Statistics and Its Interface Volume 4 (2011) 137–148

Threshold variable selection via a L_1 penalty approach*

Qian Jiang and Yingcun Xia†

Selecting the threshold variable is a key step in building a general threshold autoregressive (TAR) model. Based on a general smooth threshold autoregressive (STAR) model, we propose to select the threshold variable by the recently developed L_1-penalizing approach. Moreover, by penalizing the direction of the coefficient vector instead of the coefficients themselves, the threshold variable is more accurately selected. Oracle properties of the estimator are obtained. Its advantage is shown with both numerical and real data analysis.

Keywords and phrases: Smooth threshold AR model, Variable selection, Adaptive lasso, Oracle property.

1. INTRODUCTION

Tong's threshold autoregressive (TAR) model (see, e.g., Tong and Lim, 1980) is one of the most popular models in the analysis of time series in biology, finance, economy and many other areas. It assumes different AR models in different regions of the state space divided according to some threshold variable y_{t-d}, $d \geq 1$. A typical two-regime threshold autoregressive (TAR) model is

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + \left(b_0 + \sum_{j=1}^{p} b_j y_{t-j} \right) I_r(y_{t-d}) + \varepsilon_t,$$

where I_r is an indicator function such that

$$I_r(x) = \begin{cases} 1 & \text{if } x > r \\ 0 & \text{if } x \leq r. \end{cases}$$

In Chan and Tong (1986, esp., P187), a more data driven model, smooth threshold autoregressive (STAR) model of the form

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + \left(b_0 + \sum_{j=1}^{p} b_j y_{t-j} \right) F\left(\frac{y_{t-d} - r}{c} \right) + \varepsilon_t$$

was proposed, where $F(.)$ is any sufficiently smooth function with a rapidly decaying tail. However, the most commonly discussed TAR or STAR models use one lagged value y_{t-d} as the threshold variable and most existing studies focus on either model specification or parameter estimation with the delay parameter d chosen by hypothesis testing. See, e.g., the van Dijk, Teräsvirta and Franses (2002) for a review. It is obvious that the selection of the threshold variable is essential in building a TAR model.

In this paper, we study the following STAR(p, q) model

$$y_t = \left(a_0 + \sum_{j=1}^{p} a_j y_{t-j} \right) + \left(b_0 + \sum_{j=1}^{p} b_j y_{t-j} \right) \Phi\left(\theta_0 + \sum_{j=1}^{q} \theta_j y_{t-j} \right) + \varepsilon_t, \tag{1}$$

where we set F equal to the standard Gaussian distribution for simplicity of discussion although this is not essential. $\{\varepsilon_t\}$ is assumed to be a white noise with finite variance σ^2, and be independent of the past observations $\{y_s, s < t\}$. The threshold variable $z_t = \theta_0 + \sum_{j=1}^{q} \theta_j y_{t-j}$ is a linear function of lagged endogenous variables.

One advantage of the proposed model is in the selection of threshold variable. For example, if θ_k are all zeros except for $k = j$, then the selected threshold variable is y_{t-j}. We have the following result about the stationarity of the model, for which the proof can be found in the Appendix.

Lemma 1.1. *If*

$$\sup_{0 \leq u \leq 1} \sum_{j=1}^{p} |a_j + b_j u| < 1, \tag{2}$$

there exists a strictly stationary solution $\{y_t\}$ from the model (1).

We propose to use the recently developed L_1 regularization approaches which tend to produce a parsimonious number of nonzero coefficients for z_t, thus leading to a simple way of selecting the significant/threshold variables without testing the $2^q - 1$ subsets of $\{y_{t-1}, y_{t-2}, \ldots, y_{t-q}\}$. The lasso penalty can perform model selection as well as estimation. However, its variable selection may be inconsistent (see, e.g., [12]). Fan and Li (2001) proposes the SCAD penalty which is shown to have oracle properties. But the concavity of the

*The research was partially supported by the Education Department of Nature Science Research of Guizhou Province (Grant No. 2010028) and the Nomarch Foundation of Guizhou Province (Grant No. 2010025), China.

†Corresponding author.

penalty function may result in a local minima problem. In this paper, we adopt the adaptive lasso penalty proposed in the paper of Zou [12], which is convex and leads to a variable selection estimator with the oracle properties. Moreover, we propose a direction adaptive lasso method. By penalizing the direction of the coefficient vector instead of the coefficients themselves, the threshold variable is more accurately selected, especially when the sample size is not large enough. Note that the norm of the coefficient vector implies the threshold shape, which should not be penalized. Our penalization on the direction can achieve this goal while the direct penalization on the coefficient cannot. Both numerical and real data analysis are provided to illustrate its advantage. The oracle properties of the resulting estimators are also obtained.

The rest of the article is organized as follows. In Section 2 we derive the consistency and asymptotic normality of the conditional LS estimator. In Section 3 we propose to use the adaptive lasso method to select the threshold variable and estimate the unknown parameters simultaneously. As in the regression case, the adaptive lasso estimator has the oracle properties. In Section 4, we propose the direction adaptive lasso method and show its oracle properties. Section 5 is devoted to simulation and real data analysis. The study compares the LS estimator, adaptive lasso estimator and the proposed direction adaptive lasso estimator in two data generating processes and one real data set: the Canadian Lynx Data.

2. THE CONDITIONAL LEAST SQUARES ESTIMATOR

Let $a = (a_0, a_1, \ldots, a_p)^\top$, $b = (b_0, b_1, \ldots, b_p)^\top$, $\theta = (\theta_0, \theta_1, \ldots, \theta_q)^\top$, we rewrite model (1) as

$$y_t = x_t^\top a + (x_t^\top b)\Phi(s_t^\top \theta) + \varepsilon_t, \tag{3}$$

where

$$x_t^\top = (1, y_{t-1}, \ldots, y_{t-p}),\ s_t^\top = (1, y_{t-1}, \ldots, y_{t-q}),$$

for $t = m+1, \ldots, T$ and $m = \max(p, q)$. The unknown parameters $\eta = (a^\top, b^\top, \theta^\top)^\top = (\eta_1, \ldots, \eta_L)^\top$ $(L = 2p+q+3)$ is assumed to be in an open set Θ of $\mathbb{R}^{\otimes(2p+q+3)}$. Denote $\theta = (\theta_0, \vartheta^\top)^\top = (\theta_0, \theta_1, \ldots, \theta_q)^\top$ with $\vartheta = (\theta_1, \ldots, \theta_q)^\top \in \mathbb{R}^q$ and the true value $\vartheta_0 = (\theta_{10}, \ldots, \theta_{q0})^\top$. Denote the true value of η by $\eta_0 = (\boldsymbol{a}_0^\top, \boldsymbol{b}_0^\top, \boldsymbol{\theta}_0^\top)^\top$. For ease of exposition, we use the boldfaced letter to denote a vector if there exists the same notation for a scalar. For example, $\boldsymbol{a}_0$ denotes the true value of the vector $a = (a_0, a_1, \ldots, a_p)^\top$ and $\boldsymbol{\theta}_0$ denotes the true value of vector $\theta = (\theta_0, \theta_1, \ldots, \theta_q)^\top$. Let K be the index set of those $j \in I \equiv \{1, \ldots, q\}$ with $\theta_{j0} \neq 0$ and κ be the number of components of K and denote $\bar{K} = I \backslash K$.

For each t, we refer to the lagged variables of y_t in the set $\{y_{t-j}, j \in K\}$ as the significant threshold variables and define the transition variable z_t as

$$z_t = s_t^\top \theta = \theta_0 + \theta_1 y_{t-1} + \cdots + \theta_q y_{t-q}. \tag{4}$$

Denote by $\mathscr{F}_t = \sigma(y_1, \ldots, y_t)$ $(t \geq 1)$ the $\sigma-$field generated by $y_s, 1 \leq s \leq t$ and by $\mathscr{F}_0$ the trivial $\sigma-$field. Define

$$l_t = (1, \tilde{l}_t^\top)^\top,\ \ \tilde{l}_t = (y_{t-1}, \ldots, y_{t-m})^\top \tag{5}$$

and

$$\begin{aligned} g(\eta, \tilde{l}_t) &= g(\eta, \mathscr{F}_{t-1}) \equiv E_\eta(y_t|\mathscr{F}_{t-1}) \\ &= x_t^\top a + (x_t^\top b)\Phi(s_t^\top \theta),\ \ t \geq 1. \end{aligned}$$

Given a set of observations $\{y_1, \ldots, y_T\}$, the conditional least squares (LS) estimator minimizes the objective function

$$\begin{aligned} Q_T(\eta) &= \sum_{t=m+1}^{T} (y_t - E_\eta(y_t|\mathscr{F}_{t-1}))^2 \\ &= \sum_{t=m+1}^{T} \left\{y_t - x_t^\top a - (x_t^\top b)\Phi(s_t^\top \theta)\right\}^2, \end{aligned} \tag{6}$$

with respect to η. Let η_T^{LS} denote the least squares estimator.

