Oracally efficient two-step estimation for additive regression

Shujie Ma
University of California at Riverside

Lijian Yang
Soochow University and Michigan State University

Contents

1 Introduction and Overview of Additive Regression 1

2 SBK in Additive Models 5
  2.1 The SBK estimator .............................................. 5
  2.2 Application to Boston housing data ......................... 9

3 SBK in Partially Linear Additive Models (PLAM) 10
  3.1 Application to Boston housing data ......................... 13

4 SBK in Additive Coefficient Models (ACM) 13
  4.1 Application to Cobb-Douglas model ......................... 19

5 SBS in Additive Models 20

6 Future Research 23

1 Introduction and Overview of Additive Regression

Linear regression is one of the most widely used technique for studying the relationship between a scalar variable $Y$ and a vector of independent variables $X = (X_1, \ldots, X_d)^T$. Given a data set $(Y_i, X_i^T)^T$ of $n$ subjects or experimental units, where $X_i = (X_{i1}, \ldots, X_{id})^T$, a linear model has the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_d X_{id} + \varepsilon_i, i = 1, \ldots, n,$$

where $\varepsilon_i$ is an unobserved random variable which adds noise to this relationship. Linear regression has gained its popularity because of its simplicity and
easy-to-interpret nature, but it has suffered from inflexibility in modeling possible complicated relationships between \(Y\) and \(X\). To avoid the strong linearity assumption and capture possible nonlinear relationships, nonparametric models were proposed and have gained much attention in the last three decades. In nonparametric models, the response \(Y\) depends on the explanatory variables \(X\) through a nonlinear function \(m(\cdot)\) such that

\[
Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{2}
\]

The functional form of \(m(\cdot)\) is not predetermined, which is estimated from the data, so that we can let the data speak for themselves. Under smoothness condition, the unknown function can be estimated nonparametrically by such methods as kernel and spline smoothing.

Nonparametric modeling imposes no specific model structure and enables one to explore the data more flexibly, but it does not perform well when the dimension of the predictor vector in the model is high. The variances of the resulting estimates tend to be unacceptably large due to the sparseness of data, which is the so-called "curse of dimensionality". To overcome these difficulties, Stone (1985a) proposed additive models. In model (2), the unknown function \(m(\cdot)\) is replaced by sum of univariate functions, so an additive model is given as

\[
Y_i = m(X_{i1}, \ldots, X_{id}) + \sigma(X_{i1}, \ldots, X_{id}) \varepsilon_i, \quad m(x_1, \ldots, x_d) = c + \sum_{\alpha=1}^{d} m_{\alpha}(x_{\alpha}), \tag{3}
\]

where \(m\) and \(\sigma\) are the mean and standard deviation of the response \(Y_i\) conditional on the predictor vector \(X_i\), and each \(\varepsilon_i\) is white noise conditional on \(X_i\). By definition of conditional mean and variance

\[
m(X_i) = E(Y_i | X_i), \quad \sigma^2(X_i) = \text{var}(Y_i | X_i), \quad i = 1, \ldots, n
\]

and so the error term \(\varepsilon_i = \{Y_i - m(X_i)\} \sigma^{-1}(X_i)\) accommodates the most general form of heteroskedasticity, as we do not assume independence of \(\varepsilon_i\) and \(X_i\) but only \(E(\varepsilon_i | X_i) \equiv 0, E(\varepsilon_i^2 | X_i) \equiv 1\). For identifiability, it is commonly assumed that \(Em_{\alpha}(X_{ia}) \equiv 0, \alpha = 1, \ldots, d\). Some other restrictions can also solve the identifiability problem such as by letting \(m_{\alpha}(0) = 0\), for \(\alpha = 1, \ldots, d\). Because the unknown functions \(m_{\alpha}(\cdot), 1 \leq \alpha \leq d\), are one-dimensional, the problem associated with the so-called "curse of dimensionality" is solved.

In model (3), each predictor \(X_{\alpha}, 1 \leq \alpha \leq d\), is required to be a continuous variable. In order to incorporate discrete variables, different forms of semiparametric models have been proposed, including partially linear additive models (PLAM) given as

\[
Y_i = m(X_i, T_i) + \sigma(X_i, T_i) \varepsilon_i, m(x, t) = c_{00} + \sum_{l=1}^{d_1} c_{0l} t_l + \sum_{\alpha=1}^{d_2} m_{\alpha}(x_{\alpha}) \tag{4}
\]
in which the sequence \( \{ Y_i, X_i^T, T_i^T \}_{i=1}^n = \{ Y_i, X_{i1}, ..., X_{id_2}, T_{i1}, ..., T_{id_1} \}_{i=1}^n \). For identifiability, we let both the additive and linear components be centered, i.e., \( E m_\alpha (X_{i\alpha}) \equiv 0, \alpha = 1, ..., d_2, E T_{il} \equiv 0, l = 1, ..., d_1 \). Model (4) is more parsimonious and easier to interpret than purely additive models (3) by allowing a subset of predictors \((T_{il})_{l=0}^{d_1}\) to be discrete, and more flexible than linear models (1) by allowing nonlinear relationships. To allow the coefficients of the linear predictors to change with some other variable, Xue and Yang (2006a), Xue and Yang (2006b), and Yang et al. (2006) proposed an additive coefficient model (ACM) that allows a response variable \( Y \) to depend linearly on some regressors, with coefficients as smooth additive functions of other predictors, called tuning variables. Specifically,

\[
Y_i = \sum_{l=1}^{d_1} m_l (X_i) T_{il}, \quad m_l (X_i) = m_{0l} + \sum_{\alpha=1}^{d_2} m_{\alpha l} (X_{i\alpha}), \quad 1 \leq l \leq d_1, \quad (5)
\]

in which the predictor vector \((X_i^T, T_i^T)^T\) consists of the tuning variables \( X_i = (X_{i1}, ..., X_{id_2})^T \) and linear predictors \( T_i = (T_{i1}, ..., T_{id_1})^T \).

Model (5)’s versatility for econometric applications is illustrated by the following example: Consider the forecasting of the U.S. GDP annual growth rate, which is modeled as the total factor productivity (TFP) growth rate plus a linear function of the capital growth rate and the labor growth rate, according to the classic Cobb-Douglas model (Cobb and Douglas (1928)). As pointed out in Li and Racine (2007) (p. 302), it is unrealistic to ignore the nonneutral effect of R&D spending on the TFP growth rate and on the complementary slopes of capital and labor growth rates. Thus, a smooth coefficient model should fit the production function better than the parametric Cobb-Douglas model. Indeed, Figure 1 (see Liu and Yang 2010) shows that a smooth coefficient model has much smaller rolling forecast errors than the parametric Cobb-Douglas model, based on data from 1959 to 2002. In addition, Figure 2 (see Liu and Yang 2010) shows that the TFP growth rate is a function of R&D spending, not a constant.

(Insert Figure 1 about here)

(Insert Figure 2 about here)

The additive model (3), the PLAM (4) and the ACM (5) achieve dimension reduction through representing the multivariate function of the predictors by sum of additive univariate functions. People have been making great efforts to develop statistical tools to estimate these additive functions. In review of literature, there are four types of kernel-based estimators: the classic backfitting estimators of Hastie and Tibshirani (1990) and Opsomer and Ruppert (1997); marginal integration estimators of Fan et al. (1997), Linton and Nielsen (1995), Linton and Härdle (1996), Kim et al. (1999), Sperlich et al. (2002), and Yang et al. (2003) and a kernel-based method of estimating rate to optimality of Hengartner and Sperlich (2005); the smoothing backfitting estimators of Mammen et al. (1999); and the two-stage estimators, such as one step backfitting of the integration estimators of Linton (1997), one step backfitting of the projection estimators of Horowitz et al. (2006) and one Newton step from the

Satisfactory estimators of the additive functions should be (i) computationally expedient; (ii) theoretically reliable; and (iii) intuitively appealing. The kernel procedures mentioned above satisfy criterion (iii) and partly (ii) but not (i), since they are computationally intensive when sample size $n$ is large, as illustrated in the Monte-Carlo results of Xue and Yang (2006b) and Wang and Yang (2007). Kim et al. (1999) introduces a computationally efficient marginal integration estimator for the component functions in additive models, which provides a reduction in computation of order $n$. Spline approaches are fast to compute, thus satisfying (i), but they do not satisfy criterion (ii) as they lack limiting distribution. By combining the best features of both kernel and spline methods, Wang and Yang (2007), Ma and Yang (2011), and Liu and Yang (2010) proposed a “spline-backfitted kernel smoothing” (SBK) method for the additive autoregressive model (3), the PLAM (4) and the ACM (5), respectively.

The SBK estimator is essentially as fast and accurate as a univariate kernel smoothing, satisfying all three criteria (i)–(iii), and is oracle efficient such that it has the same limiting distribution as the univariate function estimator by assuming that other parametric and nonparametric components are known. The SBK method is proposed for time series data, which has a geometrically $\alpha$-mixing distribution. The SBK estimation method has several advantages compared to most of the existing methods. First, as pointed out in Sperlich et al. (2002), the estimator of Linton (1997) mixed up different projections, making it uninterpretable if the real data generating process deviates from additivity, while the projections in both steps of the SBK estimator are with respect to the same measure. Second, the SBK method is computationally expedient, since the pilot spline estimator is thousands of times faster than the pilot kernel estimator in Linton (1997), as demonstrated in Table 2 of Wang and Yang (2007). Third, the SBK estimator is shown to be as efficient as the “oracle smoother” uniformly over any compact range, whereas Linton (1997) proved such “oracle efficiency” only at a single point. Moreover, the regularity conditions needed by the SBK estimation procedure are natural and appealing and close to being minimal. In contrast, higher order smoothness is needed with growing dimensionality of the regressors in Linton and Nielsen (1995). Stronger and more obscure conditions are assumed for the two-stage estimation proposed by Horowitz and Mammen (2004).

