Journal of Econometrics 189 (2015) 285–296

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

LASSO estimation of threshold autoregressive models

Ngai Hang Chan [a,b], Chun Yip Yau [b], Rong-Mao Zhang [c,*]

[a] Southwestern University of Finance and Economics, China
[b] Chinese University of Hong Kong, Hong Kong
[c] Zhejiang University, Hangzhou, China

ARTICLE INFO

Article history:
Available online 2 April 2015

JEL classification:
C22

Keywords:
Group lasso
Information criterion
Least angle regression (LARS)
Multiple regimes

ABSTRACT

This paper develops a novel approach for estimating a threshold autoregressive (TAR) model with multiple-regimes and establishes its large sample properties. By reframing the problem in a regression variable selection context, a least absolute shrinkage and selection operator (LASSO) procedure is proposed to estimate a TAR model with an unknown number of thresholds, where the computation can be performed efficiently. It is further shown that the number and the location of the thresholds can be consistently estimated. A near optimal convergence rate of the threshold parameters is also established. Simulation studies are conducted to assess the performance in finite samples. The results are illustrated with an application to the quarterly US real GNP data over the period 1947–2009.

1. Introduction

A time series $\{Y_t\}$ is said to follow a threshold autoregressive (TAR) model if it satisfies

$$Y_t = \phi_o^{(j)} + \sum_{i=1}^{p} \phi_i^{(j)} Y_{t-i} + \sigma_j \eta_t, \quad \text{for } r_{j-1} < Y_{t-d} \le r_j, \tag{1.1}$$

where σ_j's are positive constants $j = 1, \ldots, m+1$ and d is a positive integer known as the delay parameter, which is assumed to be known. The vector $\mathbf{r} = (r_1, \ldots, r_m)$, $-\infty < r_1 \le r_2 \le \cdots \le r_m < \infty$, is the threshold parameter that partitions the process into $m+1$ regimes. By convention, $r_0 = -\infty$ and $r_{m+1} = \infty$. The error $\{\eta_t\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with zero mean and unit variance.

Proposed by Tong (1978), the TAR model constitutes an important class of non-linear time series models. This model has many desirable features including: asymmetric limit cycles, amplitude dependent frequencies, jump resonance and chaos, among others, which are observed in many empirical studies and cannot be explained by linear time series models. Moreover, the regime-changing autoregressive structure of the TAR model offers a simple and convenient means for interpretation. As a result, the TAR model has attracted considerable attention and has been widely used in diverse areas such as biological sciences, econometrics, environmental sciences, finance, hydrology, physics and population dynamics. Excellent surveys on TAR models can be found in Tong (1990, 2010).

Probabilistic properties and estimation theory of TAR models are well explored. For example, Chan and Tong (1985), Chen and Tsay (1991), Brockwell et al. (1992), Liu and Susko (1992) and An and Huang (1996), among others, studied the stationarity and ergodicity of TAR models. The large sample theory of the least squares estimation (LSE) of two-regime TAR models was established in Chan (1993), who showed that the estimated threshold is n-consistent and its limiting distribution is the minimizer of a one-dimensional two-sided compound Poisson process. Recently, Chan and Kutoyants (2010) derived the asymptotic distribution and compared the asymptotic efficiency of maximum likelihood and Bayesian estimates for the thresholds, and Li and Ling (2012) established the limiting distributions of the least squares estimates for the thresholds of TAR model with multiple-regimes.

Despite the theoretical advances and practical relevance of non-linear time series, the TAR model has not enjoyed the same popularity as the linear counterpart. One of the reasons is due to the high computational costs associated with the estimation procedure. When the TAR models has more than two regimes, the global minimum of the least squares criterion requires multi-parameter grid-based search over all possible values of all threshold parameters (Li and Ling (2012)). Tsay (1989) proposed a graphical approach to determine the number and location of the thresholds. More recently, Gonzalo and Pitarakis (2002) suggested a sequential

* Corresponding author.
E-mail address: rmzhang@zju.edu.cn (R.-M. Zhang).

http://dx.doi.org/10.1016/j.jeconom.2015.03.023
0304-4076/

estimation procedure for multiple thresholds. Coakley et al. (2003) provided a different estimation approach, which relies on the computational advantages of QR factorizations of matrices. Nevertheless, the computational burden is still growing exponentially with the sample size. Moreover, the results of Gonzalo and Pitarakis (2002) and Coakley et al. (2003) assume that the σ_j's are equal. A computation efficient procedure which can be applied in general situations and large sample size is still missing.

In this paper we focus on tackling the computational challenge inherited in the estimation of multiple threshold models. Motivated by the computationally efficient least angle regression by Efron et al. (2004) and its close connection to the well known Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996), we reframe the problem of estimating multiple-regime TAR models as a model selection problem, so that an efficient algorithm can be developed to give a computationally feasible solution. The fast algorithm allows the estimation to be performed in $O(m^3+mn)$ order, where m is the number of thresholds. We also show that the number and the location of the thresholds can be consistently estimated. Finally, a near optimal convergence rate of the threshold parameters is established.

In change-point analysis, the use of LASSO on estimating the number and the location of change-points has been considered by Harchaoui and Levy-Leduc (2010) and Bleakley and Vert (2011) in i.i.d. data and Chan et al. (2014) in time series. Although threshold models are similar to change-point models, the change occurs in the state of the lag-d observation Y_{t-d} rather than at the time index. Therefore the methods in change-point analysis are not directly applicable for threshold estimation. To see this, let $(Y_{1-d}, Y_{2-d}, \ldots, Y_{n-d})^{\mathrm{T}}$ be the observations and $(Y_{\pi(1)}, Y_{\pi(2)}, \ldots, Y_{\pi(n)})^{\mathrm{T}}$ be the order statistics of the observations, from the smallest to the largest. In other words, $\pi(i)$ is the time index of the $(i-d)$th smallest element in $\mathbf{Y}$. Then the sorted observations $\boldsymbol{Y}_n^0 = (Y_{\pi(1)+d}, \ldots, Y_{\pi(n)+d})^{\mathrm{T}}$ can be viewed as a change-point model.

However, this sequence is no longer stationary. Thus, the ordinary large deviation principle for stationary processes, which is the key to establish the theoretical properties for change-point estimations, cannot be used in threshold models. See Hansen (2000) for more details. The computation and theoretical treatment for threshold models require different tools.

From the computational point of view, this paper modifies the LASSO procedure in change-points to the framework of threshold models. In particular, a new design matrix is constructed by a sorting operation to reformulate the threshold model into a high dimensional regression model. More importantly, from the statistical standpoint, we develop a large deviation probability bound for the quantity $\frac{1}{\sqrt{n}}\left|\sum_{t=1}^{n} Y_t^j I(Y_{t-d} \le x)\epsilon_t\right|$ using a bracketing entropy technique to show the consistency and the convergence rate of the threshold estimator, where $\varepsilon_t = \sigma_j \eta_t I(r_{j-1} < Y_{t-d} \le r_j)$, $t = 1, \ldots, n$. As a result, the conditions required and the convergence rate obtained for the threshold estimators are different from that of the change-point estimators. See Section 2 and Appendix A for details.

This paper is organized as follows. Section 2 presents the estimation procedure and theoretical results. Simulation studies and data applications are given Sections 3 and 4 respectively. We conclude in Section 5. The proofs are given in Appendix A.

2. Estimation

In this section, we introduce a two-step procedure for consistent and computationally efficient threshold estimation. The first step obtains a set of potential thresholds by a LASSO estimation. The second step is a model selection procedure that selects the consistency estimates of the thresholds from the set of potential thresholds. The two steps are discussed in Sections 2.1 and 2.2 respectively. Section 2.3 describes the implementation of the two-step procedure.

2.1. One-step group LASSO estimate

In this subsection, we introduce the LASSO estimation procedure by reformulating the TAR model as a high dimensional variable selection problem. Then we study the asymptotic theory of the estimator. For notational simplicity we assume that $\mathbf{Y} = (Y_{1-d}, \ldots, Y_n)^T$ are observed. Let $(Y_{\pi(1)}, Y_{\pi(2)}, \ldots, Y_{\pi(n)})^{\mathrm{T}}$ be the order statistics of $(Y_{1-d}, Y_{2-d}, \ldots, Y_{n-d})^{\mathrm{T}}$, from the smallest to the largest.

Let $\boldsymbol{Y}_n^0 = (Y_{\pi(1)+d}, \ldots, Y_{\pi(n)+d})^{\mathrm{T}}$ be the sorted observations, $\boldsymbol{\theta}(n) = (\boldsymbol{\theta}_{\pi(1)}^T, \ldots, \boldsymbol{\theta}_{\pi(n)}^T)^{\mathrm{T}}$ and $\boldsymbol{\varepsilon}(n) = (\varepsilon_{\pi(1)+d}, \ldots, \varepsilon_{\pi(n)+d})^{\mathrm{T}}$ be the corresponding parameter vectors and error terms. Let $\boldsymbol{X}_n$ be an $n \times np$ matrix defined by

$$\boldsymbol{X}_n = \begin{pmatrix} \boldsymbol{Y}_{\pi(1)}^{\mathrm{T}} & 0 & 0 & \ldots & 0 \\ \boldsymbol{Y}_{\pi(2)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(2)}^{\mathrm{T}} & 0 & \ldots & 0 \\ \boldsymbol{Y}_{\pi(3)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(3)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(3)}^{\mathrm{T}} & \ldots & 0 \\ \vdots & & & \ddots & \vdots \\ \boldsymbol{Y}_{\pi(n)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(n)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(n)}^{\mathrm{T}} & \ldots & \boldsymbol{Y}_{\pi(n)}^{\mathrm{T}} \end{pmatrix}, \tag{2.1}$$

where $\mathbf{Y}_{\pi(j)}^{\mathrm{T}} = (1, Y_{\pi(j)+d-1}, Y_{\pi(j)+d-2}, \ldots, Y_{\pi(j)+d-p})$. Set $\boldsymbol{\phi}_j = (\phi_0^{(j)}, \phi_1^{(j)}, \ldots, \phi_p^{(j)})$. If $\boldsymbol{\theta}_{\pi(1)} = \boldsymbol{\phi}_1$, $\boldsymbol{\theta}_{\pi(t_j)} = \boldsymbol{\phi}_{j+1} - \boldsymbol{\phi}_j$ for $t_j \in \{2, \ldots, n\}$, $j = 1, \ldots, m$ satisfying $Y_{\pi(t_j-1)} \le r_j < Y_{\pi(t_j)}$, and $\boldsymbol{\theta}_{\pi(j)} = 0$ for $j \in \{2, \ldots, n\} \setminus \{t_1, \ldots, t_m\}$, then it can be seen that model (1.1) can be expressed as a high dimensional regression model as

$$\boldsymbol{Y}_n^0 = \boldsymbol{X}_n \boldsymbol{\theta}(n) + \boldsymbol{\varepsilon}(n). \tag{2.2}$$

The parameter $\boldsymbol{\theta}_{\pi(j)}$ can be interpreted as the change in the AR parameter vector when the value of the lag-d observation is switching from $Y_{\pi(j)}$ to $Y_{\pi(j)+1}$. Since only m of the vectors $\boldsymbol{\theta}_{\pi(j)}$s in $\boldsymbol{\theta}(n)$ are non-zero, we look for a sparse solution to the high dimension regression model (2.2). A well-known solution to this problem is given by the group lasso estimation (Yuan and Lin (2006)). Thus, we propose to estimate $\boldsymbol{\theta}(n)$ by the following group LASSO equation:

$$\widehat{\boldsymbol{\theta}}(n) = \underset{\boldsymbol{\theta}(n)}{\operatorname{argmin}} \frac{1}{n} \|\boldsymbol{Y}_n^0 - \boldsymbol{X}_n \boldsymbol{\theta}(n)\|^2 + \lambda_n \sum_{i=1}^{n} \|\boldsymbol{\theta}_i\|, \tag{2.3}$$

where λ_n is a regularization parameter and $\|\cdot\|$ is the l_2-norm. Note that when $\widehat{\boldsymbol{\theta}}_{\pi(j)} \ne 0$, $j \ge 2$, there is a change in the autoregressive parameters $\{\widehat{\boldsymbol{\phi}}_l,\ l = 1, 2, \ldots, m\}$ at level $Y_{\pi(j)}$. Thus the threshold r_j, $j = 1, 2, \ldots, m$ can be estimated from the $Y_{\pi(j)}$, $(j \ge 2)$ when $\widehat{\boldsymbol{\theta}}_{\pi(j)}$ is non-zero. Denote the estimates of the thresholds by

$$\mathcal{A}_n = \{Y_{\pi(j-1)} : \widehat{\boldsymbol{\theta}}_{\pi(j)} \ne 0,\ j \ge 2\}. \tag{2.4}$$

