Identification of Nonlinear Additive Autoregressive Models

Jianhua Z. Huang †

The Wharton School, University of Pennsylvania, Philadelphia, USA.

Lijian Yang

Michigan State University, East Lansing, USA.

Summary. We propose a lag selection method for nonlinear additive autoregressive models based on spline estimation and the BIC criterion. The additive structure of the autoregression function is used to overcome the "curse of dimensionality", while the spline estimators effectively take into account such a structure in estimation. A stepwise procedure is suggested to implement the proposed method. Comprehensive Monte Carlo study demonstrates good performance of the proposed method and substantial computational advantage over existing local polynomial based methods. Consistency of the BIC based lag selection method is established under the assumption that the observations are from a stochastic process that is strictly stationary and strongly mixing, which provides the first theoretical result of this kind for spline smoothing of weakly dependent data.

Key Words: BIC, lag selection, nonlinear time series, nonparametric method, splines, stochastic regression, variable selection.

1. Introduction

For virtually every time series modelling approach, there is a need to select significant explanatory lagged variables. The classic approach is to search for the optimal $\text{AR}(p)$ model via criteria such as AIC, FPE or BIC; see, for instance, Akaike (1969, 1970). Although the AIC, FPE and BIC are well-established criteria for selecting significant variables, their proper use is restricted to data sets that closely follow some linear $\text{AR}(p)$ structure. Many of the time series data of practical interests, however, exhibit nonlinearity.

Nonparametric methods have found significant applications in modelling nonlinearity in time series since the work of Robinson (1983). Györfi, Härdle, Sarda and Vieu (1989), Bosq (1998) systematically extended the results of kernel based smoothing to dependent data under various mixing assumptions. The important issue of lag selection has also been addressed using kernel based nonparametric extensions. Cheng and Tong (1992), Vieu (1994), Yao and Tong (1994) used a cross-validation (CV) approach, while Tjøstheim and Auestad (1990,1994b) used a nonparametric version of the final prediction error (FPE) criterion of Akaike (1969, 1970), all based on the Nadaraya-Watson estimator. Tschernig and Yang (2000), Yang and Tschernig (2002) introduced FPE with local linear estimators, with automatic bandwidth choice provided by the method of Yang and Tschernig (1999).

†*Address for correspondence:* Jianhua Huang, The Wharton School, Statistics Department, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, U.S.A. Email: **jianhua@wharton.upenn.edu**

All the aforementioned local polynomial kernel based lag selection methods are computationally intensive due to the local nature of kernel smoothing, and their ability to identify the exact set of lags suffers from the "curse of dimensionality". The "curse of dimensionality" refers to the inaccuracy when estimating multivariate regression functions nonparametrically. This limitation has led many authors to consider an additive form for the regression function as a compromise between the general nonparametric regression model and the simple linear model (Stone 1985, Hastie and Tibshirani 1990, Yang, Härdle and Nielsen 1999). Chen and Tsay (1993) suggested use of the adaptive backfitting BRUTO algorithm of Hastie (1989) to automatically determine the order and lags of an additive autoregressive model. This method, however, lacks theoretical justification, and no systematic simulation study has been done to evaluate its performance.

In this paper we propose a method based on spline estimation and the BIC criterion to select significant lags in nonlinear additive autoregression. The additive structure in the autoregression function is effectively taken into account through the use of polynomial spline global smoothing. Extensive Monte-Carlo study has demonstrated that our method is computationally fast and has good accuracy in identifying significant lags. Compared with local polynomial kernel based methods for lag selection, our method is simple and easy to implement. Recently, Rech, Teräsvirta, and Tschernig (2001) proposed a variable selection technique based on polynomial approximations, which shares the same simplicity as our method, although no theoretical justification is provided. In contrast, the intuitive appeal of polynomial spline smoothing is enhanced in this paper by a rigorous proof of the consistency of the BIC lag selection rule.

Lewis and Stevens (1991) used the multivariate adaptive regression splines (MARS) of Friedman (1991) to build adaptive spline autoregressive models. Although MARS can perform variable selection automatically, its ability to identify the set of significant variables (or lags) is unclear. In our simulation study, we find that MARS as well as BRUTO, using the default values of tuning parameters, tends to over-fit, that is, select more variables than necessary in the model. When we adjust the tuning parameters to penalize the degrees of freedom as strongly as in BIC (not as usually recommended), the performance of MARS and BRUTO improves, but is still not as good as our proposed method. These empirical findings explain in part the lack of theoretical justification of MARS and BRUTO for variable selection.

The application of the proposed method is not restricted to nonlinear autoregression. In fact, the general framework of nonlinear stochastic regression (Yao and Tong 1994) is adopted in this paper. Our method is applicable in selecting significant variables in regression models that may include endogenous variables (lagged variables) as well as exogenous variables. The consistency of the BIC criterion is established under the assumption that the observations are from a stochastic process that is strictly stationary and strongly mixing (α-mixing). Note that stronger mixing conditions (i.e., β-mixing) have been used to show the consistency of the CV or FPE method in the literature.

We organize our paper as follows. In Section 2 we set up the proper stochastic additive regression model and formulate the lag selection problem. As a necessary preparation, we describe in Section 3 the polynomial spline estimator for additive regression. In Section 4 the selection of significant variables (or lags) using spline estimation with the BIC criterion is proposed and the consistency of BIC is established in Section 5. Section 6 describes an implementation of the proposed method based on a stepwise procedure. Results of the Monte-Carlo study are reported in Section 7 and application of the proposed method to quarterly US unemployment rate time series is given in Section 8. All technical proofs are

given in the Appendix.

2.   The Model

We adopt the general framework of nonparametric stochastic regression. Let (X_t, Y_t), $t = 0, \pm 1, \ldots$, denote a (strictly) stationary time series with $X_t = (X_{t1}, \ldots, X_{td})$ being $\mathbb{R}^d$-valued ($d \geq 1$) and Y_t being real-valued. In particular, X_t may consist of lagged values of Y_t. Let $\mu(x) = E(Y_t | X_t = x)$, $x \in \mathbb{R}^d$, denote the regression function. Then we can write

$$Y_t = \mu(X_t) + \epsilon_t, \qquad t = 0, \pm 1, \ldots, \tag{1}$$

with $E(\epsilon_t | X_t) = 0$. When X_t consists of lagged values of Y_t, (1) becomes a nonparametric autoregressive model. The X_t can also include some exogenous variables. In this formulation, ϵ_t can be conditional homoskedastic ($\mathrm{var}(\epsilon_t | X_t)$ is a constant) or conditional heteroskedastic ($\mathrm{var}(\epsilon_t | X_t)$ is not a constant).

The goal of this paper is to determine, without assuming that μ is known, a proper subset of variables $\{X_{ti}, i \in s\}$, $s \subset \{1, \ldots, d\}$, with the cardinality of s (denoted as $\#(s)$) as small as possible which provides (almost) the same information on Y_t as $X_t = (X_{t1}, \ldots, X_{td})$, i.e.,

$$E(Y_t | X_{ti}, i \in s) = E(Y_t | X_t), \qquad a.s.$$

The selected variables are called significant variables. If X_t consists of only lagged values of Y_t, the selected lags are called significant lags. Since we do not assume the regression function μ to have a known parametric form, our method is nonparametric in nature. It is well-known that nonparametric estimation suffers from the "curse of dimensionality". One way to overcome the difficulty is to impose some structure on the unknown regression function. In this paper we will consider additive models. We assume that, for a collection of significant variables an additive model holds, and we want to determine the significant variables (or lags) from the data.

