Electronic Journal of Statistics
Vol. 7 (2013) 381–411
ISSN: 1935-7524
DOI: 10.1214/13-EJS769

Consistency of minimum description length model selection for piecewise stationary time series models

Richard A. Davis[*]

Department of Statistics
Columbia University
New York, NY 10027
USA
e-mail: rdavis@stat.columbia.edu

and

Chun Yip Yau[†]

Department of Statistics
Chinese University of Hong Kong
Shatin, N.T.
Hong Kong
e-mail: cyyau@sta.cuhk.edu.hk

Abstract: This paper establishes the consistency of the minimum description length (MDL) model selection procedure by [10, 11] for a class of non-stationary time series models. We consider a time series model in which the observations are viewed as coming from stationary segments. In other words, the data are assumed to come from a general time series model in which the parameters change at break-points. Each of these segments is modeled by a pre-specified family of parametric stationary time series models. [10, 11] formulated the above problem and used the minimum description length (MDL) principle to estimate the number of break-points, the location of the break-points, the order of the parametric model and the parameter values in each of the segments. The procedure performed well on a variety of examples. In this paper we show consistency of their minimal MDL model selection procedure under general regularity conditions on the likelihood function. Results about the rate of convergence of the break-point-location estimator are also given. Applications are considered for detecting changes in independent random variables, and in ARMA and GARCH processes.

AMS 2000 subject classifications: Primary 62F10, 62M10.
Keywords and phrases: GARCH, minimum description length (MDL) principle, multiple change points, non-stationary time series, piecewise-stationary processes.

Received August 2012.

[*] This research is supported in part by NSF Grant DMS-1107031

[†] Supported in part by HKSAR-RGC Grants CUHK405012 and Direct Grant CUHK2060445

1. Introduction

There has been considerable development in non-linear time series modeling during the past 20 years. One prominent subject in non-linear time series modeling is the "change-point" or the "structural breaks" model. In this model, a non-stationary time series can be partitioned into a number of segments of different stationary processes. At each break-point, the stationary process experiences either a change in the mean, variance, correlation structure or other dependence features. Davis, Lee and Rodriguez-Yam [10, 11] proposed the Automatic Segmentation (Auto-Seg) procedure for modeling such kind of piecewise-stationary time series. This procedure simultaneously estimates the number of break points, the location of break points, and the parametric model in each segment.

The main idea of Auto-Seg procedure is to model non-stationary time series by segmenting the time series into blocks of different stationary time series. Here, the model for a non-stationary time series is described by the locations of the change-points and the parametric model in each of the segments. To select a model, the minimum description length (MDL) model selection criterion is employed to estimate simultaneously the number of change-points, the locations of the change-points, the model for each segment and its parameters. The procedure works as follows. Given the locations of the change-points and the parametric models in each segments, a MDL can be evaluated. The MDL can be regarded as the negative of the sum of the log-likelihood for each of the segments plus a penalty term which penalizes the *size* of the model. Then the best model is selected by minimizing the MDL over the change-point locations and the parametric models in each segments. While this minimization problem is difficult, the genetic algorithm can be employed to produce near optimal solutions. For details about the MDL and the genetic algorithm in this setting, see [10, 11]. Simulation studies gave promising results for the estimation of the number of break-points and their locations for various families of the time series models.

Theoretical results are available for a special case of the Auto-Seg procedure, Auto-PARM (Automatic Piecewise AutoRegressive Modeling), proposed by [10], where the parametric family used for modeling the stationary processes is restricted to pure AR models. When the number of change-points is known, [10] showed that the estimated change-point locations are strongly consistent. Davis, Hancock and Yao [9] proved that AutoPARM's estimate of the number of change-points and the change-point locations are weakly consistent under conditional maximum likelihood estimation. The issue of consistency for the more general Auto-Seg procedure, which applies to any stationary time series with likelihood function being well defined, remains open.

In this paper we consider the strong consistency properties of the Auto-Seg procedure. We show that, under some regularity conditions on the likelihood function, the number of change-points, the locations of the change-points, the parametric models and the parameters in each of the segments can be consistently estimated.

Section 2 begins by reviewing the piecewise stationary models and the Auto-Seg procedure. Section 3 contains the main results about the strong consistency of the Auto-Seg procedure. In Section 4, we consider the application of the main results to independent data, and in ARMA and GARCH processes.

2. Setting and assumptions

Before embarking on our consistency results, we first review the Automatic Segmentation (Auto-Seg) modeling procedure developed by [11].

The Auto-Seg procedure applies to the class of piecewise stationary time series models, which is the class of time series $(Y_1, Y_2, \ldots, Y_n)$ that can be partitioned into stationary segments by m unknown distinct break-points $\tau_1, \tau_2, \ldots, \tau_m$. Set $\tau_0 = 0$, $\tau_{m+1} = n$, and let $\lambda_j = \tau_j/n, j = 0, \ldots m+1$ be the normalized break-points. Note that $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_m < \lambda_{m+1} = 1$. The asymptotic result is based on increasing n with the λ_j's being fixed.

Given the break-points $\tau_1, \ldots, \tau_m$, the observed time series can be segmented into $m+1$ pieces of stationary time series. The j-th piece of $\{Y_t\}$ is modeled by a stationary time series $\mathbf{x_j} = \{X_{t,j}\}_{t\in\mathbb{Z}}$ such that

$$Y_t = X_{t-\tau_{j-1},j} \qquad \text{for} \quad \tau_{j-1} + 1 \le t \le \tau_j \, . \tag{2.1}$$

Intuitively, the observation first starts as the stationary process $\{X_{t,1}\}$. After the structural break at τ_1, a new stationary time series $\{X_{t,2}\}$ is observed, and so on. In other words, the observed time series $(Y_1, Y_2, \ldots, Y_n)$ can be written as $(X_{1,1}, \ldots, X_{n_1,1}, X_{1,2}, \ldots, X_{n_2,2}, \ldots, X_{n_{m+1},m+1})$, where $n_j = \tau_j - \tau_{j-1}$ for $j = 1, \ldots, m+1$ and $n = n_1 + n_2 + \ldots + n_{m+1}$. Assume also that each of the $m+1$ segments is associated with a sequence of unobserved "past observations" $\{X_{t,j}, t \le 0\}$ such that $\{X_{t,j}, -\infty < t < \infty\}$ is a stationary process for $j = 1, \ldots, m+1$. Similar to [1] and [3], we assume that the series $\{X_{t,j}\}$, $j = 1, \ldots, m+1$, are independent, although this is not essential (see [11]).

Given the location of the break-points, each segment of stationary time series is modeled by an element in a pre-specified finite class of models $\mathcal{M}$. Each element in $\mathcal{M}$ is a model associated with an integer-valued vector parameter ξ of dimension c, which represents the order of the model. Let ξ_j be the c_j-dimensional vector specifying a model for the j-th piece. Given ξ_j, the model depends on a real-valued parameter $\theta_j = \theta(\xi_j)$ of dimension $d_j = d_j(\xi_j)$. The joint probability distribution for the j-th segment, $\mathbf{x_j}$, is completely determined by $\theta(\xi_j)$. In other words, we regard ξ_j as a parameter specifying the order of the model and θ_j as the parameters of the specified model. Assume that $\theta_j \in \Theta_j \equiv \Theta_j(\xi_j)$, where $\Theta_j \subset \mathbb{R}^{d_j}$ is a compact parameter space. Define $\psi_j = (\xi_j, \theta_j)$ to be the parameter set of the j-th piece.

Example 1. If the parametric model for the stationary segments is chosen from the class of Gaussian ARMA models ($\mathcal{M}$), then the ARMA(2,1) model

$$(1 - \phi_1 B - \phi_2 B^2)(X_{t,j} - \mu) = (1 - \vartheta B) Z_{t,j}, \quad Z_{t,j} \sim N(0, \sigma^2) \, ,$$

may be specified by $\xi_j = (2, 1)$, $c_j = 2$, $\theta_j = (\mu, \sigma^2, \phi_1, \phi_2, \vartheta)$ and $d_j = 5$. □

Let $f_{\xi_j}(x_{i,j}|x_{s,j}, s < i; \theta_j)$ be the conditional density function for the i-th observation of the j-th piece, given all the past observations. Note that the functional form of $f_{\xi_j}(x_{i,j}|x_{s,j}, s < i; \theta_j)$ is specified by ξ_j and the parameter values of $f_{\xi_j}(x_{i,j}|x_{s,j}, s < i; \theta_j)$ is given by θ_j. If all the past observations $\{x_{t,j}, t < 0\}$ are known, then the conditional log-likelihood function for an observation $x_{i.j}$ given the past is defined by $l_j(\psi_j; x_{i,j}|x_{s,j}, s < i) \equiv \log f_{\xi_j}(x_{i,j}|x_{s,j}, s < i; \theta_j)$. The conditional log-likelihood of the j-th piece, $\mathbf{x_j} \equiv \{X_{t,j}, t = 1, 2, \ldots, n_j\}$, given all the past observations, is given by

$$L_n^{(j)}(\psi_j; \mathbf{x_j}) = \sum_{i=1}^{n_j} l_j((\xi_j, \theta_j); x_{i,j}|x_{s,j}, s < i)\,.$$

Of course, the past observations $\{x_{t,j}\}_{t<0}$ is unknown in practice. Thus, for any observation $x_{i,j}$, its "observed past", is in fact $(\ldots, 0, 0, x_{1,1}, x_{2,1}, \ldots, x_{n_1,1}, x_{1,2}, \ldots, x_{n_{j-1},j-1}, x_{1,j}, \ldots, x_{i-1,j})$, or equivalently, $\boldsymbol{y_{i,j}} \equiv (\ldots, 0, 0, y_1, y_2, y_3 \ldots, y_{\tau_{j-1}+i-1})$. The observed likelihood for the j-th piece is then given by

$$\tilde{L}_n^{(j)}(\psi_j; \mathbf{x_j}) = \sum_{i=1}^{n_j} l_j((\xi_j, \theta_j); x_{i,j}|\boldsymbol{y_{i,j}})\,.$$

Note that $l_j((\xi_j, \theta_j); x_{i,j}|\boldsymbol{y_{i,j}})$ is obtained by replacing the true past observations $\{x_{i,j}\}$ by $\boldsymbol{y_{i,j}}$ in $l_j((\xi_j, \theta_j); x_{i,j}|x_{s,j}, s < i)$, which is not same as the true conditional distribution given the entire past history of the time series.

Denote the location vector and the parameter vector by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{m+1})$, respectively. The vector $(m, \boldsymbol{\lambda}, \boldsymbol{\psi})$ specifies completely a model of a non-stationary time series $\{Y_t\}_{t=1,\ldots,n}$ defined in (2.1). According to [11], the MDL for this model is given by

$$\begin{aligned}\mathrm{MDL}(m, \boldsymbol{\lambda}, \boldsymbol{\psi}) \quad &= \quad \log m + (m+1)\log n + \sum_{j=1}^{m+1}\sum_{k=1}^{c_j} \log \xi_{k,j} \\ &\qquad + \sum_{j=1}^{m+1} \frac{d_j}{2}\log n_j - \sum_{j=1}^{m+1} \tilde{L}_n^{(j)}(\psi_j; \mathbf{x_j})\,,\end{aligned}$$

where $\xi_j = (\xi_{1,j}, \ldots, \xi_{c_j,j})$. Note that the MDL model selection procedure is closely connected to penalized maximum likelihood estimation since the MDL can be regarded as the minus log-likelihood plus a penalty term of order $\log n$. The best model is selected by minimizing the MDL with respective to $(m, \boldsymbol{\lambda}, \boldsymbol{\psi})$. To ensure identifiability of the change-points, when we search for the change-points, we assume that there exists a $\epsilon_\lambda > 0$ such that $\min_{1\le j\le m+1}(|\lambda_j - \lambda_{j-1}|) > \epsilon_\lambda$. That is, we propose a constraint $\boldsymbol{\lambda} \in A_{\epsilon_\lambda}$, where

$$A_{\epsilon_\lambda}^m = \{\boldsymbol{\lambda} \in (0,1)^m, 0 < \lambda_1 < \ldots < \lambda_m < 1, \lambda_i - \lambda_{i-1} \ge \epsilon_\lambda, i = 1, \ldots, m\}\,. \tag{2.2}$$

Under this restriction the number of change-points is bounded by $M = [1/\epsilon_\lambda] + 1$.

The estimates of the number of change-points, the locations of the change-points and the parameters in each of the segments are given by the vector $(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}})$, where

$$(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) = \arg \min_{\substack{m \le M, \\ \boldsymbol{\psi} \in \mathcal{M}, \\ \boldsymbol{\lambda} \in A^m_{\epsilon_\lambda}}} \mathrm{MDL}(m, \boldsymbol{\lambda}, \boldsymbol{\psi}) \,, \tag{2.3}$$

$\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} = (\hat{\lambda}_1, \dots, \hat{\lambda}_{\hat{m}})$ and $\hat{\boldsymbol{\psi}}_{\boldsymbol{n}} = (\hat{\psi}_1, \dots, \hat{\psi}_{\hat{m}+1})$. Note that $\hat{\psi}_j = (\hat{\xi}_j, \hat{\theta}^{(j)}_n)$, where

$$\hat{\theta}^{(j)}_n = \arg \max_{\theta_j \in \Theta_j(\hat{\xi}_j)} \tilde{L}^{(j)}_n((\hat{\xi}_j, \theta_j); \hat{\mathbf{x}}_{\mathbf{j}}) \,,$$

with $\hat{\mathbf{x}}_{\mathbf{j}} = \{y_t; [n\hat{\lambda}_{j-1}] < t \le [n\hat{\lambda}_j]\}$ denotes the estimated j-th segment of the time series.

We first consider the situation where a portion of the data within the j-th stationary segment is chosen for parameter estimation. Let λ_u and λ_d be in $[0, 1]$ with $\lambda_d < \lambda_u$ and $\lambda_u - \lambda_d > \epsilon_\lambda$. To simplify notation, denote

$$\sup_{\lambda_d, \lambda_u} = \sup_{\substack{\lambda_d \in [0,1], \lambda_u \in [0,1] \\ \lambda_u - \lambda_d > \epsilon_\lambda}} \,. \tag{2.4}$$

Define, for $j = 1, \dots, m+1$, the true and the observed likelihood formed by a portion of the j-th segment respectively by

$$L^{(j)}_n(\psi_j, \lambda_d, \lambda_u; \mathbf{x}_j) = \sum_{i=[n_j\lambda_d]+1}^{[n_j\lambda_u]} l_j(\psi_j; x_{i,j} | x_{l,j}, l < i) \,, \tag{2.5}$$

$$\tilde{L}^{(j)}_n(\psi_j, \lambda_d, \lambda_u; \mathbf{x}_j) = \sum_{i=[n_j\lambda_d]+1}^{[n_j\lambda_u]} l_j(\psi_j; x_{i,j} | \mathbf{y}_{\mathbf{i},\mathbf{j}}) \,. \tag{2.6}$$

In practice, the observed likelihood $\tilde{L}^{(j)}_n(\psi_j, \lambda_d, \lambda_u; \mathbf{x}_{\mathbf{j}})$ is used to approximate the true likelihood $L^{(j)}_n(\psi_j, \lambda_d, \lambda_u; \mathbf{x}_{\mathbf{j}})$. The following assumption is used to control the quality of this approximation.

