Journal of Multivariate Analysis 162 (2017) 1–15

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Statistical inference for generalized additive partially linear models

Rong Liu [a,*], Wolfgang K. Härdle [b,c], Guoyi Zhang [d]

[a] *Department of Mathematics and Statistics, University of Toledo, Toledo, OH, United States*
[b] *Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Germany*
[c] *School of Business, Singapore Management University, Singapore*
[d] *Department of Mathematics and Statistics, The University of New Mexico, NM, United States*

ARTICLE INFO

Article history:
Received 23 September 2016
Available online 18 August 2017

Keywords:
B-spline
Default
Empirical likelihood
Kernel estimator
Link function
Mixing

ABSTRACT

The class of Generalized Additive Models (GAMs) is a powerful tool which has been well studied. It helps to identify additive regression structure that can be determined even more sharply via test procedures when some component functions have a parametric form. Generalized Additive Partially Linear Models (GAPLMs) enjoy the simplicity of GLMs and the flexibility of GAMs because they combine both parametric and nonparametric components. We use the hybrid spline-backfitted kernel estimation method, which combines the best features of both spline and kernel methods, to make fast, efficient and reliable estimation under an α-mixing condition. In addition, simultaneous confidence corridors (SCCs) for testing overall trends and empirical likelihood confidence regions for parameters are provided under an independence condition. The asymptotic properties are obtained and simulation results support the theoretical properties. As an illustration, we use GAPLM methodology to improve the accuracy ratio of the default predictions for 19,610 German companies. The quantlet for this paper are available on https://github.com.

1. Introduction

The class of Generalized Additive Models (GAMs) provides an effective semiparametric regression tool for high-dimensional data; see [6]. For a response Y and a predictor vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$, the pdf of Y_i conditional on $\mathbf{X}_i$ with respect to a fixed σ-finite measure is from an exponential family, viz.

$$f\left(Y_i \mid \mathbf{X}_i, \phi\right) = \exp\left[\left\{Y_i m\left(\mathbf{X}_i\right) - b\left\{m\left(\mathbf{X}_i\right)\right\}\right\} / a\left(\phi\right) + h\left(Y_i, \phi\right)\right].$$

The function b is a given function which relates $m(\mathbf{x})$ to the conditional variance function $\sigma^2(\mathbf{x}) = \operatorname{var}(Y \mid \mathbf{X} = \mathbf{x})$ via the equation $\sigma^2(\mathbf{x}) = a(\phi)\, b''\{m(\mathbf{x})\}$, in which $a(\phi)$ is a nuisance parameter that quantifies overdispersion. For theoretical developments, it is not necessary to assume that the data $(Y_1, \mathbf{X}_1^\top), \ldots, (Y_n, \mathbf{X}_n^\top)$ come from such an exponential family, but only that the conditional mean and variance are linked by the relation

$$\operatorname{var}(Y \mid \mathbf{X} = \mathbf{x}) = a(\phi)\, b''[(b')^{-1}\{\mathrm{E}(Y \mid \mathbf{X} = \mathbf{x})\}].$$

[*] Corresponding author.
E-mail addresses: rong.liu@utoledo.edu (R. Liu), haerdle@wiwi.hu-berlin.de (W.K. Härdle), gzhang@unm.edu (G. Zhang).

http://dx.doi.org/10.1016/j.jmva.2017.07.011
0047-259X/© 2017 Elsevier Inc. All rights reserved.

More specifically, the model is

$$\mathrm{E}\left(Y \mid \mathbf{X}\right) = b'\left\{c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha)\right\}, \tag{1}$$

where b' is the derivative of function b. Model ((1)) can be used, e.g., in scoring methods and analyzing default of companies; here $Y = 1$ denotes default and $b' = e^y/1 + e^y$ is the link function. Fitting Model (1) to such a default data set leads to estimated component functions $\hat{m}_1, \ldots, \hat{m}_d$; see, e.g., [11,25]. Plotting these functions with simultaneous confidence corridors (SCCs) as developed by [25], one can check the functional form and therefore obtain simpler parameterizations of $m_1, \ldots, m_d$.

The typical approach is to perform a preliminary (nonparametric) analysis on the influence of the component functions, and one may improve the model by introducing parametric components. This will lead to simplification, more interpretability and higher precision in statistical calibration. With these thoughts in mind, GAMs can be extended to Generalized Additive Partially Linear Models (GAPLM), in which

$$\mathrm{E}\left(Y \mid \mathbf{T}, \mathbf{X}\right) = b'\left\{m\left(\mathbf{T}, \mathbf{X}\right)\right\}, \tag{2}$$

with $m\left(\mathbf{T}, \mathbf{X}\right) = \boldsymbol{\beta}^\top \mathbf{T} + \sum_{\alpha=1}^{d_2} m_\alpha(X_\alpha)$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{d_1})^\top$, $\mathbf{T} = \left(T_0, \ldots, T_{d_1}\right)^\top$, and $\mathbf{X} = \left(X_1, \ldots, X_{d_2}\right)^\top$, where $T_0 = 1$ and $T_k \in \mathbb{R}$ for all $k \in \{1, \ldots, d_1\}$. In this paper, we assume that

$$\mathrm{var}\left(Y \mid \mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x}\right) = a\left(\phi\right) b''[\left(b'\right)^{-1}\left\{\mathrm{E}\left(Y \mid \mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x}\right)\right\}].$$

We can write (2) in the usual regression form $Y_i = b'\left\{m\left(\mathbf{T}_i, \mathbf{X}_i\right)\right\} + \sigma\left(\mathbf{T}_i, \mathbf{X}_i\right)\varepsilon_i$ with white noise ε_i that satisfies $\mathrm{E}\left(\varepsilon_i \mid \mathbf{T}_i, \mathbf{X}_i\right) = 0$, $\mathrm{E}(\varepsilon_i^2 \mid \mathbf{T}_i, \mathbf{X}_i) = 1$. For identifiability, we impose the condition

$$\forall_{\alpha \in \{1, \ldots, d_2\}} \quad \mathrm{E}\left\{m_\alpha(X_\alpha)\right\} = 0. \tag{3}$$

As in most works on nonparametric smoothing, estimation of the functions $m_1, \ldots, m_{d_2}$ is conducted on compact sets. Without loss of generality, let the compact set be $\varkappa = [0, 1]^{d_2}$.

Some estimation methods for Model (2) have been proposed, but are either computationally expensive or lacking theoretical justification. The kernel-based backfitting and marginal integration methods, e.g., in [5,9,24], are computationally expensive. More advanced non- and semi-parametric models (without link function) have also been studied, e.g., partially linear models and varying-coefficient models; see [10,14,16,20,23]. In [20], a nonconcave penalized quasi-likelihood method was proposed with polynomial spline smoothing for estimation of $m_1, \ldots, m_{d_2}$, and deriving quasi-likelihood based estimators for the linear parameter $\boldsymbol{\beta} \in \mathbb{R}^{1+d_1}$.

To our knowledge, [20] is a pilot paper since it establishes the asymptotic normality of the estimators for the parametric components in GAPLMs with independent observations. However, the asymptotic normality of the estimators of the nonparametric component functions $m_1, \ldots, m_{d_2}$ and SCCs remains to be proved. Recently, [12] studied more complicated Generalized Additive Coefficient Models by using a two-step spline method, but an iid assumption is required for the asymptotic properties of the estimation and inference of m_α, and the asymptotic normality of parameter estimates has not been shown either. Nonparametric analysis of deviance tools was developed in [4], which can be used to test the significance of the nonparametric term in generalized partially linear models with univariate nonparametric component function. Empirical likelihood based confidence regions for the parameter $\boldsymbol{\beta}$ and point-wise confidence intervals for the nonparametric term in generalized partially linear models were also provided in [8].

The spline-backfitted kernel (SBK) estimation introduced in [21] combines the advantages of both kernel and spline methods and the result is balanced in terms of theory, computation, and interpretation. The basic idea is to pre-smooth the component functions by spline estimation and then use the kernel method to improve the accuracy of the estimation on a specific m_α. In this paper, we extend the SBK method to calibrate Model (2) with additive nonparametric components and as a result, we obtain oracle efficiency and asymptotic normality of the estimators for both the parametric and nonparametric components under α- mixing condition, which complicates the derivation of theoretical properties. With the stronger iid assumption, we provide an empirical likelihood (EL) based confidence region for the parameter $\boldsymbol{\beta}$ due to the advantages of EL such as increase in coverage accuracy, easy implementation, avoiding estimating variances and Studentizing automatically; see [8]. In addition, we provide SCCs for the nonparametric component functions based on the maximal deviation distribution in [2], so that one can test the hypothesis of the shape for nonparametric terms.

The paper is organized as follows. In Section 2, we discuss the details of (2). In Section 3, the oracle estimator and its asymptotic properties are introduced. In Section 4, the SBK estimator is introduced and the asymptotics for both the parametric and nonparametric component estimations is given. In addition, SCCs for testing overall trends and entire shapes are considered. In Section 5, we apply the methods to simulated and real data examples. All technical proofs are given in Appendix.

2. Model assumptions

The space of α-centered square integrable functions on [0, 1] is defined as in [18], viz.

$$\mathcal{H}_\alpha^0 = \{g : \mathrm{E}\{g(X_\alpha)\} = 0, \mathrm{E}\{g^2(X_\alpha)\} < \infty\}.$$

Next define the model space $\mathcal{M}$, a collection of functions on $\mathbb{R}^{d_2}$ as

$$\mathcal{M} = \left\{ g\left(\mathbf{x}\right) = \sum_{\alpha=1}^{d_2} g_\alpha(\mathbf{x}) : g_\alpha \in \mathcal{H}_\alpha^0 \right\}.$$

The constraints that $\mathrm{E}\left\{g_\alpha(X_\alpha)\right\} = 0$ for all $\alpha \in \{1, \ldots, d_2\}$ ensure the unique additive representation of m_α as expressed in (3). Denote the empirical expectation by E_n, i.e., $\mathrm{E}_n(\varphi) = \sum_{i=1}^n \varphi\left(\mathbf{X}_i\right)/n$. For functions $g_1, g_2 \in \mathcal{M}$, the theoretical and empirical inner products are defined respectively as $\langle g_1, g_2\rangle = \mathrm{E}\left\{g_1\left(\mathbf{X}\right) g_2\left(\mathbf{X}\right)\right\}$, $\langle g_1, g_2\rangle_n = \mathrm{E}_n\left\{g_1\left(\mathbf{X}\right) g_2\left(\mathbf{X}\right)\right\}$. The corresponding induced norms are $\|g_1\|_2^2 = \mathrm{E}\{g_1^2\left(\mathbf{X}\right)\}$, $\|g_1\|_{2,n}^2 = \mathrm{E}_n\{g_1^2\left(\mathbf{X}\right)\}$. More generally, we set $\|g\|_r^r = \mathrm{E}|\{g\left(\mathbf{X}\right)|^r\}$.

