Spline-based sieve maximum likelihood estimation in the partly linear model under monotonicity constraints

Minggen Lu
School of Community Health Sciences, University of Nevada, Reno, NV 89557, USA

A R T I C L E   I N F O

Article history:
Received 15 June 2009
Available online 29 July 2010

AMS subject classifications:
62G08

Keywords:
Empirical process
Generalized Rosen algorithm
Maximale likelihood method
Monotone B-splines
Monte Carlo

A B S T R A C T

We study a spline-based likelihood method for the partly linear model with monotonicity constraints. We use monotone B-splines to approximate the monotone nonparametric function and apply the generalized Rosen algorithm to compute the estimators jointly. We show that the spline estimator of the nonparametric component achieves the possible optimal rate of convergence under the smooth assumption and that the estimator of the regression parameter is asymptotically normal and efficient. Moreover, a spline-based semiparametric likelihood ratio test is established to make inference of the regression parameter. Also an observed profile information method to consistently estimate the standard error of the spline estimator of the regression parameter is proposed. A simulation study is conducted to evaluate the finite sample performance of the proposed method. The method is illustrated by an air pollution study.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

In this paper, we consider spline-based maximum likelihood estimation of the partly linear model under monotonicity constraints. A general partly linear model takes the form

\[ Y = \psi(Z) + X^T \beta + \varepsilon, \]

where \( X^T = (x_1, \ldots, x_d)^T \) and \( Z \) are explanatory variables, \( \beta \) is a \( d \times 1 \) vector of the unknown regression parameters, \( \psi \) is an unknown function, the error term \( \varepsilon \) is normally distributed with mean 0 and finite variance \( \sigma^2 \), and \( (X, Z) \) and \( \varepsilon \) are independent. The partly linear model is an extension of a standard linear model without having to specify the functional forms of some predictor variables. It can be an appropriate choice when the response variable \( Y \) is assumed to be linearly associated with covariate \( X \), but the relationship between \( Y \) and \( Z \) may be nonlinear.

The partly linear model (1) has been extensively studied by many authors, see for example, Bianco and Boente [1], Engle et al. [5], Green et al. [6], Green and Silverman [7], Robinson [25], and Schimek [27] among many others. Many authors have also investigated the asymptotic behaviors of the estimates of the regression parameter and the smooth nonparametric component using smoothing spline, kernel smoothing, or local linear smoother methods. Heckman [9] explored the asymptotic properties of the estimate of \( \beta \) using the penalized likelihood estimation method. Chen [2] used piecewise polynomials to approximate \( \psi \) and showed that the estimate of \( \beta \) can achieve a rate of convergence \( n^{-1/2} \) with smallest possible asymptotic variance. Chen and Shiau [3] studied the asymptotic behaviors of two data-driven efficient estimators of \( \beta \) using the spline estimation method. Mammen and van der Geer [19] applied the empirical process theory to study the asymptotic properties of the penalized quasi-likelihood estimator of \( \beta \). Speckman [32] investigated the theoretical properties of the kernel smoothing approach for the partly linear model. Hamilton and Truong [8] used the local linear smoother method to derive the asymptotic distributions of the estimates of \( \beta \) and \( \psi \), which generalized the results of [32].

E-mail address: minggenl@unr.edu.

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
In many studies, there is a monotonic relationship between one or more of covariates and the response variable, for example, the dose–response relationship. Huang [10] considered the isotonic regression approach for estimation of the partly linear model when \( \psi \) is assumed to be a smooth monotone function. Under the assumption that the error \( \varepsilon \) is normally distributed, the estimator of \( \beta \) was shown to be asymptotically efficient among all regular estimators. The limiting distribution of the isotonic estimator of the monotone nonparametric function \( \psi \) at a fixed point was also established. To the best of our knowledge, however, there is no systematic study for the spline–based estimator of \( (\beta, \psi) \) when \( \psi \) is subject to be monotone. Therefore, it would be preferable to develop a practical spline procedure for the partly linear model under monotonicity constraints on \( \psi \) and study the asymptotic properties of the estimates.

The spline estimation of an unknown monotone function has been studied by many researchers. For example, Ramsay [24] defined monotone \( l \)-splines and discussed the computational and inferential issues of the method. The proposed monotone spline approach was used in many applications, such as response variable transformation in nonlinear regression and use of monotone splines to model the dose–response relationship. Kelly and Rice [14] proposed a nonparametric smoothing method to study dose–response curves under a monotonicity constraint. Shen [30] introduced a spline–based sieve maximum likelihood estimation method for the baseline function and the regression parameter in the proportional odds regression model with right-censored data and Case 2 interval-censored data. Leitenstorfer and Tutz [16] used a monotone \( B \)-spline smoothing procedure within a generalized additive model framework to investigate the influence of the air pollutant on respiratory mortality. Lu et al. [17] proposed a nonparametric monotone \( l \)-spline method for panel count data with proportional mean model. The spline–based estimators exceed the nonparametric estimators proposed by Wellner and Zhang [35] in the rate of convergence and the finite sample performance.

In this manuscript, monotone \( B \)-splines are applied to approximate the nondecreasing function \( \psi(Z) \), i.e.

\[
\psi(Z) \approx \sum_{j=1}^{k_n} a_j B_j(Z),
\]

subject to the constraints \( a_1 \leq \cdots \leq a_{k_n}. \) The nondecreasing constraints on the coefficients \( a_j, j = 1, \ldots, k_n \), guarantee the resulting spline to be monotone by Schumaker [29]. This approach follows the idea of the sieve method for the estimation of the infinite-dimensional parameter \( \psi \). In sieve estimation a sequence of subspaces (sieves) that depend on the sample size \( n \) are used to approximate the original space such that the resulting estimation problem over sieves becomes less complicated. In the model presented here, the sieves are the collections of monotone splines and the original space is the set of bounded nondecreasing smooth functions. By using monotone \( B \)-splines to approximate \( \psi \), we can estimate the spline coefficients \( \alpha = (\alpha_1, \ldots, \alpha_{k_n}) \) and the regression parameter \( \beta \) simultaneously. The generalized Rosen algorithm proposed by Jamshidian [13] is applied for computing the estimates of \( \alpha = (\alpha_1, \ldots, \alpha_{k_n}) \) and \( \beta \). We show that the estimator of \( \beta \) is asymptotically normal and efficient and the estimator of \( \psi \) achieves the possible optimal rate of convergence under the smooth condition. We also develop a spline–based likelihood ratio test and a spline–based Wald test for the inference of \( \beta \).

The rest of the paper is organized as follows: The spline maximum likelihood estimator \( (\hat{\beta}_n, \hat{\psi}_n) \) and the numerical algorithm are presented in Section 2. Asymptotic results are given in Section 3. A Monte Carlo simulation study and an illustrating example from an air pollution study are displayed in Section 4. The summary of our findings and some related topics are discussed in Section 5. Finally, the proofs of asymptotic results are sketched in the Appendix.

2. Method and algorithm

Let \( (\hat{\beta}_0, \hat{\psi}_0) \) be the true value of the parameter \( (\beta, \psi) \). Assume \( \beta_0 \) belongs to a convex and compact subset \( \Theta \subset \mathbb{R}^d \) and \( \psi_0 \) is a smooth monotone function. Let \( (Y_1, Z_1, X_1), \ldots, (Y_n, Z_n, X_n) \) be a random sample. The log–likelihood for this random sample is

\[
I_n(\beta, \psi) = -\sum_{i=1}^{n} (Y_i - X_i^T \beta - \psi(Z_i))^2 / (2\sigma^2),
\]

subject to \( \beta \in \Theta \) and \( \psi \) being monotone. Let \( Z_{(1)} \leq \cdots \leq Z_{(n)} \) be the ordered values of \( Z \)'s and \( \psi_1 = \psi(Z_{(1)}) \). Huang [10] defined the semiparametric likelihood estimator of \( (\hat{\beta}_0, \hat{\psi}_0) \) as the maximizer of \( I_n(\beta, \psi) \) subject to \( \beta \in \Theta \) and \( \psi_1 \leq \cdots \leq \psi_n \). Indeed, the semiparametric maximum likelihood estimation of \( \beta \) and \( \psi \) is a semiparametric isotonic regression problem. The semiparametric estimation method can be implemented based on the profile likelihood method in which the estimator of \( \psi \) is defined as a nondecreasing step function with jumps only occurring at observation points. In this manuscript, we propose to estimate \( \psi \) using monotone \( B \)-splines instead of the step function in order to achieve faster rate of convergence and better finite sample performance of the estimate of \( \psi \).

