STAT 461 Final Project - Identification of Non-linear Additive Autoregressive Model

Xiaolong Liu, Siyu Xie

May 2024

1 Introduction

In Chen and Tsay (1993), they proposed **functional coefficient autoregressive (FAR)** models.

$$x_t = f(x_{t-1}, \cdots, x_{t-p}) + a_t$$

This model focuses on exploring the nonlinear feature of a time series in the process of model building. This is achieved by generalizing directly the linear autoregressive (AR) models and exploiting local characteristics of a given time series. A major difficulty in applying nonparametric methods to nonlinear time series analysis like this is the "curse of dimensionality" since a direct application of nonparametric methods to estimate $f(\cdot)$ would require $p-$dim smoothing. This is difficult when p is large, especially when the number of observation is relatively small. A simple but effective way to overcome this difficulty is to apply an additive model that only requires lower dimensional smoothing.

In this project, we want to study a specific realization of functional-coefficient autoregressive models - **nonlinear additive autoregressive (NAAR)** model in Huang and Yang (2004) focusing on lag selection using spline methods.

$$x_t = \mu_0 + \sum_{i=1}^{p} f_i(x_{t-i}) + a_t$$

The authors propose a method that uses polynomial spline global smoothing to enhance the model's ability to capture non-linear dependencies in time series data. This methodology leverages spline estimation techniques coupled with the Bayes Information Criterion (BIC) to effectively address the "curse of dimensionality". An extensive Monte Carlo simulation validates the method's computational efficiency and superior performance in identifying significant lags. These simulations also demonstrate the method's robustness compared to traditional lag selection techniques, which often fail to balance model accuracy and complexity effectively.

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 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

$$Y_t = \mu(X_t) + \epsilon_t, \tag{1}$$

with $E(\epsilon_t|X_t) = 0$.

The goal of this paper is to determine, without assuming that μ is known, a proper subset of variables $\{X_{t,i}|i \in S\}$, $S \subseteq \{1, \ldots, d\}$, 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.

$$E(Y_t|X_{t,i}, i \in S) = E(Y_t|X_t),$$

almost surely.

The variables selected are called significant variables. If X_t consists of only lagged values of Y_t, the lags selected are called significant lags. Since we do not assume that the regression function μ has 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 this difficulty is to impose some structure on the unknown regression function. In Huang and Yang (2004) we shall 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).

3 Additive spline estimation

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, we assume that $E\{\mu_i(X_{ti})\} = 0$, $i = 1, \ldots, d$, in equation (2). 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

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

where B_{ij}, $j = 1, \ldots, J_i$, 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 $\{\mu_0, \gamma_{ij}, j = 1, \ldots, J_i, i = 1, \ldots, d\}$ the criterion

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

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

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

where

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

4 Selection of significant variables

Consider the variable (or lag) selection problem for additive models. We assume that, for some index set $s_0 \subseteq \{1, \ldots, d\}$, the actual regression function $\mu(x) = E(Y_t|X_t = x), x = (x_1, \ldots, x_d)$, is an additive function in $x_i, i \in s_0$. If such an s_0 exists, for any s satisfying $s_0 \subseteq s \subseteq \{1, \ldots, d\}$, $\mu(x)$ 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.

Definition 4.1. *The index set s_0 of significant variables is the minimal set $s \subseteq \{1, \ldots, d\}$ such that $\mu \in H_s$. The variables $X_{t,i}, i \in s_0$, are called significant variables.*

Lemma 4.2. *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.*

Definition 4.3. *Let s be the index set of variables selected. We say that s overfits if $s > s_0$ but $s \neq s_0$, and we say that s underfits if $s_0 \subseteq s$. 'Overfitting' means that the set of selected variables includes other variables in addition to the significant variables; 'underfitting' means that the set of selected variables does not include all the significant variables.*

To define the variable selection criterion for additive models, we need, for each $s \subseteq \{1, \ldots, d\}$, an estimate of the regression function pretending that the index set of significant variables is s. Specifically, let 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 $g_i \in G_i$, where 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 G_i is $N_i = 1 + q_i + J_i$, $i = 1, \ldots, n$. The spline estimate corresponding to the index set s is

$$\hat{\mu}_s = \arg\min_{g \in G_s} \sum_{t=1}^{n} (Y_t - g(X_t))^2.$$

For each subset s of $\{1, \ldots, d\}$, define the mean-squared error of $\hat{\mu}_s$ as

