Time Series Model Selection via Adaptive Sparse Estimation

Seong-Tae Kim* Kendra C. Kirby†

Abstract

Model selection is a central agenda of autoregressive moving average (ARMA) modeling in time series data analysis. Recent advances in sparse estimation methods provide a fresh look at the time series model selection different from information criterion approaches. The adaptive LASSO method is paid attention in time series model selection due to its oracle property: the consistency of a set of non-zero parameters and its asymptotic normality. In spite of the solid theoretical property of adaptive LASSO method, this type is not a full-fledged method in time series analysis. This presentation will introduce a novel adaptive sparse method, the elastic net method, for time series model selection and investigate how the selection of initial estimates and tuning parameters in these adaptive sparse methods affects the performance of the time series model selection in various types of finite sample time series data. The performance will be assessed by examining the oracle property and the prediction accuracy. Comparison with other existing information criterion methods will be presented for both simulation studies and real data applications.

Key Words: ARMA, Sparsity, Adaptive LASSO, Adaptive Elastic Net, Model Selection

1. Introduction

Time series modeling requires a consideration of multiple aspects such as continuity, linearity, stationarity, seasonality, structural change, white noise error, and so on. Due to the complexity, classical time series modeling focuses on discrete-time, linear, stationary, no-change points, white noise error via autoregressive moving average (ARMA) model.

One of the most important modeling features in the ARMA model is a lag (model or variable) selection. Let consider an ARMA (p,q) model with seasonality, which can be modeled by an ARMA$(p,q)\times$SARMA$(P,Q)_{[s]}$ model given:

$$\begin{aligned}(1-\phi_1 L-\cdots-\phi_p L^p)(1-\Phi_1 L^s-\cdots-\Phi_P L^{sP})Y_t \\ = (1+\theta_1 L+\cdots+\theta_q L^q)(1+\Theta_1 L^s+\cdots+\Theta_Q L^{sQ})\varepsilon_t \end{aligned} \quad (1)$$

where p is the order of autoregressive terms, q is the order of moving average terms, P is the order of seasonal autoregressive terms, and Q is the order of seasonal moving average terms, s denotes the seasonal periodicity, *e.g.*, $s = 12$, is used for monthly data, and $\varepsilon_t \sim (0, \sigma^2)$. In this ARMA model, the dimension of the parameter space is determined by $(1, \phi_1, \cdots, \phi_p) \times (1, \Phi_1, \cdots, \Phi_P) + (1, \theta_1, \cdots, \theta_q) \times (1, \Theta_1, \cdots, \Theta_Q)$. Classical model selection chooses a model of which information criterion value attains the minimum among all possible subsets. As well known, there are numerous information criterion methods such as AIC (Akaike, 1973), BIC (Schwarz, 1978)

*North Carolina Agricultural and Technical State University, 1601 E. Market Street, Greensboro, NC 27411

†North Carolina Agricultural and Technical State University, 1601 E. Market Street, Greensboro, NC 27411

and their variants. If we exhaustively consider all possible subsets for a parameter space of p elements, there are 2^p subsets. Needless to say, the fairly large number of elements computationally costly for the subset selection approach.

Since Tibshirani's seminal method, *least absolute shrinkage and selection operator* (LASSO), in 1996, many sparse (shrinkage or regularized) estimation methods have been proposed to address the variable selection issue in a regression model. As we encounter the situation of 'large p, small n ($p >> n$) in the era of big data, this sparse estimation method is significantly useful to estimate nonzero parameters. The LASSO solution takes an advantage of a convex optimization in which a local optimum is not a problem. LASSO has two drawbacks: (1) the LASSO estimator does not achieve the oracle property under a certain condition; and (2) this estimator is suffered from multicollinearity among explanatory variables.

In order to address the oracle property issue of LASSO, Adaptive LASSO (ALASSO) was proposed by Zou (2006). ALASSO introduced the weight for each parameter in addition to the global tuning parameter, λ_T in LASSO. In time series analysis, LASSO and ALASSO have been applied for model selection in ARMA model. Nardi and Rinaldo (2011) applied LASSO to AR(p) model, and Chan, Yau, and Zhang (2015) applied LASSO to a threshold autoregressive model. Chen and Chan (2011) applied ALASSO to ARMA (p, q) model. Park and Sakaori (2013) applied ALASSO to autoregressive distributed lag (ADL) model. Kock(2016) extended the application of ALASSO to AR(p) with a unit root. Chan, Yau, and Zhang (2014) applied group LASSO to identify structural changes in time series. Although these articles successfully applied LASSO-type sparse estimation method, a lot of work is demanded to establish consensus of choosing the global tuning parameter, the penalty of individual parameters, the degree of weight, etc. The Elastic Net (ENET) and Adaptive ENET (AENET) methods were proposed for multicollinearity by Zou and Hastie (2005) and Zou and Zhang (2009), respectively. ENET and AENET have been applied to time series ARMA models yet.