Let \( \tau_n = (t_1, t_2, \ldots, t_{m_n}) \), with

\[
a = t_1 = \cdots = t_i < t_{i+1} < \cdots < t_{m_n+i} = \cdots = t_{m_n+2i} = b,
\]

be a sequence of knots that partition a closed interval \([a, b]\) into \( m_n + 1 \) subintervals \( l = [t_{i+1}, t_{i+1+1}) \), for \( i = 0, \ldots, m_n - 1 \), and \( l_{m_n} = [t_{m_n+1}, t_{m_n+1+1}] \). Let \( \delta_s(\tau_n, l) \) denote the class of splines of order \( l \geq 1 \) with knots \( \tau_n \). For any \( s \in \delta_s(\tau_n, l) \), according to Corollary 4.10 of Schumaker [29], there exist a class of \( B \)-spline basis functions \( \{B_i, 1 \leq i \leq k_n\} \), where
\[ k_n = m_n + l, \text{ such that } s = \sum_{i=1}^{k_0} \alpha_i Bi. \] According to Theorem 5.9 of Schumaker [29], the spline \( s \) is monotone nondecreasing on \( [a, b] \) if nondecreasing constraints are imposed on the coefficients \( \alpha = (\alpha_1, \ldots, \alpha_{k_0}) \). Thus,

\[
\mathcal{M}_n(T_n, l) = \left\{ \sum_{i=1}^{k_0} \alpha_i B_i : \alpha_1 \leq \cdots \leq \alpha_{k_0}, \max |\alpha_i| \leq L_n \text{ for some constant } L_n \right\},
\]

the subclass of \( \mathcal{S}_n(T_n, l) \), is the collection of monotone nondecreasing splines on \([a, b]\).

By replacing \( \psi(Z) \) by \( \sum_{i=1}^{k_0} \alpha_i B_i(Z) \) in the log-likelihood function \((2)\), we obtain the spline log-likelihood function

\[
l_n(\alpha, \beta) = -\frac{1}{2} \sum_{i=1}^{n} \left( Y_i - X_i^T \beta - \sum_{j=1}^{k_0} \alpha_j B_j(Z) \right)^2 / (2\sigma^2),
\]

subject to \( \alpha_1 \leq \cdots \leq \alpha_{k_0} \). The advantage of this reparametrization is that we can estimate the regression parameters \( \beta \) and coefficients \( \alpha = (\alpha_1, \ldots, \alpha_{k_0}) \) simultaneously through maximizing the spline likelihood function subject to nondecreasing constraints. The computational burden can be greatly alleviated by such fully parametric representation of spline likelihood function. See [18,28] for most recent application of splines.

Let \( \hat{\alpha}_n = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{k_0}) \) and \( \hat{\beta}_n \), be the values that maximize the spline likelihood function \((3)\). We denote the spline estimator of \( \psi \) by \( \sum_{i=1}^{k_0} \hat{\alpha}_i B_i \).

The spline likelihood estimation problem \((3)\) can be formulated as an optimization problem subject to linear inequality constraints

\[
\max_{\theta \in \Theta_\alpha \times \mathbb{R}^d} l(\theta | Y, Z, X),
\]

where \( \theta = (\alpha, \beta) \) with \( \alpha \in \Theta_\alpha = \{ \alpha : \alpha_1 \leq \cdots \leq \alpha_{k_0} \} \). In the optimization literature, Rosen algorithm [26], also known as gradient projection (GP) algorithm, is often applied for optimizing the objective function with linear constraints. Jamshidian [13] generalized the Rosen algorithm by using a general metric with norm \( \|x\| = x^T W x \), where \( W \) is a positive definite matrix. Zhang and Jamshidian [36] applied this algorithm for computing the nonparametric maximum likelihood estimators of the failure functions with doubly censored data and interval-censored data. In order to avoid a possible storage problem in updating the Hessian matrix \( H \) for the large sample size, they chose \( W = D_{\beta} \), the matrix containing only the diagonal elements of negative \( H \). This resulted in increasing the number of iterations and thereby the computing time. In our application, the dimension of the unknown parameter space is relatively small due to the use of \( B \)-splines. Therefore, we choose the negative Hessian matrix as \( W \). The use of the full Hessian matrix substantially reduces the number of iterations. As a result, the computational burden for the spline estimation is expected to be alleviated.

Now we describe the algorithm used in computing the proposed spline estimator. Let \( l(\theta) \) and \( W \) be the gradient and negative Hessian matrix of the log-likelihood with respect to \( \theta \). We denote by \( A = \{i_1, \ldots, i_m\} \), the set of indices satisfying \( \alpha_{i_j} = \alpha_{i_j+1} \). If \( m > 0 \), define the corresponding working matrix \( \hat{A}_{m \times (k_0 + d)} \), in which the \( j \)-th row is the vector with its \( j \)-th element equal to one, \((i_j + 1)\)th element equal to negative one, and the remaining components being zeros, where \( d \) is the dimension of \( \beta \). The modified generalized Rosen algorithm for partly linear model is implemented in the following steps:

**S0:** Determine the index set \( A \) and its corresponding \( A \) of the initial \( \alpha \in \Theta_\alpha \).

**S1:** Find the feasible search direction

\[
\eta = (I - W^{-1} A^T (AW^{-1} A^T)^{-1} A) W^{-1} l(\theta).
\]

**S2:** If \( \|\eta\| < \varepsilon \), for some small \( \varepsilon > 0 \), compute the Lagrange multipliers \( \lambda = (AW^{-1} A^T)^{-1} AW^{-1} l(\theta) \). Let \( \lambda_i \) be the \( i \)-th component of \( \lambda \).

- If \( \lambda_i \leq 0 \), for all \( i \in A \), then set \( \hat{\theta}_n = \theta \) and stop.
- If there is at least one \( \lambda_i > 0 \), for some \( i \in A \), determine the index corresponding to the largest \( \lambda_i \) and remove this from \( A \), accordingly modify the \( A \) and go to S1.

**S3:** Compute \( \theta_1 = \min_{\eta > 0} \eta \), if \( \eta = (\frac{1}{\eta_1 + \theta_1}) \) and find a smallest integer \( k \) such that \( l(\theta) + (1/2)^k \eta) > l(\theta) \). Then replace \( \theta \) by \( \hat{\theta} = \hat{\theta} + (1/2)^k \), accordingly modify the \( A \) and its corresponding \( A \), and then go to S1.

One choice for the initial values of the modified generalized Rosen algorithm is \( \alpha_1 = \alpha_2 = 0 \) and \( \alpha_i = i - 2 \), for \( i = 3, \ldots, k_0 \). For this case, the working matrix \( A = (1, -1, 0, \ldots, 0)_{1 \times (k_0 + d)} \).

### 3. Asymptotic results

In this section we present asymptotic results for the spline maximal likelihood estimator \( (\hat{\theta}_n, \hat{\psi}_n) \). Denote \( \hat{\theta} = (\beta, \psi) \). Assume the regression parameter space \( \Theta \) to be a convex and compact subset of \( \mathbb{R}^d \) and the parameter space for the nonparametric function \( \psi \) is taken to be

\[ \mathcal{F} = \{ \psi : \psi \text{ is monotone nondecreasing on } [0, \tau] \}. \]
Let $\| \cdot \|$ be the Euclidean distance of $\mathbb{R}^d$. For any probability measure $P$, define $L_2$-norm $\| f \|_2 = \left( \int f^2 dP \right)^{1/2}$. We study the asymptotic properties of $(\hat{\beta}_n, \hat{\psi}_n)$ with $L_2$ metric

$$d_2(\hat{\beta}_1, \hat{\beta}_2) = \| \beta_2 - \beta_1 \|^2 + \| \psi_2 - \psi_1 \|^2 = \int \| \psi_2(z) - \psi_1(z) \|^2 dF_Z(z).$$

for any $\hat{\beta}_i = (\hat{\beta}_i, \hat{\psi}_i), i = 1, 2$, where $F_Z(z)$ is the marginal probability measure of the variable $Z$.

The following regularity conditions with respect to the locations of knots, the smoothness and monotonicity of $\psi_0$, and the underlying distributions of covariates $(X, Z)$ are needed to derive the asymptotic results of the spline maximum likelihood estimator $(\hat{\beta}_n, \hat{\psi}_n)$.

(C1) The maximum spacing of the knots is assumed to be $O(n^{-\nu}), 0 < \nu < 1/2$. Moreover, the ratio of maximum and minimum spacings of knots is uniformly bounded.

(C2) The true function $\psi_0$ is strictly increasing and its $r$th derivative satisfies Lipschitz condition on $[0, \tau]$, with $r \geq 1$, that is, $\psi_0 \in C^r[0, \tau]: \| \psi^{(j)} \|_{\infty} \leq M, j = 0, \ldots, r$. $|\psi^{(r)}(z_1) - \psi^{(r)}(z_2)| \leq L|z_1 - z_2|$.

(C3) The true parameter $\beta_0$ is the interior of $\Theta$.

(C4) The support of $Z$ is an interval within $[0, \tau]$.

(C5) There exists $x_0$ such that $P(\| X \| \leq x_0) = 1$. That is, the covariate $X$ has a bounded support.

(C6) The density function of $Z$ is continuous and positive on $[0, \tau]$.

(C7) For any $\beta \neq \beta_0, P(X^T \beta \neq X^T \beta_0) > 0$.

(C8) $E(X - E(X|Z)) \otimes 2$ is positive definite, where $x \otimes 2 = xx^T$.

