Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Oracally Efficient Two-Step Estimation of Generalized Additive Model

Rong Liu [a b], Lijian Yang [a c] & Wolfgang K. Härdle [d]

[a] Center for Advanced Statistics and Econometrics Research, Soochow University, China

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

[c] Department of Statistics and Probability, Michigan State University, East Lansing, MI, 48824

[d] Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Berlin, Germany

Accepted author version posted online: 22 Jan 2013.

To cite this article: Rong Liu, Lijian Yang & Wolfgang K. Härdle (2013): Oracally Efficient Two-Step Estimation of Generalized Additive Model, Journal of the American Statistical Association, 108:502, 619-631

To link to this article: http://dx.doi.org/10.1080/01621459.2013.763726

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Oracally Efficient Two-Step Estimation of Generalized Additive Model

Rong LIU, Lijian YANG, and Wolfgang K. HÄRDLE

The generalized additive model (GAM) is a multivariate nonparametric regression tool for non-Gaussian responses including binary and count data. We propose a spline-backfitted kernel (SBK) estimator for the component functions and the constant, which are oracally efficient under weak dependence. The SBK technique is both computationally expedient and theoretically reliable, thus usable for analyzing high-dimensional time series. Inference can be made on component functions based on asymptotic normality. Simulation evidence strongly corroborates the asymptotic theory. The method is applied to estimate insolvent probability and to obtain higher accuracy ratio than a previous study. Supplementary materials for this article are available online.

KEY WORDS: Bandwidths; B-spline; Knots; Link function; Mixing; Nadaraya-Watson estimator.

1. INTRODUCTION

An effective semiparametric regression tool for high-dimensional data is the additive model introduced by Hastie and Tibshirani (1990), which stipulates that

$$\mathrm{E}(Y|\mathbf{X}) = m(\mathbf{X}),\ m(\mathbf{X}) = c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) \tag{1}$$

for a response Y and a predictor vector $\mathbf{X} = (X_1, \ldots, X_d)^\mathsf{T}$. When a dataset $\{Y_i, \mathbf{X}_i^\mathsf{T}\}_{i=1}^n = \{Y_i, X_{i1}, \ldots, X_{id}\}_{i=1}^n$ of size n is observed that follows model (1), unknown component functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ can be estimated via kernel, B-spline, and smoothing spline with a univariate convergence rate. This fact—together with the interpretability of the functions—has led not only to a remedy of the "curse of dimensionality," but also to increased practical applications of additive models. Articles on additive models and related works include, among others, Stone (1985), Stone (1994), Huang and Yang (2004), and Xue and Yang (2006a) for B-spline methods; Tjøstheim and Auestad (1994), Linton and Nielsen (1995), Linton (1997), Fan, Härdle, and Mammen (1998), Yang, Härdle, and Nielsen (1999), Xue and Yang (2006b), and Yang et al. (2006) for kernel methods; and more recently, the spline-backfitted kernel (SBK) smoothing methods of Wang and Yang (2007), Wang and Yang (2009), Liu and Yang (2010), and Ma and Yang (2011), and the spline-backfitted spline (SBS) smoothing method of Song and Yang (2010).

Certain types of responses Y, however, such as binary or Poisson responses, are much more appropriately described by the generalized additive model (GAM). In the GAM framework, the data $\{Y_i, \mathbf{X}_i^\mathsf{T}\}_{i=1}^n$ are generated according to

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = b'\{m(\mathbf{x})\}, \tag{2}$$

with $m(\mathbf{x})$ of additive structure as in (1) and a given function b' that relates $m(\mathbf{x})$ to the conditional variance function $\sigma^2(\mathbf{x}) = \mathrm{var}(Y|\mathbf{X} = \mathbf{x})$ via the equation $\sigma^2(\mathbf{x}) = a(\phi)b''\{m(\mathbf{x})\}$, in which $a(\phi)$ is a nuisance parameter that quantifies overdispersion. The inverse of b' is called the link function. For binary responses, one commonly takes $(b')^{-1}(x) = \log\{x/(1-x)\}$, the logistic link to conduct logistic regression, while for Poisson regression, $(b')^{-1}(x) = \log x$, the log link. If one takes $(b')^{-1}(x) = x$, the identity link, model (2) becomes model (1).

Model (2) has its origin in the special case where the probability density function of Y_i conditional on $\mathbf{X}_i$ with respect to a fixed σ-finite measure forms an exponential family

$$f(Y_i|\mathbf{X}_i, \phi) = \exp[\{Y_i m(\mathbf{X}_i) - b\{m(\mathbf{X}_i)\}\}/a(\phi) + h(Y_i, \phi)]. \tag{3}$$

For the theoretical development in this article, however, it is not necessary to assume that the data $\{Y_i, \mathbf{X}_i^\mathsf{T}\}_{i=1}^n$ come from such an exponential family, but only that the conditional variance and conditional mean are linked by the following equation

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

We can also write model (2) in the usual regression form

$$Y_i = b'\{m(\mathbf{X}_i)\} + \sigma(\mathbf{X}_i)\varepsilon_i \tag{4}$$

for conditional white noise ε_i that satisfies $\mathrm{E}(\varepsilon_i|\mathbf{X}_i) = 0$, $\mathrm{E}(\varepsilon_i^2|\mathbf{X}_i) = 1$. For identifiability, we need

$$\mathrm{E}\{m_\alpha(X_\alpha)\} = 0,\ 1 \le \alpha \le d \tag{5}$$

for unique additive representations of $m(\mathbf{x}) = c + \sum_{\alpha=1}^d m_\alpha(x_\alpha)$. As in most works on nonparametric smoothing, the estimation of the functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ is conducted on compact sets. Without loss of generality, let the compact set be $\chi = [0, 1]^d$.

Rong Liu is Visiting Assistant Professor, Center for Advanced Statistics and Econometrics Research, Soochow University, China, and 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, 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, Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Berlin, Germany (E-mail: *haerdle@wiwi.hu-berlin.de*). This work has been supported in part by National Science Foundation Awards DMS 0706518 and 1007594; funding from the National University of Singapore, the Jiangsu Specially-Appointed Professor Program, Jiangsu, China; and from Deutsche Forschungsgemeinschaft SFB 649 "Ökonomisches Risiko," Humboldt-Universität zu Berlin. The very constructive comments of the Associate Editor and two Reviewers are gratefully acknowledged.

Table 1. Example 4. Definitions of financial ratios

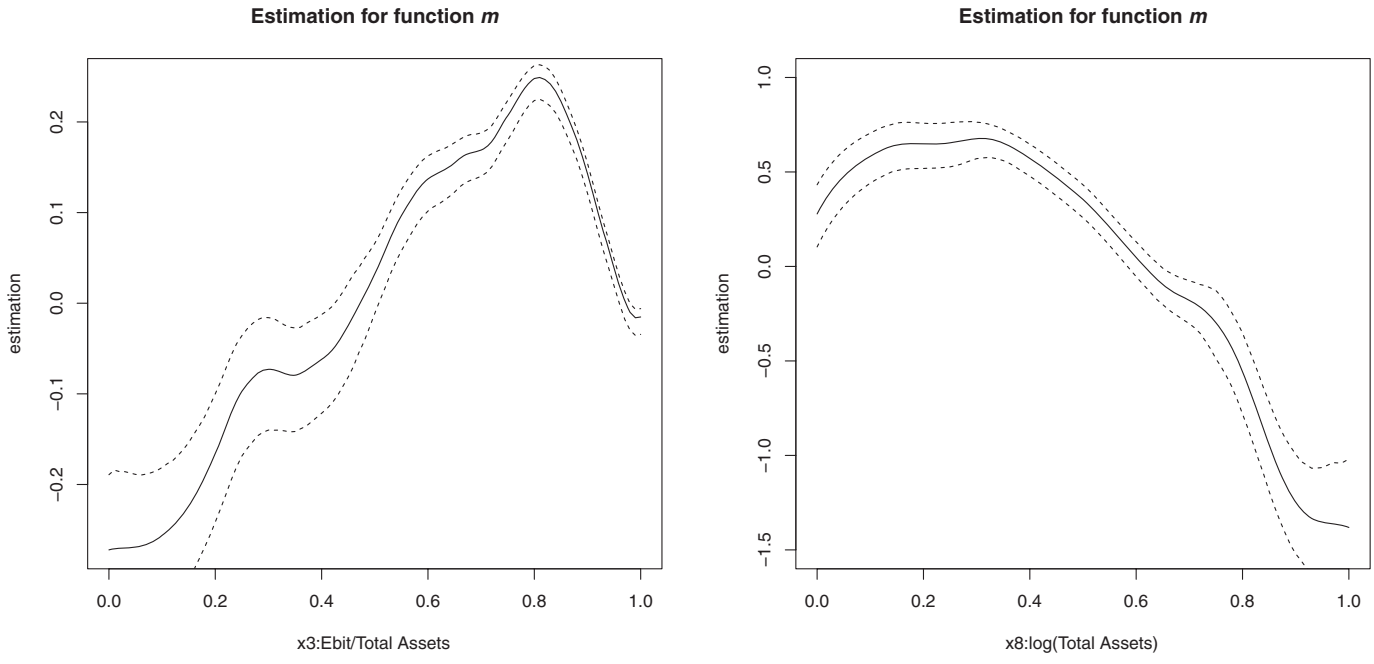
Ratio no.	Definition	Ratio no.	Definition
Z_1	Net_Income/Sales	Z_5	Cash/Total_Assets
Z_2	Operating_Income/ Total_Assets	Z_6	Inventories/Sales
Z_3	Ebit/Total_Assets	Z_7	Accounts _Payable/Sales
Z_4	Total_Liabilities/ Total_Assets	Z_8	log (Total_Assets)

Model (2) is a powerful tool for forecasting when companies will default, such as those listed in the credit reform database, provided by the Research Data Center (RDC) of the Humboldt-Universität zu Berlin. The dataset contains $d = 8$ financial ratios shown in Table 1, such as Ebit/Total_Assets and log(Total_Assets), of 18,610 solvent ($Y = 0$) and 1000 insolvent ($Y = 1$) German companies, see Härdle, Hoffmann, and Moro (2011) for details.

The company's default rate, that is, the conditional probability of $Y = 1$, is modeled as a logit function of the additive effects m_α, $1 \le \alpha \le 8$ of all the 8 financial ratios. In particular, estimates of m_3 for Ebit/Total_Assets (x_3) and m_8 for log(Total_Assets) (x_8) are shown in Figure 1.

Methods for the GAM (2) are much less developed in comparison to the additive model (1), see, for instance, the B-spline method of Stone (1986) and Xue and Liang (2010); the kernel method of Linton and Härdle (1996) and Yang, Sperlich, and Härdle (2003); and the two-stage methods of Horowitz and Mammen (2004) and Horowitz, Klemelä, and Mammen (2006). Generally speaking, the proposed kernel methods are too computationally intensive for high-dimension d, thus limiting their applicability to a small number of predictors. On the other hand, B-spline methods provide only convergence rates but no asymptotic distributions, so no measures of confidence can be assigned to the estimators. In the case of the additive model (1), the SBK method of Wang and Yang (2007) combines the advantages of both kernel and spline methods and the result is balanced in terms of theory, computation, and interpretation. The basic idea of the SBK method for the additive model (1) is to first project the data with B-splines into a space of functions with additive structure and then to apply kernel smoothing to the projected objects.