In the paper, for any compact interval $[a, b]$, we denote the space of pth order smooth functions as $C^{(p)}[a, b] = \{g : g^{(p)} \in C\left[a, b\right]\}$, and the class of Lipschitz continuous functions for constant $C > 0$ as

$$\mathrm{Lip}\left(\left[a, b\right], C\right) = \{g : \forall_{x,x' \in [a,b]}\ |g\left(x\right) - g\left(x'\right)| \leq C|x - x'|\}.$$

For any vector $\mathbf{x} = (x_1, \ldots, x_d)^\top$, we denote the supremum and p norm as $|\mathbf{x}| = \max_{1\leq\alpha\leq d}|x_\alpha|$ and $\|\mathbf{x}\|_p = (\sum_{\alpha=1}^d x_\alpha^p)^{1/p}$, respectively. In particular, we use $\|\mathbf{x}\|$ to denote the Euclidean norm, i.e., $p = 2$. We need the following assumptions.

(A1) For every $\alpha \in \{1, \ldots, d_2\}$, one has $m_\alpha \in C^{(1)}[0, 1]$; furthermore, $m_1 \in C^{(2)}\left[0, 1\right]$ and there exists a constant $C_m > 0$ such that, for all $\alpha \in \{2, \ldots, d_2\}$, $m'_\alpha \in \mathrm{Lip}\left(\left[0, 1\right], C_m\right)$.

(A2) The inverse link function b' satisfies $b' \in C^2\left(\mathbb{R}\right), b''\left(\theta\right) > 0, \theta \in \mathbb{R}$ and $C_b > \max_{\theta\in\Theta} b''\left(\theta\right) \geq \min_{\theta\in\Theta} b''\left(\theta\right) > c_b$ for constants $C_b > c_b > 0$.

(A3) The conditional variance function $\sigma^2\left(\mathbf{x}\right)$ is measurable and bounded. The errors $\epsilon_1, \ldots, \epsilon_n$ are such that $\mathrm{E}(\varepsilon_i \mid \mathcal{F}_i) = 0$, $\mathrm{E}(|\varepsilon_i|^{2+\eta}) \leq C_\eta$ for some $\eta \in (1/2, \infty)$ with the sequence of σ-fields: $\mathcal{F}_i = \sigma\{(\mathbf{X}_j) : j \leq i, \varepsilon_j, j \leq i - 1\}$ for all $i \in \{1, \ldots, n\}$.

(A4) The density function f of $\left(X_1, \ldots, X_{d_2}\right)$ is continuous and $0 < c_f \leq \inf_{\mathbf{x}\in\chi} f\left(\mathbf{x}\right) \leq \sup_{\mathbf{x}\in\chi} f\left(\mathbf{x}\right) \leq C_f < \infty$. The marginal densities f_α of X_α have continuous derivatives on [0, 1] and are uniformly bounded from above by C_f and from below by c_f.

(A5) There exist constants $K_0, \lambda_0 \in (0, +\infty)$ such that $\alpha\left(n\right) \leq K_0 e^{-\lambda_0 n}$ holds for all $n \in \mathbb{N}$, with the α-mixing coefficients for the sequence $\mathbf{Z}_1 = (\mathbf{T}_1^\top, \mathbf{X}_1^\top, \varepsilon_1)^\top, \ldots, \mathbf{Z}_n = (\mathbf{T}_n^\top, \mathbf{X}_n^\top, \varepsilon_i)^\top$ defined, for every integer $k \geq 1$, by

$$\alpha\left(k\right) = \sup_{B\in\sigma\{\mathbf{Z}_s, s\leq t\}, C\in\sigma\{\mathbf{Z}_s, s\geq t+k\}} \left|\Pr\left(B\cap C\right) - \Pr\left(B\right)\Pr\left(C\right)\right|.$$

(A5’) The variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are mutually independent and identically distributed.

(A6) There exist constants $0 < c_\delta < C_\delta < \infty$ and $0 < c_\mathbf{Q} < C_\mathbf{Q} < \infty$ such that $c_\delta \leq \mathrm{E}(|T_k|^{2+\delta} \mid \mathbf{X} = \mathbf{x}) \leq C_\delta$ for some $\delta > 0$, and $c_\mathbf{Q} I_{d_1\times d_1} \leq \mathrm{E}\left(\mathbf{T}\mathbf{T}^\top \mid \mathbf{X} = \mathbf{x}\right) \leq C_\mathbf{Q} I_{d_1\times d_1}$.

Assumptions (A1), (A2) and (A4) are standard in the GAM literature; see [19,22]. Assumptions (A3) and (A5) are the same for weakly dependent data as in [11,21], and Assumption (A6) is the same with (C5) in [20]. When categorical predictors are present, we can create dummy variables in $\mathbf{T}_i$ and Assumption (A6) is still satisfied.

3. Oracle estimators

The aim of our analysis is to provide precise estimators for the component functions m_α and parameters $\boldsymbol{\beta}$. Without loss of generality, we may focus on m_1. If all the unknown $\boldsymbol{\beta}$ and other $m_2, \ldots, m_{d_2}$ were known, we are in a comfortable situation since the multidimensional modeling problem has reduced to one dimension. As in [17] define, for each $x_1 \in [h, 1 - h]$ and $a \in A$, a local quasi log-likelihood function

$$\tilde{\ell}_{m_1}\left(a, x_1\right) = \frac{1}{n}\sum_{i=1}^n \left[Y_i\left\{a + m\left(\mathbf{T}_i, \mathbf{X}_{i\_1}\right)\right\} - b\left\{a + m\left(\mathbf{T}_i, \mathbf{X}_{i\_1}\right)\right\}\right] K_h\left(X_{i1} - x_1\right)$$

with $m\left(\mathbf{T}_i, \mathbf{X}_{i\_1}\right) = \boldsymbol{\beta}^\top \mathbf{T}_i + \sum_{\alpha=2}^{d_2} m_\alpha\left(\mathbf{X}_{i\alpha}\right)$ and $K_h\left(u\right) = K\left(u/h\right)/h$ a kernel function K with bandwidth h satisfying the following condition.

(A7) The kernel function $K \in C^1[-1, 1]$ is a symmetric pdf and $h = h_n$ satisfies $h = \mathcal{O}\{n^{-1/5}(\ln n)^{-1/5}\}$, $h^{-1} = \mathcal{O}\{n^{1/5}(\ln n)^\delta\}$ for some constant $\delta > 1/5$.

Since all the $\boldsymbol{\beta}$ and $m_2, \ldots, m_{d_2}$ are known as obtained from the oracle, one can obtain the so-called oracle estimator

$$\tilde{m}_{K,1}(x_1) = \mathrm{argmax}_{a\in A}\tilde{\ell}_{m_1}(a, x_1). \tag{4}$$

Denote $\|K\|_2^2 = \int K^2(u)\,du$, $\mu_2(K) = \int K(u)\,u^2 du$ and introduce the scale function

$$D_1(x_1) = f_1(x_1)\mathrm{E}\left\{b''\{m(\mathbf{T},\mathbf{X})\} \mid X_1 = x_1\right\}, \tag{5}$$

and the bias function

$$\begin{aligned}\mathrm{bias}_1(x_1) = {} & \mu_2(K)\left[m_1''(x_1)f_1(x_1)\mathrm{E}\left[b''\{m(\mathbf{T},\mathbf{X})\} \mid X_1 = x_1\right]\right.\\ & + m_1'(x_1)\frac{\partial}{\partial x_1}\left\{f_1(x_1)\mathrm{E}\left[b''\{m(\mathbf{T},\mathbf{X})\} \mid X_1 = x_1\right]\right\}\\ & \left.- \left\{m_1'(x_1)\right\}^2 f_1(x_1)\mathrm{E}\left[b'''\{m(\mathbf{T},\mathbf{X})\} \mid X_1 = x_1\right]\right].\end{aligned} \tag{6}$$

Lemma 1. *Under Assumptions* (A1)–(A7), *for any $x_1 \in [h, 1-h]$, as $n \to \infty$, the oracle kernel estimator $\tilde{m}_{\mathrm{K},1}(x_1)$ given in* (4) *satisfies*

$$\sup_{x_1\in[h,1-h]}|\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1)| = \mathcal{O}_{a.s.}(\ln n/\sqrt{nh}),$$

$$\sqrt{nh}\,\{\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1) - \mathrm{bias}_1(x_1)h^2/D_1(x_1)\} \rightsquigarrow \mathcal{N}[0, D_1(x_1)^{-1}v_1^2(x_1)D_1(x_1)^{-1}],$$

with $v_1^2(x_1) = f_1(x_1)E\{\sigma^2(\mathbf{T},\mathbf{X}) \mid X_1 = x_1\}\,\|K\|_2^2$.

Lemma 1 is proved in [11]. The above oracle idea applies to the parametric part as well. Define the log-likelihood function

$$\tilde{\ell}_{\beta}(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n}[Y_i\{\mathbf{a}^\top\mathbf{T}_i + m(\mathbf{X}_i)\} - b\{\mathbf{a}^\top\mathbf{T}_i + m(\mathbf{X}_i)\}], \tag{7}$$

where $m(\mathbf{X}_i) = \sum_{\alpha=1}^{d_2} m_\alpha(X_{i\alpha})$. The infeasible estimator of $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}} = \mathrm{argmax}_{\mathbf{a}\in\mathbb{R}^{1+d_1}}\tilde{\ell}_{\beta}(\mathbf{a})$. Clearly, $\nabla\tilde{\ell}_{\beta}(\boldsymbol{\beta}) = \mathbf{0}$. To maximize (7), we have

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i\mathbf{T}_i - b'\{\mathbf{a}^\top\mathbf{T}_i + m(\mathbf{X}_i)\}\mathbf{T}_i] = \mathbf{0},$$

then the empirical likelihood ratio is

$$\tilde{R}(\mathbf{a}) = \max\left\{\prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i Z_i(\mathbf{a}) = \mathbf{0}, p_i \geq 0, \sum_{i=1}^{n} p_i = 1\right\}$$

where $Z_i(\mathbf{a}) = \left[Y_i - b'\left\{\mathbf{a}^\top\mathbf{T}_i + m(\mathbf{X}_i)\right\}\right]\mathbf{T}_i$.

Theorem 1. (i) *Under Assumptions* (A1)–(A6), *as $n \to \infty$,*

$$\left|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} - [Eb''\{m(\mathbf{T},\mathbf{X})\}\mathbf{T}\mathbf{T}^\top]^{-1}\frac{1}{n}\sum_{i=1}^{n}\sigma(\mathbf{T}_i,\mathbf{X}_i)\,\varepsilon_i\mathbf{T}_i\right| = \mathcal{O}_{a.s.}\{(\ln n)^2/n\},$$

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightsquigarrow \mathcal{N}\left[\mathbf{0}, a(\phi)\,[Eb''\{m(\mathbf{T},\mathbf{X})\}\mathbf{T}\mathbf{T}^\top]^{-1}\right].$$

(ii) *Under Assumptions* (A1)–(A4), (A5') *and* (A6), $-2\ln\{\tilde{R}(\boldsymbol{\beta})\} \rightsquigarrow \chi^2_{d_1}$.

