Journal of Nonparametric Statistics, 2016
Vol. 28, No. 3, 595–616, http://dx.doi.org/10.1080/10485252.2016.1191633

Variable selection for additive model via cumulative ratios of empirical strengths total

Miao Yang[a], Lan Xue[a] and Lijian Yang[b*]

[a] *Department of Statistics, Oregon State University, Corvallis, OR, USA;* [b] *Center for Statistical Science & Department of Industrial Engineering, Tsinghua University, Beijing, People's Republic of China*

(*Received 18 May 2015; accepted 23 April 2016*)

We propose a data-driven method to select significant variables in additive model via spline estimation. The additive structure of the regression model is imposed to overcome the 'curse of dimensionality', while the spline estimators provide a good approximation to the additive components of the model. The additive components are ordered according to their empirical strengths, and the significant variables are chosen at the first crossing of a predetermined threshold by the Cumulative Ratios of Empirical Strengths Total of the components. Consistency of the proposed method is established when the number of variables are allowed to diverge with sample size, while extensive Monte-Carlo study demonstrates superior performance of the proposed method and its advantages over the BIC method of Huang and Yang [(2004), 'Identification of Nonlinear: Additive Autoregressive Models', *Journal of the Royal Statistical Society Series B*, 66, 463–477] in terms of speed and accuracy.

Keywords: additive model; B-spline; Cumulative Ratios of Empirical Strengths Total (CUREST); lag selection; variable selection

AMS Subject Classification: 62G08; 62G10; 62G20

1. Introduction

Variable selection is widely used in diverse areas such as genetics, digital imaging processing, functional data and financial data. Consequently, it has become a topic of high priority in statistics research, especially in the analysis of high-dimensional and time series data.

Statistical literature contains numerous procedures on variable selection for linear and other parametric models, and many classical approaches, such as Akaike Information Criterion, Final Prediction Error and Bayesian Information Criterion (BIC), are well established, see for instance, Akaike (1969, 1970) and Schwarz (1978). However, their proper use is restricted to data that fit the specific model structure. Nonparametric method can be applied to avoid the bias of misspecified parametric model structure, see Cheng and Tong (1992), Yao and Tong (1994), Tjøstheim and Auestad (1994b), Tschernig and Yang (2000) and Yang and Tschernig (2002) for nonparametric time series lag selection. General nonparametric models, however, suffer the 'curse of dimensionality', i.e. the inaccuracy of estimating multivariate nonparametric function based on a moderate number of observations.

*Corresponding author. Email: yanglijian@mail.tsinghua.edu.cn

Additive form of the regression function, introduced by Stone (1985), is a semiparametric tool to effectively circumvent the curse of dimensionality, which became popularised with the publication of Hastie and Tibshirani (1990). There has been a vast array of influential publications on the subject of additive model for the last two decades, see for instance, Buja, Hastie, and Tibshirani (1989), Friedman (1991), Lewis and Stevens (1991), Tjøstheim and Auestad (1990, 1994a), Chen and Tsay (1993), Linton and Nielsen (1995), Masry and Tjøstheim (1997), Fan, Härdle, and Mammen (1998), Mammen, Linton, and Nielsen (1999), Yang, Härdle, and Nielsen (1999), Yang, Park, Xue, and Härdle (2006), Wang and Yang (2007), Yu, Park, and Mammen (2008), Mammen, Støve, and Tjøstheim (2009), Wang and Yang (2009), Jiang, Fan, and Fan (2010), Lee, Mammen and Park (2010), Liu and Yang (2010), Song and Yang (2010), Ma and Yang (2011), Ma (2012), and Liu, Yang, and Härdle (2013). These works span different approaches (spline, kernel, etc.) and strategies (one- or two-step, or iterative, etc.), and have various advantages/disadvantages that have been well known and understood.

One important aspect of additive modelling, as for other semiparametric modelling, is the selection of significant variables. Huang and Yang (2004) extended the BIC method of time series lag selection Schwarz (1978) by combining it with spline smoothing to select significant variables (or lags) in nonlinear additive autoregressive (AR) model. Other well known works on additive model variable selection include Härdle and Korostelev (1996), Xue (2009), Huang, Horowitz, and Wei (2010), Xue, Qu, and Zhou (2010), Chen, Liang, and Wang (2011), Fan, Feng, and Song (2011), Liu, Wang and Liang (2011), Wang, Liu, Liang, and Carroll (2011), Jiang and Xue (2013), Ma, Song, and Wang (2013), to name a few.

Huang and Yang (2004) rigorously established consistency of their spline BIC method for fixed number of covariates. Lian (2014) adopted the extended BIC (Chen and Chen 2008) to high-dimensional additive models and established selection consistency when the number of variables is of ultra-high dimensionality. For both methods, the asymptotics are established by searching over all possible subsets of variables. However, the implementation was conducted by a directed search of stepwise addition and deletion to avoid the computationally infeasible task of full subset search.

We propose to remedy this major gap between theory and implementation by selecting variables in additive model according to their empirical strengths, i.e. their averaged squared estimated function values. By approximating the additive components via B-spline, the collection of CUmulative Ratios of Empirical Strengths Total (CUREST) of the components is a good indicator to identify significant variables through a predefined threshold value. Our method is computationally expeditious. To appreciate the difference, to select significant variables out of a pool of 20, the full search of BIC (or extended BIC) needs to perform $2^{20} = 1,048,576$ additive spline regressions, the stepwise procedure needs $20 \times 21 = 420$ such regressions, while CUREST approach only needs one such regression. In addition, the CUREST approach is adaptive to nonlinear models and selection consist for high-dimensional additive model, hence overcomes all the aforementioned problems.

The CUREST method takes the spline approach of Huang and Yang (2004) with all computational advantages, and improves on the thresholding idea in Härdle and Korostelev (1996). At a first look, the CUREST method highly resembles the nonparametric independence screening (NIS) in Fan et al. (2011), but careful examination reveals a fundamental difference: NIS assumes that the set of significant variables is independent of the set of insignificant ones, and hence carries out spline smoothing on each regressor one at a time; the CUREST method does not need independence assumption and performs merely one spline smoothing on the whole set of regressors. Thus the advantage of CUREST over NIS is that it is capable of zeroing in on a set of significant variables not necessarily independent of the rest; the advantage of NIS over CUREST being that it allows for the full dimension growing to infinity much faster than sample size, as long as the significant variables are independent of the others and of a fixed small number.

Penalised polynomial spline method in Xue (2009), Huang et al. (2010), and Xue et al. (2010), also uses polynomial splines to approximate nonparametric functions, and conducts variable selection through the penalisation on the L_2 norm of each additive component. The computation of such penalised polynomial spline estimator is more challenging due to non-smoothness of the penalty functions and often obtained by an iterative procedure such as the local linear or local quadratic algorithms. However, the proposed CUREST approach is much simpler to compute and only needs to solve one least squared problem. Similar to NIS, the advantage of the penalised polynomial spline method is that it adapts automatically to high-dimensional cases and allows the dimensionality to grow much faster with the sample size.

The rest of the article is organised as follows. In Section 2 we describe the additive stochastic regression model which includes the classic iid additive regression as well as the additive AR. Section 3 introduces spline estimators of additive components of the model. Section 4 proposes the variable selection rule and states the main theoretical results. Section 5 provides implementation details of the proposed method, while Monte-Carlo results are reported in Section 6. All technical assumptions and proofs are given in the appendix.

2. The model

Let $\{(Y_t, \mathbf{X}_t)\}_{t=1}^n$ be a sequence of stationary observations, with univariate response Y_t and d_n predictors $\mathbf{X}_t = (X_{t1}, \dots, X_{td_n})$. We allow diverging number of predictors with $d_n \to \infty$ as $n \to \infty$. With unknown conditional mean function $m(\mathbf{X}_t) = \mathsf{E}(Y_t \mid \mathbf{X}_t)$, the observations satisfy

$$Y_t = m(\mathbf{X}_t) + \varepsilon_t, \quad t = 1, \dots, n, \tag{1}$$

where $\{\varepsilon_t\}_{t=1}^n$ are uncorrelated white noises with $\mathsf{E}(\varepsilon_t \mid \mathbf{X}_t) = 0$, and ε_t is independent of the σ-field $\mathcal{F} = \sigma(\mathbf{X}_{t'},\ t' \le t)$ for $t = 1, \dots, n$. No additional assumption about ε_t is specified. It can either be conditional homoscedastic ($\mathsf{Var}(\varepsilon_t \mid \mathbf{X}_t)$ is a constant) or conditional heteroscedastic ($\mathsf{Var}(\varepsilon_t \mid \mathbf{X}_t)$ is not a constant). The variables $\mathbf{X}_t$ can consist of either exogenous variables or lagged values of Y_t.

For the additive model (Hastie and Tibshirani 1990), the regression function takes the form

$$m(\mathbf{x}) = m_0 + \sum_{l=1}^{d_n} m_l(x_l), \tag{2}$$

where m_0 is an unknown constant and $\{m_l(\cdot)\}_{l=1}^{d_n}$ are unknown nonparametric functions. For model identification purpose, we assume that $\mathsf{E}\{m_l(X_{tl})\} = 0$ for $1 \le l \le d_n$ in Equation (2).

In the literature of nonparametric smoothing, estimation of the functions $\{m_l(x_l)\}_{l=1}^{d_n}$ is often conducted on compact sets. Without loss of generality, we consider the estimation of each $m_l(\cdot)$ on $\mathcal{C}_l = [0, 1]$, for $l = 1, \dots, d_n$. Following Stone (1985) and Huang (1998), denote the additive model space

$$\mathbb{H} = \left\{ h(\mathbf{x}) = h_0 + \sum_{l=1}^{d_n} h_l(x_l), h_0 \text{ is a constant},\ \mathsf{E}\{h_l(X_l)\} = 0,\ \mathsf{E}\{h_l(X_l)\}^2 < +\infty \right\}.$$

Then the regression function m in Equation (2) is the minimiser of $\mathsf{E}\{Y_t - h(\mathbf{X}_t)\}^2$ over all $h \in \mathbb{H}$.

In what follows, denote by E_n the empirical expectation $\mathsf{E}_n(\varphi) = n^{-1} \sum_{t=1}^n \varphi(\mathbf{X}_t)$. We introduce two inner products on $\mathbb{H}$. For functions $\varphi,\ \phi \in \mathbb{H}$, the theoretical and empirical inner products are defined, respectively, as $\langle \varphi, \phi \rangle = \mathsf{E}\{\varphi(\mathbf{X})\phi(\mathbf{X})\}$ and $\langle \varphi, \phi \rangle_{2,n} = \mathsf{E}_n\{\varphi(\mathbf{X})\phi(\mathbf{X})\}$. The corresponding induced norms are $\|\varphi\|_2^2 = \mathsf{E}\{\varphi^2(\mathbf{X})\}$ and $\|\varphi\|_{2,n}^2 = \mathsf{E}_n\{\varphi^2(\mathbf{X})\}$.

3. Additive spline estimation

Spline method is applied to estimate unknown components in the regression function m. Let N_n denote a positive integer and let $p > 1$ be an integer. For each $l = 1, \ldots, d_n$, let $0 = \xi_{l0} < \xi_{l1} < \cdots < \xi_{l,N_n} < \xi_{l,N_n+1} = 1$ be a knot sequence and $\xi_{l1}, \ldots, \xi_{l,N_n}$ are called interior knots. The polynomial spline space $\mathbb{G}_l$ consists of functions that are polynomial with degree $p-1$ (or less) on the intervals $[\xi_{li}, \xi_{l,i+1}),\ i = 0, \ldots N_n - 1$ and $[\xi_{l,N_n}, \xi_{l,N_n+1}]$, and overall it is $p-2$ times continuously differentiable on $[0, 1]$. The functions in $\mathbb{G}_l$ are called polynomial splines. A polynomial spline with $p = 2$,3,4 is a piecewise linear, quadratic, cubic function, respectively. See de Boor (2001), Chapter IX for details. The space $\mathbb{G}_l$ is determined by the polynomial degree $p-1$ and the knot sequence $\{\xi_{li}\}_{i=0}^{N_n+1}$. Let $\mathbb{G}_l^0 = \{g_l \in \mathbb{G}_l, \mathsf{E}_n(g_l) = 0\}$, where $\mathsf{E}_n(g_l) = \sum_{i=1}^n g_l(X_{il})/n = 0$. It is the space of empirically centred polynomial splines. Then define the estimation space

$$\mathbb{G} = \left\{ g(\mathbf{x}) = g_0 + \sum_{l=1}^{d_n} g_l(x_l), g_0 \text{ is a constant and } g_l \in \mathbb{G}_l^0 \right\},$$

which is a direct sum of the space of constants and $\mathbb{G}_l^0,\ 1 \le l \le d_n$. Intuitively any $m \in \mathbb{H}$ can be approximated by an additive spline function in $\mathbb{G}$, which leads to the following least squares estimation.