(C9) The derivative of $h^*(z) = E(X|Z = z)$ is bounded on $[0, \tau]$.

**Remark 1.** (C1) is a mild assumption on knots and needed to derive consistency and rate of convergence of $(\hat{\beta}_n, \hat{\psi}_n)$. The condition on smoothness and monotonicity of $\psi_0$ is standard in the spline estimation. The compactness and convexity of $\Theta$ and (C3) are common in the literature of semiparametric estimation. Assumptions that are related to observation scheme of $(X, Z)$, (C4)-(C6), are needed for the entropy calculation in the proofs of Theorems 1–3. (C7) is required to establish the identifiability of the semiparametric model. (C8) is useful in the proof of the asymptotic normality. (C9) is needed to define appropriate approximately least favorable submodels in Theorems 2 and 3.

For a single observation $(Y, Z, X)$, its log density is given by

$$l(\beta, \psi) = -(Y - \psi(Z) - X^T \beta)^2/(2\sigma^2).$$

The score function for $\beta$ is

$$\dot{l}_\beta(\beta, \psi) = (Y - \psi(Z) - X^T \beta)X/\sigma^2.$$ 

Consider a parametric smooth submodel $(\beta, \psi_t)$, where $\psi_0 = \psi$ and $\psi_t = \partial \psi_t/\partial t|_{t=0} = h$. Let $\mathcal{H}$ be the class of such $h$ of bounded variation on $[0, \tau]$. The score function for $\psi$ takes the form of

$$\dot{l}_\psi(\beta, \psi) h = (Y - \psi(Z) - X^T \beta)h/\sigma^2.$$ 

The efficient score for $\beta$ at the true parameter $(\beta_0, \psi_0)$ is given by

$$I^\star_\beta = \dot{l}_\beta(\beta_0, \psi_0) - \dot{l}_\beta(\beta, \psi_0) h^*,$$

where $h^* \in \mathcal{H}^d$ satisfies $E[\dot{l}_\beta(\beta_0, \psi_0) - \dot{l}_\beta(\beta, \psi_0) h^*]I^\star_\psi(\beta_0, \psi_0) h = 0$, for all $h \in \mathcal{H}^d$. This simplifies to

$$E(Y - \psi_0(Z) - X^T \beta_0)^2(X - h^*(Z))h^*(Z) = 0,$$

for all $h \in \mathcal{H}^d$. Thus,

$$h^*(z) = E(X|Z = z).$$

So the efficient score function for $\beta$ at $(\beta_0, \psi_0)$ is

$$I^\star_\beta = (Y - \psi_0(Z) - X^T \beta_0)(X - E(X|Z))/\sigma^2.$$ 

The efficient information takes the form of

$$I(\beta_0) = E[I^\star_\beta \otimes 2 = E(X - E(X|Z)) \otimes 2/\sigma^2,$$

where $x \otimes 2 = xx^T$, for $x \in \mathbb{R}^d$.

**Theorem 1.** Let $k_n = O(n^\nu)$, for $1/(2r + 2) < \nu < 1/(2r)$. Suppose conditions (C1)–(C9) hold. Then

(a) Consistency

$$d_2((\hat{\beta}_n, \hat{\psi}_n), (\beta_0, \psi_0)) \rightarrow 0$$

in probability, as $n \rightarrow \infty$. 

(b) Rate of convergence
\[ d_2((\hat{\beta}_n, \hat{\psi}_n), (\beta_0, \psi_0)) = O_p(n^{-\min(r,v,(1-v)/2)}). \]

Thus, if \( v = 1/(1 + 2r) \), \( O_p(n^{-\min(r,v,(1-v)/2)}) = O_p(n^{-r/(1+2r)}) \), which is the optimal rate of convergence under the smooth condition.

(c) Asymptotic normality
\[ \sqrt{n}(\hat{\beta}_n - \beta_0) = n^{-1/2}I^{-1}(\beta_0)\sum_{i=1}^{n} I_i + o_p(1) \to N(0, I^{-1}(\beta_0)) \]

in distribution, as \( n \to \infty \), where \( I(\beta_0) = E(X - E(X|Z))^2/\sigma^2 \).

For any \( \beta \) in the neighborhood of \( \hat{\beta}_n \), let \( \hat{\psi}_\beta \) be the maximizer of the log-likelihood \( l_n(\beta, \psi) \). The profile log-likelihood for \( \beta \) is defined as \( pl_n(\beta) = l_n(\hat{\beta}, \hat{\psi}_\beta) \). For testing \( \beta = \beta_0 \), the likelihood ratio statistic is given by
\[ \text{lrt}_n(\beta_0) = 2pl_n(\hat{\beta}_n) - 2pl_n(\beta_0). \]

**Theorem 2 (Likelihood Ratio Inference).** If conditions (C1)–(C9) hold. Then under \( H_0 : \beta = \beta_0 \in \mathbb{R}^d \),
\[ \text{lrt}_n(\beta_0) \to \chi^2_d \]
in distribution, as \( n \to \infty \).

Although the efficient information matrix \( I(\beta_0) \) has an explicit expression, it is not trivial to directly estimate \( I(\beta_0) \). One approach is using the second derivative of the profile log-likelihood to estimate \( I(\beta_0) \). However, since there is no explicit form for the profile log-likelihood, we cannot directly differentiate the profile log-likelihood. Instead the discretized version of the observed profile information proposed by Nielsen et al. [23] and Murphy et al. [20] is applied to estimate \( I(\beta_0) \). The following theorem shows that the discretized version of the observed profile information is a consistent estimator of \( I(\beta_0) \). A general discussion of the observed information in semiparametric models can be found in [21].

**Theorem 3 (Estimation of the Standard Errors).** For every \( h_n = o_p(1) \) such that \( (\sqrt{n}h_n)^{-1} = O_p(1) \), if conditions (C1)–(C9) hold, then
\[ -h_n^{-2}(pl_n(\hat{\beta}_n + h_ne_i + h_ne_i) - pl_n(\hat{\beta}_n + h_ne_i)) \to I(\beta_0)_{ij} \]
in probability, as \( n \to \infty \), where \( e_i \) is a unit vector with ith element equal to 1 and the remaining being 0.

4. Numerical results

4.1. Simulation study

In this section a Monte Carlo simulation study is performed to evaluate the finite sample performance of the proposed spline estimation method. We generate \( n \) independently and identically distributed observations \( \{Y_i, X_i, Z_i : i = 1, \ldots, n\} \) as follows: \( Z_i \sim \text{Uniform}[0, 10] \); Given \( Z_i, X_i|Z_i \sim N(0, (Z_i/5)^2) \); and \( Y_i = \psi(Z_i) + 0.2X_i + e_i \), where \( \psi(Z) = \exp(Z/4) - 1/2 \) and \( e_i \sim N(0, 0.5^2). \) In this simulation, the cubic B-splines are used to approximate \( \psi \). To investigate the impact of the choice of knots on the proposed spline method, we compare simulation results with different selections of the number and places of knots. The number of internal knots \( m_n \) is chosen as \( n^{1/3} - 1 \) or \( n^{2/5} - 1 \). After the number of knots is fixed, the locations of the knots are determined in two ways. Let \( Z_{\text{min}} \) and \( Z_{\text{max}} \) be the minimum and maximum values of \( Z \), respectively. One method is selecting the end points of \( m_n + 1 \) equally partitioned subintervals of \( [Z_{\text{min}}, Z_{\text{max}}] \). The alternative approach is choosing the \( i/(m_n + 1) \) quantiles, \( i = 0, \ldots, m_n + 1 \), of the observations as the knots. For each scenario, 2000 Monte Carlo samples, with \( n = 50, 100, \) or 200, are generated. The simulation results show that the proposed spline method is not sensitive to the selection of knots. Therefore, we only present the results with the number of internal knots being cubic root of sample size minus 1 and the placements of knots being chosen by the quantile method.

The Monte Carlo sample bias, standard deviation, and mean squared error for the semiparametric maximum likelihood estimator proposed by Huang [10] and the spline likelihood estimator of \( \beta \) are summarized in Table 1. For the current simulation setting, we can directly compute the efficient information \( I(\beta_0) = 2/3 \) and the asymptotic standard error \( \sigma(\hat{\beta}_n) = (1.5/n)^{1/2} \). The asymptotic standard errors are included to compare Monte Carlo standard deviations in Table 1. Also, we estimate \( \sigma^2 \) by
\[ \hat{\sigma}^2 = (n - d - k_n)^{-1} \sum_{i=1}^{n} \left( Y_i - X_i\hat{\beta}_n - \sum_{j=1}^{k_n} \hat{\alpha}_j B_j(Z_i) \right)^2, \]
where \( k_n = m_n + 1 \) is the number of spline basis functions and \( d \) is the dimension of \( \beta \).
Table 1
Comparison of bias, asymptotic standard error (ASE), and mean squared error (MSE) between the spline likelihood estimator and the semiparametric maximum likelihood estimator of $\beta_0 = 0.2$, based on 2000 repeated samples, $n = 50, 100,$ and 200.