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

and the BIC as

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

The variable selection rule is to select the subset $s \subseteq \{1, \ldots, d\}$ 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 that was defined in the previous section is consistent. It is difficult to obtain a reliable nonparametric estimate of the regression function at the tail of the distribution of X_i owing to sparseness of data. Thus, we focus on the estimation of μ_i on a compact set in our theoretical analysis below. Let C_i be a compact interval contained in the range of X_i and let C be the Cartesian product of $C_i, i = 1, \ldots, d$. We require that C_i contains all the interior knots for splines in $G_i, i = 1, \ldots, d$. We modify previous equations slightly to

$$\hat{\mu}_s = \arg\min_{g \in G_s} \sum_{l=1}^{n} \{Y_l - g(X_l)\}^2 I(X_l \in C)$$

and

$$\mathrm{MSE} = \frac{1}{n} \sum_{l=1}^{n} \{Y_l - \hat{\mu}_s(X_l)\}^2 I(X_l \in C),$$

where I is the indicator function.

We introduce some additional notation 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 $\{(X_i, Y_i)\}$ is defined as

$$\alpha(n) = \sup_{B \in \sigma((X_r, Y_r), r \leq t)} \sup_{C \in \sigma((X_r, Y_r), r \geq t+n)} |P(B \cap C) - P(B)P(C)|$$

for an index set T, where $\sigma((X_t, Y_t), t \in T)$ denotes the σ-field that is generated by the random variables $(X_t, Y_t) \in T$.

We make the following assumptions on the data-generating process:

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

(b) For some positive constants c_1 and c_2, the α-mixing coefficient of $\{(X_t, Y_t)\}$ satisfies

$$\alpha(n) \le c_1 n^{-(5/2)/(\nu-1)} \quad \text{and} \quad \alpha(n) \le c_2 n^{-2\nu/(\nu-2)}.$$

(c) The density p_{X_0} of X_0 is bounded away from 0 and ∞ on C.

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

Lemma 5.1. *Suppose each $\mu_i, i \in s_0$, has bounded second derivative. In addition, suppose that the degree of splines is 1 or bigger. Then a sufficient condition for (d) is that $\gamma > 1/5$.*

Theorem 5.2. *Suppose that assumptions (a)-(d) hold. The variable selection rule consistently selects the set of significant variables, i.e., $\lim_{n\to\infty} P(\hat{s} = s_0) = 1$.*

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 $\mu_i, i \in s_0$, 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 $\gamma \ge 1/5$. It is worthwhile to point out that, if $J_n \asymp n^{1/5}$, then $\|\hat{\mu}_s - \mu\| = O_P(n^{-4/5})$, which is the optimal rate of convergence.

6 Implementation

In actual implementation of the method proposed, we first decide on a set of candidate variables for selection. These candidates can include lagged variables of a time series or exogenous variables. The selection procedure consists of three stages: forward, backward, and final selection.

Forward Stage: Starting from a null model, where $Y_t = \mu_0 + \epsilon_t$ and μ_0 is a constant, we sequentially add variables to minimize the mean-squared error. This process continues until the number of variables reaches a predetermined maximum, denoted as $S_{\max}$.

Backward Stage: We begin with the maximal set of variables from the forward stage and sequentially remove variables, also minimizing the mean-squared error, until no variables remain.

Final Selection: The final selection from the set of models generated through this process is made by optimizing the Bayesian Information Criterion (BIC).

Let d be the total number of candidate variables. If d is not excessively large, $S_{\max}$ can be set to d. For each variable addition or removal, an additive spline model is fit, where spline knots are placed between the 5th and 95th percentiles of the data, and the number of knots is selected based on the formula $\left(\lfloor k(n)^{1/5}\rfloor\right)$ for linear splines and $\left(\lfloor k(n)^{1/5}\rfloor - 1\right)$ for quadratic and cubic splines, with k typically set to 2.

This stepwise procedure, although similar to methods used in linear regression for variable selection, differs significantly in that each addition or deletion of a variable corresponds to the manipulation of an additive component of the model, rather than merely adjusting coefficients in a linear combination.

7 Other regression spline model

To provide a clearer understanding of the context in which the authors compare their method to other regression spline models, we have conducted a detailed study and summary of the methods mentioned. This section outlines the key features and distinctions among the regression spline models used in the simulation study, emphasizing the comparative analysis performed by the authors.