3.   Additive Spline Estimation

To select significant variables or lags, we need some nonparametric techniques to estimate the regression function $E(Y_t | X_{ti}, i \in s)$ for any candidate subset $\{X_{ti}, i \in s\}$ of significant variables. This section gives a description of the additive spline estimation method that is used in this paper. This method has been studied theoretically by Stone (1985, 1994) and Huang (1998, 2001).

An additive model for the regression function $\mu(x) = E(Y_t | X_t = x)$ assumes that

$$\mu(x) = \mu_0 + \sum_{i=1}^{d} \mu_i(x_i), \tag{2}$$

where μ_0 is a constant. For identification purpose, we assume that $E[\mu_i(X_{ti})] = 0, i = 1, \ldots, d$, in the above equation. To fit this model, we approximate each $\mu_i(x_i)$ by a spline function and then use the least squares method. To be specific, we can write

$$\mu_i(x_i) \approx \sum_{j=1}^{J_i} \gamma_{ji} B_{ji}(x_i), \tag{3}$$

where $B_{ji}, j = 1, \ldots, J_i$, is a basis of the space of spline functions for a given degree and knot sequence. Commonly used bases for spline functions are truncated power bases or B-spline bases (see de Boor 1978). Then, for a sample (X_t, Y_t), $t = 1, \ldots, n$, we minimize over $\{\mu_0, \gamma_{ji}, j = 1, \ldots, J_i, i = 1, \ldots, d\}$ the criterion

$$\sum_t \left\{ Y_t - \mu_0 - \sum_{i=1}^{d} \sum_{j=1}^{J_i} \gamma_{ji} B_{ji}(X_{ti}) \right\}^2. \tag{4}$$

Denote the minimizers as $\widehat{\mu}_0$ and $\widehat{\gamma}_{ji}, j = 1, \ldots, J_i, i = 1, \ldots, d$. The spline estimate of μ is given by

$$\widehat{\mu}(x) = \widehat{\mu}_0 + \sum_{i=1}^{d} \widehat{\mu}_i(x_i) \tag{5}$$

where

$$\widehat{\mu}_i(x_i) = \sum_{j=1}^{J_i} \widehat{\gamma}_{ji} B_{ji}(x_i).$$

The success of the spline method is due to the fact that polynomial splines provide good approximations to smooth functions. Indeed, rather accurate approximations can be achieved in (3) by increasing the number of knots, provided that μ_i satisfies some smoothness condition (see, for example, Chapter XII of de Boor 1978). By letting the number of knots or equivalently, the number of terms in (3), increase with the sample size, the spline estimate is consistent (Stone 1985, Huang 1998, 2001) in estimating any additive function with smooth components.

Note that the estimation method described above is very easy to implement. After the basis functions are chosen, operationally the problem reduces to a parametric linear regression problem. Standard efficient algorithms (see Miller 2002) for linear least squares regression can be employed for fitting the additive model. The simplicity of this method is advantageous for our purpose, since variable (or lag) selection requires fitting and comparing many candidate additive models. Alternative methods for fitting additive models such as the backfitting algorithm (Buja, Hastie and Tibshirani, 1989) and the integration method (Tjøstheim and Auestad 1994a, Linton and Nielsen 1995, Masry and Tjøstheim 1997, Mammen, Linton and Nielsen 1999) can also be applied, but their use for variable selection is not as straightforward as our simple spline method.

4. Selection of Significant Variables

Consider the variable (or lag) selection problem for additive models. We assume that, for some index set $s_0 \subset \{1, \ldots, d\}$, the actual regression function $\mu(x) = E(Y_t | X_t = x)$, $x = (x_1, \ldots, x_d)$, is an additive function in $x_i, i \in s_0$. Note that, if such an s_0 exists, for any s satisfying $s_0 \subset s \subset \{1, \ldots, d\}$, $\mu(x)$ is also an additive function in $x_i, i \in s$. We thus assume that s_0 has the smallest cardinality among sets with the specified property.

To give a more formal definition of the set of significant variables, we assume that the regression function is square-integrable. For each $s \subset \{1, \ldots, d\}$, let $\mathbb{H}_s$ denote the space of all square-integrable additive functions of variables $x_i, i \in s$. We view functions in $\mathbb{H}_s$ as functions of $x_i, i = 1, \ldots, d$, and thus have $\mathbb{H}_s \subset \mathbb{H}_{s'} \subset \mathbb{H}_{\{1,\ldots,d\}}$ for $s \subset s' \subset \{1, \ldots, d\}$.

Definition. The index set s_0 of significant variables is the minimal set $s \subset \{1, \ldots, d\}$ such that $\mu \in \mathbb{H}_s$. The variables $X_{ti}, i \in s_0$, are called significant variables.

Logically it is possible that for two subsets s_0 and s_0' of $\{1, \ldots, d\}$ with $s_0 \not\subset s_0'$ and $s_0' \not\subset s_0$ we have $\mu \in \mathbb{H}_{s_0}$ and $\mu \in \mathbb{H}_{s_0'}$. If this is the case, s_0 in the above definition is not unique. The following result rules out such a possibility.

LEMMA 1. *Assume that $X_t = (X_{t1}, \ldots, X_{td})$ has a joint density relative to the Lebesgue measure. Then the index set s_0 of significant variables is uniquely defined.*

There are three possible outcomes for a variable selection method.

Definition. Let s be the index set of selected variables. We say that s correct-fits, if $s = s_0$; we say that s over-fits, if $s \supset s_0$ but $s \neq s_0$; we say that s under-fits, if $s_0 \not\subset s$.

In words, "over-fitting" means that the set of selected variables includes other variables in additional to the significant variables; "under-fitting" means that the set of selected variables does not include all of the significant variables.

To define the variable selection criterion for additive models, we need, for each $s \subset \{1, \ldots, d\}$, an estimate of the regression function pretending the index set of significant variables is s. Specifically, let $\mathbb{G}_s$ denote the space of functions having the form

$$g(x) = g_0 + \sum_{i \in s} g_i(x_i),$$

with g_0 a constant and $g_i \in \mathbb{G}_i$, where $\mathbb{G}_i$ is a space of spline functions, defined on the range of X_{ti}, with degree q_i and J_i interior knots. Then the dimension of $\mathbb{G}_i$ is $N_i = 1 + q_i + J_i$, $i = 1, \ldots, n$. Taking into account the identifiability constraints, it is easily seen that the dimension of $\mathbb{G}_s$ is $N_s = 1 + \sum_{i \in s}(q_i + J_i)$. The spline estimate corresponding to the index set s is

$$\widehat{\mu}_s = \arg\min_{g \in \mathbb{G}_s} \sum_{t=1}^{n} \{Y_t - g(X_t)\}^2. \tag{6}$$

Here, we view each function in $\mathbb{G}_s$ as a function of $x_i, i = 1, \ldots, d$, and thus have $\mathbb{G}_s \subset \mathbb{G}_{s'} \subset \mathbb{G}_{\{1,\ldots,d\}}$ for $s \subset s' \subset \{1, \ldots, d\}$. For each subset s of $\{1, \ldots, d\}$, define the mean squared error of $\widehat{\mu}_s$ as

$$\mathrm{MSE}_s = \frac{1}{n} \sum_{t=1}^{n} \{Y_t - \widehat{\mu}_s(X_t)\}^2, \tag{7}$$

and the BIC criterion as

$$\mathrm{BIC}_s = \log(\mathrm{MSE}_s) + \frac{N_s}{n} \log n.$$

The BIC criterion was first proposed in Schwarz (1978) for selection of parametric models for independent and identically distributed data.

