Statistical Papers
https://doi.org/10.1007/s00362-023-01472-7

REGULAR ARTICLE

Active-set based block coordinate descent algorithm in group LASSO for self-exciting threshold autoregressive model

Muhammad Jaffri Mohd Nasir[1] · **Ramzan Nazim Khan**[2] · **Gopalan Nair**[2] · **Darfiana Nur**[3]

Received: 16 September 2022 / Revised: 3 May 2023

Abstract

Group LASSO (gLASSO) estimator has been recently proposed to estimate thresholds for the *self-exciting* threshold autoregressive model, and a group least angle regression (*gLAR*) algorithm has been applied to obtain an approximate solution to the optimization problem. Although *gLAR* algorithm is computationally fast, it has been reported that the algorithm tends to estimate too many irrelevant thresholds along with the relevant ones. This paper develops an *active-set* based block coordinate descent (*aBCD*) algorithm as an exact optimization method for gLASSO to improve the performance of estimating relevant thresholds. Methods and strategy for choosing the appropriate values of shrinkage parameter for gLASSO are also discussed. To consistently estimate relevant thresholds from the threshold set obtained by the gLASSO, the backward elimination algorithm (*BEA*) is utilized. We evaluate numerical efficiency of the proposed algorithms, along with the Single-Line-Search (*SLS*) and the *gLAR*

Ramzan Nazim Khan, Gopalan Nair, and Darfiana Nur have contributed equally to this work.

✉ Muhammad Jaffri Mohd Nasir
jaffri.mn@umk.edu.my

Ramzan Nazim Khan
nazim.khan@uwa.edu.au

Gopalan Nair
gopalan.nair@uwa.edu.au

Darfiana Nur
darfiana.nur@curtin.edu.au

[1] Faculty of Entrepreneurship and Business, Universiti Malaysia Kelantan, 16100 Kota Bharu, Kelantan, Malaysia

[2] Department of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, Perth, WA 6009, Australia

[3] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Kent Street, Perth, WA 6102, Australia

algorithms through simulated data and real data sets. Simulation studies show that the *SLS* and *aBCD* algorithms have similar performance in estimating thresholds although the latter method is much faster. In addition, the *aBCD-BEA* can sometimes outperform *gLAR-BEA* in terms of estimating the correct number of thresholds under certain conditions. The results from case studies have also shown that *aBCD-BEA* performs better in identifying important thresholds.

Keywords Karush–Kuhn–Tucker · Group LASSO · SETAR · *aBCD* algorithm · *BEA* · Sparsity conditions

1 Introduction

The $(m+1)$-regime threshold autoregressive (TAR) model of order p, or TAR(p) for the time series $\{y_t, t = 1, \cdots, n\}$, is defined as

$$y_t = \sum_{j=1}^{m+1} \left(\phi_0^{(j)} + \sum_{i=1}^{p} \phi_i^{(j)} y_{t-i} \right) I_{\mathcal{R}_j}(s_t) = \sum_{j=1}^{m+1} \boldsymbol{x}_t^T \boldsymbol{\phi}_j I_{\mathcal{R}_j}(s_t) + \varepsilon_t, \tag{1}$$

$$\varepsilon_t = \sigma \eta_t, \qquad \eta_t \overset{iid}{\sim} D(0,1), \quad t = p+1, \cdots, n, \tag{2}$$

where $\boldsymbol{x}_t^T = (1, y_{t-1}, y_{t-2}, \cdots, y_{t-p})$, $\boldsymbol{\phi}_j = (\phi_0^{(j)}, \phi_1^{(j)}, \cdots \phi_p^{(j)})^T$ is the set of parameters for regime j, $\mathcal{R}_j = (r_{j-1}, r_j]$ are the threshold intervals with conventions of $r_0 = -\infty$, $r_{m+1} = \infty$ and $\mathcal{R}_{m+1} = (r_m, \infty)$, the indicator function $I_{\mathcal{R}_j}(s_t) = 1$, if $s_t \in \mathcal{R}_j$, zero otherwise and $D(0, 1)$ is a distribution with zero mean and unit variance. Here, $\{s_t, t = p+1, \cdots, n\}$ is the threshold process (sometimes referred to as a switching variable), which controls the switching or jump between the regimes. It follows that the error term $\varepsilon_t, t = p+1, \cdots, n$, are independent and identically distributed with $E(\varepsilon_t) = 0$ and a constant variance $\text{Var}(\varepsilon_t) = \sigma^2$. In this paper, we assume $s_t = y_{t-d}$, where the integer $0 < d \leq p$ is called a delay parameter, and this subclass is called a *self-exciting* TAR (or SETAR) model.

The TAR model was initially proposed by Tong (1978) and several of the TAR sub-classes, including the *self-exciting* are discussed by Tong and Lim (1980) and Tong (1990). The TAR is an AR(p) model in each of several regimes. As such, it is a piecewise model which is linear in each regime, but the overall time series process is non-linear. The piecewise nature of the TAR model is able to capture some important non-linear phenomena, such as sudden jumps, asymmetric limit cycles and chaos, sub and higher harmonics, and amplitude dependent frequency (Tong and Lim 1980; Tong 1990). Since TAR is a piecewise linear extension of a linear AR model, its interpretation is simple and similar to the interpretation of linear models (Li and Ling 2012).

Estimation of SETAR model involves the determination of the number of regimes, thresholds, delay parameter and model order (Chen et al. 2011). The estimation procedure is usually complicated and can be computational costly, despite the well-established asymptotic theory of the SETAR model estimation via least-squares (LS) and maximum likelihood (ML) estimators; for examples, see Chan (1993), Qian (1998) and Li and Ling (2012).

Since the LS and ML functions are discontinuous in d and r_j, $j = 1, \cdots, m$, obtaining global minimum for the LS and global maximum for the ML require a multi-parameter grid search over all possible values of the r_js and d, which is computationally cumbersome, if not impossible, for large m (Li and Ling 2012; Chan et al. 2017). If d is assumed to be known, then the computational cost to estimate all m thresholds via the grid search is $O(n^m)$ (Bai and Perron 2003; Li and Ling 2012).

Some alternative techniques have been proposed to speed-up the thresholds estimation time. For $m = 1$, Li and Tong (2016) developed the nested sub-sample search (NeSS), which drastically reduces the computational cost of one-dimensional grid search algorithm for two-regime threshold models, from $O(n)$ to $O(\log n)$. For the case of unknown m, Gonzalo and Pitarakis (2002) proposed sequential estimation procedure to estimate multiple thresholds, which has linear computational cost $O(mn)$ and requires only a one-dimensional grid-search algorithm for estimating each threshold one at a time. Recently, Chan et al. (2015) proposed a fast approximation algorithm called group least angle regression (*gLAR*) for the group least absolute shrinkage and selection operator (gLASSO) estimator to locate and estimate relevant change-points for a reformulated SETAR model, which then used as a proxy for estimating thresholds. However, it was reported in Chan et al. (2017) that the *gLAR* suffers from estimating excessive irrelevant change-points/thresholds even after performing the additional step of threshold filtration procedure.

gLASSO is a type of regularization method which is a natural extension of the standard LASSO (Yuan and Lin 2006; Nardi and Rinaldo 2008). Unlike the standard LASSO which penalizes individual parameters, gLASSO imposes a penalty on the ℓ_2-norm of the set of model parameters, in order to obtain a group-wise sparse parameter estimate. Furthermore, gLASSO penalizes all sets of parameters at the same rate without evaluating the importance of each of them. Thus, it tends to overpenalize large coefficients. Despite being able to perform parameter estimation and model selection simultaneously, gLASSO has notable drawback of estimation inefficiency and selection inconsistency similar to that of the standard LASSO, if certain *sparsity conditions* are not met (Wang and Leng 2008; Bach 2008; Nardi and Rinaldo 2008).

Some differences between *gLAR* and the gLASSO are described as follows. First, gLASSO uses a set of values of shrinkage parameter λ_n along the solution path while *gLAR* computes the entire path of solutions without evaluating each value of λ_n. Second, if the design matrix of a model is not orthonormal or there is more than one covariate in a group, the path solution of gLASSO is not piecewise-linear while the path solution of *gLAR* is a piecewise-linear. Third, *gLAR* uses the average squared correlation between a group of covariates and the current residual for adding covariates into a model while gLASSO evaluates Karush–Kuhn–Tucker (KKT) conditions for the same purpose. Fourth, *gLAR* lacks a covariate removal procedure while gLASSO might remove some of covariates during the evaluation of KKT conditions (Yuan and Lin 2006; Roth and Fischer 2008; Yau and Hui 2017).

In this paper, we propose an exact optimization algorithm for the gLASSO, called the *active-set* based block coordinate descent (*aBCD*) as an alternative to *gLAR* algorithm in order to improve the estimation performance of change-points for the reformulated SETAR model. A similar algorithm known as the Single-Line-Search (SLS) has been applied by Foygel and Drton (2010) for linear regression without the use of the *active-set* strategy developed by Roth and Fischer (2008). However, they indicated that including an active-set strategy in the algorithm is a possible extension and could improve the computational time. In our change-point problem, the SLS algorithm is ineffective in controlling the estimation number of change-points due to the high-dimensionality and its behavior of cycling through all groups of parameters for each iteration causing higher computational time. On the other hand, the active-set strategy in our *aBCD* algorithm enables us to monitor and assert control over the estimation of the number of change-points up to a predetermined upper bound.

In addition, our gLASSO criteria for the change-point model in this study is a modified version of one given in Foygel and Drton (2010) and Chan et al. (2015), and we implemented a non-derivative approach of bisection method in our algorithm as an alternative to Newton's method suggested by Foygel and Drton (2010) for the root search approximation in gLASSO. Methods and strategy for choosing the appropriate values of shrinkage parameter for gLASSO are also discussed. Monte Carlo simulation and case studies compare the estimation performance between the *aBCD* and *gLAR* approaches.

Throughout this paper, we denote the true parameters with a superscript 0 and their estimates parameter with circumflex "hat" symbol on top. In particular, r_j^0 and $\widehat{r}_j$ denote the true and estimated jth thresholds, respectively; t_j^0 and $\widehat{t}_j$ denote the true and estimated jth change-points, or the location of jth thresholds, respectively; m^0 and $\widehat{m}$ denote the true and estimated number of thresholds, respectively; $\boldsymbol{\phi}_{j'}^0$ and $\widehat{\boldsymbol{\phi}}_j$ denote the respective true and estimated set of parameters, for $j^{'} = 1, 2, \cdots, m^0$ and $j = 1, 2, \cdots, \widehat{m}$; $\mathfrak{T}^0$ and $\widehat{\mathfrak{T}}$ denote the respective set of true and estimated change-points; and $\mathfrak{R}^0$ and $\widehat{\mathfrak{R}}$ denote the respective set of true and estimated thresholds. The notations $\otimes$ and I_p denote respectively, the Kronecker product operator and $(p \times p)$ identity matrix.

This paper is organized as follows. The transformation of SETAR model into a change-point model is detailed in Sect. 2. In Sect. 3, we formulate the group LASSO for the reformulated SETAR model. Discussion on main assumptions and theoretical results are given in Sect. 4. In Sect. 5, computational algorithms and post-analysis procedures are given to estimate the SETAR model. Performance of exact and approximation gLASSO algorithms is evaluated through empirical studies in Sects. 6 and 7. Final remarks are given in Sect. 8.

2 SETAR as change-point model

As stated by Hansen (2000), a threshold model is very similar to a change-point model, except the structural change of data occurs along the observation of the threshold process instead of sampling index. Thus, the threshold variable s_t plays the role of the time index t. If the threshold variable takes a set of discrete values, the TAR parameters can be estimated by first sorting the observations in ascending order of the observations of the threshold process, and subsequently applying well-known methods for change-point model.

Tsay (1989, 1998) and Bai and Perron (2003) proposed an algorithm to convert threshold model estimation into a change-point estimation problem using a particular sorting procedure known as *arranged autoregression*, which is commonly applied in both frequentist (Coakley et al. 2003; Chan et al. 2004) and Bayesian (Chen 1995; Pan et al. 2017) analyses. Under this procedure, the structure of threshold model remains unaffected despite the arrangement of threshold observations (Tsay 1998; Bai and Perron 2003). The main benefits of performing arranged autoregression is it simplifies the process of estimating thresholds by arranging and constraining possible positions of the thresholds so that observations can be appropriately grouped and separated into their respective regimes (Li and Ling 2012).

For the SETAR model, let $\mathbf{y} = (y_{p+1}, y_{p+2}, \cdots, y_n)^T$ and $\mathbf{y}_d = (y_{p+1-d}, y_{p+2-d}, \cdots, y_{n-d})^T$. Let $(y_{\pi_1}, y_{\pi_2}, \cdots, y_{\pi_N})^T$ be the order statistics of the observations in $\mathbf{y}_d$, where π_i is the original index of the ith smallest observations in $\mathbf{y}_d$ and $N := n - p$ is called an *effective sample size*. Then the vector $\mathbf{y}_\pi := (y_{\pi_1+d}, y_{\pi_2+d}, \cdots, y_{\pi_N+d})^T$ is the column vector of rearranged elements of $\mathbf{y}$, with $y_{\pi_1} \le y_{\pi_2} \le \cdots \le y_{\pi_N}$. Note that this procedure also works well for observations with tied values. The arranged autoregression data can also be expressed in a matrix form (Coakley et al. 2003) and a spread sheet form (Chan et al. 2004), which are quite useful for the estimation procedure (see Section 2.1.1 in Nasir (2020) for more details).

To understand how a SETAR can be reformulated into a change-point model, consider the following linear regression framework (Bai and Perron 2003; Qian and Su 2016),

$$y_{\pi_t+d} = \omega_{0,\pi_t} + \sum_{i=1}^{p} \omega_{i,\pi_t} y_{\pi_t+d-i} + \varepsilon_{\pi_t+d} = \mathbf{x}_{\pi_t}^T \boldsymbol{\omega}_{\pi_t} + \varepsilon_{\pi_t+d}, \quad t = 1, \cdots, N, \tag{3}$$

where $\boldsymbol{\omega}_{\pi_t} = (\omega_{0,\pi_t}, \omega_{1,\pi_t}, \cdots, \omega_{p,\pi_t})^T$ is a vector of unknown parameters and $\mathbf{x}_{\pi_t}^T = (1, y_{\pi_t+d-1}, y_{\pi_t+d-2}, \cdots, y_{\pi_t+d-p})$. For linking SETAR with (3), set

$$\boldsymbol{\omega}_{\pi_t} = \boldsymbol{\phi}_j = (\phi_0^{(j)}, \phi_1^{(j)}, \cdots, \phi_p^{(j)})^T \in \mathbb{R}^{p+1}$$

for $t = t_{j-1}, \cdots, t_j - 1$ and $j = 1, 2, \cdots, m+1$, with the conventions $t_0 = 1$ and $t_{m+1} = N + 1$, where $t_j \in (2, \cdots, N)$, for $j = 1, \cdots, m$, is the jth change or *change-point* parameter in (3), satisfying $y_{\pi_{t_{j-1}}} \le r_j < y_{\pi_{t_j}}$. Under these settings, (3) is referred to as a *partial change-point* model (Bai and Perron 2003). In this setup, one need to estimate the set of change-points $\mathfrak{T} = \{t_1, t_2, \cdots, t_m\}$, the number of thresholds m, and the regression coefficients $\boldsymbol{\omega}_{\pi_t}$, for $t \in \mathfrak{T}$.