Theorem 2.1. *If $\{y_t\}$ is a stationary ergodic sequence of integrable variables and $\tilde{l}_0$ has a positive density function almost everywhere, then as $T \to \infty$,*

$$\eta_T^{LS} \to \eta_0,\ \ a.s. \tag{7}$$

and

$$T^{1/2}(\eta_T^{LS} - \eta_0) \Rightarrow N(0, \sigma^2 U^{-1}), \tag{8}$$

where

$$\begin{aligned} U &\equiv E_{\eta_0}\left(\frac{\partial g(\tilde{l}_t, \eta_0)}{\partial \eta} \cdot \frac{\partial g(\tilde{l}_t, \eta_0)}{\partial \eta^\top}\right) \\ &= E_{\eta_0}\left(\frac{\partial g(\tilde{l}_0, \eta_0)}{\partial \eta} \cdot \frac{\partial g(\tilde{l}_0, \eta_0)}{\partial \eta^\top}\right) \end{aligned} \tag{9}$$

is positive definite.

Remark 2.2. Using the Fisher information matrix $I(\eta)$,

$$\begin{aligned} I(\eta) &= E_\eta\left\{\frac{\partial \log f}{\partial \eta} \cdot \frac{\partial \log f}{\partial \eta^\top}\right\} \\ &= \frac{1}{\sigma^2} E_\eta\left\{\frac{\partial g(\tilde{l}_t, \eta)}{\partial \eta} \cdot \frac{\partial g(\tilde{l}_t, \eta)}{\partial \eta^\top}\right\}, \end{aligned} \tag{10}$$

where $f = (\sqrt{2\pi}\sigma)^{-1} \exp\{-\frac{\varepsilon_t^2}{2\sigma^2}\}$, the result of the theorem 2.1 can be written as

$$T^{1/2}(\eta_T^{LS} - \eta_0) \Rightarrow N(0, I^{-1}(\eta_0)). \tag{11}$$

3. THE ADAPTIVE LASSO ESTIMATOR

In this section, we shrink the unnecessary coefficients of the transition variable z_t to 0 and select the true threshold variables by the adaptive lasso approach proposed by Zou [12]. We use η_T^{ADL} to denote the adaptive lasso estimator of η which minimizes

$$Q_T^{ADL}(\eta) = Q_T(\eta) + \lambda_T \sum_{j=1}^{q} \hat{w}_j |\theta_j|, \tag{12}$$

where the weight $\hat{w}_j$ is the reciprocal of an increasing function of the corresponding LS estimate of θ_j, i.e., $\hat{w}_j = 1/|\theta_j^{LS}|^\gamma$, $\lambda_T > 0$, $\gamma > 0$ are two nonnegative tuning parameters.

Let $K_T^{ADL} = \{j : \theta_j^{ADL} \neq 0, 1 \leq j \leq q\}$, where θ_j^{ADL} is the adaptive lasso estimate of θ_j. Recall that $K = \{1 \leq j \leq q : \theta_{j0} \neq 0\}$ and $\kappa = |K|$. That is, the correct model has κ significant threshold variables. For any vector/matrix A, denote by $A_{(K)}$ a sub-vector/sub-matrix of A formed by the elements at K'th rows (and K'th columns) of A. For example, if $A = (a_{ij})_{1\leq i,j\leq 5}$ and $K = \{1, 3\}$, then $A_{(K)} = (a_{ij})_{i,j=1,3}$.

Theorem 3.1. *Suppose that* $\frac{\lambda_T}{\sqrt{T}} \to 0$, *and* $\lambda_T T^{\frac{\gamma-1}{2}} \to \infty$. *Then the adaptive lasso estimates* η_T^{ADL} *satisfy the following oracle properties:*

1. *Consistency in variable selection:*

$$\lim_{T\to\infty} P(K_T^{ADL} = K) = 1.$$

2. *Asymptotic normality:*

$$\sqrt{T}(\eta_{T,(K)}^{ADL} - \eta_{0,(K)}) \Rightarrow N_{2p+\kappa+3}\big(\mathbf{0}, I^{-1}(\eta_{0,(K)})\big).$$

The second part of Theorem 3.1 implies that the final estimator can achieve the efficiency of the estimator when the true threshold variables are known and estimated with irrelevant variables being removed. Thus, as in the literature estimator η_T^{ADL} has the so-called oracle property.

4. THE DIRECTION ADAPTIVE LASSO ESTIMATOR

As $c \to +\infty$, the function $\Phi(c(x - r))$ approaches to the indicator function

$$I_r(x) = \begin{cases} 1 & \text{if } x > r, \\ 0 & \text{if } x \leq r, \end{cases}$$

which is the threshold principle of the classical two-regime TAR model. However, in the STAR(p, q) model (1), when the length of the vector $\vartheta = (\theta_1, \ldots, \theta_q)^\top$ is large, penalizing $\tilde{\theta}_j \equiv \theta_j/\|\vartheta\|$ instead of θ_j seems more desirable $(j = 1, 2, \ldots, q)$ than penalizing the coefficient vector since the latter also penalizes the length of the coefficients, which plays the role of c.

We call the estimator by adaptively penalizing the direction of coefficient vector the direction adaptive lasso estimator and denote it as η_T^{DAL}, which minimizes

$$Q_T(\eta) + \lambda_T \sum_{j=1}^{q} \tilde{w}_j |\tilde{\theta}_j| = Q_T(\eta) + \frac{\lambda_T}{l(\vartheta)} \sum_{j=1}^{q} \tilde{w}_j |\theta_j|, \tag{13}$$

where $l(\vartheta) = \sqrt{\theta_1^2 + \cdots + \theta_q^2}$, the new weight $\tilde{w}_j$ is the reciprocal of an increasing function of the corresponding LS estimate of $\tilde{\theta}_j$, i.e.,

$$\tilde{w}_j = 1/|\tilde{\theta}_j^{LS}|^\gamma = \frac{l^\gamma(\theta_T^{LS})}{|\theta_j^{LS}|^\gamma},$$

and $\lambda_T > 0$, $\gamma > 0$ are two nonnegative tuning parameters.

The oracle properties of η_T^{ADL} are provided by the following theorem.

Lemma 4.1. *As* $T \to \infty$, $\tilde{\vartheta}_T^{LS}$, *the LS estimator of* $\tilde{\vartheta}$ *satisfies*

$$\tilde{\vartheta}_T^{LS} \to \tilde{\vartheta}_0, \ \ a.s.$$

and

$$T^{1/2}(\tilde{\vartheta}_T^{LS} - \tilde{\vartheta}_0) \Rightarrow N(0, \tilde{\Sigma}),$$

where $\tilde{\vartheta}_0 = \vartheta_0/l(\vartheta_0)$ *and*

$$\tilde{\Sigma} = (\vartheta_0^\top \vartheta_0)^{-1}(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)$$

is a non-negative definite matrix with rank $q-1$. *Here,* I_q *is the* $q \times q$ *identity matrix,* $I^{-1}(\vartheta_0)$ *is submatrix composed of the last* q *rows and the last* q *columns of the inverse matrix of* $I(\eta_0)$ *defined in (10).*

Denote $K_T^{DAL} = \{j : \tilde{\theta}_j^{DAL} \neq 0, 1 \leq j \leq q\}$, where $\tilde{\theta}_j^{DAL}$ is the adaptive lasso estimate of $\tilde{\theta}_j$.

Theorem 4.2. *Suppose that* $\frac{\lambda_T}{\sqrt{T}} \to 0$, *and* $\lambda_T T^{\frac{\gamma-1}{2}} \to \infty$. *Then the direction adaptive lasso estimates* η_T^{DAL} *satisfy the following oracle properties:*

1. *Consistency in variable selection:*

$$\lim_{T\to\infty} P(K_T^{DAL} = K) = 1.$$

2. *Asymptotic normality:*

$$\sqrt{T}(\eta_{T,(K)}^{DAL} - \eta_{0,(K)}) \Rightarrow N_{2p+\kappa+3}\big(\mathbf{0}, I^{-1}(\eta_{0,(K)})\big).$$

Under the same condition as the adaptive lasso method, Theorem 4.2 indicates that the proposed direction adaptive lasso also selects the correct subset of threshold variables consistently. From the asymptotic normality, the method can estimate the non-zero parameters efficiently as if we knew in advance which variables were uninformative and were removed.

