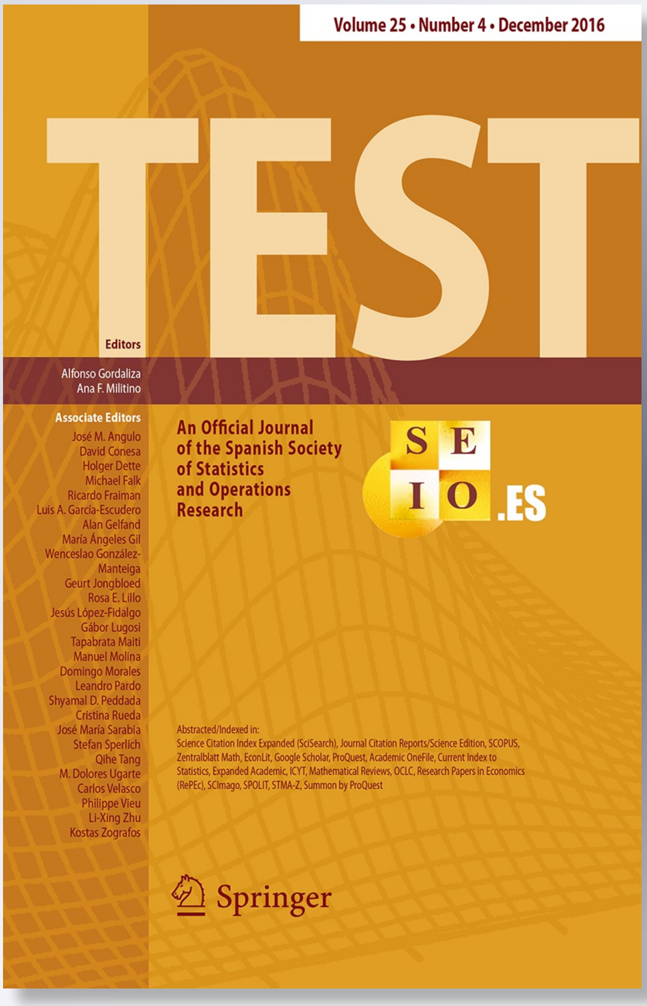
Statistical inference for generalized additive models: simultaneous confidence corridors and variable selection

Shuzhuan Zheng, Rong Liu, Lijian Yang & Wolfgang K. Härdle

TEST
An Official Journal of the Spanish Society of Statistics and Operations Research

ISSN 1133-0686
Volume 25
Number 4

TEST (2016) 25:607-626
DOI 10.1007/s11749-016-0480-8



Volume 25 • Number 4 • December 2016

TEST

An Official Journal of the Spanish Society of Statistics and Operations Research

Abstracted/Indexed in:
Science Citation Index Expanded (SciSearch), Journal Citation Reports/Science Edition, SCOPUS, Zentralblatt Math, EconLit, Google Scholar, ProQuest, Academic OneFile, Current Index to Statistics, Expanded Academic, ICYT, Mathematical Reviews, OCLC, Research Papers in Economics (RePEc), SCImago, SPOLIT, STMA-Z, Summon by ProQuest

Springer

Springer

TEST (2016) 25:607–626
DOI 10.1007/s11749-016-0480-8

ORIGINAL PAPER

Statistical inference for generalized additive models: simultaneous confidence corridors and variable selection

Shuzhuan Zheng[1,2] · **Rong Liu**[3] · **Lijian Yang**[4] · **Wolfgang K. Härdle**[5,6]

Received: 29 September 2014 / Accepted: 14 January 2016 / Published online: 23 April 2016

Abstract In spite of widespread use of generalized additive models (GAMs) to remedy the "curse of dimensionality", there is no well-grounded methodology developed for simultaneous inference and variable selection for GAM in existing literature. However, both are essential in enhancing the capability of statistical models. To this end, we establish simultaneous confidence corridors (SCCs) and a type of Bayesian information criterion (BIC) through the spline-backfitted kernel smoothing techniques proposed in recent articles. To characterize the global features of each non-parametric components, SCCs are constructed for testing their overall trends and entire shapes. By extending the BIC in additive models with identity/trivial link, an asymptotically consistent BIC approach for variable selection is built up in GAM to improve the parsimony of model without loss of prediction accuracy. Simulations and a real example corroborate the above findings.

Electronic supplementary material The online version of this article (doi:10.1007/s11749-016-0480-8) contains supplementary material, which is available to authorized users.

✉ Lijian Yang
yanglijian@tsinghua.edu.cn

[1] Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou 215006, China

[2] Department of Economics, Columbia University, New York, NY 10027, USA

[3] Department of Mathematics and Statistics, University of Toledo, Toledo, OH 43606, USA

[4] Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

[5] C.A.S.E.-Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[6] Lee Kong Chian School of Business, Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore, Singapore

Keywords BIC · Confidence corridor · Extreme value · Generalized additive mode · Spline-backfitted kernel

Mathematics Subject Classification 62G08 · 62G15 · 62G32

1 Introduction

Generalized additive model (GAM) has gained popularity on addressing the curse of dimensionality in multivariate nonparametric regressions with non-Gaussian responses. GAM was developed by Hastie and Tibshirani (1990) for blending generalized linear model with nonparametric additive regression, which stipulates that a data set $\left\{Y_i, \mathbf{X}_i^T\right\}_{i=1}^n$ consists of iid copies of $\left\{Y, \mathbf{X}^T\right\}$ that satisfies

$$\begin{aligned} \mathsf{E}(Y|\mathbf{X}) &= b'\left\{m\left(\mathbf{X}\right)\right\}, \operatorname{var}(Y|\mathbf{X}) = a\left(\phi\right) b''\left\{m\left(\mathbf{X}\right)\right\}, \\ m\left(\mathbf{X}\right) &= c + \sum_{l=1}^{d} m_l(X_l), \end{aligned} \tag{1}$$

$$Y = b'\left\{m\left(\mathbf{X}\right)\right\} + \sigma\left(\mathbf{X}\right)\varepsilon, \sigma\left(\mathbf{X}\right) = \left\{\operatorname{var}(Y|\mathbf{X})\right\}^{1/2}$$

where the response Y is one of certain types, such as Bernoulli, Poisson and so forth, the vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$ consists of the predictors, $m_l(\cdot), 1 \leq l \leq d$ are unknown smooth functions, the white noise ε satisfies that $\mathsf{E}\left(\varepsilon|\mathbf{X}\right) = 0$ and $\mathsf{E}\left(\varepsilon^2|\mathbf{X}\right) = 1$, while c is an unknown constant, $a\left(\phi\right)$ is a nuisance parameter that quantifies overdispersion, and the known inverse link function b' satisfies that $b' \in C^2\left(\mathbb{R}\right), b''\left(\theta\right) > 0, \theta \in \mathbb{R}$, see Assumption (A2) in the Appendix. In particular, if one takes the identity/trivial link, model (1) becomes a common additive model, see Huang and Yang (2004).

The joint density $f\left(\mathbf{x}\right)$ of $(X_1, \ldots, X_d)$ is assumed to be continuous and

$$0 < c_f \leq \inf_{\mathbf{x}\in[0,1]^d} f\left(\mathbf{x}\right) \leq \sup_{\mathbf{x}\in[0,1]^d} f\left(\mathbf{x}\right) \leq C_f < \infty,$$

see Assumption (A4) in the Appendix. Furthermore, for each $1 \leq l \leq d$, the marginal density function $f_l\left(x_l\right)$ of X_l has continuous derivatives on $[0, 1]$ and the same uniform bounds C_f and c_f. There exists a σ-finite measure λ on $\mathbb{R}$ such that the distribution of Y_i conditional on X_i has a probability density function $f_{Y|\mathbf{X}}\left(y; b'\left\{m\left(\mathbf{x}\right)\right\}\right)$ relative to λ whose support for y is a common Ω, and is continuous in both $y \in \Omega$ and $x \in [0, 1]^d$.

It is often the case that in model (1) the probability density function of Y_i conditional on $\mathbf{X}_i$ with respect to a fixed σ-finite measure forms an exponential family:

$$f\left(Y_i|\mathbf{X}_i, \phi\right) = \exp\left[\left\{Y_i m\left(\mathbf{X}_i\right) - b\left\{m\left(\mathbf{X}_i\right)\right\}\right\}/a\left(\phi\right) + h\left(Y_i, \phi\right)\right]. \tag{2}$$

Nonetheless, such an assumption is not necessary in this paper. Instead, we only stipulate that the conditional variance and conditional mean are linked by

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

For identifiability, one needs

$$\mathrm{E}\{m_l(X_l)\} = 0, 1 \le l \le d$$

that leads to unique additive representations of $m(\mathbf{x}) = c + \sum_{l=1}^{d} m_l(x_l)$. Without loss of generality, $\mathbf{x}$ take values in $\chi = [0,1]^d$.