By the definition of $\boldsymbol{\omega}_{\pi_t}$, the set of vectors $\{\boldsymbol{\omega}_{\pi_1}^T, (\boldsymbol{\omega}_{\pi_2} - \boldsymbol{\omega}_{\pi_1})^T, \cdots, (\boldsymbol{\omega}_{\pi_N} - \boldsymbol{\omega}_{\pi_{N-1}})^T\}^T$ exhibits a groupwise *sparse characteristic* in the sense that it contains only $(m+1)$ nonzero vectors, corresponding to the number of regimes in the SETAR model. From the sparse characteristic, one can easily locate the change-points by identifying the non-zero vectors in the set. Indeed, if $\boldsymbol{\omega}_{\pi_i} - \boldsymbol{\omega}_{\pi_{i-1}} \neq \mathbf{0}$, for some $i \geq 2$, then i is a change-point.

Let $\boldsymbol{\theta}^N := (\boldsymbol{\theta}_{\pi_1}^T, \boldsymbol{\theta}_{\pi_2}^T, \cdots, \boldsymbol{\theta}_{\pi_N}^T)^T = (\boldsymbol{\omega}_{\pi_1}^T, (\boldsymbol{\omega}_{\pi_2} - \boldsymbol{\omega}_{\pi_1})^T, \cdots, (\boldsymbol{\omega}_{\pi_N} - \boldsymbol{\omega}_{\pi_{N-1}})^T)^T$ be the transformed $N(p+1)$-dimensional row vector of parameters, in which only $(m+1)$ of the vectors $\boldsymbol{\theta}_{\pi_i}$ are non-zero. Then, (3) can be expressed as

$$y_{\pi_t+d} = \mathbf{x}_{\pi_t}^T \sum_{k=1}^{t} \boldsymbol{\theta}_{\pi_k} + \varepsilon_{\pi_t+d}, \quad \text{for } t = 1, 2, \cdots, N. \tag{4}$$

Since $\boldsymbol{\theta}^N$ is groupwise sparse, we express (4) as

$$y_{\pi_t+d} = \mathbf{x}_{\pi_t}^T \sum_{k \in \left\{i : \boldsymbol{\theta}_{\pi_i} \neq \mathbf{0}, i \leq t\right\}} \boldsymbol{\theta}_{\pi_k} + \varepsilon_{\pi_t+d} \tag{5}$$

to highlight the benefit of lower computational cost.

Define $\mathrm{I}_\Delta = \mathbf{1}_\Delta \otimes \mathrm{I}_{p+1}$ as a $N(p+1) \times N(p+1)$ block triangular matrix, where $\mathbf{1}_\Delta$ is an $(N \times N)$ lower triangular matrix of ones. Then the design matrix,

$$\begin{aligned} X = \begin{bmatrix} \mathbf{x}_{\pi_1}^T & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{x}_{\pi_2}^T & \mathbf{x}_{\pi_2}^T & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{x}_{\pi_3}^T & \mathbf{x}_{\pi_3}^T & \mathbf{x}_{\pi_3}^T & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{x}_{\pi_N}^T & \mathbf{x}_{\pi_N}^T & \mathbf{x}_{\pi_N}^T & \cdots & \mathbf{x}_{\pi_N}^T \end{bmatrix} &= \begin{bmatrix} \mathbf{x}_{\pi_1}^T & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{\pi_2}^T & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{x}_{\pi_3}^T & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{x}_{\pi_N}^T \end{bmatrix} \mathbf{I}_\Delta \\ := \left[X_{\pi,1} \; X_{\pi,2} \; X_{\pi,3} \cdots, X_{\pi,N} \right] & \end{aligned} \tag{6}$$

is a $N \times N(p+1)$ block lower triangular matrix, where $X_{\pi,k}$, is the kth block of covariates, for $k = 1, \cdots, N$. Then (4) can be written in the following high-dimensional sparse regression form:

$$\mathbf{y}_\pi = X\boldsymbol{\theta}^N + \boldsymbol{\varepsilon}_\pi. \tag{7}$$

The regression setting (7) is similar to the high-dimensional regression model for change-point problem in Chan et al. (2014) and Qian and Su (2016), except that the samples being considered here are the effective samples. Relations between the parameters in (1) and (7) can be expressed as

$$\boldsymbol{\theta}_{\pi_i} = \begin{cases} \boldsymbol{\omega}_{\pi_1} = \boldsymbol{\phi}_1, & \text{for } i = 1, \\ \boldsymbol{\omega}_{\pi_i} - \boldsymbol{\omega}_{\pi_{i-1}} = \boldsymbol{\phi}_{j+1} - \boldsymbol{\phi}_j \neq \mathbf{0}, & \text{for } i = t_j \geq 2 \; andj = 1, \cdots, m, \\ \boldsymbol{\omega}_{\pi_i} - \boldsymbol{\omega}_{\pi_{i-1}} = \mathbf{0}, & \text{for } i \in \{2, \cdots, N\} \setminus \{t_1, \cdots, t_m\}. \end{cases} \tag{8}$$

Note that $\sum_{i=1}^{t_j} \boldsymbol{\theta}_{\pi_i} = \boldsymbol{\phi}_{j+1}$.

3 Penalized estimation methods

In this paper, we aim to estimate $\boldsymbol{\theta}^N$ by solving the following penalized LS objective/loss function:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^N &= \arg\min_{\boldsymbol{\theta}^N} f(\boldsymbol{\theta}^N)\\ &:= \arg\min_{\boldsymbol{\theta}^N} \left(\frac{1}{N} \sum_{t=1}^{N} \left(y_{\pi_t+d} - \mathbf{x}_{\pi_t}^T \sum_{k=1}^{t} \boldsymbol{\theta}_{\pi_k} \right)^2 + \lambda_n \sum_{i=2}^{N} \left\| \boldsymbol{\theta}_{\pi_i} \right\|_2 \right). \end{aligned} \tag{9}$$

Note that (9) is a gLASSO optimization problem (Yuan and Lin 2006) for estimating multiple changes of parameter vectors such that $\widehat{\boldsymbol{\theta}}_{\pi(i)} \neq \mathbf{0}$, for $i \geq 2$, given an appropriate selection of the shrinkage parameter λ_n. Furthermore, the optimization (9) is similar to Equation (2.4) in Qian and Su (2016) for regression models with multiple structural breaks. Due to the convexity of (9), any local minimizer for this function is also a global minimizer, and convex optimizations methods are feasible for minimizing (9). However, multiple solutions for $\widehat{\boldsymbol{\theta}}^N$ may exist as (9) may not be strictly convex when the least squares estimator is not uniquely defined (e.g., when X is linearly dependent) (Osborne et al. 2000; Huang et al. 2012; Tibshirani 2013).

It is also worth mentioning that (9) differs from those proposed by Harchaoui and Lévy-Leduc (2010) and Chan et al. (2014) for change-point estimation, and also Chan et al. (2015) for threshold estimation, since the vector of parameters $\boldsymbol{\theta}_{\pi_1}$ is not penalized, as $t_0 = 1$ is not a candidate for a change-point in our study.

After obtaining $\widehat{\boldsymbol{\theta}}^N$, the set of estimated change-points are given by $\widehat{\mathfrak{T}} := \{t : \widehat{\boldsymbol{\theta}}_{\pi_t} \neq \mathbf{0}, t \geq 2\} = \left\{\widehat{t}_1, \widehat{t}_2, \cdots, \widehat{t}_{\widehat{m}}\right\}$, where $\widehat{m} = \text{card}(\widehat{\mathfrak{T}})$ is the estimated number of change-points. Subsequently, the set of the estimated thresholds are given as $\widehat{\mathfrak{R}} = \{\widehat{r}_1, \widehat{r}_2, \cdots, \widehat{r}_{\widehat{m}}\} := \{y_{\pi_{\widehat{t}_1^*}}, y_{\pi_{\widehat{t}_2^*}}, \cdots, y_{\pi_{\widehat{t}_{\widehat{m}}^*}}\}$, where $\widehat{t}_j^* := \widehat{t}_j - 1$, for $j = 1, 2, \cdots, \widehat{m}$. By the close relationship between (1) and (7), the estimated autoregressive parameters for all regimes can be retrieved as $\widehat{\boldsymbol{\phi}}_1 = \widehat{\boldsymbol{\theta}}_{\pi_1}$, and $\widehat{\boldsymbol{\phi}}_{j+1} = \sum_{i=1}^{\widehat{t}_j} \widehat{\boldsymbol{\theta}}_{\pi_i}$, for $j = 1, \cdots, \widehat{m}$. Algorithm-wise, coordinate descent method is also feasible for optimizing (9) due to its convexity.

For gLASSO to be consistent in selection of relevant groups, it is necessary for the design matrix X to satisfy the groupwise *irrepresentable condition*, which requires that any of relevant group of covariates is weakly correlated with any irrelevant group of covariates (Bach 2008). In the case of (7), observe that these following three consecutive blocks of covariates X_{π,t_j^0-1}, X_{π,t_j^0} and X_{π,t_j^0+1} given in (6), differ only in one row. Thus, any block corresponding to the index t_j has very high correlation with the adjacent irrelevant blocks. Furthermore, Harchaoui and Lévy-Leduc (2010) showed that similar design matrix to X with $p = 0$ does not satisfy the *irrepresentable condition* of Zhao and Yu (2006). In conclusion, a perfect estimation of number of thresholds, e.g., ($\widehat{m} = m^0$) is not possible under (9) for any λ_n under finite sample size. Since there is a possibility of overestimating m, a post-analysis is discussed in Sect. 5.3 to obtain a consistent estimator of it.

4 Assumptions and asymptotic properties

In this section, some common assumptions and conditions are stated for the consistency of estimators for SETAR parameters using gLASSO.

For $j = 1, 2, \cdots, m^0 + 1$, define $d^t_j = t^0_j - t^0_{j-1}$ and $d^r_j = r^0_j - r^0_{j-1}$. Let $d^t_{\min} = \min\limits_{1\le j\le m^0+1} \left\| d^t_j \right\|$, $d^r_{\min} = \min\limits_{1\le j\le m^0+1} \left\| d^r_j \right\|$ and $d^{\boldsymbol{\phi}}_{\min} = \min\limits_{1\le j\le m^0} \left\| \boldsymbol{\phi}^0_{j+1} - \boldsymbol{\phi}^0_j \right\|_2$.

Here, $d^t_{\min}$ denotes the minimum interval length of the regime, $d^r_{\min}$ denotes the minimum distance of two consecutive thresholds, and $d^{\boldsymbol{\phi}}_{\min}$ denotes the minimum ℓ_2-distance between consecutive parameter vectors of SETAR.

4.1 Assumptions

To establish the asymptotic theory, we impose the following assumptions.

HA1 $\{\eta_t\}$ is a sequence of real valued *independent and identically distributed random variables* with bounded, continuous and positive density, $E(\eta_t) = 0$ and $E(|\eta_t|)^{2+\tau} < \infty$, for some $\tau > 0$.

HA2 $\{y_t\}$ is a *α-mixing stationary process with geometric decaying rate* with $E(|y_t|)^{2+\tau} < \infty$.

HA3 $\{\gamma_n\}$ is a *positive and decreasing sequence* converging to zero as $n \to 0$, and satisfies $\gamma_n \ge c_* \log(N)^{(2+\tau)/\tau}/N$ for some $c_* > 0$, $N\gamma_n (d^{\boldsymbol{\phi}}_{\min})^2/(\log N) \to \infty$ and $d^r_{\min}/\gamma_n \to \infty$.

HA4 (a) $d^{\boldsymbol{\phi}}_{\min} > \upsilon_*$, for some $\upsilon_* > 0$, and (b) $m^0 < m_{\max}$, an upper bound of the number of thresholds.

HA5 The sequence of non-negative *regularization parameter* $\{\lambda_n\}$ satisfies $\lambda_n / d^{\boldsymbol{\phi}}_{\min}\gamma_n \to 0$, as $n \to \infty$.

HA6 $d^t_{\min}/N\gamma_n \to \infty$ as $n \to \infty$.

HA1–HA4 are the standard assumptions for the stability and the estimation of threshold autoregressive models, similar to those in Chan et al. (1985), Chan (1993) and Li and Ling (2012). For example, HA2 is satisfied if HA1 holds and either all roots of the polynomial $1 - \sum_{i=1}^p \phi^{(j)}_i z^i$ are outside the unit circle or $\sup_j \sum_{i=1}^p \left|\phi^{(j)}_i\right| < 1$, for each $j = 1, \cdots, m^0 + 1$. For $p = 1$, the following conditions $\phi^{(1)}_1 < 1$, $\phi^{(m+1)}_1 < 1$, $\phi^{(1)}_1 \phi^{(m+1)}_1 < 1$, or $\phi^{(1)}_0 > 0$, $\phi^{(1)}_1 = 1$, $\phi^{(m+1)}_1 < 1$, or $\phi^{(1)}_0 < 0$, $\phi^{(1)}_1 < 1$, $\phi^{(m+1)}_1 = 1$ implies that the time series is stationary and ergodic. Furthermore, strong mixing property such as α-mixing in HA2 implied that the past and distance future observations are asymptotically independent (Fan and Yao 2003; Tsay and Chen 2018).

The sequence $\{\gamma_n\}$ in HA3 controls the rate at which $\widehat{r}_j$ converges to r^0_j when the number of thresholds is correctly estimated. For example, if m^0 is known, and r^0_j is fixed, then the threshold estimator $\widehat{r}_j$ is found to be n-consistent (Qian 1998; Li and Ling 2012; Chan et al. 2015) and thus $\gamma_n = O(1/n)$. However, if m^0 and r_j are unknown and they are estimated by gLASSO via the reformulated SETAR, then $\gamma_n = (\log N)^{(2+\tau)/\tau}/N$, a much slower rate (Harchaoui and Lévy-Leduc 2010; Chan et al. 2015; Qian and Su 2016). Furthermore, HA3 requires that the minimum distance between two consecutive thresholds is bigger than γ_n (Chan et al. 2015).

HA4 (a) is necessary to ensure that all thresholds are identified by considering the changes in AR parameters. Furthermore, it plays an important role in obtaining the n-convergence rate of thresholds and its limiting (asymptotic) distribution of the threshold estimator when the number of thresholds is correctly estimated or known (Qian 1998; Li and Ling 2012; Chan et al. 2015). HA4 (b) bounds the true number of thresholds m^0 to its upper limit $m_{\max}$ for a consistent estimation of change-points/thresholds (Gonzalo and Pitarakis 2002; Qian and Su 2016; Yau et al. 2015). Note that m^0 may be allowed to increase at the (slow) rate of $O(\log(n))$ or at a much faster rate (Chan et al. 2015; Qian and Su 2016).

