Submitted to the Annals of Statistics

CONSISTENCY OF MINIMUM DESCRIPTION LENGTH MODEL SELECTION FOR PIECEWISE AUTOREGRESSIONS

By Richard A. Davis[§], Stacey Hancock[*,†,‡,¶] and Yi-Ching Yao[||]

Columbia University[§], Reed College[¶], and Academia Sinica and National Chengchi University[||]

The program Auto-PARM (Automatic Piecewise AutoRegressive Modeling), developed by Davis, Lee, and Rodriguez-Yam (2006), uses the minimum description length (MDL) principle to estimate the number and locations of change-points in a time series by fitting autoregressive models to each segment. When the number of change-points is known, Davis et al. (2006) show that the (relative) change-point location estimates are consistent when the true underlying model is segmented autoregressive. In this paper, we show that the estimates of the number of change-points and the autoregressive orders obtained by minimizing the MDL are consistent for the true values when using conditional maximum (Gaussian) likelihood variance estimates. However, if Yule-Walker variance estimates are used, the estimate of the number of change-points is not necessarily consistent. This surprising result is due to an exact cancellation of first-order terms in a Taylor series expansion in the conditional maximum likelihood case, which does not occur in the Yule-Walker case.

1. Introduction.

In recent years, there has been considerable development in non-linear time series modeling. One prominent subject in non-linear time series modeling is the "change-point" or "structural breaks" model. In this paper, we discuss *a posteriori* estimation of change-points with a fixed sample size using minimum description length as a model fitting criterion with minimal assumptions.

The majority of the early literature on change-point estimation assumes independent normal data. In their seminal paper, Chernoff and Zacks (1964) examine the problem of detecting mean changes in independent normal data

[*] Research supported by the Program for Interdisciplinary Mathematics, Ecology and Statistics (PRIMES) NSF IGERT Grant DGE 0221595, administered by Colorado State University.

[†] Research partially supported by the NSF East Asia and Pacific Summer Institute.

[‡] Research partially supported by the National Park Service.

AMS 2000 subject classifications: Primary 62F12; secondary 62M10, 62B10

Keywords and phrases: Change-point, minimum description length, autoregressive process

with unit variance. Both Yao (1988) and Sullivan (2002) estimate the number and locations of changes in the mean of independent normal data with constant variance, and Chen and Gupta (1997) examine changes in the variance of independent normal data with a constant mean. Some research considers the change-point problem without assuming normality, but still assumes independence (see, for example, Lee (1996, 1997), Hawkins (2001), and Yao and Au (1989)). Bayesian approaches have also been explored, e.g., Fearnhead (2006), Perreault et al. (2000), Stephens (1994), Yao (1984), and Zhang and Siegmund (2007).

Recent literature focuses more on detecting changes in dependent data, though the majority of this literature concerns hypothesis testing. Davis et al. (1995) derive the asymptotic distribution of the likelihood ratio test for a change in the parameter values and order of an autoregressive model, and Ling (2007) examines a general asymptotic theory on the Wald test for change-points in a general class of time series models under the no change-point hypothesis. Research on the estimation of the number and locations of the change-points includes Kühn (2001), who assumes a weak invariance principle, and Kokoszka and Leipus (2000) on the estimation of change-points in ARCH models. See Csorgo and Horvath (1997) for a comprehensive review.

This paper examines the program Auto-PARM, a method developed by Davis, Lee, and Rodriguez-Yam (2006) for estimating the number and locations of the change-points. This method does not assume independence nor a distribution on the data, e.g., normal, and does not assume a specific type of change. It models the data as a piecewise autoregressive (AR) process, and can detect changes in the mean, variance, spectrum, or other model parameters. The most important ingredient of Auto-PARM is its use of the minimum description length criterion in fitting the model.

The estimated (relative) change-point locations are shown to be consistent in Davis et al. (2006) when the number of change-points is known. However, the paper leaves open the issue of consistency when the number of change-points is unknown, which is the focus of the present study. When using conditional maximum (Gaussian) likelihood variance estimates, we show that the estimated number of change-points and the estimated AR orders are weakly consistent. However, the estimate for the number of change-points may not even be weakly consistent if we use Yule-Walker variance estimates. It is surprising to find that consistency breaks down in general for Auto-PARM when using Yule-Walker estimation, but can be saved if conditional maximum likelihood estimates are substituted for Yule-Walker.

