Journal of Business & Economic Statistics

A Publication of the American Statistical Association

Volume 35 Number 2 April 2017

Special Issue on Regime Switching and Threshold Models
Guest Editors: Kung-Sik Chan, Bruce E. Hansen, and Allan Timmermann

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://www.tandfonline.com/loi/ubes20

Threshold Estimation via Group Orthogonal Greedy Algorithm

Ngai Hang Chan, Ching-Kang Ing, Yuanbo Li & Chun Yip Yau

To cite this article: Ngai Hang Chan, Ching-Kang Ing, Yuanbo Li & Chun Yip Yau (2017) Threshold Estimation via Group Orthogonal Greedy Algorithm, Journal of Business & Economic Statistics, 35:2, 334-345, DOI: 10.1080/07350015.2015.1064820

To link to this article: http://dx.doi.org/10.1080/07350015.2015.1064820

Accepted author version posted online: 12 Aug 2015.
Published online: 13 Mar 2017.

Threshold Estimation via Group Orthogonal Greedy Algorithm

Ngai Hang Chan
Southwestern University of Finance and Economics and The Chinese University of Hong Kong, Shatin, NT, Hong Kong, Hong Kong (*nhchan@sta.cuhk.edu.hk*)

Ching-Kang Ing
Academia Sinica, Institute of Statistical Sciences, National Taiwan University, Taipei 11529, Taiwan (*cking@stat.sinica.edu.tw*)

Yuanbo Li and Chun Yip Yau
Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, Hong Kong (*jacoblyb@gmail.com; cyyau@sta.cuhk.edu.hk*)

A threshold autoregressive (TAR) model is an important class of nonlinear time series models that possess many desirable features such as asymmetric limit cycles and amplitude-dependent frequencies. Statistical inference for the TAR model encounters a major difficulty in the estimation of thresholds, however. This article develops an efficient procedure to estimate the thresholds. The procedure first transforms multiple-threshold detection to a regression variable selection problem, and then employs a group orthogonal greedy algorithm to obtain the threshold estimates. Desirable theoretical results are derived to lend support to the proposed methodology. Simulation experiments are conducted to illustrate the empirical performances of the method. Applications to U.S. GNP data are investigated.

KEY WORDS: High-dimensional regression; Information criteria; Multiple-regime; Multiple-threshold; Nonlinear time series.

1. INTRODUCTION

The $(m + 1)$-regime threshold autoregressive (TAR) model, defined by

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

where $j = 1, \ldots, m + 1$ and $\eta_t \sim \text{IID}(0, 1)$, asserts that the autoregressive structure of the process $\{Y_t\}$ is governed by the range of values of Y_{t-d}. The integer-valued quantity d is called the delay parameter and the parameters $-\infty = r_0 < r_1 < \cdots < r_m < r_{m+1} = \infty$ are known as the thresholds. The regime-changing autoregressive structure in (1.1) offers a simple and convenient means for interpretation. More importantly, realizations of this model capture important features observed in many empirical studies that cannot be explained by linear time series models, for example, asymmetric limit cycles, amplitude dependent frequencies, jump resonance and chaos, among others. As a result, the TAR model has attracted enormous attention in the literature across different fields such as finance, econometrics, and hydrology since the seminal article of Tong (1978). See, for example, the excellent surveys on the TAR models by Tong (1990, 2011), and theoretical backgrounds on inference for TAR models by Chan (1993), Li and Ling (2012), and Li, Ling, and Zhang (2016).

In spite of the well-developed asymptotic theory of estimation, estimation of TAR models incurs a high computational cost due to the irregular nature of the threshold parameters; for example, see Li and Ling (2012). In particular, to explicitly evaluate the loss functions such as least-squares criterion or likelihood functions, the thresholds $r_1, \ldots, r_m$ have to be specified first a priori. Hence, for an $(m + 1)$-regime TAR model, locating the global minimum of any loss function requires a multi-parameter grid search over all possible values of the r threshold parameters, which is computational infeasible, if not impossible. To circumvent this difficulty, Tsay (1989) developed a transformation that connects (1.1) to a change-point model and suggested a graphical approach to visually determine the number and values of the thresholds. Coakley, Fuertes, and Pérez (2003) used similar techniques to develop an estimation approach that is based on QR factorizations of matrices. When the number of thresholds r is unknown, Gonzalo and Pitarakis (2002) suggested a sequential estimation procedure for choosing r, under the assumption that all σ_j's are equal. Recently, Chan, Yau, and Zhang (2015) reframed the problem into a regression variable selection context and proposed an computationally efficient way to solve the problem by least absolute shrinkage and selection operator (LASSO) estimation.

In this article, motivated by the connection between TAR model and the high-dimensional regression model developed by Chan, Yau, and Zhang (2015), we explore an alternative procedure to efficiently estimate the number and values of thresholds. The procedure is based on a group orthogonal greedy algorithm (GOGA) for the high-dimensional regression problem. With the use of GOGA, the well-known problem of biasedness in

Journal of Business & Economic Statistics
April 2017, Vol. 35, No. 2
DOI: 10.1080/07350015.2015.1064820

the LASSO procedure (e.g., Fan and Li 2001 and Zou 2006) can be circumvented. Simulation studies demonstrate that the GOGA procedure gives much better empirical results than the LASSO procedure. On the other hand, the asymptotic theory is nontrivial since the design matrix of the high-dimensional regression model constructed from a TAR model does not satisfy some standard regularity conditions assumed in the literature. In particular, the design matrix has highly correlated columns and thus the standard "restricted eigenvalue" condition (e.g., Bickel, Ritov, and Tsybakov 2009; Ing and Lai 2011) is not applicable. Substantial arguments are needed in establishing the consistency of the number and locations of the estimated thresholds by GOGA. Further, the convergence rate of the thresholds is also established.

This article is organized as follows. In Section 2, we first review the connection between TAR models and high-dimensional regression problem. In Section 3, estimation for TAR models with GOGA is discussed. Simulation studies are given in Sections 4. Applications to U.S. GNP data are provided in Section 5. Technical details are deferred to the Appendix.

2. TAR MODEL

In this section, we introduce the connection between the TAR model and a high-dimensional regression model developed in Chan, Yau, and Zhang (2015). For notational simplicity, suppose we observe $\mathbf{Y} = (Y_{1-d}, \dots, Y_n)^T$ from model (1.1). Let $(Y_{\pi(1)}, Y_{\pi(2)}, \dots, Y_{\pi(n)})^{\mathrm{T}}$ be the order statistics of $(Y_{1-d}, Y_{2-d}, \dots, Y_{n-d})^{\mathrm{T}}$, from the smallest to the largest. For example, if $Y_5 < Y_3$ are the two smallest observations among $(Y_{1-d}, Y_{2-d}, \dots, Y_{n-d})^{\mathrm{T}}$, then $\pi(1) = 5$ and $\pi(2) = 3$. Note from (1.1) that $(Y_{\pi(1)}, Y_{\pi(2)}, \dots, Y_{\pi(n)})^{\mathrm{T}}$ governs the dependence structure of $\boldsymbol{Y}_n^0 \triangleq (Y_{\pi(1)+d}, \dots, Y_{\pi(n)+d})^{\mathrm{T}}$, since the time lag between them is the delay parameter, d. Denote the error terms corresponding to $\boldsymbol{Y}_n^0$ by $\boldsymbol{\varepsilon}(n) = (\varepsilon_{\pi(1)+d}, \dots, \varepsilon_{\pi(n)+d})^{\mathrm{T}}$. Define the design matrix $\boldsymbol{X}_n$ as

$$\boldsymbol{X}_n = \begin{pmatrix} \boldsymbol{Y}_{\pi(1)}^{\mathrm{T}} & 0 & 0 & \dots & 0 \\ \boldsymbol{Y}_{\pi(2)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(2)}^{\mathrm{T}} & 0 & \dots & 0 \\ \boldsymbol{Y}_{\pi(3)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(3)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(3)}^{\mathrm{T}} & \dots & 0 \\ \vdots & & & \ddots & \vdots \\ \boldsymbol{Y}_{\pi(n)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(n)}^{\mathrm{T}} & \boldsymbol{Y}_{\pi(n)}^{\mathrm{T}} & \dots & \boldsymbol{Y}_{\pi(n)}^{\mathrm{T}} \end{pmatrix}, \tag{2.2}$$

where $\mathbf{Y}_{\pi(j)}^{\mathrm{T}} = (1, Y_{\pi(j)+d-1}, Y_{\pi(j)+d-2}, \dots, Y_{\pi(j)+d-p})$ and p is the AR order. Note that we can assume all the regimes follow the same AR(p) model since an AR(p_*) model with $p_* < p$ can be regarded as an AR(p) model with the last few coefficient equaling to zero. Note also that the design matrix $\boldsymbol{X}_n$ is of size $n \times n(p+1)$, that is, the number of variables is greater than the number of observations. For simplicity, we assume that the delay parameter d is known. Consider the regression model

$$\boldsymbol{Y}_n^0 = \boldsymbol{X}_n \boldsymbol{\theta}(n) + \boldsymbol{\varepsilon}(n), \tag{2.3}$$

where $\boldsymbol{\theta}(n) \triangleq \{\boldsymbol{\theta}_{\pi(1)}^{\mathrm{T}}, \boldsymbol{\theta}_{\pi(2)}^{\mathrm{T}}, \dots, \boldsymbol{\theta}_{\pi(n)}^{\mathrm{T}}\}^{\mathrm{T}}$ is the regression coefficient. By expanding the quantity $\boldsymbol{X}_n\boldsymbol{\theta}(n)$, it can be shown that (1.1) is equivalent to the model (2.3) with $\boldsymbol{\theta}_{\pi(1)} = \boldsymbol{\phi}_1$ and

$$\boldsymbol{\theta}_{\pi(j)} = \begin{cases} \boldsymbol{\phi}_{k+1} - \boldsymbol{\phi}_k & \text{if } Y_{\pi(j-1)} \le r_k < Y_{\pi(j)}, \text{ for } k = 1, \dots, m \\ 0 & \text{otherwise,} \end{cases} \tag{2.4}$$

where $\boldsymbol{\phi}_k = (\phi_0^{(k)}, \phi_1^{(k)}, \dots, \phi_p^{(k)})$ is the autoregressive parameter vector in the kth segment of the TAR model and $r_1, \dots, r_m$ are the unknown thresholds.

The basic idea of model (2.3) is as follows. The jth block $\boldsymbol{\theta}_{\pi(j)}$ of the coefficient vector $\boldsymbol{\theta}(n)$ indicates a change in the autoregressive parameter when the value of the threshold variable is increased from $Y_{\pi(j-1)}$ to $Y_{\pi(j)}$. Thus, if a threshold is within the interval $[Y_{\pi(j-1)}, Y_{\pi(j)}]$, then $\boldsymbol{\theta}_{\pi(j)}$ is nonzero and it follows that the corresponding observations $Y_{\pi(j-1)+d}$ and $Y_{\pi(j)+d}$ are generated from different autoregressive processes. Otherwise, if there is no threshold within the interval $[Y_{\pi(j-1)}, Y_{\pi(j)}]$, then $\boldsymbol{\theta}_{\pi(j)}$ is zero and $Y_{\pi(j-1)+d}$ and $Y_{\pi(j)+d}$ are in the same regime. Since only m of the vectors $\boldsymbol{\theta}_{\pi(j)}$s in $\boldsymbol{\theta}(n)$ are nonzero, the solution to the high-dimensional regression model (2.3) is sparse. In Section 3, we develop a three-step procedure using GOGA to obtain an estimate $\hat{\boldsymbol{\theta}}_n$ to (2.3).