HA5 provides condition for λ_n, which depends on $d_{\min}^{\boldsymbol{\phi}}$ and γ_n. By choosing $\lambda_n = (\log N)/N$ and $d_{\min}^{\boldsymbol{\phi}} \geq (\log N)^{1/4}$, the assumptions HA3, HA4 (a), HA5 and HA6 are satisfied, leading to the convergence rate of $\frac{(\log N)^{(2+\tau)/\tau}}{N}$ in estimating $\widehat{r}_j$. With this choice, we can obtain an almost optimal rate of $1/n$ for the estimation of $\widehat{r}_j$ up to the logarithmic factor (Chan 1993; Li and Ling 2012).

Finally, HA6 is required to satisfy LASSO-type conditions such as *incoherent design*, or the *restrictive eigenvalue condition* (Nardi and Rinaldo 2008; Bickel et al. 2009), so that $\left\|\widehat{\boldsymbol{\theta}}^N - \boldsymbol{\theta}^{0N}\right\|_2 \rightarrow_p 0$ as $n \rightarrow \infty$. For example, Harchaoui and Lévy-Leduc (2010) proved that, if the distance between two consecutive non-zero parameters is equal to one, $t_k - t_l = 1$, for all k and l such that $k - l = 1$, then the *incoherent design* is not satisfied since $\liminf_{n\rightarrow\infty} \phi_{\min}(s_n \log n) \leq \frac{1}{n} \rightarrow 0$. This assumption implies that the difference between two consecutive change-points cannot be too close, and the distance is at least larger than $N\gamma_n$, and tends to infinity at different rates, as $n \rightarrow \infty$. Harchaoui and Lévy-Leduc (2010) assumed that $d_{\min}^t \geq N\gamma_n$.

4.2 Asymptotic properties

The consistency of gLASSO estimator in terms of prediction error, estimating thresholds and other model parameters are presented here. First, the following lemma provides the derivatives of $f(\boldsymbol{\theta}^N)$ which is useful for proving theoretical results and developing exact optimization for gLASSO in Sect. 5.

Lemma 4.1 Consider the gLASSO problem in (9). Let $\widehat{\boldsymbol{\theta}}^N = (\widehat{\boldsymbol{\theta}}_{\pi_1}^T, \widehat{\boldsymbol{\theta}}_{\pi_2}^T, \cdots, \widehat{\boldsymbol{\theta}}_{\pi_N}^T)^T$ be a solution. Under HA1–HA6, the KKT conditions for the solution (9) are

$$\text{(i)} \quad \sum_{l=\widehat{t}_j}^{N} \mathbf{x}_{\pi_l}\left(y_{\pi_l+d} - \mathbf{x}_{\pi_l}^T \sum_{i=1}^{l} \widehat{\boldsymbol{\theta}}_{\pi_i}\right) = \frac{N\lambda_n}{2} \frac{\widehat{\boldsymbol{\theta}}_{\pi_{\widehat{t}_j}}}{\left\|\widehat{\boldsymbol{\theta}}_{\pi_{\widehat{t}_j}}\right\|_2},$$

$$\text{for } j = 1, \cdots, \widehat{m},\ \widehat{t}_j \geq 2,\ \widehat{\boldsymbol{\theta}}_{\pi_{\widehat{t}_j}} \neq \mathbf{0}, \quad \text{and}$$

$$\text{(ii)} \quad \left\| \sum_{l=j'}^{N} \mathbf{x}_{\pi_l}\left(y_{\pi_l+d} - \mathbf{x}_{\pi_l}^T \sum_{i=1}^{l} \widehat{\boldsymbol{\theta}}_{\pi_i}\right) \right\|_2 \leq \frac{N\lambda_n}{2}, \qquad \text{for } j' = 1, 2, \cdots, N.$$

Furthermore, $\sum_{l=1}^{N} \mathbf{x}_{\pi_l}\left(y_{\pi_l+d} - \mathbf{x}_{\pi_l}^T \widehat{\boldsymbol{\theta}}_{\pi_1}\right) = \mathbf{0}$.

For a proof of this lemma, see the proof of Lemma 3.1.2 in Nasir (2020).

The following result establishes the consistency in prediction or prediction error of gLASSO.

Theorem 4.1 *Under* HA1–HA2, *if* $\lambda_n = 2(p+1)c_0\sqrt{(log\ N)/N}$, *then*

$$P\left(\frac{1}{N}\left\|\boldsymbol{X}\left(\widehat{\boldsymbol{\theta}}^{N}-\boldsymbol{\theta}^{0N}\right)\right\|_2^2 \le b_n\right) \ge 1 - c/\left(4(p+1)^2 c_0^2\ log\ N\right)^{1+\tau/2}, \qquad (10)$$

where $b_n = 2\lambda_n m^0 \underset{j}{max}\left\|\boldsymbol{\phi}_j^0 - \boldsymbol{\phi}_{j-1}^0\right\|_2 + \lambda_n\left\|\widehat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_1^0\right\|_2$, *for some* $c_0 > 2\sqrt{2}$, $c > 0$ *and* $\tau > 0$.

Proof of this theorem is given as a proof of Theorem 3.1.1 in Nasir (2020). Note that this result differs from the result obtained by Harchaoui and Lévy-Leduc (2010) for the change-in mean model, and by Chan et al. (2015) for the reformulated SETAR in their Proposition 1 and Theorem 2.1, respectively, due to the fact that our gLASSO does not penalize $\widehat{\boldsymbol{\theta}}_{\pi_1}$. The rationale is that the lowest index is not a candidate for a change-point. Consequently, the error bound obtained in our result is lower than those obtained by both of the aforementioned studies.

The following theorem establishes the consistency of the estimated thresholds $\widehat{\mathfrak{R}}$ when the number of the estimated thresholds is equal to the number of true thresholds ($\widehat{m} = m^0$).

Theorem 4.2 *Suppose that* HA1–HA6 *are satisfied. If* $\widehat{m} = card(\widehat{\mathfrak{R}}) = m^0$, *then*

$$P\left(\max_{1\le j\le m^0}\left|\widehat{r}_j - r_j^0\right| \le \gamma_n\right) \to 1 \quad as \quad n \to \infty.$$

Proof of this theorem is given as a proof of Theorem 3.1.2 in Nasir (2020), where the author uses similar arguments as in the proofs of Proposition 3 in Harchaoui and Lévy-Leduc (2010), Theorem 2.2 in Chan et al. (2014, 2015) and Theorem 3.1 in Qian and Su (2016). In particular, compared to the proof of Theorem 2.2 in Chan et al. (2015) for SETAR model, Nasir (2020) provided a different proof and with more details, to show that $P\left(\max_{1\le j\le m^0}\left|\widehat{r}_j - r_j^0\right| > \gamma_n\right) \to 0$ as $n \to \infty$. The proof of this theorem relies heavily on the inspection of the KKT conditions in *Lemma* 4.1. It can be shown that if $\left|\widehat{r}_j - r_j^0\right| > \gamma_n$, then gLASSO solutions do not satisfy the KKT conditions and the solutions are not optimal. This theorem also implies that when the sample size is large, the convergence rate of the estimated thresholds can be improved when $\widehat{m} = m^0$ (Qian and Su 2016).

In practice, the true number of thresholds m^0 is usually unknown and this requires different results for the consistency of $\widehat{\mathfrak{R}}$ (Chan et al. 2015). With that, it is shown in the following theorems that the number of estimated thresholds $\widehat{m}$ cannot be lower than the true thresholds m^0, under the HA1–HA6. Moreover, there exist $\widehat{r}_i$ sufficiently close to $r_j^0 \in \mathfrak{R}^0$, for some j, when $\widehat{m} \ge m^0$.

Let

$$d_H(\boldsymbol{A}, \boldsymbol{B}) = \sup_{b \in \boldsymbol{B}} \inf_{a \in \boldsymbol{A}} |a - b| \tag{11}$$

be a one-sided Hausdorff's distance (Boysen et al. 2009), from set $\boldsymbol{B}$ to set $\boldsymbol{A}$, measuring the maximum distance from $\boldsymbol{B}$ to the nearest point in $\boldsymbol{A}$.

Theorem 4.3 *If* HA1–HA6 *hold, then,*

$$P(\widehat{m} \geq m_0) \rightarrow 1, \ \ as\ n \rightarrow \infty.$$

Theorem 4.4 *Suppose that* HA1–HA6 *hold. If $m^0 \leq \widehat{m} = card(\widehat{\mathfrak{R}}) \leq m_{max}$, where m_{max} is the upper bound of the number of thresholds, then*

$$P\left(d_H\left(\widehat{\mathfrak{R}}, \mathfrak{R}^0\right) \leq \gamma_n\right) = P\left(\underset{r_k^0 \in \mathfrak{R}^0}{max}\ \underset{\widehat{r}_j \in \widehat{\mathfrak{R}}}{min} \left\|\widehat{r}_j - r_k^0\right\| \leq \gamma_n\right) \rightarrow 1, \ \ as\ n \rightarrow \infty.$$

Algorithm 1: Active Set - Block Coordinate Descent for Group LASSO of the reformulated SETAR

```
Data: $\mathbf{y}_\pi \in \mathbb{R}^N, \mathbf{x}_{\pi_1} \in \mathbb{R}^{p+1}, \cdots, \mathbf{x}_{\pi_N} \in \mathbb{R}^{p+1}, \lambda_n \geq 0$ , $\Delta_* \geq 0$ and $k_{\max} \geq 1$.
Result: $\widehat{\boldsymbol{\theta}}^N \leftarrow \boldsymbol{\theta}^N$ satisfying (9), and $\mathcal{B}$.
for $j = 1, 2, \cdots, N$, do
    Obtain $U_j$ and $D_j$, from $\sum_{l=j}^N \mathbf{x}_{\pi_l}\mathbf{x}_{\pi_l}^T$ using SVD, such that $\sum_{l=j}^N \mathbf{x}_{\pi_l}\mathbf{x}_{\pi_l}^T = U_j^T D_j U_j$. Write
      $D_j = \text{diag}\left(d_{j,1}, d_{j,2}, \cdots, d_{j,p+1}\right)$.
Initialize: $\boldsymbol{\theta}^N = (\boldsymbol{\theta}_{\pi_1}^T, \boldsymbol{\theta}_{\pi_2}^T, \cdots, \boldsymbol{\theta}_{\pi_N}^T)^T \leftarrow \mathbf{0}$, $\mathcal{B} = \{1\}$ and $\mathcal{B}^* = \{1, 2, \cdots N\} \setminus \{1, \cdots, 1 + \Delta_*\}$.
repeat
    repeat
        foreach $j \in \mathcal{B}$ do
            if $j = 1$ then
                Compute $\boldsymbol{\theta}_{\pi_1} = U_1^T D_1^{-1} U_1 \boldsymbol{f}_1(\mathcal{B})$.
            else
                if $(2\left\|\boldsymbol{f}_j(\mathcal{B})\right\|_2 / N) > \lambda_n$ then
                    Compute $U_j \boldsymbol{f}_j(\mathcal{B}) = (\nu_{j,1}, \nu_{j,2}, \cdots, \nu_{j,p+1})^T$, where $\boldsymbol{f}_j(\mathcal{B})$ is given in
                      (27).
                    Find the unique $u_j > 0$ satisfying (13). Then, compute
                      $\boldsymbol{\theta}_{\pi_j} = U_j^T \left(D_j + \frac{N\lambda_n}{2u_j}\mathrm{I}_{p+1}\right)^{-1} U_j \boldsymbol{f}_j(\mathcal{B})$.
                else
                    Set $\boldsymbol{\theta}_{\pi_j} = \mathbf{0}$.
    until some convergence criterion of parameters is met.
    Update $\mathcal{B} \leftarrow \mathcal{B} \setminus \left\{j \in \mathcal{B} : \boldsymbol{\theta}_{\pi_j} = \mathbf{0}\right\}$.
    Compute $\widetilde{u} = \min(\arg\max_{j' \in \mathcal{B}^*} \left\|\boldsymbol{f}_{j'}(\mathcal{B})\right\|_2)$.
    if $(2\left\|\boldsymbol{f}_{\widetilde{u}}(\mathcal{B})\right\|_2 / N) > \lambda_n$ then
        Update $\mathcal{B} \leftarrow \mathcal{B} \cup \widetilde{u}$ and $\mathcal{B}^* \leftarrow \mathcal{B}^* \setminus \{\widetilde{u} - \Delta_*, \cdots, \widetilde{u} + \Delta_*\}$.
until $(2\left\|\boldsymbol{f}_{\widetilde{u}}(\mathcal{B})\right\|_2 / N) \leq \lambda_n$ or $\text{card}(\mathcal{B}) = k_{\max}$.
```

Proofs of these Theorems 4.3 and 4.4 are given as proof of Theorems 3.1.3 and 3.1.4, respectively, in Nasir (2020). The KKT conditions in Lemma 4.1 are key to the proofs. The proof uses similar arguments as in the proof of Proposition 4 in Harchaoui and Lévy-Leduc (2010), proof of Theorem 2.3 in Chan et al. (2015) and proof of Theorem 3.2 in Qian and Su (2016). Both results are established by contradiction. These theorems imply that if the thresholds are being overestimated, then there will be a threshold which is close to the true threshold when $\widehat{m} \geq m^0$.

5 Algorithms and selection of shrinkage parameter

Following the methods, assumptions and asymptotic properties in the previous sections, we now provide two algorithms for parameter estimation. Firstly, the *aBCD* algorithm for the first-step estimation of group LASSO, and then the backward elimination algorithm (*BEA*) for the post-selection of thresholds.

5.1 Optimization via *aBCD*

Here, we implement the *active-set* strategy (Roth and Fischer 2008) to optimize (9). The main benefit of using this strategy, for the reformulated SETAR model, is that we can monitor and assert control over the estimation of the number of change-points/thresholds up to a upper bound, say $k_{\max}$, since we assume that the true number of change-points/thresholds is fixed and much smaller than the sample size. Particularly, it is designed to discard values of λ_n for which the cardinality of the active-set exceeds $k_{\max}$. Note that when λ_n decreases, the computation time for the *aBCD* algorithm increases, as an increasing number of non-zero group of parameters $\boldsymbol{\theta}_{\pi_i}$ need to be optimized one at a time.

For the reformulated SETAR model (7), the derivative of penalized least square function $f(\boldsymbol{\theta}^N)$, defined in (9), is given by

$$\sum_{l=j}^{N} \mathbf{x}_{\pi_l} \left(y_{\pi_l+d} - \mathbf{x}_{\pi_l}^T \sum_{i=1}^{l} \boldsymbol{\theta}_{\pi_i} \right) = \frac{N\lambda_n}{2} \widetilde{\boldsymbol{e}}_j, \tag{12}$$

for $j = 1, 2, \cdots, N$, where $\widetilde{\boldsymbol{e}}_j$ is the sub-gradient. Let $\mathcal{B}$ and $\mathcal{B}^*$ be two subsets of $\{1, 2, \cdots, N\}$ such that $\mathcal{B} = \{i : \boldsymbol{\theta}_{\pi_i} \neq \mathbf{0}\}$ and $\mathcal{B}^* = \{i : \boldsymbol{\theta}_{\pi_i} = \mathbf{0}\}$. We call $\mathcal{B}$ and $\mathcal{B}^*$ as the active and inactive sets, respectively. For $j = 1, 2, \cdots, N$, we then compute the singular value decomposition (SVD) of the Gram matrix $\sum_{l=j}^{N} \mathbf{x}_{\pi_l} \mathbf{x}_{\pi_l}^T = U_j^T D_j U_j$, where U_j is a $(p+1) \times (p+1)$ orthonormal matrix and $D_j = \operatorname{diag}\left(d_{j,1}, d_{j,2}, \cdots, d_{j,p+1}\right)$ is a $(p+1) \times (p+1)$ invertible diagonal matrix with $d_{j,k}$ as the eigenvalues of the Gram matrix for $k = 1, 2, \cdots, p+1$.