Model (1) has numerous applications. In corporate credit rating, for instance, one is interested in modelling how the default or non-default of a given corporate or company depends on the additive effects of the covariates in financial statements, i.e., the response $Y = 0, 1$ with 1 indicating default, 0 indicating non-default, and the predictors are selected from financial statements with a logit-link $\left(b'\right)^{-1}(x) = \log\{x/(1-x)\}$. Our method has been applied to 3472 companies in Japan within a 5-year default horizon (2005–2010), and it has been discovered that the current liabilities and stock market returns of current, 3 months and 6 months prior to default are very significant as rating factors, and the default impact of the selected factors are examined via the simultaneous confidence corridors (SCCs) in Fig. 1a–c. More details of this example are contained in Sect. 6.

The smooth functions $\{m_l(x_l)\}_{l=1}^{d}$ in (1) can be estimated by, for instance, kernel methods in Linton and Härdle (1996), Linton (1997) and Yang et al. (2003), B-spline methods in Stone (1986) and Xue and Liang (2010), and two-stage methods in Horowitz and Mammen (2004). To make statistical inference on these functions individually and collectively, however, the proper tools are nonparametric simultaneous confidence corridors (SCCs) and consistent variable selection criteria, both of which are absent in the literature.

Nonparametric SCCs methodology has become increasingly important in statistical literature, see Xia (1998), Fan and Zhang (2000), Wu and Zhao (2007), Zhao and Wu (2008), Ma et al. (2012), Wang et al. (2014), Zheng et al. (2014), Gu et al. (2014), Cai and Yang (2015) and Gu and Yang (2015) for recent theoretical works on nonparametric SCCs. Capturing global shape properties by SCCs of the functions $\{m_l(x_l)\}_{l=1}^{d}$ in GAM (1) is of prime importance. A nonparametric component can be replaced by a parametric one covered entirely within the SCCs, significantly decreasing the estimation variance, see He et al. (2002, 2005) for discussions. As far as we know, SCCs has not been established for functions $\{m_l(x_l)\}_{l=1}^{d}$ in GAM (1) due to the lack of estimators that fit in Gaussian process extreme value theory. Using the spline-backfitted kernel (SBK) smoothing of Liu et al. (2013), we extend the SCCs works of univariate nonparametric regression in Bickel and Rosenblatt (1973) and Härdle (1989) to those of GAM. The SBK smoothing has been well developed in Wang and Yang (2007), Wang and Yang (2009), Liu and Yang (2010) and Ma and Yang (2011) for the much simpler additive model (i.e., GAM with $b'(x) \equiv x$) including the construction of SCCs, but ours is the first work on SCCs on GAM with nonlinear link.

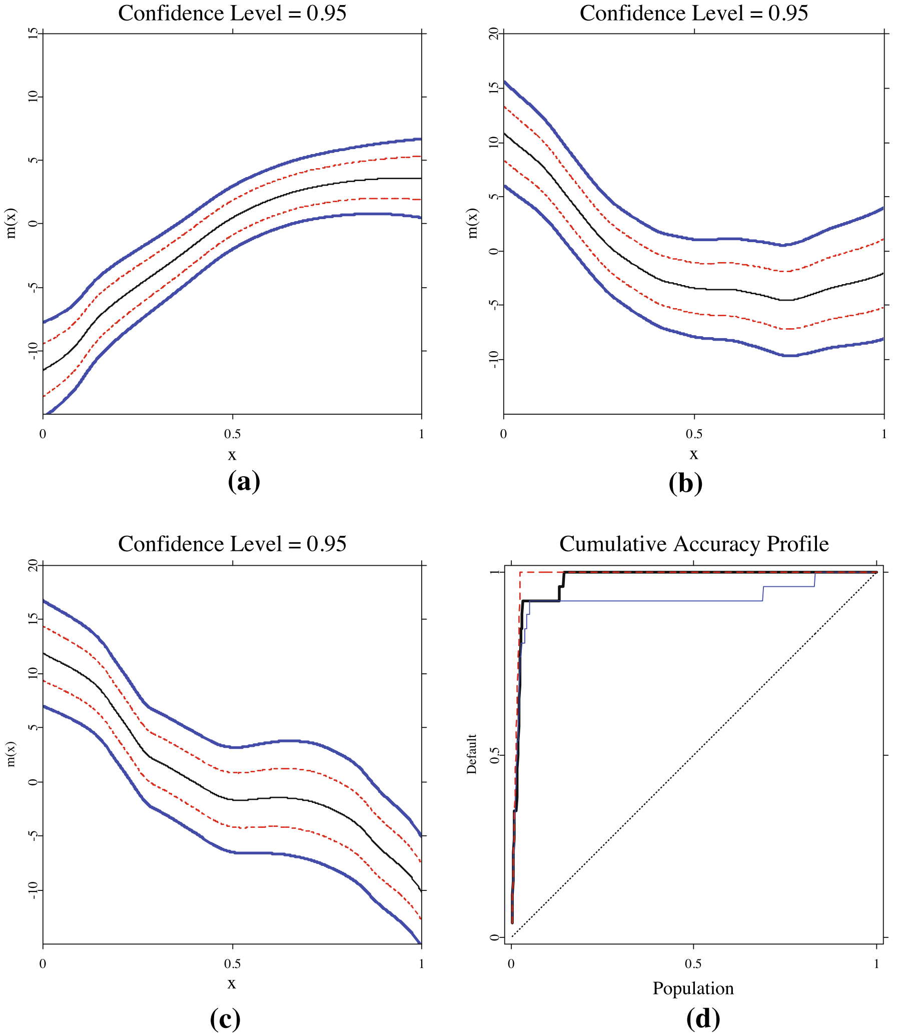


Fig. 1 Plots of the rating factors in **a**–**c** SBK estimators (*thin*), 95 % CIs (*dashed*) and 95 % SCCs (*thick*). Plot of the CAPs defined as (24) in **d** Perfect (*dashed*), GAM (*thick solid*), GLM (*thin solid*), non-informative (*dotted*). **a** Current liability. **b** 3 months earlier return. **c** 6 months earlier return. **d** The CAP curves

While variable selection for nonparametric additive model has been investigated under different settings, see Wang et al. (2008), there is lack of theoretically reliable variable selection approach for GAM. To the best of our knowledge, only Zhang and Lin (2006) proposed a sounding method named "COSSO" , which stands for components (CO) LASSO using penalized likelihood method, for selecting components in nonparametric regression with exponential families, but it leaves the asymptotic distributions and variable selection consistency to be desired. Instead, we tackle this issue by building a BIC type criterion based on spline pre-smoothing (first stage in the SBK), which is asymptotically consistent and easy to compute. Our work extends the BIC criterion for additive models (trivial link) in Huang and Yang (2004). Such an extension is challenging since a much more complicated quasi-likelihood is used in GAM with possibly nonlinear link instead of the log mean squared error for trivial link, see the Appendix for details.

The rest of paper is organized as follows. The SBK estimator and its oracle property are briefly described in Sect. 2. Asymptotic extreme value distribution of the SBK estimator is investigated in Sect. 3, which is used to construct the SCCs of component functions. Section 4 introduces a BIC criterion in the GAM setting and provides results on consistent component selection as well as the implementation, followed by the Monte Carlo simulations in Sect. 5. Section 6 illustrates the application of our SCCs and BIC methods to predict default of nearly 3, 500 listed companies in Japan. Technical assumptions and proofs are presented in the Appendix.

2 Spline-backfitted kernel smoothing in GAM

In this section, we briefly describe the spline-backfitted kernel (SBK) estimator for GAM (1) and its oracle properties obtained in Liu et al. (2013). Let $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ be i.i.d. observations following model (1). Without loss of generality, one denotes $\mathbf{x}_{\_1} = (x_2, \ldots, x_d)$ and $m_{\_1}\left(\mathbf{x}_{\_1}\right) = c + \sum_{l=2}^{d} m_l\left(x_l\right)$ and estimates $m_1\left(x_1\right)$.

As a benchmark of efficiency, we introduce the "oracle smoother" by treating the constant c and the last $d-1$ components $\{m_l\left(x_l\right)\}_{l=2}^d$ as known, then the only unknown component $m_1\left(x_1\right)$ may be estimated by the following procedure. Although the exponential family Eq. (2) does not necessarily hold, one still defines, as in Severini and Staniswalis (1994), for each $x_1 \in [h, 1-h]$ a local log-likelihood function $\tilde{l}\left(a\right) = \tilde{l}\left(a, x_1\right)$ as

$$\tilde{l}\left(a, x_1\right) = n^{-1} \sum_{i=1}^{n} \left[Y_i \left\{a + m_{\_1}\left(\mathbf{X}_{i,\_1}\right)\right\} - b\left\{a + m_{\_1}\left(\mathbf{X}_{i,\_1}\right)\right\}\right] K_h\left(X_{i1} - x_1\right), \tag{3}$$

where $a \in A$, a set whose interior contains $m_1\left([0, 1]\right)$. The oracle smoother of $m_1\left(x_1\right)$ is

$$\widetilde{m}_{\mathrm{K},1}\left(x_1\right) = \mathrm{argmax}_{a \in A}\, \tilde{l}\left(a, x_1\right).$$

Although $\widetilde{m}_{\mathrm{K},1}\left(x_1\right)$ is not a statistic since c and $\{m_l\left(x_l\right)\}_{l=2}^d$ are actually unknown, its asymptotic properties serve as a benchmark for estimators of $m_1\left(x_1\right)$ to achieve.