Given observations $(\mathbf{X}_t, Y_t)$, $t = 1, \ldots, n$, from Equation (1), the estimator of the unknown regression function m is defined as its best approximation from $\mathbb{G}$ by minimising the sum of squared errors

$$\hat{m}(\mathbf{x}) = \arg\min_{g \in \mathbb{G}} \sum_{t=1}^{n} \{Y_t - g(\mathbf{X}_t)\}^2. \tag{3}$$

It is equivalent to minimise the sum of squared errors with respect to $\gamma = \{\gamma_0, \gamma_{lj},\ 2 - p \le j \le N_n,\ 1 \le l \le d_n\}$

$$\hat{\gamma} = \arg\min_{\gamma} \sum_t \left\{ Y_t - \gamma_0 - \sum_{l=1}^{d_n} \sum_{j=2-p}^{N_n} \gamma_{lj} B_{lj}(X_{tl}) \right\}^2, \tag{4}$$

where $\{B_{lj}\}_{j=2-p}^{N_n}$ is a set of basis of the spline space $\mathbb{G}_l^0$. For example, one can use the truncated power basis of form $\{x_l, \ldots, x_l^{p-1}, (x_l - \xi_{l1})_+^{p-1}, \ldots, (x_l - \xi_{lN_n})_+^{p-1}\}$, in which $(x)_+^p = (x_+)^p$. Then the estimator of the regression function m is given by

$$\hat{m}(\mathbf{x}) = \hat{m}_0 + \sum_{l=1}^{d_n} \hat{m}_l(x_l), \tag{5}$$

where $\hat{m}_l(x_l) = \sum_{j=2-p}^{N_n} \hat{\gamma}_{lj}[B_{lj}(x_l) - \mathsf{E}_n\{B_{lj}(X_l)\}]$, and $\hat{m}_0 = \hat{\gamma}_0 + \sum_{j=2-p}^{N_n} \hat{\gamma}_{lj}\mathsf{E}_n\{B_{lj}(X_l)\}$. The estimators $\{\hat{m}_l(x_l)\}_{l=1}^{d_n}$ are empirically centred to consistently estimate the theoretically centred function components in Equation (2).

In the following, for positive numbers b_n and c_n, $n \ge 1$, $b_n \asymp c_n$ means b_n and c_n have the same order, $b_n \gg c_n$ means $b_n/c_n \to \infty$, and $b_n \le c_n$ means $b_n/c_n \to 0$. To establish the theoretical results, we need the following assumptions.

(A1) The vector stochastic process $\{(\mathbf{X}_t, Y_t)\}_{t=-\infty}^{\infty}$ is stationary and geometrically strong mixing. That is, there exist constants $c_1 > 0$ and $0 < \rho < 1$, such that $\alpha(n) \le c_1\rho^n$ for all n, with the α-mixing coefficient of $\{(\mathbf{X}_t, Y_t)\}_{t=-\infty}^{\infty}$ defined as

$$\alpha(n) = \sup\{P(B \cap C) - P(B)P(C) : B \in \sigma(\{(\mathbf{X}_t, Y_t),\ t \leq 0\}),\\ C \in \sigma(\{(\mathbf{X}_t, Y_t),\ t \geq n\})\}.$$

(A2) The noise ε_t satisfies $\mathsf{E}(\varepsilon_t \mid \mathbf{X}_t) = 0$ and $\sup_{\mathbf{x}\in[0,1]^{d_n}} \mathsf{E}\{|\varepsilon_t|^{2+\nu} \mid \mathbf{X}_t = \mathbf{x}\} < \infty$ for some $\nu > 0$.

(A3) The joint density of $\mathbf{X}$, denoted by $f_{\mathbf{X}}$, satisfies $0 < M_1^{-1} \leq f_{\mathbf{X}}(\mathbf{x}) \leq M_2 < \infty$ for some positive constants $M_1, M_2 > 1$ and all $\mathbf{x} \in [0,1]^{d_n}$. Denote $2\varepsilon_1 = 1 - (1 - M_1^{-1}M_2^{-2})^{1/2}$.

(A4) The functions $\{m_l\}_{l=1}^{d_n}$ belong to a class of functions $\mathcal{F}$, whose $(p-1)$th derivative exists and is Lipschitz continuous of order 1. That is, $\mathcal{F} = \{\varphi : |\varphi^{(p-1)}(s) - \varphi^{(p-1)}(t)| \leq K|s - t|,\ \text{for } s, t \in [0,1]\}$, for some positive constant K.

(A5) For the d_n sets of knots $\{0 = \xi_{l0} < \xi_{l1} < \cdots < \xi_{l,N_n} < \xi_{l,N_n+1} = 1\}_{l=1}^{d_n}$, there exists $c_2 > 0$ such that

$$\max_{1\leq l\leq d_n} \frac{\max(\xi_{l,j+1} - \xi_{lj},\ j = 0, \ldots N_n)}{\min(\xi_{l,j+1} - \xi_{lj},\ j = 0, \ldots N_n)} \leq c_2.$$

(A6) As $n \to \infty, N_n \to \infty, d_n \to \infty$, and $d_n^3 N_n^3 \log^2(n)/n\varepsilon_1^{2d_n} \to 0$, $d_n/\varepsilon_1^{d_n} N_n^p \to 0$.

Remark 3.1 Assumptions (A1)–(A5) are commonly used assumptions in the nonparametric regression literature. Assumption (A1) requires that the α-mixing coefficient decays to zero at an exponential rate. Similar assumptions are assumed in Huang and Shen (2004), Huang and Yang (2004) and Xue and Yang (2006). Assumptions similar to (A2)–(A5) are also considered in Huang and Yang (2004), Jiang and Xue (2013) and Lian (2014). Assumption (A2) is a moment condition. It follows from (A2) that the conditional variance of Y_t given $\mathbf{X}_t = \mathbf{x}$ is bounded on $\mathbf{x} \in [0,1]^{d_n}$. It follows from Assumptions (A3) that the marginal destiny f_l of X_l is bounded away from zero and infinity on $[0,1]$. Assumption (A4) constrains the smoothness of the original function. Assumption (A5) requires that the distances between any adjacent knots are of the same order. In Assumption (A6), the number of predictors d_n is allowed to increase at a certain rate that depends on both the number of interior knots N_n and the sample size n.

Theorem 3.1 *Suppose Assumptions* (A1)–(A6) *hold, let* $u_n^2 = \varepsilon_1^{-d_n}(d_n^2 N_n^{-2p} + d_n N_n/n)$. *Then*

$$\|\hat{m} - m\|_{2,n}^2 + \|\hat{m} - m\|_2^2 = \mathcal{O}_P\left(d_n^2 N_n^{-2p} + \frac{d_n N_n}{n}\right),$$

$$\|\hat{m}_l - m_l\|_{2,n}^2 + \|\hat{m}_l - m_l\|_2^2 = \mathcal{O}_P(u_n^2) \quad \text{for } 1 \leq l \leq d_n.$$

Theorem 3.1 established the rate of L_2-convergence for the spline estimators $\hat{m}, \{\hat{m}_l\}_{l=1}^{d_n}$ with the number of predictors diverging with sample size n. Most of the literatures on polynomial spline regression, such as, Stone (1985), Huang (1998) and Jiang and Xue (2013) established the convergence rate only for fixed number of predictors. Theorem 3.1 also provided the basis to study the orders of empirical strengths for important variables and redundant variables respectively in Theorem 4.1.

Remark 3.2 In Theorem 3.1, if $N_n \asymp (nd_n)^{1/(2p+1)}$, then we have the best balance in the bias and variance trade-off and the asymptotic variance and the square of the asymptotic bias have the same order. We call such choice of knot number as the optimal rate of N_n.

Remark 3.3 If one takes $N_n \asymp (nd_n)^{1/(2p+1)}$, then condition (A6) requires that $d_n \ll \log(n)$.

4. Selection of significant variables

A natural question arising from fitting the full model (1) is whether the variables are all relevant. The removal of redundant variables from model (1) can effectively reduce the dimensionality and is vital for the modelling of high-dimensional time series data. Let S_0 denote the set of indices of significant variables. That is, the subset of variables $\{X_{tl},\ l \in S_0\}$ provides the same information on Y_t as $\mathbf{X}_t = (X_{t1}, \ldots, X_{td_n})$ with

$$\mathsf{E}(Y_t \,|\, X_{tl}, l \in S_0) = \mathsf{E}(Y_t \,|\, \mathbf{X}_t), \quad \text{a.s.}$$

The goal of this paper is to determine S_0, the subset of relevant variables. Our idea is based on the strength of each $m_l(x_l)$ which is defined by $\lambda_l = \mathsf{E}\{m_l(X_l)\}^2,\ l = 1, \ldots, d_n$. Thus $\lambda_l = 0$ if and only if $m_l(x_l) \equiv 0$ due to continuity of $m_l,\ l = 1, \ldots, d_n$.

Order all the strengths from the largest to smallest as $\lambda_{s_1} \geqslant \lambda_{s_2} \geqslant \cdots \geqslant \lambda_{s_{d_n}} \geqslant 0$. Here s_l is the index of the variable whose strength is lth largest among d_n variables. Define its strength ratio to the cumulative total as

$$r_l = \frac{\lambda_{s_l}}{\lambda_{s_1} + \cdots + \lambda_{s_l}}, \quad l = 1, \ldots, d_n.$$

A nice property of strength ratio to the cumulative total is that it is monotone non-increasing that, $r_1 \geqslant \cdots \geq r_{d_n} \geq 0$ and $r_l = 0$ if and only if $\lambda_{s_l} = 0$. Therefore, a variable with larger strength also has larger strength ratio to cumulative total. Compared with original strength, this ratio to cumulative total is scaled between 0 and 1. It only compares the relative size of each strength and is not effected by the absolute size of the strengths. In particular, if the true additive model only contains d_0 non-zero components, then $d_0 = \max\{l : r_l > 0\}$ and $S_0 = \{s_1, \ldots, s_{d_0}\}$.

But in practice $m_l(x_l),\ l = 1, \ldots, d_n$ are all unknown. Therefore the empirical estimates of the strengths λ_l are used instead. Let $\{\hat{m}_l\}_{l=1}^{d_n}$ be the estimate of the additive components as given in Equation (5). Define the empirical strength of variable X_l as $\hat{\lambda}_l = \mathsf{E}_n\{\hat{m}_l^2\}$ for $l = 1, \ldots, d_n$. Let $\hat{\lambda}_{q_1} \geqslant \hat{\lambda}_{q_2} \geqslant \cdots \geqslant \hat{\lambda}_{q_{d_n}} \geqslant 0$ be the ordered empirical strengths and q_l be the index of the variable whose empirical strength is l th largest among d_n variables. Define the CUREST as

$$\hat{r}_l = \frac{\hat{\lambda}_{q_l}}{\hat{\lambda}_{q_1} + \cdots + \hat{\lambda}_{q_l}}, \tag{6}$$

for $l = 1, \ldots, d_n$. Accordingly define $\hat{d}_0 = \max\{l : \hat{r}_l > a_n\}$ where $a_n > 0$ is a predefined threshold value. Then *Variable selection rule by CUREST* is to select the subset $\hat{S} = \{q_1, \ldots, q_{\hat{d}_0}\}$. Our model selection result crucially depends on the choice of a_n. In general, a larger value of a_n results a model with fewer variables. Too large value of a_n will incorrectly leave out important variables, and too small value of a_n will keep redundant variables. We will consider in detail the selection of a_n in Section 5. The following asymptotic theory provides a theoretical guidance on the choice of a_n.

