exists \( \varepsilon > 0 \) such that \( h_0(t) \geq \varepsilon \) for all \( t \in [0, \tau] \).

(III) The first derivative of \( h_0(\cdot) \), denoted by \( h_0^{(1)}(\cdot) \), is Holder continuous with the exponent \( \alpha_0 \), i.e., \( |h_0^{(1)}(t_1) - h_0^{(1)}(t_2)| \leq M|t_1 - t_2|^\alpha_0 \) for some \( \alpha_0 \in (0, 1) \) and for all \( t_1, t_2 \in (0, \tau) \), and \( M \) is some constant.

Therefore, the parameter space is given by

\[
\Theta = \{ \theta(\cdot) \in C[0, \tau] : \theta(\cdot) \text{ satisfies (I)-(III)} \}.
\]

The log-likelihood function based on the data \( W = (W_1, W_2, \ldots, W_n) \) is \( l_n(\theta, \mathbf{w}) = \sum_{i=1}^n \log f(X_i, \Delta_i, \theta) \) where \( \theta(\cdot) \) denotes the hazard function. Note that the part of the density function \( p_\theta \) is not included in the likelihood since it does not involve \( \theta \) hence it does not enter into the maximum likelihood calculation. The estimate of the hazard function is obtained by maximizing \( l_n(\mathbf{w}, \theta) = \sum_{i=1}^n l(X_i, \Delta_i, \theta) \) where \( l(Y, \Delta, \theta) = \Delta \log \theta(Y) - \int_0^Y \theta(u)du. \) If the maximization is taken over the infinite dimensional space \( \Theta \), it imposes great difficulty and it could lead to inconsistent estimate (Geman and Hwang, 1982).

Alternatively, we can construct a subspace to approximate the parameter space \( \Theta \) and carry out the maximization over this much smaller finite dimensional space so called the sieve (Grenander, 1981). Here we construct the sieve as

\[
\Theta_m = \left\{ \theta_m(t) = \sum_{k=1}^m \gamma_k g_{m,k}(t) : \mathbf{y} = (\gamma_1, \gamma_2, \ldots, \gamma_m)^T \in [0, 1]^m \right\},
\]

where the dimension of the space, \( m \), is a function of \( n \), say \( m(n) \) and again \( g_{m,k}(t) = f_k(t/\tau; k, m - k + 1)/\tau. \) We further assume \( m = m(n) = o(n^\kappa) \) for some \( \kappa > 0 \) that will be determined later. It follows that \( \Theta_m \subset \Theta \) for all \( m \). For an arbitrary element in the sieve \( \theta_m \in \Theta_m \), we have \( \theta_m = \sum_{k=1}^m \gamma_k g_{m,k}(t) \) and the true hazard rate \( h_0(\cdot) \) is assumed continuous by the condition (II), so it follows from the Weierstrass theorem that there exists \( \hat{h}_m \in \Theta_m \) such that \( d(\hat{h}_m, h_0) \to 0 \) as \( m \to \infty \), that is, \( \cup_{m=1}^\infty \Theta_m \) is dense in \( \Theta \) with respect to the Hellinger metric.

Let \( \hat{h}_{m,n} = \arg \max_{\hat{h}_m \in \Theta_m} l(\hat{h}, \mathbf{w}) \) be the sieve maximum likelihood estimator. We know that \( \hat{h}_{m,n}(\cdot) \) exists and it is unique (see Remark 1). The following results further imply it also consistent and give its rate of convergence.

**Theorem 1** (Consistency). Suppose the conditions (I)–(II) hold and the sieve \( \Theta_m \) is defined as in (14), then \( \hat{h}_{m,n}(\cdot) \to_{a.s.} h_0(\cdot) \) as \( m, n \to \infty \).

**Theorem 2** (Rate of Convergence). Suppose the conditions (I)–(III) hold and the sieve \( \Theta_m \) is defined as in (14), if \( m = o(n^\kappa) \) with \( \kappa = 2/5 + 2\alpha_0 \), then \( d(\hat{h}_{m,n}, h_0) = O_p \left( n^{-\frac{1+\alpha_0}{3+10\alpha_0}} \right) \).