To define the SBK, we introduce the linear B spline basis for smoothing: $b_J\left(x\right) = \left(1 - |x - \xi_J|/H\right)_+, 0 \le J \le N+1$ where $0 = \xi_0 < \xi_1 < \cdots < \xi_N < \xi_{N+1} = 1$ are a sequence of equally spaced points, called interior knots, on interval [0, 1]. Denote by $H = (N+1)^{-1}$ the width of each subinterval $\left[\xi_J, \xi_{J+1}\right], 0 \le J \le N$ and the degenerate knots by $\xi_{-1} = 0, \xi_{N+2} = 1$. The space of l-empirically centered linear spline functions on [0, 1] is

$$G_{n,l}^{0} = \left\{ g_l : g_l(x_l) \equiv \sum_{J=0}^{N+1} \lambda_J b_J(x_l), \mathrm{E}_n\{g_l(X_l)\} = 0 \right\}, \quad 1 \le l \le d, \tag{4}$$

with empirical expectation $\mathrm{E}_n\{g_l(X_l)\} = n^{-1}\sum_{i=1}^{n} g_l(X_{li})$. The space of additive spline functions on $\chi = [0, 1]^d$ is

$$G_n^0 = \left\{ g(\mathbf{x}) = c + \sum_{l=1}^{d} g_l(x_l); c \in \mathbb{R}, \quad g_l \in G_{n,l}^0 \right\}.$$

The SBK method is defined in two steps. One first pre-estimates the unknown functions $\{m_l(x_l)\}_{l=2}^{d}$ and constants c by linear spline smoothing. We define the log-likelihood function $\widehat{L}(g)$ as

$$\widehat{L}(g) = n^{-1}\sum_{i=1}^{n} [Y_i g(\mathbf{X}_i) - b\{g(\mathbf{X}_i)\}], \quad g \in G_n^0. \tag{5}$$

According to Lemma 14 of Stone (1986), (5) has a unique maximizer with probability approaching 1. Therefore, the multivariate function $m(\mathbf{x})$ can be estimated by an additive spline function:

$$\widehat{m}(\mathbf{x}) = \mathrm{argmax}_{g \in G_n^0} \widehat{L}(g). \tag{6}$$

The spline estimator is asymptotically consistent, and can be solved efficiently via generalized linear models. However, as stated in Wang and Yang (2007) and Liu et al. (2013), spline methods only provide convergence rates but no asymptotic distributions, so no measures of confidence can be assigned to the estimators. To overcome this problem, we adapt the SBK estimator, which combines the strength of kernel smoothing with regression spline. One then rewrites $\widehat{m}(\mathbf{x}) = \hat{c} + \sum_{l=1}^{d} \widehat{m}_l(x_l)$ for $\widehat{c} \in \mathbb{R}$ and $\widehat{m}_l(x_l) \in G_{n,l}^0$ and defines a univariate quasi-likelihood function similar to $\widetilde{l}(a, x_1)$ in (3) as

$$\widehat{l}(a, x_1) = n^{-1}\sum_{i=1}^{n} \left[Y_i \left\{ a + \widehat{m}_{\_1}\left(\mathbf{X}_{i,\_1}\right) \right\} - b\left\{ a + \widehat{m}_{\_1}\left(\mathbf{X}_{i,\_1}\right) \right\} \right] K_h(X_{i1} - x_1)$$

with $\widehat{m}_{\_1}\left(\mathbf{x}_{\_1}\right) = \widehat{c} + \sum_{l=2}^{d} \widehat{m}_l(x_l)$ being the pilot spline estimator of $m_{\_1}\left(\mathbf{x}_{\_1}\right)$. Consequently, the spline-backfitted kernel (SBK) estimator of $m_1(x_1)$ is

$$\widehat{m}_{\mathrm{SBK},1}(x_1) = \mathrm{argmax}_{a \in A} \widehat{l}(a, x_1). \tag{7}$$

We now introduce some useful results and definitions from Liu et al. (2013), under Assumptions (A1)–(A7) in appendix, as $n \to \infty$,

$$\sup_{x_1 \in [0,1]} \left| \widehat{m}_{\mathrm{SBK},1}(x_1) - \widetilde{m}_{\mathrm{K},1}(x_1) \right| = \mathcal{O}_{a.s.}\left(n^{-1/2} \log n \right), \tag{8}$$

$$\begin{aligned} \widetilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1) &= \mathrm{bias}_1(x_1)\, h^2 / D_1(x_1) \\ &+ n^{-1} \sum_{i=1}^{n} K_h(X_{i1} - x_1)\, \sigma(\mathbf{X}_i)\, \varepsilon_i / D_1(x_1) + r_{\mathrm{K},1}(x_1) \end{aligned} \tag{9}$$

in which the higher order remainder $r_{\mathrm{K},1}(x_1)$ satisfies

$$\sup_{x_1 \in [h, 1-h]} \left| r_{\mathrm{K},1}(x_1) \right| = \mathcal{O}_{a.s.}\left(n^{-1/2} h^{1/2} \log n \right). \tag{10}$$

The scale function $D_1(x_1)$ and bias function $\mathrm{bias}_1(x_1)$ are defined in Liu et al. (2013) as:

$$\begin{aligned} &\sigma_b^2(x_1) = \mathrm{E}\left[b''\{m(\mathbf{X})\} | X_1 = x_1 \right],\ \sigma^2(x_1) = \mathrm{E}\left\{ \sigma^2(\mathbf{X}) | X_1 = x_1 \right\} \\ &D_1(x_1) = f_1(x_1)\, \sigma_b^2(x_1),\ v_1^2(x_1) = \|K\|_2^2\, f_1(x_1)\, \sigma^2(x_1). \\ &\mathrm{bias}_1(x_1) = \mu_2(K) \times \left\{ m_1''(x_1)\, D_1(x_1) + m_1'(x_1)\, f(x_1)\, \sigma_b^2(x_1)' \right. \\ &\qquad \left. - \left\{ m_1'(x_1) \right\}^2 f(x_1)\, \mathrm{E}\left[b'''\{m(\mathbf{X})\} | X_1 = x_1 \right] \right\} \end{aligned} \tag{11}$$

where $\|K\|_2^2 = \int K^2(u)\, du, \mu_2(K) = \int K(u)\, u^2 du$. The above Eqs. (8), (9) and (10) lead one to a simplifying decomposition of the estimation error $\widehat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1)$

$$\begin{aligned} &\sup_{x_1 \in [h, 1-h]} \left| \widehat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) - n^{-1} \sum_{i=1}^{n} K_h(X_{i1} - x_1)\, \sigma(\mathbf{X}_i)\, \varepsilon_i / D_1(x_1) \right| \\ &\quad = \mathcal{O}_{a.s.}\left(n^{-1/2} h^{1/2} \log n + n^{-1/2} \log n + h^2 \right). \end{aligned} \tag{12}$$

The decomposition in (12) is fundamental for constructing SCCs in Sect. 3, and it follows from Theorems 1 and 4 of Liu et al. (2013), which were proved under weak dependence. A similar Theorem 2 in Horowitz and Mammen (2004) for the two-stage estimator was established only for a fixed x_1, not uniformly for x_1 in the growing interval $[h, 1-h]$, and exclusively for iid data, not dependent data, see detailed discussion on page 621 of Liu et al. (2013).

3 GAM inference via simultaneous confidence corridor

In this section, we propose SCCs for GAM components based on the SBK smoothing, extending the works for univariate nonparametric function estimation in Bickel and Rosenblatt (1973) and Härdle (1989).

3.1 Main results

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

$$Q_h(\alpha) = a_h + a_h^{-1}\left[\log\left\{\sqrt{C(K)}/(2\pi)\right\} - \log\left\{-\log\sqrt{1-\alpha}\right\}\right]. \tag{13}$$

Also with $D_1(x_1)$ and $v_1^2(x_1)$ given in (11), we define

$$\sigma_n(x_1) = n^{-1/2}h^{-1/2}v_1(x_1)\, D_1^{-1}(x_1). \tag{14}$$

Theorem 1 *Under Assumptions (A1)–(A7), as $n \to \infty$*

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

A $100(1-\alpha)\,\%$ *simultaneous confidence corridor for* $m_1(x_1)$ *is*

$$\widehat{m}_{\mathrm{SBK},1}(x_1) \pm \sigma_n(x_1)\, Q_h(\alpha). \tag{15}$$

The above SCC for component function $m_1(x_1)$ resembles the SCCs in Bickel and Rosenblatt (1973) and Härdle (1989) for estimating unknown univariate nonparametric function, although it is for multivariate nonparametric regression.

