Statistica Sinica Preprint No: SS-2017-0500	
Title	Lasso-based Variable Selection of ARMA Models
Manuscript ID	SS-2017-0500
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0500
Complete List of Authors	Ngai Hang Chan Shiqing Ling and Chun Yip Yau
Corresponding Author	Ngai Hang Chan
E-mail	nhchan@sta.cuhk.edu.hk
Notice: Accepted version subject to English editing.	

Statistica Sinica

Lasso-based Variable Selection of ARMA Models

Ngai Hang Chan[†,*], Shiqing Ling[‡] and Chun Yip Yau[*]

[†]*Southwestern University of Finance and Economics*

[*]*The Chinese University of Hong Kong*

and

[‡]*The Hong Kong University of Science and Technology*

Abstract:

This paper considers the Lasso-based approach for variable selection of ARMA models. We first show that the Lasso estimator has the Knight-Fu's limit distribution under a general tuning parameter assumption. With a special restriction on tuning parameters, we show that the Lasso estimator can achieve the "oracle" properties – zero parameters are estimated to be zero exactly and other estimators are as efficient as those under the true model. The results are also extended for the non-stationary ARMA models. An algorithm is discussed. In particular, we propose a data-driven information criterion to select the tuning parameter, which is shown to be consistent with probability approaching to 1. A simulation study is carried out to assess the performance of our procedure and an example is given.

Key words and phrases: ARMA model, Lasso estimation, tuning parameter, vari-

able selection and information criterion.

1. Introduction

It is well known that identifying an ARMA (p, q) model is always a challenging issue for a given data set. The main difficulty is to select the order (p, q). Since Akaike's (1977) AIC criterion is not weakly consistent, one usually uses the BIC criterion to select (p, q); see Rissanen (1978) and Schwarz (1978). The consistency of the BIC criterion was proved by Hannan (1980). This criterion requires prior determination of two constants P and Q such that $p \leq P$ and $q \leq Q$ and uses a sequential procedure to estimate all the possible ARMA(k_1, k_2) models with $k_1 = 1, \ldots, p$ and $k_2 = 1, \ldots, q$. Furthermore, it needs to be combined with checking the adequacy of fitted models and other tests of variable selection for a final model; see Pötscher (1983) and Pötscher and Srinivasan (1994). This classical approach involves a great computational burden, in particular, when p or q is large.

This paper is written to explore the least absolute shrinkage and selection operator (Lasso) approach for the variable selection of ARMA (p, q) models and to determine the order p (or q), simultaneously. Lasso was developed by Tibshirani (1996) for selecting variables and estimating parameters. It has been extensively studied and many of its variants were proposed, for example, Fan and Li (2001) for a non-concave penalized like-

lihood, Fan and Li (2002) for Cox's proportional hazards model, Knight and Fu (2002) and Wang *et al.* (2007) for Lasso-type estimators of regression models, Yuan and Lin (2006) for model selection with grouped variables, Zou (2006) for the adaptive Lasso, and Huang *et al.* (2008) for adaptive Lasso of high-dimensional regression. In the time series setting, the Lasso approach was mainly applied for the autoregressive (AR) models. For example, Nardi and Rinaldo (2011) studied the Lasso estimator for fitting AR models; see also Wang *et al.* (2007), Song and Bickel (2011) for a large vector AR model, Liao and Phillips (2015) studied a general Lasso-type estimator for the vector-error correction models, and Kock (2016) considered adaptive Lasso for autoregressions. Chen and Chan (2011) considered adaptive Lasso for ARMA model selection and obtained asymptotic normality for the estimated parameters. As far as we know, this approach has not been considered for non-stationary ARMA models with a unit root. For stationary ARMA processes, this approach serves the same purposes as the three stage procedure suggested by Hannan and Kavalieris (1984). However, unlike that procedure, this approach does not need to specify $\max(p,q)$.

This paper presents the Lasso-type estimation in Section 2. We first show that the Lasso estimator has the Knight-Fu's limit distribution under

a general tuning parameter assumption. With a special restriction on tuning parameters, we show that the Lasso estimator can achieve the "oracle" properties – zero parameters are estimated to be zero exactly and other estimators are as efficient as those under the true model. The parameter estimators are shown to converge weakly to the Knight-Fu's distribution, which extends the asymptotic normality results in Chen and Chan (2011). The results are extended for the non-stationary ARMA models in Section 3. An algorithm is discussed in Section 4. We propose a data-driven information criterion to select the tuning parameter, which is shown to be consistent with probability approaching to 1. Simulation results are reported in Section 5 and a real example is given in Section 6. All the proofs are given in Section 7.

2. Lasso-type estimation

Assume that the time series $\{y_t\}$ is generated by the ARMA (p, q) model:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \psi_i \varepsilon_{t-i} + \varepsilon_t, \tag{2.1}$$

where ε_t is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean and variance σ^2. The unknown parameters are $\theta \equiv (\phi_1, \cdots, \phi_p, \psi_1, \cdots, \psi_q)'$ and its true value is denoted by θ_0. The parameter subspace, $\Theta \subset R^{p+q}$, is a compact set and θ_0 is an interior

point in Θ, where $R = (-\infty, \infty)$. We make the following assumption:

Assumption 2.1. *$\phi(z) \equiv 1 - \sum_{i=1}^{p} \phi_i z^i \neq 0$ and $\psi(z) \equiv 1 + \sum_{i=1}^{q} \psi_i z^i \neq 0$ when $|z| \leq 1$, and $\phi(z)$ and $\psi(z)$ have no common root with $\phi_p \neq 0$ or $\psi_q \neq 0$.*

This is the usual stationarity and invertibility condition of model (2.1). If both $\phi_p = 0$ and $\psi_q = 0$, then model (2.1) is not identifiable. Hannan (1980) mentioned this already and commented that the estimator based on the quasi- maximum likelihood estimator (MLE) does not converge in any reasonable sense. The unknown order p can be any integer larger than the true order (say p_0) when $\psi_q \neq 0$. In this case, the Lasso approach will overestimate the model and identify the order p_0 via shrinking ϕ_i to be zero for $i = p_0 + 1, \cdots, p$.

Given the observations $\{y_n, \cdots, y_1\}$ and the initial values $\{y_0, y_{-1}, y_{-2}, \cdots\}$ which are generated by models (2.1), we can write the parametric model as

$$\varepsilon_t(\theta) \;=\; y_t - \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{i=1}^{q} \psi_i \varepsilon_{t-i}(\theta)\,. \tag{2.2}$$

Here, $\varepsilon_t(\theta_0) = \varepsilon_t$. The minus conditional log-quasi-Gaussian likelihood function based on $\{\varepsilon_t(\theta) : t = 1, \cdots, n\}$ plus a penalty is

$$L_n(\theta) = \sum_{t=1}^{n} \varepsilon_t^2(\theta) + \sum_{i=1}^{\tilde{p}} \lambda_{in} |\theta_i|, \tag{2.3}$$

where $\tilde{p} = p + q$, and $\{\lambda_{in} : i = 1, \cdots, \tilde{p}\}$ are the non-negative tuning

parameters. The minimizer of $L_n(\theta)$ on Θ is called the Lasso estimator of θ_0, denoted by $\hat{\theta}_n$. When $\lambda_{in} = \lambda_n$ for all i, $\hat{\theta}_n$ reduces to the classical Lasso estimator of Tibshirani (1996). It may also suffer a significant bias; see Fan and Li (2001). Here, $\hat{\theta}_n$ based on (2.3) is the modified Lasso-type estimator as in Wang *et al.* (2007). Let

$$a_n = \max\{\lambda_{in} | i = 1, \cdots, \tilde{p}\}.$$

We have the first result as follows.

Theorem 1. *Suppose that Assumption 2.1 holds for each $\theta \in \Theta$. Then,*

(a) If $a_n/n \to 0$, then $\hat{\theta}_n \to \theta_0$ almost surely (a.s.);

(b) Furthermore, if $\lambda_{in}/\sqrt{n} \to \lambda_{i0} \geq 0$, $i = 1, \cdots, \tilde{p}$, then,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_L \text{argmin}_{u \in \mathbb{R}^p}\{V(u)\},$$

as $n \to \infty$, where $\to_L$ denotes convergence in distribution,

$$V(u) = -2u'N + u'\Omega u + \sum_{i=1}^{\tilde{p}} \lambda_{i0}[u_i \text{sgn}(\theta_{i0}) I(\theta_{i0} \neq 0) + |u_i| I(\theta_{i0} = 0)],$$

where $N \sim N(0, \sigma^2\Omega)$, $I(\cdot)$ is the indicator function and $\Omega = E[\bigtriangledown \varepsilon_t(\theta_0) \bigtriangledown' \varepsilon_t(\theta_0)]$.

