SFB 649 Discussion Paper 2014-008

Simultaneous Confidence Corridors and Variable Selection for Generalized Additive Models

Shuzhuan Zheng*
Rong Liu**
Lijian Yang*
Wolfgang Karl Härdle***

* Soochow University, China and Michigan State University, USA
** University of Toledo, USA
*** Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Simultaneous Confidence Corridors and Variable Selection for Generalized Additive Models

Shuzhuan Zheng
Center for Advanced Statistics and Econometrics Research
Soochow University
Suzhou 215006, China
and
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824 email: zheng@stt.msu.edu

Rong Liu
Department of Mathematics and Statistics
University of Toledo
Toledo, OH 43606 email: rong.liu@utoledo.edu

Lijian Yang
Center for Advanced Statistics and Econometrics Research
Soochow University
Suzhou 215006, China email: yanglijian@suda.edu.cn
and
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824 email: yang@stt.msu.edu

Wolfgang K. Härdle
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin, Germany email: haerdle@wiwi.hu-berlin.de
and
Lee Kong Chian School of Business, Singapore Management University

Author's Footnote:

Shuzhuan Zheng is Visiting Scholar, Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou 215006, China, and Ph.D. student, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824 (E-mail: zheng@stt.msu.edu). Rong Liu is Assistant Professor, Department of Mathematics and Statistics, University of Toledo, Toledo, OH 43606 (E-mail: rong.liu@utoledo.edu). Lijian Yang is Director, Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou 215006, China, and Professor, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824 (E-mail: yanglijian@suda.edu.cn; yang@stt.msu.edu). Wolfgang K. Härdle is Professor, C.A.S.E. – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, and Distinguished Visiting Professor, Lee Kong Chian School of Business, Singapore Management University (E-mail: haerdle@wiwi.hu-berlin.de). This work is supported in part by the Deutsche Forschungsgemeinschaft through the CRC 649 "Economic Risk", NSF Awards DMS 0706518, DMS 1007594, an Michigan State University Dissertation Continuation Fellowship, funding from the National University of Singapore, the Jiangsu Specially-Appointed Professor Program SR10700111, the Jiangsu Key Discipline Program (Statistics) ZY107002, National Natural Science Foundation of China award 11371272, and Research Fund for the Doctoral Program of Higher Education of China award 20133201110002.

Abstract

In spite of the widespread use of generalized additive models (GAMs), there is no well established methodology for simultaneous inference and variable selection for the components of GAM. There is no doubt that both, inference on the marginal component functions and their selection, are essential in this additive statistical models. To this end, we establish simultaneous confidence corridors (SCCs) and a variable selection criteria through the spline-backfitted kernel smoothing techniques. To characterize the global features of each component, SCCs are constructed for testing their shapes. By extending the BIC to additive models with identity/trivial link, an asymptotically consistent BIC approach for variable selection is proposed. Our procedures are examined in simulations for its theoretical accuracy and performance, and used to forecast the default probability of listed Japanese companies.

Keywords: BIC; Confidence corridor; Extreme value; Generalized additive model; Spline-backfitted kernel.

JEL Classification: C35; C52; C53; G33.

1 INTRODUCTION

The 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 $\left\{\left(\mathbf{X}_i^{\mathsf{T}}, Y_i\right)\right\}_{i=1}^n$ consists of iid copies of $\left(\mathbf{X}^{\mathsf{T}}, Y\right)$ that satisfy:

$$\mathsf{E}(Y|\mathbf{X}) = b'\left\{m\left(\mathbf{X}\right)\right\}, \mathsf{Var}(Y|\mathbf{X}) = a\left(\phi\right) b''\left\{m\left(\mathbf{X}\right)\right\}, m\left(\mathbf{X}\right) = c + \sum\nolimits_{\alpha=1}^{d} m_\alpha(X_\alpha), \tag{1}$$

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

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

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:

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

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

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

For identifiability, one needs

$$\mathsf{E}\left\{m_\alpha\left(X_\alpha\right)\right\} = 0, 1 \leq \alpha \leq d \tag{2}$$

that leads to unique additive representations of $m\left(\mathbf{x}\right) = c + \sum_{\alpha=1}^{d} m_\alpha\left(x_\alpha\right)$. 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 $(b')^{-1}(x)$ $= \log\{x/(1-x)\}$. Our method has been applied to $3,472$ 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 Figure 1 (a)-(c). More details of this example are contained in Section 6.

[Figure 1 about here.]

The smooth functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^{d}$ in (1) can be estimated by, for instance, kernel methods in Linton and Härdle (1996), Linton (1997) and Yang, Sperlich and Härdle (2003), B-spline methods in Stone (1986) and Xue and Liang (2010), and two-stage methods in Horowitz and Mammen (2004) and Horowitz et al. (2006). To make statistical inference on these functions individually and collectively, however, the proper tools are simultaneous confidence corridors (SCCs) and consistent variable selection criteria.

The SCC methodology has attracted attention in a variety of applied fields, see Xia (1998), Fan and Zhang (2000), Wu and Zhao (2007), Zhao and Wu (2008), Ma, Yang and Carroll (2012) among others. Capturing shape properties of the functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^{d}$ is of utmost importance. A smooth component covered entirely within SCC can be replaced by a parametric one, thereby improving the estimation efficiency, see He, Zhu and Fung (2002), He, Fung and Zhu (2005) for discussions. To our knowledge, SCCs have not been established due to a technological lack of estimators that fit in Gaussian process extreme value theory. Using the spline-backfitted kernel (SBK) methodology of Liu, Yang and Härdle (2013) (hereafter LYH), we extend work of univariate nonparametric regression in Bickel and Rosenblatt (1973) and Härdle (1989) to those of GAM. The SBK technique has been studied in Wang and Yang (2007), Wang and Yang (2009), Liu and Yang (2010) and Ma and Yang (2011) for the simpler additive model (i.e., GAM with $b'(x) \equiv x$) including the construction of SCC, but ours is the first work on SCC for GAM with nonlinear link.

While variable selection for nonparametric additive model has been investigated under different settings, see Wang, Li and Huang (2008), there is lack of theoretically reliable variable selection for GAM. To the best of our knowledge, only Zhang and Lin (2006) proposed the "COSSO" method for variable selection in nonparametric regression with exponential families, but without asymptotic theory. 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). This extension is challenging since a much more complicated quasi-likelihood is employed with nonlinear link instead of the log mean squared error for trivial link. The appendix gives more details.

The rest of paper is organized as follows. The SBK estimator and its oracle property are briefly described in Section 2. Asymptotic extreme value distribution of the SBK estimator is investigated in Section 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 Section 5. Section 6 illustrates the application of our SCC 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 SBK estimator for GAM (1) and its oracle properties obtained in LYH. Let $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ be i.i.d. observations following model (1). Without loss of generality, one denotes $\mathbf{x}_{\text{-}1} = (x_2, ..., x_d)$ and $m_{\text{-}1}(\mathbf{x}_{\text{-}1}) = c + \sum_{\alpha=2}^d m_\alpha(x_\alpha)$ and estimates $m_1(x_1)$.

As a benchmark of efficiency, we introduce the "oracle smoother" by treating the constant c and the last $d-1$ components $\{m_\alpha(x_\alpha)\}_{\alpha=2}^d$ as known. The only unknown component $m_1(x_1)$ is estimated by maximizing a local log-likelihood function $\widetilde{l}(a, x_1)$ for each $x_1 \in [h, 1-h]$:

$$\widetilde{l}(a, x_1) = n^{-1} \sum\nolimits_{i=1}^n [Y_i \{a + m_{\text{-}1}(\mathbf{X}_{i,\text{-}1})\} - b\{a + m_{\text{-}1}(\mathbf{X}_{i,\text{-}1})\}] K_h(X_{i1} - x_1), \tag{3}$$

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

$$\widetilde{m}_{\text{K},1}(x_1) = \underset{a \in A}{\text{argmax}}\, \widetilde{l}(a, x_1). \tag{4}$$

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

To define the SBK, we introduce the linear B spline basis for smoothing: $b_J(x) = (1 - |x - \xi_J| / H)_+$, $0 \le J \le N+1$ 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 α-empirically centered linear spline functions on $[0,1]$ is

$$G_{n,\alpha}^{0}=\left\{g_{\alpha}:g_{\alpha}\left(x_{\alpha}\right)=\sum\nolimits_{J=0}^{N+1}\lambda_{J}b_{J}\left(x_{\alpha}\right),\mathsf{E}_{n}\left\{g_{\alpha}\left(X_{\alpha}\right)\right\}=0\right\},1\leq\alpha\leq d, \tag{5}$$

with empirical expectation $\mathsf{E}_{n}\left\{g_{\alpha}\left(X_{\alpha}\right)\right\}=n^{-1}\sum_{i=1}^{n}g_{\alpha}\left(X_{\alpha i}\right)$. The space of additive spline functions on $\chi=[0,1]^{d}$ is

$$G_{n}^{0}=\left\{g\left(\mathbf{x}\right)=c+\sum\nolimits_{\alpha=1}^{d}g_{\alpha}\left(x_{\alpha}\right);c\in\mathbb{R},\ g_{\alpha}\in G_{n,\alpha}^{0}\right\}. \tag{6}$$