Variable Selection Rule. Select the subset $\widehat{s} \subset \{1, \ldots, d\}$ with the smallest BIC value.

5. Consistency of the Variable Selection Rule

In this section we show that under appropriate assumptions, the Variable Selection Rule defined in the previous section is consistent.

Note that it is hard to obtain a reliable nonparametric estimate of the regression function at the tail of the distribution of X_{ti} due to data sparseness. Thus we focus on the estimation of μ_i on a compact set in our theoretical analysis below. Let $\mathbb{C}_i$ be a compact interval

contained in the range of X_{ti} and let $\mathbb{C}$ be the Cartesian product of $\mathbb{C}_i$, $i = 1, \ldots, d$. We require that $\mathbb{C}_i$ contains all the interior knots for splines in $\mathbb{G}_i$, $i = 1, \ldots, d$. We modify (6) and (7) slightly to

$$\widehat{\mu}_s = \arg \min_{g \in \mathbb{G}_s} \sum_{t=1}^{n} \{Y_t - g(X_t)\}^2 \mathrm{I}(X_t \in \mathbb{C})$$

and

$$\mathrm{MSE}_s = \frac{1}{n} \sum_{t=1}^{n} \{Y_t - \widehat{\mu}_s(X_t)\}^2 \mathrm{I}(X_t \in \mathbb{C}),$$

and modify the BIC accordingly. We simply choose $\mathbb{C}_i$ to be the range of the data in our numerical implementation.

We next introduce some additional notations and assumptions. For two sequences of positive numbers a_n and b_n, let $a_n \lesssim b_n$ mean that a_n/b_n is bounded and $a_n \asymp b_n$ mean that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The α-mixing coefficient of the process $\{(X_t, Y_t)\}$ is defined as

$$\alpha(n) = \sup\{P(B \cap C) - P(B)P(C) : B \in \sigma(\{(X_{t'}, Y_{t'}), t' \leq t\}), C \in \sigma(\{(X_{t'}, Y_{t'}), t' \geq t + n\})\},$$

where, for an index set $\mathcal{T}$, $\sigma(\{Z_t, t \in \mathcal{T}\})$ denotes the σ-field generated by the random variables $\{Z_t, t \in \mathcal{T}\}$. Note that the right side of the above equation does not depend on t because $\{(X_t, Y_t)\}$ is stationary.

Recall that each $\mathbb{G}_i$, $i = 1, \ldots, d$, is a space of splines with J_i knots. Suppose the ratios of the differences between consecutive knots are bounded. Let $J_n = n^{\gamma}$ for $0 < \gamma < 1$. Assume $J_i \asymp J_n$ for $i = 1, \ldots, d$. Recall that s_0 is the set of significant lags. Set $\rho_{s_0} = \inf_{g \in \mathbb{G}_{s_0}} \|g - \mu\|_{\infty}$. The quantity ρ_{s_0} measures the best obtainable approximation rate for using functions in $\mathbb{G}_{s_0}$ to approximate μ.

We make the following assumptions on the data generating process.

(A1) $\sup_{x \in \mathbb{C}} E[|Y_t|^{\nu}|X_t = x] < \infty$ for some $\nu > 2$.

(A2) For some positive constants c_1 and c_2, the α-mixing coefficient of $\{(X_t, Y_t)\}$ satisfies $\alpha(n) \leq c_1 n^{-(5/2)\gamma/(1-\gamma)}$ and $\alpha(n) \leq c_2 n^{-2\nu/(\nu-2)}$.

(A3) The density p_{X_0} of X_0 is bounded away from 0 and ∞ on $\mathbb{C}$.

(A4) $\lim_{n \to \infty} \rho_{s_0} = 0$ and $\limsup_{n \to \infty} \rho_{s_0}^2/(J_n/n) < \infty$.

A moment condition as in (A1) is commonly used in the literature. It follows from (A1) that the conditional variance of Y_t given $X_t = x$ is bounded on $x \in \mathbb{C}$. Assumption (A2) requires that the α-mixing coefficient decays algebraically to zero. Stronger conditions involving β-mixing coefficient have been used in Yao and Tong (1994), Tjøstheim and Auestad (1994b), and Tschernig and Yang (2000) to show consistency of a certain lag selection criterion.

Assumption (A3) is a mild condition on the marginal density of X_t (note that X_t is stationary). It is usually assumed that p_{X_0} is continuous or continuously differentiable in asymptotic analysis of the local polynomial method; see, for example, Tschernig and Yang (2000). In (A4), the quantities ρ_{s_0} and J_n/n measure respectively the magnitude of the bias and variance for $\widehat{\mu}_{s_0}$ (see Lemma A.2). Thus (A4) requires that the squared bias is not asymptotically dominated by the variance of the estimator (i.e., there is no over-smoothing). Assumption (A4) can be replaced by a smoothness condition of the regression function and a requirement on the knot number. To be precise, write

$$\mu(x) = \mu_0 + \sum_{i \in s_0} \mu_i(x_i),$$

where $E[\mu_i(X_{t,i})] = 0$, $i \in s_0$. Recall that we require the knot number satisfies $J_n \asymp n^\gamma$, $0 < \gamma < 1$.

LEMMA 2. *Suppose each μ_i, $i \in s_0$, has bounded second derivative. In addition, suppose that the degree of splines is 1 or bigger. Then a sufficient condition for* (A4) *is that $\gamma \geq 1/5$.*

Here is our main theoretical result.

THEOREM 1. *Suppose Assumptions* (A1)–(A4) *hold. The Variable Selection Rule consistently selects the set of significant variables, that is, $\lim_{n \to \infty} P(\widehat{s} = s_0) = 1$.*

Observe that the consistency of the Variable Selection Rule holds for a wide range of choices for the number of knots J_n. If each μ_i, $i \in s_0$, has bounded second derivative, and splines of degree 1 or bigger are used, then it is sufficient to have $J_n \asymp n^\gamma$ with $\gamma \geq 1/5$. It is worthwhile to point out that if $J_n \asymp n^{1/5}$, then $\|\widehat{\mu}_{s_0} - \mu\| = O_P(n^{-4/5})$, which is the optimal rate of convergence (Stone 1985).

There has been extensive study on variable selection for parametric linear models (Miller 2002). Many consistent variable selection criteria have been suggested; see Rao and Wu (1989) for a survey. It would be interesting to point out the difference between Theorem 1 and existing consistency results for parametric models. In our setting, adding a variable to the model corresponds to adding a function. If the function is approximated by a spline function, then adding a variable corresponds to adding a set of bases for splines. In addition, in order for the spline estimate to be consistent, it is necessary to let the number of knots (or the number of spline basis functions for each variable) increase with the sample size.

6.   Implementation

In actual implementation of the proposed method, one first decides on a set of candidate variables to be selected. The candidate variables can be the lagged variables of a time series and/or some exogenous variables. Since a full search through all possible subsets of variables is in general computationally too costly, we propose a stepwise procedure. The procedure consists of three stages: a forward stage, a backward stage, and a final selection stage.