5. COMPUTATIONAL ISSUES

For the adaptive lasso and direction adaptive lasso estimator, we apply the local quadratic approximation (LQA) proposed in Fan and Li (2001) to our implementation. Suppose we have an initial value $\boldsymbol{\theta}_0 = (\theta_{00}, \theta_{01}, \ldots, \theta_{0q})^\top$ that is close to the optimization solution, except for a constant, we can equivalently get the adaptive lasso estimator through minimizing

$$Q_T(\eta) + \frac{\lambda_T}{2}\theta^\top \Sigma \theta,$$

and get the direction adaptive lasso estimator through minimizing

$$Q_T(\eta) + \frac{\lambda_T}{2l(\theta)}\theta^\top \Sigma \theta,$$

where $\Sigma \equiv \Sigma(\boldsymbol{\theta}_0) = \mathrm{diag}(v)$ with $\boldsymbol{\theta}_0$ being the value of the last step, and for the adaptive lasso,

$$v = (0, w_1/|\theta_{01}|, \ldots, w_q/|\theta_{0q}|)^\top, \quad w_i = 1/|\theta_i^{LS}|^\gamma,$$

for the direction adaptive lasso,

$$v = (0, \tilde{w}_1/|\theta_{01}|, \ldots, \tilde{w}_q/|\theta_{0q}|)^\top, \quad \tilde{w}_i = 1/|\tilde{\theta}_i^{LS}|^\gamma.$$

Remark 5.1. Under the assumption that $\theta_0 \neq 0$, the transition variable

$$z_t = \theta_0 + \theta_1 y_{t-1} + \cdots + \theta_q y_{t-q} \tag{14}$$

can also be equivalently written as

$$z_t = \frac{1 + \tau_1 y_{t-1} + \cdots + \tau_q y_{t-q}}{c} \tag{15}$$

with

$$c = 1/\theta_0, \quad \tau_j = \theta_j/\theta_0, \quad j = 1, \ldots, q.$$

In the numerical experiments, we use this form to evaluate the estimation accuracy.

Specifically, when we evaluate the MSE of the estimate of $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_q)^\top$, we use $(\hat{\tau}, \hat{c}) = (\hat{\tau}_1, \ldots, \hat{\tau}_q, \hat{c})$ instead. That is, we evaluate the deviation of $(\hat{\tau}, \hat{c})$ from the true value (τ_0, c_0) with $\tau_0 = (\tau_{10}, \ldots, \tau_{q0}) = (\theta_{10}/\theta_{00}, \ldots, \theta_{q0}/\theta_{00})$ and $c_0 = 1/\theta_{00}$.

M-folder cross validation (CV) and Bayesian information criterion (BIC) are used to select the tuning parameter $\rho = (\lambda, \gamma)$ and $\gamma \in \{0.5, 1, 2\}$ which is consistent with the choice of γ in Zou [12]. For the BIC, the criterion is

$$\mathrm{BIC}_\rho = \log(\mathrm{RSS}_\rho) + \mathrm{df}(\rho) \times \frac{\log(T-m)}{T},$$

where

$$\mathrm{RSS}_\rho = T^{-1} \sum_{t=m+1}^{T} \left\{ y_t - x_t^\top a - (x_t^\top b)\Phi(s_t^\top \theta) \right\}^2$$

and $\mathrm{df}(\tau) = 2p + 3 + \hat{q}$ with $\hat{q}$ being the number of nonzero coefficients identified by the estimate. For the M-folder CV, denote the full data set by T, and denote the cross-validation training and test set by $T - T^\nu$ and $T^\nu, \nu = 1, \ldots, M$, respectively. For each ρ and ν, we find the estimate using the training set and find a ρ that minimizes

$$CV(\rho) = \sum_{\nu=1}^{M} \sum_{y_k \in T^\nu} (y_k - \hat{y}_k)^2,$$

where $\hat{y}$ is the corresponding fitted value.

6. NUMERICAL EXPERIMENTS

Our aim of numerical experiments is to show the performance of using the L_1-penalization to select the threshold variables. Moreover, the finite sample performance of the LS estimator, adaptive lasso estimator and the proposed direction adaptive lasso estimator are also compared. We summarize the results in the following aspects. (1) Estimation accuracy. Mean squared error (MSE) is examined. For $r = 1, \ldots, R$, let

$$\mathrm{MSE}_r = \sum_{i=0}^{p} (\hat{a}_i^r - a_{i0})^2 + \sum_{i=0}^{p} (\hat{b}_i^r - b_{i0})^2 + \sum_{i=1}^{q} (\hat{\tau}_i^r - \tau_{i0})^2 + (\hat{c}^r - c_0)^2.$$

and $\mathrm{MSE} = \sum_{r=1}^{R} \mathrm{MSE}_r/R$. The standard deviation MSE_r over the R simulation replications is also measured. (2) The average number of correctly selected 0 coefficients of the threshold variable.

We use the following three setups for tuning parameter selection.

Setup 1 Two folder CV.
λ take a set of values with exponentially increasing gaps, say, $\lambda = n^{db}, db = lb + (N-1)d$, with $lb > 0, d = \frac{ub-lb}{N-1}, ub < 0.5$, where the integer N is the number of choices of λ, and lb and ub are chosen such that (λ, γ) satisfies

$$\frac{\lambda}{\sqrt{n}} \to 0, \quad \frac{\lambda}{\sqrt{n}} \cdot n^{\gamma/2} \to \infty.$$

as $n \to \infty$.

Setup 2 Five folder CV and $\lambda = 0.5i, i = 1, 2, \ldots, 20$.

Setup 3 BIC and $\lambda = 0.5i, i = 1, 2, \ldots, 20$.

Example 6.1. In the simulation, the following two STAR models

Model 1: $p = 2, q = 2$, the true threshold variable set is $\{y_{t-2}\} \subseteq \{y_{t-1}, y_{t-2}\}$. The model is

$$y_t = (8 - 0.4y_{t-1} + 0.5y_{t-2}) + (-10 + 0.3y_{t-1} - 0.4y_{t-2})\Phi(-5 + 6y_{t-2}) + \varepsilon_t.$$

Table 1.1. Estimation results for Model 1 under Setup 1

n	Method	MSE	S.d.	Avg. no. of 0 coeff.
50	LS	5.0885	230.7513	0
	AL	1.3200	46.6351	0.56
	DAL	0.6407	27.2812	0.76
100	LS	1.1944	58.5926	0
	AL	0.1322	1.3404	0.58
	DAL	0.2261	5.9606	0.66
200	LS	0.0446	0.4138	0
	AL	0.0353	0.2849	0.74
	DAL	0.0401	0.3846	0.84
500	LS	0.0113	0.0962	0
	AL	0.0108	0.0946	0.76
	DAL	0.0111	0.0964	0.78

Table 1.2. Estimation results for Model 1 under Setup 2

n	Method	MSE	S.d.	Avg. no. of 0 coeff.
50	LS	4.2117	224.9379	0
	AL	1.1895	35.3070	0.58
	DAL	0.4380	8.1961	0.72
100	LS	45.8695	2908.6	0
	AL	0.2163	6.7845	0.62
	DAL	0.1372	2.5903	0.70
200	LS	0.0499	0.8037	0
	AL	0.0398	0.3985	0.60
	DAL	0.0427	0.5044	0.64

Model 2: $p = 2, q = 4$, the true threshold variable set is $\{y_{t-1}, y_{t-3}\} \subseteq \{y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}\}$. The model is

$$y_t = (2 + 0.5y_{t-1} - 0.4y_{t-2}) + (-1.5 - 0.4y_{t-1} + 0.2y_{t-2})\Phi(-10 + 5y_{t-1} + 3y_{t-3}) + \varepsilon_t,$$

where ε_t is simulated from $N(0,1)$. In the second model setup, we let the order $q = 4$ which is bigger than the largest lag of the true threshold variables.

A total of 50 simulation replications are conducted for each model setup. For every simulated data, we find the LS, adaptive lasso and the direction adaptive lasso estimates. The calculation results are summarized in Tables 1 and 2. We can see from Tables 1.1, 1.2, 2.1 and 2.2 that the DAL method can indeed improve the estimation efficiency over the direct adaptive Lasso method. The DAL method is also more powerful in eliminating the unimportant variables.

Example 6.2 (The Canadian Lynx data)**.** To further illustrate the performance of the proposed method in selecting the threshold variable set, we examine one popular studied real data set. Following Tong (1990), we transform the data by taking base-10 logarithm to the original data, and denoted the transformed time series by y_t. Now assume that the time series follows the STAR(p,q) model. Applying different estimation methods to the data, we have the results listed in Table 3.

Both biological facts and previous statistical data analysis suggest that the significant threshold variable can be y_{t-2} or y_{t-3} or both. (See, e.g., Tong [9] section 7.2, Fan and Yao [4]). Both the adaptive Lasso and the direction adaptive Lasso tend to lend support to the above suggestion.