Let $|\mathcal{A}_n|$ be the cardinality of the set $\mathcal{A}_n$ and denote the elements of $\mathcal{A}_n$ by $Y_{\pi(\hat{t}_1)}, \ldots, Y_{\pi(\hat{t}_{|\mathcal{A}_n|})}$. Note that $\hat{m} := |\mathcal{A}_n|$ is the estimated number of thresholds and $Y_{\pi(\hat{t}_j)}$ is the jth estimated threshold. In view of the relationship between $\boldsymbol{\phi}_j$ and $\boldsymbol{\theta}_j$ in (2.2), the AR parameters in each segment can be estimated by

$$\widehat{\boldsymbol{\phi}}_1 = \widehat{\boldsymbol{\theta}}_{\pi(1)} \quad \text{and} \quad \widehat{\boldsymbol{\phi}}_j = \sum_{l=1}^{\hat{t}_j} \widehat{\boldsymbol{\theta}}_{\pi(l)} \quad j = 1, \ldots, \hat{m}. \tag{2.5}$$

In this paper, it is shown that the group LASSO estimate of the autoregressive parameters given by (2.5) is consistent in terms of the prediction error. Further, the estimates of the thresholds defined by (2.4) are consistent and their convergence rates are nearly optimal (except for a factor of $\log n$), when ε_1 has moments of all orders. To assess the performance of the procedure, several simulations and applications are presented in Sections 3 and 4.

Although (2.2) is very similar to the LASSO procedure for high dimension regression problem, the techniques used to develop the asymptotic properties are different from ordinary regression problems. In the classical high-dimension regression setting, the explanatory variables are not strongly correlated with each other. For example, the *restricted eigenvalue* conditions (e.g., Bickel et al. (2009)) require any subset of explanatory variables of certain sizes to form a positive definite design matrix. However, in the threshold time series setting, the restricted eigenvalue assumption does not hold. As a result, the consistency property in the classical setting cannot be established by means of this assumption. Instead, it is shown in Theorem 2.3 that the number of the thresholds will be overestimated. Therefore, we propose a two-step procedure to yield a consistent set of estimated thresholds, see Theorems 2.4 and 2.5.

Let $\{r_i^0, i = 1, \ldots, m_0\}$ be the true thresholds and $\boldsymbol{\phi}_j^0$ be the true AR parameter vector in the jth regime, $j = 1, \ldots, m_0 + 1$. For simplicity we assume that all $\boldsymbol{\phi}_j$s are of dimension p for some fixed p, see Remark 2.3 for implementations in general. To discuss the asymptotic results for the LASSO procedure, first impose the following assumptions:

Assumptions.

H1: $\{\eta_t\}$ has a bounded, continuous and positive density function and $\mathrm{E}|\eta_1|^{2+\iota} < \infty$ for some $\iota > 0$.

H2: $\{Y_t\}$ is a α-mixing stationary process with a geometric decaying rate and $\mathrm{E}|Y_t|^{2+\iota} < \infty$.

H3: $\min_{1\le i\le m_0+1} \|\boldsymbol{\phi}_i^0 - \boldsymbol{\phi}_{i-1}^0\| > \nu$, for some $\nu > 0$.

H4: there exist constants l, u such that $r_i \in [l, u]$, $1 \le i \le m_0$ and $\min_{2\le i\le m_0} |r_i - r_{i-1}|/\gamma_n \to \infty$ for some $\gamma_n \to 0$ such that $m_0\gamma_n \to 0$, $m_0\gamma_n^{-1}(n\gamma_n)^{-\iota/2}(\log n)^{2(2+\iota)} \to 0$ and $\gamma_n/\lambda_n \to \infty$ as $n \to \infty$, where $\{\lambda_n\}$ is the regularization parameter sequence given in (2.3).

Assumptions H1–H3 are regularity assumptions for standard asymptotic theories, see for example Chan (1993) and Li and Ling (2012). Assumption H2 is satisfied if $\{\varepsilon_i\}$ in each segment is i.i.d. sequence and $\sup_j \sum_{i=1}^p |\phi_{j,i}| < 1$, see for example Remark B of Chan (1993). Assumption H3 is a trivial condition to ensure that a change occurs at r_j. Lastly, Assumption H4 gives the conditions on the growth rate of the size of each regime related to the regularization parameter λ_n. Most of the literatures assume that $\min_i |r_i - r_{i-1}| > \delta$ for some $\delta > 0$. Assumption H4 only requires the size of each regime to be bigger than γ_n. Thus the number of thresholds m may tend to infinity. For example, when the number of thresholds m_0 is assumed fixed and r_js are constants, γ_n can be taken to be any values such that $\gamma_n = o(1)$ and $\gamma_n^{-1}(n\gamma_n)^{-\iota/2}(\log n)^{2(2+\iota)} \to 0$ and Assumption H3 holds. Given γ_n, the order of λ_n can be chosen accordingly, details will be given in Section 2.3.

Let $(\boldsymbol{\theta}^0(n), \boldsymbol{\theta}_j^0)$ be the true value of $(\boldsymbol{\theta}(n), \boldsymbol{\theta}_j)$. First we give the consistency result of the regression parameters in terms of the prediction error.

Theorem 2.1. *Under Assumptions* H1 *and* H2, *if* $a_n = (4p)^{-1}\sqrt{n}\lambda_n \to \infty$ *and* $m_0 \le m_n$ *for some* $m_n = o(\lambda_n^{-1})$, *then with some* $C > 0$ *and probability greater than* $1 - C[\exp(-a_n/15) + n^{-\iota/2}(\log n)^{4+2\iota}]$,

$$\frac{1}{n}\|\boldsymbol{X}_n(\widehat{\boldsymbol{\theta}}(n) - \boldsymbol{\theta}^0(n))\|_2^2 \le 2\lambda_n m_n \max_i \|\boldsymbol{\phi}_i^0 - \boldsymbol{\phi}_{i-1}^0\|. \tag{2.6}$$

The next theorem is about the consistency results for the estimate of the thresholds when the number of thresholds m_0 is known.

Theorem 2.2. *Let* $\hat{r}_j = Y_{\pi(\hat{\iota}_j)}$ *be the jth estimated threshold given in* (2.4). *Suppose Assumptions* H1–H4 *hold with* $\iota > 2$. *Then, on the set* $|\mathcal{A}_n| = m_0$ *with* m_0 *being known,*

$$P\left\{\max_{1\le j\le m_0} |\hat{r}_j - r_j^0| \le \gamma_n\right\} \to 1, \quad \text{as } n \to \infty. \tag{2.7}$$

Remark 2.1. The quantity γ_n in (2.7) can be regarded as the convergence rate of the threshold parameters $\hat{r}_j$s. The optimal rate of convergence for $\hat{r}_j$ is found to be $O(1/n)$ when the number of thresholds m_0 is known and the r_js are constants (Li and Ling (2012)). In this case, from Assumption H3, γ_n can be taken as $O(\log^{4+c} n/n^{\frac{\iota}{2+\iota}})$ for some $c > 0$. When ϵ_1 has finite moments of any order, ι can be arbitrarily large, and the order of γ_n is close to $O(1/n)$, which implies that the convergence rate of the LASSO estimator is nearly optimal.

Note that Theorem 2.2 requires that the cardinality of $\mathcal{A}_n$ equals to the true number m_0, which may be unrealistic because we usually do not have this information a priori. Let $\mathcal{A} = \{r_1^0, r_2^0, \ldots, r_{m_0}^0\}$ be the set of true thresholds and $d_H(A, B)$ be the Hausdorff distance between two sets A and B given by

$$d_H(A, B) = \max_{b\in B} \min_{a\in A} |b - a|,$$

where $d_H(A, \emptyset) = d_H(\emptyset, B) = 1$, and $\emptyset$ is the empty set, see Boysen et al. (2009). The following theorem shows that, when the true number m_0 of the thresholds is unknown, the threshold parameter m will be overestimated. However, each of the true thresholds can be identified within a γ_n neighborhood.

Theorem 2.3. *Under the conditions of Theorem* 2.2, *then as* $n \to \infty$,

$$P\{|\mathcal{A}_n| \ge m_0\} \to 1 \tag{2.8}$$

and

$$P\{d_H(\mathcal{A}_n, \mathcal{A}) \le \gamma_n\} \to 1, \tag{2.9}$$

where γ_n *is given in* H3.

2.2. Two-step estimation procedure

Theorem 2.3 provides a set of threshold estimates, which is usually larger than the true number m_0. Two immediate issues arise: (i) how to improve the estimate from Theorem 2.3 so that it would better approximate m_0, the true number of thresholds, and (ii) how to estimate the thresholds with a nearly optimal rate? These two issues are dealt with in this subsection.

From Theorem 2.3, with probability tending to 1, all the true thresholds are identified by $\mathcal{A}_n$ within a γ_n neighborhood. Therefore, the thresholds can be consistently estimated by choosing the "best possible subset" of thresholds in $\mathcal{A}_n$ according to some prescribed information criterion (IC). Given m and the thresholds $\mathbf{r} = (r_1, \ldots, r_m)$, an information criterion $IC(m, \mathbf{r})$ typically consists of a sum of a goodness-of-fit measure and a penalty term that accounts for the model complexity. Specifically, let $\boldsymbol{Y}_t = (1, Y_{t-1}, \ldots, Y_{t-p})^T$ and

$$\widehat{\widehat{\boldsymbol{\phi}}}_j = \left(\sum_{t=1}^n \boldsymbol{Y}_t\boldsymbol{Y}_t^{\mathrm{T}} I(r_{j-1} < Y_{t-d} \le r_j)\right)^{-1} \times \sum_{t=1}^n \boldsymbol{Y}_t Y_t I(r_{j-1} < Y_{t-d} \le r_j)$$

be the least squares estimator and $S_n(r_{j-1}, r_j) = \sum_{t=1}^{n}(Y_t - \widehat{\widehat{\boldsymbol{\phi}}}_j \boldsymbol{Y}_t)^2 I(r_{j-1} < Y_{t-d} \le r_j)$ be the residual sum of squares for the jth regime. Consider a general information criterion of the form

$$IC(m, \mathbf{t}) = S_n(r_1, r_2, \ldots, r_m) + m\omega_n, \tag{2.10}$$

where the least squares criterion $S_n(r_1, r_2, \ldots, r_m) = \sum_{j=1}^{m+1} S_n(r_{j-1}, r_j)$ is the goodness-of-fit measure and ω_n is the penalty term. We estimate the number and locations of the thresholds by solving

$$(\hat{\hat{m}}, \hat{\hat{\mathbf{r}}}) = \arg \min_{\substack{m\in(0,1,\ldots,|\mathcal{A}_n|),\\ \mathbf{r}=(r_1,\ldots,r_m)\subset\mathcal{A}_n}} IC(m, \mathbf{r}). \tag{2.11}$$

Some commonly used information criteria that take similar forms as (2.10) are the AIC of Li and Ling (2012) for selecting the number of regimes, BIC of Yao (1988) and the MDL of Davis et al. (2006) for change-point detection. In these papers, the best subset of thresholds or change-points is chosen over all possible locations, which could be computationally challenging when m_0 is large. In contrast, the minimizing domain in (2.11) is a much smaller set, namely over the set $\mathcal{A}_n$. In practice, all possible subsets of $\mathcal{A}_n$ have to be evaluated to yield the threshold estimates. The following theorem shows that $(\hat{\hat{m}}, \hat{\hat{\mathbf{r}}})$ in (2.11) are consistent estimates of m_0 and the thresholds.