In the forward stage, one starts from the null model (i.e., $Y_t = \mu_0 + \epsilon_t$, where μ_0 is a constant), adds one variable at a time to the current model, choosing among the various candidate variables not yet selected by minimizing the mean squared error (MSE, see Section 4). This addition process stops when the number of selected variables equals some pre-specified number, say, $S_{\max}$. The constant $S_{\max}$ is the maximal number of variables allowed in the model.

The backward stage follows the forward stage. In this stage, one starts with the maximal set of variables selected in the last step of the forward stage, deletes one variable at a time by also minimizing the MSE, and stops when no variable remains in the model. After the forward and the backward stages, one obtains a collection of "good" models. The final model is chosen from this collection by minimizing the BIC criterion.

Let d denote the total number of candidate variables to be selected from. It is necessary to require that $S_{\max} \leq d$. If d is not very big, we can set $S_{\max} = d$ and the forward stage of our procedure is not necessary.

For each step in the forward/backward stage, we fit an additive spline model. In our implementation, the knots are equally placed between the 5% and 95% sample quantiles of

Table 1. Dynamics of the time series in simulation study.

AR1	$Y_t = 0.5Y_{t-1} + 0.4Y_{t-2} + 0.1\xi_t$
AR2	$Y_t = -0.5Y_{t-1} + 0.4Y_{t-2} + 0.1\xi_t$
AR3	$Y_t = -0.5Y_{t-6} + 0.5Y_{t-10} + 0.1\xi_t$
NLAR1	$Y_t = -0.4(3 - Y_{t-1}^2)/(1 + Y_{t-1}^2) + 0.6\{3 - (Y_{t-2} - 0.5)^3\}/\{1 + (Y_{t-2} - 0.5)^4\} + 0.1\xi_t$
NLAR2	$Y_t = \{0.4 - 2\exp(-50Y_{t-6}^2)\}Y_{t-6} + \{0.5 - 0.5\exp(-50Y_{t-10}^2)\}Y_{t-10} + 0.1\xi_t$
NLAR3	$Y_t = \{0.4 - 2\cos(40Y_{t-6})\exp(-30Y_{t-6}^2)\}Y_{t-6} + \{0.55 - 0.55\sin(40Y_{t-10})\exp(-10Y_{t-10}^2)\}Y_{t-10} + 0.1\xi_t$
NLAR1U1	$Y_t = -0.4(3 - Y_{t-1}^2)/(1 + Y_{t-1}^2) + 0.1\xi_t$
NLAR1U2	$Y_t = 0.6\{3 - (Y_{t-2} - 0.5)^3\}/\{1 + (Y_{t-2} - 0.5)^4\} + 0.1\xi_t$

the data. Let $\lfloor x \rfloor$ denote the smallest integer bigger than or equal to x. Partly motivated by the asymptotic result, the number of knots is set to be $\lfloor (kn)^{1/5} \rfloor$ for linear splines and $\lfloor (kn)^{1/5} \rfloor - 1$ for quadratic and cubic splines, where k is a tuning constant whose default value is 2 in our implementation. It has been observed in our simulation study that the lag selection results are not very sensitive to the choice of tuning constant.

Similar stepwise procedures have been used for variable selection in linear regression. There are however noteworthy differences between our method and the method for linear regression. For our method, adding or deleting one variable corresponds to adding or deleting an additive component of the model that consists of a linear combination of several (B-spline) basis terms.

7.   Simulation study

We have conducted Monte Carlo simulations to evaluate the performance of the proposed method and to compare with other methods. In particular, our proposed method of spline fitting with BIC criterion is compared with spline fitting with AIC or GCV criteria. For an index set of variables $s \subset \{1, \ldots, d\}$, these criteria are defined as

$$\text{AIC}_s = \log(\text{MSE}_s) + 2\frac{N_s}{n} \qquad \text{and} \qquad \text{GCV}_s = \frac{\text{MSE}_s}{(1 - N_s/n)^2}. \tag{8}$$

When N_s/n is small, which is usually the case, it can be seen from the approximation $1/(1-x)^2 \approx 1 + 2x$ that AIC and GCV are similar. Our method is also compared with the local linear FPE method of Tschernig and Yang (2000), MARS (Friedman 1991) and BRUTO (Hastie 1989).

In our simulation study, we considered eight additive autoregressive (AR) processes with some autoregression functions being linear and some nonlinear. Six of these processes (AR1-AR3, NLAR1-NLAR3) were used in Tschernig and Yang (2000) to evaluate their lag selection method. We also considered two other processes (NLAR1U2 and NLAR1U2) each with one significant lag in our simulation study. The dynamics of these processes are described by the equations given in Table 1, where ξ_t are i.i.d. $N(0, 1)$ random variables.

These processes differ in the shape of the conditional mean function and the lag vector. We used the computing software Splus for all of our simulations with the same initial random seed. The Splus functions "`mars()`" and "`bruto()`" in the "mda" library contributed by Trever Hastie and Robert Tibshirani were used for the MARS and BRUTO simulation (the "mda" library was downloaded from StatLib: "`http://lib.stat.cmu.edu/`"). For sample sizes $n = 100, 200, 500$, realizations of size $n+400$ were generated and the last n observations

Table 2. Simulation results for lag selection using spline-fitting with BIC or AIC criteria. For each setup, the first, second, and third columns give respectively the number of under-fitting, correct-fitting (in bold face), and over-fitting over 100 simulation runs.

Models	n	BIC degree =1			BIC degree =2			BIC degree =3			AIC degree =1		
AR1	100	69	**28**	3	69	**28**	3	85	**15**	0	92	**8**	0
	200	18	**82**	0	19	**81**	0	32	**68**	0	0	**0**	100
	500	0	**100**	0	0	**100**	0	0	**100**	0	0	**0**	100
AR2	100	52	**41**	7	50	**47**	3	70	**29**	1	83	**16**	1
	200	10	**90**	0	11	**89**	0	27	**73**	0	0	**0**	100
	500	0	**100**	0	0	**100**	0	0	**100**	0	0	**0**	100
AR3	100	10	**87**	3	9	**89**	2	23	**77**	0	26	**74**	0
	200	0	**100**	0	0	**99**	1	2	**97**	1	0	**0**	100
	500	0	**100**	0	0	**100**	0	0	**100**	0	0	**0**	100
NLAR1	100	0	**83**	17	0	**92**	8	0	**99**	1	0	**27**	73
	200	0	**95**	5	0	**99**	1	0	**100**	0	0	**34**	66
	500	0	**85**	15	0	**100**	0	0	**100**	0	0	**9**	91
NLAR2	100	33	**64**	3	38	**59**	3	62	**37**	1	12	**14**	74
	200	2	**97**	1	4	**96**	0	11	**89**	0	0	**32**	68
	500	0	**100**	0	0	**100**	0	0	**100**	0	0	**43**	57
NLAR3	100	21	**73**	6	19	**75**	6	27	**72**	1	8	**22**	70
	200	1	**99**	0	0	**100**	0	3	**97**	0	0	**35**	65
	500	0	**100**	0	0	**100**	0	0	**100**	0	0	**36**	64
NLAR1U1	100	0	**97**	3	0	**95**	5	0	**100**	0	0	**40**	60
	200	0	**99**	1	0	**99**	1	0	**100**	0	0	**46**	54
	500	0	**100**	0	0	**100**	0	0	**100**	0	0	**54**	46
NLAR1U2	100	0	**97**	3	0	**98**	2	0	**100**	0	0	**34**	66
	200	0	**99**	1	0	**98**	2	0	**100**	0	0	**37**	63
	500	0	**100**	0	0	**100**	0	0	**100**	0	0	**43**	57

were taken as the observed time series. This, together with the form of the conditional mean function, ensure that the realizations behave like strictly stationary and geometrically β-mixing, thus more than fulfilling our assumption (A1). We generated 100 replications for each of the above processes and carried out lag selection for each replication. The lags were searched from $\{1, \ldots, 10\}$ for all methods. In implementing the proposed method, we set the maximum number of variables allowed in the model to be $S_{\max} = 10$. We have documented the over-fit, correct-fit and under-fit frequencies for all processes in Table 2 for the spline fitting with BIC and AIC criterion and in Table 3 for MARS and BRUTO. Here under-fitting refers to the selection of the correct variables, not the number of variables in the model. For instance, at $n = 100$, the 92 under-fits for AIC with spline-fitting for the first AR model often included more than two lagged variables, but missed at least one of the correct lags of $t - 1$ and $t - 2$.

