Statistics and Its Interface Volume 10 (2017) 343–353

Information criterion of seriously over-fitting change-point models

Chi Tim Ng* and Chun Yip Yau

It is shown that a general class of information criteria is able to rule out seriously over-fitting change-point models where the number of change points is comparable to the sample size. Equivalently speaking, it is not necessary to impose a pre-specified upper bound on the number of change points when we search for the optimal solution as in Bardet, Kengne, and Wintenberger (2012). For the time series with finite but unknown number of change points, the model with consistently estimated number of change points tends to be preferred to any other models (even seriously over-fitting) under such a class of information criteria. The results hold under a broad class of time series model introduced in Bardet and Wintenberger (2009) that includes ARMA-GARCH as a special case. Since exhaustive search of all possible change-point models for the optimal information criterion value is computationally infeasible, it is common to impose certain restrictions on the searching range. The applications of the information criterion to the restricted search of the optimal model are also discussed.

AMS 2000 subject classifications: Primary 62M10; secondary 62M09.
Keywords and phrases: ARMA-GARCH, Bayesian information criterion, Causal process, Change point, Consistency.

1. INTRODUCTION

The change-point detection has received considerable attention in various disciplines including econometrics, meteorology, engineering, and biomedicine. For example, [6] studied the abrupt climate changes in the historical data. [14] considered the segmentation of music video as a change-point detection problem. [16] studied the change-points in the copy number variation data and identify the location of abnormality in the chromosome of a tumor cell. [1] applied the change-point detection method to the volatility of the financial markets.

By assuming that the number of change points is known, change-point detection can be done by maximizing the likelihood function or other objective functions like sum of squares. Under this framework, [2] and [3] proposed a dynamic programming algorithm for the maximum likelihood estimation and developed an asymptotic theory of the maximum likelihood estimator. See also [24] for a summary of this approach. Methods are also developed to test the existence of change-points, for example, [22].

If the true number of change points is unknown, [4] suggested to select the optimal change-point model based on the Bayesian information criterion and established the asymptotic theory under a broad class of time series models. Let n be the sample size of the data. Since it is computationally infeasible to compare all 2^{n-1} change-point models using the Bayesian information criterion, it is necessary to impose certain restrictions on the change-point models. For example, [4] only select the optimal model from those with $\leq K_{\max}$ change points, where $K_{\max}$ is a user-specified fixed upper bound. Some other authors only compare the models obtained from the sequential algorithms, see e.g., the sequential cusum method of [17], [19], and [7], the pruned exact linear time (PELT) of [18], the simultaneous multiscale change-point estimator (SMUCE) of [13], and the pruned dynamic programming segmentation algorithm (SEG) of [25]. Recently, [16] considered the penalized likelihood method of change-point detection based on the LASSO penalty of [27]. This method has also been studied in [8], [9], [15], and [26]. It takes a tuning parameter λ as input and generates a piecewise constant function as output. The Bayesian information criterion can further be used to compare the piecewise constant functions generated from different values of λ. It is noteworthy that when λ is close to zero, the number of change points in the generated piecewise constant function can be comparable to the sample size n. The theory developed in [4] cannot handle the "seriously" over-fitting model with diverging number of change points directly and thus cannot guarantee that the Bayesian information criterion prefers the consistent model (in terms of correct identification of the number of change points) to the over-fitting model generated from small λ. All the above-mentioned methods actually generate a solution set (change-point models) from which the optimal solution is further chosen by certain information criteria, especially the Bayesian information criterion and its variants.

Information criterion can serve as a general method of comparing the models obtained from any solution-set-generating methods only if the theory of the "seriously" over-fitting model with diverging number of change points is available. The objective of this paper is to establish the

*Corresponding author.

consistency theory of the number of change points without placing any fixed upper bound on the number of change points in the solution. The theory developed in this paper is applicable to all situations provided that the solution set is shown to be including certain consistent models with probability goes to one. The ideas are elaborated in section 4 in this paper.

This paper is organized as follows. In section 2, the change-point model is specified and a class of information criteria encompassing Bayesian information criterion and Akaike information criterion as special cases are described. The new consistency theory of the number of change points is established in section 3. The implications of the new theory on the change-point detection are discussed in section 4. Simulation studies are given in section 5. The finite-sample performance of a variety of criteria including Bayesian information criterion and Akaike information criterion are compared. The application of information criterion is illustrated in section 6 via the US monthly purchasing manager index data.

2. CHANGE-POINT MODEL

Consider the model of [23], which is closely related to those used in [4] and [5]. Let $X_1, X_2, \ldots, X_n$ be the observable time series generated from the following multidimensional causal process,

$$\begin{aligned} X_t &= M_t \epsilon_t + f_t \\ M_t &= \mathcal{M}(\theta^0_{M,t}, X_{t-1}, X_{t-2}, \ldots) \\ f_t &= \mathcal{F}(\theta^0_{f,t}, X_{t-1}, X_{t-2}, \ldots) . \end{aligned}$$

Here, $\theta^0_t = (\theta^0_{M,t}, \theta^0_{f,t})$, $t = 1, 2, \ldots, n$ is p-dimensional vector-valued and is piecewise constant with k change points. Suppose that k is unknown but finite. Following [4] and [20], assume that for some $0 = q^{(0)} < q^{(1)} < q^{(2)} < \ldots < q^{(k)} < q^{(k+1)} = 1$, $\theta^0_t = \theta^0_{[nq^{(\ell)}]}$ for $[nq^{(\ell-1)}] \le t < [nq^{(\ell)}]$, $\ell = 1, \ldots, k+1$. Here, $[x]$ is the greatest integer less than or equal to x. It is convenient to set $[nq^{(0)}] = 1$ and $[nq^{(k+1)}] = n+1$. The random variables ϵ_t are independent and identically-distributed with mean zero and variance one. $\mathcal{M}$ and $\mathcal{F}$ are two given functions and are used to describe the evolution of the volatility and mean respectively. For convenience, assume that the stochastic process X_t, $t = [nq^{(1)}]-1, [nq^{(1)}]-2, \ldots, 2, 1, 0, -1, -2, \ldots$ is strictly stationary.

For any sequence of p-dimensional vectors $\theta = \{\theta_t = (\theta_{M,t}, \theta_{f,t}) : t = 1, 2, \ldots, n\}$, define the quasi log-likelihood function

(1)

$$\begin{aligned} L(\theta) &= \sum_{t=1}^{n} C_t(\theta_t) \\ &= -\frac{1}{2} \sum_{t=1}^{n} \left\{ \log |H_t(\theta_{M,t})| + tr(H_t^{-1}(\theta_{M,t}) S_t(\theta_{f,t})) \right\} . \end{aligned}$$

Here, $H_t(\theta_{M,t}) = M_t(\theta_{M,t}) M_t^T(\theta_{M,t})$, $S_t(\theta_{f,t}) = (X_t - f_t(\theta_{f,t}))(X_t - f_t(\theta_{f,t}))^T$ and $(M_t, f_t)(\theta)$ are obtained recursively by the relationships,

$$\begin{aligned} M_t(\theta_{M,t}) &= \mathcal{M}(\theta_{M,t}, X_{t-1}, X_{t-2}, \ldots, X_1, 0, 0, \ldots) , \\ f_t(\theta_{f,t}) &= \mathcal{F}(\theta_{f,t}, X_{t-1}, X_{t-2}, \ldots, X_1, 0, 0, \ldots) , \end{aligned}$$

for $t = 1, 2, \ldots, n$. In the above recursive formulas, the observations before $t = 1$ are truncated and replaced by zero. Note that certain assumptions on the positive definiteness of $H_t(\theta_{M,t})$ are required. Roughly speaking, X_t cannot be a deterministic sequence. Moreover, there should be no deterministic relationship between the components in X_t. The required assumptions are stated in the next section.