Although the oracle estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{m}_{K,1}(x_1)$ enjoy the desirable theoretical properties in Theorem 1 and Lemma 1, they are not feasible statistics as their computation is based on the knowledge of unavailable component functions $m_2, \ldots, m_{d_2}$.

4. Spline-backfitted kernel estimators

In practice, $m_2, \ldots, m_{d_2}$ are of course unknown and need to be approximated. We obtain the spline-backfitted kernel estimators by using estimations of $m_2, \ldots, m_{d_2}$ and the unknown $\boldsymbol{\beta}$ by splines and we employ them to estimate $m_1(x_1)$ as in (4). First, we introduce the linear spline basis as in [10]. Let $0 = \xi_0 < \xi_1 < \cdots < \xi_N < \xi_{N+1} = 1$ denote a sequence of equally spaced points, called interior knots, on $[0, 1]$. Denote by $H = 1/(N+1)$ the width of each subinterval $\left[\xi_J, \xi_{J+1}\right]$ for each $j \in \{0, \ldots, N\}$ and denote the degenerate knots $\xi_{-1} = 0$, $\xi_{N+2} = 1$. We need the following assumption.

(A8) The number of interior knots $N \sim n^{1/4}\ln n$, i.e., $c_N n^{1/4}\ln n \leq N \leq C_N n^{1/4}\ln n$ for some constants $c_N, C_N > 0$.

Following [11], for each $j \in \{0, \ldots, N\}$, define the linear B-spline basis as follows:

$$b_J(x) = (1 - |x - \xi_J|/H)_+ = \begin{cases} (N+1)\,x - J + 1 \text{if } \xi_{J-1} \le x \le \xi_J, \\ J + 1 - (N+1)\,x \text{if } \xi_J \le x \le \xi_{J+1}, \\ 0 \qquad \text{otherwise.} \end{cases}$$

Let also the space of α-empirically centered linear spline functions on $[0, 1]$ be defined, for each $\alpha \in \{1, \ldots, d_2\}$, as

$$G^0_{n,\alpha} = \left\{ g_\alpha : g_\alpha(X_\alpha) = \sum_{J=0}^{N+1} \lambda_J b_J(X_\alpha),\ \mathrm{E}_n\{g_\alpha(X_\alpha)\} = 0 \right\},$$

and let the space of additive spline functions on $\boldsymbol{\chi}$ be

$$G^0_n = \left\{ g(\mathbf{x}) = \sum_{\alpha=1}^{d_2} g_\alpha(X_\alpha) : g_\alpha \in G^0_{n,\alpha} \right\}.$$

Define the log-likelihood function be given, for any $g \in G^0_n$, by

$$\hat{L}(\boldsymbol{\beta}, g) = \frac{1}{n} \sum_{i=1}^{n} [Y_i\{\boldsymbol{\beta}^\top \mathbf{T}_i + g(\mathbf{X}_i)\} - b\{\boldsymbol{\beta}^\top \mathbf{T}_i + g(\mathbf{X}_i)\}], \tag{8}$$

which according to Lemma 14 of [19], has a unique maximizer with probability approaching 1. The multivariate function $m(\mathbf{x})$ is then estimated by the additive spline function $\hat{m}(\mathbf{x})$ with

$$\hat{m}(\mathbf{t}, \mathbf{x}) = \hat{\boldsymbol{\beta}}^\top \mathbf{t} + \hat{m}(\mathbf{x}) = \mathrm{argmax}_{g \in G^0_n} \hat{L}(\boldsymbol{\beta}, g).$$

Since $\hat{m}(\mathbf{x}) \in G^0_n$, one can write $\hat{m}(\mathbf{x}) = \sum_{\alpha=1}^{d_2} \hat{m}_\alpha(x_\alpha)$ for $\hat{m}_\alpha(X_\alpha) \in G^0_{n,\alpha}$. Next define the log-likelihood function

$$\hat{\ell}_{m_1}(a, x_1) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \left\{ a + \hat{m}(\mathbf{T}_i, \mathbf{X}_{i\_1}) \right\} - b \left\{ a + \hat{m}(\mathbf{T}_i, \mathbf{X}_{i\_1}) \right\} \right] K_h(X_{i1} - x_1), \tag{9}$$

where $\hat{m}(\mathbf{T}_i, \mathbf{X}_{i\_1}) = \hat{\boldsymbol{\beta}}^\top \mathbf{T}_i + \sum_{\alpha=2}^{d_2} \hat{m}_\alpha(X_{i\alpha})$. Define the SBK estimator as

$$\hat{m}_{\mathrm{SBK},1}(x_1) = \mathrm{argmax}_{a \in A} \hat{\ell}_{m_1}(a, x_1). \tag{10}$$

Theorem 2. *Under Assumptions* (A1)–(A8), *as* $n \to \infty$, $\hat{m}_{\mathrm{SBK},1}(x_1)$ *is oracally efficient,*

$$\sup_{x_1 \in [0,1]} |\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{K,1}(x_1)| = \mathcal{O}_{a.s.}(n^{-1/2} \ln n).$$

The following corollary is a consequence of Lemma 1 and Theorem 2.

Corollary 1. *Under Assumptions* (A1)–(A8), *as* $n \to \infty$, *the SBK estimator* $\hat{m}_{\mathrm{SBK},1}(x_1)$ *given in* (10) *satisfies*

$$\sup_{x_1 \in [h, 1-h]} |\hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1)| = \mathcal{O}_{a.s.}(\ln n / \sqrt{nh})$$

and for any $x_1 \in [h, 1-h]$, *with* $\mathrm{bias}_1(x_1)$ *as in* (6) *and* $D_1(x_1)$ *in* (5)

$$\sqrt{nh}\,\{\hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) - \mathrm{bias}_1(x_1) h^2 / D_1(x_1)\} \rightsquigarrow \mathcal{N}[0, D_1(x_1)^{-1} v_1^2(x_1) D_1(x_1)^{-1}].$$

Denote $a_h = \sqrt{-2 l_{n,h}}$, $C(K) = \left\| K' \right\|_2^2 \|K\|_2^{-2}$ and for any $\alpha \in (0, 1)$, the quantile

$$Q_h(\alpha) = a_h + a_h^{-1} [\ln\{\sqrt{C(K)}/(2\pi)\} - \ln\{-\ln \sqrt{1-\alpha}\}].$$

Also with $D_1(x_1)$ and $v_1^2(x_1)$ given in (5), define $\sigma_n(x_1) = n^{-1/2} h^{-1/2} v_1(x_1) D_1^{-1}(x_1)$.

Theorem 3. *Under Assumptions* (A1)–(A4), (A5'), (A6)–(A8), *as* $n \to \infty$,

$$\lim_{n \to \infty} \Pr \left\{ \sup_{x_1 \in [h, 1-h]} \left| \hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) \right| / \sigma_n(x_1) \le Q_h(\alpha) \right\} = 1 - \alpha.$$

A $100 \times (1-\alpha)\,\%$ *simultaneous confidence band for* $m_1(x_1)$ *is* $\hat{m}_{\mathrm{SBK},1}(x_1) \pm \sigma_n(x_1) Q_h(\alpha)$.

In fact, $\hat{\boldsymbol{\beta}}$ obtained by maximizing (8) is equivalent to $\hat{\boldsymbol{\beta}}_{\mathrm{SBK}} = \mathrm{argmax}_{\mathbf{a} \in \mathbb{R}^{1+d_1}} \hat{\ell}_{\boldsymbol{\beta}}(\mathbf{a})$ with

$$\hat{\ell}_{\boldsymbol{\beta}}(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} [Y_i\{\mathbf{a}^\top \mathbf{T}_i + \hat{m}(\mathbf{X}_i)\} - b\{\mathbf{a}^\top \mathbf{T}_i + \hat{m}(\mathbf{X}_i)\}]$$

in which $\hat{m}\left(\mathbf{X}_i\right)=\sum_{\alpha=1}^{d_2}\hat{m}_\alpha\left(X_{i\alpha}\right)$. The empirical likelihood ratio is

$$\hat{R}\left(\mathbf{a}\right)=\max\left\{\prod_{i=1}^{n}np_i:\sum_{i=1}^{n}p_i\hat{Z}_i\left(\mathbf{a}\right)=\mathbf{0},p_1\geq 0,\ldots,p_n\geq 0,\sum_{i=1}^{n}p_i=1\right\}$$

where $\hat{Z}_i\left(\mathbf{a}\right)=\left[Y_i-b'\left\{\mathbf{a}^\top\mathbf{T}_i+\hat{m}\left(\mathbf{X}_i\right)\right\}\right]\mathbf{T}_i$. Similar to Theorem 2, the main result shows that the difference between $\hat{\boldsymbol{\beta}}$ and its infeasible counterpart $\tilde{\boldsymbol{\beta}}$ is asymptotically negligible.

Theorem 4. (i) *Under Assumptions* (A1)–(A6) *and* (A8), *as* $n\to\infty$, $\hat{\boldsymbol{\beta}}$ *is oracally efficient, i.e.,* $\sqrt{n}\,(\hat{\beta}_k-\tilde{\beta}_k)\xrightarrow{p}0$ *for all* $k\in\{0,\ldots,d_1\}$ *and hence*

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\rightsquigarrow\mathcal{N}[\mathbf{0},a\left(\phi\right)[Eb''\left\{m\left(\mathbf{T},\mathbf{X}\right)\right\}\mathbf{T}\mathbf{T}^\top]^{-1}].$$

(ii) *Under Assumptions* (A1)–(A4), (A5'), (A6) *and* (A8), *as* $n\to\infty$, $\sup|-2\ln\hat{R}\left(\boldsymbol{\beta}\right)+2\ln\tilde{R}\left(\boldsymbol{\beta}\right)|=\mathcal{O}_p\left(1\right)$, *and hence* $-2\ln\{\hat{R}(\boldsymbol{\beta})\}\rightsquigarrow\chi^2_{d_1}$.

As a reviewer pointed out, an obvious advantage of GAPLM over GAM is the capability of including categorical predictors. Since m_α is not a function of $\mathbf{T}$ in GAPLM, we can simply create dummy variables to represent the categorical effects and use spline estimation. [13] proposed spline estimation combined with categorical kernel functions to handle the case when function m_α depends on categorical predictors.

5. Examples

We have applied the SBK procedure to both simulated (Example 1) and real (Example 2) data and implemented our algorithms with the following rule-of-thumb number of interior knots

$$N=N_n=\min(\lfloor n^{1/4}\ln n\rfloor+1,\lfloor n/4d-1/d\rfloor-1),$$

which satisfies (A8), i.e., $N=N_n\sim n^{1/4}\ln n$, and ensures that the number of parameters in the linear least squares problem is less than $n/4$, i.e., $1+d\left(N+1\right)\leq n/4$. The bandwidth of h_α is computed as in [11] in an asymptotically optimal way.

