Variable Selection and Identification of High-Dimensional Nonparametric Additive Nonlinear Systems

Biqiang Mu, Wei Xing Zheng, *Fellow, IEEE*, and Er-Wei Bai, *Fellow, IEEE*

***Abstract*—This paper considers variable selection and identification of dynamic additive nonlinear systems via kernel-based nonparametric approaches, where the number of variables and additive functions may be large. Variable selection aims to find which additive functions contribute and which do not. The proposed variable selection consists of two successive steps. At the first step, one estimates each additive function by kernel-based nonparametric identification approaches without suffering from the curse of dimensionality. At the second step, a nonnegative garrote estimator is applied to identify which additive functions are nonzero by utilizing the obtained nonparametric estimates of each function. Under weak conditions, the nonparametric estimates of each additive function can achieve the same asymptotic properties as for 1D nonparametric identification based on kernel functions. It is also established that the nonnegative garrote estimator turns a consistent estimate for each additive function into a consistent variable selection with probability one as the number of samples tends to infinity. Two simulation examples are presented to verify the effectiveness of the variable selection and identification approaches proposed in the paper.**

***Index Terms*—Additive nonlinear systems, asymptotic normality, backfitting estimator, high-dimensional systems, nonnegative garrote estimator, set convergence, variable selection.**

I. Introduction

Linear system identification aims at searching for a linear system by means of fitting a set of input-output data generated from a practical system in some optimal sense. Its theory is considerably matured and a number of identification methods have been developed in the literature [1], [2]. Their effective ness however depends on whether the behavior of practical systems is linear or not. These methods will achieve good results if the practical system is indeed linear or is well approximated by a linear system. When systems are strongly nonlinear, these methods will lead to a large modeling error. Accordingly, development of identification methods for a nonlinear system becomes very necessary.

Nonlinear system identification can be roughly divided into two categories, parametric approaches and nonparametric approaches, according to the available *a priori* information. If the structure of an unknown system is available *a priori*, then the nonlinear system can be expressed by some nonlinear functions together with some unknown parameters. In this case, the system is actually characterized by these parameters and the resulting identification problem is a nonlinear optimization problem. This kind of methods are referred to as parametric approaches. If little *a priori* information on the structure of nonlinear systems is accessible, then identification approaches under such a setting are called nonparametric approaches. Nonparametric nonlinear identification is much harder.

Owing to lack of *a priori* structure information of the system under consideration, any nonparametric approach has to rely on a general structure. For example, the following nonlinear autoregressive systems with exogenous inputs (NARX)

$$y_k = f(y_{k-1}, \ldots, y_{k-s}, u_{k-1}, \ldots, u_{k-t}) + \varepsilon_k, \quad k = 1, \ldots, n \tag{1}$$

is widely used in the literature, where y_k, u_k and ε_k are the output, the input, and the observation noise at time k, respectively, and the integers s and t are the orders of AR-part and X-part of the system, respectively. It is clear that the dynamic behavior of the NARX system is completely characterized by the unknown $d = (s + t)$-dimensional nonlinear function $f(\cdot)$. The system (1) includes many common nonlinear systems as special cases, for example, Hammerstein systems [3]–[6] and Wiener systems [7]–[10]. A nonparametric approach for estimating nonlinear systems is usually implemented in a point-wise way without *a priori* structure information. For any given point x of interest, the value $f(x)$ is estimated by a weighted average of the observation points in a neighborhood of x. According to the

Manuscript received May 20, 2016; accepted August 17, 2016. Date of publication September 2, 2016; date of current version April 24, 2017. This work was supported in part by the President Fund of Academy of Mathematics and Systems Science, CAS under Grant 2015-hwyxqnrcmbq, the National Key Basic Research Program of China (973 Program) under Grant 2014CB845301, the National Natural Science Foundation of China under Grant 61603379, the Australian Research Council under Grant DP120104986, and the National Science Foundation under Grant CNS-1239509. Recommended by Associate Editor M. Verhaegen. *(Corresponding author: Wei Xing Zheng.)*

B. Mu is with the Key Laboratory of Systems and Control of CAS, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China and the School of Computing, Engineering and Mathematics, Western Sydney University, Sydney, NSW 2751, (e-mail: bqmu@amss.ac.cn).

W. X. Zheng is with the School of Computing, Engineering and Mathematics, Western Sydney University, Sydney, NSW 2751, Australia (e-mail: w.zheng@westernsydney.edu.au).

E.-W. Bai is with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA and the School of Electronics, Electrical Engineering and Computer Science, Queens University, Belfast, U.K (e-mail: er-wei-bai@uiowa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2016.2605741

TABLE I
SAMPLE SIZE n VERSUS DIMENSION d

d	1	2	3	4	5
n	100	1273	$1.91\mathrm{E}+04$	$3.24\mathrm{E}+05$	$6.08\mathrm{E}+06$
d	6	7	8	9	10
n	$1.24\mathrm{E}+08$	$2.71\mathrm{E}+09$	$6.31\mathrm{E}+10$	$1.55\mathrm{E}+12$	$4.02\mathrm{E}+13$

weight functions chosen, nonparametric approaches include direct weight optimization [11], spline smoothing [12], kernel estimators [13], [14], local polynomial estimators [15] and others. For instance, for the NARX system (1), a recursive multivariate kernel estimator was introduced in [16] based on multivariate kernel functions and stochastic approximation, and the corresponding recursive estimate was shown to converge to the true values with probability one. At the same time, the minimum mean squared error (MMSE) estimator for the NARX system (1) was studied in [17], where it was shown that local linear estimators (LLEs) are a linear asymptotic MMSE estimator for a wide class of nonlinear systems. Later, a recursive LLE (RLLE) for the NARX system (1) was proposed in [18], and the strong consistency and the asymptotical mean squared error properties were established.

Unfortunately, for a NARX system (1), any nonparametric approach mentioned above is only feasible for low-dimensional nonlinear systems. When the dimension $d = s + t$ of $f(\cdot)$ (the number of variables) is large, identification becomes more and more harder due to *the curse of dimensionality*. To clearly illustrate this point, let us cite an example given in [19]. Consider a d-dimensional regression function $f(\cdot)$. For simplicity, let d variables be distributed uniformly in $[-1, 1]$. Suppose that we are interested in some point $x_0 \in [-1, 1]^d$. To reliably estimate $f(x_0)$, there must be enough observations near x_0 due to the influence of noise and uncertainty. For ease of presentation, suppose that the neighborhood of x_0 is a ball of radius 0.1 centered at x_0. Thus, the probability that a sample is in the neighborhood of x_0 is $\frac{\pi^d 0.1^d}{2^d \Gamma(d/2+1)}$, where $\Gamma(\cdot)$ is the Gamma function. Suppose that 10 points in the neighborhood are adequate. Then on average to have 10 or more points in the neighborhood, the sample length n has to satisfy $n\frac{\pi^d 0.1^d}{2^d \Gamma(d/2+1)} \geq 10$ or $n \geq \frac{10\cdot 20^d \cdot \Gamma(d/2+1)}{\pi^d}$.

To feel the relationship between the required sample length n and the dimension d of the function, Table I presents the required number of samples in terms of the dimension d. It is seen that the required sample length n increases exponentially with d. This phenomenon is called *the curse of dimensionality*, which is a core problem for all local average approaches (not restricted to identification). The reason why this happens is the sparsity of a high-dimensional space.

According to the above analysis, any local averaged based nonparametric approach is infeasible due to the curse of dimensionality when the number of variables is large. To overcome this difficulty, additive nonlinear models have been proposed in the literature [20] in which each variable separately contributes to the output by a 1D function. An additive nonlinear system is an extension of linear systems and replaces each linear term by a 1D unknown nonparametric nonlinear function. Recall that in linear identification, one generally does not believe that the model is linear. Rather, a linear model is a good first-order approximation and so it can uncover important properties. Additive nonlinear models are obviously more general approximations. Since its inception, additive nonlinear models have been extensively studied in terms of fitting data to the system [20], [21] and becoming the mostly applied nonlinear models in the literature. For example, in medical applications, additive models are used for studying factors effecting patterns of insulin-dependent diabetes mellitus in children [22], for investigating dependence of the level of serum C-peptide on various heart attacks to establish the intensity of ischaemic heart disease risk factors in high-incidence regions [20] and for evaluating treatment efficacy in clinical trials [23]. In environmental research, additive models are adopted to predict the atmospheric ozone concentration [24] and to study the rainfalls [25]. Additive models are also utilized in many other areas, e.g., for economics and consumer behavior in microeconomics [26]. Though extensively used, identification of such additive systems has not received much attention in the field of system identification. While many works on the additive (static) models have emerged in the regression analysis area [20], there has been little result on identification of dynamic additive nonlinear systems. The main difficulties lie in interconnection of all the additive functions and unavailability of the structural information of the additive functions. To the best of our knowledge, the literature in the system identification field only includes identification of additive systems with very restrictive structures [21], [27], [28]. It is unclear whether the identification method in [21], [27], [28] can be extended to higher-order additive nonlinear systems up to now. Further, variable selection for an additive nonlinear system is virtually untouched in the literature. This is an important topic. Due to lack of *a priori* information, the structure of the model has to be assumed to be rich enough to contain the true but unknown nonlinear system. In other words, the model may include many variables or functions that do not contribute. A consequence is that the resultant model is sensitive in terms of prediction and analysis. The goal of variable selection is to identify those variables that do not contribute, and once identified, those variables will be removed from the model.

This paper aims at variable selection and identification of high-dimensional dynamic additive nonlinear systems based on nonparametric kernel function approaches. One of the goals is to correctly find all the variables that contribute without suffering from the curse of dimensionality. The existing literatures [29], [30] on variable selection of additive nonlinear models are mainly based on a spline approximation to the additive functions, in which each function is represented by a linear combination of spline basis functions. Thus, variable selection for additive nonlinear models becomes a linear problem but an approximated one [31]. To the best of our knowledge, there has not been any result reported on variable selection for additive nonlinear models if functions are unknown and nonparametric. Variable selection proposed in this paper is composed of two successive steps: 1) performing nonparametric identification of dynamic additive nonlinear systems based on kernel functions; 2) applying nonnegative garrote estimators [32], [33] to find the nonzero functions by the nonparametric estimates of the additive functions obtained at Step 1). The nonparametric identification at Step 1) is conducted in an iterative way and does not suffer from the curse of dimensionality since only 1D and 2D kernel estimations are involved. Further, the estimate for each additive function can reach the optimal rate of convergence as if other functions are exactly known. That is to say, each additive function can be estimated with the same accuracy as a 1D function. The nonnegative garrote estimator can turn a consistent initial nonparametric estimation into an estimate that is consistent not only in terms of estimation but also in terms of variable selection.

The rest of the paper is organized as follows. Static additive nonlinear models, additive nonlinear autoregressive system with exogenous input (ANARX), and nonparametric identification for 1D nonlinear function based on kernel functions are successively introduced in Section II. In Section III, variable selection algorithms for a static additive nonlinear model are proposed based on smooth backfitting kernel estimators and nonnegative garrote estimators, respectively. Section IV extends variable selection for the static additive nonlinear model in Section III to the ANARX, and the detailed identification algorithms are also given. Furthermore, the corresponding convergence properties on kernel-based nonparametric identification and variable selection for the ANARX are established. Section V presents two simulation examples to show the performance of the variable selection methods proposed in this paper for the static additive nonlinear models and the ANARX, respectively. Section VI provides some concluding remarks. Finally, the main theoretical proofs are collected in the Appendix.

Notation: $P(\cdot)$ represents the probability of a set and E denotes the expectation. Let $\{M_n\}$ be a sequence of real numbers or random variables. $M_n = o(1)$ means that $M_n \to 0$ as $n \to \infty$ if $\{M_n\}$ is a deterministic sequence, and indicates $M_n \to 0$ with probability one as $n \to \infty$ if $\{M_n\}$ is a random process. Similarly, $M_n = O(1)$ means that $\{M_n\}$ is bounded by a finite positive number if $\{M_n\}$ is a deterministic sequence, and represents that $\{M_n\}$ is bounded uniformly over n with probability one by a finite positive number if $\{M_n\}$ is a random process. Meanwhile, let $\{M_n\}$ be a sequence of random variables. Then, $M_n = o_P(1)$ means that $\{M_n\}$ converges in probability to zero, i.e., $\forall \epsilon > 0$, $P(|M_n| > \epsilon) \to 0$ as $n \to \infty$, while $M_n = O_P(1)$ represents that $\{M_n\}$ is bounded in probability (or stochastically bounded), i.e., $\forall \epsilon > 0\ \exists L > 0$ such that $P(|M_n| > L) < \epsilon$, $\forall n$. All integrals are taken over the support of the relevant variables, so the lower and upper limits of the definite integral are omitted throughout the paper.

II. Problem Formulation

Our goal is to study variable selection and nonparametric identification of additive nonlinear models. First, we consider a static additive nonlinear model described by

$$y_k = f_0 + \sum_{j=1}^{d} f_j(x_{kj}) + \varepsilon_k,\ k = 1, \ldots, n \tag{2}$$

where $\{y_k, x_k, k = 1, \ldots, n\}$ with $x_k = [x_{k1}, \ldots, x_{kd}]$ are n input-output samples from a random vector $\{\mathbf{Y}, \mathbf{X}\}$ with $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_d]$, d the number (dimension) of variables, f_0 an unknown constant term, $f_j(\cdot)$'s unknown 1D additive functions, and ε_k an observation noise. For identifiability, assume that $f_0 = E\mathbf{Y}$ and $Ef_j(\mathbf{X}_j) = 0$ for $j = 1, \ldots, d$ if the process $\{y_k, x_k\}$ is stationary. Denote by $f_j = [f_j(x_{1j}), f_j(x_{2j}), \ldots, f_j(x_{nj})]^T$ the column vectors composed of the values of the additive functions $f_j(\cdot)$, $j = 1, \ldots, d$ at the observation points $\{x_{1j}, \ldots, x_{nj}\}$. Thus, the model (2) is expressible as

$$Y = f_0\mathbf{1}_n + \sum_{j=1}^{d} f_j + \varepsilon, \tag{3}$$

where $Y = [y_1, \ldots, y_n]^T$, $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^T$, and $\mathbf{1}_n$ represents an n-dimensional column vector with all elements being 1. This model is widely referenced in the statistical literature and used in practice.

The second model under consideration is an additive nonlinear autoregressive system with exogenous input (ANARX), which is more popular to the system identification community and described as follows:

$$\begin{aligned} y_k = \ & f_0 + f_1(y_{k-1}) + \cdots + f_s(y_{k-s}) + f_{s+1}(u_{k-1}) \\ & + \cdots + f_{s+t}(u_{k-t}) + \varepsilon_k,\ k = 1, \ldots, n, \end{aligned} \tag{4}$$

where u_k is the input, y_k is the output observation corrupted by the noise ε_k, and s and t are the delayed orders of the autoregressive part and the exogenous part, respectively. Similar to the setting in the model (2), f_0 is an unknown constant term, and $f_j(\cdot), j = 1, \ldots, s+t$ are some unknown univariate nonlinear functions. Let the regressor vector $\phi_k = [y_k, \ldots, y_{k-s+1}, u_k, \ldots, u_{k-t+1}]^T$ and the function $f(\phi_k) = f_0 + f_1(y_k) + \cdots + f_s(y_{k-s+1}) + f_{s+1}(u_k) + \cdots + f_{s+t}(u_{k-t+1})$. The system (4) can be rewritten in a compact form as $y_{k+1} = f(\phi_k) + \varepsilon_{k+1}$. The well-known Hammerstein systems extensively studied in the literature [3]–[6] are a special case of the model (4). Similarly, for identifiability, assume that $f_0 = Ey_k, Ef_j(y_k) = 0,\ j = 1, \ldots, s,\ Ef_{s+l}(u_k) = 0,\ l = 1, \ldots, t$ if the process $\{\phi_k\}$ is stationary. Further, let $Y = [y_1, y_2, \ldots, y_n]^T$, $f_j = [f_j(y_{1-j}), f_j(y_{2-j}), \ldots, f_j(y_{n-j})]^T$, $j = 1, \ldots, s$, $f_{s+l} = [f_{s+l}(u_{1-l}), f_{s+l}(u_{2-l}), \ldots, f_{s+l}(u_{n-l})]^T$, $l = 1, \ldots, t$ and $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^T$. The ANARX model (4) can be rewritten in the form of the model (3) as

$$Y = f_0\mathbf{1}_n + \sum_{j=1}^{d} f_j + \varepsilon,\ \ d \overset{\triangle}{=} s + t. \tag{5}$$

As a result, the static additive model (3) can be viewed as a special case of the ANARX model (4). In particular, by defining

$$x_{kj} = \begin{cases} y_{k-j}, & j = 1, \ldots, s,\ k = 1, 2, \ldots, n \\ u_{k-j+s}, & j = s+1, \ldots, d,\ k = 1, 2, \ldots, n \end{cases}$$

the model (2) becomes a special case of (4).