3. INFORMATION CRITERION

A class of information criterion is defined as follows. Let Θ be a p-dimensional compact space. For any $k' = 1, 2, \ldots, n$ and $1 = t^{(0)} < t^{(1)} < t^{(2)} < \ldots < t^{(k')} < t^{(k'+1)} = n+1$, define $\hat{\theta}(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')})$ as the maximum point of $L(\theta)$ within Θ^n subjected to the constraints that $\theta_t = \theta_{t^{(\ell-1)}}$ for $t^{(\ell-1)} \le t < t^{(\ell)}$, $\ell = 1, \ldots, k'+1$. Since Θ is compact, the maximum of $L(\theta)$ must exists and $L(\hat{\theta}(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')}))$ is well-defined. However, it should be noted that when $t^{(\ell)} - t^{(\ell-1)}$ is less than the number of unknown parameters, the corresponding θ is not unique. In such cases, choose any one of such θ values. The proofs of the main results in the next section do not depend on such a choice. Let D_n be a monotonic increasing deterministic sequence. Define the information criterion as

$$\begin{aligned} & IC(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')}) \\ = \; & -2L(\hat{\theta}(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')})) + D_n k' \psi(n) . \end{aligned}$$

Below are some examples of the function $\psi(n)$,

- Akaike information criterion (AIC): $\psi(n) = 2p$,
- Small-sample-corrected AIC (AICc): $2p\left(1 + \frac{p+1}{n-p-1}\right)$,
- Bayesian information criterion (BIC): $\psi(n) = p \log(n)$,
- Minimum description length (MDL): $\psi(n) = \log(k') + (k'+1)(\log(n) + \log(p)) + \frac{p+2}{2} \sum_{j=1}^{k'+1} \log(n_j)$,

where n_j is the length of the j-th segment. Roughly speaking, all these four information criteria discourage over-fitting ($k' > k$) by imposing a large penalty term $D_n k' \psi(n)$ on large k'. Traditionally, $D_n = 1$ is chosen in the regression analysis literature particularly when the number of covariates is finite. Recently, even greater D_n is chosen for the high-dimensional problems, see e.g., [28] and [12]. The choice of D_n for the change-point problems is discussed in Theorem 3.1.

The objective of this section is to show that the optimal model with smallest $IC(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')})$ has $k' = k$. [4] considered the optimal model among those with $k' \le K_{\max}$, a pre-specified fixed upper bound. In this paper, we further

prove that such restricted optimal model of [4] indeed has smaller IC value than any other models even with $k' > K_{\max}$.

3.1 Notation and assumptions

First, notation to be used in the main results is introduced. Similar notation is also used in [4], [5], and [23]. For $\ell = 0, 1, \ldots, k$, $t = 0, \pm 1, \pm 2, \ldots$, let ℓ be the regime at which t is located, and $\epsilon_t^{(\ell)}$ be independent and identically distributed such that for $[nq^{(\ell)}] \le t < [nq^{(\ell+1)}]$, $\epsilon_t^{(\ell)} = \epsilon_t$. Define $X_t^{(\ell)}$, $t = 0, \pm 1, \pm 2, \ldots$ via

$$\begin{aligned} X_t^{(\ell)} &= M_t^{(\ell)} \epsilon_t^{(\ell)} + f_t^{(\ell)}\,, \\ M_t^{(\ell)} &= \mathcal{M}(\theta^0_{M,[nq^{(\ell-1)}]}, X_{t-1}^{(\ell)}, X_{t-2}^{(\ell)}, \ldots)\,, \\ f_t^{(\ell)} &= \mathcal{F}(\theta^0_{f,[nq^{(\ell-1)}]}, X_{t-1}^{(\ell)}, X_{t-2}^{(\ell)}, \ldots)\,. \end{aligned}$$

Let $(\tilde{X}, \tilde{M}, \tilde{f})_t = (X, M, f)_t^{(0)}$ for $t < [nq^{(1)}]$ and $(\tilde{X}, \tilde{M}, \tilde{f})_t = (X, M, f)_t^{(\ell)}$ for $[nq^{(\ell)}] \le t < [nq^{(\ell+1)}]$. For $[nq^{(\ell)}] \le t < [nq^{(\ell+1)}]$, define

$$\begin{aligned} \tilde{M}_{M,t}(\theta_t) &= \mathcal{M}(\theta_{M,t}, X_{t-1}^{(\ell)}, X_{t-2}^{(\ell)}, \ldots)\,, \\ \tilde{f}_{f,t}(\theta_t) &= \mathcal{F}(\theta_{F,t}; X_{t-1}^{(\ell)}, X_{t-2}^{(\ell)}, \ldots)\,. \\ \tilde{C}_t(\theta_t) &= -\frac{1}{2}\left\{\log|\tilde{H}_t(\theta_{M,t})| + tr[\tilde{H}_t^{-1}(\theta_{M,t})\tilde{S}_t(\theta_{f,t})]\right\}\,. \end{aligned}$$

Here, $\tilde{H}_t(\theta_{M,t}) = \tilde{M}_t(\theta_{M,t})\tilde{M}_t^T(\theta_{M,t})$ and $\tilde{S}_t(\theta_{f,t}) = (\tilde{X}_t - \tilde{f}_t(\theta_{f,t}))(\tilde{X}_t - \tilde{f}_t(\theta_{f,t}))^T$.

The following assumptions are used.

(A1) For any $\epsilon > 0$,

$$\sup_{\theta \in \Theta^n} \max_{1 \le a < b \le n} (b-a)^{-1/2} \left| \sum_{t=a}^{b} \left\{ \partial^m \tilde{C}_t(\theta_t) - E[\partial^m \tilde{C}_t(\theta_t)] \right\} \right|$$

is $o_p(n^\epsilon)$. Here, ∂^m is any m-th order partial derivative w.r.t. components of θ_t, where $m \le 3$.

Following the arguments of Lemma 3.1 in [21], condition (A1) holds under the extra assumption that for all $s > 0$,

$$E\left| \sum_{t=a}^{b} \left\{ \partial^m \tilde{C}_t(\theta_t) - E[\partial^m \tilde{C}_t(\theta_t)] \right\} \right|^{2s} \le K_s (b-a)^s$$

for some $K_s > 0$. This can be satisfied easily by the time series fulfilling certain mixing conditions (i.e. weakly dependent, roughly speaking), for example, stationary time series with geometrically decaying covariance structure. It is possible to establish condition (A1) rigorously under some assumptions. However, this involves proofs that would be too lengthy for this paper aiming at the information criterion theory. In this paper, we assume condition (A1) without proof.

(A2) The smallest eigenvalue of the Hessian matrix $E[\nabla^2 \tilde{C}_t(\theta_t)]$ is bounded below by some positive constant not depending on θ and t. This condition indeed implies the identifiability of the model.

Conditions (A3) and (A4) below are similar to those adopted by [4], [5], and [23]. They are satisfied by many commonly used models including the GARCH model and ARMA model with finite autoregressive order and moving average order.

(A3) The following Lipschitz conditions are satisfied. There exists a sequence of non-negative real numbers $(\alpha, \beta, \gamma)_j^{(m)}$, $j = 1, 2, \ldots$ such that for all m, index set with cardinality ≤ 3,

$$\begin{aligned} &\sup_{\vartheta \in \Theta} \left\| \partial_{\theta_M}^{(m)} \mathcal{M}(\vartheta_M, x_1, x_2, \ldots) - \partial_{\theta_M}^{(m)} \mathcal{M}(\vartheta_M, y_1, y_2, \ldots) \right\| \\ &\le \sum_{j=1}^{\infty} \alpha_j^{(m)} \|x_j - y_j\| \\ &\sup_{\vartheta \in \Theta} \left\| \partial_{\theta_f}^{(m)} \mathcal{F}(\vartheta_f, x_1, x_2, \ldots) - \partial_{\theta_f}^{(m)} \mathcal{F}(\vartheta_f, y_1, y_2, \ldots) \right\| \\ &\le \sum_{j=1}^{\infty} \beta_j^{(m)} \|x_j - y_j\| \\ &\sup_{\vartheta \in \Theta} \left\| \partial_{\theta_M}^{(m)} \mathcal{H}(\vartheta_M, x_1, x_2, \ldots) - \partial_{\theta_M}^{(m)} \mathcal{H}(\vartheta_M, y_1, y_2, \ldots) \right\| \\ &\le \sum_{j=1}^{\infty} \gamma_j^{(m)} \|x_j - y_j\|\,. \end{aligned}$$

Here, $\mathcal{H} = \mathcal{M}\mathcal{M}^T$ and $\vartheta = (\vartheta_M, \vartheta_f)$. The sequence $(\alpha, \beta, \gamma)_j^{(m)}$ is bounded above by some $O(j^{-\kappa})$ quantity with $\kappa > 2$. In addition,

$$\sum_{j=1}^{\infty} \alpha_j^{(0)} + [E\epsilon_0^\rho]^{1/\rho} \sum_{j=1}^{\infty} \beta_j^{(0)} < 1$$

for some $\rho > 4$.