Assumption 1 (k). For any $j = 1, 2, \dots, m+1$ and fixed ξ_j, the function l_j is two-time continuously differentiable with respective to θ_j and the first and second derivatives $L^{'(j)}_n$, $\tilde{L}^{'(j)}_n$ and $L^{''(j)}_n$, $\tilde{L}^{''(j)}_n$, respectively, of the function defined in (2.5) and (2.6), satisfy

$$\sup_{\lambda_d, \lambda_u} \sup_{\theta_j \in \Theta_j(\xi_j)} \left| \frac{1}{n} L^{(j)}_n((\xi_j, \theta_j), \lambda_d, \lambda_u, \mathbf{x}_{\mathbf{j}}) - \frac{1}{n} \tilde{L}^{(j)}_n((\xi_j, \theta_j), \lambda_d, \lambda_u, \mathbf{x}_{\mathbf{j}}) \right| = o\left(n^{\frac{1}{k}-1}\right) \,,$$

$$\sup_{\lambda_d, \lambda_u} \sup_{\theta_j \in \Theta_j(\xi_j)} \left| \frac{1}{n} L^{'(j)}_n((\xi_j, \theta_j), \lambda_d, \lambda_u, \mathbf{x}_{\mathbf{j}}) - \frac{1}{n} \tilde{L}^{'(j)}_n((\xi_j, \theta_j), \lambda_d, \lambda_u, \mathbf{x}_{\mathbf{j}}) \right| = o\left(n^{\frac{1}{k}-1}\right) \,,$$

$$\sup_{\lambda_d, \lambda_u} \sup_{\theta_j \in \Theta_j(\xi_j)} \left| \frac{1}{n} L^{''(j)}_n((\xi_j, \theta_j), \lambda_d, \lambda_u, \mathbf{x}_{\mathbf{j}}) - \frac{1}{n} \tilde{L}^{''(j)}_n((\xi_j, \theta_j), \lambda_d, \lambda_u, \mathbf{x}_{\mathbf{j}}) \right| = o(1) \,,$$

almost surely.

For example, by Assumption 1(2) we mean the above assumption is satisfied with $k = 2$. Next, some regularity conditions for the conditional log-likelihood function is needed for the standard properties of maximum likelihood estimation.

Assumption 2 (k). For $j = 1, \ldots, m+1$, and any fixed ξ_j, there exists an $\epsilon > 0$ such that

$$\begin{aligned} \sup_{\theta_j \in \Theta_j(\xi_j)} E|l_j((\xi_j, \theta_j); x_{1,j}|x_{l,j}, l < 1)|^{k+\epsilon} &< \infty\,, \\ \sup_{\theta_j \in \Theta_j(\xi_j)} E|l'_j((\xi_j, \theta_j); x_{1,j}|x_{l,j}, l < 1)|^{k+\epsilon} &< \infty\,, \\ \sup_{\theta_j \in \Theta_j(\xi_j)} E|l''_j((\xi_j, \theta_j); x_{1,j}|x_{l,j}, l < 1)| &< \infty\,. \end{aligned}$$

Assumption 3. For each $j = 1, \ldots, m+1$ and any fixed ξ_j,

$$\begin{aligned} \sup_{\theta_j \in \Theta_j(\xi_j)} \left| \frac{1}{n} L_n^{(j)}((\xi_j, \theta_j); \mathbf{x}_j) - L_j((\xi_j, \theta_j)) \right| &\xrightarrow{a.s.} 0\,, \\ \sup_{\theta_j \in \Theta_j(\xi_j)} \left| \frac{1}{n} L_n'^{(j)}((\xi_j, \theta_j); \mathbf{x}_j) - L'_j((\xi_j, \theta_j)) \right| &\xrightarrow{a.s.} 0\,, \\ \sup_{\theta_j \in \Theta_j(\xi_j)} \left| \frac{1}{n} L_n''^{(j)}((\xi_j, \theta_j); \mathbf{x}_j) - L''_j((\xi_j, \theta_j)) \right| &\xrightarrow{a.s.} 0\,, \end{aligned}$$

where

$$\begin{aligned} L_j((\xi_j, \theta_j)) &:= E(l_j((\xi_j, \theta_j); x_{1,j}|x_{l,j}, l < 1))\,, \\ L'_j((\xi_j, \theta_j)) &:= E(l'_j((\xi_j, \theta_j); x_{1,j}|x_{l,j}, l < 1))\,, \\ L''_j((\xi_j, \theta_j)) &:= E(l''_j((\xi_j, \theta_j); x_{1,j}|x_{l,j}, l < 1))\,. \end{aligned}$$

In practice, the likelihood has to be defined in terms of the estimated location of the break-points. Even if the estimated location of the change-points is very close to the true ones, the two ends of the j-th estimated segment may contain observations from the $(j-1)$-th or the $(j+1)$-th piece of the true piecewise stationary process. To establish the rate of convergence of the location estimators, one extra assumption is needed to control the effect at the two ends of the fitted segments.

Assumption 4 (w). For $j = 1, \ldots, m+1$, any fixed ψ and any sequence $\{g(n)\}_{n \geq 1}$ of integers that satisfies $g(n) > cn^w$ for some $c > 0$ when n is sufficiently large, then

$$\frac{1}{g(n)} \sum_{i=n-g(n)+1}^{n} l_j(\psi; x_{i,j}|x_{l,j}, l < i) \overset{a.s.}{\to} E(l_j(\psi; x_{1,j}|x_{l,j}, l < 1))\,, \quad (2.7)$$

$$\frac{1}{g(n)} \sum_{i=n-g(n)+1}^{n} l'_j(\psi; x_{i,j}|x_{l,j}, l < i) \overset{a.s.}{\to} E(l'_j(\psi; x_{1,j}|x_{l,j}, l < 1))\,. \quad (2.8)$$

It will be shown below in Lemma 1 that (2.7) and (2.8) hold under Assumption 2(2) if the following assumption is satisfied.

Assumption 4*. For each j,

$$\{l_j(\psi; x_{i,j}|x_{l,j}, l < i); i \in \mathbb{Z}\} \quad \text{and} \quad \{l'_j(\psi; x_{i,j}|x_{l,j}, l < i); i \in \mathbb{Z}\}$$

are strongly mixing sequences of random variables with geometric rate.

Lastly we discuss an assumption on the models in $\mathcal{M}$ which allows the consistency of model selection within each stationary piece of time series. It involves the issue of model unidentifiability, which will be investigated based on the recent work of [2]. Let ξ_b and ξ_s correspond to two models in $\mathcal{M}$. We say that ξ_b is a bigger model than ξ_s if for every $\theta_s \in \Theta(\xi_s)$, there exists a (possibly non-unique) $\theta_b^* \in \Theta(\xi_b)$ such that for every $\mathbf{x} = \{x_i; i \in \mathbb{Z}^+\}$, the conditional densities are equal almost everywhere, i.e.,

$$f_{\xi_b}(\cdot|\mathbf{x}; \theta_b^*) = f_{\xi_s}(\cdot|\mathbf{x}; \theta_s)\,. \tag{2.9}$$

Assumption 5. A) For the j-th stationary piece of the time series, there exists a model $\xi_j^o \in \mathcal{M}$ with parameter $\theta_j^o \in \mathbb{R}^{d_j}$, which satisfies $(\xi_j^o, \theta_j^o) = \arg\max_{\xi,\theta} E(l_j((\xi, \theta);\ x_{1,j}|x_{l,j}, l < 1))$. Also, the model ξ_j^o is uniquely identifiable. That is, if there exists a θ_j^* such that $f_{\xi_j^o}(\cdot|\mathbf{x}; \theta_j^*) = f_{\xi_j^o}(\cdot|\mathbf{x}; \theta_j^o)$ almost everywhere for every $\mathbf{x}$, then $\theta_j^* = \theta_j^o$. Moreover, if there exists a model $\psi_s = (\xi_s, \theta_s)$ with $\xi_s \neq \xi_j^o$, $\theta_s \in \mathbb{R}^{d_s}$ such that $f_{\xi_s}(\cdot|\mathbf{x}; \theta_s) = f_{\xi_j^o}(\cdot|\mathbf{x}; \theta_j^o)$ almost everywhere for any $\mathbf{x}$, then $d_s > d_j$.

B) Suppose that ξ_b is a bigger model than ξ_s, with ξ_b and ξ_s associated with parameter vectors $\theta_b \in \Theta(\xi_b) \subset \mathbb{R}^{d_b}$ and $\theta_s \subset \Theta(\xi_s) \subset \mathbb{R}^{d_s}$ respectively. Then θ_b can be partitioned (possibly after some 1-1 continuous transformation) into three sub-vectors, $\theta_b = (\beta, \zeta, \pi)$, where $\beta \in \Theta_\beta \subset \mathbb{R}^{d_\beta}$, $\zeta \in \Theta_\xi \subset \mathbb{R}^{d_s}$, $\pi \in \Theta_\pi \subset \mathbb{R}^{d_\pi}$, Θ_β, Θ_ζ and Θ_π are compact, $d_\beta \geq 1$, $d_\pi \geq 0$, and $d_b = d_\beta + d_s + d_\pi$. In such partition, the vector $\theta_b^* = (0, \theta_s, \pi)$ satisfies (2.9) for any $\pi \in \Theta_\pi$. Moreover, for any given $\pi \in \Theta_\pi$, the vector $\theta_b^* = (0, \theta_s, \pi)$ is the unique vector satisfying (2.9) in the neighborhood $V_\delta(\pi) = \{\theta_b = (\beta, \zeta, \pi) : |\beta| < \delta, |\zeta - \theta_s| < \delta\}$ for some $\delta > 0$.

Assumption 5A) ensures that the true model is of the simplest form in the family $\mathcal{M}$, in the sense that the model cannot be expressed by another model $\psi_s = (\xi_s, \theta_s)$ in $\mathcal{M}$ where θ_s is of smaller dimension. Assumption 5B), which follows [2], tackles the problem of model unidentification. The problem of model unidentification arises when there are more than one θ_b^* satisfying (2.9) provided that (ξ_s, θ_s) is the true model. For example, every ARMA(1,1) model with the same AR and MA coefficient is a white noise, or an ARMA(0,0) model. In this case, there is no unique "true value" in the unidentifiable ARMA(1,1) model, so a Taylor's expansion around the true value of parameter cannot be applied as in the standard theory of likelihood inference. Following [2], after a rearrangement of the parameter vector, the partition (β, ζ) of θ_b is identifiable and the parameter π is unidentified when $\beta = 0$. This allows for the application of a Taylor's expansion around the point $(0, \zeta)$, rather than around the true parameter, which is key to establish the asymptotic properties of the estimators. In fact, this idea has been used previously in [13]. The example below illustrates Assumption 5 for the family of ARMA(p, q) models.

Remark 1. The model class $\mathcal{M}$ is used to specify an objective function for the estimation procedure and it is not necessary that the data in fact has the joint distributions as specified by $\mathcal{M}$. For example, in Section 4, the time series is only assumed to be generated from an ARMA model with some independent and identically distributed noise sequences, while the model class $\mathcal{M}$ assumes Gaussian noise sequences. In such cases the estimation procedure can be interpreted as quasi-likelihood estimation, see [15]. The key point is that Assumption 5 and subsequent conditions hold for the underlying time series. □

The example below illustrates Assumption 5 for the family of ARMA(p, q) models.

Example 2. Consider the family of ARMA(p, q) models satisfying $\Phi(B)(X_t - \mu) = \Theta(B)Z_t$, $Z_t \sim IID(0, \sigma^2)$, where $\Phi(B) = 1 + \sum_{k=1}^{p} \phi_k B^k$ and $\Theta(B) = 1 + \sum_{k=1}^{q} \vartheta_k B^k$ are polynomials in the lag operator B with roots outside the unit circle, and ϕ_p, $\vartheta_q \neq 0$. Recall from Example 1 that $\xi = (p, q)$ specifies the order of the model and $\theta = (\phi_1, \ldots, \phi_p, \vartheta_1, \ldots, \vartheta_q, \mu, \sigma^2)$ specifies the parameters of the model.

First we show that $\xi_b = (p_b, q_b)$ is a bigger model than $\xi_s = (p_s, q_s)$ if $p_s \leq p_b$ and $q_s \leq q_b$. To see this, suppose that the model $\xi_s = (p_s, q_s)$ is associated with the parameter $\theta_s = (\phi_{s,1}, \ldots, \phi_{s,p_s}, \vartheta_{s,1}, \ldots, \vartheta_{s,q_s}, \mu, \sigma^2)$, and the AR and MA polynomials are denoted by $\Phi_s(B) = 1 + \sum_{k=1}^{p_s} \phi_{s,k} B^k$ and $\Theta_s(B) = 1 + \sum_{k=1}^{q_s} \vartheta_{s,k} B^k$ respectively. Consider without loss of generality that $p_b - p_s = p_d$, $q_b - q_s = q_d$ and $p_d - q_d = r$ for some positive integers p_d, q_d and r, i.e., the difference of order between the AR polynomials from models ξ_b and ξ_s is $q_d + r$, and the order difference for the MA polynomials is q_d. Then, consider the ARMA(p_b, q_b) model with

$$\begin{aligned} \Phi_b(B) &= \Phi_s(B)(1 + \pi_1 B + \ldots + \pi_{q_d} B^{q_d})(1 + 0B + \ldots + 0B^r)\,, \qquad (2.10) \\ \Theta_b(B) &= \Theta_s(1 + \pi_1 B + \ldots + \pi_{q_d} B^{q_d})\,, \end{aligned}$$

where $(\pi_1, \ldots, \pi_{q_d})$ is arbitrary. It can be seen that the common factor $(1 + \pi_1 B + \ldots + \pi_{q_d} B^{q_d})$ cancels and the ARMA(p_b, q_b) is the same as the model (ξ_s, θ_s). Thus (2.9) holds and ARMA(p_b, q_b) is a bigger model than ARMA(p_s, q_s).

Next we verify Assumption 5 for ARMA models. First, Assumption 5A) follows from the theory of quasi-likelihood for ARMA models in [7] and [15] and the fact that the polynomials $\Phi(B)$ and $\Theta(B)$ involve non-zero leading coefficients and do not have common zeros. For Assumption 5B), we construct the parameter partition $\theta_b = (\beta, \zeta, \pi)$ which satisfies the required properties. To begin, let $\beta = (\beta_1, \ldots, \beta_{q_d + r})$, $\zeta = (\zeta_1, \ldots, \zeta_{p_s + q_s + 2})$ and $\pi = (\pi_1, \ldots, \pi_{q_d})$.

Reparameterizes $\Phi_b(B)$ and $\Theta_b(B)$ as

$$\begin{aligned}\Phi_b(B) &= (1+\zeta_1 B+\ldots+\zeta_{p_s}B^{p_s})(1+(\beta_1+\pi_1)B+\ldots \\ &\quad +(\beta_{q_d}+\pi_{q_d})B^{q_d})\times(1+\beta_{q_d+1}B+\ldots+\beta_{q_d+r}B^{q_d+r}) \qquad (2.11)\\ \Theta_b(B) &= (1+\zeta_{p_s+1}B+\ldots+\zeta_{p_s+q_s}B^{q_s})(1+\pi_1 B+\ldots+\pi_{q_d}B^{q_d})\,.\end{aligned}$$

Note that the reparameterization (2.11) can be obtained from linear transformation on the usual parameterization $\Phi_b(B) = 1+\sum_{k=1}^{p_b}\phi_{b,k}B^k$ and $\Theta_b(B) = 1+\sum_{k=1}^{q_b}\vartheta_{b,k}B^k$. If $\beta = 0$ and $\zeta = \theta_s$, then (2.11) reduces to (2.10). Thus the vector $\theta_b^* = (0, \theta_s, \pi)$ satisfies (2.9) for any π. Moreover, for any fixed π and sufficiently small $\delta > 0$, the unique factorization of polynomial implies that the vector $\theta_b^* = (0, \theta_s, \pi)$ is the only element in $V_\delta(\pi) = \{\theta_b = (\beta, \zeta, \pi) : |\beta| < \delta, |\zeta - \theta_s| < \delta\}$ such that $\Phi_b(B)$ and $\Theta_b(B)$ in (2.11) can be reduced to (2.10). Thus $\theta_b^* = (0, \theta_s, \pi)$ is the unique vector satisfying (2.9) in $V_\delta(\pi)$, and Assumption 5 holds for ARMA models. □.