Wang and Wang (2011) applied the SBK method to survey data. As an extension, Song and Yang (2010) proposed a spline backfitted spline (SBS) approach in the framework of additive autoregressive models. The SBS achieves the oracle efficiency as the SBK method, and is more computationally efficient. Asymptotically simultaneous confidence bands can be constructed for each functional curve by the proposed SBK and SBS methods. In the following sections, we will discuss the SBK method with applications to the additive model (3), the PLAM (4) and the ACM (5), and the SBS method for the additive model (3).
2 SBK in Additive Models

In model (3), if the last $d - 1$ component functions were known by “oracle”, one could create $\{ Y_{i1}, X_{i1} \}_{i=1}^n$ with $Y_{i1} = Y_i - c - \sum_{\alpha=2}^d m_{\alpha} (X_{i\alpha}) = m_1 (X_{i1}) + \sigma \epsilon_{i1}$, from which one could compute an “oracle smoother” to estimate the only unknown function $m_1 (x_1)$, thus effectively bypassing the “curse of dimensionality”. The idea of Linton (1997) was to obtain an approximation to the unobservable variables $Y_{i1}$ by substituting $m_{\alpha} (X_{i\alpha}), i = 1, ..., n, \alpha = 2, ..., d$ with marginal integration kernel estimates and arguing that the error incurred by this “cheating” is of smaller magnitude than the rate $O (n^{-2/5})$ for estimating function $m_1 (x_1)$ from the unobservable data. Wang and Yang (2007) modify the procedure of Linton (1997) by substituting $m_{\alpha} (X_{i\alpha}), i = 1, ..., n, \alpha = 2, ..., d$ with spline estimators, specifically, Wang and Yang (2007) propose a two-stage estimation procedure: first they pre-estimate $\{ m_{\alpha} (x_{\alpha}) \}_{\alpha=2}^d$ by its pilot estimator through an under smoothed centered standard spline procedure, and next they construct the pseudo response $\hat{Y}_{i1}$ and approximate $m_1 (x_1)$ by its Nadaraya-Watson estimator.

The SBK estimator achieves its seemingly surprising success by borrowing the strengths of both spline and kernel: spline does a quick initial estimation of all additive components and removes them all except the one of interests; kernel smoothing is then applied to the cleaned univariate data to estimate with asymptotic distribution. The proposed estimators achieve uniform oracle efficiency. The two-step estimating procedure accomplishes the well-known “reducing bias by undersmoothing” in the first step using spline and “averaging out the variance” in the second step with kernel, both steps taking advantage of the joint asymptotics of kernel and spline functions.

2.1 The SBK estimator

In this section, we describe the spline-backfitted kernel estimation procedure. Let $\{ Y_i, X_i^T \}_{i=1}^n = \{ Y_{i1}, X_{i1}, ..., X_{id} \}_{i=1}^n$ be observations from a geometrically $\alpha$-mixing process following model (3). We assume that the predictor $X_{\alpha}$ is distributed on a compact interval $[a_{\alpha}, b_{\alpha}], \alpha = 1, ..., d$. Without loss of generality, we take all intervals $[a_{\alpha}, b_{\alpha}] = [0, 1], \alpha = 1, ..., d$. Denote by $\| \varphi_{\alpha} \|_2$ the theoretical $L_2$ norm of a function $\varphi_{\alpha}$ on $[0, 1], \| \varphi_{\alpha} \|_2^2 = \int_0^1 \varphi_{\alpha}^2 (x_{\alpha}) f (x_{\alpha}) \, dx_{\alpha}$. We pre-select an integer $N = N_n \sim n^{2/5} \log n$, see Assumption (A6) in Wang and Yang (2007). Next, we define for any $\alpha = 1, ..., d$, the first order B spline function (page 89, De Boor (2001)), or say the constant B spline function as the indicator function $I_J (x_{\alpha})$ of the $(N + 1)$ equally-spaced subintervals of the finite interval $[0, 1]$ with length $H = H_n = (N + 1)^{-1}$, that is

$$ I_{J, \alpha} (x_{\alpha}) = \begin{cases} 1 & JH \leq x_{\alpha} < (J + 1)H, \quad J = 0, 1, ..., N, \\ 0 & \text{otherwise}, \end{cases} $$
Define the following centered spline basis
\[ b_{J,\alpha} (x_\alpha) = I_{J+1,\alpha} (x_\alpha) - \frac{\|I_{J+1,\alpha}\|_2}{\|I_{J,\alpha}\|_2} I_{J,\alpha} (x_\alpha), \forall \alpha = 1, \ldots, d, J = 1, \ldots, N, \] (6)
with the standardized version given for any \( \alpha = 1, \ldots, d, \)
\[ B_{J,\alpha} (x_\alpha) = \frac{b_{J,\alpha} (x_\alpha)}{\|b_{J,\alpha}\|_2}, \forall J = 1, \ldots, N. \]

Define next the \((1 + dN)\)-dimensional space \( G = G[0, 1] \) of additive spline functions as the linear space spanned by \{ \( b_{J,\alpha} (x_\alpha) \), \( \alpha = 1, \ldots, d, J = 1, \ldots, N \) \}, while \( G_n \subset \mathbb{R}^n \) is spanned by \{ \( b_{J,\alpha} (X_i\alpha) \) \}_{i=1}^n, \( \alpha = 1, \ldots, d, J = 1, \ldots, N \). As \( n \to \infty \), the dimension of \( G_n \) becomes \( 1 + dN \) with probability approaching one. The spline estimator of additive function \( m (x) \) is the unique element \( \hat{m} (x) = \hat{m}_n (x) \) from the space \( G \) so that the vector \( \{ \hat{m} (X_1), \ldots, \hat{m} (X_n) \} \) best approximates the response vector \( Y \). To be precise, we define
\[ \hat{m} (x) = \hat{\lambda}'_0 + \sum_{\alpha=1}^d \sum_{J=1}^N \hat{\lambda}'_{J,\alpha} I_{J,\alpha} (x_\alpha), \] (7)
where the coefficients \( \left( \hat{\lambda}'_0, \hat{\lambda}'_{1,1}, \ldots, \hat{\lambda}'_{N,d} \right) \) are solutions of the least squares problem
\[ \left\{ \hat{\lambda}'_0, \hat{\lambda}'_{1,1}, \ldots, \hat{\lambda}'_{N,d} \right\}^T = \text{argmin} \sum_{i=1}^n \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^d \sum_{J=1}^N \lambda_{J,\alpha} I_{J,\alpha} (X_i\alpha) \right\}^2. \]

Simple linear algebra shows that
\[ \hat{m} (x) = \hat{\lambda}_0 + \sum_{\alpha=1}^d \sum_{J=1}^N \hat{\lambda}_{J,\alpha} B_{J,\alpha} (x_\alpha), \] (8)
where \( \left( \hat{\lambda}_0, \hat{\lambda}_{1,1}, \ldots, \hat{\lambda}_{N,d} \right) \) are solutions of the following least squares problem
\[ \left\{ \hat{\lambda}_0, \hat{\lambda}_{1,1}, \ldots, \hat{\lambda}_{N,d} \right\}^T = \text{argmin} \sum_{i=1}^n \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^d \sum_{J=1}^N \lambda_{J,\alpha} B_{J,\alpha} (X_i\alpha) \right\}^2, \] (9)
while (7) is used for data analytic implementation, the mathematically equivalent expression (8) is convenient for asymptotic analysis.
The pilot estimators of the component functions and the constant are

\[
\hat{m}_\alpha(x_\alpha) = \sum_{j=1}^{N} \hat{\lambda}_{j,\alpha} B_{j,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N} \hat{\lambda}_{j,\alpha} B_{j,\alpha}(X_{i\alpha}),
\]

\[
\hat{m}_c = \hat{\lambda}_0 + n^{-1} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} \sum_{j=1}^{N} \hat{\lambda}_{j,\alpha} B_{j,\alpha}(X_{i\alpha}).
\]

(10)

These pilot estimators are then used to define new pseudo-responses \(\hat{Y}_{i1}\), which are estimates of the unobservable “oracle” responses \(Y_{i1}\). Specifically,

\[
\hat{Y}_{i1} = Y_i - \hat{c} - \sum_{\alpha=2}^{d} \hat{m}_\alpha(X_{i\alpha}), \quad Y_{i1} = Y_i - c - \sum_{\alpha=2}^{d} m_\alpha(X_{i\alpha}),
\]

(11)

where \(\hat{c} = \overline{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i\), which is a \(\sqrt{n}\)-consistent estimator of \(c\) by Central Limit Theorem. Next, we define the spline-backfitted kernel (SBK) estimator of \(m_1(x_1)\) as \(\hat{m}_{\text{SBK},1}(x_1)\) based on \(\{\hat{Y}_{i1}, X_{i1}\}_{i=1}^{n}\), which attempts to mimic the would-be Nadaraya-Watson estimator \(\hat{m}_{\text{K},1}(x_1)\) of \(m_1(x_1)\) based on \(\{Y_{i1}, X_{i1}\}_{i=1}^{n}\) if the unobservable “oracle” responses \(\{Y_{i1}\}_{i=1}^{n}\) were available

\[
\hat{m}_{\text{SBK},1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1) \hat{Y}_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)}, \quad \hat{m}_{\text{K},1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1) Y_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)},
\]

(12)

where \(\hat{Y}_{i1}\) and \(Y_{i1}\) are defined in (11). Similarly, the spline-backfitted local linear (SBLL) estimator \(\hat{m}_{\text{SBLL},1}(x_1)\) based on \(\{\hat{Y}_{i1}, X_{i1}\}_{i=1}^{n}\) mimics the would-be local linear estimator \(\hat{m}_{\text{LL},1}(x_1)\) based on \(\{Y_{i1}, X_{i1}\}_{i=1}^{n}\)

\[
\{\hat{m}_{\text{SBLL},1}(x_1), \hat{m}_{\text{LL},1}(x_1)\} = (1, 0) \left(Z^T W Z\right)^{-1} Z^T W \left(\hat{Y}_{1}, Y_1\right),
\]