The proofs of these two theorems are given in the Appendix A.
4. Numerical examples

4.1. Simulated data

We conducted two simulation studies to investigate the empirical performance of the proposed Bernstein polynomial based regression model and to compare it with some popular nonparametric or semiparametric models. As mentioned earlier, we focus on the estimation of the survival function, which should give clinical practitioners more information under nonproportional hazards. Therefore, we use the integrated absolute error (IAE) defined by

\[
IAE = \int_0^T |\hat{S}(t) - S_0(t)|dt \approx \frac{1}{J} \sum_{j=1}^J |\hat{S}(t_j) - S_0(t_j)|
\]

as the criterion to measure the performance of different methods studied, where \(t_0 = 0, t_j = t_{j-1} + \frac{1}{J}, t_J \) is a time point such that \(S_0(t_J) \leq 0.001\), \(\hat{S}(\cdot)\) denotes the estimated survival function, \(S_0(\cdot)\) is the true survival function, and \(J = 1000\) is the number of equally-spaced time points at which we evaluate the survival functions. For both studies, we generated the random censoring time \(C\) following an exponential distribution with rate \(\lambda_c\) (i.e., with the hazard \(h_c(t) = \lambda_c\)). We adjusted the value of \(\lambda_c\) in different scenarios to approximately control the censoring rate \(R_c\) to take values \(\{0, 0.30, 0.50\}\), which represent completely observed, moderately censored, and severely censored data scenarios, respectively. We set the order of Bernstein polynomials \(m = \lceil n^{0.5} \rceil\), where \(\lceil \cdot \rceil\) denotes rounding up to the nearest integer. As discussed at the end of the Appendix A, the choice of the order closely relates to the degree of smoothness or differentiability of the true hazard function. Basically, the rate \(n^{2/3}\) is corresponding to the true hazard function that is twice (or more) differentiable, whereas the rate becomes \(n^{1/3}\) for the true hazard function that is only differentiable to the first degree. An order that is too small will likely result in biased estimates, while a large order will introduce too much variation. So our choice of \(m = \lceil n^{0.5} \rceil\) is a crude compromise to account for such bias–variance tradeoff.

4.1.1. Binary covariate example: crossing survival curves

The first simulation study is based on a hypothetical example of a randomized clinical trial with two treatment arms motivated by a real data set (see Section 4.2.1). The event (or failure) times in both control and treatment groups were generated by log-normal distributions. In the control group the mean and standard deviation on the log-scale are \(-0.1\) and \(0.5\), respectively, while these parameter are respectively set to \(0\) and \(0.25\) for the treatment group. The event times in both groups were subject to independent random censoring with hazard rate \(\lambda_c\) adjusted to get appropriate censoring rates. The parameter values of log-normal models were deliberately chosen to allow the survival curves for two groups to cross each other during the trial. The sample size is set at \(n = 100\) with equal sizes \((n_0 = n_1 = 50)\) for each treatment arm. For each censoring rate, we generated \(N = 1000\) Monte Carlo replicates of the data. Note that the Cox PH model is inappropriate in this setting for obvious reasons and hence not compared against our method. Instead, survival functions were estimated using our proposed Bernstein polynomial approach, the Kaplan–Meier method and the HARE method.

The Monte Carlo pointwise means of the estimated survival curves using the three methods are shown in Fig. 1 of the electronic supplementary material. It is evident that the proposed model using Bernstein polynomials provides a very good fit to data generated from a model with crossing survival curves. As the Monte Carlo means approximately coincide with the true survival functions, the estimated curves obtained by the Bernstein polynomial model appear to be empirically unbiased under all three censoring scenarios. The performances of the three methods in terms of IAE are also shown in Fig. 1. Generally, in the control group the estimated survival functions by the Bernstein polynomial model have very similar IAE with the ones from both the Kaplan–Meier estimator and the HARE method, while the Bernstein polynomial method performs slightly better than the other two methods for the treatment group. This advantage of the Bernstein polynomial method is more clearly summarized in Table 1. Overall, the Bernstein polynomial method has lower median IAE than the other two methods. For each of 1000 Monte Carlo replicates, the method that has the smallest IAE is recorded. The lower panel of Table 1 shows the percentage out of all 1000 replicates that a particular method has the smallest IAE among the three methods being compared. As shown in Table 1, the Bernstein polynomial method has the smallest IAE in majority of the cases across all three censoring scenarios.