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

$$\widehat{L}\left(g\right)=n^{-1}\sum\nolimits_{i=1}^{n}\left[Y_{i}g\left(\mathbf{X}_{i}\right)-b\left\{g\left(\mathbf{X}_{i}\right)\right\}\right],g\in G_{n}^{0}. \tag{7}$$

According to Lemma 14 of Stone (1986), (7) has a unique maximizer with probability approaching 1. Therefore, the multivariate function $m\left(\mathbf{x}\right)$ can be estimated by an additive spline function:

$$\widehat{m}\left(\mathbf{x}\right)=\underset{g\in G_{n}^{0}}{\operatorname{argmax}}\,\widehat{L}\left(g\right). \tag{8}$$

The spline estimator is asymptotically consistent, and can be calculated efficiently. However, no measure of confidence can be assigned to the spline estimator, see Wang and Yang (2007) and LYH. To overcome this problem, we adapt the SBK estimator, which combines the strength of kernel smoothing with regression spline. One then rewrites $\widehat{m}\left(\mathbf{x}\right)=\hat{c}+\sum_{\alpha=1}^{d}\widehat{m}_{\alpha}\left(X_{i\alpha}\right)$ for $\widehat{c}\in\mathbb{R}$ and $\widehat{m}_{\alpha}\left(x_{\alpha}\right)\in G_{n,\alpha}^{0}$ and defines a univariate quasi-likelihood function similar to $\widetilde{l}\left(a,x_{1}\right)$ in (3) as

$$\widehat{l}\left(a,x_{1}\right)=n^{-1}\sum\nolimits_{i=1}^{n}\left[Y_{i}\left\{a+\widehat{m}_{\text{-}1}\left(\mathbf{X}_{i,\text{-}1}\right)\right\}-b\left\{a+\widehat{m}_{\text{-}1}\left(\mathbf{X}_{i,\text{-}1}\right)\right\}\right]K_{h}\left(X_{i1}-x_{1}\right), \tag{9}$$

with $\widehat{m}_{\text{-}1}\left(\mathbf{x}_{\text{-}1}\right)=\widehat{c}+\sum_{\alpha=2}^{d}\widehat{m}_{\alpha}\left(x_{\alpha}\right)$ being the pilot spline estimator of $m_{\text{-}1}\left(\mathbf{x}_{\text{-}1}\right)$. Consequently, the SBK estimator of $m_{1}\left(x_{1}\right)$ is

$$\widehat{m}_{\mathrm{SBK},1}\left(x_{1}\right)=\underset{a\in A}{\operatorname{argmax}}\,\widehat{l}\left(a,x_{1}\right). \tag{10}$$

We now introduce some useful results and definitions from LYH, under Assumptions (A1)-(A7) in appendix, as $n\to\infty$,

$$\sup_{x_{1}\in[0,1]}\left|\widehat{m}_{\mathrm{SBK},1}\left(x_{1}\right)-\widetilde{m}_{\mathrm{K},1}\left(x_{1}\right)\right|=\mathcal{O}_{a.s.}\left(n^{-1/2}\log n\right), \tag{11}$$

$$\widetilde{m}_{\mathrm{K},1}(x_1)-m_1(x_1)=\mathrm{bias}_1(x_1)\,h^2/D_1(x_1)+n^{-1}\sum\nolimits_{i=1}^{n}K_h(X_{i1}-x_1)\,\sigma(X_i)\,\varepsilon_i/D_1(x_1)+r_{K,1}(x_1) \quad (12)$$

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

$$\sup_{x_1\in[h,1-h]}|r_{\mathrm{K},1}(x_1)|=\mathcal{O}_{a.s.}\left(n^{-1/2}h^{1/2}\log n\right). \quad (13)$$

The scale function $D_1(x_1)$ and bias function $\mathrm{bias}_1(x_1)$ are defined in LYH as:

$$\sigma_b^2(x_1)=\mathsf{E}\left[b''\{m(\mathbf{X})\}\,|X_1=x_1\right],\ \sigma^2(x_1)=\mathsf{E}\left\{\sigma^2(\mathbf{X})\,|X_1=x_1\right\} \quad (14)$$

$$D_1(x_1)=f_1(x_1)\,\sigma_b^2(x_1),v_1^2(x_1)=\|K\|_2^2\,f_1(x_1)\,\sigma^2(x_1). \quad (15)$$

$$\begin{gathered}\mathrm{bias}_1(x_1)=\mu_2(K)\times \quad (16)\\ \left\{m_1''(x_1)\,D_1(x_1)+m_1'(x_1)\,f(x_1)\,\sigma_b^2(x_1)'-\{m_1'(x_1)\}^2 f(x_1)\,\mathsf{E}\left[b'''\{m(\mathbf{X})\}\,|X_1=x_1\right]\right\}\end{gathered}$$

where $\|K\|_2^2=\int K^2(u)\,du$, $\mu_2(K)=\int K(u)\,u^2du$. The above equations (11), (12) and (13) lead one to a simplifying decomposition of the estimation error $\widehat{m}_{\mathrm{SBK},1}(x_1)-m_1(x_1)$

$$\begin{gathered}\sup_{x_1\in[h,1-h]}\left|\widehat{m}_{\mathrm{SBK},1}(x_1)-m_1(x_1)-n^{-1}\sum\nolimits_{i=1}^{n}K_h(X_{i1}-x_1)\,\sigma(\mathbf{X}_i)\,\varepsilon_i/D_1(x_1)\right| \quad (17)\\ =\mathcal{O}_{a.s.}\left(n^{-1/2}h^{1/2}\log n+n^{-1/2}\log n+h^2\right).\end{gathered}$$

A decomposition such as (17) has not appeared in the literature for any other estimators of $m_1(x_1)$, and it is fundamental for constructing SCCs in section 3.

3 GAM INFERENCE VIA SCC

In this section, we propose SCCs for GAM components.

3.1 Main Results

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

$$Q_h(\alpha)=a_h+a_h^{-1}\left[\log\left\{\sqrt{C(K)}/(2\pi)\right\}-\log\left\{-\log\sqrt{1-\alpha}\right\}.\right] \quad (18)$$

Also with $D_1(x_1)$ and $v_1^2(x_1)$ given in (15), we define

$$\sigma_n(x_1) = n^{-1/2} h^{-1/2} v_1(x_1) D_1^{-1}(x_1). \tag{19}$$

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

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

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

$$\widehat{m}_{\mathrm{SBK},1}(x_1) \pm \sigma_n(x_1) Q_h(\alpha). \tag{20}$$

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 high dimensional nonparametric regression.

3.2 Implementation

To construct the SCC for $m_1(x_1)$ in (20), one needs to select the bandwidth h first, and then evaluate $m_{\mathrm{SBK},1}(x_1)$, $Q_h(\alpha)$ and $\sigma_n(x_1)$ given in (10), (18) and (19).

Assumption (A6) requires that the bandwidth for SCC be slightly smaller than the mean square optimal bandwidth h_{opt} (minimizing AMISE) in LYH, we therefore have taken $h = h_{\mathrm{opt}}(\log n)^{-1/4}$ as a data-driven undersmoothing bandwidth for SCC construction to fulfill Assumption (A6). Recent articles on SCC for time series, such as Wu and Zhao (2007), Zhao and Wu (2008), have used similar undersmoothing bandwidths.