This is the Knight-Fu's style asymptotic property; see Knight and Fu (2000). When all $\lambda_{i0} = 0$, $i = 1, \cdots, \tilde{p}$, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_L$

$N(0, \sigma^2\Omega^{-1})$, which is the same as the limit distribution of the usual conditional LSE. However, when some $\theta_{i0} = 0$, the Lasso estimator $\hat{\theta}_{in}$ cannot shrink to zero exactly, and the estimators of other parameters cannot achieve the same efficiency as those of the LSE with restriction on $\theta_{i0} = 0$. To achieve the "oracle" properties in the sense that zero coefficients are estimated as zero exactly and the non-zero coefficients are estimated as efficiently as the conditional LSE with restriction on zero coefficients to be zero, we consider the following tuning parameters:

$$\lambda_{in} = \frac{\lambda_n}{\sqrt{n}|\tilde{\theta}_{in}|}, \tag{2.4}$$

where $\lambda_n > 0$, $\sqrt{n}(\tilde{\theta}_{in} - \theta_{i0}) \to_L \xi_i$ as $n \to \infty$ and ξ_i is a random variable with $P(\xi_i = 0) = 0$. Obviously, we can take $\tilde{\theta}_n = \hat{\theta}_n^{OLS}$, the ordinary least squares estimator, which minimizes (2.3) with $\lambda_{in} = 0$ for all is. Furthermore, since the consistent estimators are available, we restrict our parametric space to be $\Theta_n = \{\theta : \|\theta\| \leq \delta_n\}$, where $\|\cdot\|$ is the Euclidean norm, $\delta_n \to 0$ as $n \to \infty$ and define

$$\hat{\theta}_n^o = \text{argmin}_{\theta \in \Theta_n} L_n(\theta).$$

Equation (2.4) reduces the n tuning parameters in (2.3) to one tuning parameter and renders the computational implementation of the minimization problem easy. This kind of tuning parameters was suggested by Wang *et*

al. (2006) and Zou (2006). The latter called it the adaptive Lasso penalty.

Let θ_{10} be the subset of θ_0 with no-zero elements and θ_{20} be the subset of θ_0 with zero elements, and their corresponding estimators in $\hat{\theta}_n^o$ are denoted by $\hat{\theta}_{1n}^o$ and $\hat{\theta}_{2n}^o$. We have the following result:

Theorem 2. *Suppose that Assumption 2.1 holds for each $\theta \in \Theta_n$, λ_{in} is defined as in (2.4), $\lambda_n/\sqrt{n} \to \lambda_0$, and $\lambda_n \to \infty$. Then, it follows that*

(a) $P(\hat{\theta}_{2n}^o = \mathbf{0}) \to 1$ if $\lambda_0 = 0$;

(b) $\sqrt{n}(\hat{\theta}_{1n}^o - \theta_{10}) \to_L N(B_1, \sigma^2\Omega_1^{-1})$,

as $n \to \infty$, where Ω_1 and B_1 are the submatrix of Ω and the subvector of $\Omega^{-1/2}\lambda_0(\ |\phi_{10}|^{-1}, \cdots, |\phi_{p0}|^{-1}, |\psi_{10}|^{-1}, \cdots, |\psi_{q0}|^{-1})'$, respectively, corresponding to θ_{10}.

When $\lambda_0 = 0$, the Lasso estimator $\hat{\theta}_n$ achieves its "oracle" properties. The penalty function in (2.3) and the tuning parameters in (2.4) can be replaced by other ones, e.g. $\lambda_{in} = (\lambda_n/\sqrt{n}|\tilde{\theta}_{in}|)^\omega$ with $\omega > 0$. As long as similar conditions on the tuning parameters in Liao and Phillips (2015) were satisfied, Theorem 1 and 2 still hold. In the AR model, Nardi and Rinaldo (2011) and Song and Bickel (2011) allowed the order p to approach ∞ when the sample size $n \to \infty$. In the ARMA model, this issue seems to be challenging. The main difficulty lies in the fact that the Lasso estimator $\hat{\theta}_n$ may not be consistent and the objective function does not have a quadratic

approximating form. We did not see any current technique in "large p small n" which can be applied for the ARMA models. This remains an open problem for future research.

3. Extension to non-stationary ARMA models

This section considers the non-stationary ARMA(p, q) model with AR polynomial $\phi(z)$ and MA polynomial $\psi(z)$. The notations used in this section should not be confused with those in Section 2. Assume that the true AR polynomial $\phi_0(z) = 1 - \sum_{i=1}^{p} \phi_{i0} z^i$ has a unit root $+1$, i.e. $\phi_0(1) = 0$, and other roots lie outside the unit circle. Denote $c = -\phi(1)$, $\phi_i^* = -\sum_{k=i+1}^{p} \phi_k$ and $w_t = y_t - y_{t-1}$. We can rewrite model (2.1) as

$$w_t = c y_{t-1} + \sum_{i=1}^{p-1} \phi_i^* w_{t-i} + \sum_{j=1}^{q} \psi_j \varepsilon_{t-j} + \varepsilon_t. \tag{3.1}$$

Assume the following condition is satisfied:

Assumption 3.1 *$\phi^*(z) \equiv 1 - \sum_{i=1}^{p-1} \phi_i^* z^i \neq 0$ and $\psi(z) = 1 + \sum_{i=1}^{q} \psi_i z^i \neq 0$ when $|z| \leq 1$, and $\phi^*(z)$ and $\psi(z)$ have no common root with $\phi_{p-1}^* \neq 0$ or $\psi_q \neq 0$.*

Let $\theta = (\phi_1^*, \cdots, \phi_{p-1}^*, \psi_1, \cdots, \psi_q)'$. The unknown parameter vector is $(c, \theta')'$ and its true value is denoted by $(0, \theta_0')'$. Assume θ lies in a compact set $\Theta \subset R^{p+q-1}$ and its true value θ_0 is an interior point. The full parameter space is now $\Theta_n = [-\delta/n, \delta/n] \times \Theta$, where δ is a small positive number.

The residual from model (3.1) is as follows:

$$\varepsilon_t(c,\theta) \;=\; w_t - cy_{t-1} - \sum_{i=1}^{p-1}\phi_i^* w_{t-i} - \sum_{j=1}^{q}\psi_j\varepsilon_{t-j}(c,\theta). \tag{3.2}$$

The minus conditional log-quasi-Gaussian likelihood function based on $\{\varepsilon_t(c,\theta) : t = 1,\cdots,n\}$ plus a penalty is

$$\tilde{L}_n(c,\theta) = \sum_{t=1}^{n}\varepsilon_t^2(c,\theta) + \sum_{i=1}^{p+q-1}\lambda_{in}|\theta_i|. \tag{3.3}$$

The Lasso estimator of $(0,\theta_0)$ is the minimizer of $\tilde{L}_n(c,\theta)$ on Θ_n:

$$(\hat{c}_n,\hat{\gamma}_n) = \mathrm{argmin}_{\Theta_n}\tilde{L}_n(c,\theta).$$

We should mention that $\hat{c}_n$ is only the local minimizer of $\tilde{L}_n(c,\theta)$ and its global minimizer is not clear. This phenomenon has been well observed in the literature about the unit root problem. The LSE of c in Phillips (1987) can serve as its initial value. The following theorem gives the asymptotic properties of $(\hat{c}_n,\hat{\gamma}_n)$.

Theorem 3. *Suppose that Assumption 3.1 holds for each $\theta\in\Theta$.*

(a) If $\max\{\lambda_{in} : i = 1,\cdots,p+q-1\}/n \to 0$ as $n\to\infty$, then $\hat{\theta}_n \to_p \theta_0$.