4.1.2. Continuous covariate example

For models with a continuous covariate, say \(Z\), we studied the performance of the proposed Bernstein polynomial method when data were generated from incorrect semiparametric or parametric models. Specifically, the event time \(T\) for a given \(Z\) was generated by the conditional log-normal nonlinear heteroscedastic model

\[
\log T = \mu(Z) + \varepsilon,
\]

where the mean function \(\mu(Z) = \cos(\pi Z)\) and \(\varepsilon|Z \sim N(0, \sigma(Z))\) with \(\sigma(Z) = |Z|\). For simplicity the covariate was generated as \(Z \sim \text{Unif}(0, 1)\) and the censoring time \(C\) was generated from an exponential distribution. Together with the Bernstein polynomial model and the HARE method, we fitted the data with the Cox PH model and the parametric AFT model both with the mean function \(\mu(Z)\) misspecified as a linear function of \(Z\). Also, the baseline distribution of the parametric
Fig. 1. Boxplots of the integrated absolute error (IAE) for the binary covariate example based on 1000 MC replicates. The solid dot in the box represents the median IAE value. BP: the Bernstein polynomial estimator; KM: the Kaplan–Meier estimator; HARE: the Hazard Regression estimator.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Treatment</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td><strong>0%</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>30%</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>50%</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Median IAE (× 100)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP</td>
<td>1.49</td>
<td>1.89</td>
<td>2.43</td>
</tr>
<tr>
<td>KM</td>
<td>1.79</td>
<td>2.26</td>
<td>3.05</td>
</tr>
<tr>
<td>HARE</td>
<td>1.96</td>
<td>2.27</td>
<td>2.58</td>
</tr>
<tr>
<td><strong>Smallest IAE achieved</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP</td>
<td>64%</td>
<td>61%</td>
<td>54%</td>
</tr>
<tr>
<td>KM</td>
<td>3%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>HARE</td>
<td>33%</td>
<td>37%</td>
<td>46%</td>
</tr>
</tbody>
</table>

AFT model is misspecified to be exponential distribution. Again, we set the sample size $n = 100$ and the order of Bernstein polynomial $m = \lceil n^{0.5} \rceil = 10$.

In Fig. 2 of the electronic supplementary material it can be clearly observed that the Bernstein polynomial model provides the most accurate estimates for the conditional survival curves in this case, whereas the survival curve estimates obtained from the misspecified Cox PH model and the misspecified parametric AFT model are generally biased throughout all censoring rates. The HARE method also provides empirically unbiased estimates, but it appears that the HARE method has relatively larger variability than the Bernstein polynomial method. The results displayed in Fig. 2 of the electronic supplementary material are also summarized in Fig. 2 and Table 2 in terms of IAE evaluated at $Z = 0.5$. The Bernstein polynomial model outperforms the Cox PH model and the parametric AFT model by a large margin, and it also provides relatively smaller IAE than the HARE method for this particular scenario. As shown in the lower panel of Table 2, the Bernstein polynomial model has the smallest IAE among the four methods more than 70% of the time out of 1000 MC replicates. Thus,
Fig. 2. Boxplots of the integrated absolute error (IAE) of the conditional survival function evaluated at $Z = \mu_Z = 0.5$. BP denotes the Bernstein polynomial estimator, PH denotes the Cox proportional hazard estimator, AFT denotes the parametric AFT estimator with exponential baseline, and HARE denotes the Hazard Regression estimator.

Table 2
Performances of the four methods in terms of integrated absolute error (IAE) for the simulated scenario with continuous covariate. BP: Bernstein polynomial; PH: Cox proportional hazard model; AFT: parametric AFT model; HARE: hazard regression.