Section 2 begins by reviewing the Auto-PARM procedure, then provides some preliminaries needed for the consistency proofs. Included in the preliminaries are subsections on the functional law of the iterated logarithm applied to sample autocovariance functions and on conditional maximum likelihood estimation for piecewise autoregressive processes. Section 3 contains the main results, starting by addressing the simple case of an AR process with no change-points then moving to the more general piecewise AR process. Lemma 3.1 states consistency under conditional maximum likelihood estimation when there are no change-points. Theorem 3.1 provides a case where the Auto-PARM estimate of the number of change-points using Yule-Walker estimation is not consistent. Theorem 3.2 returns to conditional maximum likelihood estimation and shows consistency under a piecewise autoregressive model. Some of the technical details are relegated to the appendix.

2. Background and Preliminaries.

Before embarking on our consistency results, we first discuss the modeling procedure Auto-PARM and some preliminary topics. Of central importance to the proofs is the functional law of the iterated logarithm for stationary time series and conditional maximum likelihood estimation.

2.1. *Automatic Piecewise Autoregressive Modeling (Auto-PARM).*

Davis et al. (2006) develop a procedure for modeling a non-stationary time series by segmenting the series into blocks of different autoregressive processes. The modeling procedure, referred to as Auto-PARM (*A*utomatic *P*iecewise *A*uto*R*egressive *M*odeling), uses a minimum description length (MDL) model selection criterion to estimate the number of change-points, the locations of the change-points, and the autoregressive model orders.

The class of piecewise autoregressive models that Auto-PARM fits to an observed time series with n observations is as follows. For $k = 1, \ldots, m$, denote the change-point between the kth and $(k+1)$st autoregressive processes as τ_k, where $\tau_0 := 0$ and $\tau_{m+1} := n$. Let $\{\epsilon_{k,t}\}$, $k = 1, \ldots, m+1$, be independent sequences of independent and identically distributed (iid) random variables with mean zero and unit variance. Then for given initial values X_{-P+1}, …, X_0 with P a preassigned upper bound on the AR order, AR coefficient parameters $\phi_{k,j}$, $k = 1, \ldots, m+1$, $j = 0, 1, \ldots, p_k$, and noise parameters σ_1, …, σ_{m+1}, the *piecewise autoregressive process* $\{X_t\}$ is defined as

$$X_t = \phi_{k,0} + \phi_{k,1} X_{t-1} + \cdots + \phi_{k,p_k} X_{t-p_k} + \sigma_k \epsilon_{k,t-\tau_{k-1}} \tag{2.1}$$

for $t \in (\tau_{k-1}, \tau_k]$, where $\boldsymbol{\psi}_k := (\phi_{k,0}, \phi_{k,1}, \ldots, \phi_{k,p_k}, \sigma_k)$ is the parameter vector corresponding to the causal AR(p_k) process in the kth segment. Notice that in each segment, the subscripting on the new noise sequence is restarted to time one. Denoting the mean of $\{X_t\}$ in the kth segment by μ_k, the intercept $\phi_{k,0}$ equals $\mu_k(1-\phi_{k,1}-\cdots-\phi_{k,p_k})$, and for $t \in (\tau_{k-1}, \tau_k]$, we can express the model as

$$X_t - \mu_k = \phi_{k,1}(X_{t-1}-\mu_k) + \cdots + \phi_{k,p_k}(X_{t-p_k}-\mu_k) + \sigma_k \epsilon_{k,t-\tau_{k-1}}.$$

To ensure identifiability of the change-point locations, the model assumes that $\boldsymbol{\psi}_j \neq \boldsymbol{\psi}_{j+1}$ for every $j = 1, \ldots, m$. That is, between consecutive segments, at least one of the AR coefficients, the process mean, the white noise variance, or the AR order must change.