5.1. Example 1

The data are generated from the model

$$\Pr(Y=1\mid\mathbf{T}=\mathbf{t},\mathbf{X}=\mathbf{x})=b'\left\{\boldsymbol{\beta}^\top\mathbf{T}+\sum_{\alpha=1}^{d_2}m_\alpha(X_\alpha)\right\},\quad b'\left(x\right)=\frac{e^x}{1+e^x}$$

with $d_1=2,d_2=5,\boldsymbol{\beta}=\left(\beta_0,\beta_1,\beta_2\right)^\top=\left(1,1,1,\right)^\top,m_1\left(x\right)=m_2\left(x\right)=m_3\left(x\right)=\sin\left(2\pi x\right),m_4\left(x\right)=\Phi\left(6x-3\right)-0.5$ and $m_5\left(x\right)=x^2-1/3$, where Φ is the standard normal cdf. The predictors are generated by transforming the following vector autoregression (VAR) equation for $0\leq r_1,r_2<1$ and all $i\in\{1,\ldots,n\}$, viz. $\mathbf{Z}_0=\mathbf{0}$, and

$$\mathbf{Z}_i=r_1\mathbf{Z}_{i-1}+\boldsymbol{\varepsilon}_i,\boldsymbol{\varepsilon}_i\sim\mathcal{N}\left(0,\Sigma\right),\Sigma=\left(1-r_2\right)\mathbf{I}_{d\times d}+r_2\mathbf{1}_d\mathbf{1}_d^\top,\quad d=d_1+d_2,$$
$$\mathbf{T}_i=\left(1,Z_{i1},\ldots,Z_{id_1},\right)^\top,X_{i\alpha}=\Phi\left(\sqrt{1-r_1^2}Z_{i\alpha}\right),\quad 1+d_1\leq\alpha\leq d_1+d_2,$$

with stationary $\mathbf{Z}_i=\left(Z_{i1},\ldots,Z_{id}\right)^\top\sim\mathcal{N}[0,(1-r_1^2)^{-1}\Sigma]$, $\mathbf{1}_d=\left(1,\ldots,1\right)^\top$ and $\mathbf{I}_{d\times d}$ is the $d\times d$ identity matrix. The X is transformed from Z to satisfy Assumption (A4). In this study, we selected four scenarios: (a) $r_1=0,r_2=0$; (b) $r_1=0.5$, $r_2=0$; (c) $r_1=0,r_2=0.5$; (d) $r_1=0.5,r_2=0.5$. The parameter r_1 controls the dependence between observations and r_2 controls the correlation between variables. In the selected scenarios, $r_1=0$ indicates independent observations and $r_1=0.5$ α-mixing observations, $r_2=0$ indicates independent variables and $r_2=0.5$ correlated variables within each observation. Define the empirical relative efficiency of $\hat{\beta}_1$ with respect to $\tilde{\beta}_1$ as $\mathrm{EFF}_r(\hat{\beta}_1)=\{\mathrm{MSE}(\tilde{\beta}_1)/\mathrm{MSE}(\hat{\beta}_1)\}^{1/2}$.

Table 1 shows the mean of bias, variances, MSEs and EFFs of $\hat{\beta}_1$ for $R=1000$ with sample sizes $n\in\{500,1000,2000,4000\}$. The results show that the estimator works as the asymptotic theory indicates, see Theorem 4(i).

Fig. 1 shows the kernel densities of $\hat{\beta}_1$s for $n\in\{500,1000,2000,4000\}$ from 1000 replications, again the theoretical properties are supported.

Table 2 shows the simulation results of the empirical likelihood confidence interval for β with $n\in\{500,1000,2000,4000\}$, and $r_1=0,r_2=0$ from 1000 replications. The mean and standard deviation of $-2\ln\{\hat{R}\left(\boldsymbol{\beta}\right)\}+2\ln\tilde{R}\{(\boldsymbol{\beta})\}$ (DIFF) support the oracle efficiency in Theorem 4 (ii). The performance of empirical likelihood confidence interval are compared with the wald-type one and it is clear that they have similar performance but empirical likelihood confidence interval has better coverage ratio and shorter average length.

Next for $\alpha\in\{1,\ldots,5\}$, let $X^i_{\alpha,\min}$, $X^i_{\alpha,\max}$ denote the smallest and largest observations of the variable X_α in the ith replication, respectively. The component functions $m_1,\ldots,m_5$ are estimated on equally spaced points such that $0=x_0<$

Table 1
The mean of 10 × bias, 100 × variances, 100 × MSEs and EFFs of $\hat{\beta}_1$ from 1000 replications.

r	n	$10 \times \overline{\text{BIAS}}$	$100 \times \overline{\text{VARIANCE}}$	$100 \times \overline{\text{MSE}}$	$\overline{\text{EFF}}\left(\hat{\beta}_1\right)$
$r_1 = 0$ $r_2 = 0$	500	1.509	2.018	4.298	0.8436
	1000	0.727	1.197	1.726	0.8749
	2000	0.408	0.626	0.793	0.9189
	4000	0.240	0.282	0.339	0.9534
$r_1 = 0.5$ $r_2 = 0$	500	1.473	3.136	5.306	0.8392
	1000	0.834	1.287	1.983	0.8873
	2000	0.476	0.674	0.901	0.9294
	4000	0.260	0.202	0.270	0.9665
$r_1 = 0$ $r_2 = 0.5$	500	1.327	3.880	5.642	0.8475
	1000	0.699	1.851	2.339	0.8856
	2000	0.665	0.739	1.182	0.9353
	4000	0.390	0.290	0.442	0.9479
$r_1 = 0.5$ $r_2 = 0.5$	500	1.635	4.230	6.903	0.8203
	1000	0.901	1.190	2.002	0.8758
	2000	0.529	0.806	1.086	0.9304
	4000	0.209	0.366	0.410	0.9483

Table 2
Coverage ratios and average length of the empirical likelihood confidence interval (EL) and Wald-type confidence interval for β_1 for n = 500, 1000, 2000, 4000 with $r_1 = 0$ from 1000 replications. DIFF= $-2\ln\{\hat{R}(\boldsymbol{\beta})\} + 2\ln\{\tilde{R}(\boldsymbol{\beta})\}$ is the difference between $-2\ln\{\hat{R}(\boldsymbol{\beta})\}$ and $-2\ln\{\tilde{R}(\boldsymbol{\beta})\}$.

		$n = 500$	$n = 1000$	$n = 2000$	$n = 4000$
Coverage ratio	EL	0.923	0.941	0.946	0.951
	Wald	0.918	0.934	0.944	0.948
Average length	EL	1.2675	0.9474	0.7105	0.5339
	Wald	1.4073	1.0447	0.7480	0.5625
DIFF	MEAN	0.1213	0.1023	0.0981	0.0726
	SD	0.5199	0.4703	0.3667	0.3242

$\cdots < x_{100} = 1$ and the estimator of m_α in the rth sample as $\hat{m}_{\text{SBK},\alpha,r}$. The (mean) average squared errors (ASE and MASE) are:

$$\text{ASE}(\hat{m}_{\text{SBK},\alpha,r}) = \frac{1}{101}\sum_{t=0}^{100}\left\{\hat{m}_{\text{SBK},\alpha,r}(x_t) - m_\alpha(x_t)\right\}^2,$$

$$\text{MASE}(\hat{m}_{\text{SBK},\alpha}) = \frac{1}{R}\sum_{r=1}^{R}\text{ASE}(\hat{m}_{\text{SBK},\alpha,r}).$$

In order to examine the efficiency of $\hat{m}_{\text{SBK},\alpha}$ relative to the oracle estimator $\tilde{m}_{K,\alpha}(x_\alpha)$, both are computed using the same data-driven bandwidth $\hat{h}_{\alpha,\text{opt}}$, described in Section 5 of [11]. Define the empirical relative efficiency of $\hat{m}_{\text{SBK},\alpha}$ with respect to $\tilde{m}_{K,\alpha}$ as

$$\text{EFF}_r\left(\hat{m}_{\text{SBK},\alpha}\right) = \left[\frac{\sum_{t=0}^{100}\left\{\tilde{m}_{K,\alpha}(x_t) - m_\alpha(x_t)\right\}^2}{\sum_{t=0}^{100}\left\{\hat{m}_{\text{SBK},\alpha,r}(x_t) - m_\alpha(x_t)\right\}^2}\right]^{1/2}.$$

EFF measures the relative efficiency of the SBK estimator to the oracle estimator. For increasing sample size, it should increase to 1 by Theorem 2. Table 3 shows the MASEs of $\tilde{m}_{K,1}$, $\hat{m}_{\text{SBK},1}$ and the average of EFFs from 1000 replications for $n \in \{500, 1000, 2000, 4000\}$. It is clear that the MASEs of both SBK estimator and the oracle estimator decrease when sample sizes increase, and the SBK estimator performs as well asymptotically as the oracle estimator, see Theorem 2.

To have an impression of the actual function estimates, for $r_1 = 0, r_2 = 0.5$ with sample size $n \in \{500, 1000, 2000, 4000\}$, we have plotted the SBK estimators and their 95% asymptotic SCCs (red solid lines), point-wise confidence intervals (red dashed lines), oracle estimators (blue dashed lines) for the true functions m_1 (thick black lines) in Fig. 2. Here we use $r_1 = 0$ because we want to give the 95% asymptotic SCCs, which need the observations be iid to satisfy Assumption (A5'). As expected by theoretical results, the estimation is closer to the real function and the confidence band is narrower as sample size increasing.

To compare the prediction performance of GAM and GAPLM, we introduce CAP and AR first. For any score function S, one defines its alarm rate $F(s) = \Pr(S \le s)$ and the hit rate $F_{\text{D}}(s) = \Pr(S \le s \mid D)$ where D represents the conditioning event of "default". Define the Cumulative Accuracy Profile (CAP) curve, for each $u \in (0, 1)$, as