<table>
<thead>
<tr>
<th>Method</th>
<th>Median IAE (× 100)</th>
<th>Smallest IAE achieved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
<td>30%</td>
</tr>
<tr>
<td>BP</td>
<td>2.29</td>
<td>2.63</td>
</tr>
<tr>
<td>PH</td>
<td>5.22</td>
<td>5.70</td>
</tr>
<tr>
<td>AFT</td>
<td>7.83</td>
<td>10.13</td>
</tr>
<tr>
<td>HARE</td>
<td>3.23</td>
<td>3.80</td>
</tr>
</tbody>
</table>

overall we find the performance of the Bernstein polynomial based model relatively robust especially when the dependence of the event time on other covariates is through a non-linear function.

4.2. Real data applications

We studied two well known datasets with potentially nonproportional hazards: the gastric cancer data and the Veterans Administration lung cancer data. The first example demonstrates the application of the proposed Bernstein polynomial model with a binary covariate, while the second one is an example of application with multiple categorical covariates and a continuous covariate.

4.2.1. Gastric cancer data

In a gastric cancer study reported by Stablein et al. (1981), $n = 90$ patients with locally advanced gastric carcinoma were randomized to two treatment groups (45 patients per group). One group only received chemotherapy while the other group received radiotherapy together with the same chemotherapy. The dataset has been studied by various authors (see e.g., Cai and Sun, 2003, and references within) and became a classic example for studying nonproportional hazards because survival curves for two treatment groups cross during the trial. As shown by the Kaplan–Meier curves in Fig. 3, before the crossing point at approximately 1000 days the patients in the group receiving only chemotherapy had better survival rates while the benefit of combination treatment of chemotherapy and radiotherapy started to emerge at a later stage of the study. We estimated the survival functions using the model described in Section 2.1 with the order $m = \lceil n^{0.5} \rceil = 10$. The results in Fig. 3 indicate that the estimated survival curves obtained by the proposed Bernstein polynomial model are close to the ones obtained by the Kaplan–Meier estimator. However, it is evident that the Bernstein polynomial based estimates provide much smoother survival curves which allows for a better estimation of the crossing time of the survival curves. The smooth estimated curves cross at $t = 952$ days.

4.2.2. Veterans Administration data

In the Veterans Administration lung cancer study (Prentice, 1973), $n = 137$ male patients were assigned to two treatment groups: standard chemotherapy and test chemotherapy. Together with treatment assignment, 5 other baseline covariates
were recorded. Following previous works (see e.g., Peng and Huang, 2007) we only used Karnofsky performance score and cell type as important covariates other than the treatment assignment. The reason for this choice can also be justified by following the initial analysis (using the Cox PH model) reported in Therneau and Grambsch (2000, Chap. 6). Also, an earlier analysis based on the Cox PH model indicated that Karnofsky score and cell type are the two predictors that have significantly large test statistics for nonproportionality. Karnofsky score is a continuous variable that takes value from 0 to 100 while cell type is a categorical variable with four levels: squamous cell, small cell, adenocarcinoma and large cell.

Since there are both categorical and continuous covariates, we first considered the stratified approach. However, given the fact that the sample size \( n = 137 \) of this study is not very large, there is a concern that there may not be sufficient data in a particular subgroup to fit the Bernstein polynomial model with continuous covariate. For example, there are only 9 patients with small cell type who received the standard chemotherapy, so Karnofsky scores of these patient might not be representative enough to span the interval \( (0, 100) \). Therefore, we analyzed the data using the partially linear coefficient model described at the end of Section 2.2 with the order \( m = \left\lfloor \frac{10}{0.5} \right\rfloor = 12 \). This model is a simplification of the stratified approach under the assumption that there is no interaction between Karnofsky score and cell type. At the risk of making this assumption, we were able to model the effect of the continuous covariate Karnofsky score using the entire data. The results of the estimated conditional survival contours are displayed in Fig. 4. Generally, patients with higher Karnofsky performance scores have higher survival rates at any given time point for all groups. But such association differs across different treatment groups and more importantly across different cell types. Overall, the patients with small cell type in the test treatment group underwent the sharpest decline in survival rates, while the patients with squamous cell type receiving the test chemotherapy had the best survival profiles among all groups. The most significant treatment difference is also observed for the patients with small cell type. In this subgroup, the patients receiving the standard chemotherapy appear to have better survival rates than the ones receiving the test chemotherapy. On the contrary, patients receiving the test treatment had better survival rates than the ones receiving the standard treatment for the subgroup with squamous cell type. For the patients with adenocarcinoma or large cell type, survival contours are similar across the treatment groups.