3.2 Implementation

To satisfy Assumption (A4), one could use the transformed $U_{il} = F_{nl}(X_{il})$ instead of X_{il} as predictors for each $l = 1, \ldots, d$ and $i = 1, \ldots, n$, where F_{nl} is the empirical distribution of $(X_{1l}, \ldots, X_{nl})$. We still use symbol X instead of U to avoid involving new symbols, but the X variates have been transformed in simulation study and applications.

To construct the SCC for $m_1(x_1)$ in (15), one needs to select the bandwidth h and the number of knots N to evaluate $m_{\mathrm{SBK},1}(x_1)$, $Q_h(\alpha)$ and $\sigma_n(x_1)$ given in (7), (13) and (14).

Assumption (A6) requires that the bandwidth for SCCs be different from the mean square optimal bandwidth $h_{\mathrm{opt}} \sim n^{-1/5}$ (minimizing AMISE) in Liu et al. (2013). This is due to the two conflicting goals in SCCs construction: coverage of the true curve and narrowness of the corridor, are not quantifiable in a single measure to minimize, such as the mean integrated squared error. We, therefore, take $h = h_{\mathrm{opt}}(\log n)^{-1/4}$, as a data-driven undersmoothing bandwidth for SCCs construction to fulfill Assumption (A6), where h_{opt} is computed as in Liu et al. (2013), page 623–624. Recent articles on SCCs for time series, such as Wu and Zhao (2007), Zhao and Wu (2008), have used similar undersmoothing bandwidths.

For a given l and a chosen bandwidth h, one can easily estimate $m_{\mathrm{SBK},1}(x_1)$ and $Q_h(\alpha)$ as in (7), (13). To evaluate $\sigma_n(x_1)$, one needs to estimate $v_1(x_1)$ and $D_1^{-1}(x_1)$ given in (11), i.e., estimating $f(x_1)$, $\sigma_b^2(x_1)$ and $\sigma^2(x_1)$. The density function $f(x_1)$

is estimated by $\widehat{f}(x_1) = n^{-1}\sum_{i=1}^{n} K_{h_{\mathrm{ROT}}}(X_{i1} - x_1)$, where h_{ROT} is the rule-of-thumb bandwidth in equation (5.8), page 200 of Fan and Yao (2003), namely $h_{\mathrm{ROT}} = \left(8\sqrt{\pi}/3\right)^{1/5} \mu_2(K) \|K\|_2^{2/5} n^{-1/5}\hat{\sigma}$, in which $\hat{\sigma}$ is the sample standard deviation of $\{X_{i1}\}_{i=1}^{n}$. We further illustrate the spline estimates of $\sigma_b^2(x_1)$ and $\sigma^2(x_1)$ below:

One partitions $\min_i X_{i1} = t_{1,0} < \cdots < t_{1,N+1} = \max_i X_{i1}$ where N is the number of spline interior knots, i.e.,

$$\max\left(1, \min\left(\left\lfloor n^{1/4}\log n + 1\right\rfloor, \lfloor n/4d - 1/d\rfloor - 1\right)\right), \tag{16}$$

which satisfies Assumption (A7) in the Appendix. Then $\sigma_b^2(x_1)$ can be estimated as $\sum_{k=0}^{3}\widehat{a}_{1,k}x_1^k + \sum_{k=4}^{N+3}\widehat{a}_{1,k}\left(x_1 - t_{l,k-3}\right)_+^3$ where $\left\{\widehat{a}_{1,k}\right\}_{k=0}^{N+3}$ minimize

$$\sum_{i=1}^{n}\left[b''\{\widehat{m}(\mathbf{X}_i)\} - \left\{\sum_{k=0}^{3} a_{1,k}X_{i1}^k + \sum\nolimits_{k=4}^{N+3} a_{1,k}(X_{i1} - t_{k-3})_+^3\right\}\right]^2, \tag{17}$$

and $\sigma^2(x_1)$ can be estimated as $\sum_{k=0}^{3}\widehat{a}_{1,k}x_1^k + \sum_{k=4}^{N+3}\widehat{a}_{1,k}\left(x_1 - t_{l,k-3}\right)_+^3$ where $\left\{\widehat{a}_{1,k}\right\}_{k=0}^{N+3}$ minimize

$$\sum_{i=1}^{n}\left[\left[Y_i - b'\{\widehat{m}(\mathbf{X}_i)\}\right]^2 - \left\{\sum_{k=0}^{3} a_{l,k}X_{i1}^k + \sum_{k=4}^{N+3} a_{l,k}(X_{i1} - t_{k-3})_+^3\right\}\right]^2. \tag{18}$$

The resulted estimate $\hat{\sigma}_n(x_1)$ of $\sigma_n(x_1)$, using (17) and (18) satisfies $\sup_{x_1\in[h,1-h]}\left|\hat{\sigma}_n(x_1) - \sigma_n(x_1)\right| = \mathcal{O}_p\left(n^{-\gamma}\right)$ for some $\gamma > 0$, see Liu et al. (2013) Sect. 5 for details. This consistency and Slutsky's theorem ensure that

$$\mathrm{P}\left\{\sup_{x_1\in[h,1-h]}\left|\widehat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1)\right|/\hat{\sigma}_n(x_1) \leq Q_h(\alpha)\right\} \to 1 - \alpha$$

as $n \to \infty$, and therefore $\widehat{m}_{\mathrm{SBK},1}(x_1) \pm \hat{\sigma}_n(x_1) Q_h(\alpha)$ is a $100(1-\alpha)\,\%$ simultaneous confidence corridor for $m_1(x_1)$. The SCCs constructions of other components $m_2(x_2), \ldots, m_d(x_d)$ are similar. It is worth while to emphasize that, based on extensive simulation experiments, the estimators $\widehat{m}_{\mathrm{SBK},1}(x_1)$, $\widehat{Q}_h(\alpha)$, $\widehat{f}(x_1)$ and $\hat{\sigma}_n(x_1)$ remain stable if h and N slightly vary.

4 Variable selection in GAM

In this section, we propose a Bayesian Information Criterion (BIC) for component function selection based on spline smoothing in step one of the SBK estimation for GAM and an efficient implementation follows.

4.1 Main results

According to Stone (1985), p. 693, the space of l-centered square integrable functions on [0, 1] is defined as

$$\mathcal{H}_l^0 = \left\{ g : \mathrm{E}\left\{ g\left(X_l\right) \right\} = 0, \mathrm{E}\left\{ g^2\left\{X_l\right\} \right\} < \infty, 1 \le l \le d \right\}, \tag{19}$$

and the model space $\mathcal{M}$ is

$$\mathcal{M} = \left\{ g\left(\mathbf{x}\right) = c + \sum_{l=1}^{d} g_l\left(\mathbf{x}_l\right); c \in \mathbb{R}, g_l \in \mathcal{H}_l^0, 1 \le l \le d \right\}. \tag{20}$$

To introduce the proposed BIC, let $\{1, \ldots, d\}$ denote the complete set of indices of d tuning variables $(X_1, \ldots, X_d)$. For each subset $S \subset \{1, \ldots, d\}$, define a corresponding model space $\mathcal{M}_S$ for S as

$$\mathcal{M}_S = \left\{ g\left(\mathbf{x}\right) = c + \sum_{l \in S} g_l\left(\mathbf{x}_l\right); c \in \mathbb{R}, g_l \in \mathcal{H}_l^0, l \in S \right\},$$

with $\mathcal{H}_l^0$ given in (19), and the space of the additive spline functions as

$$G_{n,S}^0 = \left\{ g\left(\mathbf{x}\right) = c + \sum_{l \in S} g_l\left(x_l\right); c \in \mathbb{R}, g_l \in G_{n,l}^0, l \in S \right\},$$

with $G_{n,l}^0$ given in (4). Following Definition 1 of Huang and Yang (2004), the set S_0 of significant variables is defined as the minimal set $S \subset \{1, \ldots, d\}$ such that $m \in \mathcal{M}_S$. According to Lemma 1 of Huang and Yang (2004), the set S_0 is uniquely defined. Standard theory of Hilbert space and subspace projection implies that the set S_0 is also the minimal set $S \subset \{1, \ldots, d\}$ such that $\mathrm{E}\{m\left(\mathbf{X}\right) - m_S\left(\mathbf{X}\right)\}^2 = 0$ in which the least squares projection of function m in $\mathcal{M}_S$ is

$$m_S = \underset{g \in \mathcal{M}_S}{\operatorname{argmin}}\, \mathrm{E}\left\{ m\left(\mathbf{X}\right) - g\left(\mathbf{X}\right) \right\}^2. \tag{21}$$

To identify S_0, one computes for an index set S the BIC as

$$\mathrm{BIC}_S = -2\widehat{L}\left(\widehat{m}_S\right) + \frac{N_S}{n}\left(\log n\right)^3 \tag{22}$$

where $\widehat{L}\left(\cdot\right)$ is given in (5), $\widehat{m}_S\left(\mathbf{x}\right) \in G_{n,S}^0$ is the pilot spline estimator as in (6), $N_S = 1 + (N+1)\,\#\left(S\right)$ with N the number of interior knots as defined in (16), $\#\left(S\right)$ the cardinality of S.