3. Main results

In this section we present the main results of the paper. Theorem 1 shows the strong consistency of the MDL procedure when the number of change-points is known. Theorem 2 gives a rate of convergence of the change-point location estimates. Then Theorem 1 is extended to the case of an unknown number of change-points in Theorem 3. Finally, Theorems 4 and 5 give the weak consistency analog of Theorems 2 and 3, where weaker moment conditions are required. To lighten notation, when the model order ξ_j is fixed we suppress ξ_j in specifying the model parameters, e.g., we use $L_j(\theta_j)$ instead of $L_j(\psi_j)$ or $L_j((\xi_j, \theta_j))$, Θ_j instead of $\Theta_j(\xi_j)$, when there is no confusion. We may also suppress the j in n_j, $x_{i,j}$ and $\mathbf{x_j}$.

Proposition 1. *Under Assumption 1(1), 2(1) and 3, for any fixed ξ_j, we have*

$$\sup_{\lambda_d,\lambda_u}\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}\tilde{L}_n^{(j)}(\theta_j,\lambda_d,\lambda_u;\mathbf{x}_j)-(\lambda_u-\lambda_d)L_j(\theta_j)\right| \xrightarrow{a.s.} 0\,, \qquad (3.1)$$

$$\sup_{\lambda_d,\lambda_u}\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}\tilde{L}_n'^{(j)}(\theta_j,\lambda_d,\lambda_u;\mathbf{x}_j)-(\lambda_u-\lambda_d)L_j'(\theta_j)\right| \xrightarrow{a.s.} 0\,, \qquad (3.2)$$

$$\sup_{\lambda_d,\lambda_u}\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}\tilde{L}_n''^{(j)}(\theta_j,\lambda_d,\lambda_u;\mathbf{x}_j)-(\lambda_u-\lambda_d)L_j''(\theta_j)\right| \xrightarrow{a.s.} 0\,, \qquad (3.3)$$

when the supremum notation is defined in (2.4).

In Proposition 1, λ_d and λ_u are restricted to the interval $[0, 1]$, i.e., we search for the maximum inside the stationary piece of time series. For our application to piecewise stationary processes, an extension has to be made so that λ_d and λ_u are allowed to be slightly outside $[0, 1]$, that is, the j-th estimated segment covers part of the $j-1$-th or the $j+1$-th piece of stationary time series. For any real-valued functions $f_n(\lambda_d, \lambda_u)$ on $\mathbb{R}^2$, we use the terminology

$$\sup_{\underline{\lambda_d}, \overline{\lambda_u}} f_n(\lambda_d, \lambda_u) \overset{a.s.}{\to} 0 \tag{3.4}$$

to denote

$$\sup_{\substack{-h_n < \lambda_d < \lambda_u < 1+k_n \\ \lambda_u - \lambda_d > \epsilon_\lambda}} f_n(\lambda_d, \lambda_u) \overset{a.s.}{\to} 0 \,,$$

for any pre-specified positive-valued sequences h_n and k_n which are converging to 0 as $n \to \infty$. The notions such as "$\overset{p}{\to} 0$", "$= O_p(n)$" or "$= O(n)$ almost surely" are defined similarly.

Proposition 2. *Proposition 1 holds with* $\sup_{\lambda_d, \lambda_u}$ *replaced by* $\sup_{\underline{\lambda_d}, \overline{\lambda_u}}$.

Next we discuss the issue of model selection for each of the stationary pieces of the time series. Suppose the data follows a model with parameter $\psi^o = (\xi^o, \theta^o)$, one way to compare a model specified by $\xi \in \mathcal{M}$ to ψ^o is by the Kullback-Leibler (KL) distance, defined by

$$D(f_{\xi^o}; \theta^o | f_\xi; \theta^*) := \mathrm{E}_{\psi^o} \left(\log \frac{f_{\xi^o}(x_1 | x_l, l < 1; \theta^o)}{f_\xi(x_1 | x_l, l < 1; \theta^*)} \right) , \tag{3.5}$$

where

$$\theta^* = \arg \min_{\theta \in \Theta(\xi)} D(f_{\xi^o}; \theta^o | f_\xi; \theta) = \arg \max_{\theta \in \Theta(\xi)} E_{\psi^o}(f_\xi(x_1 | x_l, l < 1; \theta)) , \tag{3.6}$$

and E_{ψ^o} is the expectation under the model in $\mathcal{M}$ with parameter ψ^o. Thus, in (3.6) we find a model in ξ which is closest to ψ^o in terms of KL distance.

KL distance is a measure of discrepancy between two probability densities. It can be shown by Jensen's inequality that it is non-negative and equal to zero if and only if the two densities are equal almost everywhere. For $\psi = (\xi, \theta)$, recall that

$$L_j(\psi) = E_{\psi_j^o} \left(\log f_\xi(x_1 | x_l, l < 1; \theta) \right) .$$

It follows from (3.5) that the maximum of $L_j(\psi)$ is achieved at $L_j(\psi_j^o) = L_j((\xi_b, \theta_b^*))$ whenever the true model $\xi = \xi^o$ or a bigger model $\xi = \xi_b$ is used. For other models $\xi \neq \xi^o$ not bigger than ξ^o, we have

$$L_j(\psi_j^o) > L_j((\xi, \theta^*)) \tag{3.7}$$

by Jensen's inequality. The following proposition shows the convergence of the likelihood function to $L_j((\xi, \theta^*))$ and the consistency of the parameter estimates.

Proposition 3. *Let* $\psi_j^o = (\xi_j^o, \theta_j^o)$ *be the true model parameter. Suppose that a model* ξ_j *is specified for estimation. Define*

$$\begin{aligned} \hat{\theta}_n &\equiv \hat{\theta}_n^{(j)}(\lambda_d, \lambda_u) = \arg \max_{\theta_j \in \Theta_j(\xi_j)} \tilde{L}_n^{(j)}((\xi_j, \theta_j), \lambda_d, \lambda_u; \mathbf{x_j}) , \qquad (3.8) \\ \theta_j^* &= \arg \max_{\theta \in \Theta(\xi_j)} L_j((\xi_j, \theta)) . \end{aligned}$$

If Assumptions 1(1) and 2(1) and 3 hold, then

$$\sup_{\underline{\lambda}_d,\overline{\lambda}_u}\left|\frac{1}{n}L_n^{(j)}((\xi_j,\hat{\theta}_n),\lambda_d,\lambda_u;\mathbf{x_j})-(\lambda_u-\lambda_d)L_j((\xi_j,\theta_j^*))\right|\xrightarrow{a.s.}0\,, \tag{3.9}$$

where the supremum is defined in (3.4)*. Moreover, if $\xi_j=\xi_j^o$ and Assumption 5A) holds, then*

$$\sup_{\underline{\lambda}_d,\overline{\lambda}_u}\left|\hat{\theta}_n(\lambda_d,\lambda_u)-\theta_j^o\right|\xrightarrow{a.s.}0\,. \tag{3.10}$$

If, instead, the specified model ξ_j is bigger than ξ_j^o. If Assumption 5B) holds, then the partition $\hat{\theta}_n\equiv\hat{\theta}_n(\lambda_d,\lambda_u)=(\hat{\beta}_n^{(j)}(\lambda_d,\lambda_u),\hat{\zeta}_n^{(j)}(\lambda_d,\lambda_u),\hat{\pi}_n^{(j)}(\lambda_d,\lambda_u))$ satisfies

$$\sup_{\underline{\lambda}_d,\overline{\lambda}_u}\left|\hat{\beta}_n^{(j)}(\lambda_d,\lambda_u)\right|\xrightarrow{a.s.}0\,, \tag{3.11}$$

$$\sup_{\underline{\lambda}_d,\overline{\lambda}_u}\left|\hat{\zeta}_n^{(j)}(\lambda_d,\lambda_u)-\theta_j^o\right|\xrightarrow{a.s.}0\,. \tag{3.12}$$

From (3.7) and (3.9), it can be seen that if a wrong model is specified, i.e., the selected model is not equal to or bigger than the true one, then the likelihood will be greater than the likelihood of a correctly specified model by an order of n, when n is sufficiently large. As we have seen from Section 2, MDL can be regarded as the minus log-likelihood plus a penalty term of order $\log n$. Therefore, the likelihood function dominates and hence the MDL procedure is able to select the correct order of model or a bigger model in $\mathcal{M}$. $D(f_{\xi^o};\theta^o|f_{\hat{\xi}};\theta^*)>0$, where D is defined in (3.5). On the other hand, we say that $\hat{\xi}$ *overestimates* ξ^o if $D(f_{\xi^o};\theta^o|f_{\hat{\xi}};\theta^*)=0$ and the model parameter θ^* has a higher dimension than θ^o. Note that if $\hat{\xi}$ is not underestimating the true model ξ^o, then $D(f_{\xi^o};\theta^o|f_{\hat{\xi}};\theta^*)=0$ and the true probability density of the segment can be identified. Theorem 1 gives a preliminary result about the convergence of the change-point location estimates and the order parameter estimates when the number of change-points is known.

Theorem 1. *Let $\mathbf{y}=\{y_t;t=1,\ldots,n\}$ be observations from a piecewise stationary process specified by the vector $(m_o,\boldsymbol{\lambda^o},\boldsymbol{\psi^o})$ and satisfying Assumptions 1(1), 2(1) and 3. Suppose the number of change-points m_o is known. We estimate the locations and the model parameter by*

$$\{\boldsymbol{\hat{\lambda}_n},\boldsymbol{\hat{\psi}_n}\}=\arg\min_{\boldsymbol{\psi},\boldsymbol{\lambda}\in A_\epsilon^{m_o}}\frac{2}{n}\mathrm{MDL}(m_o,\boldsymbol{\lambda},\boldsymbol{\psi})\,,$$

where $A_{\epsilon_\lambda}^{m_o}$ is defined in (2.2)*. Then, $\boldsymbol{\hat{\lambda}_n}\overset{a.s.}{\to}\boldsymbol{\lambda^o}$ and for each segment the estimated model does not underestimate the true model.*

Proof. Let B be the probability one set in which Proposition 2 and 3 holds. We will show that for each $\omega\in B$, $\boldsymbol{\hat{\lambda}_n}\to\boldsymbol{\lambda^o}$ and $\boldsymbol{\hat{\psi}_n}\to\boldsymbol{\psi^o}$. To begin, for any $\omega \in B$, suppose on the contrary that $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} \nrightarrow \boldsymbol{\lambda^o}$. Because the values of $\boldsymbol{\lambda}$ are bounded, there exists a subsequence $\{n_k\}$ such that $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} \to \boldsymbol{\lambda^*} \neq \boldsymbol{\lambda^o}$ along the subsequence. Note that $\boldsymbol{\lambda^*} \in A_{\epsilon_\lambda}^{m_o}$ since $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} \in A_{\epsilon_\lambda}^{m_o}$ for all n. Recall that $(\hat{\xi}_j, \hat{\theta}_n^{(j)})$ is the estimator for the model order and model parameters for the j-th segment. Since $\mathcal{M}$ is a finite set, without lost of generality we can assume that $\hat{\xi}_j$ converges to ξ_j^*, say, along $\{n_k\}$. Similarly, as $\Theta_j \equiv \Theta_j(\xi_j)$ is compact for every ξ_j, we can assume that $\hat{\theta}_n^{(j)}$ converges to θ_j^*, say, along $\{n_k\}$. To lighten notations we replace n_k by n. It follows that for all sufficiently large n,

$$\frac{2}{n}\mathrm{MDL}(m_o, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) = c_n - \frac{1}{n}\sum_{j=1}^{m+1} L_n^{(j)}((\xi_j^*, \hat{\theta}_n^{(j)}), \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y}),$$

where c_n is deterministic with order $O(\log(n)/n)$.

For each limiting estimated interval $I_j^* = (\lambda_{j-1}^*, \lambda_j^*)$, $j = 1, \ldots, m+1$, there are two possible cases. First, I_j^* is nested in the true i-th interval $(\lambda_{i-1}^o, \lambda_i^o)$. Second, I_j^* covers (fully or partly) $k+2$ $(k \geq 0)$ true intervals $(\lambda_{i-1}^o, \lambda_i^o), \ldots, (\lambda_{i+k}^o, \lambda_{i+k+1}^o)$. We consider each of these two cases separately.

Case 1. If $\lambda_{i-1}^o \leq \lambda_{j-1}^* < \lambda_j^* \leq \lambda_i^o$. In particular, if $\lambda_{i-1}^o < \lambda_{j-1}^* < \lambda_j^* < \lambda_i^o$, then for sufficiently large n the estimated j-th segment will be a part of the stationary processes from the true i-th segment. If $\lambda_j^* = \lambda_i^o$ or $\lambda_{j-1}^* = \lambda_{i-1}^o$, then as $\hat{\lambda}_{j-1} \to \lambda_{i-1}^o$ and $\hat{\lambda}_j \to \lambda_i^o$, the estimated segment can only include a decreasing proportion of observations from the adjacent segments. Then $\max(\hat{\lambda}_j - \lambda_i^o, 0)$ and $\max(\lambda_{i-1}^o - \hat{\lambda}_{j-1}, 0)$ play the role of h_n and k_n in (3.4). So, we have from Proposition 2 that

$$\frac{1}{n}L_n^{(j)}(\hat{\theta}_n^{(j)}, \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y}) \xrightarrow{a.s.} (\lambda_j^* - \lambda_{j-1}^*)L_i((\xi_j^*, \theta_j^*)). \tag{3.13}$$

In particular, if $\xi_j^* = \xi_i^o$, then θ_j^* is in fact θ_i^o, the true parameter value of the i-th segment. Then the last quantity on (3.13) is in fact $(\lambda_j^* - \lambda_{j-1}^*)L_i((\xi_i^o, \theta_i^o))$. If ξ_j^* underestimates ξ_i^o, then $D(f_{\xi_i^o}; \theta_i^o | f_{\xi_j^*}; \theta_j^*) > 0$, which implies

$$L_i((\xi_i^o, \theta_i^o)) - L_i((\xi_j^*, \theta_j^*)) > 0. \tag{3.14}$$

Case 2. If $\lambda_{i-1}^o \leq \lambda_{j-1}^* < \lambda_i^o < \ldots < \lambda_{i+k}^o < \lambda_j^* \leq \lambda_{i+k+1}^o$ for some $k \geq 0$, then for sufficiently large n the estimated j-th segment contains observations from at least two pieces of different stationary processes and is thus non-stationary. We will partition the likelihood by the true configuration of the series, i.e.,

$$\begin{aligned}
&\frac{1}{n}L_n^{(j)}(\hat{\theta}_n^{(j)}, \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y}) \\
= \; & \frac{1}{n}L_n^{(j)}(\hat{\theta}_n^{(j)}, \hat{\lambda}_{j-1}, \lambda_i^o, \mathbf{y}) + \frac{1}{n}\sum_{l=i}^{i+k-1} L_n^{(j)}(\hat{\theta}_n^{(j)}, \lambda_l^o, \lambda_{l+1}^o, \mathbf{y}) \\
&+ \frac{1}{n}L_n^{(j)}(\hat{\theta}_n^{(j)}, \lambda_{i+k}^o, \hat{\lambda}_j, \mathbf{y}).
\end{aligned} \tag{3.15}$$

Each of the likelihood functions in (3.15) involves observations from one piece of the stationary time series. From Proposition 1 and the fact that $L_l((\xi_l^o, \theta_l^o)) \geq L_j((\xi_j^*, \theta_j^*))$ for all $l = i, \ldots, i+k+1$, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} L_n^{(j)}(\hat{\theta}_n^{(j)}, \hat{\lambda}_{j-1}, \lambda_i^o, \mathbf{y}) &\leq (\lambda_i^o - \lambda_{j-1}^*) L_i((\xi_i^o, \theta_i^o))\,, \\
\lim_{n\to\infty} \frac{1}{n} L_n^{(j)}(\theta_j^*, \lambda_l^o, \lambda_{l+1}^o, \mathbf{y}) &\leq (\lambda_{l+1}^o - \lambda_l^o) L_{l+1}((\xi_{l+1}^o, \theta_{l+1}^o))\,, \\
\lim_{n\to\infty} \frac{1}{n} L_n^{(j)}(\theta_j^*, \lambda_{i+k}^o, \hat{\lambda}_j, \mathbf{y}) &\leq (\lambda_j^* - \lambda_{i+k}^o) L_{i+k+1}((\xi_{i+k+1}^o, \theta_{i+k+1}^o))\,.
\end{aligned}$$

Note that strict inequalities hold for at least one of the above equations since (ξ_j^*, θ_j^*) cannot correctly specify the model for all different segments. Thus

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} L_{n,j}(\hat{\theta}_n^{(j)}, \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y}) \\
<\; & (\lambda_i^o - \lambda_{j-1}^*) L_i((\xi_i^o, \theta_i^o)) + \sum_{l=i}^{i+k-1} (\lambda_{l+1}^o - \lambda_l^o) L_l((\xi_l^o, \theta_l^o)) \\
& + (\lambda_j^* - \lambda_{i+k}^o) L_{i+k+1}((\xi_{i+k+1}^o, \theta_{i+k+1}^o))\,. \qquad (3.16)
\end{aligned}$$

Now, as the number of estimated segments is equal to the true number of segments and $\lambda^* \neq \boldsymbol{\lambda^o}$, there is at least one segment in which case 2 applies. Thus for sufficiently large n,

$$\begin{aligned}
& \frac{2}{n}\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) \\
> \; & \frac{c_n}{n} - \sum_{i=1}^{m+1} (\lambda_i^o - \lambda_{i-1}^o) L_i((\xi_i^o, \theta_i^o)) \;\; [(3.16) \text{ for at least one piece}] \\
= \; & \frac{2}{n}\mathrm{MDL}(m, \boldsymbol{\lambda^o}, \boldsymbol{\psi^o}) \qquad [\text{Definition of MDL.}] \\
\geq \; & \frac{2}{n}\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}})\,, \qquad [\text{Property of the estimator.}] \qquad (3.17)
\end{aligned}$$

which is a contradiction. Hence $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} \to \boldsymbol{\lambda^o}$ for all $\omega \in B$.