In this article, we extend the SBK method to model (2). The desired aim is to achieve oracle efficiency. If all the nonparametric functions of the last $d - 1$ variables, $\{m_\alpha(x_\alpha)\}_{\alpha=2}^d$, and the constant c were known by an "oracle," one could simply plug these in and estimate the only unknown functions $m_1(x_1)$ by maximizing the log-likelihood function with kernel weights computed from variable X_1. This estimator of $m_1(x_1)$ is called an "oracle smoother" or "infeasible estimator," and it does not suffer from the "curse of dimensionality" since the smoothing operation involves w.l.o.g. only X_1. The proposed SBK method pre-estimates functions $\{m_\alpha(x_\alpha)\}_{\alpha=2}^d$ and constant c by linear splines and then uses these estimates as proxies for the unknown functions $\{m_\alpha(x_\alpha)\}_{\alpha=2}^d$ and constant c. The main contribution is proving that the error caused by this approximation is uniformly negligible of order $\mathcal{O}_{\text{a.s.}}(n^{-1/2}\log n)$ (see Theorem 4). Consequently, the SBK estimator is uniformly (over the data range) asymptotically equivalent to the "oracle smoother," automatically inheriting all oracle efficiency properties of the latter. Our proof relies on "reducing bias by undersmoothing" and "averaging out the variance," accomplished with the joint asymptotics of kernel and spline functions for realizations of geometrically strongly mixing time series. These results are established under substantially greater technical difficulty than existing works on additive model such as Wang and Yang (2007), Wang and Yang (2009), Liu and Yang (2010), Ma and Yang (2011), and Song and Yang (2010). The additional complication is due to the lack of decomposition of spline



(a) (b)

Figure 1. Example 4. Estimators (a) $\hat{m}_{\text{SBK},3}(x_3)$ and (b) $\hat{m}_{\text{SBK},8}(x_8)$ and asymptotic 95% pointwise confidence intervals.

estimation error into the sum of a bias and a noise term when the link function $(b')^{-1}$ is nonlinear.

The asymptotic distribution (Theorem 2) for the two-stage estimator in Horowitz and Mammen (2004) was valid only for iid data, while our result for the SBK estimator is proved under weak dependence (Corollary 1). It is also worth noting that although Horowitz and Mammen (2004) had used the B-spline estimator for the first stage in simulation, their proof is valid only for using the orthogonal series estimator in stage one. In addition, our Theorem 4 would allow one to construct a simultaneous band of the function $m_1(x_1)$, which has never been done, based on the SBK estimator by further research on the maximal deviation distribution (see Härdle 1989) of the uniformly $\mathcal{O}_{\text{a.s.}}(n^{-1/2}\log n)$ equivalent, but structurally simpler infeasible estimator. Another contribution beyond Horowitz and Mammen (2004) is establishing that the spline-backfitted estimator of the baseline constant c is within a negligible error of order $o_p(n^{-1/2})$ of the infeasible estimator and thus also oracally efficient. As far as we know, our estimator of the baseline constant c is the only one that has an asymptotic distribution with $n^{-1/2}$ rate.

The article is organized as follows. In Section 2, we discuss the assumptions of model (2). In Section 3, we introduce the oracle smoother or infeasible estimator for $m_1(x_1)$ and for c and state their asymptotics. In Section 4, we introduce the SBK estimator for $m_1(x_1)$ and the spline-backfitted estimator for c and present their asymptotic oracle efficiencies by showing that they differ from their infeasible counterparts only negligibly. In Section 5, we describe implementation steps of the estimators. In Section 6, we apply the methods to simulated and real examples. All technical proofs are given in the Appendix.

2. MODEL ASSUMPTIONS

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

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

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

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

in which c is finite constant. The constraints that $\mathrm{E}\{g_\alpha(X_\alpha)\} = 0$, $1 \le \alpha \le d$ ensure unique additive representation of m_α as expressed in (5), but are not necessary for the definition of space $\mathcal{M}$. In what follows, denote by E_n the empirical expectation, $\mathrm{E}_n\varphi = \sum_{i=1}^n \varphi(\mathbf{X}_i)/n$. We introduce two inner products on $\mathcal{M}$. For functions $g_1, g_2 \in \mathcal{M}$, the theoretical and empirical inner products are defined, respectively, as $\langle g_1, g_2\rangle = \mathrm{E}\{g_1(\mathbf{X})g_2(\mathbf{X})\}$, $\langle g_1, g_2\rangle_n = \mathrm{E}_n\{g_1(\mathbf{X})g_2(\mathbf{X})\}$. The corresponding induced norms are $\|g_1\|_2^2 = \mathrm{E}g_1^2(\mathbf{X})$, $\|g_1\|_{2,n}^2 = \mathrm{E}_n g_1^2(\mathbf{X})$. More generally, we define $\|g\|_r^r = \mathrm{E}|g(\mathbf{X})|^r$.

Throughout the article, for any compact interval $[a, b]$, we denote the space of pth order smooth function as $C^{(p)}[a, b] = \{g|g^{(p)} \in C[a, b]\}$ and the class of Lipschitz continuous functions for constant $C > 0$ as $\mathrm{Lip}([a, b], C) = \{g||g(x) - g(x')| \le C|x - x'|, \forall x, x' \in [a, b]\}$. We mean by "$\sim$" both sides having the same order as $n \to \infty$. For any vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T}$, we denote the supremum and p norms as $|\mathbf{x}| = \max_{1\le\alpha\le d}|x_\alpha|$ and $\|\mathbf{x}\|_p = (\sum_{\alpha=1}^d x_\alpha^p)^{1/p}$. In particular, we use $\|\mathbf{x}\|$ to denote the Euclidean norm.

We need the following assumptions on the data-generating process.

(A1) *The additive component functions* $m_\alpha \in C^{(1)}[0, 1]$, $1 \le \alpha \le d$ *with* $m_1 \in C^{(2)}[0, 1]$, $m'_\alpha \in \mathrm{Lip}([0, 1], C_m) = 2 \le \alpha \le d$ *for some constant* $C_m > 0$.

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

(A3) *The conditional variance function* $\sigma^2(\mathbf{x})$ *is measurable and bounded. The errors* $\{\varepsilon_i\}_{i=1}^n$ *satisfy* $\mathrm{E}(\varepsilon_i|\mathcal{F}_i) = 0$, $\mathrm{E}(|\varepsilon_i|^{2+\eta}) \le C_\eta$ *for some* $\eta \in (1/2, +\infty)$ *and the sequence of* σ*-fields* $\mathcal{F}_i = \sigma\{(\mathbf{X}_j), j \le i; \varepsilon_j, j \le i - 1\}$ *for* $i = 1, \ldots, n$.

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

$$0 < c_f \le \inf_{\mathbf{x}\in\chi} f(\mathbf{x}) \le \sup_{\mathbf{x}\in\chi} f(\mathbf{x}) \le C_f < \infty.$$

The marginal densities $f_\alpha(x_\alpha)$ *of* X_α *have continuous derivatives on* $[0, 1]$ *as well as the uniform upper bound* C_f *and lower bound* c_f.

(A5) *Constants* $K_0, \lambda_0 \in (0, +\infty)$ *exist such that* $\alpha(n) \le K_0 e^{-\lambda_0 n}$ *holds for all n, with the* α*-mixing coefficients for* $\{\mathbf{Z}_i = (\mathbf{X}_i^\mathsf{T}, \varepsilon_i)\}_{i=1}^n$ *defined as*

$$\alpha(k) = \sup_{B\in\sigma\{\mathbf{Z}_s, s\le t\}, C\in\sigma\{\mathbf{Z}_s, s\ge t+k\}} |\mathrm{P}(B \cap C) - \mathrm{P}(B)\mathrm{P}(C)|, \quad k \ge 1.$$

Assumptions (A1), (A2), and (A4) are standard in the GAM literature, see Stone (1986), Xue and Liang (2010), while Assumptions (A3) and (A5) are the same for weakly dependent data as in Wang and Yang (2007) and Liu and Yang (2010). Assumption (A2) implies that a compact interval A exists whose interior contains $m_1([0, 1])$ and that Θ's interior contains $A + m_{-1}([0, 1]^{d-1})$ where $m_{-1}(\mathbf{x}_{-1}) = c + \sum_{\alpha=2}^d m_\alpha(x_\alpha)$ with $x_{-1} = (x_2, \ldots, x_d)$.

3. ORACLE SMOOTHERS

We now introduce what is known as the oracle smoother in Wang and Yang (2007) as a benchmark for evaluating the estimators. If the last $d - 1$ components $\{m_\alpha(x_\alpha)\}_{\alpha=2}^d$ were w.l.o.g. known by an "oracle," then the only unknown component $m_1(x_1)$ may be estimated by the following procedure. Although the exponential family Equation (3) 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 $\tilde{l}(a) = \tilde{l}(a, x_1)$, $a \in A$ as

$$n^{-1}\sum_{i=1}^n [Y_i\{a + m_{-1}(\mathbf{X}_{i-1})\} - b\{a + m_{-1}(\mathbf{X}_{i-1})\}]K_h(X_{i1} - x_1) \tag{6}$$

with $m_{-1}(\mathbf{X}_{i-1}) = c + \sum_{\alpha=2}^d m_\alpha(\mathbf{X}_{i\alpha})$ and define the oracle smoother of $m_1(x_1)$ as

$$\tilde{m}_{\mathrm{K},1}(x_1) = \arg\max_{a\in A} \tilde{l}(a, x_1). \tag{7}$$

in which $K_h(u) = K(u/h)/h$ for a kernel function K and bandwidth h that satisfy

(A6) *The kernel function K is a symmetric probability density, supported on $[-1, 1]$ and $K \in \mathrm{Lip}([-1, 1], C_K)$ for some positive constant $C_K > 0$. A constant $c_h > 0$ exists such that the bandwidth $h = h_n$ satisfies $h = \mathcal{O}(n^{-1/5})$, $h^{-1} = \mathcal{O}(n^{1/5}(\log n)^{c_h})$.*

We denote $\|K\|_2^2 = \int K^2(u)du$, $\mu_2(K) = \int K(u)u^2du$, and the higher-order error of $\tilde{m}_{\mathrm{K},1}(x_1)$ as

$$r_{\mathrm{K},1}(x_1) = \tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1) - \mathrm{bias}_1(x_1)h^2/D_1(x_1) - n^{-1}\sum_{i=1}^{n} K_h(X_{i1} - x_1)\sigma(\mathbf{X}_i)\varepsilon_i/D_1(x_1),$$

with the scale function $D_1(x_1)$ and bias function $\mathrm{bias}_1(x_1)$ defined as

$$D_1(x_1) = f_1(x_1)\mathrm{E}\{b''\{m(\mathbf{X})\}|X_1 = x_1\}, \tag{8}$$

$$\begin{aligned}\mathrm{bias}_1(x_1) = {} & \mu_2(K)[m_1''(x_1)f_1(x_1)\mathrm{E}[b''\{m(\mathbf{X})\}|X_1 = x_1] \\ & + m_1'(x_1)\frac{\partial}{\partial x_1}\{f_1(x_1)\mathrm{E}[b''\{m(\mathbf{X})\}|X_1 = x_1]\} \\ & - \{m_1'(x_1)\}^2 f_1(x_1)\mathrm{E}[b'''\{m(\mathbf{X})\}|X_1 = x_1]].\end{aligned} \tag{9}$$

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

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

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

Theorem 2. Under Assumptions (A1)–(A6), for any $x_1 \in [h, 1-h]$, as $n \to \infty$, the oracle kernel smoother $\tilde{m}_{\mathrm{K},1}(x_1)$ given in (7) satisfies

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

in which

$$v_1^2(x_1) = f_1(x_1)\mathrm{E}\{\sigma^2(\mathbf{X})|X_1 = x_1\}\|K\|_2^2. \tag{10}$$

The same oracle idea applies to the constant as well. Define the log-likelihood function

$$\tilde{l}_c(a) = n^{-1}\sum_{i=1}^{n}\left[Y_i\left\{a + m_{\_c}(\mathbf{X}_i)\right\} - b\left\{a + m_{\_c}(\mathbf{X}_i)\right\}\right],$$

where $m_{\_c}(\mathbf{X}) = \sum_{\alpha=1}^{d} m_\alpha(X_\alpha)$. The infeasible estimator of c is $\tilde{c} = \arg\max_{a\in A}\tilde{l}_c(a)$. Clearly, $\tilde{l}_c'(\tilde{c}) = 0$.