Now we summarize our simulation results.

- The selection procedure based on spline fitting with BIC criterion (referred to hereafter as the proposed method) performs very well and is robust for all processes. When the sample size increases from 100 to 200 and 500, the frequency of correct-fitting (the middle number of the triplet from the table) increases to 100 or close to 100 in all situations simulated. This corroborates the asymptotic consistency result.

Table 3. Simulation results for lag selection using MARS and BRUTO. For each setup, the first, second, and third columns give respectively the number of under-fitting, correct-fitting (in bold face), and over-fitting over 100 simulation runs. The constant a specifies the cost per degree of freedom change.

| | | MARS | | | | | | BRUTO | | | | | |
Models	n		$a = 2$			$a = \log n$			$a = 2$			$a = \log n$	
AR1	100	17	**29**	54	27	**61**	12	6	**35**	59	18	**68**	14
	200	0	**31**	69	0	**79**	21	0	**21**	79	0	**75**	25
	500	0	**22**	78	0	**68**	32	0	**10**	90	0	**79**	21
AR2	100	22	**21**	57	30	**56**	14	4	**26**	70	14	**61**	25
	200	1	**24**	75	1	**71**	28	0	**16**	84	0	**71**	29
	500	0	**19**	81	0	**67**	33	0	**13**	87	0	**66**	34
AR3	100	4	**32**	64	5	**74**	21	2	**53**	45	8	**86**	6
	200	0	**20**	80	0	**72**	28	0	**50**	50	0	**98**	2
	500	0	**17**	83	0	**71**	29	0	**58**	42	0	**99**	1
NLAR1	100	0	**29**	71	0	**65**	35	0	**48**	52	1	**85**	14
	200	0	**25**	75	0	**78**	22	0	**52**	48	0	**89**	11
	500	0	**26**	74	0	**62**	38	0	**34**	66	0	**93**	7
NLAR2	100	12	**23**	65	33	**45**	22	16	**61**	23	94	**6**	0
	200	0	**16**	84	0	**65**	35	0	**60**	40	6	**93**	1
	500	0	**18**	82	0	**56**	44	0	**47**	53	0	**96**	4
NLAR3	100	5	**14**	81	10	**55**	35	4	**37**	59	11	**78**	11
	200	0	**16**	84	0	**65**	35	0	**49**	51	0	**93**	7
	500	0	**21**	79	0	**65**	35	0	**49**	51	0	**98**	2
NLAR1U1	100	0	**35**	65	0	**77**	23	0	**7**	93	0	**33**	67
	200	0	**19**	81	0	**66**	34	0	**1**	99	0	**38**	62
	500	0	**11**	89	0	**66**	34	0	**0**	100	0	**25**	75
NLAR1U1	100	0	**27**	73	0	**68**	32	0	**0**	100	0	**0**	100
	200	0	**22**	78	0	**62**	38	0	**0**	100	0	**0**	100
	500	0	**12**	88	0	**46**	54	0	**0**	100	0	**0**	100

- For the proposed method, using linear, quadratic or cubic splines gives similar results, except that cubic splines give slightly worse results for AR1, AR2, NLAR2 for sample size $n = 100$.

- Spline fitting with the AIC criterion over-fits, that is, very often it chooses more variables than are in the true model. We have also observed that the GCV criterion behaves similarly as AIC (results not shown). This is not surprising in light of the similarity of the two criteria as explained after (8).

- The proposed method almost always outperforms the local linear FPE method of Tschernig and Yang (2000) (see pages 472-473 of the cited paper). The only exceptions are cubic spline fitting for AR1 and NLAR2 for sample size $n = 100$. Moreover, we note that the proposed method is computationally much faster than the local linear FPE method. We would like to point out that running 100 simulations with the proposed method takes less than 10 minutes on a pentium PC, for sample size as large as $n = 500$, while it took days to run 100 simulations of $n = 100$ with the local linear FPE method of Tschernig and Yang (2000). For this comparison, we recoded our method using Xplore, the same software that the local linear method of Tschernig and Yang (2000) was originally coded. In addition to having an advantage in computation speed, the proposed method is also much easier to program because of its simplicity.

- MARS uses a modified GCV to select models (see Friedman 1991 for details). In "mars()", there is a tuning parameter (denoted as a) that specifies the cost per degree of freedom change. The default value of a is 2, corresponding to a model selection criterion that is similar to AIC. It has been recommended to use $2 \leq a \leq 4$ (Friedman 1991 and Stone et al. 1997). The results for $a = 2$ and $a = \log n$ are reported in Table 3. We see that using $a = 2$ (AIC type of penalty) always yields substantial over-fitting. It is interesting to see that even when we change the cost parameter a to $\log n$ (BIC type of penalty), a value much higher than what is usually recommended, MARS still over-fits for about one-third of the simulations for all processes. The performance of MARS for $a = 3$ or 4 is between that for $a = 2$ and $a = \log n$ (the detailed results are not reported).

- The BRUTO algorithm combines backfitting and adaptive smoothing parameter selection and uses a modified GCV for model selection (see Hastie 1989 for details). Chen and Tsay (1993) used it to automatically select the significant variables for additive models. In "bruto()", there is a tuning parameter (denoted as a) that specifies the cost per degree of freedom change. The default value $a = 2$ corresponds to a penalty similar to that in AIC. We observe that, similar to MARS, when the cost parameter $a = 2$, BRUTO tends to over-fit. When the cost parameter is set to $a = \log n$, BRUTO performs quite well for some processes, but very bad for other processes (i.e., NLAR1U1 and NLAR1U2).

8. Real data example

In addition to the Monte Carlo evidence of the effectiveness of the proposed variable selection method, we further illustrate the practical usefulness of the method for building a parsimonious additive autoregressive model for quarterly US unemployment rate time

series. We also carry out a rolling forecasting exercise based on the identified additive autoregression model.

The data set analyzed here is the non-seasonally adjusted quarterly series of US unemployment rate from the first quarter of 1948 to the first quarter of 2003, denoted as $\{R_t\}_{t=1}^{221}$. It is obtained from US Bureau of Labor Statistics, and covers unemployed persons (in the labor force) of 16 years and older of all ethnic origins, races and sexes, without distinction of industries or occupations. The fourth difference of the data is taken in order to eliminate seasonality. The resulting difference series is denoted as $\{Y_t\}_{t=1}^{217}$, $Y_t = R_{t+4} - R_t, t = 1, ..., 217$. We leave out the last 10 periods of the data (i.e., $\{Y_t\}_{t=208}^{217}$) for the forecasting exercise and use the rest of the series for model building.