On the other hand, if $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} \to \boldsymbol{\lambda^o}$ but in some of the segments the estimated model underestimates the true model, then in Case 1 (3.14) holds for those segments and the contradiction (3.17) arises. Thus the estimated model does not underestimates the true model. This completes the proof of Theorem 1. □

Corollary 1. *Under the conditions of Theorem 1, if the number of change-points is unknown and is estimated from the data using (2.3), then*

A) The number of change-points cannot be under-estimated. That is, $\hat{m} \geq m_o$ for n large almost surely.

B) When $\hat{m} > m$, $\boldsymbol{\lambda^o}$ must be a subset of the limit points of $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}$, in the sense that given any $\omega \in B$, $\epsilon > 0$ and $\lambda_j^o \in \boldsymbol{\lambda^o}$, there exists a $\hat{\lambda}_k \in \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}$ such that $|\lambda_j^o - \hat{\lambda}_k| < \epsilon$ for sufficiently large n. In other words, the true change-point locations can be identified.

C) The order of the model in each segment cannot be under-estimated.

Proof. Notice that in the proof of Theorem 1, the assumption of known number of change-points is only used to ensure that case 2 applies, i.e., $\lambda^o_{i-1} \leq \lambda^*_{j-1} < \lambda^o_i < \ldots < \lambda^o_{i+k} < \lambda^*_j \leq \lambda^o_{i+k+1}$, for at least one j. In fact, no matter how many segments $\boldsymbol{\lambda^*}$ contains, contradiction (3.17) arises whenever case 2 applies. From this observation, A) and B) follow. Thus we can assume that only case 1 applies. If any of the model segments is underestimated, then (3.14) still arises and leads to the contradiction (3.17), proving C). □

The next result is about the convergence rate of the change-point estimates.

Theorem 2. *Suppose that Assumption 1(p), 2(q), 3 and 4(w) hold with $p \geq 2$, $q \geq 4$ and $w = \max\left(\frac{1}{p}, \frac{2}{q}\right) \leq \frac{1}{2}$. Let $\boldsymbol{\lambda^o} = (\lambda^o_1, \ldots, \lambda^o_{m_o})$ be the true change-point configuration. Then, with $(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}})$ defined in (2.3), for each $j = 1, 2, \ldots, m_o$, there exists a $\hat{\lambda}_{i_j} \in \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}$, $1 \leq i_j \leq \hat{m}$ such that*

$$|\lambda^o_j - \hat{\lambda}_{i_j}| = o\left(n^{w-1}\right) , \tag{3.18}$$

almost surely. Alternatively, (3.18) holds if Assumption 1(p), 2(q), 3 and 4 are satisfied with $p \geq 2$ and $q \geq 4$.*

Proof. From Corollary 1, we can assume that $\hat{m} \geq m$ and for each λ^o_j there exists a $\hat{\lambda}_{i_j}$ such that $|\lambda^o_j - \hat{\lambda}_{i_j}| = o(1)$ a.s. This theorem gives a bound for the convergent rate of the change-point location estimators. We prove this by contradiction. Let B be the probability one set in which Theorem 1 holds. For each $\omega \in B$, suppose that for some λ^o_l there does not exist any $\hat{\lambda}_{i_l}$ such that (3.18) holds. Then there exists a subsequence n_k, and a constant c such that either one of

$$\begin{array}{ll} i) & \lambda^o_l - \hat{\lambda}_{i_l} > c n_k^{w-1} \qquad or \\ ii) & \hat{\lambda}_{i_l} - \lambda^o_l > c n_k^{w-1} \end{array} \tag{3.19}$$

holds, where $\hat{\lambda}_{i_l}$ is the location estimate closest to λ^o_l. From Corollary 1 and the boundedness of $\boldsymbol{\lambda_n}$, we can further assume that $\hat{\lambda}_{i_l-1} \overset{a.s.}{\to} \lambda^*_{i_l-1}$, $\hat{\lambda}_{i_l} \overset{a.s.}{\to} \lambda^o_l \equiv \lambda^*_{i_l}$ and $\hat{\lambda}_{i_l+1} \overset{a.s.}{\to} \lambda^*_{i_l+1}$ with $\lambda^o_{l-1} \leq \lambda^*_{i_l-1} < \lambda^o_l < \lambda^*_{i_l+1} \leq \lambda^o_{l+1}$. Without loss of generality we replace $\{n_k\}$ by $\{n\}$.

Let $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{n}} = \{\hat{\lambda}_1, \ldots, \hat{\lambda}_{i_l-1}, \lambda^o_l, \hat{\lambda}_{i_l+1} \ldots, \hat{\lambda}_m\}$. A contradiction arises if we show that for all sufficiently large n,

$$MDL(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) > MDL(\hat{m}, \tilde{\boldsymbol{\lambda}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) \, . \tag{3.20}$$

Note that as the number of estimated change-points and the number of the models are bounded, $\hat{m}$ and the orders of the models in each segment can be chosen to be the same along the subsequence.

The difference, $MDL(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) - MDL(\hat{m}, \tilde{\boldsymbol{\lambda}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}})$, is either

$$\begin{array}{ll} i) & \sum_{k=n_l-[n(\lambda^o_l-\hat{\lambda}_{i_l})]+1}^{n_l} \left(l_{i_l}(\hat{\theta}_{i_l}; x_{k,l}|\mathbf{y_{k,l}}) - l_{i_l+1}(\hat{\theta}_{i_l+1}; x_{k,l}|\mathbf{y_{k,l}}) \right) , \qquad or \\ ii) & \sum_{k=1}^{[n(\hat{\lambda}_{i_l}-\lambda^o_l)]} \left(l_{i_l+1}(\hat{\theta}_{i_l+1}; x_{k,l}|\mathbf{y_{k,l}}) - l_{i_l}(\hat{\theta}_{i_l}; x_{k,l}|\mathbf{y_{k,l}}) \right) , \end{array}$$

according to the two cases in (3.19). By Assumption 1(p), the above can be expressed as either one of

$$\begin{aligned} &i) \quad \sum{}' \left(l_{i_l}(\hat{\theta}_{i_l}; x_{k,l}|x_{j,l}, j<k) - l_{i_l+1}(\hat{\theta}_{i_l+1}; x_{k,l}|x_{j,l}, j<k) \right) + o\left(n^{\frac{1}{p}}\right) \\ &ii) \quad \sum{}'' \left(l_{i_l+1}(\hat{\theta}_{i_l+1}; x_{k,l}|x_{j,l}, j<k) - l_{i_l}(\hat{\theta}_{i_l}; x_{k,l}|x_{j,l}, j<k) \right) + o\left(n^{\frac{1}{p}}\right), \end{aligned} \tag{3.21}$$

where $p \geq 2$, $\sum' = \sum_{k=n_l-[n(\lambda_l^o-\hat{\lambda}_{i_l})]+1}^{n_l}$ and $\sum'' = \sum_{k=1}^{[n(\hat{\lambda}_{i_l}-\lambda_l^o)]}$. By Proposition 3, we have for case (i) in (3.21) that $l_{i_l}(\hat{\theta}_{i_l}; \cdot) \overset{a.s.}{\to} l_l(\theta_l^o; \cdot)$ and $l_{i_l+1}(\hat{\theta}_{i_l+1}; \cdot) \overset{a.s.}{\to} l_l(\theta_{l+1}^o; \cdot)$. Similarly, for case (ii), we have $l_{i_l}(\hat{\theta}_{i_l}; \cdot) \overset{a.s.}{\to} l_{l+1}(\theta_l^o; \cdot)$ and $l_{i_l+1}(\hat{\theta}_{i_l+1}; \cdot) \overset{a.s.}{\to} l_{l+1}(\theta_{l+1}^o; \cdot)$. Moreover, in case (i) where $n_l - [n(\lambda_l^o - \hat{\lambda}_i)] < k \leq n_l$, the observations are from the l-th segment and

$$E_{\theta_l^o}(l_l(\theta_l^o; x_{k,l}|x_{s,l}, s<k) - l_l(\theta_{l+1}^o; x_{k,l}|x_{s,l}, s<k)) > 0 \tag{3.22}$$

by Jensen's inequality. Similarly, in case (ii) where $n_l < k \leq [n(\hat{\lambda}_i - \lambda_l^o)]$, the observations are from the $(l+1)$-th segment and we have

$$E_{\theta_{l+1}^o}(l_{l+1}(\theta_{l+1}^o; x_{k,l+1}|x_{s,l+1}, s<k) - l_{l+1}(\theta_l^o; x_{k,l+1}|x_{s,l+1}, s<k)) > 0. \tag{3.23}$$

Using the ergodic theorem and (3.22) for (i) and Assumption 4(w) with $g(n) = n^w$ and (3.23) for (ii), we see that the first quantities in both cases of (3.21) are positive and of order $O(n^w)$ but not smaller. Since $w \geq \frac{1}{p}$ by construction, the two quantities in (3.21) are positive for sufficiently large n, yielding the contradiction (3.20). On the other hand, if Assumption 4* holds, then (i) and (ii) in the Lemma 1 below can be applied respectively to the two cases of (3.21) with $r = q$ and $g(n) = n^w$ and the same conclusion follows. This completes the proof of Theorem 2. □

Lemma 1. *If $\{X_t\}$ is a sequence of stationary, zero-mean strongly mixing process with geometric rate, and $E(|X_1|^{r+\epsilon}) < \infty$ for some $2 \leq r < \infty$ and $\epsilon > 0$, then*

$$\begin{aligned} &i) \quad \tfrac{1}{g(n)} \textstyle\sum_{t=1}^{g(n)} X_t \xrightarrow{a.s.} \mu\,, \\ &ii) \quad \tfrac{1}{g(n)} \textstyle\sum_{t=n-g(n)+1}^{n} X_t \xrightarrow{a.s.} \mu\,, \end{aligned}$$

for any sequence $\{g(n)\}_{n\geq 1}$ of integers that satisfies $g(n) > cn^{2/r}$ for some $c > 0$ when n is sufficiently large. Moreover,

$$\begin{aligned} &iii) \quad \textstyle\sum_{t=1}^{s(n)} X_t = O(n^{2/r})\,, \\ &iv) \quad \textstyle\sum_{t=n-s(n)+1}^{n} X_t = O(n^{2/r})\,, \end{aligned}$$

almost surely, for any sequence $\{s(n)\}_{n\geq 1}$ satisfying $s(n) = O(n^{2/r})$.

Using the convergence rate of the change-point estimator obtained in Theorem 2, the following lemma shows that the convergence rate of the maximum likelihood estimator is not affected even when the estimated piece may not be fully inside a stationary piece of a time series but involves part of the adjacent stationary pieces.

Lemma 2. *Suppose that Assumption 1(p), 2(q) and 3 hold with $p \geq 2, q \geq 4$ and $\omega = \max\left(\frac{1}{p}, \frac{2}{q}\right) \leq \frac{1}{2}$. If the true model for the j-th piece, ξ_j^o, $j = 1, \ldots, m+1$, is specified and Assumption 5A) holds, then*

$$\hat{\theta}_n^{(j)}(\hat{\lambda}_{j-1}, \hat{\lambda}_j) - \theta_j^o = O\left(\sqrt{\frac{\log\log n}{n}}\right) \qquad \text{a.s.}\,, \tag{3.24}$$

where θ_j^o is the true parameter vector and $\hat{\theta}_n^{(j)}(\hat{\lambda}_{j-1}, \hat{\lambda}_j)$ is defined in (3.8) with $\xi_j = \xi_j^o$.

Suppose the specified model ξ_j is bigger than the true model ξ_j^o and Assumption 5B) holds, then we have the partition $\hat{\theta}_n \equiv \hat{\beta}_n^{(j)}(\hat{\lambda}_{j-1}, \hat{\lambda}_j)$, $\hat{\zeta}_n^{(j)}(\hat{\lambda}_{j-1}, \hat{\lambda}_j)$, $\hat{\pi}_n^{(j)}(\hat{\lambda}_{j-1}, \hat{\lambda}_j))$, with

$$\hat{\beta}_n^{(j)}(\hat{\lambda}_{j-1}, \hat{\lambda}_j) = O\left(\sqrt{\frac{\log\log n}{n}}\right) \qquad \text{a.s.}\,, \tag{3.25}$$

$$\hat{\zeta}_n^{(j)}(\hat{\lambda}_{j-1}, \hat{\lambda}_j) - \theta_j^o = O\left(\sqrt{\frac{\log\log n}{n}}\right) \qquad \text{a.s.}\,. \tag{3.26}$$

Finally we come to the main results of this paper, the consistency of MDL model selection procedure.

Theorem 3. *Let $\mathbf{y} = \{y_t; t = 1, \ldots, n\}$ be a piecewise stationary process specified by the vector $\{m_o, \boldsymbol{\lambda^o}, \boldsymbol{\psi^o}\}$ and satisfy Assumptions 1(2), 2(4), 3, 5 and either Assumptions 4(0.5) or 4*. For the estimator $\{\hat{m}_n, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}\}$ defined in (2.3), we have*

$$\hat{m}_n \xrightarrow{a.s.} m_o, \quad \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} \xrightarrow{a.s.} \boldsymbol{\lambda^o}, \quad \hat{\boldsymbol{\psi}}_{\boldsymbol{n}} \xrightarrow{a.s.} \boldsymbol{\psi^o}\,.$$

Proof. Again the proof will be by contradiction. Fix any $\omega \in B$ where B is a probability one set satisfying the conclusions of Theorems 1, 2 and Lemma 2. Suppose that $\hat{m}_n \nrightarrow m$ on ω. As the estimated number of change-points is bounded, there exists a m^* and a subsequence n_k such that $\hat{m}_n = m^* \neq m_o$ for sufficiently large k. From Corollary 1, $m^* > m_o$. As the relative change-points take values in a compact interval [0, 1], we can assume that there exists a limiting partition $\boldsymbol{\lambda^*}$ such that $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} \rightarrow \boldsymbol{\lambda^*} := (\lambda_1^*, \ldots, \lambda_{m^*}^*)$. From Corollary 1 again, $\boldsymbol{\lambda^o} = \{\lambda_1^o, \ldots, \lambda_{m_o}^o\}$ is a subset of $\{\lambda_1^*, \ldots, \lambda_{m^*}^*\}$. Thus every segment in the limiting partition $\boldsymbol{\lambda^*}$ is contained in exactly one of the stationary segments of the true partition $\boldsymbol{\lambda^o}$. As the number of models in the family $\mathcal{M}$ is finite, by taking further subsequences we can assume that $\hat{\xi}_j = \xi_j^*$ for sufficiently large

n_k. From Corollary 1 again, ξ_j^* must be no less than the dimension of the true model. Without loss of generality we replace n_k by n.