(A4) For $\vartheta \in \Theta$, the smallest eigenvalue of $\mathcal{M}(\vartheta_M, x_1, x_2, \ldots)$ is bounded below by some positive constant.

More technical assumptions are used in [4], [5], and [23]. To avoid repeating all these assumptions here, (A5) below is used instead for technical convenience. Indeed, following the lines of [4], [5], and [23], (A5) can be established rigorously under mild conditions. Therefore, condition (A5) is weak enough.

(A5) For any $0 < \nu < \rho$, we have

$$\begin{aligned} \mathrm{E}\|\tilde{X}_t\|^\nu < \infty \text{ and } \mathrm{E}\|X_t\|^\nu < \infty\,. \\ \mathrm{E}\|\tilde{M}_t\|^\nu < \infty \text{ and } \mathrm{E}\|M_t\|^\nu < \infty \\ \mathrm{E}\sup_{\Theta} \|\partial^m \tilde{M}_t(\theta_{M,t})\|^\nu < \infty \text{ and } \mathrm{E}\sup_{\Theta} \|\partial^m M_t(\theta_{M,t})\|^\nu < \infty \\ \mathrm{E}\sup_{\Theta} \|\partial^m \tilde{f}_t(\theta_{f,t})\|^\nu < \infty \text{ and } \mathrm{E}\sup_{\Theta} \|\partial^m f_t(\theta_{f,t})\|^\nu < \infty \end{aligned}$$

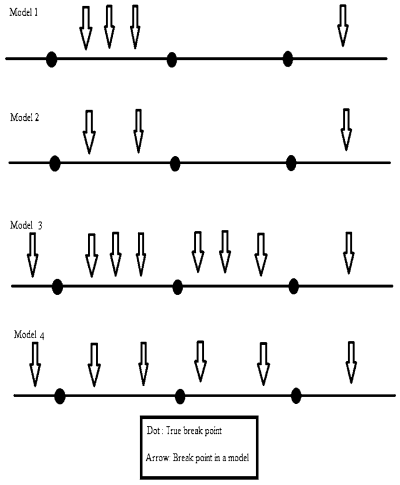

Figure 1. Merging of constant regimes.

3.2 Main theorem

The following theorem suggests that information criterion with appropriately chosen D_n and $\psi(n)$ can select the change-point model consistently in terms of the number of change points. This means that the consistent model is preferred to any other change-point models including those with diverging number of change points.

Theorem 3.1. *Choose $D_n = n^{\varpi}$ and $\psi(n) = O(n^{\varpi^*})$ for some $\varpi > 0$ and $\varpi^* > 0$ so that $\varpi + \varpi^* < 1$. Under conditions (A1)–(A5), $IC(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')})$ is minimized at $k' = k$ and $(t^{(1)}, t^{(2)}, \ldots, t^{(k')}) = (t_\dagger^{(1)}, t_\dagger^{(2)}, \ldots, t_\dagger^{(k)})$, where*

$$\begin{aligned}&(t_\dagger^{(1)}, t_\dagger^{(2)}, \ldots, t_\dagger^{(k)})\\ =\ &\operatorname{argmax}_{u^{(1)},u^{(2)},\ldots,u^{(k)}} L(\hat{\theta}(k, u^{(1)}, u^{(2)}, \ldots, u^{(k)}))\end{aligned}$$

with probability going to one as $n \to \infty$.

Proof. For arbitrarily given k' and $1 = t^{(0)} < t^{(1)} < t^{(2)} < \ldots < t^{(k')} < t^{(k'+1)} = n+1$, the ℓ-th regime $t^{(\ell-1)} \le t < t^{(\ell)}$ is called a constant regime if it contains no true change point.

Consider the change in the IC after merging adjacent constant regimes. The idea is illustrated in Figure 1. In this example, there are $k = 3$ true change points. In Model 1, there are five regimes. Regime 1 and 4 contain true change points. Regime 2 and 3 are consecutive constant regimes. Model 2 is essentially the same as Model 1 except that regime 2 and 3 are merged. After merging consecutive constant regimes, there is at most one constant regime between two consecutive non-constant regimes. Moreover, the number of non-constant regimes is at most k. Therefore, the merged model (Model 2) contains at most $2k$ change points, or equivalently $2k + 1$ regimes. If there are 3 true change points, the number of change points in the merged model must be bounded by 6, see Model 3 (before merging) and Model 4 (after merging). Let

$$\begin{aligned}\hat{\theta}^{(\ell)} &= \operatorname{argmax}_{\vartheta\in\Theta} \sum_{t=t^{(\ell-1)}}^{t^{(\ell)}-1} C_t(\vartheta)\,,\\ \tilde{\theta}^{(\ell)} &= \operatorname{argmax}_{\vartheta\in\Theta} \sum_{t=t^{(\ell-1)}}^{t^{(\ell)}-1} \tilde{C}_t(\vartheta)\,,\\ \hat{\theta}^{(\ell+1)} &= \operatorname{argmax}_{\vartheta\in\Theta} \sum_{t=t^{(\ell)}}^{t^{(\ell+1)}-1} C_t(\vartheta)\,,\\ \tilde{\theta}^{(\ell+1)} &= \operatorname{argmax}_{\vartheta\in\Theta} \sum_{t=t^{(\ell)}}^{t^{(\ell+1)}-1} \tilde{C}_t(\vartheta)\,,\\ \hat{\theta}^{(\ell,\ell+1)} &= \operatorname{argmax}_{\vartheta\in\Theta} \sum_{t=t^{(\ell-1)}}^{t^{(\ell+1)}-1} C_t(\vartheta)\,,\\ \tilde{\theta}^{(\ell,\ell+1)} &= \operatorname{argmax}_{\vartheta\in\Theta} \sum_{t=t^{(\ell-1)}}^{t^{(\ell+1)}-1} \tilde{C}_t(\vartheta)\,.\end{aligned}$$

It suffices to show that for any two adjacent constant regimes ℓ and $\ell+1$,

$$\begin{aligned}&\Delta(t^{(\ell-1)}, t^{(\ell)}, t^{(\ell+1)})\\ =\ &\sum_{t=t^{(\ell-1)}}^{t^{(\ell)}-1} C_t(\hat{\theta}^{(\ell)}) + \sum_{t=t^{(\ell)}}^{t^{(\ell+1)}-1} C_t(\hat{\theta}^{(\ell+1)})\\ &- \sum_{t=t^{(\ell-1)}}^{t^{(\ell+1)}-1} C_t(\hat{\theta}^{(\ell,\ell+1)})\end{aligned}$$

is bounded by an $O_p(n^\delta)$ quantity for any $0 < \delta < \varpi + \varpi^*$ and such a bound is uniform of the starting points and end points of the regimes, i.e. $(t^{(\ell-1)}, t^{(\ell)}, t^{(\ell+1)})$. If this is true, the IC of any model must be greater than that of its merged model. Therefore, any models that can further be merged must not be optimal. Since the number of change points in the merged model is bounded by $K_{\max} = 2k$, which is fixed (not to be confused with k'), the existing results of [4] is then applicable. Though such a value of $K_{\max}$ is unknown, it should be noted that the results of [4] only requires that $K_{\max}$ is fixed.