Given the estimate $\hat{\boldsymbol{\theta}}_n$, unknown thresholds can be identified by the positions of nonzero elements in $\hat{\boldsymbol{\theta}}(n)$. In particular, the estimates of the thresholds are given by

$$\mathcal{A}_n = \{Y_{\pi(j-1)} : \widehat{\boldsymbol{\theta}}_{\pi(j)} \neq 0, \ j \geq 2\}. \tag{2.5}$$

Thus, $\hat{m} \triangleq \#(\mathcal{A}_n)$, the cardinality of the set $\mathcal{A}_n$, is the estimated number of thresholds. Denote the elements of $\mathcal{A}_n$ by $Y_{\pi(\hat{t}_1)}, \dots, Y_{\pi(\hat{t}_{\hat{m}})}$ so that $Y_{\pi(\hat{t}_k)}$ is the kth estimated threshold. Finally, the autoregressive parameter for the first and the kth regime can be estimated by $\hat{\phi}_1 = \hat{\boldsymbol{\theta}}_{\pi(1)}$ and $\hat{\phi}_k = \sum_{j=1}^{\hat{t}_{k+1}} \hat{\boldsymbol{\theta}}_{\pi(j)}$, respectively, where $k = 2, \dots \hat{m} + 1$.

Since many computationally efficient procedures for high dimensional regression have been developed in recent decades (e.g., Efron et al. 2004; Fan and Li 2001; Fan and Lv 2008; Buhlmann, Kalisch, and Maathuis 2009; Cho and Fryzlewicz 2012; Ing and Lai 2011), estimating the TAR model under the framework of (2.3) can be much more effective than the traditional approaches of a full combinatorial search. Chan, Yau, and Zhang (2015) proposed a group LASSO procedure to solve (2.3). The computation is conducted efficiently by a group LARS algorithm that approximately solves the group LASSO problem. In the next section, we develop a GOGA algorithm which yields improved thresholds estimates.

3. THREE-STEP PROCEDURE: GOGA + HDIC + TRIM

In this section, we introduce a three-step procedure originally proposed by Ing and Lai (2011) to estimate the high-dimensional regression model (2.3). Some modifications are introduced to the current setting. The first step employs the GOGA to build a solution path, which includes one additional group of variables sequentially. The second and third steps incorporate a high-dimensional information criterion (HDIC) to consistently select relevant groups of variables on the solution path.

3.1 Group Orthogonal Greedy Algorithm

Let $\boldsymbol{X}_{n,j} = (0, \ldots, 0, \boldsymbol{Y}_{\pi(j)}, \boldsymbol{Y}_{\pi(j+1)}, \ldots, \boldsymbol{Y}_{\pi(n)})^{\mathrm{T}}$ for $j = 1, \ldots, n$. From (2.4), the design matrix can be expressed as $\boldsymbol{X}_n = \{\boldsymbol{X}_{n,1}, \boldsymbol{X}_{n,2}, \ldots, \boldsymbol{X}_{n,n}\}$. Replacing $\boldsymbol{Y}_n^0$ by $\boldsymbol{Y}_n^0 - \overline{\boldsymbol{Y}}_n^0$ and $\boldsymbol{X}_{n,i}$ by $\boldsymbol{X}_{n,i} - \overline{\boldsymbol{X}}_{n,i}$, for $i = 1, \ldots, n$, where $\overline{\boldsymbol{Y}}_n^0 = n^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y}_n^0$ and $\overline{\boldsymbol{X}}_{n,i} = n^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\boldsymbol{X}_{n,i}$, we can assume that $\boldsymbol{Y}_n^0$ and $\boldsymbol{X}_{n,i}$ have mean zeros. Here, $\mathbf{1}$ is an n-dimension vector with all elements equaling to 1. The GOGA algorithm is described as follows:

Group Orthogonal Greedy Algorithm (GOGA)

1. Initializing $k = 0$ with fitted value $\widehat{\boldsymbol{Y}}^{(0)} = 0$ and active set $\hat{J}_0 = \emptyset$.
2. Compute the current residuals $\boldsymbol{U}^{(k)} := \boldsymbol{Y}_n^0 - \widehat{\boldsymbol{Y}}^{(k)}$. Regress $\boldsymbol{U}^{(k)}$ against each of $\boldsymbol{X}_{n,j}$ for $j = 1, \ldots, n$ and obtain the most correlated group
$$\hat{j}_{k+1} = \arg\min_{1\le j\le n} \left\| \boldsymbol{U}^{(k)} - \boldsymbol{X}_{n,j}\left(\boldsymbol{X}_{n,j}^{\mathrm{T}}\boldsymbol{X}_{n,j}\right)^{-1}\boldsymbol{X}_{n,j}^{\mathrm{T}}\boldsymbol{U}^{(k)} \right\|^2. \quad (3.6)$$
Include the $\hat{j}_{k+1}$th group to update the active set as $\hat{J}_{k+1} = \{\hat{j}_1, \ldots, \hat{j}_k, \hat{j}_{k+1}\}$.
3. Transform the matrix $\boldsymbol{X}_{n,\hat{j}_{k+1}}$ into the orthogonalized version $\boldsymbol{X}_{n,\hat{j}_{k+1}}^{\perp}$ as follows:
   (a) Denote $\mathbf{H}_{\{j\}}^{\perp} = \boldsymbol{X}_{n,j}^{\perp}(\boldsymbol{X}_{n,j}^{\perp}{}^{\mathrm{T}}\boldsymbol{X}_{n,j}^{\perp})^{-1}\boldsymbol{X}_{n,j}^{\perp}{}^{\mathrm{T}}$. The first column of $\boldsymbol{X}_{n,\hat{j}_{k+1}}^{\perp}$, denoted as $\boldsymbol{X}_{n,\hat{j}_{k+1},1}^{\perp}$, is computed by
$$\boldsymbol{X}_{n,\hat{j}_{k+1},1}^{\perp} = \boldsymbol{X}_{n,\hat{j}_{k+1},1} - \sum_{i=1}^{k}\mathbf{H}_{\{\hat{j}_i\}}^{\perp}\boldsymbol{X}_{n,\hat{j}_{k+1},1}, \quad (3.7)$$
   that is, the residual of projecting the first column of $\boldsymbol{X}_{n,\hat{j}_{k+1}}$ onto the space spanned by $\{\boldsymbol{X}_{n,\hat{j}_i}^{\perp}\}_{i=1,\ldots,k}$.
   (b) For $l = 2, \ldots, p+1$, the lth column of $\boldsymbol{X}_{n,\hat{j}_{k+1}}^{\perp}$, $\boldsymbol{X}_{n,\hat{j}_{k+1},l}^{\perp}$, is computed by
$$\boldsymbol{X}_{n,\hat{j}_{k+1},l}^{\perp} = \boldsymbol{X}_{n,\hat{j}_{k+1},l} - \sum_{i=1}^{k}\mathbf{H}_{\{\hat{j}_i\}}^{\perp}\boldsymbol{X}_{n,\hat{j}_{k+1},l} - \sum_{u=1}^{l-1}\mathbf{H}_{\{\hat{j}_{k+1},u\}}^{\perp}\boldsymbol{X}_{n,\hat{j}_{k+1},l}. \quad (3.8)$$
   Thus, the lth column $\boldsymbol{X}_{n,\hat{j}_{k+1},l}^{\perp}$ is orthogonal to both $\boldsymbol{X}_{n,\hat{j}_1}^{\perp}, \boldsymbol{X}_{n,\hat{j}_2}^{\perp}, \ldots, \boldsymbol{X}_{n,\hat{j}_k}^{\perp}$ and $\boldsymbol{X}_{n,\hat{j}_{k+1},1}^{\perp}, \ldots, \boldsymbol{X}_{n,\hat{j}_{k+1},l-1}^{\perp}$.
4. Updated the fitted value $\widehat{\boldsymbol{Y}}^{(k+1)}$ by
$$\widehat{\boldsymbol{Y}}^{(k+1)} = \widehat{\boldsymbol{Y}}^{(k)} + \mathbf{H}_{\{\hat{j}_{k+1}\}}^{\perp}\boldsymbol{U}^{(k)}. \quad (3.9)$$
5. Repeat Steps 2–4 until iteration $k = K_n$, where K_n is a prespecified upper bound. The resulting variables selected is given by $\hat{J}_{K_n} = \{\hat{j}_1, \ldots, \hat{j}_{K_n}\}$.

There are two advantages to obtain the orthogonalized predictors $\boldsymbol{X}_{n,i}^{\perp}$ sequentially. First, the difficulties from inverting high-dimensional matrices in the usual implementation of ordinary least squares can be circumvented by a componentwise linear regression. Second, the algorithm is more efficient since the same group cannot be selected repeatedly. If each group contains only one variable, then the GOGA reduces to the ordinary orthogonal greedy algorithm. Note that the most correlated group (3.6) can be rewritten as
$$\hat{j}_{k+1} = \arg\min_{1\le j\le n} 1 - r_j^2, \quad (3.10)$$
where r_j is the correlation coefficient between $\boldsymbol{X}_{n,j}$ and the current residual $\boldsymbol{U}^{(k)}$. In other words, OGA chooses the predictor that is most correlated with $\boldsymbol{U}^{(k)}$ at the kth step. This is similar to the LARS algorithm (Efron et al. 2004). However, the LARS algorithm selects the best predictor that has as much correlation with the current residual as the best predictor at the previous iteration does. In the high-dimension regression model (2.3), the design matrix in (2.2) has a very special structure: the adjacent groups are highly correlated since they differ by only one entry. As a consequence, LARS algorithm tends to select many irrelevant groups around each important group. The results could be even worse when the number of thresholds is large. On the other hand, the GOGA's "greedy" manner can avoid including too many irrelevant groups in the active set. After a group $\boldsymbol{X}_{n,j}$ is selected, then the groups near $\boldsymbol{X}_{n,j}$ would not be selected easily because their contributions to further reducing current residuals are relatively small compared to the groups that are far away from $\boldsymbol{X}_{n,j}$.

Ing and Lai (2011) first used the OGA to solve a general high-dimensional regression problem. They establish the sure screening property (i.e., including all relevant variables with probability approaching 1) of the OGA with a relative small number of iterations, $K_n = O((n/\log p)^{1/2})$, where n is the sample size and p is the number of variables. Their extensive simulation studies indicate that GOGA provides a better performance than other penalization methods such as SIS-SCAD, ISIS-SCAD, and adaptive Lasso. In the theoretical results of Ing and Lai (2011), it is assumed that the regression predictors are not highly correlated with each other, which is reasonable in a practical variable selection context. Since the design matrix $\boldsymbol{X}_n$ in (2.3) has the special property that adjacent groups are highly correlated, the assumption in Ing and Lai (2011) is not guaranteed, however. Indeed, as adjacent columns in $\boldsymbol{X}_n$ only differs by one entry, they cannot be distinguished asymptotically, and thus the usual "consistency" of model selection does not hold. Therefore, the theoretical properties of the solution for (2.3) is different from that in Ing and Lai (2011) and should incorporate information from TAR models. To establish the asymptotic theory, we impose the following assumptions, which are similar to Chan and Tsay (1998) and Chan, Yau, and Zhang (2015):

Assumptions.