Theorem 3. Under Assumptions (A1)–(A5), as $n \to \infty$,

$$\tilde{c} - c = [\mathrm{E}b''\{m(\mathbf{X})\}]^{-1}n^{-1}\sum_{i=1}^{n}\sigma(\mathbf{X}_i)\varepsilon_i + \mathcal{O}_{\mathrm{a.s.}}(n^{-1}(\log n)^2).$$

Although the oracle smoother $\tilde{m}_{\mathrm{K},1}(x_1)$ enjoys the desirable theoretical properties in Theorems 1 and 2, it is not a statistic, as its computation is based on the knowledge of unavailable functions $\{m_\alpha(x_\alpha)\}_{\alpha=2}^d$ and the unknown constant c, the same can be said of $\tilde{c}$. These benchmarks, however, motivate the spline-backfitted estimators that we will introduce in the next section.

4. SPLINE-BACKFITTED KERNEL ESTIMATORS

In this section, we describe how the unknown functions $\{m_\alpha(x_\alpha)\}_{\alpha=2}^d$ and constants c can be pre-estimated by linear splines and how the estimates are used to construct the SBK estimator. First, we introduce the space of linear splines as in Liu and Yang (2010). Let $0 = \xi_0 < \xi_1 < \cdots < \xi_N < \xi_{N+1} = 1$ denote a sequence of equally spaced points, called interior knots, on interval $[0, 1]$. Denote by $H = (N+1)^{-1}$ the width of each subinterval $[\xi_J, \xi_{J+1}]$, $0 \le J \le N$, and denote the degenerate knots by $\xi_{-1} = 0$, $\xi_{N+2} = 1$. Assume that

A7. *The number of interior knots satisfies $N \sim n^{1/4}\log n$, that is, $c_N n^{1/4}\log n \le N \le C_N n^{1/4}\log n$ for some positive constants c_N, C_N.*

For $J = 0, \ldots, N+1$, define the linear B-spline basis as

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

the space of α-empirically centered linear spline functions on $[0, 1]$ as

$$G_{n,\alpha}^0 = \left\{g_\alpha : g_\alpha(x_\alpha) = \sum_{J=0}^{N+1}\lambda_J b_J(x_\alpha), \mathrm{E}_n\{g_\alpha(x_\alpha)\} = 0\right\}, \quad 1 \le \alpha \le d,$$

and the space of additive spline functions on χ as

$$G_n^0 = \left\{g(\mathbf{x}) = c + \sum_{\alpha=1}^{d} g_\alpha(x_\alpha); \quad c \in R, g_\alpha \in G_{n,\alpha}^0\right\},$$

which is equipped with the empirical inner product $\langle\cdot,\cdot\rangle_{2,n}$. Define the log-likelihood function as

$$\hat{L}(g) = n^{-1}\sum_{i=1}^{n}[Y_i g(\mathbf{X}_i) - b\{g(\mathbf{X}_i)\}], \; g \in G_n^0, \tag{11}$$

which, according to lemma 14 of Stone (1986), has a unique maximizer with probability approaching 1. The multivariate function $m(\mathbf{x})$ is then estimated by the additive spline function

$$\hat{m}(\mathbf{x}) = \arg\max_{g\in G_n^0}\hat{L}(g).$$

Since $\hat{m}(\mathbf{x}) \in G_n^0$, one can write $\hat{m}(\mathbf{x}) = \hat{c} + \sum_{\alpha=1}^{d}\hat{m}_\alpha(x_\alpha)$ for $\hat{c} \in \mathbb{R}$ and $\hat{m}_\alpha(x_\alpha) \in G_{n,\alpha}^0$. Next define the log-likelihood function

$$\hat{l}(a) = \frac{1}{n}\sum_{i=1}^{n}[Y_i\{a + \hat{m}_{\_1}(\mathbf{X}_{i\_1})\} - b\{a + \hat{m}_{\_1}(\mathbf{X}_{i\_1})\}]K_h(X_{i1} - x_1), \tag{12}$$

where $\hat{m}_{\_1}(\mathbf{X}_{i\_1}) = \hat{c} + \sum_{\alpha=2}^{d}\hat{m}_\alpha(X_{i\alpha})$. Define the SBK estimator as

$$\hat{m}_{\mathrm{SBK},1}(x_1) = \arg\max_{a\in A}\hat{l}(a). \tag{13}$$

Theorem 4. Under Assumptions (A1)–(A7), as $n \to \infty$, $\hat{m}_{\mathrm{SBK},1}(x_1)$ is oracally efficient,

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

Theorem 4 follows from (A.17) and Lemmas A.15 and A.16. The following corollary is a consequence of Theorems 1, 2, and 4.

Corollary 1. Under Assumptions (A1)–(A7), as $n \to \infty$, the SBK estimator $\hat{m}_{\mathrm{SBK},1}(x_1)$ given in (13) satisfies

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

and for any $x_1 \in [h, 1-h]$, with $\mathrm{bias}_1(x_1)$ as in (9) and $D_1(x_1)$ in (8),

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

The estimator $\hat{m}_{\mathrm{SBK},1}(x_1)$ is called oracally efficient because it differs from the infeasible oracle smoother $\tilde{m}_{\mathrm{K},1}(x_1)$ by a term uniformly of order $O(n^{-1/2} \log n)$, which is negligible compared with the $O(\log n/\sqrt{nh})$ magnitude of $\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1)$ according to Theorem 1. Consequently, $\hat{m}_{\mathrm{SBK},1}(x_1)$ as an estimator of $m_1(x_1)$ is as efficient as $\tilde{m}_{\mathrm{K},1}(x_1)$. We agree with one referee that analog of such oracle efficiency does not hold for linear regression model, and the reason is that the kernel and spline smoothing methods have no parametric counterparts.

Define next the spline-backfitted estimator $\hat{c} = \arg\max_{a \in A} \hat{l}_c(a)$ with

$$\hat{l}_c(a) = n^{-1} \sum_{i=1}^{n} [Y_i \{a + \hat{m}_{\_c}(\mathbf{X}_i)\} - b\{a + \hat{m}_{\_c}(\mathbf{X}_i)\}]$$

in which $\hat{m}_{\_c}(\mathbf{X}_i) = \sum_{\alpha=1}^{d} \hat{m}_\alpha(X_{i\alpha})$. Similar to Theorem 4, the main result shows that the difference between $\hat{c}$ and its infeasible counterpart $\tilde{c}$ is asymptotically negligible.

Theorem 5. Under Assumptions (A1)–(A5) and (A7), as $n \to \infty$, $\hat{c}$ is oracally efficient, that is, $\sqrt{n}(\hat{c} - \tilde{c}) \xrightarrow{p} 0$ and hence

$$\sqrt{n}(\hat{c} - c) \xrightarrow{\mathcal{L}} N\big(0, a(\phi)^{1/2} [\mathrm{E}\, b''\{m(\mathbf{X})\}]^{-1/2}\big).$$

The recent work of Ravikumar et al. (2009) on sparse additive model suggests possible extension of the SBK method in this article to dimension d higher than n, by adapting the penalty term of Ravikumar et al. (2009) to B-spline basis and adding it to the spline log-likelihood function $\hat{L}(g)$ in (11). Further investigation is needed to understand the theoretical properties and performance of such extension.

5. IMPLEMENTATION

We implement our procedures with the following rule-of-thumb number of interior knots

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

which satisfies (A8), that is, $N = N_n \sim n^{1/4} \log n$ and ensures that the number of parameters in the linear least squares problem is less than $n/4$, that is, $1 + d(N+1) \le n/4$.

According to Corollary 1, the asymptotic distribution of the estimator $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ depends not only on the functions $\mathrm{bias}_\alpha(x_\alpha)/D_\alpha(x_\alpha)$ and $D_\alpha(x_\alpha)^{-1} v_\alpha^2(x_\alpha) D_\alpha(x_\alpha)^{-1}$, but also crucially on the choice of bandwidths h_α. Define the optimal bandwidth of h_α, denoted by $h_{\alpha,\mathrm{opt}}$, as the minimizer of the asymptotic mean integrated squared errors (AMISE) of $\{\hat{m}_\alpha(x_a), \alpha = 1, \ldots, d\}$:

$$\mathrm{AMISE}\,(\hat{m}_\alpha) = \int \big[\big\{\mathrm{bias}_\alpha(x_\alpha) h_\alpha^2 \big/ D_\alpha(x_\alpha)\big\}^2 + D_\alpha(x_\alpha)^{-1} v_\alpha^2(x_\alpha) D_\alpha(x_\alpha)^{-1}/(nh_\alpha)\big] f_\alpha(x_\alpha) dx_\alpha.$$

By letting $d\,\mathrm{AMISE}\,(\hat{m}_\alpha)/dh_\alpha = 0$, one obtains an optimal bandwidth $h_{\alpha,\mathrm{opt}}$:

$$h_{\alpha,\mathrm{opt}} = \left\{ \frac{n^{-1} \int D_\alpha(x_\alpha)^{-1} v_\alpha^2(x_\alpha) D_\alpha(x_\alpha)^{-1} f_\alpha(x_\alpha) dx_\alpha}{4 \int \{\mathrm{bias}_\alpha(x_\alpha)/D_\alpha(x_\alpha)\}^2 f_\alpha(x_\alpha) dx_\alpha} \right\}^{1/5},$$

which is approximated by

$$\hat{h}_{\alpha,\mathrm{opt}} = \left\{ \frac{n^{-1} \sum_{i=1}^{n} D_\alpha(X_{i\alpha})^{-1} v_\alpha^2(X_{i\alpha}) D_\alpha(X_{i\alpha})^{-1}}{4 \sum_{i=1}^{n} \{\mathrm{bias}_\alpha(X_{i\alpha})/D_\alpha(X_{i\alpha})\}^2} \right\}^{1/5},$$

where

$$D_\alpha(x_\alpha) = f_\alpha(x_\alpha) \mathrm{E}[b''\{m(\mathbf{X})\} | X_\alpha = x_\alpha]$$

and

$$\begin{aligned} v_\alpha^2(x_\alpha) &= f_\alpha(x_\alpha) \mathrm{E}\{\sigma^2(\mathbf{X}) | X_\alpha = x_\alpha\} \| K \|_2^2, \\ \mathrm{bias}_\alpha(x_\alpha) &= \mu_2(K) \Big\{ m_\alpha''(x_\alpha) f_\alpha(x_\alpha) \mathrm{E}[b''\{m(\mathbf{X})\} | X_\alpha = x_\alpha] \\ &\quad + m_\alpha'(x_\alpha) \frac{\partial}{\partial x_\alpha} \{ f_\alpha(x_\alpha) \mathrm{E}[b''\{m(\mathbf{X})\} | X_\alpha = x_\alpha] \} \\ &\quad - \{m_\alpha'(x_\alpha)\}^2 f_\alpha(x_\alpha) \mathrm{E}[b'''\{m(\mathbf{X})\} | X_\alpha = x_\alpha] \Big\}. \end{aligned}$$

The following estimation methods for the terms $m_\alpha'(x_\alpha)$, $m_\alpha''(x_\alpha)$, $f_\alpha(x_\alpha)$, $\mathrm{E}\{\sigma^2(\mathbf{X}) | X_\alpha = x_\alpha\}$, $\mathrm{E}[b''\{m(\mathbf{X})\} | X_\alpha = x_\alpha]$, $\mathrm{E}[b'''\{m(\mathbf{X})\} | X_\alpha = x_\alpha]$, and $\frac{\partial}{\partial x_\alpha} f_\alpha(x_\alpha) \mathrm{E}[b''\{m(\mathbf{X})\} | X_\alpha = x_\alpha]$ are proposed. The final bandwidth is denoted as $\hat{h}_{\alpha,\mathrm{opt}}$.