For a given α and a chosen bandwidth h, one can easily estimate $m_{\mathrm{SBK},1}(x_1)$ and $Q_h(\alpha)$ as in (10), (18). To evaluate $\sigma_n(x_1)$, one needs to estimate $v_1(x_1)$ and $D_1^{-1}(x_1)$ given in (15), i.e., estimating $f(x_1), \sigma_b^2(x_1)$ and $\sigma^2(x_1)$. The estimation of the density function $f(x_1)$ is trivial, namely, $\widehat{f}(x_1) = n^{-1}\sum_{i=1}^n K_h(X_{i1} - x_1)$. 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., $N = N_n = \min\left(\left[n^{1/4}\log n\right] + 1, [n/4d - 1/d] - 1\right)$ which satisfies the assumption (A7) in the Appendix. Then $\sigma_b^2(x_1)$ can be estimated as $\sum_{k=0}^{3} \widehat{a}_{1,k}^{k} x_1^k + \sum_{k=4}^{N+3} \widehat{a}_{1,k}(x_1 - t_{\alpha,k-3})^3$ where $\{\widehat{a}_{1,k}\}_{k=0}^{N+3}$

minimize

$$\sum_{i=1}^{n} \left[b'' \left\{ \widehat{m} \left(\mathbf{X}_i \right) \right\} - \left\{ \sum\nolimits_{k=0}^{3} a_{1,k} X_{i1}^{k} + \sum\nolimits_{k=4}^{N+3} a_{1,k} \left(X_{i1} - t_{k-3} \right)^3 \right\} \right]^2, \tag{21}$$

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

$$\sum_{i=1}^{n} \left[\left[Y_i - b' \left\{ \widehat{m} \left(\mathbf{X}_i \right) \right\} \right]^2 - \left\{ \sum_{k=0}^{3} a_{\alpha,k} X_{i1}^{k} + \sum_{k=4}^{N+3} a_{\alpha,k} \left(X_{i1} - t_{k-3} \right)^3 \right\} \right]^2. \tag{22}$$

The resulted estimate $\hat{\sigma}_n \left(x_1 \right)$ of $\sigma_n \left(x_1 \right)$, using (21) and (22) satisfies $\sup_{x_1 \in [h,1-h]} \left| \hat{\sigma}_n \left(x_1 \right) - \sigma_n \left(x_1 \right) \right| = \mathcal{O}_p \left(n^{-\gamma} \right)$ for some $\gamma > 0$, see LYH Section 5 for details. This consistency and Slutzky's theorem ensure that $\mathrm{P} \left\{ \sup_{x_1 \in [h,1-h]} \left| \widehat{m}_{\mathrm{SBK},1} \left(x_1 \right) - m_1 \left(x_1 \right) \right| / \hat{\sigma}_n \left(x_1 \right) \leq Q_h \left(\alpha \right) \right\} \to 1 - \alpha$ as $n \to \infty$, and therefore

$$\widehat{m}_{\mathrm{SBK},1} \left(x_1 \right) \pm \hat{\sigma}_n \left(x_1 \right) Q_h \left(\alpha \right) \tag{23}$$

is a $100 \left(1 - \alpha \right) \%$ simultaneous confidence corridor for $m_1 \left(x_1 \right)$. The SCC constructions of other components $m_2 \left(x_2 \right), ..., m_d \left(x_d \right)$ are similar. It is worthwhile to emphasize that, in general, the estimators $\widehat{m}_{\mathrm{SBK},1} \left(x_1 \right)$, $\widehat{Q}_h \left(\alpha \right), \widehat{f} \left(x_1 \right)$ and $\hat{\sigma}_n \left(x_1 \right)$ remain stable if h slightly varies.

4 VARIABLE SELECTION IN GAM

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

4.1 Main Results

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

$$\mathcal{H}^0 = \left\{ g : \mathsf{E} \left\{ g \left(X_\alpha \right) \right\} = 0, \mathsf{E} \left\{ g^2 \left(X_\alpha \right) \right\} < \infty, 1 \leq \alpha \leq d \right\}, \tag{24}$$

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

$$\mathcal{M} = \left\{ g \left(\mathbf{x} \right) = c + \sum\nolimits_{\alpha=1}^{d} g_\alpha \left(\mathbf{x}_\alpha \right); g_\alpha \in \mathcal{H}^0, 1 \leq \alpha \leq d \right\}, \tag{25}$$

where c is a finite constant.

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

$$\mathcal{M}_S=\left\{g\left(\mathbf{x}\right)=c+\sum\nolimits_{\alpha\in S}g_\alpha\left(\mathbf{x}_\alpha\right);c\in\mathbb{R},g_\alpha\in\mathcal{H}^0,\alpha\in S\right\}, \tag{26}$$

with $\mathcal{H}^0$ given in (24), and the space of the additive spline functions as

$$G^0_{n,S}=\left\{g\left(\mathbf{x}\right)=c+\sum\nolimits_{\alpha\in S}g_\alpha\left(x_\alpha\right);c\in\mathbb{R},g_\alpha\in G^0_{n,\alpha},\alpha\in S\right\}, \tag{27}$$

with $G^0_{n,\alpha}$ given in (5). Define the least squares projection of function m in $\mathcal{M}_S$ as

$$m_S=\operatorname*{argmin}_{g\in\mathcal{M}_S}\mathsf{E}\left\{m\left(\mathbf{X}\right)-g\left(\mathbf{X}\right)\right\}^2 \tag{28}$$

and define the set S_0 of significant variables as the minimal set $S\subset\{1,\ldots,d\}$ such that $\mathsf{E}\{m\left(\mathbf{X}\right)-m_S\left(\mathbf{X}\right)\}^2=0$, which is uniquely defined according to Lemma 1 of Huang and Yang (2004).

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

$$\mathrm{BIC}_S=-2\widehat{L}\left(\widehat{m}_S\right)+\frac{N_S}{n}\left(\log n\right)^3 \tag{29}$$

where $\widehat{L}\left(\cdot\right)$ is given in (7), $\widehat{m}_S\left(\mathbf{x}\right)\in G^0_{n,S}$ is the pilot spline estimator as in (8), $N_S=1+\left(N+1\right)\#\left(S\right)$ with N the number of interior knots, $\#\left(S\right)$ the cardinality of S. In practice, $N=N_n$ can be taken as

$$\min\left(\left[n^{1/4}\log n\right]+1,\left[n/4d-1/d\right]-1\right), \tag{30}$$

which satisfies the assumption (A7) in the Appendix.

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

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

According to Theorem 2, the variable selection rule based on the BIC in (29) 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

The proposed BIC is implemented without a greedy search through all possible subsets. Instead, the forward stepwise regression procedure is used with minimizing BIC as the criterion.

5 MONTE CARLO SIMULATION

This section studies the performance of the proposed procedures, reporting also the computational costs, the consistency of selecting variables via BIC and the global coverage precision of the SCC. The data are generated from

$$\mathrm{P}(Y=1|\mathbf{X}=\mathbf{x}) = b'\left\{c+\sum\nolimits_{\alpha=1}^{d} m_\alpha\left(X_\alpha\right)\right\}, b'\left(x\right)=\frac{e^x}{1+e^x} \tag{31}$$

with $d=10, c=0, m_3\left(x\right)=m_4\left(x\right)=m_5\left(x\right)=\sin\left(\pi x\right),\ m_6\left(x\right)=x, m_7\left(x\right)=e^x-\left(e-e^{-1}\right)$ and $m_\alpha\left(x\right)=0$ for $\alpha=1,2,8,9,10$. The set S_0 is therefore $S_0=\{3,4,5,6,7\}$.

The predictors are generated by

$$X_{i\alpha}=2\Phi\left(Z_{i\alpha}\right)-1,\ \mathbf{Z}_i=\left(Z_{i1},...,Z_{id}\right)\sim\mathrm{N}\left(0,\Sigma\right), 1\leq i\leq n, 1\leq\alpha\leq d, \tag{32}$$

where Φ is the standard normal c.d.f. and $\Sigma=\left(1-r\right)\mathbf{I}_{d\times d}+r\mathbf{1}_d\mathbf{1}_d^T$. The parameter r $(0\leq r<1)$ controls the correlation between $Z_{i\alpha}$,$1\leq\alpha\leq d$.

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 SCC global coverage for components.

[Table 1 about here.]

[Table 2 about here.]

Table 1 shows the simulation results from 100 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 $m_\alpha\left(X_\alpha\right), \alpha=3,...,7$, is quite satisfactory. The selection accuracy becomes higher as the sample size increases and the correlation decreases. The accuracy and computing time of COSSO are also listed for comparison (Platform: R; PC: Intel 3.4 GHz processor and 16 GB RAM). It is obvious that the BIC significantly outperforms the COSSO in terms of both accuracy and computing time. To examine the computing advantage of BIC for large d, we extend $d = 10$ to 50 by using $m_3, ..., m_7$ as above and all the other component functions are 0. The BIC is vastly faster than COSSO for $d = 50$. All of these findings confirm what is expected according to the asymptotic theory.

The SCC global coverage for $m_\alpha(x_\alpha), \alpha = 3, ..., 7$ is reported in Table 2. It turns out that the empirical coverage approaches the nominal confidence levels as n increases, and a better coverage occurs when the correlation is lower.

The above studies evidently indicate the reliability of our methodology, such as a high selection accuracy of the BIC and a desired global coverage of the SCC, which make their applications for credit rating modelling in the following section sensible.