(b) Furthermore, if $\lambda_{in}/\sqrt{n}\to\lambda_{i0}\geq 0$, $i = 1\cdots,p+q-1$, then

$$(i)\qquad n\hat{c}_n \to_L \phi^*(1)[\int_0^1 B^2(\tau)\,d\tau]^{-1}\int_0^1 B(\tau)\,dB(\tau),$$

$$(ii)\qquad \sqrt{n}(\hat{\theta}_n-\theta_0)\to_L \mathrm{argmin}_{u\in\mathbb{R}^p}\{V(u)\},$$

as $n \to \infty$, where $B(\tau)$ is a standard Brownian motion,

$$V(u) = -2u'N + u'\Omega u + \sum_{i=1}^{p+q-1} \lambda_{i0}[u_i \mathrm{sgn}(\theta_{i0}) I(\theta_{i0} \neq 0) + |u_i| I(\theta_{i0} = 0)],$$

$N \sim N(0, \sigma^2\Omega)$ and $\Omega = E[\partial\varepsilon_t(0,\theta_0)/\partial\theta' \partial\varepsilon_t(0,\theta_0)/\partial\theta]$.

As for the stationary case, to achieve the "oracle" properties of the Lasso estimator, we consider the following tuning parameters:

$$\lambda_{in} = \frac{\lambda_n}{\sqrt{n}|\tilde{\theta}_{in}|}, \tag{3.4}$$

where $\lambda_n > 0$, $\sqrt{n}(\tilde{\theta}_{in} - \theta_{i0}) \to_L \xi_i$ as $n \to \infty$ and ξ_i is a random variable with $P(\xi_i = 0) = 0$. Furthermore, we restrict our parameter space of θ to be $\Theta_{1n} = \{\theta : \|\theta\| \leq \delta_n\}$, where $\delta_n \to 0$ as $n \to \infty$. The Lasso estimator of (c, θ) is as follows:

$$(\hat{c}_n, \hat{\theta}_n^o) = \mathrm{argmin}_{(c,\theta) \in [-\delta/n, \delta/n] \times \Theta_{1n}} L_n(c, \theta).$$

Let θ_{10} be the subset of θ_0 with no-zero elements and θ_{20} be the subset of θ_0 with zero elements, and their corresponding estimators in $\hat{\theta}_n^o$ be denoted by $\hat{\theta}_{1n}^o$ and $\hat{\theta}_{2n}^o$.

Theorem 4. *Suppose that Assumption 3.1 holds for each $\theta \in \Theta_n$, λ_{in} is defined as in (3.4), $\lambda_n/\sqrt{n} \to \lambda_0$, and $\lambda_n \to \infty$. Then,*

(a) $P(\hat{\theta}_{2n}^o = \mathbf{0}) \to 1$ if $\lambda_0 = 0$,

---

(b) $\sqrt{n}(\hat{\theta}^o_{1n} - \theta_{10}) \to_L N(B_1, \sigma^2\Omega_1^{-1})$,

as $n \to \infty$, *where* Ω_1 *and* B_1 *are the submatrix of* Ω *and the subvector of* $\Omega^{-1/2}\lambda_0(|\phi^*_{10}|^{-1}, \cdots, |\phi^*_{p-1,0}|^{-1}, |\psi_{10}|^{-1}, \cdots, |\psi_{q0}|^{-1})'$, *respectively, corresponding to* θ_{10}.

Since our main target is the variable selection in the ARIMA model, we do not impose the penalty on the parameter c. If a penalty is imposed on c, some oracle properties should be similar to those of Kock (2016) for model (3.1) with $q = 0$. Koch (2016) also obtained some oracle properties when $c \in (-2, 0)$ for model (3.1) with $q = 0$. These results should be able to extend for model (3.1).

4. Algorithm

In the empirical implementation, the tuning parameter λ_n in (2.3) is important. If we take

$$\lambda_n = h \log n \ \textit{or} \ h \log\log n,$$

then the conditions of λ_n in Theorem 3 are satisfied with $\lambda_0 = 0$, where $h > 0$ is a constant. Here, we use the data-driven information criterion (IC) to select the tuning parameter h; see Liao and Phillips (2015). For each h,

denote $\hat{\theta}_n^o$ by $\hat{\theta}_n^o(h)$ and

$$S_n(h) = \sum_{t=1}^{n} \varepsilon_t^2[\hat{\theta}_n^o(h)].$$

Let $d(h)$ and d_0 be the non-zero number of components in $\hat{\theta}_n^o(h)$ and θ_0, respectively. Define

$$IC(h) = S_n(h) + d(h)\log n.$$

The tuning parameter is selected by

$$h_n = \text{argmin}_{h\in[0,h_{\max}]} IC(h), \tag{4.1}$$

where $h_{\max}$ is a positive constant. By Theorem 3, $d(h) = d_0$ for any $h > 0$ with probability approaching to 1. If $d(h_n) > d_0$, then

$$S_n(h_n) - S_n(h) = O_p(1)\,,$$

by Theorem 3. It follows that

$$IC(h_n) - IC(h) = O_p(1) + [d(h_n) - d_0]\log n \to \infty,$$

with probability approaching to 1 as $n \to \infty$. Thus, the model based on the tuning parameter h_n cannot be overfitted, i.e., we must have $d(h_n) \le d_0$. Note that

$$S_n(0)/n \to_p \sigma^2,$$

as $n \to \infty$. If the model is underfitted [i.e., $d(h_n) < d_0$], then

$$S_n(h)/n \to_p C > \sigma^2,$$

as $n \to \infty$, where C is a positive constant. Since $[d(h) - d(h_n)] \log n/n \to 0$, we have $IC(h_n) \leq IC(0) < IC(h)$ as $n \to \infty$. Thus, the tuning parameter selected by (4.1) will not underfit the model. It follows that $P(d(h_n) = d_0) \to 1$ as $n \to \infty$. Thus, our estimator based on the tuning parameter h_n will achieve the "oracle" properties.

Since the objective function (2.3) with λ_{in} defined by (2.4) is a non-convex function, we need to use an iterative approach to search for its minimizer. First, we use the usual conditional LSE of θ_0 as the initial value $\tilde{\theta}_n$. First, note that $\varepsilon_t(\theta)$ can be approximated as follows

$$\varepsilon_t(\theta) \approx \left[\varepsilon_t(\theta_0) - \theta_0' \frac{\partial \varepsilon_t(\theta_0)}{\partial \theta}\right] + \theta' \frac{\partial \varepsilon_t(\theta_0)}{\partial \theta}.$$

Let

$$\tilde{y}_t(\theta) = \varepsilon_t(\theta) - \theta' \frac{\partial \varepsilon_t(\theta)}{\partial \theta} \ \ and \ \tilde{x}_t(\theta) = -\frac{\partial \varepsilon_t(\theta)}{\partial \theta}.$$

Then, we use the following local quadratic function to approximate (2.3):

$$Q(\theta) = \|Y(\theta_n^{(m)}) - X(\theta_n^{(m)})\theta\|^2 + \lambda_n \sum_{i=1}^{p+q} \frac{|\theta_i|}{\sqrt{n}|\theta_{in}^{(m)}|},$$

where $\theta_n^{(m)}$ is the minimizer of $Q(\theta)$ at the m-th iteration, starting from $\theta_n^{(0)} = \tilde{\theta}_n$, $Y(\theta_n^{(m)}) = (\tilde{y}_1(\theta_n^{(m)}), \cdots, \tilde{y}_n(\theta_n^{(m)}))'$ and $X(\theta_n^{(m)}) = (\tilde{x}_1(\theta_n^{(m)}), \cdots, \tilde{x}_n(\theta_n^{(m)}))'$.

The minimizer of $Q(\theta)$ is denoted by $\theta_n^{(m+1)}$. Since $Q(\theta)$ is a convex function in terms of θ, the usual Lasso algorithm can be applied in each iteration, such as those in Tibshirani (1996), Fu (1998), Fan and Li (2001) and Cai *et al.* (2005), among others. In this iteration, we need to set up a threshold for accuracy of estimators. When the estimator is less than this threshold, it will be shrunk to zero exactly and the sparse solution is achieved.

5. Simulation Study

In this section, we investigate the finite sample performance of the proposed Lasso procedure for model identifications. In all simulation experiments, the algorithm described in Section 4 is applied with $h_{\max} = 50$ and 500 replications are used.