APPENDIX A. PROOFS

Proof of Lemma 1.1. For $x = (x_1, \ldots, x_m)^\top$, $m = \max(p,q)$, denote $\Phi(x) = \Phi(\theta_0 + \sum_{j=1}^q \theta_j x_j)$ thus $0 \le \Phi(x) \le 1$ and we have

$$\begin{aligned}
|g(\eta, x)| &= \left| \left(a_0 + \sum_{j=1}^p a_j x_j \right) + \left(b_0 \Phi(x) + \sum_{j=1}^p b_j \Phi(x) x_j \right) \right| \\
&= \left| (a_0 + b_0\Phi(x)) + \sum_{j=1}^p (a_j + b_j\Phi(x)) x_j \right| \\
&\le |a_0 + b_0\Phi(x)| + \left| \sum_{j=1}^p (a_j + b_j\Phi(x)) x_j \right| \\
&\le \sum_{j=1}^p |a_j + b_j\Phi(x)||x_j| + C \\
&\le \sum_{j=1}^p |a_j + b_j\Phi(x)| \max\{|x_1|, \ldots, |x_p|\} + C
\end{aligned}$$

When

$$\sup_{0 \le u \le 1} \sum_{j=1}^p |a_j + b_j u| < 1,$$

the model is geometrically ergodic by the Theorem 3.2 of An and Huang (1996). Hence, there exists a stationary distribution F such that the time series y_t given by (1) and initiated at $\tilde{l}_0 = (y_{-1}, \ldots, y_{-m+1})^\top \sim F$ is strictly stationary. □

Proof of Theorem 2.1. The proof that U is positive definite is the same as the proof given by Chan and Tong (1986) in its Appendix II, we thus omit it here.

To show the consistency and asymptotic normality, we follow from the standard method proposed in Klimko and Nelson [8].

First, note that η_T^{LS} is actually obtained by solving the equations

$$\frac{\partial Q_T(\eta)}{\partial \eta_j} = 0, \quad j = 1, 2, \ldots, L, \tag{16}$$

and if we denote the difference $u_t(\eta)$ by

$$u_t(\eta) = y_t - g(\eta, \mathscr{F}_{t-1}),$$

then $\{u_t(\eta_0)\}$ is a sequence of martingale differences.

Table 2.1. Estimation results for Model 2 under Setup 1

n	Method	Estimation accuracy MSE	Estimation accuracy S.d.	Model complexity Avg. 0 no.	Model complexity $\hat{\theta}_2 = 0$ and $\hat{\theta}_4 \neq 0$	Model complexity $\hat{\theta}_4 = 0$ and $\hat{\theta}_2 \neq 0$	Model complexity $\hat{\theta}_2 = 0$ and $\hat{\theta}_4 = 0$
50	LS	0.1136	1.0532	0	–	–	–
	AL	0.5348	14.2598	0.88	0.30	0.14	0.22
	DAL	0.1828	4.9394	1.44	0.14	0.14	0.58
100	LS	0.0677	0.7656	0.02	0.02	0	0
	AL	0.2207	5.3545	0.92	0.28	0.16	0.24
	DAL	0.0710	0.9065	1.3	0.08	0.10	0.56
200	LS	0.0274	0.2856	0	–	–	–
	AL	0.0882	1.6219	1.32	0.26	0.10	0.48
	DAL	0.0302	0.3619	1.68	0.10	0.06	0.76
500	LS	0.0098	0.0795	0.02	0	0.02	0
	AL	0.0124	0.1393	1.50	0.14	0.08	0.64
	DAL	0.0103	0.1007	1.82	0.04	0.10	0.84

Table 2.2. Estimation results for Model 2 under Setup 3

n	Method	Estimation accuracy MSE	Estimation accuracy S.d.	Model complexity Avg. 0 no.	Model complexity $\hat{\theta}_2 = 0$ and $\hat{\theta}_4 \neq 0$	Model complexity $\hat{\theta}_4 = 0$ and $\hat{\theta}_2 \neq 0$	Model complexity $\hat{\theta}_2 = 0$ and $\hat{\theta}_4 = 0$
50	LS	0.1531	2.6703	0.02	0	0.02	0
	AL	9.2426	596.30	1.28	0.22	0.18	0.44
	DAL	0.1932	3.2533	1.62	0.16	0.06	0.70
100	LS	0.0678	0.7654	0	–	–	–
	AL	0.0801	1.0342	1.28	0.12	0.24	0.46
	DAL	0.0683	0.8363	1.72	0.06	0.10	0.78
200	LS	0.0293	0.3022	0	–	–	–
	AL	0.0302	0.3418	1.52	0.16	0.12	0.62
	DAL	0.0299	0.3301	1.82	0.12	0.02	0.84

Table 3. Results for Example 6.2 under Setup 1

p	q	Method	threshold variable(s)	p	q	Method	threshold variable(s)
2	2	AL	y_{t-2}	3	2	AL	y_{t-2}
		DAL	y_{t-2}			DAL	y_{t-2}
	3	AL	$y_{t-1}, y_{t-2}, y_{t-3}$		3	AL	y_{t-2}
		DAL	y_{t-1}, y_{t-3}			DAL	y_{t-2}
	4	AL	y_{t-2}, y_{t-4}		4	AL	$y_{t-1}, y_{t-2}, y_{t-3}$
		DAL	y_{t-2}			DAL	y_{t-3}
	5	AL	y_{t-2}, y_{t-4}		5	AL	$y_{t-2}, y_{t-3}, y_{t-4}$
		DAL	y_{t-2}			DAL	y_{t-2}
4	2	AL	y_{t-2}	5	2	AL	y_{t-2}
		DAL	y_{t-2}			DAL	y_{t-2}
	3	AL	y_{t-3}		3	AL	y_{t-2}
		DAL	y_{t-3}			DAL	y_{t-2}
	4	AL	y_{t-3}		4	AL	y_{t-3}
		DAL	y_{t-3}			DAL	y_{t-3}
	5	AL	y_{t-3}		5	AL	y_{t-3}
		DAL	y_{t-3}			DAL	y_{t-3}

Now, we expand $T^{-1/2}\partial Q_T(\eta)/\partial\eta$ in a Taylor series at η_0 and suppose that η_T^{LS} satisfies (16), we have

(17)
$$\begin{aligned} 0 &= T^{-\frac{1}{2}}\frac{\partial Q_T(\eta_T^{LS})}{\partial\eta} \\ &= T^{-\frac{1}{2}}\frac{\partial Q_T(\eta_0)}{\partial\eta} + T^{-1}(U_T + D_T(\eta^*))\cdot T^{\frac{1}{2}}(\eta_T^{LS}-\eta_0), \end{aligned}$$

where

$$U_T \equiv \frac{\partial^2 Q_T(\eta_0)}{\partial\eta\partial\eta^\top},$$

(18)
$$\begin{aligned} D_T(\eta^*) &\equiv \frac{\partial^2 Q_T(\eta^*)}{\partial\eta\partial\eta^\top} - U_T \\ &= \frac{\partial^2 Q_T(\eta^*)}{\partial\eta\partial\eta^\top} - \frac{\partial^2 Q_T(\eta_0)}{\partial\eta\partial\eta^\top}, \end{aligned}$$

and η^* being an appropriate intermediate point between η_0 and η_T^{LS}.

We claim that

(19)
$$(2T)^{-1}U_T \to U, \quad \text{a.s.}$$

In fact, denote $(U_T)_{ij}$ as the (i,j)-th element of the matrix U_T, we have

$$\begin{aligned} \frac{1}{2}(U_T)_{ij} = &\left(\sum_{t=m+1}^{T}\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_i}\cdot\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_j}\right) \\ &- \left(\sum_{t=m+1}^{T}\frac{\partial^2 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j}u_t(\eta_0)\right). \end{aligned}$$

By the strong law of large numbers for martingales, we get

(20)
$$\frac{1}{T}\sum_{t=m+1}^{T}\frac{\partial^2 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j}u_t(\eta_0)\to 0, \quad \text{a.s.},$$

and by the ergodic theorem we have

$$\frac{1}{T}\sum_{t=m+1}^{T}\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_i}\cdot\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_j}\to U_{ij} \quad \text{a.s.},$$

thus

$$\frac{1}{2T}(U_T)_{ij}\to U_{ij}, \quad \text{a.s.}$$

Similar to (20), we have

$$\frac{1}{T}\frac{\partial Q_T(\eta_0)}{\partial\eta} = -\frac{2}{T}\sum_{t=m+1}^{T}\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta}u_t(\eta_0)\to 0, \quad \text{a.s.}$$