Note that $U_j^T U_j = \mathrm{I}_{p+1}$, $\left\| U_j \right\|_2 = \mu_{\max}(U_j) = 1$, $\left\| U_j \boldsymbol{x} \right\|_2 = \|\boldsymbol{x}\|_2$, for $\boldsymbol{x} \in \mathbb{R}^{p+1}$, and $\left(\sum_{l=j}^{N} \mathbf{x}_{\pi_l} \mathbf{x}_{\pi_l}^T \right)^{-1} = U_j^T D_j^{-1} U_j$, where $\mu_{\max}(.)$ is a maximum eigenvalue of the matrix, and they require the Gram matrices to be well-behaved for the properties to hold. The computations of SVD is not expensive since the decomposition only involves the jth sum of $(p+1)\times(p+1)$ Gram matrices. In addition, these decomposed matrices can be pre-computed once and stored for later use.

The step-by-step procedure to perform *aBCD* for gLASSO is summarized in Algorithm 1. It is worth mentioning that our algorithm differs from the one provided by Chan et al. (2014) in their supplementary material. In the initial estimation step, we set $\mathcal{B} = \{1\}$ and $\mathcal{B}^* = \{1, 2, \cdots N\} \setminus \{1, \cdots, 1 + \Delta_*\}$ and $\boldsymbol{\theta}^N = (\boldsymbol{\theta}_{\pi_1}^T, \boldsymbol{\theta}_{\pi_2}^T, \cdots, \boldsymbol{\theta}_{\pi_N}^T)^T = \mathbf{0}$, where $\Delta_* \geq 0$ be an integer which allows a gap between two estimated change-points (Δ_* is discussed further in Remark 5.1).

The next step is to evaluate the KKT conditions and estimating $\boldsymbol{\theta}_{\pi_j}^T \in \boldsymbol{\theta}^N$, for each $j \in \mathcal{B}$ until convergence. Given $\mathcal{B}$, existence of solution for $\{\boldsymbol{\theta}_{\pi_j}; j \in \mathcal{B}\}$ is given in Theorem 5.1. Once the parameters converge, we remove any index $j \in \mathcal{B}$ which satisfies $\boldsymbol{\theta}_{\pi_j} = \mathbf{0}$. In the final step, we check for any *violation of KKT* for $\left\| \boldsymbol{f}_{j'}(\mathcal{B}) \right\|_2$, $j' \in \mathcal{B}^*$. Specifically, we look for any j' that satisfies $\max(2 \left\| \boldsymbol{f}_{j'}(\mathcal{B}) \right\|_2 /N) > \lambda_n$. Let $\widetilde{u} = \min(\arg\max_{j' \in \mathcal{B}^*} \left\| \boldsymbol{f}_{j'}(\mathcal{B}) \right\|_2)$. If $\left(2 \left\| \boldsymbol{f}_{\widetilde{u}}(\mathcal{B}) \right\|_2 /N\right) > \lambda_n$, then we update $\mathcal{B} \leftarrow \mathcal{B} \cup \widetilde{u}$ and $\mathcal{B}^* \leftarrow \mathcal{B}^* \setminus \{\widetilde{u} - \Delta_*, \cdots, \widetilde{u} + \Delta_*\}$ and the previous steps of the optimization procedure are repeated. The algorithm halted when the conditions $(2 \left\| \boldsymbol{f}_{\widetilde{u}}(\mathcal{B}) \right\|_2 /N) \leq \lambda_n$ or $\mathrm{card}(\mathcal{B}) = k_{\max}$ are met.

Theorem 5.1 *If* $\left(2 \left\| \boldsymbol{f}_j(\mathcal{B}) \right\|_2 /N\right) > \lambda_n$, *then there exist* $u_j > 0$, *for* $j \in \mathcal{B} \setminus \{1\}$ *satisfying the nonlinear equation*

$$g(u_j) = \sum_{k=1}^{p+1} \frac{\nu_{j,k}^2}{\left(d_{j,k} u_j + \frac{N\lambda_n}{2}\right)^2} = 1, \tag{13}$$

where $U_j \boldsymbol{f}_j(\mathcal{B}) = (\nu_{j,1}, \nu_{j,2}, \cdots, \nu_{j,p+1})^T$ *and* $\boldsymbol{f}_j(\mathcal{B}) = \sum_{l=j}^{N} \boldsymbol{x}_{\pi(l)} y_{\pi_l + d} - \boldsymbol{g}_j$. *Furthermore,*

$$\boldsymbol{\theta}_{\pi_j} = \begin{cases} U_j^T \left(D_j + \frac{N\lambda_n}{2u_j} I_{p+1}\right)^{-1} U_j \boldsymbol{f}_j(\mathcal{B}), & \text{for } j \in \mathcal{B} \setminus \{1\} \\ & \quad \text{and } \left(2 \left\| \boldsymbol{f}_j(\mathcal{B}) \right\|_2 /N\right) > \lambda_n, \\ \mathbf{0}, & \text{for } j \in \mathcal{B} \setminus \{1\} \\ & \quad \text{and } \left(2 \left\| \boldsymbol{f}_j(\mathcal{B}) \right\|_2 /N\right) \leq \lambda_n, \\ U_j^T D_j^{-1} U_j \boldsymbol{f}_j(\mathcal{B}), & \text{for } j = 1, \end{cases} \tag{14}$$

where $\boldsymbol{g}_j = \sum_{\substack{i \in \mathcal{B} \\ i \neq j}} \left\{ \sum_{h=max(i,j)}^{N} \boldsymbol{x}_{\pi_h} \boldsymbol{x}_{\pi_h}^T \right\} \boldsymbol{\theta}_{\pi_i}$ *if* $card(\mathcal{B}) > 1$, *otherwise* $\boldsymbol{g}_j = \mathbf{0}$.

The proof of Theorem 5.1 is given in the Appendix, and it implies that conditions (I) and (II) in Lemma 4.1 are satisfied for $j \in \mathcal{B} \setminus \{1\}$. Since the minimizer of gLASSO (9) is convex, the objective function $f(\boldsymbol{\theta}^N)$ will keep decreasing for every iteration and eventually the parameter set $\boldsymbol{\theta}_{\pi_j}$ will converge to global minimum, as shown in Corollary 1 and Theorem 3 of Foygel and Drton (2010). Also, this theorem implies that root search method, such as the Newton–Raphson or bisection, can be used to search u_j. Note that Foygel and Drton (2010) and Nasir (2020) showed that the function $g(u_j)$ is strictly decreasing. In our empirical studies, we used bisection approach to solve for the optimal u_j. Further explanation on u_j is given in Remark 5.2.

Remark 5.1 The quantity $\Delta_* \geq 0$ is an integer for removing $2\Delta_*$ neighboring indices, i.e. $\widetilde{u} - \Delta_*, \cdots, \widetilde{u} - 1, \widetilde{u} + 1, \cdots \widetilde{u} + \Delta_*$ from the inactive-set $\mathcal{B}^*$. The rational for this removal is that once an index is estimated, consecutive indices are not considered as candidates for change-points. By removing these indices points, fewer irrelevant points will be selected into the active-set $\mathcal{B}$. Furthermore, the removal of the points may caused the *aBCD* algorithm to speed up as being observed in Sect. 6. The choice of Δ_* may depend on the length of the time series, where a sufficiently large Δ_* can be set if n is large. This strategy has been implemented by Chan et al. (2014) for the structural break autoregressive (SBAR) through their *gLAR* algorithm. However, Δ_* cannot be set too large as this might remove some of important change-points, especially when n is small.

Remark 5.2 In our empirical study, the root u_j of (13) is obtained using the bisection method, in which the property $\text{sign}(g(a_\star) - 1) \neq \text{sign}(g(b_\star) - 1)$ has to be satisfied for some $a_\star$ and $b_\star$ such that $a_\star \leq u_j < b_\star$. In the case of SETAR model, $a_\star = 10^{-5}$ and $b_\star = 10^5$ are deemed to be adequate based on results of simulation studies in Nasir (2020). However, occasionally the sign property may not hold for some particular j when $n \leq 300$, even when the initial interval is increased. The problem might be caused by unstable parameter convergence during the *aBCD* iterations under small sample size. To overcome this issue, the quantity u_j is temporarily replaced with 1 when this situation occurs.

5.2 Selecting shrinkage parameter and full gLASSO algorithm

For the reformulated SETAR model (7) with homoscedastic variance, we consider the following BIC (Wang et al. 2009):

$$\text{BIC}(\lambda) = N \log\left(\frac{\text{RSS}_\lambda}{N}\right) + \text{card}\left(\mathcal{A}_\lambda\right) \log(N) c_n, \tag{15}$$

where

$$\text{RSS}_\lambda = \sum_{t=1}^{N} \left(y_{\pi_t + d} - \mathbf{x}_{\pi_t}^T \sum_{k=1}^{t} \widehat{\boldsymbol{\theta}}_{\pi_k} \right)^2$$

is the residual sum-of-squares (RSS), $\mathcal{A}_\lambda = \{k : \widehat{\boldsymbol{\theta}}_{\pi_k} \neq \mathbf{0}\}$ is the set of indices corresponding to the set of non-zero estimated set of parameters and $c_n > 0$ is some positive constant. The first and the second terms in the RHS of (15) are known as the *goodness-of-fit* and *criterion penalty*, respectively. Since $\widehat{\boldsymbol{\theta}}^N$ could be group-wise sparse, we can take an advantage of the sparse feature to reduce the computational time for the residual sum-of-squares by replacing $\sum_{k=1}^{t} \widehat{\boldsymbol{\theta}}_{\pi_k}$ with $\sum_{k \in \{i : \widehat{\boldsymbol{\theta}}_{\pi_i} \neq \mathbf{0}, i \leq t\}} \widehat{\boldsymbol{\theta}}_{\pi_k}$.

In our simulation studies using SETAR models, we found that the change-points/thresholds are underestimated when $c_n \geq 1$. This issue is caused by the tendency of gLASSO to estimate an excessive amount of irrelevant change-points along with the important ones, causing the criterion penalty term of the RHS of (15) to become excessively large for $c_n \geq 1$. To circumvent this issue, we set c_n to a very low value, e.g., $c_n \leq 0.01$, so that all of the important change-points are eventually selected at a particular range for λ. Furthermore, this strategy is equivalent to achieving prediction accuracy rather than consistent model selection.

We now provide a strategy for choosing the appropriate values of λ_n. The main purpose of this strategy is to estimate only a small percentage of sets of non-zero parameters and the location of change-points. We choose grid of k_0 values for λ_n: $\lambda_1 = \lambda_{\max}, \lambda_2, \cdots, \lambda_{k_0} = \lambda_{\min}, \lambda_1 > \lambda_2 > \cdots > \lambda_{k_0}$.

Let $\widetilde{\mathcal{B}}_i$ be an active set corresponding to each $\lambda_i \in \{\lambda_n\}$, with convention $\widetilde{\mathcal{B}}_0 = \emptyset$. For each λ_i, we compute $\widetilde{\mathcal{B}}_i := \mathcal{B}$ and the corresponding $\text{BIC}(\lambda_i)$, where $\mathcal{B}$ is obtained from the *aBCD* algorithm and $\text{BIC}(\lambda_i)$ is given in (15). At the end, we choose a $\widetilde{\mathcal{B}}_i$ with the lowest BIC, denoted as $\widehat{\mathcal{B}}_* = \arg\min_{\widetilde{\mathcal{B}}_i}(v_i)$, where $v_i = \text{BIC}(\lambda_i)$. Finally, the thresholds are estimated using indices in $\widehat{\mathcal{B}}_*$, by $\widehat{\mathfrak{R}} = \{y_{\pi_{l-1}} : l \in \widehat{\mathcal{B}}_* \setminus \{1\}\}$.

The upper bound $k_{\max}$ is crucial to control how many change-points are estimated by the *aBCD* algorithm. Specifically, when the BCD iterations with a particular λ_i yields $\text{card}(\mathcal{B}) \geq k_{\max}$, this indicates that the current λ_i has overestimated the number of change-points and we may ignore the corresponding output. The full procedure to run gLASSO for the reformulated SETAR model is given in Algorithm 2.

Algorithm 2: Complete algorithm for the group LASSO of the reformulated SETAR

```
Data: $\mathbf{y}_\pi \in \mathbb{R}^N, \mathbf{x}_{\pi_1} \in \mathbb{R}^{p+1}, \cdots, \mathbf{x}_{\pi_N} \in \mathbb{R}^{p+1}, k_0, k_{\max} \geq 1$ and $c_n > 0$.
Result: The threshold set $\widehat{\mathfrak{R}}$.
Initialize: Set $i = 1$ and $\widetilde{\mathcal{B}}_0 = \emptyset$. Setup a grid of shrinkage parameter :
    $\{\lambda_1 = \lambda_{\max}, \lambda_2, \cdots, \lambda_{k_0} = \lambda_{\min}\}$, such that $\lambda_1 > \lambda_2 > \cdots > \lambda_{k_0}$ .
while $i \leq k_0$, do
    Apply Algorithm 1 with $\lambda_i$ and $k_{\max}$, and obtain $\widehat{\boldsymbol{\theta}}^N$ and $\mathcal{B}$. Then set
     $\widetilde{\mathcal{B}}_i := \mathcal{B}$.
    if card$(\widetilde{\mathcal{B}}_i) < k_{\max}$, then
        Compute $v_i = \text{BIC}(\lambda_i)$, where $\text{BIC}(\lambda_i)$ is given in (15).
    Update $i \leftarrow i + 1$.
Compute $\widehat{\mathcal{B}}_* = \arg\min_{\widetilde{\mathcal{B}}_i}(v_i)$.
Generate $\widehat{\mathfrak{R}} = \{y_{\pi_{l-1}} : l \in \widehat{\mathcal{B}}_* \setminus \{1\}\} := \{\widehat{r}_1, \cdots, \widehat{r}_{\widehat{m}}\}$, where $\widehat{m} = \text{card}(\widehat{\mathfrak{R}})$.
```

5.3 Post-analysis for SETAR

We now focus on obtaining consistent estimators of thresholds for SETAR model. Given a set of the estimated thresholds $\widehat{\mathfrak{R}} = (\widehat{r}_1, \cdots, \widehat{r}_{\widehat{m}})^T$ obtained from gLASSO and $\widehat{m} = \text{card}(\widehat{\mathfrak{R}})$, we define the information criterion similar to (15) as

$$\text{tBIC}(\widehat{m}, \widehat{\mathfrak{R}}) = N\log(s(\widehat{r}_1, \widehat{r}_2, \cdots, \widehat{r}_{\widehat{m}})/N) + \widehat{m}\log(N)c_{\text{E}}, \tag{16}$$

where $c_{\text{E}} \geq 0$ is the criterion constant (see Remark 5.3 for details regarding selection of c_{E}) and $s(\widehat{r}_1, \widehat{r}_2, \cdots, \widehat{r}_{\widehat{m}}) = \sum_{j=1}^{\widehat{m}+1} s(\widehat{r}_{j-1}, \widehat{r}_j)$ is the joint residual sum-of-squares (jRSS) function, with

$$s(\widehat{r}_{j-1}, \widehat{r}_j) = \sum_{t=p+1}^{n} \left(y_t - \boldsymbol{x}_t^T \widehat{\boldsymbol{\phi}}_j\right)^2 I_{(\widehat{r}_{j-1}, \widehat{r}_j]}(y_{t-d}), \tag{17}$$

the residual sum-of-squares function for jth regime. Recall that $\boldsymbol{x}_t = (1, y_{t-1}, y_{t-2}, \cdots y_{t-p})^T$ and

$$\widehat{\boldsymbol{\phi}}_j = \sum_{t=p+1}^{n} \left[\left(\boldsymbol{x}_t \boldsymbol{x}_t^T\right) I_{(\widehat{r}_{j-1}, \widehat{r}_j]}(y_{t-d})\right]^{-1} \sum_{t=p+1}^{n} (\boldsymbol{x}_t y_t) I_{(\widehat{r}_{j-1}, \widehat{r}_j]}(y_{t-d}) \tag{18}$$

as the parameter estimate for the jth regime.