Given an observed time series $X_1, \ldots, X_n$, Auto-PARM obtains the best-fitting model by finding the best combination of the number of change-points m, the change-point locations $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$, and the AR orders $\boldsymbol{p} = (p_1, \ldots, p_{m+1})$ according to the MDL criterion. When estimating the change-points, it is necessary to require sufficient separation between the change-point locations in order to be able to estimate the AR parameters. We define “relative change-points” $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ such that $\lambda_k = \tau_k/n$ for $k = 0, \ldots, m+1$. Defined as such, we take throughout the convention that $\lambda_k n$ is an integer. Let $\delta > 0$ be a preassigned lower bound for the relative length of each of the fitted segments, and define

$$A_m^\delta = \{(\lambda_1, \ldots, \lambda_m) : 0 < \lambda_1 < \cdots < \lambda_m < 1, \\ \lambda_k - \lambda_{k-1} \geq \delta, k = 1, \ldots, m+1\}. \tag{2.2}$$

(Note that the total number of change-points is bounded by $M = M_\delta := [1/\delta]-1$ where $[x]$ denotes the integer part of x.) Estimates are then obtained by minimizing the MDL over $0 \leq m \leq M$, $0 \leq \boldsymbol{p} \leq P$, and $\boldsymbol{\lambda} \in A_m^\delta$, where P is a preassigned upper bound for p_k. Using results from information theory [20] and standard likelihood approximations, we define the minimum description length for a piecewise autoregressive model [8] as

$$\begin{aligned} &\mathrm{MDL}(m, \tau_1, \ldots, \tau_m; p_1, \ldots, p_{m+1}) \\ &\quad = \log^+ m + (m+1)\log n + \sum_{k=1}^{m+1} \log^+ p_k \\ &\qquad \sum_{k=1}^{m+1} \frac{p_k+2}{2}\log n_k + \sum_{k=1}^{m+1} \frac{n_k}{2}\log(2\pi\hat{\sigma}_k^2), \end{aligned} \tag{2.3}$$

where $\log^+ x = \max\{\log x, 0\}$, $n_k = \tau_k - \tau_{k-1}$ is the number of observations in the kth segment and $\hat{\sigma}_k^2$ is the white noise variance estimate in the kth segment. Then, the parameter estimates are denoted by

$$\hat{m}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{p}} = \underset{0\leq m\leq M,\, 0\leq \boldsymbol{p}\leq P,\, \boldsymbol{\lambda}\in A_m^{\delta}}{\arg\min} \left\{ \frac{2}{n}\mathrm{MDL}(m, \boldsymbol{\lambda}; \boldsymbol{p}) \right\}. \tag{2.4}$$

The only dependence on the AR parameter estimates in the MDL is through the white noise variance estimates, $\hat{\sigma}_k^2$, which only involve sample autocovariance functions (ACVFs). In this paper, we show that the estimates of the number of change-points and the autoregressive orders obtained by minimizing the MDL are consistent when this white noise variance estimate is obtained by conditional maximum likelihood methods. In practice, Auto-PARM uses Yule-Walker estimates for the AR white noise variance. However, consistency of the estimates for the number of change-points does not necessarily hold when using Yule-Walker variance estimates.

Minimizing the criterion function in (2.4) over non-negative valued integers can be a difficult computational problem. Davis et al. (2006) use a genetic algorithm to perform the optimization. Although there is no guarantee that the procedure will find the global minimizer, the minimum produced by Auto-PARM seems to perform quite well. Of course in our results we will assume that the global minimizer is computable.

2.2. *Functional law of the iterated logarithm.* Throughout the consistency proofs in this paper, we use the functional law of the iterated logarithm (FLIL) on the sample ACVF and sample means of autoregressive processes. Therefore, we will first describe how to apply the FLIL to AR processes and discuss sufficient conditions in order for the FLIL to hold.