Our variable selection rule takes the subset $\widehat{S} \subset \{1, \ldots, d\}$ that minimizes BIC_S.

Theorem 2 *Under Assumptions (A1)–(A5), (A7),* $\lim_{n\to\infty} P\left(\widehat{S}=S_0\right)=1$.

According to Theorem 2, the variable selection rule based on the BIC in (22) is consistent. The nonparametric version BIC was firstly established in Huang and Yang (2004) for additive autoregression model, and adapted to additive coefficient model by Xue and Yang (2006), to single index model by Wang and Yang (2009). Our proposed BIC differs from all of the above as it is based on quasi-likelihood rather than mean squared error, which makes the technical proof of consistency much more challenging. To the best of our knowledge, it is the first theoretically reliable information criterion in this setting.

4.2 Implementation

We have not implemented the BIC variable selection by a greedy search through all possible subsets. Instead, a forward stepwise procedure is used with minimizing BIC as the criterion since it is more common that only a few variables are significant among many variables. We have also experimented with backward as well as forward–backward stepwise procedures which have yielded similar outcomes in simulation examples.

5 Simulation

This section studies under simulated setting the performance of the proposed procedures including the computational cost of the SBK, the consistency of selecting variables via BIC and the coverage frequency of the SCCs. The data are generated from

$$\mathrm{P}(Y=1|\mathbf{X}=\mathbf{x})=b'\left\{c+\sum_{l=1}^{d}m_l\left(X_l\right)\right\},b'\left(x\right)=\frac{e^x}{1+e^x} \tag{23}$$

with $d=10, c=0, m_3\left(x\right)=\sin\left(4\pi x\right), m_4\left(x\right)=m_5\left(x\right)=\sin\left(\pi x\right), m_6\left(x\right)=x, m_7\left(x\right)=e^x-\left(e-e^{-1}\right)$ and $m_l\left(x\right)=0$ for $l=1,2,8,9,10$. The predictors are generated by

$$X_{il}=2\Phi\left(Z_{il}\right)-1,\ \mathbf{Z}_i=\left(Z_{i1},\ldots,Z_{id}\right)\sim\mathrm{N}\left(0,\Sigma\right),\quad 1\le i\le n,\quad 1\le l\le d,$$

where Φ is the standard normal c.d.f. and $\Sigma=(1-r)\,\mathbf{I}_{d\times d}+r\mathbf{1}_d\mathbf{1}_d^T$. The parameter r $(0\le r<1)$ controls the correlation between $Z_{il,}1\le l\le d$. To examine the computing advantage of BIC for large d, we have also included results for $d=50$ with $m_3,\ldots,m_7$ as above and all the other component functions are 0.

COSSO is a penalized likelihood method proposed in Zhang and Lin (2006) for LASSO type component selection and nonparametric regression in exponential families. In what follows, the performance of BIC and COSSO is firstly compared, followed by a computational comparison between the SBK and a kernel method in GAM, and it ends with a report on the SCCs coverage frequency for components function (the

Table 1 Simulation comparison of the proposed BIC method and COSSO with $d = 10, 50$

d	r	n	Computing time			Accuracy					
			BIC	COSSO	Ratio	BIC			COSSO		
10	0	250	0.17	1.85	10.9	25	441	34	98	327	75
		500	0.41	4.33	10.6	6	476	28	42	414	44
		1000	0.66	20.14	30.5	2	491	7	26	455	19
	0.5	250	0.18	1.91	10.6	165	298	37	204	221	75
		500	0.42	4.43	10.5	11	452	37	89	359	52
		1000	0.67	20.64	30.8	1	493	6	67	401	32
50	0	250	1.00	–	–	312	78	110	–	–	–
		500	1.43	59.77	41.8	106	327	67	124	207	169
		1000	3.32	268.24	80.8	2	465	33	20	426	54
	0.5	250	1.04	–	–	319	65	116	–	–	–
		500	1.55	60.87	39.2	297	174	29	203	145	152
		1000	3.48	274.25	78.8	47	428	25	52	356	92

Computing time is in seconds and the ratio is the computing time of COSSO over that of BIC. For $d = 50$ and $n = 250$, COSSO becomes unstable to the point of crashing. Accuracy (the last 6 columns) gives for BIC and COSSO the numbers of underfitting, correct fitting, and overfitting out of 500 replications

frequency that SCCs covering the entire curve on the domain). We have tried numbers of knots different from the one in (16) with similar results, so our conclusion is that the performance of BIC is rather insensitive to the number of knots.

Table 1 shows the simulation results from 500 replications, where the outcome is defined in accuracy as correct fitting, if $\widehat{S} = S_0$; overfitting, if $S_0 \subset \widehat{S}$; and underfitting, if $S_0 \not\subseteq \widehat{S}$. It is clear that the performance of BIC on selecting 5 significant variables $m_l\left(X_l\right), l = 3, \ldots, 7$, is quite satisfactory. The selection accuracy becomes higher as the sample size increases and/or the correlation decreases; it is poorer with higher dimension d $(= 50)$ but still high when sample size $n = 1000$. The accuracy and computing time of COSSO are also listed for comparison (Platform: R; PC: Intel 3.1 GHz processor and 8 GB RAM). It is shown in Table 1 that the BIC significantly outperforms the COSSO in terms of accuracy and computing time, and the advantage in computing time widens significantly for $d = 50$.

In addition to the above comparison for model selection, we have also conducted numerical comparison between COSSO and our proposed SBK estimation method in terms of probability prediction. The proposed SBK method has higher prediction accuracy in almost all cases, see Table 4 in the Supplement. Comparison regarding SCC has not been made against COSSO because it does not produce one.

The SCCs coverage frequency for $m_l\left(x_l\right), l = 1, \ldots, 7$ is reported in Table 2. Among the zero functions, we have omitted the results for m_8, m_9 and m_{10} because the results are very similar to m_1 and m_2. The empirical coverage approaches the nominal confidence levels as n increases, and better coverage occurs when the correlation is lower. The coverage frequencies vary slightly when d increases, the numerical results of which have not been included for brevity. We have also compared the coverage

Table 2 The 95 % SCCs coverage frequency for $m_l(x)$, $l = 1, 2, \ldots, 7$ from 2000 replications

r	n	l						
		1	2	3	4	5	6	7
0.0	250	0.9305	0.9250	0.9235	0.9250	0.9235	0.9240	0.9230
	500	0.9455	0.9475	0.9430	0.9405	0.9425	0.9440	0.9530
	1000	0.9515	0.9520	0.9475	0.9455	0.9480	0.9510	0.9485
0.5	250	0.9215	0.9185	0.9120	0.9145	0.9205	0.9210	0.9185
	500	0.9420	0.9405	0.9330	0.9325	0.9375	0.9385	0.9415
	1000	0.9485	0.9505	0.9420	0.9475	0.9455	0.9430	0.9445

frequency of SCC and method VOT (Volume of Tube) in the same setup of the simulation 1 in Wiesenfarth et al. (2012), which considered only the case of trivial link function. The performance of our proposed SCC is quite similar to the VOT method Wiesenfarth et al. (2012), see Table 3 in the Supplement.

The above studies evidently indicate the reliability of our methodology, such as high selection accuracy of the BIC and desired coverage frequency of the SCCs. It ensures their applications for credit rating modelling in the following section.

6 Application

We now return to forecast default probabilities of the listed companies in Japan. The data taken from the Risk Management Institute, National University of Singapore include the comprehensive financial statements and the credit events (default or bankruptcy) from 2005 to 2010 of 3583 Japanese firms.

Berg (2007) found that the liability status was important to indicate the creditworthiness of a company, while Bernhardsen (2001) and Ryser and Denzler (2009) proposed to consider the "leverage effect" expressed by the financial statement ratios. Therefore, we have pooled two situations by considering X_1: Current liability, X_2: Current stock return, X_3: Long-term borrow, X_4: Short-term borrow, X_5: Total asset, X_6: Non-current liability, X_7: 3 months earlier (stock) return, X_8: 6 months earlier (stock) return, X_9: Current ratio, X_{10}: Net liability to shareholder equity, X_{11}: Shareholder equity to total liability and equity, X_{12}: TCE ratio, X_{13}: Total debt to total asset, X_{14}: Quick ratio.