7.1 Multivariate adaptive regression spline (MARS)

Friedman (1991) propose a model called Multivariate Adaptive Regression Splines (MARS). It is a non-parametric regression technique that extends linear models by incorporating automatic variable selection and

non-linearities. MARS is particularly effective for high-dimensional data and capable of modeling complex interactions between variables.

The MARS model builds its form by using basis functions, typically piecewise polynomials, combined in a way that captures interactions and non-linear relationships in the data. The general form of the MARS model can be expressed as:

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x) \tag{6}$$

where:

- $f(x)$ is the response variable.
- $B_m(x)$ are the basis functions which are products of hinge functions defined as $\max(0, x - c)$ or $\max(0, c - x)$, where c is a constant.
- β_m are coefficients estimated during the training process.
- M is the number of basis functions used in the model.

The model construction involves two key steps:

1. **Forward Pass:** Sequentially adds terms to the model to improve fit.
2. **Backward Pass:** Prunes terms using a criterion like cross-validation to prevent overfitting.

The main difference is in the backward fitting part, they used generalized cross-validation for pruning.

7.2 Adaptive Backfitting BRUTO Algorithm

The BRUTO (Bias Reduction in Unbiased Transformations) algorithm, developed in Friedman and Silverman (1989), is an enhancement of the Multivariate Adaptive Regression Splines (MARS) model. One of its key features is adaptive backfitting, which refines the model by iteratively adjusting basis functions and their coefficients to improve generalization and reduce overfitting.

Adaptive backfitting in BRUTO involves several key steps:

1. **Initialization**: Start with a set of basis functions from the forward pass of the MARS algorithm.
2. **Backfitting**: Iteratively adjust each basis function coefficient by refitting it to the residuals of the model excluding the current basis function. This is expressed as:

$$\hat{\beta}_j = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{k \neq j} \beta_k B_k(x_i) - \beta B_j(x_i) \right)^2$$

   where $B_j(x)$ is the j-th basis function, and β_j is its coefficient.
3. **Assessment and Adjustment**: Evaluate each term's contribution using a cost function that includes a complexity penalty. Remove terms that do not significantly improve model performance.
4. **Iteration**: Repeat the backfitting and adjustment until the model converges or no significant improvements are made.

The adaptive backfitting technique provides several benefits:

- Reduces the likelihood of overfitting by ensuring that each model term justifies its presence through rigorous validation.
- Enhances the model's ability to generalize to new data by focusing on robust, significant predictors.
- Maintains flexibility to model complex interactions and nonlinearities while controlling model complexity.

8 Simulation Results

In this section, we first review the findings presented in the original study. The authors implemented the proposed methodology and evaluated its performance against the Akaike Information Criterion (AIC). They observed that AIC tends to overfit more than the Bayesian Information Criterion (BIC), likely due to its lesser degree of penalization. Additionally, the proposed lag selection method was compared with the MARS and BRUTO algorithms, as documented in Chen and Tsay (1993), leading to insightful conclusions.

Following this review, we describe our efforts to replicate the study and contrast some of our findings with those reported in the original paper. Our experimentation with the data revealed that the effectiveness of the method heavily depends on the assumptions made about the model structure and the underlying data, which we found particularly intriguing.

8.1 Methods and Results from the Paper

The paper employed Monte Carlo simulations to assess the efficacy of the proposed method. The dynamics of the time series used in these simulations are defined as follows:

$$
\begin{aligned}
\text{AR1} \quad & Y_t = 0.5Y_{t-1} + 0.4Y_{t-2} + 0.1\xi_t, \\
\text{AR2} \quad & Y_t = -0.5Y_{t-1} + 0.4Y_{t-2} + 0.1\xi_t, \\
\text{AR3} \quad & Y_t = -0.5Y_{t-6} + 0.5Y_{t-10} + 0.1\xi_t, \\
\text{NLAR1} \quad & Y_t = -0.4\left(\frac{3 - Y_{t-1}^2}{1 + Y_{t-1}^2}\right) + 0.6\left(\frac{3 - (Y_{t-2} - 0.5)^3}{1 + (Y_{t-2} - 0.5)^4}\right) + 0.1\xi_t, \\
\text{NLAR2} \quad & Y_t = \left(0.4 - 2e^{-50Y_{t-6}^2}\right) Y_{t-6} + \left(0.5 - 0.5e^{-50Y_{t-10}^2}\right) Y_{t-10} + 0.1\xi_t, \\
\text{NLAR3} \quad & Y_t = \left(0.4 - 2\cos(40Y_{t-6})e^{-30Y_{t-6}^2}\right) Y_{t-6} + \left(0.55 - 0.55\sin(40Y_{t-10})e^{-10Y_{t-10}^2}\right) Y_{t-10} + 0.1\xi_t, \\
\text{NLAR1U1} \quad & Y_t = -0.4\left(\frac{3 - Y_{t-1}^2}{1 + Y_{t-1}^2}\right) + 0.1\xi_t, \\
\text{NLAR1U2} \quad & Y_t = 0.6\left(\frac{3 - (Y_{t-2} - 0.5)^3}{1 + (Y_{t-2} - 0.5)^4}\right) + 0.1\xi_t,
\end{aligned}
$$