The purpose of this article is to investigate the variable selection in the ARMA model via sparse regression methods. The sparse estimation methods including LASSO, ALASSO, ENET, and AENET, will be compared to traditional information criterion methods under various simulation settings such as different types of parameter spaces, choice of the global tuning parameter, choice of the penalty of individual parameters, degree of penalty weight, and signal to noise ratio.

In section 2, an AR model and its ALASSO estimator will be introduced. In section 3, the finite sample properties will be examined via Monte Carlo simulation studies. In section 4, simulation result will be presented for three time series models. In section 5, simulation results will discussed along with theoretical properties. Lastly, in section 6, conclusion and future studies will be provided.

2. Sparse Estimation Methods in Time Series

Consider the model selection problem in a stationary AR(p) model, which allows seasonal autoregressive terms:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \tag{2}$$

Through this paper, we discuss the model selection of an AR model which is introduced for a model simplicity. The same points can be easily extended to the ARMA

model as in Chen and Chan (2011). In this section, we first discuss the properties of parameter space in the ARMA model, which bears more restrictions compared to the parameter space of a regular linear regression model. Consequently, we introduce selected sparse estimation methods for the lag selection in (2), and discuss their statistical properties.

2.1 Properties of Parameter Space in Time Series

In a regular linear regression, the parameter space related issues may include the space length such as $p < n, p > n$, the sparsity determined the number of zero coefficients, and the multicollinearity among independent variables. In a time series regression, serial correlation and stationarity pose some critical restriction on the parameter space. In this section, we discuss the properties of parameter space in time series model.

In order to specify the parameter space of an AR mode, let consider a stationary AR(2)$\times$ SAR(1)$_{[12]}$ model with $\phi_1 = 0.6, \phi_2 = 0.3$, and $\Phi_1 = 0.7$. Using the representation in (1) yields

$$(1 - 0.6L - 0.3L^2)(1 - 0.7L^{12})Y_t = \varepsilon_t \tag{3}$$

Rearranging equation (3) in terms of a linear regression model yields

$$Y_t = 0.6Y_{t-1} + 0.3Y_{t-2} + 0.7Y_{t-12} - 0.42Y_{t-13} - 0.21Y_{t-14} + \varepsilon_t. \tag{4}$$

The first noticeable property is a sparse parameter space. For this example, the nonzero parameter space of the AR process can be defined as $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_{12}, \phi_{13}, \phi_{14})$. However, the estimation process of this parameter space in time series analysis intrinsically requires the consideration of all lagged variables between Y_{t-1} and Y_{t-14} of which parameter space is $\boldsymbol{\phi} = \{0.6, 0.3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.7, -0.42, -0.21\}$.

Another important issue to this parameter space naturally arises when one stops at the lag of 14. If all lagged values after Y_{t-14} have no effect on Y_t, then the length of the parameter space is obviously 14. In a regular regression model, the number of independent variables which are included in the model are predetermined. The maximal length of the parameter space, p^*, however, is usually unknown so that we appropriately determine this value in a time series regression model. The value of p^* should be fairly large in order not miss any important lagged effects. On the other hand, the maximal value of p cannot be greater than the sample size T.Hence, there are two different choices for the maximal value of p, p^*. The first choice is the fixed value, and the second choice is the function of sample size such that $p^* = T^\nu,\ o < \nu < 1$. In general, the growth rate of p^* is much slower than the growth rate of T.

The serial correlation in time series data is an underlying feature, which is closely related to multicollinearity. Therefore, it is natural to consider ENET and AENET to achieve a good selection performance without losing valuable information.

The last issue of the time series parameter space we can come up with is stationarity. Stationarity is one of the most salient properties in the time series parameter space (Dickey and Fuller, 1979). In the regular linear regression model, we can select fairly large coefficient values which result in the better variable selection. However, the stationarity condition does not allow arbitrarily large coefficient values in the

ARMA model. This property costs the performance of sparse regression methods as pointed out in Tibshirani (1996).