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>$n = 50$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Semiparametric</td>
<td>Spline</td>
<td>Semiparametric</td>
</tr>
<tr>
<td>Bias $\times 10^3$</td>
<td>-1.0755</td>
<td>0.1236</td>
<td>-0.1478</td>
</tr>
<tr>
<td>SD $\times 10^3$</td>
<td>90.0991</td>
<td>73.1878</td>
<td>53.3625</td>
</tr>
<tr>
<td>ASE $\times 10^3$</td>
<td>61.2372</td>
<td>61.2372</td>
<td>43.3012</td>
</tr>
<tr>
<td>MSE $\times 10^3$</td>
<td>8.2656</td>
<td>5.3564</td>
<td>2.8475</td>
</tr>
</tbody>
</table>

Table 2
Comparison of bias and mean squared error (MSE) between the spline likelihood estimator (SL) and the semiparametric maximum likelihood estimator (ML) of $\psi_0(Z) = \exp(Z/4) - 1/2$, based on 2000 repeated samples, $n = 50, 100,$ and 200.

<table>
<thead>
<tr>
<th>$Z$</th>
<th>$\psi_0(Z)$</th>
<th>$n = 50$</th>
<th>$n = 100$</th>
<th>$n = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ML</td>
<td>SL</td>
<td>ML</td>
</tr>
<tr>
<td>Z = 1</td>
<td>0.7840</td>
<td>-0.0479</td>
<td>0.0122</td>
<td>-0.0192</td>
</tr>
<tr>
<td>Bias</td>
<td>SD</td>
<td>MSE</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>1.487</td>
<td>0.2451</td>
<td>0.0623</td>
<td>0.0377</td>
<td>0.1581</td>
</tr>
<tr>
<td>Z = 2</td>
<td>1.6170</td>
<td>-0.0552</td>
<td>0.0001</td>
<td>-0.0337</td>
</tr>
<tr>
<td>Bias</td>
<td>SD</td>
<td>MSE</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>2.2182</td>
<td>0.2891</td>
<td>0.0675</td>
<td>-0.0801</td>
<td>0.1523</td>
</tr>
<tr>
<td>Z = 3</td>
<td>2.9903</td>
<td>-0.1086</td>
<td>0.0271</td>
<td>-0.0531</td>
</tr>
<tr>
<td>Bias</td>
<td>SD</td>
<td>MSE</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>3.9816</td>
<td>0.3265</td>
<td>0.1185</td>
<td>-0.1086</td>
<td>0.1646</td>
</tr>
<tr>
<td>Z = 4</td>
<td>5.2546</td>
<td>-0.1670</td>
<td>0.0270</td>
<td>-0.0705</td>
</tr>
<tr>
<td>Bias</td>
<td>SD</td>
<td>MSE</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>5.2546</td>
<td>0.3479</td>
<td>0.1698</td>
<td>-0.2119</td>
<td>0.1826</td>
</tr>
<tr>
<td>Z = 7</td>
<td>6.8890</td>
<td>-0.2828</td>
<td>0.0333</td>
<td>-0.1222</td>
</tr>
<tr>
<td>Bias</td>
<td>SD</td>
<td>MSE</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>6.8890</td>
<td>0.4863</td>
<td>0.3165</td>
<td>-0.2828</td>
<td>0.1931</td>
</tr>
<tr>
<td>Z = 8</td>
<td>8.9877</td>
<td>-0.3786</td>
<td>0.0372</td>
<td>-0.1890</td>
</tr>
<tr>
<td>Bias</td>
<td>SD</td>
<td>MSE</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>8.9877</td>
<td>0.5534</td>
<td>0.4496</td>
<td>-0.3786</td>
<td>0.2427</td>
</tr>
</tbody>
</table>

The simulation results show that the sample biases are small and the standard deviations and the mean squared errors decrease when the sample size $n$ increases for both the semiparametric maximum likelihood estimator and the spline estimator. The same pattern is also observed for the estimation of $\sigma^2$. Moreover, the standard errors derived from the asymptotic theory are close to the corresponding standard deviations based on the Monte Carlo simulations, which provides an numerical justification for the asymptotic normality result in Theorem 1.

To compare both the semiparametric maximum likelihood estimator and the spline-based estimator for $\psi(Z)$ in detail, we calculate the estimates of $\psi(Z)$ at the points $= 1, \ldots, 9$. Table 2 displays the pointwise sample biases and the mean squared errors for the estimators of $\psi(Z) = \exp(Z/4) - 1/2$, based on 2000 Monte Carlo samples. It can be seen from Table 2 that the biases of the spline estimator are much smaller than those of the semiparametric estimator. Clearly, the spline likelihood estimator has smaller variability than the semiparametric maximum likelihood estimator and the variances of both estimators decrease when the sample size increases.
For each sample size, we evaluate the powers of tests resulting from asymptotic results in Section 3 from 2000 replications. Let $\hat{\beta}_{\text{ml}}$ and $\hat{\beta}_{\text{sl}}$ be the semiparametric maximum likelihood estimator and the spline likelihood estimator, respectively. For hypothesis $H_0 : \beta = \beta_0$, we compare the powers of four test statistics, $T_1 = (\hat{\beta}_{\text{ml}} - \beta_0)^2 / \text{se}(\hat{\beta}_{\text{ml}})^2$, $T_2 = (\hat{\beta}_{\text{sl}} - \beta_0)^2 / \text{se}(\hat{\beta}_{\text{sl}})^2$, $T_3 = (\hat{\beta}_{\text{sl}} - \beta_0)^2 / \hat{I}(\hat{\beta}_{\text{sl}})$, and $T_4 = 2(p_{\text{ml}}(\hat{\beta}_{\text{ml}}) - p_{\text{sl}}(\hat{\beta}_{\text{sl}}))$, where $\text{se}(\hat{\beta}_{\text{ml}})$ and $\text{se}(\hat{\beta}_{\text{sl}})$ are bootstrap standard errors and $\hat{I}(\hat{\beta}_{\text{sl}})$ is the discretized version of the observed profile information. Under $H_0$, Wald test statistics $T_1$, $T_2$, and $T_3$ and the likelihood ratio test statistics $T_4$, follow $\chi^2$ distribution with degree of freedom 1. All the tests are run at 5% significance level; the 95th percentile of the $\chi^2$ distribution with degree of freedom 1 are used as the critical point for all tests. The powers are computed as the proportions of $H_0$ being rejected in 2000 simulations. For each Monte Carlo sample, 100 bootstrap samples are generated to estimate the standard errors of $\hat{\beta}_{\text{ml}}$ and $\hat{\beta}_{\text{sl}}$. The results are summarized in Table 3 and Fig. 1. As described in Table 3, the spline likelihood ratio test with the test statistic $T_4$ and the spline Wald test based on the observed information method with the test statistic $T_3$ are almost identical and have the highest power. The semiparametric Wald test based on the bootstrap method with the test statistic $T_1$ has the lowest power. As expected, the power increases as the sample size or effect size increase. For all sample sizes, the symmetry of the power curve around the true parameter $\beta = 0.2$ is pronounced. Furthermore, the sizes of all tests are close to nominal level 5%.

### 4.2. A real data example

We apply the proposed method to a study where air pollution at a road is related to the traffic volume and meteorological variables, measured at Alnabru in Oslo, Norway, between October 2001 and August 2003 by Norwegian Public Roads Administration. In this paper, we studied the partly linear model

\[ Y = X\beta + \psi(Z) + \epsilon, \]

where the dependent variable $Y$ is the hourly value of logarithm of the concentration of NO$_2$ (particles), $Z$ is the logarithm of the number of cars per hour, $X$ is the two meter above ground temperature (°C), and $\psi$ is an unspecified smooth monotone increasing function. The sample size is 500. Since there is empirical evidence that more cars result in higher concentration of NO$_2$, it is reasonable to assume the monotonicity of $\psi$. In this study we want to test the association between the air pollution and the traffic volume and some meteorological variable, the ground temperature. The null hypothesis is $H_0 : \beta = 0$.

We consider both the semiparametric estimation and the spline estimation. For the spline estimation, the results are similar for different combinations of the number and placements of knots as described in the simulation section, which shows that the selection of the number and locations of knots is insensitive to the choice of knots in real application. Therefore, we only present the results with the number of interior knots being the cubic root of the distinct observation times and the places of knots being determined by quantile. The spline Wald test and the spline likelihood ratio test are used for the inference of $\beta$. For comparison purpose, the semiparametric Wald test is also included. The asymptotic
standard error of the spline estimator of \( \beta \) is estimated by the observed profile information approach and the bootstrap method. The standard error of the semiparametric estimator is estimated by the bootstrap method only. For each scenario, 1000 replications of bootstrap samples are generated in this study. We also plot the spline estimator and step function estimator of \( \psi(Z) \). The results are summarized in Table 4 and Fig. 2.