Next, we show that for any $\delta>0$ such that $\|\eta^*-\eta_0\|\le\delta$,

(21)
$$\lim_{T\to\infty}\sup_{\delta\to 0}\frac{|D_T(\eta^*)_{ij}|}{T\delta}<\infty,\ 1\le i,j\le L, \ \text{a.s.}$$

In fact,

$$\begin{aligned} &|D_T(\eta^*)_{ij}| \\ &= \left|\frac{\partial^2 Q_T(\eta^*)}{\partial\eta_i\partial\eta_j} - \frac{\partial^2 Q_T(\eta_0)}{\partial\eta_i\partial\eta_j}\right| \\ &\le \left|\sum_{t=m+1}^{T}\left\{\frac{\partial g(\tilde{l}_t,\eta^*)}{\partial\eta_i}\cdot\frac{\partial g(\tilde{l}_t,\eta^*)}{\partial\eta_j} - \frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_i}\cdot\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_j}\right\}\right| \\ &\quad + \left|\sum_{t=m+1}^{T}\left\{\frac{\partial^2 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j}u_t(\eta_0) - \frac{\partial^2 g(\tilde{l}_t,\eta^*)}{\partial\eta_i\partial\eta_j}u_t(\eta^*)\right\}\right|. \end{aligned}$$

And from the Taylor expansion,

$$u_t(\eta^*) = u_t(\eta_0) + \frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta^\top}(\eta_0-\eta^*)(1+o_p(1)),$$

$$\frac{\partial g(\tilde{l}_t,\eta^*)}{\partial\eta_i} = \frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_i} + \frac{\partial g^2(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta^\top}(\eta^*-\eta_0)(1+o_p(1)),$$

and

$$\frac{\partial^2 g(\tilde{l}_t,\eta^*)}{\partial\eta_i\partial\eta_j} = \frac{\partial^2 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j} + \frac{\partial^3 g(\tilde{l}_t,\eta_0)}{\partial\eta_i\partial\eta_j\partial\eta^\top}(\eta^*-\eta_0)(1+o_p(1)).$$

Note that

$$\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta} = \left(x_t, x_t\Phi_t, (x_t^\top b)\varphi_t s_t\right)^\top,$$

where $\Phi_t\equiv\Phi(s_t^\top\theta), \varphi_t\equiv\varphi(s_t^\top\theta)$ with $\varphi(\cdot)$ being the standard normal pdf are both continuous for all $\eta\in\Theta$. Since $\{y_t\}$ is a stationary ergodic sequence of integrable variables, $u_t(\eta_0)$ is a sequence of martingale differences, by the martingale convergence theorem, it is easy to see that (21) is satisfied.

The conditions of theorem 2.1 of [8] are thereby satisfied. We get the strong consistency (7) from (19), (20) and (21) by the theorem 2.1 of [8].

Next, we prove the asymptotic normality (8): $T^{1/2}(\eta_T^{LS}-\eta_0)\Rightarrow N(0,\sigma^2U^{-1})$.

In view of (17), (19) and the proved consistency result, we only need to show that

(22)
$$\frac{1}{2}T^{-\frac{1}{2}}\frac{\partial Q_T(\eta_0)}{\partial\eta}\Rightarrow N(0,\sigma^2U).$$

In fact, using the Cramer-Wold method, to show (22), it suffices to prove $\forall\ h=(h_1,\ldots,h_L)^\top\in\mathbb{R}^L$,

(23)
$$\frac{1}{2}T^{-\frac{1}{2}}h^\top\frac{\partial Q_T(\eta_0)}{\partial\eta}\Rightarrow N(0,v),$$

where $v=\sigma^2E_{\eta_0}(\sum_{k=1}^{L}h_k\frac{\partial g(\tilde{l}_t,\eta_0)}{\partial\eta_k})^2$. Note that $\partial Q_T(\eta_0)/\partial\eta=-2\sum_{t=m+1}^{T}u_t(\eta_0)\partial g(\tilde{l}_t,\eta_0)/\partial\eta$, let

$$f_1(\tilde{l}_t,h,\eta)\equiv-\sum_{k=1}^{L}h_k\frac{\partial g(\tilde{l}_t,\eta)}{\partial\eta_k},$$

it follows that

$$\frac{1}{2}T^{-\frac{1}{2}}h^\top\frac{\partial Q_T(\eta_0)}{\partial\eta}=T^{-\frac{1}{2}}\sum_{t=m+1}^{T}f_1(\tilde{l}_t,h,\eta_0)u_t(\eta_0). \tag{24}$$

Define

$$Y_t=\frac{f_1(\tilde{l}_t,h,\eta_0)u_t(\eta_0)}{\sigma\sqrt{E_{\eta_0}(f_1^2(\tilde{l}_t,h,\eta_0))}}=\frac{f_1(\tilde{l}_t,h,\eta_0)u_t(\eta_0)}{\sqrt{v}},$$

$$V_T^2=\sum_{t=m+1}^{T}E(Y_t^2|\mathscr{F}_{t-1}),\quad \sigma_T^2=EV_T^2,$$

we claim that

(1) $V_T^2/\sigma_T^2\to 1$ in probability. This is shown by

$$V_T^2=\sum_{t=m+1}^{T}E(Y_t^2|\mathscr{F}_{t-1})=\left(\sum_{t=m+1}^{T}f_1^2\right)/Ef_1^2,\quad \sigma_T^2=T-m$$

and the ergodic theorem.

(2) Lindeberg condition: for any $\epsilon>0$,

$$\frac{1}{\sigma_T^2}\sum_{t=m+1}^{T}E(Y_t^2I(|Y_t|\geq\epsilon\sigma_T))\to 0$$

is satisfied. This is shown by noting that

$$\begin{aligned}Y_{T,t}&\equiv\frac{Y_t}{\sigma_T}=\frac{Y_t}{\sqrt{T-m}}\\&=\frac{f_1(\tilde{l}_t,h,\eta_0)u_t(\eta_0)}{\sqrt{T-m}\sigma\sqrt{E(f_1^2(\tilde{l}_t,h,\eta_0))}}\leq\frac{C}{\sqrt{T-m}}\to 0\end{aligned}$$

as $T\to\infty$ where $C>0$ is some finite constant. By the martingale CLT, we have

$$\sum_{t=m+1}^{T}Y_t/\sqrt{T}\Rightarrow N(0,1) \tag{25}$$

and (22) is proved.

We therefore complete the proof of consistency and asymptotic normality of η_T^{LS}. □

Remark A.1. The result (19) can be written as

$$(2T)^{-1}\left(\frac{\partial^2Q_T(\eta_0)}{\partial\eta\partial\eta^\top}\right)\to\sigma^2I(\eta_0),\quad\text{a.s.} \tag{26}$$

and the result (22) can be written as

$$\frac{1}{2}T^{-\frac{1}{2}}\frac{\partial Q_T(\eta_0)}{\partial\eta}\Rightarrow N(0,\sigma^4I(\eta_0)). \tag{27}$$

Proof of Theorem 3.1. The proof is an application of the same method used to show the oracle properties of the adaptive lasso estimator in Zou [12] to our case.

Step 1. We first show the asymptotic normality.

Let $\eta=\eta_0+u/\sqrt{T}$, $u=(u_1,\ldots,u_L)^\top$, $L=2p+3+q$, and

$$\Psi_T(u)=Q_T(\eta_0+u/\sqrt{T})+\lambda_T\sum_{j=1}^{q}\hat{w}_j\left|\theta_{j0}+\frac{u_{2p+3+j}}{\sqrt{T}}\right|.$$

Suppose $\hat{u}_T=\arg\min_u\Psi_T(u)$, then

$$\eta_T^{ADL}=\eta_0+\hat{u}_T/\sqrt{T}\text{ or }\hat{u}_T=\sqrt{T}(\eta_T^{ADL}-\eta_0)$$

since

$$\eta_T^{ADL}=\arg\min Q_T(\eta)+\lambda_T\sum_{j=1}^{q}\hat{w}_j|\theta_j|.$$

Denote $V_T(u)\equiv\Psi_T(u)-\Psi_T(\mathbf{0})$, we have

$$\begin{aligned}V_T(u)&=\{Q_T(\eta_0+u/\sqrt{T})-Q_T(\eta_0)\}\\&\quad+\left\{\lambda_T\sum_{j=1}^{q}\hat{w}_j\left(\left|\theta_{j0}+\frac{u_{2p+3+j}}{\sqrt{T}}\right|-|\theta_{j0}|\right)\right\}\\&\equiv H_T(u)+P_T(u),\end{aligned} \tag{28}$$

where the loss function term

$$H_T(u)=Q_T(\eta_0+u/\sqrt{T})-Q_T(\eta_0)$$

and the penalty term

$$P_T(u)=\lambda_T\sum_{j=1}^{q}\hat{w}_j\left(\left|\theta_{j0}+\frac{u_{2p+3+j}}{\sqrt{T}}\right|-|\theta_{j0}|\right).$$

Note that

$$\begin{aligned}&Q_T(\eta_0+u/\sqrt{T})-Q_T(\eta_0)\\&\quad=\frac{1}{\sqrt{T}}u^\top\frac{\partial Q_T(\eta_0)}{\partial\eta}+\frac{1}{2T}u^\top\frac{\partial^2Q_T(\eta_0)}{\partial\eta\partial\eta^\top}u(1+o_p(1)).\end{aligned}$$

From the results (26) and (27), we know that as $T\to\infty$,

$$\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_0)}{\partial\eta}\Rightarrow W\sim N(\mathbf{0},4\sigma^4I(\eta_0))$$

and

$$\frac{1}{2T}\frac{\partial^2Q_T(\eta_0)}{\partial\eta\partial\eta^\top}\to\sigma^2I(\eta_0)\ \text{ a.s.}$$

Thus the loss function term

$$H_T(u)\Rightarrow u^\top W+\sigma^2u^\top I(\eta_0)u.$$

Now we consider the limiting behavior of the penalty term.