Let

$$\begin{aligned}&\tilde{\Delta}(t^{(\ell-1)}, t^{(\ell)}, t^{(\ell+1)})\\ =\ &\sum_{t=t^{(\ell-1)}}^{t^{(\ell)}-1} \tilde{C}_t(\tilde{\theta}^{(\ell)}) + \sum_{t=t^{(\ell)}}^{t^{(\ell+1)}-1} \tilde{C}_t(\tilde{\theta}^{(\ell+1)})\\ &- \sum_{t=t^{(\ell-1)}}^{t^{(\ell+1)}-1} \tilde{C}_t(\tilde{\theta}^{(\ell,\ell+1)})\,.\end{aligned}$$

Using the fact that for any functions $f(\theta)$ and $g(\theta)$, the supremum norm $\|f\|_\infty = \sup_\theta |f(\theta)|$ fulfills $-\|f-g\|_\infty \le$

$$\|f\|_\infty - \|g\|_\infty \le \|f-g\|_\infty,$$

$$\begin{aligned}(2)\qquad & |\Delta(t^{(\ell-1)},t^{(\ell)},t^{(\ell+1)}) - \tilde\Delta(t^{(\ell-1)},t^{(\ell)},t^{(\ell+1)})| \\ \le\ & 2\sup_{\vartheta\in\Theta}\sum_{t=t^{(\ell-1)}}^{t^{(\ell)}-1}|C_t(\vartheta)-\tilde C_t(\vartheta)| \\ & +2\sup_{\vartheta\in\Theta}\sum_{t=t^{(\ell)}}^{t^{(\ell+1)}-1}|C_t(\vartheta)-\tilde C_t(\vartheta)| \\ \le\ & 2\sup_{\vartheta\in\Theta}\sum_{t=1}^{n}|C_t(\vartheta)-\tilde C_t(\vartheta)| \\ \le\ & \sup_{\theta\in\Theta^n}\sum_{t=1}^{n}|C_t(\theta_t)-\tilde C_t(\theta_t)|\,.\end{aligned}$$

The items in (2) are handled in the following.

Bound of $\tilde\Delta(t^{(\ell-1)},t^{(\ell)},t^{(\ell+1)})$: Arbitrarily choose $0<\epsilon<\delta/2$. Conditions (A1) and (A2) guarantees that standard arguments based on the Taylor expansion can be used to establish

$$\sup_{t^{(\ell)}-t^{(\ell-1)}\ge n^\delta} n^\epsilon (t^{(\ell)}-t^{(\ell-1)})^{1/2}|\tilde\theta^{(\ell)}-\theta_t^0| = o_p(1)$$

and

$$\sup_{t^{(\ell)}-t^{(\ell-1)}\ge n^\delta}\sup_{\vartheta\in\Theta}(t^{(\ell)}-t^{(\ell-1)})^{-1}\left|\sum_{t=t^{(\ell-1)}}^{t^{(\ell)}-1}\nabla^2\tilde C_t(\vartheta)\right| = O_p(1)\,.$$

Moreover, condition (A2) and the compactness of the parameter space Θ give $|\tilde\theta^{(\ell)}-\theta_t^0| = O_p(1)$ and

$$\sup_{t^{(\ell)}-t^{(\ell-1)}<n^\delta}\sup_{\vartheta\in\Theta}\sum_{t=t^{(\ell-1)}}^{t^{(\ell)}-1}\nabla^2\tilde C_t(\vartheta) = O_p(n^\delta)\,.$$

Combining these results, the Taylor expansion at θ_t around $\tilde\theta^{(\ell)}$ suggests that $\sum_{t=t^{(\ell-1)}}^{t^{(\ell)}-1}[\tilde C_t(\tilde\theta^{(\ell)})-\tilde C_t(\theta_t^0)] \le O_p(n^\delta)$. Similarly, we have $\sum_{t=t^{(\ell)}}^{t^{(\ell+1)}-1}[\tilde C_t(\tilde\theta^{(\ell+1)})-\tilde C_t(\theta_t^0)] \le O_p(n^\delta)$ and $\sum_{t=t^{(\ell-1)}}^{t^{(\ell+1)}-1}[\tilde C_t(\tilde\theta^{(\ell,\ell+1)})-\tilde C_t(\theta_t^0)] \le O_p(n^\delta)$. Since regimes ℓ and $\ell+1$ are adjacent constant regimes, canceling out the terms $\tilde C_t(\theta_t^0)$ yields $\tilde\Delta^{(\ell,\ell+1)} = O_p(n^\delta)$.

Bound of $\sup_{\theta\in\Theta^n}\sum_{t=1}^n|C_t(\theta_t)-\tilde C_t(\theta_t)|$: In condition (A3), we assume $\kappa>2$ which is stronger than that used in [23]. Under such stronger condition, together with (A2), (A4), and (A5), following the lines in the proof of Lemma C.6 in [23], we have for all $0<\nu<\rho$,

$$(3)\qquad \sum_{t=1}^n\left\{E\|X_t-\tilde X_t\|^\nu\right\}^{1/\nu} = O(1)\,,$$

$$(4)\qquad \sum_{t=1}^n\left\{E\sup_{\theta\in\Theta^n}\|M_t(\theta_{M,t})-\tilde M_t(\theta_{M,t})\|^\nu\right\}^{1/\nu} = O(1)\,,$$

$$(5)\qquad \sum_{t=1}^n\left\{E\sup_{\theta\in\Theta^n}\|f_t(\theta_{f,t})-\tilde f_t(\theta_{f,t})\|^\nu\right\}^{1/\nu} = O(1)\,.$$

Then, the required bound can be established as follows,

$$\begin{aligned}& 2E\sup_{\theta\in\Theta^n}\sum_{t=1}^n|C_t(\theta_t)-\tilde C_t(\theta_t)| \\ \le\ & E\sup_{\theta\in\Theta^n}\sum_{t=1}^n\Big\{\log|H_t(\theta_{M,t})\tilde H^{-1}(\theta_{M,t})| \\ & +tr(H_t^{-1}(\theta_{M,t})S_t(\theta_{f,t})-\tilde H_t^{-1}(\theta_{M,t})\tilde S_t(\theta_{f,t}))\Big\} \\ \le\ & E\sup_{\theta\in\Theta^n}\sum_{t=1}^n\Big\{|\tilde H_t(\theta_{M,t})|^{-1}\cdot|H_t(\theta_{M,t})-\tilde H_t(\theta_{M,t})|\Big\} \\ & +E\sup_{\theta\in\Theta^n}\sum_{t=1}^n\Big\{tr[H_t^{-1}(\theta_{M,t})(H_t(\theta_{M,t})-\tilde H_t(\theta_{M,t})) \\ & \cdot\tilde H_t^{-1}(\theta_{M,t})\tilde S_t(\theta_{f,t})]\Big\} \\ & +E\sup_{\theta\in\Theta^n}\sum_{t=1}^n\Big\{tr[H_t^{-1}(\theta_{M,t})(S_t(\theta_{f,t})-\tilde S_t(\theta_{f,t}))]\Big\}\,.\end{aligned}$$

Using

$$\begin{aligned}& H_t(\theta_{M,t})-\tilde H_t(\theta_{M,t}) \\ =\ & M_t(\theta_{M,t})(M_t(\theta_{M,t})-\tilde M_t(\theta_{M,t}))^T \\ & +(M_t(\theta_{M,t})-\tilde M_t(\theta_{M,t}))\tilde M_t^T(\theta_{M,t})\end{aligned}$$

and

$$\begin{aligned}& S_t(\theta_{f,t})-\tilde S_t(\theta_{f,t}) \\ =\ & (X_t-f_t(\theta_{f,t}))(X_t-\tilde X_t-f_t(\theta_{f,t})+\tilde f_t(\theta_{f,t}))^T \\ & +(X_t-\tilde X_t-f_t(\theta_{f,t})+\tilde f_t(\theta_{f,t}))(\tilde X_t-\tilde f_t(\theta_{f,t}))\,,\end{aligned}$$

Holder's inequality, (3)–(5), condition (A4), and condition (A5), it can be shown that

$$E\sup_{\theta\in\Theta^n}\sum_{t=1}^n|C_t(\theta_t)-\tilde C_t(\theta_t)| = O(1)\,.$$

From the inequality (2) and the bounds obtained above, we have $\Delta(t^{(\ell-1)},t^{(\ell)},t^{(\ell+1)}) = O_p(n^\delta)$ for any $0<\delta<\varpi$. This completes the proof. □