(13)

in which the oracle and pseudo-response vectors are \(Y_1 = (Y_{11},...,Y_{n1})^T\), \(\hat{Y}_1 = (\hat{Y}_{11},...,\hat{Y}_{n1})^T\), and the weight and design matrices are

\[
W = \text{diag} \{K_h(X_{i1} - x_1)\}_{i=1}^{n}, \quad Z^T = \begin{pmatrix} 1 & \cdots & 1 \\ X_{i1} - x_1 & \cdots & X_{n1} - x_1 \end{pmatrix}.
\]

The asymptotic properties of the smoothers \(\hat{m}_{\text{K},1}(x_1)\) and \(\hat{m}_{\text{LL},1}(x_1)\) are well-developed. Under Assumptions (A1)-(A5) in Wang and Yang (2009), according to Theorem 4.2.1 of Härdle (1990), one has for any \(x_1 \in [h, 1-h]\),

\[
\sqrt{n h} \left\{ \hat{m}_{\text{K},1}(x_1) - m_1(x_1) - b_K(x_1) h^2 \right\} \overset{D}{\to} N \left\{ 0, v^2(x_1) \right\},
\]

\[
\sqrt{n h} \left\{ \hat{m}_{\text{LL},1}(x_1) - m_1(x_1) - b_{\text{LL}}(x_1) h^2 \right\} \overset{D}{\to} N \left\{ 0, v^2(x_1) \right\},
\]
where
\[
\begin{align*}
  b_K(x_1) &= \int u^2 K(u) \, du \{m''_1(x_1) f_1(x_1) / 2 + m'_1(x_1) f'_1(x_1)\} f_1^{-1}(x_1), \\
  b_{LL}(x_1) &= \int u^2 K(u) \, du \{m''_1(x_1) / 2\} \\
  v^2(x_1) &= \int K^2(u) \, du E[\sigma^2(X_1,...,X_d) | X_1 = x_1] f_1^{-1}(x_1).
\end{align*}
\]

The equation for \(\hat{m}_{K,1}(x_1)\) requires additional Assumption (A7) in Wang and Yang (2009). The next two theorems state that the asymptotic magnitude of difference between \(\hat{m}_{SBK,1}(x_1)\) and \(\hat{m}_{K,1}(x_1)\) is of order \(o_p\left(n^{-2/5}\right)\), both pointwise and uniformly, which is dominated by the asymptotic size of \(\hat{m}_{K,1}(x_1) - m_1(x_1)\).

Hence \(\hat{m}_{SBK,1}(x_1)\) will have the same asymptotic distribution as \(\hat{m}_{K,1}(x_1)\) The same is true for \(\hat{m}_{SBLL,1}(x_1)\) and \(\hat{m}_{LL,1}(x_1)\).

**Theorem 1.** Under Assumptions (A1) to (A6) and (A2') in Wang and Yang (2009), the estimators \(\hat{m}_{SBK,1}(x_1)\) and \(\hat{m}_{SBLL,1}(x_1)\) given in (12) and (13) satisfy
\[
|\hat{m}_{SBK,1}(x_1) - \hat{m}_{K,1}(x_1)| + |\hat{m}_{SBLL,1}(x_1) - \hat{m}_{LL,1}(x_1)| = O_P\left(n^{-2/5}\right).
\]

Hence with \(b_K(x_1), b_{LL}(x_1)\) and \(v^2(x_1)\) as defined in (14), for any \(x_1 \in [h, 1-h]\)
\[
\sqrt{nh} \left\{ \hat{m}_{SBLL,1}(x_1) - m_1(x_1) - b_{LL}(x_1) h^2 \right\} \overset{D}{\rightarrow} N\left\{0, v^2(x_1)\right\},
\]
and with the additional assumption (A7) in Wang and Yang (2009), one has
\[
\sqrt{nh} \left\{ \hat{m}_{SBK,1}(x_1) - m_1(x_1) - b_K(x_1) h^2 \right\} \overset{D}{\rightarrow} N\left\{0, v^2(x_1)\right\}.
\]

**Theorem 2.** Under Assumptions (A1) to (A6) and (A2') in Wang and Yang (2009), the estimators \(\hat{m}_{SBK,1}(x_1)\) and \(\hat{m}_{SBLL,1}(x_1)\) given in (12) and (13) satisfy
\[
\sup_{x_1 \in [0,1]} \{|\hat{m}_{SBK,1}(x_1) - \hat{m}_{K,1}(x_1)| + |\hat{m}_{SBLL,1}(x_1) - \hat{m}_{LL,1}(x_1)|\} = O_P\left(n^{-2/5}\right).
\]

Hence for any \(z\),
\[
\lim_{n \to \infty} P \left\{ \log (h^{-2}) \right\}^{1/2} \left( \sup_{x_1 \in [h, 1-h]} \sqrt{nh} \frac{v(x_1)}{v(x_1)} |\hat{m}_{SBLL,1}(x_1) - m_1(x_1)| - d_n \right) < z \right\} = \exp \{-2 \exp (-z)\},
\]
in which \(d_n = \{\log (h^{-2})\}^{1/2} + \{\log (h^{-2})\}^{-1/2} \log \{c(K') (2\pi)^{-1} c^{-1}(K)\} \).

With the additional assumption (A7) in Wang and Yang (2009), it is also true that
\[
\lim_{n \to \infty} P \left\{ \log (h^{-2}) \right\}^{1/2} \left( \sup_{x_1 \in [h, 1-h]} \sqrt{nh} \frac{v(x_1)}{v(x_1)} |\hat{m}_{SBK,1}(x_1) - m_1(x_1)| - d_n \right) < z \right\} = \exp \{-2 \exp (-z)\}.
\]
For any $\alpha \in (0, 1)$, an asymptotic $100 (1 - \alpha)\%$ confidence band for $m_1 (x_1)$ over interval $[h, 1 - h]$ is

$$\hat{m}_{SBLL1} (x_1) \pm v (x_1) (nh)^{-1/2} \left[ d_n - \log^{-1/2} (2h) \log \left\{ \frac{\log (1 - \alpha)}{2} \right\} \right].$$

Remark 1. Similar estimators $\hat{m}_{SBK, \alpha} (x_\alpha)$ and $\hat{m}_{SBLL, \alpha} (x_\alpha)$ can be constructed for $m_\alpha (x_\alpha), 2 \leq \alpha \leq d$ with same oracle properties.

### 2.2 Application to Boston housing data

Wang and Yang (2009) applied the proposed method to the well-known Boston housing data, which contains 506 different houses from a variety of locations in Boston Standard Metropolitan Statistical Area in 1970. The median value and 13 sociodemographic statistics values of the Boston houses were first studied by Harrison and Rubinfeld (1978) to estimate the housing price index model. Breiman and Friedman (1985) did further analysis to deal with the multi-collinearity for overfitting by using a stepwise method. The response and explanatory variables of interest are:

- **MEDV**: Median value of owner-occupied homes in $\$1000$’s
- **RM**: average number of rooms per dwelling
- **TAX**: full-value property-tax rate per $\$10,000$
- **PTRATIO**: pupil-teacher ratio by town school district
- **LSTAT**: proportion of population that is of "lower status" in %.

In order to ease off the trouble caused by big gaps in the domain of variables TAX and LSTAT, logarithmic transformation is done for both variables before fitting the model. Wang and Yang (2009) fitted an additive model as follows:

$$MEDV = \mu + m_1 (RM) + m_2 (\log(TAX)) + m_3 (PTRATIO) + m_4 (\log(LSTAT)) + \varepsilon.$$

In Figure 3 (see Wang and Yang 2009), the univariate function estimates and corresponding confidence bands are displayed together with the “pseudo data points” with pseudo response as the backfitted response after subtracting the sum function of the remaining three covariates. All the function estimates are represented by the dotted lines, “data points” by circles, and confidence bands by upper and lower thin lines. The kernel used in SBLL estimator is quartic kernel, $K(u) = \frac{15}{16} (1 - u^2)^2$ for $-1 < u < 1$.

(Insert Figure 3 about here)

The proposed confidence bands are used to test the linearity of the components. In Figure 3 the straight solid lines are the least squares regression lines. The first figure shows that the null hypothesis $H_0 : m_1 (RM) = a_1 + b_1 RM$, will be rejected since the confidence bands with 0.99 confidence couldn’t totally cover the straight regression line, i.e., the p-value is less than 0.01. Similarly the linearity of the component functions for log (TAX) and log (LSTAT)
are not accepted at the significance level 0.01. While the least square straight line of variable PTRATIO in the upper right figure totally falls between the upper and lower 95% confidence bands, thus the linearity null hypothesis $\mathcal{H}_0: m_3(PTRATIO) = a_3 + b_3PTRATIO$ is accepted at the significance level 0.05.

3 SBK in Partially Linear Additive Models (PLAM)

Wang and Yang (2009) fitted an additive model using RM, log (TAX), PTRATIO and log(LSTAT) as predictors to test the linearity of the components and found that only PTRATIO is accepted at the significance level 0.05 for the linearity hypothesis test. Based on the conclusion drawn from Wang and Yang (2009), a PLAM can be fitted as

$$\text{MEDV} = c_{00} + c_{01} \times \text{PTRATIO} + m_1(\text{RM}) + m_2(\log(\text{TAX})) + m_3(\log(\text{LSTAT})) + \varepsilon. \quad (15)$$

PLAMs contain both linear and nonlinear additive components, and they are more flexible compared to linear models and more efficient compared to general nonparametric regression models. A general form of PLAMs is given in (4). In the PLAM (4), if the regression coefficients $\{c_{0l}\}^d_0$ and the component functions $\{m_\beta(x_\beta)\}^{d_2}_{\beta=1, \beta \neq \alpha}$ were known by “oracle”, one could create $\{Y_{i\alpha}, X_{i\alpha}\}^{n}_{i=1}$ with $Y_{i\alpha} = Y_i - c_{00} - \sum^{d_1}_{l=1} c_{0l}^T I_l - \sum^{d_2}_{\beta=1, \beta \neq \alpha} m_\beta(X_{i\beta}) = m_\alpha(X_{i\alpha}) + \sigma(X_{i\alpha} T_{i\alpha}) \varepsilon_i$, from which one could compute an “oracle smoother” to estimate the only unknown function $m_\alpha(x_\alpha)$, bypassing the “curse of dimensionality”. A major theoretical innovation is to resolve the dependence between $T$ and $X$, making use of Assumption (A5) in Ma and Yang (2011), which is not needed in Wang and Yang (2007). Another significant innovation is the $\sqrt{n}$-consistency and asymptotic distribution of estimators for parameters $\{c_{0l}\}^{d_1}_{l=0}$, which is trivial for the additive model of Wang and Yang (2007).