A1: $\{\eta_t\}$ has a bounded, continuous and positive density function and $\mathrm{E}|\eta_1|^{2+\iota} < \infty$ for some $\iota > 0$.

A2: $\{Y_t\}$ is a α-mixing stationary process with a geometric decaying rate and $\mathrm{E}|Y_t|^{2+\iota} < \infty$.

A3: $\min_{2\le i\le m_0+1} ||\boldsymbol{\phi}_i^0 - \boldsymbol{\phi}_{i-1}^0|| > \nu$ for some $\nu > 0$, where $\boldsymbol{\phi}_i^0$ is the true AR parameter vector in the ith regime, $i = 1, \ldots, m_0 + 1$. Also, there exists $Z_i = (1, z_{p-1,i}, \ldots, z_{0,i})^{\mathrm{T}}$, where $z_{p-d,i} = r_{i-1}$, such that $(\boldsymbol{\phi}_{i-1}^0 - \boldsymbol{\phi}_i^0)^{\mathrm{T}} Z_i \neq 0$ for $i = 2, \ldots, m_0 + 1$.

A4: There exist constants l, u such that $r_i \in [l, u]$ for $i = l, 2, \ldots, m_0$, where m_0 is unknown and not fixed. Furthermore, for some γ_n satisfying $\gamma_n/n \to 0$, $m_0\gamma_n/n \to$

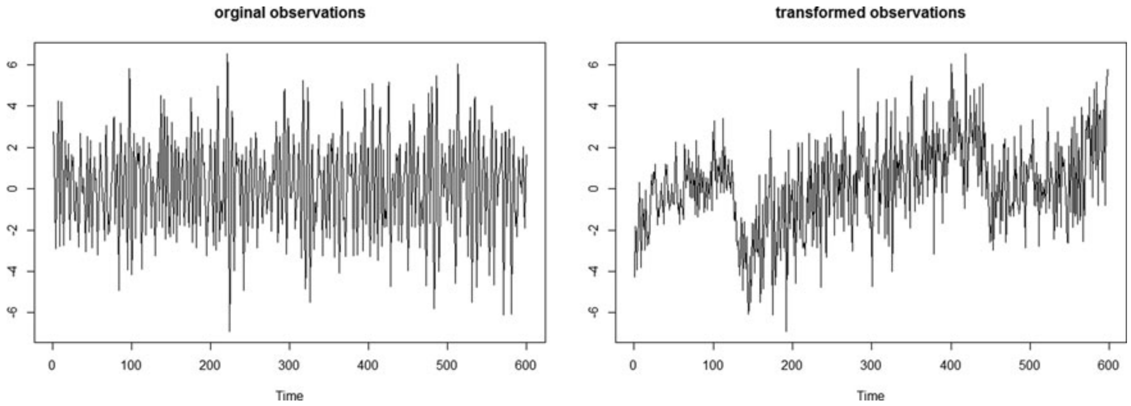


Figure 1. Example 1: Left: time series plot of a realization of (4.1). Right: time series plot of the transformed observations $\boldsymbol{Y}_n^0$. Sample size $n = 600$.

0 and $m_0 n \gamma_n^{-1}(\gamma_n)^{-\iota/2}(\log n)^{2(2+\iota)} \to 0$, it holds that $\min_{2\le i\le m_0} |r_i - r_{i-1}|/(\gamma_n/n) \to \infty$.

The following theorem ensures that the variables selected by the GOGA will be close to the true one so that consistency of the thresholds $\hat{r}_1, \dots, \hat{r}_m$ can be obtained.

Theorem 3.1. Suppose that Assumptions A1–A4 hold and let $K_n = O((n/\log n)^{1/2})$. Suppose that the groups $\hat{J}_{K_n} = \{\hat{j}_1, \hat{j}_2, \dots, \hat{j}_{K_n}\}$ are selected by the GOGA at the end of K_n iterations. Let $\mathcal{J} = \{j_1^0, j_2^0, \dots, j_{m_0}^0\}$ be the index set of true non-zero elements in $\boldsymbol{\theta}(n)$. Then as $n \to \infty$,

$$P\{d_H(\hat{J}_{K_n}, \mathcal{J}) \le \gamma_n\} \to 1, \tag{3.11}$$

where $\gamma_n = \log(n)$ and $d_H(A, B)$ is the Hausdorff distance between two sets A and B given by $d_H(A, B) = \max_{b\in B} \min_{a\in A} |b - a|$, and $d_H(A, \emptyset) = d_H(\emptyset, B) = 1$ if $\emptyset$ is an empty set. It follows that

$$P\left\{d_H(\hat{\boldsymbol{r}}, \boldsymbol{r}_0) \le \frac{\gamma_n}{n}\right\} \to 1, \tag{3.12}$$

where $\hat{\boldsymbol{r}}$ and $\boldsymbol{r}_0$ are the sets of estimated and true thresholds, respectively.

3.2 High-Dimensional Information Criterion

Note that Theorem 3.1 only ensures that each true threshold can be identified by an estimated threshold up to a γ_n/n neighborhood. It is possible that the set $\hat{\boldsymbol{r}}$ is greater than $\boldsymbol{r}_0$, thus the consistency of $\hat{m} = \#(\hat{\boldsymbol{r}})$ for the number of threshold is not guaranteed. One natural remedy is to select the best subset among $\hat{\boldsymbol{r}}$ according to some criteria. Based on Ing and Lai (2011), we prescribe the second and third steps of the threshold estimation procedure in this subsection. In particular, the second step selects the model along the solution path of GOGA based on a HDIC. The third step further trims the model to eliminate irrelevant variables by HDIC.

Specifically, for a nonempty subset J of $\{1, \dots, n\}$, let $\hat{\sigma}_J^2 = n^{-1}\|\boldsymbol{Y}_n^0 - \widehat{\boldsymbol{Y}}_{n;J}^0\|^2$, where $\widehat{\boldsymbol{Y}}_{n;J}^0$ is the fitted value by projecting $\boldsymbol{Y}_n^0$ onto the space spanned by $\{\boldsymbol{X}_{n,j}\}$, $j \in J$. Define the HDIC by

$$\begin{aligned} \mathrm{HDIC}(J) = n \log \hat{\sigma}_J^2 + \#(J) \log n(\log n \\ - \log\log n), \end{aligned} \tag{3.13}$$

$$\hat{k}_n = \arg \min_{1\le k\le K_n} \mathrm{HDIC}(\hat{J}_k), \tag{3.14}$$

in which $\hat{J}_k = \{\hat{j}_1, \dots, \hat{j}_k\}$ contains the first k groups along the GOGA solution path. Similar to other information criteria, in (3.13), the $n \log \hat{\sigma}_J^2$ is the lack of fit term and $\#(J) \log n(\log n - \log\log n)$ is the penalty term for model complexity. The HDIC reduces to the usual BIC without the factor $\log n - \log\log n$. Thus, the factor $\log n - \log\log n$ can be interpreted as the additional adjustment for a high-dimensional problem. Note also that the penalty term of the HDIC in Ing and Lai (2011) is $\log n(\log n(p + 1))$, where $n(p + 1)$ is the total number of columns in $\boldsymbol{X}_n$. It is different from the term $\log n(\log n - \log\log n)$ in (3.13) because groups of variable are selected. In the derivation of the HDIC in Section 4.2 of Ing and Lai (2011), the penalization factor $\log p - \log\log p$ is established, where p is the number of variables. Since the $\log\log p$ term is relatively small, only the $\log p$ term was used

Table 1. Percentage of correct identification of the number of thresholds ($\%m_0$), bias and empirical standard deviation (ESD) of the GOGA + HDIC + Trim

n	$\%m_0$		r_1	r_2
600	100	Bias	0.000	0.004
	(97)		(0.010)	(0.012)
		ESD	0.024	0.022
			(0.056)	(0.018)
900	100	Bias	0.000	0.002
	(96)		(0.006)	(0.008)
		ESD	0.015	0.015
			(0.037)	(0.011)
1200	100	Bias	0.000	0.002
	(95.7)		(0.004)	(0.007)
		ESD	0.011	0.012
			(0.027)	(0.009)

NOTE: The corresponding values of the two-step estimation in Chan, Yau, and Zhang (2015) are given in the parentheses. Replication = 1000.

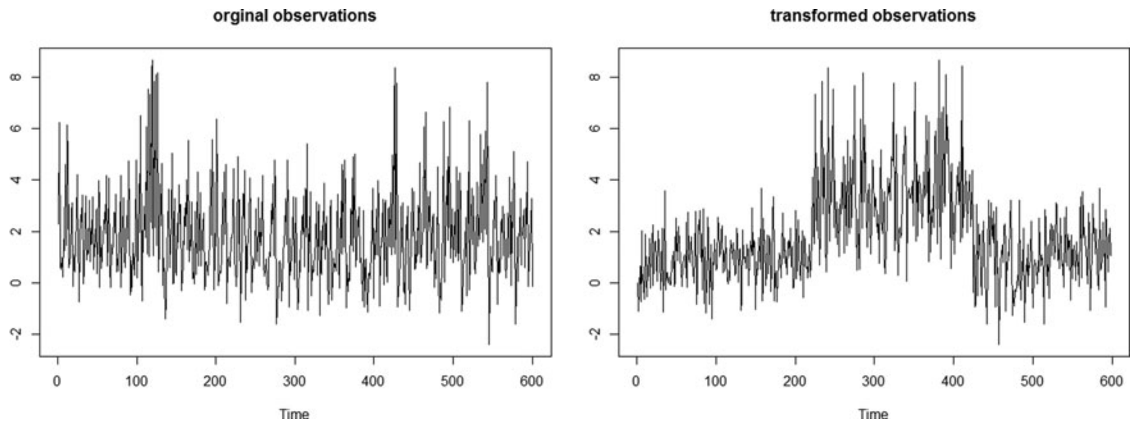


Figure 2. Example 2: Left: time series plot of a realization of (4.2). Right: time series plot of the transformed observations $\boldsymbol{Y}_n^0$. Sample size $n = 600$.

in defining the HDIC. However, extensive simulation evidence suggests that keeping the $-\log\log p$ term yields good finite sample performance in the threshold estimation context. Therefore, as there are n groups of variables, we have $p = n$ and the term $\log n - \log\log n$ in our definition of HDIC.

In terms of the computational complexity, the second step involves relatively little cost because $\hat{\sigma}_J^2$ can be readily computed in the kth GOGA iteration. Therefore, the HDIC along the GOGA solution path at each step can be obtained efficiently.