For sufficiently large n, the MDL for the model $\{\hat{m}_n, \boldsymbol{\hat{\lambda}_n}, \boldsymbol{\hat{\psi}_n}\}$ is given by

$$C_1 - \sum_{j=1}^{m^*+1} \tilde{L}_n^{(j)}((\xi_j^*, \hat{\theta}_n^{(j)}); \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y})\,, \tag{3.27}$$

where $C_1 = O(\log n/n)$. As the limiting partition $\boldsymbol{\lambda^*}$ is finer than the true partition $\boldsymbol{\lambda^o}$, we assume without loss of generality that the k-th true segment contains $d > 1$ segments from $\boldsymbol{\lambda^*}$, say from the $(i+1)$-th to the $(i+d)$-th segments, i.e., $\lambda_{k-1}^o = \lambda_i^*$ and $\lambda_k^o = \lambda_{i+d}^*$. Consider fitting one model over the d segments and let

$$\begin{aligned}
\tilde{\theta}_n &= \arg\max_{\theta} \tilde{L}_n^{(i+1)}((\xi_k^o, \theta), \hat{\lambda}_i, \hat{\lambda}_{i+d}, \mathbf{y})\,,\\
\boldsymbol{\tilde{\lambda}_n} &= \{\hat{\lambda}_1, \ldots, \hat{\lambda}_{i-1}, \hat{\lambda}_i, \hat{\lambda}_{i+d}, \hat{\lambda}_{i+d+1}, \ldots \hat{\lambda}_{m^*-1}, \hat{\lambda}_{m^*}\}\,,\\
\boldsymbol{\tilde{\xi}_n} &= \{\xi_1^*, \ldots, \xi_{i-1}^*, \xi_i^*, \xi_k^o, \xi_{i+d+1}^*, \ldots \xi_{m^*}^*, \xi_{m^*+1}^*\}\,,\\
\boldsymbol{\tilde{\theta}_n} &= \{\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(i-1)}, \hat{\theta}_n^{(i)}, \tilde{\theta}_n, \hat{\theta}_n^{(i+d+1)}, \ldots, \hat{\theta}_n^{(m^*)}, \hat{\theta}_n^{(m^*+1)}\}\,,
\end{aligned}$$

and $\boldsymbol{\tilde{\psi}_n} = (\boldsymbol{\tilde{\xi}_n}, \boldsymbol{\tilde{\theta}_n})$. Note that for sufficiently large n, the MDL for the model $\{m^* - d + 1, \boldsymbol{\tilde{\lambda}_n}, \boldsymbol{\tilde{\psi}_n}\}$ is given by

$$C_2 - \sum_{\substack{1 \le j \le m^*+1\\ j \ne i+1, \ldots, i+d}} \tilde{L}_n^{(j)}((\xi_j^*, \hat{\theta}_n^{(j)}); \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y}) - \tilde{L}_n^{(i+1)}((\xi_k^o, \tilde{\theta}_n); \hat{\lambda}_i, \hat{\lambda}_{i+d}, \mathbf{y})\,, \tag{3.28}$$

where $C_1 - C_2 = O(\log n/n)$ and is positive because C_1 contains codes for more segments and no segment is underestimated. The difference between the MDLs in (3.27) and (3.28) is

$$C_1 - C_2 + \tilde{L}_n^{((i+1))}((\xi_k^o, \tilde{\theta}_n); \hat{\lambda}_i, \hat{\lambda}_{i+d}, \mathbf{y}) - \sum_{j=i+1}^{i+d} \tilde{L}_n^{(j)}((\xi_j^*, \hat{\theta}_n^{(j)}); \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y})\,. \tag{3.29}$$

A contradiction arises if the quantity in (3.29) is positive for sufficiently large n. As $C_1 - C_2 = O(\log n/n)$ is positive, it suffices to show that the difference between the last two terms in (3.29) is of order $o(\log n/n)$.

We first consider the case that for each $j = i+1, \ldots, i+d$, the model are correctly specified, i.e., $\xi_j^* = \xi_l^o$ if the j-th estimated segment is within the l-th true segment. From Lemma 2 we have $\hat{\theta}_n^{(j)} - \theta_j^* \overset{a.s.}{\to} 0$, where $\theta_j^* = \theta_l^o$ if $\xi_j^* = \xi_l^o$. By the definition of $\hat{\theta}_n^{(j)}$ we have $\tilde{L}_n'^{(j)}((\xi_j^*, \hat{\theta}_n^{(j)}); \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y}) = 0$, $j = i, \ldots, i+d$. A Taylor's series expansions of $\tilde{L}_n^{(j)}((\xi_j^*, \theta_j^*); \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y})$ around $\hat{\theta}_n^{(j)}$ gives

$$\begin{aligned}
\tilde{L}_n^{(j)}((\xi_j^*, \theta_j^*); \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y}) &= \tilde{L}_n^{(j)}((\xi_j^*, \hat{\theta}_n^{(j)}); \hat{\lambda}_{j-1}, \hat{\lambda}_j, \mathbf{y})\\
&\quad + (\hat{\theta}_n^{(j)} - \theta_j^*)^T \tilde{L}_n''^{(j)}((\xi_j^*, \theta_j^+); \hat{\lambda}_{j-1}; \hat{\lambda}_j, \mathbf{y})(\hat{\theta}_n^{(j)} - \theta_j^*)\,,
\end{aligned} \tag{3.30}$$

with $|\theta_j^+ - \theta_j^*| < |\hat\theta_n^{(j)} - \theta_j^*|$. Similarly, we have

$$\begin{aligned}\tilde L_n^{(i+1)}((\xi_k^o,\theta_k^o);\hat\lambda_i,\hat\lambda_{i+d},\mathbf{y}) &= \tilde L_n^{(i+1)}((\xi_k^o,\tilde\theta_n);\hat\lambda_i,\hat\lambda_{i+d},\mathbf{y})\\ &\quad + (\tilde\theta_n-\theta_k^o)^T\tilde L_n''^{(i+1)}((\xi_k^o,\tilde\theta^+);\hat\lambda_i;\hat\lambda_{i+d},\mathbf{y})(\tilde\theta_n-\theta_k^o)\,,\end{aligned}\tag{3.31}$$

with $|\tilde\theta^+ - \theta_k^o| < |\tilde\theta_n - \theta_k^o|$. On the other hand, as the model cannot be underestimated for each segment, we have

$$f_{\xi_i^*}(\cdot;\theta_i^*) = \ldots = f_{\xi_{i+d}^*}(\cdot;\theta_{i+d}^*) = f_{\xi_k^o}(\cdot;\theta_k^o)\,,$$

almost everywhere. Thus

$$\begin{aligned}\tilde L_n^{(i+1)}((\xi_k^o,\theta_k^o);\hat\lambda_i,\hat\lambda_{i+d},\mathbf{y}) &= \sum_{j=i+1}^{i+d}\tilde L_n^{(j)}((\xi_j^*,\theta_j^*);\hat\lambda_{j-1},\hat\lambda_j,\mathbf{y}) &(3.32)\\ &= \sum_{j=[n\hat\lambda_i]+1}^{[n\hat\lambda_{i+d}]} l_k((\xi_k^o,\theta_k^o),y_j|y_l,l<j)\,.\end{aligned}$$

Thus, using (3.30) to (3.32), the last two terms in (3.29) reduces to

$$\begin{aligned}&(\tilde\theta_n-\theta_k^o)^T\frac{1}{n}\tilde L_n''^{(i+1)}((\xi_k^o,\tilde\theta^+);\hat\lambda_i,\hat\lambda_{i+d},\mathbf{y})(\tilde\theta_n-\theta_k^o)\\ &-\sum_{j=i+1}^{i+d}(\hat\theta_n^{(j)}-\theta_j^*)^T\frac{1}{n}\tilde L_n''^{(j)}((\xi_j^*,\theta_j^+);\hat\lambda_{j-1},\hat\lambda_j,\mathbf{y})(\hat\theta_n^{(j)}-\theta_j^*)\,.\end{aligned}\tag{3.33}$$

From Proposition 2 and Lemma 2, the quantity in (3.33) is of order $O(\log\log n/n) = o(\log n/n)$, and thus the quantity in (3.29) is positive for sufficiently large n, contradicting (2.3). For the case where ξ_j^* is a bigger model than ξ_l^o, using Assumption 5B) we may apply Taylor's expansion on the first two components of $\hat\theta_n^{(j)}$ as in the proof of Lemma 2, and the last two terms in (3.29) are also of order $O\big(\frac{\log\log n}{n}\big)$ and the same conclusion follows. This completes the proof of Theorem 3. □

The strong consistency of the MDL model selection procedure requires the existence of the q-th ($q>4$) moments of the likelihood and the score functions. The moment condition is mainly used in Lemma 1(ii) for the almost sure convergence of the average of the observations in the end of a sequence, where the ergodic theorem does not apply. The weak consistency of MDL model selection procedure can be shown under weaker conditions as the moment condition can be avoided.

Assumption 1*. Under the notation of Assumption 1, for $j = 1,\ldots,m+1$

$$\begin{aligned}\sup_{\lambda_d,\lambda_u}\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}L_n^{(j)}(\theta_j,\lambda_d,\lambda_u,\mathbf{x_j}) - \frac{1}{n}\tilde L_n^{(j)}(\theta_j,\lambda_d,\lambda_u,\mathbf{x_j})\right| &= O_p\left(n^{-1}\right)\,,\\ \sup_{\lambda_d,\lambda_u}\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}L_n'^{(j)}(\theta_j,\lambda_d,\lambda_u,\mathbf{x_j}) - \frac{1}{n}\tilde L_n'^{(j)}(\theta_j,\lambda_d,\lambda_u,\mathbf{x_j})\right| &= O_p\left(n^{-1}\right)\,,\\ \sup_{\lambda_d,\lambda_u}\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}L_n''^{(j)}(\theta_j,\lambda_d,\lambda_u,\mathbf{x_j}) - \frac{1}{n}\tilde L_n''^{(j)}(\theta_j,\lambda_d,\lambda_u,\mathbf{x_j})\right| &= o_p\left(1\right)\,.\end{aligned}$$

Theorem 4. *Suppose that Assumptions 1(1), 1*, 2(1) and 3 hold. Using the notation of Theorem 2, for each λ_j^o, there exists a $\hat{\lambda}_{i_j}$, $1 \leq i_j \leq \hat{m}$, such that for any $\delta > 0$,*

$$|\lambda_j^o - \hat{\lambda}_{i_j}| = O_p\left(n^{\delta-1}\right) .$$

Proof. From Corollary 1 we can assume that $\hat{m} \geq m_o$ and for each λ_j^o there exists a $\hat{\lambda}_{i_j}$ such that $|\lambda_j^o - \hat{\lambda}_{i_j}| = o(1)$ a.s., where $1 < i_1 < i_2 < \ldots < i_m < \hat{m}$. By construction, for every $k = 0, \ldots, \hat{m}-1$, we have $|\hat{\lambda}_{k+1} - \hat{\lambda}_k| > \epsilon_\lambda$, so $\hat{\lambda}_{i_j}$ is the estimated location of change-point closest to λ_j^o for sufficiently large n. We have to prove that, for any $\delta > 0$, there exists a $c > 0$ such that

$$P(\exists l, |\lambda_l^o - \hat{\lambda}_{i_l}| > cn^{\delta-1}) \to 0 .$$

By the definition of $(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}})$, it suffices to show that

$$P(MDL(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) < MDL(\hat{m}, \tilde{\boldsymbol{\lambda}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}), \exists l, |\lambda_l^o - \hat{\lambda}_{i_l}| > cn^{\delta-1}) \to 0 ,$$

where $\tilde{\boldsymbol{\lambda}}$ is the same as $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}$, except that the $\hat{\lambda}_{i_l}$ is replaced by λ_l^o for any l that satisfies $|\lambda_l^o - \hat{\lambda}_{i_l}| > cn^{\delta-1}$. As the number of change-points is bounded, it suffices to prove that, for each fixed l,

$$P(MDL(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) < MDL(\hat{m}, \tilde{\boldsymbol{\lambda}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}), |\lambda_l^o - \hat{\lambda}_{i_l}| > cn^{\delta-1}) \to 0 . \tag{3.34}$$

Given that $|\lambda_l^o - \hat{\lambda}_{i_l}| > cn^{\delta-1}$ and Assumption 1* holds, similar to (3.21), we have that $MDL(\hat{m}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}}) - MDL(\hat{m}, \tilde{\boldsymbol{\lambda}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}})$ is equal to either one of

$$\begin{aligned} &i) \quad \textstyle\sum' \Big(l_{i_l}(\hat{\theta}_{i_l}; x_{k,l}|x_{j,l}, j < k) - l_{i_l+1}(\hat{\theta}_{i_l+1}; x_{k,l}|x_{j,l}, j < k) \Big) + O_p(1), \\ &ii) \quad \textstyle\sum'' \Big(l_{i_l+1}(\hat{\theta}_{i_l+1}; x_{k,l}|x_{j,l}, j < k) - l_{i_l}(\hat{\theta}_{i_l}; x_{k,l}|x_{j,l}, j < k) \Big) + O_p(1) , \end{aligned} \tag{3.35}$$

where $\sum' = \sum_{k=n_l-[n(\lambda_l^o - \hat{\lambda}_{i_l})]+1}^{n_l}$ and $\sum'' = \sum_{k=1}^{[n(\hat{\lambda}_{i_l} - \lambda_l^o)]}$. By Proposition 3 and the ergodic theorem, the first term of case (ii) in (3.35) is positive and of order no less than $O(n^\delta)$ almost surely. For case (i) in (3.35), the ergodic theorem cannot be applied directly as the summation does not begin from 1. To tackle this, we can apply the ergodic theorem for the sum in (i) with $\sum_{k=n_l-[n(\lambda_l^o - \hat{\lambda}_{i_j})]+1}^{n_l}$ replaced by $\sum_{k=1}^{[n(\lambda_l^o - \hat{\lambda}_{i_j})]}$ and conclude that the sum is of order $O_p(n^\delta)$. Then the stationarity of the l-th piece implies that the original sum is also positive and of order $O_p(n^\delta)$. Therefore the quantities in both cases of (3.35) are positive with probability going to 1, and (3.34) follows. □

Theorem 5. *Let $\mathbf{y} = \{y_t; t = 1, \ldots, n\}$ be observations from a piecewise stationary process specified by the vector $(m_o, \boldsymbol{\lambda^o}, \boldsymbol{\psi^o})$ and satisfying Assumptions 1(1), 1*, 2(2), 3 and 5. The estimator $(\hat{m}_n, \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \hat{\boldsymbol{\psi}}_{\boldsymbol{n}})$ is defined in (2.3). Then we have*

$$\hat{m}_n \xrightarrow{P} m_o, \quad \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} \xrightarrow{P} \boldsymbol{\lambda^o}, \quad \hat{\boldsymbol{\psi}}_{\boldsymbol{n}} \xrightarrow{P} \boldsymbol{\psi^o} .$$