We denote by $I_r$, the $r \times r$ identity matrix, $0_{r \times s}$ the zero matrix of dimension $r \times s$, and $\text{diag}(a, b)$ the $2 \times 2$ diagonal matrix with diagonal entries $a, b$. Let $\{Y_i, X^T_i, T^T_i\}^n_{i=1}$ be a sequence of strictly stationary observations from a geometrically $\alpha$-mixing process following model (4), where $Y_i$ and $(X_i, T_i) = \{(X_{i1}, \ldots, X_{id_2})^T, (T_{i1}, \ldots, T_{id_1})^T\}$ are the $i$-th response and predictor vector. Define next the space $G$ of partially linear additive spline functions as the linear space spanned by $\{1, t_l, b_{l, \alpha}(x_\alpha) : 1 \leq l \leq d_1, 1 \leq \alpha \leq d_2, 1 \leq J \leq N + 1\}$. Let $\{1, \{t_l, b_{l, \alpha}(X_{i\alpha})\}^{n}_{i=1} : 1 \leq l \leq d_1, 1 \leq \alpha \leq d_2, 1 \leq J \leq N + 1\}$ span the space $G_n \subset R^n$, where $b_{l, \alpha}$ is defined in (6). As $n \to \infty$, with probability approaching 1, the dimension of $G_n$ becomes $\{1 + d_1 + d_2(N + 1)\}$. The spline estimator of $m(x, t)$ is the unique element $\hat{m}(x, t) = \hat{m}_n(x, t)$ from $G$ so that $\{\hat{m}(X_i, T_i)\}^{n}_{i=1}$ best approximates the response vector $Y$. To be
would-be Nadaraya-Watson estimator for if the unobservable responses Based on of theoretically centered pseudo responses where the coefficients minimize

\[
\sum_{i=1}^{n} \left( Y_i - c_0 - \sum_{l=1}^{d_1} c_l T_{il} - \sum_{\alpha=1}^{d_2} \sum_{J=1}^{N+1} c_{J,\alpha} b_{J,\alpha} (x_{i\alpha}) \right)^2.
\]

Pilot estimators of \( e^T = \{ c_{0l} \}_{l=0}^{d_1} \) and \( m_{\alpha} (x_{i\alpha}) \) are \( \hat{e}^T = \{ \hat{c}_{0l} \}_{l=0}^{d_1} \) and \( \hat{m}_\alpha (x_{i\alpha}) = \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} b_{J,\alpha} (x_{i\alpha}) - n^{-1} \sum_{i=1}^{n} \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} b_{J,\alpha} (X_{i\alpha}) \), which are used to define pseudo responses \( \hat{Y}_{i\alpha} \), estimates of the unobservable “oracle” responses \( Y_{i\alpha} \):

\[
\begin{align*}
\hat{Y}_{i\alpha} &= Y_i - c_0 - \sum_{l=1}^{d_1} \hat{c}_l T_{il} - \sum_{\beta=1, \beta \neq \alpha} m_\beta (X_{i\beta}), \\
Y_{i\alpha} &= Y_i - c_0 - \sum_{l=1}^{d_1} c_l T_{il} - \sum_{\beta=1, \beta \neq \alpha} m_\beta (X_{i\beta}).
\end{align*}
\]

(16)

Based on \( \{ \hat{Y}_{i\alpha}, X_{i\alpha} \}_{i=1}^{n} \), the SBK estimator \( \hat{m}_{\mathrm{SBK},\alpha} (x_{i\alpha}) \) of \( m_{\alpha} (x_{i\alpha}) \) mimics the would-be Nadaraya-Watson estimator \( \hat{m}_{K,\alpha} (x_{i\alpha}) \) of \( m_{\alpha} (x_{i\alpha}) \) based on \( \{ Y_{i\alpha}, X_{i\alpha} \}_{i=1}^{n} \), if the unobservable responses \( \{ Y_{i\alpha}, X_{i\alpha} \}_{i=1}^{n} \) were available

\[
\begin{align*}
\hat{m}_{\mathrm{SBK},\alpha} (x_{i\alpha}) &= \left\{ n^{-1} \sum_{i=1}^{n} K_h (X_{i\alpha} - x_{i\alpha}) \hat{Y}_{i\alpha} \right\} / \hat{f}_\alpha (x_{i\alpha}), \\
\hat{m}_{K,\alpha} (x_{i\alpha}) &= \left\{ n^{-1} \sum_{i=1}^{n} K_h (X_{i\alpha} - x_{i\alpha}) Y_{i\alpha} \right\} / \hat{f}_\alpha (x_{i\alpha}).
\end{align*}
\]

(17)

with \( \hat{Y}_{i\alpha}, Y_{i\alpha} \) in (16), \( \hat{f}_\alpha (x_{i\alpha}) = n^{-1} \sum_{i=1}^{n} K_h (X_{i\alpha} - x_{i\alpha}) \) an estimator of \( f_\alpha (x_{i\alpha}) \).

Define the Hilbert space

\[
\mathcal{H} = \left\{ p (x) = \sum_{\alpha=1}^{d_2} p_\alpha (x_{i\alpha}), E p_\alpha (X_{i\alpha}) = 0, E^2 p_\alpha (X_{i\alpha}) < \infty \right\}
\]

of theoretically centered \( L_2 \) additive functions on \([0, 1]^{d_2}\), while denote by \( \mathcal{H}_n \) its subspace spanned by \( \{ B_{J,\alpha} (x_{i\alpha}) \}, 1 \leq \alpha \leq d_2, 1 \leq J \leq N + 1 \}. \ Denote

\[
\begin{align*}
\text{Proj}_{\mathcal{H}} T_l &= \left. p_l (X) = \arg\min_{p \in \mathcal{H}} E \{ T_l - p (X) \}^2 \right| T_l = T_l - \text{Proj}_{\mathcal{H}} T_l, \\
\text{Proj}_{\mathcal{H}_n} T_l &= \left. \arg\min_{p \in \mathcal{H}_n} E \{ T_l - p (X) \}^2 \right| T_l = T_l - \text{Proj}_{\mathcal{H}_n} T_l,
\end{align*}
\]

for \( 1 \leq l \leq d_1 \), where \( \text{Proj}_{\mathcal{H}} T_l \) and \( \text{Proj}_{\mathcal{H}_n} T_l \) are orthogonal projections of \( T_l \) unto subspaces \( \mathcal{H} \) and \( \mathcal{H}_n \) respectively. Denote next in vector form

\[
\begin{align*}
\tilde{T}_n &= \left\{ \tilde{T}_{l,n} \right\}_{1 \leq l \leq d_1}, \tilde{T} = \left\{ \tilde{T}_l \right\}_{1 \leq l \leq d_1}.
\end{align*}
\]

(18)
Without loss of generality, let $\alpha = 1$. Under Assumptions (A1)-(A5) and (A7) in Ma and Yang (2011), it is straightforward to verify (as in Bosq (1998)) that as $n \to \infty$,

$$\sup_{x_1 \in [h, 1-h]} |\hat{m}_{K,1} (x_1) - m_1 (x_1)| = o_p \left( n^{-2/5} \log n \right),$$

$$\sqrt{n} \{ \hat{m}_{K,1} (x_1) - m_1 (x_1) - b_1 (x_1) h^2 \} \overset{D}{\to} N \{ 0, \nu_1^2 (x_1) \},$$

where, $b_1 (x_1) = \int u^2 K (u) \, du \{ m_1'' (x_1) f_1 (x_1) / 2 + m_1' (x_1) f_1' (x_1) \} f_1^{-1} (x_1), \tag{19}$

$$\nu_1^2 (x_1) = \int K^2 (u) \, du \, E \{ \sigma^2 (X, T) \mid X_1 = x_1 \} f_1^{-1} (x_1).$$

It is shown in Li (2000) and Schimek (2000) that the spline estimator $\hat{m}_1 (x_1)$ in the first step uniformly converges to $m_1 (x_1)$ with certain convergence rate, but lacks asymptotic distribution. Theorem 3 below states that the difference between $\hat{m}_{SBK,1} (x_1)$ and $\hat{m}_{K,1} (x_1)$ is $o_p \left( n^{-2/5} \right)$ uniformly, dominated by the asymptotic uniform size of $\hat{m}_{K,1} (x_1) - m_1 (x_1)$. So $\hat{m}_{SBK,1} (x_1)$ has identical asymptotic distribution as $\hat{m}_{K,1} (x_1)$.

**Theorem 3.** Under Assumptions (A1)-(A7) in Ma and Yang (2011), as $n \to \infty$, the SBK estimator $\hat{m}_{SBK,1} (x_1)$ given in (17) satisfies

$$\sup_{x_1 \in [0, 1]} |\hat{m}_{SBK,1} (x_1) - \hat{m}_{K,1} (x_1)| = o_p \left( n^{-2/5} \right).$$

Hence with $b_1 (x_1)$ and $\nu_1^2 (x_1)$ as defined in (19), for any $x_1 \in [h, 1-h]$, $\sqrt{n} \{ \hat{m}_{SBK,1} (x_1) - m_1 (x_1) - b_1 (x_1) h^2 \} \overset{D}{\to} N \{ 0, \nu_1^2 (x_1) \}.$

Instead of Nadaraya-Watson estimator, one can use local polynomial estimator, see Fan and Gijbels (1996). Under Assumptions (A1)-(A7), for any $\alpha \in (0, 1)$, an asymptotic 100 $(1 - \alpha)$ % confidence intervals for $m_1 (x_1)$ is

$$\hat{m}_{SBK,1} (x_1) - \hat{b}_1 (x_1) h^2 \pm z_{\alpha/2} \hat{v}_1 (x_1) (nh)^{-1/2}$$

where $\hat{b}_1 (x_1)$ and $\hat{v}_1^2 (x_1)$ are estimators of $b_1 (x_1)$ and $\nu_1^2 (x_1)$ respectively.