2.2 Selected Sparse Estimation Methods

In this article, four significant sparse estimation methods are selected for the lag selection of the AR model. The LASSO is the breakthrough sparse estimation method with L_1 penalty. The other three sparse methods were proposed to overcome the limitations of LASSO.
The LASSO estimator is the minimizer of the following objective function

$$\widehat{\boldsymbol{\phi}}^{LASSO} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left\{ ||Y - \mathbf{X}\boldsymbol{\phi}||_2^2 + \lambda_T \sum_{j=1}^{p^*} |\phi_j| \right\} \tag{5}$$

where $Y = (Y_p, \cdots, Y_T)$ is a $(T-p) \times 1$ vector, $X = (Y_{-1}, \cdots, Y_{-p})$ is a $(T-p) \times p$ matrix with $Y_{-j} = (Y_{p-j+1}, Y_{p-j+2}, \cdots, Y_{T-j})$ for $j = 1, 2, \cdots, p$, λ_T is the global tuning parameter controlling the degree of sparsity of parameters.
Zou (2006) pointed out that the LASSO estimator does not achieve the consistency to the true parameter space if $\lambda_T = O(T^{1/2})$. Consequently, the LASSO estimator does not hold the oracle property that an estimator attains the consistency and asymptotic normality. Zou proposed the adaptive LASSO (ALASSO) estimator to overcome the inconsistency of the LASSO estimator. The adaptive LASSO estimator, $\hat{\boldsymbol{\phi}}^{ALASSO}$, is obtained by minimizing the following objective function with constraint:

$$\widehat{\boldsymbol{\phi}}^{ALASSO} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left\{ ||Y - X\boldsymbol{\phi}||_2^2 + \lambda_T \sum_{j=1}^{p^*} \lambda_j |\phi_j| \right\} \tag{6}$$

where $\lambda = (\lambda_1, \cdots, \lambda_p)$ is a $p \times 1$ vector of known data-driven weights. In the weights $\lambda = |\hat{\phi}|^{-\eta}$, $\hat{\phi}$ can be the least squares estimator (Zou, 2006 and Chen and Chan, 2011), the ridge estimator (Chen and Chan, 2011), or the lasso estimator (Chen and Chan, 2011), and $\eta > 0$. If $\lambda_j = 1$ for all j, then the solution in (6) is the LASSO estimator for $\boldsymbol{\phi}$.
The LASSO estimator incorporates an $L_1$1penalty, which achieves a sparsity of parameter estimation, but does not deal with multicollinearity well (Chun and Keles (2009), Zou and Zhang (2009)). In order to address both the sparsity of the parameter space and the multicollinearity of explanatory variable, Zou and Hastie (2005) proposed the elastic net estimator (ENET):

$$\widehat{\boldsymbol{\phi}}^{ENET} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left\{ \frac{1}{2} ||Y - X\boldsymbol{\phi}||_2^2 + \lambda_T \left[\frac{1}{2}(1-\alpha) \sum_{j=1}^{p^*} \phi_j^2 + \alpha \sum_{j=1}^{p^*} |\phi_j| \right] \right\} \tag{7}$$

in which $\alpha \in [0, 1]$, and the penalty term is a convex combination of L_1 and L_2 penalties. When $\alpha = 1$, the ENET problem becomes the LASSO penalty estimation, and $\alpha = 0$ makes it the ridge estimation problem. Since ENET did not achieve the oracle property, Zou and Zhang (2009) proposed the adaptive elastic net (AENET)

method to simultaneously address the oracle property and the multicollinearity. The AENET estimator is the solution to the following objective function

$$\widehat{\boldsymbol{\phi}}^{AENET} = \left(1 + \frac{\lambda 2}{T}\right) \times \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \left\{ ||Y - X\boldsymbol{\phi}||_2^2 + \lambda_{2T} \sum_{j=1}^{p^*} \phi_j^2 + \lambda_{1T} \sum_{j \in \widehat{\mathcal{A}}^{ENET}} \lambda_j |\phi_j| \right\} \tag{8}$$

where $\widehat{\mathcal{A}}^{ENET} = \{j : \widehat{\boldsymbol{\phi}}^{ENET} \neq 0\}$, which is the set of non-zero ENET estimators.

2.3 Asymptotic Properties of Adaptive LASSO Estimator

In the stationary ARMA model, the oracle property of ALASSO has been proven by Chen and Chan (2011). Let $\mathcal{A}$ be the set of nonzero parameters such as $\mathcal{A} = \{j : \phi_j \neq 0\} = \text{supp}(\boldsymbol{\phi}) \subset I = \{1, 2, \cdots, \text{p}\}$, and let $\hat{\mathcal{A}}$ be the set of nonzero parameter estimators such as $\hat{\mathcal{A}} = \{j : \hat{\phi}_j \neq 0\}$.
Statistical assumptions are as follows:

A1. For AR(p) model in (2), the lag polynomial, $\phi(z)$ has the roots outside the unit circle where $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$.