If $j \in K$, i.e., $\theta_{j0} \neq 0$, from the result of the theorem 2.1,

$$\hat{w}_j = 1/|\theta_j^{LS}|^\gamma \to |\theta_{j0}|^{-\gamma}, \quad \text{a.s.}$$

and

$$\sqrt{T}\left(\left|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}\right| - |\theta_{j0}|\right) \to u_{2p+3+j}\text{sgn}(\theta_{j0}).$$

Since $\lambda_T/\sqrt{T} \to 0$, we have

$$\frac{\lambda_T}{\sqrt{T}}\hat{w}_j\sqrt{T}\left(\left|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}\right| - |\theta_{j0}|\right) \to 0.$$

If $j \in \bar{K}$, i.e., $\theta_{j0} = 0$, then $\sqrt{T}\big(|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}| - |\theta_{j0}|\big) = |u_{2p+3+j}|$. Since $\sqrt{T}\theta_j^{LS} = O_p(1)$ and $\lambda_T T^{(\gamma-1)/2} \to \infty$, we have

$$\frac{\lambda_T}{\sqrt{T}}\hat{w}_j = \lambda_T T^{\frac{\gamma-1}{2}}|\sqrt{T}\theta_j^{LS}|^{-\gamma} \to \infty.$$

Therefore, by Slutsky's theorem, we have $V_T(u) \Rightarrow V(u)$ for every u, where

$$V(u) = \begin{cases} (u_{(K)})^\top W_{(K)} + \sigma^2 (u_{(K)})^\top I(\eta_{0,(K)}) u_{(K)}, \\ \qquad\qquad \text{if } u_{2p+3+j} = 0, \forall j \in \bar{K} \\ \infty, \qquad \text{otherwise,} \end{cases}$$

where $u_{(K)}$ and $W_{(K)}$ are the j-th ($j \in \{2p+3+k : k \in \bar{K}\}$) elements deleted from u and W respectively.

Note that $V_T(u)$ is convex, and the unique minimum of $V(u)$ is

$$u_{min} = \begin{pmatrix} -\frac{1}{2\sigma^2} I^{-1}(\eta_{0,(K)}) W_{0,(K)} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ denotes that the other corresponding components $u_{2p+3+j}, j \in \bar{K}$ are all 0 in the vector u.

Following the epi-convergence property of Geyer [6], which is also used in Zou [12], we have

$$\hat{u}_{T,(K)} \Rightarrow -\frac{1}{2\sigma^2} I^{-1}(\eta_{0,(K)}) W_{(K)}$$

and the other components $\to \mathbf{0}$, i.p..

Finally, recall that $W_{(K)} \sim N(\mathbf{0}, 4\sigma^4 I(\eta_{0,(K)}))$, we get

$$\sqrt{T}(\eta_{T,(K)}^{ADL} - \eta_{0,(K)}) \Rightarrow N\big(\mathbf{0}, I^{-1}(\eta_{0,(K)})\big). \tag{29}$$

Step 2. Now we prove the consistency.

If $j \in K$, then $\theta_j^{ADL} \to \theta_{j0}$ i.p., thus $P(j \in K_T^{ADL}) \to 1$. Thus we only need to show that $\forall j \in \bar{K},\ P(j \in K_T^{ADL}) \to 0$.

By the KKT optimality conditions,

$$\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{ADL})}{\partial \theta_j} + \frac{\lambda_T}{\sqrt{T}}\hat{w}_j \text{sgn}(\theta_j^{ADL}) = 0.$$

Note that

$$\left|\frac{\lambda_T}{\sqrt{T}}\hat{w}_j \text{sgn}(\theta_j^{ADL})\right| = \frac{\lambda_T}{\sqrt{T}} T^{\gamma/2} |\sqrt{T}\theta_j^{LS}|^{-\gamma} \to \infty, \quad \text{i.p.},$$

whereas

$$\begin{aligned} &\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{ADL})}{\partial \theta_j} \\ &= \frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_0)}{\partial \theta_j} + \frac{1}{T}\frac{\partial^2 Q_T(\eta_0)}{\partial \theta_j^2}\sqrt{T}(\theta_j^{ADL} - \theta_{j0})(1 + o_p(1)) \\ &\Rightarrow \text{ some normal distribution} \end{aligned}$$

by (29) and Slutsky's theorem. Thus, for $j \in \bar{K}$,

$$P(j \in K_T^{ADL}) \le P\left(\left|\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{ADL})}{\partial \theta_j}\right| = \frac{\lambda_T}{\sqrt{T}}\hat{w}_j\right) \to 0.$$

This completes the proof. □

Proof of Lemma 4.1. Recall that $\vartheta = (\theta_1, \ldots, \theta_q)^\top$, denote

$$g(\vartheta) = (\vartheta^\top \vartheta)^{-1/2} = \frac{1}{\sqrt{\theta_1^2 + \cdots + \theta_q^2}}, \tag{30}$$

then

$$\tilde{\vartheta} = \frac{\vartheta}{l(\theta)} = \frac{\vartheta}{(\vartheta^\top \vartheta)^{1/2}} \equiv \vartheta g(\vartheta).$$

From the asymptotic result of ϑ_T^{LS}, we have

$$g(\vartheta_T^{LS}) \to g(\vartheta_0).$$

Thus

$$\tilde{\vartheta}_T^{LS} = \vartheta_T^{LS} g(\vartheta_T^{LS}) \to \tilde{\vartheta}_0 = \vartheta_0 g(\vartheta_0) \ \text{ a.s.}$$

Next we will show the asymptotic normality. From (11), we know that

$$\sqrt{T}(\theta_T^{LS} - \theta_0) \Rightarrow N(0, I^{-1}(\theta_0)),$$

where $I^{-1}(\vartheta_0)$ is submatrix composed of the last q rows and the last q columns of the inverse matrix of $I(\eta_0)$ defined in (10). Thus,

$$\begin{aligned} &\sqrt{T}(\tilde{\vartheta}_T^{LS} - \tilde{\vartheta}_0) \\ &\quad = \sqrt{T}(\vartheta_T^{LS} g(\vartheta_T^{LS}) - \vartheta_0 g(\vartheta_0)) \\ &\quad = \sqrt{T}\big(\vartheta_T^{LS} g(\vartheta_T^{LS}) - \vartheta_0 g(\vartheta_T^{LS}) + \vartheta_0 g(\vartheta_T^{LS}) - \vartheta_0 g(\vartheta_0)\big) \\ &\quad = \sqrt{T}(\vartheta_T^{LS} - \vartheta_0) g(\vartheta_T^{LS}) + \vartheta_0 \sqrt{T}(g(\vartheta_T^{LS}) - g(\vartheta_0)) \\ &\quad \Rightarrow \text{some normal distribution} \end{aligned}$$

by the Slutsky theorem and the continuous mapping theorem.

It is easy to see that the mean of the asymptotic normal distribution is $\mathbf{0}$. We now provide the asymptotic covariance matrix $\tilde{\Sigma}$ and show that its rank is $q - 1$.