Though notation-wise, the above two models look similar, we comment that there are some fundamental differences. In fact, theoretical analysis of variable selection and identification are much harder for the ANARX model (4) than that of the static model (2). The main difficulty lies in the dynamics of the ANARX system so that the strict stationarity and strong mixing conditions required are easily satisfied for the static model but not necessarily so for the ANARX model unless some additional assumptions are imposed.

Since this paper intends to give the nonparametric estimation of each additive function based on kernel functions, here we first briefly introduce the two commonly used kernel-based nonparametric estimators: the kernel estimator and the local linear estimator, for a 1D case $y_k = f(x_k) + \epsilon_k, x_k \in \mathbb{R}$. High-dimensional cases can be defined similarly.

The nonparametric approach for estimating nonlinear functions is usually implemented in a point-wise way and the idea is that only local observation points are useful for estimating a nonlinear function at some points due to lack of *a priori* structure information. For any given point x of interest, the value $f(x)$ is estimated by a weighted average of the observation points in a neighborhood of x. For kernel-based nonparametric estimators, the weighted average is implemented by a kernel function which gives the points near x bigger weights and the points far from x smaller weights. Kernel functions usually have a shape of a probability density function. It follows that nonparametric estimates at a point can use only a part of all the observation points due to locally weighted average, but all the observations are used for estimating the whole function. A common assumption on a kernel function $K(\cdot)$ is given as follows.

Assumption 1: The kernel $K(\cdot)$ is nonnegative, bound, has compact support, and satisfies $\int K(x)\mathrm{d}x = 1$ and $\int x^2 K(x)\mathrm{d}x < \infty$. Also, $K(\cdot)$ is symmetric about zero, and is Lipschitz continuous, i.e., there exists a positive real number C_1 such that $|K(s) - K(t)| \leq C_1|s - t|$.

The well-known kernel functions including Epanechnikov, uniform, triangle, biweight, etc. [34] satisfy Assumption 1.

The (Nadaraya-Watson) kernel estimator

The (Nadaraya-Watson) kernel estimator, proposed in [13], [14], for estimating a univariate nonlinear function at some point $x \in \mathbb{R}$ of interest, is given by

$$\widehat{f}(x) = \sum_{k=1}^{n} K_h(x_k - x)y_k \Big/ \sum_{k=1}^{n} K_h(x_k - x)$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and h is the bandwidth of $K_h(\cdot)$. In many cases (for example, calculating the prediction error $\frac{1}{n}\sum_{k=1}^{n}(y_k - \widehat{f}(x_k))^2$), we are more interested in the values of $\widehat{f}(\cdot)$ at the observations $\{x_1, \ldots, x_n\}$. Then $\widehat{f}(x_k) = s_k Y$, where $s_k = [K_h(x_1 - x_k), \ldots, K_h(x_n - x_k)]/\sum_{i=1}^{n} K_h(x_i - x_k)$ and $Y = [y_1, y_2, \ldots, y_n]^T$. Denote $\widehat{f} = [\widehat{f}(x_1), \ldots, \widehat{f}(x_n)]^T$. We have $\widehat{f} = SY$, where $S = [s_1^T, \ldots, s_n^T]^T$. Note that $\widehat{f}$ is a linear transformation of the output vector Y and S depends only on the kernel functions $K(\cdot)$ and the observation points $\{x_1, \ldots, x_n\}$. So S is called a linear smoother matrix corresponding to the kernel estimator.

The local linear estimator

Suppose that we are interested in the value of $f(\cdot)$ at x. By the Taylor expansion, $f(x_k) \approx f(x) + f'(x)(x_k - x)$ for the observations x_k in a neighborhood of x. As a result, the prediction error criterion $\sum_{k=1}^{n}(y_k - f(x_k))^2$ can be approximated by $\sum_{k=1}^{n}(y_k - f(x) - f'(x)(x_k - x))^2$. Noting that the Taylor expansion only holds in a small neighborhood of x, we add the kernel function to control the size of the neighborhood and hence the criterion function becomes $\sum_{k=1}^{n}(y_k - f(x) - f'(x)(x_k - x))^2 K_h(x_k - x)$. Therefore, $f(\cdot)$ and its derivative $f'(\cdot)$ at x can be estimated by minimizing

$$\sum_{k=1}^{n} \big(y_k - a - b(x_k - x)\big)^2 K_h(x_k - x) \tag{6}$$

over two parameters (a, b) if $x \in \mathbb{R}$. The solution of (6) is called the local linear (LL) estimator. Given that the objective function (6) is a quadratic function over (a, b), the LL estimator has an explicit form

$$[\widehat{f}(x), \widehat{f}'(x)]^T = (X^T W X)^{-1} X^T W Y, \tag{7}$$

where

$$X^T = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 - x & \cdots & x_n - x \end{bmatrix},$$

$$W = \operatorname{diag}\,[K_h(x_1 - x), \ldots, K_h(x_n - x)].$$

It is observed that the LL estimator $\widehat{f}(x)$ of $f(x)$ is also a linear combination of the output vector Y, and hence $\widehat{f} = [\widehat{f}(x_1), \ldots, \widehat{f}(x_n)]^T$ also has the form of $\widehat{f} = SY$, where each row of S corresponds to the weights of the LL estimator and S is referred to as a linear smoother matrix corresponding to the LL estimator. By some straightforward calculations, it is seen that the kernel estimator is the minimizer of $\sum_{k=1}^{n}(y_k - a)^2 K_h(x_k - x)$ over a and hence the kernel estimator is also called the local constant estimator sometimes.

III. Variable Selection and Identification of a Static Additive Nonlinear Model

In this section, we first consider the variable selection of the static additive nonlinear model (2).

A. Variable Selection and Set Convergence Analysis

Recall that variable selection is to find which 1D function $f_j(\cdot)$ contributes and which one does not. This problem is very important when the dimension d is very large. Denote the index set of nonzero functions and its complement respectively by

$$\mathcal{I} = \{j : f_j(\cdot) \not\equiv 0\},\ \mathcal{I}^c = \{j : f_j(\cdot) \equiv 0\} = \{1, \ldots, d\} \setminus \mathcal{I}$$

where $f_j(\cdot) \not\equiv 0$ represents that $f_j(\cdot)$ is not identical to zero, namely, the measure that $f_j(\cdot)$ is unequal to zero is greater than 0. Variable selection proposed herein contains two steps. First, a consistent estimate $\widehat{f}_j$ of f_j is sought, which will be discussed later. Then, the nonnegative garrote estimator is adopted here to identify the set $\mathcal{I}$ by

$$\min_{c} \frac{1}{2}\Big\|Y - \sum_{j=1}^{d} c_j \widehat{f}_j\Big\|^2 + \lambda_n \sum_{j=1}^{d} c_j \tag{8}$$

over $c = [c_1, \ldots, c_d]^T$ with the constraints $c_j \geq 0$, $j = 1, \ldots, d$, where $\widehat{f}_j$ is a consistent estimate of f_j for $j = 1, \ldots, d$, namely, $\|\widehat{f}_j - f_j\|/n = o(n^{-\beta})$, $\beta > 0$, $j = 1, \ldots, d$, and $\lambda_n > 0$ is a tuning parameter. Denote the minimizer of (8) by $\widehat{c} = [\widehat{c}_1, \ldots, \widehat{c}_d]^T$, which implies that the function $f_j(\cdot) \equiv 0$ if $\widehat{c}_j = 0$; otherwise $f_j(\cdot) \not\equiv 0$. Thus, the nonnegative garrote estimate of the jth additive nonlinear function is given by

$$\widehat{f}_j^{\mathrm{NG}} = \widehat{c}_j \widehat{f}_j,\ j = 1, 2, \ldots, d. \tag{9}$$

The theorem to be given below shows that the nonnegative garrote estimator will produce consistent variable selections given that the $\widehat{f}_j$'s are consistent estimates and the tuning parameter λ_n is appropriately chosen. The nonnegative garrote estimator was first proposed in [32], [33] for linear models and later extended to additive nonlinear models [30]. However, in [30] the nonlinear additive functions were approximated by splines which essentially reduces a nonlinear problem to a linear problem, but an approximated one. In the current paper, $f_j(\cdot)$'s are nonparametric and no attempt is made to approximate them by linear combinations of basis functions.

Assumption 2: $\|(f_{\mathcal{I}}^T f_{\mathcal{I}}/n)^{-1}\| < \infty$, where $f_{\mathcal{I}}$ is the matrix composed of the column vectors f_j with $j \in \mathcal{I}$ and $\|f_j\|/\sqrt{n} < \infty$ for $j \in \mathcal{I}$.

Actually, the condition $\|(f_{\mathcal{I}}^T f_{\mathcal{I}}/n)^{-1}\| < \infty$ in Assumption 2 is a natural extension of the persistent excitation condition of linear models to the additive nonlinear models. To see this, let $f_j(x_{kj}) = \beta_j x_{kj}$ with $\beta_j \neq 0$ for $j \in \mathcal{I}$. Thus the conditions

$\|(f_{\mathcal{I}}^T f_{\mathcal{I}}/n)^{-1}\| < \infty$ becomes

$$\frac{1}{n}\|\beta^T X^T X\beta\| > 0 \tag{10}$$

where X is the design matrix with its columns composed of the observation points $[x_{1j}, x_{2j}, \ldots, x_{nj}]^T$ for $j \in \mathcal{I}$ and $\beta = \mathrm{diag}[\beta_1, \beta_2, \ldots, \beta_n]$. Since $\beta_j \neq 0$ for $j \in \mathcal{I}$, β is nonsingular. Thus the formula (10) results in $\frac{1}{n}\|X^T X\| > 0$. It is well-known that this is the persistent excitation condition of linear models. If the condition $\|f_j\|/\sqrt{n} < \infty$ is contradicted for some $j \in \mathcal{I}$, then $\frac{1}{n}\sum_{k=1}^n f_j(x_{kj})^2 \to \infty$. This means that the sequence $\{f_j(x_{1j}), f_j(x_{2j}), \ldots\}$ diverges. In this case, the output sequence $\{y_1, y_2, \ldots,\}$ also diverges and hence in this case the identifiability is lost. Meanwhile, when the observation points $\{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ come from a continuous distribution with compact support, which is an assumption used in the next subsection for identifying the additive functions, we can see that

$$\frac{\|f_j\|}{\sqrt{n}} = \left(\frac{1}{n}\sum_{k=1}^n f_j(x_{kj})^2\right)^{1/2} \to \left(\int f^2(x)p_j(x)\mathrm{d}x\right)^{1/2}$$

and the condition $\|f_j\|/\sqrt{n} < \infty$ for $j \in \mathcal{I}$ holds for any function such that the integral on the right-hand side is finite and this is easily satisfied.

Theorem 1: Consider the model (3). Suppose that Assumption 2 holds and $\|\widehat{f}_j - f_j\|^2/n = O_P(\delta_n^2),\ j = 1, \ldots, d$, where $\delta_n \to 0$. If the tuning parameter λ_n satisfies $\lambda_n/n \to 0$ and $\delta_n = o(\lambda_n/n)$, then we have as $n \to \infty$:

1) $P(\widehat{c}_j = 0) \to 1$ for any $j \in \mathcal{I}^c$,
2) $P(\widehat{c}_j > 0) \to 1$ for any $j \in \mathcal{I}$ and $\sup_{j\in\mathcal{I}}\|f_j - \widehat{f}_j^{\mathrm{NG}}\|^2/n = O_P(\lambda_n^2/n^2)$.

Proof: See the Appendix. ■

It is seen from Theorem 1 that the nonnegative garrote estimate (9) of the nonzero additive functions converges at a different and slower rate than its initial consistent estimation.

B. Consistent Estimates $f_j(\cdot)$'s of Static Additive Nonlinear Models

From the previous subsection, if consistent estimates of f_j's can be obtained, then the variable selection problem is resolved. In this subsection, we focus on various ways to find consistent estimates $\widehat{f}_j$'s for the static model (2).

Some notation used in Section II for univariate regression function will be expanded to corresponding d-dimensional vector expressions. For example, the bandwidth $h = [h_1, \ldots, h_d]^T$, $p(\cdot)$ is the joint density of $\mathbf{X}$ and $p_j(\cdot)$ is the marginal density of $\mathbf{X}_j$ for $j = 1, \ldots, d$. For convenience, we also use the notation $v_{-j} = [v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d]^T \in \mathbb{R}^{d-1}$ for any vector $v = [v_1, v_2, \ldots, v_d]^T \in \mathbb{R}^d$.

The simplest case is that the inputs x_{ki}'s and x_{lj}'s are statistically independent for $k \neq l$ or $i \neq j$. Then both the kernel estimator and the LL estimator introduced above can be directly applied to identify the additive nonlinear function $f_j(\cdot)$ at the observation points $\{x_{1j}, \ldots, x_{nj}\}$. The estimate $\widehat{f}_j$ of f_j is given by $\widehat{f}_j = S_j Y,\ j = 1, \ldots, d$, where S_j is either the kernel or LL linear smoother matrix corresponding to fitting the data $\{y_k, x_{kj}, k = 1, \ldots, n\}$ as a univariate nonparametric identification. For identifiability, it is required that $Ef_j(\mathbf{X}_j) = 0$. Hence, we obtain the simple smoother estimator

$$\widehat{f}_j = S_j Y - E(S_j Y),\ j = 1, \ldots, d. \tag{11}$$

The convergence of (11) can be derived. The problem is that a simple smoother estimator like (11) does not work when x_{ki}'s are correlated.

To overcome the problem of simple smoother estimators, we propose a smooth backfitting algorithm. Two versions are presented in this subsection, based on the idea of the kernel estimator and the LL estimator, respectively.

Let $\mathcal{F}_j^l$ be the σ-algebra generated by the random variables $\{y_k, x_{k1}, \ldots, x_{kd}, 0 \leq j \leq k \leq l\}$. The process $\{y_k, x_{k1}, \ldots, x_{kd}, k \geq 0\}$ is called strongly mixing [35] if

$$\sup_{l, A\in\mathcal{F}_0^l, B\in\mathcal{F}_{l+k}^\infty} |P(AB) - P(A)P(B)| \ \triangleq\ \alpha(k) \to 0\text{as } k \to \infty.$$

Let us first give the required conditions.