Selecting the rating factors via the BIC given in (22), we have found that X_1: Current liabilities, X_7: 3 months earlier return, X_8: 6 months earlier return are significant. Similar rating covariates were also discovered in Shina and Moore (2003), Berg (2007) and Ryser and Denzler (2009). However, Berg (2007) selected 23 variables which led to a non-parsimonious GAM. In contrast, Ryser and Denzler (2009) had found that 3 financial ratios (capital turnover, long-term debt ratio, return on total capital) were significant based on the blockwise cross-validation (CV) method which is nonetheless extremely time consuming in comparison to the proposed BIC.

Figure 1a–c depicts the SBK estimator of the factor's default impact curve on domain, while a shoal of 95 % CIs and the 95 % SCCs present, respectively, the pointwise and global uncertainty of the whole curve. The SBK estimators indicate overall monotonicities of each rating factors, and the SCCs turn out to be fairly narrow to warrant the global nonlinearities of the factors' curves which reveal the underlying nonlinear features in different segments of domain.

As for the model evaluations, the cumulative accuracy profile (CAP) is plotted in Fig. 1d. For any score function S, one defines its alarm rate $F(s) = P(S \leq s)$ and the hit rate $F_{\mathrm{D}}(s) = P(S \leq s|\,\mathrm{D})$ where D represents the conditioning event of "default". One then defines the CAP curve as

$$\mathrm{CAP}(u) = F_{\mathrm{D}}\left(F^{-1}(u)\right), \quad u \in (0, 1), \tag{24}$$

which is the percentage of default-infected obligators that are found among the first (according to their scores) $100u$ % of all obligators. A satisfactory model's CAP would be expected to approach to that of the perfect model (i.e., $\mathrm{CAP}_{\mathrm{P}}(u) = \min(u/p, 1), u \in (0, 1)$ where p is the unconditional default probability) and always better than the noinformative. In contrast, a noninformative rating method with zero discriminatory power displays a diagonal line $\mathrm{CAP}_{\mathrm{N}}(u) \equiv u, u \in (0, 1)$. See details of the CAP in Engelmann et al. (2003).

The AR is the ratio of two areas a_R and a_P. The area between the given CAP curve and the noninformative diagonal $\mathrm{CAP}_{\mathrm{N}}(u) \equiv u$ is a_R, whereas a_P is the area between the perfect CAP curve $\mathrm{CAP}_{\mathrm{P}}(u)$ and the noninformative diagonal $\mathrm{CAP}_{\mathrm{N}}(u)$. Thus,

$$\mathrm{AR} = \frac{a_R}{a_P} = \frac{2\int_0^1 \mathrm{CAP}(u)\,\mathrm{d}u - 1}{1 - p}, \tag{25}$$

where CAP (u) is given in (24). 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.

Using both GAM and GLM obtained from first 2000 companies to predict the default rate of the rest 1583 companies, the accuracy ratio is 97.56 % for GAM, much higher than the 89.76 % for GLM. We have also applied the COSSO method to the same data, and the following error message has appeared "Error in solve.QP(GH$H, GH$H %*% old.theta - GH$G, t(Amat), bvec): matrix D in quadratic function is not positive definite!", which once again has illustrated the advantage of the proposed BIC procedure over the existing method.

Acknowledgments This work is supported in part by the Jiangsu Specially Appointed Professor Program SR10700111, the Jiangsu Key Discipline Program (Statistics) ZY107002, ZY107992 National Natural Science Foundation of China award 11371272, Research Fund for the Doctoral Program of Higher Education of China award 20133201110002, United States NSF awards DMS 0706518, DMS 1007594, an Michigan State University Dissertation Continuation Fellowship, funding from the National University of Singapore, the Deutsche Forschungsgemeinschaft (DFG) via SFB 649 "Economic Risk", and the International Research Training Group (IRTG) 1792. The helpful comments from two Reviewers and an Associate Editor are gratefully acknowledged.

7 Appendix

In what follows, we take $\|\cdot\|$ and $\|\cdot\|_\infty$ as the Euclidean and supremum norms, respectively, i.e., for any $\mathbf{x} = (x_1, x_2, \ldots, x_d)^{\ T} \in \mathbb{R}^d$, $\|\mathbf{x}\| = \left(\sum_{l=1}^d x_l^2\right)^{1/2}$ and $\|\mathbf{x}\|_\infty = \max\limits_{1 \le l \le d} |x_l|$. For any interval $[a, b]$, denote the space of pth order smooth function by $C^{(p)}[a, b] = \left\{g \middle| g^{(p)} \in C\,[a, b]\right\}$, and the class of Lipschitz continuous functions by

$$\mathrm{Lip}\,([a, b]\,, C) = \left\{g \middle| \left|g\,(x) - g\left(x'\right)\right| \le C\left|x - x'\right|, \quad \forall x, x' \in [a, b]\right\}$$

for constant $C > 0$. Lastly, define the following latent regression errors

$$\xi_i = Y_i - b'\left\{m\left(\mathbf{X}_i\right)\right\} = \sigma\left(\mathbf{X}_i\right)\varepsilon_i, \quad 1 \le i \le n. \tag{26}$$

7.1 Technical assumptions

We need the following technical assumptions:

(A1) *The additive component functions* $m_l \in C^{(1)}\,[0, 1]\,, 1 \le l \le d$*:* $m_1 \in C^{(2)}\,[0, 1]$, $m_l' \in \mathrm{Lip}\,([0, 1]\,, C_m)\,, 2 \le l \le d$ *for some constant* $C_m > 0$.

(A2) *The inverse link function* b' *satisfies that* $b' \in C^2\,(\mathbb{R})\,, b''\,(\theta) > 0, \theta \in \mathbb{R}$*. For a compact interval* Θ *whose interior contains* $m\left([0, 1]^d\right)$, $C_b > \max_{\theta \in \Theta} b''\,(\theta) \ge \min_{\theta \in \Theta} b''\,(\theta) > c_b$ *for constants* $0 < c_b < C_b < \infty$.

(A3) *The conditional variance function* $\sigma^2\,(\mathbf{x})$ *is continuous and positive for* $\mathbf{x} \in [0, 1]^d$*. The errors* $\{\varepsilon_i\}_{i=1}^n$ *satisfy that* $\mathrm{E}\,(\varepsilon_i|\mathbf{X}_i) = 0$, $\mathrm{E}\left(|\varepsilon_i|^{2+\eta}\right) \le C_\eta$ *for some* $\eta \in (1/2, 1]$.

(A4) *The joint density* $f\,(\mathbf{x})$ *of* $(X_1, \ldots, X_d)$ *is continuous and*

$$0 < c_f \le \inf_{\mathbf{x} \in [0,1]^d} f\,(\mathbf{x}) \le \sup_{\mathbf{x} \in [0,1]^d} f\,(\mathbf{x}) \le C_f < \infty.$$

For each $1 \le l \le d$*, the marginal density function* $f_l\,(x_l)$ *of* X_l *has continuous derivatives on* $[0, 1]$ *and the same uniform bounds* C_f *and* c_f*. There exists a* σ *-finite measure* λ *on* $\mathbb{R}$ *such that the distribution of* Y_i *conditional on* $\mathbf{X}_i$ *has a probability density function* $f_{Y|\mathbf{X}}\left(y; b'\left\{m\,(\mathbf{x})\right\}\right)$ *relative to* λ *whose support for* y *is a common* Ω*, and is continuous in both* $y \in \Omega$ *and* $\mathbf{x} \in [0, 1]^d$.

(A5) $\left\{\mathbf{Z}_i = \left(\mathbf{X}_i^T, \varepsilon_i\right)\right\}_{i=1}^n$ *are independent and identically distributed.*

(A6) *The kernel function* $K\,(x)$ *is a symmetric probability density function supported on* $[-1, 1]$ *and* $\in C^1[-1, 1]$*. The bandwidth* $h = h_n$ *satisfies that* $h = o\left(n^{-1/5}(\log n)^{-1/5}\right)$, $h^{-1} = \mathcal{O}\left(n^{1/5}\,(\log n)^\delta\right)$ *for some constant* $\delta > 1/5$.

(A7) *The number of interior knots* N satisfies $c_N n^{1/4} \log n \le N \le C_N n^{1/4} \log n$ *for some constants*$c_N, C_N > 0$.

Assumptions (A1)–(A7) are standard in GAM, see Stone (1986), Xue and Yang (2006). The i.i.d. feature is technically acceptable if the data are collected across a large number of sections, for instance, our real example in Sect. 6. Assumptions (A5), (A6) are more restrictive than in Liu et al. (2013) for the purpose of constructing simultaneous confidence corridor, but are unnecessary for Theorem 2 on the consistency of BIC. All these assumptions are satisfied by the simulation example in Sect. 5.