Both the semiparametric method and the spline-based method yield the same result that the NO\(_2\) concentration tends to be lower when the temperature increases, \( p < 0.0001 \) and \( p < 0.0001 \), respectively. When the temperature increases by 1 \(^\circ\)C, the hourly value of the concentration of NO\(_2\) will decrease by 1.66\% or 1.74\% by the semiparametric estimation or the spline estimation, respectively. As seen in Fig. 2, the spline estimator smoothes out the step function estimator of \( \psi(Z) \).

**Table 4**

<table>
<thead>
<tr>
<th>Method</th>
<th>( \hat{\beta} )</th>
<th>se(( \hat{\beta} ))</th>
<th>( \chi^2 )</th>
<th>( p )-value</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semiparametric bootstrap</td>
<td>-0.0167</td>
<td>0.00236</td>
<td>50.428</td>
<td>&lt;0.0001</td>
<td>(-0.0214, -0.0121)</td>
</tr>
<tr>
<td>Spline bootstrap</td>
<td>-0.0176</td>
<td>0.00248</td>
<td>50.285</td>
<td>&lt;0.0001</td>
<td>(-0.0224, -0.0127)</td>
</tr>
<tr>
<td>Spline observed information</td>
<td>-0.0176</td>
<td>0.00305</td>
<td>33.206</td>
<td>&lt;0.0001</td>
<td>(-0.0235, -0.0116)</td>
</tr>
<tr>
<td>Spline likelihood ratio test</td>
<td>-0.0176</td>
<td>NA</td>
<td>32.039</td>
<td>&lt;0.0001</td>
<td>(-0.0276, -0.0076)</td>
</tr>
</tbody>
</table>

Fig. 1. Power curves of Wald tests and likelihood ratio test, based on 2000 repeated samples, \( n = 50, 100, \) and 200.
5. Final remarks and future problems

The proposed monotone $B$-spline method shows not only good theoretical properties but also desirable finite sample performance. The spline estimator of $\psi$ achieves the rate of convergence faster than $n^{1/3}$ when the true function $\psi_0$ is sufficiently smooth. Moreover, the estimate of $\beta$ is asymptotically normal and efficient. Furthermore, the spline estimator of $\psi$ has the smaller variability than its alternative proposed by Huang [10]. Finally, the spline method is robust to the selection of knots in our simulation setting. The monotone $B$-spline method presented here can be applied to other semiparametric models, for instance, Cox proportional hazard model for current status data [11], proportional odds regression model [12], and hazard regression [15]. As a concluding remark, the proposed method provides a useful approach in application to semiparametric models with monotonicity constraints on the nonparametric component. In this manuscript, we have used the pre-specified number and places of knots. One may investigate further to adaptively select the number and placements of knots. It would also be desirable to study other algorithms to compute the estimator with monotonicity constraints. Finally, the limiting distribution of the spline estimator of $\psi$ needs to be explored in the future.

Appendix

For simplicity we assume that $X \in \mathbb{R}$. The general case can be proved similarly. Let $P_{\beta, \psi}$ be the distribution of $(Y, Z, X)$ under parameter $\vartheta = (\beta, \psi)$ and $p_{\beta, \psi}$ be the corresponding density. Also define $P_0 \equiv P_{\beta_0, \psi_0}$ and $p_0 \equiv p_{\beta_0, \psi_0}$. Given a random sample $X_1, \ldots, X_n$ with probability measure $P$ on a measurable space $(\mathcal{X}, \mathcal{A})$, for a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, define $P f = \int f \, dP$ as the expectation of $f$ under $P$ and $P_n f = n^{-1} \sum_{i=1}^n f(X_i)$ as the expectation of $f$ under the empirical measure $P_n$. We write $G_f = \sqrt{n}(P_n - P_0) f$ for the empirical process evaluated at $f$ and $\|G_n\|_F = \sup_{f \in F} |G_n f|$ for any measurable class of functions $F$.

A.1. Proof of Theorem 1(a) (consistency)

Let $M(\vartheta) = P_l(\vartheta)$ and $M_n(\vartheta) = P_{nl}(\vartheta)$. Recall that $\mathcal{F}$ is the class of monotone nondecreasing functions on $[0, \tau]$. Define $\mathcal{L}_1 = \{(\beta, \psi) : (\beta, \psi) \in \Theta \times \mathcal{F}\}$. According to Theorem 2.7.5 of van der Vaart [33], for any $\varepsilon > 0$, the logarithm of the bracketing number of $\mathcal{F}$ computed with $L_2(P)$ is bounded by $1/\varepsilon$, up to a constant. Hence, $\mathcal{F}$ is a $P$-Donsker class. Furthermore, $X$ has a bounded support and $\Theta$ is compact. Therefore, we can show that $\mathcal{L}_1$ is $P$-Glivenko–Cantelli. It yields $\sup_{(\beta, \psi) \in \Theta \times \mathcal{F}} |M_n(\beta, \psi) - M(\beta, \psi)| = o_p(1)$. Thus, we have uniform convergence of $M_n$ to $M$ on $\Theta \times \mathcal{F}$.

A straightforward algebra yields

$$M(\vartheta_0) - M(\vartheta) = P(g + h)^2/(2\sigma^4),$$

where $g = X\beta - X\beta_0$ and $h = \psi - \psi_0$. Note that

$$(Pgh)^2 = \sigma^4(\beta - \beta_0)^2[P l_\psi(\beta_0, \psi_0) h l_\beta(\beta_0, \psi_0)]^2.$$
Since $\hat{P}_n(\beta_0, \psi_0)h\hat{I}_n(\beta_0, \psi_0) = 0$, for any $h$, we have

$$[\hat{P}_n(\beta_0, \psi_0)h\hat{I}_n(\beta_0, \psi_0)]^2 = [\hat{P}_n(\beta_0, \psi_0)h(\hat{I}_n(\beta_0, \psi_0) - I_0(\beta_0, \psi_0))]^2.$$ 

By Cauchy–Schwarz inequality and the fact that $\hat{P}(\hat{I}_n(\beta_0, \psi_0) - I_0(\beta_0, \psi_0))^2 = CP(\hat{I}_n(\beta_0, \psi_0))^2$, for $0 < C < 1$, we obtain

$$[\hat{P}_n(\beta_0, \psi_0)h\hat{I}_n(\beta_0, \psi_0)]^2 \leq CP(\hat{I}_n(\beta_0, \psi_0))^2P(\hat{I}_n(\beta_0, \psi_0)h).$$

Therefore, $(\hat{p}gh)^2 \leq CP^2Ph^2$, for $0 < C < 1$. According to Lemma A.6 of Murphy and van der Vaart [22], there exists some $C > 0$ such that

$$P(g + h)^2 \geq Cd_2^2(\hat{\theta}, \beta_0).$$

Hence, $M(\hat{\theta}_n) - M(\beta_0) \geq Cd_2^2(\hat{\theta}, \beta_0)$, for $C > 0$. Then, it implies $\sup_{\theta \in \Theta} M(\hat{\theta}_n) \leq \sup_{\theta \in \Theta} M(\hat{\theta}_n) - C \varepsilon^2 < M(\beta_0).$

Following the same lines as those in Lemma A1 of Lu et al. [17], we can show that there exists a $\psi_{0,n} \in M_n$ of order $l \geq r + 2$ such that $\|\psi_0 - \psi_{0,n}\| = O(n^{-r'})$, for $1/(2r + 2) < \nu < 1/(2r)$. Denote $\hat{\theta}_n = (\hat{\beta}_n, \hat{\psi}_n)$ and $\theta_{0,n} = (\beta_0, \psi_{0,n})$. We have

$$M_n(\hat{\theta}_n) - M_n(\theta_{0,n}) \geq M_n(\theta_{0,n}) - M_n(\theta_0) = \eta_1 + \eta_2,$$

where $\eta_1 = (P_n - P)(l(\theta_{0,n}) - l(\theta_0))$ and $\eta_2 = M(\theta_{0,n}) - M(\theta_0)$.

We write $\eta_1$ as

$$\eta_1 = n^{-r'}P_n(l(\theta_{0,n}) - l(\theta_0))/n^{r' + 2},$$

for $0 < \varepsilon < 1/2 - r$.

Define class

$$\mathcal{L}_2 = \{l(\beta_0, \psi) - l(\beta_0, \psi_0) : \psi \in \mathcal{F}, \|\psi - \psi_0\| \leq \eta\}.$$

The fact that $\mathcal{F}$ is $P$-Donsker and conditions (C2) and (C5) yield $\mathcal{L}_2$ is $P$-Donsker. Furthermore, by the boundedness of $\psi$ and $\psi_0$, we have

$$P\left(\frac{1}{\psi_0 - \psi_0}/n^{r' + 2}\right)^2 \leq \frac{C\psi_0 - \psi_0}{r' + 2} = 0.$$

as $n \to \infty$, for $\eta = O(n^{-r'})$. According to Lemma 19.24 of van der Vaart [33], we obtain $(P_n - P)(l(\beta_0, \psi_{0,n}) - l(\beta_0, \psi_0))/n^{r' + 2} = o_p(n^{-1/2})$, and hence $\eta_1 = o_p(n^{-1/2}) = o_p(n^{-2r'})$. Furthermore, $\eta_2 = -C\psi_0 - \psi_0 = -C\psi_0 - o_p(1)$. We conclude that

$$M_n(\hat{\theta}_n) - M_n(\theta_{0,n}) > -O_p(n^{-2r'}) \geq -O_p(n^{-2\min(1 - (n^{-2}))/2}) = o_p(1).$$

The uniform convergence of $M_n$ to $M$ on $\Theta \times \mathcal{F}$ implies $M_n(\theta_0) \to M(\theta_0)$, in probability. It follows that $M_n(\hat{\theta}_n) \geq M(\theta_0) - o_p(1)$.