The following corollary provides the asymptotic distribution of $\hat{m}_{SBK} (x)$.

**Corollary 1.** Under Assumptions (A1)-(A7) in Ma and Yang (2011) and $m_\alpha \in C^{(2)} [0, 1], 2 \leq \alpha \leq d_2$. Let $\hat{m}_{SBK} (x) = \sum_{a=1}^{d_2} \hat{m}_{SBK,\alpha} (x_\alpha), b (x) = \sum_{a=1}^{d_2} b_\alpha (x_\alpha), \nu_\alpha (x) = \sum_{a=1}^{d_2} \nu_{\alpha}^2 (x_\alpha),$ for any $x \in [0, 1]^{d_2}$, with SBK estimators $\hat{m}_{SBK,\alpha} (x_\alpha),$ $1 \leq \alpha \leq d_2$, defined in (17), and $b_\alpha (x_\alpha), \nu_{\alpha}^2 (x_\alpha)$ similarly defined as in (19), as $n \to \infty$,

$$\sqrt{n} \{ \hat{m}_{SBK} (x) - \sum_{\alpha=1}^{d_2} m_\alpha (x_\alpha) - b (x) h^2 \} \overset{D}{\to} N \{ 0, \nu_\alpha (x) \}.$$
Next theorem describes the asymptotic behavior of estimator $\hat{c}$ for $c$.

**Theorem 4.** Under Assumptions (A1)-(A6) in Ma and Yang (2011), as $n \to \infty$, $\|\hat{c} - c\| = O_p\left(n^{-1/2}\right)$. With the additional Assumption A8 in Ma and Yang (2011),

$$\sqrt{n} (\hat{c} - c) \rightarrow_d N \left( 0, \sigma_0^2 \left\{ \begin{array}{c} 1 \\ 0_{d_1} \\ 0_{d_2} \\ \Sigma^{-1} \end{array} \right\} \right),$$

for $\Sigma = \text{cov} \left( \tilde{T} \right)$ with random vector $\tilde{T}$ defined in (18).

### 3.1 Application to Boston housing data

The Boston housing data were studied by Ma and Yang (2011) by fitting model (15). Figure 4 (see Ma and Yang (2011)) shows the univariate nonlinear function estimates (dashed lines) and corresponding simultaneous confidence bands (thin lines) together with the "pseudo data points" (dots) with pseudo response as the backfitted response after subtracting the sum function of the remaining covariates. The confidence bands are used to test the linearity of the nonparametric components. In Figure 4 the straight solid lines are the least squares regression lines through the pseudo data points. The first figure confidence band with 0.999999 confidence level does not totally cover the straight regression line, i.e., the $p$-value is less than 0.000001. Similarly the linearity of the component functions for $\log(\text{TAX})$ and $\log(\text{LSTAT})$ are rejected at the significance levels 0.017 and 0.007, respectively. The estimators $\hat{c}_{00}$ and $\hat{c}_{01}$ of $c_{00}$ and $c_{01}$ are 33.393 and $-0.58845$ and both are significant with $p$-values close to 0. The correlation between the estimated and observed values of MEDV is 0.89944, much higher than 0.80112 obtained by Wang and Yang (2009). This improvement is due to fitting the variable PTRATIO directly as linear with the higher accuracy of parametric model instead of treating it unnecessarily as a nonparametric variable. In other words, the PLAM fits the housing data much better than the additive model of Wang and Yang (2009).

### 4 SBK in Additive Coefficient Models (ACM)

To estimate the additive function components in the model (5), we introduce the similar idea as in the previous two sections for additive models (3) and PLAMs (4). If all the nonparametric functions of the last $d_2 - 1$ variables, $\{m_{\alpha l}(x_\alpha)\}_{l=1, \alpha=2}^{d_1, d_2}$ and all the constants $\{m_{0l}\}_{l=1}^{d_1}$ were known by “oracle”, one could define a new variable $Y_{,1} = \sum_{l=1}^{d_1} m_{1l}(X_1) T_l + \sigma(X, T) \varepsilon = Y - \sum_{l=1}^{d_1} \left\{ m_{0l} + \sum_{\alpha=2}^{d_2} m_{\alpha l}(X_\alpha) \right\} T_l$ and estimate all functions $\{m_{1l}(x_1)\}_{l=1}^{d_1}$ by linear regression of $Y_{,1}$ on $T_1, \ldots, T_{d_1}$ with kernel weights computed from variable.
$X_1$. In stead of using the Nadaraya-Watson estimating method in the second step, Liu and Yang (2010) proposed to pre-estimate the functions $\{m_{\alpha l}(x_{i\alpha})\}_{l=1,\alpha=2}^{d_1,d_2}$ and constants $\{m_{\alpha l}\}_{l=1}^{d_1}$ by linear spline and then use these estimates as substitutes to obtain an approximation $\hat{Y}_{1}$ to the variable $Y_{1}$, and construct “oracle” estimators based on $\hat{Y}_{1}$.

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

$$\mathcal{H}_\alpha^0 = \{g : E\{g(X_{\alpha})\} = 0, E\{g^2(X_{\alpha})\} < +\infty\}, 1 \leq \alpha \leq d_2.$$  

Next define the model space $\mathcal{M}$, a collection of functions on $\chi \times R^{d_1}$ as

$$\mathcal{M} = \left\{ g(\mathbf{x},t) = \sum_{l=1}^{d_1} g_{l}(\mathbf{x}) t_l; \quad g_l(\mathbf{x}) = g_{0l} + \sum_{\alpha=1}^{d_2} g_{\alpha l}(x_{i\alpha}); g_{\alpha l} \in \mathcal{H}_\alpha^0 \right\},$$

in which $\{g_{0l}\}_{l=1}^{d_1}$ are finite constants. The constraints that $E\{g_{\alpha l}(X_{i\alpha})\} = 0$, $1 \leq \alpha \leq d_2$ ensure unique additive representation of $m_l$ as expressed in (5) but are not necessary for the definition of space $\mathcal{M}$.

For any vector $\mathbf{x} = (x_1, x_2, \cdots, x_{d_2})$, denote the deleted vector as $\mathbf{x}_{-1} = (x_2, \cdots, x_{d_2})$ and for the random vector $\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots, X_{id_2})$ the deleted vector $\mathbf{X}_{i,-1} = (X_{i2}, \cdots, X_{id_2})$, $1 \leq i \leq n$. For any $1 \leq l \leq d_1$, write $m_{-1,l}(\mathbf{x}_{-1}) = m_{0l} + \sum_{\alpha=2}^{d_2} m_{\alpha l}(x_{i\alpha})$. Denote the vector of pseudo-responses $\mathbf{Y}_1 = (Y_{1,1}, \cdots, Y_{n,1})^T$ in which

$$Y_{i,1} = Y_{i} - \sum_{l=1}^{d_1} \{m_{0l} + m_{-1,l}(\mathbf{X}_{i,-1})\} T_{il} = \sum_{l=1}^{d_1} m_{1l}(X_{i1}) T_{il} + \sigma(\mathbf{X}_i, \mathbf{T}_i) \varepsilon_i.$$

These would be the “responses” have the unknown functions $\{m_{-1,l}(\mathbf{x}_{-1})\}_{l=1}^{d_1}$ been given. In that case, one could “estimate” all the coefficient functions in $x_1$, the vector function $m_{-1, \cdot}(x_1) = \{m_{11}(x_1), \cdots, m_{1d_1}(x_1)\}^T$ by solving a kernel weighted least squares problem

$$\tilde{m}_{K,1, \cdot}(x_1) = \{\tilde{m}_{K,11}(x_1), \cdots, \tilde{m}_{K,1d_1}(x_1)\}^T = \arg\min_{\lambda=(\lambda_1)_{1 \leq \lambda \leq d_1}} L(\lambda, m_{-1, \cdot}, x_1)$$

in which

$$L(\lambda, m_{-1, \cdot}, x_1) = \sum_{i=1}^{n} \left( Y_{i,1} - \sum_{l=1}^{d_1} \lambda_l T_{il} \right)^2 K_h(X_{i1} - x_1).$$
Alternatively, one could rewrite the above kernel oracle smoother in matrix form

\[ \tilde{m}_{K,1.}(x_1) = \left( \frac{C_K^T W_1 C_K}{n} \right)^{-1} \frac{1}{n} C_K^T W_1 Y_1 = \left( \frac{1}{n} C_K^T W_1 C_K \right)^{-1} \frac{1}{n} C_K^T W_1 Y_1 \]

(20)

in which

\[ T_i = (T_{i1}, \ldots, T_{id_1})^T, C_K = \{ T_1, \ldots, T_n \}^T, \]

\[ W_1 = \text{diag} \{ K_h (X_{11} - x_1), \ldots, K_h (X_{n1} - x_1) \}. \]

Likewise, one can define the local linear oracle smoother of \( m_{1.}(x_1) \) as

\[ \tilde{m}_{LL,1.}(x_1) = (I_{d_1 \times d_1}, 0_{d_1 \times d_1}) \left( \frac{1}{n} C_{LL,1}^T W_1 C_{LL,1} \right)^{-1} \frac{1}{n} C_{LL,1}^T W_1 Y_1, \]

(21)

in which

\[ C_{LL,1} = \left\{ T_1 (X_{11} - x_1), \ldots, T_n (X_{n1} - x_1) \right\}^T. \]