A2. Innovations satisfies that for all t $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, and $E(\varepsilon_t^{2+\delta}) < \infty$ for some $\delta > 0$.

A3. The maximal lag order, p^*, increases to infinity at a rate $c \log T \leq p^* \leq (\log T)^b$ for some $1 < b < \infty$ and for some positive constant c.

A4. For $\eta > 0$, the global tuning parameter increases to infinity with

$$\lambda_T T^{(\eta-1)/2} \to \infty \quad and \quad \lambda_T T^{-1/2} \to 0. \tag{9}$$

A5. As $T \to \infty$, $\frac{1}{T} X^T X \to \mathbf{C}$ where $\mathbf{C}$ is a positive definite nonrandom matrix. Similary, for $\mathcal{A}$, as $T \to \infty$, $\frac{1}{T} X_{\mathcal{A}}^T X_{\mathcal{A}} \to \mathbf{C}_{\mathcal{A}}$.

A6. As $T \to \infty$, $\frac{\boldsymbol{\varepsilon}^T \mathbf{X}}{\sqrt{T}} \underset{d}{\longrightarrow} N(\mathbf{0}, \sigma^2 \mathbf{C})$.

Assumptions A1 and A2 allows to consider only a stationary AR(p) process with a constant autocovariance. Assumption A3 imposes the growth rate of the autoregression order to be bounded by the sample size (T). The rate should not be neither too fast nor too slow compared to T. Assumption 4 is critical to attain the asymptotic consistency of the adaptive LASSO estimator. Zou (2006) showed that if $\lambda_T T^{-1/2} \to \lambda_0 < \infty$, the LASSO estimator is not consistent as $T \to \infty$. In variable selection, a proposed method is theoretically evaluated by the oracle property which consists of two parts: the consistency and asymptotic normality of the estimator obtained by the proposed estimator (Fan and Li, 2001).

Theorem 2.1. *When Assumption A1-A6 hold for model in (2), the oracle property of the adaptive LASSO estimator is attained as follows:*
1. Asymptotic normality

$$\sqrt{T}(\hat{\boldsymbol{\phi}}^{ALASSO} - \boldsymbol{\phi}) \underset{d}{\longrightarrow} N(\mathbf{0}, \sigma^2 \mathbf{C}_{\mathcal{A}}^{-1}) \; as \; T \to \infty \tag{10}$$

2. Consistency

$$lim_{T\to\infty} P(\hat{\mathcal{A}} = \mathcal{A}) = 1. \tag{11}$$

Theorem 2.1 for AR(p) model can be proved by following Chen and Chan (2011) and Zou (2006).

3. Simulation Studies

In the previous section, we introduced sparse estimation methods which can be applied to time series data, and reviewed the oracle property of the adaptive LASSO estimator. Time series modeling is usually implemented with finite sample sizes, few hundred to few thousand observations. Our objective is to compare the selected sparse estimation methods to the traditional variable selection methods such as AIC and BIC.

The finite sample properties of these estimation methods are especially affected by various simulation settings such as the sample size, T, the maximal number of parameters included in the ARMA model, p^*, the global tuning parameter, λ_T, λ_{1T} or λ_{2T}, weights of individual parameters, λ_j, degree of weights, η, signal to noise ratio by innovation variance.

The global tuning parameter λ_T (or λ_{1T} and λ_{2T} in AENET) can be chosen by two different approaches: (1) λ_T is determined as a function of p and T; and (2) as a minimizer of BIC, AIC or time series cross-validation. Nardi and Rinaldo (2011) chose the value of λ_T as a function of the sample size and the maximal number of coefficients such as $\lambda_T = \sqrt{\frac{\log T \log p}{T}}$, which is used for the L_1 penalty in LASSO and ENET. BIC is used to determine the data-driven value of the tuning parameters as many other suggested: LASSO in Zou et al. (2007), SCAD in Wang et al. (2007), and AENET in Zou and Zhang (2009).

The weights of individual parameters, λ_j, chose root-n consistent estimators such as least squares estimator (LS), ridge regression estimator (Ridge), and LASSO estimator. In the adaptive methods, ALASSO and AENET, these three types of estimators are considered.

The degree of weights, η, must be greater than zero, but we chose $\eta = 1$ or 2 followed by others' convention (Chen and Chan, 2011, Zou, 2006, and Zou and Zhang 2009). Of course, one can choose the data-drive value of η but it may cause another uncertainty.

To investigate the finite sample property, we chose two different sample sizes, $T =$ 200 and 1000. The innovation error, ε_t was generated from the standard normal distribution, $\varepsilon \sim iidN(0, 1)$.