Theorem 2.4. *If the information criteria (IC) is defined in (2.10) with ω_n satisfying $\lim_{n\to\infty} m_0\omega_n/(n \min_{1\le i\le m_0} |r_i^0 - r_{i-1}^0|) = 0$ and $\lim_{n\to\infty} \omega_n/(n\gamma_n m_0) \ge 3 + C_0$ for the constant C_0 given in (A.18), then under conditions* H1–H4 *with $\iota > 2$, the minimizer $(\widehat{\widehat{m}}, \widehat{\widehat{\mathbf{r}}})$ of (2.11) satisfies*

$$P\{\widehat{\widehat{m}} = m_0\} \to 1 \tag{2.12}$$

and there exists a constant $B > 0$ such that

$$P\left\{\max_{1\le i\le m_0} |\widehat{\widehat{r}}_i - r_i^0| \le Bm_0\gamma_n\right\} \to 1. \tag{2.13}$$

Remark 2.2. In Theorem 2.4, the penalty ω_n should satisfy $\lim_{n\to\infty} \omega_n/(n\gamma_n m_0) \ge 3 + C_0$. Although the constant C_0 is difficult to be determined in practice, it suffices to choose a ω_n in the information criterion such that $\omega_n/(n\gamma_n m_0) \to \infty$ as $n \to \infty$. In case m_0 is bounded and the noise η_1 has finite moment of all orders, Remark 2.1 indicates that $n\gamma_n m_0$ is nearly $O(1)$. Therefore, the usual consistent model selection criteria such as *BIC* ($\omega_n = \log n$) are valid.

When η_1 does not have finite moment of all orders, the maximum ι such that $E|\eta_1|^{2+\iota} < \infty$ may be obtained from tail index estimators such as Hill's estimator. Then γ_n can be chosen according to Remark 2.1. Hence, ω_n can in turn be chosen to satisfy $\omega_n/(n\gamma_n m_0) \to \infty$ as $n \to \infty$.

When $|\mathcal{A}_n|$ is large, it is possible to achieve further computational efficiency by using the well-known backward elimination algorithm (BEA) in regression analysis. Intuitively, the BEA starts with the set of thresholds $\mathcal{A}_n$, then removes the "most redundant" thresholds that corresponds to the largest reduction of the *IC*. The preceding step is repeated successively until no further removal is possible. Specifically, the BEA goes as follows.

1. Set $K = |\mathcal{A}_n|$, $\mathbf{r}_K := (r_{K,1}, \ldots, r_{K,K}) = \mathcal{A}_n$ and $V_K^* = IC(K, \mathcal{A}_n)$.
2. For $i = 1, \ldots, K$, compute $V_{K,i} = IC(K-1, \mathbf{t}_K \setminus \{t_{K,i}\})$. Set $V_{K-1}^* = \min_i V_{K,i}$.
3. • If $V_{K-1}^* > V_K^*$, then the estimated locations of thresholds are $\mathcal{A}_n^* = \mathbf{r}_K$.
   • If $V_{K-1}^* \le V_K^*$ and $K = 1$, then $\mathcal{A}_n^* = \emptyset$. That is, there is no threshold in the time series.
   • If $V_{K-1}^* \le V_K^*$ and $K > 1$, then set $j = \arg\min_i V_{K,i}$, $\mathbf{r}_{K-1} := \mathbf{r}_K \setminus \{r_{K-1,j}\}$ and $K = K - 1$. Go to step 2.

For the estimate $\mathcal{A}_n^* =: (\hat{r}_i^*, \ldots, \hat{r}_{|\mathcal{A}_n^*|}^*)$ obtained from BEA, we have the same consistency results as Theorem 2.4.

Theorem 2.5. *Under the conditions of Theorem 2.4, as $n \to \infty$,*

$$P\{|\mathcal{A}_n^*| = m_0\} \to 1 \quad \textit{and} \quad P\left\{\max_{1\le i\le m_0} |\hat{r}_i^* - r_i^0| \le Bm_0\gamma_n\right\} \to 1.$$

In summary, the one-step LASSO procedure cannot estimate the true set of thresholds $\mathcal{A}$ consistently. However, we have an estimate $\mathcal{A}_n$, from which there exists a subset $\mathcal{A}_n^*$ that estimates $\mathcal{A}$ consistently. By estimating $\mathcal{A}_n^*$ from a given $\mathcal{A}_n$ by the second-step selecting procedure (2.11), Theorems 2.4 and 2.5 state that we are able to estimate the true number m_0 with probability approaching one, and true thresholds can be consistently estimated with a rate γ_n.

Remark 2.3. To simplify notations and to facilitate the presentation, we assume that the autoregressive order p is known and is the same for all regimes. This assumption can be relaxed so that the order of each regime is an unknown integer less than p^*. In this case each regime can be regarded as a p^*th order autoregressive model with the last few coefficients equaling zero. The results on consistency and the convergent rate remain valid, while the proof only requires slight modification with more cumbersome notation. In practice, first estimate the thresholds by the two-step procedure using a sufficiently large order p^* for each regime, then identify the order of each regime by applying standard procedures such as BIC or C_p on each estimated segment.

Remark 2.4. For simplicity we have assumed that d is known. In general, following Chan (1993) and Li and Ling (2012), the estimation procedure can be repeated for various d to select the best model with respect to an information criterion.

2.3. Computational implementations

The block coordinate decent algorithm and the least angle regression (LARS) algorithm are two commonly used procedures for the implementation of LASSO. The block coordinate decent algorithm gives an exact optimization solution by iteratively solving the normalizing equations until convergence. On the other hand, the least angle regression (LARS) algorithm gives an approximation to the LASSO solution by iteratively incorporating new thresholds. The two algorithms for threshold estimation can be implemented effectively by the detailed algorithms given in Chan et al. (2014) through the design matrix $\boldsymbol{X}_n$ in (2.1).

Due to the specific structure of the design matrix $\boldsymbol{X}_n$, the computation of the LASSO algorithms can be much faster than the ordinary high-dimensional regression problem. In particular, the LARS implementation can be performed in the order of $O(K^3 + Kn)$, where $K = K(n)$ is an upper bound of thresholds. This is more efficient than the order $O(n^2)$ of the classical binary segmentation type algorithms, especially when the sample size is large. We refer to Chan et al. (2014) and Bleakley and Vert (2011) for the computation complexities for the two algorithms.

To choose the tuning parameters, for the block coordinate decent algorithm, one simple way is to select a λ_n that overestimates the number of thresholds. This can be achieved when there are more selected thresholds than expected for a given time

Table 1
Percentage of correct identification of the number of thresholds ($\%m_0$), bias and empirical standard deviation (ESD) of the two-step procedure for model (3.1). The corresponding values of LSE estimation in Li and Ling (2012) are given in the parenthesis. The number of thresholds is assumed known in Li and Ling (2012).

n	$\%m_0$		β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	r_1	r_2
300	78.1	Bias	0.058	−0.029	0.000	0.055	−0.034	0.020	0.018	0.023
			(0.032)	(0.012)	(0.000)	(0.015)	(−0.018)	(0.009)	(−0.018)	(−0.013)
		ESD	0.315	0.0162	0.148	0.116	0.202	0.104	0.089	0.036
			(0.292)	(0.151)	(0.150)	(0.361)	(0.208)	(0.104)	(0.076)	(0.053)
600	99.6	Bias	0.011	0.003	0.005	0.009	−0.007	0.005	0.001	0.013
			(0.007)	(0.002)	(0.004)	(0.012)	(−0.002)	(0.001)	(−0.007)	(−0.007)
		ESD	0.196	0.099	0.095	0.226	0.148	0.076	0.028	0.019
			(0.192)	(0.100)	(0.089)	(0.219)	(0.143)	(0.074)	(0.023)	(0.016)
900	99.5	Bias	0.012	0.006	0.002	0.002	−0.007	0.004	0.001	0.008
			(0.002)	(0.000)	(0.001)	(0.019)	(0.001)	(−0.000)	(−0.003)	(−0.006)
		ESD	0.158	0.082	0.076	0.195	0.116	0.060	0.018	0.013
			(0.156)	(0.084)	(0.076)	(0.192)	(0.116)	(0.061)	(0.016)	(0.010)
1200	99.5	Bias	0.008	0.003	−0.001	0.007	−0.006	0.002	0.000	0.006
			(0.009)	(0.003)	(0.002)	(0.009)	(−0.003)	(0.002)	(−0.003)	(0.004)
		ESD	0.139	0.073	0.065	0.158	0.099	0.051	0.012	0.011
			(0.140)	(0.073)	(0.066)	(0.160)	(0.099)	(0.050)	(0.014)	(0.008)

series, or when clusters of thresholds are observed. Note that the second stage of the proposed procedure guarantees consistent estimation from the over-estimated set of thresholds in the first stage of the LASSO procedure. Therefore, when the penalty is small enough such that all the true thresholds are captured in the first step, the second step procedure will select consistent estimates of the thresholds. Alternatively, the tuning parameters can be chosen by gradually calibrating the penalty terms λ_n until the second step estimations give stabilized results. When the LARS implementation is used, the choice of λ_n is translated to K, the maximum number of thresholds to be estimated. The particular value of K can be similarly chosen as λ_n.

Note that the block coordinate decent algorithm returns the solution only when the iterative optimization reaches convergence. Since there is no guarantee when the iterations converge, the LARS implementation is a more stable computation device. Empirical results show that the LARS algorithm usually provides good approximation to the LASSO solution, see Yuan and Lin (2006), Bleakley and Vert (2011) and Chan et al. (2014). Simulation studies not given here show that both algorithms are accurate and robust with respect to various choices of penalty constants λ_n and K, while the LARS algorithm is faster and gives a very similar solution as the exact block coordinated descent algorithm. Therefore, in our simulation studies and data applications in Sections 3 and 4, the LARS algorithm is used.

3. Simulation results

In this section, we report the simulation results to assess the finite sample behavior of the two-step procedure. Two sets of simulation studies are conducted. In the first set we study the finite sample performance of some simple threshold models. Comparisons to Li and Ling (2012) and Gonzalo and Pitarakis (2002) are made. The second set of simulations investigates the performance of the procedure for long time series with many thresholds, where existing methods may be inadequate. In the simulation experiments, the group-LARS algorithm is employed in the first step and the backward elimination algorithm is used in the second step. In the LARS algorithm, we need to specify the maximum number of thresholds K. In Section 3.1, $K = 20$ is used. In Section 3.2 for a long time series with more thresholds, $K = 40$ is used.

3.1. Comparisons to simulation studies in the literature

In this section we compare the proposed method to other methods with two examples. First, we compare the two-step LASSO procedure with the results of Li and Ling (2012) for the model

$$Y_t = \begin{cases} 1 - 0.4Y_{t-1} + \epsilon_t, & \text{if } y_{t-1} \le -0.8, \\ 0.6 + Y_{t-1} + \epsilon_t, & \text{if } -0.8 < y_{t-1} \le 0.5, \\ -1 - 0.2Y_{t-1} + \epsilon_t, & \text{if } 0.5 \le y_{t-1}, \end{cases} \tag{3.1}$$

where $\epsilon_t \sim$ i.i.d. N(0,1). In the example, 1000 realizations are simulated from model (3.1) and estimated by the two-step procedure. The results are compared to those given in Li and Ling (2012), where the global optimum of the criterion function can be obtained by evaluating all possible combinations of thresholds when the sample size is small. The results in Li and Ling (2012) serve as a benchmark showing the performance of the two-step LASSO procedure against the optimal solution.

Estimation results for the two-step procedure and those reported in Li and Ling (2012) are summarized in Table 1. The percentage (%) of correct estimation for the number of regimes, the bias and standard error of the threshold estimates and the parameters for each segment of AR models are reported. The bias and standard error are computed for the cases where the estimated number of thresholds equals to the true value. Note that the results in Li and Ling (2012) assume that the number of thresholds is known. From the table, the two-step procedure gives the correct number of thresholds in over 99% of the 1000 realizations when the number of observation is greater than 600. Also, the bias and the standard deviation of the parameters estimates are similar, although the bias of the estimates of βs in the two-step procedure tend to be greater for small sample sizes.