Finally, the third step of the threshold estimation procedure trims the solution set $\hat{J}_{\hat{k}_n}$ in the second step to eliminate irrelevant groups as follows. Define

$$\hat{N}_n = \{\hat{j}_l : \text{HDIC}(\hat{J}_{\hat{k}_n} - \{\hat{j}_l\}) > \text{HDIC}(\hat{J}_{\hat{k}_n}), \quad 1 \le l \le \hat{k}_n\} \text{ if } \hat{k}_n > 1, \tag{3.15}$$

and $\hat{N}_n = \{\hat{j}_1\}$ if $\hat{k}_n = 1$. The reason of conducting the third step is that the second-step procedure only chooses the best subset of variables *along the GOGA solution path*. While it guarantees that all relevant variables are selected, irrelevant groups may be present. Therefore, the third step is required to eliminate the irrelevant variable by HDIC again. Note that HDIC($\hat{J}_{\hat{k}_n}$) is already obtained in the preceding step, obtaining $\hat{N}_n$ only requires the computation of $\hat{k}_n - 1$ ordinary least-squares regressions. Since $\hat{k}_n$ is usually not large, the third step can be quickly computed. The following theorem shows that the three-step procedure, GOGA + HDIC + Trim, has the consistency property, that is, with probability approaching 1, the number of threshold is correctly estimated and all the true threshold can be identified up to a neighborhood shrinking to zero. The proof follows from similar arguments in Chan, Yau, and Zhang (2015) and is thus omitted.

Table 2. Percentage of correct identification of the number of thresholds (%m_0), bias and empirical standard deviation (ESD) of the GOGA + HDIC + Trim

n	%m_0		r_1	r_2
600	100 (90)	Bias	0.004 (0.005)	0.001 (0.009)
		ESD	0.016 (0.025)	0.019 (0.036)
900	100 (89.3)	Bias	0.003 (0.007)	0.001 (0.008)
		ESD	0.011 (0.027)	0.011 (0.011)
1200	100 (91.4)	Bias	0.001 (0.000)	0.000 (0.005)
		ESD	0.008 (0.026)	0.008 (0.007)

NOTE: The corresponding values of the two-step estimation in Chan, Yau, and Zhang (2015) are given in the parentheses. Replication = 1000.

Theorem 3.2. Assume that Assumption A1–A4 hold. For the index set $\hat{N}_n = \{\hat{j}_1, \hat{j}_2, \ldots, \hat{j}_{\#(\hat{N}_n)}\}$ that satisfies (3.15), there exist a constant $B > 0$ and $\gamma_n = O(\log n)$ such that as $n \to \infty$,

$$P\{|\hat{N}_n| = m_0\} \to 1 \text{ and } P\left\{\max_{1 \le i \le m_0} |\hat{j}_i - j_i^0| \le B\gamma_n\right\} \to 1.$$

Equivalently, we have

$$P\left\{\max_{1 \le i \le m_0} |\hat{r}_i - r_i^0| \le B\frac{\gamma_n}{n}\right\} \to 1.$$

4. SIMULATION STUDIES

In this section, we explore the finite sample performance of GOGA + HDIC + Trim. We also consider the two-step procedure proposed by Chan, Yau, and Zhang (2015), which uses LASSO to build the solution path. We first compare the performances of the two methods in several simple threshold models and then demonstrate the advantage of GOGA + HDIC + Trim in complicated cases, which contain larger number of thresholds. In applying the GOGA + HDIC + Trim, $K_n = \sqrt{n/\log n}$ is employed in the first step.

Example 1. In this example, 1000 realizations are simulated from the model

$$Y_t = \begin{cases} 2 + 0.8Y_{t-1} - 0.2Y_{t-2} + \epsilon_t, & \text{if } y_{t-1} \le -1.5, \\ 1.9Y_{t-1} - 0.81Y_{t-2} + \epsilon_t, & \text{if } -1.5 < y_{t-1} \le 1.5, \\ -2 + 1.32Y_{t-1} - 0.81Y_{t-2} + \epsilon_t, & \text{if } 1.5 \le y_{t-1}. \end{cases} \tag{4.1}$$

Following the transformation in Section 2.1, we obtain the response $\boldsymbol{Y}_n^0$. The original observations $\{Y_t\}$ and the transformed

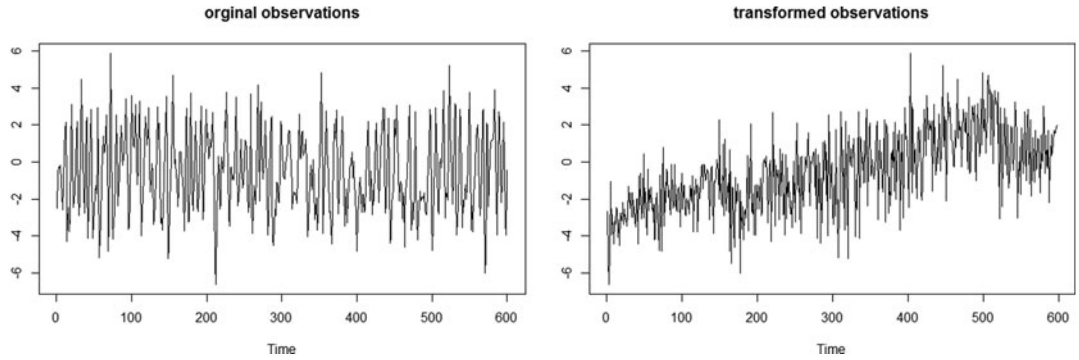


Figure 3. Example 3: Left: time series plot of a realization of (4.3). Right: time series plot of the transformed observations $\boldsymbol{Y}_n^0$. Sample size $n = 600$.

observations Y_n^0 are plotted in Figure 1 for a sample with size $n = 600$. Note that the right plot shows a clear regime switching structure, which suggests that the model may contain two thresholds. We apply the GOGA–HDIC–Trim and the LASSO procedure to estimate the threshold. The results are summarized in Table 1. The percentage (%) of correct estimation for the number of regimes, and the bias and standard error of the threshold estimates are reported. The bias and standard error are computed for the cases where the estimated number of thresholds equals to the true value. From Table 1, the GOGA + HDIC + Trim method has 100% accuracy in detecting the two thresholds and has lower bias and standard errors of the first threshold estimates in all the three different sample sizes. For the second threshold estimate, GOGA + HDIC + Trim has three times smaller bias. This is consistent with the finding that LASSO estimation tends to have a larger bias in the regression coefficients.

Example 2. In this example, 1000 realizations are generated from

$$Y_t = \begin{cases} 1 + 0.1Y_{t-1} + \epsilon_t, & \text{if } y_{t-1} \le 1, \\ 1 + 0.5Y_{t-1} + 0.8Y_{t-2} + \epsilon_t, & \text{if } 1 < y_{t-1} \le 2.5, \\ 2 + 0.1Y_{t-1} - 0.6Y_{t-2} + \epsilon_t, & \text{if } 2.5 \le y_{t-1}. \end{cases} \quad (4.2)$$

Similar to Figure 1, the transformed variable $\boldsymbol{Y}_n^0$ in Figure 2 also suggests that three regimes exist in the data. Table 2 reports the estimation results of the GOGA + HDIC + Trim and the LASSO procedure. In this case, GOGA + HDIC + Trim has a much better performance than the LASSO procedure in terms of the number and accuracy of the thresholds estimates. In particular, the percentage of correct identification is 100 for GOGA + HDIC + Trim but is only 90 for LASSO. Also, GOGA + HDIC + Trim attains lower bias and standard errors in both two thresholds for all sample sizes.

Example 3. In this example, the time series

$$Y_t = \begin{cases} 0.8Y_{t-1} - 0.2Y_{t-2} + \epsilon_t, & \text{if } y_{t-1} \le -2, \\ 1.9Y_{t-1} - 0.81Y_{t-2} + \epsilon_t, & \text{if } -2 < y_{t-1} \le 2, \\ 0.6Y_{t-1} - Y_{t-2} + \epsilon_t, & \text{if } 2 \le y_{t-1}, \end{cases} \quad (4.3)$$

is more difficult to handle because the intercepts in all segments are zero and each segment has a similar structure. In particular, it is observed from the right plot of Figure 3 that the regime switching pattern is more obscure. Performances of the two methods are summarized in Table 3. Surprisingly, the results of GOGA + HDIC + Trim are still very promising. The number of thresholds is perfectly identified in such a difficult situation. In contrast, only 70% of the LASSO estimates identify the correct number of thresholds. Moreover, GOGA + HDIC + Trim

Table 3. Percentage of correct identification of the number of thresholds ($\%m_0$), bias and empirical standard deviation (ESD) of the GOGA + HDIC + Trim

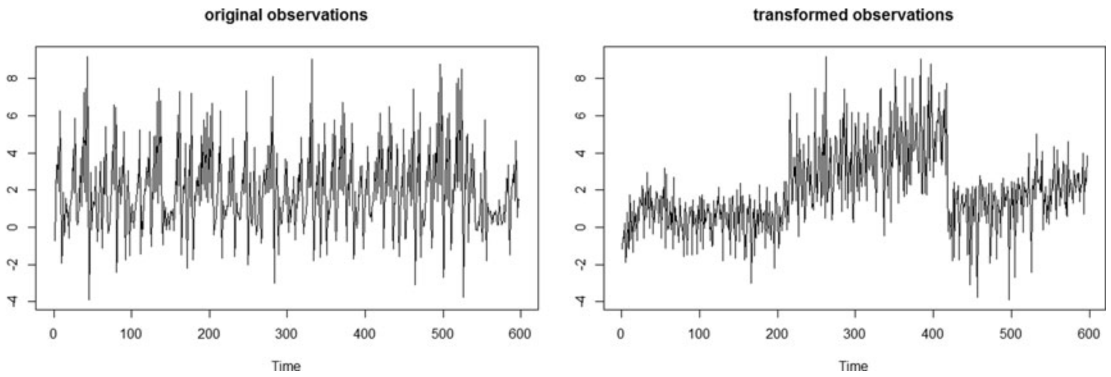
n	$\%m_0$		r_1	r_2
600	100	Bias	0.010	0.013
	(71.2)		(0.002)	(0.020)
		ESD	0.043	0.048
			(0.056)	(0.107)
900	100	Bias	0.007	0.013
	(72.5)		(0.000)	(0.015)
		ESD	0.029	0.037
			(0.036)	(0.069)
1200	100	Bias	0.006	0.008
	(71)		(0.002)	(0.012)
		ESD	0.022	0.027
			(0.033)	(0.053)

NOTE: The corresponding values of the two-step estimation in Chan, Yau, and Zhang (2015) are given in the parentheses. Replication = 1000.

Table 4. Percentage of correct identification of the number of thresholds ($\%m_0$), bias and empirical standard deviation (ESD) of the GOGA + HDIC + Trim

n	$\%m_0$		r_1	r_2
600	99.1	Bias	0.016	0.010
	(93.3)		(0.013)	(0.012)
		ESD	0.029	0.023
			(0.040)	(0.084)
900	99.6	Bias	0.012	0.007
	(96)		(0.007)	(0.010)
		ESD	0.021	0.013
			(0.025)	(0.012)
1200	99.2	Bias	0.010	0.005
	(96.6)		(0.007)	(0.008)
		ESD	0.016	0.011
			(0.020)	(0.010)

NOTE: The corresponding values of the two-step estimation in Chan, Yau, and Zhang (2015) are given in the parentheses. Replication = 1000.


original observations

transformed observations

Figure 4. Example 4: Left: time series plot of a realization of (4.4). Right: time series plot of the transformed observations $\boldsymbol{Y}_n^0$. Sample size $n = 600$.

achieves lower bias and standard deviations in all thresholds and in all sample size.