Denote \( \mu_2 (K) = \int u^2 K(u) \, du, \| K \|^2_2 = \int K(u)^2 \, du, Q_1 (x_1) = \{ q_1 (x_1) \}_{l,l'=1}^{d_1} = E(\mathbf{T}\mathbf{T}^T | X_1 = x_1), \) and define the following bias and variance coefficients

\[ b_{LL,l,l',1}(x_1) = \frac{1}{2} \mu_2 (K) m_{11}''(x_1) f_1(x_1) q_{ll',1}(x_1), \]

\[ b_{K,l,l',1}(x_1) = \frac{1}{2} \mu_2 (K) \left[ 2m_{11}''(x_1) \frac{\partial}{\partial x_1} \{ f_1(x_1) q_{ll',1}(x_1) \} + m_{11}''(x_1) f_1(x_1) q_{ll',1}(x_1) \right], \]

\[ \Sigma_1 (x_1) = \| K \|^2_2 f_1(x_1) E \{ \mathbf{T}\mathbf{T}^T \sigma^2 (\mathbf{X}, \mathbf{T}) | X_1 = x_1 \}, \]

\[ \{ v_{l,l',1}(x_1) \}_{l,l'=1}^{d_1} = Q_1 (x_1)^{-1} \Sigma_1 (x_1) Q_1 (x_1)^{-1}. \]

(22)

**Theorem 5.** Under Assumptions (A1) to (A5) and (A7) in Liu and Yang (2010), for any \( x_1 \in [h, 1 - h] \), as \( n \rightarrow \infty \), the oracle local linear smoother \( \tilde{m}_{LL,1.}(x_1) \) given in (21) satisfies

\[ \sqrt{nh} \left[ \tilde{m}_{LL,1.}(x_1) - m_{1.}(x_1) - \left\{ \sum_{l=1}^{d_1} b_{LL,l,l',1}(x_1) \right\}_{l,l'=1}^{d_1} h^2 \right] \]

\[ \rightarrow N \left( 0, \{ v_{l,l',1}(x_1) \}_{l,l'=1}^{d_1} \right). \]
With Assumption (A6) in addition, the oracle kernel smoother \( \tilde{m}_{K,1.}(x_1) \) in (20) satisfies
\[
\sqrt{nh} \left( \tilde{m}_{K,1.}(x_1) - m_{1.}(x_1) - \left\{ \sum_{t=1}^{d_1} b_{K,l',1}(x_1) \right\}_{l'=1}^{d_1} h^2 \right) 
\rightarrow N \left( 0, \{v_{l,l',1}(x_1)\}_{l,l'=1}^{d_1} \right).
\]

**Theorem 6.** Under Assumptions (A1) to (A5) and (A7) in Liu and Yang (2010), as \( n \to \infty \), the oracle local linear smoother \( \tilde{m}_{LL,1.}(x_1) \) given in (21) satisfies
\[
\sup_{x_1 \in [h,1-h]} \left| \tilde{m}_{LL,1.}(x_1) - m_{1.}(x_1) \right| = O_p \left( \log n / \sqrt{nh} \right).
\]

With Assumption (A6) in addition, the oracle kernel smoother \( \tilde{m}_{K,1.}(x_1) \) in (20) satisfies
\[
\sup_{x_1 \in [h,1-h]} \left| \tilde{m}_{K,1.}(x_1) - m_{1.}(x_1) \right| = O_p \left( \log n / \sqrt{nh} \right).
\]

**Remark 1.** The above theorems hold for \( \tilde{m}_{LL,\alpha.}(x_\alpha) \) and \( \tilde{m}_{K,\alpha.}(x_\alpha) \) similarly constructed as \( \tilde{m}_{LL,1.}(x_1) \) and \( \tilde{m}_{K,1.}(x_1) \), for any \( \alpha = 2, \ldots, d_2 \).

The same oracle idea applies to the constants as well. Define the would-be “estimators” of constants \( (m_0l)_{1 \leq l \leq d_1} \) as the following least squares solution
\[
\tilde{m}_0 = (\tilde{m}_0l)_{1 \leq l \leq d_1} = \arg \min \sum_{i=1}^{n} \left\{ Y_{ic} - \sum_{l=1}^{d_1} m_0l T_{il} \right\}^2,
\]
in which the oracle responses are
\[
Y_{ic} = Y_i - \sum_{l=1}^{d_1} \sum_{\alpha=1}^{d_2} m_{\alpha l} (X_{i\alpha}) T_{il} = \sum_{l=1}^{d_1} m_0l T_{il} + \sigma (X_i, T_i) \varepsilon_i. \tag{23}
\]

The following result provides optimal convergence rate of \( \tilde{m}_0 \) to \( m_0 \), which are needed for removing the effects of \( m_0 \) for estimating the functions \( \{m_{1l}(x_1)\}_{l=1}^{d_1} \).

**Proposition 1.** Under Assumptions (A1)-(A5) and (A8) in Liu and Yang (2010), as \( n \to \infty \), \( \sup_{1 \leq l \leq d_1} |\tilde{m}_{0l} - m_{0l}| = O_p \left( n^{-1/2} \right) \).

Although the oracle smoothers \( \tilde{m}_{LL,\alpha.}(x_\alpha) \), \( \tilde{m}_{K,\alpha.}(x_\alpha) \) possess the theoretical properties in Theorems 5 and 6, they are not useful statistics as they are
computed based on the knowledge of unavailable functions \( \{ m_{o\alpha}(x_\alpha) \}_{l=1,\alpha=2}^{d_1,d_2} \) and constants \( \{ m_{0\alpha} \}_{l=1}^{d_1} \). They do, however, motivate the spline-backfitted estimators that we introduce next.

In the following, we describe how the unknown functions \( \{ m_{o\alpha}(x_\alpha) \}_{l=1,\alpha=2}^{d_1,d_2} \) and constants \( \{ m_{0\alpha} \}_{l=1}^{d_1} \) can be pre-estimated by linear spline and how the estimates are used to construct the “oracle estimators”. Define the space of \( \alpha \)-empirically centered linear spline functions on \([0, 1]\) as

\[
G_{n,\alpha}^0 = \left\{ g_\alpha : g_\alpha(x_\alpha) \equiv \sum_{j=0}^{N+1} \lambda_j b_j(x_\alpha), E_n \{ g_\alpha(X_\alpha) \} = 0 \right\}, \quad 1 \leq \alpha \leq d_2,
\]

and the space of additive spline coefficient functions on \( \chi \times R^{d_1} \) as

\[
G_n^0 = \left\{ g(x, t) = \sum_{l=1}^{d_1} g_l(x) t_i; \quad g_l(x) = g_{0l} + \sum_{\alpha=1}^{d_2} g_{\alpha l}(x_\alpha); \quad g_{0l} \in R, g_{\alpha l} \in G_{n,\alpha}^0 \right\},
\]

which is equipped with the empirical inner product \( \langle \cdot, \cdot \rangle_{2,n} \).

The multivariate function \( m(x, t) \) is estimated by an additive spline coefficient function

\[
\hat{m}(x, t) = \sum_{l=1}^{d_1} \hat{m}_l(x) t_i = \arg\min_{g \in G_n^0} \sum_{i=1}^{n} \{ Y_i - g(X_i, T_i) \}^2. \tag{24}
\]

Since \( \hat{m}(x, t) \in G_n^0 \), one can write \( \hat{m}_l(x) = \hat{m}_{0l} + \sum_{\alpha=1}^{d_2} \hat{m}_{\alpha l}(x_\alpha) \); for \( \hat{m}_{0l} \in R \) and \( \hat{m}_{\alpha l}(x_\alpha) \in G_{n,\alpha}^0 \). Simple algebra shows that the following oracle estimators of the constants \( m_{0\alpha} \) are exactly equal to \( \hat{m}_{0l} \), in which the oracle pseudo-responses \( \hat{Y}_{ic} = Y_i - \sum_{l=1}^{d_1} \sum_{\alpha=1}^{d_2} \hat{m}_{\alpha l}(X_{i\alpha}) T_{il} \) which mimick the oracle responses \( Y_{ic} \) in (23)

\[
\hat{m}_0 = (\hat{m}_{0\alpha})_{1 \leq i \leq d_1}^T = \arg\min_{(\lambda_{0\alpha}, \ldots, \lambda_{0d_1})} \sum_{i=1}^{n} \left\{ \hat{Y}_{ic} - \sum_{l=1}^{d_1} \lambda_{0l} T_{il} \right\}^2. \tag{25}
\]

**Proposition 2.** Under Assumptions (A1) to (A5) and (A8) in Liu and Yang (2010), as \( n \to \infty, \sup_{1 \leq i \leq d_1} |\hat{m}_{0l} - m_{0l}| = O_p(n^{-1/2}), \) hence \( \sup_{1 \leq i \leq d_1} |\hat{m}_{0l} - m_{0l}| = O_p(n^{-1/2}) \) following Proposition 1.

Define next the oracle pseudo-responses

\[
\hat{Y}_{i1} = Y_i - \sum_{l=1}^{d_1} \left( \hat{m}_{0l} + \sum_{\alpha=2}^{d_2} \hat{m}_{\alpha l}(X_{i\alpha}) \right) T_{il}
\]
and \( \hat{Y}_1 = \left( \hat{Y}_{11}, \ldots, \hat{Y}_{n1} \right)^T \), with \( \hat{m}_{0l} \) and \( \hat{m}_{nl} \) defined in (25) and (24) respectively. The spline-backfitted kernel (SBK) and spline-backfitted local linear (SBLL) estimators are

\[
\hat{m}_{\text{SBK},l} \left( x_1 \right) = \left( C_K^T W_1 C_K \right)^{-1} C_K^T W_1 \hat{Y}_1 = \left( \frac{1}{n} C_K^T W_1 C_K \right)^{-1} \frac{1}{n} C_K^T W_1 \hat{Y}_1 = \left( \frac{1}{n} C_K^T W_1 C_K \right)^{-1} \frac{1}{n} C_K^T W_1 \hat{Y}_1,
\]

(26)

\[
\hat{m}_{\text{SBLL},l} \left( x_1 \right) = \left( I_{d_l \times d_l}, 0_{d_l \times d_l} \right) \left( \frac{1}{n} C_{\text{LL},l}^T W_1 C_{\text{LL},l} \right)^{-1} \frac{1}{n} C_{\text{LL},l}^T W_1 \hat{Y}_1.
\]

(27)

The following theorem states that the asymptotic uniform magnitude of difference between \( \hat{m}_{\text{SBK},l} \left( x_1 \right) \) and \( \hat{m}_{\text{LL},l} \left( x_1 \right) \) is of order \( o_p \left( n^{-2/5} \right) \), which is dominated by the asymptotic size of \( \hat{m}_{\text{K},l} \left( x_1 \right) - m_{1,l} \left( x_1 \right) \). As a result, \( \hat{m}_{\text{SBK},l} \left( x_1 \right) \) will have the same asymptotic distribution as \( \hat{m}_{\text{K},l} \left( x_1 \right) \). The same is true for \( \hat{m}_{\text{SBLL},l} \left( x_1 \right) \) and \( \hat{m}_{\text{LL},l} \left( x_1 \right) \).