Proof. Following similar lines as in Lemma 2, it can be shown that, under Assumption 1*, 2(2) and 3, the analog of Lemma 2 holds with convergence in probability instead of almost surely, namely that

$$\hat{\theta} - \theta = O_p\left(n^{-\frac{1}{2}}\right), \tag{3.36}$$

where $\{\hat{\theta}, \theta\}$ is either $\{\hat{\theta}_n^{(j)}(\lambda_d, \lambda_u), \theta_j^o\}$, $\{\hat{\beta}_n^{(j)}(\lambda_d, \lambda_u), 0\}$ or $\{\hat{\zeta}_n^{(j)}(\lambda_d, \lambda_u), \theta_j^o\}$, using the notations in Lemma 2. Following Corollary 1 and Theorem 4, it suffices to prove that for any integer $d = 1, \ldots, M - m_o$, any $\delta > 0$ and any sequence $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{n}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{m_o})$ such that $|\lambda_j^o - \tilde{\lambda}_j| = O(n^{\delta-1})$ for $j = 1, \ldots, m_o$,

$$\arg \min_{\substack{\boldsymbol{\psi}, \boldsymbol{\lambda} \in A_{\epsilon_\lambda}^{(m_o+d)} \\ \tilde{\boldsymbol{\lambda}}_{\boldsymbol{n}} \subset \boldsymbol{\lambda}}} \left[\frac{2}{n} MDL(m_o + d, \boldsymbol{\lambda}, \boldsymbol{\psi})\right] - \frac{2}{n} MDL(m_o, \tilde{\boldsymbol{\lambda}}_{\boldsymbol{n}}, \boldsymbol{\psi^o}) \tag{3.37}$$

is positive with probability approaching 1. Denote $\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_{m_o+d+1})$ to be the minimizer for the first term in (3.37). Note that $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{n}} \subset \hat{\boldsymbol{\lambda}}_{\boldsymbol{n}}$ by construction. Similar to the proof in Theorem 3 it suffices to consider only the case where each $\hat{\xi}_l$, $l = 1, \ldots, m_o + d$ is associated with a unique parameter vector θ_l^* that correctly identifies the model. Otherwise the Taylor's expansion can be applied with respect to the first two components of $\hat{\theta}_n^{(j)}$. Using Assumption 5 and the Taylor's series expansions on the likelihood function, the quantity in (3.37) can be expressed as

$$\begin{aligned} & C_1 - C_2 \\ + \quad & \frac{1}{n}\left(\sum_{j=1}^{m_o+1} \tilde{L}_n^{(j)}((\xi_j^o, \theta_j^o), \tilde{\lambda}_{j-1}, \tilde{\lambda}_j; \mathbf{y}) - \sum_{l=1}^{m_o+d+1} \tilde{L}_n^{(l)}((\hat{\xi}_l, \theta_l^*), \hat{\lambda}_{l-1}, \hat{\lambda}_l; \mathbf{y})\right) \quad (3.38) \\ - \quad & \sum_{l=1}^{m_o+d+1} (\hat{\theta}_n^{(l)} - \theta_l^*)^T \frac{1}{n} \tilde{L}_n^{\prime\prime(l)}((\hat{\xi}_l; \theta_l^+); \hat{\lambda}_{l-1}, \hat{\lambda}_l, \mathbf{y})(\hat{\theta}_n^{(l)} - \theta_l^*) \quad (3.39) \end{aligned}$$

where $C_1 - C_2$ is positive and of order $O(\log n/n)$, and $|\theta_l^+ - \theta_l^*| < |\hat{\theta}_n^{(l)} - \theta_l^*|$. Using similar arguments as in showing (3.32) in Theorem 3, it follows that the quantity in (3.38) is exactly 0. Also, from (3.36), the summation in (3.39) is of order $O_p(n^{-1})$. Therefore $C_1 - C_2$, which is of order $O(\log n/n)$, dominates the expression and the quantity in (3.37) is indeed positive with probability approaching 1. This completes the proof of Theorem 5. □

4. Applications

4.1. Independent and identically distributed (iid) random variables

If the time series $\{Y_t\}$ satisfies (2.1) and each of the stationary pieces $\{\mathbf{X_j}\}$, $j = 1, \ldots, m + 1$ is a sequence of iid random variables, the break-point corresponds to a change in marginal distribution of the process. In this case, Assumption 1(k) is trivial for all k since the conditional likelihood reduces to the marginal likelihood. Assumption 4* is also satisfied as iid process must be mixing. Therefore, given Assumption 5, Theorems 3 and 5 hold if Assumption 2(2) and Assumption 2(4) are satisfied respectively. [14] studied weak consistency of multiple change-point detection for iid observations. They assumed that the number of change-points is known but there can be a common parameter throughout the whole series. Therefore, the results of [14] have been extended to strong consistency with unknown number of change-points in the special case that there is no common parameter throughout the series.

4.2. ARMA processes

Suppose, for $j = 1, \dots, m+1$, the j-th piece $\{X_{t,j}\}$ is an ARMA(p,q) process defined by

$$\Phi(B)(X_{t,j} - \mu_j) = \Theta(B) Z_{t,j}\,, \tag{4.1}$$

where $Z_{t,j} \sim \mathrm{IID}(0,\sigma_j^2)$, $\Phi(B) = 1 - \phi_{1,j}B - \ldots - \phi_{p,j}B^p$ and $\Theta(B) = (1 + \vartheta_{1,j}B + \ldots + \vartheta_{q,j}B^q)$ are polynomials with roots bounded outside the unit circle with a small distance δ_a. This ensures the compactness of the parameter space for the AR and MA parameters. For the compactness of the whole parameter space, we assume that there exists a large constant $K > 0$ such that $-K < \mu_j < K$ and $1/K < \sigma_j^2 < K$ for all j. Assume that the distribution of the white noise $\{Z_{t,j}\}$ is absolutely continuous with respect to the Lebesgue measure.

It is well known that there exists a causal representation

$$Z_{t,j} = \sum_{k=0}^{\infty} \psi_{k,j}(X_{t-k,j} - \mu_j)\,,$$

where the $\psi_{k,j}$'s are uniquely determined by $\phi_{1,j}, \dots, \phi_{p,j}, \vartheta_{1,j}, \dots, \vartheta_{q,j}$. Let $\psi_j = (\xi_j, \theta_j)$ with $\xi_j = (p_j, q_j)$ and $\theta_j = (\mu_j, \sigma_j^2, \phi_{1,j}, \dots, \phi_{p_j,j}, \vartheta_{1,j}, \dots, \vartheta_{q_j,j})$.

The exact and the observed Gaussian log-likelihood are given by

$$\begin{aligned} L_n^{(j)}((\xi_j,\theta_j),\mathbf{x_j}) &= -\frac{1}{2}\sum_{i=1}^{n_j}\left(\log\sigma_j^2 + \frac{1}{\sigma_j^2}\left(\sum_{k=0}^{\infty}\psi_{k,j}\tilde{x}_{i-k,j}\right)^2\right)\,, \\ \tilde{L}_n^{(j)}((\xi_j,\theta_j),\mathbf{x_j}) &= -\frac{1}{2}\sum_{i=1}^{n_j}\left(\log\sigma_j^2 + \frac{1}{\sigma_j^2}\left(\sum_{k=0}^{i}\psi_{k,j}\tilde{x}_{i-k,j} + \sum_{k=0}^{\infty}\psi_{i+k,j}\tilde{y}_{N_j-k}\right)^2\right)\,, \end{aligned}$$

respectively, where $\tau_{j-1} = n_1 + \ldots + n_{j-1}$, $y_{-i} := \mu_j$ for $i \geq 0$, $\tilde{x}_{i-k,j} = x_{i-k,j} - \mu_j$, and $\tilde{y}_{\tau_{j-1}-k} = y_{N_j-k} - \mu_j$.

Proposition 4. *Suppose that each segment of the piecewise stationary time series process follows an ARMA(p,q) model defined in (4.1).*

i) If $E(Z_{t,j}^{4+\delta}) < \infty$ for some $\delta > 0$, then we have weak consistency of MDL model selection, i.e., Theorem 5 holds.

ii) If $E(Z_{t,j}^{8+\delta}) < \infty$ for some $\delta > 0$, then we have strong consistency of MDL model selection, i.e., Theorem 3 holds.

Proof. i) For Theorem 5 to hold we need to verify Assumption 1(1), 1*, 2(2), 3 and 5 for ARMA models. For Assumption 1, by noting that there exists a $C > 0$ and a $\rho \in (0,1)$ such that $|\psi_{k,j}| < C\rho^k$ (e.g., [6]), we have

$$\begin{aligned} & -2\sigma_j^2 \left| L_n^{(j)}((\xi_j,\theta_j),\lambda_d,\lambda_u,\mathbf{x_j}) - \tilde{L}_n^{(j)}((\xi_j,\theta_j),\lambda_d,\lambda_u,\mathbf{x_j})\right| \\ \leq\ & \sum_{i=1}^{n_j}\left|\left(\sum_{l=0}^{\infty}\psi_{i+l,j}(y_{\tau_{j-1}-l}-x_{-l,j})\right)\zeta_{i,\tau_{j-1}}\right| \\ \leq\ & C\left|\sum_{i=1}^{n_j}\left(\rho^i\sum_{l=0}^{\infty}\rho^l|y_{\tau_{j-1}-l}-x_{-l,j}|\right)\zeta_{i,\tau_{j-1}}\right| \\ \leq\ & C\eta_{N_j}\sum_{i=1}^{n_j}\rho^i\zeta_{i,\tau_{j-1}}\,, \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} \eta_{\tau_{j-1}} &= \sum_{l=0}^{\infty}\rho^l|y_{\tau_{j-1}-l}-x_{-l,j}|\,, \\ \zeta_{i,\tau_{j-1}} &= \left|2\sum_{k=0}^{i}\psi_{k,j}\tilde{x}_{i-k,j}+\sum_{k=0}^{\infty}\psi_{i+k,j}(\tilde{y}_{\tau_{j-1}-k}+\tilde{x}_{-k,j})\right|\,. \end{aligned}$$

Note that Assumption 1* holds as $C\eta_{\tau_{j-1}}\sum_{i=1}^{n_j}\rho^i\zeta_{i,\tau_{j-1}} = O_p(1)$.

Also, when the $2p$-th moment of $Z_{t,j}$ exists, the $2p$-th moment of $\eta_{\tau_{j-1}}$ and $\zeta_{i,\tau_{j-1}}$ exist. It follows from Borel-Cantelli lemma that $\eta_{\tau_{j-1}} = \eta_{[\lambda_j n]-1} = O(n^{1/2p})$, since for any $K > 0$,

$$\sum_{n\geq 1}P(\eta_n > Kn^{1/2p}) = \sum_{n\geq 1}P(\eta_n^{2p} > K^{2p}n) \leq \frac{1}{K^{2p}}\mathrm{E}(\eta_1^{2p}) < \infty\,.$$

Similarly, we have $\zeta_{i,\tau_{j-1}} = O(n^{1/2p})$ and thus $C\eta_{\tau_{j-1}}\sum_{i=1}^{n_j}\rho^i\zeta_{i,\tau_{j-1}} = O(n^{1/p})$ almost surely. As $\mathrm{E}(Z_{t,j}^{4+\delta}) < \infty$, we can take $p = 2+\delta/2$ and it follows that Assumption 1(2) holds. Assumption 2(2) follows from the assumed compactness of the parameter spaces and the moment assumption on $Z_{t,j}$. Assumption 3 can be verified by the ergodic theorem and the compactness of the parameter space. Assumption 5 is verified in Example 2 of Section 3. Thus all the conditions for Theorem 5 are fulfilled.

ii) For Theorem 3 to hold we verify Assumptions 1(2), 2(4), 3, 4(0.5) and 5 for ARMA models. Similar to the argument in the proof of i), if the $(8+\delta)$-th moment of $Z_{t,j}$ exists, then Assumptions 1(2),2(4), 3 and 5 hold. Therefore, it remains to verify Assumption 4(0.5). Here $-2l_j(\psi_j;x_{i,j}|x_{l,j},l<i) = \log\sigma_j^2 + \left(\sum_{k=0}^{\infty}\psi_{k,j}(x_{i-k,j}-\mu_j)\right)^2/\sigma_j^2$. It is well known that (e.g., p.99 of [12]) causal ARMA processes are strongly mixing with geometric rate if the distribution of the white noise $\{Z_t\}$ is absolutely continuous with respect to Lebesgue measure. For an AR(p) process, $\psi_{k,j} = 0$ for $k > p$. Thus $l_j(\psi_j;x_{i,j}|x_{l,j},l<i)$ is also

strongly mixing with the same geometric rate as it is a function of finite number of the strongly mixing $x'_{i,j}$s (e.g., Theorem 14.1 of [8]). Thus Assumption 4* holds. For a general ARMA(p,q) process, we will verify Assumption 4(0.5) by approximating the likelihood of an ARMA process by that of a high order AR process. We give the details of the verification for (2.7), while (2.8) follows similarly. Replacing n_j by n and $x_{i,j}-\mu_j$ by x_i, it suffices to show that

$$A_n := \frac{1}{g(n)} \sum_{i=n-g(n)+1}^{n} \left(\sum_{l=0}^{\infty} \psi_{l,j} x_{i-l} \right)^2 \overset{a.s.}{\to} E\left(\left(\sum_{l=0}^{\infty} \psi_{l,j} x_{i-l} \right)^2 \right) =: \mu_A \,.$$

Since $H_{i,m} := \sum_{l=0}^{m} \psi_{l,j} x_{i-l,j}$, $i \in \mathbb{Z}$, is a strongly mixing process with geometric rate for any fixed m, Lemma 1 implies

$$B_{n,m} := \frac{1}{g(n)} \sum_{i=n-g(n)+1}^{n} \left(\sum_{l=0}^{m} \psi_{l,j} x_{i-l} \right)^2 \overset{a.s.}{\to} E\left(\left(\sum_{l=0}^{m} \psi_{l,j} x_{i-l} \right)^2 \right) =: \mu_{B_m} \,.$$

By taking m large enough, μ_{B_m} can be arbitrary close to μ_A. Therefore Assumption 4(0.5) holds if $g(n) > cn^{1/2}$ for some $c > 0$ and

$$\lim_{m\to\infty} \lim_{n\to\infty} |A_n - B_{n,m}| = 0, \quad \text{a.s.} \,. \tag{4.3}$$

To check (4.3), note that by Cauchy-Schwartz inequality and Lemma 1, we have

$$\begin{aligned}
& |A_n - B_{n,m}| \\
= & \left| \frac{1}{g(n)} \sum_{i=n-g(n)+1}^{n} \left(2H_{i,m} \sum_{l=m+1}^{\infty} \psi_{l,j} x_{i-l} + \left(\sum_{l=m+1}^{\infty} \psi_{l,j} x_{i-l} \right)^2 \right) \right| \\
\leq & \frac{1}{g(n)} \sum_{i=n-g(n)+1}^{n} \left(2|H_{i,m}| \sum_{l=m+1}^{\infty} |\psi_{l,j}||x_{i-l}| + \sum_{l=m+1}^{\infty} |\psi_{l,j}| \sum_{l=m+1}^{\infty} |\psi_{l,j}| x_{i-l}^2 \right) \\
\leq & 4 \sum_{l=m+1}^{\infty} |\psi_{l,j}| \frac{1}{g(n)} \sum_{i=n-g(n)+1}^{n} (H_{i,m}^2 + x_{i-l}^2) + \\
& \left(\sum_{l=m+1}^{\infty} |\psi_{l,j}| \right) \sum_{l=m+1}^{\infty} |\psi_{l,j}| \left(\frac{1}{g(n)} \sum_{i=n-g(n)+1}^{n} x_{i-l}^2 \right) \\
\leq & 4 \sum_{l=m+1}^{\infty} |\psi_{l,j}| E(H_{i,m}^2) + \left(4 + \sum_{l=m+1}^{\infty} |\psi_{l,j}| \right) \times \qquad (4.4) \\
& \left(E(x_1^2) \sum_{l=1}^{g(n)} |\psi_{m+l,j}| + \frac{1}{g(n)} \sum_{k=0}^{\infty} \left(\sum_{l=1}^{g(n)} |\psi_{m+l+k,j}| \right) x_{n-g(n)-m-k}^2 \right) + o(1) \,,
\end{aligned}$$

almost surely. To ensure that Lemma 1 is applicable in the above calculation with $g(n) > cn^{1/2}$ for some $c > 0$ when n is sufficiently large, the $(4+\delta)$-th moment of $B_{n,m}$, or the $(8+\delta)$-th moment for the ARMA process is required. As $|\psi_{k,j}| < C\rho^k$, the quantity in (4.4) can be arbitrary small for large m, thus (4.3) follows. □