Note that $\partial g(\vartheta)/\partial\vartheta = -(\vartheta^\top\vartheta)^{-3/2}\vartheta$ and $\vartheta_T^{LS} - \vartheta_0 = O_p(T^{-1/2})$, we have

$$\begin{aligned}
&\vartheta_0\sqrt{T}(g(\vartheta_T^{LS}) - g(\vartheta_0)) \\
&\quad = \vartheta_0\sqrt{T}\frac{\partial g(\vartheta_0)}{\partial\vartheta^\top}(\vartheta_T^{LS} - \vartheta_0) + O_p(T^{-1/2}) \\
&\quad = -\vartheta_0\vartheta_0^\top\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)(\vartheta_0^\top\vartheta_0)^{-3/2} + O_p(T^{-1/2}).
\end{aligned}$$

Denote $Z_{T,1} = \sqrt{T}(\vartheta_T^{LS} - \vartheta_0)g(\vartheta_T^{LS})$ and $Z_{T,2} = -\vartheta_0\vartheta_0^\top\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)(\vartheta_0^\top\vartheta_0)^{-3/2}$, we next calculate the covariance matrix of $Z_{T,1} + Z_{T,2}$.

$$\begin{aligned}
&\mathrm{Var}(Z_{T,1} + Z_{T,2}) \\
&\quad = \mathrm{E}(Z_{T,1} + Z_{T,2})(Z_{T,1} + Z_{T,2})^\top \\
&\quad = \mathrm{E}\Big(\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)^\top g^2(\vartheta_T^{LS})\Big) \\
&\qquad - \mathrm{E}\Big(\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)^\top\vartheta_0\vartheta_0^\top \\
&\qquad \times (\vartheta_0^\top\vartheta_0)^{-3/2}g(\vartheta_T^{LS})\Big) \\
&\qquad - \mathrm{E}\Big(\vartheta_0\vartheta_0^\top\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)^\top \\
&\qquad \times (\vartheta_0^\top\vartheta_0)^{-3/2}g(\vartheta_T^{LS})\Big) \\
&\qquad + \mathrm{E}\Big(\vartheta_0\vartheta_0^\top\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)\sqrt{T}(\vartheta_T^{LS} - \vartheta_0)^\top\vartheta_0\vartheta_0^\top \\
&\qquad \times (\vartheta_0^\top\vartheta_0)^{-3}\Big).
\end{aligned}$$

Since as $T \to \infty$, $\sqrt{T}(\vartheta_T^{LS} - \vartheta_0) \Rightarrow N(0, I^{-1}(\vartheta_0))$ and $g(\vartheta_T^{LS}) \to g(\vartheta_0)$, a.s., we thus get the limiting covariance matrix

$$\begin{aligned}
\tilde{\Sigma} = {} & I^{-1}(\vartheta_0)(\vartheta_0^\top\vartheta_0)^{-1} - I^{-1}(\vartheta_0)\vartheta_0\vartheta_0^\top(\vartheta_0^\top\vartheta_0)^{-2} \\
& - \vartheta_0\vartheta_0^\top I^{-1}(\vartheta_0)(\vartheta_0^\top\vartheta_0)^{-2} + \vartheta_0\vartheta_0^\top I^{-1}(\vartheta_0)\vartheta_0\vartheta_0^\top(\vartheta_0^\top\vartheta_0)^{-3}.
\end{aligned}$$

Recall that $\tilde{\vartheta}_0 = \vartheta_0(\vartheta_0^\top\vartheta_0)^{-1/2}$, we have

$$\begin{aligned}
\tilde{\Sigma} = {} & \{I^{-1}(\vartheta_0)(\vartheta_0^\top\vartheta_0)^{-1} - I^{-1}(\vartheta_0)\tilde{\vartheta}_0\tilde{\vartheta}_0^\top(\vartheta_0^\top\vartheta_0)^{-1}\} \\
& - \{\tilde{\vartheta}_0\tilde{\vartheta}_0^\top I^{-1}(\vartheta_0)(\vartheta_0^\top\vartheta_0)^{-1} \\
& - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top I^{-1}(\vartheta_0)\tilde{\vartheta}_0\tilde{\vartheta}_0^\top(\vartheta_0^\top\vartheta_0)^{-1}\} \\
= {} & (\vartheta_0^\top\vartheta_0)^{-1}I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top) \\
& - (\vartheta_0^\top\vartheta_0)^{-1}\tilde{\vartheta}_0\tilde{\vartheta}_0^\top I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top) \\
= {} & (\vartheta_0^\top\vartheta_0)^{-1}(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top).
\end{aligned}$$

Notice that the $q \times q$ matrix $I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top$ is an idempotent matrix due to the relationship $\tilde{\vartheta}_0^\top\tilde{\vartheta}_0 = 1$. That is, $(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)^2 = I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top$ given $\tilde{\vartheta}_0^\top\tilde{\vartheta}_0 = 1$. We thus have

$$\mathrm{rank}(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top) = q - 1.$$

Denote $A = I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top = A^\top$, $B = I^{-\frac{1}{2}}(\vartheta_0)$ and $C = AB$ then

$$\tilde{\Sigma} = (I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top)I^{-1}(\vartheta_0)(I_q - \tilde{\vartheta}_0\tilde{\vartheta}_0^\top) = CC^\top.$$

From the Sylvester's inequality, we get

$$\begin{aligned}
\mathrm{rank}(\tilde{\Sigma}) &= \mathrm{rank}(CC^\top) = \mathrm{rank}(C) \\
&= \mathrm{rank}(AB) \le \min\{\mathrm{rank}(A), \mathrm{rank}(B)\} = q - 1 \\
\mathrm{rank}(\tilde{\Sigma}) &= \mathrm{rank}(AB) \ge \mathrm{rank}(A) + \mathrm{rank}(B) - q = q - 1.
\end{aligned}$$

Therefore, we show that the rank of the matrix $\tilde{\Sigma}$ is $q-1$. □

Proof of Theorem 4.2. The proof is very similar to that of theorem 3.1 and the only difference concerns the treatment of the penalty term.

Let $\eta = \eta_0 + u/\sqrt{T}$, $u = (u_1, \ldots, u_L)^\top$, $L = 2p + 3 + q$, and

$$\begin{aligned}
&\Psi_T(u) \\
&\quad = Q_T\Big(\eta_0 + u/\sqrt{T}\Big) \\
&\qquad + \lambda_T\sum_{j=1}^{q}\tilde{w}_j\left|\left(\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}\right)g\left(\vartheta_0 + \frac{u_{2p+4:2p+3+q}}{\sqrt{T}}\right)\right|,
\end{aligned}$$

where $g(\vartheta)$ is defined in (30) and the q-dimensional sub-vector $u_{2p+4:2p+3+q}$ is composed of the components $u_{2p+4}, u_{2p+5}, \ldots, u_{2p+3+q}$ of the vector u. We denote $u_{2p+4:2p+3+q}$ as $\tilde{u}$.

It follows that the penalty term

$$P_T(u) = \lambda_T\sum_{j=1}^{q}\tilde{w}_j\left(\left|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}\right|g\left(\vartheta_0 + \frac{\tilde{u}}{\sqrt{T}}\right) - |\theta_{j0}|g(\vartheta_0)\right).$$

Since $g'(\vartheta) = -(\vartheta^\top\vartheta)^{-3/2} = -(g(\vartheta))^3$, from the Taylor expansion of g, we have

$$g\left(\vartheta_0 + \frac{\tilde{u}}{\sqrt{T}}\right) = g(\vartheta_0) - (g(\vartheta_0))^3\frac{\tilde{u}^\top\vartheta_0}{\sqrt{T}}(1 + o_p(1)).$$

If $j \in K$, i.e., $\tilde{\theta}_{j0} \neq 0$, from the result of the lemma 4.1,

$$\tilde{w}_j = 1/|\tilde{\theta}_j^{LS}|^\gamma \to |\tilde{\theta}_{j0}|^{-\gamma}, \quad \text{a.s.}$$

and

$$\begin{aligned}
&\sqrt{T}\left(\left|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}\right|g\left(\vartheta_0 + \frac{\tilde{u}}{\sqrt{T}}\right) - |\theta_{j0}|g(\vartheta_0)\right) \\
&\quad = \sqrt{T}\left(\left|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}\right| - |\theta_{j0}|\right)g(\vartheta_0) \\
&\qquad - \left|\theta_{j0} + \frac{u_{2p+3+j}}{\sqrt{T}}\right|(g(\vartheta_0))^3\tilde{u}^\top\vartheta_0(1 + o_p(1)) \\
&\quad \to u_{2p+3+j}\mathrm{sgn}(\theta_{j0})g(\vartheta_0) - |\theta_{j0}|(g(\vartheta_0))^3\tilde{u}^\top\vartheta_0 \\
&\quad = g(\vartheta_0)(u_{2p+3+j}\mathrm{sgn}(\tilde{\theta}_{j0}) - |\tilde{\theta}_{j0}|\tilde{u}^\top\tilde{\vartheta}_0).
\end{aligned}$$

Since $\lambda_T/\sqrt{T} \to 0$, we have

$$\frac{\lambda_T}{\sqrt{T}}\tilde{w}_j\sqrt{T}\left(\left|\theta_{j0}+\frac{u_{2p+3+j}}{\sqrt{T}}\right|g\left(\vartheta_0+\frac{\tilde{u}}{\sqrt{T}}\right)-|\theta_{j0}|g(\vartheta_0)\right)\to 0.$$