6 APPLICATION

We now return to forecast default probabilities of the listed companies in Japan. The data provided by 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 $3,583$ 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 (29), we have found that X_1: Current liabilities, X_7: 3 months earlier return, X_8: 6 months earlier return are selected. 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 1 (a)-(c) depicts the SBK estimator of the factor's default impact curve on domain, while a shoal of 95% CIs and the 95% SCC 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 used. For any credit scoring method S, one defines its alarm rate $F(s) = \mathrm{P}(S \leq s)$ and the hit rate $F_{\mathrm{D}}(s) = \mathrm{P}(S \leq s \,|\mathrm{D})$ where D represents the conditioning event of "default". One then defines the CAP curve for S as

$$\mathrm{CAP}(u) = F_{\mathrm{D}}\left\{F^{-1}(u)\right\}, u \in (0,1), \tag{33}$$

which is the percentage of default-infected obligators that are found among the first (according to their scores) $100u\%$ of all obligators. The perfect scoring method P assigns low scores first to all default-infected obligators and thus $\mathrm{CAP}_{\mathrm{P}}(u) = \min(u/p, 1), u \in (0,1)$ where p is the unconditional default probability, whereas the completely noninformative scoring 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, Hayden and Tasche (2003).

A satisfactory scoring method's CAP curve would be expected to approach $\mathrm{CAP}_{\mathrm{P}}(u)$ and always better than the noinformative $\mathrm{CAP}_{\mathrm{N}}(u)$, and one uses the Accuracy Ratio (AR) to quantify its position. The AR is the ratio of the area a_R enclosed between the given CAP curve and the noninformative diagonal curve $\mathrm{CAP}_{\mathrm{N}}(u) \equiv u$, and the total area a_P enclosed between the perfect CAP curve $\mathrm{CAP}_{\mathrm{P}}(u)$ and the noninformative diagonal curve $\mathrm{CAP}_{\mathrm{N}}(u)$. Thus

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

where $\mathrm{CAP}(u)$ is given in (33). 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.

We have applied both GAM and GLM to the first 2000 companies and computed default probabilities of the remaining 1583 companies, and used the default probabilities as scores. Figure 1 (d) depicts the CAPs of GAM (thick solid) and GLM (thin solid), as well as the $\mathrm{CAP}_{\mathrm{P}}(u)$ (dashed) and $\mathrm{CAP}_{\mathrm{N}}(u)$ (dotted). Visually it is clear that GAM has much higher accuracy ratio than GLM, which is in fact the case: the AR is 97.56% for GAM, much higher than the 89.76% for GLM. Nonetheless, we failed to apply the COSSO for the same data.

APPENDIX

In what follows, we take $\|\cdot\|$ and $\|\cdot\|_\infty$ as the Euclidean and supremum norms, respectively, i.e., for any $\mathbf{x} = (x_1, x_2, ..., x_d)^\mathsf{T} \in \mathbb{R}^d, \|\mathbf{x}\| = \left(\sum_{\alpha=1}^{d} x_\alpha^2\right)^{1/2}$ and $\|\mathbf{x}\|_\infty = \max_{1\leq\alpha\leq d}|x_\alpha|$. For any interval $[a, b]$, denote the space of p-th order smooth function by $C^{(p)}[a, b] = \left\{g\left|g^{(p)} \in C\left[a, b\right]\right.\right\}$, and the class of Lipschitz continuous functions by $\text{Lip}\left(\left[a, b\right], C\right) = \left\{g \left|\left|g\left(x\right) - g\left(x'\right)\right| \leq C\left|x - x'\right|, \forall x, x' \in \left[a, b\right]\right.\right\}$ for constant $C > 0$. Lastly, define the following latent regression errors

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

We need the following technical assumptions:

(A1) *The additive component functions* $m_\alpha \in C^{(1)}\left[0, 1\right], 1 \leq \alpha \leq d$: $m_1 \in C^{(2)}\left[0, 1\right], m'_\alpha \in \text{Lip}\left(\left[0, 1\right], C_m\right)$, $2 \leq \alpha \leq d$ *for some constant* $C_m > 0$.

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

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

(A4) *The joint density* $f\left(\mathbf{x}\right)$ *of* $\left(X_1, ..., X_d\right)$ *is continuous:* $0 < c_f \leq \inf_{\mathbf{x}\in[0,1]^d} f\left(\mathbf{x}\right) \leq \sup_{\mathbf{x}\in[0,1]^d} f\left(\mathbf{x}\right) \leq C_f < \infty$. *The marginal density function* $f_\alpha\left(x_\alpha\right)$ *of* X_α *have continuous derivatives on* $[0, 1]$ *and the 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* $f_{Y|\mathbf{X}}\left(y; b'\left\{m\left(\mathbf{x}\right)\right\}\right)$ *relative to* λ *whose support for* y *is a common* Ω, *and is uniformly continuous in* $\mathbf{x} \in \left[0, 1\right]^d$ *for* $y \in \Omega$.

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

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

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

Assumptions (A1)-(A7) are standard in GAM, see Stone (1986), Xue and Liang (2010). Assumptions (A5), (A6) are more restrictive than in LYH for the purpose of constructing SCCs, but are unnecessary for Theorem 2 on the consistency of BIC.

A.1. Preliminaries

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

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

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

$$d_h = \left(-2\log h\right)^{1/2} + \left(-2\log h\right)^{-1/2}\left\{\sqrt{C\left(K\right)}/\left(2\pi\right)\right\} \tag{A.3}$$

with $C\left(K\right)$ given in (18).

Lemma A.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 wrt 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 wrt the Lebesgue measure $\mu^{(2)}$, which one denotes as $f_{T,X_1}\left(t, x_1\right)$.

Lemma A.2 *Under Assumptions (A1)-(A5), for ξ_i in (A.1), the distribution of $\left(\xi_i, X_{i1}\right)$ has joint pdf wrt $\mu^{(2)}$ 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 therefore $f_{Y|\mathbf{X}}\left(y; t\right) f_{T,X_1}\left(t, x_1\right)$. For any $\left(z, x_1\right) \in \mathbb{R} \times [0,1]$, and $\triangle z, \triangle x_1 > 0$, one has

$$\mathrm{P}\left[\left(\xi_i, X_{i1}\right) \in \left(z - \triangle z, z + \triangle z\right) \times \left(x_1 - \triangle x_1, x_1 + \triangle x_1\right)\right] =$$

$$\mathrm{P}\left[\left(Y_i - T_i, X_{i1}\right) \in \left(z - \triangle z, z + \triangle z\right) \times \left(x_1 - \triangle x_1, x_1 + \triangle x_1\right)\right] =$$

$$= \int_{\Omega} d\lambda\left(y\right) \int_{y-\tau \in (z-\triangle z, z+\triangle z)} d\tau \int_{\chi_1 \in (x_1 - \triangle x_1, x_1 + \triangle x_1)} f_{Y|\mathbf{X}}\left(y;\tau\right) f_{T,X_1}\left(\tau, \chi_1\right) d\chi_1.$$

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

$$\left|\mathrm{P}\left[\left(\xi_i, X_{i1}\right) \in \left(z - \triangle z, z + \triangle z\right) \times \left(x_1 - \triangle x_1, x_1 + \triangle x_1\right)\right] - \left\{\int_{\Omega} f_{Y|\mathbf{X}}\left(y; y - z\right) f_{T,X_1}\left(y - z, x_1\right) d\lambda\left(y\right)\right\}\right.$$
$$\left. \times \mu^{(2)}\left[\left(z - \triangle z, z + \triangle z\right) \times \left\{\left(x_1 - \triangle x_1, x_1 + \triangle x_1\right) \cap [0,1]\right\}\right]\right| = \scriptstyle{\mathcal{O}}(1)$$

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

For theoretical analysis, we write $c_{J,\alpha} = \mathsf{E}\, b_J\left(X_\alpha\right) = \int b_J\left(x_\alpha\right) f_\alpha\left(x_\alpha\right) dx_\alpha$ and define the centered B spline basis $b_{J,\alpha}\left(x_\alpha\right)$ and the standardized B spline basis $B_{J,\alpha}\left(x_\alpha\right)$ respectively as

$$\begin{aligned} b_{J,\alpha}\left(x_\alpha\right) &= b_J\left(x_\alpha\right) - \frac{c_{J,\alpha}}{c_{J-1,\alpha}} b_{J-1}\left(x_\alpha\right), \\ B_{J,\alpha}\left(x_\alpha\right) &= \frac{b_{J,\alpha}\left(x_\alpha\right)}{\left\{\int b_{J,\alpha}^2\left(x_\alpha\right) f_\alpha\left(x_\alpha\right) dx_\alpha\right\}^{1/2}}, 1 \le J \le N + 1, \end{aligned} \tag{A.4}$$

so that $\mathsf{E}\, B_{J,\alpha}\left(X_\alpha\right) \equiv 0$, $\mathsf{E}\, B_{J,\alpha}^2\left(X_\alpha\right) \equiv 1$.