Rio (1995) shows that the FLIL holds for stationary strong mixing sequences under the following condition. Suppose $\{X_t\}_{t\in\mathbb{Z}}$ is a strictly stationary and strong mixing sequence of real-valued mean zero random variables, with sequence of strong mixing coefficients $\{\alpha_n\}_{n>0}$. Define the strong mixing function $\alpha(\cdot)$ by $\alpha(u) = \alpha_{[u]}$, and denote the quantile function of $|X_1|$ by Q. Then the FLIL holds for the sequence $\{X_t\}$ if

$$\int_0^1 \alpha^{-1}(u) Q^2(u) du < \infty,$$

where f^{-1} denotes the inverse of the monotonic function f. This condition simplifies if the process is strong mixing at a geometric rate. In this case, the FLIL holds if

$$\mathbb{E}\left(X_1^2 \log^+ |X_1|\right) < \infty$$

(see [19] for proof). Therefore, assuming

$$\mathbb{E}\left(X_1^4(\log^+|X_1|)^2\right) < \infty \tag{2.5}$$

and strong mixing with a geometric rate function allows us to apply the FLIL of Rio to the sample ACVF calculated using the change-point locations that minimize the MDL. In other words, e.g., for an AR(p) process $\{X_t\}$ with mean μ and small $\delta > 0$, we have

$$\sup_{0\le\lambda<\lambda+\delta\le\lambda'\le 1} \frac{\tilde{\gamma}_{\lambda:\lambda'}(i,j) - \gamma(|i-j|)}{\sqrt{\frac{2}{n}\log\log n}} < C \quad \text{a.s.} \tag{2.6}$$

where $C < \infty$ is a constant,

$$\tilde{\gamma}_{\lambda:\lambda'}(i,j) := \frac{1}{(\lambda'-\lambda)n}\sum_{t=\lambda n+1}^{\lambda' n}(X_{t-i}-\mu)(X_{t-j}-\mu), \tag{2.7}$$

and $\gamma(h)$ is the true ACVF of the process. (Throughout λ and λ' are assumed to be such that λn and $\lambda' n$ are integers.) Note that (2.6) also holds if $\tilde{\gamma}_{\lambda:\lambda'}(i,j)$ is replaced by

$$\begin{aligned}\hat{\gamma}_{\lambda:\lambda'}(i,j) := \frac{1}{(\lambda'-\lambda)n}\sum_{t=\lambda n+1}^{\lambda' n}&(X_{t-i}-\overline{X}_{\lambda n+1-i:\lambda' n-i})\cdot \\ &(X_{t-j}-\overline{X}_{\lambda n+1-j:\lambda' n-j})\end{aligned} \tag{2.8}$$

where $\overline{X}_{a:b} := \sum_{t=a}^{b} X_t/(b-a+1)$. It follows that

$$\sup_{|i-j|\le P,\ 0\le\lambda<\lambda+\delta\le\lambda'\le 1} |\tilde{\gamma}_{\lambda:\lambda'}(i,j)-\gamma(|i-j|)| = O(\sqrt{\log\log n/n}), \tag{2.9}$$

$$\sup_{|i-j|\le P,\ 0\le\lambda<\lambda+\delta\le\lambda'\le 1} |\hat{\gamma}_{\lambda:\lambda'}(i,j)-\gamma(|i-j|)| = O(\sqrt{\log\log n/n}) \tag{2.10}$$

for some upper bound $P < \infty$.

When applying the FLIL to the sample ACVF of a piecewise autoregressive process, we will need to assume throughout this paper that the stationary process generated by the parameter values and the iid noise sequence for each of the segments

(A1) is causal and strongly mixing at a geometric rate, and
(A2) satisfies the moment condition (2.5).

As commented in Remark 2.1 of [7], there are many sufficient conditions on the distribution of the noise in order to ensure that $\{X_t\}$ is strongly mixing. One such condition is for $\{\epsilon_t\}$ to be iid with a common distribution function which has a nontrivial absolutely continuous component [see Athreya and Pantula (1986a, b)]. Under this condition, it can be shown [cf. Theorems 16.0.1 and 16.1.5 in Meyn and Tweedie (1993)] that the strong mixing function $\alpha(u)$ decays at a geometric rate.