4.3. GARCH processes

A GARCH(p,q) process is defined by the equations

$$x_t = \sigma_t \epsilon_t \,, \quad \sigma_t^2 \quad = \quad \omega + \sum_{i=1}^{p} \alpha_i x_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \,, \tag{4.5}$$

where

$$\omega > 0, \quad \alpha_i \geq 0, \quad 1 \leq i \leq p, \quad \beta_j \geq 0, \quad 1 \leq j \leq q$$

are constants, and $\{\epsilon_i, \ -\infty \leq i \leq \infty\}$ is a sequence of iid random variables with mean 0 and variance 1. To ensure certain mixing conditions we assume that ϵ_t is absolutely continuous with Lebesgue density being strictly positive in a neighborhood of zero, and $E|\epsilon_t|^s < \infty$ for some $s > 0$ (see [16] and [5]). In this section we will closely follow [4],which gives a comprehensive treatment in the estimation theory of GARCH model. Here the model $\mathcal{M}$ can be index by a two-dimensional parameter $\xi = (p,q)$ for integers $p \geq 0$ and $q \geq 0$. The parameter vector is $\theta = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)$. Following [4], given $\xi = (p,q)$, we restrict the parameter space to the compact set

$$\begin{aligned} \Theta(\xi) \quad = \quad & \{\theta : \sum_{i=1}^{q} \beta_i \leq \rho_o \ \text{ and } \ \underline{u} \leq \min(\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q) \\ & \leq \max(\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q) \leq \overline{u}\} \,, \end{aligned}$$

for some $0 < \underline{u} < \overline{u}$, $0 < \rho_o < 1$ and $q\underline{u} < \rho_o$. It is shown in Lemma 3.1. of [4] that σ_t^2 can be expressed as

$$\sigma_t^2 = c_0(\theta) + \sum_{i=1}^{\infty} c_i(\theta) x_{t-i}^2$$

with some deterministic coefficients $c_i(\theta), \ i = 1, 2, \ldots$, with $\sup_\theta |c_i(\theta)| < C_2 \rho^i$. for some $C_2 > 0$ and $\rho \in (0,1)$. Using the notation of Section 2, the true and the observed log-likelihood function for GARCH model are given respectively by

$$\begin{aligned} L_n^{(j)}(\theta_j) = \sum_{i=1}^{n_j} l(\theta_j; x_{i,j} | x_{l,j}, l < i) = -\frac{1}{2} \sum_{i=1}^{n_j} \left(\log \sigma_{i,j}^2 + \frac{x_{i,j}^2}{\sigma_{i,j}^2} \right) \,, \\ \tilde{L}_n^{(j)}(\theta_j) = \sum_{i=1}^{n_j} l(\theta_j; x_{i,j} | \mathbf{y_{i,j}}) = -\frac{1}{2} \sum_{i=1}^{n_j} \left(\log \tilde{\sigma}_{i,j}^2 + \frac{x_{i,j}^2}{\tilde{\sigma}_{i,j}^2} \right) \,. \end{aligned}$$

where

$$l_j(\theta; x_{i,j} | x_{l,j}, l < i) \quad = \quad -\frac{1}{2} \left(\log \sigma_{i,j}^2 + \frac{x_{i,j}^2}{\sigma_{i,j}^2} \right) \,,$$

$$l_j(\theta; x_{i,j}|\mathbf{y_{i,j}}) = -\frac{1}{2}\left(\log\tilde{\sigma}^2_{i,j} + \frac{x^2_{i,j}}{\tilde{\sigma}^2_{i,j}}\right),$$
$$\tilde{\sigma}^2_{i,j} = c_0(\theta) + \sum_{k=1}^{\tau_{j-1}+i-1} c_i(\theta)y^2_{N_j+i-k},$$

where $\tau_{j-1} = n_1 + \ldots + n_{j-1}$ for $j > 1$.

Proposition 5. *Suppose that each segment of the piecewise stationary time series process follows a GARCH(p,q) model defined in (4.5).*

i) If $E(X^{4+\delta}_{t,j}) < \infty$ for some $\delta > 0$, then we have weak consistency of MDL model selection, i.e., Theorem 5 holds.

ii) If $E(X^{8+\delta}_{t,j}) < \infty$ for some $\delta > 0$, then we have strong consistency of MDL model selection, i.e., Theorem 3 holds.

Proof. i) For Theorem 5. to hold we need to verify Assumptions 1(1),1*, 2(2), 3 and 5 for GARCH models. By a similar argument as in the proof of Lemma 5.8 and 5.9 in [4],

$$\begin{aligned}&\sup_{\lambda_d,\lambda_u}\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}L^{(j)}_n(\theta_j,\lambda_d,\lambda_u,\mathbf{x_j}) - \frac{1}{n}\tilde{L}^{(j)}_n(\theta_j,\lambda_d,\lambda_u,\mathbf{x_j})\right|\\ \leq\ & \frac{K_1}{n}(\sum_{k=1}^{n_j}\rho^k)\sum_{l=0}^{\infty}\rho^l(y^2_{\tau_{j-1}-l} + x^2_{-l,j})\\ &+\frac{K_2}{n}(\sum_{k=1}^{n_j}\rho^k(y^2_{\tau_{j-1}-k} + x^2_{-k,j}))\sum_{l=\tau_{j-1}+1}^{\tau_j}\rho^l\sup_{\theta_j\in\Theta_j}\left|\frac{y^2_l}{\sigma^2_{l,j}}\right|,\end{aligned}$$

for some $K_1, K_2 > 0$ and $\rho \in (0,1)$. It follows that the above term is of order $O_p(n^{-1})$ if $E(\log^+ y_1) < \infty$, and of order $O(n^{1/p-1})$ if $E(y_1^{4p}) < \infty$. Thus the first equation in Assumption 1 and 1* are verified. The second and the third equations follow similarly. Thus we conclude that Assumption 1(1) and 1(2) are satisfied when the $(4+\delta)$-th and the $(8+\delta)$-th moments of the GARCH process exist respectively. Similarly, it can be checked that Assumption 1* holds when the $(2+\delta)$-th moment exists.

Next, by (5.15) of [4] and Holder's inequality, it can be checked that Assumption 2(2) and 2(4) hold if the $(4+\delta)$-th and the $(8+\delta)$-th moments of the GARCH process exist respectively. Moreover, Assumptions 3 and 5A) are exactly the content of Lemma 5.4 to 5.6 of [4]. Assumption 5B) can be checked following similar lines as that of ARMA models. Thus all the conditions for Theorem 5 are fulfilled.

ii) We verify Assumptions 1(2), 2(4), 3, 4* and 5 for Theorem 3 to hold. It has been shown in the proof of i) that if the $(8+\delta)$-th moment of the GARCH process exists, then Assumptions 1(2),2(4),3 and 5 hold. Finally, Assumption 4* holds since $l_j(\theta_j; x_{i,j}|x_{l,j}, l < i)$ is a function of $x_{i,j}$ and $\sigma^2_{i,j}$ only, and the sequences $(x^2_{t,j})_{t\in\mathbb{Z}}$ and $(\sigma^2_{t,j})_{t\in\mathbb{Z}}$ are strongly mixing with geometric rate ([5], Theorem 3.4.2). The mixing property of $l'_j(\theta_j; x_{i,j}|x_{l,j}, l < i)$ can be shown similarly as the arguments in [5] can be extended to show the strongly mixing of $({\sigma^2}'_{t,j})_{t\in\mathbb{Z}}$. This completes the proof of Proposition 5. □

Appendix A: Proofs of propositions and lemmas

Proof of Proposition 1. We only prove (3.1), since (3.2) and (3.3) follow similarly. From Assumption 1(1), it suffices to show (3.1) for $L_n^{(j)}$ instead of $\tilde{L}_n^{(j)}$. Let $Q_{[0,1]}$ be the set of rational numbers in [0,1]. For any pair $r_1, r_2 \in Q_{[0,1]}$ with $r_1 < r_2$, we have by Assumption 3 that

$$\begin{aligned}
&\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}L_n^{(j)}(\theta_j, r_1, r_2; \mathbf{x}) - (r_2-r_1)L_j(\theta_j)\right| \\
=\ &\sup_{\theta_j\in\Theta_j}\left| r_2\left(\frac{1}{nr_2}\sum_{i=1}^{[nr_2]} l_j(\theta_j; x_i|x_l, l<i) - L_j(\theta_j)\right)\right. \\
&\left. - r_1\left(\frac{1}{nr_1}\sum_{i=1}^{[nr_1]} l_j(\theta_j; x_i|x_l, l<i) - L_j(\theta_j)\right)\right| \\
\xrightarrow{a.s.}\ & 0\,. \qquad\qquad (A.1)
\end{aligned}$$

Let B_{r_1,r_2} be the probability one set of ω's for which (A.1) holds. Set

$$B = \bigcap_{r_1,r_2\in Q_{[0,1]}} B_{r_1,r_2}\,,$$

and note that $P(B) = 1$. Moreover for any $\omega \in B$ and any $\lambda \in [0,1]$, choose $r_d, r_u \in Q_{[0,1]}$ such that $r_d \le \lambda \le r_u$. Hence

$$\begin{aligned}
&\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}\sum_{i=1}^{[n\lambda]} l_j(\theta_j; x_i|x_l, l<i) - \frac{1}{n}\sum_{i=1}^{[nr_d]} l_j(\theta_j; x_i|x_l, l<i)\right| \\
\le\ &\sup_{\theta_j\in\Theta_j}\frac{1}{n}\sum_{i=[nr_d]+1}^{[nr_u]} |l_j(\theta_j; x_i|x_l, l<i)| \\
\longrightarrow\ &(r_u - r_d)\sup_{\theta_j\in\Theta_j} E|l_j(\theta_j; x_i|x_l, l<i)|\,.
\end{aligned}$$

From Assumption 2(1), we have $\sup_{\theta_j\in\Theta_j} E|l_j(\theta_j; x_i|x_l, l<i)| < \infty$. So by making $(r_u - r_d)$ arbitrarily small, $L_n^{(j)}(\theta_j, 0, \lambda; \mathbf{x})/n \overset{a.s.}{\to} \lambda L_j(\theta_j)$ uniformly in $\theta_j \in \Theta_j$. By the same argument we have

$$\sup_{\theta_j\in\Theta_j}\left|\frac{1}{n}L_n^{(j)}(\theta_j, \lambda_d, \lambda_u; \mathbf{x}) - (\lambda_u - \lambda_d)L(\theta_j)\right| \xrightarrow{a.s.} 0\,. \qquad (A.2)$$

for any λ_d and λ_u in $[0,1]$ with $\lambda_d < \lambda_u$. Now we show that the convergence in (A.2) is uniform in λ_d, λ_u with $\lambda_u - \lambda_d > \epsilon_\lambda$. For any fixed positive $\epsilon < \epsilon_\lambda$, choose a large m_1 with $r_0, r_1, \ldots, r_{m_1} \in Q_{[0,1]}$ such that $0 = r_0 < r_1 < \ldots < r_{m_1} = 1$, and $\max_{i=1,\ldots,m}(r_i - r_{i-1}) \leq \epsilon$. Then for any $\lambda_d, \lambda_u \in [0,1]$, we can find j and k such that $j < k$, $r_{j-1} < \lambda_d < r_j$ and $r_{k-1} < \lambda_u < r_k$. Then we have

$$\begin{aligned}
& \left| \frac{1}{n} L_n^{(j)}(\theta_j, \lambda_d, \lambda_u; \mathbf{x_j}) - (\lambda_u - \lambda_d) L_j(\theta_j) \right| \\
\leq \quad & \left| \frac{1}{n} L_n^{(j)}(\theta_j, \lambda_d, \lambda_u; \mathbf{x_j}) - \frac{1}{n} L_n^{(j)}(\theta_j, r_{j-1}, r_k; \mathbf{x_j}) \right| + \\
& \left| \frac{1}{n} L_n^{(j)}(\theta_j, r_{j-1}, r_k; \mathbf{x_j}) - (r_k - r_{j-1}) L_j(\theta_j) \right| + \\
& \left| (r_k - r_{j-1}) L_j(\theta_j) - (\lambda_u - \lambda_d) L_j(\theta_j) \right| \, .
\end{aligned}$$

For large n the first and the third term are almost surely bounded by

$$\sup_{\theta_j \in \Theta_j} \left| (r_k - r_{k-1}) L_j(\theta_j) + (r_j - r_{j-1}) L_j(\theta_j) \right| < 2\epsilon \sup_{\theta_j \in \Theta_j} E|l_j(\theta_j; x_1 | x_l, l < 1)| \, .$$

By (A.1), the second term is bounded by ϵ for sufficiently large n. It follows that

$$\begin{aligned}
& \sup_{\lambda_d, \lambda_u} \sup_{\theta_j \in \Theta_j} \left| \frac{1}{n} L_n^{(j)}(\theta_j, \lambda_d, \lambda_u; \mathbf{x}_j) - (\lambda_u - \lambda_d) L_j(\theta_j) \right| \\
< \quad & 2\epsilon \sup_{\theta_j \in \Theta_j} E|l_j(\theta_j; x_1 | x_l, l < 1)| + \epsilon + 2\epsilon \sup_{\theta_j \in \Theta_j} E|l_j(\theta_j; x_1 | x_l, l < 1)| \, ,
\end{aligned}$$

for sufficiently large n, almost surely. The proposition follows as ϵ is arbitrary and independent of λ_d and λ_u. □

Proof of Proposition 2. By setting

$$\grave{\lambda}_d = \max(0, \lambda_d), \ \ddot{\lambda}_d = \min(0, \lambda_d), \ \grave{\lambda}_u = \min(1, \lambda_u), \ \ddot{\lambda}_u = \max(1, \lambda_u) \, , \quad \text{(A.3)}$$

we can consider the stationary and the leftover pieces from the adjacent segments separately. By elementary algebra,

$$\begin{aligned}
& \frac{1}{n_j} L_{n_j}^{(j)}(\theta_j, \lambda_d, \lambda_u, \mathbf{x_j}) - (\lambda_u - \lambda_d) L_j(\theta_j) \qquad \text{(A.4)} \\
= \quad & \frac{1}{n_j} L_{n_j}^{(j)}(\theta_j, \grave{\lambda}_d, \grave{\lambda}_u, \mathbf{x_j}) - (\grave{\lambda}_u - \grave{\lambda}_d) L_j(\theta_j) - L_j(\theta_j)(\ddot{\lambda}_u - 1 - \ddot{\lambda}_d) \\
& + \frac{1}{n_j} \sum_{i = n_{j-1} + [n_j(\ddot{\lambda}_d)] + 1}^{n_{j-1}} l_j(\theta_j; x_{i,j-1} | x_{l,j-1}, l < i) \\
& + \frac{1}{n_j} \sum_{i=1}^{[n_j(\ddot{\lambda}_u - 1)]} l_j(\theta_j; x_{i,j+1} | x_{l,j+1}, l < i) \, .
\end{aligned}$$

Since $0 \leq \grave{\lambda}_d < \grave{\lambda}_u \leq 1$, the sum of the first two terms converges to zero almost surely by Proposition 1. Moreover, for any $\delta > 0$, $\max(|\ddot{\lambda}_d|, |\ddot{\lambda}_u - 1|) < \delta$ for sufficiently large n, thus the third term is bounded by $2\delta|L_j(\theta_j)|$, and the fourth term is bounded by

$$\frac{1}{n_j} \sum_{i=n_{j-1}-[n_j\delta]}^{n_{j-1}} |l_j(\theta_j; x_{i,j-1}|x_{l,j-1}, l < i)| \xrightarrow{a.s.} \delta E|l_j(\theta_j; x_{1,j-1}|x_{l,j-1}, l < 1)|\,.$$