4. IMPLICATIONS OF THE NEW CONSISTENCY RESULTS

Even though in practice it is computational infeasible to compare each of 2^{n-1} possible configurations $(k,t^{(1)},t^{(2)},\ldots,t^{(k)})$, Theorem 3.1 guarantees that the information criterion with appropriately chosen penalty terms D_n and $\psi(n)$ can be used to select the consistent change-point model from the solution set obtained by any solution-set-generating methods under certain conditions. Suppose

that there is a solution-set-generating method that takes a parameter λ as input and generates a piecewise constant function with configuration $(k'_\lambda, t^{(1)}_\lambda, t^{(2)}_\lambda, \ldots, t^{(k'_\lambda)}_\lambda)$ as output. If we are able to show that the consistent model with $k' = k$ and

$$\begin{aligned}&(t^{(1)}, t^{(2)}, \ldots, t^{(k')})\\ =\ &\text{argmax}_{u^{(1)},u^{(2)},\ldots,u^{(k)}}\, L(\hat{\theta}(k, u^{(1)}, u^{(2)}, \ldots, u^{(k)}))\end{aligned}$$

is included in the solution set with probability goes to one,

$$\lambda^* = \text{argmin}_\lambda\, IC(k'_\lambda, t^{(1)}_\lambda, t^{(2)}_\lambda, \ldots, t^{(k'_\lambda)}_\lambda)$$

must correspond to this consistent model since it is optimal in terms of the IC value according to Theorem 3.1. It is out of the scope of this paper to develop a solution-set-generating method fulfilling such requirement, but it is an interesting future research topic.

The conditions for establishing (i) and (ii) below have been widely discussed in the literature, e.g., [21], [4], [8], and [9];

(i) at least a model with $k' = k$ is included in the solution set and

(ii) $\Delta = \max_{1\le j\le k} \min_{1\le i\le k'} |n^{-1}t^{(i)} - q^{(j)}| \to^p 0$.

However, it is noteworthy that (i) and (ii) do not necessarily imply that λ^* gives the consistent model. It is difficult to rule out the possibility that some models with $k' \neq k$ has smaller IC value than some consistent but not optimal models.

To illustrate the ideas, consider the LASSO penalized likelihood method of [16]. In the cases where the time series is a sequence of independent Normal random variables with constant variance and time-varying mean, it takes λ as input and minimizes

$$\sum_{t=1}^{n}(X_t - \theta_t)^2 + \lambda\sum_{t=1}^{n-1}|\theta_t - \theta_{t+1}|$$

over $\theta \in \Theta^n$ for some compact space $\Theta \subset R$. Let $\mathcal{A}_n$ be the set of change points in the solution. [8] and [9] shown that if λ is chosen appropriately, Bayesian information criterion can select the consistent model with probability goes to one from the models with all change points belong to $\mathcal{A}_n$. However, in the finite-sample cases, it is difficult to tell if a particular value of λ falls within the range of consistency. Therefore, it is necessary to test many different possible values of λ, including those much smaller than the lower bound of the range. The Bayesian information criterion theory established in [8] and [9] cannot handle small λ below such lower bound. Theorem 3.1 in this paper fills some missing gaps by showing that the all over-fitting models are not optimal. However, it should be noted that though there exists λ corresponding to a consistent model, such consistent model is not necessarily the model with optimal IC value. It is still difficult to rule out the possibility that some inconsistent model have even better IC value.

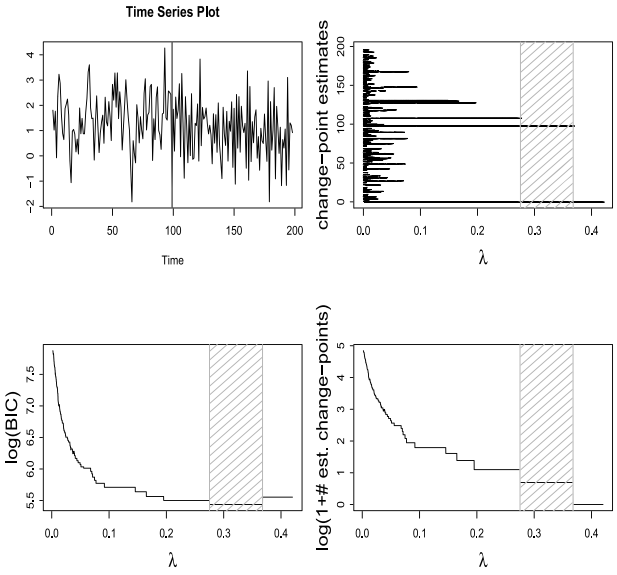

Figure 2. Time series plot, estimated change-points, BIC and the number of estimated change-points along the solution path for a realization of (6). The estimated change-point is in vertical line. The shaded area is the range of λs that achieves the minimum BIC.

5. SIMULATION STUDIES

Following the discussion in Section 4, in this section we conduct simulation experiments to illustrate the consistency of change-point estimation using different information criteria along the solution path of LASSO-type penalized likelihood methods. In particular, it is demonstrated that the over-fitting model generated from small penalization parameter λ does not affect the determination of the optimal solution.

5.1 Single change-point model

We consider the piecewise stationary autoregressive model

(6)
$$X_t = \begin{cases} 1 + 0.3X_{t-1} + 0.1X_{t-2} + \epsilon_t\,, & \text{if } 1 \le t \le [n/2]\,,\\ 1 - 0.3X_{t-1} + 0.2X_{t-2} + \epsilon_t\,, & \text{if } [n/2] < t \le n\,,\end{cases}$$

where $\epsilon_t \sim N(0,1)$. Using the notations in Section 3, we may express $\theta_t = (\phi_{0,t}, \phi_{1,t}, \phi_{2,t})$, where $\phi_{0,t}, \phi_{1,t}, \phi_{2,t}$ are the coefficient of the autoregressive model. Also, the number of change-points is $k = 1$, $q^{(1)} = 0.5$, $M^0_t = M^1_t = 1$ for all $t = 1, \ldots, n$, $\theta^0_0 = (1, 0.3, 0.1, 1)$, $\theta^1_0 = (1, -0.3, 0.2, 1)$, $f^0_t = 1 + 0.3X_{t-1} + 0.1X_{t-2}$ for $1 \le t < [nq^{(1)}]$ and $f^1_t = 1 - 0.3X_{t-1} + 0.2X_{t-2}$ for $[nq^{(1)}] \le t \le n$, $\tilde{H}_t(\theta_{M,t}) = 1$, and $S_t(\theta_{f,t}) = (X_t - \phi_{0,t} - \phi_{1,t}X_{t-1} - \phi_{2,t}X_{t-2})^2$. A realization of (6) is given in Figure 2.

The group LASSO solution path are the parameters that

Table 1. Estimation results for Model 6. Proportion of correct estimated number of breakpoints (Corr. #), the averages (ave) and the empirical standard deviations (e.s.d.) of the change-point estimates. True change-point is at $t = n/2$

		ϖ=0.05			ϖ=0.1			ϖ=0.2		
IC	n	Corr. #	ave	e.s.d.	Corr. #	ave	e.s.d.	Corr. #	ave	e.s.d.
	200	0.860	93.80	9.164	0.965	93.58	9.479	0.930	93.48	8.404
AIC	400	0.890	195.12	8.404	0.965	194.05	9.500	1.000	194.12	9.403
	800	0.925	394.24	8.054	0.995	393.79	8.980	1.000	392.04	10.46
	200	0.920	93.99	9.427	0.970	93.39	9.086	0.905	93.49	8.365
AICc	400	0.910	195.11	8.363	0.970	194.08	9.483	1.000	194.05	9.395
	800	0.935	394.07	8.600	0.995	393.56	8.270	1.000	392.04	10.46
	200	0.845	93.22	8.467	0.725	93.33	8.426	0.150	94.30	7.159
BIC	400	0.995	194.05	9.412	0.940	193.77	9.400	0.555	194.68	8.323
	800	1.000	392.04	10.46	1.000	392.04	10.46	0.850	391.25	10.82
	200	0.825	92.98	8.99	0.695	93.41	8.321	0.150	94.30	7.159
MDL	400	0.980	193.85	9.32	0.910	193.95	9.088	0.460	194.86	8.243
	800	1.000	392.04	10.46	0.995	391.99	10.47	0.825	391.16	10.95

minimize

$$-2L(\hat{\theta}(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')})) + \lambda \sum_{t=1}^{n-1} \|\theta_t - \theta_{t+1}\|_2 \,,$$

for different values of λ, where $\|\cdot\|_2$ is the Euclidean norm and $L(\theta)$ is given in (1). Note that group LASSO has to be used instead of the ordinary LASSO since each segment contains involves three parameters. The group LASSO estimation is implemented by the R-package *gglasso*. Figure 2 plots a typical solution path for a group LASSO estimation for a realization of (6) with $n = 200$. Also, the log of BIC for each value of λ is depicted. It can be seen that BIC successfully pick up the change-point around $n = 100$. Also, when λ is very small, the number of estimated change-points and BIC are both very large, which is consistent to our finding that the BIC is able to handle "seriously" over-fitting model.