5.1 AR model

We first consider the AR(6) model

$$X_t = 0.6X_{t-1} - 0.4X_{t-2} - 0.3X_{t-6} + \epsilon_t \,, \tag{5.1}$$

where $\epsilon_t \overset{i.i.d.}{\sim} N(0,1)$. The Lasso procedure is applied using an AR model with a maximum lag of 10. The results are provided in Table 1. The true model is correctly identified in over 90% of the replications. Moreover, the accuracy of the parameter estimates are very high.

Table 1: Proportion of correct model identification (Corr.), the averages (ave) and the empirical standard deviations (e.s.d.) of the parameter estimates.

n	Corr.		ϕ_1	ϕ_2	ϕ_6
		True	0.600	-0.400	-0.300
500	0.932	ave.	0.592	-0.394	-0.288
		e.s.d.	0.038	0.040	0.037
		True	0.600	-0.400	-0.300
1000	0.960	ave.	0.597	-0.396	-0.295
		e.s.d.	0.027	0.027	0.025

5.2 MA model

Next, we consider the MA(5) model

$$X_t = \epsilon_t + 0.5\epsilon_{t-1} + 0.3\epsilon_{t-3} - 0.4\epsilon_{t-5}\,, \tag{5.2}$$

where $\epsilon_t \overset{i.i.d.}{\sim} N(0,1)$. The Lasso procedure is applied to an MA model with a maximum lag of 10. The results are provided in Table 2. Note that estimation of an MA model is more difficult than that of an AR models, as the regressors of the Lasso procedure are not directly observable and are obtained by the iterative procedure in Section 2. Nevertheless, the true model is correctly identified in almost 90% of the replications. The accuracy of the parameter estimates are also very high.

5.3 ARMA model

In this subsection, we investigate the ARMA(5,4) model

$$X_t = 0.5X_{t-1} + 0.3X_{t-2} - 0.3X_{t-5} + \epsilon_t + 0.5\epsilon_{t-1} - 0.4\epsilon_{t-2} + 0.4\epsilon_{t-4}\,, \tag{5.3}$$

where $\epsilon_t \overset{i.i.d.}{\sim} N(0,1)$. The Lasso procedure is applied using an ARMA model with maximum lags of (5,5). The results are provided in Table 3. The identification of ARMA model is more difficult than that of pure AR or pure MA models. Although the percentage of correctly identifying the ARMA(5,4) model with exactly six non-zero coefficients is only around 50%

Table 2: Proportion of correct model identification (Corr.), the averages (ave) and the empirical standard deviations (e.s.d.) of the parameter estimates.

n	Corr.		θ_1	θ_3	θ_5
		True	0.500	0.300	-0.400
500	0.896	ave.	0.487	0.283	-0.381
		e.s.d.	0.044	0.047	0.050
		True	0.500	0.300	-0.400
1000	0.908	ave.	0.494	0.291	-0.391
		e.s.d.	0.030	0.034	0.031

Table 3: Proportion of correct model identification (Corr.), the averages (ave) and the empirical standard deviations (e.s.d.) of the parameter estimates.

n	Corr.		ϕ_1	ϕ_2	ϕ_5	θ_1	θ_2	θ_4
		True	0.500	0.300	-0.300	0.500	-0.400	0.400
500	0.518	ave.	0.608	0.185	-0.251	0.363	-0.383	0.353
		e.s.d.	0.253	0.170	0.089	0.263	0.155	0.098
		True	0.500	0.300	-0.300	0.500	-0.400	0.400
1000	0.784	ave.	0.544	0.255	-0.276	0.444	-0.398	0.385
		e.s.d.	0.162	0.119	0.060	0.169	0.083	0.055
		True	0.500	0.300	-0.300	0.500	-0.400	0.400
2000	0.906	ave.	0.515	0.283	-0.292	0.481	-0.399	0.391
		e.s.d.	0.085	0.069	0.028	0.085	0.037	0.027

for small n, the percentage increases dramatically as sample size grows. In particular, the percentage is close to 90% when $n = 2000$. The accuracy of the parameter estimates also improves as the sample size grows.

We remark that when the true order of the ARMA process is (p_0, q_0), then the Lasso procedure cannot be applied with both $p > p_0$ and $q > q_0$. The reason is as follows. The initial value of the Lasso procedure is obtained

by fitting an ARMA(p, q) model to a ARMA(p_0, q_0) process. When $p > p_0$ and $q > q_0$, there will be at least one pair of redundant factors in the AR and MA polynomials which can cancel each other. For example, if the true model is $(1 - \phi_0 B)X_t = (1 - \theta_0 B)\epsilon_t$, i.e., $p_0 = q_0 = 1$, but $p = q = 2$ is used in the Lasso procedure, then the resulting ARMA(p, q) model is typically of the form

$$(1 - \hat{\phi}_1 B)(1 - \hat{\phi}_2 B)X_t = (1 - \hat{\theta}_1 B)(1 - \hat{\theta}_2 B)\epsilon_t \,,$$

where $\hat{\phi}_1$ and $\hat{\theta}_1$ are close to the true parameters ϕ_0 and θ_0, respectively, and $\hat{\phi}_2 \approx \hat{\theta}_2$ nearly cancel each other. With these redundant factors, the initial estimates are not consistent estimates of the true ARMA coefficients and thus the Lasso procedure is not applicable. This is an identification issue similar to that mentioned by Hannan (1980). Nevertheless, using the sequential estimation procedure in Pötscher (1990), the quantity $r_0 = \max(p_0, q_0)$ can be consistently estimated from the data. Thus, the Lasso procedure can be applied with the model ARMA(r_0, r_0), which avoids the problem of redundant factors.

To illustrate the practicality of combining the sequential estimation procedure and the proposed Lasso procedure, the following simulation experiment is conducted. For completeness, we briefly outline the sequential procedure of Pötscher (1990):

1. Given a time series $(x_1, \ldots, x_n)$, an integer r and a function $C(n)$ satisfying $\lim_{n\to\infty} \frac{C(n)}{n} = 0$ and $\liminf_{n\to\infty} \frac{C(n)}{\log\log n} > 2$, define the model selection criteria

$$\psi(r) = \log \hat{\sigma}_n^2(r) + \frac{2rC(n)}{n},$$

where $\hat{\sigma}_n^2(r)$ is the estimator of the white noise variance σ^2 obtained by maximizing the Gaussian likelihood from fitting an ARMA(r, r) model to the data. In our implementation, the BIC with $C(n) = \log n$ is used.

2. The estimator $\hat{r}$ of $r_0 = \max(p_0, q_0)$ is given by the first local minimum of $\psi(r)$, i.e., the integer $\hat{r}$ that satisfies

$$\psi(r) > \psi(r+1) \quad \text{for} \quad 0 \le r \le \hat{r}\,, \quad \text{and} \quad \psi(\hat{r}) \le \psi(\hat{r}+1)\,.$$

We repeat the simulation study using the ARMA(5,4) model in (5.3) with Lasso procedure applied using an ARMA model with maximum lags of $(\hat{r}, \hat{r})$ instead of (5,5). The results are provided in Table 4. It can be seen that the sequential procedure for estimating r_0 is highly accurate when n reaches 1000. Therefore, the Lasso procedure is highly probable to be applied with the true $r_0 = 5$, and thus the results in Table 3 and Table 4 give equally good results for $n = 1000$ and 2000.