**Theorem 7.** Under Assumptions (A1) to (A5), (A7) and (A8) in Liu and Yang (2010), as \( n \to \infty \), the SBK estimator \( \hat{m}_{\text{SBK},l} \left( x_1 \right) \) in (26) and the SBLL estimator \( \hat{m}_{\text{SBLL},l} \left( x_1 \right) \) in (27) satisfy

\[
\sup_{x_1 \in [0,1]} \{ |\hat{m}_{\text{SBK},l} \left( x_1 \right) - \hat{m}_{\text{K},l} \left( x_1 \right) | + |\hat{m}_{\text{SBLL},l} \left( x_1 \right) - \hat{m}_{\text{LL},l} \left( x_1 \right) | \} = o_p \left( n^{-2/5} \right).
\]

The following corollary provides the asymptotic distributions of \( \hat{m}_{\text{SBLL},l} \left( x_1 \right) \) and \( \hat{m}_{\text{K},l} \left( x_1 \right) \). The proof of this corollary is straightforward from Theorems 5 and 7.

**Corollary 2.** Under Assumptions (A1) to (A5), (A7) and (A8) in Liu and Yang (2010), for any \( x_1 \in [h, 1-h] \), as \( n \to \infty \), the SBLL estimator \( \hat{m}_{\text{SBLL},l} \left( x_1 \right) \) in (27) satisfies

\[
\sqrt{nh} \left[ \hat{m}_{\text{SBLL},l} \left( x_1 \right) - m_{1,l} \left( x_1 \right) - \left\{ \sum_{l=1}^{d_l} b_{\text{LL},l,l',l} \left( x_1 \right) \right\}_{l',l=1}^{d_l} \right] \to N \left( 0, \left\{ v_{l,l',l} \left( x_1 \right) \right\}_{l,l'=1}^{d_l} \right)
\]

and with the additional Assumption (A6), the SBK estimator \( \hat{m}_{\text{SBK},l} \left( x_1 \right) \) in (26) satisfies

\[
\sqrt{nh} \left[ \hat{m}_{\text{K},l} \left( x_1 \right) - m_{1,l} \left( x_1 \right) - \left\{ \sum_{l=1}^{d_l} b_{\text{K},l,l',l} \left( x_1 \right) \right\}_{l',l=1}^{d_l} \right] \to N \left( 0, \left\{ v_{l,l',l} \left( x_1 \right) \right\}_{l,l'=1}^{d_l} \right)
\]

where \( b_{\text{LL},l,l',l} \left( x_1 \right) \), \( b_{\text{K},l,l',l} \left( x_1 \right) \) and \( v_{l,l',l} \left( x_1 \right) \) are defined as (22).
Remark 2. For any $\alpha = 2, \ldots, d$, the above theorem and corollary hold for $\hat{m}_{SBK, \alpha}(x_\alpha)$ and $\hat{m}_{SBLL, \alpha}(x_\alpha)$ similarly constructed, i.e.,

$$
\hat{m}_{SBK, \alpha}(x_\alpha) = \left( \frac{1}{n} C_T^T W_\alpha C_K \right)^{-1} \frac{1}{n} C_K^T W_\alpha \hat{Y}_\alpha,
$$

where $\hat{Y}_\alpha = Y_i - \sum_{t=1}^{d_1} \left\{ \hat{m}_0 + \sum_{1 \leq \alpha' \leq d_2, \alpha' \neq \alpha} \hat{m}_{\alpha'}(X_{i\alpha}) \right\}$.

4.1 Application to Cobb-Douglas model

Liu and Yang (2010) applied the SBLL procedure to a varying coefficient extension of the Cobb-Douglas model for the U.S. GDP that allows nonneutral effects of the R&D on capital and labor as well as in total factor productivity (TFP). Denoted by $Q_t$ the US GDP at year $t$, $K_t$ the US capital at year $t$, $L_t$ the US labor at year $t$, $X_t$ the growth rate of ratio of R&D expenditure to GDP at year $t$, all data have been downloaded from the Bureau of Economic Analysis (BEA) website for years, $t = 1959, \ldots, 2002$ ($n = 44$). The standard Cobb-Douglas production function (Cobb and Douglas (1928)) is $Q_t = A_t K_t^{\beta_1} L_t^{1-\beta_1}$, where $A_t$ is the Total Factor Productivity (TFP) of year $t$, $\beta_1$ is a parameter determined by technology. Define the following stationary time series variables

$$
Y_t = \log Q_t - \log Q_{t-1}, T_{1t} = \log K_t - \log K_{t-1}, T_{2t} = \log L_t - \log L_{t-1},
$$

then the Cobb-Douglas equation implies the following simple regression model

$$
Y_t = (\log A_t - \log A_{t-1}) + \beta_1 T_{1t} + (1 - \beta_1) T_{2t}.
$$

According to Solow (1957), the total factor productivity $A_t$ has an almost constant rate of change, thus one might replace $\log A_t - \log A_{t-1}$ with an unknown constant and arrive at the following model

$$
Y_t - T_{2t} = \beta_0 + \beta_1 (T_{1t} - T_{2t}). \tag{28}
$$

As technology growth is one of the biggest sub-sections of TFP, it is reasonable to examine the dependence of both $\beta_0$ and $\beta_1$ on technology rather than treating them as fixed constants. We use exogenous variables $X_t$ (Growth rate of ratio of R&D expenditure to GDP at year $t$) to represent technology level and model $Y_t - T_{2t} = m_1(X_t) + m_2(X_t) (T_{1t} - T_{2t})$ where $m_l(X_t) = m_{a_l} + \sum_{a=1}^{d_2} m_{a_l}(X_{t-a+1})$, $l = 1, 2$, $X_t = (X_{t-1}, \ldots, X_{t-d_2+1})$. Using the BIC of Xue and Yang (2006b) for additive coefficient model with $d_2 = 5$, the following reduced model is considered optimal

$$
Y_t - T_{2t} = c_1 + m_{41}(X_{t-3}) + \{c_2 + m_{52}(X_{t-4})\} (T_{1t} - T_{2t}). \tag{29}
$$
The rolling forecast errors of GDP by SBLL fitting of model (29) and linear fitting of (28) are shown in Figure 1. The averaged squared prediction error (ASPE) for model (29) is

$$\frac{1}{9} \sum_{t=1994}^{2002} [Y_t - T_{2t} - \hat{c}_1 - \hat{m}_{\text{SBLL},11}(X_{t-3}) - \hat{c}_2 - \hat{m}_{\text{SBLL},11}(X_{t-4})] (T_{1t} - T_{2t})^2 = 0.001818,$$

which is about 60% of the corresponding ASPE (0.003097) for model (28). The in-sample averaged squared estimation error (ASE) for model (29) is $5.2399 \times 10^{-5}$, which is about 68% of the in-sample ASE ($7.6959 \times 10^{-5}$) for model (28).

In model (29), $\hat{c}_1 + \hat{m}_{\text{SBLL},11}(X_{t-3})$ estimates the TFP growth rate, which is shown as a function of $X_{t-3}$ in Figure 2. It is obvious that the effect of $X_{t-3}$ is positive when $X_{t-3} \leq 0.02$, but negative when $X_{t-3} > 0.02$. On average, the higher R&D investment spending causes faster GDP growing. However, overspending on R&D often leads to high losses (Tokic (2003)).

Liu and Yang (2010) also computed the average contribution of R&D to GDP growth for 1964-2001, which is about 40%. The GDP and estimated TFP growth rates is shown in Figure 5 (see Liu and Yang 2010), it is obvious that TFP growth is highly correlated to the GDP growth.

(Insert Figure 5 about here)

### 5 SBS in Additive Models

In this section, we describe the spline-backfitted spline estimation procedure for model (3). Let $0 = t_0 < t_1 < \ldots < t_{N+1} = 1$ be a sequence of equally spaced knots, dividing $[0,1]$ into $(N+1)$ subintervals of length $h = \frac{1}{N+1}$ with a preselected integer $N \sim n^{1/5}$ given in Assumption (A5) of Song and Yang (2010), and let $0 = t_0 < t_1 < \ldots < t_{N^*} = 1$ be another sequence of equally-spaced knots, dividing $[0,1]$ into $N^*$ subintervals of length $H = \frac{N}{N^*} \sim n^{2/5} \log n$ is another preselected integer, see Assumption (A5) in Song and Yang (2010). Denote by $G_\alpha$ the linear space spanned by \{1, $b_J(x_{i\alpha})\}_{J=1}^{N+1}$, whose elements are called linear splines, piecewise linear functions of $x_\alpha$ which are continuous on $[0,1]$ and linear on each subinterval $[t_J, t_{J+1}]$, $0 \leq J \leq N$. Denote by $G_{n,\alpha} \subset \mathbb{R}^n$ the corresponding subspace of $\mathbb{R}^n$ spanned by \{1, $b_J(X_{i\alpha})\}_{J=1}^{N+1}$. Similarly, define the \{1 + dN^*\}-dimensional space $G^* = G^*([0,1]^d)$ of additive constant spline functions as the linear space spanned by \{1, $I_{J^*}(x_{i\alpha})\}_{J^*=1}^{N^*}$, while denote by $G^*_{n,\alpha} \subset \mathbb{R}^n$ the corresponding subspace spanned by \{1, $I_{J^*}(X_{i\alpha})\}_{J^*=1}^{N^*}$. As $n \to \infty$, with probability approaching one, the dimension of $G_{n,\alpha}$ becomes $N + 2$, and the dimension of $G^*_{n,\alpha}$ becomes $1 + dN^*$.  