Next, we compare the estimations for the number of thresholds using the two-step Lasso procedure and the sequential model selection approach by Gonzalo and Pitarakis (2002). The sequential model selection approach is based on the criterion

$$IC_n(\gamma_1, \ldots, \gamma_m) = \log\left(\mathbf{y}'\mathbf{y} - \sum_{j=1}^{m+1} \mathbf{y}'\mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y}\right) + \frac{\lambda_n}{n} p(m+1), \tag{3.2}$$

where $\mathbf{X}_j$ is the $n \times p$ design matrix formed by the variables $I(\gamma_j < y_{t-d} \le \gamma_{j+1})$ and $y_{t-k}I(\gamma_j < y_{t-d} \le \gamma_{j+1})$, $k = 1, \ldots, p$, $t = 1, \ldots, n$. The sequential procedure begins by testing for the single best γ_1 over all possible values that minimizes $IC_n(\gamma_1)$. Once γ_1 is specified, the time series is divided into two regimes. Then the previous testing procedure is applied to each of the regimes to find possible regimes. The procedure is continued until no more threshold is detected.

Table 2
Correct decision frequencies of two-step procedures under information criterion BIC, BIC2 and BIC3 for the linear model: $y_t = \rho y_{t-1} + \epsilon_t$. The corresponding correct decision frequencies of sequential testing method are given in the parentheses. Replication = 1000.

ρ	$n = 200$			$n = 400$			$n = 600$		
	BIC	BIC2	BIC3	BIC	BIC2	BIC3	BIC	BIC2	BIC3
0.5	88.5	99.0	100.0	87.8	99.0	100.0	87.9	99.0	100.0
	(88.4)	(99.3)	(99.8)	(92.4)	(99.2)	(100.0)	(93.5)	(99.7)	(100.0)
0.7	87.9	99.6	100.0	86.0	98.8	100.0	87.6	99.3	100.0
	(87.8)	(98.8)	(99.8)	(91.1)	(99.5)	(100.0)	(92.1)	(99.7)	(100.0)
0.9	86.6	98.8	99.9	88.7	99.8	100.0	86.3	99.4	100.0
	(85.6)	(98.7)	(99.9)	(90.9)	(99.4)	(100.0)	(91.9)	(99.6)	(100.0)
1.0	47.7	91.1	98.8	52.4	92.1	99.5	56.3	94.0	99.5
	(50.0)	(89.8)	(98.7)	(56.9)	(93.2)	(99.7)	(60.1)	(94.2)	(99.7)

Table 3
Correct decision frequencies of two-step procedures under information criteria BIC, BIC2 and BIC3 for the linear model: $y_t = \rho y_{t-1} I(y_{t-1} \le 0) - \rho y_{t-1} I(y_{t-1} > 0) + \epsilon_t$. The corresponding correct decision frequencies of sequential testing method are given in parentheses. Replication = 1000.

ρ	$n = 200$			$n = 400$			$n = 600$		
	BIC	BIC2	BIC3	BIC	BIC2	BIC3	BIC	BIC2	BIC3
−0.4	69.5	96.6	95.7	72.4	97.3	100.0	75.8	98.5	100.0
	(100.0)	(99.3)	(95.1)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)
−0.25	69.6	65.3	40.2	72.3	94.8	81.5	73.4	98.2	97.6
	(94.2)	(72.4)	(43.4)	(99.9)	(97.0)	(84.2)	(100.0)	(99.6)	(97.1)
−0.15	49.4	21.7	6.50	60.7	44.7	18.7	69.9	66.2	34.1
	(63.3)	(25.2)	(7.4)	(84.5)	(46.6)	(18.2)	(93.5)	(60.5)	(36.3)
−0.10	30.1	8.40	1.70	39.7	17.1	38	51.9	25.3	5.5
	(38.7)	(9.90)	(1.8)	(53.2)	(16.5)	(3.5)	(66.5)	(25.4)	(6.0)
−0.05	15.7	2.30	0.5	19.3	3.1	0.1	24.9	3.6	0.5
	(19.0)	(2.70)	(0.2)	(21.6)	(2.5)	(0.2)	(22.7)	(2.8)	(0.2)

Table 4
Distribution (%) of estimated number of the thresholds for model 3 using two-step procedure under information criteria BIC, BIC2 and BIC3. The corresponding values for the sequential testing procedure are given in parentheses. Replications = 1000.

$\hat{m}$	$n = 400$				$n = 600$				$n = 800$			
	0	1	2	≥3	0	1	2	≥3	0	1	2	≥3
BIC	0.8	86.6	12.1	0.5	0.2	88.4	10.8	0.6	0.0	95.1	4.0	0.9
	(0.8)	(80.5)	(18.7)	(0.0)	(0.1)	(90.0)	(10.0)	(0.0)	(0.1)	(94.4)	(5.4)	(0.0)
BIC2	1.2	89.5	9.0	0.3	0.5	97.1	1.6	0.8	0.0	98.6	0.9	0.5
	(1.3)	(91.1)	(7.6)	(0.0)	(0.2)	(96.3)	(3.5)	(0.0)	(0.1)	(98.4)	(1.5)	(0.0)
BIC3	2.8	90.3	6.7	0.2	0.4	95.1	3.9	0.6	0.5	97.2	1.9	0.4
	(1.3)	(91.5)	(7.1)	(0.0)	(0.4)	(96.4)	(3.3)	(0.0)	(0.2)	(98.5)	(1.3)	(0.0)

Table 5
Distribution (%) of estimated number of the thresholds for model 4 using two-step procedure under information criteria BIC, BIC2 and BIC3. The corresponding values for the sequential testing procedure are given in parentheses. Replications=1000.

$\hat{m}$	$n = 400$				$n = 600$				$n = 800$			
	≤1	2	3	≥4	≤1	2	3	≥4	≤1	2	3	≥4
BIC	0.6	82.1	17.2	0.1	1.1	81.7	16.8	0.4	0.8	86.8	12.4	0.0
	(0.0)	(79.7)	(20.3)	(0.0)	(0.0)	(85.4)	(14.6)	(0.0)	(0.0)	(88.1)	(11.9)	(0.0)
BIC2	0.4	92.0	7.3	0.3	0.9	91.7	7.4	0.0	0.4	92.2	7.4	0.0
	(0.0)	(98.1)	(1.9)	(0.0)	(0.0)	(99.0)	(1.0)	(0.0)	(0.0)	(99.0)	(1.0)	(0.0)
BIC3	0.5	95.2	4.3	0.0	1.1	94.6	4.3	0.0	0.6	94.5	4.9	0.0
	(0.0)	(99.1)	(0.9)	(0.0)	(0.0)	(99.3)	(0.7)	(0.0)	(0.0)	(99.8)	(0.2)	(0.0)

To fairly compare with the sequential model selection approach, first apply the first-step Lasso procedure to get a set of possible thresholds $\mathcal{A}_n$, then the sequential model selection approach is applied to $\mathcal{A}_n$, using criterion (3.2). Tables 2–5 summarize the results for the following four models:

1. $y_t = \rho y_{t-1} + \epsilon_t$.
2. $y_t = \rho y_{t-1} I(y_{t-1} \le 0) - \rho y_{t-1} I(y_{t-1} > 0) + \epsilon_t$.
3. $y_t = \begin{cases} -3 + 0.5y_{t-1} - 0.9y_{t-2} + \epsilon_t, & y_{t-2} \le 1.5, \\ 2 + 0.3y_{t-1} + 0.2\rho y_{t-2} + \epsilon_t, & y_{t-2} > 1.5. \end{cases}$
4. $y_t = \begin{cases} 2.7 + 0.8y_{t-1} - 0.2y_{t-2} + \epsilon_t, & y_{t-2} \le 5, \\ 6 + 1.9y_{t-1} - 1.2\rho y_{t-2} + \epsilon_t, & 5 < y_{t-2} \le 12, \\ 1 + 0.7y_{t-1} - 0.3\rho y_{t-2} + \epsilon_t, & y_{t-2} > 12. \end{cases}$

Following Gonzalo and Pitarakis (2002), $\log n$ (BIC), $2\log n$ (BIC2) and $3\log n$ (BIC3) are taken as the model complexity penalty term λ_n in the simulations. From these tables, we see that the LASSO procedure is comparable to the sequential model selection approach. This implies that the sequential approach of threshold estimation over all possible values of threshold is approximately equivalent to that over the smaller subset $\mathcal{A}_n$ obtained from the LASSO procedure. In view of its computational efficiency, the LASSO procedure serves as a useful alternative for multiple-threshold estimation.

It can be seen that BIC2 and BIC3 work better than BIC from Table 2. On the other hand, BIC outperforms BIC2 and BIC3 for the model in Table 3, especially for small values of ρ. One possi-

Table 6
Threshold estimations from the two-step estimation procedure for the threshold model (3.3). The model parameter C_i, average number of observations (# Obs) in each true regime, mean and standard error (SE, in (10^{-3})) of the threshold estimates are given in the columns. The average computing time for one estimation and the percentage of correct estimation of the number of thresholds are denoted by "Time" and "% $\hat{m} = 8$". Replications=1000.

	Scenario 1				Scenario 2				Scenario 3			
T	10,000				30,000				50,000			
Time	4 s				7 s				16 s			
%, $\hat{m} = 8$	91.7				93.9				99.8			
	C_i	*#Obs*	Mean	SE (10^{-3})	C_i	*#Obs*	Mean	SE (10^{-3})	C_i	*#Obs*	Mean	SE (10^{-3})
$r_1 = -3.5$	−4.5	1079	−3.50	1.6	2.0	2814	−3.501	29.1	−0.6	543	−3.524	60.5
$r_2 = -2.5$	2.5	1019	−2.499	1.2	3.0	1289	−2.499	0.99	1.6	519	−2.497	10.7
$r_3 = -1.5$	−2.0	1068	−1.504	9.9	4.0	1362	−1.499	0.79	−0.6	2840	−1.486	23.7
$r_4 = -0.5$	2.3	1019	−0.499	12.4	9.0	1351	−0.499	0.78	1.6	8080	−0.500	0.20
$r_5 = 0.5$	1.0	1145	0.4983	6.3	8.0	1306	0.501	1.16	−0.6	13,264	0.500	0.19
$r_6 = 1.5$	3.0	1245	1.5037	27.6	11.0	1337	1.457	58.8	1.6	13,699	1.500	11.54
$r_7 = 2.5$	1.6	1153	2.5010	1.0	9.0	1599	2.502	3.41	−0.6	8012	2.501	2.26
$r_8 = 3.5$	−0.5	961	3.5015	7.6	12.0	2103	3.501	0.63	1.6	2489	3.500	2.82
$r_9 = \infty$	1.5	1313	–	–	9.0	16,840	–	–	−0.6	555	–	–

ble explanation for this discrepancy is that the magnitudes of the penalty of the criteria are given in the order BIC < BIC2 < BIC3 ($\log n < 2\log n < 3\log n$). For Table 2, the true model is an AR model without threshold. Thus, BIC2 and BIC3, with greater penalties for model complexity, tend to select the true model with no threshold. On the other hand, the true model in Table 3 has one threshold. When ρ is small, it is more likely that BIC2 and BIC3 over-penalize and choose the model without threshold. Therefore, BIC performs better in these cases. Nevertheless, BIC2 and BIC3 seem to perform better in general.