Next we conduct large scale simulations to study the biases and variance of the change-point estimators. We compare several information criteria of the form

$$\begin{aligned} & IC(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')}) \\ = \; & -2L(\hat{\theta}(k', t^{(1)}, t^{(2)}, \ldots, t^{(k')})) + D_n k' \psi(n) \,, \end{aligned}$$

along the group LASSO solution path. In particular, we compare

- Akaike information criterion (AIC): $\psi(n) = 2p$,
- Small-sample-corrected AIC (AICc): $2p\left(1 + \frac{p+1}{n-p-1}\right)$,
- Bayesian information criterion (BIC): $\psi(n) = p\log(n)$,
- Minimum description length (MDL): $\psi(n) = \log(k') + (k'+1)(\log(n) + \log(p)) + \frac{p+2}{2}\sum_{j=1}^{k'+1}\log(n_j)$,

where n_j is the length of the j-th segment.

Table 1 reports the change-point estimation results for $D_n = n^{\varpi}$, $\varpi \in \{0.05, 0.1, 0.2\}$, $n \in \{200, 400, 800\}$ with 200 replications. First we interpret the performance of the BIC criterion. Table 1 suggests that the performance using the penalty $\varpi = 0.05$ is the best in terms of detecting the correct number of change-points. Also, the performance of change-point estimation improves as sample size increases. Moreover, when the sample size is large, the detection performance is not very sensitive to the choice of ϖ. On the other hand, the penalty $\varpi = 0.2$ appears to be too heavy especially for small sample size $n = 200$; the percentage of correct estimated number of change-point is only 15%.

Comparing the performance of different criteria in Table 1, we see that AIC and AICc perform similarly, and BIC and MDL perform similarly. This is because both AIC and AICc have penalty $\psi(n)$ of constant order, while both BIC and MDL have penalty $\psi(n)$ of order $\log(n)$. As AIC and AICc impose a smaller penalty, they perform better than BIC and MDL for small sample sizes. However, BIC and MDL tends to perform better as the sample size grows. In summary, using AIC or AICc with $\varpi = 0.2$, and BIC or MDL with $\varpi = 0.05$ give good performance in most cases.

5.2 Single change-point model with changes near the end-points

We consider the piecewise stationary autoregressive model

$$(7) \qquad X_t = \begin{cases} 0.75X_{t-1} + \epsilon_t \,, & \text{if } 1 \le t \le [\eta n] \,, \\ -0.5X_{t-1} + \epsilon_t \,, & \text{if } [\eta n] < t \le n \,, \end{cases}$$

where $\epsilon_t \sim N(0,1)$. The specification of k, q, M_t^0, M_t^1, θ_0^0, θ_0^1, f_t^0, f_t^1, $\tilde{H}_t(\theta_{M,t}) = 1$, and $S_{f,t}(\theta_t)$ are similar to that in Section 5.1. We consider the cases $\eta = 0.1$ and 0.9, that is, the change-point is close to the boundary of the data set. Model 7 with $n = 1024$ and $[\eta n] = 50$ was studied in [10]. The realizations of (7) with $\eta = 0.1$ is given in Figure 3. We consider the performance of change-point estimation using the four criteria along the group LASSO solution path. See Figure 3 for a typical solution path of a realization of (7) with $n = 200$ and $\eta = 0.1$.

Table 2. Estimation results for Model 7. Proportion of correct estimated number of breakpoints (Corr. #), the averages (ave) and the empirical standard deviations (e.s.d.) of the change-point estimates. True change-point is at $t = [\eta n]$

IC	n	ϖ=0.05 Corr. #	ϖ=0.05 ave	ϖ=0.05 e.s.d.	ϖ=0.1 Corr. #	ϖ=0.1 ave	ϖ=0.1 e.s.d.	ϖ=0.2 Corr. #	ϖ=0.2 ave	ϖ=0.2 e.s.d.
					η=0.1					
	200	0.655	24.85	15.44	0.800	25.83	15.95	0.865	25.34	12.58
AIC	400	0.680	42.35	6.45	0.900	43.41	8.26	1.00	46.88	15.15
	800	0.755	81.89	4.61	0.930	83.22	6.42	1.00	84.57	10.50
	200	0.715	26.45	18.34	0.845	25.77	15.59	0.865	25.57	12.60
AICc	400	0.735	42.27	6.47	0.905	43.60	8.47	1.000	46.97	15.15
	800	0.775	82.03	4.69	0.935	82.20	6.40	1.000	84.57	10.50
	200	0.835	24.98	10.08	0.795	26.48	13.69	0.620	23.85	8.11
BIC	400	1.000	47.09	15.19	0.965	47.35	15.84	0.815	46.26	12.19
	800	1.000	84.84	11.05	0.995	85.77	12.15	0.970	86.82	13.95
	200	0.795	26.48	13.69	0.710	26.18	13.65	0.505	23.15	8.30
MDL	400	0.965	47.36	15.84	0.935	47.54	16.16	0.690	45.52	12.13
	800	0.995	85.67	12.13	0.990	86.38	13.10	0.910	86.42	12.81
					η=0.9					
	200	0.370	172.66	16.86	0.680	168.36	24.60	0.815	166.06	22.70
AIC	400	0.460	354.40	8.50	0.795	352.81	22.53	0.960	350.22	20.42
	800	0.470	715.69	6.02	0.800	715.14	7.59	1.000	712.255	11.18
	200	0.485	171.57	20.44	0.745	168.36	23.79	0.810	165.31	23.10
AICc	400	0.560	351.88	26.33	0.810	352.48	22.83	0.970	349.99	20.43
	800	0.515	715.86	5.85	0.820	715.06	7.42	1.000	711.13	15.25
	200	0.780	165.46	22.65	0.645	167.24	17.41	0.335	169.36	14.25
BIC	400	0.945	345.33	29.85	0.885	344.86	25.89	0.590	346.25	19.48
	800	0.990	709.55	16.54	0.980	707.10	18.61	0.880	702.63	22.94
	200	0.640	167.06	17.41	0.500	168.77	14.88	0.195	171.80	9.99
MDL	400	0.895	344.75	25.79	0.765	344.99	22.57	0.440	348.42	14.78
	800	0.985	707.49	18.38	0.965	704.65	21.00	0.740	701.68	24.12

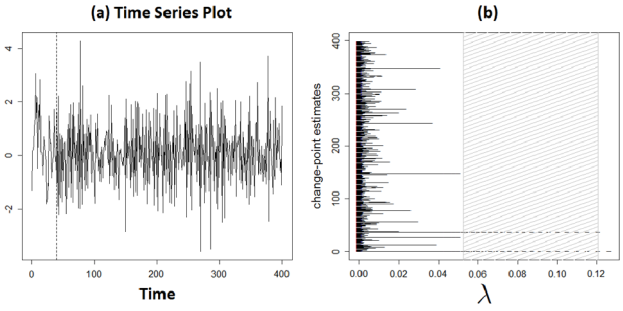

Figure 3. Time series plot, estimated change-points, and the number of estimated change-points along the solution path for a realization of (7) with $\eta = 0.1$. The estimated change-point is in vertical line. The shaded area is the range of λs that achieves the minimum BIC.