20
The additive function $m(x)$ has a multivariate additive regression spline (MARS) estimator $\hat{m}(x) = \hat{m}_n(x)$, the unique element of $G^*$ so that the vector $\{\hat{m}(x_1), \ldots, \hat{m}(x_n)\}^T \in G_n^*$ best approximates the response vector $Y$. Precisely

$$\hat{m}(x) = \text{argmin}_{g \in G^*} \sum_{i=1}^{n} \{Y_i - g(X_i)\}^2 = \lambda_0' + \sum_{a=1}^{d} \sum_{\alpha^*} \lambda_{\alpha^*,a}^i I_{\alpha^*}(x_\alpha),$$

where $(\lambda_0', \lambda_{1,1}', \ldots, \lambda_{N^*,d}')$ is the solution of the least squares problem

$$\{\lambda_0', \lambda_{1,1}', \ldots, \lambda_{N^*,d}'\}^T = \text{argmin}_{R(\mathbb{R}^n + 1)} \sum_{i=1}^{n} \left\{Y_i - \lambda_0 - \sum_{\alpha=1}^{d} \sum_{\alpha^*} \lambda_{\alpha^*,a} I_{\alpha^*}(X_{i\alpha})\right\}^2.$$

Estimators of each component function and the constant are derived as

$$\hat{m}_\alpha(x_\alpha) = \sum_{\alpha^*} \lambda_{\alpha^*,a}^i \left\{I_{\alpha^*}(x_\alpha) - n^{-1} \sum_{i=1}^{n} I_{\alpha^*}(X_{i\alpha})\right\},$$

$$\hat{m}_c = \lambda_0 + n^{-1} \sum_{i=1}^{n} \sum_{\alpha=1}^{d} \sum_{\alpha^*} \lambda_{\alpha^*,a} I_{\alpha^*}(X_{i\alpha}) = \hat{c} = Y.$$

These pilot estimators are used to define pseudo-responses $\hat{Y}_{i\alpha}, \forall 1 \leq \alpha \leq d$, which approximate the “oracle” responses $Y_{i\alpha}$. Specifically, we define

$$\hat{Y}_{i\alpha} = Y_i - \hat{c} - \sum_{\beta=1,\beta \neq \alpha}^{d} \hat{m}_{\beta}(X_{i\beta})$$

where $\hat{c} = Y_n = n^{-1} \sum_{i=1}^{n} Y_i$, which is a $\sqrt{n}$-consistent estimator of $c$ by central limit theorem for strongly mixing sequences. Correspondingly, we denote vectors

$$Y_\alpha = \{\hat{Y}_{1\alpha}, \ldots, \hat{Y}_{n\alpha}\}^T, Y_{1\alpha} = \{Y_{1\alpha}, \ldots, Y_{n\alpha}\}^T. \quad (30)$$

We define the spline-backfitted spline (SBS) estimator of $m_\alpha(x_\alpha)$ as $\hat{m}_{sbs}(x_\alpha)$ based on $\{\hat{Y}_{i\alpha}, X_{i\alpha}\}^{n}_{i=1}$, which attempts to mimic the would-be spline estimator $\hat{m}_{\alpha,S}(x_\alpha)$ of $m_\alpha(x_\alpha)$ based on $\{Y_{i\alpha}, X_{i\alpha}\}^{n}_{i=1}$ if the unobservable “oracle” responses $\{Y_{i\alpha}\}^{n}_{i=1}$ were available. Then,

$$\hat{m}_{\alpha,SBS}(x_\alpha) = \text{argmin}_{g_{\alpha} \in G_{\alpha}} \sum_{i=1}^{n} \{\hat{Y}_{i\alpha} - g_{\alpha}(X_{i\alpha})\}^2,$$

$$\hat{m}_{\alpha,S}(x_\alpha) = \text{argmin}_{g_{\alpha} \in G_{\alpha}} \sum_{i=1}^{n} \{Y_{i\alpha} - g_{\alpha}(X_{i\alpha})\}^2. \quad (31)$$

21
Theorem 8. Under Assumptions (A1) to (A5) in Song and Yang (2010), as \( n \to \infty \), the SBS estimator \( \hat{m}_{\alpha,SBS}(x) \) and the oracle smoother \( \hat{m}_{\alpha,S}(x) \) given in (31) satisfy

\[
\sup_{x \in [0,1]} |\hat{m}_{\alpha,SBS}(x) - \hat{m}_{\alpha,S}(x)| = o_p \left( n^{-2/5} \right).
\]

Theorem 8 provides that the maximal deviation of \( \hat{m}_{\alpha,SBS}(x) \) from \( \hat{m}_{\alpha,S}(x) \) over \([0,1]\) is of the order \( O_p \left( n^{-2/5} (\log n)^{-1} \right) = o_p \left( n^{-2/5} (\log n)^{1/2} \right) \), which is needed for the maximal deviation of \( \hat{m}_{\alpha,SBS}(x) \) from \( m_{\alpha}(x) \) over \([0,1]\) and the maximal deviation of \( \hat{m}_{\alpha,S}(x) \) from \( m_{\alpha}(x) \) to have the same asymptotic distribution, of order \( n^{-2/5} (\log n)^{1/2} \). The estimator \( \hat{m}_{\alpha,SBS}(x) \) is therefore asymptotically oracally efficient, i.e., it is asymptotically equivalent to the oracle smoother \( \hat{m}_{\alpha,S}(x) \) and in particular, the next theorem follows. The simultaneous confidence band given in (32) has width of order \( n^{-2/5} (\log n)^{1/2} \) at any point \( x_{\alpha} \in [0,1] \), consistent with published works on nonparametric simultaneous confidence bands such as Bosq (1998) and Claeskens and Van Keilegom (2003).

Theorem 9. Under Assumptions (A1)-(A5) in Song and Yang (2010), for any \( p \in (0,1) \), as \( n \to \infty \), an asymptotic 100 \((1-p)\)% simultaneous confidence band for \( m_{\alpha}(x) \) is

\[
\hat{m}_{\alpha,SBS}(x) \pm 2\hat{\sigma}_{\alpha}(x) \left\{ 3\Delta^T (x) \Xi_j(x) \Delta (x) \log (N + 1)/2\hat{f}_{\alpha}(x) nh \right\}^{1/2} \times \left[ 1 - \{2 \log (N + 1)\}^{-1} \left( \log (p/4) + \frac{1}{2} \left( \log \log (N + 1) + \log 4\pi \right) \right) \right], \tag{32}
\]

where \( \hat{\sigma}_{\alpha}(x) \) and \( \hat{f}_{\alpha}(x) \) are some consistent estimators of \( \sigma_{\alpha}(x) \) and \( f_{\alpha}(x) \), \( j(x) = \min \{ \lfloor x_{\alpha}/h \rfloor, N \} \), \( \delta(x) = \{ x_{\alpha} - t_{j(x)} \}/h \), and

\[
\Delta (x) = \begin{pmatrix}
 c_j(x) 1 - \delta(x) \\
 c_j(x) \delta(x)
\end{pmatrix}, \quad c_j = \begin{cases}
 \sqrt{2} & j = -1, N \\
 1 & 0 \leq j \leq N - 1
\end{cases},
\]

\[
\Xi_j = \begin{pmatrix}
 l_{j+1,j+1} & l_{j+1,j+2} \\
 l_{j+2,j+1} & l_{j+2,j+2}
\end{pmatrix}, \quad 0 \leq j \leq N,
\]

where terms \( \{l_{ik}\}_{|i-k|\leq1} \) are the entries of the inverse of the \( (N + 2) \times (N + 2) \)
matrix \( M_{N+2} \)

\[
M_{N+2} = \begin{pmatrix}
1 & \sqrt{2}/4 & 0 \\
\sqrt{2}/4 & 1 & 1/4 \\
1/4 & 1 & \ddots \\
\ddots & \ddots & 1/4 \\
0 & 1/4 & \sqrt{2}/4 & 1
\end{pmatrix}.
\]

6 Future Research

Fan and Jiang (2005) provides generalized likelihood ratio (GLR) tests for additive models using the backfitting estimator. Similar GLR test based on the two-step estimator is feasible for future research. The SBS method can also be applied to the PLAMs (4) and the ACMs (5). The two-step estimating procedure can be extended to generalized additive, partially linear additive, and additive coefficient models. Ma et al. (2012, Forthcoming) proposed a one-step penalized spline estimation and variable selection procedure in PLAMs with clustered/longitudinal data. The procedure is fast to compute, but lacks asymptotic distributions for the additive function components. Thus no confidence measures can be established. As another future work, we target to apply the two-step estimation to the analysis of clustered/longitudinal data, and to establish the oracle efficiency of the estimators. The confidence bands of each additive function can be constructed based on the same idea in Section 5.

References


Figure 1: Errors of GDP forecasts for a smooth coefficient model (solid line); Cobb-Douglas model (dotted line).
Figure 2: Estimation of TFP growth rate function.
Figure 3: Linearity test for the Boston housing data. Plots of null hypothesis curves of $H_0: m(x_\alpha) = a_\alpha + b_\alpha \cdot x_{\alpha}$, $\alpha = 1, 2, 3, 4$ (solid line), linear confidence bands (upper and lower thin lines), the linear spline estimator (dotted line) and the data (circle).
Figure 4: Plots of the least squares regression estimator (solid line), confidence bands (upper and lower thin lines), the spline estimator (dashed line) and the data (dot)
Figure 5: Estimation of function $\hat{c}_1 + \hat{m}_{SBLL,41} (X_{t-3})$: GDP growth rate–dotted line; $\hat{c}_1 + \hat{m}_{SBLL,41} (X_{t-3})$–solid line.