3.2. Long time series with multiple regimes

In this section we demonstrate the performance of the two-step procedures for several long time series with a large number of thresholds. With the rapid growth of high frequency data, time series with length over 10,000 are often encountered. As the time series becomes longer, more information is available and a model with more regimes may be used to explain the structures of the data generating mechanism. However, the computation burden will be heavy using existing methods as the combination of possible threshold values grows exponentially with the sample size. The two-step procedure developed in this paper, which inherits the computational efficiency from the LASSO, is well-suited for such situations. Consider three scenarios of long time series with 8 thresholds, with series lengths ranging from 10,000 to 50,000. Different patterns of the regimes are studied. In particular, three scenarios are considered: (1) the number of observations in each regime are uniformly located, (2) more than half of the observations are located in one regime and (3) around 1% of observations are located in the regimes with extremal values. Since the main focus is about estimating threshold values, for illustration, we consider the following threshold autoregressive model.

$$Y_t = \begin{cases} C_1 - 0.6Y_{t-1} + \epsilon_t, & \text{if } y_{t-1} \le -3.5, \\ C_2 + 0.3Y_{t-1} + 0.9Y_{t-2} + \epsilon_t, & \text{if } -3.5 \le y_{t-1} \le -2.5, \\ C_3 - 0.9Y_{t-1} + \epsilon_t, & \text{if } -2.5 \le y_{t-1} \le -1.5, \\ C_4 + 0.7Y_{t-1} + 0.5Y_{t-2} + \epsilon_t, & \text{if } -1.5 \le y_{t-1} \le -0.5, \\ C_5 + 0.1Y_{t-1} + \epsilon_t, & \text{if } -0.5 \le y_{t-1} \le 0.5, \\ C_6 - 0.9Y_{t-1} + \epsilon_t, & \text{if } 0.5 \le y_{t-1} \le 1.5, \\ C_7 + 0.9Y_{t-1} + \epsilon_t, & \text{if } 1.5 \le y_{t-1} \le 2.5, \\ C_8 - 0.8Y_{t-1} - 0.2Y_{t-2} + \epsilon_t, & \text{if } 2.5 \le y_{t-1} \le 3.5, \\ C_9 - 1.1Y_{t-1} + \epsilon_t, & \text{if } 3.5 \le y_{t-1}, \end{cases} \tag{3.3}$$

where $\epsilon_t \sim$ i.i.d. N(0,1) and the constants C_is determine the percentage of observations in each regime. In this example, 1000 realizations are simulated from model (3.3) with values of $\mathbf{t} = (t_1, t_2, \ldots, t_8)$ estimated by the two-step procedure. The estimation results are reported in Table 6. The percentage (%) of the estimated number of segments, the mean and standard error of the threshold estimates are reported. Despite the length of the time series, the computation of a two-step estimation procedure can be completed within 20 s. All computations are performed using the program **R** on a laptop with an Intel Core i5 480M processor. The estimation accuracy is also extremely high. On the other hand, the implementation of the sequential approach of Gonzalo and Pitarakis (2002) requires a computation order of n^2, which becomes formidable for long time series.

4. Applications to real data

In this section we illustrate the two-step procedure on fitting a multiple-regime TAR model for the growth rate of the quarterly US real GNP data over the period 1947–2012. Given the quarterly GNP data $\{y_t\}_{t=1,\ldots,261}$ from 1947 to the first quarter of 2012, the growth rate is defined as

$$x_t = 100(\log y_t - \log y_{t-1}) \quad t = 2, \ldots, 261.$$

The GNP $\{y_t\}$ and the growth rate $\{x_t\}$ are displayed in Fig. 1. This data set has been investigated previously by Li and Ling (2012) and they argued that a three-regime model encompassing bad, good and normal times fits the data reasonable well. The two threshold are 1.20 and 2.43 respectively.

The two-step procedure is implemented on the growth rate series using the criterion

$$\text{AIC} = \sum_{j=1}^{m+1} n_j \log \hat{\sigma}_j^2 + 2(p+1)(m+1),$$

which was used in Li and Ling (2012). In the first step, the LARS algorithm is used with $K = 10$ with AR order $p^* = 12$. In the second step, the backward elimination algorithm is used for model selection. The two-step procedure identified three thresholds at 1.23, 1.83 and 2.55 respectively. The AIC of the selected model from two-step procedure model is 1050.7 and the AIC of Li and Ling's model is 1148.4. The four-regime model may give a better fit to the data. Note that the thresholds 1.23 and 2.55 are in line with the thresholds 1.20 and 2.43 found in Li and Ling (2012). The threshold 1.83 suggests that the normal time regime may be further divided into two sub-regimes.

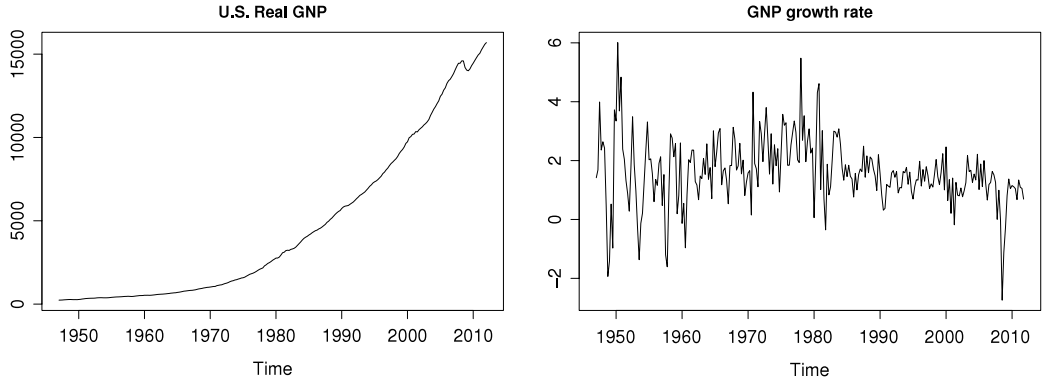

Fig. 1. Left: US GNP data from 1947 to 2012. Right: Growth rate of the US GNP data.

Acknowledgments

We would like to thank the editor and two anonymous referees for helpful comments, which led to an improved version of this paper. This research was supported in part by grants from the General Research Fund of HKSAR-RGC-GRF Nos. 400313 and 400410 (Chan), NSFC Nos. 11171074, 11371318 and the Fundamental Research Funds for the Central Universities (Zhang), HKSAR-RGC-ECS 405012, HKSAR-RGC-GRF 405113 (Yau). Part of this research was conducted when NH Chan was the Visiting Chang-Jiang Professor at Renmin University of China (RUC). Research support from RUC is gratefully acknowledged.

Appendix A. Proofs

Proofs of the main results are presented in this section. The proofs of the lemmas are given in the supplementary materials (see Appendix B). In the following proofs, the constant C denotes a generic constant which may be different from line to line.

Lemma A.1. *Under Assumptions* H1 *and* H2, *there exist positive constants* c_0, C *and* C' *such that when n is large, for any* $y \geq 6 + 1/c_0$,

$$P\left(\max_{x\in R}\max_{1\leq j\leq p}\frac{1}{\sqrt{n}}\left|\sum_{t=1}^{n}Y_t^j I(Y_{t-d}\leq x)\varepsilon_t\right|\geq y\right)$$
$$\leq C\exp\left\{-\frac{c_0 y^2}{2(2+y)}\right\}+C'n^{-\iota/2}(\log n)^{4+2\iota},$$

where $\varepsilon_t = \sigma_j\eta_t I(r_{j-1} < Y_{t-d} \leq r_j)$, $t = 1,\ldots,n$ *and* $Y_t^0 = 1$, $Y_t^j = Y_{t-j}$ *for* $j = 1,\ldots,p$.

Proof of Theorem 2.1. By the definition of $\widehat{\boldsymbol{\theta}}(n)$, we have

$$\frac{1}{n}\|\boldsymbol{Y}_n^0 - \boldsymbol{X}_n\widehat{\boldsymbol{\theta}}(n)\|_2^2 + \lambda_n\sum_{i=1}^{n}\|\widehat{\boldsymbol{\theta}}_i\|$$
$$\leq \frac{1}{n}\|\boldsymbol{Y}_n^0 - \boldsymbol{X}_n\boldsymbol{\theta}^0(n)\|_2^2 + \lambda_n\sum_{i=1}^{n}\|\boldsymbol{\theta}_i^0\|. \tag{A.1}$$

Let $\mathcal{A} = \{i : \boldsymbol{\theta}_i^0 \neq 0\}$. Then, applying $\boldsymbol{Y}_n^0 = \boldsymbol{X}_n\boldsymbol{\theta}(n) + \boldsymbol{\varepsilon}(n)$ to (A.1) gives

$$\frac{1}{n}\|\boldsymbol{X}_n(\boldsymbol{\theta}^0(n) - \widehat{\boldsymbol{\theta}}(n))\|_2^2$$
$$\leq \frac{2}{n}(\widehat{\boldsymbol{\theta}}(n) - \boldsymbol{\theta}^0(n))^{\mathrm{T}}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\eta}(n) + \lambda_n\sum_{i=1}^{n}\|\boldsymbol{\theta}_i^0\| - \lambda_n\sum_{i=1}^{n}\|\widehat{\boldsymbol{\theta}}_i\|$$
$$= 2\sum_{j=1}^{n}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^0)^{\mathrm{T}}\Big(\frac{1}{n}\sum_{i=j}^{n}\boldsymbol{Y}_{\pi(i)}\varepsilon_{\pi(i)+d}\Big)$$
$$+\lambda_n\sum_{i\in\mathcal{A}}(\|\boldsymbol{\theta}_i^0\| - \|\widehat{\boldsymbol{\theta}}_i\|) - \lambda_n\sum_{i\in\mathcal{A}^c}\|\widehat{\boldsymbol{\theta}}_i\|$$
$$\leq 2p\Big(\sum_{j=1}^{n}\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^0\|\Big)\Big(\max_{1\leq l\leq p}\Big|\frac{1}{n}\sum_{i=1}^{n}Y_{i-1}^l\varepsilon_i I(Y_{i-d}\geq Y_{\pi(j)})\Big|\Big)$$
$$+\lambda_n\sum_{i\in\mathcal{A}}(\|\boldsymbol{\theta}_i^0\| - \|\widehat{\boldsymbol{\theta}}_i\|) - \lambda_n\sum_{i\in\mathcal{A}^c}\|\widehat{\boldsymbol{\theta}}_i\|.$$

Thus, by Lemma A.1, when $a_n = \sqrt{n}\lambda_n/(4p) \to \infty$, we have with probability greater than $1 - C[\exp(-a_n/15) + n^{-\iota/2}(\log n)^{4+2\iota}]$ such that

$$\frac{1}{n}\|\boldsymbol{X}_n(\boldsymbol{\theta}^0(n) - \widehat{\boldsymbol{\theta}}(n))\|_2^2$$
$$\leq \lambda_n\sum_{j=1}^{n}\|\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^0\| + \lambda_n\sum_{i\in\mathcal{A}}(\|\boldsymbol{\theta}_i^0\| - \|\widehat{\boldsymbol{\theta}}_i\|) - \lambda_n\sum_{i\in\mathcal{A}^c}\|\widehat{\boldsymbol{\theta}}_i\|$$
$$\leq \lambda_n\sum_{i\in\mathcal{A}}\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^0\| + \lambda_n\sum_{i\in\mathcal{A}}(\|\boldsymbol{\theta}_i^0\| - \|\widehat{\boldsymbol{\theta}}_i\|)$$
$$\leq 2\lambda_n\sum_{i\in\mathcal{A}}\|\boldsymbol{\theta}_i^0\| \leq 2\lambda_n m_n\max_i\|\boldsymbol{\phi}_i^0 - \boldsymbol{\phi}_{i-1}^0\|, \tag{A.2}$$

which implies (2.6). □

Let Y_i^j, η_i be defined as in Lemma A.1 and $I(a \leq x \leq b) = -I(b \leq x \leq a)$ if $b < a$. To show Theorem 2.2, we need to establish a uniformly continuous theorem for the process $g(x) = \sum_{i=1}^{n}Y_i^j\varepsilon_i I(Y_{i-d}\leq x)$.

Lemma A.2. *Let* $\sigma^2(t) = \mathrm{E}\Big(\Big(Y_i^j I(r_{l-1}^0 \leq Y_{i-d} \leq r_{l-1}^0 + t)\Big)^2\Big)$. *Under Assumptions* H1, H2, *there exists a constant* C *such that for any* $a_n, y > 0$ *and any* $1 \leq l \leq m_0 + 1$, $0 \leq j \leq p$,

$$P\left\{\max_{|t|\geq a_n}\Big[n\sigma^2(t)\Big]^{-1}\left|\sum_{i=1}^{n}Y_i^j I(r_{l-1}^0 \leq Y_{i-d} \leq r_{l-1}^0 + t)\varepsilon_i\right| \geq 2y\right\}$$
$$\leq C\left(n^{-\iota/2}(\log n)^{4+2\iota} + y^{-2}(na_n)^{-1}\log(1/a_n)\right). \tag{A.3}$$

The following lemma concerns the Karush–Kuhn–Tucker (KKT) condition of the group LASSO estimator given in (2.3). The proof can be directly deduced by applying the KKT condition to $\widehat{\boldsymbol{\theta}}(n)$ (the solution of (2.3)), and the details are omitted.