$$\text{CAP}(u) = F_{\text{D}}\{F^{-1}(u)\}, \tag{11}$$

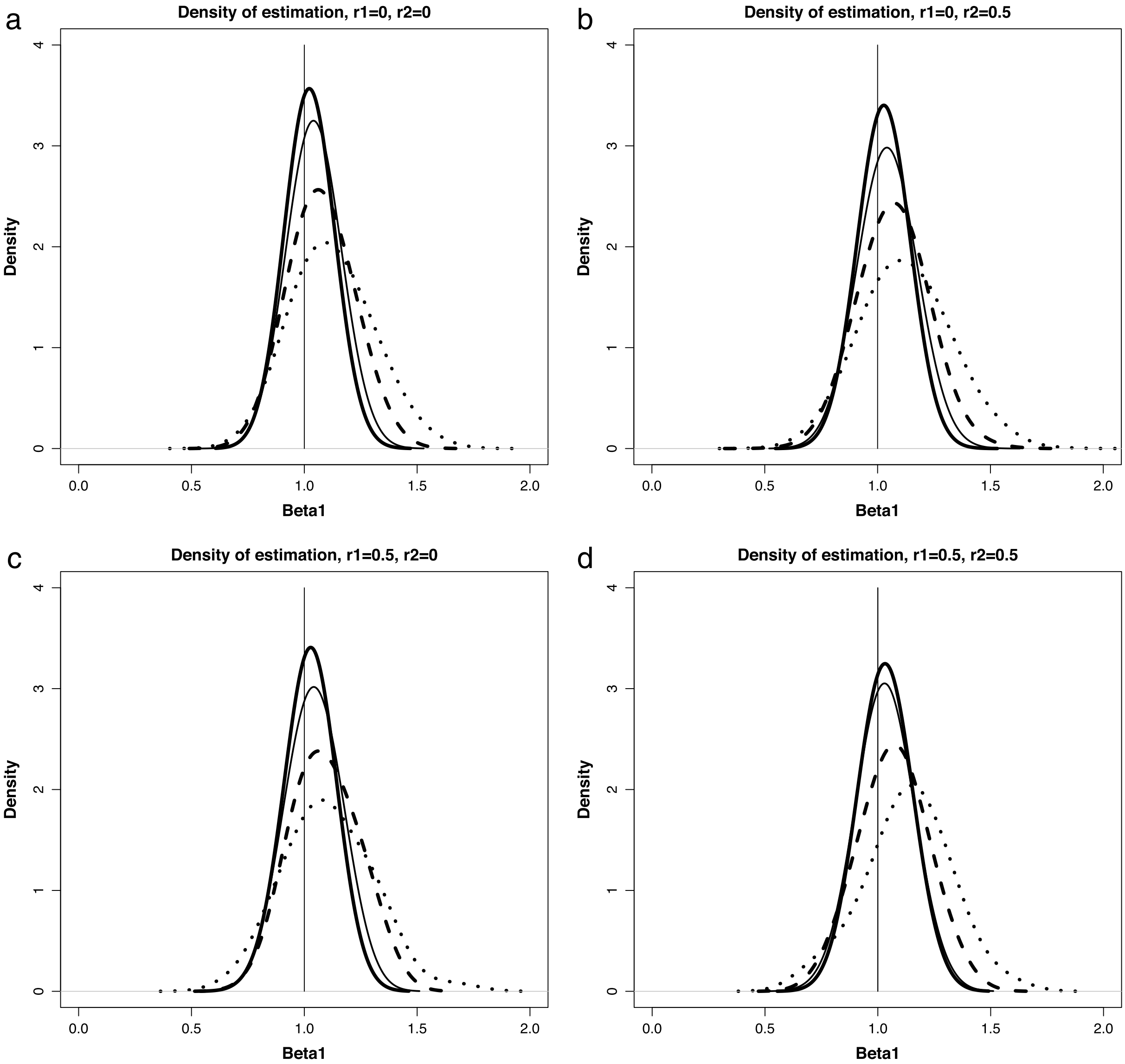

Fig. 1. Plots of densities for $\hat{\beta}_1$ with $n = 500$ (dotted line), $n = 1000$ (dashed line), $n = 2000$ (thin solid line), $n = 4000$ (thick solid line) for (a) $r_1 = 0, r_2 = 0$, (b) $r_1 = 0, r_2 = 0.5$, (c) $r_1 = 0.5, r_2 = 0$, (d) $r_1 = 0.5, r_2 = 0.5$ from 1000 replications.

which is the percentage of default-infected obligators that are found among the first (according to their scores) $100 \times u\%$ of all obligators. A perfect rating method assigns all lowest scores to exactly the defaulters, so its CAP curve linearly increases up and then stays at 1; in other words, $\mathrm{CAP_P}(u) = \min(u/p, 1)$ for all $u \in (0, 1)$, where p denotes the unconditional default probability. In contrast, a noninformative rating method with zero discriminatory power displays a diagonal line $\mathrm{CAP_N}(u) = u$ for all $u \in (0, 1)$. The CAP curve of a given scoring method S always locates between these two extremes and gives information about its performance.

The area between the CAP curve and the noninformative diagonal $\mathrm{CAP_N}(u) \equiv u$ is a_R, whereas a_P is the area between the perfect CAP curve $\mathrm{CAP_P}(u)$ and the noninformative diagonal $\mathrm{CAP_N}(u)$. Thus the CAP can be measured for example by Accuracy Ratio (AR): the ratio of a_R and a_P, viz.

$$\mathrm{AR} = \frac{a_R}{a_P} = \frac{2}{1-p}\left\{\int_0^1 \mathrm{CAP}(u)\,du - 1\right\},$$

where $\mathrm{CAP}(u)$ is given in (11). The AR takes value in [0, 1], with value 0 corresponding to the noninformative scoring, and 1 the perfect scoring method. A higher AR indicates an overall higher discriminatory power of a method. Table 4 shows the average and standard deviations of the ARs from 1000 replications using k-fold cross-validation with $k \in \{2, 10, 100\}$ for

Table 3
The 100×MASEs of $\tilde{m}_{\mathrm{K},1}$, $\hat{m}_{\mathrm{SBK},1}$ and $\overline{\mathrm{EFF}}$s for $n \in \{500, 1000, 2000, 4000\}$ from 1000 replications.

r	n	$100 \times \mathrm{MASE}\left(\tilde{m}_{\mathrm{K},\alpha}\right)$	$100 \times \mathrm{MASE}\left(\hat{m}_{\mathrm{SBK},\alpha}\right)$	$\overline{\mathrm{EFF}}\left(\hat{m}_{\mathrm{SBK},1}\right)$
$r_1 = 0$ $r_2 = 0$	500	4.482	4.603	0.9501
	1000	2.418	2.503	0.9809
	2000	1.582	1.613	0.9854
	4000	1.212	1.247	0.9923
$r_1 = 0.5$ $r_2 = 0$	500	4.060	4.322	0.9445
	1000	2.592	2.649	0.9767
	2000	1.746	1.714	0.9832
	4000	1.194	1.218	0.9936
$r_1 = 0$ $r_2 = 0.5$	500	4.845	6.348	0.8827
	1000	2.935	3.559	0.8755
	2000	1.951	2.177	0.9494
	4000	1.515	1.648	0.9795
$r_1 = 0.5$ $r_2 = 0.5$	500	5.656	7.114	0.8722
	1000	2.804	3.570	0.8951
	2000	1.886	2.089	0.9478
	4000	1.525	1.634	0.9744

Table 4
The mean and standard deviation (in parentheses) of Accuracy Ratio (AR) values for GLM, GAM, GAPLM for $r_1 = 0$, $r_2 = 0$ from 1000 replications.

n		$k = 2$	$k = 10$	$k = 100$
500	GLM	0.6287 (0.0436)	0.6412 (0.0397)	0.6438 (0.0390)
	GAM	0.6222 (0.0732)	0.6706 (0.0393)	0.6756 (0.0400)
	GAPLM	0.6511 (0.0479)	0.6828 (0.0377)	0.6861 (0.0391)
1000	GLM	0.6429 (0.0282)	0.6476 (0.0268)	0.6488 (0.0268)
	GAM	0.6735 (0.0438)	0.6863 (0.0326)	0.6929 (0.0261)
	GAPLM	0.6861 (0.0298)	0.6968 (0.0254)	0.7001 (0.0258)
2000	GLM	0.6474 (0.0204)	0.6513 (0.0195)	0.6519 (0.0188)
	GAM	0.6842 (0.0615)	0.6984 (0.0286)	0.7000 (0.0185)
	GAPLM	0.6984 (0.0204)	0.7067 (0.0178)	0.7057 (0.0178)
4000	GLM	0.6507 (0.0134)	0.6522 (0.0136)	0.6529 (0.0132)
	GAM	0.6889 (0.0243)	0.6968 (0.0403)	0.7079 (0.0164)
	GAPLM	0.7056 (0.0130)	0.7110 (0.0124)	0.7119 (0.0119)

$r_1 = 0, r_2 = 0$ and $n \in \{500, 1000, 2000, 4000\}$. In each replication, we randomly divide the set of observations into k equal size folds and use the remaining $k - 1$ folds as training data set to make prediction for each fold. After we obtain all the predictions for each observation in the data set, we compute the CAP and AR based on above formula. It is clear that GAPLM has best predication accuracy.

Finally, to show the estimation performance when $\mathbf{T}$ has categorical variables, we generate data using the same model above but add one more categorical variable, i.e., $d_1 = 3$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top = (1, 1, 1, 1)^\top$, $T_3 = \{0, 1\}$ with probability 0.5 for $T_3 = 1$ and independent with the other variables T and X. Table 5 shows the bias, variances, MSEs and EFFs of $\hat{\beta}_3$ for $R = 1000$ with sample sizes $n \in \{500, 1000, 2000, 4000\}$. The results show that the estimator works as the asymptotic theory indicates.

5.2. Example 2

The credit reform database, provided by the Research Data Center (RDC) of the Humboldt Universität zu Berlin, was studied using a GAM in [11]. The data set contains $d = 8$ financial ratios, which are shown in Table 6, of 19,610 German companies (18,610 solvent and 1000 insolvent). The time period ranges from 1997 to 2002 and in the case of the insolvent companies the information was gathered two years before the insolvency took place. The last annual report of a company before it went bankrupt receives the indicator $Y = 1$ and for the rest (solvent) $Y = 0$. In the original data set, the variables are labeled as Z_α. In order to satisfy the Assumption (A4) in [11], we need the transformation $X_{i\alpha} = F_{n\alpha}\left(Z_{i\alpha}\right)$ for all $\alpha \in \{1, \ldots, 8\}$, where $F_{n\alpha}$ is the empirical cdf of the data $X_{1\alpha}, \ldots, X_{n\alpha}$. See [3,11] for more details of this data set.

Using a GAM and the SBK method, we clearly see via the SCCs that the shape of $m_2\left(x_2\right)$ is linear. Fig. 3(a) shows that a linear line is covered by the SCCs of $\hat{m}_2$. We additionally show the SCCs for another component function of ln(Total_Assets) in Fig. 3(b). The SCCs do not cover a linear line. In fact, among the eight financial ratios considered, only x_2 yields a linear influence. To improve the precision in statistical calibration and interpretability, we can use GAPLM with parametric $m_2\left(x_2\right) = \beta_2 x_2$.