where ξ_t are independent and identically distributed normal random variables $N(0, 1)$. These processes vary in the shape of the conditional mean function and the lag vector. The time series models AR1 through NLAR3 are adapted from Tschernig and Yang (2000) for evaluating their lag selection method.

Data were generated of size $n + 400$, with the last n observations used for analysis at sample sizes $n = 100, 200, 500$. This setup ensures that the data realizations are strictly stationary and geometrically β-mixing, thus meeting the model assumptions. For each (model, data_size) combination, 100 replications were generated for each of the described processes. Lag selection was conducted for each replication, searching within the range $\{1, \ldots, 10\}$. The frequencies of overfitting, correct fitting, and underfitting across all processes were documented in tables 1 and 2.

In this context, "underfitting" refers to failing to select all relevant variables rather than selecting too few variables. For example, at $n = 100$, there were 92 instances of underfitting using the AIC with spline fitting for the first autoregressive model (AR1). These instances often included more than two lagged variables, yet missed at least one of the critical lags ($t - 1$ or $t - 2$), essential in the data generating process.

8.2 Main Results

The primary findings of the simulation study are summarized as follows:

1. The spline fitting selection procedure using the Bayesian Information Criterion (BIC), referred to hereafter as the proposed method, demonstrates excellent performance and robustness across all simulated

Table 2. Simulation results for lag selection using spline fitting with the BIC or AIC†

Model	*n*	*Results for the BIC*									*Results for the AIC, degree = 1*		
		degree = 1			*degree = 2*			*degree = 3*					
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

†For each set-up, the first, second and third columns give respectively the numbers of underfitting, correct fitting and overfitting over 100 simulation runs.

Figure 1: Simulation Results from Paper - Comparison with AIC

processes. As the sample size increases from 100 to 200 and 500, the frequency of correct fits reaches or approaches 100%, confirming the asymptotic consistency of the method.

2. For the proposed method, employing linear, quadratic, or cubic splines yields similar outcomes, with some exceptions in models AR1, AR2, and NLAR2 at a sample size of $n = 100$.

3. Spline fitting with the Akaike Information Criterion (AIC) consistently leads to overfitting.

4. In the `mars()` function, the tuning parameter (denoted as `a`) specifies the cost per degree-of-freedom change. Setting $a = 2$ mirrors an AIC-like penalty, consistently resulting in substantial overfitting. Even with $a = \log(n)$, akin to a BIC-like penalty, overfitting is still prevalent, indicating that a values typically recommended ($2 \leq a \leq 4$) are insufficient.

5. The `bruto()` function also features a tuning parameter (denoted as `a`) for the cost per degree-of-freedom change. At the default value of $a = 2$, akin to an AIC penalty, BRUTO tends to overfit. When adjusted to $a = \log(n)$, BRUTO's performance varies significantly; it performs well for some processes but poorly for others (e.g., NLAR1U1 and NLAR1U2).

Table 3. Simulation results for lag selection using the MARS and BRUTO algorithms†

Model	*n*	*Results for MARS*						*Results for BRUTO*					
		$a=2$			$a=log(n)$			$a=2$			$a=log(n)$		
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
NLAR1U2	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 each set-up, the first, second and third columns give respectively the number of underfitting, correct fitting and overfitting over 100 simulation runs. The constant a specifies the cost per degree-of-freedom change.

Figure 2: Simulation Results from Paper - Other Methods

8.3 Replication Results

To validate the method and deepen our understanding of the empirical findings, we replicated the study described in Huang and Yang (2004) from scratch, focusing on the realization of the algorithm to select significant variables. The code is available here.

Following the methodology outlined in Huang and Yang (2004), we utilized R to simulate variable selection for the series previously described. The results using the Bayesian Information Criterion (BIC) are documented in table 1 for degree=3, with results for degrees 1 and 2 included in the appendix. Results using the Akaike Information Criterion (AIC) are shown in table 2.