THEOREM 4.1 *Suppose that Assumptions* (A1)–(A6) *hold, and the number of important variables $d_0 = |S_0|$ is fixed. Then one has* $\max_{l \in S_0} |\hat{\lambda}_l - \lambda_l| = \mathcal{O}_P(u_n)$ *and* $\max_{l \in S_0^C} \hat{\lambda}_l = \mathcal{O}_P(\varepsilon_1^{-d_n} u_n^2)$.

Theorem 4.1 shows the consistency of $\hat{\lambda}_l$ as the estimators of λ_l, $1 \leq l \leq d_n$.

THEOREM 4.2 *Under the same conditions of Theorem* 4.1. *In addition, we assume*

(A7) $\min_{l \in S_0} \lambda_l = \lambda_{\min} \geq c_3$ *for positive constant* c_3.

(A8) *The threshold a_n satisfies $a_n \to 0, \varepsilon_1^{-d_n} u_n^2 / a_n \to 0$ as $n \to \infty$.*

Then CUREST consistently selects the set of significant variables. That is, $\lim_{n\to\infty} P(\hat{S} = S_0) = 1$.

Remark 4.1 Assumption (A8) gives conditions on the orders of a_n that result in consistent model selection. In particular, if one sets $N_n \asymp (nd_n)^{1/(2p+1)}$, which is the optimal rate of N_n for fixed d_n, then the threshold $a_n \to 0$, but $a_n \gg \varepsilon_1^{-2d_n} d_n^{(2p+2)/(2p+1)} n^{-2p/(2p+1)}$.

5. Implementation

In this section, we describe the actual implementation of the method proposed. Theorem 4.2 provides theoretical guidance on the choice of a_n. But in practice, we need a data-driven procedure to select an appropriate value for a_n. We propose to select the threshold a_n through an additional BIC step. Given the data set $\{(Y_t, \mathbf{X}_t)\}_{t=1}^n$ from model (1), the proposed CUREST model selection is implemented via the following procedure.

Step 1. At each X_{tl}, evaluate the B-spline basis of degree $p-1$ (default $p=2$) where equally spaced knots are used with the number of knots J_n selected from

$$\mathcal{C} = \{J_n \in \mathbb{N} : 0.5n^{-1/(2p+1)} \le J_n \le \min(2n^{-1/(2p+1)}, (n/4-1)d_n^{-1})\}. \qquad (7)$$

And $\hat{J}_n^{\mathrm{Opt}}$ is the one minimising the BIC value with $\hat{J}_n^{\mathrm{Opt}} = \arg\min_{J_n\in\mathcal{C}} \mathsf{BIC}(J_n)$, where $\mathsf{BIC} = \log(\mathsf{MSE}) + p_n \log(n)/n$ and $\mathsf{MSE} = n^{-1}\sum_{t=1}^n \{Y_t - \hat{m}(\mathbf{X}_t)\}^2$ with $\hat{m}$ defined in Equation (3) and $p_n = d_n(\hat{J}_n^{\mathrm{Opt}} + p - 1) + 1$ is the number of estimated parameters.

Step 2. Find the estimators of each additive component $\{\hat{m}_l(x_l)\}_{l=1}^{d_n}$ and their empirical strengths $\{\hat{\lambda}_l\}_{l=1}^{d_n}$ through Equations (5) and (6). Order the variables by their empirical strengths and let q_l be the index of the variable whose empirical strength is lth largest among d_n variables.

Step 3. For variables in the order of $q_1, \ldots, q_{d_n}$, calculate their values of CUREST until the variable $q_{\hat{d}_0+1}$ whose CUREST is less than a_n, where a_n is a predefined threshold and the choice of it will be discussed in the following remark. Here $\hat{d}_0$ is the number of variables whose CUREST is larger than a_n.

Step 4. Consider models ranging from $\{q_1, \ldots, q_{\hat{d}_1}\}$ to $\{q_1, \ldots, q_{\hat{d}_2}\}$, where $\hat{d}_1 = \min_{q_l}\{\hat{\lambda}_{q_l} > 2a_n\}$, $\hat{d}_2 = \min_{q_l}\{\hat{\lambda}_{q_l} > 0.5a_n\}$. For each model, we estimate an additive model only with those variables. Calculate the BIC value of a model with variables indexed by S as $\mathsf{BIC}_S = \log(\mathsf{MSE}_S) + p_S \log(n)/n$, where $\mathsf{MSE}_S = n^{-1}\sum_{t=1}^n \{Y_t - \hat{m}_S(\mathbf{X}_t)\}^2$, p_S is the dimension of corresponding estimation space $\mathbb{G}_S$.

Step 5. Select the model with the smallest BIC value.

Remark 5.1 According to our experience, only a small number of knots is needed to give adequate approximations of most smooth functions. The range given in Equation (7) to search for the optimal knot number works well in all numerical examples in the paper. In Equation (7), the constraint that $J_n \le (n/4-1)d_n^{-1}$ ensures that the number of terms in the least square problem (4), $1 + d_n J_n$, is no greater than $n/4$, which is necessary when the sample size n is moderate and dimension d_n is high.

Remark 5.2 The selection of the threshold a_n is crucial in the implementation of CUREST. In general, a larger a_n leads to a smaller model. A too large value of a_n can erroneously delete important variables from the model, and a too small value of a_n can unnecessarily keep redundant variables. Theorem 4.2 gives useful theoretical guidance on the choice of a_n. In particular, if we set $N_n \asymp (nd_n)^{1/(2p+1)}$, the optimal rate for fixed and consider $d_n \asymp \{\log n\}^{0.75}$, which is the rate we used in our simulation, then the asymptotic order condition on a_n reduces to $a_n \to 0$ and $a_n \gg \varepsilon_1^{-2\log^{0.75}(n)}(\log(n))^{(6p+6)/(8p+4)}n^{-2p/(2p+1)}$. We have used $a_n = n^{0.1-2p/(2p+1)}(\log n)^{3/2}$. It works well in our simulation studies. This choice also works for data with different values of d_n as long as $d_n \ll \log(n)$.

6. Simulation study

In this section, we analyse some simulated data to illustrate the numerical performance of the proposed CUREST method. Since the CUREST method can be applied to both AR and regression models, we organise the simulation study according to these two kinds of models.

6.1. *AR models*

In this section, we examine seven additive AR processes similar to Huang and Yang (2004). These models contain both linear and nonlinear structures. The dynamics of these processes are described by the following equations, where ξ_t are i.i.d. $N(0,1)$ random variables.

(1) **ARI** $Y_t = 0.5Y_{t-1} + 0.4Y_{t-2} + 0.1\xi_t$.
(2) **ARII** $Y_t = -0.5Y_{t-1} + 0.4Y_{t-2} + 0.1\xi_t$.
(3) **ARIII** $Y_t = -0.5Y_{t-5} + 0.5Y_{t-9} + 0.1\xi_t$.
(4) **NLARI** Additive nonlinear AR(2) model

$$\begin{aligned} Y_t = &-0.4(3 - Y_{t-1}^2)/(1 + Y_{t-1}^2) \\ &+ 0.6\{3 - (Y_{t-2} - 0.5)^3\}/\{1 + (Y_{t-2} - 0.5)^4\} + 0.1\xi_t. \end{aligned}$$

(5) **NLARII** Additive exponential AR model

$$Y_t = \{0.4 - 2\exp(-50Y_{t-6}^2)\}Y_{t-6} + \{0.5 - 0.5\exp(-50Y_{t-8}^2)\}Y_{t-8} + 0.1\xi_t.$$

(6) **NLARIII** Additive exponential AR model with sine and cosine terms

$$\begin{aligned} Y_t = &\{0.4 - 2\cos(40Y_{t-6})\exp(-30Y_{t-6}^2)\}Y_{t-6} \\ &+ \{0.55 - 0.55\sin(40Y_{t-8})\exp(-10Y_{t-8}^2)\}Y_{t-8} + 0.1\xi_t. \end{aligned}$$

(7) **NLARIUI** $Y_t = -0.4(3 - Y_{t-1}^2)/(1 + Y_{t-1}^2) + 0.1\xi_t$.

These processes differ in shape of the conditional mean function and the lag vector. All the three linear processes are close to the border of nonstationarity. These processes are chosen to have different auto-correlation patterns, and the significant lags have different signal sizes. For sample sizes $n = 50, 100, 250, 500$, realisations of size $n + 420$ are generated and the last $n + 20$ observations are taken as the observed time series. Here, the first 400 realisations are the 'burn-in' period, the last 20 observations are used for prediction and the rest n observations are used for model estimation. We generated 500 data sets from each of the above processes and carried out lag selection for each replication using linear splines with $p = 2$. The lags were selected from $\{1, 2, \ldots, d_n\}$, where $d_n = \lceil 3\log^{0.75} n \rceil$, the smallest integer greater than $3\log^{0.75} n$. To be specific, $d_n = 9, 10, 11, 12$ when $n = 50, 100, 250, 500$, respectively.

To graphically illustrate how the CUREST works, Figure 1 plots the CUREST values of the first d_n lags from NLARI model under different sample sizes. In Figure 1, the points denote the CUREST values and the dotted line locates the threshold value a_n described in Section 5. Figure 1 clearly shows that the relevant lags (1,2) always have larger CUREST values than the irrelevant ones, and the predefined threshold a_n correctly separates two relevant lags from the irrelevant ones.

Table 1 reports the frequencies of overfitting, correct-fitting and underfitting for CUREST in all models. An outcome is defined as correct-fitting, if $\hat{\mathcal{S}} = \mathcal{S}_0$; overfitting, if $\mathcal{S}_0 \subset \hat{\mathcal{S}}$; and underfitting, if $\mathcal{S}_0 \not\subset \hat{\mathcal{S}}$. In Table 1, the 'Selection Result' columns give the frequencies of underfitting (U), correct-fitting (C) and overfitting (O) over 500 simulations. In addition, to assess the prediction accuracy of the selected model by CUREST, we also considered linear polynomial spline estimation of the full (FULL) and oracle (ORACLE) models. The full model is an AR model that contains all d_n lags and the oracle model only contains the relevant lags. The mean squared prediction error (MSPE) columns report the MSPE for CUREST model, full model and oracle model, respectively.

Table 1 shows that CUREST method performs reasonably well for all seven AR models. The percentage of correct-fitting increases as the sample size increases. When the sample size is 500, the percentage of correct-fitting is 100% or very close to 100% for all seven models. This supports the asymptotic consistency result in Theorem 4.2. Furthermore, Table 1 shows that the MSPEs of CUREST are almost always smaller than those from FULL, and close to those from ORACLE. It indicates that CUREST can not only result in a parsimonious model, but also has better prediction performance than the full model. When sample size increases to 500, the MSPEs of all three methods are close to 0.01, the variance of the random error.

Figure 2 plots several estimated functional components using CUREST and FULL in one run for ARIII and NLARI models with $n = 250$. It shows that CUREST (a and b in each plot) estimates the non-zero functional components very well, and it gives exactly zero estimates for null functional components, which verifies Theorem 3.1.

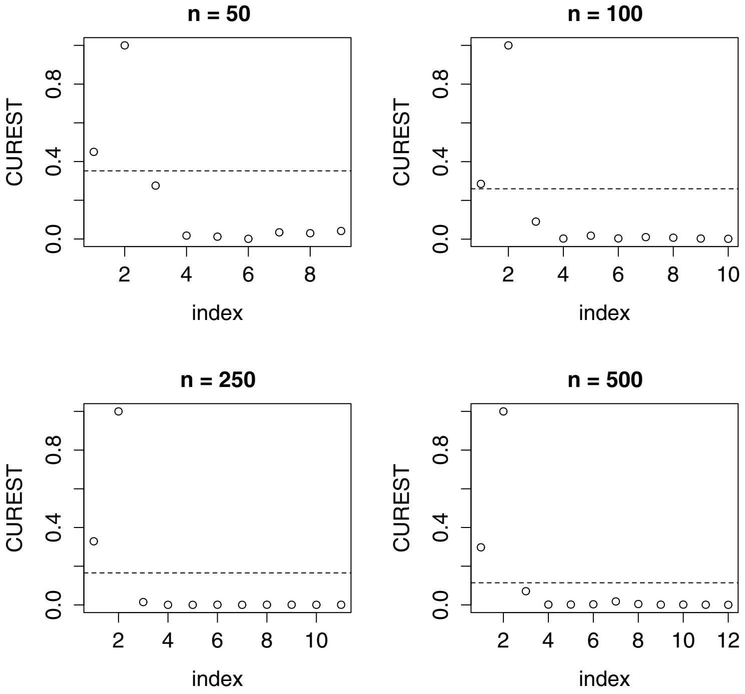

Figure 1. CUREST values in NLARI model for different sample sizes. In each plot, the dots denote the CUREST values for each lag, and the dotted line locates the predefined threshold a_n.