5. Conclusions and discussions

In this paper we present a nonparametric regression model for right-censored data based on Bernstein polynomials. We have demonstrated several advantages of our approach compared to other available methods in the similar context. The
The most remarkable feature of the proposed method is that the log-likelihood, its gradient, and the Hessian matrix all take a relatively simple form which makes it easy to compute the estimator in practice. This is due to the unique differentiability property of Bernstein polynomials. Additionally, we show that the general simple form of the log-likelihood function holds even in the presence of categorical and continuous covariates, so the proposed method can be implemented in a unified way for all these cases. Under some mild conditions, the proposed sieve maximum likelihood estimator is shown to be consistent and the corresponding rate of convergence is obtained. Through several simulation experiments with both categorical and continuous covariates, we empirically demonstrated that the proposed Bernstein polynomial model has reasonably robust performance compared to other semiparametric models particularly when the semiparametric assumptions (e.g., PH, AFT etc.) are violated. Also, the proposed method provides similar or slightly better estimates than the HARE model in the simple situations with only binary covariates, but this advantage of the proposed method over the HARE model starts to be more prominent when the dependence of event time on covariates is relatively complex. In general, the computation algorithm involved in the proposed method was stable in the simulation studies, it reached convergence for all generated data across
all scenarios studied. On the contrary, the HARE model failed to reach convergence in several occasions, especially when the censoring rate was chosen to be large. For example, in 39 out of 1000 MC replicates the HARE model did not converge for the binary covariate example when the censoring rate was 50%.

In the numerical examples throughout the paper, we used the asymptotic rate \( m = [n^{0.5}] \) as the default choice for the order of Bernstein polynomial. Alternatively, following \( \text{Kooperberg et al. (1995)} \) we also tried to use the Bayesian Information Criterion (BIC) to choose an optimal order. Unfortunately, the BIC approach always chose the smallest order in the given range for our method. The order \( m \) essentially plays the role of bandwidth as in kernel smoothing or number of knots as in spline smoothing. So the choice of \( m \) could also be determined by more popular data-driven approaches such as cross-validation. However, at least to the best of our efforts, we were unable to find a widely applicable cross-validation method for censored data in the regression setting. This is certainly an issue that deserves more future studies.

Lastly, although the model can contain multiple categorical covariates, only one continuous covariate is used throughout this paper. The reason for this is that in many clinical trial settings, the number of important continuous variables is usually low, and in most cases only one continuous covariate is used. The case of multiple continuous covariates could be handled by incorporating multivariate Bernstein polynomial. For the partially linear coefficient model, additional continuous covariates can be included easily through an additive structure. However, both approaches would lead to relatively high dimensions when the number of continuous covariate is large. This is certainly a drawback of the proposed method compared to other methods such as HARE. One way around this problem would be to use a single index model to reduce the dimension of continuous covariate to a scalar and then fit the model described in this paper.