The LASSO-type estimation is known to have a strength to identify a sparse parameter space. We considered three different AR(p) models with sparse space with weak coefficients, moderately sparse space with strong coefficients, and dense parameter spaces with mild coefficients:

1. $Y_t = 0.2Y_{t-1} + 0.1Y_{t-3} + 0.2Y_{t-5} + 0.3Y_{t-10} + 0.1Y_{t-15} + \varepsilon_t$
2. $Y_t = 0.8Y_{t-1} + 0.7Y_{t-5} - 0.56Y_{t-6} + \varepsilon_t$
3. $Y_t = 0.3Y_{t-1} + 0.25Y_{t-2} + 0.2Y_{t-3} + 0.15Y_{t-4}$

The first model is designed for a sparse and weak parameter space, $\boldsymbol{\phi} = (0.2, 0, 0.1, 0, 0.2, 0, 0, 0, 0, 0.3, 0, 0, 0, 0, 0.1)$ as in Nardi and Rinaldo (2011). The second model is designed for a moderately sparse parameter space, $\boldsymbol{\phi} = (0.8, 0, 0, 0, 0, 0.7, -0.56)$ (Chen and Chan 2011). The third model is designed for a dense parameter space, $\boldsymbol{\phi} = (0.3, 0.25, 0.2, 0.15)$. In the actual estimation process, additional sparsity will be included depending on the maximal length of p.
Under various situations mentioned above, the sparse estimation methods were compared to the AR models selected by BIC and AIC. All simulation work was performed using R packages, glmnet, gcdnet, and FitAR. Computing resource is Dell workstation equipped with dual core Intel Xeon CPU E5-2697 vs 2.60 GHz CPU and 128GB ram.

4. Finite Sample Properties via Simulation Studies

In this conference proceedings, simulation results for the three AR models with the sample size 200 are shown as a comparison pivot. Other cases are compared to these pivot results. For all tables presented in this article, the first column represents each of the selected methods: LASSO, ALASSO, ENET, AENET, BIC, and AIC. The second and third columns (L1 and L2) is for the selection methods for L_1 and L_2 penalty: Fix or BIC where FIX means $\lambda_T = \sqrt{\frac{\log T \log p}{T}}$. The fourth and fifth columns (P1 and P2) is for the selection methods for the initial weight of individual parameters: OLS, Ridge, LASSO, and ENET. Our main interest in the presented tables is to see how correctly the non-zero coefficients are captured by the selected estimation methods. Hence, the presentation focuses on the true positive (TN), the true negative (TN), the false positive (FP), and the false negative (FN), as well as, the sensitivity and the specificity.

Table 1: Simulation Results for $\boldsymbol{\phi} = (0.2, 0, 0.1, 0, 0.2, 0, 0, 0, 0, 0.3, 0, 0, 0, 0, 0.1)$, $T = 200$, $p^* = 15$

Method	L1	L2	P1	P2	TP	TN	FP	FN	Sensitivity	Specificity
LASSO	FIX				2.59(1.17)	9.7(0.68)	0.3(0.68)	2.41(1.17)	0.52(0.23)	0.97(0.07)
LASSO	BIC				4.86(0.36)	1.40(1.12)	8.6(1.12)	0.14(0.36)	0.97(0.07)	0.14(0.11)
ALASSO	BIC		OLS		2.05(0.81)	9.95(0.22)	0.05(0.22)	2.95(0.81)	0.41(0.16)	1.00(0.02)
ALASSO	BIC		RIDGE		2.53(0.93)	9.85(0.42)	0.16(0.42)	2.47(0.93)	0.51(0.19)	0.98(0.04)
ALASSO	BIC		LASSO		1.89(0.81)	9.96(0.19)	0.04(0.19)	3.11(0.81)	0.38(0.16)	1.00(0.02)
ENET	FIX	BIC			2.60(1.18)	9.70(0.69)	0.30(0.69)	2.40(1.18)	0.52(0.24)	0.97(0.07)
ENET	BIC	BIC			4.87(0.36)	1.40(1.10)	8.60(1.10)	0.13(0.36)	0.97(0.07)	0.14(0.11)
AENET	BIC	BIC	OLS	ENET	2.10(0.82)	9.95(0.22)	0.05(0.22)	2.90(0.82)	0.42(0.16)	0.99(0.02)
AENET	BIC	BIC	RIDGE	ENET	2.60(0.95)	9.83(0.43)	0.17(0.43)	2.41(0.95)	0.52(0.19)	0.98(0.04)
AENET	BIC	BIC	LASSO	ENET	1.96(0.83)	9.96(0.19)	0.04(0.19)	3.04(0.83)	0.39(0.17)	1.00(0.02)
BIC					2.75(1.12)	9.65(0.66)	0.36(0.66)	2.25(1.12)	0.55(0.22)	0.96(0.07)
AIC					3.84(0.81)	8.01(1.46)	1.99(1.46)	1.16(0.81)	0.77(0.16)	0.80(0.15)

NOTE: TP=True Positive, TN = True Negative, FP=False Positive, FN=False Negative, Sensitivity=TP/(TP+FN), Specificity=TN/(TN+FP); Mean(SD) represents the mean and the standard deviation obtained from 1000 iterations.