Using the proposed method (spline fitting with BIC criterion), MARS and BRUTO (both with BIC type penalty), we construct an additive autoregression model

$$Y_t = f_{i_1}(Y_{t-i_1}) + \cdots + f_{i_k}(Y_{t-i_k}) + \epsilon_t,$$

where the significant lags $\{i_1, \ldots, i_k\}$ are chosen from $\{1, \ldots, 8\}$. We also use the BIC to choose a linear autoregressive model, which constrains each of the f_{i_j} in the above equation to be a linear function. The selected significant lags using various methods are presented in Table 4. We see that the proposed method, with linear, quadratic or cubic splines, always picks a parsimonious model with lags 1 and 2, while MARS, BRUTO, and BIC for linear AR all pick more lags.

The fitness of the selected models is measured by the coefficient of determination

$$R^2 = 1 - \sum_{t=11}^{207}(Y_t - \widehat{Y}_t)^2 \bigg/ \sum_{t=11}^{207}(Y_t - \bar{Y})^2,$$

where $\widehat{Y}_t$ is the fitted value at time period t, and $\bar{Y} = \sum_{t=11}^{207} Y_t/196$. The R^2 values are high for all selected models, suggesting all models fit the data well. The fitted models are used for producing out-of-sample, rolling forecasts for time periods 208 to 217. The forecasting performance is measured in terms of the MSPE (mean squared prediction error) and MAPE (mean absolute prediction error), which are defined as

$$\text{MSPE} = \frac{1}{10}\sum_{t=208}^{217}(Y_t - \widehat{Y}_t)^2 \quad \text{and} \quad \text{MAPE} = \frac{1}{10}\sum_{t=208}^{217}|Y_t - \widehat{Y}_t|,$$

where $\widehat{Y}_t$ are forecasts produced by the selected model. From Table 4 we see that the forecasting performance of models selected by the proposed method using splines of degree one, two or three is comparable to that of the MARS model, and is superior to the BRUTO model and the linear AR model. However, the model selected by our method, which has significant lags 1 and 2, is easier to interpret than the more complicated MARS model which has 1,2,4,,5 as the significant lags. Note that the BRUTO model and the linear AR model also use many lag variables in spite of their relatively poor forecasting performance.

In conclusion, for the quarterly US unemployment rate, we have found evidence that our proposed method can identify a parsimonious nonlinear additive autoregression model with good forecasting performance. This conclusion does not depend on what degree of splines we use in estimation. The results of the out-of-sample forecasting exercise also suggests that nonlinear models (identified by the proposed method or MARS) provide better description of dynamics of the quarterly US unemployment rate time series than linear AR models.

Table 4. US unemployment data. Results on model selection and out-of-sample prediction.

	Proposed method					
	degree=1	degree=2	degree=3	MARS	BRUTO	Linear AR
Selected lags	1,2	1,2	1,2	1,2,4,5	1,2,4,5,8	1,2,4,5,8
R^2 (in sample)	0.876	0.874	0.878	0.892	0.864	0.864
MSPE	0.023	0.031	0.031	0.024	0.057	0.058
MAPE	0.122	0.125	0.128	0.132	0.159	0.161

Acknowledgment

Appendix: Proofs

In this appendix, we provide the proofs of all the technical results.

Proof of Lemma 1. Suppose $\mu(x) = \mu_{s_0}(x_{s_0}) = \mu_{s'_0}(x_{s'_0})$. Write $\mu_{s_0}(x_{s_0}) = \mu_0 + \sum_{i \in s_0} \mu_i(x_i)$ and $\mu_{s'_0}(x_{s'_0}) = \mu_0 + \sum_{i' \in s'_0} \mu_{i'}(x_i)$, where $E[\mu_i(X_{t,i})] = 0$ and $E[\mu_{i'}(X_{t,i})] = 0$. Using the same argument as in the proof of Lemma 3.2 of Stone (1994), we can show that μ should have a unique representation as an element of $\mathbb{H}_{\{1,\ldots,d\}}$. Therefore, $\mu_{s_0} = \mu_{s'_0}$, a.s.
□

Proof of Lemma 2. Set $\rho_{n,i} = \inf_{g \in \mathbb{G}_i} \|g - \mu_i\|_\infty$ for $i \in s_0$. Then $\rho_{s_0} \leq \sum_{i \in s_0} \rho_{n,i}$. According to Theorem XII.1 of de Boor (1978, page 170), $\rho_{n,i} \lesssim J_i^{-2} = n^{-2\gamma}$, $i \in s_0$, and thus $\rho_{s_0} \lesssim n^{-2\gamma}$. As a result, if $\gamma \geq 1/5$, then (A4) holds. □

We now state some useful results to facilitate our proof of Theorem 1. Let us first introduce two inner-products as in Huang (1998). Define the theoretical inner product by $\langle f, g \rangle = E[f(X_t)g(X_t)\mathrm{I}(X_t \in \mathbb{C})]$ for square-integrable functions f and g and denote the theoretical norm by $\|g\|^2 = \langle g, g \rangle$. Similarly, define the empirical inner product by $\langle f, g \rangle_n = (1/n) \sum_{t=1}^n [f(X_t)g(X_t)\mathrm{I}(X_t \in \mathbb{C})]$ and denote the corresponding empirical norm by $\|g\|_n^2 = \langle g, g \rangle_n$. Let $Y(\cdot)$ denote a function on $\mathbb{C}$ interpolating the observed values (X_t, Y_t), that is, it satisfies $Y(X_t) = Y_t$. Then the least squares estimate $\widehat{\mu}_s$ is the orthogonal projection of $Y(\cdot)$ on $\mathbb{G}_s$ relative to the empirical inner product. Let $\mathrm{Proj}_{s,n}$ and Proj_s denote respectively the orthogonal projection onto $\mathbb{G}_s$ and $\mathbb{H}_s$ relative to the theoretical inner product. Denote $\mu_{s,n}^* = \mathrm{Proj}_{s,n} \mu$ and $\mu_s^* = \mathrm{Proj}_s \mu$. Set $N_s = \dim(\mathbb{G}_s)$ and $\rho_s = \inf_{g \in \mathbb{G}_s} \|g - \mu_s^*\|_\infty$. The following results are proved in Huang (1998, 2002).

LEMMA A.1. *Under Assumptions* (A2) *and* (A3), $\sup_{g \in \mathbb{G}_{\{1,\ldots,d\}}} |\|g\|_n^2 / \|g\|^2 - 1| = o_P(1)$.

LEMMA A.2. *Under Assumptions* (A1)–(A3), $\|\widehat{\mu}_s - \mu_{s,n}^*\|_n + \|\widehat{\mu}_s - \mu_{s,n}^*\| = O_P(\sqrt{N_s/n})$ *and* $\|\mu_{s,n}^* - \mu_s^*\|_n + \|\mu_{s,n}^* - \mu_s^*\| = O(\rho_s)$.

We now give a formal characterization of under-fitting. For $s \in \{1, \ldots, d\}$, denote $c(s, \mu) = \|\mathrm{Proj}_s \mu - \mu\|$. Since $\mu \in \mathbb{H}_{s_0}$, $\mathrm{Proj}_{s_0} \mu = \mu$ and thus $c(s_0, \mu) = 0$.