1. The derivative functions $m_\alpha'(X_{i\alpha})$ and $m_\alpha''(X_{i\alpha})$ are estimated as

$$\sum_{k=1}^{3} k \hat{a}_{\alpha,l,k} X_{i\alpha}^{k-1} + 3 \sum_{k=4}^{N+3} \hat{a}_{\alpha,l,k} (X_{i\alpha} - t_{\alpha,k-3})_+^2$$

and

$$\sum_{k=2}^{3} k(k-1) \hat{a}_{\alpha,l,k} X_{i\alpha}^{k-2} + 6 \sum_{k=4}^{N+3} \hat{a}_{\alpha,l,k} (X_{i\alpha} - t_{\alpha,k-3})_+$$

where $\{\hat{a}_{\alpha,l,k}\}_{k=0}^{N+3}$ maximize:

$$\sum_{i=1}^{n} \left[Y_i \left\{ \sum_{k=0}^{3} a_{\alpha,l,k} X_{i\alpha}^k + \sum_{k=4}^{N+3} a_{\alpha,l,k} (X_{i\alpha} - t_{\alpha,k-3})_+^3 \right\} - b \left\{ \sum_{k=0}^{3} a_{\alpha,l,k} X_{i\alpha}^k + \sum_{k=4}^{N+3} a_{\alpha,l,k} (X_{i\alpha} - t_{\alpha,k-3})_+^3 \right\} \right],$$

where $\min_i X_{i\alpha} = t_{\alpha,0} < \cdots < t_{\alpha,N+1} = \max_i X_{i\alpha}$.

2. $\mathrm{E}[b''\{m(\mathbf{X})\}|X_\alpha = x_\alpha]$ is estimated as $\sum_{k=0}^{3}\hat{a}_{\alpha,l,k}^{k}x_\alpha^k + \sum_{k=4}^{N+3}\hat{a}_{\alpha,l,k}(x_\alpha - t_{\alpha,k-3})_+^3$ by minimizing

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

$\frac{\partial}{\partial x_\alpha}f_\alpha(x_\alpha)\mathrm{E}[b''\{m(\mathbf{X})\}|X_\alpha = x_\alpha]$ and $\mathrm{E}[b'''\{m(\mathbf{X})\}|X_\alpha = x_\alpha]$ are estimated by $f_\alpha(x_\alpha)\sum_{k=1}^{3}k\hat{a}_{\alpha,l,k}^{k}x_\alpha^{k-1} + 3\sum_{k=4}^{N+3}\hat{a}_{\alpha,l,k}(x_\alpha - t_{\alpha,k-3})_+^2$ and $\sum_{k=0}^{3}\hat{a}_{\alpha,l,k}^{k}x_\alpha + \sum_{k=4}^{N+3}\hat{a}_{\alpha,l,k}(x_\alpha - t_{\alpha,k-3})_+^3$ by minimizing $\sum_{i=1}^{n}[b'''\{\hat{m}(\mathbf{X}_i)\} - \{\sum_{k=0}^{3}a_{\alpha,l,k}X_{i\alpha}^k + \sum_{k=4}^{N+3}a_{\alpha,l,k}(X_{i\alpha} - t_{k-3})_+^3\}]^2$.

3. $\mathrm{E}\{\sigma^2(\mathbf{X})|X_\alpha = x_\alpha\}$ is estimated by $\sum_{k=0}^{3}\hat{a}_{\alpha,l,k}^{k}x_\alpha + \sum_{k=4}^{N+3}\hat{a}_{\alpha,l,k}(x_\alpha - t_{\alpha,k-3})_+^3$ by minimizing

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

4. The density function $f_\alpha(x_\alpha)$ is estimated by $n^{-1}\sum_{i=1}^{n}K_{h_\alpha}(X_{i\alpha} - x_\alpha)$ with a rule-of-the-thumb bandwidth h_α.

6. EXAMPLES

We have applied the estimation procedure described in the previous section to both simulated (Examples 1, 2, and 3) and real (Example 4) data. The R package for SBK estimation is provided in the online supplementary materials.

6.1 Example 1

The data are generated from the model

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

with $d = 5, c = 0, m_1(x) = -\sin(2\pi x)$, $m_2(x) = \Phi(6x - 3) - 0.5$, and $m_3(x) = m_4(x) = m_5(x) = 2x - 1$, where Φ is the standard normal distribution function. The predictors are generated by transforming the following vector autoregression equation for $0 \le a, r < 1$,

$$\begin{aligned} X_{t\alpha} &= \Phi(\sqrt{1-a^2}Z_{t\alpha}), 2 \le t \le n, 1 \le \alpha \le d \\ \mathbf{Z}_t &= a\mathbf{Z}_{t-1} + \varepsilon_i, \varepsilon_i \sim N(0, \Sigma), 2 \le t \le n, \\ \Sigma &= (1-r)\mathbf{I}_{d\times d} + r\mathbf{1}_d\mathbf{1}_d^{\mathsf{T}}, \end{aligned}$$

with stationary $\mathbf{Z}_t = (Z_{t1}, \ldots, Z_{td})^{\mathsf{T}} \sim N\{0, (1-a^2)^{-1}\Sigma\}$, $\mathbf{1}_d = (1, \ldots, 1)^{\mathsf{T}}$ and $\mathbf{I}_{d\times d}$ is the $d \times d$ identity matrix. Higher values of a correspond to stronger dependence among the observations, and in particular, if $a = 0$, the data are iid The r controls the correlation of X_{t1} and X_{t2}. In this study, we have experimented with two cases: $r = 0, a = 0$ and $r = 0.5$, $a = 0.5$ to cover various scenarios. For $\alpha = 1, \ldots, d$, let $X_{\alpha,\min}^i$, $X_{\alpha,\max}^i$ denote the smallest and largest observations of the variable X_α in the ith replication. The component functions $\{m_\alpha\}_{\alpha=1}^{d}$ are estimated on equally spaced points $\{x_t\}_{t=0}^{100}$ with $0 = x_0 < \cdots < x_{100} = 1$ and the estimator of m_α in the rth sample as $\hat{m}_{\mathrm{SBK},\alpha,r}$. We define the (mean) average squared error (ASE and MASE) as

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

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

To examine the efficiency of the SBK estimator $\hat{m}_{\mathrm{SBK},\alpha}$ relative to the "oracle smoother" $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$, both are computed using the same data-driven bandwidth $\hat{h}_{\alpha,\mathrm{opt}}$ described in Section 5, and define the empirical relative efficiency of $\hat{m}_{\mathrm{SBK},\alpha}$ with respect to $\tilde{m}_{\mathrm{K},\alpha}$ as

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

To compare with the existing estimation method, we also compute the MASEs and EFFs of the GAM estimator $\hat{m}_{\mathrm{GAM},\alpha}$ from the R package. Table 2 shows the MASEs of $\tilde{m}_{\mathrm{K},\alpha}$, $\hat{m}_{\mathrm{SBK},\alpha}$, and $\hat{m}_{\mathrm{GAM},\alpha}$, together with $\overline{\mathrm{EFF}}(\hat{m}_{\mathrm{SBK},\alpha})$, $\overline{\mathrm{EFF}}(\hat{m}_{\mathrm{GAM},\alpha})$, which are the means of the EFFs for $\alpha = 1, 2$ and $R = 100$. It is apparent that the SBK estimator performs as well asymptotically as the oracle estimator, see Theorem 4. For this example of low dimension ($d = 5$), the performance of the SBK estimator $\hat{m}_{\mathrm{SBK},1}$ of m_1 is comparable to the GAM estimator $\hat{m}_{\mathrm{GAM},1}$ for sample sizes $n = 500$, 1000, and significantly better than $\hat{m}_{\mathrm{GAM},1}$ for larger sample sizes $n = 2000$, 4000. The performance of the SBK estimator $\hat{m}_{\mathrm{SBK},2}$ of m_2 is clearly better than the GAM estimator $\hat{m}_{\mathrm{GAM},2}$ for all combinations of r, a, and n.

6.2 Example 2

We now examine a variation of Example 1, with the same design variables and link function but a higher dimension $d = 10$, $m_\alpha(x_\alpha) = -\sin(2\pi x_\alpha)$, $\alpha = 1, \ldots, 10$. We have run 100 replications for sample sizes $n = 500$, 1000, 2000, 4000. The MASEs and EFFs of $\tilde{m}_{\mathrm{K},1}$, $\hat{m}_{\mathrm{SBK},1}$, and $\hat{m}_{\mathrm{GAM},1}$ are shown in Table 3. As expected, increases in sample size reduce MASE for all estimators and across all combinations of r and a values. For this example of higher dimension ($d = 10$), the performance of the SBK estimator $\hat{m}_{\mathrm{SBK},1}$ is significantly better than the GAM estimator $\hat{m}_{\mathrm{GAM},1}$ for all combinations of r, a, and n except $r = 0.5, a = 0.5, n = 500$, and much more markedly for larger sample sizes $n = 2000$, 4000.

The convergence properties are displayed in Figure 2(a) showing the kernel density estimator of the simulated efficiencies for $\alpha = 1$ and sample sizes $n = 500$, 1000, 2000, 4000 for $r = 0, a = 0$. The vertical line at efficiency $= 1$ is the standard line for the comparison of $\hat{m}_{\mathrm{SBK},1}$ and $\tilde{m}_{\mathrm{K},1}$. One can clearly see that the center of the density plots is moving toward the standard line 1.0 with a narrower spread when sample size increases, which confirms the result of Theorem 4. The basic graphic pattern of Figure 2(b) with $r = 0, a = 0.5$; (c) with $r = 0.5, a = 0$; and (d) with $r = 0.5, a = 0.5$ are similar to that for the iid case, with slightly slower convergence and slightly poorer efficiency.

To have an impression of the actual function estimates, for $r = 0.5$, $a = 0.5$ with sample sizes $n = 500$, 1000, 2000, 4000, we have plotted the SBK estimators and their asymptotic 95% pointwise confidence intervals (solid lines) and

Table 2. Example 1. The MASEs and $\overline{\text{EFF}}$s of $\tilde{m}_{\text{K},\alpha}$, $\hat{m}_{\text{SBK},\alpha}$, $\hat{m}_{\text{GAM},\alpha}$ for $\alpha = 1, 2$, $d = 5$ and $n = 500, 1000, 2000, 4000$