Lemma A.3. *Let* $\widehat{\boldsymbol{\theta}}(n)$ *and* $\boldsymbol{\theta}(n)$ *be defined as in* (2.3). *Under the conditions of Theorem* 1.1, *we have*

$$\sum_{l=\hat{t}_j}^{n}\boldsymbol{Y}_{\pi(l)}(Y_{\pi(l)+d} - \sum_{i=1}^{l}\widehat{\boldsymbol{\theta}}_{\pi(i)}^T\boldsymbol{Y}_{\pi(l)})$$
$$+\frac{1}{2}n\lambda_n\widehat{\boldsymbol{\theta}}_{\pi(\hat{t}_j)}/\|\widehat{\boldsymbol{\theta}}_{\pi(\hat{t}_j)}\| = 0, \quad \text{if } \widehat{\boldsymbol{\theta}}_{\pi(\hat{t}_j)} \neq 0 \tag{A.4}$$

and

$$\left\| \sum_{l=j}^{n} \boldsymbol{Y}_{\pi(l)} (Y_{\pi(l)+d} - \sum_{i=1}^{l} \widehat{\boldsymbol{\theta}}_{\pi(i)}^{T} \boldsymbol{Y}_{\pi(l)}) \right\| \le n\lambda_n/2, \quad \textit{for all } j. \tag{A.5}$$

Furthermore, $\sum_{i=1}^{t} \widehat{\boldsymbol{\theta}}_{\pi(i)} = \widehat{\boldsymbol{\phi}}_j$ *for* $\hat{t}_j < t \le \hat{t}_j$, $j = 1, 2, \ldots, |\mathcal{A}_n|$.

Proof of Theorem 2.2. Let $A_{ni} = \{|\hat{r}_i - r_i^0| > \gamma_n\}$, $i = 1, 2, \ldots, m_0$, then

$$P\left\{ \max_{1\le i\le m_0} |\hat{r}_i - r_i^0| > \gamma_n \right\} \le \sum_{i=1}^{m_0} P\{|\hat{r}_i - r_i^0| > \gamma_n\} = \sum_{i=1}^{m_0} P(A_{ni}).$$

Put $C_n = \{\max_{1\le i\le m_0} |\hat{r}_i - r_i^0| \le \min_i |r_i^0 - r_{i-1}^0|\}$. To prove Theorem 2.2, it suffices to show that

$$\sum_{i=1}^{m_0} P(A_{ni}C_n) \quad \text{and} \quad \sum_{i=1}^{m_0} P(A_{ni}C_n^c) \to 0,$$

where C_n^c denotes the complement of the set C_n. We only give the outline of the proof for $\sum_{i=1}^{m_0} P(A_{ni}C_n) \to 0$.

First consider the case when m_0 is fixed. From the definition, we see that in C_n,

$$r_{i-1}^0 < \hat{r}_i < r_{i+1}^0, \quad \text{for all } 1 \le i \le m_0.$$

Next, we split A_{ni} into two cases ((i) $\hat{r}_i < r_i^0$ and (ii) $\hat{r}_i > r_i^0$) to show $P(A_{ni}C_n) \to 0$.

When $\hat{r}_i < r_i^0$, note that for any given r_i^0, $i = 1, \ldots, m_0$, there exists a t_i^0 such that $Y_{\pi(t_i^0-1)} \le r_i^0 < Y_{\pi(t_i^0)}$ and $\sum_{i=1}^{n} Y_i^j I(Y_{i-d} \le r_i^0) = \sum_{i=1}^{n} Y_i^j I(Y_{i-d} \le Y_{\pi(t_i^0)})$. Applying Lemma A.3 with r_i^0 and $\hat{r}_i$, i.e., t_i^0 and $\hat{t}_i$, we get

$$\left\| \sum_{l=\hat{t}_i}^{t_i^0-1} \boldsymbol{Y}_{\pi(l)} (Y_{\pi(l)+d} - \widehat{\boldsymbol{\phi}}_{i+1}^{\mathrm{T}} \boldsymbol{Y}_{\pi(l)}) \right\| \le n\lambda_n.$$

Let $\boldsymbol{Y}_t = (1, Y_{t-1}, \ldots, Y_{t-d})^T$, then

$$\left\| \sum_{l=1}^{n} \boldsymbol{Y}_l (Y_l - \widehat{\boldsymbol{\phi}}_{i+1}^{\mathrm{T}} \boldsymbol{Y}_l) I(\hat{r}_i < Y_{l-d} \le r_i^0) \right\| \le n\lambda_n.$$

This implies that

$$\begin{aligned} n\lambda_n \ge & \left\| \sum_{l=1}^{n} \boldsymbol{Y}_l \varepsilon_l I(\hat{r}_i < Y_{l-d} \le r_i^0) \right. \\ & + \sum_{l=1}^{n} \boldsymbol{Y}_l (\boldsymbol{\phi}_i^{0\mathrm{T}} - \boldsymbol{\phi}_{i+1}^{0\mathrm{T}}) \boldsymbol{Y}_l I(\hat{r}_i < Y_{l-d} \le r_i^0) \\ & \left. + \sum_{l=1}^{n} \boldsymbol{Y}_l (\boldsymbol{\phi}_{i+1}^{0\mathrm{T}} - \widehat{\boldsymbol{\phi}}_{i+1}^{\mathrm{T}}) \boldsymbol{Y}_l I(\hat{r}_i < Y_{l-d} \le r_i^0) \right\| \\ =: & \|\Sigma_{n,1} + \Sigma_{n,2} + \Sigma_{n,3}\|. \end{aligned}$$

It follows that on $\hat{r}_i < r_i^0$,

$$\begin{aligned} P(A_{ni}C_n) \le & P\left(\left\{ \frac{1}{3}\|\Sigma_{n,2}\| \le n\lambda_n \right\} \cap \{|\hat{r}_i - r_i^0| > \gamma_n\} \right) \\ & + P\left(\left\{ \|\Sigma_{n,1}\| > \frac{1}{3}\|\Sigma_{n,2}\| \right\} \cap \{|\hat{r}_i - r_i^0| > \gamma_n\} \right) \\ & + P\left(\left\{ \|\Sigma_{n,3}\| > \frac{1}{3}\|\Sigma_{n,2}\| \right\} \cap A_{ni}C_n \right) \\ = & P(A_{ni1}) + P(A_{ni2}) + P(A_{ni3}). \end{aligned}$$

Since $\mathrm{E}|Y_t^j \varepsilon_t|^{2+\iota} < \infty$ for some $\iota > 2$, applying the same argument in (S.18) of the supplementary materials to $\{\boldsymbol{Y}_t \boldsymbol{Y}_t^T\}$, it is easy to see that when $(m_0/n)\gamma_n^{-1} \log \gamma_n^{-1} \to 0$,

$$\left\| \sup_{|t|\ge\gamma_n} \frac{\sum_{l=1}^{n} \left(\boldsymbol{Y}_l \boldsymbol{Y}_l^T I(r_i^0 - t < Y_{l-d} \le r_i^0) - \mathrm{E}\boldsymbol{Y}_l \boldsymbol{Y}_l^T I(r_i^0 - t < Y_{l-d} \le r_i^0) \right)}{\sum_{l=1}^{n} \mathrm{E}\boldsymbol{Y}_l \boldsymbol{Y}_l^T I(r_i^0 - t < Y_{l-d} \le r_i^0)} \right\| = o(1). \tag{A.6}$$

This implies

$$\begin{aligned} \frac{1}{3}\|\Sigma_{n2}\| \ge & \frac{n}{3} \|\mathrm{E}[(\boldsymbol{Y}_{t_{i-1}^0} \boldsymbol{Y}_{t_{i-1}^0}^{\mathrm{T}}) I(r_i^0 - \gamma_n < Y_{i-d} \le r_i^0)] \\ & \times (\boldsymbol{\phi}_i^0 - \boldsymbol{\phi}_{i+1}^0)\| =: c_0 n\gamma_n > 0. \end{aligned}$$

Thus, by $\gamma_n/\lambda_n \to \infty$, we have $P(A_{ni1}) \to 0$.

By Lemma A.2 and assumption $(m_0/n)\gamma_n^{-1} \log \gamma_n^{-1} \to 0$ again, we have

$$P\left(\|\Sigma_{n1}\| \ge \frac{1}{3} c_0 n\gamma_n \right) \to 0, \quad \text{in probability.} \tag{A.7}$$

Combining this with (A.6) yields $P(A_{ni2}) \to 0$.

Next, we show $P(A_{ni3}) \to 0$. Note that for $(r_i^0 + r_{i+1}^0)/2$, we can find a t_i^* : $t_i^0 < t_i^* < t_{i+1}^0$ such that $Y_{\pi(t_i^*-1)} \le (r_i^0 + r_{i+1}^0)/2 < Y_{\pi(t_i^*)}$, applying Lemma A.3 to t_i^0 and t_i^*, we have

$$\left\| \sum_{l=t_i^0}^{t_i^*-1} \boldsymbol{Y}_{\pi(l)} (\boldsymbol{\phi}_{i+1}^{0\mathrm{T}} - \widehat{\boldsymbol{\phi}}_{i+1}^{\mathrm{T}}) \boldsymbol{Y}_{\pi(l)} \right\| \le n\lambda_n + \left\| \sum_{l=t_i^0}^{t_i^*-1} \boldsymbol{Y}_{\pi(l)} \varepsilon_{\pi(l)} \right\|. \tag{A.8}$$

Since $|r_{i+1}^0 - r_i^0| \ge 4\gamma_n$, using (A.7), it follows that for any $x > 0$,

$$\begin{aligned} \left\| \sum_{l=t_i^0}^{t_i^*-1} \boldsymbol{Y}_{\pi(l)} \varepsilon_{\pi(l)} \right\| &= \left\| \sum_{l=1}^{n} \boldsymbol{Y}_l \varepsilon_l I(r_i^0 < Y_{l-d} \le (r_{i+1}^0 + r_i^0)/2) \right\| \\ &\le xn(r_{i+1}^0 - r_i^0). \end{aligned} \tag{A.9}$$

On the other hand, the same argument as in (A.6) shows in probability that

$$\begin{aligned} & \left\| \sum_{l=t_i^0}^{t_i^*-1} \boldsymbol{Y}_{\pi(l)} (\boldsymbol{\phi}_{i+1}^{0\mathrm{T}} - \widehat{\boldsymbol{\phi}}_{i+1}^{\mathrm{T}}) \boldsymbol{Y}_{\pi(l)} \right\| \\ & \ge \|n[\mathrm{E}\boldsymbol{Y}_l \boldsymbol{Y}_l^T I(r_i^0 < Y_{l-d} \le (r_i^0 + r_{i+1}^0)/2)](\boldsymbol{\phi}_{i+1}^0 - \widehat{\boldsymbol{\phi}}_{i+1})\|. \end{aligned}$$

Combining this with (A.8) and (A.9) gives in probability that for any $x > 0$,

$$\begin{aligned} & \|[\mathrm{E}\boldsymbol{Y}_l \boldsymbol{Y}_l^T I(r_i^0 < Y_{l-d} \le (r_i^0 + r_{i+1}^0)/2)](\boldsymbol{\phi}_{i+1}^0 - \widehat{\boldsymbol{\phi}}_{i+1})\| \\ & \le \lambda_n + x(r_{i+1}^0 - r_i^0). \end{aligned} \tag{A.10}$$

As a result, using (A.6), we have in probability that

$$\begin{aligned} & \left\| \sum_{l=\hat{t}_i}^{t_i^0-1} \boldsymbol{Y}_{\pi(l)} (\boldsymbol{\phi}_{i+1}^{0\mathrm{T}} - \widehat{\boldsymbol{\phi}}_i^{\mathrm{T}}) \boldsymbol{Y}_{\pi(l)} \right\| \\ & \le C[n\lambda_n(r_i^0 - \hat{r}_i)/(r_{i+1}^0 - r_i^0) + xn(r_i^0 - \hat{r}_i)]. \end{aligned} \tag{A.11}$$

However, (A.6) implies in probability that

$$\frac{1}{3} \left\| \sum_{l=\hat{t}_i}^{t_i^0-1} \boldsymbol{Y}_{\pi(l)} (\boldsymbol{\phi}_i^{0\mathrm{T}} - \boldsymbol{\phi}_{i+1}^{0\mathrm{T}}) \boldsymbol{Y}_{\pi(l)} \right\| \ge c_0' n(r_i^0 - \hat{r}_i). \tag{A.12}$$

Since $\lambda_n/(r^0_{i+1}-r^0_i)\to 0$, (A.11) and (A.12) show that $P(A_{ni3})\to 0$. Thus, $P(A_{ni}C_n\cap\{\hat r_i < r^0_i\})\to 0$. Similarly, we can show that $P(A_{ni}C_n\cap\{\hat r_i > r^0_i\})\to 0$. Hence, $P(A_{ni}C_n)\to 0$.