Assumption 3:

i) The d-dimensional variables $\mathbf{X}{=}[\mathbf{X}_1, \ldots, \mathbf{X}_d]$ has compact support $\mathsf{I} = \mathsf{I}_1 \times \cdots \times \mathsf{I}_d$ for bounded interval $\mathsf{I}_j, j = 1, \ldots, d$. The joint density $p(v)$ of $\mathbf{X}$ and the densities $p^l(x^0, y^0)$ of $(x_k, x_{k+l}), l = 1, \ldots,$ are uniformly bounded. Furthermore, $p(v) > 0$ on the support I.
ii) For some $\theta > 2$, $E|y_k|^\theta < \infty$. Let $\sigma^2 = \mathrm{Var}(\varepsilon_k)$.
iii) The second-order derivatives $f_j''(\cdot)$ of the additive functions $f_j(\cdot), j = 1, \ldots, d$, exist and are Lipschitz continuous. The first partial derivatives of $p(v)$ exist and are continuous.
iv) The conditional densities $p_{\mathbf{X}|\mathbf{Y}}(x|y)$ of $\mathbf{X}$ given $\mathbf{Y}$ and $p_{x_k, x_{k+l}|y_k, y_{k+l}}(x^0, x^l|y^0, y^l)$ of (x_k, x_{k+l}) given $(y_k, y_{k+l}), l = 1, \ldots,$ exist and are bounded from above.
v) The process $\{y_k, x_{k1}, \ldots, x_{kd}\}$ is strongly mixing with $\sum_{k=1}^\infty k^b\alpha(k)^{1-2/\xi} < \infty$ for some $2 < \xi \leq \theta$ and $b > 1 - 2/\xi$.
vi) The mixing coefficients satisfy $\sum_{k=1}^\infty \varphi(k, j, l) < \infty$ and $\sum_{k=1}^\infty \phi(k, j, l) < \infty$ for $j = 1, \ldots, d, l = 1, 2,$ where $\varphi(k, j, l) = (kL_1(k))/r_1(k))(kT_k^2/h_j^l \log k)^{1/4}$ $\alpha(r_1(k))$ with $r_1(k) = (kh_j^l/T_k \log k)^{1/2}$ and $L_1(k) = (kT_k^2/h_j^{l+2}\log k)^{l/2}$ with $T_k{=}(k \log k(\log\log(k))^{1+\delta})^{1/\theta}$ for some $0 < \delta < 1$, while $\phi(k, j, l) = (kL_2(k)/r_2(k))$ $\times(k/h_j^l \log k)^{1/4}\alpha(r_2(k))$ with $r_2(k){=}(kh_j^l/\log k)^{1/2}$ and $L_2(k) = (k/h_j^{l+2} \log k)^{l/2}$.

The assumptions v) and vi) that the mixing coefficients need to satisfy seem a little complicated, but the mixing coefficients decaying exponentially to zero (i.e., $\alpha(k) = O(\rho^k)$ for some $0 < \rho < 1$) satisfy the assumptions v) and vi), which include many common random processes.

Smooth backfitting algorithms are a projection, which directly projects the samples $\{y_k, k = 1, \ldots, n\}$ onto the space of additive functions with a multivariate kernel density weight. In fact, it will be seen from a brief proof of the algorithms given in the Appendix that the backfitting algorithms are a method of alternating projections with a linear iterative form and the norm of the resulting linear projection operator is smaller than unity. This interpretation will enable us to understand the convergence of the backfitting algorithms. For simplicity of notation in the following derivations, let us denote

$$\widehat{p}_j(v_j) = \frac{1}{n}\sum_{k=1}^{n} K_{h_j}(v_j - x_{kj}), \tag{12}$$

$$\widehat{p}_j^{j}(v_j) = \frac{1}{n}\sum_{k=1}^{n} K_{h_j}(v_j - x_{kj})(v_j - x_{kj}),$$

$$\widehat{p}_j^{jj}(v_j) = \frac{1}{n}\sum_{k=1}^{n} K_{h_j}(v_j - x_{kj})(v_j - x_{kj})^2,$$

$$\widehat{p}_{jl}(v_j, v_l) = \frac{1}{n}\sum_{k=1}^{n} K_{h_j}(v_j - x_{kj})K_{h_l}(v_l - x_{kl}), \tag{13}$$

$$\widehat{p}_{jl}^{l}(v_j, v_l) = \frac{1}{n}\sum_{k=1}^{n} K_{h_j}(v_j - x_{kj})K_{h_l}(v_l - x_{kl})(v_l - x_{kl}),$$

$$\widehat{p}_{jl}^{jl}(v_j, v_l) = \frac{1}{n}\sum_{k=1}^{n} K_{h_j}(v_j - x_{kj})K_{h_l}(v_l - x_{kl}) \times (v_j - x_{kj})(v_l - x_{kl}).$$

We now discuss smooth backfitting kernel estimators. Let $v = [v_1, \ldots, v_d] \in \mathsf{I}$, where I is the support of the input variables. The smooth backfitting kernel estimator is defined as the minimizer of the criterion function

$$\frac{1}{2n}\int \sum_{k=1}^{n}\left(y_k - \bar{f}_0 - \sum_{j=1}^{d}\bar{f}_j(v_j)\right)^2 \prod_{r=1}^{d} K_{h_r}(v_r - x_{kr})\mathrm{d}v \tag{14}$$

where the minimization runs over all additive functions $\bar{f}(v) = \bar{f}_0 + \sum_{j=1}^{d}\bar{f}_j(v_j)$ with constraints $\int \bar{f}_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j = 0$. Using the Lagrange multipliers, the constrained functional optimization (14) is transformed into the unconstrained functional optimization

$$\frac{1}{2n}\int \sum_{k=1}^{n}\left(y_k - \bar{f}_0 - \sum_{j=1}^{d}\bar{f}_j(v_j)\right)^2 \prod_{r=1}^{d} K_{h_r}(v_r - x_{kr})\mathrm{d}v + \sum_{j=1}^{d}\lambda_j \int \bar{f}_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j. \tag{15}$$

Let $\{\widetilde{f}_0, \widetilde{f}_1(v_1), \ldots, \widetilde{f}_d(v_d)\}$ be the minimizer of (15). According to the method of variation, we find that $\{\widetilde{f}_0, \widetilde{f}_j(v_j), j = 1, \ldots, d\}$ satisfy the system of equations

$$\widetilde{f}_0 = \frac{1}{n}\sum_{k=1}^{n} y_k - \sum_{j=1}^{d}\int \widetilde{f}_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j, \tag{16}$$

$$\int \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \widetilde{f}_0 - \sum_{j=1}^{d}\widetilde{f}_j(v_j)\right)\prod_{r=1}^{d} K_{h_r}(v_r - x_{kr})\mathrm{d}v_{-j} - \lambda_j\widehat{p}_j(v_j) = 0. \tag{17}$$

Note the constraint $\int \widetilde{f}_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j = 0$. We have $\widetilde{f}_0 = \frac{1}{n}\sum_{k=1}^{n} y_k$. At the same time, (17) is simplified as

$$\widehat{f}_j(v_j)\widehat{p}_j(v_j) - \widetilde{f}_0\widehat{p}_j(v_j) - \widetilde{f}_j(v_j)\widehat{p}_j(v_j) - \sum_{l\neq j}\widetilde{f}_l(v_l)\widehat{p}_{jl}(v_j, v_l)\mathrm{d}v_l - \lambda_j\widehat{p}_j(v_j) = 0$$

where $\widehat{f}_j(v_j)$ is the univariate kernel estimator of the jth function $f_j(\cdot)$

$$\widehat{f}_j(v_j) \triangleq \frac{\frac{1}{n}\sum_{k=1}^{n} K_{h_j}(v_j - x_{kj})y_k}{\widehat{p}_j(v_j)}. \tag{18}$$

By moving the terms, one obtains

$$\widetilde{f}_j(v_j) = \widehat{f}_j(v_j) - \sum_{l\neq j}\int \widetilde{f}_l(v_l)\frac{\widehat{p}_{jl}(v_j, v_l)}{\widehat{p}_j(v_j)}\mathrm{d}v_l - \widetilde{f}_0 - \lambda_j$$

and by using the constraints $\int \widetilde{f}_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_i = 0$, one derives that the Lagrange multipliers satisfy $\lambda_j = 0$. As a result, the system of equations is simplified as $(j = 1, \ldots, d)$

$$\widetilde{f}_j(v_j) = \widehat{f}_j(v_j) - \sum_{l\neq j}\int \widetilde{f}_l(v_l)\frac{\widehat{p}_{jl}(v_j, v_l)}{\widehat{p}_j(v_j)}\mathrm{d}v_l - \overline{Y} \tag{19}$$

where the sample mean $\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} y_k$ of Y is a $\sqrt{n}$-consistent estimation of $E\mathbf{Y}$, which is faster than the convergence rate $O_P(1/n^{2/5})$ of nonparametric identification. In the following, the backfitting algorithm for solving (19) is provided. One starts with an arbitrary initial guess $\widetilde{f}_j^{(0)}(v_j)$ for $\widetilde{f}_j(v_j)$, for example, one can choose the univariate kernel estimators: $\widetilde{f}_j^{(0)}(v_j) = \widehat{f}_j(v_j)$. The jth function at the kth step is updated as follows:

$$\widetilde{f}_j^{(k)}(v_j) = \widehat{f}_j(v_j) - \sum_{l<j}\int \widetilde{f}_l^{(k)}(v_l)\frac{\widehat{p}_{jl}(v_j, v_l)}{\widehat{p}_j(v_j)}\mathrm{d}v_l - \sum_{l>j}\int \widetilde{f}_l^{(k-1)}(v_l)\frac{\widehat{p}_{jl}(v_j, v_l)}{\widehat{p}_j(v_j)}\mathrm{d}v_l - \overline{Y} \tag{20}$$

and the algorithm iterates over k until a predetermined convergence criterion is satisfied. The integrals are computed by numerical integrals.

The backfitting algorithm (20) can be implemented on a grid in the support of $\mathbf{X}$. The merit of doing this is that the size of the grid can be fixed and will not increase with the sample size n, particularly when n is large. This means that the computational complexity of the backfitting algorithm (20) is not relevant to the sample size n. Denote the preset grid of the jth variable by $v_j^0 = [v_{1j}^0, \ldots, v_{mj}^0]^T$, where m is the number of the grid points.

The Smooth backfitting kernel estimator (SBKE)

Step 1: Use the observation points $\{y_k, x_{kj}\}_{k=1,\ldots,n}^{j=1,\ldots,d}$ and the kernel function $K(\cdot)$ to calculate the values of 1D density estimates $\widehat{p}_j$ of $p_j(\cdot)$ and kernel estimates $\widehat{f}_j$ of $f_j(\cdot)$ at the points $\{v_{ij}^0\}_{i=1,\ldots,m}^{j=1,\ldots,d}$ and 2D density estimates $\widehat{p}_{jl}, j \neq l$ of $p_{jl}(\cdot)$ at the points $\{(v_{ij}^0, v_{rl}^0)\}_{i,r=1,\ldots,m}^{j,l=1,\ldots,d}$ by the formulas (12), (18), and (13).

Step 2: Initiate the estimates: set $\widetilde{f}_j^{(0)} = \widehat{f}_j, j = 1, \ldots, d$.

Step 3: Iterate for k: from $j = 1$ to d, successively calculate the estimates $f_j^{(k)}(\cdot)$ of $f_j(\cdot)$ at the points

$\{v^0_{ij}\}^{j=1,\ldots,d}_{i=1,\ldots,m}$ via (20), where the integrals are approximated by the numerical methods in terms of the values on the resulting grids.

Step 4: Stop if a preset ignorance criterion is satisfied; otherwise, continue to iterate as at Step 3.

Step 5: If one needs to calculate the estimated values of $f_j(\cdot)$ at the original observation points, then the interpolation technique can be used to achieve this with the help of the values on the grids produced at Step 4.

Theorem 2: Suppose that Assumptions 1 and 3 hold and the bandwidths $h_j \to 0, nh_j \to \infty$ as $n \to \infty$. Then, with probability tending to 1, the solution to (19) exists and is unique. Furthermore, there exist constants $0 < \gamma < 1$ and $\bar\xi > 0$ such that, with probability approaching to 1, the following inequality holds for all $j = 1, \ldots, d$:

$$\int \left(\widetilde{f}_j^{(k)}(v_j) - \widetilde{f}_j(v_j)\right)^2 p_j(v_j)\mathrm{d}v_j$$

$$\leq \bar\xi\gamma^{2k}\left(1 + \sum_{j=1}^{d}\int \left(\widetilde{f}_j^{(0)}(v_j)\right)^2 p_j(v_j)\mathrm{d}v_j\right)$$

where the functions $\widetilde{f}_1^{(0)}(v_1), \ldots, \widetilde{f}_d^{(0)}(v_d)$ are the initial values of the backfitting algorithm (20).

Suppose further that $n^{1/5}h_j \xrightarrow[n\to\infty]{} \psi_j$ for some constants $\psi_j > 0$. Then, the following convergence holds in distribution for any $v \in \mathsf{I}$:

$$n^{2/5}\begin{bmatrix}\widetilde{f}_1(v_1) - f_1(v_1)\\ \widetilde{f}_2(v_2) - f_2(v_2)\\ \vdots \\ \widetilde{f}_d(v_d) - f_d(v_d)\end{bmatrix} \xrightarrow[n\to\infty]{}$$

$$\mathcal{N}\left[\begin{bmatrix}\psi_1^2\beta_1(v_1)\\ \psi_2^2\beta_2(v_2)\\ \vdots\\ \psi_d^2\beta_d(v_d)\end{bmatrix}, \begin{bmatrix} w_1(v_1) & 0 & \cdots & 0\\ 0 & \ddots & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & w_d(v_d)\end{bmatrix}\right]$$

where $\beta_j(v_j), j = 1, \ldots, d$ on $\mathbb{R}$ with constraints $\int \beta_j(v_j)p_j(v_j)\mathrm{d}v_j = 0$ is the solution to

$$(\beta_0, \beta_1(\cdot), \ldots, \beta_d(\cdot))$$

$$= \arg\min_{\beta_0,\ldots,\beta_d}\int(\beta(v) - \beta_0 - \beta_1(v_1) - \cdots - \beta_d(v_d))^2\, p(v)\mathrm{d}v$$

with a constant β_0 and $\beta(v) = \sum_{j=1}^{d}(f_j'(v_j)\frac{\partial}{\partial v_j}\log p(v) + \frac{1}{2}f_j''(v_j))\int t^2K(t)\mathrm{d}t$, and where $w_j(v_j) = \frac{\sigma^2\int K^2(t)\mathrm{d}t}{\psi_j p_j(v_j)}$, $j = 1, \ldots, d$. Furthermore, for any $v \in \mathsf{I}$,

$$n^{2/5}\left(\widetilde{f}(v) - f(v)\right) \to \mathcal{N}\left(\sum_{j=1}^{d}\psi_j^2\beta_j(v_j), \sum_{j=1}^{d}w_j(v_j)\right),$$

where $\widetilde{f}(v) = \overline{Y} + \sum_{j=1}^{d}\widetilde{f}_j(v_j)$ and $f(v) = f_0 + \sum_{j=1}^{d}f_j(v_j)$.