2.3. *Conditional maximum likelihood.* In the proofs that follow, we use conditional maximum likelihood estimates (equivalent to conditional least squares) for the AR parameters. Here we will further discuss the assumptions (A1) and (A2) necessary for the FLIL to hold in this case. When fitting an AR(p) process to observations $X_1, \dots, X_n$, conditional maximum likelihood estimation uses a definition of the sample ACVF that includes initial values $X_{-p+1}, \dots, X_0$. (Note that as a piecewise autoregressive model of order up to P is fitted to the observations, we treat the first P observations as the initial values.) In other words, conditional maximum likelihood estimates use

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n}(X_t - \overline{X}_{1:n})(X_{t-h} - \overline{X}_{1-h:n-h})$$

for the sample ACVF. For a stationary process $\{X_t\}$ satisfying (A1) and (A2), the FLIL holds for $\hat{\gamma}(h)$. While we may assume that the X_t in the first segment of a piecewise AR model are stationary, the X_t in each of the other segments (if any) cannot be stationary. In order to apply the FLIL to this sample ACVF within any given segment of a piecewise AR model when using conditional maximum likelihood estimation, we must show that the FLIL holds for $\hat{\gamma}(h)$ when we condition on any initial values $X_{-p+1}, \dots, X_0$. The following argument shows that for an AR(p) process $\{X_t\}$ with initial values $X_{-p+1}, \dots, X_0$, there exists a (causal) stationary AR(p) process $\{X'_t\}$ generated by the same AR coefficients and the same noise sequence such that as $t \to \infty$, $X_t - X'_t$ tends to zero at a geometric rate a.s. Thus, if $\{X'_t\}$ satisfies (2.9) and (2.10), so does $\{X_t\}$ where in (2.9) and (2.10), $\gamma(h) = \lim_{t\to\infty} \operatorname{Cov}(X_t, X_{t-h}) = \operatorname{Cov}(X'_s, X'_{s-h})$ for all s.

Suppose $\{X_t\}_{t=1}^{\infty}$ follows an AR(p) process with mean μ and AR coefficients $\phi_1, \dots, \phi_p$ conditioned on some initial values $X_{-p+1}, \dots, X_0$. Then this conditional process can be expressed as

$$X_t - \mu = \sum_{j=0}^{t-1} \psi_j \, \sigma\epsilon_{t-j} + a_{0t}(X_0 - \mu) + \cdots + a_{p-1,t}(X_{-p+1} - \mu),$$

where by the causality assumption, the ψ_j satisfy $\sum_{j=0}^{\infty} \psi_j B^j = (1-\sum_{j=1}^{p} \phi_j B^j)^{-1}$ (B the backward shift operator), $\{\epsilon_t\}$ is an iid noise sequence with mean zero and unit variance, and a_{it} is a function, depending on t, of sums and products of $\phi_1, \ldots, \phi_p$ for $i = 0, \ldots, p-1$. Defining the stationary process

$$X'_t - \mu = \sum_{j=0}^{\infty} \psi_j \, \sigma\epsilon_{t-j}, \ -\infty < t < \infty,$$

which satisfies

$$X'_t - \mu = \sum_{j=1}^{p} \phi_j (X'_{t-j} - \mu) + \sigma\epsilon_t, \ -\infty < t < \infty,$$

and

$$\begin{aligned} X'_t - \mu \quad &= \quad \sum_{j=0}^{t-1} \psi_j \sigma\epsilon_{t-j} + a_{0t}(X'_0 - \mu) \\ &\qquad + \cdots + a_{p-1,t}(X'_{-p+1} - \mu), \ t > 0, \end{aligned}$$

it follows that

$$X_t - X'_t = a_{0t}(X_0 - X'_0) + \cdots + a_{p-1,t}(X_{-p+1} - X'_{-p+1}).$$

Since each coefficient a_{it} tends to zero at a geometric rate, we have that as $t \to \infty$,

$$X_t - X'_t = O(\rho^t) \text{ a.s.} \tag{2.11}$$

for some constant $0 < \rho < 1$ depending on the AR coefficients $\phi_1, \ldots, \phi_p$.

3. Main Results. Consistency results require an assumption of a true model, in our case the piecewise autoregressive model (2.1). Throughout this paper we denote the true value of a parameter with a zero in the subscript or superscript when necessary (except for the AR coefficient parameters $\phi_{k,j}$ and white noise variances σ_k^2). Thus, we denote the true number of change-points by m_0, where the AR order in the kth segment is denoted by p_k^0, $k = 1, \ldots, m_0 + 1$. The change-point between the kth and $(k+1)$st AR processes is denoted by τ_k^0, or relative change-point λ_k^0. The $\lambda_k^0, k = 1, ..., m_0$, are allowed to depend on the sample size n subject to the condition that $\lambda_k^0 - \lambda_{k-1}^0 \geq \delta, k = 1, ..., m_0 + 1 (\lambda_0^0 := 0$ and $\lambda_{m_0+1}^0 := 1)$.