Therefore,

$$M(\theta_0) - M(\hat{\theta}_n) \leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_p(1) \leq \sup_{\theta \in \Theta \times \mathcal{F}} |M_n - M(\theta) + o_p(1) \to 0|,$$

in probability. The last inequality holds because of the uniform convergence of $M_n$ to $M$ on $\Theta \times \mathcal{F}$.

For every $\varepsilon > 0$, by $\sup_{d_2(\theta, \theta_0) \geq 0} M(\theta) < M(\theta_0)$, there exists a number $\eta > 0$, such that $M(\theta) < M(\theta_0) - \eta$, for every $\theta$ with $d_2(\theta, \theta_0) \geq \varepsilon$. Thus, the event $\{d_2(\hat{\theta}_n, \theta_0) \geq \varepsilon\}$ is contained in the event $\{M(\hat{\theta}_n) < M(\theta_0) - \eta\}$. The probability of latter event converges to 0 by the preceding display. This completes the proof of $d_2(\hat{\theta}_n, \theta_0) = o_p(1)$.

A.2. Proof of Theorem 1(b) (rate of convergence)

We apply Theorem 3.4.1 of van der Vaart and Wellner [34] to prove the rate of convergence. Denote the regression function by $g(z) = X\beta + \psi(z)$. Let $(\beta_0, \psi_0)$ be the true parameter. Denote $g_0(z) = X\beta_0 + \psi_0(z)$. In the proof of consistency, we show that there exists $\psi_0 \in M_n$ of order $l \geq r + 2$ such that $\|\psi_0 - \psi_0\| = O(n^{-r'})$, for $1/(2r + 2) < \nu < 1/(2r)$. Let $g_n(z) = X\beta_0 + \psi_n(z)$. Also denote the estimate of $g_0(z)$ by $\hat{g}_n(z) = X\hat{\beta}_n + \hat{\psi}_n(z)$. Define $l(g) = -1/(2\sigma^2)(Y - g)^2$ and $M(g) = \hat{P}(g)$. First we need to find $\phi_0(\eta)/\eta$ is decreasing in $\eta$ and

$$E_{\eta/2 \leq \nu \leq \eta} |G_0 l(g) - G_n l(g_0)| \leq C \phi_0(\eta).$$

Define class

$$\mathcal{L}_3 = \{l(g) - l(g_0), \psi \in M_n \text{ and } g - g_0 \geq \eta \leq \eta\}.$$
For any ε ≥ 0 and ε ≤ η, by the calculation of Shen and Wong [31], the logarithm of the bracketing number of $\mathcal{M}_n$ computed with $L_2(P)$ can be bounded by $k_n \log(\eta/\epsilon)$, up to a constant. Furthermore, by conditions (C2) and (C5), we can show that, for some $C > 0$,

$$J_l(\eta, \mathcal{L}_3, \|\cdot\|_{P, B}) \leq Ck_n^{1/2} \eta,$$

where $\|\cdot\|_{P, B}$ is the Bernstein norm defined as $\|f\|_{P, B} = \{2P(|e|^l - |f|)\}^{1/2}$ in [34]. Moreover, some algebra leads to $\|l(g) - l(g_0)\|_{P, B} \leq Cn^{1/2}$, for some $C > 0$ and any $l(g) - l(g_0) \in \mathcal{L}_3$. According to Lemma 3.4.3 of van der Vaart and Wellner [34], we obtain

$$E_P\|G_n\|_{\mathcal{L}_3} \leq J_l(\eta, \mathcal{L}_3, \|\cdot\|_{P, B}) \left(1 + \frac{J_l(\eta, \mathcal{L}_3, \|\cdot\|_{P, B})}{n^{1/2} \eta^2}\right) \leq C(k_n^{1/2} \eta + n^{1/2}).$$

Hence, we choose $\phi_n(\eta) = \eta^{1/2} + k_n^{1/2} \eta$. Clearly, $\phi_n(\eta)/\eta$ is decreasing in $\eta$. Therefore, by Theorem 3.4.1 of van der Vaart and Wellner [34], choosing the distance $d_n$ defined in the theorem to be $d_n^2(\hat{g}_n, g_0) = \|g_n - g_0\|^2$, we have $r_n^2(M(g_n) - M(\hat{g}_n)) = O_P(1)$, where $r_n$ satisfies

$$r_n^2(k_n^{1/2} r_n^{-1} + k_n^{1/2}) = O(n^{1/2}).$$

Note that

$$n^{1-\nu} \phi_n(1/n^{(1-\nu)/2}) = 2n^{1/2},$$

and that, if $(1 - \nu)/2 \geq rv$,

$$n^{2rv} \phi_n(1/n^{rv}) = n^{1/2} \{n^{rv-(1-\nu)/2} + n^{2rv-(1-\nu)}\} \leq 2n^{1/2}.$$ 

It follows that $r_n = n^{\min(rv, (1-\nu)/2)}$. Note that

$$M(g_n) - M(g_0) = M(\hat{g}_n) - M(g_0) + M(g_0) - M(g_0) = -\|\psi_n - \psi_0\|^2/(2\sigma^2) + \|\hat{g}_n - g_0\|^2/(2\sigma^2),$$

and

$$\|\hat{g}_n - g_0\|^2 \geq \|\hat{g}_n - g_0\|^2 - \|g_n - g_0\|^2 = \|\hat{g}_n - g_0\|^2 - \|\psi_n - \psi_0\|^2/(2\sigma^2).$$

It follows

$$\|\hat{g}_n - g_0\|^2 \leq O_P(r_n^{-2}) + O_P(n^{-2rv}) = O_P(r_n^{-2}).$$

Because $\|g_n - g_0\|^2 = O_P(n^{-2rv})$, we have $\|\hat{g}_n - g_0\|^2 = O_P(r_n^{-2})$. In the proof of consistency we have already shown that $\|\hat{g}_n - g_0\|^2 = 2n^2(M(\hat{g}_n) - M(\hat{g}_n)) \geq C(n^2(\hat{\theta}_n, \theta))$, for $C > 0$. Hence, $r_n^2d_n^2(\hat{g}_n, \theta_0) = O_P(1)$.

A.3. Proof of Theorem 1(c) (asymptotic normality)

In this section we show that the spline estimator $\hat{g}_n$ for $\beta_n$ is asymptotically efficient. Since $h^*(z) = E(X|Z = z)$ has the bounded derivative on $[0, \tau]$, according to Jackson’s theorem for polynomial in [4], there exists a spline $\varphi_n$ with order $l \geq 2$ and knots $T_n$ satisfying

$$0 = t_1 = \ldots = t_l < t_{l+1} < \ldots < t_{m+1} < t_{m+1+l} = \ldots = t_{m+2l} = \tau,$$

such that

$$\|h^* - \varphi_n\|_{\infty} = O_P(n^{-rv}).$$

for $1/(2r + 2) < v < 1/(2r)$. Choosing small enough $s$ will lead to $\hat{\psi}_n + s\varphi_n \in \mathcal{M}_n$. Hence,

$$\frac{d}{ds} \sum_{i=1}^n (Y_i - (\hat{\psi}_n - s\varphi_n) - X_i(\hat{\beta}_n + s))^2|_{s=0} = 0.$$

Let $\hat{h}_n = Y - \hat{\psi}_n - X\hat{\beta}_n$ and $h_0 = Y - \psi_0 - X\beta_0$. We have

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\psi}_n - X_i\hat{\beta}_n)(X_i - \varphi_n) = P_n\hat{g}_n(X - E(X|Z)) + P_n\hat{g}_n(E(X|Z) - \varphi_n) = I_{h_3} + I_{h_4},$$

We decompose $I_{h_3}$ to

$$P_n\hat{g}_n(X - E(X|Z)) + P_n(\psi_0 - \hat{\psi}_n)(X - E(X|Z)) = \sigma^2 P_n\hat{g}_n + \Delta_{h_1} - \Delta_{h_2}.$$
function where consistency and rate of convergence of $\psi$ complete the asymptotic normality. Thus, $I_3 = \sigma^2 P \hat{r}^2 - \hat{\beta}_0 (\sigma^2 I(\beta_0) + o_p(1))$.