Table 4: Proportion of correct r_0 estimation (Corr.r_0), proportion of correct model identification (Corr.), the averages (ave) and the empirical standard deviations (e.s.d.) of the parameter estimates.

n	Corr.r_0	Corr.		ϕ_1	ϕ_2	ϕ_5	θ_1	θ_2	θ_4
			True	0.500	0.300	-0.300	0.500	-0.400	0.400
500	0.678	0.412	ave.	0.637	0.127	-0.191	0.334	-0.360	0.400
			e.s.d.	0.247	0.189	0.137	0.258	0.178	0.134
			True	0.500	0.300	-0.300	0.500	-0.400	0.400
1000	0.968	0.8	ave.	0.547	0.249	-0.273	0.441	-0.392	0.384
			e.s.d.	0.160	0.141	0.07	0.163	0.086	0.059
			True	0.500	0.300	-0.300	0.500	-0.400	0.400
2000	0.996	0.902	ave.	0.510	0.288	-0.294	0.485	-0.397	0.390
			e.s.d.	0.088	0.068	0.030	0.088	0.044	0.027

5.4 Non-stationary ARMA model

Finally, we investigate the non-stationary ARIMA(5,1,4) model

$$(1 - 0.5B - 0.3B^2 + 0.3B^5)(1 - B)X_t = \epsilon_t + 0.5\epsilon_{t-1} - 0.4\epsilon_{t-2} + 0.4\epsilon_{t-4}, \quad (5.4)$$

where $\epsilon_t \overset{i.i.d.}{\sim} N(0, 1)$. The Lasso procedure discussed in Section 3 for non-stationary ARIMA model is applied with maximum lags of (5,5). The results are provided in Table 5.1. Note that the AR and MA coefficients of Model (5.4) are the same as that in Model (5.3). Moreover, the Lasso procedure successfully shrinks the estimate of c to zero. Therefore, the fitting of Model (5.4) is essentially a fitting of Model (5.3) using the differenced data. Hence, the proportion of correct model identification, estimated coefficients and their empirical standard deviations in Table 5 and Table 3 are very similar.

6. Real Examples

Chan (2010) analyzed the monthly interest rate on three-month government Treasury bills from 1950 to 1988; see Figure 1. The series is of length $n = 461$. Based on preliminary analysis on the ACF and PACF graphs, several ARMA models with lag 6 are fitted to the differenced log-series. In particular, three models, AR(6), MA(6) and ARMA(6,6) models, are

Table 5: Proportion of correct model identification (Corr.), the averages (ave) and the empirical standard deviations (e.s.d.) of the parameter estimates.

n	Corr.		ϕ_1	ϕ_2	ϕ_5	θ_1	θ_2	θ_4	c
		True	0.500	0.300	-0.300	0.500	-0.400	0.400	0
500	0.530	ave.	0.612	0.182	-0.247	0.363	-0.382	0.358	0
		e.s.d.	0.247	0.179	0.091	0.257	0.152	0.094	0
		True	0.500	0.300	-0.300	0.500	-0.400	0.400	0
1000	0.776	ave.	0.546	0.254	-0.275	0.442	-0.397	0.383	0
		e.s.d.	0.161	0.122	0.057	0.166	0.074	0.047	0
		True	0.500	0.300	-0.300	0.500	-0.400	0.400	0
2000	0.900	ave.	0.507	0.293	-0.294	0.488	-0.398	0.392	0
		e.s.d.	0.081	0.064	0.031	0.082	0.039	0.028	0

compared. It is concluded that AR(6) models performed the best in terms of AIC. Since only three models are considered, it is likely that certain ARMA models within lag (6,6) may give better fitting.

We revisit this data set by applying the Lasso procedure with ARMA model with maximum lag (6,6). Similar to Section 4, the tuning parameter $h_{\max} = 50$ is used. The computation is conducted using `R` in a laptop with a 1.44GHz processor with 4GB RAM. The computation time used for the Lasso procedure takes 43.37 seconds. The procedure arrives at the ARMA(6,6) model

$$X_t = -0.429X_{t-6} + \epsilon_t + 0.432\epsilon_{t-1} + 0.232\epsilon_{t-6}\,, \tag{6.1}$$

where X_t is the difference log-interest rate and ϵ_t is white noise. This model has only 3 parameters and is more parsimonious than the AR(6) model selected in Chan (2010). The tuning parameter selected according to (4.1) is $h_n = 2.5$, corresponding to $\lambda_n = h_n \log(n) = 15.3$. Indeed, the same model is selected over the range $\lambda_n \in [15.3, 27.6]$. Outside this range, for $\lambda_n \in [3.1, 15.3)$, only an additional parameter ϕ_3 is selected; for $\lambda_n \in (27.6, 58.3]$, only the parameter ψ_6 is deleted. Thus, the effect of the tuning parameter λ_n on the Lasso procedure is reasonably stable.

As there are 12 parameters in an ARMA(6,6) model, there are $2^{12} = 4096$ possible models. We evaluated the BIC of all of the 4096 models. The

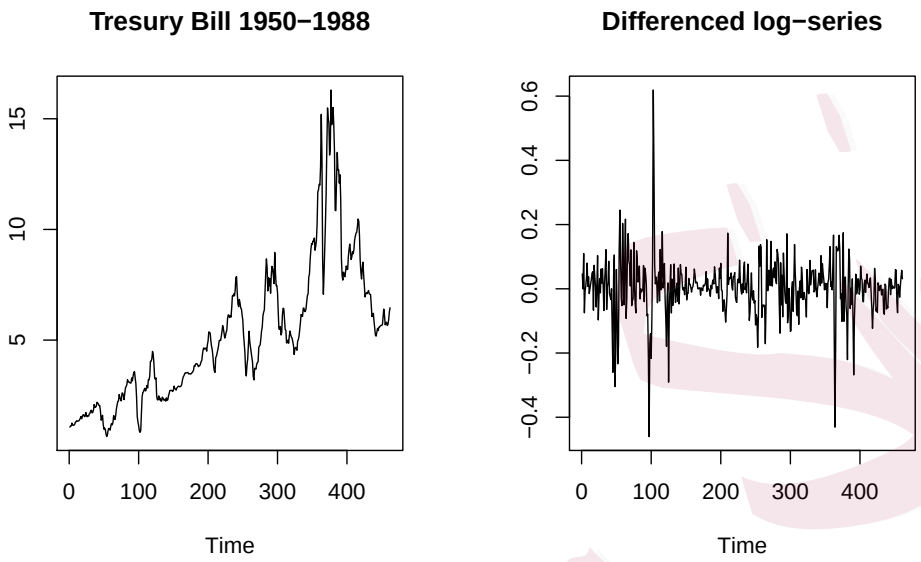

Figure 1: Monthly interest rate on three-month government Treasury bills, 1950-1988.

computation time used for the estimation of the 4096 models is 2240.66 seconds. The ARMA(6,6) model (6.1) achieves the lowest BIC among all models. In conclusion, the shrinkage effect of the proposed Lasso procedure successfully selects a parsimonious ARMA which best describes the treasury bills series.

7. Conclusions

This paper proposes a Lasso-based approach for the order determination of stationary and nonstationary ARMA models. As discussed in Liao and Phillips (2015), it is possible to extend the results to the vector error correction ARMA model or the partially non-stationary ARMA model of Yap and Reinsel (1995a).

8. Appendix: Proofs

Proof of Theorem 1. By the ergodic theorem and a piece-wise argument, see Ling and McAleer (2010), it is readily shown that

$$\max_{\Theta}\left|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2(\theta)-E\varepsilon_t^2(\theta)\right|=o(1), a.s.$$

Furthermore, since $a_n/n\to 0$, we have

$$\max_{\Theta}\left|\frac{1}{n}L_n(\theta)-E\varepsilon_t^2(\theta)\right|=o(1), a.s.$$

Note from (2.2) that we can express $\varepsilon_t(\theta)=\varepsilon_t+\kappa_t(\theta)$ where κ_t only depends on y_ks and ε_ks for $k<t$, and $\kappa_t(\theta)=0$ if and only if $\theta=\theta_0$. Therefore, $E(\varepsilon_t^2(\theta))=E(\varepsilon_t^2)+E(\kappa_t^2(\theta))$ has a unique minimizer at $\theta=\theta_0$. Using a similar approach as for Theorem 2.1 (a) in Ling and McAleer (2010), we can show that (a) holds.

By (a) of this theorem, we have $\hat{\theta}_n \to \theta_0$ a.s. when $n \to \infty$. Denote $\hat{u}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$.