Table 1. Simulation results of model selection via CUREST.

		Selection result (%)			MSPE		
Model	n	U	C	O	CUREST	FULL	ORACLE
ARI	50	75.2	20.4	4.4	0.022	0.027	0.018
	100	27.0	63.4	9.6	0.013	0.022	0.012
	250	0.2	98.4	1.4	0.010	0.011	0.010
	500	0.0	99.8	0.2	0.010	0.011	0.010
ARII	50	59.6	33.8	6.6	0.014	0.016	0.012
	100	30.8	60.8	8.4	0.012	0.014	0.011
	250	0.4	98.0	1.6	0.010	0.011	0.010
	500	0.0	99.6	0.4	0.010	0.011	0.010
ARIII	50	12.8	80.0	7.2	0.012	0.015	0.011
	100	7.8	87.6	4.6	0.011	0.014	0.011
	250	0.0	100.0	0.0	0.010	0.011	0.010
	500	0.0	100.0	0.0	0.010	0.010	0.010
NLARI	50	47.4	36.8	15.8	0.043	0.036	0.030
	100	1.4	94.6	4.0	0.015	0.020	0.015
	250	0.0	100.0	0.0	0.012	0.014	0.012
	500	0.0	100.0	0.0	0.012	0.013	0.012
NLARII	50	83.0	11.0	6.0	0.017	0.020	0.016
	100	58.0	32.4	9.6	0.015	0.018	0.014
	250	15.0	84.0	1.0	0.012	0.014	0.012
	500	0.0	98.8	1.2	0.011	0.012	0.011
NLARIII	50	64.8	22.0	13.2	0.026	0.028	0.020
	100	15.8	69.0	15.2	0.019	0.024	0.018
	250	0.0	98.0	2.0	0.017	0.019	0.017
	500	0.0	100.0	0.0	0.016	0.017	0.016
NLARIUI	50	0.0	98.8	1.2	0.012	0.017	0.012
	100	0.0	100.0	0.0	0.011	0.014	0.011
	250	0.0	100.0	0.0	0.010	0.011	0.010
	500	0.0	100.0	0.0	0.010	0.011	0.010

Notes: The columns of U, C, O summarise the percentage of underfitting (U), correct-fitting (C) and overfitting (O) in 500 replications, respectively. The MSPE reports the averaged MSPE for CUREST, FULL, and ORACLE in 500 replications.

Table 2 compares the execution time and variable selection accuracy between CUREST and the BIC method in Huang and Yang (2004). The BIC method was implemented in the stepwise fashion as suggested in Huang and Yang (2004), and set $\lceil d_n/2 \rceil$ to be the total number of candidate variables to be selected from. Both methods are carried out under the same circumstance, that is, for model NLARI, using linear splines ($p = 2$), and 500 replications. Table 2 clearly shows that the proposed CUREST procedure is more computationally efficient. The BIC method in Huang and Yang (2004) always takes much longer time than CUREST. When sample size is $n = 1000$ and $d_n = 100$, BIC can take 187 times longer to compute than CUREST. In terms of variable selection accuracy, Table 2 shows that CUREST also outperforms BIC and the frequency of correct fitting for CUREST is always close to 100% regardless of sample size n and dimension d_n. However, BIC has a noticeably higher percentage of overfitting especially when the number of candidate lags are large ($d_n = 0.1n$), which is consistent with the theoretical results in Chen and Chen (2008). Instead of a directed stepwise search, a full search over all sub-models may improve the selection accuracy of the BIC approach. But with dimension as high as $d_n = 50$, the full search over 2^{50} subsets becomes infeasible. Figure 3 plots the CUREST values and the stepwise BIC values in one run when $n = 250, d_n = 25$. In Figure 3, the CUREST values are well ordered and two relevant lags (1,2) are clearly distinguished from the rest. However, the stepwise BIC method selects an overfitted model with three lags.

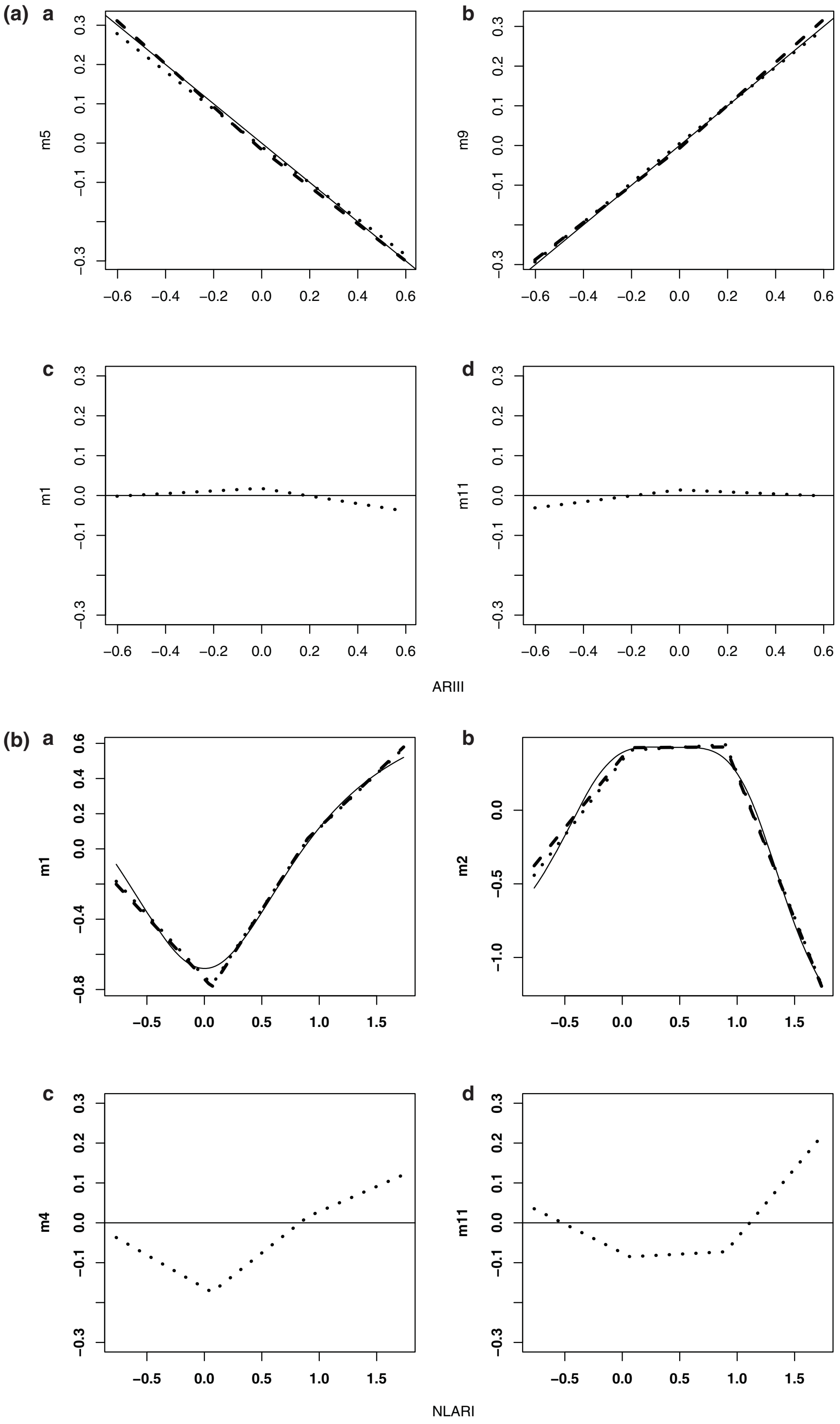

Figure 2. The estimated component functions using CUREST (dashed line), FULL (dotted line), and the true component functions (solid line) in one run for ARIII and NLARI models with $n = 250$. Shown are the significant lags (a and b) and two unimportant lags (c and d).

Table 2. Comparison of execution time and accuracy of CUREST and BIC for NLAR I model under different d_n and n.

		Execution time (s)			Accuracy (%)					
		CUREST	BIC	Ratio	CUREST			BIC		
d_n	n				U	C	O	U	C	O
$\lceil 3\log^{0.75} n\rceil$	250	17.41	55.21	3.17	0.0	100.0	0.0	0.0	91.6	8.4
	500	31.90	73.56	2.31	0.0	100.0	0.0	0.0	96.2	3.8
	1000	71.43	159.56	2.23	0.0	100.0	0.0	0.0	97.8	2.2
$0.1n$	250	17.64	376.42	21.34	0.0	100.0	0.0	0.0	77.6	22.4
	500	61.29	4085.94	66.66	0.0	99.8	0.2	0.0	75.8	24.2
	1000	337.54	63197.89	187.23	0.0	99.6	0.4	0.0	72.6	27.4

Notes: Execution time records the time spent for 500 replications and the ratio of BIC over CUREST. In each accuracy setup, three columns give the frequencies of underfitting (U), correct-fitting (C), and overfitting (O), respectively.

6.2. *Regression models*

In this section we illustrate the performance of the proposed method by considering additive regression models (2). The covariates $\{\mathbf{X}_t = (X_{t1},\dots,X_{td_n})^\mathsf{T}\}_{t=1}^n$ are generated from the vector AR equation $X_{tl} = \Phi\{(1-a^2)^{1/2}Z_{tl}\} - 1/2,\ 1 \le l \le d_n$ with stationary distribution $\mathbf{Z}_t = (Z_{t1},\dots,Z_{td_n})^\mathsf{T} \sim N(\mathbf{0}_{d_n},(1-a^2)^{-1}\Sigma)$, and

$$\mathbf{Z}_1 \sim N(\mathbf{0}_{d_n},(1-a^2)^{-1}\Sigma), \quad \mathbf{Z}_t = a\mathbf{Z}_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(\mathbf{0}_{d_n},\Sigma), \quad 2 \le t \le n,$$
$$\Sigma = (1-\rho)\mathbf{I}_{d_n\times d_n} + \rho\mathbf{1}_{d_n}\mathbf{1}_{d_n}^\mathsf{T}, \quad 0 \le a < 1,\ 0 \le \rho < 1,$$

where Φ is the standard normal distribution function, $\mathbf{1}_{d_n} = (1,\dots,1)^\mathsf{T}$, and $\mathbf{I}_{d_n\times d_n}$ is the $d_n \times d_n$ identity matrix. So $\{\mathbf{X}_t\}_{t=1}^n$ is geometrically α-mixing with marginal distribution $U[-0.5,0.5]$. Larger value of a corresponds to stronger dependence among the observations over time, and in particular, if $a=0$, $\{\mathbf{X}_t\}_{t=1}^n$ are iid. The parameter ρ controls correlation among the d_n covariates $\mathbf{X}_t = (X_{t1},\dots,X_{t,d_n})$. We consider $a = \rho = 0.5$ where both the dependence over time and correlation between the covariates are moderately strong. We will test the method through the following regression models:

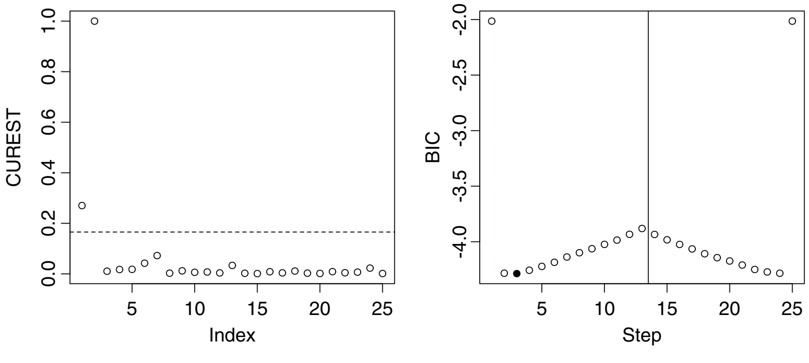

Figure 3. Plots of CUREST values (left panel) and stepwise BIC (right panel) in one run when $n = 250, d_n = 25$ for NLARI model. The points in left panel are each lags' CUREST values, and the dotted line presents the threshold a_n. The points in right panel are the BIC values in each step and the solid vertical line separates the forward and backward stages. In this run, the CUREST will correctly identify the true model since the important lags (1,2) are the only ones with their CUREST values greater than a_n; however the BIC method falsely picks an overfitted model with three lags (the solid point).