A similar bound can be established for the last term. Since δ is arbitrary, the term in (A.4) converges to zero uniformly in λ_d and λ_u in the sense of (3.4). □

Proof of Proposition 3. Note that by the definition of $\hat{\theta}_n$, we have $\tilde{L}_n^{(j)}((\xi_j, \hat{\theta}_n), \lambda_d, \lambda_u; \mathbf{x_j}) \geq \tilde{L}_n^{(j)}((\xi_j, \theta_j^*), \lambda_d, \lambda_u; \mathbf{x_j})$ for every λ_d, λ_u and n. Together with the uniform convergence of $\tilde{L}_n^{(j)}/n$ to $L_n^{(j)}/n$ and $L_n^{(j)}/n$ to L_j from Assumption 1(1) and Assumption 3 respectively, we have

$$\begin{aligned}
& (\lambda_u - \lambda_d)(L_j((\xi_j, \theta_j^*)) - L_j((\xi_j, \hat{\theta}_n)) \\
\leq\ & \sup_{\underline{\lambda_d}, \overline{\lambda_u}} \bigg((\lambda_u - \lambda_d) L_j((\xi_j, \theta_j^*)) - \frac{1}{n}\tilde{L}_n^{(j)}((\xi_j, \theta_j^*), \lambda_d, \lambda_u; \mathbf{x_j}) \\
& + \frac{1}{n}\tilde{L}_n^{(j)}((\xi_j, \hat{\theta}_n), \lambda_d, \lambda_u; \mathbf{x_j}) - (\lambda_u - \lambda_d) L_j((\xi_j, \hat{\theta}_n)) \bigg) \\
=\ & \sup_{\underline{\lambda_d}, \overline{\lambda_u}} \bigg((\lambda_u - \lambda_d) L_j((\xi_j, \theta_j^*)) - \frac{1}{n}L_n^{(j)}((\xi_j, \theta_j^*), \lambda_d, \lambda_u; \mathbf{x_j}) \\
& + \frac{1}{n}L_n^{(j)}((\xi_j, \hat{\theta}_n), \lambda_d, \lambda_u; \mathbf{x_j}) - (\lambda_u - \lambda_d) L_j((\xi_j, \hat{\theta}_n)) \bigg) + o(1) \\
\leq\ & 2 \sup_{\underline{\lambda_d}, \overline{\lambda_u}} \sup_{\theta_j \in \Theta_j} \left| \frac{1}{n}\tilde{L}_n^{(j)}((\xi_j, \theta_j), \lambda_d, \lambda_u; \mathbf{x_j}) - (\lambda_u - \lambda_d) L_j((\xi_j, \theta_j)) \right| + o(1) \\
\xrightarrow{a.s.}\ & 0\,.
\end{aligned}$$

In the foregoing relations, the first inequality is obtained by definition of the maximum likelihood estimator, the second equality follows from Assumption 1(1), and the last convergence follows from Proposition 2. Since $L_j((\xi_j, \theta_j^*))$ is the maximum value over $L_j((\xi_j, \cdot))$ and $\lambda_u - \lambda_d > 0$, it follows that

$$|L_j((\xi_j, \hat{\theta}_n)) - L_j((\xi_j, \theta_j^*))| \overset{a.s.}{\to} 0\,. \tag{A.5}$$

Combining (A.5) with Proposition 2 and Assumption 2(1), (3.9) follows. If $\xi_j = \xi_j^o$ and Assumption 5A) holds, then $L_j((\xi_j, \theta_j))$ has a unique maximum at θ_j^o. It follows from (A.5) that (3.10) holds. Similarly, if $\xi_j = \xi_b$ is a bigger model than ξ_j^o and Assumption 5B) holds, then for any $\hat{\pi}_n(\lambda_d, \lambda_u)$, $L_j((\xi_b, \cdot))$ has a unique maximum at $(0, \theta_j^o, \hat{\pi}_n(\lambda_d, \lambda_u))$ over the set $V_\delta(\hat{\pi}_n(\lambda_d, \lambda_u))$ for some $\delta > 0$, thus (3.11) and (3.12) follow from (A.5). □

Proof of Lemma 1. (i) is an immediate consequence of the ergodic theorem. For (ii), since $\{X_t\}$ is strongly mixing with geometric rate, from [18], we have

$$E\left|\sum_{t=1}^{n} X_t\right|^r \leq K n^{\frac{r}{2}}$$

for some constant K. Therefore, by the stationarity of $\{X_t\}$ and Markov's inequality, for any fixed $\delta > 0$,

$$\begin{aligned} P\left(\left|\sum_{t=n-g(n)+1}^{n} (X_t - \mu)\right| > \delta g(n)\right) &= P\left(\left|\sum_{t=1}^{g(n)} (X_t - \mu)\right| > \delta g(n)\right) \\ &\leq \frac{E\left|\sum_{t=1}^{g(n)} X_t\right|^{r+\epsilon}}{(\delta g(n))^{r+\epsilon}} \\ &= O(g(n)^{-(r+\epsilon)/2}) = O(n^{-1-\epsilon/r}) . \end{aligned}$$

Therefore

$$\sum_{n=1}^{\infty} P\left(\left|\frac{1}{g(n)} \sum_{t=n-g(n)+1}^{n} (X_t - \mu)\right| > \epsilon\right) \leq \infty ,$$

and (ii) follows from the Borel-Cantelli lemma. For (iii) and (iv), define $a(n) = \max(n^{2/r}, s(n))$. Note that

$$-\sum_{t=1}^{a(n)} |X_t| \leq \sum_{t=1}^{s(n)} X_t \leq \sum_{t=1}^{a(n)} |X_t| .$$

Since $|X_t| - E(|X_t|)$ is also a strongly mixing sequence with geometric rate, the terms that sandwich $\sum_{t=1}^{s(n)} X_t$ in (38) are of order $O(a(n))$ almost surely. So (iii) follows from (i) and the fact that $a(n) = O(n^{2/r})$. (iv) follows similarly and therefore Lemma 1 follows. □

Proof of Lemma 2. We only prove (3.25) and (3.26), while (3.24) can be shown similarly. For notational simplicity, denote $(\hat{\lambda}_{j-1}, \hat{\lambda}_j)$ by (λ_d, λ_u). Let $\hat{\gamma}_n^{(j)} \equiv (\hat{\beta}_n^{(j)}, \hat{\zeta}_n^{(j)}) \equiv (\hat{\beta}_n^{(j)}(\lambda_d, \lambda_u), \hat{\zeta}_n^{(j)}(\lambda_d, \lambda_u))$ be the first two components of $\hat{\theta}_n$. Define also $\theta_n^o \equiv (0, \theta_j^o, \hat{\pi}_n^{(j)})$ and $\gamma_n^o \equiv (0, \theta_j^o)$. For the model ξ_j with parameter vector θ partitioned as $\theta = (\beta, \zeta, \pi)$, let $\tilde{L}_n'^{(j)}(\theta, \lambda_d, \lambda_u, \mathbf{x_j})$ and $\tilde{L}_n''^{(j)}(\theta, \lambda_d, \lambda_u, \mathbf{x_j})$ be respectively the first and second partial derivatives of $\tilde{L}_n^{(j)}(\theta, \lambda_d, \lambda_u, \mathbf{x_j})$ with respect to (β, ζ). From (3.11) and (3.12), a Taylor series expansion on $\tilde{L}_n'^{(j)}(\hat{\theta}_n, \lambda_d, \lambda_u, \mathbf{x_j})$ around γ_n^o, with $\hat{\pi}_n^{(j)}$ keeping fixed, can be applied to give

$$\tilde{L}_n'^{(j)}(\hat{\theta}_n, \lambda_d, \lambda_u, \mathbf{x_j}) = \tilde{L}_n'^{(j)}(\theta_n^o, \lambda_d, \lambda_u, \mathbf{x_j}) + \tilde{L}_n''^{(j)}(\theta_n^+, \lambda_d, \lambda_u, \mathbf{x_j})(\hat{\gamma}_n - \gamma_n^o) , \quad (A.6)$$

where $\theta_n^+ = (\gamma_n^+, \hat{\pi}_n)$, $\gamma_n^+ \in \mathbb{R}^{d_\beta + d_s}$ and $|\gamma_n^+ - \gamma_n^o| \leq |\hat{\gamma}_n - \gamma_n^o|$. By the definition of $\hat{\theta}_n$, we have $L_n'^{(j)}(\hat{\theta}_n, \lambda_d, \lambda_u, \mathbf{x_j}) = 0$. Therefore (A.6) is equivalent to

$$\frac{1}{n}\tilde{L}_n''^{(j)}(\theta_n^+, \lambda_d, \lambda_u, \mathbf{x}_j)(\hat{\gamma}_n - \gamma_n^o) = -\frac{1}{n}\tilde{L}_n'^{(j)}(\theta_n^o, \lambda_d, \lambda_u, \mathbf{x}_j)\,. \tag{A.7}$$

Combining Assumption 1(p), Theorem 2 and Lemma 1(iii)(iv) with $w = \max(\frac{1}{p}, \frac{1}{q}) \leq \frac{1}{2}$, we have

$$\begin{aligned}
&\tilde{L}_n'^{(j)}(\theta_n^o, \lambda_d, \lambda_u, \mathbf{x_j}) = L_n'^{(j)}(\theta_n^o, \lambda_d, \lambda_u, \mathbf{x_j}) + O(n^{\frac{1}{p}}) \quad \text{[Assumption 1}(p)\text{]}\\
&= \sum_{i=[n\grave{\lambda}_d]+1}^{[n\grave{\lambda}_u]} l_j'((\xi_j, \theta_n^o); x_{i,j}|x_{l,j}, l < i) + O(n^\omega) \quad \text{[Theorem 2 and Lemma 1]}\\
&= \sum_{i=1}^{[n\grave{\lambda}_u]} l_j'((\xi_j, \theta_n^o); x_{i,j}|x_{l,j}, l < i) - \sum_{i=1}^{[n\grave{\lambda}_d]} l_j'((\xi_j, \theta_n^o); x_{i,j}|x_{l,j}, l < i) + O(n^\omega)\,,
\end{aligned}$$

where $\grave{\lambda}_d$ and $\grave{\lambda}_u$ are defined in (A.3). Since, for any fixed π and $\mathbf{x} = \{x_i; i \in \mathbb{Z}^+\}$,

$$f_{\xi_j}(\cdot|\mathbf{x}; (0, \theta_j^o, \pi)) = f_{\xi_j^o}(\cdot|\mathbf{x}; \theta_j^o)$$

almost everywhere, it follows that $E_{\psi_j^o}(l_j'((\xi_j, (0, \theta_j^o, \pi); x_{i,j}|x_{l,j}, l < i)) = 0$. Therefore, the sequence $(l_j'((\xi_j, (0, \theta_j^o, \pi)); x_{i,j}|x_{l,j}, l < i))_{i \in N}$ is a stationary ergodic zero-mean martingale difference sequence with finite second moment. The law of iterative logarithm for martingales from [17] implies now that both

$$\sum_{i=1}^{[n\lambda_u]} l_j'((\xi_j, (0, \theta_j^o, \pi)); x_{i,j}|x_{l,j}, l < i) \quad \text{and} \quad \sum_{i=1}^{[n\grave{\lambda}_d]} l_j'((\xi_j, (0, \theta_j^o, \pi)); x_{i,j}|x_{l,j}, l < i)$$

are of order $O(\sqrt{n \log\log n})$. Thus we have

$$\tilde{L}_n'^{(j)}((0, \theta_j^o, \pi), \lambda_d, \lambda_u, \mathbf{x}_j) = O(\sqrt{n \log\log n}) \quad \text{a.s.}\,. \tag{A.8}$$

By the compactness of $\Theta(\xi_j)$ and the continuity of l_j', it follows that (A.8) holds uniformly in π. Thus we have

$$\tilde{L}_n'^{(j)}(\theta_n^o, \lambda_d, \lambda_u, \mathbf{x}_j) = O(\sqrt{n \log\log n}) \quad \text{a.s.}\,. \tag{A.9}$$

From Assumption 5B) and Jensen's Inequality, the expectation of the likelihood, $L_j((\xi_j, (\gamma_b, \pi)))$, obtains its unique maximum at $\gamma_b = (0, \theta_j^o)$. Thus $L_j''((\xi_j, (0, \theta_j^o, \pi)))$ is positive definite. Also, the positive definiteness holds uniformly in π by the compactness of $\Theta(\xi_j)$ and the assumed continuity of L_j''. Together with Proposition 2 and the fact that $|\theta_n^+ - \theta_n^o| \overset{a.s.}{\to} 0$, we have almost surely,

$$\frac{1}{n}\tilde{L}_n''^{(j)}(\theta_n^+, \lambda_d, \lambda_u, \mathbf{x}_j) \quad \text{is positive definite}\,. \tag{A.10}$$

Combining (A.7), (A.9) and (A.10), the equations (3.25) and (3.26) follow. □

References

[1] Adak, S. (1998). Time-Dependent Spectral Analysis of Nonstationary Time Series. *Journal of the American Statistical Association* **93** 1488–1501. MR1666643

[2] Andrews, D., Cheng, X. (2010). Estimation and Inference with Weak, Semi-strong, and Strong Identification. *Cowles Foundation Paper No. 1773.*

[3] Aue, Alexander, H ormann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *Annals of Statistics* **37(6B)** 4046–4087. MR2572452

[4] Berkes, I., Horvath, L., Kokoszka, P. (2003). GARCH processes: structure and estimation. *Bernoulli.* **9(2)** 201–227. MR1997027

[5] Boussama, F. (1998). *Ergodicité, mélange et estimation dans le modelés GARCH*. Ph.D. Thesis, Université 7 Paris.

[6] Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Method*. Springer, New York. MR1093459

[7] Bardet, J.M., Wintenberger, O. (2009). Asymptotic normality of the quasi-maximum likelihood estimator for multidensional causal processes. *Annals of Statistics* **37(5B)** 2730–2759. MR2541445

[8] Davidson, J. (1994). *Stochastic Limit Theory.* Oxford: Oxford University Press. MR1430804

[9] Davis, R.A., Hancock, S. and Yao, Y.C. (2010). Consistency of Minimum description length model selection for piecewise autoregressions. *Preprint.*

[10] Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2006). Structural Break Estimation for nonstationary time series models. *Journal of American Statistical Association.* **101** 223–239. MR2268041

[11] Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2008). Break Detection for a Class of Nonlinear Time Series Models. *Journal of Time Series Analysis.* **29** 834–867. MR2450899

[12] Doukhan, P. (1994). *Mixing properties and examples, Lecture Notes in Statistics*, **85**, Springer-Verlag. MR1312160

[13] Hannan, E. J. (1980). The Estimation of the Order of an ARMA Process. *Annals of Statistics.* **8** 1071–1081. MR0585705

[14] He, H. and Severini, T.A.(2010). Asymptotic Properties of Maximum Likelihood Estimators in Models with Multiple Change Points. *Bernoulli.* **16(3)**, 759–779. MR2730647

[15] Jeantheau, T.(1998). Strong consistency of estimators for multivariate arch models. *Econometric Theory.* **14**, 70–86. MR1613694

[16] Lindner, A.M. (2009). Stationarity, Mixing, Distributional Properties and Moments of GARCH(p,q) Processes. In *Handbook of Financial Time Series.* Editors: Mikosch, T., Kreis, J., Davis, R.A., Andersen, T.G., Springer-Verlag.

[17] Stout, W.F.(1974). *Almost sure convergence*. New York: Academic Press. MR0455094

[18] Yokoyama, R. (1980). Moment bounds for stationary mixing sequences. *Z.Wahrsheinlichkeitstheor. Verw. Geb.* **52**, 45–57. MR0568258