One can rewrite with slight abuse of notations the log-likelihood $\widehat{L}\left(g\right)$ given in (7) as

$$\widehat{L}\left(g\right) = \widehat{L}\left(\boldsymbol{\lambda}\right) = n^{-1} \sum\nolimits_{i=1}^{n} \left[Y_i \boldsymbol{\lambda}^{\mathsf{T}} \mathbf{B}\left(\mathbf{X}_i\right) - b\left\{\boldsymbol{\lambda}^{\mathsf{T}} \mathbf{B}\left(\mathbf{X}_i\right)\right\}\right], \tag{A.5}$$

with $g\left(\mathbf{X}_i\right) = \boldsymbol{\lambda}^{\mathsf{T}} \mathbf{B}\left(\mathbf{X}_i\right) \in G_n^0$, $\boldsymbol{\lambda} = \left(\lambda_0, \lambda_{J,\alpha}\right)_{1 \le J \le N+1, 1 \le \alpha \le d}^{\mathsf{T}} \in \mathbb{R}^{N_d}$ with $N_d = \left(N + 1\right) d + 1$, $\mathbf{B}\left(\mathbf{x}\right) = \left\{1, B_{1,1}\left(x_1\right), ..., B_{N+1,d}\left(x_d\right)\right\}^{\mathsf{T}}$ and $B_{J,\alpha}\left(x_\alpha\right)$ as given in (A.4). It is straightforward to verify that the gradient and Hessian of $\widehat{L}\left(\boldsymbol{\lambda}\right)$ are

$$\begin{aligned} \nabla \hat{L}\left(\boldsymbol{\lambda}\right) &= n^{-1} \sum\nolimits_{i=1}^{n} \left[Y_i \mathbf{B}\left(\mathbf{X}_i\right) - b'\left\{\boldsymbol{\lambda}^{\mathsf{T}} \mathbf{B}\left(\mathbf{X}_i\right)\right\} \mathbf{B}\left(\mathbf{X}_i\right)\right], \\ \nabla^2 \hat{L}\left(\boldsymbol{\lambda}\right) &= -n^{-1} \sum\nolimits_{i=1}^{n} b''\left\{\boldsymbol{\lambda}^{\mathsf{T}} \mathbf{B}\left(\mathbf{X}_i\right)\right\} \mathbf{B}\left(\mathbf{X}_i\right) \mathbf{B}\left(\mathbf{X}_i\right)^{\mathsf{T}}. \end{aligned} \tag{A.6}$$

Proposition A.1 *Under Assumptions (A1)-(A5) and (A7), for $m \in M$ with M given in (25) and $\widehat{m}$ as in (8), as $n \to \infty$, $\left\|m - \widehat{m}\right\|_{2,n} + \left\|m - \widehat{m}\right\|_2 = \mathcal{O}_{a.s.}\left(N^{1/2} n^{-1/2} \log n\right)$ and $\left\|m - \widehat{m}\right\|_\infty = \mathcal{O}_{a.s.}\left(N n^{-1/2} \log n\right)$. With probability approaching 1, the Hessian matrix $\nabla^2 \hat{L}\left(\boldsymbol{\lambda}\right)$ satisfies that $\nabla^2 \hat{L}\left(\boldsymbol{\lambda}\right) < \mathbf{0}, \forall \boldsymbol{\lambda}$ and $\nabla^2 \hat{L}\left(\boldsymbol{\lambda}\right) \le -c_b c_V \mathbf{I}$ if $\boldsymbol{\lambda}^{\mathsf{T}} \mathbf{B}\left(\mathbf{X}_i\right) \in \Theta, 1 \le i \le n$.*

PROOF. See Lemma A.13 of LYH, Assumption (A2), equation (A.6) and Lemma A.11 of LYH.

A.2. Proof of Theorem 1

PROOF. 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 (A.1), then (12) and (13) 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 (11), lead to

$$\begin{aligned}&\sup_{x_1\in[h,1-h]} \left|\widehat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) - D_1^{-1}(x_1)\widehat{\varepsilon}_n(x_1)\right| \\ &= \mathcal{O}_{a.s.}\left(h^2 + n^{-1/2}h^{1/2}\log n + n^{-1/2}\log n\right) = \mathcal{O}_{a.s.}\left(h^2 + n^{-1/2}\log n\right).\end{aligned} \tag{A.7}$$

Using $v_1(x_1)$ given in (15), one can standardize $\widehat{\varepsilon}_n(x_1)$ and then replace x_1 by t to obtain

$$\widehat{\zeta}_n(t) = (nh)^{1/2} v_1^{-1}(t)\widehat{\varepsilon}_n(t) = (nh)^{1/2} v_1^{-1}(t)\left\{n^{-1}\sum_{i=1}^n K_h(X_{i1} - t)\xi_i\right\}. \tag{A.8}$$

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

$$\left(X_{i1}', \xi_i'\right) = T(X_{1i}, \xi_i) = \left\{F_{X_1}(X_{1i}), F_{\xi|X_1}(\xi_i|X_{1i})\right\}.$$

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 $\widehat{\xi}_n(t)$ as

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

By the strong approximation theorem in Tusnady (1977), there exists a version of the 2-dimensional Brownian Bridge $B_n\left(x_1', \xi'\right)$ such that

$$\sup_{x_1,s}|Z_n(x_1,\xi) - B_n\{T(x_1,\xi)\}| = \mathcal{O}_{a.s.}\left(n^{-1/2}\overset{2}{\log} n\right). \tag{A.9}$$

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

$$\sup_{t\in[0,1]}\left|\widehat{\xi}_n(t)-M_h(t)/\|K\|_2^2\right|=\mathcal{o}_p\left\{(\log n)^{-1/2}\right\}, \tag{A.10}$$

for a version of the $M_h(t)$ given in (A.2). Similar result can be found in Xia (1998).

Furthermore, (A.7) and (A.8) imply that

$$\sup_{x_1\in[h,1-h]}\left|\sigma_n^{-1}(x_1)\left\{\widehat{m}_{\mathrm{SBK},1}(x_1)-m_1(x_1)\right\}-\widehat{\xi}_n(x_1)\right|=\mathcal{O}_{a.s.}\left(n^{1/2}h^{5/2}+h^{1/2}\log n\right), \tag{A.11}$$

with $\sigma_n(x)$ given in (19). Under Assumption (A6), which entails that $(-2\log h)^{1/2}$ is of the same order as $(\log n)^{1/2}$, (A.10) and (A.11) can show that

$$\begin{aligned}&\sup_{x_1\in[h,1-h]}(-2\log h)^{1/2}\left|\sigma_n^{-1}(x_1)\left|\widehat{m}_{\mathrm{SBK},1}(x_1)-m_1(x_1)\right|-\left|M_h(t)\right|/\|K\|_2^2\right| \\ &=\mathcal{O}_{a.s}\left\{(\log n)^{1/2}\times\left(n^{1/2}h^{5/2}+h^{1/2}\log n\right)\right\}+\mathcal{o}_p(1)=\mathcal{o}_p(1).\end{aligned} \tag{A.12}$$

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

A.3. Proof of Theorem 2

Prior to proving Theorem 2, we restate Proposition A.1 for any index set $S\subset\{1,2,\ldots d\}$.

Denote by $\boldsymbol{\lambda}=(\lambda_0,\lambda_{J,\alpha})^{\mathsf{T}}_{1\le J\le N+1,1\le\alpha\le d}$ an arbitrary vector. For any S, $N_S=1+(N+1)\#(S)$, denote

$$\boldsymbol{\lambda}_S=(\lambda_0,\lambda_{J,\alpha})^{\mathsf{T}}_{1\le J\le N+1,\alpha\in S}\in\mathbb{R}^{N_S},\mathbf{B}_S(\mathbf{x})=\left\{1,B_{J,\alpha}(x_\alpha)\right\}^{\mathsf{T}}_{1\le J\le N+1,\alpha\in S}, \tag{A.13}$$

and with slight abuse of notations

$$\widehat{L}_S(\boldsymbol{\lambda}_S)=\widehat{L}_S\left\{\boldsymbol{\lambda}_S^{\mathsf{T}}\mathbf{B}_S(\mathbf{x})\right\}=n^{-1}\sum\nolimits_{i=1}^{n}\left[Y_i\boldsymbol{\lambda}_S^{\mathsf{T}}\mathbf{B}_S(\mathbf{X}_i)-b\left\{\boldsymbol{\lambda}_S^{\mathsf{T}}\mathbf{B}_S(\mathbf{X}_i)\right\}\right] \tag{A.14}$$

whose maximizer is $\widehat{m}_S=\widehat{\boldsymbol{\lambda}}_S^{\mathsf{T}}\mathbf{B}_S(\mathbf{x})$.