$d = 5$	n	MASE($\tilde{m}_{\text{K},\alpha}$)	MASE($\hat{m}_{\text{SBK},\alpha}$)	MASE($\hat{m}_{\text{GAM},\alpha}$)	$\overline{\text{EFF}}$($\hat{m}_{\text{SBK},\alpha}$)	$\overline{\text{EFF}}$($\hat{m}_{\text{GAM},\alpha}$)
$r = 0$,	500	0.04482	0.04603	0.04693	0.9501	0.9441
$a = 0$,	1000	0.02418	0.02503	0.02502	0.9809	0.9850
$\alpha = 1$	2000	0.01582	0.01613	0.02679	0.9854	0.5735
	4000	0.01212	0.01247	0.02440	0.9923	0.4805
$r = 0$,	500	0.04060	0.04322	0.04398	0.9445	0.9451
$a = 0.5$,	1000	0.02592	0.02649	0.02654	0.9767	0.9745
$\alpha = 1$	2000	0.01746	0.01714	0.02885	0.9832	0.5821
	4000	0.01194	0.01218	0.02437	0.9936	0.4783
$r = 0.5$,	500	0.04845	0.06348	0.06218	0.8827	0.8871
$a = 0$,	1000	0.02935	0.03559	0.03485	0.8755	0.8826
$\alpha = 1$	2000	0.01951	0.02177	0.03418	0.9494	0.5416
	4000	0.01515	0.01648	0.03020	0.9795	0.4916
$r = 0.5$,	500	0.05656	0.07114	0.07057	0.8722	0.8736
$a = 0.5$,	1000	0.02804	0.03570	0.03488	0.8951	0.8972
$\alpha = 1$	2000	0.01886	0.02089	0.03360	0.9478	0.5413
	4000	0.01525	0.01634	0.02986	0.9744	0.4955
$r = 0$,	500	0.01875	0.02973	0.03372	0.7276	0.5213
$a = 0$,	1000	0.01074	0.01641	0.01644	0.7699	0.5712
$\alpha = 2$	2000	0.00700	0.00800	0.00879	0.8199	0.6877
	4000	0.00267	0.00307	0.00401	0.9573	0.7416
$r = 0$,	500	0.01553	0.02766	0.02801	0.6958	0.5252
$a = 0.5$,	1000	0.01064	0.01668	0.01761	0.7401	0.5986
$\alpha = 2$	2000	0.00615	0.00809	0.00850	0.8368	0.7846
	4000	0.00360	0.00458	0.00504	0.8521	0.7637
$r = 0.5$,	500	0.02202	0.03303	0.04222	0.6816	0.4185
$a = 0$,	1000	0.01312	0.01667	0.02254	0.7789	0.5091
$\alpha = 2$	2000	0.00794	0.01019	0.01090	0.7960	0.7454
	4000	0.00495	0.00594	0.00633	0.8920	0.7467
$r = 0.5$,	500	0.02121	0.03173	0.04538	0.7186	0.4478
$a = 0.5$,	1000	0.01427	0.01728	0.02272	0.7943	0.6703
$\alpha = 2$	2000	0.00800	0.00905	0.00993	0.8078	0.7661
	4000	0.00491	0.00514	0.00658	0.9242	0.7475

Table 3. Example 2. The MASEs and $\overline{\text{EFF}}$s of $\tilde{m}_{\text{K},1}$, $\hat{m}_{\text{SBK},1}$, $\hat{m}_{\text{GAM},1}$ for $d = 10$, $n = 500, 1000, 2000, 4000$

$d = 10$	n	MASE($\tilde{m}_{\text{K},1}$)	MASE($\hat{m}_{\text{SBK},1}$)	MASE($\hat{m}_{\text{GAM},1}$)	$\overline{\text{EFF}}$($\hat{m}_{\text{SBK},1}$)	$\overline{\text{EFF}}$($\hat{m}_{\text{GAM},1}$)
$r = 0$,	500	0.06066	0.06810	0.07280	0.9347	0.8149
$a = 0$	1000	0.03767	0.04421	0.05070	0.9809	0.6132
	2000	0.02097	0.02368	0.04509	0.9729	0.4089
	4000	0.01482	0.01630	0.04484	0.9873	0.3831
$r = 0$,	500	0.06589	0.07275	0.07519	0.8831	0.8373
$a = 0.5$	1000	0.04100	0.04328	0.04905	0.8967	0.5563
	2000	0.02561	0.02776	0.04574	0.9532	0.6203
	4000	0.01571	0.01733	0.04200	0.9736	0.4150
$r = 0.5$,	500	0.14958	0.16693	0.18963	0.9075	0.8766
$a = 0$	1000	0.08199	0.08734	0.09359	0.9135	0.8738
	2000	0.04976	0.05327	0.05946	0.9880	0.8856
	4000	0.02989	0.03176	0.04182	0.9920	0.7270
$r = 0.5$,	500	0.15187	0.16416	0.15622	0.9012	0.9273
$a = 0.5$	1000	0.08447	0.08762	0.08947	0.9382	0.9141
	2000	0.05846	0.06022	0.06586	0.9635	0.7803
	4000	0.02838	0.02942	0.04180	0.9895	0.6462

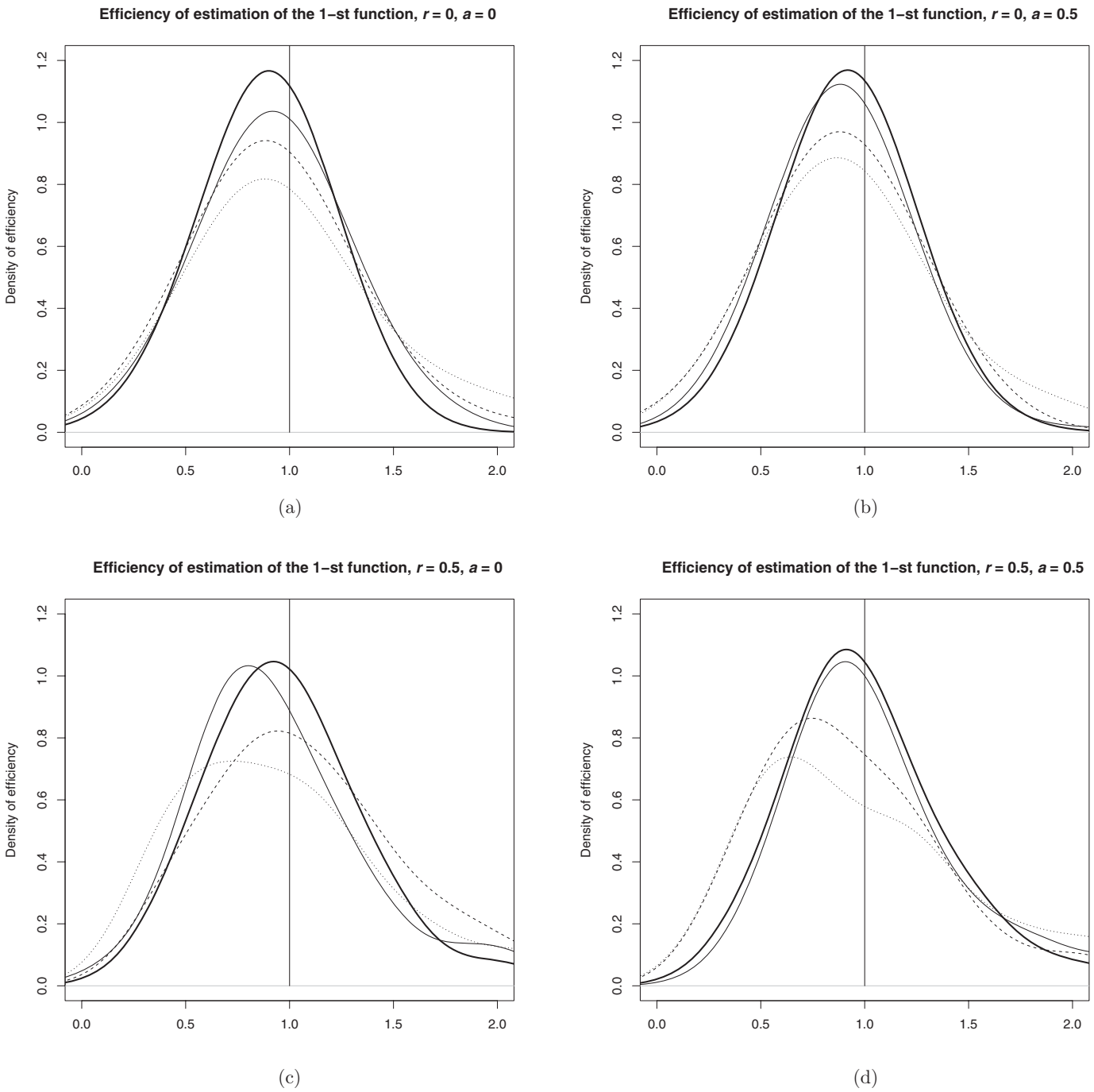


Figure 2. Example 2. Empirical distributions of relative efficiency of $n = 500$ (dotted line), $n = 1000$ (dashed line), $n = 2000$ (thin solid line), $n = 4000$ (thick solid line) for (a) $r = 0, a = 0$; (b) $r = 0, a = 0.5$; (c) $r = 0.5, a = 0$; (d) $r = 0.5, a = 0.5$.

oracle estimators (dashed lines) for the true functions m_1 (thick lines) in Figure 3. The results are satisfactory and show that the estimator works well as the asymptotic theory indicates and that performance improves with increasing sample size.

6.3 Example 3

We now examine a sparse high-dimensional version of Example 2, with $d = 50$ and $m_\alpha(x_\alpha) = -\sin(2\pi x_\alpha)$, $\alpha = 1, \ldots, 5$ and $m_\alpha(x_\alpha) = 0$, $\alpha = 6, \ldots, 50$. We report the efficiency and computing time of the SBK estimator $\hat{m}_{\mathrm{SBK},1}$ and how it compares with $\hat{m}_{\mathrm{GAM},1}$.

Numerical comparison of the efficiencies are in Table 4, while computing time comparison is given in Table 5. These two tables show that $\hat{m}_{\mathrm{SBK},1}$ is asymptotically more efficient and computationally much faster than $\hat{m}_{\mathrm{GAM},1}$ in this sparse high-dimensional case and much more strikingly so for larger sample sizes $n = 2000$, 4000. These numerical observations corroborate with the oracle efficiency in Theorem 4 and what is known about SBK method for additive model, see Wang and Yang (2007) and Liu and Yang (2010).