Example 4. In this example, 1000 realizations are simulated from the model

$$Y_t = \begin{cases} 1 + 0.1Y_{t-1} - 0.5Y_{t-2} + 0.5\epsilon_t, & \text{if } y_{t-1} \le 1, \\ 1 + 0.5Y_{t-1} + 0.8Y_{t-2} + \epsilon_t, & \text{if } 1 < y_{t-1} \le 2.5, \\ 2 + 0.1Y_{t-1} - 0.6Y_{t-2} + 0.2Y_{t-3} + \epsilon_t, & \text{if } 2.5 \le y_{t-1}, \end{cases} \tag{4.4}$$

see Figure 4. Compare to the previous examples, this model involves higher AR orders and conditional heteroscedasticity. The estimation results of GOGA + HDIC + Trim and the LASSO procedure are summarized in Table 4. It can be seen that percentage of correct identification of GOGA + HDIC + Trim is always larger than that of the LASSO procedure under three different settings of sample size n. Also, GOGA + HDIC + Trim achieves lower bias and standard deviations in most cases.

Example 5. In this last example, we use the following model to explore the performance of the two methods when the number of thresholds is large:

$$Y_t = \begin{cases} -4.5 - 0.6Y_{t-1} + \epsilon_t, & \text{if } y_{t-1} \le -3.5, \\ 2.5 + 0.3Y_{t-1} + 0.9Y_{t-2} + \epsilon_t, & \text{if } -3.5 < y_{t-1} \le -2.5, \\ -2 - 0.9Y_{t-1} + \epsilon_t, & \text{if } -2.5 < y_{t-1} \le -1.5, \\ 2.3 + 0.7Y_{t-1} + 0.5Y_{t-2} + \epsilon_t, & \text{if } -1.5 < y_{t-1} \le -0.5, \\ 1 + 0.1Y_{t-1} + \epsilon_t, & \text{if } -0.5 < y_{t-1} \le 0.5, \\ 3 + 0.9Y_{t-1} + \epsilon_t, & \text{if } 0.5 < y_{t-1} \le 1.5, \\ 1.6 - 0.9Y_{t-1} + \epsilon_t, & \text{if } 1.5 < y_{t-1} \le 2.5, \\ -0.5 - 0.8Y_{t-1} - 0.2Y_{t-2} + \epsilon_t, & \text{if } 2.5 < y_{t-1} \le 3.5, \\ 1.5 - 1.1Y_{t-1} + \epsilon_t, & \text{if } 3.5 \le y_{t-1}. \end{cases} \tag{4.5}$$

The transformed observations are plotted in Figure 5 and the estimation results are reported in Table 5. From Figure 5, the boundaries between each segments are indistinguishable especially for the last two thresholds. When the sample size is as small as 2000, each segment has only around 200 observations. With such a little information, GOGA + HIDC + Trim still has over 84% chance to detect the correct number of thresholds with small bias and standard errors. When the number of observations increases to 3000, GOGA + HDIC + Trim attains nearly 100% detection accuracy, and achieves low bias and small standard deviation for all thresholds estimates. Again, GOGA + HDIC + Trim outperforms the LASSO procedure in terms of the number and accuracy of the threshold estimates.

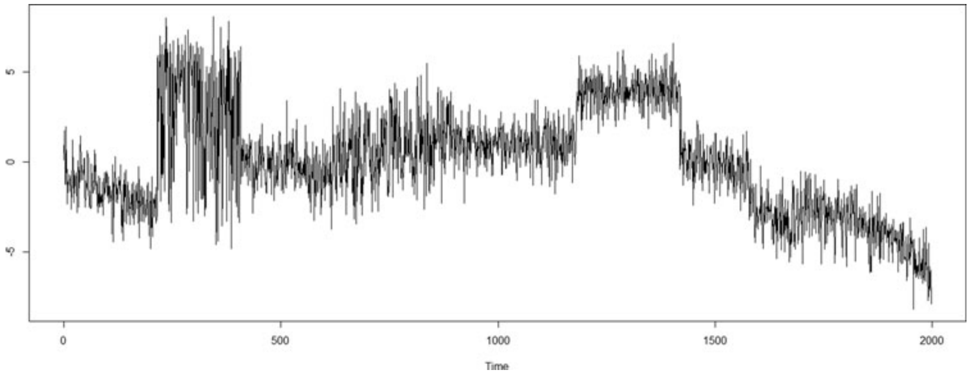


Figure 5. Example 5: time series plot of the transformed observations $\boldsymbol{Y}_n^0$ based on (4.5). Sample size $n = 2000$.

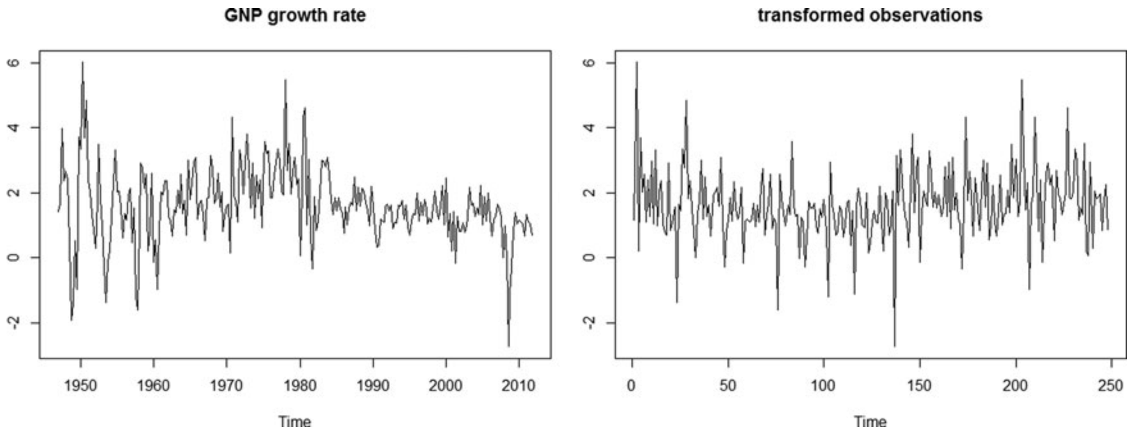


Figure 6. Left: growth rate of the U.S. GNP data from 1947 to 2012. Right: plot of the transformed observations.

Table 5. Percentage of correct identification of the number of thresholds ($\%m_0$), bias and empirical standard deviation (ESD) of the GOGA + HDIC + Trim

n	$\%m_0$		r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
5000	100	Bias	0.004	0.001	0.004	0.000	0.003	0.003	0.002	0.002
	(54.7)		(0.003)	(0.003)	(0.002)	(0.004)	(0.003)	(0.096)	(0.059)	(0.030)
		ESD	0.005	0.002	0.007	0.007	0.001	0.014	0.003	0.007
			(0.004)	(0.003)	(0.012)	(0.014)	(0.012)	(0.170)	(0.050)	(0.137)
3000	99.3	Bias	0.004	0.000	0.002	0.000	0.002	0.002	0.002	0.019
	(41)		(0.004)	(0.004)	(0.003)	(0.002)	(0.004)	(0.056)	(0.622)	(0.283)
		ESD	0.006	0.007	0.008	0.014	0.005	0.008	0.013	0.050
			(0.006)	(0.004)	(0.019)	(0.046)	(0.051)	(0.117)	(0.440)	(0.516)
2000	84.1	Bias	0.006	0.000	0.004	0.000	0.003	0.003	0.000	0.027
	(20.4)		(0.006)	(0.006)	(0.002)	(0.005)	(0.003)	(0.043)	(0.981)	(0.758)
		ESD	0.009	0.011	0.014	0.020	0.008	0.012	0.018	0.067
			(0.009)	(0.008)	(0.020)	(0.026)	(0.017)	(0.025)	(0.230)	(0.475)

NOTE: The corresponding values of the two-step estimation in Chan, Yau, and Zhang (2015) are given in the parentheses. Replication = 1000.

The LASSO procedure is particularly disappointing in this scenario in view of the fact that it has only 54.7% detection accuracy even for $n = 5000$. One possible reason for the failure of LASSO is that too many irrelevant variables around each true threshold are selected in the first step. The maximum number of threshold K is not sufficient to include groups around all true thresholds when the number of true threshold is large, hence the number of thresholds is under-estimated. As discussed in Section 3.1, GOGA avoids this problem and thus better empirical performance is achieved.

5. APPLICATIONS TO REAL DATA

In this section, we apply the GOGA–HDIC–Trim procedure to a multiple-regime TAR model for the growth rate of the quarterly U.S. 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

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

see Figure 6. This dataset has been investigated previously by Li and Ling (2012) and Chan, Yau, and Zhang (2015). Li and Ling (2012) used a three-regime model encompassing bad, good, and normal times and the two threshold are 1.20 and 2.43, respectively. The two-step procedure of Chan, Yau, and Zhang (2015) identified three thresholds at 1.23, 1.83, and 2.55, respectively. They also show that the four-regime model may give a better fit to the data. Setting $K_n = 6$ and AR order $p^* = 11$ in GOGA–HDIC–Trim, the estimated thresholds are 1.23, 1.65, and 2.23, which are very quite close to the estimates in Chan, Yau, and Zhang (2015). This example not only illustrates the usefulness of GOGA + HDIC + Trim, but also reinforces the notion that a four-regime model may give a better explanation to this GNP dataset.

APPENDIX: PROOF OF THEOREMS

First, define two types of group orthogonal greedy algorithms, population GOGA, and its generalized version, which are crucial to the study of the "sample version" of GOGA introduced in Section 3.1. For notational simplicity, we denote $\boldsymbol{X}_{n,j}$ by $\boldsymbol{X}_j$. Suppose the true model is $\boldsymbol{u} = \sum_{j=1}^n \boldsymbol{X}_j \boldsymbol{b}_j$. Let H_J be the projection matrix associated with the linear space spanned by X_j, $j \in J \subseteq \{1, \ldots, n\}$. Since

$$\frac{\left\| \boldsymbol{u} - \boldsymbol{X}_j \left(\boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{X}_j\right)^{-1} \boldsymbol{X}_j^{\mathrm{T}} \boldsymbol{u} \right\|^2}{\|\boldsymbol{u}\|^2} = 1 - \frac{\boldsymbol{u}^{\mathrm{T}} H_{\{j\}} \boldsymbol{u}}{\|\boldsymbol{u}\|^2},$$

the criterion (3.6) is equivalent to

$$\hat{j}_{k+1} = \arg\max_{1 \le j \le n} \boldsymbol{U}^{(k)\mathrm{T}} H_{\{j\}} \boldsymbol{U}^{(k)}.$$

This representation will be used in the following proof. First, we introduce the population GOGA as follows. Let $\boldsymbol{U}^{(0)} = \boldsymbol{u}$, $\widetilde{j}_1 = \arg\max_{1\le j\le n} \boldsymbol{U}^{(0)\mathrm{T}} H_{\{j\}} \boldsymbol{U}^{(0)}$ and $\boldsymbol{U}^{(1)} = \left(\boldsymbol{I} - \boldsymbol{H}_{\{\widetilde{j}_1\}}\right)\boldsymbol{u}$. At the mth iteration, we have

$$\begin{aligned}\widetilde{j}_m &= \arg\max_{1\le j\le n} \boldsymbol{U}^{(m-1)\mathrm{T}} H_{\{j\}} \boldsymbol{U}^{(m-1)}\,,\\ U^{(m)} &= (\boldsymbol{I} - \boldsymbol{H}_{\{\widetilde{j}_1,\ldots,\widetilde{j}_m\}})\boldsymbol{u}, \qquad \text{(A.1)}\end{aligned}$$

and the active set $\widetilde{J}_m = \{\widetilde{j}_1, \ldots, \widetilde{j}_m\}$. The population GOGA approximates $\boldsymbol{u}$ by $\boldsymbol{H}_{\widetilde{J}_m}\boldsymbol{u}$.