Proof: See the Appendix. ■

We now discuss the smooth backfitting local linear estimator, which is obtained by projecting $\{y_k, k = 1, \ldots, n\}$ onto the space of additive functions similarly to the ideas of the LL estimator for 1D functions. This means that it is the minimizer of the criterion

$$\frac{1}{2n}\int\sum_{k=1}^{n}\left(y_k - \bar f_0 - \sum_{j=1}^{d}\bar f_j(v_j) - \sum_{j=1}^{d}\bar\theta_j(v_j)(v_j - x_{kj})\right)^2$$

$$\times\prod_{r=1}^{d}K_{h_r}(v_r - x_{kr})\mathrm{d}v, \tag{21}$$

where the minimization runs over all additive functions $\bar f(v) = \bar f_0 + \sum_{j=1}^{d}\bar f_j(v_j)$ with the constraints $\int \bar f_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j + \int\bar\theta_j(v_j)\widehat{p}_j^j(v_j)\mathrm{d}v_j = 0,\ j = 1, \ldots, d$. Here, the $-\bar\theta_j(\cdot)$ s can be regarded as the first-order derivative of $f_j(\cdot)$ s. This is similar to the idea that is used in the LL estimator for univariate nonparametric identification. Using the Lagrange multipliers, the constrained functional optimization (21) is transformed into the unconstrained functional optimization

$$\frac{1}{2n}\int\sum_{k=1}^{n}\left(y_k - \bar f_0 - \sum_{j=1}^{d}\bar f_j(v_j) - \sum_{j=1}^{d}\bar\theta_j(v_j)(v_j - x_{kj})\right)^2$$

$$\times\prod_{r=1}^{d}K_{h_r}(v_r - x_{kr})\mathrm{d}v + \sum_{j=1}^{d}\lambda_j\left(\int\bar f_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j\right.$$

$$\left. + \int\bar\theta_j(v_j)\widehat{p}_j^j(v_j)\mathrm{d}v_j\right). \tag{22}$$

Let $\{\widetilde{f}_0, \widetilde{f}_j(v_j), \widetilde{\theta}_j(v_j), j = 1, \ldots, d\}$ be the minimizer of (22). According to the method of variation, we find that $\{\widetilde{f}_0, \widetilde{f}_j(v_j), \widetilde{\theta}_j(v_j), j = 1, \ldots, d\}$ satisfy the system of equations

$$\widetilde{f}_0 = \frac{1}{n}\sum_{k=1}^{n}y_k - \sum_{j=1}^{d}\int\widetilde{f}_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j$$

$$-\sum_{j=1}^{d}\int\widetilde{\theta}_j(v_j)\widehat{p}_j^j(v_j)\mathrm{d}v_j,$$

$$\int\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \widetilde{f}_0 - \sum_{j=1}^{d}\widetilde{f}_j(v_j) - \sum_{j=1}^{d}\widetilde{\theta}_j(v_j)(v_j - x_{kj})\right)$$

$$\times\prod_{r=1}^{d}K_{h_r}(v_r - x_{kr})\mathrm{d}v_{-j} - \lambda_j\widehat{p}_j(v_j) = 0, \tag{23}$$

$$\int\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \widetilde{f}_0 - \sum_{j=1}^{d}\widetilde{f}_j(v_j) - \sum_{j=1}^{d}\widetilde{\theta}_j(v_j)(v_j - x_{kj})\right)$$

$$\times\prod_{r=1}^{d}K_{h_r}(v_r - x_{kr})\mathrm{d}v_{-j}(v_j - x_{kj}) - \lambda_j\widehat{p}_j^j(v_j) = 0. \tag{24}$$

Using the constraints $\int\widetilde{f}_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j + \int\widetilde{\theta}_j(v_j)\widehat{p}_j^j(v_j)\mathrm{d}v_j = 0,\ j = 1, \ldots, d$, results in $\widetilde{f}_0 = \frac{1}{n}\sum_{k=1}^{n}y_k \triangleq \overline{Y}$. Equation

(23) is simplified as

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{n}K_{h_j}(v_j-x_{kj})y_k-\widetilde{f}_0\widehat{p}_j(v_j)-\widetilde{f}_j(v_j)\widehat{p}_j(v_j)\\&\quad-\sum_{l\neq j}\int\widetilde{f}_l(v_l)\widehat{p}_{jl}(v_j,v_l)\mathrm{d}v_l-\widetilde{\theta}_j(v_j)\widehat{p}_j^{j}(v_j)\\&\quad-\sum_{l\neq j}\int\widetilde{\theta}_l(v_l)\widehat{p}_{jl}^{l}(v_j,v_l)\mathrm{d}v_l-\lambda_j\widehat{p}_j(v_j)=0.\end{aligned}$$

Moving the terms arrives at

$$\begin{aligned}A(v_j)&\triangleq\widehat{p}_j(v_j)\widetilde{f}_j(v_j)+\widehat{p}_j^{j}(v_j)\widetilde{\theta}_j(v_j)\\&=\frac{1}{n}\sum_{k=1}^{n}K_{h_j}(v_j-x_{kj})y_k-\sum_{l\neq j}\int\widetilde{f}_l(v_l)\widehat{p}_{jl}(v_j,v_l)\mathrm{d}v_l\\&-\sum_{l\neq j}\int\widetilde{\theta}_l(v_l)\widehat{p}_{jl}^{l}(v_j,v_l)\mathrm{d}v_l-\widehat{p}_j(v_j)(\widetilde{f}_0+\lambda_j).\end{aligned}$$

Applying the constraint $\int\widetilde{f}_j(v_j)\widehat{p}_j(v_j)\mathrm{d}v_j+\int\widetilde{\theta}_j(v_j)\widehat{p}_j^{j}(v_j)\mathrm{d}v_j=0$ obtains the Lagrange multipliers $\lambda_j=0$. This entails

$$\begin{aligned}A(v_j)&\triangleq\widehat{p}_j(v_j)\widetilde{f}_j(v_j)+\widehat{p}_j^{j}(v_j)\widetilde{\theta}_j(v_j)\\&=\frac{1}{n}\sum_{k=1}^{n}K_{h_j}(v_j-x_{kj})y_k-\sum_{l\neq j}\int\widetilde{f}_l(v_l)\widehat{p}_{jl}(v_j,v_l)\mathrm{d}v_l\\&-\sum_{l\neq j}\int\widetilde{\theta}_l(v_l)\widehat{p}_{jl}^{l}(v_j,v_l)\mathrm{d}v_l-\widehat{p}_j(v_j)\overline{Y}.\end{aligned}\tag{25}$$

Similarly, (24) is also simplified as

$$\begin{aligned}B(v_j)&\triangleq\widehat{p}_j^{j}(v_j)\widetilde{f}_j(v_j)+\widehat{p}_j^{jj}(v_j)\widetilde{\theta}_j(v_j)\\&=\frac{1}{n}\sum_{k=1}^{n}K_{h_j}(v_j-x_{kj})(v_j-x_{kj})y_k\\&\quad-\sum_{l\neq j}\int\widetilde{f}_l(v_l)\widehat{p}_{jl}^{j}(v_j,v_l)\mathrm{d}v_l\\&\quad-\sum_{l\neq j}\int\widetilde{\theta}_l(v_l)\widehat{p}_{jl}^{jl}(v_j,v_l)\mathrm{d}v_l-\widehat{p}_j^{j}(v_j)\overline{Y}.\end{aligned}\tag{26}$$

Set $C(v_j)\triangleq\widehat{p}_j(v_j)\widehat{p}_j^{jj}(v_j)-(\widehat{p}_j^{j}(v_j))^2$. Thus, we obtain

$$\widetilde{f}_j(v_j)=\big(\widehat{p}_j^{jj}(v_j)A(v_j)-\widehat{p}_j^{j}(v_j)B(v_j)\big)/C(v_j),\tag{27}$$

$$\widetilde{\theta}_j(v_j)=\big(-\widehat{p}_j^{j}(v_j)A(v_j)+\widehat{p}_j(v_j)B(v_j)\big)/C(v_j).\tag{28}$$

Similar to the smooth backfitting kernel estimator, the system of equations (25)–(28) is iteratively solved by the backfitting algorithm described as follows:

$$\begin{aligned}A^{(k)}(v_j)&=\frac{1}{n}\sum_{k=1}^{n}K_{h_j}(v_j-x_{kj})y_k\\&-\sum_{l<j}\int\widetilde{f}_l^{(k)}(v_l)\widehat{p}_{jl}(v_j,v_l)\mathrm{d}v_l-\sum_{l<j}\int\widetilde{\theta}_l^{(k)}(v_l)\widehat{p}_{jl}^{l}(v_j,v_l)\mathrm{d}v_l\\&-\sum_{l>j}\int\widetilde{f}_l^{(k-1)}(v_l)\widehat{p}_{jl}(v_j,v_l)\mathrm{d}v_l\\&-\sum_{l>j}\int\widetilde{\theta}_l^{(k-1)}(v_l)\widehat{p}_{jl}^{l}(v_j,v_l)\mathrm{d}v_l-\widehat{p}_j(v_j)\overline{Y},\end{aligned}\tag{29}$$

$$\begin{aligned}B^{(k)}(v_j)&=\frac{1}{n}\sum_{k=1}^{n}K_{h_j}(v_j-x_{kj})(v_j-x_{kj})y_k\\&-\sum_{l<j}\int\widetilde{f}_l^{(k)}(v_l)\widehat{p}_{jl}^{j}(v_j,v_l)\mathrm{d}v_l-\sum_{l<j}\int\widetilde{\theta}_l^{(k)}(v_l)\widehat{p}_{jl}^{jl}(v_j,v_l)\mathrm{d}v_l\\&-\sum_{l>j}\int\widetilde{f}_l^{(k-1)}(v_l)\widehat{p}_{jl}^{j}(v_j,v_l)\mathrm{d}v_l\\&-\sum_{l>j}\int\widetilde{\theta}_l^{(k-1)}(v_l)\widehat{p}_{jl}^{jl}(v_j,v_l)\mathrm{d}v_l-\widehat{p}_j^{j}(v_j)\overline{Y},\end{aligned}\tag{30}$$

$$\widetilde{f}_j^{(k)}(v_j)=\big(\widehat{p}_j^{jj}(v_j)A^{(k)}(v_j)-\widehat{p}_j^{j}(v_j)B^{(k)}(v_j)\big)/C(v_j),\tag{31}$$

$$\widetilde{\theta}_j^{(k)}(v_j)=\big(-\widehat{p}_j^{j}(v_j)A^{(k)}(v_j)+\widehat{p}_j(v_j)B^{(k)}(v_j)\big)/C(v_j)\tag{32}$$

where the initial values $\widetilde{f}_j^{(0)}(v_j),\widetilde{\theta}_j^{(0)}(v_j)$ can be chosen to be the 1D LL estimator of $\{y_k,k=1,\ldots,n\}$ onto $\{x_{kj},k=1,\ldots,n\}$.

Like the iterative algorithm for the smooth backfitting kernel estimator, the smooth backfitting LL estimator is also implemented on a fixed grid to reduce the computational complexity. Denote the preset grid of the jth variable by $v_j^0=[v_{1j}^0,\ldots,v_{mj}^0]^T$, where m is the number of the grid points.

The Smooth backfitting local linear estimator (SBLL)

Step 1: Use the observation points $\{y_k,x_{kj}\}_{k=1,\ldots,n}^{j=1,\ldots,d}$ and the kernel function $K(\cdot)$ to calculate the values of 1D density estimates $\widehat{p}_j$, $\widehat{p}_j^{j}$, $\widehat{p}_j^{jj}$ and LL estimates $\widehat{f}_j$, $\widehat{\theta}_j$ of $f_j(\cdot)$, $\theta_j(\cdot)$ at the points $\{v_{ij}^0\}_{i=1,\ldots,m}^{j=1,\ldots,d}$ and 2D density estimates $\widehat{p}_{jl}$, $\widehat{p}_{jl}^{l}$, $\widehat{p}_{jl}^{j}$, $\widehat{p}_{jl}^{jl}$, $j\neq l$ at the points $\{(v_{ij}^0,v_{rl}^0)\}_{i,r=1,\ldots,m}^{j,l=1,\ldots,d}$.

Step 2: Initiate the estimates: set $\widetilde{f}_j^{(0)}=\widehat{f}_j$, $\widetilde{\theta}_j^{(0)}=\widehat{\theta}_j$, $j=1,\ldots,d$.

Step 3: Iterate for k: from $j=1$ to d, successively calculate the estimates of $f_j(\cdot)$, $\theta_j(\cdot)$ at the points $\{v_{ij}^0\}_{i=1,\ldots,m}^{j=1,\ldots,d}$ according to (29)–(32), where the integrals are approximated by the numerical methods via the values on the resulting grids.

Step 4: Stop if a preset ignorance criterion is satisfied; otherwise, continue to iterate as at Step 3.

Step 5: If one needs to obtain the estimated values of $f_j(\cdot)$ s and their derivatives $\theta_j(\cdot)$ s at the original observation points, then the interpolation technique can be used to achieve this with the help of the values on the grids produced at Step 4.

Theorem 3: Suppose that Assumptions 1 and 3 hold and the bandwidths $h_j \to 0, nh_j \to \infty$ as $n \to \infty$. Then, with probability tending to 1, the solution to (25)–(28) exists and is unique. Furthermore, there exist constants $0 < \gamma < 1$ and $\bar{\xi} > 0$ such that, with probability approaching to 1, the following inequalities hold for any $j = 1, \ldots, d$:

$$\int \left(\widetilde{f}_j^{(k)}(v) - \widetilde{f}_j(v)\right)^2 p_j(v_j)\mathrm{d}v_j \le \bar{\xi}\gamma^{2k}\Gamma,$$

$$\int (\widetilde{\theta}_j^{(k)}(v) - \widetilde{\theta}_j(v))^2 p_j(v)\mathrm{d}v_j \le \bar{\xi}\gamma^{2k}\Gamma$$

where $\Gamma = 1 + \sum_{j=1}^{d} \int((\widetilde{f}_j^{(0)}(v_j))^2 + (\widetilde{\theta}_j^{(0)}(v_j))^2)p_j(v_j)\mathrm{d}v_j$ and the functions $\widetilde{f}_j^{(0)}(v_j), \widetilde{\theta}_j^{(0)}(v_j), j = 1, \ldots, d$ are the initial values of the smooth backfitting LL estimator (29)–(32).

Suppose further that $n^{1/5}h_j \to \psi_j$ for some constants $\psi_j > 0$. Then, the smooth backfitting LL estimator converges in distribution for any $v \in \mathsf{I}$

$$n^{2/5}\begin{bmatrix} \widetilde{f}_1(v_1) - f_1(v_1) + \nu_1 \\ \widetilde{f}_2(v_2) - f_2(v_2) + \nu_2 \\ \vdots \\ \widetilde{f}_d(v_d) - f_d(v_d) + \nu_d \end{bmatrix} \xrightarrow[n\to\infty]{}$$

$$\mathcal{N}\left[\begin{bmatrix} \psi_1^2\alpha_1(v_1) \\ \psi_2^2\alpha_2(v_2) \\ \cdots \\ \psi_d^2\alpha_d(v_d) \end{bmatrix}, \begin{bmatrix} w_1(v_1) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & w_d(v_d) \end{bmatrix}\right]$$

where

$$\nu_j = \int f_j(v_j)K_{h_j}(v_j - t)p_j(t)\mathrm{d}t\mathrm{d}v_j,$$

$$\alpha_j(v_j) = \frac{\int t^2 K(t)\mathrm{d}t}{2}\left(f_j''(v_j) - \int f_j''(v_j)p_j(v_j)\mathrm{d}v_j\right),$$

$$w_j(v_j) = \frac{\sigma^2 \int K(t)^2\mathrm{d}t}{\psi_j p_j(v_j)}.$$

Furthermore, for any $v \in \mathsf{I}$

$$n^{2/5}(\widetilde{f}(v) - f(v)) \to \mathcal{N}\left(\sum_{j=1}^{d}\psi_j^2\alpha_j(v_j), \sum_{j=1}^{d} w_j(v_j)\right)$$

where $\widetilde{f}(v) = \overline{Y} + \sum_{j=1}^{d}\widetilde{f}_j(v_j)$ and $f(v) = f_0 + \sum_{j=1}^{d} f_j(v_j)$.