7.2 Preliminaries

Throughout this section, C denotes some generic positive constant unless stated otherwise. Define

$$M_h\left(t\right)=h^{-1/2}\int_0^1 K\left\{\left(x-t\right)/h\right\}dW\left(x\right) \tag{27}$$

where $W\left(x\right)$ is a Wiener process defined on $(0,\infty)$ and denote

$$d_h=\left(-2\log h\right)^{1/2}+\left(-2\log h\right)^{-1/2}\left\{\sqrt{C\left(K\right)}/\left(2\pi\right)\right\}$$

with $C\left(K\right)$ given in (13).

Lemma 1 *Under Assumption (A6), for any $x\in\mathbb{R}$*

$$\lim_{n\to\infty}\mathrm{P}\left[\left(-2\log h\right)^{1/2}\left\{\sup_{t\in[h,1-h]}\left|M_h\left(t\right)\right|/\left\|K\right\|_2^2-d_h\right\}<x\right]=e^{-2e^{-x}}.$$

Proof One simply applies the same steps in proving Lemma 2.2 of Härdle (1989).

Denote by T_i the random variable $b'\left\{m\left(\mathbf{X}_i\right)\right\}$, and the Lebesgue measure on $\mathbb{R}^d$ as $\mu^{(d)}$. By Assumption (A4), $\mathbf{X}_i$ has pdf w.r.t. the Lebesgue measure $\mu^{(d)}$, and Assumptions (A1) and (A2) ensure that functions b' and m are at least C^1, thus the random vector $\left(T_i,X_{i1}\right)$ has a joint pdf w.r.t. the Lebesgue measure $\mu^{(2)}$, which one denotes as $f_{T,X_1}\left(t,x_1\right)$. □

Lemma 2 *Under Assumptions (A1)–(A5), for ξ_i in (26), the distribution of $\left(\xi_i,X_{i1}\right)$ has joint pdf w.r.t. $\mu^{(2)}$ as*

$$f_{\xi,X_1}\left(z,x_1\right)=\int_\Omega f_{Y|\mathbf{X}}\left(y;y-z\right)f_{T,X_1}\left(y-z,x_1\right)d\lambda\left(y\right).$$

Proof The joint pdf of $\left(Y_i,T_i,X_{i1}\right)$ wrt $\lambda\times\mu^{(2)}$ is $f_{Y|\mathbf{X}}\left(y;t\right)f_{T,X_1}\left(t,x_1\right)$. For any $\left(z,x_1\right)\in\mathbb{R}\times[0,1]$, and $\triangle z,\triangle x_1>0$, one has

$$\begin{aligned}&P\left[\left(\xi_i,X_{i1}\right)\in\left(z-\triangle z,z+\triangle z\right)\times\left(x_1-\triangle x_1,x_1+\triangle x_1\right)\right]\\&\quad=P\left[\left(Y_i-T_i,X_{i1}\right)\in\left(z-\triangle z,z+\triangle z\right)\times\left(x_1-\triangle x_1,x_1+\triangle x_1\right)\right]\\&\quad=\int_\Omega d\lambda\left(y\right)\int_{y-\tau\in\left(z-\triangle z,z+\triangle z\right)}d\tau\int_{\chi_1\in\left(x_1-\triangle x_1,x_1+\triangle x_1\right)}f_{Y|\mathbf{X}}\left(y;\tau\right)f_{T,X_1}\left(\tau,\chi_1\right)d\chi_1.\end{aligned}$$

Applying dominated convergence theorem, one has as max $(\triangle z, \triangle x_1) \to 0$,

$$\begin{aligned}
&P\left[(\xi_i, X_{i1}) \in (z - \triangle z, z + \triangle z) \times (x_1 - \triangle x_1, x_1 + \triangle x_1)\right] \\
&= \left\{ \int_\Omega f_{Y|\mathbf{X}}(y; y - z)\, f_{T,X_1}(y - z, x_1)\, d\lambda(y) \right\} \\
&\quad \times \mu^{(2)}\left[(z - \triangle z, z + \triangle z) \times \{(x_1 - \triangle x_1, x_1 + \triangle x_1) \cap [0,1]\}\right] + o(\triangle z \triangle x_1)
\end{aligned}$$

hence the joint pdf of (ξ_i, X_{i1}) wrt $\mu^{(2)}$ is $\int_\Omega f_{Y|\mathbf{X}}(y; y - z)\, f_{T,X_1}(y - z, x_1)\, d\lambda(y)$.

For theoretical analysis, we write $c_{J,l} = \mathrm{E}\, b_J(X_l) = \int b_J(x_l)\, f_l(x_l)\, dx_l$ and define the centered B spline basis $b_{J,l}(x_l)$ and the standardized B spline basis $B_{J,l}(x_l)$ respectively as

$$\begin{aligned}
b_{J,l}(x_l) &= b_J(x_l) - \frac{c_{J,l}}{c_{J-1,l}} b_{J-1}(x_l), \\
B_{J,l}(x_l) &= \frac{b_{J,l}(x_l)}{\left\{\int b_{J,l}^2(x_l)\, f_l(x_l)\, dx_l\right\}^{1/2}}, \quad 1 \le J \le N + 1,
\end{aligned} \tag{28}$$

so that $\mathrm{E}\, B_{J,l}(X_l) \equiv 0$, $\mathrm{E}\, B_{J,l}^2(X_l) \equiv 1$.

With slight abuse of notations the log-likelihood $\widehat{L}(g)$ in (5) is

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

with $g(\mathbf{X}_i) = \lambda^T \mathbf{B}(\mathbf{X}_i) \in G_n^0$, $\lambda = \left(\lambda_0, \lambda_{J,l}\right)^T_{1 \le J \le N+1, 1 \le l \le d} \in \mathbb{R}^{N_d}$ with $N_d = (N+1)d + 1$, $\mathbf{B}(\mathbf{x}) = \left\{1, B_{1,1}(x_1), \ldots, B_{N+1,d}(x_d)\right\}^T$ and $B_{J,l}(x_l)$ as given in (28). It is straightforward to verify that the gradient and Hessian of $\widehat{L}(\lambda)$ are

$$\begin{aligned}
\nabla \hat{L}(\lambda) &= n^{-1} \sum_{i=1}^{n} \left[Y_i \mathbf{B}(\mathbf{X}_i) - b'\left\{\lambda^T \mathbf{B}(\mathbf{X}_i)\right\} \mathbf{B}(\mathbf{X}_i)\right], \\
\nabla^2 \hat{L}(\lambda) &= -n^{-1} \sum_{i=1}^{n} b''\left\{\lambda^T \mathbf{B}(\mathbf{X}_i)\right\} \mathbf{B}(\mathbf{X}_i)\, \mathbf{B}(\mathbf{X}_i)^T.
\end{aligned} \tag{29}$$

□

Proposition 1 *Under Assumptions (A1)–(A5) and (A7), for $m \in M$ with M given in (20) and $\widehat{m}$ as in (6), as $n \to \infty$, $\|m - \widehat{m}\|_{2,n} + \|m - \widehat{m}\|_2 = \mathcal{O}_{a.s.}\left(N^{1/2} n^{-1/2} \log n\right)$ and $\|m - \widehat{m}\|_\infty = \mathcal{O}_{a.s.}\left(N n^{-1/2} \log n\right)$. With probability approaching 1, the Hessian matrix $\nabla^2 \hat{L}(\lambda)$ satisfies that $\nabla^2 \hat{L}(\lambda) < \mathbf{0}$, $\forall \lambda$ and $\nabla^2 \hat{L}(\lambda) \le -c_b c_V \mathbf{I}$ if $\lambda^T \mathbf{B}(\mathbf{X}_i) \in \Theta$, $1 \le i \le n$.*

Proof See Lemma A.13 of Liu et al. (2013), Assumption (A2), Eq. (29) and Lemma A.11 of Liu et al. (2013). □

7.3 Proof of Theorem 1

Define a stochastic process $\widehat{\varepsilon}_n(x_1) = n^{-1}\sum_{i=1}^n K_h(X_{i1} - x_1)\xi_i$, $x_1 \in [0,1]$ with ξ_i given in (26), then (9) and (10) show that

$$\sup_{x_1\in[h,1-h]}\left|\widetilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1) - D_1^{-1}(x_1)\widehat{\varepsilon}_n(x_1)\right| = \mathcal{O}_{a.s.}\left(h^2 + n^{-1/2}h^{1/2}\log n\right),$$

which, together with (8), lead to

$$\begin{aligned}&\sup\nolimits_{x_1\in[h,1-h]}\left|\widehat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) - D_1^{-1}(x_1)\widehat{\varepsilon}_n(x_1)\right|\\ &\quad = \mathcal{O}_{a.s.}\left(h^2 + n^{-1/2}h^{1/2}\log n + n^{-1/2}\log n\right) = \mathcal{O}_{a.s.}\left(h^2 + n^{-1/2}\log n\right).\end{aligned} \tag{30}$$