Let $h = \sum_{i=0}^{\text{card}(\widehat{\mathfrak{R}})} \text{card}(\widehat{\mathfrak{R}})!/(i!(\text{card}(\widehat{\mathfrak{R}}) - i)!)$ and $\mathcal{P}(\widehat{\mathfrak{R}}) := \{\widehat{\mathfrak{R}}_0^*, \widehat{\mathfrak{R}}_1^*, \cdots, \widehat{\mathfrak{R}}_h^*\}$be the power set of thresholds where $\widehat{\mathfrak{R}}_0^* = \emptyset$, the empty set. One way to select the number of thresholds is by the minimization

$$\widehat{\widehat{\mathfrak{R}}} = \arg\min_{\widehat{\mathfrak{R}}_j^* \subseteq \widehat{\mathfrak{R}}} \ \text{tBIC}\left(\text{card}\left(\widehat{\mathfrak{R}}_j^*\right), \widehat{\mathfrak{R}}_j^*\right), \quad j \in \{0, 1, 2, \cdots, h\}. \tag{19}$$

We write $\widehat{\widehat{\mathfrak{R}}} = (\widehat{\widehat{r}}_1, \cdots, \widehat{\widehat{r}}_{\widehat{\widehat{m}}})$ with $\widehat{\widehat{m}} = \text{card}(\widehat{\widehat{\mathfrak{R}}})$. The minimization (19) implies that all possible subsets in $\widehat{\mathfrak{R}}$ are accounted. Therefore, the computation of the criterion is of order of h, which can be prohibitive if $\widehat{\mathfrak{R}}$ is a large set.

Chan et al. (2015) suggested an application of the backward elimination algorithm, or *BEA* to further improve the computational time for estimating thresholds. Note that this algorithm is part of well-known stepwise selection approach for regression (Weisberg 2005, p. 222). The algorithm iteratively removes a threshold from the set $\widehat{\mathfrak{R}}$ one at a time, to lower the tBIC given in (16), until no further reduction in tBIC is possible.

Given $\widehat{\widehat{\mathfrak{R}}}$, we can estimate all the parameters for each regime of SETAR model by (18). The steps for performing *BEA* and parameter estimation for SETAR are summarized in Algorithm 3.

The following theorem given the consistency result in estimating threshold via *BEA*.

Theorem 5.2 *Under the conditions of* HA1–HA4, *and when card*$(\widehat{\mathfrak{R}}) \geq m^0$, *the BEA satisfies*

$$P\left(\widehat{\widehat{m}} = m^0\right) \to 1$$

```
Algorithm 3: Backward Elimination Algorithm and parameter estimates for SETAR.
Data: $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{x}_{p+1} \in \mathbb{R}^{p+1}, \cdots, \boldsymbol{x}_n \in \mathbb{R}^{p+1}$, $c_E$ and $\widehat{\mathfrak{R}}$.
Result: $\widehat{\widehat{\boldsymbol{\phi}}}_j s$, $\widehat{\widehat{\mathfrak{R}}}$ and $\widehat{\widehat{m}}$.
Initialize: $k_0 = card(\widehat{\mathfrak{R}})$.
repeat
    Compute $v^*_{k_0} = tBIC(k_0, \widehat{\mathfrak{R}})$, where $tBIC()$ is given in (16).
    for $i = 1, \cdots, k_0$, do
        Compute $v_{k_0,i} = tBIC(k_0 - 1, \widehat{\mathfrak{R}} \setminus \{\widehat{r}_i\})$, $\widehat{r}_i \in \widehat{\mathfrak{R}}$.
    Set $v^*_{k_0-1} = min_i\ (v_{k_0,i})$.
    if $v^*_{k_0-1} < v^*_{k_0}$ then
        Compute $j = arg\ min_i\ v_{k_0,i}$.
        Update $\widehat{\mathfrak{R}} \leftarrow \widehat{\mathfrak{R}} \setminus \left\{\widehat{r}_{k_0,j}\right\}$ and $k_0 \leftarrow k_0 - 1$.
until $v^*_{k_0-1} \geq v^*_{k_0}$ or $k_0 = 0$.
Set $\widehat{\widehat{\mathfrak{R}}} := \widehat{\mathfrak{R}}$ and $\widehat{\widehat{m}} := card(\widehat{\mathfrak{R}})$.
for $j = 1, \cdots, k_0 + 1$, do
    Compute $\widehat{\widehat{\boldsymbol{\phi}}}_j =$
    $\sum_{t=p+1}^{n} \left[\left(\boldsymbol{x}_t \boldsymbol{x}_t^T\right) I_{(\widehat{\widehat{r}}_{j-1}, \widehat{\widehat{r}}_j]}(y_{t-d})\right]^{-1} \sum_{t=p+1}^{n} (\boldsymbol{x}_t y_t) I_{(\widehat{\widehat{r}}_{j-1}, \widehat{\widehat{r}}_j]}(y_{t-d}), \widehat{\widehat{r}}_j \in \widehat{\widehat{\mathfrak{R}}}$
    with conventions $\widehat{\widehat{r}}_0 = -\infty$ and $\widehat{\widehat{r}}_{k_0+1} = +\infty$.
```

and there exist a constant $b > 0$ such that

$$P\left(\max_{1 \leq j \leq m^0} \left|\widehat{\widehat{r}}_j - r_j^0\right| \leq b m^0 \gamma_n\right) \to 1.$$

Theorem 5.2 can be proved using similar lines as in the proof of Theorem 2.5 in Chan et al. (2014, 2015). The idea of the tBIC is simple. Assume that all relevant thresholds are in $\widehat{\mathfrak{R}}$ and Theorem 4.4 holds. If the estimated number of thresholds is lower than m^0, then the goodness-of-fit dominates the criterion, which leads to $P(\widehat{\widehat{m}} < m^0) \to 0$. Otherwise, if the estimated number of thresholds is higher than m^0, then the criterion penalty dominates the criterion instead, which leads to $P(\widehat{\widehat{m}} > m^0) \to 0$.

Remark 5.3 As shown by Gonzalo and Pitarakis (2002) and Chan et al. (2015), $c_E = 2, 3$ usually works better than $c_E = 1$ in correctly estimating the number of thresholds via BIC provided that model coefficients for each regime are sufficiently large. Alternatively, one can consider replacing the default penalty term $\log(N)c_E$ with N^δ with $\delta \in (1/2, 3/4)$; refer to Remark 7 in Ciuperca (2011). The latter implies that as sample size increases, so does c_E.

6 Simulation studies

In this section, we compare the performance between the *SLS* (Algorithms A1 and 2) and *aBCD* (Algorithms 1 and 2) algorithms, along with the two ensemble algorithms of *aBCD-BEA* (Algorithms 1, 2 and 3) and *gLAR-BEA* (Algorithms A2 and 3). Both *SLS* (Algorithm A1) and *gLAR* (Algorithm A2) algorithms are given in the Supplementary Materials. These algorithms were coded using the `R` language in conjunction with the `Cpp` language through `Rcpp` package (Eddelbuettel and Francois 2011) to considerably speed up the run time of these algorithms in the `R` statistical environment. Simulation studies were conducted on multiple personal computers without parallelization, each running on a four-core Intel *i7* processor with base clock speed of at least 3.5 GHz. Discussion on the choice of λ_n, $k_{\max}$ and Δ_* for these studies is provided in Remark 6.1.

Remark 6.1 The appropriate range of values for λ_n can be difficult to determine in practice. In this section, we determine that $\lambda_{\max} = 0.5$, $\lambda_{\min} = 0.01$ and $20 \leq k_0 \leq 40$ are deemed to be appropriate for estimating a moderate number of relevant change-points/thresholds. Meanwhile, the best value for both $k_{\max}$ and Δ_* can be evaluated in practice using grid-search approach and BIC, e.g., through *BEA* algorithm, but was not considered here for the sake of comparison purposes and reducing computational costs. Some of these quantities used in our empirical studies may be different in some previous studies.

6.1 Comparison study: *SLS* and *aBCD* algorithms

First, we evaluate the performance of *SLS* (Algorithms A1 and 2) algorithm and *aBCD* algorithm (Algorithms 1 and 2) using datasets generated by the three models given below.

Model 1 Three regime SETAR(1) with the non-zero intercepts is defined as

$$y_t = \begin{cases} 1 - 0.4y_{t-1} + \varepsilon_t, & \text{if} \quad y_{t-1} \in (-\infty, -0.8], \\ 0.6 + y_{t-1} + \varepsilon_t, & \text{if} \quad y_{t-1} \in (-0.8, 0.5], \\ -1 - 0.2y_{t-1} + \varepsilon_t, & \text{if} \quad y_{t-1} \in (0.5, \infty), \end{cases} \tag{20}$$

$t = 2, 3, \cdots, n$, where $\varepsilon_t \overset{i.i.d}{\sim} N(0, 1)$. The model was introduced by Li and Ling (2012). Although the coefficient associated with the term y_{t-1} in the second regime of (20) is exactly one, the overall process $\{y_t\}$ is not a unit-root process since the stationarity of multiple regime SETAR with $p = 1$ depends on the first and the last regimes (Chan et al. 1985; Li and Ling 2012).

Model 2 Three regime SETAR(2) with the zero intercepts is defined as

$$y_t = \begin{cases} 0.8y_{t-1} - 0.2y_{t-2} + \varepsilon_t, & \text{if} \quad y_{t-1} \in (-\infty, -2], \\ 1.9y_{t-1} - 0.81y_{t-2} + \varepsilon_t, & \text{if} \quad y_{t-1} \in (-2, 2], \\ 0.6y_{t-1} - y_{t-2} + \varepsilon_t, & \text{if} \quad y_{t-1} \in (2, \infty), \end{cases} \tag{21}$$

$t = 3, 4, \cdots, n$, where $\varepsilon_t \overset{i.i.d}{\sim} N(0, 1)$. This model was considered by Chan et al. (2017). In TAR literature, the nonzero intercepts in at least one regime can provide different levels and variability in the series structure, as well as asymmetry and a multimodal distribution (Niglio and Vitale 2015). With the zero intercepts, identification of important thresholds might be challenging since both level and variability of the time series will be limited.

Model 3 The nine regime SETAR(2) with the non-zero intercepts is defined as

$$
\begin{aligned}
y_t = & (-4.5 - 0.6y_{t-1})\, I_{(-\infty,-3.5]}(y_{t-1}) \\
& + (2.5 + 0.3y_{t-1} + 0.9y_{t-2})\, I_{(-3.5,-2.5]}(y_{t-1}) \\
& + (-2.0 - 0.9y_{t-1})\, I_{(-2.5,-1.5]}(y_{t-1}) \\
& + (2.3 + 0.7y_{t-1} + 0.5y_{t-2})\, I_{(-1.5,-0.5]}(y_{t-1}) \\
& + (1.0 + 0.1y_{t-1})\, I_{(-0.5,0.5]}(y_{t-1}) + (3.0 - 0.9y_{t-1})\, I_{(0.5,1.5]}(y_{t-1}) \\
& + (1.6 + 0.9y_{t-1})\, I_{(1.5,2.5]}(y_{t-1}) + (-0.5 - 0.8y_{t-1} - 0.2y_{t-2})\, I_{(2.5,3.5]}(y_{t-1}) \\
& + (1.5 - 1.1y_{t-1})\, I_{(3.5,\infty)}(y_{t-1}) + \varepsilon_t
\end{aligned}
\tag{22}
$$

$t = 2, 3, \cdots, n$, where $\varepsilon_t \overset{i.i.d}{\sim} N(0, 1)$. This model with the same parameters was considered in Chan et al. (2015) and Chan et al. (2017), with the exception that in Chan et al. (2017), one of coefficients in the sixth and seventh regimes had opposite signs to (22).

Three different values were pre-set for λ_n for the three models. We ran 1000 replication, where for each replication, all methods shared the same dataset for a fair comparison. For both methods, the convergence criterion is assumed to be met when $\left\| \widehat{\boldsymbol{\theta}}_N^{[l+1]} - \widehat{\boldsymbol{\theta}}_N^{[l]} \right\|_1 < 0.001$, for $l = 1, 2, \cdots$, where $\widehat{\boldsymbol{\theta}}_N^{[l]} = (\widehat{\boldsymbol{\theta}}_{\pi_1}^{[l]T}, \cdots, \widehat{\boldsymbol{\theta}}_{\pi_N}^{[l]T})^T$ is the estimates of $\boldsymbol{\theta}_N$ after lth iteration. For the *aBCD* algorithm, we set $k_{\max}$ =10,000 to allow as many threshold estimates as possible.

Table 1 shows the comparison between the algorithms' performance of *SLS*, *aBCD* with $\Delta = 0$ and *aBCD* with $\Delta = 10$ for three different models. First, we observe that the results for *SLS* and *aBCD* with $\Delta = 0$ are similar for all models, with a few exceptions that the *aBCD* gives equal or smaller average Hausdorff distance, one less change points estimate in average and much faster convergence compared to the former for Models 1 and 3. In the case of Model 2, both methods had issues with the convergence in the sense that both *SLS* and *aBCD* alternate indefinitely between a few sets of solutions. From this results, *aBCD* is preferable due to its computational speed while having comparable average Hausdorff distances with *SLS*.