Proof: The proof follows from the same steps as what presented in [36]. ■

IV. Variable Selection and Identification for the ANARX System

In this section, we extend the results on variable selection and nonparametric identification of the static additive nonlinear model (2) to the ANARX system (4). Recall $\phi_k = [y_k, \ldots, y_{k-s+1}, u_k, \ldots, u_{k-t+1}]^T$ and the function $f(\phi_k) = f_0 + f_1(y_k) + \cdots + f_s(y_{k-s+1}) + f_{s+1}(u_k) + \cdots + f_{s+t}(u_{k-t+1})$. The system (4) can be rewritten in a compact form as $y_{k+1} = f(\phi_k) + \varepsilon_{k+1}$ and further it has the vector form

$$Y = f_0\mathbf{1}_n + \sum_{j=1}^{d} f_j + \varepsilon. \tag{33}$$

Define $G(\phi_k) = \left[f(\phi_k), y_k, \ldots, y_{k-s+2}, 0, u_k, \ldots, u_{k-t+2}\right]^T$, $\eta_k = [\varepsilon_k, 0, \ldots, 0, u_k, 0, \ldots, 0]^T$. Thus, we get

$$\phi_{k+1} = G(\phi_k) + \eta_{k+1}. \tag{34}$$

The range of ϕ_{k+1} taking values in $\mathbb{R}^{s+t}$ is denoted by $\mathcal{H}$. Note that $\{\eta_{k+1}\}$ is a sequence of independent and identically distributed (i.i.d.) random vectors and independent of $\{\phi_j, j \le k\}$ under Assumption 4iii) given below. For any $A \in \mathcal{B}^{s+t}$, where $\mathcal{B}^{s+t}$ is the Borel σ-algebra on $\mathcal{H}$, one derives that

$$\begin{aligned} P(\phi_{k+1} \in A \mid \phi_k, \ldots, \phi_0) &= P(\phi_{k+1} \in A \mid \phi_k) \\ &= P(\phi_1 \in A \mid \phi_0). \end{aligned}$$

This means that $\{\phi_k\}$ is a time-homogeneous Markov chains and its k-step transition probability is defined by

$$P_k(\phi, A) = P(\phi_k \in A \mid \phi_0 = \phi), \ \forall\, A \in \mathcal{B}^{s+t},\ \phi \in \mathcal{H}.$$

A major difficulty is that Assumption 3 is easily satisfied for the static model (2) but not obviously for the ANARX system (4) because of its dynamics, mainly on the strong mixing condition of $\{\phi_k\}$. Providing the conditions on the system structure $f_j(\cdot)$, the input u_k, and the noise ε_k so as to guarantee the strongly mixing is very useful to practical applications. Therefore, the conditions on the system structure $f_j(\cdot)$, the input u_k, and the noise ε_k such that Assumption 3 holds are given in the following assumption, which are commonly used for identifying nonlinear systems based on kernel functions in the literature [18], [27].

Assumption 4:

i) The dynamic difference equation $y_k = f_0 + f_1(y_{k-1}) + \cdots + f_s(y_{k-s})$ has an exponentially stable equilibrium point, i.e., $\|y_k\| \le M_1\rho^k\|y_0\|$ for some $M_1 < \infty, 0 < \rho < 1$, and any $k \ge 0$. Also, the system (4) is locally controllable at the equilibrium point.

ii) The second-order derivatives of the functions $f_j(\cdot), j = 1, \ldots, d$ exist and are Lipschitz continuous.

iii) Both the input $\{u_k\}$ and the noise $\{\varepsilon_k\}$ are a sequence of i.i.d. random variables with compact support. Also, both of them have a density function on their resulting supports. Let $\sigma^2 = \mathrm{Var}(\varepsilon_k)$.

Theorem 4: Consider the ANARX system (4) under Assumptions 1 and 4. We have

1) The Markov chain $\{\phi_k\}$ is geometrically ergodic, i.e., there exist a constant $0 < \rho < 1$ and a unique invariant measure P_{IV} such that $\|P_k(\phi, \cdot) - P_{\mathrm{IV}}\|_{\mathrm{var}} \le \bar{g}(\phi)\rho^k$, where $\|\cdot\|_{\mathrm{var}}$ denotes the total variation norm and $\bar{g}(\phi)$ is integrable with respect to P_{IV}. As a result, the process $\{\phi_k\}$ is strictly stationary and strongly mixing with a geometric convergence rate, i.e., the mixing coefficients of $\{\phi_k\}$ satisfy $\alpha(k) = O(\rho^k)$.

2) Consider the SBKE given by (20). Then, the convergence results of Theorem 2 hold for the ANARX system (4).

3) Consider the SBLL given by (29)–(32). Then, the convergence results of Theorem 3 hold for the ANARX system (4).

4) Further with an additional Assumption 2, apply variable selection using the nonnegative garrote estimator (8) for the ANARX system (4). Then the results of Theorem 1 hold for the ANARX system if the SBKE or the SBLL is used as the initial consistent estimate in the nonnegative garrote estimator. That is to say, the nonnegative garrote estimator will correctly find the set of the nonzero functions with probability one as $n \to \infty$ for the ANARX system (4).

Proof: The first part follows from [37, Theorem A 1.6, page 458] via some corresponding modifications. The main steps of the idea consist of: (a) To show that the chain $\{\phi_k\}$ defined by (34) is μ-irreducible, aperiodic, and that any μ-non null compact set is small, where μ is the Lebesgue measure on $(\mathcal{H}, \mathcal{B}^d)$. (b) To show that the chain $\{\phi_k\}$ is geometrically ergodic by using the drift criterion for geometric ergodicity. This implies that Assumption 3 v)-vi) are satisfied. The proofs for the rest parts of the theorem are parallel to the proofs of Theorems 2 and 3, thus omitted. This completes the proof. ■

To allow a little abuse of notation, we still use the acronyms SBKE and SBLL to denote the whole variable selection procedure including the smooth backfitting estimator and the nonnegative garrote estimator. The variable selection algorithm for the ANARX system (4) can now be stated step by step as follows. Let us take the SBKE as an example. The implementation of the SBLL is similar.

Step 1: The smooth backfitting kernel estimator (SBKE)

1) Collect the data $\{y_k, u_k, k = 1, \ldots, n\}$, and generate the design matrix X with its (i, j)-element $x_{kj} (k = 1, \ldots, n, j = 1, \ldots, d)$ defined as

$$x_{kj} = \begin{cases} y_{k-j}, & j = 1, \ldots, s, \\ u_{k-j+s}, & j = s+1, \ldots, d. \end{cases}$$

2) Preset d 1D grids by the range of each column of X, where the jth grid is denoted by $v_j^0 = [v_{1j}^0, \ldots, v_{mj}^0]^T$ and m is the number of 1D grid points.
3) Use the data $\{y_k, x_{kj}\}_{k=1,\ldots,n}^{j=1,\ldots,d}$ and the selected kernel function $K(\cdot)$ to calculate the values of 1D density estimates $\widehat{p}_j$ of $p_j(\cdot)$ and kernel estimates $\widehat{f}_j$ of $f_j(\cdot)$ at the points $\{v_{ij}^0\}_{i=1,\ldots,m}^{j=1,\ldots,d}$ and 2D density estimates $\widehat{p}_{jl}, j \neq l$ of $p_{jl}(\cdot)$ at the points $\{(v_{ij}^0, v_{rl}^0)\}_{i,r=1,\ldots,m}^{j,l=1,\ldots,d}$ by the formulas (12), (18), and (13).
4) Initiate the estimates: set $\widetilde{f}_j^{(0)} = \widehat{f}_j, j = 1, \ldots, d$.
5) Iterate for k: from $j = 1$ to d, successively calculate the estimates $f_j^{(k)}(\cdot)$ of $f_j(\cdot)$ at the points $\{v_{ij}^0\}_{i=1,\ldots,m}^{j=1,\ldots,d}$ via (20), where the integrals can be calculated by the function `trapz` in Matlab.
6) Stop if a preset ignorance criterion is satisfied; otherwise, continue to iterate as at 5), where the ignorance criterion given in [38] is used for the simulation in Section V. That is, if for all $j = 1, \ldots, d$,

$$\frac{\sum_{i=1}^{m} \left(\widetilde{f}^{(k+1)}(v_{ij}^0) - \widetilde{f}^{(k)}(v_{ij}^0)^2 \right)^2}{\sum_{i=1}^{m} \widetilde{f}^{(k)}(v_{ij}^0)^2 + 0.0001} < 0.0001,$$

then stop.

7) Use the interpolation technique to calculate the estimated values of $f_j(\cdot)$ at the original observation points by the values on the grids produced at 6).

Step 2: The nonnegative garrote estimator

1) Solve the optimization problem

$$\min_c \frac{1}{2} \left\| Y - \sum_{j=1}^{d} c_j \widehat{f}_j \right\|^2 \text{ s.t. } \sum_{j=1}^{d} c_j \leq \kappa_n \tag{35}$$

for a given tuning parameter κ_n and the consistent estimate $\widehat{f}_j, j = 1, \ldots, d$ obtained at Step 1 to get the solution $\widehat{c} = [\widehat{c}_1, \ldots, \widehat{c}_d]^T$. The choice of κ_n will be given below.

2) The functions with the indices such that $\widehat{c}_j > 0$ are taken as the nonzero functions, otherwise as the zero functions.

Note that the optimization problem (35) at Step 2 1), which is equivalent to the original problem (8) in the paper, is a constrained linear least squares problem. Thus, this optimization problem can be effectively solved by the numerical algorithm, for example, the function `lsqlin` in Matlab. So the choice of λ_n is transformed to that of κ_n, which controls the size of the selected coefficients. In the following, a data-driven choice for κ_n is provided. To be consistent and compare with the method in [30], here the L-curve criterion is used to implement the automatic selection of κ_n. The L-curve criterion is based on the so-called L-curve, which is a parametric plot of $(\zeta(\kappa_n), \omega(\kappa_n))$, where $\zeta(\kappa_n)$ and $\omega(\kappa_n)$ measure, respectively, the size of the regularized solution and the corresponding residual [39]. The L-curve has a distinct L-shaped corner located exactly where the solution $\widehat{c}$ changes from being dominated by the regularization error to being dominated by the noise.

These explanations mean that the "optimal" tuning parameter corresponds to the "corner" (maximum curvature) of the L-curve. In the following, an algorithm for finding the "corner" is given.

Step 1: Generate a gird $\kappa = \{\kappa_n^l\}_{l=1}^{Q}$ in an ascending order in the interval $[1, d]$ for κ_n.

Step 2: Calculate $\zeta(\kappa) = \sum_{j=1}^{d} \widehat{c}_j, \ \omega(\kappa) = \frac{1}{n} \|Y - \widehat{Y}\|_2^2$ on the grid κ, where $\widehat{Y} = \sum_{j=1}^{d} \widehat{c}_j \widehat{f}_j$ is the prediction for Y and $\widehat{c}$ is the solution of (35).

Step 3: Calculate

$$A_l = \arctan\left(\frac{\omega(\kappa_n^l) - \omega(\kappa_n^{l-1})}{\zeta(\kappa_n^l) - \zeta(\kappa_n^{l-1})} \right), \ l = 2, \ldots, Q.$$

Step 4: Calculate $D_l = A_l - A_{l-1}, \ l = 3, \ldots, Q$.

Step 5: The "optimal" tuning parameter $\widehat{\kappa}_n$ is defined by

$$\widehat{\kappa}_n = \kappa_n^{\widehat{l}}, \text{ where } \widehat{l} = \arg \max_{l=3,\ldots,Q} D_l - 1.$$

For each point except two endpoints on the plot $(\zeta(\kappa), \omega(\kappa))$, one can obtain two associated angles by connecting the adjacent two points and extending to the horizontal axis. The difference between the two angles at a point can be regarded as a measure of the resulting curvature of the L-curve, and hence the point with the maximum difference can be thought of as the point of the maximum curvature of the L-curve.

TABLE II
TOTAL APPEARANCE FREQUENCY OF VARIABLE SELECTIONS FOR EXAMPLE 1

Independent variables ($q = 0$)										
Method	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
$n = 100$										
SBKE	100	0	0	100	0	99	0	0	100	0
SBLL	100	0	0	100	0	100	0	0	100	0
NGPS	100	0	0	100	0	98	0	0	100	0
$n = 200$										
SBKE	100	0	0	100	0	100	0	0	100	0
SBLL	100	0	0	100	0	100	0	0	100	0
NGPS	100	0	0	100	0	100	0	0	100	0
Dependent variables with correlation 0.5 ($q = 1$)										
Method	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
$n = 100$										
SBKE	98	0	0	100	0	98	0	0	100	0
SBLL	99	0	0	100	0	99	0	0	100	0
NGPS	97	1	0	100	1	99	0	0	100	0
$n = 200$										
SBKE	100	0	0	100	0	100	0	0	100	0
SBLL	100	0	0	100	0	100	0	0	100	0
NGPS	99	0	0	100	0	100	0	0	100	0

TABLE III
STATISTICAL ANALYSIS OF VARIABLE SELECTIONS FOR EXAMPLE 1

Independent variables ($q = 0$)							
Method	NS	NCS	NIS	PIN	PCS	GoF	True GoF
$n = 100$							
SBKE	3.99	3.99	0	99%	99%	0.65	0.58
	(0.10)	(0.10)	(0)			(0.03)	(0.03)
SBLL	4	4	0	100%	100%	0.67	0.58
	(0)	(0)	(0)			(0.03)	(0.03)
NGPS	3.98	3.98	0	98%	98%	0.63	0.58
	(0.14)	(0.14)	(0)			(0.03)	(0.03)
$n = 200$							
SBKE	4	4	0	100%	100%	0.61	0.57
	(0)	(0)	(0)			(0.02)	(0.02)
SBLL	4	4	0	100%	100%	0.62	0.57
	(0)	(0)	(0)			(0.02)	(0.02)
NGPS	4	4	0	100%	100%	0.83	0.57
	(0)	(0)	(0)			(0.02)	(0.02)
Dependent variables with correlation 0.5 ($q = 1$)							
Method	NS	NCS	NIS	PIN	PCS	GoF	True GoF
$n = 100$							
SBKE	3.96	3.96	0	96%	96%	0.55	0.47
	(0.20)	(0.20)	(0)			(0.04)	(0.04)
SBLL	3.98	3.98	0	98%	98%	0.57	0.47
	(0.14)	(0.14)	(0)			(0.04)	(0.04)
NGPS	3.98	3.96	0.02	96%	94%	0.53	0.47
	(0.25)	(0.20)	(0.14)			(0.04)	(0.04)
$n = 200$							
SBKE	4	4	0	100%	100%	0.51	0.47
	(0)	(0)	(0)			(0.01)	(0.03)
SBLL	4	4	0	100%	100%	0.52	0.47
	(0)	(0)	(0)			(0.01)	(0.03)
NGPS	3.99	3.99	0	99%	99%	0.50	0.47
	(0.10)	(0.10)	(0)			(0.01)	(0.03)

V. SIMULATION EXAMPLES

For comparison, the proposed SBKE and SBLL variable selection approaches and the variable selection method based on P-splines in [30], denoted by NGPS, are all applied to the following examples.

Example 1: Consider an additive nonlinear model

$$y_k = \sum_{j=1}^{10} f_j(x_{kj}) + \varepsilon_k,\ k = 1, \ldots, n,$$

where $f_1(x) = 3x,\ f_4(x) = 4x^2,\ f_6(x) = 1.5\sin(2\pi x)/(2 - \sin(2\pi x)),\ f_9(x) = 2\cos(2\pi x)$, and $f_j(x) = 0$ for other six additive functions. So the number of nonzero functions is four. The variables x_{kj} are generated by $x_{kj} = \frac{W_{kj}+qU_k}{1+q},\ j = 1, \ldots, 10$, where W_{kj} and U_k are independently generated from the uniform $[0, 1]$ and the number q plays a role in controlling the correlation among variables since the correlation coefficients of x_{ki} and x_{kj} are equal to $\mathrm{Corr}(x_{ki}, x_{kj}) = q^2/(1 + q^2)$ for any $1 \le i \ne j \le 10$. The variables are mutually independent if $q = 0$, while the case for $q = 1$ leads to the variables with correlation 0.5. The observation noise $\{\varepsilon_k\}$ is a sequence of i.i.d. zero-mean Gaussian random variables with variance 1, and the resulting signal-to-noise ratios (SNR) are 6.61 dB for the independent case and 4.13 dB for the dependence case. The sample sizes are taken as $n = 100$ and $n = 200$ data points, respectively. The sample size $n = 100$ or 200 is very small for estimating a 10-dimensional nonparametric nonlinear model. The results presented below are based on 100 Monte-Carlo tests.