From Theorem 8.11.1 of Brockwell and Davis (1991), we have $\frac{1}{n}\sum_{t=1}^n \partial\varepsilon_t(\theta_0)/\partial\theta\partial\varepsilon_t(\theta_0)/\partial\theta' \to_p \Omega$, $\sum_{t=1}^n [\partial\varepsilon_t(\theta_0)/\partial\theta]\varepsilon_t/\sqrt{n} \to_L N(0, \sigma^2\Omega)$, and Ω is positive definite. Moreover, as $\partial^2\varepsilon_t(\theta_0)/\partial\theta\partial\theta'$ involves information up to time $t-1$, we have $\frac{1}{n}\sum_{t=1}^n \frac{\partial^2\varepsilon_t(\theta_0)}{\partial\theta\partial\theta'}\varepsilon_t \to_p E(\frac{\partial^2\varepsilon_t(\theta_0)}{\partial\theta\partial\theta'}\varepsilon_t) = 0$ by ergodic theorem. Hence, by Taylor's expansion and ergodic theorem, we have

$$\begin{aligned} L_n(\hat{\theta}_n) - L_n(\theta_0) &= 2\hat{u}_n' D_n + \hat{u}_n'[\Omega + o_p(1)]\hat{u}_n \\ &+ \sum_{i=1}^{\tilde{p}} \frac{\lambda_{in}}{\sqrt{n}}[\hat{u}_{in}\mathrm{sgn}(\theta_{i0})I(\theta_{i0} \neq 0) + |\hat{u}_{in}|I(\theta_{i0} = 0)], \quad (8.1) \end{aligned}$$

where $D_n = \sum_{t=1}^n [\partial\varepsilon_t(\theta_0)/\partial\theta]\varepsilon_t/\sqrt{n} \to_L N(0, \sigma^2\Omega)$ as $n \to \infty$. Note that Ω is positive definite and $\lambda_{in}/\sqrt{n} \to \lambda_{i0}$, when $n \to \infty$. From (8.1), we see that $\hat{u}_n = O_p(1)$; Otherwise, if $\hat{u}_n$ is unbounded in probability, we will have $P(L_n(\hat{\theta}_n) - L_n(\theta_0) > \eta) > 1 - \epsilon$ for some η and any ϵ, which is a contradiction to the definition of $\hat{\theta}_n$. Thus, (8.1) reduces to

$$L_n(\hat{\theta}_n) - L_n(\theta_0) = V_n(\hat{u}_n) + o_p(1), \quad (8.2)$$

where

$$V_n(u) = 2u'D_n + u'\Omega u + \sum_{i=1}^{\tilde{p}} \lambda_{i0}[u_i\mathrm{sgn}(\theta_{i0})I(\theta_{i0} \neq 0) + |u_i|I(\theta_{i0} = 0)].$$

Since $\hat{\theta}_n$ minimizes the left hand side of (8.2), $\hat{u}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ minimizes

the right hand side of (8.2). As $V_n(u)$ is convex in u, it follows that

$$\hat{u}_n = \text{argmin}_{u\in\mathbb{R}^p}\{V_n(u)\} + o_p(1). \tag{8.3}$$

It is easy to see that the finite dimensional distributions of $V_n(u)$ converge to those of $V(u)$. Since both $V_n(u)$ and $V(u)$ are convex functions in terms of u, we claim that $\text{argmin}_{u\in\mathbb{R}^p}\{V_n(u)\} \to_L \text{argmin}_{u\in\mathbb{R}^p}\{V(u)\}$ as $n \to \infty$. Hence, by (8.3), the conclusion holds. This completes the proof. □

Proof of Theorem 2. Note that the existence of $\hat{\theta}_n^o$ is guaranteed since $\hat{\theta}_n^o$ is the minimizer of the continuous function $L_n(\theta)$ over the compact subset. Denote $\hat{u}_n = \sqrt{n}(\hat{\theta}_n^o - \theta_0)$. Then

$$\hat{u}_n = \text{argmin}_{u\in\mathbb{R}^p} L_n(\theta_0 + \frac{u}{\sqrt{n}}).$$

By Taylor's expansion and ergodic theorem, we have

$$\begin{aligned} L_n(\theta_0 + \frac{\hat{u}_n}{\sqrt{n}}) - L_n(\theta_0) &= 2\hat{u}_n' D_n + \hat{u}_n'[\Omega + o_p(1)]\hat{u}_n \\ &\quad + \frac{\lambda_n}{\sqrt{n}} \sum_{i=1}^{\tilde{p}} \frac{\hat{u}_{in}}{\theta_{i0} + o_p(1)} \text{sgn}(\theta_{i0}) I(\theta_{i0} \neq 0) \\ &\quad + \lambda_n \sum_{i=1}^{\tilde{p}} \left| \frac{\hat{u}_{in}}{\sqrt{n}\tilde{\theta}_{in}} \right| I(\theta_{i0} = 0)\,, \end{aligned} \tag{8.4}$$

where $D_n = \sum_{t=1}^n [\partial\varepsilon_t(\theta_0)/\partial\theta]\varepsilon_t/\sqrt{n} \to_L D \equiv N(0, \sigma^2\Omega)$ as $n \to \infty$. Since Ω is positive definite, we see that $\hat{u}_n = O_p(1)$. Otherwise, we will have $P(L_n(\hat{\theta}_n^o) - L_n(\theta_0) > \eta) > 1-\epsilon$ for some η and any ϵ, which is a contradiction

to the definition of $\hat{\theta}_n^o$. Thus,

$$L_n(\theta_o + \frac{\hat{u}_n}{\sqrt{n}}) - L_n(\theta_0) = V_n(\hat{u}_n) + o_p(1),$$

where

$$\begin{aligned} V_n(u) &= 2u'D_n + u'\Omega u + \frac{\lambda_n}{\sqrt{n}} \sum_{i=1}^{\tilde{p}} \frac{u_i}{\theta_{i0}} \text{sign}(\theta_{i0}) I(\theta_{i0} \neq 0) \\ &\quad + \lambda_n \sum_{i=1}^{\tilde{p}} \left| \frac{u_i}{\sqrt{n}\tilde{\theta}_{in}} \right| I(\theta_{i0} = 0)\,. \end{aligned} \tag{8.5}$$

Since $\lambda_n \to \infty$, $\lambda_n/\sqrt{n} \to \lambda_0$ and $\sqrt{n}\tilde{\theta}_{in} I(\theta_{i0} = 0) \to_L \xi_i I(\theta_{i0} = 0)$ when $n \to \infty$, we can show that $V_n(u) \longrightarrow_d V(u)$ for every u, where

$$V(u) = \begin{cases} 2u'D + u'\Omega u + \lambda_0 \sum_{i=1}^{\tilde{p}} \frac{u_i}{\theta_{i0}} \text{sign}(\theta_{i0}) I(\theta_{i0} \neq 0)\,, \\ \qquad \text{if } \sum_{i=1}^{\tilde{p}} u_i I(\theta_{i0} = 0) = 0\,, \\ \infty, \text{ otherwise}\,, \end{cases}$$

$V_n(u)$ is convex, and the unique minimum of $V(u)$ is u^*, where the subvector of u^* consisting of the component corresponding to $\theta_{i0} \neq 0$ is $\Omega_1^{-1} D_1 + B_1$, while the component of u^* corresponding to $\theta_{i0} = 0$ is 0. Note that

$$\hat{u}_n = \text{argmin}_{u \in \mathbb{R}^p} V_n(u) + o_p(1).$$

Using the argmax theorem as in Knight and Fu (2000), we conclude that (b) of Theorem 2 holds and $\hat{u}_{in} = \sqrt{n}\hat{\theta}_{in} \to_d 0$ if $\theta_{i0} = 0$.