If $j \in \bar{K}$, i.e., $\tilde{\theta}_{j0} = 0$, then

$$\begin{aligned}&\sqrt{T}\left(\left|\theta_{j0}+\frac{u_{2p+3+j}}{\sqrt{T}}\right|g\left(\vartheta_0+\frac{\tilde{u}}{\sqrt{T}}\right)-|\theta_{j0}|g(\vartheta_0)\right)\\&\quad=|u_{2p+3+j}|g\left(\vartheta_0+\frac{\tilde{u}}{\sqrt{T}}\right)\to|u_{2p+3+j}|g(\vartheta_0).\end{aligned}$$

When $\tilde{\theta}_{j0} = 0$, we have $\sqrt{T}\tilde{\theta}_j^{LS} = \sqrt{T}(\tilde{\theta}_j^{LS} - \tilde{\theta}_{j0}) = O_p(1)$ from the asymptotical normality result of lemma 4.1. It follows that

$$\frac{\lambda_T}{\sqrt{T}}\tilde{w}_j=\lambda_T T^{\frac{\gamma-1}{2}}|\sqrt{T}\tilde{\theta}_j^{LS}|^{-\gamma}\to\infty$$

since $\lambda_T T^{(\gamma-1)/2} \to \infty$.

Therefore, using the same notations as in the proof of theorem 3.1 and by Slutsky's theorem, we have $V_T(u) \Rightarrow V(u)$ for every u, where

$$V(u)=\begin{cases}(u_{(K)})^\top W_{(K)}+\sigma^2(u_{(K)})^\top I(\eta_{0,(K)})u_{(K)}, & \\ \qquad\qquad\text{if } u_{2p+3+j}=0,\forall j\in\bar{K}\\ \infty, \qquad\text{otherwise,}\end{cases}$$

and get the same asymptotic normality result.

As for the variable selection consistency, we only need to show that

$$\forall j\in\bar{K},\ P(j\in K_T^{DAL})\to 0.$$

Recall that the objective function of the direction adaptive lasso estimator is

$$Q_T(\eta)+\lambda_T\sum_{i=1}^{q}\tilde{w}_i|\tilde{\theta}_i|=Q_T(\eta)+\lambda_T g(\vartheta)\sum_{i=1}^{q}\tilde{w}_i|\theta_i|.$$

For $j \in \bar{K}$, consider the event $j \in K_T^{DAL}$. By the KKT optimality conditions, we have

(31)

$$\begin{aligned}0&=\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{DAL})}{\partial\theta_j}-\frac{\lambda_T}{\sqrt{T}}(g(\vartheta_T^{DAL}))^3\theta_j^{DAL}\sum_{i=1}^{q}\tilde{w}_i|\theta_i^{DAL}|\\&\quad+\frac{\lambda_T}{\sqrt{T}}\tilde{w}_j g(\vartheta_T^{DAL})\mathrm{sgn}(\tilde{\theta}_j^{DAL})\\&=\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{DAL})}{\partial\theta_j}-\frac{\lambda_T}{\sqrt{T}}g(\vartheta_T^{DAL})\tilde{\theta}_j^{DAL}\sum_{i=1}^{q}\tilde{w}_i|\tilde{\theta}_i^{DAL}|\\&\quad+\frac{\lambda_T}{\sqrt{T}}\tilde{w}_j g(\vartheta_T^{DAL})\mathrm{sgn}(\tilde{\theta}_j^{DAL})\\&=\left\{\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{DAL})}{\partial\theta_j}\right.\\&\quad\left.-\frac{\lambda_T}{\sqrt{T}}g(\vartheta_T^{DAL})\tilde{\theta}_j^{DAL}\sum_{i\in K}\tilde{w}_i|\tilde{\theta}_i^{DAL}|\right\}\\&\quad+\left\{\frac{\lambda_T}{\sqrt{T}}\tilde{w}_j g(\vartheta_T^{DAL})\mathrm{sgn}(\tilde{\theta}_j^{DAL})\right.\\&\quad\left.-\frac{\lambda_T}{\sqrt{T}}g(\vartheta_T^{DAL})\tilde{\theta}_j^{DAL}\sum_{i\in\bar{K}}\tilde{w}_i|\tilde{\theta}_i^{DAL}|\right\}\\&\equiv S_{T1}+S_{T2}\end{aligned}$$

We first claim that the term

$$\begin{aligned}S_{T1}&=\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{DAL})}{\partial\theta_j}\\&\quad-\frac{\lambda_T}{\sqrt{T}}g(\vartheta_T^{DAL})\tilde{\theta}_j^{DAL}\sum_{i\in K}\tilde{w}_i|\tilde{\theta}_i^{DAL}|\end{aligned}$$

(32) $\Rightarrow$ some normal distribution

In fact,

$$\frac{1}{\sqrt{T}}\frac{\partial Q_T(\eta_T^{DAL})}{\partial\theta_j}\Rightarrow\quad\text{some normal distribution}$$

and

$$\frac{\lambda_T}{\sqrt{T}}g(\vartheta_T^{DAL})\tilde{\theta}_j^{DAL}\sum_{i\in K}\tilde{w}_i|\tilde{\theta}_i^{DAL}|\to 0$$

as for $i \in K$, $\tilde{w}_i \to |\theta_{i0}|^{-\gamma}, \tilde{\theta}_j^{DAL} \to_p 0, \tilde{\theta}_i^{DAL} \to_p \tilde{\theta}_{i0}$ and $\lambda_T/\sqrt{T} \to 0$. By Slutsky's theorem, we get (32).

We next show that $S_{T2} \to_p \infty$. Note that

$$\begin{aligned}S_{T2}&=\frac{\lambda_T}{\sqrt{T}}\tilde{w}_j g(\vartheta_T^{DAL})\mathrm{sgn}(\tilde{\theta}_j^{DAL})\\&\quad-\frac{\lambda_T}{\sqrt{T}}g(\vartheta_T^{DAL})\tilde{\theta}_j^{DAL}\sum_{i\in\bar{K}}\tilde{w}_i|\tilde{\theta}_i^{DAL}|\\&=\lambda_T T^{\frac{\gamma-1}{2}}g(\vartheta_T^{DAL})\left\{\frac{1}{|\sqrt{T}\tilde{\theta}_j^{LS}|^\gamma}\mathrm{sgn}(\tilde{\theta}_j^{DAL})\right.\\&\quad\left.-\tilde{\theta}_j^{DAL}\sum_{i\in\bar{K}}\frac{1}{|\sqrt{T}\tilde{\theta}_i^{LS}|^\gamma}|\tilde{\theta}_i^{DAL}|\right\}\\&\to_p\infty\end{aligned}$$

since $\lambda_T T^{\frac{\gamma-1}{2}} \to \infty$ and $\forall j \in \bar{K}, \sqrt{T}\tilde{\theta}_j^{LS} = O_p(1)$.

Therefore, for $j \in \bar{K}$,

$$P(j\in K_T^{DAL})\le P(|S_{T1}|=|S_{T2}|)\to\quad 0.$$

This completes the proof. □

Received 8 October 2010

REFERENCES

[1] An, H. Z. and Huang, F. C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statist. Sinica* **6** 943–956. MR1422412

[2] Chan, K. S. and Tong, H. (1986). On estimating thresholds in autoregressive models. *J. Time Ser. Anal.* **7** 179–190. MR0857248

[3] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[4] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series. Nonparametric and Parametric Methods*. Springer-Verlag, New York. MR1964455

[5] Fu, W. (1998). Penalized regressions: the bridge versus the Lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

[6] Geyer, C. (1994). On the asymptotics of constrained M-estimation. *Ann. Statist.* **22** 1993–2010. MR1329179

[7] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[8] Klimko, L. A. and Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* **6** 629–642. MR0494770

[9] Tong, H. (1990). *Nonlinear Time Series. A Dynamical System Approach*. Oxford University Press, New York. MR1079320

[10] Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *J. Roy. Statist. Soc. Ser. B* **42** 245–292.

[11] van Dijk, D., Teräsvirta, T. and Franses, P. H. (2002). Smooth transition autoregressive models - a survey of recent developments. *Econometric Rev.* **21** 1–47. MR1893981

[12] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Qian Jiang
Department of Mathematics and Statistics
Guizhou College of Finance and Economics
China

Department of Statistics and Applied Probability
National University of Singapore
Singapore
E-mail address: jiangqian@nus.edu.sg

Yingcun Xia
Department of Statistics and Applied Probability
Risk Management Institute
National University of Singapore
Singapore
E-mail address: staxyc@nus.edu.sg