Table II shows the total appearance frequency of variable selection by using the SBKE, SBLL, and NGPS, while the corresponding statistical analysis is summarized in Table III. They are the average of the number of the selected variables (NS), the average of the number of the correctly selected variables (NCS), the average of the number of incorrectly selected variables (NIS), the percentage of tests where all the correct variables are contained in the selected variables (PIN), the percentage of tests where the selected variables are exactly the correct variables (PCS), the average of goodness-of-fits (GoF), and the average of true GoF, respectively. The GoF is defined by $\mathrm{GoF} = 1 - \sqrt{\sum_{k=1}^n (y_k - \widehat{y}_k)^2 / \sum_{k=1}^n (y_k - \overline{Y})^2}$, where n is the sample size, $\overline{Y} = \frac{1}{n}\sum_{k=1}^n y_k, \widehat{y}_k = \sum_{j\in\widehat{\mathcal{I}}} \widehat{f}_j(x_{kj}), \widehat{\mathcal{I}}$ is the set of nonzero functions identified via the nonnegative garrote estimator, and $\widehat{f}_j$ is the estimate of f_j obtained by rerunning the nonparametric identification methods given in Section III for only the selected variables after finishing the variable selection process by the nonnegative garrote estimator. Similarly, the average of true GoF defined by $1 - \sqrt{\sum_{k=1}^n \varepsilon_k^2 / \sum_{k=1}^n (y_k - \overline{Y})^2}$. The values in the parentheses are the resulting standard errors.

The simulation results indicate that in the independent case the percentages of the SBKE, SBLL, and NGPS finding the correct variables are 99%, 100%, and 98% for the sample sizes 100, respectively, and all the approaches identify the correct variables 100% when $n = 200$. In the dependent case, the SBKE, SBLL, and NGPS that find the correct variables are 96%, 98%, and 94% for $n = 100$, respectively. When $n = 200$, the resulting percentages increase to 100%, 100%, and 99%, respectively. This example shows that the SBKE and SBLL outperform the NGPS for a small sample size.

Example 2: Consider an ANARX system

$$y_k = \sum_{j=1}^{5} f_j(y_{k-j}) + \sum_{l=1}^{5} f_{5+l}(u_{k-l}) + \varepsilon_k, \qquad (36)$$

where $f_2(x) = -1.5\sin(0.8x),\quad f_3(x) = -4\exp(-0.1x^2) + 2.5,\quad f_7(x) = 2x^3 - 1,\quad f_9(x) = 3\sin(2\pi x)/(2 - \sin(2\pi x)),$

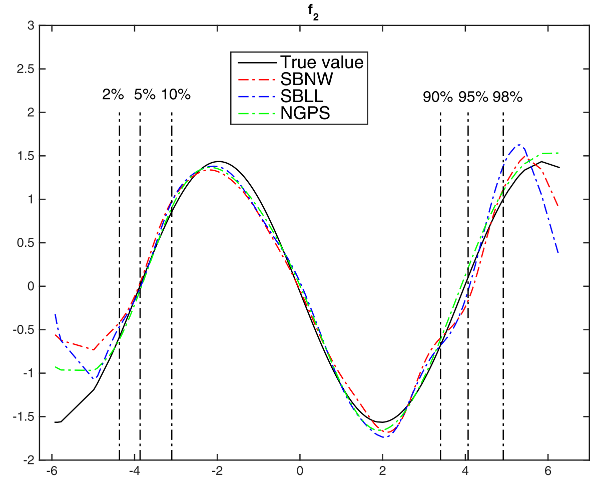


Fig. 1. Nonparametric estimation of $f_2(\cdot)$ for Example 2.

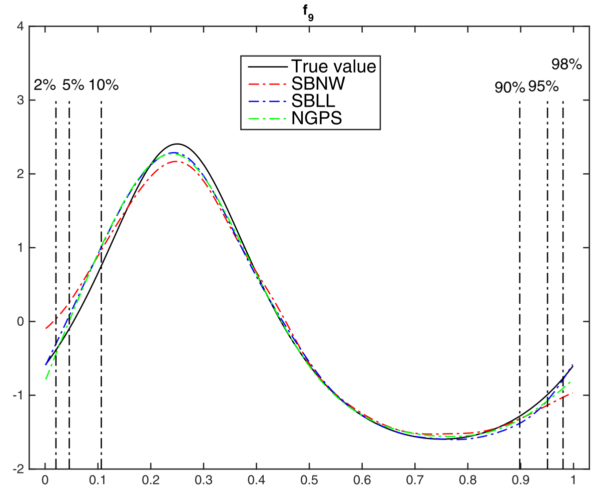


Fig. 2. Nonparametric estimation of $f_9(\cdot)$ for Example 2.

while the other additive nonlinear functions are zero. The input u_k is a sequence of i.i.d. uniform random variables over $[0, 1]$, and the noise $\{\varepsilon_k\}$ is a sequence of i.i.d. uniform random variables over $[-1.73, 1.73]$. Under this setting, the resulting SNR is 6.83 dB. The sample size is $n = 500$, which is a relatively small sample size for estimating a 10-dimensional nonlinear dynamic system. The results presented below are based on 100 Monte-Carlo tests.

To illustrate the performance of the nonparametric approaches SBKE, SBLL, and NGPS, the identification results for the two nonzero functions $f_2(\cdot)$ and $f_9(\cdot)$ based on a realization of (36) are plotted in Figs. 1 and 2. Note that the estimates in Figs. 1 and 2 have been subtracted with their resulting means to make them have zero means. Since nonparametric approaches have a boundary effect due to less data at the boundary, which is a common problem and cannot be easily avoided, we plot the positions of the 2%, 5%, 10%, 90%, 95%, and 98% quantiles of the domain of each function by the black dashed line in Figs. 1 and 2 to illustrate this. The number of the observation points corresponding to these quantiles are 10, 25, and 50, respectively. These amounts of samples are not sufficient to obtain a reliable estimate for a 1D function according to the relationship in Table I. Indeed, it is seen that the SBKE, SBLL, and NGPS estimates at the boundary deviate from the true values. On the other hand, these estimates perform well and have no obvious differences in the main region of the domain, for example, from the 10% quantile to 90% quantile, where some small fluctuations are because these estimates are based on one random realization of (36) and the sample size is small. Note that the identification results for the other two nonzero functions $f_3(\cdot)$ and $f_7(\cdot)$ are also similar, but are not shown here due to limited space.

TABLE IV
TOTAL APPEARANCE FREQUENCY OF VARIABLE SELECTIONS FOR EXAMPLE 2

Method	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
SBKE	0	100	100	0	0	0	99	0	100	0
SBLL	0	100	100	0	0	0	99	0	100	0
NGPS	0	100	100	0	0	0	91	0	100	0

TABLE V
STATISTICAL ANALYSIS OF VARIABLES SELECTIONS FOR EXAMPLE 2

Method	NS	NCS	NIS	PIN	PCS	GoF	True GoF
SBKE	3.99	3.99	0	99%	99%	0.61	0.59
	(0.10)	(0.10)	(0)			(0.02)	(0.02)
SBLL	3.99	3.99	0	99%	99%	0.61	0.59
	(0.10)	(0.10)	(0)			(0.02)	(0.02)
NGPS	3.91	3.91	0	91%	91%	0.55	0.59
	(0.29)	(0.29)	(0)			(0.02)	(0.02)

Table IV displays the total appearance frequency of variable selection by using the SBKE, SBLL, and NGPS, and the corresponding statistical analysis is outlined in Table V with its entries defined as the same as in Table III. The simulation results reveal that the SBKE, SBLL, and NGPS identify the correct variables with the percentage 99%, 99%, and 91%, respectively. This shows that the boundary effect of nonparametric estimates at Step 1 will not greatly influence the subsequent variable selection procedure, which is the final goal of the paper, since the estimates for the other zero functions are very close to zero.

In summary, the two-step variable selection methods (the SBKE and SBLL) proposed in the paper perform well in a small sample size and better than the NGPS.

VI. CONCLUSION

In this paper we have investigated the variable selection problem for high-dimensional dynamic additive nonlinear systems. This is the first time that such problem is tackled by nonparametric kernel approaches since the existing methods are mainly based on spline approximations. Our proposed methods are implemented by two subsequent steps: nonparametric identification of all the additive functions, followed by nonnegative garrote estimation. In the stage of nonparametric identification based on kernel functions, the algorithms including SBKE and SBLL have been provided and both of them do not suffer from the curse of dimensionality since only 1D and 2D kernel estimations are involved. Further, they achieve the convergence and asymptotic normality of nonparametric identification under weak conditions, and especially, the SBLL can ensure that the estimate for each additive function acquires the same asymptotic properties as if other functions are exactly known. We have proved that the nonnegative garrote estimator can convert a consistent nonparametric estimate for the additive functions into a consistent estimate for the set of the nonzero functions. Therefore, the variable selection methods provided in this paper can find the correct nonzero functions with probability one under weak conditions as the sample size approaches infinity.

Appendix
Main Theoretical Proofs

Proof of Theorem 1: According to the convex optimization theory with constraints [40], the solution $\widehat{c}_j,\ j=1,\dots,d$ to the constrained optimization (8) satisfies the Karush-Kuhn-Tucker (KKT) conditions

$$-\widehat{f}^T(Y-\widehat{f}\widehat{c})+\lambda_n\mathbf{1}_d+u=0, \tag{37}$$

$$u_j\widehat{c}_j=0 \text{ for all } j=1,\dots,d, \tag{38}$$

$$-\widehat{c}_j\le 0 \text{ for all } j=1,\dots,d, \tag{39}$$

$$u_j\ge 0 \text{ for all } j=1,\dots,d, \tag{40}$$

where $\widehat{f}=[\widehat{f}_1,\dots,\widehat{f}_d]$, $\widehat{c}=[\widehat{c}_1,\dots,\widehat{c}_d]^T$, and $u=[u_1,\dots,u_d]^T$. Divide the index set $\{1,\dots,d\}$ into the following subsets in terms of the sets $\widehat{c}$ and $\mathcal{I}$:

$$\begin{aligned}
\Phi_{01}&=\{j:\widehat{c}_j=0, f_j(\cdot)\ne 0\},\\
\Phi_{00}&=\{j:\widehat{c}_j=0, f_j(\cdot)=0\},\\
\Phi_{11}&=\{j:\widehat{c}_j>0, f_j(\cdot)\ne 0\},\\
\Phi_{10}&=\{j:\widehat{c}_j>0, f_j(\cdot)=0\}.
\end{aligned}$$

It is clear that $\Phi_{00}\cup\Phi_{10}=\mathcal{I}^c$ and $\Phi_{01}\cup\Phi_{11}=\mathcal{I}$. For convenience, we denote $\widehat{\mathcal{I}}=\Phi_{11}\cup\Phi_{10}=\{j:\widehat{c}_j>0\}$ and $\widehat{\mathcal{I}}^c=\Phi_{01}\cup\Phi_{00}=\{1,\dots,d\}\setminus\widehat{\mathcal{I}}=\{j:\widehat{c}_j=0\}$.

First, we prove that $P(|\Phi_{10}|>0)\to 0$ as $n\to\infty$, where $|\Phi_{10}|$ denotes the cardinality of the set Φ_{10}. The formula (37) can be rewritten as

$$\lambda_n\begin{bmatrix}\mathbf{1}_{|\widehat{\mathcal{I}}|}\\ \mathbf{1}_{|\widehat{\mathcal{I}}^c|}\end{bmatrix}+\begin{bmatrix}u_{\widehat{\mathcal{I}}}\\ u_{\widehat{\mathcal{I}}^c}\end{bmatrix}+\begin{bmatrix}\widehat{f}_{\widehat{\mathcal{I}}}^T\widehat{f}_{\widehat{\mathcal{I}}} & \widehat{f}_{\widehat{\mathcal{I}}}^T\widehat{f}_{\widehat{\mathcal{I}}^c}\\ \widehat{f}_{\widehat{\mathcal{I}}^c}^T\widehat{f}_{\widehat{\mathcal{I}}} & \widehat{f}_{\widehat{\mathcal{I}}^c}^T\widehat{f}_{\widehat{\mathcal{I}}^c}\end{bmatrix}\begin{bmatrix}\widehat{c}_{\widehat{\mathcal{I}}}\\ \widehat{c}_{\widehat{\mathcal{I}}^c}\end{bmatrix}=\begin{bmatrix}\widehat{f}_{\widehat{\mathcal{I}}}^TY\\ \widehat{f}_{\widehat{\mathcal{I}}^c}^TY\end{bmatrix}.$$

By (38)–(40), it is clear that $\widehat{c}_j>0, u_j=0$ for $j\in\widehat{\mathcal{I}}$ and $\widehat{c}_j=0$ for $j\in\widehat{\mathcal{I}}^c$. It follows that

$$\lambda_n\begin{bmatrix}\mathbf{1}_{|\widehat{\mathcal{I}}|}\\ \mathbf{1}_{|\widehat{\mathcal{I}}^c|}\end{bmatrix}+\begin{bmatrix}0\\ u_{\widehat{\mathcal{I}}^c}\end{bmatrix}+\begin{bmatrix}\widehat{f}_{\widehat{\mathcal{I}}}^T\widehat{f}_{\widehat{\mathcal{I}}} & \widehat{f}_{\widehat{\mathcal{I}}}^T\widehat{f}_{\widehat{\mathcal{I}}^c}\\ \widehat{f}_{\widehat{\mathcal{I}}^c}^T\widehat{f}_{\widehat{\mathcal{I}}} & \widehat{f}_{\widehat{\mathcal{I}}^c}^T\widehat{f}_{\widehat{\mathcal{I}}^c}\end{bmatrix}\begin{bmatrix}\widehat{c}_{\widehat{\mathcal{I}}}\\ 0\end{bmatrix}=\begin{bmatrix}\widehat{f}_{\widehat{\mathcal{I}}}^TY\\ \widehat{f}_{\widehat{\mathcal{I}}^c}^TY\end{bmatrix}.$$

Taking the first $|\widehat{\mathcal{I}}|$ rows leads to $\lambda_n\mathbf{1}_{|\widehat{\mathcal{I}}|}+\widehat{f}_{\widehat{\mathcal{I}}}^T\widehat{f}_{\widehat{\mathcal{I}}}\widehat{c}_{\widehat{\mathcal{I}}}=\widehat{f}_{\widehat{\mathcal{I}}}^TY$. This indicates that

$$\begin{aligned}
\widehat{c}_{\widehat{\mathcal{I}}}=\begin{bmatrix}\widehat{c}_{\Phi_{11}}\\ \widehat{c}_{\Phi_{10}}\end{bmatrix}&=\begin{bmatrix}\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}/n & \widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{10}}/n\\ \widehat{f}_{\Phi_{10}}^T\widehat{f}_{\Phi_{11}}/n & \widehat{f}_{\Phi_{10}}^T\widehat{f}_{\Phi_{10}}/n\end{bmatrix}^{-1}\\
&\quad\times\begin{bmatrix}\widehat{f}_{\Phi_{11}}^TY/n-\lambda_n\mathbf{1}_{m_{11}}/n\\ \widehat{f}_{\Phi_{10}}^TY/n-\lambda_n\mathbf{1}_{m_{10}}/n\end{bmatrix}.
\end{aligned}$$

Denote

$$\begin{aligned}
F&=\widehat{f}_{\widehat{\mathcal{I}}}^T\widehat{f}_{\widehat{\mathcal{I}}}/n,\ F_{ij}=\widehat{f}_{\Phi_{1i}}^T\widehat{f}_{\Phi_{1j}}/n \text{ for } i,j=0,1,\\
\widetilde{F}&=F_{00}-F_{01}F_{11}^{-1}F_{10}\\
&=\widehat{f}_{\Phi_{10}}^T(I-\widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)\widehat{f}_{\Phi_{10}}/n.
\end{aligned}$$