In the case of *aBCD* algorithm with $\Delta = 10$, it has faster convergence although with a slightly higher average Hausdorff distance for Model 1, and having no convergence issue for Model 2 when compared to the same algorithm with $\Delta = 0$. For Model 3, results of the *aBCD* algorithm with both $\Delta = 0$ and $\Delta = 10$ are the same.

Table 1 Results of comparison between *SLS*, *aBCD* with $\Delta = 0$ and *aBCD* with $\Delta = 10$ for three models with $n = 1200$

	SLS				*aBCD* with $\Delta = 0$			*aBCD* with $\Delta = 10$		
	λ	d_H	$\#(\bar{\mathcal{B}})$	$\bar{\mathcal{T}}$	d_H	$\#(\bar{\mathcal{B}})$	$\bar{\mathcal{T}}$	d_H	$\#(\bar{\mathcal{B}})$	$\bar{\mathcal{T}}$
Model 1	0.1	0.016	8	3.000	0.016	7	0.119	0.019	6	0.036
Model 2	0.4	*NA*			*NA*			0.033	5	0.047
Model 3	0.1	2.039	5	2.286	2.007	4	0.023	2.007	4	0.023

The acronym *NA* indicates that the method has convergence issue; $\#(\bar{\mathcal{B}})$ is the average number of estimated change-point/threshold candidates; d_H is the Hausdorff distance equation from (11); $\bar{\mathcal{T}}$ is the average time in minutes to complete 1000 simulations

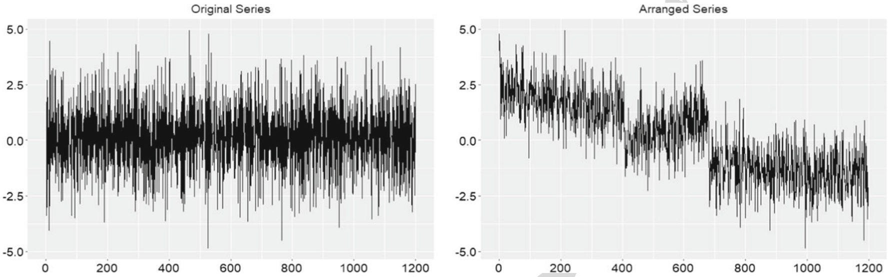

Fig. 1 Plots of a realization of original series (left) and arranged series (right) generated from (20), with $n = 1200$

6.2 Comparison study: *aBCD-BEA* and *gLAR-BEA* algorithms

In this section, the performance of two ensemble algorithms, the *aBCD-BEA* and *gLAR-BEA* for a two-step threshold estimation procedure are compared through three simulation studies. The first-step estimation procedure for the ensemble algorithms involves the application of *aBCD* (Algorithms 1 and 2) and *gLAR* (Algorithm A1) to estimate threshold candidates through the estimation of change-points. The second-step estimation procedure uses the *BEA* (Algorithm 3) to exclude any irrelevant thresholds from the set of the threshold candidates obtained in the first step procedure. For each simulation study, we generate 1000 datasets from each model (Models 1–3), where for every replication, each method shared the same dataset for a fair comparison. For the gLASSO estimation via *aBCD* algorithm, we generate decreasing sequence of twenty points $0.5, 0.474, 0.448, \cdots, 0.01$ for the shrinkage parameter λ_n, and the BIC penalty constant c_n in (15) is set to 0.01.

First, we are comparing both ensemble methods using data generated from Model 1. For this simulation study, we consider sample sizes $n = 300, 750,$ and 1200. For all sample sizes, we set $k_{\max} = 5, 10, 15, 20$ with $\Delta_* = 10$ for both *aBCD* and *gLAR* algorithms. Further, we set $c_{\mathrm{E}} = 3$ in the *BEA* step.

Figure 1 shows an example of plots of original (left) and the corresponding arranged time series (right) generated from (20). The original time series plot appears stationary while the plot of arranged series indicates abrupt switching patterns at two locations, as expected.

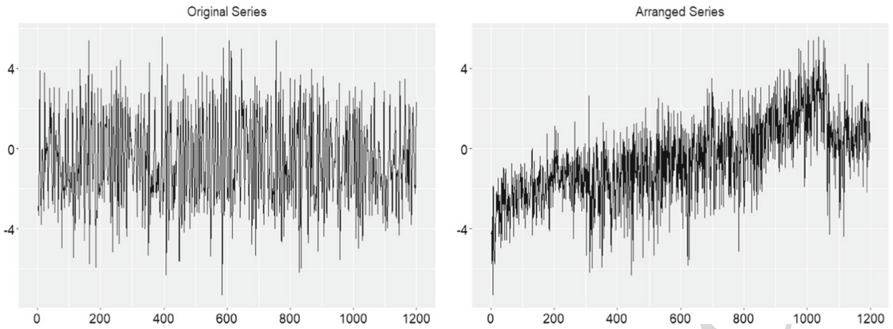

Fig. 2 The line plots of an original series (left) and the corresponding arranged series (right) generated from (21), with $n = 1200$

From Table 2, we observe that when $k_{\max} = 5$, the percentages of correctly estimating the number of thresholds are significantly lower, having much higher rate for both average Hausdorff distances and underestimation issue is more severe for *aBCD-BEA* compared to *gLAR-BEA* regardless of sample size, indicating there are at least one or more relevant thresholds are regularly not estimated by the *aBCD* algorithm under the preset $k_{\max}$.

When $k_{\max} = 10$, the percentages of correctly estimating the number of thresholds are comparable between both methods for each sample size. However, The average Hausdorff distances under *aBCD-BEA* are slighly higher then the ones generated by *gLAR-BEA*. The percentages are close to 100% when we increase the sample size to 1200 for both methods. When we increase the $k_{\max}$ to 15 and 20, we observe the percentages are lower for *gLAR-BEA* as compared to *aBCD-BEA* for all sample sizes, especially when $n = 300$. It is not surprising that *aBCD* is computationally slower than *gLAR* due to its behavior of estimating parameters until convergence.

Comparing with Chan et al. (2015), for $n = 300$, we obtain a higher percentages of correctly estimated number of thresholds for both ensemble algorithms with $k_{\max} \geq 10$, compared to their result 78.1%. Note that this comparison might depend on the values of Δ_* and c_E, which were not specified in their paper.

Next, we evaluate both ensemble methods using data generated from Model 2. We considered the same settings as in the previous simulation study for the sample size n, $k_{\max}$, Δ and c_E. Figure 2 shows an example of plots of original (left) and the corresponding arranged time series (right) generated from (21).

The plot of original time series in Fig. 2 shows that the series looks somehow stationary. From the plot of the arranged series in Fig. 2, structural changes in the series are difficult to identify due to vague switching patterns. Furthermore, the switching appears to be smooth rather than abrupt, unlike the previous model. Therefore, the threshold estimation for Model (21) might be challenging. We are interested to know whether both ensemble algorithms are able to identify correct thresholds for this model.

From Table 3, we observe that when $k_{\max} = 5$, the percentages of correct estimation of the number of thresholds are comparable for both methods when $n = 300$. The underestimation issue occurred by *aBCD-BEA* were more severe when the sample size increases to 750 and 1200 under the preset $k_{\max} = 5$. Meanwhile, *gLAR-BEA* does

Table 2 Result of two-step estimation procedures for 1000 samples generated from (20) with various sample sizes with $k_{\max} = 5, 10, 15, 20$ and $\Delta_* = 10$

		aBCD	*aBCD-BEA*					*gLAR*	*gLAR-BEA*				
K	n	$\#(\mathcal{B})$	< 2	$= 2$	> 2	$\bar{d}_H$	$\bar{\mathcal{T}}$	$\#(\mathcal{B})$	< 2	$= 2$	> 2	$\bar{d}_H$	$\bar{\mathcal{T}}$
5	300	4	76.9	23.1	0	1.286	0.042	5	12.6	86.1	1.3	0.147	0.002
	750	4	78.8	21.2	0	1.574	0.180	5	19	80.6	0.4	0.291	0.008
	1200	4	84.6	15.4	0	1.679	0.434	5	27.4	72.4	0.2	0.421	0.017
10	300	7	8.2	90.1	1.7	0.092	0.063	10	8.5	89.7	4.5	0.053	0.003
	750	7	1.5	98	0.5	0.040	0.213	10	0	98.7	1.3	0.019	0.013
	1200	7	1.7	97.8	0.5	0.034	0.477	10	0.1	99.1	0.8	0.013	0.031
15	300	12	6	89.1	4.9	0.054	0.080	15	6.5	82.9	10.6	0.051	0.005
	750	11	0	99.1	0.9	0.019	0.235	15	0	97.9	2.1	0.019	0.020
	1200	10	0	99.3	0.7	0.012	0.505	15	0	99.1	0.9	0.012	0.047
20	300	15	6.3	84.3	9.4	0.051	0.101	20	7.1	77.5	15.4	0.049	0.007
	750	11	0	98.5	1.5	0.019	0.260	20	0	95.5	4.5	0.019	0.027
	1200	10	0	99.3	0.7	0.012	0.529	20	0	98.6	1.4	0.012	0.064

$\#(\bar{\mathcal{B}})$ is the average number of estimated change-point/threshold candidates from the first step estimation procedure. For the second step estimation procedure, $(< 2, = 2, > 2)$ are the estimated number of thresholds in percentages, $\bar{d}_H$ is the average Hausdorff distance equation using (11) and $\bar{\mathcal{T}}$ is the average time in minutes to complete 1000 simulations

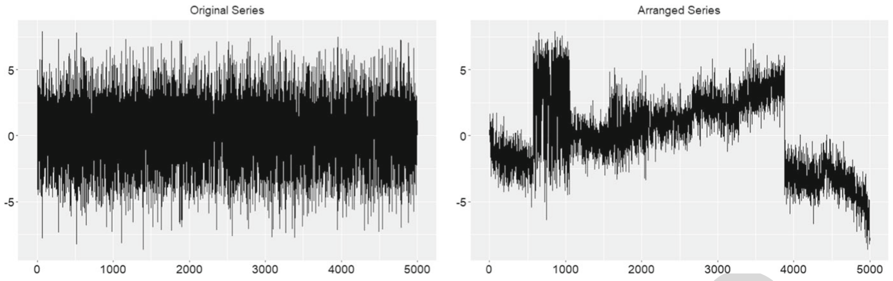

Fig. 3 The line plots of one realization of original series (left) and the corresponding arranged series (right) generated from nine-regime SETAR(2) of (22), with $n = 5000$

not suffer any underestimation issue and able to retain the rate of correct estimation for more than 90% regardless of the sample size.

When we increase $k_{\max}$ to 10, 15 and 20, we notice that the underestimation (< 2) rate for the *aBCD-BEA* has dropped significantly while having much higher percentages on the correct estimation of the number of thresholds compared to *gLAR-BEA* for all sample sizes. In addition, the *gLAR-BEA* suffered a more severe percentages decrease on the correct estimation of the number of thresholds ($= 2$) for all sample sizes compared to *aBCD-BEA* despite estimating more thresholds during the first step estimation procedure. Interestingly, the average Hausdorff distances under *aBCD-BEA* are lower than the ones generated by *gLAR-BEA* for all sample sizes and $k_{\max}$ despite the underestimation issue.

Finally, we now evaluate both ensemble methods again using data generated from Model 3. Previously, Chan et al. (2015) applied their version of *gLAR-BEA* for n =10,000 with $k_{\max} = 40$, but Δ_* and c_{E} were not specified. Figure 3 shows an original and the corresponding arranged time series generated from this model. The original time series exhibit no obvious trend indicating stationarity of the series, but the plot of arranged series shows an abrupt switching pattern which corresponds to the multiple structural changes or breaks in the series.

For this simulation study, we consider sample sizes $n = 2500$, 5000 and 7500, with 1000 replications for each sample size. We fix $\Delta_* = 20$ and choose $k_{\max} = 20$ and 40 for both *aBCD* and *gLAR* algorithms, with $c_{\mathrm{E}} = 3$ for *BEA*.

The results in Table 4 show that for $k_{\max} = 20$, the *aBCD* estimates 15–16 thresholds on average while *gLAR* always estimates exactly 20 thresholds for all sample sizes. Furthermore, *aBCD-BEA* struggles to achieve at least 90% of correct estimation for the number of thresholds for all sample sizes, and having much severe underestimation issue and larger average Hausdorff distances compared to the ones obtained by *gLAR-BEA*.

As we increase $k_{\max}$ to 40, we observe that the percentages of correct estimation for the number of thresholds under *aBCD-BEA* are substantially increased and exceed 96% for all sample sizes. Meanwhile, the performance of *gLAR-BEA* is comparable to *aBCD-BEA* especially for $n = 5000$ and 7500, with a few exceptions where the former having slightly lower average Hausdorff distances and much lower computational times.

Table 3 Result of two-step estimation procedures for 1000 samples generated from (21) with various sample sizes with $k_{\max} = 5, 10, 15, 20$ and $\Delta_* = 10$

		aBCD	*aBCD-BEA*					*gLAR*	*gLAR-BEA*				
K	n	$\#(\bar{\mathcal{B}})$	< 2	$= 2$	> 2	$\bar{d}_H$	$\bar{\mathcal{T}}$	$\#(\bar{\mathcal{B}})$	< 2	$= 2$	> 2	$\bar{d}_H$	$\bar{\mathcal{T}}$
5	300	3	7.1	91.1	1.8	0.090	0.036	5	0	91	9	0.113	0.002
	750	3	13.6	85.7	0.7	0.040	0.102	5	0	93.9	6.1	0.072	0.008
	1200	3	18.1	81.4	0.5	0.027	0.197	5	0	96.2	3.8	0.062	0.018
10	300	8	0	94.1	5.9	0.075	0.063	10	0	87.8	12.2	0.091	0.003
	750	8	1.1	96.7	2.2	0.032	0.132	10	0	94.6	5.4	0.042	0.013
	1200	8	2.1	96.6	1.3	0.019	0.239	10	0	95.5	4.5	0.028	0.031
15	300	12	0	92.5	7.5	0.074	0.140	15	0	86.1	13.9	0.086	0.005
	750	11	0	96.9	3.1	0.030	0.183	15	0	93	7	0.041	0.019
	1200	12	0.2	98.2	1.6	0.019	0.305	15	0	94.5	5.5	0.026	0.046
20	300	17	0	90.6	9.4	0.073	0.232	20	0	85.2	14.8	0.082	0.007
	750	14	0	96.3	3.6	0.030	0.282	20	0	92.2	7.8	0.040	0.028
	1200	16	0	97.6	2.4	0.019	0.360	20	0	93.1	6.9	0.026	0.064

$\#(\bar{\mathcal{B}})$ is the average number of estimated change-point/threshold candidates from the first step estimation procedure. For the second step estimation procedure, $(< 2, = 2, > 2)$ are the estimated number of thresholds in percentages, $\bar{d}_H$ is the average Hausdorff distance equation using (11) and $\bar{\mathcal{T}}$ is the average time in minutes to complete 1000 simulations