Our replication confirms the findings reported in the original paper. However, we observed some intriguing outcomes when altering the simulation procedure or modifying the data generating processes:

1. We experimented with including all 10 lags, treating each basis as an individual variable. Although this approach seems counterintuitive, it produced surprisingly effective results, outperforming those obtained using the AIC. For instance, for the process AR2 at $n = 500$, we observed 65 correct selections and 35 cases of overfitting.

2. The effectiveness of the method heavily depends on the assumptions underlying the data generating processes. When we modified the coefficients in the NLAR2 model to $Y_t = \{-2\exp(-50Y_{t-6}^2)\}Y_{t-6} + \{0.5 - 0.5\exp(-50Y_{t-10}^2)\}Y_{t-10} + 0.1\xi_t$, even minor changes, such as altering the term related to Y_{t-6}, significantly impaired the model's ability to select relevant variables.

This experience underscores the critical importance of thoroughly examining assumptions when employing time series analysis methods. Throughout this quarter, our extensive engagement with statistical testing for various models has reaffirmed the significance of verifying theoretical assumptions, enriching our understanding and application of time series modeling.

Table 1: Replication Results - BIC criterion, degree=3

Model	n	num_underfit	num_correct	num_overfit
AR1	100	63	34	3
	200	4	96	0
	500	0	99	1
AR2	100	45	53	2
	200	8	91	1
	500	0	100	0
AR3	100	13	85	2
	200	1	94	5
	500	0	99	1
NLAR2	100	39	59	2
	200	1	97	2
	500	0	97	3
NLAR3	100	11	86	3
	200	1	96	3
	500	0	98	2
NLAR1U1	100	5	95	0
	200	2	98	0
	500	0	100	0
NLAR1U2	100	7	93	0
	200	0	100	0
	500	9	90	1

Table 2: Replication Results - AIC criterion, degree=3

Model	n	num_underfit	num_correct	num_overfit
AR1	100	16	37	47
	200	1	43	56
	500	0	49	51
AR2	100	22	33	45
	200	1	51	48
	500	0	51	49
AR3	100	8	22	70
	200	0	44	56
	500	0	36	64
NLAR2	100	5	34	61
	200	0	31	69
	500	0	34	66
NLAR3	100	0	19	81
	200	0	33	67
	500	0	31	69
NLAR1U1	100	30	48	22
	200	42	42	16
	500	30	49	21
NLAR1U2	100	34	32	34
	200	50	25	25
	500	55	8	37

References

Rong Chen and Ruey S Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308, 1993.

J. H. Friedman. Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, 19:1–141, 1991.

Jerome H. Friedman and Bernard W. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21, 1989.

Jianhua Z. Huang and Lijian Yang. Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(2):463–477, 2004. doi: 10.1111/j.1369-7412.2004.05500.x.

Rolf Tschernig and Lijian Yang. Nonparametric lag selection for time series. *Journal of Time Series Analysis*, 21(4):457–487, 2000.

Appendix

Table 3: Replication Results - BIC criterion, degree=1

Model	n	num_underfit	num_correct	num_overfit
AR1	100	59	38	3
	200	8	91	1
	500	0	99	1
AR2	100	54	46	0
	200	11	88	1
	500	0	99	1
AR3	100	17	81	2
	200	1	96	3
	500	0	100	0
NLAR2	100	53	46	1
	200	1	97	2
	500	0	100	0
NLAR3	100	13	82	5
	200	0	99	1
	500	0	100	0
NLAR1U1	100	3	97	0
	200	0	100	0
	500	0	100	0
NLAR1U2	100	6	94	0
	200	5	95	0
	500	9	91	0

Table 4: Replication Results - BIC criterion, degree=2

Model	n	num_underfit	num_correct	num_overfit
AR1	100	58	37	5
	200	5	94	1
	500	0	99	1
AR2	100	58	42	0
	200	8	91	1
	500	0	100	0
AR3	100	14	81	5
	200	0	98	2
	500	0	100	0
NLAR2	100	41	55	4
	200	3	95	2
	500	0	100	0
NLAR3	100	13	85	2
	200	0	100	0
	500	0	100	0
NLAR1U1	100	5	95	0
	200	0	100	0
	500	0	100	0
NLAR1U2	100	8	92	0
	200	8	92	0
	500	8	92	0