The generalization of the population GOGA considers an additional parameter $0 \le \xi \le 1$. At the ith step, instead of (A.1), the generalization of the population GOGA replaces $\widetilde{j}_i$ by $\widetilde{j}_{i,\xi}$, where $\widetilde{j}_{i,\xi}$ is any $1 \le l \le n$ satisfying

$$\boldsymbol{U}^{(i-1)\mathrm{T}} H_{\{l\}} \boldsymbol{U}^{(i-1)} \ge \xi \max_{1\le j\le n} \boldsymbol{U}^{(i-1)\mathrm{T}} H_{\{j\}} \boldsymbol{U}^{(i-1)}. \qquad \text{(A.2)}$$

Without loss of generality, we assume that each group $\boldsymbol{X}_j$ contains only two variables. Thus, the matrix $\boldsymbol{X}_n$ in (2.2) can be denoted as

$$\boldsymbol{X}_n = \begin{pmatrix} a_1 & b_1 & 0 & 0 & 0 & 0 & \ldots & 0 & 0\\ a_2 & b_2 & a_2 & b_2 & 0 & 0 & \ldots & 0 & 0\\ a_3 & b_3 & a_3 & b_3 & a_3 & b_3 & \ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ a_{n-1} & b_{n-1} & a_{n-1} & b_{n-1} & a_{n-1} & b_{n-1} & \ldots & a_{n-1} & b_{n-1}\\ a_n & b_n & a_n & b_n & a_n & b_n & \ldots & a_n & b_n \end{pmatrix}, \qquad \text{(A.3)}$$

where $\boldsymbol{X}_j$ is the submatrix containing the $(2j-1)$th and $(2j)$th columns of $\boldsymbol{X}_n$.

Now we prove an inequality for the generalization of population GOGA.

Lemma A.1. Let $0 \le \xi \le 1$, $m \ge 1$, and $\widetilde{J}_{m,\xi} = \{\widetilde{j}_{1,\xi}, \widetilde{j}_{2,\xi}, \ldots, \widetilde{j}_{m,\xi}\}$ be the active set selected by the generalization of population GOGA at the end of the mth iteration. Then,

$$\left\|\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m,\xi}}\boldsymbol{u}\right\|^2 \le \left(\sum_{j=1}^n \|\boldsymbol{a}_j\|\right)^2 (1 + m\xi)^{-1},$$

where $\boldsymbol{a}_j = \boldsymbol{b}_j^{\mathrm{T}}\left(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j\right)\boldsymbol{A}_j$ and $\boldsymbol{A}_j\boldsymbol{A}_j = \left(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j\right)^{-1}$.

Proof. For $J \subseteq \{1, \ldots, n\}$, $i \in 1, \ldots, n$ and $m \ge 1$, define $\boldsymbol{v}_{J,i} = n^{-1/2}\boldsymbol{A}_i\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}_J)\boldsymbol{u}$. Note that

$$\begin{aligned}\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m,\xi}}\right)\boldsymbol{u}\right\|^2 &\le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u} - \boldsymbol{X}_{\widetilde{j}_{m,\xi}}\left(\boldsymbol{X}_{\widetilde{j}_{m,\xi}}{}^{\mathrm{T}}\boldsymbol{X}_{\widetilde{j}_{m,\xi}}\right)^{-1}\right.\\ &\qquad \left.\times\ \boldsymbol{X}_{\widetilde{j}_{m,\xi}}{}^{\mathrm{T}}\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2\\ &\le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2 - n\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},\widetilde{j}_{m,\xi}}\right\|^2\\ &\le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2 - n\xi\max_{1\le j\le n}\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right\|^2. \qquad \text{(A.4)}\end{aligned}$$

Note that for $\forall i, j \in \{1, \ldots, n\}$,

$$\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m,\xi}}\right)\boldsymbol{u}\right\|^2 \le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2 - n\xi\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},i}\right\|^2,$$

and

$$\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m,\xi}}\right)\boldsymbol{u}\right\|^2 \le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2 - n\xi\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right\|^2.$$

Therefore,

$$\begin{aligned}\left(\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m,\xi}}\right)\boldsymbol{u}\right\|^2\right)^2 &\le \left(\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2\right)^2\\ &\quad + (n\xi)^2\left(\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},i}\right\|^2\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right\|^2\right)\\ &\quad -2\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2 n\xi\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},i}\right\|\\ &\quad \times\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right\|\\ &= \left(\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2 - n\xi\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},i}\right\|\right.\\ &\quad \left.\times\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right\|\right)^2.\end{aligned}$$

Thus, we have

$$\begin{aligned}\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m,\xi}}\right)\boldsymbol{u}\right\|^2 &\le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2\\ &\quad -n\xi\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},i}\right\|\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right\|.\end{aligned}$$

Since this is true for $\forall i, j \in \{1, \ldots, n\}$,

$$\begin{aligned}\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m,\xi}}\right)\boldsymbol{u}\right\|^2 &\le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2\\ &\quad -n\xi\max_{1\le i,j\le n}\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},i}\right\|\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right\|. \qquad \text{(A.5)}\end{aligned}$$

On the other hand,

$$\begin{aligned}\left(\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2\right)^2 &= \left(\sum_{j=1}^n \boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right)^2\\ &= n\left(\sum_{j=1}^n \boldsymbol{a}_j\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right)^2\\ &= n\sum_{i,j=1}^n\left(\boldsymbol{a}_i\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},i}\right)\left(\boldsymbol{a}_j\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right)\\ &\le n\max_{1\le i,j\le n}\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},i}\right\|\left\|\boldsymbol{v}_{\widetilde{J}_{m-1,\xi},j}\right\|\\ &\quad \times\left(\sum_{j=1}^n\|\boldsymbol{a}_j\|\right)^2. \qquad \text{(A.6)}\end{aligned}$$

It follows from (A.5) and (A.6) that

$$\begin{aligned}\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m,\xi}}\right)\boldsymbol{u}\right\|^2 &\le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2\\ &\quad \times\left(1 - \frac{\xi}{(\sum_{j=1}^n\|\boldsymbol{a}_j\|)^2}\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\widetilde{J}_{m-1,\xi}}\right)\boldsymbol{u}\right\|^2\right). \qquad \text{(A.7)}\end{aligned}$$

Combining (A.7) with Lemma 3.1 of Temlyakov (2000) yields the desired conclusion. □

Lemma A.2. For the matrix $\boldsymbol{X}_n$ defined in (A.3), denote $\boldsymbol{X}_{j,k}$ $(1 \le j \le n, k = 1, 2)$ as the kth columns of the jth group. Then, there exists $M > 0$ such that

$$\max_{1\le\#(J)\le K_n, i\notin J, k=1,2} ||(\boldsymbol{X}_J^{\mathrm{T}}\boldsymbol{X}_J)^{-1}\boldsymbol{X}_J^{\mathrm{T}}\boldsymbol{X}_{i,k}||_1 < M \quad a.s. \text{ as } n \to \infty$$

Proof. For notational simplicity, let $J = \{1, 2, \ldots, p\}$ and $i = p + s$, where $p < K_n$. That is, we select the first p groups of $\boldsymbol{X}_n$ as J and consider $\boldsymbol{X}_{p+s} = (\boldsymbol{X}_{p+s,1}, \boldsymbol{X}_{p+s,2})$ which is s groups away from J. Since the best predictor of $\boldsymbol{X}_{p+s}$ in terms of $\boldsymbol{X}_J$ is through $\boldsymbol{X}_p$, the regression coefficients of other variables except $\boldsymbol{X}_p$ is zero. Thus, we only need to calculate $(\boldsymbol{X}_p^{\mathrm{T}}\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^{\mathrm{T}}\boldsymbol{X}_{p+s}$ instead of the complicated inverse $(\boldsymbol{X}_J^{\mathrm{T}}\boldsymbol{X}_J)^{-1}$. Therefore,

$$(\boldsymbol{X}_J^{\mathrm{T}}\boldsymbol{X}_J)^{-1}\boldsymbol{X}_J^{\mathrm{T}}\boldsymbol{X}_{p+s} = \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ m11 & m12 \\ m21 & m22 \end{pmatrix}, \tag{A.8}$$

where
$$\begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} = \begin{pmatrix} \sum_{j=p}^{n} a_j^2 & \sum_{j=p}^{n} a_j b_j \\ \sum_{j=p}^{n} a_j b_j & \sum_{j=p}^{n} b_j^2 \end{pmatrix}^{-1} \times \begin{pmatrix} \sum_{j=p+s}^{n} a_j^2 & \sum_{j=p+s}^{n} a_j b_j \\ \sum_{j=p+s}^{n} a_j b_j & \sum_{j=p+s}^{n} b_j^2 \end{pmatrix}.$$

Let $\boldsymbol{q}_j = (a_j\ b_j)^{\mathrm{T}}$ and

$$\boldsymbol{Q} = \begin{pmatrix} \sum_{j=p}^{n} a_j^2 & \sum_{j=p}^{n} a_j b_j \\ \sum_{j=p}^{n} a_j b_j & \sum_{j=p}^{n} b_j^2 \end{pmatrix}.$$

Let $\boldsymbol{\Sigma} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{Q}$ be the population covariance matrix. Then, it follows that for sufficiently large n,

$$\boldsymbol{Q}^{-1}\left(\boldsymbol{Q} - \sum_{j=p}^{p+s-1} \boldsymbol{q}_j\boldsymbol{q}_j^{\mathrm{T}}\right) = \boldsymbol{I} - \boldsymbol{Q}^{-1}\sum_{j=p}^{p+s-1} \boldsymbol{q}_j\boldsymbol{q}_j^{\mathrm{T}}$$
$$= \boldsymbol{I} - n^{-1}(\boldsymbol{\Sigma}^{-1} + O_p(1))\sum_{j=p}^{p+s-1} \boldsymbol{q}_j\boldsymbol{q}_j^{\mathrm{T}}. \tag{A.9}$$

Combining (A.9), $s \le n$ and the ergodicity of TAR model, we have that $m_{ij},\ i, j = 1, 2$, are bounded by a constant almost surely. The other cases can be proved similarly. Thus, the proof of Lemma A.2 is completed. □

Proof of Theorem 3.1. For a given $0 \le \xi \le 1$, let $\widetilde{\xi} = \sqrt{2/(1-\xi)}$. Define $\boldsymbol{v}_{J,i} = n^{-1/2}\boldsymbol{A}_i\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}_J)\boldsymbol{u}$ and $\hat{\boldsymbol{v}}_{J,i} = n^{-1/2}\boldsymbol{A}_i\boldsymbol{X}_i(\boldsymbol{I} - \boldsymbol{H}_J)y$, where $y = \sum_{j=1}^{n}\boldsymbol{X}_j\boldsymbol{b}_j + \epsilon$, and the white-noise $\epsilon_t = \sigma_j\eta_t$ if Y_{t-d} belongs to the jth regime. Define also $\sigma_U = \max\{\sigma_1, \sigma_2, \ldots, \sigma_{m_0+1}\}$. Construct the sets A_n and B_n as

$$A_n = \left\{\max_{(J,i):\#(J)\le m-1, i\notin J} |\hat{\boldsymbol{v}}_{J,i}^{\mathrm{T}}\hat{\boldsymbol{v}}_{J,i} - \boldsymbol{v}_{J,i}^{\mathrm{T}}\boldsymbol{v}_{J,i}| \le C^2\sigma_U^2 n^{-1}\log n\right\},$$
$$B_n = \left\{\min_{0\le k\le m-1}\max_{1\le i,j\le n} \left\|\boldsymbol{v}_{\hat{J}_k,i}\right\|\left\|\boldsymbol{v}_{\hat{J}_k,j}\right\| > \widetilde{\xi}^2 C^2\sigma_U^2 n^{-1}\log n\right\},$$

where $\hat{J}_m = \{\hat{j}_1, \ldots, \hat{j}_m\}$ is the index set associated with the groups selected by the sample GOGA .