Appendix A

The proof of consistency is built within the framework of Theorem 1 of \( \text{Geman and Hwang (1982)} \). Following \( \text{Geman and Hwang (1982)} \), the following notations and definitions will be used: (1) for \( h \in \Theta_m \) and \( \forall \varepsilon > 0 \), an open ball is defined as \( B_m(h, \varepsilon) = \{g : g \in \Theta_m \text{ and } d(h, g) < \varepsilon\} \); (2) \( E_0 g(X) = \int g(x)p_0(x)d\mu(x) \) and \( Q(h, g) = E_h \log p_h(X) \); (3) for a function \( g : \Theta \rightarrow \mathbb{R}, g(B) = \sup_{b \in B} g(b) \) for any \( B \subseteq \Theta \); (4) the maximum entropy set is defined as \( A_m = \{h : h \in \Theta_m \text{ and } Q(h_0, h) = Q(h_0, \Theta_m)\} \); (5) for \( C_m \subset A, C_m \rightarrow h \) means \( \sup_{g \in C_m} d(h, g) \rightarrow 0 \).

**Lemma 1 (Geman and Hwang, 1982).** Assume that the sieve \( \{\Theta_m\} \) satisfies following conditions:

B1. \( f(x, B_m(h, \varepsilon)) \) is measurable in \( x \) for \( \forall m, \forall h \in \Theta_m, \forall \varepsilon > 0 \) and \( f(x, h) \) is upper-semicontinuous in \( h \) on \( \Theta_m \);

B2. \( \exists \varepsilon > 0 \text{ such that } E_0\log f(X, B_m(h, \varepsilon)) < \infty \text{ for } \forall m, \forall h \in \Theta_m \);

B3. \( \Theta_m \) is compact for \( \forall m \);

B4. \( A_m \rightarrow h_0 \) as \( m \rightarrow \infty \).

Then the sieve maximum likelihood estimator \( \hat{h}_m(\cdot) \rightarrow a.s. h_0(\cdot) \) as \( m, n \rightarrow \infty \).

**Proof of Theorem 1. (Consistency):** Since the functional \( f(x, \delta, \theta) = \theta(x)^{\delta}e^{-\int_0^\theta \theta^{(u)}du} \) is continuous with respect to \( \theta(\cdot) \), it immediately follows that B1 is satisfied. Also, when \( \theta \in \Theta_m \), the function \( \theta(x, y_m) \) can be regarded as a continuous mapping from \([0, L] \times \Theta_m \) to \( \Theta_m \). As \([0, L] \times \Theta_m \) is a compact space, it follows that \( \Theta_m \) is also compact, so B3 holds.

For any \( h \in \Theta_m \), it is easy to verify the discrepancy \( D(h_0, h) = E_0 \left\{ \log \frac{f(X, \Delta, h_0)}{f(X, \Delta, h)} \right\} \geq 0 \) holds for right-censored data. So we have \( E_0 \log f(X, \Delta, h_0) \leq E_0 \Delta \log h_0(X) \). As \( h_0(X) \) is bounded above and also bounded away from 0 by the condition (II), we have \( E_0 \log f(X, \Delta, h) < \infty \) for \( \forall h \in \Theta_m \). This shows that B2 is satisfied.

Let \( \tilde{h}_m(t) = \sum_{k=1}^m h_0 \left( \frac{t}{m} \right)^k (t/\tau)^{-k} (1 - t/\tau)^{-m-k} \), clearly \( \tilde{h}_m \in \Theta_m \). And from the Weierstrass theorem, \( \|h_0(\cdot) - \tilde{h}_m(\cdot)\|_\infty = \sup_{0 \leq t \leq \tau} |h_0(t) - \tilde{h}_m(t)| \rightarrow 0 \) as \( m \rightarrow \infty \) (Lorentz, 1953). The discrepancy between \( f(x, \delta, h_0) \) and \( f(x, \delta, \tilde{h}_m) \)