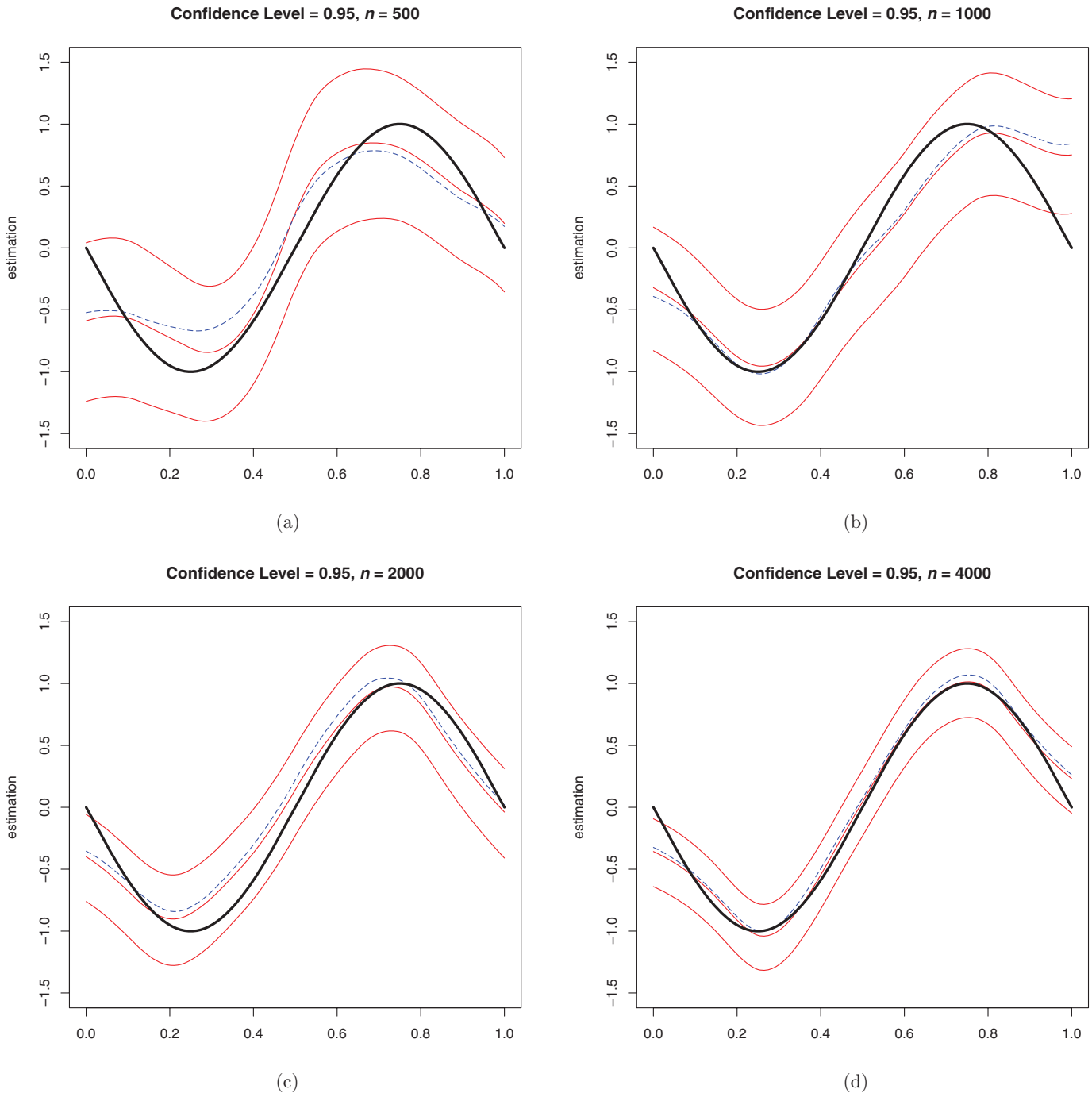


Figure 3. Example 2. Plots of $m_1(x_1)$ (thick line), $\tilde{m}_{\mathrm{K},1}(x_1)$ (dashed line), asymptotic 95% pointwise confidence intervals and $\hat{m}_{\mathrm{SBK},1}(x_1)$ (solid lines) for $r = 0.5$, $a = 0.5$ and (a) $n = 500$, (b) $n = 1000$, (c) $n = 2000$, (d) $n = 4000$.

6.4 Example 4

We revisit the credit reform data discussed in the introduction. After excluding the missing values, it contains financial information from 18,610 solvent ($Y = 0$) and 1000 insolvent ($Y = 1$) German companies. The time period ranges from 1997 to 2002, and in the case of the insolvent companies, the information was gathered 2 years before the insolvency took place. To satisfy Assumption (A4), we make the following transformation: $X_{i\alpha} = F_{n\alpha}(Z_{i\alpha})$, $\alpha = 1, \ldots, 8$, where $F_{n\alpha}$ is the empirical cdf for the data $\{X_{i\alpha}\}_{i=1}^{n}$.

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 cumulative accuracy profile (CAP) curve as

$$\mathrm{CAP}(u) = F_{\mathrm{D}}(F^{-1}(u)), u \in (0, 1), \tag{14}$$

which is the percentage of default-infected obligators that are found among the first (according to their scores) $100u\%$ of all obligators. A perfect rating method assigns all lowest scores to exactly the defaulters, so its CAP curve linearly increases up and then stays at 1, in other words, $\mathrm{CAP}_{\mathrm{P}}(u) = \min(u/p, 1)$,

Table 4. Example 3. The MASEs and $\overline{\text{EFF}}$s of $\tilde{m}_{\text{K},1}$, $\hat{m}_{\text{SBK},1}$, $\hat{m}_{\text{GAM},1}$ for $d = 50$, $n = 500, 1000, 2000, 4000$

$d = 50$	n	MASE($\tilde{m}_{\text{K},1}$)	MASE($\hat{m}_{\text{SBK},1}$)	MASE($\hat{m}_{\text{GAM},1}$)	$\overline{\text{EFF}}(\hat{m}_{\text{SBK},1})$	$\overline{\text{EFF}}(\hat{m}_{\text{GAM},1})$
$r = 0$,	500	0.06570	0.09497	0.12028	0.7342	0.6313
$a = 0$	1000	0.04222	0.05785	0.06385	0.8910	0.7345
	2000	0.02531	0.02617	0.03149	0.9235	0.7619
	4000	0.01392	0.01547	0.02967	0.9734	0.5138
$r = 0$,	500	0.08799	0.17948	0.33984	0.5542	0.3342
$a = 0.5$	1000	0.05533	0.06993	0.10952	0.6743	0.5817
	2000	0.03237	0.04170	0.04483	0.8066	0.7294
	4000	0.01839	0.01893	0.03170	0.9756	0.5718
$r = 0.5$,	500	0.06932	0.10792	0.14825	0.6976	0.5497
$a = 0$	1000	0.04075	0.04501	0.06064	0.8789	0.7101
	2000	0.02458	0.02559	0.03168	0.9754	0.7515
	4000	0.01514	0.01648	0.02982	0.9645	0.5610
$r = 0.5$,	500	0.08334	0.19293	0.33611	0.5780	0.3610
$a = 0.5$	1000	0.05197	0.07392	0.11177	0.8005	0.4697
	2000	0.03097	0.03880	0.04224	0.9234	0.7816
	4000	0.01871	0.01930	0.03139	0.9937	0.5975

$u \in (0, 1)$, where p denotes the unconditional default probability. In contrast, a noninformative rating method with zero discriminatory power displays a diagonal line $\text{CAP}_{\text{N}}(u) \equiv u$, $u \in (0, 1)$. The CAP curve of a given scoring method S always locates between these two extremes.

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

$$\text{AR} = \frac{a_R}{a_P} = \frac{2\int_0^1 \text{CAP}(u)du - 1}{1 - p}, \tag{15}$$

with $\text{CAP}(u)$ given in (14). The AR takes value in [0, 1], with 0 corresponding to the noninformative scoring, and 1 the perfect scoring method; a higher AR indicates higher discriminatory power of a method. In this study, we compute the GAM SBK score $S = b'\{\hat{c} + \sum_{\alpha=1}^{8} \hat{m}_{\text{SBK},\alpha}(X_\alpha)\}$, $b'(x) = e^x/(1 + e^x)$.

For the RDC data, our analysis has the AR value 62.46%, better than the AR value 60.51% obtained in Härdle, Hoffmann, and Moro (2011). This is clearly due to the fact that our credit score function depends on each variate X_α, $1 \le \alpha \le d$, nonparametrically via the GAM. The score function used in Härdle, Hoffmann, and Moro (2011), on the other hand, is a linear function of X_α, $1 \le \alpha \le d$, thus lacking flexibility. Our AR value of 62.46% is also higher than the AR value 58.69% obtained using the GAM procedure in R. We can also estimate the functions $m_\alpha(x_\alpha)$ for X_α. The effects of $X_3 = \text{Ebit/Total\_Assets}$ and $X_8 = \log(\text{Total\_Assets})$, which are estimates for $m_3(x_3)$ and $m_8(x_8)$, respectively, are shown in Figure 1. It is no surprise that the estimator $\hat{m}_{\text{SBK},8}(x_8)$ for $m_8(x_8)$ decreases as x_8 value increases. It means that a company with more Total_Assets has smaller probability of insolvent. While as x_3 value increases, the estimator $\hat{m}_{\text{SBK},3}(x_3)$ for $m_3(x_3)$ increases for most part but decreases at the end. So generally, companies with higher Ebit/Total_Assets ratio have greater probability of insolvency. It appears that companies with extremely high Ebit/Total_Assets ratio have smaller probability of insolvency; the underlying reason of which requires further investigation.

Table 5. Example 3. The average computing time (in seconds per replication) of $\hat{m}_{\text{SBK},1}$, $\hat{m}_{\text{GAM},1}$ for $d = 50$, $n = 500, 1000, 2000, 4000$

$d = 50$	n	SBK	GAM	ratio
$r = 0$,	500	0.16	0.86	1:5.3
$a = 0$	1000	0.36	2.61	1:7.2
	2000	0.80	6.52	1:8.1
	4000	2.11	41.2	1:19.5
$r = 0.5$,	500	0.17	0.94	1:5.5
$a = 0.5$	1000	0.37	3.22	1:8.6
	2000	0.85	7.73	1:9.1
	4000	2.21	46.5	1:20.9

APPENDIX

A.1 Preliminaries

In the proofs that follow, we use "$\mathcal{U}$" and "u" to denote sequences of random variables that are uniformly "$\mathcal{O}$" and "o" of certain order.

Lemma A.1. (Sunklodas 1984, theorem 1). Let $\{\xi_i\}_{i=1}^n$ be an α-mixing sequence with $\text{E}\xi_n = 0$. Denote $d_\delta = \max_{1 \le i \le n}\{\text{E}|\xi_i|^{2+\delta}\}$, $0 < \delta \le 1$, $S_n = \sum_{i=1}^n \xi_i$, $\sigma_n^2 \stackrel{\text{def}}{=} \text{E}S_n^2 \ge c_0 n$ for some $c_0 \in (0, +\infty)$. If $\alpha(n) \le K_0 \exp(-\lambda_0 n)$, $\lambda_0 > 0$, $K_0 > 0$, then $c_1 = c_1(K_0, \delta)$, $c_2 = c_2(K_0, \delta)$ exist such that

$$\Delta_n = \sup_z \left| \text{P}\left\{\sigma_n^{-1} S_n < z\right\} - \Phi(z)\right| \le c_1 \frac{d_\delta}{c_0 \sigma_n^\delta} \left\{ \log\left(\sigma_n / c_0^{1/2}\right)/\lambda \right\}^{1+\delta} \tag{A.1}$$

for any λ with $\lambda_1 \le \lambda \le \lambda_2$, where

$$\lambda_1 = c_2 \left\{ \log\left(\sigma_n / c_0^{1/2}\right) \right\}^b / n, \; b > 2(1 + \delta)/\delta;$$
$$\lambda_2 = 4(2 + \delta)\delta^{-1} \log\left(\sigma_n / c_0^{1/2}\right).$$

Lemma A.2. (Bernstein's inequality, Bosq 1998, theorem 1.4). Let $\{\xi_i\}$ be a zero-mean real-valued process. Suppose that there exists $c > 0$ such that for $i = 1, \ldots, n$, $k \geq 3$, $\mathrm{E}|\xi_i|^k \leq c^{k-2}k!\mathrm{E}\,\xi_i^2 < +\infty$, $m_r = \max_{1\leq i\leq n} \|\xi_i\|_r$, $r \geq 2$. Then for each $n > 1$, integer $q \in [1, n/2]$, each $\varepsilon > 0$ and $k \geq 3$

$$\mathrm{P}\left\{\left|\sum_{i=1}^{n}\xi_i\right| > n\varepsilon\right\} \leq a_1 \exp\left(-\frac{q\varepsilon^2}{25m_2^2 + 5c\varepsilon}\right) + a_2(k)\alpha\left(\left[\frac{n}{q+1}\right]\right)^{\frac{2k}{2k+1}},$$

where

$$a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\varepsilon^2}{25m_2^2 + 5c\varepsilon}\right), \quad a_2(k) = 11n\left(1 + \frac{5m_k^{2k/(2k+1)}}{\varepsilon}\right).$$

Denote the theoretical inner product of b_J and 1 with respect to the αth marginal density $f_\alpha(x_\alpha)$ as $c_{J,\alpha} = \langle b_J(X_\alpha), 1\rangle = \int b_J(x_\alpha)f_\alpha(x_\alpha)dx_\alpha$ and define the centered B-spline basis $b_{J,\alpha}(x_\alpha)$ and the standardized B-spline basis $B_{J,\alpha}(x_\alpha)$ as

$$b_{J,\alpha}(x_\alpha) = b_J(x_\alpha) - \frac{c_{J,\alpha}}{c_{J-1,\alpha}}b_{J-1}(x_\alpha),$$
$$B_{J,\alpha}(x_\alpha) = \frac{b_{J,\alpha}(x_\alpha)}{\|b_{J,\alpha}\|_2}, \quad 1 \leq J \leq N+1,$$

so that $\mathrm{E}\,B_{J,\alpha}(X_\alpha) = 0$, $\mathrm{E}\,B_{J,\alpha}^2(X_\alpha) = 1$.