$I_4$ can be written as

$$\left( P_n - P \right) \hat{h}_n (E(X|Z) - \varphi_n) + P \hat{h}_n (E(X|Z) - \varphi_n).$$

According to the bracketing number calculation in [31], for any $\varepsilon > 0$ and $\varepsilon \leq \eta$, the logarithm of bracketing number of the class of splines on $[0, \tau]$ computed with $L_2(P)$ is bounded by $k_n \log(\eta/\varepsilon)$, up to a constant. We can show that the bracket integral of the class of $\hat{h}_n (E(X|Z) - \psi)$ for $\psi$ ranging over the class of splines on $[0, \tau]$ with $\|E(X|Z) - \psi\|_\infty \leq \eta$ is $k_n^{1/2} \eta$, up to a constant. By Lemma 3.4.3 of van der Vaart and Wellner [34], we have

$$E_P \|G\| \leq C (k_n^{1/2} \eta + n^{-1/2} k_n) = o_p(1),$$

for $\eta = o_p(n^{-r\varepsilon})$. Furthermore,

$$\|\hat{h}_n (E(X|Z) - \psi)\|_{2, b}^2 \leq C \|\hat{h}_n (E(X|Z) - \psi)\|^2 \leq C \|E(X|Z) - \psi\|_\infty^2 \hat{h}_n^2 \leq C \eta^2,$$

for any $\|E(X|Z) - \psi\|_\infty \leq \eta$. Therefore,

$$\left( P_n - P \right) \hat{h}_n (E(X|Z) - \varphi_n) = o_p(n^{-1/2}).$$

The second term in $I_4$ can be written as

$$P(\psi_0 - \hat{\psi}_n) (E(X|Z) - \varphi_n) + PX(\beta_0 - \hat{\beta}_n) (E(X|Z) - \varphi_n).$$

The rate of convergence of $\hat{\beta}_n$, $\hat{\psi}_n$ and $\|E(X|Z) - \varphi\|_\infty$ lead to

$$P(\psi_0 - \hat{\psi}_n) (E(X|Z) - \varphi_n) = o_p(n^{-1/2}),$$

and

$$PX(\beta_0 - \hat{\beta}_n) (E(X|Z) - \varphi_n) = o_p(n^{-1/2}).$$

Thus, $I_4 = o_p(n^{-1/2}).$

Therefore,

$$\sqrt{n} (\hat{\beta}_n - \beta_0) (\sigma^2 I(\beta_0) + o_p(1)) = \sigma^2 \sqrt{n} \hat{r}^2 + o_p(1).$$

Thus, the result follows from Central Limit Theorem, Slutsky’s Lemma, and nonsingularity of information $I(\beta_0)$. This completes the asymptotic normality.

### A.4. Proof of Theorem 2 (likelihood ratio test)

The proof of Theorem 2 is based on Theorem 3.1 of Murphy and van der Vaart [22]. Let $\hat{\psi}_0$ be the maximum likelihood estimator of $\psi_0$, given $\beta = \beta_0$. We can deduce $\|\hat{\psi}_0 - \psi_0\|_2 = o_p(n^{-1/2})$ using the similar arguments in the proofs of consistency and rate of convergence of $\hat{\beta}_n$, $\hat{\psi}_n$. Define the approximately least favorable submodel

$$\Psi_t(\beta, \psi) = (t, \psi_t(\beta, \psi)).$$

where

$$\psi_t(\beta, \psi) = \psi + (\beta - t) h^* \circ \psi_0^{-1} \circ \psi.$$

Since $h^*$ and $\psi_0^{-1}$ are bounded and Lipschitz, $\Phi_t(\beta, \psi)$ is a valid parameter for $t$ sufficiently close to $\beta$.

Let $p(t, \beta, \psi)$ and $l(t, \beta, \psi)$ be density and log density functions under parameter value $(t, \psi_t(\beta, \psi))$. We have the score function

$$\tilde{l}(t, \beta, \psi) = \partial l(t, \beta, \psi)/\partial t = (Y - \psi_t - X t)(X - h^* \circ \psi_0^{-1} \circ \psi)/\sigma^2.$$  

Note that $\psi_t(t, \beta, \psi)$ converges to $\psi_0$ as $(t, \beta, \psi)$ tends to $(\beta_0, \beta_0, \psi_0)$. We have

$$\tilde{l}(t, \beta, \psi) \rightarrow (Y - \psi_0 - X \beta_0)(X - E(X|Z))/\sigma^2 = I^*_\beta.$$
as \((t, \beta, \psi)\) tends to \((\beta_0, \beta_0, \psi_0)\). We can show that

\[
\mathcal{L}_4 = \{ (t, \beta, \psi) : \psi \in \mathcal{F}, |t - \beta_0| < \eta, |\beta - \beta_0| < \eta \}
\]

is a Donsker class by the fact that \(\mathcal{F}\) is \(P\)-Donsker and \(\psi_0\) and \(X\) are bounded and \(\Theta\) is compact.

Now we verify condition (3.14) in [22]. The term \(p^{-1}(t \cdot \psi(t, \beta, \psi)) \partial^2 p(t \cdot \psi(t, \beta, \psi)) / \partial^2 t\) can be written as

\[
(X - h^* \circ \psi^{-1}_0 \circ \psi)^2 \{ (Y - \psi_t - X t)^2 - \sigma^2 \} / \sigma^4.
\]

Using the same arguments as before, we can conclude that the class of these functions is \(P\)-Donsker, and hence \(P\)-Glivenko–Cantelli. Furthermore, as \((t, \beta, \psi)\) approaches to \((\beta_0, \beta_0, \psi_0)\), \(p^{-1}(t \cdot \psi(t, \beta, \psi)) \partial^2 p(t \cdot \psi(t, \beta, \psi)) / \partial^2 t\) converges to

\[
(X - E(X|Z))^2 (\sigma^2 - \sigma^2) / \sigma^4
\]

with mean 0. Thus, condition (3.14) holds.

Finally, we verify the “unbiasedness” condition

\[
\sqrt{n} \hat{p}_0 (\beta_0, \beta_0, \hat{\psi}_0) = o_p(1).
\]

For notational convenience, abbreviate \(\hat{p}(\beta_0, \beta_0, \psi)\) to \(\hat{p}(\psi)\). We have

\[
P_0 \hat{p}(\psi_0) = (P_0 - P_{\beta_0, \hat{\psi}_0}) \hat{p}(\psi_0) + (P_0 - P_{\beta_0, \hat{\psi}_0}) (\hat{p}(\psi_0) - \hat{p}(\psi_0)) = I_{n_3} + I_{n_6}.
\]

The above decomposition holds due to the fact \(P_{\beta, \hat{\psi}} (\beta, \hat{\psi}) = 0\), for all \((\beta, \hat{\psi})\). Since \(\hat{p}(\psi_0)\) is the efficient score function for \(\beta, P_0 (\hat{p}(\psi_0) \hat{p}(\beta_0, \psi_0)(\psi_0 - \hat{\psi}_0)) = 0\). Furthermore,

\[
P_0 \hat{p}(\psi_0) (P_0 - P_{\beta_0, \hat{\psi}_0}) / P_0 = (P_0 - P_{\beta_0, \hat{\psi}_0}) \hat{p}(\psi_0).
\]

Therefore, we can write \(I_{n_3}\) as

\[
P_0 \left\{ \hat{p}(\psi_0) \left[ (P_0 - P_{\beta_0, \hat{\psi}_0}) / P_0 - \hat{p}(\beta_0, \psi_0)(\psi_0 - \hat{\psi}_0) \right] \right\}.
\]

By Taylor expansion, we have

\[
I_{n_3} = -P_0 \left[ \hat{p}(\psi_0)(2p(\beta_0, \psi_0))^{-1} \partial^2 p(\beta_0, \psi_0 + t^*)(\psi_0 - \hat{\psi}_0) / \partial t^2 \right] .
\]

Let \(h = \hat{\psi}_0 - \psi_0\) and \(Q = Y - (\psi_0 + \theta t) - X \beta_0 \). \(\partial^2 p(\beta_0, \psi_0 + \theta t) / \partial^2 t\) can be written as

\[
\exp(-Q^2 / 2 \sigma^2) \{ Q^2 - \sigma^2 \} \mu / \sigma^4.
\]

By conditions (C5) and (C9) and the fact that \(\psi_0\) and \(h\) are bounded, we have \(P_0 Q^2\) is bounded, and hence,

\[
|I_{n_3}| \leq CP_0 (\hat{\psi}_0 - \psi_0)^2 = o_p(n^{-1/2}).
\]

The last equality holds due to the rate of convergence of \(\hat{\psi}_0\). Now write \(I_{n_6}\) as

\[
\int (\hat{p}(\psi_0)) (p_0 - P_{\beta_0, \hat{\psi}_0}) d\mu = \Delta_{n_3} + \Delta_{n_4},
\]

where

\[
\Delta_{n_3} = -\int (\hat{l}(\psi_0) - \hat{l}(\psi_0)) \hat{l}(\beta_0, \psi_0)(\hat{\psi}_0 - \psi_0) p(\beta_0, \psi_0) d\mu,
\]

\[
\Delta_{n_4} = -\int 1/2 (\hat{l}(\psi_0) - \hat{l}(\psi_0)) \partial^2 p(\beta_0, \psi_0 + t^*(\psi_0 - \hat{\psi}_0)) / \partial t^2 d\mu.
\]