PROPOSITION A.2 *Under Assumptions (A1)-(A5) and (A7), for $m_S\in M_S$ given in (28), $\widehat{m}_S$ in (A.14), as $n\to\infty$, $\|m_S-\widehat{m}_S\|_{2,n}+\|m_S-\widehat{m}_S\|_2=\mathcal{O}_{a.s.}\left(N^{1/2}n^{-1/2}\log n\right)$ and $\|m_S-\widehat{m}_S\|_\infty=\mathcal{O}_{a.s.}\left(Nn^{-1/2}\log n\right)$.*

Next, we consider two cases "underfitting" and "overfitting" for the index set S to establish Theorem 2.

Definition: if $S \supset S_0$ and $S \neq S_0$, then S overfits, while S is underfitting if $S_0 \cap S \neq S_0$ with S_0 given in Theorem 2. We shall show that $\lim_{n\to\infty} \mathrm{P}\left(\mathrm{BIC}_S - \mathrm{BIC}_{S_0} > 0\right) = 1$ in both situations.

Proof. **I: overfitting, i.e.,** $S \supset S_0$ **and** $S \neq S_0$.

Let $\boldsymbol{\lambda}_{S_0S} = \left\{\boldsymbol{\lambda}_{S_0}, \left(\lambda_{J,\alpha'}\right)^{\mathsf{T}}_{1\leq J\leq N+1, \alpha' \in S\setminus S_0}\right\}$, $\widehat{\boldsymbol{\lambda}}_{S_0S} = \left\{\widehat{\boldsymbol{\lambda}}_{S_0}, \left(\lambda_{J,\alpha'}\right)^{\mathsf{T}}_{1\leq J\leq N+1, \alpha' \in S\setminus S_0}\right\}$ with $\lambda_{J,\alpha'} \equiv 0$ and $\widehat{\boldsymbol{\lambda}}_S$ (or $\widehat{\boldsymbol{\lambda}}_{S_0}$) as the MLE of (A.14) (or when $S = S_0$). Note that $\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_{S_0S}\right) = \widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right)$.

Using Taylor's expansion, $\exists$ a vector $\widetilde{\boldsymbol{\lambda}}_S$ between $\widehat{\lambda}_S$ and $\widehat{\boldsymbol{\lambda}}_{S_0S}$, i.e., $\widetilde{\boldsymbol{\lambda}}_S = \mathbf{t}\widehat{\boldsymbol{\lambda}}_S + \left(\mathbf{I}_{Ns} - \mathbf{t}\right)\widehat{\boldsymbol{\lambda}}_{S_0S}$ with a $N_s \times N_s$ diagonal matrix $\mathbf{t}$ whose diagonal elements are in $[0,1]$ s.t.

$$\begin{aligned} \widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) &= \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_{S_0S}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) \\ = \left(\widehat{\boldsymbol{\lambda}}_{S_0S} - \widehat{\boldsymbol{\lambda}}_S\right)^{\mathsf{T}} \triangledown \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) &+ \frac{1}{2}\left(\widehat{\boldsymbol{\lambda}}_{S_0S} - \widehat{\boldsymbol{\lambda}}_S\right)^{\mathsf{T}} \triangledown^2 \widehat{L}_S\left(\widetilde{\boldsymbol{\lambda}}_S\right)\left(\widehat{\boldsymbol{\lambda}}_{S_0S} - \widehat{\boldsymbol{\lambda}}_S\right). \end{aligned} \tag{A.15}$$

Since $\triangledown \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) = 0$ and $\triangledown^2 \widehat{L}_S\left(\widetilde{\boldsymbol{\lambda}}_S\right)$ is given in (A.6), for $\widetilde{m}_S = \widetilde{\boldsymbol{\lambda}}_S^{\mathsf{T}}\mathbf{B}_S\left(\mathbf{x}\right)$, one has

$$\begin{aligned} &\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) \\ = -\left(2n\right)^{-1}\sum\nolimits_{i=1}^{n} b''&\left\{\widetilde{\boldsymbol{\lambda}}_S^{\mathsf{T}}\mathbf{B}_S\left(\mathbf{X}_i\right)\right\}\left\{\left(\widehat{\boldsymbol{\lambda}}_{S_0S} - \widehat{\boldsymbol{\lambda}}_S\right)^{\mathsf{T}}\mathbf{B}_S\left(\mathbf{X}_i\right)\right\}\left\{\left(\widehat{\boldsymbol{\lambda}}_{S_0S} - \widehat{\boldsymbol{\lambda}}_S\right)^{\mathsf{T}}\mathbf{B}_S\left(\mathbf{X}_i\right)\right\}^{\mathsf{T}} \\ &= -\left(2n\right)^{-1}\sum\nolimits_{i=1}^{n} b''\left\{\widetilde{m}_S\left(\mathbf{X}_i\right)\right\}\left\{\widehat{m}_{S_0}\left(\mathbf{X}_i\right) - \widehat{m}_S\left(\mathbf{X}_i\right)\right\}^2. \end{aligned} \tag{A.16}$$

By $m = m_{S_0} = m_S \in M_{S_0} \subset M_S$, Proposition A.2 implies that $\left\|\widehat{m}_{S_0} - m\right\|_{2,n} = \mathcal{O}_{a.s.}\left(N^{1/2}n^{-1/2}\log n\right)$ and $\left\|\widehat{m}_S - m\right\|_{2,n} = \mathcal{O}_{a.s.}\left(N^{1/2}n^{-1/2}\log n\right)$, thus

$$\left\|\widehat{m}_{S_0} - \widehat{m}_S\right\|_{2,n}^2 = \mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right). \tag{A.17}$$

Similarily, one has $\left\|\widetilde{m}_S - m\right\|_\infty = o_{a.s.}\left(1\right)$, which warrants for large n that $\widetilde{m}_S \in \Theta$ with Θ given in Assumption (A.2), so (A.16) implies that

$$0 \geq \widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right) \geq -\frac{C_b}{2}\left\|\widehat{m}_{S_0} - \widehat{m}_S\right\|_{2,n}^2. \tag{A.18}$$

As a result, BIC given in (29) shows that

$$\begin{aligned} \mathrm{BIC}_S - \mathrm{BIC}_{S_0} &= 2\left\{\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)\right\} + \frac{N_S - N_{S_0}}{n}\log^3 n \\ &\geq -C_b\left\|\widehat{m}_{S_0} - \widehat{m}_S\right\|_{2,n}^2 + \left(N+1\right)n^{-1}\log^3 n, \end{aligned} \tag{A.19}$$

which implies by (A.17) that $\lim_{n\to\infty} \mathrm{P}\left(\mathrm{BIC}_S - \mathrm{BIC}_{S_0} > 0\right) = 1$.

II: underfitting, i.e., $S_0 \cap S \neq S_0$.

Let $S' = S_0 \cup S$ and denote by $\widehat{\boldsymbol{\lambda}}_{S_0}$, $\widehat{\boldsymbol{\lambda}}_S$ and $\widehat{\boldsymbol{\lambda}}_{S'}$ the MLEs in (A.14) for S_0, S and S', respectively. Since S' overfits S_0, similarly to (A.18), one has

$$0 \geq \widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \geq -\frac{C_b}{2}\left\|\widehat{m}_{S_0} - \widehat{m}_{S'}\right\|_{2,n}^2, \tag{A.20}$$

Define a set $\Theta_{S'} = \left\{\boldsymbol{\lambda}_{S'} : \boldsymbol{\lambda}_{S'}^{\mathsf{T}}\mathbf{B}_{S'}\left(\mathbf{X}_i\right) \in \Theta, 1 \leq i \leq n\right\}$. which is compact and convex in $\mathbb{R}^{N_S}$. By definition,

$$\max_{1\leq i\leq n}\left|\widehat{\boldsymbol{\lambda}}_{S'}^{\mathsf{T}}\mathbf{B}_{S'}\left(\mathbf{X}_i\right) - m\left(\mathbf{X}_i\right)\right| \leq \left\|\widehat{m}_{S'} - m\right\|_\infty = \mathcal{O}_{a.s.}\left(Nn^{-1/2}\log n\right),$$

so for large n, with probability approaching 1, $\widehat{\boldsymbol{\lambda}}_{S'} \in \Theta_{S'}$, so Proposition A.1 ensures that, with probability approaching 1, the Hessian matrix $\nabla^2\widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \leq -c_b c_V \mathbf{I}_{N_{S'}}$, while $\nabla\widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) = 0$ and $\nabla^2\widehat{L}_{S'}\left(\boldsymbol{\lambda}_{S'}\right) \leq \mathbf{0}$, $\forall\ \boldsymbol{\lambda}_{S'}$. Thus, with probability approaching 1, there exists a constant $c_1 > 0$ such that

$$\widehat{L}_{S'}\left(\boldsymbol{\lambda}_{S'}\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \leq \left\{\begin{array}{cc} -2^{-1}c_b c_V\left\|\boldsymbol{\lambda}_{S'} - \widehat{\boldsymbol{\lambda}}_{S'}\right\|^2, & \text{if } \boldsymbol{\lambda}_{S'} \in \Theta_{S'} \\ \max\limits_{\boldsymbol{\lambda}_{S'}\in\partial\Theta_{S'}}\widehat{L}_S\left(\boldsymbol{\lambda}_S\right) - \widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \leq -c_1, & \text{otherwise}\end{array}\right. . \tag{A.21}$$