\[
D(h_0, \tilde{h}_m) = E_0 \log \frac{f(X, \Delta, h_0)}{f(X, \Delta, \tilde{h}_m)} = E_0 \Delta \log \frac{h_0(X)}{\tilde{h}_m(X)} + E_0 \int_0^X \left[ h_0(u) - \tilde{h}_m(u) \right] du \leq E_0 \Delta \frac{h_0(X) - \tilde{h}_m(X)}{h_0(X)} + \tau \|h_0(\cdot) - \tilde{h}_m(\cdot)\|_\infty \leq E_0 \left[ \frac{|h_0(X) - \tilde{h}_m(X)|}{h_0(X)} + \tau \|h_0(\cdot) - \tilde{h}_m(\cdot)\|_\infty \right] = \left( E_0 \frac{\tau}{h_0(X)} + \tau \right) \|h_0(\cdot) - \tilde{h}_m(\cdot)\|_\infty.
\]
Again, due to the fact that $h_0$ is bounded away from 0 (see the condition (II)) and $\sup_{0 \leq u \leq 1} |h_0(u) - \tilde{h}_m(u)| \to 0$ as $m, n \to 0$, there exists $\tilde{h}_m \in \Theta_m$ such that $D(h_0, \tilde{h}_m) \to 0$, this implies that B4 is satisfied. Finally, by Remark 1 in Section 2, it follows that the set consisting of sieve maximum likelihood estimators is non-empty. By Lemma 1, $\tilde{h}_m(\cdot) \to a.s. h_0(\cdot)$ as $m, n \to \infty$. This completes the proof of consistency.

The proof of the convergence rate is built on a result given by Shen and Wong (1994).

**Lemma 2 (Shen and Wong, 1994).** For the sieve maximum likelihood estimator

$$\hat{h}_m = \arg \max_{h \in \Theta_m} \sum_{i=1}^{n} l(X_i, \Delta_i, h),$$

if following two conditions hold:

C1. $\inf_{d(h_0, h_m) \leq \varepsilon} E_0[l(X, \Delta, h_0) - l(X, \Delta, h_m)] \geq c_2 \varepsilon^{2\alpha}$;

C2. $\sup_{d(h_0, h_m) \leq \varepsilon} \text{VAR}_0[l(X, \Delta, h_0) - l(X, \Delta, h_m)] \leq c_3 \varepsilon^{2\beta}$, and additionally,

C3. for function class $\mathcal{F}_m = \{l(h, \cdot) - l(h_m, \cdot) : h \in \Theta_m\}$ the uniform metric entropy satisfies $H(\varepsilon, \mathcal{F}_m) \leq c_4 n^{2\alpha} \varepsilon^{-\tau}$ for the case of $0^+ < \tau < 2$, then

$$d(\hat{h}_m, h_0) = O_p(\max(n^{-\alpha}, d(\pi_n h_0, h_0), D^{1/2\alpha}(\pi_n h_0, h_0))),$$

where $D(\pi_n h_0, h_0) = E_0(l((X, \Delta, h_0) - l(X, \Delta, \pi_n h_0)))$, $\pi_n h_0$ is an approximation of $h_0$ in the sieve $\Theta_m$, and $\tau_0 = \frac{1 - 2\alpha}{4\alpha - \min(\alpha, \beta)(2 - \tau)}$ if $0 < \tau < 2$ (for other values of $\tau$, the values of $\tau_0$ can be found in Shen and Wong, 1994).

**Proof of Theorem 2. (Rate of Convergence):** For the Bernstein polynomial

$$\tilde{h}_m(t) = \sum_{k=1}^{m} h_0(k \tau / m) \left( \frac{m-1}{k-1} \right) (t/\tau)^{k-1}(1 - t/\tau)^{m-k},$$

it follows from the Theorem 1.6.2 of Lorentz (1953) and the condition (II) that the approximation error $|\tilde{h}_m(t) - h_0(t)| \leq \frac{3}{4}(m-1)^{-\frac{1+\alpha_0}{2}} \approx \frac{3}{4}m^{-\frac{1+\alpha_0}{2}}$ when $m$ is large enough. Assume $m = o(n^2)$, we have

$$\sup_{0 \leq t \leq \tau} |\tilde{h}_m(t) - h_0(t)| \leq \frac{3}{4} n^{(1+\alpha_0)/2}.$$

First, since the discrepancy

$$D(h_0, h_m) = E_0 \left\{ \log \frac{f(X, \Delta, h_0)}{f(X, \Delta, h_m)} \right\} = 2E_0 \left\{ - \log \frac{f(X, \Delta, h_m)}{f(X, \Delta, h_0)} \right\}$$

$$\geq 2E_0 \left\{ 1 - \sqrt{\frac{f(X, \Delta, h_m)}{f(X, \Delta, h_0)}} \right\} = 2d(h_0, h_m)^2,$$

the condition C1 is satisfied with $\alpha = 1$.