For the case $m_0\to\infty$, it can be seen from Lemma A.2 that, under suitable choices of C_1 and C_2, the rate of convergence of the $P(A_{ni})$s can be fast enough such that $m_0P(A_{ni})\to 0$ for all $i=1,\ldots,m_0$. □

Proof of Theorem 2.3. First, we show $|\mathcal{A}_n|\ge m_0$ by contradiction. Let $I_{\min}=\min_{1\le i\le m_0+1}|r_i-r_{i-1}|$. Suppose that $|\mathcal{A}_n|<m_0$, then there exist some $r^0_{i_0}$ and $\hat r_{l_0}\in\mathcal{A}_n\cup\{0,\infty\}$ with $r^0_{i_0}-r^0_{i_0-1}\vee\hat r_{l_0}\ge I_{\min}/3$ and $r^0_{i_0+1}\wedge\hat r_{l_0+1}-r^0_{i_0}\ge I_{\min}/3$, where $\hat r_0=0$ and $\hat r_{|\mathcal{A}_n|+1}=\infty$. Similar to (A.8), applying Lemma A.3 to $(r^0_{i_0-1}\vee\hat r_{l_0},r^0_{i_0}]$ and $(r^0_{i_0},r^0_{i_0+1}\wedge\hat r_{l_0+1}]$, we get that

$$\left\|\sum_{l=1}^n \boldsymbol{Y}_l(\boldsymbol{\phi}^{0\mathrm{T}}_{i_0}-\widehat{\boldsymbol{\phi}}^{\mathrm{T}}_{l_0+1})\boldsymbol{Y}_lI(r^0_{i_0-1}\vee\hat r_{l_0}<Y_{l-d}\le r^0_{i_0})\right\|$$
$$\le n\lambda_n+\left\|\sum_{l=1}^n\boldsymbol{Y}_l\varepsilon_lI(r^0_{i_0-1}\vee\hat r_{l_0}<Y_{l-d}\le r^0_{i_0})\right\|$$

and

$$\left\|\sum_{l=1}^n \boldsymbol{Y}_l(\boldsymbol{\phi}^{0\mathrm{T}}_{i_0+1}-\widehat{\boldsymbol{\phi}}^{\mathrm{T}}_{l_0+1})\boldsymbol{Y}_lI(r^0_{i_0}<Y_{l-d}\le r^0_{i_0+1}\wedge\hat r_{l_0+1})\right\|$$
$$\le n\lambda_n+\left\|\sum_{l=1}^n\boldsymbol{Y}_l\varepsilon_lI(r^0_{i_0}<Y_{l-d}\le r^0_{i_0+1}\wedge\hat r_{l_0+1})\right\|.$$

Thus, a similar argument to (A.10) shows that for any $x>0$, in probability,

$$\|[\mathrm{E}(\boldsymbol{Y}_l\boldsymbol{Y}_l^TI(r^0_{i_0-1}\vee\hat r_{l_0}<Y_{l-d}\le r^0_{i_0}))](\boldsymbol{\phi}^{0\mathrm{T}}_{i_0}-\widehat{\boldsymbol{\phi}}^{\mathrm{T}}_{l_0+1})\|$$
$$\le\lambda_n+x(r^0_{i_0}-r^0_{i_0-1}\vee\hat r_{l_0})$$

and

$$\|[\mathrm{E}(\boldsymbol{Y}_l\boldsymbol{Y}_l^TI(r^0_{i_0}<Y_{l-d}\le r^0_{i_0+1}\wedge\hat r_{l_0+1}))](\boldsymbol{\phi}^{0\mathrm{T}}_{i_0+1}-\widehat{\boldsymbol{\phi}}^{\mathrm{T}}_{l_0+1})\|$$
$$\le\lambda_n+x(r^0_{i_0+1}\wedge\hat r_{l_0+1}-r^0_{i_0}).$$

Since $r^0_{i_0}-r^0_{i_0}\vee\hat r_{l_0}\ge I_{\min}/3$ and $r^0_{i_0+1}\wedge\hat r_{l_0+1}-r^0_{i_0}\ge I_{\min}/3$ and $I_{\min}/\lambda_n\to\infty$, by the arbitrariness of x, it follows that $\|\boldsymbol{\phi}^0_{i_0}-\widehat{\boldsymbol{\phi}}_{l_0+1}\|\xrightarrow{p}0$ and $\|\boldsymbol{\phi}^0_{i_0+1}-\widehat{\boldsymbol{\phi}}_{l_0+1}\|\xrightarrow{p}0$. This contradicts with the fact that $\min_{1\le i\le m_0+1}\|\boldsymbol{\phi}^0_i-\boldsymbol{\phi}^0_{i-1}\|>\nu$, for some $\nu>0$. Thus, $|\mathcal{A}_n|\ge m_0$ in probability. This proves (2.8).

Next, we show (2.9). Let $\hat R_{n,k}=\{\hat r_1,\hat r_2,\ldots,\hat r_k\}$. Then, we only need to show

$$P\{d_H(\mathcal{A}_n,\mathcal{A})>\gamma_n,m_0\le|\mathcal{A}_n|\le n\}$$
$$=\sum_{k=m_0}^n P(\{d_H(\hat R_{n,k},\mathcal{A})>\gamma_n\})P(|\mathcal{A}_n|=k)\to 0,$$

as $n\to\infty$. By Theorem 2.2, we have $P(d_H(\hat R_{n,m_0},\mathcal{A})>\gamma_n)\to 0$. It suffices to show

$$\max_{k>m_0}P(d_H(\hat R_{n,k},\mathcal{A})>\gamma_n)\to 0. \tag{A.13}$$

Given r^0_i, define $\tau_i=\{j:\hat r_j\in\hat R_{n,k}\text{ and }|\hat r_j-r^0_i|=\min_{r\in\hat R_{n,k}}|r-r^0_i|\}$ and

$$E_{n,k,i,1}=\{\forall 1\le l\le k,|\hat r_l-r^0_i|\ge\gamma_n\text{ and }\hat r_l<r^0_i\},$$
$$E_{n,k,i,2}=\{\forall 1\le l\le k,|\hat r_l-t^0_i|\ge\gamma_n\text{ and }\hat r_l>r^0_i\},$$
$$E_{n,k,i,3}=\{|\hat r_{\tau_i}-r^0_i|\ge\gamma_n,|\hat r_{\tau_i+1}-r^0_i|\ge\gamma_n\text{ and }\hat r_{\tau_i}<r^0_i<\hat r_{\tau_i+1}\},$$
$$E_{n,k,i,4}=\{|\hat r_{\tau_i}-r^0_i|\ge\gamma_n,|\hat r_{\tau_i-1}-r^0_i|\ge\gamma_n\text{ and }\hat r_{\tau_i-1}<r^0_i<\hat r_{\tau_i}\}.$$

Then,

$$\max_{k>m_0}P(d_H(\hat R_{n,k},\mathcal{A})>n\gamma_n)$$
$$=\max_{k>m_0}P(\cup_{i=1}^{m_0}\cup_{j=1}^4\{E_{n,k,i,1}\cup E_{n,k,i,2}\cup E_{n,k,i,3}\}).$$

Let t^0_i and $\hat t_i$ be time points such that $Y_{\pi(t^0_i-1)}\le r^0_i<Y_{\pi(t^0_i)}$ and $Y_{\pi(\hat t_i-1)}\le\hat r_i<Y_{\pi(\hat t_i)}$. Applying Lemma A.3 to t^0_i and $\hat t_i$, then by Lemmas A.2 and the same arguments used in Theorem 2.2, and Proposition 6 of Harchaoui and Levy-Leduc (2010), it can be shown that $\max_{k>m_0}P(\cup_{i=1}^{m_0}E_{n,k,i,j})\to 0$ for $1\le j\le 4$. This gives (2.9) and completes the proof of Theorem 2.3. □

Next we prove Theorems 2.4 and 2.5. The idea of proving the consistency is as follows: when the estimated number of thresholds m is less than m_0, the effect of the goodness-of-fit term S_n dominates, which results in $P(\widehat{\widehat m}<m_0)\to 0$. On the other hand, when $m>m_0$, the effect of the penalty term ω_n dominates and leads to $P(\widehat{\widehat m}>m_0)\to 0$. Therefore, $P(\widehat{\widehat m}=m_0)\to 1$. The following technical lemma will be used to prove Theorems 2.4 and 2.5.

Lemma A.4. *Under the conditions of Theorem 2.4, we have for $m<m_0$, there exists a constant $\nu>0$ such that*

$$\lim_{n\to\infty}P\left\{S_n(\tilde r_1,\tilde r_2,\ldots,\tilde r_m)>\sum_{t=1}^n\varepsilon_t^2+\nu n\left(\min_{1\le j\le m_0}|r^0_j-r^0_{j-1}|\right)\right\}=1,$$

where $(\tilde r_1,\tilde r_2,\ldots,\tilde r_m)=\arg\min_{(r_1,r_2,\ldots,r_m)}S_n(r_1,r_2,\ldots,r_m)$.

Proof of Theorem 2.4. We prove the first conclusion by showing (a) $P(\widehat{\widehat m}<m_0)\to 0$ and (b) $P(\widehat{\widehat m}>m_0)\to 0$. For the first claim (a) $P(\widehat{\widehat m}<m_0)\to 0$, note that Theorem 2.3 implies that there exist points $\hat r_{ni}\in\mathcal{A}_n$, $i=1,\ldots,m_0$ such that $\max_{1\le i\le m_0}|\hat r_{ni}-r^0_i|\le\gamma_n$. Thus, it is enough to show that if $m<m_0$, then

$$IC(\hat{\hat m},\hat{\hat r})\ge S_n(\hat r_{n1},\ldots,\hat r_{nm_0})+m_0\omega_n,\quad\text{in probability.} \tag{A.14}$$

Let $R_n(m_0)=\{(r_1,r_2,\ldots,r_{m_0}):|r_i-r^0_i|\le\gamma_n,i=1,2,\ldots,m_0\}$. For any $\mathbf{r}\in R_n(m_0)$, rewrite $S_n(r_1,r_2,\ldots,r_{m_0})$ as

$$\begin{aligned}&S_n(r_1,r_2,\ldots,r_{m_0})\\&=\sum_{i=1}^n(Y_i-\widehat{\widehat{\boldsymbol{\phi}}}_1^{\mathrm{T}}\boldsymbol{Y}_i)^2I(-\infty<Y_{i-d}\le r^0_1-\gamma_n)\\&\quad+\sum_{j=2}^{m_0}\sum_{i=1}^n(Y_i-\widehat{\widehat{\boldsymbol{\phi}}}_j^{\mathrm{T}}\boldsymbol{Y}_i)^2I(r^0_{j-1}+\gamma_n<Y_{i-d}\le r^0_j-\gamma_n)\\&\quad+\sum_{i=1}^n(Y_i-\widehat{\widehat{\boldsymbol{\phi}}}_{m_0+1}^{\mathrm{T}}\boldsymbol{Y}_i)^2I(r^0_{m_0}+\gamma_n<Y_{i-d}<\infty)\\&\quad+\sum_{j=1}^{m_0}\sum_{i=1}^n(Y_i-\widehat{\widehat{\boldsymbol{\phi}}}_j^{\mathrm{T}}\boldsymbol{Y}_i)^2I(r^0_j-\gamma_n<Y_{i-d}\le r_j)\\&\quad+\sum_{j=1}^{m_0}\sum_{i=1}^n(Y_i-\widehat{\widehat{\boldsymbol{\phi}}}_{j+1}^{\mathrm{T}}\boldsymbol{Y}_i)^2I(r_j<Y_{i-d}\le r^0_j+\gamma_n)\\&=:L_1+L_2+\cdots+L_5.\end{aligned} \tag{A.15, A.16, A.17}$$