Next, define a new vector $\widehat{\boldsymbol{\lambda}}_{SS'} = \left\{\widehat{\boldsymbol{\lambda}}_S, \left(\lambda_{J,\alpha'}\right)_{1\leq J\leq N+1,\alpha'\in S'\setminus S}^{\mathsf{T}}\right\}$ with $\lambda_{J,\alpha'} \equiv 0$ and note that $\widehat{\boldsymbol{\lambda}}_{SS'}^{\mathsf{T}}$ $\mathbf{B}_{S'}\left(\mathbf{x}\right) \equiv \widehat{m}_S\left(\mathbf{x}\right), \widehat{\boldsymbol{\lambda}}_{S'}^{\mathsf{T}}\ \mathbf{B}_{S'}\left(\mathbf{x}\right) \equiv \widehat{m}_{S'}\left(\mathbf{x}\right)$, so applying Lemma A.5 of Wang and Yang (2007), there exists a constant $C_0 > 0$ such that

$$\left\|\widehat{\boldsymbol{\lambda}}_{SS'} - \widehat{\boldsymbol{\lambda}}_{S'}\right\|^2 \geq C_0^{-1}\left\|\widehat{m}_S - \widehat{m}_{S'}\right\|_2^2.$$

Applying Proposition A.2 entails that

$$\left|\left\|\widehat{m}_S - \widehat{m}_{S'}\right\|_2^2 - \left\|m_S - m_{S'}\right\|_2^2\right| = \mathcal{O}_{a.s.}\left(Nn^{-1}\overset{2}{\log} n\right)$$

while the definitions of underfitting and overfitting lead to

$$\left\|m_S - m_{S'}\right\|_2^2 = \left\|m_S - m\right\|_2^2 = c_S > 0$$

and thus

$$\left\|\widehat{\boldsymbol{\lambda}}_{SS'} - \widehat{\boldsymbol{\lambda}}_{S'}\right\|^2 \geq C_0^{-1}c_S + \mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right). \tag{A.22}$$

Note next that

$$\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)-\widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right)=\widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{SS'}\right)-\widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right) \tag{A.23}$$

which according to (A.21), is bounded by

$$\leq\begin{cases} -2^{-1}c_b c_V\left\|\widehat{\boldsymbol{\lambda}}_{SS'}-\widehat{\boldsymbol{\lambda}}_{S'}\right\|^2, & \text{if } \widehat{\boldsymbol{\lambda}}_{SS'}\in\Theta_{S'} \\ \max\limits_{\boldsymbol{\lambda}_{S'}\in\partial\Theta_{S'}}\widehat{L}_S\left(\boldsymbol{\lambda}_S\right)-\widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right)\leq -c_1, & \text{otherwise}\end{cases}$$

which is, according to (A.22), bounded by

$$\begin{aligned} &\leq \max\left(-2^{-1}c_b c_V C_0^{-1}c_S,-c_1\right)+\mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right) \\ &= -c_2+\mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right),\ \text{for a constant } c_2>0.\end{aligned}$$

The above bound, together with (A.17), (A.20) and (A.23) lead to $\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right)-\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)$

$$\begin{aligned} &= \left\{\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right)-\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_{S'}\right)\right\}-\left\{\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)-\widehat{L}_{S'}\left(\widehat{\boldsymbol{\lambda}}_{S'}\right)\right\} \\ &\geq c_2+\mathcal{O}_{a.s.}\left(Nn^{-1}\log^2 n\right). \end{aligned} \tag{A.24}$$

Finally, (A.24) implies that

$$\mathrm{BIC}_S-\mathrm{BIC}_{S_0}=2\left\{\widehat{L}_{S_0}\left(\widehat{\boldsymbol{\lambda}}_{S_0}\right)-\widehat{L}_S\left(\widehat{\boldsymbol{\lambda}}_S\right)\right\}+\frac{N_S-N_{S_0}}{n}\log^3 n\geq c_2+o_p(1), \tag{A.25}$$

and thus $\lim_{n\to\infty}\mathrm{P}\left(\mathrm{BIC}_S-\mathrm{BIC}_{S_0}>0\right)=1$.

References

[1] Berg, D. (2007), "Bankruptcy Prediction by Generalized Additive Models," *Applied Stochastic Models in Business and Industry*, 23, 129–143.

[2] Bernhardsen, E. (2001), *A Model of Bankruptcy Prediction*, Norges Bank WP, 2001.

[3] Bickel, P.J. and Rosenblatt, M. (1973), "On Some Global Measures of the Deviations of Density Function Estimates," *Annals of Statistics*, 1, 1071–1095.

[4] Engelmann, B., Hayden, E. and D. Tasche. (2003), "Testing Rating Accuracy," *Risk*, 16, 82–86.

[5] Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.* 26, 943–971.

[6] Fan, J. and Zhang, W. Y. (2000), "Simultaneous Confidence Bands and Hypothesis Testing in Varying-coefficient Models," *Scandinavian Journal of Statistics*, 27, 715–731.

[7] Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis* 29, 163–179.

[8] Härdle, W., Hoffmann, L. and Moro, R. (2011), *Learning Machines Supporting Bankruptcy prediction. Statistical Tools in Finance and Insurance* (2nd ed.), Cizek, Härdle, Weron, Springer Verlag.

[9] Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

[10] He, X., Fung, W. and Zhu, Z. (2005), "Robust Estimation in Generalized Partial Linear Models for Clustered Data," *Journal of the American Statistical Association*, 100, 1176–1184.

[11] He, X., Zhu, Z and Fung, W. (2002), "Estimation in A Semiparamtric Model for Longitudinal Data with Unspecified Dependence Structure," *Biometrika*, 89, 579–590.

[12] Horowitz, J. and Mammen, E. (2004), "Nonparametric Estimation of An Additive Model with A Link Function," *Annals of Statistics*, 32, 2412–2443.

[13] Horowitz, J., Klemelä, J. and Mammen, E. (2006), "Optimal Estimation in Additive Regression," *Bernoulli*, 12, 271–298.

[14] Huang, J. Z. and Yang, L. (2004), "Identification of Nonlinear Additive Autoregression Models," *Journal of the Royal Statistical Society: Series B*, 66, 463–477.

[15] Linton, O. B. (1997), "Efficient Estimation of Additive Nonparametric Regression Models," *Biometrika*, 84, 469–473.

[16] Linton, O. B. and Härdle, W. (1996), "Estimation of Additive Regression Models with Known Links," *Biometrika*, 83, 529–540.

[17] Liu, R. and Yang, L. (2010), "Spline-backfitted Kernel Smoothing of Additive Coefficient Model," *Econometric Theory*, 26, 29–59.

[18] Liu, R., Yang, L and Härdle, W. (2013), "Oracally Efficient Two-Step Estimation of Generalized Additive Model," *Journal of the American Statistical Association*, 108, 619–631.

[19] Ma, S. and Yang, L. (2011), "Spline-backfitted Kernel Smoothing of Partially Linear Additive model," *Journal of Statistical Planning and Inference*, 141, 204–219.

[20] Ma, S., Yang, L. and Carroll, R. J. (2012), " Simultaneous Confidence Band for Sparse Longitudinal Regression," *Statistica Sinica*, 22, 95–122.

[21] Ryser, M. and Denzler, S. (2009), "Selecting Credit Rating Models: A Cross-validation-based Comparison of Discriminatory Power," *Financ Mark Portf Manag*, 23, 187–203.

[22] Shina, Y. and Moore, W. (2003), "Explaining Credit Rating Differences between Japanese and U.S. Agencies," *Review of Financial Economics*, 12, 327–344.

[23] Song, Q. and Yang, L. (2010), "Oracally Efficient Spline Smoothing of Nonlinear Additive Autoregression Model with Simultaneous Confidence Band," *Journal of Multivariate Analysis*, 101, 2008–2025.

[24] Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *Annals of Statistics*, 13, 689–705.

[25] Stone, C. J. (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *Annals of Statistics*, 14, 590–606

[26] Tusnady, G. (1977), "A Remark on the Approximation of the Sample Distribution Function in the Multidimensional Case," *Period. Math. Hungar.*, 8, 53–55.

[27] Wang, L. and Yang, L. (2007), "Spline-backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model," *Annals of Statistics*, 35, 2474–2503.