In Table 1, the number of true non-zero parameter is 5. As the mean values of TP is closer to 5, the selected model is considered as a better model. At the same time, as the mean value of FP is closer to 0, the selected model is considered as

Table 2: Simulation Results for $\phi = (0.8, 0, 0, 0, 0, 0.7, -0.56)$, $T = 200, p^* = 15$

Method	L1	L2	P1	P2	TP	TN	FP	FN	Sensitivity	Specificity
LASSO	FIX				2.00(0.05)	11.02(0.81)	0.98(0.81)	1.00(0.05)	0.67(0.02)	0.92(0.07)
LASSO	BIC				3.00(0)	3.86(1.6)	8.14(1.6)	0(0)	1.00(0)	0.32(0.13)
ALASSO	BIC		OLS		2.99(0.08)	11.98(0.16)	0.02(0.16)	0.01(0.08)	1.00(0.03)	1.00(0.01)
ALASSO	BIC		RIDGE		3.00(0)	11.02(1.01)	0.98(1.01)	0(0)	1.00(0)	0.92(0.08)
ALASSO	BIC		LASSO		2.99(0.1)	11.99(0.08)	0.01(0.08)	0.01(0.1)	1.00(0.03)	1.00(0.01)
ENET	FIX	BIC			2.00(0.03)	10.87(0.84)	1.13(0.84)	1.00(0.03)	0.67(0.01)	0.91(0.07)
ENET	BIC	BIC			3.00(0)	3.34(1.45)	8.66(1.45)	0(0)	1.00(0)	0.28(0.12)
AENET	BIC	BIC	OLS	ENET	2.99(0.1)	11.95(0.24)	0.05(0.24)	0.01(0.1)	1.00(0.03)	1.00(0.02)
AENET	BIC	BIC	RIDGE	ENET	3.00(0)	10.48(1.4)	1.52(1.4)	0(0)	1.00(0)	0.87(0.12)
AENET	BIC	BIC	LASSO	ENET	2.98(0.13)	11.99(0.12)	0.01(0.12)	0.02(0.13)	0.99(0.04)	1.00(0.01)
BIC					3.00(0)	11.42(1.08)	0.59(1.08)	0(0)	1.00(0)	0.95(0.09)
AIC					3.00(0)	9.33(1.88)	2.67(1.88)	0(0)	1.00(0)	0.78(0.16)

NOTE: TP=True Positive, TN = True Negative, FP=False Positive, FN=False Negative, Sensitivity=TP/(TP+FN), Specificity=TN/(TN+FP); Mean(SD) represents the mean and the standard deviation obtained from 1000 iterations.

a better model. The higher the sensitivity and specificity, the better the selected model. AIC achieved the relatively high true positive rate, but it was suffered from a high false positive rate. LASSO and ENET with the fixed tuning parameter and AENET with the weight of the ridge estimator achieved similar TP and FP to BIC. They achieved a high specificity (0.52) and a low sensitivity (0.97). The sensitivities of AENET with OLS and LASSO LASSO are worse than those for ALASSO. and ENET estimators with the selection of λ_T by BIC are suffered from high false positive rates.

In Table 2, the number of true non-zero parameter is 3. As the mean values of TP is closer to 3, the selected model is considered as a better model. At the same time, as the mean value of FP is closer to 0, the selected model is considered as a better model. Unlike the result in Table 1, the overall performance of the selected methods is pretty impressive. The result in 2 shows that sparse estimation methods such as ALASSO and AENET outperform the information criterion methods. AIC achieved the relatively high true positive rate, but it was suffered from a high false positive rate. ALASSO and AENET with the weight of the OLS and LASSO estimators achieved the best performance as their sensitivity and specificity are close to one. BIC achieved the highest sensitivity but the slightly low specificity. Again, LASSO and ENET estimators with the selection of $\lambda)T$ by BIC are suffered from high false positive rates.