(1) *Model* 1 $Y_t = 0.5 + X_{t1} + X_{t5} + X_{t7} + X_{t8} + 0.1\xi_t$.

(2) *Model* 2

$$Y_t = 0.5 + \sqrt{3}X_{t2} + \frac{1}{\sqrt{6}}(2X_{t4} - 1)^2 + \frac{\sin(2\pi X_{t6})}{2 - \sin(2\pi X_{t6})} + \frac{1}{\sqrt{2}}\cos(2\pi X_{t9}) + 0.1\xi_t.$$

(3) *Model* 3

$$Y_t = 0.5 + (1 - 2X_{t3})\exp(-X_{t3}^2) + 4\frac{1 - X_{t5}^2}{1 + X_{t5}^2} + \sqrt{2}g(X_{t8}) + 6\Phi(3X_{t9}) + 0.1\xi_t,$$

where

$$g(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\{\sin(2\pi x)\}^2 + 0.4\{\cos(2\pi x)\}^2 + 0.5\{\sin(2\pi x)\}^3.$$

In above three models, ξ_t are i.i.d. $N(0, 1)$ random variables.

Similar to Section 6.1, we consider sample sizes 50,100,250,500 with the number of variables $d_n = 9, 10, 11, 12$ respectively. In addition, we also considered two more challenging cases with a larger number of covariates. In particular, we considered $d_n = 0.1n$ with $d_n = 25, 50$ for $n = 250$ and 500, respectively. A total of 500 data sets are generated from each of above three models. Figure 4 takes Model 3 as an example to illustrate how CUREST method works. It plots the CUREST values of each covariate and the location of the threshold a_n (dotted line). In Figure 4,

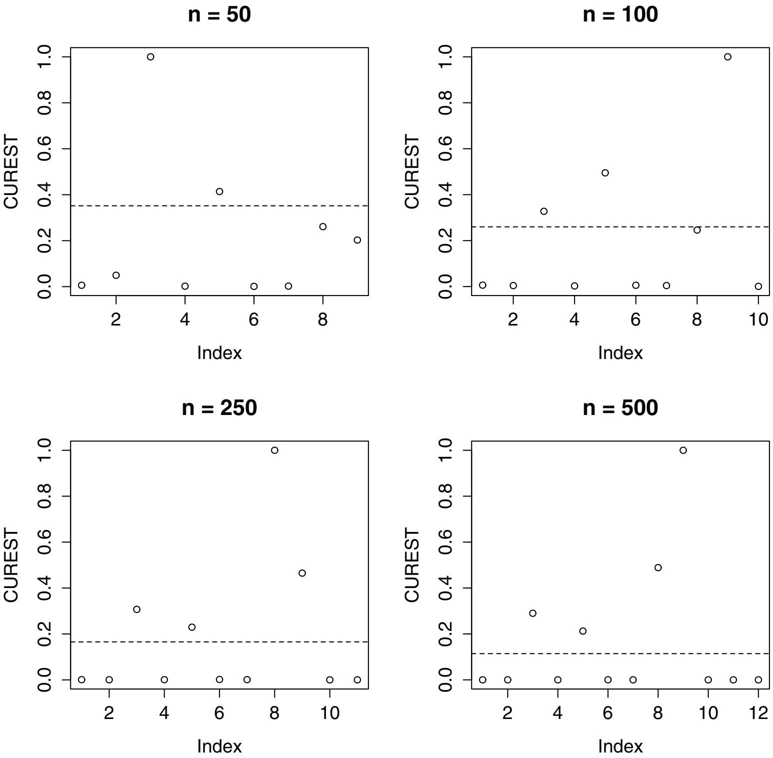

Figure 4. CUREST values in Model 3 under different sample sizes. In each plot, the dots denote the CUREST values for each variable, and the dotted line is the predefined threshold a_n.

the relevant variables (3,5,8,9) all have higher CUREST values than the irrelevant ones, and the threshold value a_n correctly separate the two groups especially when the sample size n is large.

Table 3 summarises the variable selection and prediction results for all three models. There are two extra rows with $n = 250^*, 500^*$ for each model, which are the results when $d_n = 0.1n$. It shows that the correct-fitting frequency converges to 100% as n increases, and the MSPE of the selected model invariably outperforms the full model and is closer to the oracle model.

Figures 5 and 6 plot the spline estimates of the additive components in Models 2 and 3 when $n = 100$. Panels a, b, d, and e display the relevant variables, and panels c and f show several irrelevant ones. The plots demonstrate the attractive property of spline estimates as well as the difference between the important variables from the insignificant ones.

7. Real data analysis

In this section, we apply the CUREST method to analyse the US unemployment rate. The data set we consider is the quarterly US unemployment rate time series from the first quarter of year 1948 to the last quarter of year 2000, denoted as $\{R_t\}_{t=1}^{212}$, obtained from the website of US Bureau of Labor Statistics and covers unemployed persons (in the labour 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 $\{Y_t\}_{t=1}^{208},\ Y_t = R_{t+4} - R_t,\ 1 \le t \le 208$. We leave out the last 20 periods of the data (i.e. $\{Y_t\}_{t=189}^{208}$) for prediction exercise and use the rest of the series as training set.

We consider an additive AR model of form

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

where the lags $\{i_1, \ldots, i_k\}$ are selected from $\{1, 2, \ldots, d_n\}$ with $d_n = 11$ using either CUREST or stepwise BIC method. In CUREST, the threshold $a_n = n^{0.1-2p/(2p+1)}(\log n)^{3/2}$, the same choice

Table 3. Simulation results of model selection via CUREST.

		Selection result (%)			MSPE		
Model	n	U	C	O	CUREST	FULL	ORACLE
Model 1	50	26.4	73.6	0.0	0.030	0.018	0.012
	100	0.0	100.0	0.0	0.011	0.013	0.011
	250	0.0	100.0	0.0	0.010	0.011	0.010
	500	0.0	100.0	0.0	0.010	0.011	0.010
	250*	0.0	100.0	0.0	0.010	0.012	0.010
	500*	0.0	100.0	0.0	0.010	0.013	0.010
Model 2	50	74.2	25.8	0.0	0.234	0.194	0.139
	100	25.0	75.0	0.0	0.142	0.143	0.123
	250	0.2	99.8	0.0	0.115	0.122	0.115
	500	0.0	100.0	0.0	0.111	0.114	0.111
	250*	0.2	99.8	0.0	0.113	0.140	0.113
	500*	0.0	100.0	0.0	0.109	0.136	0.109
Model 3	50	64.2	35.8	0.0	0.333	0.276	0.202
	100	6.6	93.4	0.0	0.184	0.205	0.176
	250	0.0	100.0	0.0	0.160	0.170	0.160
	500	0.0	100.0	0.0	0.160	0.164	0.160
	250*	0.0	100.0	0.0	0.163	0.198	0.163
	500*	0.0	100.0	0.0	0.162	0.200	0.162

Notes: The columns of U, C, O summarise the percentage of underfitting (U), correct-fitting (C), and overfitting (O) in 500 replications, respectively. The MSPE reports the averaged MSPEs for CUREST, FULL, and ORACLE in 500 replications. For each model, two extra rows with $n = 250^*, 500^*$ give the results when $d_n = 0.1n$.

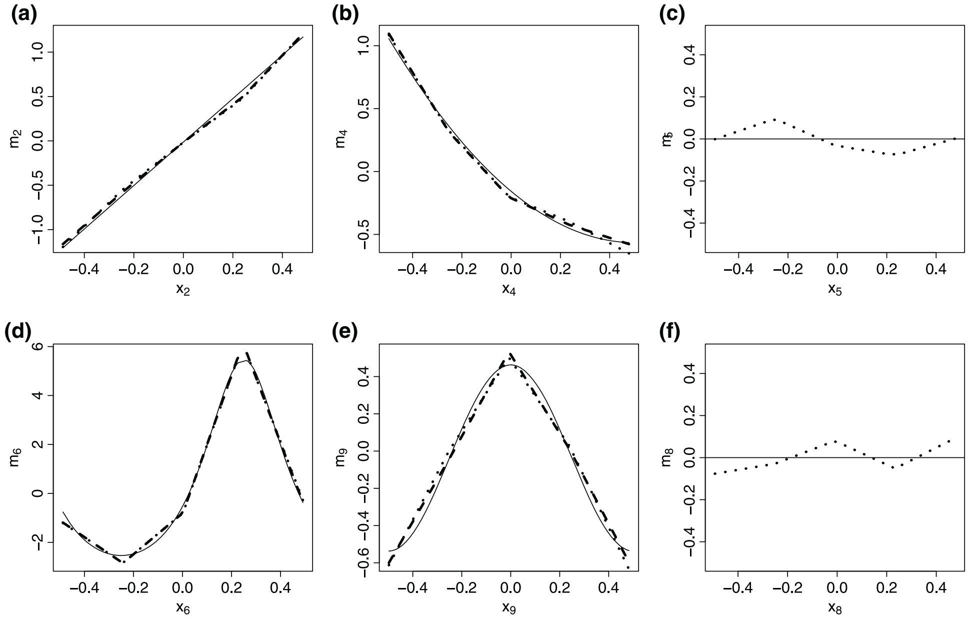

Figure 5. The estimated component functions using selected CUREST model (dashed line), full model (dotted line), and the true component functions (solid line) in one run for Model 2 with $n = 100$. Shown are the significant variables (a, b, d, and e) and two unimportant lags (c and f).

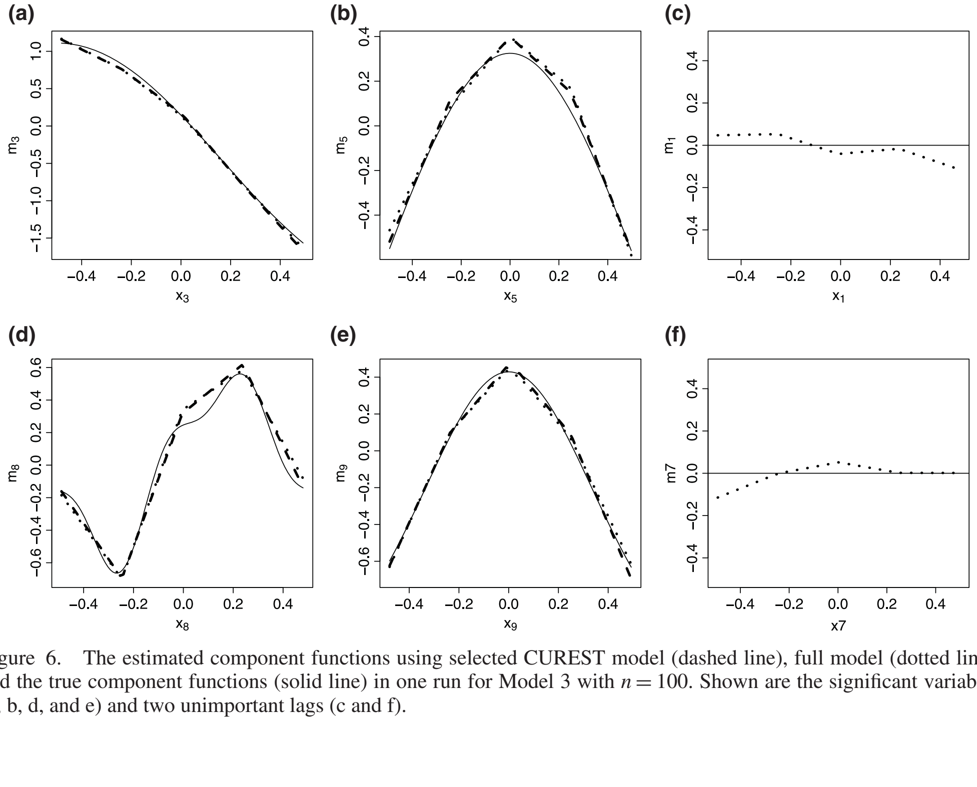

Figure 6. The estimated component functions using selected CUREST model (dashed line), full model (dotted line), and the true component functions (solid line) in one run for Model 3 with $n = 100$. Shown are the significant variables (a, b, d, and e) and two unimportant lags (c and f).