Davis et al. (2006) shows that when the true number of change-points, m_0, is known, the estimated change-point locations, $\hat{\lambda}_k$, are strongly consistent for the true change-point locations, λ_k^0. That is, $\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^0 \to \mathbf{0}$ a.s. as

$n \to \infty$. We will show that the estimated number of change-points, $\hat{m}$, and the estimated AR orders, $\hat{\boldsymbol{p}}$, are also consistent estimators for m_0 and $\boldsymbol{p}^0$, respectively, under conditional maximum likelihood estimation.

We begin by focusing on the case where there are no change-points in the underlying process. The first lemma proves that, when using conditional maximum likelihood variance estimates in the MDL, the estimated number of change-points is strongly consistent for the true number of change-points. Under this simple case, we will also show that consistency does not hold when we substitute Yule-Walker estimates for the conditional maximum likelihood estimates.

Assume that the true process follows the causal AR(p) model (we use p here rather than p^0 for simplicity) with mean μ and no change-points ($m_0 = 0$)

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \sigma\epsilon_t, \quad t = 1, \ldots, n, \tag{3.1}$$

where $\phi_0 = \mu(1 - \phi_1 - \cdots - \phi_p)$ and the noise sequence $\{\epsilon_t\}$ is iid with mean zero and unit variance.

LEMMA 3.1. *Assume the true process $\{X_t\}$ follows the AR(p) model given in (3.1) with no change-points and initial values $X_{-P+1}, \ldots, X_0$. Under assumptions (A1) and (A2), for any $m \geq 1$, with probability one,*

$$\mathrm{MDL}(0; p) < \min_{\boldsymbol{\lambda} \in A_m^\delta} \mathrm{MDL}(m, \boldsymbol{\lambda}; p, \ldots, p)$$

for n large, where $\mathrm{MDL}(0; p)$ denotes the MDL when fitting an AR(p) model with no change-points and $\mathrm{MDL}(m, \boldsymbol{\lambda}; p, \ldots, p)$ denotes the MDL when fitting a piecewise AR(p) model with $m \geq 1$ change-points, both of which use conditional maximum likelihood variance estimates.

PROOF. Assuming the AR order p is known, we will fit the following two models to the data set:

***Model 1*:** Fit an AR(p) model with no change-points.
***Model 2*:** Fit a piecewise AR(p) model with $m \geq 1$ relative change-points, $\boldsymbol{\lambda} \in A_m^\delta$, where A_m^δ is defined in (2.2).

The MDL for Model 1 is

$$\mathrm{MDL}(0; p) \quad = \quad \frac{p+4}{2} \log n + \log^+ p + \frac{n}{2}\left[\log(2\pi) + \log \hat{\sigma}^2\right],$$

where $\hat{\sigma}^2$ is the conditional maximum likelihood estimate of the AR(p) noise variance over the entire data set. The MDL for Model 2 is

$$\begin{aligned}\mathrm{MDL}(m, \boldsymbol{\lambda}; p, \dots, p) &= \log m + (m+1)\left(\frac{p+4}{2}\log n + \log^{+} p\right) \\ &\quad + \frac{p+2}{2}\sum_{k=1}^{m+1}\log(\lambda_k - \lambda_{k-1}) \\ &\quad + \frac{n}{2}\left[\log(2\pi) + \sum_{k=1}^{m+1}(\lambda_k - \lambda_{k-1})\log\hat{\sigma}_k^2\right],\end{aligned}$$

where $\hat{\sigma}_k^2$ is the conditional maximum likelihood estimate of the AR(p) noise variance in the kth fitted segment, $k = 1, \dots, m+1$.