Table 4 Result of two-step estimation procedures for 1000 samples generated from (22) with various sample sizes with $k_{\max} = 20, 40$ and $\Delta_* = 20$

		aBCD	*aBCD-BEA* $(\%\widehat{\widehat{m}})$					*gLAR*	*gLAR-BEA* $(\%\widehat{\widehat{m}})$				
$k_{\max}$	n	$\#(\bar{\mathcal{B}})$	< 8	$= 8$	> 8	$\bar{d}_H$	$\bar{\mathcal{T}}$	$\#(\bar{\mathcal{B}})$	< 8	$= 8$	> 8	$\bar{d}_H$	$\bar{\mathcal{T}}$
20	2500	16	11	87.7	1.3	0.177	0.901	20	0.9	94.9	4.2	0.071	0.282
	5000	15	19.5	78.4	2.1	0.243	2.029	20	4.3	94.5	1.2	0.077	1.114
	7500	15	21.2	75.3	3.5	0.262	3.681	20	11.3	88.2	0.5	0.135	2.254
40	2500	22	0	96.8	3.2	0.054	2.362	40	0	90.5	9.5	0.045	0.659
	5000	24	0	98.5	1.5	0.038	4.634	40	0	97.8	2.2	0.023	2.375
	7500	27	0.1	97	2.9	0.032	7.489	40	0	98.7	1.3	0.015	4.658

$\#(\bar{\mathcal{B}})$ is the average number of estimated change-point/threshold candidates from the first step estimation procedure. For the second step estimation procedure, $(< 2, = 2, > 2)$ are the estimated number of thresholds in percentages, $\bar{d}_H$ is the average Hausdorff distance equation using (11) and $\bar{\mathcal{T}}$ is the average time in minutes to complete 1000 simulations

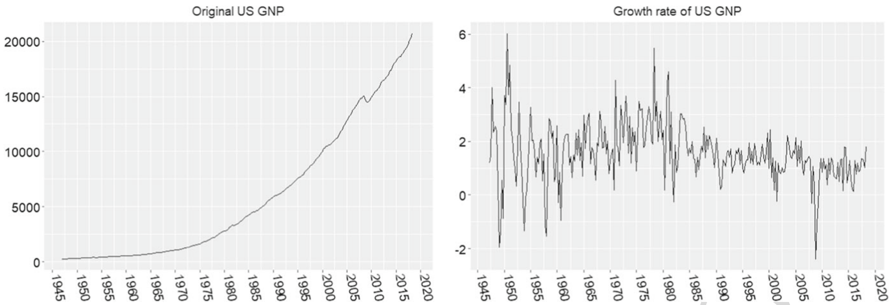

Fig. 4 The plots of an original series (left) and the corresponding growth rate (right) of quarterly US GNP time series, from 1947 to 2018

In conclusion, *gLAR-BEA* is definitely much faster than *aBCD-BEA* since *gLAR* algorithm does not estimate set of parameters in cyclical manner till convergence as compared to *aBCD*. However, we observe that the performance of *gLAR-BEA* in estimating correct number of thresholds tends to decrease especially for Models 1 and 2 when $k_{\max}$ increases as too many irrelevant thresholds estimated by *gLAR* might cause the *BEA* to choose model with the overestimated thresholds.

On the other hand, *aBCD-BEA* is somehow has a better robust and do not suffer much from the same issue. In addition, we observe *aBCD-BEA* has higher percentage of correct estimation number of thresholds compared to *gLAR-BEA* for sufficiently large $k_{\max}$. The average Hausdorff distances obtained by both *aBCD-BEA* and *gLAR-BEA* are acceptable under sufficiently large $k_{\max}$.

7 Case studies

In this section, the performance of two ensemble algorithms, the *aBCD-BEA* and *gLAR-BEA* for a two-step threshold estimation procedure are compared through two case studies. Both *aBCD-BEA* and *gLAR-BEA* are applied and several statistics, along with the estimated thresholds obtained by both *aBCD* and *gLAR* are reported. Similar setup from Sect. 6.2 is applied for the shrinkage parameter λ_n.

7.1 Case study 1: US GNP data

The quarterly growth series of United States (US) gross national product (GNP) was obtained from https://fred.stlouisfed.org/series/GDP. This data has previously been analyzed by Li and Ling (2012), Chan et al. (2015) and Chan et al. (2017) using different estimation methods and periods of time series. In this study, we select the series starting from the first quarter of 1947 to the first quarter of 2018, with a total of 286 observations, and aim to compare and evaluate results of *aBCD-BEA* and *gLAR-BEA*.

Table 5 A summary of two-step threshold estimates using *aBCD-BEA* and *gLAR-BEA* with $k_{\max} = 10$ and $\Delta_* = 10$ for the growth rate of US GNP time series (1947–2018); Bolded values indicate equal selection of threshold for both *aBCD* and *gLAR*; Estimated threshold in the first step (ETH1), estimated thresholds in the second step (ETH2), number of observation in each regime (#Obs.), Bayesian information criterion (BIC), joint sum-of-squared error (jSSE), individual sum-of-squared error (SSE)

	aBCD	*gLAR*
ETH1	(1.361, **1.629**, 1.940, **2.137**, 2.514, **3.292**)	(− 0.298, 0.373, 0.833, 1.023, **1.629**, 1.840, **2.137**, 2.377, **3.292**)
	aBCD-BEA	*gLAR-BEA*
ETH2	(1.361, 1.940, **2.137**, 2.514, **3.292**)	(**1.629**, **2.137**)
#Obs	(122, 68, 21, 19, 30, 14)	(156, 55, 63)
BIC	− 108.24	− 99.18
jSSE	110.6	155.45
SSE	(55.58, 32.09, 8.02, 5.52, 7.21, 2.19)	(82.52, 32.62, 40.30)

We compute the growth rate by the following operation:

$$y_t = 100(\log(x_t) - \log(x_{t-1})), \quad t = 2, \cdots, 286,$$

where x_t is the original observation and y_t is the growth rate, and the plots of these two series are shown in Fig. 4. Here, $p = 11$ is chosen similar to the setup in Chan et al. (2015). Using likelihood ratio test of Chan and Tong (1990) with $p = 11$, via `tlrt` function of `TSA` package in `R`, we determine that the delay parameter d is 6, based on the highest test statistic of the ratio. The selected value of the delay parameter coincides with the value used by the aforementioned studies.

We utilize both ensemble algorithms, *aBCD-BEA* and *gLAR-BEA*. For both procedures, we set $k_{\max} = 10$ and $\Delta_* = 10$. For the BEA, we set the information criterion penalty $c_E = 5$.

Table 5 provides details on the comparison. Using change-points/thresholds estimated by *aBCD*, the *BEA* only removes one value from the threshold set. On the other hand, the *BEA* removes seven values from the *gLAR* threshold set. *aBCD-BEA* eventually retains five thresholds instead of two via *gLAR-BEA*, and the BIC and jRSS suggested that the five thresholds from the former method provide a better fit for the growth rate compared to the two thresholds from the latter method. This also indicates that five thresholds estimated by *aBCD-BEA* may provide better explanation for the non-linearity of the growth rate of US GNP compared to the two thresholds estimated by *gLAR-BEA*.

It is worth mentioning that some of the estimated thresholds via *aBCD-BEA* are close to the values obtained in previous studies. For example, our estimated thresholds 1.361, 1.940, 2.137 are close to those obtained by Chan et al. (2017) which are 1.23, 1.65, 2.23.

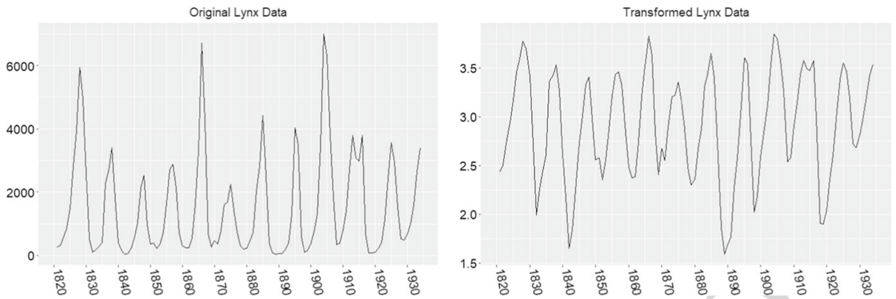

Fig. 5 The plots of original (left) and the logarithmically (base 10) transformed (right) Canadian lynx trapping time series, from 1821 to 1934

7.2 Case study 2: lynx trapping data

Next, we analyze the annual Canadian lynx trapping time series in the MacKenzie river, Canada for the period 1821–1934. The series contains 114 observations and it is obtained using the `lynx` command in `R`. The non-linearity of the series has been initially observed in Tong and Lim (1980). To explain the non-linearity, SETAR models with at most two thresholds had been previously applied and discussed and it is assumed that the number of thresholds is fixed (Tong and Lim 1980; Tsay 1989; Geweke and Terui 1993; Chen et al. 2011; Li and Tong 2001; Tong 1990; Lopes and Salazar 2006). We are not aware of any previous literature attempting to estimate SETAR model without fixing the number of thresholds a priori for this series.

Prior to analyzing the time series, we follow the recommendation of Bulmer (1974) and Tong and Lim (1980) to logarithmically (base 10) transform the time series $\{x_t\}$ to $\{y_t\}$:

$$y_t = \log_{10}(x_t), \quad t = 1, 2, \cdots, 114, \tag{23}$$

and the two series were plotted in Fig. 5. We observe that both plots exhibit strong cyclical pattern and the data transformation achieves stationarity.

Next, the delay parameter d has to be specified. Previously, Tong and Lim (1980) set $d = 2$, justified by the pre-determined predator–prey cycles of approximately 2 years between lynx and its prey (Bulmer 1974; Tong and Lim 1980). Tsay (1989) applied an F-test with two different AR orders and conclude that: if AR orders are 9 and 11, then $d = 2$ and $d = 3$ give the highest F-values, respectively. In other studies, Geweke and Terui (1993) and Chen et al. (2011) concluded, via Bayesian inference, that $d = 3$ gives the highest probability of marginal posterior distribution, and Li and Tong (2001), via classical inference identify $d = 3$ using corrected Akaike information criterion (AICc).

In our case, we run the likelihood ratio test of Chan and Tong (1990) via `tlrt` function in `R` (Cryer and Chan 2008) with different AR orders of $p = 3, 8, 12$ and 16. When AR order increases up to $p = 16$, the ratio gives the highest priority for $d = 3$. Based on this, we choose $d = 3$ for our SETAR model.

Table 6 A summary of two-step threshold estimates using *aBCD-BEA* and *gLAR-BEA*, with $k_{\max} = 7$ and $\Delta_* = 10$, for the transformed Canadian lynx trapping time series (1821–1934); Estimated threshold in the first step (ETH1), estimated thresholds in the second step (ETH2), number of observation in each regime (#Obs.), Bayesian information criterion (BIC), joint sum-of-squared error (jSSE), individual sum-of-squared error (SSE)

	aBCD	*gLAR*
ETH1	(2.538, 2.894, 3.340, 3.490, 3.629)	(2.033, 2.556, 2.719, 3.111, 3.359, 3.533, 3.800)
	aBCD-BEA	*gLAR-BEA*
ETH2	(2.894, 3.340, 3.490, 3.629)	(3.359)
#Obs	(55, 23, 11, 11, 6)	(76, 30)
BIC	−348.12	−337.81
jSSE	1.648	3.513
SSE	(0.890, 0.359, 0.260, 0.139, 0.000)	(1.978, 1.535)

We set autoregressive order $p = 8$ for our SETAR model, as in Tong and Lim (1980). We then applied *aBCD-BEA* and *gLAR-BEA* with $k_{\max} = 7$ and $\Delta_* = 10$. For BEA, the criterion penalty is set at $c_{\mathrm{E}} = 5$.

The results of the two-step estimation methods are given in Table 6. From the results, we observe that the *aBCD* and *gLAR* estimate five and seven change-points, respectively. However, none of these change-points are commonly estimated by both methods and this is might be due to *gLAR*'s tendency to estimate way more irrelevant thresholds compared to *aBCD*. In the final threshold estimate, the *aBCD-BEA* and *gLAR-BEA* retain four and one thresholds, respectively. The BIC results indicate that four thresholds estimated by *aBCD-BEA* yield lower jSSE and BIC, indicating better fit.

The estimated threshold 2.894, via *aBCD-BEA*, is very close to the one obtained by Li and Tong (2001) (2.946) via classical inference, and by both Geweke and Terui (1993) and Chen et al. (2011) via Bayesian inference (3.00 and 2.94, respectively). Note that Li and Tong (2001), Geweke and Terui (1993) and Chen et al. (2011) only consider two-regime SETAR models with $d = 3$. The remaining three thresholds that we have estimated earlier may provide important information for the additional non-linear behavior of the transformed lynx time series.

Lopes and Salazar (2006) reported several root mean squared errors (RMSE) for four different nonlinear models, where their two-regime smooth logistic transition autoregressive model with $d = 3$ and $p = 11$ or LSTAR(11) had the lowest RMSE (0.153) among all those four models. Our computed RMSE, using $\sqrt{\sum_{t=p+1}^{n}(\widehat{y}_t - y_t)^2/N}$, for our five-regime SETAR(8) is 0.136, which is lower than the RMSE of LSTAR(11) model obtained by Lopes and Salazar (2006), indicating our five-regime model fits better than their two-regime LSTAR(11) model.

8 Final remarks

In this paper, we have developed an *active-set* based block coordinate descent to exactly optimize the group LASSO for the threshold model without orthogonalizing the design matrix. Furthermore, the backward elimination algorithm is utilized to consistently estimate relevant thresholds from the threshold set obtained by the group LASSO. Empirical studies using this univariate model shows that the *aBCD* algorithm estimates less irrelevant thresholds compared to the approximation group LASSO algorithms of *gLAR*. Furthermore, the *aBCD-BEA* performs better in terms of correctly estimating the number of thresholds in simulation studies, and in identifying important thresholds in case studies compared to the *gLAR-BEA*. Codes for the datasets and algorithms are available in https://github.com/jaffrinasir/Algorithms. Note that the *aBCD* algorithm can be extended for multivariate SETAR model and the details are given in Nasir (2020).

It is possible to further improve the performance of estimating relevant thresholds in the first-step procedure by introducing appropriate adaptive weights for gLASSO (Wang and Leng 2008), or non-convex penalization approaches such as the group smooth clipped absolute deviation (SCAD) and the group minimax concave penalty (MCP) suggested by Huang et al. (2012). In addition, it maybe possible to speed up the computation of *aBCD* using parallel computing or majorization-minimization (MM) techniques (Bradley et al. 2011; Yang and Zou 2014a, b; Jiang and Huang 2014) and also study the predictive performance of gLASSO for change-point/threshold estimation. We leave these extensions to future work.

Acknowledgements This research is a part of the first author's Ph.D. study at The University of Western Australia. The author would like to thank The Ministry of Higher Education Malaysia and Universiti Malaysia Kelantan for sponsoring his Ph.D. study; and The University of Western Australia for approving travel grants to Flinders University for a research visit, and Canberra for attending Australian Statistical Conference.