For all $1 \le q \le m$ and on the set $A_n \cap B_n$, we have

$$\begin{aligned}
\boldsymbol{v}_{\hat{J}_{q-1},\hat{j}_q}^{\mathrm{T}}\boldsymbol{v}_{\hat{J}_{q-1},\hat{j}_q} &\ge -\left|\hat{\boldsymbol{v}}_{\hat{J}_{q-1},\hat{j}_q}^{\mathrm{T}}\hat{\boldsymbol{v}}_{\hat{J}_{q-1},\hat{j}_q} - \boldsymbol{v}_{\hat{J}_{q-1},\hat{j}_q}^{\mathrm{T}}\boldsymbol{v}_{\hat{J}_{q-1},\hat{j}_q}\right| + \hat{\boldsymbol{v}}_{\hat{J}_{q-1},\hat{j}_q}^{\mathrm{T}}\hat{\boldsymbol{v}}_{\hat{J}_{q-1},\hat{j}_q} \\
&\ge -\max_{(J,i):\#(J)\le m-1, i\notin J}\left|\hat{\boldsymbol{v}}_{J,i}^{\mathrm{T}}\hat{\boldsymbol{v}}_{J,i} - \boldsymbol{v}_{J,i}^{\mathrm{T}}\boldsymbol{v}_{J,i}\right| + \hat{\boldsymbol{v}}_{\hat{J}_{q-1},\hat{j}_q}^{\mathrm{T}}\hat{\boldsymbol{v}}_{\hat{J}_{q-1},\hat{j}_q} \\
&\ge -C^2\sigma_U^2 n^{-1}\log n + \max_{1\le j\le n}\left(\hat{\boldsymbol{v}}_{\hat{J}_{q-1},j}^{\mathrm{T}}\hat{\boldsymbol{v}}_{\hat{J}_{q-1},j}\right) \\
&\ge -2C^2\sigma_U^2 n^{-1}\log n + \max_{1\le j\le n}\left(\boldsymbol{v}_{\hat{J}_{q-1},j}^{\mathrm{T}}\boldsymbol{v}_{\hat{J}_{q-1},j}\right) \\
&\ge \xi\max_{1\le j\le n}\boldsymbol{v}_{\hat{J}_{q-1},j}^{\mathrm{T}}\boldsymbol{v}_{\hat{J}_{q-1},j}.
\end{aligned} \tag{A.10}$$

That is, on the set $A_n \cap B_n$, $\hat{J}_m$ is also the set of groups chosen by the generalization of the population GOGA (A.2). Therefore, Lemma A.1 implies the upper bound

$$\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 I_{A_n\cap B_n} \le \left(\sum_{j=1}^{n}||\boldsymbol{a}_j||\right)^2(1 + m\xi)^{-1}. \tag{A.11}$$

Since $\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 \le \left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\hat{J}_k}\right)\boldsymbol{u}\right\|^2$ for $0 \le k \le m-1$, we have that on B_n^c,

$$\begin{aligned}
\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 &\le \min_{0\le k\le m-1}\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\hat{J}_k}\right)\boldsymbol{u}\right\|^2 \\
&\le \min_{0\le k\le m-1}\sqrt{n\max_{1\le i,j\le n}\left\|\boldsymbol{v}_{\hat{J}_k,i}\right\|\left\|\boldsymbol{v}_{\hat{J}_k,j}\right\|\left(\sum_{j=1}^{n}||\boldsymbol{a}_j||\right)^2} \\
&\le \widetilde{\xi}C\sigma_U(\log n)^{1/2}\sum_{j=1}^{n}||\boldsymbol{a}_j||. \quad \text{(Lemma A.1)}
\end{aligned}$$

Define

$$C_n = \left\{\min_{0\le k\le m-1}\max_{1\le i,j\le n}\left\|\boldsymbol{v}_{\hat{J}_k,i}\right\|\left\|\boldsymbol{v}_{\hat{J}_k,j}\right\| < \widetilde{\xi}^2 C^2\sigma_U^2 n^{-2}\log n\right\},$$

Similar to the preceding calculations, it can be shown that, on C_n,

$$\begin{aligned}
\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 &\le \min_{0\le k\le m-1}\sqrt{n\max_{1\le i,j\le n}\left\|\boldsymbol{v}_{\hat{J}_k,i}\right\|\left\|\boldsymbol{v}_{\hat{J}_k,j}\right\|\left(\sum_{j=1}^{n}||\boldsymbol{a}_j||\right)^2} \\
&\le \widetilde{\xi}C\sigma_U(n^{-1}\log n)^{1/2}\sum_{j=1}^{n}||\boldsymbol{a}_j||,
\end{aligned} \tag{A.12}$$

With the specific structure of the matrix X_n , it can be shown that for $m > K_n$ ($K_n >> m_0$),

$$\lim_{n\to\infty} P\left(\min_{0\le k\le m-1}\max_{1\le i,j\le n}\left\|\boldsymbol{v}_{\hat{J}_k,i}\right\|\left\|\boldsymbol{v}_{\hat{J}_k,j}\right\| < \widetilde{\xi}^2 C^2\sigma_U^2 n^{-2}\log n\right) = 1\,,$$

Denote $a = \frac{\sum_{j=1}^{n}||\boldsymbol{a}_j||}{\sqrt{n}}$. For a given m , it follows from (A.11) and (A.12) that

$$\begin{aligned}
n^{-1}\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 I_{A_n} &\le \frac{a^2}{1 + m\xi}I_{\{m<K_n\}} \\
&+ a\widetilde{\xi}C\sigma_U\frac{(\log n)^{1/2}}{n}I_{\{m\ge K_n\}}.
\end{aligned} \tag{A.13}$$

Since A_n decreases as m increases, we have for all $1 \le m \le K_n$ that

$$n^{-1}\left\|\left(\boldsymbol{I} - \boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 I_{\mathcal{A}} \le \frac{a^2}{1 + m\xi}I_{\{m<K_n\}} + a\widetilde{\xi}C\sigma_U\frac{(\log n)^{1/2}}{n}I_{\{m\ge K_n\}},$$

where $\mathcal{A}$ corresponds to the set A_n with $m = K_n$. Note that

$$\begin{aligned}
&\max_{\#(J)\le K_n-1, i\notin J}\left|\hat{\boldsymbol{v}}_{J,i}^{\mathrm{T}}\hat{\boldsymbol{v}}_{J,i} - \boldsymbol{v}_{J,i}^{\mathrm{T}}\boldsymbol{v}_{J,i}\right| \\
&= \max_{\#(J)\le K_n-1, i\notin J} n^{-1}\left|\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}_J)\boldsymbol{H}_{\{i\}}(\boldsymbol{I} - \boldsymbol{H}_J)\boldsymbol{u}\right. \\
&\quad\left. -\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}_J)\boldsymbol{H}_{\{i\}}(\boldsymbol{I} - \boldsymbol{H}_J)\boldsymbol{y}\right| \\
&= \max_{\#(J)\le K_n-1, i\notin J} n^{-1}\left|\epsilon^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}_J)\boldsymbol{H}_{\{i\}}(\boldsymbol{I} - \boldsymbol{H}_J)\epsilon\right. \\
&\quad\left. +2\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}_J)\boldsymbol{H}_{\{i\}}(\boldsymbol{I} - \boldsymbol{H}_J)\epsilon\right|.
\end{aligned} \tag{A.14}$$

Denote each group $\boldsymbol{X}_i$ as $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}\quad \boldsymbol{X}_{i2})$ and $n(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1} = \begin{pmatrix} w_{i,11} & w_{i,12} \\ w_{i,21} & w_{i,22} \end{pmatrix}$. Define also $v_{i1} = \boldsymbol{X}_{i1}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{H}_J)\epsilon$ and $v_{i2} = \boldsymbol{X}_{i2}^{\mathrm{T}}(\boldsymbol{I} -$

$\boldsymbol{H}_J)\epsilon$. The first part of (A.14) can be expressed as

$$n^{-1}\epsilon^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_J)\boldsymbol{X}_i(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_J)\epsilon = w_{i,11}\frac{v_{i1}^2}{n^2} + w_{i,22}\frac{v_{i2}^2}{n^2} + 2w_{i,12}\frac{v_{i1}v_{i2}}{n^2}.$$

Using Lemma A.2, we can show that

$$\begin{aligned}\max_{\#(J)\le K_n-1, i\notin J} n^{-1}|v_{i1}| &= \max_{\#(J)\le K_n-1, i\notin J} n^{-1}|\boldsymbol{X}_{i1}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{X}_J(\boldsymbol{X}_J^{\mathrm{T}}\boldsymbol{X}_J)^{-1}\boldsymbol{X}_J^{\mathrm{T}})\epsilon| \\ &= \max_{\#(J)\le K_n-1, i\notin J} n^{-1}|\boldsymbol{X}_{i1}^{\mathrm{T}}\epsilon - \boldsymbol{X}_{i1}^{\mathrm{T}}\boldsymbol{X}_J(\boldsymbol{X}_J^{\mathrm{T}}\boldsymbol{X}_J)^{-1}\boldsymbol{X}_J^{\mathrm{T}}\epsilon| \\ &\le \max_{\#(J)\le K_n-1, i\notin J}\ \max_{1\le i\le n, j=1,2} |n^{-1}\boldsymbol{X}_{ij}^{\mathrm{T}}\epsilon| \\ &\quad \times(1+||\boldsymbol{X}_{i1}^{\mathrm{T}}\boldsymbol{X}_J(\boldsymbol{X}_J^{\mathrm{T}}\boldsymbol{X}_J)^{-1}\boldsymbol{X}_J^{\mathrm{T}}||) \\ &\le \max_{1\le i\le n, j=1,2} |n^{-1}\boldsymbol{X}_{ij}^{\mathrm{T}}\eta|(1+M)\sigma_U. \qquad \text{(A.15)}\end{aligned}$$