Let $\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in A_m^{\delta}} \left\{\frac{2}{n}\mathrm{MDL}(m, \boldsymbol{\lambda}; p, \dots, p)\right\}$, and consider the quantity

$$\begin{aligned}&\frac{2}{n}[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \dots, p) - \mathrm{MDL}(0; p)] \\ &\quad = \frac{2\log m}{n} + m(p+4)\frac{\log n}{n} + \frac{2m\log^{+} p}{n} \\ &\quad\quad + \frac{p+2}{n}\sum_{k=1}^{m+1}\log(\hat{\lambda}_k - \hat{\lambda}_{k-1}) \\ &\quad\quad + \sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\hat{\sigma}_k^2 - \log\hat{\sigma}^2.\end{aligned} \tag{3.2}$$

We will show that (3.2) is strictly positive for n large with probability one. By assumption, $\delta \leq \hat{\lambda}_k - \hat{\lambda}_{k-1} < 1$, and hence the sum of the first four terms in (3.2) is $m(p+4)\log n/n + O(1/n)$. Since there are no change-points in the true process, $\hat{\sigma}_k^2 \to \sigma^2$ as $n \to \infty$ for all $k = 1, \dots, m+1$. Thus, since $\hat{\sigma}^2$ also converges to σ^2, the quantity

$$\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\hat{\sigma}_k^2 - \log\hat{\sigma}^2 \tag{3.3}$$

converges to zero as $n \to \infty$. We will use the FLIL on a Taylor series expansion of (3.3) to show that this quantity is of order $\log\log n/n$. Since $\log n/n > \log\log n/n$ for n large, the lemma follows.

Defining

$$\mathbf{X}_{a:b} = (X_a\ X_{a+1} \cdots X_b)^T \quad \text{and} \tag{3.4}$$

(3.5)

$$\mathbf{N}^p_{a:b} = \begin{pmatrix} 1 & X_{a-1} & X_{a-2} & \dots & X_{a-p} \\ 1 & X_a & X_{a-1} & \dots & X_{a-p+1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{b-1} & X_{b-2} & \dots & X_{b-p} \end{pmatrix} = \begin{pmatrix} 1 & \\ 1 & \\ \vdots & \mathbf{X}_{a-1:b-1} \dots \mathbf{X}_{a-p:b-p} \\ 1 & \end{pmatrix},$$

note that $n\hat{\sigma}^2$ is the squared norm of the difference between $\mathbf{X}_{1:n}$ and its projection onto the subspace spanned by $(1, \dots, 1)^T$ and $\mathbf{X}_{1-i:n-i}$, $i = 1, \dots, p$, i.e.,

$$n\hat{\sigma}^2 = \left\| \mathbf{X}_{1:n} - \mathrm{P}_{\boldsymbol{N}^p_{1:n}} \left(\mathbf{X}_{1:n} \right) \right\|^2, \tag{3.6}$$

where $\mathrm{P}_{\boldsymbol{N}^p_{1:n}}(\mathbf{X}_{1:n})$ is the projection of $\mathbf{X}_{1:n}$ onto the $(p+1)$-dimensional column space of $\boldsymbol{N}^p_{1:n}$. This is the same as the squared norm of the difference between $\mathbf{X}^*_{1:n}$ and its projection onto the subspace spanned by $\mathbf{X}^*_{1-i:n-i}$, $i = 1, \dots, p$, where $\mathbf{X}^*_{1:n}$ is the component of $\mathbf{X}_{1:n}$ orthogonal to $(1, \dots, 1)^T$, i.e.,

$$\mathbf{X}^*_{1:n} = \mathbf{X}_{1:n} - (\overline{X}_{1:n})(1, \dots, 1)^T,$$

and $\mathbf{X}^*_{1-i:n-i}, i = 1, \dots, p$ are defined similarly. It follows that

$$\hat{\sigma}^2 = G_p \left(\hat{\gamma}(i,j) : i, j = 0, \dots, p \right), \tag{3.7}$$

where

$$\hat{\gamma}(i,j) := \hat{\gamma}_{0:1}(i,j) = \frac{1}{n} \sum_{t=1}^{n} (X_{t-i} - \overline{X}_{1-i:n-i})(X_{t-j} - \overline{X}_{1-j:n-j}), \tag{3.8}$$

is the sample ACVF (cf. (2.8)), and

$$\begin{aligned} & G_p \left(u_{ij} : i, j = 0, \dots, p \right) = \\ & \qquad u_{00} - (u_{01}, \dots, u_{0p}) \left[\{u_{ij}\}^p_{i,j=1} \right]^{-1} \begin{pmatrix} u_{01} \\ \vdots \\ u_{0p} \end{pmatrix}. \end{aligned} \tag{3.9}$$