LEMMA A.3. *If s under-fits, then $c(s, \mu) > 0$.*

Proof of Lemma A.3. Under-fitting means that $s_0 \not\subset s$ or equivalently $s \cap s_0 \neq s_0$. We consider two cases: (i) $s \cap s_0 = s$ and (ii) $s \cap s_0 \neq s$.

Case (i). It is necessary that $s \subset s_0$ and $s \neq s_0$. If $c(s, \mu) = \| \text{Proj}_s \mu - \mu \| = 0$, then $\mu = \text{Proj}_s \mu \in \mathbb{H}_s$, which contradicts with the minimal property of s_0. Thus $c(s, \mu) > 0$.

Case (ii). Note that $s \cap s_0 \subset s_0$ and $s \cap s_0 \neq s_0$. If $c(s, \mu) = \| \mu - \text{Proj}_s \mu \| = 0$, then $\mu = \text{Proj}_s \mu \in \mathbb{H}_s \cap \mathbb{H}_{s_0} = \mathbb{H}_{s \cap s_0}$, which contradicts with the minimal property of s_0. Thus $c(s, \mu) > 0$. □

Proof of Theorem 1. We show that, for any s such that $s \neq s_0$, $\lim_{n \to \infty} P(\text{BIC}_s > \text{BIC}_{s_0}) = 1$. By the law of large numbers for stationary processes,

$$\frac{1}{n} \sum_{t=1}^{n} [\{Y_t - \mu(X_t)\}^2 \text{I}(X_t \in \mathbb{C})] \to \sigma_0^2 = E[\{Y_0 - \mu(X_0)\}^2 \text{I}(X_t \in \mathbb{C})], \qquad n \to \infty.$$

It follows from Lemma A.2 and (A4) that $\|\widehat{\mu}_{s_0} - \mu\|_n = o_P(1)$. Hence,

$$\text{MSE}_{s_0} = \frac{1}{n} \sum_{t=1}^{n} [\{Y_t - \widehat{\mu}_{s_0}(X_t)\}^2 \text{I}(X_t \in \mathbb{C})] = \sigma_0^2 (1 + o_P(1)).$$

Over-fitting. We first consider over-fitting. Suppose $s \supset s_0$ and $s \neq s_0$. Using the orthogonal projection properties of $\widehat{\mu}_s$ and $\widehat{\mu}_{s_0}$ and applying Lemma A.1, $\text{MSE}_{s_0} - \text{MSE}_s = \|\widehat{\mu}_s - \widehat{\mu}_{s_0}\|_n^2 = \|\widehat{\mu}_s - \widehat{\mu}_{s_0}\|^2 (1 + o_P(1))$. Since $\mu_s^* = \mu_{s_0}^* = \mu$ and $\mathbb{G}_s \supset \mathbb{G}_{s_0}$, $\rho_s = \inf_{g \in \mathbb{G}_s} \|g - \mu_s^*\|_\infty \leq \inf_{g \in \mathbb{G}_{s_0}} \|g - \mu_{s_0}^*\|_\infty = \rho_{s_0}$. Note that $N_s \asymp J_n \asymp N_{s_0}$. It follows from Lemma A.2 and (A4) that $\|\widehat{\mu}_s - \widehat{\mu}_{s_0}\| \leq \|\widehat{\mu}_s - \mu_s^*\| + \|\widehat{\mu}_{s_0} - \mu_{s_0}^*\| = O_P(\sqrt{J_n/n})$. Thus, $(\text{MSE}_{s_0} - \text{MSE}_s) / \text{MSE}_{s_0} = O_P(J_n/n) = o_P(1)$. Therefore,

$$\begin{aligned} \text{BIC}_s - \text{BIC}_{s_0} &= \log\left(1 + \frac{\text{MSE}_s - \text{MSE}_{s_0}}{\text{MSE}_{s_0}}\right) + \frac{N_s - N_{s_0}}{n} \log n \\ &= \frac{\text{MSE}_s - \text{MSE}_{s_0}}{\text{MSE}_{s_0}} (1 + o_P(1)) + \frac{N_s - N_{s_0}}{n} \log n \geq -O_P\left(\frac{J_n}{n}\right) + \frac{J_n}{n} \log n. \end{aligned}$$

Consequently, $\lim_{n \to \infty} P(\text{BIC}_s - \text{BIC}_{s_0} > 0) = 1$.

Under-fitting. It is necessary that $s \cap s_0 \neq s_0$ for under-fitting. According to Lemma A.3, $c(s, \mu) > 0$. We consider two cases: (i) $s \cap s_0 = s$ and (ii) $s \cap s_0 \neq s$. We shall show that, for both cases, $\text{MSE}_s - \text{MSE}_{s_0} \geq c^2(s, \mu) + o_P(1)$. As a consequence,

$$\begin{aligned} \text{BIC}_s - \text{BIC}_{s_0} &= \log\left(1 + \frac{\text{MSE}_s - \text{MSE}_{s_0}}{\text{MSE}_{s_0}}\right) + \frac{N_s - N_{s_0}}{n} \log n \\ &\geq \log\left(1 + \frac{c^2(s, \mu) + o_P(1)}{\sigma_0^2}\right) + o_P(1), \end{aligned}$$

which implies that $\lim_{n \to \infty} P(\text{BIC}_s - \text{BIC}_{s_0} > 0) = 1$.

Case (i). Suppose $s \cap s_0 \neq s_0$ and $s \cap s_0 = s$. Thus $s \subset s_0$. Using the orthogonal projection properties of $\widehat{\mu}_s$ and $\widehat{\mu}_{s_0}$ and applying Lemma A.1, $\text{MSE}_s - \text{MSE}_{s_0} = \|\widehat{\mu}_s - \widehat{\mu}_{s_0}\|_n^2 = \|\widehat{\mu}_s - \widehat{\mu}_{s_0}\|^2 (1 + o_P(1))$. Recall $\mu_{s,n}^* = \text{Proj}_{s,n} \mu$ and $\mu \in \mathbb{H}_{s_0}$. It follows from Lemma A.2 and (A4) that $\|\widehat{\mu}_s - \mu_{s,n}^*\| = o_P(1)$ and $\|\widehat{\mu}_{s_0} - \mu\| = o_P(1)$. Thus, by the triangle inequality, $\|\widehat{\mu}_s - \widehat{\mu}_{s_0}\| \geq \|\mu_{s,n}^* - \mu\| - \|\widehat{\mu}_s - \mu_{s,n}^*\| - \|\widehat{\mu}_{s_0} - \mu\| \geq \|\mu_{s,n}^* - \mu\| - o_P(1)$. Since $\mathbb{G}_s \subset \mathbb{H}_s$, $\|\mu_{s,n}^* - \mu\| = \|\text{Proj}_{s,n} \mu - \mu\| \geq \|\text{Proj}_s \mu - \mu\| = c(s, \mu) > 0$. Hence, $\text{MSE}_s - \text{MSE}_{s_0} \geq c(s, \mu)^2 + o_P(1)$.