In Table 3, all four true parameters are non-zero parameter. As the mean values of TP is closer to 4, the selected model is considered as a better model. At the same time, as the mean value of FP is closer to 0, the selected model is considered as a better model. LASSO and ENET with the fixed tuning parameter achieved the best performance as their sensitivity and specificity are over 90%. LASSO and ENET with the BIC tuning parameter were suffered from the low specificity below 20%. ALASSO and AENET with the weight of the Ridge estimator and BIC achieved a similar performance as their specificity was almost perfect but their sensitivity was below 80% are close to one.

Table 3: Simulation Results for $\boldsymbol{\phi} = (0.3, 0.25, 0.2, 0.15)$, $T = 200$, $p^* = 15$

Method	L1	L2	P1	P2	TP	TN	FP	FN	Sensitivity	Specificity
LASSO	FIX				3.69(0.53)	10.46(0.88)	0.54(0.88)	0.31(0.53)	0.92(0.13)	0.95(0.08)
LASSO	BIC				3.96(0.20)	2.12(1.36)	8.88(1.36)	0.04(0.20)	0.99(0.05)	0.19(0.12)
ALASSO	BIC		OLS		2.38(0.65)	10.95(0.23)	0.06(0.23)	1.62(0.65)	0.59(0.16)	1.00(0.02)
ALASSO	BIC		RIDGE		2.95(0.70)	10.69(0.54)	0.31(0.54)	1.05(0.70)	0.74(0.17)	0.97(0.05)
ALASSO	BIC		LASSO		2.33(0.63)	10.97(0.17)	0.03(0.17)	1.67(0.63)	0.58(0.16)	1.00(0.02)
ENET	FIX	BIC			3.71(0.52)	10.43(0.92)	0.57(0.92)	0.29(0.52)	0.93(0.13)	0.95(0.08)
ENET	BIC	BIC			3.96(0.20)	2.13(1.36)	8.88(1.36)	0.04(0.20)	0.99(0.05)	0.19(0.12)
AENET	BIC	BIC	OLS	ENET	2.56(0.70)	10.92(0.27)	0.08(0.27)	1.44(0.70)	0.64(0.18)	0.99(0.02)
AENET	BIC	BIC	RIDGE	ENET	3.14(0.71)	10.59(0.63)	0.41(0.63)	0.86(0.71)	0.78(0.18)	0.96(0.06)
AENET	BIC	BIC	LASSO	ENET	2.48(0.70)	10.95(0.21)	0.05(0.21)	1.52(0.70)	0.62(0.17)	1.00(0.02)
BIC					3.17(0.67)	10.63(0.67)	0.37(0.67)	0.83(0.67)	0.79(0.17)	0.97(0.06)
AIC					3.60(0.52)	9.22(1.40)	1.78(1.40)	0.41(0.52)	0.90(0.13)	0.84(0.13)

NOTE: TP=True Positive, TN = True Negative, FP=False Positive, FN=False Negative, Sensitivity=TP/(TP+FN), Specificity=TN/(TN+FP); Mean(SD) represents the mean and the standard deviation obtained from 1000 iterations.

5. Discussion

The results in the previous section showed if the coefficients in the model are strong (different from zero), then ALASSO and AENET outperformed AIC and BIC in variable selection. The sparse estimation methods achieved low (good) false positive rates but were suffered from high (bad) false negative rates. In other words, the sparse estimation methods tend to excessively exclude true non-zero coefficients. This result reflects the characteristic of sparse estimation methods, which shrink some coefficient values to zero. Thus, when a true value of a coefficient is close to zero, there is a high chance for this to be zero, which depends on the selected value of the tuning parameter in the estimation process.

AIC and BIC showed similar performance in the selection of tuning parameters, which was also reported in Chen and Chan (2011). The L_1 penalty selected by BIC resulted in the least favorite performance. The results of ALASSO and ANENT with the BIC tuning parameters performed with respect to the sensitivity and specificity well. The cross-validation (CV) and the generalized cross-validation (GCV) may be considered to select the tuning parameters. Nardi and Rinaldo (2011) demonstrated the choice of λ_T using CV in 'lars' package. However, one cannot randomly select some values from time series data for k-fold CV since this scheme ignores serial dependency of the time series (Bergmeir, Hyndman, and Koo, 2015). Therefore, the CV for time series data is different from the CV for a regular regression model.

Regarding the weight for the individual variable, the OLS and LASSO estimators showed similar results. However, the ridge estimator outperformed for the sparse, weak parameter space as can be seen in Table 1. ALASSO and AENET showed similar performance for all three models. The selected tuning parameter for L_2- penalty by BIC was very close to zero. The ENET estimator was used as the initial weight for the individual variables in the AENET estimation. In 'gcdnet' package, one can choose two penalty functions for the initial weight in AENET. We used OLS, Ridge, and LASSO estimates as the initial weight for the L_1- penalty term and ENET estimates for L_2-penalty term. Zou and Zhang (2009) used the ENET for the initial weight of L_1-penalty. We need further investigation on the role ENET estimates for the L_1-penalty.