[28] Wang, L., Li, H., and Huang, J. (2008), "Variable Selection in Nonparametric Varying-coefficient Models for Analysis of Repeated Measurements," *Journal of the American Statistical Association*, 103, 1556–1569.

[29] Wu, W. and Zhao, Z. (2007), "Inference of Trends in Time Series," *Journal of the Royal Statistical Society: Series B*, 69, 391–410.

[30] Xia, Y. (1998), "Bias-corrected Confidence Bands in Nonparametric Regression," *Journal of the Royal Statistical Society: Series B*, 60, 797–811.

[31] Xue, L. and Liang, H. (2010), “Polynomial Spline Estimation for A Generalized Additive Coefficient Model,” *Scandinavian Journal of Statistics*, 37, 26–46.

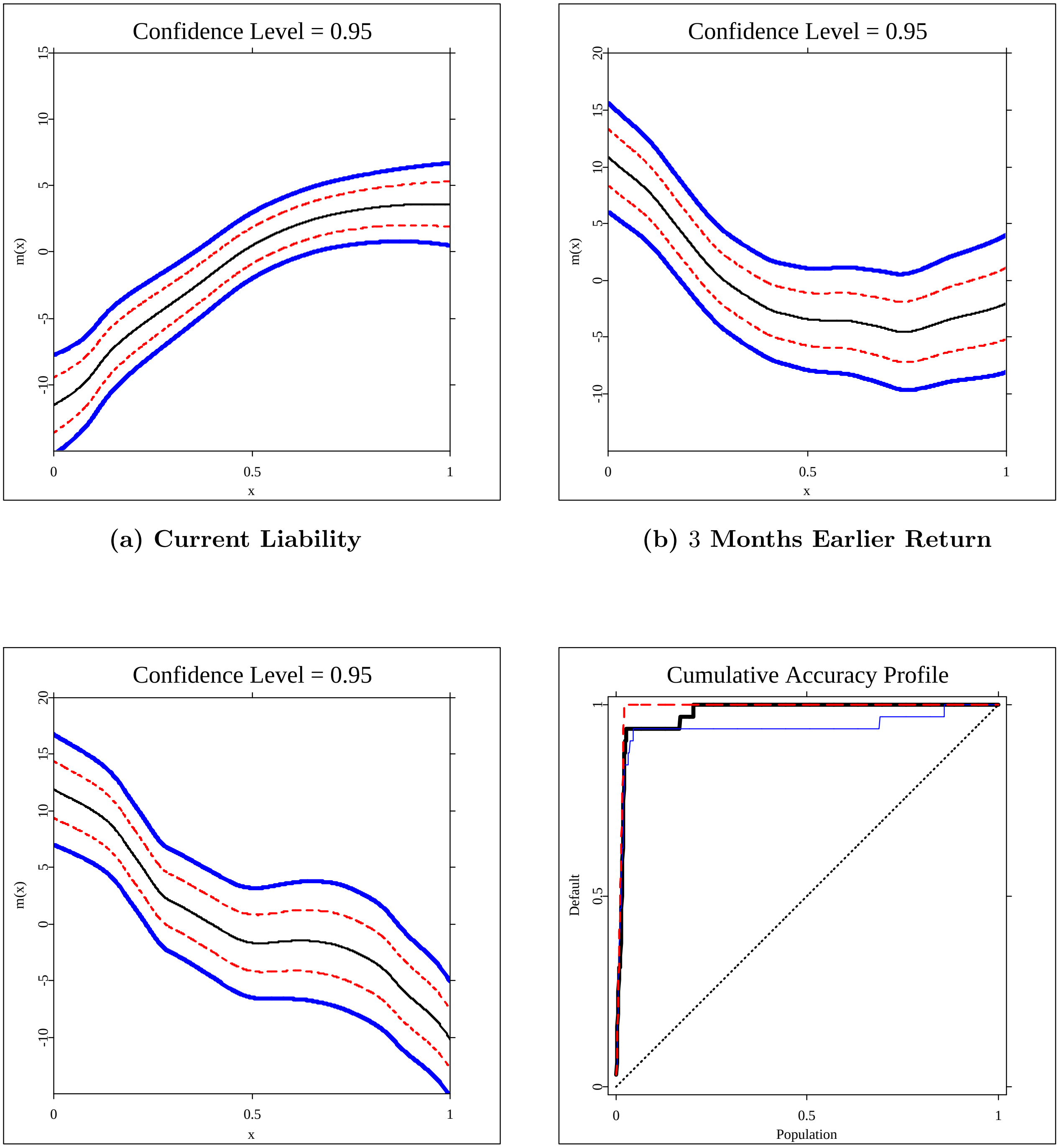
[32] Xue, L. and Yang, L. (2006), “Additive Coefficient Modeling via Polynomial Spline,” *Statistica Sinica*, 16, 1423–1446.

[33] Yang, L., Sperlich, S. and Härdle, W. (2003), “Derivative Estimation and Testing in Generalized Additive Models,” *Journal of Statistical Planning and Inference*, 115, 521–542.

[34] Zhang, H. and Lin, Y. (2006), “Component Selection and Smoothing for Nonparametric Regression in Exponential Families,” *Statistica Sinica*, 16, 1021–1042.

[35] Zhao, Z. and Wu, W. (2008), “Confidence Bands in Nonparametric Time Series Regression,” *Annals of Statistics*, 36, 1854–1878.

Figure 1: Plots of the rating factors in (a)-(c): SBK estimators (thin), 95% CIs (dashed) and 95% SCCs (thick). Plot of the CAPs in (d): Perfect (dashed), GAM (thick solid), GLM(thin solid), noninformative(dotted).



(a) **Current Liability**

(b) 3 **Months Earlier Return**

(c) 6 **Months Earlier Return**

(d) **The CAP Curves**

			Computing Time			Accuracy					
d	r	n	BIC	COSSO	Ratio	BIC			COSSO		
10	0	250	0.43	5.90	13.7	4	90	6	20	66	14
		500	0.58	11.50	19.8	0	95	5	7	85	8
		1000	1.05	27.89	26.6	0	99	1	6	90	4
	0.5	250	0.51	6.12	12.0	31	62	7	42	44	14
		500	0.64	12.90	20.1	3	91	6	17	73	10
		1000	1.12	29.43	26.3	0	99	1	12	82	6
50	0	250	1.89	–	–	87	10	3	–	–	–
		500	2.76	209.34	76	14	77	9	21	46	33
		1000	5.14	531.96	103	0	96	4	1	89	10
	0.5	250	2.02	–	–	90	4	6	–	–	–
		500	2.87	215.64	75	63	32	5	44	26	30
		1000	5.39	545.43	101	10	86	4	10	71	19

Table 1: Simulation results for the proposed BIC method and COSSO with $d = 10$ and 50. For each setup, the first, second, and third columns under Accuracy give respectively the frequencies of under fitting, correct fitting, and over fitting over 100 replications. 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.

r	n	α				
		3	4	5	6	7
0.0	250	0.925	0.920	0.920	0.930	0.925
	500	0.955	0.945	0.945	0.950	0.955
	1000	0.950	0.945	0.950	0.945	0.950
0.5	250	0.910	0.910	0.915	0.920	0.920
	500	0.935	0.925	0.930	0.930	0.935
	1000	0.945	0.945	0.940	0.950	0.945

Table 2: The 95% SCC global coverage for $m_a\left(x\right)$, $\alpha = 3, ..., 7$ from 200 replications

SFB 649 Discussion Paper Series 2014

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Principal Component Analysis in an Asymmetric Norm" by Ngoc Mai Tran, Maria Osipenko and Wolfgang Karl Härdle, January 2014.

002 "A Simultaneous Confidence Corridor for Varying Coefficient Regression with Sparse Functional Data" by Lijie Gu, Li Wang, Wolfgang Karl Härdle and Lijian Yang, January 2014.

003 "An Extended Single Index Model with Missing Response at Random" by Qihua Wang, Tao Zhang, Wolfgang Karl Härdle, January 2014.

004 "Structural Vector Autoregressive Analysis in a Data Rich Environment: A Survey" by Helmut Lütkepohl, January 2014.

005 "Functional stable limit theorems for efficient spectral covolatility estimators" by Randolf Altmeyer and Markus Bibinger, January 2014.

006 "A consistent two-factor model for pricing temperature derivatives" by Andreas Groll, Brenda López-Cabrera and Thilo Meyer-Brandis, January 2014.

007 "Confidence Bands for Impulse Responses: Bonferroni versus Wald" by Helmut Lütkepohl, Anna Staszewska-Bystrova and Peter Winker, January 2014.

008 "Simultaneous Confidence Corridors and Variable Selection for Generalized Additive Models" by Shuzhuan Zheng, Rong Liu, Lijian Yang and Wolfgang Karl Härdle, January 2014.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".