Using $v_1(x_1)$ given in (11), one can standardize $\widehat{\varepsilon}_n(x_1)$ to obtain

$$\begin{aligned}\widehat{\zeta}_n(x_1) &= (nh)^{1/2}v_1^{-1}(x_1)\widehat{\varepsilon}_n(x_1)\\ &= (nh)^{1/2}v_1^{-1}(x_1)\left\{n^{-1}\sum_{i=1}^n K_h(X_{i1} - x_1)\xi_i\right\}.\end{aligned} \tag{31}$$

Assumptions (A5), (A8) imply that the following Rosenblatt transformation to the two-dimensional sequence $\{X_{i1}, \xi_i\}_{i=1}^n$ produces $\left\{X'_{i1}, \xi'_i\right\}_{i=1}^n$ with $\left(X'_{i1}, \xi'_i\right)$ uniformly distributed on $[0,1]^2$:

$$\left(X'_{i1}, \xi'_i\right) = T(X_{i1}, \xi_i) = \left\{F_{X_1}(X_{i1}), F_{\xi|X_1}(\xi_i|X_{i1})\right\}.$$

Denote $Z_n(x_1, \xi) = \sqrt{n}\{F_n(x_1, \xi) - F(x_1, \xi)\}$ where $F_n(x_1, \xi)$ is the empirical distribution of $\{X_{i1}, \xi_i\}_{i=1}^n$, one can rewrite $\widehat{\zeta}_n(x_1)$ as

$$\widehat{\zeta}_n(x_1) = h^{-1/2}v_1^{-1}(x_1)\int\int K\{(u - x_1)/h\}\xi dZ_n(u, \xi).$$

By the strong approximation theorem in Tusnady (1977), there exists a version of the two-dimensional Brownian Bridge $B_n\left(x'_1, \xi'\right)$ such that

$$\sup_{x_1,\xi}|Z_n(x_1, \xi) - B_n\{T(x_1, \xi)\}| = \mathcal{O}_{a.s.}\left(n^{-1/2}\log^2 n\right).$$

Applying standard techniques used in Bickel and Rosenblatt (1973), Härdle (1989), one can show that

$$\sup_{t\in[h,1-h]}\left|\widehat{\zeta}_n(t) - M_h(t)/\|K\|_2^2\right| = o_p\left\{(\log n)^{-1/2}\right\}, \tag{32}$$

for a version of the $M_h(t)$ given in (27). Similar result can be found in Xia (1998). Furthermore, (30) and (31) imply that

$$\begin{aligned}&\sup_{x_1\in[h,1-h]}\left|\sigma_n^{-1}(x_1)\left\{\widehat{m}_{\mathrm{SBK},1}(x_1)-m_1(x_1)\right\}-\widehat{\zeta}_n(x_1)\right|\\&\quad=\mathcal{O}_{a.s.}\left(n^{1/2}h^{5/2}+h^{1/2}\log n\right),\end{aligned}\tag{33}$$

with $\sigma_n(x)$ given in (14). Under Assumption (A6), which entails that $(-2\log h)^{1/2}$ is of the same order as $(\log n)^{1/2}$, (32) and (33) can show that

$$\begin{aligned}&\sup_{x_1\in[h,1-h]}(-2\log h)^{1/2}\left|\sigma_n^{-1}(x_1)\left|\widehat{m}_{\mathrm{SBK},1}(x_1)-m_1(x_1)\right|-|M_h(x_1)|/\|K\|_2^2\right|\\&=\mathcal{O}_{a.s}\left\{(\log n)^{1/2}\times\left(n^{1/2}h^{5/2}+h^{1/2}\log n\right)\right\}+o_p(1)=o_p(1).\end{aligned}$$

Finally, Theorem 1 follows from Lemma 1 and Slutsky's Theorem.

7.4 Proof of Theorem 2

See the Supplement.

References

Berg D (2007) Bankruptcy prediction by generalized additive models. Appl Stoch Models Bus Ind 23:129–143

Bernhardsen E (2001) A model of bankruptcy prediction. Norges Bank, WP

Bickel PJ, Rosenblatt M (1973) On some global measures of the deviations of density function estimates. Ann Stat 1:1071–1095

Cai L, Yang L (2015) A smooth simultaneous confidence band for conditional variance function. TEST 24:632–655

Engelmann B, Hayden E, Tasche D (2003) Testing rating accuracy. Risk 16:82–86

Fan J, Yao Q (2003) Nonlinear Time Series: Nonparametric and Parametric Methods. Springer-Verlag, Berlin

Fan J, Zhang WY (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. Scand J Stat 27:715–731

Gu L, Wang L, Härdle W, Yang L (2014) A simultaneous confidence corridor for varying coefficient regression with sparse functional data. TEST 23:806–843

Gu L, Yang L (2015) Oracally efficient estimation for single-index link function with simultaneous confidence band. Electr J Stat 9:1540–1561

Härdle W (1989) Asymptotic maximal deviation of M-smoothers. J Multivariate Anal 29:163–179

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman and Hall, London

He X, Fung W, Zhu Z (2005) Robust estimation in generalized partial linear models for clustered data. J Am Stat Assoc 100:1176–1184

He X, Zhu Z, Fung W (2002) Estimation in a semiparamtric model for longitudinal data with unspecified dependence structure. Biometrika 89:579–590

Horowitz J, Mammen E (2004) Nonparametric estimation of an additive model with a link function. Ann Stat 32:2412–2443

Huang JZ, Yang L (2004) Identification of nonlinear additive autoregression models. J R Stat Soc Ser B Stat Methodol 66:463–477

Linton OB (1997) Efficient estimation of additive nonparametric regression models. Biometrika 84:469–473

Linton OB, Härdle W (1996) Estimation of additive regression models with known links. Biometrika 83:529–540

Liu R, Yang L (2010) Spline-backfitted kernel smoothing of additive coefficient model. Econom Theory 26:29–59

Liu R, Yang L, Härdle W (2013) Oracally efficient two-step estimation of generalized additive model. J Am Stat Assoc 108:619–631

Ma S, Yang L (2011) Spline-backfitted kernel smoothing of partially linear additive model. J Stat Plan Inference 141:204–219

Ma S, Yang L, Carroll RJ (2012) Simultaneous confidence band for sparse longitudinal regression. Stat Sin 22:95–122

Ryser M, Denzler S (2009) Selecting credit rating models: a cross-validation-based comparison of discriminatory power. Financ Mark Portf Manag 23:187–203

Severini T, Staniswalis J (1994) Quasi-likelihood estimation in semiparametric models. J Am Stat Assoc 89:501–511

Shina Y, Moore W (2003) Explaining credit rating differences between Japanese and U.S. agencies. Rev Finan Econ 12:327–344

Stone CJ (1985) Additive regression and other nonparametric models. Ann Statist 13:689–705

Stone CJ (1986) The dimensionality reduction principle for generalized additive models. Ann Statist 14:590–606

Tusnady G (1977) A remark on the approximation of the sample distribution function in the multidimensional case. Period Math Hungar 8:53–55

Wang J, Liu R, Cheng F, Yang L (2014) Oracally efficient estimation of autoregressive error distribution with simultaneous confidence band. Ann Stat 42:654–668

Wang L, Yang L (2007) Spline-backfitted kernel smoothing of nonlinear additive autoregression model. Ann Stat 35:2474–2503

Wang L, Yang L (2009) Spline estimation of single index model. Stat Sin 19:765–783

Wang L, Li H, Huang J (2008) Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. J Am Stat Assoc 103:1556–1569

Wiesenfarth M, Krivobokova T, Klasen S, Sperlich S (2012) Direct Simultaneous Inference in Additive Models and its Application to Model Undernutrition. J Am Stat Assoc 107:1286–1296

Wu W, Zhao Z (2007) Inference of trends in time series. J R Stat Soc Ser B Stat Methodol 69:391–410

Xia Y (1998) Bias-corrected confidence bands in nonparametric regression. J R Stat Soc Ser B Stat Methodol 60:797–811

Xue L, Yang L (2006) Additive coefficient modeling via polynomial spline. Stat Sin 16:1423–1446

Yang L, Sperlich S, Härdle W (2003) Derivative estimation and testing in generalized additive models. J Stat Plan Inference 115:521–542

Zhang H, Lin Y (2006) Component selection and smoothing for nonparametric regression in exponential families. Stat Sin 16:1021–1042

Zhao Z, Wu W (2008) Confidence bands in nonparametric time series regression. Ann Stat 36:1854–1878

Zheng S, Yang L, Härdle W (2014) A smooth simultaneous confidence corridor for the mean of sparse functional data. J Am Stat Assoc 109:661–673