Table 4. Results for US unemployment data.

Model	Lags	MSPE	MAPE
CUREST	$\{1,2\}$	0.030	0.134
Full	$\{1,2,\ldots,11\}$	0.054	0.198
AR(2)	$\{1,2\}$	0.040	0.152
AR(11)	$\{1,2,\ldots,11\}$	0.046	0.180

Note: Comparison of prediction performance of CUREST model, full model, AR model selected by BIC (AR(2)) and full AR model (AR(11)).

for simulation studies. We also consider a linear AR model of order j : AR(j), where the optimal order j is selected by the BIC method. All three variable selection procedures consistently select lags 1 and 2 as relevant variables for both additive and linear models.

In addition, we compare the performance of selected models by their prediction performances. The prediction performance is measured in terms of the MSPE and mean absolute prediction error (MAPE), which are defined as

$$\mathsf{MSPE} = \sum_{t=189}^{208} (Y_t - \hat{Y}_t)^2, \quad \mathsf{MAPE} = \sum_{t=189}^{208} |Y_t - \hat{Y}_t|,$$

where $\hat{Y}_t$ are predictors produced by the selected model.

Table 4 reports the results on lag selection and prediction performance for the additive model selected by CUREST and the optimal linear AR model. For comparison, we also considered a full additive AR model and a full linear AR model with all 11 lags. It shows that CUREST performs the best with the smallest prediction error. It illustrates that the parsimonious additive AR model identified by proposed CUREST is not only of simpler structure, but also has better prediction performance than the full model using all 11 lags. Furthermore, both additive models have noticeably smaller prediction errors than the linear ones, which indicates that the quarterly US unemployment data contains nonlinear structures that cannot be fully explained by the linear AR models.

8. Discussion

The additive model is a powerful semiparametric tool for identifying nonlinear associations in data with complicated structures. In this paper, we proposed a data-driven method to select significant variables in additive model through a newly proposed CUREST. Model selection consistency have been established under very broad assumptions on the data-generating process, which allows data to be weakly dependent and the dimensionality of the full model to grow with the sample size. Our empirical example also demonstrated its superior performance over the BIC method in terms of speed and accuracy.

However, above numerical and theoretical properties of the proposed variable selection model is established based on the assumption that the underlying regression model is additive, may not be applicable for models that are not additive, such as, models with interactions or of serious confounding covariates.

Disclosure statement

No potential conflict of interest was reported by the authors.

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*, New York: Springer-Verlag.

Buja, A., Hastie, T., and Tibshirani, R. (1989), 'Linear Smoothers and Additive Models', *Annals of Statistics*, 17, 453–510.

Chen, J., and Chen, Z. (2008), 'Extended Bayesian Information Criteria for Model Selection with Large Model Spaces', *Biometrika*, 95, 759–771.

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

Chen, R., Liang, H., and Wang, J. (2011), 'Determination of Linear Components in Additive Models', *Journal of Nonparametric Statistics*, 23, 367–383.

Cheng, B., and Tong, H. (1992), 'On Consistent Nonparametric Order Determination and Chaos (with Discussion)', *Journal of the Royal Statistical Society Series B*, 54, 427–474.

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

Fan, J., Härdle, W., and Mammen, E. (1998), 'Direct Estimation of Low-dimensional Components in Additive Models', *Annals of Statistics*, 26, 943–971.

Fan, J., Feng, Y., and Song, R. (2011), 'Nonparametric Independence Screening in Sparse Ultra-high-dimensional Additive Models', *Journal of the American Statistical Association*, 106, 544–557.

Friedman, J.H. (1991), 'Multivariate Adaptive Regression Splines (with Discussion)', *Annals of Statistics*, 19, 1–67.

Härdle, W., and Korostelev, A. (1996), 'Search for Significant Variables in Nonparametric Additive Regression', *Biometrika*, 83, 541–549.

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

Huang, J.Z. (1998), 'Functional ANOVA Models for Generalized Regression', *Journal of Multivariate Analysis*, 67, 49–71.

Huang, J.Z., and Shen, H. (2004), 'Functional Coefficient Regression Models for Nonlinear Time Series: A Polynomial Spline Approach', *Scandinavian Journal of Statistics*, 31, 515–534.

Huang, J.Z., and Yang, L. (2004), 'Identification of Non-linear Additive Autoregressive Models', *Journal of the Royal Statistical Society Series B*, 66, 463–477.

Huang, J.Z., J.L. Horowitz, and Wei, F. (2010), 'Variable Selection in Nonparametric Additive Models', *Annals of Statistics*, 38, 2282–2313.

Jiang, S., and Xue, L. (2013), 'Lag Selection in Stochastic Additive Models', *Journal of Nonparametric Statistics*, 25, 129–146.

Jiang, J., Fan, Y., and Fan, J. (2010), 'Estimation in Additive Models with Highly or Nonhighly Correlated Covariates', *Annals of Statistics*, 38, 1403–1432.

Lee, Y.K., Mammen, E., and Park, B.U. (2010), 'Backfitting and Smooth Backfitting for Additive Quantile Models', *Annals of Statistics*, 38, 2857–2883.

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.

Lian, H. (2014), 'Semiparametric Bayesian Information Criterion for Model Selection in Ultra-high Dimensional Additive Models', *Journal of Multivariate Analysis*, 123, 304–310.

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

Liu, R., and Yang, L. (2010), 'Spline-Backfitted Kernel Smoothing of Additive Coefficient Model', *Economic Theory*, 26, 29–59.

Liu, X., Wang, L., and Liang, H. (2011), 'Estimation and Variable Selection for Semiparametric Additive Partial Linear Models', *Statistica Sinica*, 21, 1225–1248.

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

Ma, S. (2012), 'Two-step Spline Estimating Equations for Generalized Additive Partially Linear Models with Large Cluster Sizes', *Annals of Statistics*, 40, 2943–2972.

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

Ma, S., Song, Q., and Wang, L. (2013), 'Simultaneous Variable Selection and Estimation in Semiparametric Modeling of Longitudinal/Clustered Data', *Bernoulli*, 19, 252–274.

Mammen, E., Linton, O., and Nielsen, J. (1999), 'The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions', *Annals of Statistics*, 27, 1443–1490.

Mammen, E., Støve, B., and Tjøstheim, D.T. (2009), 'Nonparametric Additive Models for Panels of Time Series', *Economic Theory*, 25, 442–481.

Masry, E., and Tjøstheim, D.T. (1997), 'Additive Nonlinear ARX Time Series and Projection Estimates', *Economic Theory*, 13, 214–252.

Schwarz, G. (1978), 'Estimating the Dimension of a Model', *Annals of Statistics*, 6, 461–464.

Song, Q., and Yang, L. (2010), ‘Oracally Efficient Spline Smoothing of Nonlinear Additive Autoregression Models with Simultaneous Confidence Band’, *Journal of Multivariate Analysis*, 101, 2008–2025.

Stone, C.J. (1985), ‘Additive Regression and Other Nonparametric Models’, *Annals of Statistics*, 13, 689–705.

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.

Wang, L., and Yang, L. (2007), ‘Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model’, *Annals of Statistics*, 35, 2474–2503.

Wang, J., and Yang, L. (2009), ‘Efficient and Fast Spline-Backfitted Kernel Smoothing of Additive Models’, *Annals of the Institute of Statistical Mathematics*, 61, 663–690.

Wang, L., Liu, X., Liang, H., and Carroll, R.J. (2011), ‘Estimation and Variable Selection for Generalized Additive Partial Linear Models’, *Annals of Statistics*, 39, 1827–1851.

Xue, L. (2009), ‘Variable Selection in Additive Models’, *Statistica Sinica*, 19, 1281–1296.

Xue, L., Qu, A., and Zhou, J. (2010), ‘Consistent Model Selection for Marginal Generalized Additive Model for Correlated Data’, *Journal of the American Statistical Association*, 105, 1518–1530.

Xue, L., and Yang, L. (2006), ‘Additive Coefficient Modeling via Polynomial Spline’, *Statistica Sinica*, 16, 1212–1227.

Yang, L., and Tschernig, R. (2002), ‘Non- and Semiparametric Identification of Seasonal Nonlinear Autoregression Models’, *Economic Theory*, 18, 1408–1448.

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., Park, B.U., Xue, L., and Härdle, W. (2006), ‘Estimation and Testing for Varying Coefficients in Additive Models with Marginal Integration’, *Journal of the American Statistical Association*, 101, 1212–1227.

Yao, Q., and Tong, H. (1994), ‘On Subset Selection in Non-parametric Stochastic Regression’, *Statistica Sinica*, 4, 51–70.

Yu, K., Park, B.U., and Mammen, E. (2008), ‘Smooth Backfitting in Generalized Additive Models’, *Annals of Statistics*, 36, 228–260.

Appendix

A.1. *Technical lemmas*

The B-spline basis is used in the proofs. With $J_n = N_n + p$, we denote the B-spline basis of $\mathbb{G}_l$ by $\mathbf{b}_l = (b_{l0}, b_{l1}, \ldots, b_{l,J_n})$. For $1 \le l \le d_n$, let $\mathbf{B}_l = (B_{l1}, \ldots, B_{l,J_n})$ with

$$B_{lj} = \sqrt{N_n}\left(b_{lj} - \frac{\mathsf{E}(b_{lj})}{\mathsf{E}(b_{l0})} b_{l0}\right), \quad j = 1, \ldots, J_n. \tag{A1}$$

Let $\mathbf{B} = (1, B_{11}, \ldots, B_{1,J_n}, \ldots, B_{d_n,1}, \ldots, B_{d_n,J_n})^{\mathsf{T}}, R_n = 1 + d_n J_n$. Then $\mathbf{B}$ is a set of basis for $\mathbb{G}$. The following Lemma A.1 to A.5 focus on the properties of the B-spline basis $\mathbf{B}$, and are extension work of those in Xue and Yang (2006) to a high-dimensional situation.

Lemma A.1 *For any* $1 \le l \le d_n$, *and the spline basis* B_{lj} *of Equation* (A1), *one has*

(i) $\mathsf{E}(B_{lj}) = 0,\ \mathsf{E}|B_{lj}|^k \asymp N_n^{k/2-1}$ *for* $k > 1,\ j = 1, \ldots, J_n$;
(ii) $\| \sum_{j=1}^{J_n} c_{lj} B_{lj}(x_l) \|_2^2 \ge C \sum_{j=1}^{J_n} c_{lj}^2$.

Lemma A.2 *Let* ε_1 *be defined in Assumption* (A3)

$$\left\| \sum_{l=1}^{d_n} \sum_{j=1}^{J_n} c_{lj} B_{lj}(x_l) \right\|_2^2 \ge C \varepsilon_1^{d_n} \sum_{l=1}^{d_n} \sum_{j=1}^{J_n} c_{lj}^2.$$

The proof follows directly from Lemma 6 in Huang (1998).

Lemma A.3 $\langle \mathbf{B}, \mathbf{B} \rangle = (\langle B_i, B_j \rangle)_{i,j=1}^{R_n}$, $\langle \mathbf{B}, \mathbf{B} \rangle_{2,n} = (\langle B_i, B_j \rangle_{2,n})_{i,j=1}^{R_n}$, $D = \operatorname{diag}(\langle \mathbf{B}, \mathbf{B} \rangle)$, $Q_n = \sup |D^{-1/2}(\langle \mathbf{B}, \mathbf{B} \rangle - \langle \mathbf{B}, \mathbf{B} \rangle_{2,n}) D^{-1/2}|$, *where the sup is taken over all the elements in the random matrix. Then* $Q_n = \mathcal{O}_P(\sqrt{d_n (N_n log^2(n)/n)})$.

Proof We only consider the diagonal terms.