For the RDC data, the in-sample AR value obtained from GAPLM is 62.89%, which is very close to the AR value 63.05% obtained from GAM in [11] and higher than the AR value 60.51% obtained from SVM in [3]. To compare the prediction performance, we use the AR introduced in Example 1. Then we randomly divide the data set into $k \in \{2, 10\}$ folds and

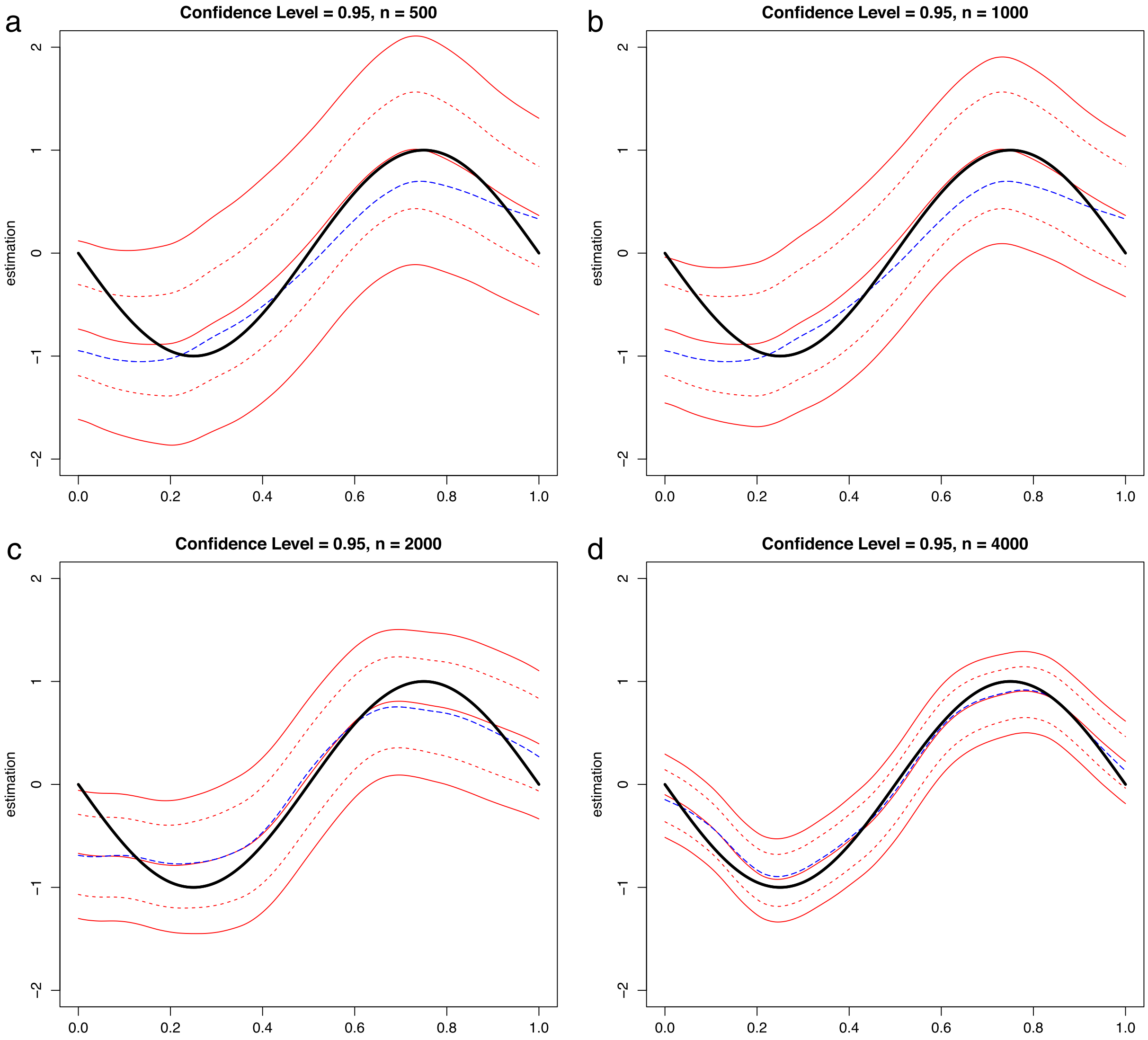

Fig. 2. Plots of m_1 (thick black line), $\tilde{m}_{K,1}$ (blue dashed line), asymptotic 95% point-wise confidence intervals (red dashed line), $\hat{m}_{SBK,1}$ and 95% simultaneous confidence bands (red solid line) for $r_1 = 0$, $r_2 = 0.5$ and (a) $n = 500$, (b) $n = 1000$, (c) $n = 2000$, (d) $n = 4000$.

obtain the prediction for each observation using the remaining $k - 1$ folds as training set. Based on the prediction of all the observations, we can compute prediction AR value. Table 7 shows the mean and standard deviation of the prediction AR values from 100 replications. GAPLM has higher prediction AR value than GAM for 99 replications when $k = 2$ and 100 times when $k = 10$. It is clear that GAPLM has best prediction accuracy due to the better statistical calibration.

Acknowledgments

Financial support from the Deutsche Forschungsgemeinschaft (DFG) via SFB 649 “Economic Risk”, and International Research Training Group (IRTG) 1792 was gratefully acknowledged. The authors thank the Editor-in-Chief, the Associate Editor and two referees for their comments and suggestions which have led to substantial improvement of this work.

Appendix A

A.1. Preliminaries

In the proofs that follow, we use “$\mathcal{U}$” and “$\mathcal{U}$” to denote sequences of random variables that are uniformly “$\mathcal{O}$” and “$\mathcal{O}$ ” of certain order. Denote the theoretical inner product of b_J and 1 with respect to the αth marginal density $f_\alpha(X_\alpha)$ as

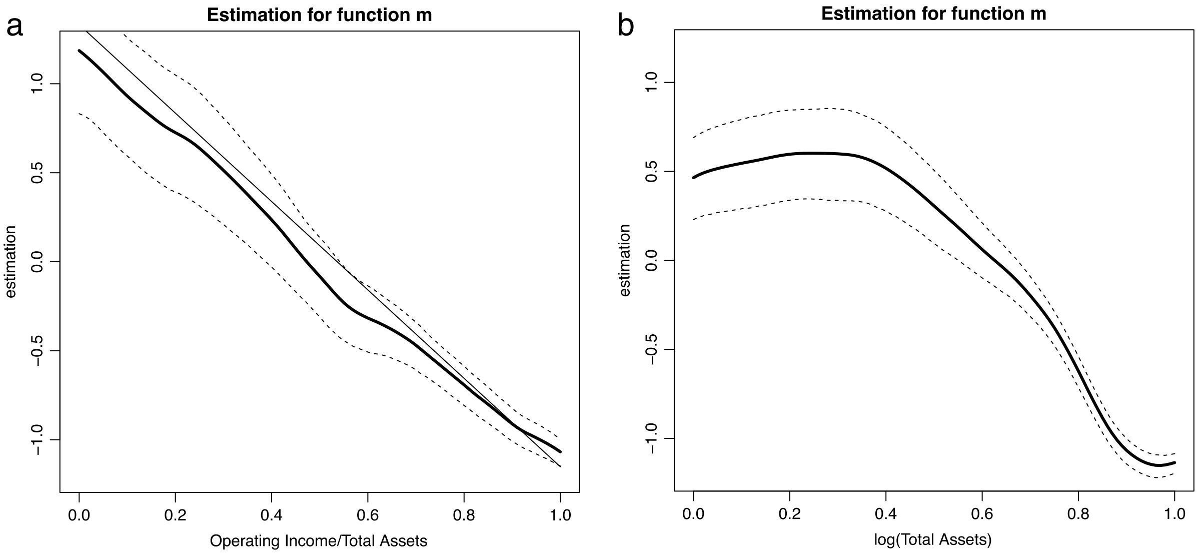

Fig. 3. Plots of estimations of component functions (a) $\hat{m}_{\mathrm{SBK},2}(x_2)$ and (b) $\hat{m}_{\mathrm{SBK},8}(x_8)$ and asymptotic 95% simultaneous confidence bands.

Table 5
The mean of 10 × bias, 100 × variances, 100 × MSEs and EFFs of $\hat{\beta}_3$ from 1000 replications.

r	n	$10\times\overline{\mathrm{BIAS}}$	$100\times\overline{\mathrm{VARIANCE}}$	$100\times\overline{\mathrm{MSE}}$	$\overline{\mathrm{EFF}}(\hat{\beta}_3)$
$r_1 = 0$ $r_2 = 0$	500	1.476	10.129	12.309	0.7634
	1000	0.770	4.437	5.031	0.8343
	2000	0.448	1.846	2.047	0.8929
	4000	0.315	0.937	1.037	0.9572
$r_1 = 0.5$ $r_2 = 0$	500	1.336	10.329	12.115	0.7445
	1000	0.833	4.221	4.916	0.8267
	2000	0.423	1.952	2.132	0.8832
	4000	0.302	0.944	1.036	0.9436
$r_1 = 0$ $r_2 = 0.5$	500	1.441	10.154	12.231	0.7556
	1000	0.803	4.446	5.114	0.8430
	2000	0.489	2.136	2.376	0.8785
	4000	0.328	0.924	1.032	0.9572
$r_1 = 0.5$ $r_2 = 0.5$	500	1.475	11.014	13.190	0.7794
	1000	0.812	4.464	5.124	0.8314
	2000	0.524	1.970	2.245	0.8852
	4000	0.302	0.966	1.058	0.9529

Table 6
Definitions of financial ratios.

Ratio No.	Definition	Ratio No.	Definition
Z_1	Net_Income/Sales	Z_5	Cash/Total_Assets
Z_2	Operating_Income/Total_Assets	Z_6	Inventories/Sales
Z_3	Ebit/Total_Assets	Z_7	Accounts_Payable/Sales
Z_4	Total_Liabilities/Total_Assets	Z_8	ln(Total_Assets)

Table 7
The mean and standard deviation (in parentheses) of AR values for GLM, GAM, GAPLM for k-fold cross-validation with $k \in \{2, 10\}$ from 1000 replications.

	$k = 2$	$k = 10$
GLM	0.5627 (0.0271)	0.5751 (0.00162)
GAM	0.5888 (0.0405)	0.6123 (0.00219)
GAPLM	0.5928 (0.0408)	0.6164 (0.00196)

$c_{J,\alpha} = \langle b_J(X_\alpha), 1\rangle = \int b_J(X_\alpha) f_\alpha(X_\alpha) dx_\alpha$ and define the centered B-spline basis $b_{J,\alpha}\left(x_\alpha\right)$ and the standardized B-spline basis, for each $J \in \{1, \ldots, N+1\}$, as

$$b_{J,\alpha}(X_\alpha) = b_J(X_\alpha) - \frac{c_{J,\alpha}}{c_{J-1,\alpha}} b_{J-1}(X_\alpha), \quad B_{J,\alpha}(X_\alpha) = \frac{b_{J,\alpha}\left(x_\alpha\right)}{\|b_{J,\alpha}\|_2},$$

so that $\mathrm{E}\{B_{J,\alpha}(X_\alpha)\} = 0$ and $\mathrm{E}\{B^2_{J,\alpha}(X_\alpha)\} = 1$. Theorem A.2 in [21] shows that under Assumptions . (A1)–(A5) and (A7), constants $c_0\left(f\right), C_0(f), c_1\left(f\right)$ and $C_1(f)$ exist depending on the marginal densities $f_1, \ldots, f_d$ such that $c_0\left(f\right) H \le c_{J,\alpha} \le C_0\left(f\right) H$ and

$$c_1\left(f\right) H \le \|b_{J,\alpha}\|_2^2 \le C_1(f) H. \tag{A.1}$$

Lemma A.1 ([1], *p. 149*). *For any* $m \in C^1[0,1]$ *with* $m' \in \mathrm{Lip}\left([0,1], C_\infty\right)$, *there exist a constant* $C_\infty > 0$ *and a function* $g \in G_n^{(0)}[0,1]$ *such that* $\|g - m\|_\infty \le C_\infty H^2$.