Table 2 reports the change-point estimation results for $D_n = n^{\varpi}$, $\varpi \in \{0.05, 0.1, 0.2\}$, $n \in \{200, 400, 800\}$ with 200 replications. Given that $\eta = 0.1$ or 0.9, the shorter segment length is only 20 when $n = 200$. Surprisingly, the performance of the change-point estimation is in general satisfactory even in this challenging set up. Similar to the results in the previous subsection, the performance of BIC using the penalty $\varpi = 0.05$ is the best in terms of detecting the correct number of change-points. Also, the performance of change-point estimation improves and becomes less sensitive to the choice of ϖ as sample size increases. On the other hand, AIC and AICc perform poorly for $\varpi = 0.05$ but perform quite well for $\varpi = 0.2$. Nevertheless, BIC and MDL tend to give a more stable results across different values of ϖ when sample size is large.

5.3 Multiple change-point model

In this subsection we consider the multiple change-point model

(8)
$$X_t = \begin{cases} 1 + 0.4X_{t-1} + 0.1X_{t-2} + \epsilon_t\,, & \text{if } 1 \le t \le [0.2n], \\ 1 - 0.5X_{t-1} + 0.2X_{t-2}\epsilon_t\,, & \text{if } [0.2n] < t \le [0.5n], \\ 2 + 0.4X_{t-1} - 0.1X_{t-2} + \epsilon_t\,, & \text{if } [0.5n] < t \le [0.8n], \\ 0.5 - 0.4X_{t-1} + 0.2X_{t-2} + \epsilon_t\,, & \text{if } [0.8n] < t \le n, \end{cases}$$

where $\epsilon_t \sim N(0,1)$. The specification of k, q, M_t^j, θ_0^j, f_t^j for $j = 1, 2, 3, 4$, $\tilde{H}_t(\theta_{M,t}) = 1$, and $S_t(\theta_{f,t})$ are similar to that in Section 5.1. A realization of (8) is given in Figure 4. We consider the performance of change-point estimation using the four criteria along the group LASSO solution path. See Figure 4 for a typical solution path of a realization of (8) with $n = 400$.

Table 3. Estimation results for Model 8. Proportion of correct estimated number of breakpoints (Corr. #), the averages (ave) and the empirical standard deviations (e.s.d.) of the change-point estimates. True change-point is at $t = [\eta n]$

IC	n	ϖ=0.05 Corr. #	ϖ=0.05 ave	ϖ=0.05 e.s.d.	ϖ=0.1 Corr. #	ϖ=0.1 ave	ϖ=0.1 e.s.d.	ϖ=0.2 Corr. #	ϖ=0.2 ave	ϖ=0.2 e.s.d.
First change-point $t^{(1)} = [0.2n]$										
	400	0.795	75.58	5.79	0.905	74.82	6.10	0.93	73.49	8.57
AIC	800	0.820	155.51	4.80	0.950	155.38	4.81	1.00	154.55	5.96
	1200	0.800	235.63	4.61	0.965	235.13	5.27	1.00	233.71	6.55
	400	0.820	73.45	8.72	0.615	73.91	8.78	0.16	75.59	4.81
BIC	800	0.990	153.97	8.07	0.915	153.25	8.81	0.395	398.41	5.92
	1200	1.000	233.28	7.31	0.980	232.28	8.60	0.625	232.84	7.56
Second change-point $t^{(2)} = [0.5n]$										
	400		198.82	1.70		198.81	1.64		198.70	1.80
AIC	800		398.99	1.43		398.98	1.36		398.86	1.85
	1200		598.99	1.71		598.99	1.63		598.97	1.63
	400		198.72	1.85		198.64	1.90		198.53	1.76
BIC	800		398.86	1.86		398.76	1.86		398.41	2.33
	1200		598.97	1.63		598.98	1.65		598.92	1.70
Third change-point $t^{(3)} = [0.8n]$										
	400		318.44	1.26		318.26	1.42		318.11	1.94
AIC	800		638.39	1.34		638.32	1.43		638.18	1.66
	1200		958.65	0.86		958.60	0.94		958.41	1.27
	400		317.98	2.15		317.73	2.86		317.03	3.58
BIC	800		638.17	1.67		638.14	1.69		637.75	2.10
	1200		958.41	1.27		958.37	1.30		958.22	1.48

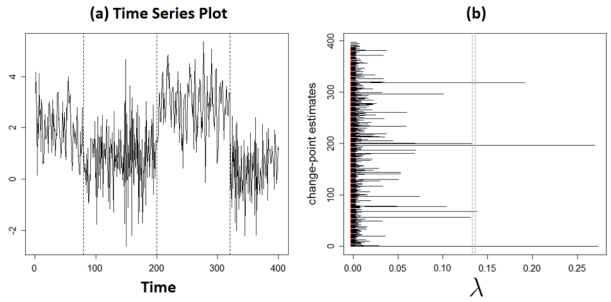

Figure 4. Time series plot, estimated change-points, and the number of estimated change-points along the solution path for a realization of (8). The estimated change-points are in vertical lines. The shaded area is the range of λs that achieves the minimum BIC.

When LASSO type methods are used for multiple change-point models, the best model obtained from the solution path contains all the true change-points in a small neighborhood, but over-estimation may occur. On the other hand, a backward elimination algorithm based on the same information criterion can be employed to extract an improved set of change-point estimators from the over-estimated set; see [8]. Therefore, we apply the backward elimination algorithm on the model selected by the group LASSO solution path.

Table 3 reports the change-point estimation results for $D_n = n^{\varpi}$, $\varpi \in \{0.05, 0.1, 0.2\}$, $n \in \{400, 800, 1200\}$ with 200 replications. The percentage of correctly identifying three change-points are over 90% for most cases except AIC with $\varpi = 0.05$ and BIC with $\varpi = 0.2$. Similar to the results in the previous subsection, the performance of BIC using the penalty $\varpi = 0.05$ and AIC using the penalty $\varpi = 0.2$ are the best in terms of detecting the correct number of change-points. Again, the performance of change-point estimation improves and becomes less sensitive to the choice of ϖ as sample size increases.

6. DATA ANALYSIS

We applied the proposed procedure in Section 4 on the US monthly purchasing managers index (PMI) data between January 1948 and July 2015. Differencing the log of the PMI series results in a seemingly piecewise stationary time series $y = \{y_1, \ldots, y_{810}\}$; see Figure 5. [11] employ autoregressive models for similar datasets. Following their approach, we applied the proposed procedure using autoregressive (AR) model with order 2. Based on the results in Section 5, AIC, AICc with $\varpi = 0.2$ and BIC, MDL with $\varpi = 0.05$, which give the best performance, are used in the analysis. Figure 5 depicts a typical solution path of the estimation procedure using BIC with $\varpi = 0.05$. Interestingly, all of the four criteria give the same change-point at $t = 401$, which corresponds to June 1981. This change-point could be associated with the early 1980s recession in the US.

To investigate the goodness of fit of each segment, Table 4 reports the AICs for various AR models for the estimated

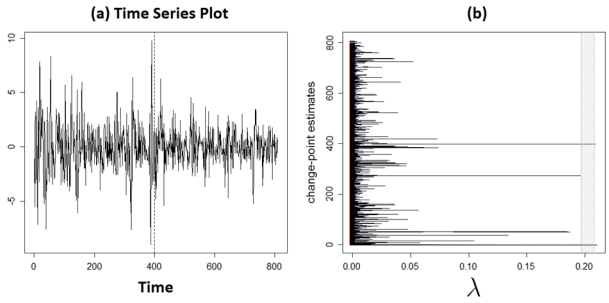

Figure 5. Left: Differenced log purchasing managers index (PMI) series. Right: Solution path from the group LASSO procedure. The time axis are the index 1 to 810 of the difference series $\{y_1, \ldots, y_{810}\}$.