Lemma A.3. (Wang and Yang 2007, theorem A.2). Under Assumptions (A1)–(A5) and (A7), one has the following:

1. constants $c_0(f)$, $C_0(f)$, $c_1(f)$, and $C_1(f)$ exist depending on the marginal densities $f_\alpha(x_\alpha)$, $1 \leq \alpha \leq d$, such that $c_0(f)H \leq c_{J,\alpha} \leq C_0(f)H$ and
$$c_1(f)H \leq \|b_{J,\alpha}\|_2^2 \leq C_1(f)H, \tag{A.2}$$
2. uniformly for $J, J' = 1, \ldots, N+1$
$$\mathrm{E}\{B_{J,\alpha}(X_{i\alpha})B_{J',\alpha}(X_{i\alpha})\} \sim \begin{cases} 1 & J' = J \\ -1/3 & |J'-J| = 1 \\ 1/6 & |J'-J| = 2 \\ 0 & |J'-J| > 2 \end{cases}$$
$$\mathrm{E}\,|B_{J,\alpha}(X_{i\alpha})B_{J',\alpha}(X_{i\alpha})|^k \sim \begin{cases} H^{1-k} & |J'-J| \leq 2 \\ 0 & |J'-J| > 2 \end{cases}, \quad k \geq 1.$$

Lemma A.4. (de Boor 2001, p. 149). A constant $C_\infty > 0$ exists such that for any $m \in C^1[0,1]$ with $m' \in \mathrm{Lip}([0,1], C_\infty)$, there is a function $g \in G_n^{(0)}[0,1]$ such that $\|g - m\|_\infty \leq C_\infty H^2$.

Lemma A.5. (Wang and Yang 2007, lemma A.2). Constants $c_0, C_0 > 0$ exist such that for any $\lambda = (\lambda_0, \lambda_{J,\alpha})_{1\leq J\leq N+1, 1\leq\alpha\leq d}^{\mathsf{T}} \in \mathbb{R}^{1+d(N+1)}$,

$$c_0\left(\lambda_0^2 + \sum_{J,\alpha}^{2}\lambda_{J,\alpha}^2\right) \leq \left\|\lambda_0 + \sum_{J,\alpha}\lambda_{J,\alpha}B_{J,\alpha}\right\|_2^2 \leq C_0\left(\lambda_0^2 + \sum_{J,\alpha}^{2}\lambda_{J,\alpha}^2\right).$$

Lemma A.6. (Xue and Yang 2006a, lemma A.4). Under Assumptions (A2), (A4), and (A6), as $n \to \infty$, the uniform supremum of the rescaled difference between $\langle g_1, g_2\rangle_{2,n}$ and $\langle g_1, g_2\rangle_2$ is

$$A_n = \sup_{g_1,g_2\in G_n^{(0)}[0,1]} \frac{|\langle g_1, g_2\rangle_{2,n} - \langle g_1, g_2\rangle_2|}{\|g_1\|_2\|g_2\|_2} = \mathcal{O}_{\mathrm{a.s.}}\left(\frac{\log n}{n^{1/2}H^{1/2}}\right).$$

A.2 Oracle Smoothers

Lemma A.7. Under Assumptions (A1)–(A6), as $n \to \infty$,

$$\sup_{x_1\in[h,1-h]}\left|\tilde{l}'\{m_1(x_1)\} - \mathrm{bias}_1(x_1)h^2 - n^{-1}\sum_{i=1}^{n}K_h(X_{i1} - x_1)\sigma(\mathbf{X}_i)\varepsilon_i\right| = \mathcal{O}_{\mathrm{a.s.}}(n^{-1/2}h^{1/2}\log n),$$

where $\mathrm{bias}_1(x_1)$ is defined in (9).

Proof. See the online supplementary materials. □

Lemma A.8. Under Assumptions (A2) and (A4)–(A6), as $n \to \infty$,

$$\sup_{x_1\in[h,1-h]}|\tilde{l}''(m_1(x_1)) + D_1(x_1)| = \mathcal{O}_{\mathrm{a.s.}}(\log n/\sqrt{nh}),$$

where $D_1(x_1)$ is defined in (8).

Proof. See the online supplementary materials. □

Lemma A.9. Under Assumptions (A1)–(A3), (A5), and (A7), a constant C exists such that, as $n \to \infty$,

$$\sup_{x_1\in[h,1-h]}|\mathrm{cov}(\xi_{i,n}, \xi_{j,n})| \leq Ch^{-\frac{1+\eta}{2+\eta}}\alpha(j-i)^{\frac{\eta}{2+\eta}} \text{ for } i \neq j$$

Proof. See the online supplementary materials. □

Proof of Theorems 1 and 2. See the online supplementary materials. □

Proof of Theorem 3. According to the mean value theorem, a constant $\bar{c}$ between c and $\tilde{c}$ exists such that $(\tilde{c} - c)\tilde{l}_c''(\bar{c}) = \tilde{l}_c'(\tilde{c}) - \tilde{l}_c'(c) = -\tilde{l}_c'(c)$, where $-\tilde{l}_c''(\bar{c}) = n^{-1}\sum_{i=1}^{n}b''\{\bar{c} + m_{-c}(\mathbf{X}_i)\} > c_b > 0$ according to (A2) and where $m_{-c}(\mathbf{X}) = \sum_{\alpha=1}^{d}m_\alpha(X_\alpha)$ and then the infeasible estimator is $\tilde{c} = \arg\max_{a\in A}\tilde{l}_c(a)$. Clearly, $\tilde{l}_c'(\tilde{c}) = 0$ and

$$\tilde{l}_c'(c) = n^{-1}\sum_{i=1}^{n}[Y_i - b'\{c + m_{-c}(\mathbf{X}_i)\}] = n^{-1}\sum_{i=1}^{n}\sigma(\mathbf{X}_i)\varepsilon_i = \mathcal{O}_{\mathrm{a.s.}}(n^{-1/2}\log n)$$

by Bernstein's inequality. Similarly, $\tilde{l}_c''(c) = -n^{-1}\sum_{i=1}^{n}b''\{c + m_{-c}(\mathbf{X}_i)\}$ converges to $-\mathrm{E}\,b''\{m(\mathbf{X})\}$ almost surely at the rate of $n^{-1/2}\log n$. These imply that $|\tilde{c} - c| = \mathcal{O}_{\mathrm{a.s.}}(n^{-1/2}\log n)$ and plugging it into $(\tilde{c} - c) = -\tilde{l}_c'(c)/\tilde{l}_c''(\bar{c})$, Theorem 3 is proved. □

A.3 Spline-Backfitted Kernel Estimators

In this section, we present the proof of Theorem 4. We write any $g \in G_n^0$ as $g = \lambda^{\mathsf{T}}\mathbf{B}(\mathbf{X}_i)$ with vector $\lambda = (\lambda_0, \lambda_{J,\alpha})_{1\leq J\leq N+1, 1\leq\alpha\leq d}^{\mathsf{T}} \in R^{N_d}$ where $N_d = (N+1)d + 1$ is the dimension of the additive spline space G_n^0, and

$$\mathbf{B}(\mathbf{x}) = \{1, B_{1,1}(x_1), \ldots, B_{N+1,d}(x_d)\}^{\mathsf{T}}$$

its standardized basis. We denote with a slight abuse of notation $\hat{L}(g) = \hat{L}(\lambda) = n^{-1}\sum_{i=1}^{n}[Y_i\lambda^{\mathsf{T}}\mathbf{B}(\mathbf{X}_i) - b\{\lambda^{\mathsf{T}}\mathbf{B}(\mathbf{X}_i)\}]$, which yields the gradient and Hessian formulas

$$\nabla\hat{L}(\lambda) = n^{-1}\sum_{i=1}^{n}[Y_i\mathbf{B}(\mathbf{X}_i) - b'\{\lambda^{\mathsf{T}}\mathbf{B}(\mathbf{X}_i)\}\mathbf{B}(\mathbf{X}_i)],$$
$$\nabla^2\hat{L}(\lambda) = -n^{-1}\sum_{i=1}^{n}b''\{\lambda^{\mathsf{T}}\mathbf{B}(\mathbf{X}_i)\}\mathbf{B}(\mathbf{X}_i)\mathbf{B}(\mathbf{X}_i)^{\mathsf{T}}.$$

The multivariate function $m(\mathbf{x})$ is estimated by an additive spline function

$$\hat{m}(\mathbf{x}) = \hat{m}_0 + \sum_{\alpha=1}^{d} \hat{m}_\alpha(x_\alpha) = \hat{\lambda}^\mathsf{T}\mathbf{B}(\mathbf{x}),$$
$$\hat{\lambda} = (\hat{\lambda}_0, \hat{\lambda}_{J,\alpha})^\mathsf{T}_{\substack{1\le\alpha\le d\\ 1\le J\le N+1}} = \arg\max_\lambda \hat{L}(\lambda).$$

Lemma 14 of Stone (1986) ensures that with probability approaching 1, $\hat{\lambda}$ exists uniquely and that $\nabla\hat{L}(\hat{\lambda}) = \mathbf{0}$. In addition, Lemma A.4 and (A1) provide a vector $\bar{\lambda}$ and an additive spline function $\bar{m}$ such that

$$\bar{m}(\mathbf{x}) = \bar{\lambda}^\mathsf{T}\mathbf{B}(\mathbf{x}),\ \|\bar{m} - m\|_\infty \le C_\infty H^2. \tag{A.3}$$

We first establish technical lemmas before proving Theorems 4 and 5.

Lemma A.10. Under Assumptions (A1)–(A5) and (A7), as $n \to \infty$,

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

Proof. See the online supplementary materials. □

Define the following matrices:

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

and similar matrices

$$\mathbf{V}_b = \mathrm{E}\,b''\{m(\mathbf{X})\}\mathbf{B}(\mathbf{X})\mathbf{B}(\mathbf{X})^\mathsf{T} = \begin{bmatrix} v_{b,00} & v_{b,0,J,\alpha} \\ v_{b,0,J',\alpha'} & v_{b,J,\alpha,J',\alpha'} \end{bmatrix}_{N_d\times N_d}$$
$$\mathbf{S}_b = \mathbf{V}_b^{-1} = \begin{bmatrix} s_{b,00} & s_{b,0,J,\alpha} \\ s_{b,0,J',\alpha'} & s_{b,J,\alpha,J',\alpha'} \end{bmatrix}_{N_d\times N_d}. \tag{A.4}$$

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

$$\mathbf{V}_b(\lambda) = \mathrm{E}\,b''\{\lambda^\mathsf{T}\mathbf{B}(\mathbf{X})\}\mathbf{B}(\mathbf{X})\mathbf{B}(\mathbf{X})^\mathsf{T} = \begin{bmatrix} v_{b,00}(\lambda) & v_{b,0,J,\alpha}(\lambda) \\ v_{b,0,J',\alpha'}(\lambda) & v_{b,J,\alpha,J',\alpha'}(\lambda) \end{bmatrix}_{N_d\times N_d},$$
$$\mathbf{S}_b(\lambda) = \mathbf{V}_b^{-1}(\lambda) = \begin{bmatrix} s_{b,00}(\lambda) & s_{b,0,J,\alpha}(\lambda) \\ s_{b,0,J',\alpha'}(\lambda) & s_{b,J,\alpha,J',\alpha'}(\lambda) \end{bmatrix}_{N_d\times N_d}$$
$$\mathbf{V}_{n,b}(\lambda) = -\nabla^2\hat{L}(\lambda), \quad \mathbf{S}_{n,b}(\lambda) = \mathbf{V}_{n,b}^{-1}(\lambda). \tag{A.5}$$