To show (a) of Theorem 2, by the first-order optimality conditions, if

$\theta_{i0} = 0$ and $\hat{\theta}_{in} \neq 0$, then

$$T_n \equiv 2\sum_{t=1}^{n} \frac{\partial \varepsilon_t(\hat{\theta}_n)}{\partial \theta_i} \varepsilon_t(\hat{\theta}_n) + \frac{\lambda_n \text{sign}(\hat{\theta}_{in})}{|\sqrt{n}\tilde{\theta}_{in}|} = 0\,. \tag{8.6}$$

Note that, if $\theta_{i0} = 0$, then

$$\begin{aligned}
\frac{2}{\sqrt{n}}\sum_{t=1}^{n} \frac{\partial \varepsilon_t(\hat{\theta}_n)}{\partial \theta_i} \varepsilon_t(\hat{\theta}_n) &= \frac{2}{\sqrt{n}}\sum_{t=1}^{n} \frac{\partial \varepsilon_t(\theta_0)}{\partial \theta_i} \varepsilon_t(\theta_0) \\
&\quad +2E\left[\frac{\partial \varepsilon_t(\theta_0)}{\partial \theta_i}\frac{\partial \varepsilon_t(\theta_0)}{\partial \theta'}\right]\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1) \\
&= \frac{2}{\sqrt{n}}\sum_{t=1}^{n} \frac{\partial \varepsilon_t(\theta_0)}{\partial \theta_i} \varepsilon_t(\theta_0) \\
&\quad +2E\left[\frac{\partial \varepsilon_t(\theta_0)}{\partial \theta_i}\frac{\partial \varepsilon_t(\theta_0)}{\partial \theta'_{11}}\right]\sqrt{n}(\hat{\theta}_{1n} - \theta_{10}) + o_p(1) \\
&\to_d \quad \text{some normal distribution}\,,
\end{aligned}$$

and $\frac{\lambda_n/\sqrt{n}}{\sqrt{n}\tilde{\theta}_{in}} \to_p 0$, where θ_{11} is the unknown parameter vector of θ_{10}. Thus,

$$P(\hat{\theta}_{in} I\{\theta_{i0} = 0\} \neq 0) \leq P(\frac{1}{\sqrt{n}} T_n = 0) \to 0\,,$$

as $n \to \infty$, that is, (a) holds. This completes the proof. □

Proof of Theorem 3. Denote

$$\varepsilon_t(\theta) = w_t - \sum_{i=1}^{p-1} \phi_i^* w_{t-i} + \sum_{j=1}^{q} \psi_j \varepsilon_{t-j}(\theta). \tag{8.7}$$

Then, $\varepsilon_t(0, \theta_0) = \varepsilon_t(\theta_0) = \varepsilon_t$. First, $\varepsilon_t(c, \theta)$ has the following expansion:

$$\varepsilon_t(c, \theta) = c\sum_{i=1}^{t} \beta_{i-1} y_{t-i} + \sum_{i=0}^{\infty} \beta_i w_{t-i} = c\sum_{i=1}^{t} \beta_{i-1} y_{t-i} + \varepsilon_t(\theta), \tag{8.8}$$

where β_i is the coefficient in the representation: $\psi^{-1}(z) = \sum_{i=0}^{\infty} \beta_i z^i$ with $\beta_i = O(\rho^i)$ and $\rho \in (0,1)$. Thus,

$$\varepsilon_t^2(c,\theta) = \varepsilon_t^2(\theta) + 2c\left(\sum_{i=1}^{t} \beta_{i-1} y_{t-i}\right)\varepsilon_t(\theta) + c^2\left[\sum_{i=1}^{t} \beta_{i-1} y_{t-i}\right]^2. \quad (8.9)$$

By Taylor's expansion,

$$\begin{aligned}\varepsilon_t(\theta) &= \varepsilon_t + (\theta-\theta_0)'\frac{\partial \varepsilon_t(\theta_0)}{\partial \theta} + \frac{1}{2}(\theta-\theta_0)'\frac{\partial^2 \varepsilon_t(\theta^*)}{\partial\theta\partial\theta'}(\theta-\theta_0) \\ &= \varepsilon_t + (\theta-\theta_0)'\frac{\partial \varepsilon_t(\theta_0)}{\partial \theta} + O(1)\|\theta-\theta_0\|^2\xi_t,\end{aligned}$$

where θ^* is between θ and θ_0, and $\xi_t = \sum_{i=1}^{\infty} \rho^i |w_{t-i}|$ with $\rho \in (0,1)$. By Lemma 1 of Yap and Reinsel (1995a),

$$\sum_{i=1}^{t} \beta_{i-1} y_{t-i} = \psi^{-1}(1) y_{t-1} + r_{t-1}, \quad (8.10)$$

where r_{t-1} is a function in terms of random variables $\{\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots, \varepsilon_1\}$ and Er_{t-1}^2 is uniformly bounded in t. Since $c \in [-\delta/n, \delta/n]$, we have

$$c\sum_{t=1}^{n} r_{t-1}\varepsilon_t(\theta) = O_p(1)\|\theta-\theta_0\|. \quad (8.11)$$

By Lemma 3.4.3 of Chan and Wei (1988), we have $\sum_{t=1}^{n} y_{t-1}\partial\varepsilon_t(\theta_0)/\partial\theta = o_p(n^{-3/2})$. Thus, we can show that

$$c\sum_{t=1}^{n} y_{t-1}\varepsilon_t(\theta) = c\sum_{t=1}^{n} y_{t-1}\varepsilon_t + \sqrt{n}\|\theta-\theta_0\| o_p(1) + O_p(1)\sqrt{n}\|\theta-\theta_0\|^2. \quad (8.12)$$

By (8.10)-(8.12), we have

$$\begin{aligned}c\sum_{t=1}^{n}\left(\sum_{i=1}^{t}\beta_{i-1}y_{t-i}\right)\varepsilon_t(\theta) &= \psi^{-1}(1)c\sum_{t=1}^{n} y_{t-1}\varepsilon_t + \sqrt{n}\|\theta-\theta_0\| o_p(1) + O_p(1)[\|\theta-\theta_0\| \\ &\quad + \sqrt{n}\|\theta-\theta_0\|^2]. \quad (8.13)\end{aligned}$$

As explained below Lemma 1 of Yap and Reinsel (1995a), $\sum_{t=1}^n r_{t-1}^2$ and $\sum_{t=1}^n y_{t-1}r_{t-1}$ are of order smaller than $\sum_{t=1}^n y_{t-1}^2$. Thus, by (8.10), we have

$$c^2\sum_{t=1}^n\left(\sum_{i=1}^t \beta_{i-1}y_{t-1}\right)^2 = \psi^{-2}(1)c^2\sum_{t=1}^n y_{t-1}^2 + o_p(1). \tag{8.14}$$

By (8.9) and (8.13)-(8.14), it follows that

$$\begin{aligned}\sum_{t=1}^n \varepsilon_t^2(c,\theta) &= \sum_{t=1}^n \varepsilon_t^2(\theta) + 2\psi^{-1}(1)c\sum_{t=1}^n y_{t-1}\varepsilon_t + \psi^{-2}(1)c^2\sum_{t=1}^n y_{t-1}^2 \\ &\quad +o_p(1) + \sqrt{n}\|\theta-\theta_0\|o_p(1) \\ &\quad +O_p(1)[\|\theta-\theta_0\| + \sqrt{n}\|\theta-\theta_0\|^2]. \end{aligned}\tag{8.15}$$

It is known that

$$\frac{1}{n^2}\sum_{t=1}^n y_{t-1}^2 \to_L \phi^{*-2}(1)\psi^2(1)\sigma^2\int_0^1 B^2(\tau)\,d\tau\,, \tag{8.16}$$

$$\frac{1}{n^2}\sum_{t=1}^n y_{t-1}\varepsilon_t \to_L \phi^{*-1}(1)\psi(1)\sigma^2\int_0^1 B(\tau)\,dB(\tau)\,, \tag{8.17}$$

see, e.g., Yap and Reinsel (1995b). Denote

$$L_n(\theta) = \sum_{t=1}^n \varepsilon_t^2(\theta) + \sum_{i=1}^{p+q-1}\lambda_{in}|\theta_i|. \tag{8.18}$$

By (8.15)-(8.18), we have

$$\frac{1}{n}\sup_{(c,\theta)\in\Theta_n}\left|\tilde{L}_n(c,\theta) - L_n(\theta)\right| = o_p(1).$$

Furthermore, as for Theorem 1, we have $\hat{\theta}_n \to_p \theta_0$. Thus, (a) holds.