A.2. Oracle estimators

Proof of Theorem 1. (i) According to the Mean Value Theorem, a vector $\bar{\boldsymbol\beta}$ between $\boldsymbol\beta$ and $\tilde{\boldsymbol\beta}$ exists such that $(\tilde{\boldsymbol\beta} - \boldsymbol\beta)\nabla^2\tilde\ell_{\boldsymbol\beta}(\bar{\boldsymbol\beta}) = \nabla\tilde\ell_{\boldsymbol\beta}(\tilde{\boldsymbol\beta}) - \nabla\tilde\ell_{\boldsymbol\beta}\left(\boldsymbol\beta\right) = -\nabla\tilde\ell_{\boldsymbol\beta}\left(\boldsymbol\beta\right)$ since $\nabla\tilde\ell_{\boldsymbol\beta}(\tilde{\boldsymbol\beta}) = \mathbf{0}$, where

$$-\nabla^2\tilde\ell_{\boldsymbol\beta}(\bar{\boldsymbol\beta}) = n^{-1}\sum_{i=1}^n b''\{\bar{\boldsymbol\beta}^\top T_i + m\left(\mathbf{X}_i\right)\}\mathbf{T}_i\mathbf{T}_i^\top > c_b c_{\mathbf{Q}} \mathbf{I}_{d_1\times d_1}$$

with $c_b > 0$ according to (A2), and then the infeasible estimator is $\tilde{\boldsymbol\beta} = \mathrm{argmax}_{\mathbf{a}\in\mathbb{R}^{1+d_1}}\tilde\ell_{\boldsymbol\beta}\left(\mathbf{a}\right)$.

$$\nabla\tilde\ell_{\boldsymbol\beta}\left(\boldsymbol\beta\right) = \frac{1}{n}\sum_{i=1}^n [Y_i\mathbf{T}_i - b'\{\boldsymbol\beta^\top\mathbf{T}_i + m\left(\mathbf{X}_i\right)\}\mathbf{T}_i] = \frac{1}{n}\sum_{i=1}^n \sigma\left(\mathbf{T}_i, \mathbf{X}_i\right)\varepsilon_i\mathbf{T}_i.$$

We have $|n^{-1}\sum_{i=1}^n \sigma\left(\mathbf{T}_i, \mathbf{X}_i\right)\varepsilon_i\mathbf{T}_i| = \mathcal{O}_{a.s.}(n^{-1/2}\ln n)$ by Bernstein's Inequality as Lemma A.2 in [11], so $|\tilde{\boldsymbol\beta} - \boldsymbol\beta| = \mathcal{O}_{a.s.}(n^{-1/2}\ln n)$ according to $\tilde{\boldsymbol\beta} - \boldsymbol\beta = -\{\nabla^2\tilde\ell_{\boldsymbol\beta}(\bar{\boldsymbol\beta})\}^{-1}\nabla\tilde\ell_{\boldsymbol\beta}\left(\boldsymbol\beta\right)$. Then

$$\nabla^2\tilde\ell_{\boldsymbol\beta}(\bar{\boldsymbol\beta}) \overset{a.s.}{\to} \nabla^2\tilde\ell_{\boldsymbol\beta}\left(\boldsymbol\beta\right) = -\frac{1}{n}\sum_{i=1}^n b''\{\boldsymbol\beta^\top\mathbf{T}_i + m\left(\mathbf{X}_i\right)\}\mathbf{T}_i\mathbf{T}_i^\top,$$

which converges to $-\mathrm{E}[b''\left\{m\left(\mathbf{T}, \mathbf{X}\right)\right\}\mathbf{T}\mathbf{T}^\top]$ almost surely at the rate of $n^{-1/2}\ln n$. So

$$\left|\tilde{\boldsymbol\beta} - \boldsymbol\beta - [\mathrm{E}b''\left\{m\left(\mathbf{T}, \mathbf{X}\right)\right\}\mathbf{T}\mathbf{T}^\top]^{-1}\frac{1}{n}\sum_{i=1}^n \sigma\left(\mathbf{T}_i, \mathbf{X}_i\right)\varepsilon_i\mathbf{T}_i\right| = \mathcal{O}_{a.s.}\{n^{-1}(\ln n)^2\}.$$

Since $n^{-1}\sum_{i=1}^n \sigma\left(\mathbf{T}_i, \mathbf{X}_i\right)\varepsilon_i\mathbf{T}_i \rightsquigarrow \mathcal{N}[\mathbf{0}, a\left(\phi\right)[\mathrm{E}b''\{m(\mathbf{T}, \mathbf{X})\}\mathbf{T}\mathbf{T}^\top]^{-1}]$ by the Central Limit Theorem, an application of Slutsky's Lemma completes the proof of Theorem 1(i).

(ii) The proof is trivial based on the properties of empirical likelihood ratio for GLMs; see Theorem 3.2 in [15] and Corollary 1 in [7]. □

A.3. Spline-backfitted kernel estimators

In this section, we present the proofs of Theorems 2–4. We write any $g \in G_n^0$ as $g = \boldsymbol\lambda^\top\mathbf{B}\left(\mathbf{X}_i\right)$ with vector $\boldsymbol\lambda_g = \left(\lambda_{J,\alpha}\right)^\top_{1\le J\le N+1, 1\le\alpha\le d_2} \in \mathbb{R}^{(N+1)d_2}$ the dimension of the additive spline space G_n^0, and

$$\mathbf{B}\left(\mathbf{x}\right) = \left(B_{1,1}(x_1), \ldots, B_{N+1,1}(x_1), \ldots, B_{1,d_2}\left(x_{d_2}\right), \ldots, B_{N+1,d_2}\left(x_{d_2}\right)\right)^\top.$$

Denote $\mathbf{B}\left(\mathbf{t}, \mathbf{x}\right) = \left(1, t_1, \ldots, t_{d_1}, B_{1,1}(x_1), \ldots, B_{N+1,1}\left(x_1\right), \ldots, B_{1,d_2}\left(x_{d_2}\right), \ldots, B_{N+1,d_2}\left(x_{d_2}\right)\right)^\top$,

$$\boldsymbol\lambda = (\boldsymbol\lambda_{\boldsymbol\beta}^\top, \boldsymbol\lambda_{\mathbf{g}}^\top)^\top = \left(\lambda_0, \lambda_k, \lambda_{J,\alpha}\right)^\top_{1\le J\le N+1, 1\le\alpha\le d_2, 1\le k\le d_1} \in \mathbb{R}^{N_d}$$

with $N_d = 1 + d_1 + \left(N+1\right)d_2$ and

$$\hat{L}(\boldsymbol\lambda_{\boldsymbol\beta}, g) = \hat{L}\left(\boldsymbol\lambda\right) = \frac{1}{n}\sum_{i=1}^n [Y_i\{\boldsymbol\lambda^\top\mathbf{B}(\mathbf{T}_i, \mathbf{X}_i)\} - b\{\boldsymbol\lambda^\top\mathbf{B}(\mathbf{T}_i, \mathbf{X}_i)\}],$$

which yields the gradient and Hessian formulas

$$\nabla \hat{L}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}[Y_i\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i) - b'\{\boldsymbol{\lambda}^\top\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)\}\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)],$$

$$\nabla^2 \hat{L}(\boldsymbol{\lambda}) = -\frac{1}{n}\sum_{i=1}^{n} b''\{\boldsymbol{\lambda}^\top\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)\}\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)\,\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)^\top.$$

The multivariate function $m(\mathbf{t},\mathbf{x})$ is estimated by

$$\hat{m}(\mathbf{t},\mathbf{x}) = \hat{\beta}_0 + \sum_{k=1}^{d_1}\hat{\beta}_k t_k + \sum_{\alpha=1}^{d_2}\hat{m}_\alpha(X_\alpha) = \hat{\boldsymbol{\lambda}}^\top\mathbf{B}(\mathbf{t},\mathbf{x}),$$

$$\hat{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\lambda}}_{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\lambda}}_{\mathbf{g}}^\top)^\top = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\lambda}}_{\mathbf{g}}^\top)^\top = (\hat{\beta}_k, \hat{\lambda}_{J,\alpha})^\top_{0\le k\le d_1, 1\le\alpha\le d_2, 1\le J\le N+1} = \mathrm{argmax}_{\boldsymbol{\lambda}}\hat{L}(\boldsymbol{\lambda}).$$

Lemma 14 of Stone [19] ensures that with probability approaching 1, $\hat{\boldsymbol{\lambda}}$ exists uniquely and that $\nabla\hat{L}(\hat{\boldsymbol{\lambda}}) = \mathbf{0}$. In addition, Lemma A.1 and (A1) provide a vector $\bar{\boldsymbol{\lambda}} = (\boldsymbol{\beta}^\top, \bar{\boldsymbol{\lambda}}_{\mathbf{g}}^\top)^\top$ and an additive spline function $\bar{m}$ such that

$$\bar{m}(\mathbf{x}) = \bar{\boldsymbol{\lambda}}_{\mathbf{g}}^\top\mathbf{B}(\mathbf{x}), \quad \|\bar{m} - m\|_\infty \le C_\infty H^2. \tag{A.2}$$

We first establish technical lemmas before proving Theorems 2 and 4.

Lemma A.2. *Under Assumptions* (A1)–(A6) *and* (A8), *as* $n\to\infty$,

$$|\nabla\hat{L}(\bar{\boldsymbol{\lambda}})| = \mathcal{O}_{a.s.}(H^2 + n^{-1/2}\ln n), \quad \|\nabla\hat{L}(\bar{\boldsymbol{\lambda}})\| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n).$$

Proof. See Online Supplement. □

Define the following matrices:

$$\mathbf{V} = \mathrm{E}\mathbf{B}(\mathbf{T},\mathbf{X})\,\mathbf{B}(\mathbf{T},\mathbf{X})^\top, \quad \mathbf{S} = \mathbf{V}^{-1}, \quad \mathbf{V}_n = n^{-1}\sum_{i=1}^{n}\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)\,\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i)^\top, \quad \mathbf{S}_n = \mathbf{V}_n^{-1},$$

$$\mathbf{V}_b = \mathrm{E}b''\{m(\mathbf{T},\mathbf{X})\}\,\mathbf{B}(\mathbf{T},\mathbf{X})\,\mathbf{B}(\mathbf{T},\mathbf{X})^\top = \begin{bmatrix} v_{b,00} & v_{b,0,k} & v_{b,0,J,\alpha} \\ v_{b,0,k'} & v_{b,k,k'} & v_{b,J,\alpha,k'} \\ v_{b,0,J',\alpha'} & v_{b,J',\alpha',k} & v_{b,J,\alpha,J',\alpha'} \end{bmatrix}_{N_d\times N_d}$$

where $N_d = (N+1)\,d_2 + 1 + d_1$, and

$$\mathbf{S}_b = \mathbf{V}_b^{-1} = \begin{bmatrix} s_{b,00} & s_{b,0,k} & s_{b,0,J,\alpha} \\ s_{b,0,k'} & s_{b,k,k'} & s_{b,J,\alpha,k'} \\ s_{b,0,J',\alpha'} & s_{b,J',\alpha',k} & b,J,\alpha,J',\alpha' \end{bmatrix}_{N_d\times N_d}. \tag{A.3}$$