$\xi = (\mathsf{E}_n - \mathsf{E})\{B_{lj}^2(x_l)\} = n^{-1}\sum_{t=1}^n \{B_{lj}^2(X_l) - \mathsf{E}B_{lj}^2(X_l)\} = n^{-1}\sum_{t=1}^n \xi_t$, $\mathsf{E}(\xi_t) = 0$,

$$\mathsf{E}|\xi_t|^k \leq 2^{k-1}[\mathsf{E}\{|B_{lj}^2|^k\} + \{\mathsf{E}|B_{lj}^2|\}^k] \leq 2^{k-1}\{C^k N_n^k + (CN_n)^k\} \leq c^k N_n^k \quad \text{for } k \geq 3,$$

$\mathsf{E}|\xi_t|^2 = \mathsf{E}[\{B_{lj}^2(X_l) - \mathsf{E}B_{lj}^2\}^2] = \mathsf{E}(B_{lj}^4) - \{\mathsf{E}(B_{lj}^2)\}^2$;

Since $\{\mathsf{E}(B_{lj}^2)\}^2 \leq c$, $\mathsf{E}(B_{lj}^4) \geq M_1^{-1}\int B_{lj}^4(x_l)\,\mathrm{d}x_l \geq cN_n^2$, one has $\mathsf{E}|\xi_t|^2 \geq cN_n^2$, $\mathsf{E}|\xi_t|^k \leq c^k N_n^k \leq (cN_n)^{k-2}k!\mathsf{E}|\xi_t|^2$.

Apply Theorem 1.4 in Bosq (1998) to $\sum_{t=1}^n \xi_t$ with Crammer constant $c_r = cN_n$.

That is, for any $\epsilon > 0$, $q \in [1, n/2]$ and $k \geq 3$, one has

$$P\left(\frac{1}{n}\sum_{t=1}^n \xi_t \geq \epsilon\sqrt{d_n\frac{N_n\log^2(n)}{n}}\right)$$
$$\leq a_1 \exp\left(-\frac{q\epsilon^2 d_n \frac{N_n\log^2(n)}{n}}{25m_2^2 + 5c_r\epsilon\sqrt{d_n\frac{N_n\log^2(n)}{n}}}\right) + a_2(k)\alpha\left(\left[\frac{n}{q+1}\right]\right)^{2k/(2k+1)},$$

where

$$a_1 = 2\frac{n}{q} + 2\left(1 + \frac{\epsilon^2 d_n\frac{N_n\log^2(n)}{n}}{25m_2^2 + 5c_r\epsilon\sqrt{d_n\frac{N_n\log^2(n)}{n}}}\right),$$

$$a_2(k) = 11n\left(1 + \frac{5m_p^{k/(2k+1)}}{\epsilon\sqrt{d_n\frac{N_n\log^2(n)}{n}}}\right), \quad m_2^2 = \mathsf{E}\xi_t^2, \quad m_p = \|\xi_t\|_p.$$

Observe that $5c_r\epsilon\sqrt{d_n(N_n\log^2(n)/n)} = 5\epsilon cN_n\sqrt{d_n(N_n\log^2(n)/n)} = 5c\epsilon\sqrt{d_n(N_n^3\log^2(n)/n)} = o(1)$, $a_2(k) = \mathcal{O}(n(N_n^{k/(2k+1)}/\sqrt{d_n(N_n\log^2(n)/n)})) = \mathcal{O}(n^{3/2}(N_n^{-1/2(2k+1)}/d_n^{1/2}\log(n))) = o(n^{3/2})$.

Take $q = n/c_0\log n$, then $a_1 = \mathcal{O}(n/q) = \mathcal{O}(\log n)$.

Also note that, by Assumption (A1), one has

$$\exp\left(-\frac{q\epsilon^2 d_n\frac{N_n\log^2(n)}{n}}{25m_2^2 + 5c_r\epsilon\sqrt{d_n\frac{N_n\log^2(n)}{n}}}\right) = \exp\left(-\frac{q\epsilon^2 d_n\frac{N_n\log^2(n)}{n}}{25m_2^2}(1 + o(1))\right),$$

$$\alpha\left(\left[\frac{n}{q+1}\right]\right) \leq c\rho^{[n/(q+1)]} \leq c_1\exp\left(\left[\frac{n}{q+1}\right]\log\rho\right)$$
$$\leq c_1\exp(\log\rho \cdot c_0 \cdot \log n).$$

For n large enough,

$$P\left(\frac{1}{n}\sum_{t=1}^n \xi_t \geq \epsilon\sqrt{d_n\frac{N_n\log^2(n)}{n}}\right) \leq c\log n\exp\left(-\frac{q\epsilon^2 d_n\frac{N_n\log^2(n)}{n}}{50m_2^2}\right)$$
$$+ c_1 n^{-3/2}\exp(\log\rho \cdot c_0 \cdot \log n)$$
$$= cn^{-\epsilon^2 d_n N_n/50m_2^2}\log n + c_1 n^{c_0\log\rho - 3/2}.$$

Taking c_0, ϵ, m large enough, one has that

$$\sum_{n=1}^{\infty} P\left(\sup|(\langle\mathbf{B},\mathbf{B}\rangle - \langle\mathbf{B},\mathbf{B}\rangle_{2,n})| \geq \epsilon\sqrt{d_n\frac{N_n\log^2(n)}{n}}\right)$$
$$\leq \sum_{n=1}^{\infty} R_n^2\left\{cn^{-\epsilon^2 d_n N_n/50m_2^2}\log n + c_1 n^{c_0\log\rho - 3/2}\right\} < \sum_{n=1}^{\infty} R_n^2 n^{-3/2} < \infty.$$

Then the lemma follows from Borel–Cantelli Lemma. ■

Lemma A.4 $\sup_{\phi_1,\phi_2\in\mathbb{G}} |(\langle\phi_1,\phi_2\rangle_{2,n} - \langle\phi_1,\phi_2\rangle)/\|\phi_1\|\|\phi_2\|| = \mathcal{O}_P(d_n^{3/2}N_n^{3/2}log(n)/n^{1/2}\varepsilon_1^{d_n}) \triangleq \mathcal{O}_P(\tau_n)$.

Proof With vector notation, we can write $\phi_1 = \mathbf{a}_1^\top\mathbf{B}, \phi_2 = \mathbf{a}_2^\top\mathbf{B}$ for $R_n \times 1$ vectors.

$$
\begin{aligned}
&|\langle\phi_1,\phi_2\rangle_{2,n} - \langle\phi_1,\phi_2\rangle| \\
&\quad = |\langle\mathbf{a}_1^\top\mathbf{B},\mathbf{a}_2^\top\mathbf{B}\rangle_{2,n} - \langle\mathbf{a}_1^\top\mathbf{B},\mathbf{a}_2^\top\mathbf{B}\rangle| = \left|\left\langle\sum_{i=1}^{R_n} a_{1i}B_i, \sum_{j=1}^{R_n} a_{2j}B_j\right\rangle_{2,n} - \left\langle\sum_{i=1}^{R_n} a_{1i}B_i, \sum_{j=1}^{R_n} a_{2j}B_j\right\rangle\right| \\
&\quad = \left|\sum_{i,j=1}^{R_n} a_{1i}a_{2j}(\langle B_i,B_j\rangle_{2,n} - \langle B_i,B_j\rangle)\right| \le CQ_n\sum_{i,j=1}^{R_n}|a_{1i}a_{2j}| = CQ_n\sum_{i=1}^{R_n}|a_{1i}|\sum_{j=1}^{R_n}|a_{2j}| \\
&\quad \le CQ_n\sqrt{R_n\sum_{i=1}^{R_n}a_{1i}^2}\sqrt{R_n\sum_{j=1}^{R_n}a_{2j}^2} = CQ_nR_n\sqrt{\mathbf{a}_1^\top\mathbf{a}_1\mathbf{a}_2^\top\mathbf{a}_2}.
\end{aligned}
$$

On the other hand, $\|\phi_1\|\|\phi_2\| = \sqrt{\mathbf{a}_1^\top\langle\mathbf{B},\mathbf{B}\rangle\mathbf{a}_1}\sqrt{\mathbf{a}_2^\top\langle\mathbf{B},\mathbf{B}\rangle\mathbf{a}_2} \ge C\varepsilon_1^{d_n}\sqrt{\mathbf{a}_1^\top\mathbf{a}_1\mathbf{a}_2^\top\mathbf{a}_2}$. Then $|(\langle\phi_1,\phi_2\rangle_{2,n} - \langle\phi_1,\phi_2\rangle)/\|\phi_1\|\|\phi_2\|| \le |CQ_nR_n\sqrt{\mathbf{a}_1^\top\mathbf{a}_1\mathbf{a}_2^\top\mathbf{a}_2}/C\varepsilon_1^{d_n}\sqrt{\mathbf{a}_1^\top\mathbf{a}_1\mathbf{a}_2^\top\mathbf{a}_2}| = CQ_n(d_nN_n/\varepsilon_1^{d_n}) = \mathcal{O}_P(d_n^{3/2}N_n^{3/2}\log(n)/n^{1/2}\varepsilon_1^{d_n})$. ■

Lemma A.5 *Let* ε_1 *be defined in Assumption* (A3)

$$
\left\|\sum_{l=1}^{d_n}\sum_{j=1}^{J_n} c_{lj}B_{lj}(x_l)\right\|_{2,n}^2 \ge C\varepsilon_1^{d_n}\sum_{l=1}^{d_n}\sum_{j=1}^{J_n} c_{lj}^2.
$$

Lemma A.6 *Let* ε_1 *be defined in Assumption* (A3). *For* $h(\mathbf{x}) = h_0 + \sum_{l=1}^{d_n} h_l(x_l) \in \mathbb{H}$

$$
\begin{aligned}
&\|h\|_2^2 \ge C\varepsilon_1^{d_n}\left\{h_0^2 + \sum_{l=1}^{d_n}\|h_l\|_2^2\right\}, \\
&\|h\|_{2,n}^2 \ge C\varepsilon_1^{d_n}\left\{h_0^2 + \sum_{l=1}^{d_n}\|h_l\|_{2,n}^2\right\} \quad \textit{with probability approaching to } 1.
\end{aligned}
$$

Proof The first inequality follows directly from Lemma 2 in Huang (1998). For $1 \le l \le d_n$, $\|h_l\|_{2,n}^2 = (1/n)\sum_{t=1}^n h_l^2(X_{tl}) = \|h_l\|_2^2 + \mathcal{O}_P(n^{-1/2})$ by central limit theorem. Therefore

$$
\begin{aligned}
C\varepsilon_1^{d_n}\left\{h_0^2 + \sum_{l=1}^{d_n}\|h_l\|_{2,n}^2\right\} &= C\varepsilon_1^{d_n}\left[h_0^2 + \sum_{l=1}^{d_n}\{\|h_l\|_2^2 + \mathcal{O}_P(n^{-1/2})\}\right] \\
&= C\varepsilon_1^{d_n}\left\{h_0^2 + \sum_{l=1}^{d_n}\|h_l\|_2^2\right\} + \mathcal{O}_P\left(\frac{\varepsilon_1^{d_n}d_n}{\sqrt{n}}\right),
\end{aligned}
$$

which completes the proof. ■

A.2. *Proof of Theorem 3.1*

In this part we will illustrate that the B-spline estimators are consistent for both the regression function m and additive components m_l, $1 \le l \le d_n$.

Let $\mathbf{Y} = (Y_1,\ldots,Y_n)^\top, \mathbf{m} = \{m(\mathbf{X}_1),\ldots,m(\mathbf{X}_n)\}^\top, \mathbf{E} = (\varepsilon_1,\ldots,\varepsilon_n)^\top$. Then $\mathbf{Y} = \mathbf{m} + \mathbf{E}$. Let $\hat{m} = \mathbf{Proj}_n\mathbf{Y}$=$\mathbf{Proj}_n\mathbf{m}$+$\mathbf{Proj}_n\mathbf{E} \triangleq \tilde{\mathbf{m}} + \tilde{\mathbf{e}}$, where $\mathbf{Proj}_n\mathbf{V}$ is the projection of vector $\mathbf{V}$ on the estimation space $\mathbb{G}$ with respect to the empirical inner product. According to de Boor (2001), for $1 \le l \le d_n$, there exists $C > 0$ and $g_l \in \mathbb{G}_l^0$ such that

$\|g_l - m_l\|_\infty \le CN_n^{-p}$. Let $m_n(\mathbf{x}) = m_0 + \sum_{l=1}^{d_n} g_l(x_l) \in \mathbb{G}$. One has

$$\|m - m_n\|_2 \le \sum_{l=1}^{d_n} \|g_l - m_l\|_2 \le \sum_{l=1}^{d_n} \|g_l - m_l\|_\infty \le Cd_nN_n^{-p}. \tag{A2}$$

Similarly $\|m - m_n\|_{2,n} \le Cd_nN_n^{-p}$. Then by the projection property, one has $\|m - \tilde{m}\|_{2,n} \le \|m - m_n\|_{2,n} \le Cd_nN_n^{-p}$ a.s.. It also implies $\|\tilde{m} - m_n\|_{2,n} \le \|\tilde{m} - m\|_{2,n} + \|m - m_n\|_{2,n} \le Cd_nN_n^{-p}$. By Lemma A.4, $\|\tilde{m} - m_n\|_2 \le \|\tilde{m} - m_n\|_{2,n} \times (1 + \tau_n) = \mathcal{O}_P(d_nN_n^{-p})$. Together with Equation (A2), one has $\|\tilde{m} - m\|_2 = \mathcal{O}_P(d_nN_n^{-p})$.