Moreover, for a sufficiently large constant $C > 0$, we have

$$\begin{aligned}&P\left(\max_{\#(J)\le K_n-1, i\notin J} n^{-1}|v_{i1}| > C\sigma_U\left(\frac{\log n}{n}\right)^{1/2}\right) \\ &\le P\left(\max_{1\le i\le n, j=1,2}\left|\frac{\sum_{t=1}^n X_{ij,t}\eta_t}{n^{1/2}}\right| > C(\log n)^{1/2}(1+M)^{-1}\right) \\ &\to 0, \qquad \text{(A.16)}\end{aligned}$$

as $n\to\infty$. The last convergence (A.16) follows from the invariant principle and the fact that $\{\sum_{t=1}^n X_{ij,t}\eta_t\}_{i=1,2,\ldots}$ is a partial sum process and $\{\eta_t\}_{t=1,2,\ldots}$ is an independent sequence. Similarly, we can show that

$$P\left(\max_{\#(J)\le K_n-1, i\notin J} n^{-1}|v_{i2}| > C\sigma_U\left(\frac{\log n}{n}\right)^{1/2}\right) \to 0 \text{ as } n\to\infty. \qquad \text{(A.17)}$$

Since $w_{i,11}$, $w_{i,12}$ and $w_{i,22}$, $1\le i\le n$, are $O_p(1)$, it follows from (A.16) and (A.17) that

$$\begin{aligned}&P\left(\max_{\#(J)\le K_n-1, i\notin J} n^{-1}\left|\epsilon^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_J)\mathbf{H}_{\{i\}}(\boldsymbol{I}-\boldsymbol{H}_J)\epsilon\right|\right. \\ &\left. > C^2\sigma_U^2\frac{\log n}{n}\right) \to 0, \qquad \text{(A.18)}\end{aligned}$$

as $n\to\infty$. Since most coefficients are zero-vector and only m_0 groups are nonzero, the second part of (A.14) is equal to

$$2n^{-1}\sum_{j=j_1}^{j_{m_0}}\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_J)\boldsymbol{X}_i(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_J)\epsilon. \qquad \text{(A.19)}$$

For each component of (A.19), we have

$$\begin{aligned}&2n^{-1}\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_J)\boldsymbol{X}_i(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_J)\epsilon \\ &= 2\left(w_{i,11}\frac{v_{i1}}{n}\frac{a_{j1}}{n} + w_{i,22}\frac{v_{i2}}{n}\frac{a_{j2}}{n} + w_{i,21}\frac{v_{i1}}{n}\frac{a_{j2}}{n} + w_{i,12}\frac{v_{i2}}{n}\frac{a_{j1}}{n}\right),\end{aligned}$$

where $\max_{\#(J)\le K_n-1, i\notin J}\left|\frac{a_{jk}}{n}\right| = \max_{\#(J)\le K_n-1, i\notin J}|n^{-1}\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_{ik} - n^{-1}\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{H}_J \boldsymbol{X}_{ik}|$, $k = 1, 2$, are bounded. Combining with (A.16) and (A.17), we have

$$\begin{aligned}&P\left(\max_{\#(J)\le K_n-1, i\notin J} n^{-1}|2\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{H}_J)\mathbf{H}_{\{i\}}(\boldsymbol{I}-\boldsymbol{H}_J)\epsilon|\right. \\ &\left. > C^2\sigma_U^2\frac{\log n}{n}\right) \to 0, \qquad \text{(A.20)}\end{aligned}$$

as $n\to\infty$. Finally, it follows from (A.18) and (A.20) that

$$\lim_{n\to\infty} P(\mathcal{A}^c) = 0. \qquad \text{(A.21)}$$

Moreover, for $m \le K_n$, it follows from (A.13) that

$$\lim_{n\to\infty} P(A_n^c(m)) \le \lim_{n\to\infty} P(\mathcal{A}^c) = 0, \qquad \text{(A.22)}$$

$$\begin{aligned}&n^{-1}\left\|\left(\boldsymbol{I}-\boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 I_{A_n(m)} \\ &\le \frac{a^2}{1+m\xi} I_{\{m<K_n\}+a\tilde{\xi}C\sigma_U\frac{(\log n)^{1/2}}{n}I_{\{m\ge K_n\}}}. \qquad \text{(A.23)}\end{aligned}$$

It follows from (A.21) that,

$$\lim_{n\to\infty} P(F_n) = 0, \qquad \text{(A.24)}$$

where $F_n = \{n^{-1}||(\boldsymbol{I}-\boldsymbol{H}_{\hat{J}_m})\boldsymbol{u}||^2 > \frac{a^2}{1+m\xi}I_{\{m<K_n\}} + a\tilde{\xi}C\sigma_U\frac{(\log n)^{1/2}}{n} I_{\{m\ge K_n\}}\}$.

For $J \subseteq \{1,\ldots,n\}$ and $j\in J$, define $\tilde{\mathbf{b}}_j(J)$ be the parameter of $\boldsymbol{X}_j$ in the best linear predictor $\sum_{i\in J}\boldsymbol{X}_i\tilde{\boldsymbol{b}}_i(J)$ of $\boldsymbol{u}$ that minimizes $||\boldsymbol{u} - \sum_{i\in J}\boldsymbol{X}_i\boldsymbol{\lambda}_i||^2$. Let $\tilde{\boldsymbol{b}}_j(J) = 0$ if $j\notin J$. Thus,

$$n^{-1}\left\|\left(\boldsymbol{I}-\boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 = n^{-1}\left\|\sum_{j\in\hat{J}_m\cup\mathcal{J}}\boldsymbol{X}_j\left(\boldsymbol{b}_j - \tilde{\boldsymbol{b}}_j(\hat{J}_m)\right)\right\|^2. \qquad \text{(A.25)}$$

Now suppose that a true group j_s^0 has not been selected, that is, $\hat{J}_m^c\cap\mathcal{J} = \{j_s^0\}$, where $m > K_n$. We can conclude that there must exist at least one group $\hat{j}_l\in\hat{J}_m$ contained in the neighbourhood of j_s^0, that is, $|l-s| < \gamma_n = \log(n)$. Otherwise, on $\{\hat{J}_m^c\cap\mathcal{J}\ne\emptyset$ and $|l-s| > \log(n)\}$, it follows from (A.25) that

$$\begin{aligned}&n^{-1}\left\|\left(\boldsymbol{I}-\boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 \ge n^{-1}\left(\min_{j\in\mathcal{J}}||\boldsymbol{b}_j||^2\right)\lambda_{min} \\ &\times\left(\boldsymbol{X}_{\hat{J}_m\cup\mathcal{J}}^{\mathrm{T}}\boldsymbol{X}_{\hat{J}_m\cup\mathcal{J}}\right) = \frac{D\log(n)}{n}, \qquad \text{(A.26)}\end{aligned}$$

where D is a positive constant and $\lambda_{min}(\boldsymbol{X}_{\hat{J}_m\cup\mathcal{J}}^{\mathrm{T}}\boldsymbol{X}_{\hat{J}_m\cup\mathcal{J}})$ is the smallest eigenvalue of $\boldsymbol{X}_{\hat{J}_m\cup\mathcal{J}}^{\mathrm{T}}\boldsymbol{X}_{\hat{J}_m\cup\mathcal{J}}$. For a sufficiently large n, we obtain that

$$n^{-1}\left\|\left(\boldsymbol{I}-\boldsymbol{H}_{\hat{J}_m}\right)\boldsymbol{u}\right\|^2 \ge \frac{D\log(n)}{n} \ge a\tilde{\xi}C\sigma_U\frac{(\log n)^{1/2}}{n}. \qquad \text{(A.27)}$$

Since we already select more than K_n variables using the sample GOGA, this implies that $\{d_H(\hat{J}_m,\mathcal{J}) > \log(n)\}\subseteq F_n$. Thus,

$$\lim_{n\to\infty} P(d_H(\hat{J}_{K_n},\mathcal{J}) \le \log(n)) = 1.$$

□

ACKNOWLEDGMENTS

We would like to thank an Associate Editor and two anonymous referees for their thoughtful and useful comments, which led to an improved version of this article. Research supported in part by HKSAR-RGC-GRF Nos: 400313, 14300514 and HKSAR-RGC-CRF: CityU8/CRG/12G (Chan); Academia Sinica Investigator Award (Ing), HKSAR-RGC-ECS: 405012 and HKSAR-RGC-GRF: 405113, 14601015 (Yau). Part of this research was conducted while N. H. Chan was visiting Renmin University of China (RUC) and Southwestern University of Finance and Economics (SWUFE). Research supported by the School of Statistics at RUC and SWUFE is gratefully acknowledged.

[Received February 2015. Revised June 2015.]

REFERENCES

Bickel, P., Ritov, Y., and Tsybakov, A. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *Annals of Statistics*, 4, 1705–1732. [335]

Buhlmann, P., Kalisch, M., and Maathuis, M. (2009), "Variable Selection for High-Dimensional Models: Partially Faithful Distributions and the pc-simple Algorithm," *Biometrika*, 97, 1–19. [335]

Chan, K. S. (1993), "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model," *Annals of Statistics*, 21, 520–533. [334]

Chan, K. S., and Tsay, R. S. (1998), "Limiting Properties of the Least Squares Estimator of a Continuous Threshold Autoregressive Model," *Biometrika*, 85, 413–426. [336]

Chan, N. H., Yau, C. Y., and Zhang, R. (2015), "Lasso Estimation for Threshold Autoregressive Models," *Journal of Econometrics*, 189, 285–296. [334,335,336,337,338,339,341]

Cho, H., and Fryzlewicz, P. (2012), "High-Dimensional Variable Selection Via Tilting," *Journal of Royal Statistical Society*, Series B, 74, 593–622. [335]

Coakley, J., Fuertes, A.-M., and Pérez, M.-T. (2003), "Numerical Issues in Threshold Autoregressive Modeling Of Time Series," *Journal of Economic Dynamics and Control*, 27, 2219–2242. [334]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407–499. [335,336]

Fan, J., and Li, R. (2001), "Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of American Statistical Association*, 96, 1348–1360. [335]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of Royal Statistical Society*, Series B, 70, 849–911. [335]

Gonzalo, J., and Pitarakis, J.-Y. (2002), "Estimation and Model Selection Based Inference in Single and Multiple Threshold Models," *Journal of Econometrics*, 110, 319–352. [334]

Ing, C. K., and Lai, T. L. (2011), "A Stepwise Regression Method and Consistent Model Selection for High-Dimensional Sparse Linear Models," *Statistica Sinica*, 21, 1473–1513. [335,336,337]

Li, D., and Ling, S. (2012), "On the Least Squares Estimation of Multiple-Regime Threshold Autoregressive Models," *Journal of Econometrics*, 167, 240–253. [334,341]

Li, D., Ling, S., and Zhang, R. (2016), "On a Threshold Double Autoregressive Model," *Journal of Business and Economic Statistics*, 34, 68–80. [334]

Temlyakov, V. (2000), "Weak Greedy Algorithms," *Advances in Computational Mathematics*, 12, 213–227. [342]

Tong, H. (1978), "On a Threshold Model," in *Pattern Recognition and Signal Processing. NATO ASI Series E: Applied Sc. (Vol. 29)*, eds. C. Chen, Oxford: Oxford University Press, pp. 575–586. [334]

——— (1990), *Non-linear Time Series: a Dynamical System Approach*. Oxford: Oxford University Press. [334]

——— (2011), "Threshold Models in Time Series Analysis–30 Years On," *Statistics and its Interface*, 4, 107–118. [334]

Tsay, R. S. (1989), "Testing and Modeling Threshold Autoregressive Processes," *Journal of the American Statistical Association*, 84, 231–240. [334]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [335]