For any vector $\boldsymbol{\lambda}\in\mathbb{R}^{N_d}$, denote

$$\mathbf{V}_b(\boldsymbol{\lambda}) = \mathrm{E}b''\{\boldsymbol{\lambda}^\top\mathbf{B}(\mathbf{T},\mathbf{X})\}\mathbf{B}(\mathbf{T},\mathbf{X})\,\mathbf{B}(\mathbf{T},\mathbf{X})^\top, \quad \mathbf{S}_b(\boldsymbol{\lambda}) = \mathbf{V}_b^{-1}(\boldsymbol{\lambda})$$

$$\mathbf{V}_{n,b}(\boldsymbol{\lambda}) = -\nabla^2\hat{L}(\boldsymbol{\lambda}), \quad \mathbf{S}_{n,b}(\boldsymbol{\lambda}) = \mathbf{V}_{n,b}^{-1}(\boldsymbol{\lambda}). \tag{A.4}$$

Lemma A.3. *Under Assumptions* (A2) *and* (A4), *one has*

$$c_{\mathbf{V}}\mathbf{I}_{N_d} \le \mathbf{V} \le C_{\mathbf{V}}\mathbf{I}_{N_d}, \quad c_{\mathbf{S}}\mathbf{I}_{N_d} \le \mathbf{S} \le C_{\mathbf{S}}\mathbf{I}_{N_d}, \quad c_{\mathbf{V},b}\mathbf{I}_{N_d} \le \mathbf{V}_b \le C_{\mathbf{V},b}\mathbf{I}_{N_d}, \quad c_{\mathbf{S},b}\mathbf{I}_{N_d} \le \mathbf{S}_b \le C_{\mathbf{S},b}\mathbf{I}_{N_d}.$$

Under Assumptions (A2), (A4), (A5) *and* (A8), *as* $n\to\infty$ *with probability increasing to* 1

$$c_{\mathbf{V}}\mathbf{I}_{N_d} \le \mathbf{V}_n(\boldsymbol{\lambda}) \le C_{\mathbf{V}}\mathbf{I}_{N_d}, \quad c_{\mathbf{S}}\mathbf{I}_{N_d} \le \mathbf{S}_n(\boldsymbol{\lambda}) \le C_{\mathbf{S}}\mathbf{I}_{N_d} \quad c_{\mathbf{V},b}\mathbf{I}_{N_d} \le \mathbf{V}_{n,b}(\boldsymbol{\lambda}) \le C_{\mathbf{V},b}\mathbf{I}_{N_d}, \quad c_{\mathbf{S},b}\mathbf{I}_{N_d} \le \mathbf{S}_{n,b}(\boldsymbol{\lambda}) \le C_{\mathbf{S},b}\mathbf{I}_{N_d}.$$

Proof. Using Lemma A.7 in [14] and the boundedness of the function b'. □

Define three vectors $\boldsymbol{\Phi}_b$, $\boldsymbol{\Phi}_v$, $\boldsymbol{\Phi}_r$ as

$$\boldsymbol{\Phi}_b = \left(\Phi_{b,J,\alpha}\right)^\top_{0\le k\le d_1, 1\le\alpha\le d_2, 1\le J\le N+1} = -\mathbf{S}_b n^{-1}\sum_{i=1}^{n}\left[b'\{m(\mathbf{T}_i,\mathbf{X}_i)\} - b'\{\bar{m}(\mathbf{T}_i,\mathbf{X}_i)\}\right]\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i),$$

$$\boldsymbol{\Phi}_v = \left(\Phi_{v,J,\alpha}\right)^\top_{0\le k\le d_1, 1\le\alpha\le d_2, 1\le J\le N+1} = -\mathbf{S}_b n^{-1}\sum_{i=1}^{n}[\sigma(\mathbf{T}_i,\mathbf{X}_i)\,\varepsilon_i]\,\mathbf{B}(\mathbf{T}_i,\mathbf{X}_i),$$

and

$$\boldsymbol{\Phi}_r = \left(\Phi_{r,J,\alpha}\right)^{\top}_{0\le k\le d_1,1\le\alpha\le d_2,1\le J\le N+1} = \hat{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}} - \boldsymbol{\Phi}_b - \boldsymbol{\Phi}_v.$$

Lemma A.4. *Under Assumptions* (A1)–(A6) *and* (A8), *as* $n \to \infty$,

$$\|\hat{\boldsymbol{\lambda}} - \bar{\boldsymbol{\lambda}}\| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n), \tag{A.5}$$

$$\|\boldsymbol{\Phi}_r\| = \mathcal{O}_p(H^{-3/2}n^{-1}\ln n), \tag{A.6}$$

$$\|\boldsymbol{\Phi}_b\| = \mathcal{O}_{a.s.}(H^2), \quad \|\boldsymbol{\Phi}_v\| = \mathcal{O}_{a.s.}(H^{-1/2}n^{-1/2}\ln n).$$

Proof. See Online Supplement. □

Lemma A.5. *Under Assumptions* (A1)–(A6) *and* (A8), *as* $n \to \infty$,

$$\left\|\hat{m} - \bar{m}\right\|_{2,n} + \left\|\hat{m} - \bar{m}\right\|_2 = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n), \left\|\hat{m} - m\right\|_{2,n} + \left\|\hat{m} - m\right\|_2 = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n).$$

Proof. Lemma A.3 implies

$$\left\|\hat{m} - \bar{m}\right\|_{2,n} + \left\|\hat{m} - \bar{m}\right\|_2 \le 2C_{\mathbf{V}}\|\hat{\boldsymbol{\lambda}}_{\mathbf{g}} - \bar{\boldsymbol{\lambda}}_g\| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\ln n).$$

The claim follows from the fact that $\|\bar{m} - m\|_\infty + \|\bar{m} - m\|_2 + \|\bar{m} - m\|_{2,n} = \mathcal{O}(H^2)$ by (A.2). □

Proof of Theorem 2. According to (9) and the Mean Value Theorem, a $\bar{m}_{\mathrm{K},1}(x_1)$ between $\hat{m}_{\mathrm{SBK},1}(x_1)$ and $\tilde{m}_{\mathrm{K},1}(x_1)$ exists such that

$$\hat{\ell}'_{m_1}\{\hat{m}_{\mathrm{SBK},1}(x_1), x_1\} - \hat{\ell}'\{\tilde{m}_{\mathrm{K},1}(x_1), x_1\} = \hat{\ell}''_{m_1}\{\bar{m}_{\mathrm{K},1}(x_1), x_1\}\{\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1)\}.$$

Then according to $\hat{\ell}'_{m_1}\{\hat{m}_{\mathrm{SBK},1}(x_1), x_1\} = 0$, one has

$$\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1) = -\frac{\hat{\ell}'_{m_1}\{\tilde{m}_{\mathrm{K},1}(x_1), x_1\}}{\hat{\ell}''_{m_1}\{\bar{m}_{\mathrm{K},1}(x_1), x_1\}}.$$

The theorem then follows Lemmas A.15 and A.16 in [11] with small modification including variable **T**. □

Proof of Theorem 3. It follows Theorem 2 and the same proof of Theorem 1 in [25]. □

Proof of Theorem 4. See Online Supplement. □

Appendix B. Supplementary data

gaplmsbk.R: R package containing code to perform SBK estimation for component functions in generalized additive partially linear model available at https://github.com.

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2017.07.011.

References

[1] C. de Boor, A Practical Guide To Splines, Springer, New York, 2001.
[2] W.K. Härdle, Asymptotic maximal deviation of *M*-smoothers, J. Multivariate Anal. 29 (1989) 163–179.
[3] W.K. Härdle, L. Hoffmann, R. Moro, Learning Machines Supporting Bankruptcy Prediction, in: Cizek, Härdle, Weron (Eds.), Statistical Tools in Finance and Insurance, second ed., Springer, Berlin, 2011.
[4] W.K. Härdle, L. Huang, Analysis of deviance for hypothesis testing in generalized partially linear models, J. Bus. Econom. Statist. (2017) accepted. DOI: http://dx.doi.org/10.1080/07350015.2017.1330693.
[5] W.K. Härdle, E. Mammen, M. Müller, Testing parametric versus semiparametric modelling in generalized linear models, J. Amer. Statist. Assoc. 93 (1998) 1461–1474.
[6] T.J. Hastie, R.J. Tibshirani, Generalized Additive Models, Chapman & Hall, London, 1990.
[7] E. Kolaczyk, Empirical likelihood for generalized linear models, Statist. Sinica 4 (1994) 199–218.
[8] H. Liang, Y. Qin, X. Zhang, D. Ruppert, Empirical-likelihood-based inferences for generalized partially linear models, Scand. J. Stat. 36 (2009) 433–443.
[9] O.B. Linton, J.P. Nielsen, A kernel method of estimating structured nonparametric regression based on marginal integration, Biometrika 82 (1995) 93–100.
[10] R. Liu, L. Yang, Spline-backfitted kernel smoothing of additive coefficient model, Econom. Theory 26 (2010) 29–59.
[11] R. Liu, L. Yang, W.K. Härdle, Oracally efficient two-step estimation of generalized additive model, J. Amer. Statist. Assoc. 108 (2013) 619–631.
[12] S. Ma, R.J. Carroll, H. Liang, S. Xu, Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates, Ann. Statist. 43 (2015) 2102–2131.
[13] S. Ma, S. Racine, L. Yang, Spline regression in the presence of xategorical predictors, J. App. Econom. 30 (2015) 705–717.
[14] S. Ma, L. Yang, Spline-backfitted kernel smoothing of partially linear additive nodel, J. Statist. Plann. Inference 141 (2011) 204–219.

[15] A. Owen, Empirical Likelihood, Chapman & Hall/CRC, London, 2001.
[16] B. Park, E. Mammen, W.K. Härdle, S. Borak, Time series modelling with semiparametric factor dynamics, J. Amer. Statist. Assoc. 104 (2009) 284–298.
[17] T. Severini, J. Staniswalis, Quasi-likelihood estimation in semiparametric models, J. Amer. Statist. Assoc. 89 (1994) 501–511.
[18] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.
[19] C.J. Stone, The dimensionality reduction principle for generalized additive models, Ann. Statist. 14 (1986) 590–606.
[20] L. Wang, X. Liu, H. Liang, R.J. Carroll, Estimation and variable selection for generalized additive partial linear models, Ann. Statist. 39 (2011) 1827–1851.
[21] L. Wang, L. Yang, Spline-backfitted kernel smoothing of nonlinear additive autoregression model, Ann. Statist. 35 (2007) 2474–2503.
[22] L. Xue, H. Liang, Polynomial spline estimation for a generalized additive coefficient model, Scand. J. Stat. 37 (2010) 26–46.
[23] L. Xue, L. Yang, Additive coefficient modeling via polynomial spline, Statist. Sinica 16 (2006) 1423–1446.
[24] L. Yang, S. Sperlich, W.K. Härdle, Derivative estimation and testing in generalized additive models, J. Statist. Plann. Inference 115 (2003) 521–542.
[25] S. Zheng, R. Liu, L. Yang, W.K. Härdle, Statistical inference for generalized additive models: Simultaneous confidence corridors and variable selection, TEST 25 (2016) 607–626.