The maximal number of parameters included in the model, p^*, generates a sparse parameter space. The choice of p^* is usually depends on the sample size, T. In the simulation study, we chose fixed values: $p^* = 15$ and 30 for $T = 200$ and 1000. As we increased p^*, the performance became slightly worse. Nardi and Rinald (2011) assumed that $p^* = o(T^{1/2})$ and Chen and Chan (2011) assumed that $p^* = 10 \log_{10} T$. The effect of p^* definitely needs more studies since a choice of p^* affects selection performance and computing time.

Although we did not presented the results with the sample size 1000, a larger sample size demonstrated better performance for all three models. The effects of the signal-to-noise is not addressed in this article. However, we expect that a smaller value of σ in the white noise term will result in better performance.

6. Conclusion

In this study, we investigated the variable selection via sparse estimation methods in the AR model. This study showed that the sparse estimation methods could be an alternative solution to the information criterion methods. In particular, the performance of ALASSO and AENET was better than or similar to the BIC performance. When the coefficients are quite distant from zero, the sparse estimation methods clearly outperformed the information criterion methods.

Despite of the interesting results, this study bears several limitations. First, simulation studies mainly relied on R packages 'glmnet' and 'gcdnet'. In order to fairly compare the selected methods, our own programming is desirable. Second, the cross-validation is a common method for the choice of the global tuning parameter. However, the time series data cannot be randomly shuffled for a cross-validation. Several time series cross-validation methods have been proposed, but their theoretical properties are not fully exploited. Cross-validation is adopted for the sparse estimation methods in time series analysis yet, which deserves a future study. Another future study will focus on the forecasting performance of the sparse estimation methods.

Future studies can be performed in several directions. First, the adaptive elastic net estimator in the ARMA model needs be investigated for their oracle property along with time series assumptions. Second, the selection methods of tuning parameters need more close investigation. In particular, the time series cross-validation methods should be compared to BIC. Third, the performance of selected methods need be compared regarding the forecasting performance in addition to sensitivity and specificity.

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Càski, Budapest: Akademiai Kiàdo, 267-281.

Bergmeir, C., Hyndman, R. J., and Koo, B., (2015), "A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction, *Monash University Department of Econometrics and Business Statistics Working Paper*, 10, 15.

Chan, N. H., Yau, C. Y., and Zhang, R. M., (2015), "LASSO estimation of threshold autoregressive models," *Journal of Econometrics*, 189(2), 285–296.

Chan, N.H., Yau, C.Y. and Zhang, R.M., (2014), "Group LASSO for structural break time series," *Journal of the American Statistical Association*, 109(506), 590–599.

Chen, K., and Chan, K.(2011), "Subset ARMA selection via the adaptive Lasso," *Statistics and its Interface*, 4, 197–205.

Chun, H., and Keles, S. (2009), " Expression quantitative trait loci mapping with multivariate sparse partial least squares regression." *Genetics*, 182(1), 79–90.

Dickey, D. A., and Fuller, W. A., (1979), "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, 74(366a), 427–431.

Fan, J., and Li, R., (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, 96(456), 1348–1360.

Knight, K., and Fu, W., (2000), "Asymptotics for lasso-type estimators," *Annals of statistics*, 28:5, 1356–1378.

Kock, A. B., (2016), "Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions," *Econometric Theory* 32:1, 243–259.

Nardi, Y., and Rinaldo, A., (2011), "Autoregressive process modeling via the lasso procedure," *Journal of Multivariate Analysis*, 102:3, 528–549.

Park, H., and Sakaori, F., (2013), "Lag weighted lasso for time series model," *Computational Statistics*, 28:2, 493–504.

Schwarz, G., (1978), "Estimating the dimension of a model," *The annals of statistics*, 6(2), 461–464.

Tibshirani, R., (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:1, 267-288.

Wang, H., Li, G. and Tsai, C.L., (2007), "Regression coefficient and autoregressive order shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1), 63–78.

Zou, H., (2006), "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 101:476, 1418-1429.

Zou, H., and Hastie, T., (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

Zou, H., Hastie, T. and Tibshirani, R., (2007), "On the "degrees of freedom" of the lasso," *The Annals of Statistics*, 35(5), 2173–2192.

Zou, H., and Zhang, H.H., (2009), "On the adaptive elastic-net with a diverging number of parameters," *Annals of statistics*, 37:4, 1733-1751.