By noting that $(I-\widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)^T(I-\widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)=(I-\widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)$, we have that $\widetilde{F}$ is a positive semi-definite matrix. Thus

$$F^{-1}=\begin{bmatrix} * & * \\ -\widetilde{F}^{-1}F_{01}F_{11}^{-1} & \widetilde{F}^{-1}\end{bmatrix}$$

by the inverse formula of a partitioned matrix (see [1, p. 359]). This means that

$$\begin{aligned}
\widehat{c}_{\Phi_{10}}&=\widetilde{F}^{-1}\big(\widehat{f}_{\Phi_{10}}^TY/n-\lambda_n\mathbf{1}_{m_{10}}/n-F_{01}F_{11}^{-1}\widehat{f}_{\Phi_{11}}^TY/n\\
&\quad+\lambda_nF_{01}F_{11}^{-1}\mathbf{1}_{m_{11}}/n\big)=\widetilde{F}^{-1}\xi,
\end{aligned}\tag{41}$$

where

$$\begin{aligned}
\xi&=\widehat{f}_{\Phi_{10}}^TY/n-\lambda_n\mathbf{1}_{m_{10}}/n-F_{01}F_{11}^{-1}\widehat{f}_{\Phi_{11}}^TY/n\\
&\quad+\lambda_nF_{01}F_{11}^{-1}\mathbf{1}_{m_{11}}/n\\
&=\widehat{f}_{\Phi_{10}}^T\big(I-\widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T\big)Y/n-\lambda_n\mathbf{1}_{m_{10}}/n\\
&\quad+\lambda_nF_{01}F_{11}^{-1}\mathbf{1}_{m_{11}}/n.
\end{aligned}$$

Since $\widehat{f}_j$ is a consistent estimate for f_j, i.e., $\|\widehat{f}_j-f_j\|^2/n=O_P(\delta_n^2)$, $j=1,\dots,d$, we have

$$\frac{1}{\sqrt{n}}\|\widehat{f}_{\Phi_{ij}}-f_{\Phi_{ij}}\|=O_P(\delta_n) \text{ for } i,j=0,1.$$

Therefore, under Assumption 2, we obtain

$$\begin{aligned}
&\left\|\frac{1}{\sqrt{n}}\widehat{f}_{\Phi_{11}}-\frac{1}{\sqrt{n}}f_{\Phi_{11}}\right\|=O_P(\delta_n),\\
&\left\|\frac{1}{n}\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}-\frac{1}{n}f_{\Phi_{11}}^Tf_{\Phi_{11}}\right\|\\
&\le\frac{1}{n}\|\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}-\widehat{f}_{\Phi_{11}}^Tf_{\Phi_{11}}\Big\|+\frac{1}{n}\Big\|\widehat{f}_{\Phi_{11}}^Tf_{\Phi_{11}}-f_{\Phi_{11}}^Tf_{\Phi_{11}}\|\\
&\le\frac{1}{\sqrt{n}}\left\|\widehat{f}_{\Phi_{11}}\right\|\frac{1}{\sqrt{n}}\left\|\widehat{f}_{\Phi_{11}}-f_{\Phi_{11}}\right\|\\
&\quad+\frac{1}{\sqrt{n}}\left\|\widehat{f}_{\Phi_{11}}-f_{\Phi_{11}}\right\|\frac{1}{\sqrt{n}}\|f_{\Phi_{11}}\|\\
&=\left(\frac{1}{\sqrt{n}}\|f_{\Phi_{11}}\|+O_P(\delta_n)\right)O_P(\delta_n)+O_P(\delta_n)\\
&=O_P(\delta_n).
\end{aligned}\tag{42}$$

This entails

$$\begin{aligned}
\frac{1}{\sqrt{n}}\widehat{f}_{\Phi_{11}}&=\frac{1}{\sqrt{n}}f_{\Phi_{11}}+O_P(\delta_n),\\
\frac{1}{n}\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}&=\frac{1}{n}f_{\Phi_{11}}^Tf_{\Phi_{11}}+O_P(\delta_n).
\end{aligned}$$

Noting that $\|(f_{\mathcal{I}}^T f_{\mathcal{I}}/n)^{-1}\| < \infty$, we get

$$\begin{aligned}\left(\frac{1}{n}\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}\right)^{-1} &= \left(\frac{1}{n}f_{\Phi_{11}}^T f_{\Phi_{11}} + O_P(\delta_n)\right)^{-1}\\ &= \left(\frac{1}{n}f_{\Phi_{11}}^T f_{\Phi_{11}}\left(I + O_P(\delta_n)\right)\right)^{-1}\\ &= \left(\frac{1}{n}f_{\Phi_{11}}^T f_{\Phi_{11}}\right)^{-1}\left(I + O_P(\delta_n)\right)^{-1}\\ &= \left(\frac{1}{n}f_{\Phi_{11}}^T f_{\Phi_{11}}\right)^{-1}\left(I + O_P(\delta_n)\right).\end{aligned}$$

It follows that

$$\begin{aligned}\|F_{01}F_{11}^{-1}\| &\le \frac{1}{\sqrt{n}}\|\widehat{f}_{\Phi_{10}}\|\frac{1}{\sqrt{n}}\|\widehat{f}_{\Phi_{11}}\|\left\|\left(\frac{1}{n}\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}\right)^{-1}\right\|\\ &= O_P\left(\frac{1}{\sqrt{n}}\|\widehat{f}_{\Phi_{10}}\|\right) = O_P(\delta_n).\end{aligned}$$

Therefore, $\xi = \widehat{f}_{\Phi_{10}}^T(I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)Y/n - \lambda_n(1+O_P(\delta_n))\mathbf{1}_{m_{10}}/n$. As mentioned above, we have

$$\begin{aligned}&(I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)^T(I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)\\ &= I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T.\end{aligned}$$

This means that its eigenvalues are either 1 or 0, and hence we have $\|I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T\| \le 1$. It follows that $\|(I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)Y\| \le \|I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T\|\|Y\| = O_P(\sqrt{n})$, which derives that

$$\begin{aligned}&\|\widehat{f}_{\Phi_{10}}^T(I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)Y\|\\ \le\ &\|\widehat{f}_{\Phi_{10}}\|\|I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)Y\|\\ =\ &O_P(\sqrt{n}\delta_n)O_P(\sqrt{n}) = O_P(n\delta_n).\end{aligned}$$

This leads to $\xi = O_P(\delta_n) - (1+O_P(\delta_n))\lambda_n/n\mathbf{1}_{m_{10}} = -\lambda_n/n\mathbf{1}_{m_{10}} < 0$ due to $\lambda_n/n \to 0$ and $\delta_n = o(\lambda_n/n)$. Since $\widehat{c}_j > 0$ for any $j \in \Phi_{10}$, we have $\xi^T\widehat{c}_{\Phi_{10}} < 0$. However, this violates (41), which implies that $\xi^T\widehat{c}_{\Phi_{10}} = \xi^T\widetilde{F}^{-1}\xi \ge 0$. Thus, we have $P(|\Phi_{10}| > 0) \to 0$ as $n \to \infty$. This means that Φ_{10} is a null set and hence $\mathcal{I}^c = \Phi_{00} \cup \Phi_{10} = \Phi_{00}$ in the asymptotic sense.

Next, we show that $P(|\Phi_{01}| > 0) \to 0$ as $n \to \infty$. Otherwise, assume that $|\Phi_{01}| > 0$. Similar to the derivation above, the formula (37) can also be rewritten as

$$\lambda_n\begin{bmatrix}\mathbf{1}_{|\mathcal{I}|}\\ \mathbf{1}_{|\mathcal{I}^c|}\end{bmatrix} + \begin{bmatrix}u_{\mathcal{I}}\\ u_{\mathcal{I}^c}\end{bmatrix} + \begin{bmatrix}\widehat{f}_{\mathcal{I}}^T\widehat{f}_{\mathcal{I}} & \widehat{f}_{\mathcal{I}}^T\widehat{f}_{\mathcal{I}^c}\\ \widehat{f}_{\mathcal{I}^c}^T\widehat{f}_{\mathcal{I}} & \widehat{f}_{\mathcal{I}^c}^T\widehat{f}_{\mathcal{I}^c}\end{bmatrix}\begin{bmatrix}\widehat{c}_{\mathcal{I}}\\ \widehat{c}_{\mathcal{I}^c}\end{bmatrix} = \begin{bmatrix}\widehat{f}_{\mathcal{I}}^T Y\\ \widehat{f}_{\mathcal{I}^c}^T Y\end{bmatrix}.$$

Note that $\widehat{c}_{\Phi_{11}} > 0$. We have $u_{\Phi_{11}} = 0$ by the KKT conditions (38)–(40). Taking the first $|\mathcal{I}|$ rows leads to

$$\begin{aligned}\lambda_n\begin{bmatrix}\mathbf{1}_{|\Phi_{11}|}\\ \mathbf{1}_{|\Phi_{01}|}\end{bmatrix} + \begin{bmatrix}0\\ u_{\Phi_{01}}\end{bmatrix} + \begin{bmatrix}\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}} & \widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{01}}\\ \widehat{f}_{\Phi_{01}}^T\widehat{f}_{\Phi_{11}} & \widehat{f}_{\Phi_{01}}^T\widehat{f}_{\Phi_{01}}\end{bmatrix}\\ \times\begin{bmatrix}\widehat{c}_{\Phi_{11}}\\ \widehat{c}_{\Phi_{01}}\end{bmatrix} = \begin{bmatrix}\widehat{f}_{\Phi_{11}}^T Y\\ \widehat{f}_{\Phi_{01}}^T Y\end{bmatrix}.\end{aligned}\qquad(43)$$

Thus, we have

$$\begin{aligned}\begin{bmatrix}\widehat{c}_{\Phi_{11}}\\ \widehat{c}_{\Phi_{01}}\end{bmatrix} &= \begin{bmatrix}\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}/n & \widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{01}}/n\\ \widehat{f}_{\Phi_{01}}^T\widehat{f}_{\Phi_{11}}/n & \widehat{f}_{\Phi_{01}}^T\widehat{f}_{\Phi_{01}}/n\end{bmatrix}^{-1}\\ &\quad\times\begin{bmatrix}\widehat{f}_{\Phi_{11}}^T Y/n - \lambda_n\mathbf{1}_{m_{11}}/n\\ \widehat{f}_{\Phi_{01}}^T Y/n - \lambda_n\mathbf{1}_{m_{01}}/n - u_{\Phi_{01}}/n\end{bmatrix}.\end{aligned}$$

To allow an abuse of notation, denote

$$\begin{aligned}F &= \widehat{f}_{\mathcal{I}}^T\widehat{f}_{\mathcal{I}}/n,\ F_{ij} = \widehat{f}_{\Phi_{i1}}^T\widehat{f}_{\Phi_{j1}}/n \text{ for } i,j = 0,1,\\ \widetilde{F} &= F_{00} - F_{01}F_{11}^{-1}F_{10}\\ &= \widehat{f}_{\Phi_{01}}^T(I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)\widehat{f}_{\Phi_{01}}/n.\end{aligned}$$

Note that F tends to the invertible matrix $f_{\mathcal{I}}^T f_{\mathcal{I}}/n$ as $n \to \infty$. It is seen from

$$F^{-1} = \begin{bmatrix} * & * \\ -\widetilde{F}^{-1}F_{01}F_{11}^{-1} & \widetilde{F}^{-1}\end{bmatrix}$$

that $\widetilde{F}$ is also invertible when n is sufficiently large. Similar to the derivation in the procedure of proving $P(|\Phi_{10}| > 0) \to 0$ as $n \to \infty$, it follows that $\widehat{c}_{\Phi_{01}} = \widetilde{F}^{-1}\xi$, where

$$\begin{aligned}\xi =\ &\widehat{f}_{\Phi_{01}}^T Y/n - \lambda_n\mathbf{1}_{m_{01}}/n - u_{\Phi_{01}}/n - F_{01}F_{11}^{-1}\widehat{f}_{\Phi_{11}}^T Y/n\\ &+ \lambda_n F_{01}F_{11}^{-1}\mathbf{1}_{m_{11}}/n\\ =\ &\widehat{f}_{\Phi_{01}}^T(I - \widehat{f}_{\Phi_{11}}(\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}})^{-1}\widehat{f}_{\Phi_{11}}^T)Y/n - \lambda_n\mathbf{1}_{m_{01}}/n\\ &+ \lambda_n F_{01}F_{11}^{-1}\mathbf{1}_{m_{11}}/n - u_{\Phi_{01}}/n\\ =\ &O_P(\delta_n) - (1+O_P(\delta_n))\lambda_n/n\mathbf{1}_{m_{10}} - u_{\Phi_{01}}/n\\ =\ &-\lambda_n/n\mathbf{1}_{m_{10}} - u_{\Phi_{01}}/n < 0,\end{aligned}$$

since $u_{\Phi_{01}} \ge 0$ by the KKT conditions. However, by the facts that $\widehat{c}_{\Phi_{01}} = 0$ and $\widetilde{F}$ is invertible, we obtain a violation $\xi = 0$. Therefore, we have shown that $P(|\Phi_{01}| > 0) \to 0$ as $n \to \infty$, i.e., Φ_{01} is also an empty set. Consequently, $\mathcal{I} = \Phi_{11}, \mathcal{I}^c = \Phi_{00}$ in the asymptotic sense.

Taking the first $|\Phi_{11}|$ rows in (43) leads to $\lambda_n\mathbf{1}_{m_{11}} + \widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}\widehat{c}_{\Phi_{11}} = \widehat{f}_{\Phi_{11}}^T Y$ since $\widehat{c}_{\Phi_{01}} = 0$. It follows that $\widehat{c}_{\Phi_{11}} = (\widehat{f}_{\Phi_{11}}^T\widehat{f}_{\Phi_{11}}/n)^{-1}(\widehat{f}_{\Phi_{11}}^T Y/n - \lambda_n\mathbf{1}_{m_{11}}/n)$. With the same derivation as that used in (42), we have

$$\begin{aligned}\frac{1}{n}\widehat{f}_{\mathcal{I}}^T\widehat{f}_{\mathcal{I}} &= \frac{1}{n}f_{\mathcal{I}}^T f_{\mathcal{I}} + O_P(\delta_n),\\ \frac{1}{n}\widehat{f}_{\mathcal{I}}^T Y &= \frac{1}{n}f_{\mathcal{I}}^T Y + O_P(\delta_n) = \frac{1}{n}f_{\mathcal{I}}^T(f_{\mathcal{I}}\mathbf{1}_{|\mathcal{I}|} + \varepsilon) + O_P(\delta_n)\\ &= \frac{1}{n}f_{\mathcal{I}}^T f_{\mathcal{I}}\mathbf{1}_{|\mathcal{I}|} + O_P(\delta_n),\end{aligned}$$

where the last equation uses the fact $\frac{1}{n}f_{\mathcal{I}}^T\varepsilon = O_P(\frac{1}{\sqrt{n}})$. Then it follows that

$$\begin{aligned}\widehat{c}_{\Phi_{11}} &= (f_{\mathcal{I}}^T f_{\mathcal{I}}/n)^{-1}(f_{\mathcal{I}}^T Y/n - \lambda_n \mathbf{1}_{|\mathcal{I}|}/n)(1+O_P(\delta_n))\\ &= (\mathbf{1}_{|\mathcal{I}|} - (f_{\mathcal{I}}^T f_{\mathcal{I}}/n)^{-1}\mathbf{1}_{|\mathcal{I}|}\lambda_n/n)(1+O_P(\delta_n))\\ &= \mathbf{1}_{|\mathcal{I}|}(1-O(\lambda_n/n))(1+O_P(\delta_n))\\ &= \mathbf{1}_{|\mathcal{I}|}(1+O_P(\lambda_n/n)).\end{aligned}$$