Appendix

Proof of Theorem 5.1 Since the vector of parameters $\boldsymbol{\theta}^N$ might be groupwise-sparse and X in (7) is a block lower triangular matrix, (12) can be simplified as

$$\sum_{l=j}^{N} \mathbf{x}_{\pi(l)} y_{\pi_l+d} - \sum_{i\in\mathcal{B}} \left(\sum_{h=\max(i,j)}^{N} \mathbf{x}_{\pi_h}\mathbf{x}_{\pi_h}^T \right) \boldsymbol{\theta}_{\pi_i} = \frac{N\lambda_n}{2}\widetilde{\boldsymbol{e}}_j. \tag{24}$$

By splitting the second term in the L.H.S of (24), we write

$$\sum_{i\in\mathcal{B}} \left(\sum_{h=\max(i,j)}^{N} \mathbf{x}_{\pi_h}\mathbf{x}_{\pi_h}^T\right)\boldsymbol{\theta}_{\pi_i} := \boldsymbol{g}_j(\mathcal{B}) + \sum_{l=j}^{N} \left(\mathbf{x}_{\pi_l}\mathbf{x}_{\pi_l}^T\right)\boldsymbol{\theta}_{\pi_j},$$

where

$$\boldsymbol{g}_j(\mathcal{B}) = \begin{cases} \mathbf{0}, & \operatorname{card}(\mathcal{B}) \leq 1, \\ \sum_{\substack{i \in \mathcal{B} \\ i \neq j}} \left(\sum_{h=\max(i,j)}^{N} \mathbf{x}_{\pi_h} \mathbf{x}_{\pi_h}^T \right) \boldsymbol{\theta}_{\pi_i}, & \operatorname{card}(\mathcal{B}) > 1, \end{cases}$$

Since $\widetilde{\boldsymbol{e}}_j = \boldsymbol{\theta}_{\pi_j} / \left\| \boldsymbol{\theta}_{\pi_j} \right\|_2$, for all $j \in \mathcal{B} \backslash \{1\}$, (24) can be written as

$$\sum_{l=j}^{N} \mathbf{x}_{\pi(l)} y_{\pi_l + d} - \left[\sum_{\substack{i \in \mathcal{B} \\ i \neq j}} \left(\sum_{h=\max(i,j)}^{N} \mathbf{x}_{\pi_h} \mathbf{x}_{\pi_h}^T \right) \boldsymbol{\theta}_{\pi_i} + \sum_{l=j}^{N} \left(\mathbf{x}_{\pi_l} \mathbf{x}_{\pi_l}^T \right) \boldsymbol{\theta}_{\pi_j} \right] = \frac{N \lambda_n}{2} \frac{\boldsymbol{\theta}_{\pi_j}}{\left\| \boldsymbol{\theta}_{\pi_j} \right\|_2}. \tag{25}$$

The above equation can be rewritten as

$$\boldsymbol{\theta}_{\pi_j} = \left(\sum_{l=j}^{N} \mathbf{x}_{\pi_l} \mathbf{x}_{\pi_l}^T + \frac{N \lambda_n}{2 \left\| \boldsymbol{\theta}_{\pi_j} \right\|_2} \mathbf{I}_{p+1} \right)^{-1} \boldsymbol{f}_j(\mathcal{B}), \tag{26}$$

where

$$\boldsymbol{f}_j(\mathcal{B}) = \sum_{l=j}^{N} \mathbf{x}_{\pi(l)} y_{\pi_l + d} - \boldsymbol{g}_j(\mathcal{B}). \tag{27}$$

While (26) gives an explicit expression for $\boldsymbol{\theta}_{\pi_j}$, $\left\| \boldsymbol{\theta}_{\pi_j} \right\|_2$ is part of LHS for $j \in \mathcal{B} \backslash \{1\}$.

When $\left(2 \left\| \boldsymbol{f}_j(\mathcal{B}) \right\|_2 / N \right) > \lambda_n$, the Eq. (26) with $u_j = \left\| \boldsymbol{\theta}_{\pi(j)} \right\|_2 > 0$ can be written as

$$\boldsymbol{\theta}_{\pi_j} = U_j^T \left(D_j + \frac{N \lambda_n}{2 u_j} \mathbf{I}_{p+1} \right)^{-1} U_j \boldsymbol{f}_j(\mathcal{B}),$$

and observe that

$$\begin{aligned} u_j^2 &= \left\| \boldsymbol{\theta}_{\pi(j)} \right\|_2^2 \\ &= \left\| U_j^T \left(D_j + \frac{N \lambda_n}{2 u_j} \mathbf{I}_{p+1} \right)^{-1} U_j \boldsymbol{f}_j(\mathcal{B}) \right\|_2^2 = \left\| \left(D_j + \frac{N \lambda_n}{2 u_j} \mathbf{I}_{p+1} \right)^{-1} U_j \boldsymbol{f}_j(\mathcal{B}) \right\|_2^2 \\ &= \left\| \begin{pmatrix} \left(d_{j,1} + \frac{N \lambda_n}{2 u_j} \right)^{-1} & 0 & \cdots & 0 \\ 0 & \left(d_{j,2} + \frac{N \lambda_n}{2 u_j} \right)^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \left(d_{j,p+1} + \frac{N \lambda_n}{2 u_j} \right)^{-1} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_{j,p+1} \end{pmatrix} \right\|_2^2 \end{aligned}$$

$$= \left\| \begin{pmatrix} \nu_{j,1}\left(d_{j,1} + \frac{N\lambda_n}{2u_j}\right)^{-1} \\ \nu_{j,2}\left(d_{j,2} + \frac{N\lambda_n}{2u_j}\right)^{-1} \\ \vdots \\ \nu_{j,p+1}\left(d_{j,p+1} + \frac{N\lambda_n}{2u_j}\right)^{-1} \end{pmatrix} \right\|_2^2 = \sum_{k=1}^{p+1} \frac{\nu_{j,k}^2}{\left(d_{j,k} + \frac{N\lambda_n}{2u_j}\right)^2} = \sum_{k=1}^{p+1} u_j^2 \frac{\nu_{j,k}^2}{\left(d_{j,k}u_j + \frac{N\lambda_n}{2}\right)^2},$$

that is

$$1 = \sum_{k=1}^{p+1} \frac{\nu_{j,k}^2}{\left(d_{j,k}u_j + \frac{N\lambda_n}{2}\right)^2}.$$

When $\left(2 \left\| \boldsymbol{f}_j(\mathcal{B}) \right\|_2 / N\right) \le \lambda_n, \boldsymbol{\theta}_{\pi_j} = \mathbf{0}$ due to the condition (II) in Lemma 4.1. For $j = 1$, the solution in (26) to the non-penalized $\boldsymbol{\theta}_{\pi_1}$ is simply

$$\boldsymbol{\theta}_{\pi_1} = U_1^T D_1^{-1} U_1 \boldsymbol{f}_j(\mathcal{B}) = \left(\sum_{l=1}^{N} \mathbf{x}_{\pi_l} \mathbf{x}_{\pi_l}^T\right)^{-1} \boldsymbol{f}_j(\mathcal{B}), \tag{28}$$

where $\boldsymbol{f}_j(\mathcal{B}) = \sum_{l=1}^{N} \mathbf{x}_{\pi(l)} y_{\pi_l+d} - \boldsymbol{g}_1(\mathcal{B}), \boldsymbol{g}_1(\mathcal{B}) = \sum_{\substack{i\in\mathcal{B} \\ i\neq 1}} \left\{\sum_{h=\max(i,1)}^{N} \mathbf{x}_{\pi_h} \mathbf{x}_{\pi_h}^T\right\} \boldsymbol{\theta}_{\pi_i}$ if $\mathrm{card}(\mathcal{B}) > 1$, otherwise $\boldsymbol{g}_1(\mathcal{B}) = \mathbf{0}$. Hence the proof. □

References

Bach FR (2008) Consistency of the group LASSO and multiple kernel learning. J Mach Learn Res 9:1179–1225

Bai J, Perron P (2003) Computation and analysis of multiple structural change models. J Appl Econom 18(1):1–22

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of LASSO and Dantzig selector. Ann Stat 37(4):1705–1732

Boysen L, Kempe A, Liebscher V, Munk A, Wittich O (2009) Consistencies and rates of convergence of jump-penalized least squares estimators. Ann Stat 37(1):157–183

Bradley JK, Kyrola A, Bickson D, Guestrin C (2011) Parallel coordinate descent for L1-regularized loss minimization. In: Proceedings of the 28th international conference on machine learning. ICML (1998), pp 321–328

Bulmer MG (1974) A statistical analysis of the 10-year cycle in Canada. J Anim Ecol 43(3):701–718

Chan KS (1993) Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. Ann Stat 21(1):520–533

Chan KS, Tong H (1990) On likelihood ratio tests for threshold autoregression. J R Stat Soc Ser B (Stat Methodol) 52(3):469–476

Chan KS, Petruccelli JD, Tong H, Woolford SW (1985) A multiple-threshold AR(1) model. J Appl Probab 22(2):267–279

Chan WS, Wong ACS, Tong H (2004) Some nonlinear threshold autoregressive time series models for actuarial use. N Am Actuar J 8(4):37–61

Chan NH, Yau CY, Zhang RM (2014) Group LASSO for structural break time series. J Am Stat Assoc 109(506):590–599

Chan NH, Yau CY, Zhang RM (2015) LASSO estimation of threshold autoregressive models. J Econom 189(2):285–296

Chan NH, Ing CK, Li Y, Yau CY (2017) Threshold estimation via group orthogonal greedy algorithm. J Bus Econ Stat 35(2):334–345

Chen R (1995) Threshold variable selection in open-loop threshold autoregressive models. J Time Ser Anal 16(5):461–481

Chen CWS, Chi F, Gerlach R (2011a) Bayesian subset selection for threshold autoregressive moving-average models. Comput Stat 26:1–30

Chen CWS, Liu FC, So MKP (2011b) A review of threshold time series models in finance. Stat Interface 4(2):167–181

Ciuperca G (2011) Estimating nonlinear regression with and without change-points by the LAD method. Ann Inst Stat Math 63(4):717–743

Coakley J, Fuertes AM, Pérez MT (2003) Numerical issues in threshold autoregressive modeling of time series. J Econ Dyn Control 27:2219–2242

Cryer JD, Chan KS (2008) Time series analysis. With applications in R. Springer, Berlin

Eddelbuettel D, Francois R (2011) Rcpp: seamless R and C++ integration. J Stat Softw 40(8):1–18

Fan J, Yao Q (2003) Nonlinear time series: nonparametric and parametric methods. Springer-Verlag, New York

Foygel R, Drton M (2010) Exact block-wise optimization in group LASSO and sparse group LASSO for linear regression. Arxiv Preprint, pp 1–19. arXiv:1010.3320

Geweke J, Terui N (1993) Bayesian threshold autoregressive models for nonlinear time series. J Time Ser Anal 14(5):441–454

Gonzalo J, Pitarakis JY (2002) Estimation and model selection based inference in single and multiple threshold models. J Econom 110(2):319–352

Hansen BE (2000) Sample splitting and threshold estimation. Econometrica 68(3):575–603

Harchaoui Z, Lévy-Leduc C (2010) Multiple change-point estimation with a total variation penalty. J Am Stat Assoc 105(492):1480–1493

Huang J, Breheny P, Ma S (2012) A selective review of group selection in high-dimensional models. Stat Sci 27(4):481–499

Jiang D, Huang J (2014) Majorization minimization by coordinate descent for concave penalized generalized linear models. Stat Comput 24(5):871–883

Li D, Ling S (2012) On the least squares estimation of multiple-regime threshold autoregressive models. J Econom 167(1):240–253

Li WK, Tong H (2001) Time series: advanced methods. In: Smelser NJ, Baltes PB (eds) International encyclopedia of the social & behavioral sciences. Pergamon, Oxford, pp 15699–15704

Li D, Tong H (2016) Nested sub-sample search algorithm for estimation of threshold models. Stat Sin 26(4):1543–1554

Lopes HF, Salazar E (2006) Bayesian model uncertainty in smooth transition autoregressions. J Time Ser Anal 27(1):99–117

Nardi Y, Rinaldo A (2008) On the asymptotic properties of the group LASSO estimator for linear models. Electron J Stat 2:605–633

Nasir MJM (2020) LASSO-type estimations for threshold autoregressive and heteroscedastic time series models. PhD Thesis, School of Mathematics, Physics and Computing, The University of Western Australia

Niglio M, Vitale CD (2015) Threshold vector ARMA models. Commun Stat Theory Methods 44(14):2911–2923

Osborne MR, Presnell B, Turlach BA (2000) On the LASSO and its dual. J Comput Graph Stat 9(2):319–337

Pan J, Xia Q, Liu J (2017) Bayesian analysis of multiple thresholds autoregressive model. Comput Stat 32(1):219–237

Qian L (1998) On maximum likelihood estimators for a threshold autoregression. J Stat Plan Inference 75:21–46

Qian J, Su L (2016) Shrinkage estimation of regression models with multiple structural changes. Econom Theory 32:1376–1433

Roth V, Fischer B (2008) The group-LASSO for generalized linear models: uniqueness of solutions and efficient algorithms. In: Proceedings of the 25th international conference on machine learning. pp 848–855

Tibshirani RJ (2013) The LASSO problem and uniqueness. Electron J Stat 7(1):1456–1490

Tong H (1978) On a threshold model. In: Chen CH (ed) Pattern recognition and signal processing. Sijthoff and Noodhoff, Alphen aan den Rijn, pp 101–141

Tong H (1990) Non-linear time series. A dynamical system approach. Oxford University Press, New York

Tong H, Lim KS (1980) Threshold autoregression, limit cycles and cyclical data. J R Stat Soc Ser B (Stat Methodol) 42(3):245–292

Tsay RS (1989) Testing and modeling threshold autoregressive processes. J Am Stat Assoc 84(405):231–240

Tsay RS (1998) Testing and modeling multivariate threshold models. J Am Stat Assoc 93(443):1188–1202

Tsay RS, Chen R (2018) Nonlinear time series analysis. Wiley, Hoboken

Wang H, Leng C (2008) A note on adaptive group LASSO. Comput Stat Data Anal 52(12):5277–5286

Wang H, Li B, Leng C (2009) Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc Ser B (Stat Methodol) 71(3):671–683

Weisberg S (2005) Applied linear regression, 3rd edn. Wiley, Hoboken

Yang Y, Zou H (2014a) A coordinate majorization descent algorithm for ℓ_1 penalized learning. J Stat Comput Simul 84(1):84–95

Yang Y, Zou H (2014b) A fast unified algorithm for solving group-LASSO penalize learning problems. Stat Comput 25(6):1129–1141

Yau CY, Hui TS (2017) LARS-type algorithm for group LASSO. Stat Comput 27:1041–1048

Yau CY, Tang CM, Lee TCM (2015) Estimation of multiple-regime threshold autoregressive models with structural breaks. J Am Stat Assoc 110(511):1175–1186

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Stat Methodol) 68(1):49–67

Zhao P, Yu B (2006) On model selection consistency of LASSO. J Mach Learn Res 7:2541–2563

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.