Lemma A.11. Under Assumptions (A2) and (A4),

$$c_\mathbf{V}\mathbf{I}_{N_d} \le \mathbf{V} \le C_\mathbf{V}\mathbf{I}_{N_d}, c_\mathbf{S}\mathbf{I}_{N_d} \le \mathbf{S} \le C_\mathbf{S}\mathbf{I}_{N_d}, \tag{A.6}$$
$$c_{\mathbf{V},b}\mathbf{I}_{N_d} \le \mathbf{V}_b \le C_{\mathbf{V},b}\mathbf{I}_{N_d}, c_{\mathbf{S},b}\mathbf{I}_{N_d} \le \mathbf{S}_b \le C_{\mathbf{S},b}\mathbf{I}_{N_d}. \tag{A.7}$$

Under Assumption (A2), (A4), (A5), and (A7), as $n \to \infty$, with probability increasing to 1,

$$c_\mathbf{V}\mathbf{I}_{N_d} \le \mathbf{V}_n(\lambda) \le C_\mathbf{V}\mathbf{I}_{N_d}, c_\mathbf{S}\mathbf{I}_{N_d} \le \mathbf{S}_n(\lambda) \le C_\mathbf{S}\mathbf{I}_{N_d} \tag{A.8}$$
$$c_{\mathbf{V},b}\mathbf{I}_{N_d} \le \mathbf{V}_{n,b}(\lambda) \le C_{\mathbf{V},b}\mathbf{I}_{N_d}, c_{\mathbf{S},b}\mathbf{I}_{N_d} \le \mathbf{S}_{n,b}(\lambda) \le C_{\mathbf{S},b}\mathbf{I}_{N_d}. \tag{A.9}$$

Proof. For (A.6), see lemma A.9 in Wang and Yang (2007), while (A.8) follows from Lemma A.6. The statements (A.8) and (A.9) follow from (A.6) and (A.8), together with the boundedness of b'' in (A2). □

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

$$\mathbf{\Phi}_b = (\Phi_{b,0}, \Phi_{b,J,\alpha})^\mathsf{T}_{\substack{1\le J\le N+1\\ 1\le\alpha\le d}}$$
$$= -\mathbf{S}_b n^{-1}\sum_{i=1}^{n}[b'\{m(\mathbf{X}_i)\} - b'\{\bar{m}(\mathbf{X}_i)\}]\mathbf{B}(\mathbf{X}_i), \tag{A.10}$$
$$\mathbf{\Phi}_v = (\Phi_{v,0}, \Phi_{v,J,\alpha})^\mathsf{T}_{\substack{1\le J\le N+1\\ 1\le\alpha\le d}}$$
$$= -\mathbf{S}_b n^{-1}\sum_{i=1}^{n}[\sigma(\mathbf{X}_i)\varepsilon_i]\mathbf{B}(\mathbf{X}_i), \tag{A.11}$$
$$\mathbf{\Phi}_r = (\Phi_{r,0}, \Phi_{r,J,\alpha})^\mathsf{T}_{\substack{1\le J\le N+1\\ 1\le\alpha\le d}}$$
$$= \hat{\lambda} - \bar{\lambda} - \Phi_b - \mathbf{\Phi}_v. \tag{A.12}$$

Lemma A.12. Under Assumptions (A1)–(A5) and (A7), as $n \to \infty$,

$$\|\hat{\lambda} - \bar{\lambda}\| = \mathcal{O}_{\text{a.s.}}(H^2 + H^{-1/2}n^{-1/2}\log n), \tag{A.13}$$
$$\|\mathbf{\Phi}_r\| = \mathcal{O}_{\text{a.s.}}(H^{-3/2}n^{-1}\log n),$$
$$\|\mathbf{\Phi}_b\| = \mathcal{O}_{\text{a.s.}}(H^2), \|\mathbf{\Phi}_v\| = \mathcal{O}_{\text{a.s.}}(H^{-1/2}n^{-1/2}\log n). \tag{A.14}$$

Proof. See the online supplementary materials. □

Lemma A.13. Under Assumptions (A1)–(A5) and (A7), as $n \to \infty$,

$$\|\hat{m} - \bar{m}\|_\infty + \sum_{\alpha=1}^{d}\|\hat{m}_\alpha - \bar{m}_\alpha\|_\infty = \mathcal{O}_{\text{a.s.}}(H^{3/2} + H^{-1}n^{-1/2}\log n),$$
$$\|\hat{m} - \bar{m}\|_{2,n} + \|\hat{m} - \bar{m}\|_2 = \mathcal{O}_{\text{a.s.}}(H^2 + H^{-1/2}n^{-1/2}\log n),$$
$$\|\hat{m} - m\|_\infty + \sum_{\alpha=1}^{d}\|\hat{m}_\alpha - m_\alpha\|_\infty = \mathcal{O}_{\text{a.s.}}(H^{3/2} + H^{-1}n^{-1/2}\log n),$$
$$\|\hat{m} - m\|_{2,n} + \|\hat{m} - m\|_2 = \mathcal{O}_{\text{a.s.}}(H^2 + H^{-1/2}n^{-1/2}\log n).$$

Proof. See the online supplementary materials. □

In the following, denote

$$\omega(x_1) = \{\omega_{J,\alpha}(x_1)\}_{J=1,\alpha=2}^{N+1,d},$$
$$\omega_{J,\alpha}(x_1) = n^{-1}\sum_{i=1}^{n}|B_{J,\alpha}(X_{i\alpha})|K_h(X_{i1} - x_1).$$

Lemma A.14. Under Assumptions (A1)–(A7), as $n \to \infty$,

$$\sup_{\substack{x_1\in[0,1],2\le\alpha\le d\\ 1\le J\le N+1}}|\omega_{J,\alpha}(x_1) - \mathrm{E}\,\omega_{J,\alpha}(x_1)| = \mathcal{O}_{\text{a.s.}}(\log n/\sqrt{nh}) \tag{A.15}$$
$$\sup_{x_1\in[0,1]}|\omega(x_1)| = \sup_{\substack{x_1\in[0,1],2\le\alpha\le d\\ 1\le J\le N+1,}}|\omega_{J,\alpha}(x_1)| = \mathcal{O}_{\text{a.s.}}(H^{1/2}). \tag{A.16}$$

Proof. See the online supplementary materials. □

Lemma A.15. Under Assumptions (A1)–(A7), as $n \to \infty$,

$$\sup_{x_1\in[0,1]}|\hat{l}'\{\bar{m}_{\mathrm{K},1}(x_1)\}| = \mathcal{O}_{\text{a.s.}}\left(n^{-1/2}\log n\right).$$

Proof. See the online supplementary materials. □

Lemma A.16. Under Assumptions (A1)–(A7), constants c, C exist such that $0 < c \le |-\hat{l}''(a, x_1)| \le C < \infty$ a.s. for $a \in A, x_1 \in [0, 1]$.

Proof. See the online supplementary materials. □

Proof of Theorem 4. According to (12) and the mean value theorem, a $\bar{m}_{\mathrm{K},1}(x_1)$ between $\hat{m}_{\mathrm{SBK},1}(x_1)$ and $\bar{m}_{\mathrm{K},1}(x_1)$ exists such that

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

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

$$\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1) = -\frac{\hat{l}'\{\tilde{m}_{\mathrm{K},1}(x_1)\}}{\hat{l}''\{\tilde{m}_{\mathrm{K},1}(x_1)\}}. \tag{A.17}$$

The theorem then follows from Lemmas A.15 and A.16. □

Proof of Theorem 5. See the online supplementary materials. □

SUPPLEMENTARY MATERIALS

Supplement to "Oracally efficient two-step estimation of generalized additive model": Supplement containing theoretical proofs referenced in the main article. (PDF file)

gamsbk.R: R-package containing code to perform SBK estimation for component functions in generalized additive model.

[Received January 2012. Revised January 2013.]

REFERENCES

Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, New York: Springer-Verlag. [629]

de Boor, C. (2001), *A Practical Guide to Splines*, New York: Springer-Verlag. [629]

Fan, J., Härdle, W., and Mammen, E. (1998), "Direct Estimation of Low-Dimensional Components in Additive Models," *The Annals of Statistics*, 26, 943–971. [619]

Härdle, W. (1989), "Asymptotic Maximal Deviation of M-Smoothers," *Journal of Multivariate Analysis*, 29, 163–179. [621]

Härdle, W., Hoffmann, L., and Moro, R. (2011), *Learning Machines Supporting Bankruptcy Prediction. Statistical Tools in Finance and Insurance* (2nd ed.), eds. P., Cizek, W., Härdle, R., Weron, Heidelberg: Springer Verlag. [620,628]

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall. [619]

Horowitz, J., Klemelä, J., and Mammen, E. (2006), "Optimal Estimation in Additive Regression," *Bernoulli*, 12, 271–298. [620]

Horowitz, J., and Mammen, E. (2004), "Nonparametric Estimation of an Additive Model With a Link Function," *The Annals of Statistics*, 32, 2412–2443. [620,621]

Huang, J. Z., and Yang, L. (2004), "Identification of Nonlinear Additive Autoregression Models," *Journal of the Royal Statistical Society,* Series B, 66, 463–477. [619]

Linton, O. B. (1997), "Efficient Estimation of Additive Nonparametric Regression Models," *Biometrika*, 84, 469–473. [619]

Linton, O. B., and Härdle, W. (1996), "Estimation of Additive Regression Models With Known Links," *Biometrika*, 83, 529–540. [620]

Linton, O. B., and Nielsen, J. P. (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration," *Biometrika*, 82, 93–100. [619]

Liu, R., and Yang, L. (2010), "Spline-Backfitted Kernel Smoothing of Additive Coefficient Model," *Econometric Theory*, 26, 29–59. [619,620,621,622,626]

Ma, S., and Yang, L. (2011), "Spline-Backfitted Kernel Smoothing of Partially Linear Additive Model," *Journal of Statistical Planning and Inference*, 141, 204–219. [619,620]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society,* Series B, 71, 1009–1030. [623]

Severini, T., and Staniswalis, J. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511. [621]

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. [619,620]

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [619,621]

——— (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics*, 14, 590–606. [620,621,622,630]

——— (1994), "The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation," *The Annals of Statistics*, 22, 118–184. [619]

Sunklodas, J. (1984), "On the Rate of Convergence in the Central Limit Theorem for Strongly Mixing Random Variables," *Lithuanian Mathematical Journal*, 24, 182–190. [628]

Tjøstheim, D., and Auestad, B. (1994), "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association*, 89, 1398–1409. [619]

Wang, J., and Yang, L. (2009), "Efficient and Fast Spline-Backfitted Kernel Smoothing of Additive Regression Model," *Annals of the Institute of Statistical Mathematics*, 61, 663–690. [619,620]

Wang, L., and Yang, L. (2007), "Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model," *The Annals of Statistics*, 35, 2474–2503. [619,620,621,626,629,630]

Xue, L., and Liang, H. (2010), "Polynomial Spline Estimation for a Generalized Additive Coefficient Model," *Scandinavian Journal of Statistics*, 37, 26–46. [620,621]

Xue, L., and Yang, L. (2006a), "Additive Coefficient Modeling via Polynomial Spline," *Statistica Sinica*, 16, 1423–1446. [619,629]

——— (2006b), "Estimation of Semiparametric Additive Coefficient Model," *Journal of Statistical Planning and Inference*, 136, 2506–2534. [619]

Yang, L., Härdle, W., and Nielsen, J. P. (1999), "Nonparametric Autoregression With Multiplicative Volatility and Additive Mean," *Journal of Time Series Analysis*, 20, 579–604. [619]

Yang, L., Park, B. U., Xue, L., and Härdle, W. (2006), "Estimation and Testing for Varying Coefficients in Additive Models With Marginal Integration," *Journal of the American Statistical Association*, 101, 1212–1227. [619]

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. [620]