Therefore, $\widehat{f}_j^{\rm NG} = \widehat{f}_j(1+O_P(\lambda_n/n))$ for all j such that $f_j(\cdot)\neq 0$ and $P(\widehat{c}_j = 0)\to 1$ for all j such that $f_j(\cdot) = 0$. We obtain that

$$\begin{aligned}&\frac{1}{n}\|\widehat{f}_j^{\rm NG} - \widehat{f}_j\|^2 = \frac{1}{n}\|\widehat{f}_j O_P(\lambda_n/n)\|^2\\ &= \frac{1}{n}\|\widehat{f}_j\|^2 O_P(\lambda_n^2/n^2) = \frac{1}{n}\|\widehat{f}_j - f_j + f_j\|^2 O_P(\lambda_n^2/n^2)\\ &\leq \left(\frac{2}{n}\|\widehat{f}_j - f_j\|^2 + \frac{2}{n}\|f_j\|^2\right) O_P(\lambda_n^2/n^2)\\ &= \left(O_P(\delta_n^2) + \frac{2}{n}\|f_j\|^2\right) O_P(\lambda_n^2/n^2) = O_P(\lambda_n^2/n^2).\end{aligned}$$

Using the triangle inequality, we arrive at

$$\begin{aligned}\frac{1}{n}\|\widehat{f}_j^{\rm NG} - f_j\|^2 &\leq \frac{1}{n}\|\widehat{f}_j^{\rm NG} - \widehat{f}_j\|^2 + \frac{1}{n}\|\widehat{f}_j - f_j\|^2\\ &= O_P(\lambda_n^2/n^2) + O_P(\delta_n^2) = O_P(\lambda_n^2/n^2).\end{aligned}$$

This completes the proof. ■

Sketch Proof of Theorem 2: The proof follows from the same steps as what presented in [36]. The main idea is as follows. Let Ψ_j, $j = 1,\ldots,d$, be some operators acting on the additive functional space $\{f(v) = \sum_{j=1}^d f_j(v_j)\}$ such that $\Psi_j f(v) = f(v) - E(f(\mathbf{X})|\mathbf{X}_j = v_j)$, where it has been assumed that $E\mathbf{Y} = 0$, i.e., $f_0 = 0$ for simplicity of presentation. Clearly,

$$\Psi_j f(v) = \sum_{l\neq j}\left(f_l(v_l) - \int f_l(v_l)p(v_l|v_j)\mathrm{d}v_l\right), \qquad (44)$$

where $p(v_l|v_j)$ are the conditional densities of $\mathbf{X}_l$ given $\mathbf{X}_j$. Further, let $\widehat{\Psi}_j$ be the resulting estimates for Ψ_j, where $p(v_l|v_j)$ are replaced by their estimates $\widehat{p}(v_l|v_j) = \widehat{p}_{jl}(v_j,v_l)/\widehat{p}_j(v_j)$. That is

$$\widehat{\Psi}_j f(v) = \sum_{l\neq j}\left(f_l(v_l) - \int f_l(v_l)\frac{\widehat{p}_{jl}(v_j,v_l)}{\widehat{p}_j(v_j)}\mathrm{d}v_l\right). \qquad (45)$$

By applying (45), the equation (19) can be rewritten as $\widetilde{f}(v) = \widehat{\Psi}_j\widetilde{f}(v) + \widetilde{f}_j(v_j)$, where $\widetilde{f}(v) = \sum_{j=1}^d \widetilde{f}_j(v_j)$. Iterative application of this equation for the indices from $d,\ldots,1$ produces $\widetilde{f}(v) = \widehat{T}\widetilde{f}(v) + \widehat{\tau}(v)$, where $\widehat{T} = \widehat{\Psi}_d\cdots\widehat{\Psi}_1$ and

$$\widehat{\tau}(v) = \widehat{\Psi}_d\cdots\widehat{\Psi}_2\widehat{f}_1(v_1) + \cdots + \widehat{\Psi}_d\widehat{f}_{d-1}(v_{d-1}) + \widehat{f}_d(v_d). \quad (46)$$

This means that $\widetilde{f}(v) = \sum_{k=0}^{\infty}\widehat{T}^k\widehat{\tau}(v)$. So the uniqueness of the solution of the system of equations (19) depends on whether the norm of the operator $\widehat{T}$ is less than unity or not. Note from (45) that the operator $\widehat{T}$ only depends on the estimated densities $\widehat{p}_{jl}(v_j,v_l)/\widehat{p}_j(v_j)$. Using the convergence of $\widehat{p}_{jl}(v_j,v_l)/\widehat{p}_j(v_j)$ and the property of alternating projections [41], [42] yields that $\|\widehat{T}\| < \gamma$ for some constant $0 < \gamma < 1$ with probability tending to 1.

On the other hand, from the definition of the algorithm (20) one has $\widetilde{f}^{(k)}(v) = \widehat{T}\widetilde{f}^{(k-1)}(v) + \widehat{\tau}(v)$ and iterative application of the above formula derives

$$\widetilde{f}^{(k)}(v) = \sum_{l=0}^{k-1}\widehat{T}^l\widehat{\tau}(v) + \widehat{T}^k\widetilde{f}^{(0)}(v)$$

where $f^{(0)}(v)$ is the initial iterative value. This means that the backfitting algorithm (20) exponentially converges to the true values since $\|\widehat{T}\| < \gamma < 1$ with probability tending to 1.

Note from (44) that for $j\neq l$,

$$\Psi_j f_l(v_l) = f_l(v_l) - \int f_l(v_l)\frac{\widehat{p}_{jl}(v_j,v_l)}{\widehat{p}_j(v_j)}\mathrm{d}v_l \qquad (47)$$

and the second term is a function that only depends on v_j. The convergence of each additive function $\widetilde{f}_j(v_j)$ of $\widetilde{f}(v)$ is based on the following observations. The formula (47) implies that only the first term $\widehat{\Psi}_d\cdots\widehat{\Psi}_2\widehat{f}_1(v_1)$ of $\widehat{\tau}(v)$ defined in (46) depends on v_1 while the other terms are independent of v_1. Further, $\widehat{\Psi}_d,\ldots,\widehat{\Psi}_2$ are also independent of v_1. This means that $\widehat{\tau}(v)$ has a form of $\widehat{f}_1(v_1) + \widehat{\tau}_{-1}(v_2,\ldots,v_d)$, where $\widehat{\tau}_{-1}(v_2,\ldots,v_d)$ is a function that does not depend on v_1. This fact can obtain the convergence of $\widetilde{f}_1(v_1)$ and, using the similar method, can give the convergence of the other additive functions $\widetilde{f}_j(v_j), j = 2,\ldots,d$. ■

References

[1] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[2] H.-F. Chen and L. Guo, *Identification and Stochastic Adapive Control*. Boston, MA: Birkhäuser, 1991.

[3] B. Ninness and S. Gibson, "Quantifying the accuracy of Hammerstein model estimation," *Automatica*, vol. 38, no. 12, pp. 2037–2051, 2002.

[4] H.-F. Chen, "Pathwise convergence of recursive identification algorithms for Hammerstein systems," *IEEE Trans. Autom. Control*, vol. 49, no. 10, pp. 1641–1649, 2004.

[5] J. Wang, Q. Zhang, and L. Ljung, "Revisiting Hammerstein system identification through the two-stage algorithm for bilinear parameter estimation," *Automatica*, vol. 45, no. 11, pp. 2627–2633, 2009.

[6] E.-W. Bai and K. Li, "Convergence of the iterative algorithm for a general Hammerstein system identification," *Automatica*, vol. 46, no. 11, pp. 1891–1896, 2010.

[7] F. Giri, Y. Rochdi, and F.-Z. Chaoui, "An analytic geometry approach to Wiener system frequency identification," *IEEE Trans. Autom. Control*, vol. 54, no. 4, pp. 683–696, 2009.

[8] Y. Zhao, L. Y. Wang, G. G. Yin, and J.-F. Zhang, "Identification of Wiener systems with binary-valued output observations," *Automatica*, vol. 43, no. 10, pp. 1752–1765, 2007.

[9] G. Li and C. Wen, "Identification of Wiener systems with clipped observations," *IEEE Trans. Signal Process.*, vol. 60, no. 7, pp. 3845–3852, 2012.

[10] B.-Q. Mu and H.-F. Chen, "Recursive identification of MIMO Wiener systems," *IEEE Trans. Autom. Control*, vol. 58, no. 3, pp. 802–808, 2013.

[11] J. Roll, A. Nazin, and L. Ljung, "Nonlinear system identification via direct weight optimization," *Automatica*, vol. 41, no. 3, pp. 475–490, 2005.

[12] P. Craven and G. Wahba, "Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation," *Numer. Math.*, vol. 31, no. 4, pp. 377–403, 1978.

[13] E. A. Nadaraya, "On estimating regression," *Theory Probab. Appl.*, vol. 9, no. 1, pp. 141–142, 1964.

[14] G. S. Watson, "Smooth regression analysis," *Sankhyā: The Indian J. Statist., Ser. A*, vol. 26, no. 4, pp. 359–372, 1964.

[15] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Application*. London, U.K.: Chapman and Hall, 1996.

[16] W. Zhao, H.-F. Chen, and W. X. Zheng, "Recursive identification for nonlinear ARX systems based on stochastic approximation algorithm," *IEEE Trans. Autom. Control*, vol. 55, no. 6, pp. 1287–1299, 2010.

[17] E.-W. Bai, "Non-parametric nonlinear system identification: An asymptotic minimum mean squared error estimator," *IEEE Trans. Autom. Control*, vol. 55, no. 7, pp. 1615–1626, 2010.

[18] W. Zhao, W. X. Zheng, and E.-W. Bai, "A recursive local linear estimator for identification of nonlinear ARX systems: Asymptotical convergence and applications," *IEEE Trans. Autom. Control*, vol. 58, no. 12, pp. 3054–3069, 2013.

[19] E.-W. Bai, K. Li, W. Zhao, and W. Xu, "Kernel based approaches to local nonlinear non-parametric variable selection," *Automatica*, vol. 50, no. 1, pp. 100–113, 2014.

[20] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*. London, U.K.: Chapman and Hall, 1990.

[21] E.-W. Bai, "Identification of nonlinear additive FIR systems," *Automatica*, vol. 41, no. 7, pp. 1247–1253, 2005.

[22] E. Sochett, D. Daneman, C. Clarson, and R. Ehrlich, "Factors affecting and patterns of residual insulin secretion during the first year of type 1 (insulin-dependent) diabetes mellitus in children," *Diabetologia*, vol. 30, no. 7, pp. 453–459, 1987.

[23] Y. Q. Chen, C. A. Rohde, and M.-C. Wang, "Additive hazards models with latent treatment effectiveness lag time," *Biometrika*, vol. 89, no. 4, pp. 917–931, 2002.

[24] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Stat. Assoc.*, vol. 80, no. 391, pp. 580–598, 1985.

[25] W. Buytaert, R. Celleri, P. Willems, B. De Bievre, and G. Wyseure, "Spatial and temporal rainfall variability in mountainous areas: A case study from the south ecuadorian andes," *J. Hydrol.*, vol. 329, no. 3, pp. 413–421, 2006.

[26] A. Deaton, *Economics and Consumer Behavior*. Cambridge, U.K.: Cambridge University Press, 1980.

[27] E.-W. Bai and K.-S. Chan, "Identification of an additive nonlinear system and its applications in generalized Hammerstein models," *Automatica*, vol. 44, no. 2, pp. 430–436, 2008.

[28] E.-W. Bai, "Non-parametric nonlinear system identification: A data-driven orthogonal basis function approach," *IEEE Trans. Autom. Control*, vol. 53, no. 11, pp. 2615–2626, 2008.

[29] J. Huang, J. L. Horowitz, and F. Wei, "Variable selection in nonparametric additive models," *Ann. Statist.*, vol. 38, no. 4, pp. 2282–2313, 2010.

[30] A. Antoniadis, I. Gijbels, and A. Verhasselt, "Variable selection in additive models using p-splines," *Technometrics*, vol. 54, no. 4, pp. 425–438, 2012.

[31] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B, Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.

[32] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[33] M. Yuan and Y. Lin, "On the non-negative garrotte estimator," *J. Roy. Statist. Soc. Ser. B, Stat. Methodol.*, vol. 69, no. 2, pp. 143–161, 2007.

[34] W. Greblicki and M. Pawlak, *Nonparametric System Identification*. Cambridge, U.K.: Cambridge University Press, 2008.

[35] M. Rosenblatt, "A central limit theorem and a strong mixing condition," *Proc. Nat. Acad. Sci. USA*, vol. 42, no. 1, pp. 43–47, 1956.

[36] E. Mammen, O. Linton, and J. Nielsen, "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions," *Ann. Stat.*, vol. 27, no. 5, pp. 1443–1490, 1999.

[37] H. Tong, *Non-Linear Time Series: A Dynamical System Approach*. Oxford, U.K.: Oxford University Press, 1990.

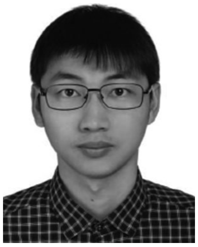
[38] J. P. Nielsen and S. Sperlich, "Smooth backfitting in practice," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 67, no. 1, pp. 43–61, 2005.

[39] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM J. Sci. Comput.*, vol. 14, no. 6, pp. 1487–1503, 1993.

[40] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.

[41] K. T. Smith, D. C. Solmon, and S. L. Wagner, "Practical and mathematical aspects of the problem of reconstructing objects from radiographs," *Bull. Amer. Math. Soc.*, vol. 83, no. 6, pp. 1227–1270, 1977.

[42] F. Deutscha and H. Hundalb, "The rate of convergence for the method of alternating projections, II," *J. Math. Anal. Appl.*, vol. 205, no. 2, pp. 381–405, 1997.



Biqiang Mu was born in Sichuan, China, in 1986. He received the B.Eng. degree in material formation and control engineering from Sichuan University in 2008 and the Ph.D. degree in operations research and cybernetics from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, in 2013.

He was a postdoctoral fellow at Wayne State University from 2013 to 2014 and also at Western Sydney University from 2015 to 2016. He is currently an assistant Professor at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. His research interests include system identification and applications.

Wei Xing Zheng (M'93–SM'98–F'14) received the B.Sc. degree in applied mathematics in 1982, the M.Sc. degree in electrical engineering in 1984, and the Ph.D. degree in electrical engineering in 1989, all from Southeast University, Nanjing, China.

He is currently a Professor with the University of Western Sydney, Australia. Over the years, he has also held various positions at Southeast University, China, Imperial College of Science, Technology and Medicine, U.K., University of Western Australia, Curtin University of Technology, Australia, Munich University of Technology, Germany, University of Virginia, USA, and University of California-Davis, USA.

Dr. Zheng was an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: FUNDAMENTAL THEORY AND APPLICATIONS, IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE SIGNAL PROCESSING LETTERS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, and IEEE TRANSACTIONS ON FUZZY SYSTEMS, and a Guest Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS. Currently, he is an Associate Editor for *Automatica*, the IEEE TRANSACTIONS ON AUTOMATIC CONTROL (the second term), IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and other scholarly journals.

Er-Wei Bai (M'90–SM'00–F'04) attended Fudan University (B.Sc. degree), Shanghai, China, Shanghai Jiaotong University (M.Eng. degree), and the University of California at Berkeley (Ph.D. degree).

He is a Professor and Chair of Electrical and Computer Engineering, and Professor of Radiology at the University of Iowa, Iowa City, where he teaches and conducts research in identification, control, signal processing, and their applications in engineering and life science. He also holds the rank of World Class Research Professor, Queen's University, Belfast, U.K.

Dr. Bai is a recipient of the President's Award for Teaching Excellence and the Board of Regents Award for Faculty Excellence.