Table 4. AICs for various AR models for the estimated segments in the differenced log-PMI series. Segment 1: $\{y_1, \ldots, y_{400}\}$*; Segment 2:* $\{y_{401}, \ldots, y_{810}\}$

	Segment 1	Segment 2
AR(0)	1846.4	1496.2
AR(1)	1811.9	1486.0
AR(2)	1813.9	1474.8
AR(3)	1813.1	1476.3
AR(4)	1813.0	1478.0
AR(5)	1814.5	1479.5
AR(6)	1815.3	1481.5

segments $y_1, \ldots, y_{400}$ and $y_{401}, \ldots, y_{810}$. It can be seen that AR(1) and AR(2) achieve the lowest AIC values for the first and second estimated segment, respectively. Therefore, the use of AR(2) model in the LASSO method is appropriate. Moreover, when AR(2) models are used in both segments, the p-values of Ljung-Box tests for the residuals of the two segments are 0.996 and 0.889, respectively, indicating the fitted models are adequate.

In practice, change-point analysis in the off-line setting is helpful for prediction. For example, if no change-point is detected in the data, then predictions will be based on the model fitted to the whole series $\{y_1, \ldots, y_n\}$. On the other hand, if change-points at $\hat{t}_1 < \hat{t}_2 < \cdots < \hat{t}_k$ are detected, then predictions based on the model fitted to $\{y_{\hat{t}_k}, \ldots, y_n\}$ is expected to perform better. To illustrate this point, we compare the out-of-sample prediction errors $E_0 = \sum_{i=1}^{8}(y_{810+i} - \hat{y}_{810+i})^2$ and $E_1 = \sum_{i=1}^{8}(y_{810+i} - \tilde{y}_{810+i})^2$, where $y_{811}, \ldots, y_{818}$ corresponds to the difference log-PMI data from August 2015 to March 2016, $\hat{y}_j$s and $\tilde{y}_j$s are predictions using the whole series and the second segment, respectively. In computing E_0, we compared the AICs of AR(0) to AR(6) fitted to the series $\{y_1, \ldots, y_{810}\}$ and the AR(4) model was selected. On the other hand, from Table 4, an AR(2) model was used in computing E_1. It is found that $E_0 = 1.89$ and $E_1 = 1.67$, which confirms that the prediction error using the post-change segment is lower than that using the whole time series.

7. CONCLUSIONS

To summarize, it is shown in Theorem 3.1 that with appropriately chosen penalty terms, information criteria can be used to select a consistent change-point model. Unlike the existing results in [4] that require a upper bound K_{max} on the number of change-points, Theorem 3.1 is able to rule out all seriously over-fitting change-point models with number of change-points exceeding K_{max}.

In practice, obtaining a change-point model with optimal information criterion value can be challenging. Though methods based on dynamic programming are available in the literature, see e.g., [18], [13], and [25], most of such methods are designed for sequences of independent observations. To guarantee the maximum $O(n^2)$ computational complexity, it is necessary that the objective function (evaluated at the parameter estimates) can be updated in $O(1)$ steps upon the arrival of a new observation. This may not be true in general for dependent non-Gaussian data. It is an interesting future research direction to study the numerical methods of finding optimal change-point model of [5].

Recently, LASSO based method is proposed in [8] and [9]. This method takes a tuning parameter λ as input and generates a piecewise constant sequence $\theta_t(\lambda)$ as output. As explained in Section 4, Theorem 3.1 guarantees that optimal change-point model can be chosen if the optimal change-point model belongs to $\{\theta_t(\lambda) : \lambda > 0\}$. However, optimality within $\{\theta_t(\lambda) : \lambda > 0\}$ may not imply global optimality. It is also interesting to establish results of the probability that the optimal change-point model belongs to $\{\theta_t(\lambda) : \lambda > 0\}$.

ACKNOWLEDGMENTS

We would like to thank the associate editor and the anonymous referee for their helpful suggestions and comments. Chi Tim Ng's work is supported by the 2013 Chonnam National University Research Program grant (No. 2013-2299) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2011-0030810). Chun Yip Yau's research has been supported in part by HKSAR-RGC-ECS 405012 and HKSAR-RGC-GRF 405113, 14601015.

Received 24 June 2015

REFERENCES

[1] Andreou, E. and Ghysels, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics* **17**, 579–600.

[2] Bai, J., Lumsdaine, R., and Stock, J. H. (1998). Testing for and dating common breaks in multivariate time series. *The Review of Economic Studies* **65**, 395–432. MR1635358

[3] Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66**, 47–78. MR1616121

[4] Bardet, J., Kengne, W., and Wintenberger, O. (2012). Multiple breaks detection in general causal time series using penalized quasi-likelihood. *Electronic Journal of Statistics* **6**, 435–477. MR2988415

[5] Bardet, J. and Wintenberger, O. (2009). Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes. *Annals of Statistics* **37**, 2730–2759. MR2541445
[6] Beaulieu, C., Chen, J., and Sarmiento, L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transaction of Royal Society A* **370**, 1228–1249.
[7] Berkes, I., Gombay, E., Horváth, L., and Kokoszka, P. (2004). Sequential change-point detection in GARCH(p,q) models. *Econometric Theory* **20**, 1140–1167. MR2101953
[8] Chan, N. H., Yau, C. Y., and Zhang, R. M. (2014). Group LASSO for structural break time series. *Journal of the American Statistical Association* **109**, 590–599. MR3223735
[9] Chan, N. H., Yau, C. Y., and Zhang, R. M. (2014). LASSO estimation of threshold autoregressive models. (To appear in Journal of Econometrics). MR3414900
[10] Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association* **101**, 223–239. MR2268041
[11] De Bondt, G. J., Dieden, H. C., Muzikarova, S., and Vincze, I. (2014). Modelling industrial new orders for the euro area. *European Central Bank, Statistics paper series* **6**.
[12] Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *The Annals of Statistics* **38**, 3567–3604. MR3065478
[13] Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society, Series B (with discussion)* **76**, 495–580. MR3210728
[14] Gillet, O., Essid, S., and Richard, G. (2007). On the correlation of automatic audio and visual segmentation of music video. *IEEE Transactions on Circuits and Systems for Video Technology*. 2007.
[15] Harchaoui, Z. and Levy-leduc, C. (2008). Catching change-points with lasso. *Advances in Neural Information Processing Systems (NIPS)*. 2008.
[16] Huang, T., Wu, B., Lizardi, P., and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics* **21**, 3811–3817.
[17] Inclan, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of American Statistical Association* **89**, 913–923. MR1294735
[18] Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changespoints with a linear coputational cost. *Journal of American Statistical Association* **107**, 1590–1598. MR3036418
[19] Kim, S., Cho, S., and Lee, S. (2000). On the cusum test for parameter changes in GARCH(1,1) models. *Communications in Statistics - Theory and Methods* **29**, 445–462. MR1749743
[20] Kokoszka, P. and Leipus, R. (2000). Change-point estimation in ARCH models. *Bernoulli* **6**, 513–539. MR1762558
[21] Lavielle, M. and Moulines, E. (2000). Least-square estimation of an unknown number of shrifts in a time series. *Journal of Time Series Analysis* **21**, 33–59. MR1766173
[22] Lee, S., Ha, J., Na, O., and Na, S. (2003). The cusum test for parameter change in time series model. *Scandinavian Journal of Statistics* **30**, 781–796. MR2155483
[23] Ng, C. T., Lee, W., and Lee, Y. (2015). Simultaneous estimation of the change points and the time-varying coefficients for time series models. *Preprint.*
[24] Perron, P. (2005). Dealing with structural breaks. *Palgrave Handbook of Econometrics, Vol. 1: Econometric Theory (Patterson, K. and Mills, T. C. eds.)*. Basingstoke, U.K.: Palgrave Macmillan.
[25] Rigaill, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. *Arxiv.org:1004.0887v1 [stat.CO]*.
[26] Rinaldo, A. (2009). Properties and refinements of the fused LASSO. *The Annals of Statistics* **37**, 2922–2952. MR2541451
[27] Tibshirani, R. (1996). Regression shrinkage and delection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288. MR1379242
[28] Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc. B* **71**, 671–683. MR2749913

Chi Tim Ng
Department of Statistics
Chonnam National University
Gwangju
ROK
E-mail address: easterlyng@gmail.com

Chun Yip Yau
Department of Statistics
Chinese University of Hong Kong
Shatin
Hong Kong
E-mail address: cyyau@sta.cuhk.edu.hk