Note that \(\hat{l}(\psi_0) - \hat{l}(\psi_0)\) can be written as

\[
- (\hat{\psi}_0 - \psi_0)(X - h^* \circ \psi_0^{-1} \circ \hat{\psi}_0) / \mu^2 - (Y - \psi_0 - X \beta_0)(h^* \circ \psi_0^{-1} \circ \hat{\psi}_0 - h^* \circ \psi_0^{-1} \circ \psi_0) / \mu^2.
\]

We have \(X - h^* \circ \psi_0^{-1} \circ \hat{\psi}_0 = X - E(X|Z) + o_p(1)\) by the consistency of \(\hat{\psi}_0\) and \(h^* \circ \psi_0^{-1}\) being Lipschitz. Furthermore, \(|h^* \circ \psi_0^{-1} \circ \hat{\psi}_0 - h^* \circ \psi_0^{-1} \circ \psi_0| < \|h^* \circ \psi_0^{-1}\| \|\psi_0 - \psi_0\|\) by the property that \(h^*\) and \(\psi\) have bounded derivatives. Hence,

\[
|\hat{l}(\psi_0) - \hat{l}(\psi_0)| \leq C |\hat{\psi}_0 - \psi_0|.
\]

for \(C > 0\) independent of \((y, x, z), \) except on an event with probability converging to 0. Boundedness of \(P_0 |Y - \psi_0 - X \beta_0|\) along with

\[
\hat{l}(\beta_0, \psi_0)(\psi_0 - \psi_0) = (Y - \psi_0 - X \beta_0)(\hat{\psi}_0 - \psi_0) / \mu^2
\]

and
yield 
\[ |\Delta_n| \leq CP_0(\hat{\psi}_0 - \psi_0)^2 = o_p(n^{-1/2}). \]

Also, the uniform boundedness of \( P_0(\hat{\psi}_0 - \psi_0)^2 \) and \( P_0 |\partial^2 p(\beta_0, \psi_0 + t^* h)/\partial t^2| \leq CP_0(\hat{\psi}_0 - \psi_0)^2 \) yield 
\[ |\Delta_{\alpha_1}| \leq CP_0(\hat{\psi}_0 - \psi_0)^2 = o_p(n^{-1/2}). \]

Thus, \( P_0(\hat{\beta}_0, \beta_0, \hat{\psi}_0) = o_p(n^{-1/2}). \) This concludes that the unbiasedness condition holds. This completes the likelihood ratio test proof.

A.5. Proof of Theorem 3 (estimation of the standard errors)

Note that \( \hat{\psi}_\beta = \arg\max_{\psi \in \mathcal{M}_n} - (Y - \psi - X\beta)^2/(2\sigma^2) \) is continuous for \( \beta \). Thus, \( \hat{\psi}_\beta \rightarrow \hat{\psi}_0 \) in probability, for any sequence \( \hat{\beta} \rightarrow \beta_0 \), in probability.

We adopt the same approximately least favorable model defined in the likelihood ratio test proof; that is, 
\[ \Psi(t, \beta, \psi) = (t, \psi_t(\beta, \psi)), \]

where 
\[ \psi_t(\beta, \psi) = \psi + (\beta - t) h^* \circ \psi_0^{-1} \circ \psi. \]

First, we apply Theorem 3.2 of Murphy and van der Vaart [21] to verify 
\[ \|\hat{\psi}_\beta - \psi_0\|_2 = O_p(\|\beta - \beta_0\| + n^{-r/(1+2r)}), \] (A.1)

where \( r \) is the order of derivative of \( \psi_0 \). For \( (\beta, \psi) \) in the neighborhood of \( (\beta_0, \psi_0) \), we have 
\[ P_0(m_{\beta, \psi} - m_{\beta_0, \psi_0}) \geq -C\|\beta - \beta_0\|^2, \]

and using the same arguments in the proof of consistency, we can show that 
\[ P_0(m_{\beta, \psi} - m_{\beta_0, \psi_0}) \leq -C\|\beta - \beta_0\|^2. \]

Hence, conditions (3.7) and (3.8) in [21] are satisfied. Define 
\[ \mathcal{L}_5 = \{m_{\beta, \psi} - m_{\beta_0, \psi_0} : \psi \in \mathcal{M}_n, d_2(\psi, \psi_0) \leq \delta, \|\beta - \beta_0\| \leq \delta \}. \]

Using the similar arguments in the proof of rate of convergence, we can show that 
\[ J_1(\delta, \mathcal{L}_5, \|\cdot\|_{P, \beta}) \leq Ck^{1/2} \delta, \]
where \( \|\cdot\|_{P, \beta} \) is the Bernstein norm. For \( (\beta, \psi) \) ranging over a neighborhood of \( (\beta_0, \psi_0) \), the boundedness of \( \psi(Z) \) and \( \psi_0(Z) \) yields 
\[ P_0(m_{\beta, \psi} - m_{\beta_0, \psi_0})^2 \leq C\|\beta - \beta_0\|^2. \]

By inequality \( e^x - x - 1 \leq 2x^2 \), for \( x \) close to \( 1 \), we can show that 
\[ \|m_{\beta, \psi} - m_{\beta_0, \psi_0}\|^2_{P, \beta} \leq C\delta^2, \]

for any \( m_{\beta, \psi} - m_{\beta_0, \psi_0} \in \mathcal{L}_5 \). Therefore, according to Lemma 3.4.3 of van der Vaart and Wellner [34], we obtain 
\[ E_P |\hat{\psi}_\beta - \psi_0|^2 |\mathcal{L}_5 \leq C\|\beta - \beta_0\|^2 + n^{-1/2} \delta^{-2} J_1(\delta, \mathcal{L}_5, \|\cdot\|_{P, \beta}). \]

Hence, we choose \( \phi_0(\delta) = k_0^{1/2} \delta + k_0 n^{-1/2} \). It is easy to see that \( \phi_0(\delta)/\delta \) is decreasing in \( \delta \). The sequence \( \delta_n = n^{r/(1+2r)} \)

satisfies \( \phi_0(\delta_n) \leq 2\sqrt{n}\delta_n \) for every \( n \). Thus, Theorem 3.2 of Murphy and van der Vaart [21] yields (A.1). The conditions of Lemma 2.2 in [21] can be verified using the same arguments in the proof of likelihood ratio test. All that remains for application of Theorem 2.1 of Murphy and van der Vaart [21] is to verify condition (2.7).

Differentiating log density function \( l(t, \beta, \psi) \) under parameter value \( (t, \psi_t(\beta, \psi)) \) yields the score function 
\[ \hat{l}(t, \beta, \psi) = \partial l(t, \beta, \psi)/(\partial \beta, \partial \psi) = (Y - \psi_t - X\beta)(X - h^* \circ \psi^{-1}_0 \circ \psi)/\sigma^2. \]

For some sequence \( \hat{\beta} \rightarrow \beta_0 \), in probability, we have 
\[ P_0(\hat{l}(\beta)) = (P_0 - P_{\beta_0} \hat{\beta}) \hat{l}(\beta_0) + (P_0 - P_{\beta_0} \hat{\beta}) \hat{l}(\beta - \beta_0). \]

Using same arguments in the proof of likelihood ratio test and (A.1), we get 
\[ \|I_{n_1}\| \leq CP_0(\hat{\psi}_\beta - \psi_0)^2 = O_p(\|\hat{\beta} - \beta_0\|^2 + n^{-1/2} \delta^{-2} J_1(\delta, \mathcal{L}_5, \|\cdot\|_{P, \beta})), \]

and 
\[ \|I_{n_2}\| \leq CP_0(\hat{\psi}_\beta - \psi_0)^2 = O_p(\|\hat{\beta} - \beta_0\|^2 + n^{-1/2} \delta^{-2} J_1(\delta, \mathcal{L}_5, \|\cdot\|_{P, \beta})), \]

This concludes that 
\[ P_0(\hat{l}(\beta_0, \beta_0, \hat{\psi}_\beta)) = o_p(\|\hat{\beta} - \beta_0\|^2 + n^{-1/2}). \]

Finally, we have 
\[ P_0(\hat{l}(\beta_0, \beta_0, \hat{\psi}_\beta)) - l(\beta_0, \beta_0, \hat{\psi}_\beta)) = -P_0 h^* \circ \psi^{-1}_0 \circ \psi^*_\beta (X - h^* \circ \psi^{-1}_0 \circ \psi^*_\beta)/\sigma^2 \]

converging to 0. Hence, 
\[ P_0(\hat{l}(\beta_0, \beta_0, \hat{\psi}_\beta)) = P_0(\hat{l}(\beta_0, \beta_0, \hat{\psi}_\beta) + o_p(\|\hat{\beta} - \beta_0\|)). \]

This concludes that condition (2.7) of Murphy and van der Vaart [21] is satisfied.
References