Next we consider the variance term, $\tilde{e}$, which can be written as $\tilde{e} = \sum_{j=1}^{R_n} a_j\phi_j$, where $\{\phi_j,\ 1 \le j \le R_n\}$ is an orthonormal basis of $\mathbb{G}$ relative to empirical inner product. Then $a_j = \langle \mathbf{E}, \phi_j\rangle_{2,n} = (1/n)\sum_{t=1}^{n} \varepsilon_t\phi_j(\mathbf{X}_t)$, $\tilde{e} = \sum_{j=1}^{R_n} a_j\phi_j = \sum_{j=1}^{R_n} \langle \mathbf{E}, \phi_j\rangle_{2,n}\phi_j$. Therefore $\|\tilde{e}\|_{2,n}^2 = \|\sum_{j=1}^{R_n} a_j\phi_j\|_{2,n}^2 = \sum_{j=1}^{R_n} a_j^2 = \sum_{j=1}^{R_n} \langle \mathbf{E}, \phi_j\rangle_{2,n}^2$.

From Assumption (A2), there exists $M > 0$, $\sigma^2(\mathbf{x}) \le M$ for all $\mathbf{x}$. One has for any $1 \le j \le R_n$

$$\begin{aligned}\mathsf{E}\{\langle \mathbf{E}, \phi_j\rangle_n^2\} &= \frac{1}{n^2}\sum_{t,t'=1}^{n} \mathsf{E}\{\phi_j(\mathbf{X}_t)\varepsilon_t\phi_j(\mathbf{X}_{t'})\varepsilon_{t'}\} = \frac{1}{n^2}\sum_{t,t'=1}^{n} \mathsf{E}[\phi_j(\mathbf{X}_t)\phi_j(\mathbf{X}_{t'})\mathsf{E}\{\varepsilon_t\varepsilon_{t'} \mid \mathbf{X}\}] \\ &= \frac{1}{n^2}\sum_{t,t'=1}^{n} \mathsf{E}\{\phi_j(\mathbf{X}_t)\sigma^2(\mathbf{X}_t)\} \le \frac{M}{n}\|\phi_j\|_{2,n}^2 = \frac{M}{n}.\end{aligned}$$

Hence $\mathsf{E}\{\|\tilde{e}\|_{2,n}^2\} \le R_n(M/n)$ and therefore $\|\tilde{e}\|_{2,n}^2 = \mathcal{O}_P(d_nN_n/n)$. Together with Lemma A.4, $\|\tilde{e}\|_2^2 = \mathcal{O}_P(d_nN_n/n)$.

In conclusion, $\|\hat{m} - m\|_{2,n}^2 + \|\hat{m} - m\|_2^2 = \mathcal{O}_P(d_n^2N_n^{-2p} + d_nN_n/n)$. By Lemma A.6 $C\varepsilon_1^{d_n}\sum_{l=1}^{d_n} \|\hat{m}_l - m_l\|_2^2 \le \|\hat{m} - m\|_2^2$, then for any $1 \le l \le d_n$

$$\|\hat{m}_l - m_l\|_2^2 \le \varepsilon_1^{-d_n}\|\hat{m} - m\|_2^2 = \mathcal{O}_P\left(\varepsilon_1^{-d_n}d_n^2N_n^{-2p} + \varepsilon_1^{-d_n}\frac{d_nN_n}{n}\right) = \mathcal{O}_P(u_n^2).$$

Hence $\|\hat{m}_l - m_l\|_2^2 = \mathcal{O}_P(u_n^2)$, similarly $\|\hat{m}_l - m_l\|_{2,n}^2 = \mathcal{O}_P(u_n^2)$, which completes the proof of Theorem 3.1.

Lemma A.7 $\max_{l\in S_0^C} \|\hat{m}_l\|_{2,n}^2 = \mathcal{O}_P(\varepsilon_1^{-d_n}u_n^2)$.

Proof By Theorem 3.1, for $l \in S_0$, $\|\hat{m}_l - m_l\|_{2,n}^2 = \mathcal{O}_P(u_n^2)$. Then $\|\sum_{l\in S_0}(\hat{m}_l - m_l)\|_{2,n}^2 = \mathcal{O}_P(u_n^2)$. Note that $\sum_{l\in S_0^C}(\hat{m}_l - m_l) = (\hat{m} - m) - \sum_{l\in S_0}(\hat{m}_l - m_l)$, therefore

$$\left\|\sum_{l\in S_0^C}(\hat{m}_l - m_l)\right\|_{2,n} \le \|\hat{m} - m\|_{2,n} + \left\|\sum_{l\in S_0}(\hat{m}_l - m_l)\right\|_{2,n},$$

which yields to

$$\left\|\sum_{l\in S_0^C}(\hat{m}_l - m_l)\right\|_{2,n}^2 = \mathcal{O}_P\left(\left(d_n^2N_n^{-2p} + \frac{d_nNn}{n}\right) + u_n^2\right) = \mathcal{O}_P(u_n^2).$$

Applying Lemma A.6, one has $\sum_{l\in S_0^C} \|(\hat{m}_l - m_l)\|_{2,n}^2 \le C\varepsilon_1^{-d_n}\|\sum_{l\in S_0^C}(\hat{m}_l - m_l)\|_{2,n}^2 = \mathcal{O}_P(\varepsilon_1^{-d_n}u_n^2)$. On the other hand, $\max_{l\in S_0^C} \|\hat{m}_l\|_{2,n}^2 \le \sum_{l\in S_0^C} \|(\hat{m}_l - m_l)\|_{2,n}^2$. Combine these together, we have $\max_{l\in S_0^C} \|\hat{m}_l\|_{2,n}^2 = \mathcal{O}_P(\varepsilon_1^{-d_n}u_n^2)$. ■

A.3. *Proof of Theorem 4.1*

According to Theorem 3.1, one has $\|\hat{m}_l - m_l\|_{2,n} + \|\hat{m}_l - m_l\|_2 = \mathcal{O}_P(u_n)$ for $1 \le l \le d_n$.

If $m_l = 0$, then $\lambda_l = 0$, hence $|\hat{\lambda}_l| = \|\hat{m}_l\|_{2,n}^2 = \|\hat{m}_l - m_l\|_{2,n}^2 = \mathcal{O}_P(u_n^2)$.

If $m_l \neq 0$, note that by definition $\|m_l\|_{2,n}^2 - \|m_l\|_2^2 = n^{-1}\sum_{t=1}^{n}[m_l^2(X_{tl}) - \mathsf{E}\{m_l^2(X_{tl})\}]$ which according to central limit theorem for α-mixing processes, converges to a Gaussian distribution at a rate of $n^{-1/2}$, hence

$\|m_l\|_{2,n}^2 - \|m_l\|_2^2 = \mathcal{O}_P(n^{-1/2})$ and $\|m_l\|_{2,n} = \mathcal{O}_P(1)$. Consequently

$$\begin{aligned}|\hat{\lambda}_l - \lambda_l| &= |\|\hat{m}_l\|_{2,n}^2 - \|m_l\|_2^2| \leq |\|\hat{m}_l\|_{2,n}^2 - \|m_l\|_{2,n}^2| + |\|m_l\|_{2,n}^2 - \|m_l\|_2^2| \\ &= |\|\hat{m}_l - m_l + m_l\|_{2,n}^2 - \|m_l\|_{2,n}^2| + \mathcal{O}_P(n^{-1/2}) \\ &= |\|\hat{m}_l - m_l\|_{2,n}^2 + 2\langle \hat{m}_l - m_l, m_l\rangle_{2,n}| + \mathcal{O}_P(n^{-1/2}) \\ &\leq \|\hat{m}_l - m_l\|_{2,n}^2 + 2\|\hat{m}_l - m_l\|_{2,n}\|m_l\|_{2,n} + \mathcal{O}_P(n^{-1/2}).\end{aligned}$$

One therefore obtains $|\hat{\lambda}_l - \lambda_l| = \mathcal{O}_P(u_n^2) + \mathcal{O}_P(u_n) + \mathcal{O}_P(n^{-1/2}) = \mathcal{O}_P(u_n)$. The second equation follows from Lemma A.7. This completes the proof of Theorem 4.1.

A.4. *Proof of Theorem 4.2*

According to Lemma A.7, $\max_{l\in S_0^C} |\hat{\lambda}_l| = \max_{l\in S_0^C} \|\hat{m}_l\|_{2,n}^2 = \mathcal{O}_P(\varepsilon_1^{-d_n} u_n^2) = o(1)$. On the other hand, the definition of S_0 ensures that $\min_{l\in S_0} \lambda_l = \lambda_{\min} > 0$. Therefore by Theorem 4.1, $\min_{l\in S_0} |\hat{\lambda}_l| \geq \lambda_{\min} + \mathcal{O}_P(u_n)$. Thus ordering by empirical strengths $\hat{\lambda}_l$ puts each $l \in S_0$ ahead of any $l \notin S_0$ with probability approaching 1. Denote $d_0 = |S_0|$, i.e. number of elements in the true model.

We next show that as $n \to \infty$

$$P(\hat{d}_0 < d_0) \to 0, \tag{A3}$$

$$P(\hat{d}_0 > d_0) \to 0. \tag{A4}$$

To prove Equation (A3), note that $\hat{d}_0 < d_0$ implies that $q_{\hat{d}_0}, q_{\hat{d}_0+1} \in S_0$ and thus

$$\hat{r}_{\hat{d}_0+1} = \frac{\hat{\lambda}_{q_{\hat{d}_0+1}}}{\hat{\lambda}_{q_1} + \cdots + \hat{\lambda}_{q_{\hat{d}_0+1}}} \geqslant \frac{\lambda_{q_{\hat{d}_0+1}} + \mathcal{O}_P(u_n)}{\sum_{l\in S_0} \lambda_l + \mathcal{O}_P(u_n)} \geqslant \frac{\lambda_{q_{\hat{d}_0+1}}}{\sum_{l\in S_0} \lambda_l} + \mathcal{O}_P(u_n) > a_n,$$

with probability approaching 1, since by definition $a_n \ll u_n$. This contradicts the definition of $\hat{d}_0 = \max\{l : \hat{r}_l > a_n\}$.

To prove Equation (A4), note that $\hat{d}_0 > d_0$ implies that $q_{\hat{d}_0} \notin S_0$, then $\lambda_{q_{\hat{d}_0}} = 0$ and $\hat{\lambda}_{q_{\hat{d}_0}} = \mathcal{O}_P(\varepsilon_1^{-d_n} u_n^2)$. Therefore

$$\hat{r}_{\hat{d}_0} = \frac{\hat{\lambda}_{q_{\hat{d}_0}}}{\hat{\lambda}_{q_1} + \cdots + \hat{\lambda}_{q_{\hat{d}_0}}} \leq \frac{\mathcal{O}_P(\varepsilon_1^{-d_n} u_n^2)}{\lambda_{\min} + \mathcal{O}_P(u_n)} = \mathcal{O}_P(\varepsilon_1^{-d_n} u_n^2) \leq a_n,$$

with probability approaching 1, since by definition $\varepsilon_1^{-d_n} u_n^2 \ll a_n$. This again contradicts the definition of $\hat{d}_0 = \max\{l : \hat{r}_l > a_n\}$.

Equations (A3) and (A4) together complete the proof of Theorem 4.2.