The condition C2 can be verified with $\beta = 1$ following the similar steps as in the Example 2 of Shen and Wong (1994),

$$\text{VAR}_0[l(X, \Delta, h_0) - l(X, \Delta, h_m)] \leq E_0[l(X, \Delta, h_m) - l(X, \Delta, h_0)]^2$$

$$\leq 4E_0 \left\{ \log \left( 1 + \left( \sqrt{\frac{f(X, \Delta, h_m)}{f(X, \Delta, h_0)}} - 1 \right) \right) \right\}^2$$

$$\leq 4B E_0 \left\{ \sqrt{\frac{f(X, \Delta, h_m)}{f(X, \Delta, h_0)}} - 1 \right\}^2 = 8B d(h_0, h_m)^2,$$

for some $B \geq 1$ if $\frac{f(X, \Delta, h_m)}{f(X, \Delta, h_0)} \geq \eta > -1$, which is satisfied because $f(X, \Delta, h_0)$ is bounded above and away from zero.

Let $\mathcal{F}_m = \{l(h, \cdot) - l(h_m, \cdot) : h \in \Theta_m\}$, where $\tilde{h}_m(t) = \sum_{k=1}^{m} h_0 \left( k \tau / m \right) \left( \frac{m-1}{k-1} \right) (t/\tau)^{k-1}(1 - t/\tau)^{m-k}$. Further, let $N(\varepsilon, \mathcal{F}_m)$ denote the minimum number of $\varepsilon$-balls in uniform metric required to cover $\mathcal{F}_m$, then $H(\varepsilon, \mathcal{F}_m) = \log N(\varepsilon, \mathcal{F}_m)$ is the $L_\infty$-metric entropy of the space $\mathcal{F}_m$. From Kolmogorov and Tikhomirov (1959) and the condition (III), we have

$$H(\varepsilon, \mathcal{F}_m) \leq H(\varepsilon, \Theta_m) \leq c_4 \varepsilon^{-1/(1+\alpha_0)}.$$

So the condition C3 is satisfied with $\tau_0 = 0$ and $r = \frac{1}{1+\alpha_0}$. Consequently,

$$\tau_0 = \frac{1 - 2\tau_0}{4\alpha - \min(\alpha, \beta)(2 - r)} = \frac{1}{4 - \left( 2 - \frac{1}{1+\alpha_0} \right)} = \frac{1 + \alpha_0}{3 + 2\alpha_0}.$$
This leads to the rate of convergence
\[ d(\hat{h}_m, h_0) = O_p\left(\max(n^{-2\alpha_0}, d(\hat{h}_m, h_0), K^{1/2\alpha}(f(\cdot, \tilde{h}_m), f(\cdot, h_0)))\right) \]
\[ = O_p\left(n^{-\frac{1+\alpha_0}{3+2\alpha_0}}\right). \]

If we let \( \kappa = \frac{2}{3+2\alpha_0} \), we have the rate of convergence \( d(\hat{h}_m, h_0) = O_p\left(n^{\frac{1+\alpha_0}{3+2\alpha_0}}\right) \). Note that \( \alpha_0 \) is the exponent of the Holder continuity of \( h_0^{(1)} \). If \( \alpha_0 = 1 \), \( h_0 \) has finite second derivative, then the rate of convergence \( d(\hat{h}_m, h_0) = O_p(n^{-2/5}) \) with \( \kappa = 2/5 \). On the other hand, \( h_0 \) only has finite first derivative when \( \alpha_0 = 0 \), this in turn gives \( d(\hat{h}_m, h_0) = O_p(n^{-1/3}) \) with \( \kappa = 2/3 \). In our empirical applications we have used \( \kappa = 0.5 \), which is corresponding to the case with \( \alpha_0 = 0.5 \).

**Appendix B. Supplementary data**

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.08.019.

**References**