Note that on $(-\infty,r^0_1-\gamma_n]$, $(r^0_{j-1}+\gamma_n,r^0_j-\gamma_n]$ and $(r^0_m+\gamma_n,\infty)$, $\widehat{\widehat{\boldsymbol{\phi}}}_j$, $1\le j\le m_0$ is the LSE of $\boldsymbol{\beta}^0_j$. It follows that

$$L_1+L_2+L_3\le\sum_{i=1}^n\varepsilon_i^2I(Y_{i-d}\in(-\infty,r^0_1-\gamma_n]\cup(r^0_{j-1}+\gamma_n,r^0_j-\gamma_n]\cup(r^0_m+\gamma_n,\infty)).$$

Since $|r_i - r_i^0| \le \gamma_n$, it can be shown that there exists a constant $C_0 > 0$ such that

$$L_4 + L_5 = \sum_{j=1}^{m_0}\sum_{i=1}^{n} \varepsilon_i^2 I(Y_{i-d} \in (r_j^0 - \gamma_n, r_j] \cup (r_j, r_j^0 + \gamma_n]) + C_0 m_0 n\gamma_n.$$

Thus,

$$S_n(r_1, r_2, \ldots, r_{m_0}) = \sum_{t=1}^{n} \varepsilon_t^2 + C_0 m_0 n\gamma_n$$

holds uniformly for all $\mathbf{r} \in R_n(m_0)$. This implies

$$S_n(\widehat{r}_{n1}, \widehat{r}_{n2}, \ldots, \widehat{r}_{nm_0}) \le \sum_{t=1}^{n} \varepsilon_t^2 + C_0 m_0 n\gamma_n. \tag{A.18}$$

However, by Lemma A.4, we know that if $m < m_0$, then

$$S_n(\widehat{\widehat{r}}_1, \widehat{\widehat{r}}_2, \ldots, \widehat{\widehat{r}}_m) \ge \sum_{t=1}^{n} \varepsilon_t^2 + \nu n \left(\min_{1\le i\le m_0} |r_i^0 - r_{i-1}^0| \right). \tag{A.19}$$

Combining (A.18) with (A.19) yields

$$\begin{aligned} IC(\hat{\hat{m}}, \hat{\hat{t}}) &= S_n(\widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_m) + m\omega_n \\ &\ge \sum_{t=1}^{n} \varepsilon_t^2 + \nu n \left(\min_{1\le i\le m_0} |r_i^0 - r_{i-1}^0| \right) + m\omega_n \\ &\ge S_n(\widehat{r}_{n1}, \ldots, \widehat{r}_{nm_0}) + m_0\omega_n \\ &\quad + \nu n \left(\min_{1\le i\le m_0} |r_i^0 - r_{i-1}^0| \right) - C_0 m_0 n\gamma_n - (m_0 - m)\omega_n \\ &\ge S_n(\widehat{r}_{n1}, \ldots, \widehat{r}_{nm_0}) + m_0\omega_n, \end{aligned}$$

where the last inequality follows from the conditions $m_0\omega_n/(n \min_{1\le i\le m_0} |r_i^0 - r_{i-1}^0|) \to 0$ and $(\min_{1\le i\le m_0} |r_i^0 - r_{i-1}^0|)/(m_0\gamma_n) \to \infty$. This gives (A.14) and $P\{\widehat{\widehat{m}} < m_0\} \to 0$ as $n \to \infty$.

Next, we establish claim (b): $P\{\widehat{\widehat{m}} > m_0\} \to 0$. To this end, it is enough to show that if $m > m_0$, then $IC(m, \widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_m) > IC(m_0, \widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_{m_0})$.

Observe that when $m > m_0$,

$$\begin{aligned} S_n(\widehat{r}_{n1}, \ldots, \widehat{r}_{nm_0}) &\ge S_n(\widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_{m_0}) \ge S_n(\widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_m) \\ &\ge S_n(\widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_m, r_1^0, \ldots, r_{m_0}^0). \end{aligned} \tag{A.20}$$

Following the same argument as in Lemma A.4, we have in probability that

$$S_n(\widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_m, r_1^0, \ldots, r_{m_0}^0) \ge \sum_{i=1}^{n} \varepsilon_i^2 - (m + m_0)n\gamma_n. \tag{A.21}$$

Thus, by Eqs. (A.18), (A.20) and (A.21), it follows that in probability

$$S_n(\widehat{\widehat{r}}_1, \widehat{\widehat{r}}_2, \ldots, \widehat{\widehat{r}}_{m_0}) - S_n(\widehat{\widehat{r}}_1, \widehat{\widehat{r}}_2, \ldots, \widehat{\widehat{r}}_m) \le [m + m_0(1 + C_0)]n\gamma_n.$$

This gives

$$\begin{aligned} &IC(m, \widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_m) - IC(m_0, \widehat{\widehat{r}}_1, \ldots, \widehat{\widehat{r}}_{m_0}) \\ &\ge (m - m_0)\omega_n - [m + m_0(1 + C_0)]n\gamma_n > 0 \end{aligned} \tag{A.22}$$

in probability since $\lim_{n\to\infty} \omega_n/(m_0 n\gamma_n) \ge C_0 + 3$. Thus, $P\{\widehat{\widehat{m}} > m_0\} \to 0$.

For the second conclusion, note that if there exists an $1 \le i \le m_0$ such that $|\widehat{\widehat{r}}_i - r_i^0| \ge 2C_0 m_0\gamma_n/\nu =: Bm_0\gamma_n$, where C_0 is given in (A.18), then there exists an r_l^0 such that $|\widehat{\widehat{r}}_i - r_l^0| \ge Bm_0\gamma_n$ holds for all $1 \le i \le m_0$. Consequently, it can be seen from Lemma A.4 that

$$S_n(\widehat{\widehat{r}}_1, \widehat{\widehat{r}}_2, \ldots, \widehat{\widehat{r}}_{m_0}) \ge \sum_{t=1}^{n} \varepsilon_t^2 + \nu Bm_0 n\gamma_n = \sum_{t=1}^{n} \varepsilon_t^2 + 2C_0 m_0 n\gamma_n.$$

This contradicts (A.18) and $P\{\max_{1\le i\le m_0} |\widehat{\widehat{r}}_i - r_i^0| \le Bm_0\gamma_n\} \to 1$. □

Proof of Theorem 2.5. By (A.19), we have that $P\{|\mathcal{A}_n^*| < m_0\} \to 0$ as $n \to \infty$. Next, we turn to show that as $n \to \infty$,

$$P\{|\mathcal{A}_n^*| > m_0\} \to 0.$$

Given $|\mathcal{A}_n^*| = m > m_0$,

$$S_n(\mathcal{A}_n^*) = \min_{(r_1, r_2, \ldots, r_m) \subseteq \mathcal{A}_n} S_n(r_1, r_2, \ldots, r_m),$$

when combining with (A.18) means that

$$\begin{aligned} S_n(\hat{r}_1^*, \hat{r}_2^*, \ldots, \hat{r}_m^*) &\le S_n(\widehat{r}_1, \ldots, \widehat{r}_m) \le S_n(\widehat{r}_1, \widehat{r}_2, \ldots, \widehat{r}_{m_0}) \\ &\le \sum_{t=1}^{n} \varepsilon_t^2 + C_0 m_0 \gamma_n/\nu. \end{aligned} \tag{A.23}$$

On the other hand, similar to (A.21), we have in probability

$$\begin{aligned} S_n(\hat{r}_1^*, \hat{r}_2^*, \ldots, \hat{r}_m^*) &\ge S_n(\hat{r}_1^*, \hat{r}_2^*, \ldots, \hat{r}_m^*, r_1^0, \ldots, r_{m_0}^0) \\ &\ge \sum_{t=1}^{n} \varepsilon_t^2 - (m + m_0)n\gamma_n. \end{aligned} \tag{A.24}$$

Using Eq. (A.23), (A.24) as in Theorem 2.4, we have $P\{|\mathcal{A}_n^*| > m_0\} \to 0$. The second conclusion of Theorem 2.5 follows as in Theorem 2.4. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2015.03.023.

References

An, H.Z., Huang, F.C., 1996. The geometrical ergodicity of non-linear autoregressive models. Statist. Sinica 6, 943–956.

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37, 1705–1732.

Bleakley, K., Vert, J.-P., 2011. The group fused Lasso for multiple change-point detection. http://hal.archives-ouvertes.fr/docs/00/60/21/21/PDF/techreport.pdf, hal-00602121, version1, 21 Jun 2011.

Boysen, L., Kempe, A., LiebsCher, V., Munk, A., Wittich, O., 2009. Consistencies and rates of convergence of jump-penalized least squares estimators. Ann. Statist. 37, 157–183.

Brockwell, P.J., Liu, J., Tweedie, R.L., 1992. On the existence of stationary threshold autoregressive moving-average processes. J. Time Ser. Anal. 13, 95–107.

Chan, K.S., 1993. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. Ann. Statist. 21, 520–533.

Chan, N.H., Kutoyants, A., 2010. On parameter estimations of threshold autoregressive models. Stat. Inference Stoch. Process. 15, 81–104.

Chan, K.S., Tong, H., 1985. On the use of the deterministic Lyapunov function for the ergodicity of stochastic difference equations. Adv. Appl. Probab. 17, 666–678.

Chan, N.H., Yau, C.Y., Zhang, R.M., 2014. Group LASSO for structural break time series. J. Amer. Statist. Assoc. 109, 590–599.

Chen, R., Tsay, R.S., 1991. On the ergodicity of TAR(1) processes. Ann. Appl. Probab. 1, 613–634.

Coakley, J., Fuertes, A., Pérez, M., 2003. Numerical issues in threshold autoregressive modeling of time series. J. Econom. Dynam. Control 27, 2219–2242.

Davis, R.A., Lee, T.C.M., Rodriguez-Yam, G.A., 2006. Structure break estimation for nonstationary time series models. J. Amer. Statist. Assoc. 101, 223–239.

Efron, B., Johnstone, I., Hastie, T., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32, 407–499.

Gonzalo, J., Pitarakis, J.-Y., 2002. Estimation and model selection based inference in single and multiple threshold models. J. Econometrics 110, 319–352.

Hansen, B., 2000. Sample splitting and threshold estimation. Econometrica 68, 575–603.

Harchaoui, Z., Levy-Leduc, C., 2010. Multiple change-point estimation with a total variation penalty. J. Amer. Statist. Assoc. 105, 1480–1493.
Li, D., Ling, S., 2012. On the least squares estimation of multiple-regime threshold autoregressive models. J. Econometrics 167, 240–253.
Liu, J., Susko, E., 1992. On strict stationarity and ergodicity of a nonlinear ARMA model. J. Appl. Probab. 29, 363–373.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288.
Tong, H., 1978. On a threshold model. In: Chen, C.H. (Ed.), Pattern Recognition and Signal Processing. Sijthoff and Noordhoff, Amsterdam, pp. 101–141.
Tong, H., 1990. Non-linear Time Series: A Dynamical System Approach. Oxford University Press, New York.
Tong, H., 2010. Threshold models in time series analysis—30 years on. Stat. Interface 4, 129–130.
Tsay, R.S., 1989. Testing and modeling threshold autoregressive processes. J. Amer. Statist. Assoc. 84, 231–240.
Yao, Y.C., 1988. Estimating the number of change-points via Schwarz' criterion. Statist. Probab. Lett. 6, 181–189.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 49–67.