Case (ii). Suppose $s \cap s_0 \neq s_0$ and $s \cap s_0 \neq s$. Let $s \cap s_0 = s'$. By the properties of the orthogonal projections, $\mathrm{MSE}_s - \mathrm{MSE}_{s'} = -\|\widehat{\mu}_s - \widehat{\mu}_{s'}\|_n^2$ and $\mathrm{MSE}_{s'} - \mathrm{MSE}_{s_0} = \|\widehat{\mu}_{s'} - \widehat{\mu}_{s_0}\|_n^2$. Combining these two equations and applying Lemma A.1 we obtain that $\mathrm{MSE}_s - \mathrm{MSE}_{s_0} = \|\widehat{\mu}_{s'} - \widehat{\mu}_{s_0}\|^2 - \|\widehat{\mu}_s - \widehat{\mu}_{s'}\|^2 + o_P(1)$ By Lemma A.2 and (A4), $\|\widehat{\mu}_s - \mu_{s,n}^*\| = o_P(1)$, $\|\widehat{\mu}_{s'} - \mu_{s',n}^*\| = o_P(1)$, and $\|\widehat{\mu}_{s_0} - \mu_{s_0,n}^*\| = o_P(1)$. Thus, it follows from the triangle inequality that $\|\widehat{\mu}_s - \widehat{\mu}_{s'}\| \leq \|\mu_{s,n}^* - \mu_{s',n}^*\| + \|\widehat{\mu}_s - \mu_{s,n}^*\| + \|\widehat{\mu}_{s'} - \mu_{s',n}^*\| = \|\mu_{s,n}^* - \mu_{s',n}^*\| + o_P(1)$ and $\|\widehat{\mu}_{s'} - \widehat{\mu}_{s_0}\| \geq \|\mu_{s'}^* - \mu_{s_0,n}^*\| - \|\widehat{\mu}_{s'} - \mu_{s',n}^*\| - \|\widehat{\mu}_{s_0} - \mu_{s_0,n}^*\| \geq \|\mu_{s',n}^* - \mu_{s_0,n}^*\| - o_P(1)$. Therefore

$$\mathrm{MSE}_s - \mathrm{MSE}_{s_0} \geq \|\mu_{s',n}^* - \mu_{s_0,n}^*\|^2 - \|\mu_{s,n}^* - \mu_{s',n}^*\|^2 + o_P(1). \tag{A.1}$$

Since $\mu_{s,n}^*$, $\mu_{s_0,n}^*$, and $\mu_{s',n}^*$ are orthogonal projections onto $\mathbb{G}_s$, $\mathbb{G}_{s_0}$, and $\mathbb{G}_{s'}$ respectively, $\|\mu_{s',n}^* - \mu_{s_0,n}^*\|^2 = \|\mu - \mu_{s',n}^*\|^2 - \|\mu - \mu_{s_0,n}^*\|^2$ and $\|\mu_{s,n}^* - \mu_{s',n}^*\|^2 = \|\mu - \mu_{s',n}^*\|^2 - \|\mu - \mu_{s,n}^*\|^2$ Thus

$$\|\mu_{s',n}^* - \mu_{s_0,n}^*\|^2 - \|\mu_{s,n}^* - \mu_{s',n}^*\|^2 = \|\mu - \mu_{s,n}^*\|^2 - \|\mu - \mu_{s_0,n}^*\|^2. \tag{A.2}$$

Since $\mathbb{G}_s \subset \mathbb{H}_s$,

$$\|\mu - \mu_{s,n}^*\| = \|\mu - \mathrm{Proj}_{s,n}\mu\| \geq \|\mu - \mathrm{Proj}_s\mu\| = c(s, \mu). \tag{A.3}$$

On the other hand, $\|\mu - \mu_{s_0,n}^*\| = \rho_{s_0} = o(1)$ by Condition (A4). Therefore, combining (A.1)–(A.3), we obtain that $\mathrm{MSE}_s - \mathrm{MSE}_{s_0} \geq c^2(s, \mu) + o_P(1)$. $\square$

References

Akaike, H. (1969) Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, **21**, 243–247.

Akaike, H. (1970) Statistical Predictor Identification. *Annals of the Institute of Statistical Mathematics*, **22**, 203–217.

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

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models (with discussion). *The Annals of Statistics*, **17**, 453–555.

Chen, R. and Tsay, R. S. (1993) Nonlinear additive ARX models. *Journal of the American Statistical Association*, **88**, 955–967.

Cheng, B. and Tong, H. (1992) On consist non-parametric order determination and chaos (with discussion). *Journal of the Royal Statistical Society*, **B, 54**, 427–474.

de Boor, C. (1978) *A Practical Guide to Splines.* Springer-Verlag, New York.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, **19** 1–141.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989) *Nonparametric Curve Estimation from Time Series.* Springer-Verlag, New York, Heidelberg.

Hastie, T. J. (1989). Discussion of "Flexible Parsimonious Smoothing and Additive Modeling" by J. Friedman and B. Silverman, *Technometrics*, **31** 23-29.

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

Huang, J. Z. (1998) Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, **26**, 242–272.

Huang, J. Z. (2001) Concave extended linear modeling: A theoretical synthesis. *Statistica Sinica*, **11**, 173–197.

Huang, J. Z. (2002) The use of polynomial splines in nonlinear time series modeling. Technical Report, Department of Statistics, University of Pennsylvania.

Lewis, P. A. W. and Stevens, J. G. (1991) Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *Journal of the American Statistical Association*, **86**, 864–877.

Linton, O. and Nielsen, J. P. (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93–100.

Mammen, E. Linton, O. and Nielsen, J. (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics,* **5**, 1443–1490.

Masry, E., and Tjøstheim, D. T. (1997) Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**, 214–252

Miller, A. *Subset Selection in Regression*, 2nd ed. Chapman & Hall/CRC, Boca Raton, Florida.

Rao, C. R. and Wu, Y. (1989) A strong consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369–374.

Rech, G., Teräsvirta, T. and Tschernig, R. (2001) A simple variable selection technique for nonlinear models. *Communications in Statistics, Part A — Theory and Methods*, **30**, 1227–1241

Robinson, P. M. (1983) Nonparametric estimators for time series. *Journal of Time Series Analysis*, **4**, 185–207.

Stone, C. J. (1985) Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689–705.

Stone, C. J. (1994) The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *The Annals of Statistics*, **22**, 118–184.

Stone, C. J., Hansen, M. H., Kooperberg, C. and Truong, Y. K. (1997) Polynomial splines and their tensor products in extended linear modeling (with discussion), *The Annals of Statistics*, **25**, 1371–1470.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Tjøstheim, D. and Auestad, B. (1990) Identification of Nonlinear Time Series: First-Order Characterization and Order Determination. *Biometrika*, **77**, 669–687.

Tjøstheim, D. and Auestad, B. (1994a) Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association,* **89**, 1398–1409.

Tjøstheim, D. and Auestad, B. (1994b). Nonparametric identification of nonlinear time series: selecting significant lags. *Journal of the American Statistical Association*, **89**, 1410–1419.

Tschernig, R. and Yang, L. (2000) Nonparametric lag selection for time series. *Journal of Time Series Analysis*, **21**, 457–487.

Vieu, P. (1994) Order choice in nonlinear autoregressive models. *Statistics*, **24**, 1–22.

Yang, L., Härdle, W. and Nielsen, J. P. (1999) Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis*, **20**, 579–604.

Yang, L. and Tschernig, R. (1999) Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society*, **B, 61**, 793–815.

Yang, L. and Tschernig, R. (2002) Non- and semiparametric identification of seasonal nonlinear autoregression models. *Econometric Theory*, **18**, 1408-1448.

Yao, Q. and Tong, H. (1994) On subset selection in non-parametric stochastic regression. *Statistica Sinica*, **4**, 51–70.