Note that $\tilde{L}_n(0,\theta_0) = L_n(\theta_0)$. Denote $\hat{u}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$. As for Theorem 2, using (8.15) and (8.18), we have the following expansion

$$\begin{aligned}
\tilde{L}_n(\hat{c}_n,\hat{\theta}_n) - \tilde{L}_n(0,\theta_0) &= L_n(\hat{\theta}_n) - L_n(\theta_0) + \sum_{t=1}^{n}[\varepsilon_t^2(\hat{c}_n,\hat{\theta}_n) - \varepsilon_t^2(\hat{\theta}_n)] \\
&= 2\hat{u}_n' D_n + \hat{u}_n'[\Omega + o_p(1)]\hat{u}_n \\
&\quad +2\psi^{-1}(1)\hat{c}_n \sum_{t=1}^{n} y_{t-1}\varepsilon_t + \psi^{-2}(1)\hat{c}_n^2 \sum_{t=1}^{n} y_{t-1}^2 \\
&\quad +o_p(\hat{u}_n) + O_p(1)(\|\hat{u}_n\| + \|\hat{u}_n\|^2)/\sqrt{n} \\
&\quad + \sum_{i=1}^{p+q-1} \frac{\lambda_{in}}{\sqrt{n}}[\hat{u}_{in}\mathrm{sgn}(\theta_{i0})I(\theta_{i0}\neq 0) + |\hat{u}_{in}|I(\theta_{i0}=0)]\,.
\end{aligned}$$

From the previous equation, as for Theorem 2, we have $\hat{u}_n = O_p(1)$ and hence

$$\begin{aligned}
\tilde{L}_n(\hat{c}_n,\hat{\theta}_n) - \tilde{L}_n(0,\theta_0) &= 2\psi^{-1}(1)\hat{c}_n \sum_{t=1}^{n} y_{t-1}\varepsilon_t + \psi^{-2}(1)\hat{c}_n^2 \sum_{t=1}^{n} y_{t-1}^2 \\
&\quad +2\hat{u}_n' D_n + \hat{u}_n'\Omega\hat{u}_n + o_p(1) + \sum_{i=1}^{p+q-1} \lambda_{i0} \\
&\quad \cdot[\hat{u}_{in}\mathrm{sgn}(\theta_{i0})I(\theta_{i0}\neq 0) + |\hat{u}_{in}|I(\theta_{i0}=0)]\,. \qquad (8.19)
\end{aligned}$$

Since $\hat{u}_n$ and $\hat{c}_n$ are the minimizer of $L_n(c,\theta)$, from the previous equation, we can see that

$$\hat{c}_n = \psi(1)(\sum_{t=1}^{n} y_{t-1}^2)^{-1} \sum_{t=1}^{n} y_{t-1}\varepsilon_t + o_p(1), \qquad (8.20)$$

$$\hat{u}_n = \mathrm{argmin}_{u\in\mathbb{R}^p}\{V_n(u)\} + o_p(1)\,, \qquad (8.21)$$

where

$$\begin{aligned} V_n(u) &= 2u'D_n + u'\Omega u \\ &\quad + \sum_{i=1}^{p+q-1} \lambda_{i0}[u_i \mathrm{sgn}(\theta_{i0}) I(\theta_{i0} \neq 0) + |u_i| I(\theta_{i0} = 0)] \,. \end{aligned}$$

By (8.16)-(8.17), (8.20)-(8.21) and the continuous mapping theorem, the conclusion (b)-(i) holds. Similar to Theorem 2, it can be shown that the conclusion (b)-(ii) holds. This completes the proof. □

Proof of Theorem 4. Denote $\hat{u}_n = \sqrt{n}(\hat{\theta}_n^o - \theta_0)$. As for Theorem 3, we can show that

$$\begin{aligned} L_n(\hat{c}_n, \hat{\theta}_n^o) - L_n(0, \theta_0) &= 2\psi^{-1}(1)\hat{c}_n \sum_{t=1}^{n} y_{t-1}\varepsilon_t + \psi^{-2}(1)\hat{c}_n^2 \sum_{t=1}^{n} y_{t-1}^2 \\ &\quad + 2\hat{u}_n' D_n + \hat{u}_n'[\Omega + o_p(1)]\hat{u}_n \\ &\quad + \frac{\lambda_n}{\sqrt{n}} \sum_{i=1}^{\tilde{p}} \frac{\hat{u}_{in}}{\theta_0 + o_p(1)} \mathrm{sgn}(\theta_{i0}) I(\theta_{i0} \neq 0) \\ &\quad + \lambda_n \sum_{i=1}^{\tilde{p}} \left| \frac{\hat{u}_{in}}{\sqrt{n}\tilde{\theta}_{in}} \right| I(\theta_{i0} = 0) \,, \end{aligned} \tag{8.22}$$

where $D_n = \sum_{t=1}^{n} [\partial \varepsilon_t(0, \theta_0)/\partial \theta]\varepsilon_t/\sqrt{n} \to_L N(0, \sigma^2\Omega)$ as $n \to \infty$. Similar to the argument for Theorem 3, we can show that the conclusion holds and the details are omitted. This completes the proof. □

Acknowledgements

We would like to thank the Editor, an Associate Editor and three anonymous referees for their thoughtful and useful comments, which led to an improved version of this paper. Research supported in part by HKSAR-RGC-GRF Nos 14300514 and 14325216 (Chan); the Theme-Based Research Fund of HKSAR-RGC-TRF No. T32-101/15-R (Chan); HKSAR-RGC-GRF Nos 16307516, 16500117 and 16500915 (Ling); HKSAR-RGC-GRF Nos 405113, 14305517 and 14601015 (Yau).

References

Akaike, H. (1977). On entropy maximisation principle. In *Applications of Statistics* (P. R. Krishnaiah, ed.), pp. 27–41. North Holland, Amsterdam.

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods, 2nd eds.* Springer, New York.

Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Model selection for multivariate failure time data. *Biometrika* 92, 303–316.

Chan, N. H. and Wei, C. Z.(1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.* **16**, 367–401.

Chan, N. H. (2010). *Time Series: Applications to Finance with R and S-Plus, 2nd eds.* Wiley, New York.

Chen, K. and Chan, K.S. (2011). Subset ARMA selection via the adaptive Lasso. *Statistics and*

its Interface, **4**, 197–205.

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

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74–99.

Fu, W. J. (1998). Penalized regression: the bridge versus the lasso. *J. Comput. Graph. Statist.* **7**, 397–416.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8** , 1071–1081.

Hannan, E.J. and Kavalieris, L. (1984). A method for autoregressive-moving average estimation. *Biometrika*, **71**, 273–280.

Huang, J., Ma, S., and Zhang, C. (2008). Adaptive lasso for sparse high dimensional regression. *Statistica Sinica* **18**, 1603–1618.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–1378.

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

Liao, Z.P. and Phillips, P.C.B. (2015). Automated estimation of vector error correction models. *Econometric Theory*, **31**, 581–646.

Ling, S. and McAleer, M. (2010). A general asymptotic theory for time-series models. *Statistica Neerlandica* **64**, 97–111.

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

Phillips, P.C.B. (1987). Time series regression with a unit root. *Econometrica* **55**, 277–301.

Pötscher, B.M. (1983). Order estimation in ARMA-models by Lagrangian multiplier tests. *Ann. Statist.* **11**, 872–885.

Pötscher, B.M. (1990). Estimation of autoregressive moving-average order given an infinite number of models and approximation of spectral densities. *J. Time Ser. Anal.* **11**, 165–179.

Pötscher, B. M. and Srinivasan, S. (1994). A comparison of order estimation procedures for ARMA models. *Statist. Sinica* **4**, 29–50.

Rissanen, J.(1978). Modeling by shortest data description. *Automatica* **14**, 465–471.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Song, S. and Bickel, P.J. (2011). Large vector autoregressions. Working paper, University of California, Berkeley.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–288.

Wang, H., Li, G. and Tsai, C.-L. (2007) Regression coefficients and autoregressive order shrink-

age and selection via the lasso, *J. R. Statist. Soc. B* **69**, 63-78.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* **68**, 49–67.

Yap, S.F. and Reinsel, G.C. (1995a). Estimation and testing for unit roots in a partially nonstationary vector autoregressive moving average model. *J. Amer. Statist. Assoc.* **90**, 253–267.

Yap, S.F. and Reinsel, G.C. (1995b). Results on estimation and testing for a unit root in the nonstationary autoregressive moving-average model. *J. Time Ser. Anal.* **16**, 339-353.

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

Department of Statistics, The Chinese University of Hong Kong

E-mail: nhchan@sta.cuhk.edu.hk

Department of Mathematics, The Hong Kong University of Science and Technology

E-mail: maling@usk.hk

Department of Statistics, The Chinese University of Hong Kong

E-mail: cyyau@sta.cuhk.edu.hk