Similarly, for $k = 1, \dots, m+1$,

$$\begin{aligned} & (\hat{\lambda}_k - \hat{\lambda}_{k-1}) n\hat{\sigma}^2_k \\ & \quad = \left\| \mathbf{X}_{\hat{\lambda}_{k-1}n+1:\hat{\lambda}_k n} - \mathrm{P}_{\boldsymbol{N}^p_{\hat{\lambda}_{k-1}n+1:\hat{\lambda}_k n}} \left(\mathbf{X}_{\hat{\lambda}_{k-1}n+1:\hat{\lambda}_k n} \right) \right\|^2, \end{aligned} \tag{3.10}$$

and thus,

$$\hat{\sigma}_k^2 = G_p\left(\hat{\gamma}_k(i,j) : i,j = 0,\ldots,p\right), \tag{3.11}$$

where

$$\begin{aligned}\hat{\gamma}_k(i,j) &:= \hat{\gamma}_{\hat{\lambda}_{k-1}:\hat{\lambda}_k}(i,j)\\ &= \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=\hat{\lambda}_{k-1}n+1}^{\hat{\lambda}_k n} (X_{t-i} - \overline{X}_{\hat{\lambda}_{k-1}n+1-i:\hat{\lambda}_k n-i}) \cdot \\ &\qquad (X_{t-j} - \overline{X}_{\hat{\lambda}_{k-1}n+1-j:\hat{\lambda}_k n-j})\end{aligned} \tag{3.12}$$

is the sample ACVF in the kth segment (cf. (2.8)).

While $\{X_t\}$ is not assumed to be stationary, recall the argument in Section 2.3 showing that $\{X_t\}$ can be approximated by a stationary AR(p) process $\{X'_t\}$ generated by the same AR coefficients and the same noise sequence in such a way that (2.11) holds, i.e., as $t \to \infty$, $X_t - X'_t = O(\rho^t)$ a.s. where $0 < \rho < 1$ depends on the AR coefficients. Let

$$\mu := \lim_{t\to\infty} \mathbb{E}(X_t) \quad \text{and} \quad \gamma(h) := \lim_{t\to\infty} \operatorname{Cov}(X_t, X_{t-h}). \tag{3.13}$$

Note that

$$\mu = \mathbb{E}(X'_t) \quad \text{and} \quad \gamma(h) = \operatorname{Cov}(X'_t, X'_{t-h}) \text{ for all } t. \tag{3.14}$$

Without loss of generality, we take $\mu = 0$ since the estimates of $\hat{\sigma}_k^2$ are location invariant, and consider the μ-centered sample ACVF (cf. (2.7)),

$$\tilde{\gamma}(i,j) := \tilde{\gamma}_{0:1}(i,j) = \frac{1}{n}\sum_{t=1}^{n}(X_{t-i} - \mu)(X_{t-j} - \mu) = \frac{1}{n}\sum_{t=1}^{n} X_{t-i}X_{t-j} \tag{3.15}$$

and

$$\tilde{\gamma}_k(i,j) := \tilde{\gamma}_{\hat{\lambda}_{k-1}:\hat{\lambda}_k}(i,j) = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=\hat{\lambda}_{k-1}n+1}^{\hat{\lambda}_k n} X_{t-i}X_{t-j}. \tag{3.16}$$

Let $n\tilde{\sigma}^2$ denote the squared norm of the difference between $\mathbf{X}_{1:n}$ and its projection onto the subspace spanned by $\mathbf{X}_{1-i:n-i}$, $i = 1,\ldots,p$, and for each $k = 1,\ldots,m+1$, let $(\hat{\lambda}_k - \hat{\lambda}_{k-1})n\tilde{\sigma}_k^2$ denote the squared norm of the difference between $\mathbf{X}_{\hat\lambda_{k-1}n+1:\hat\lambda_k n}$ and its projection onto the subspace spanned by $\mathbf{X}_{\hat\lambda_{k-1}n+1-i:\hat\lambda_k n-i}$, $i=1,\dots,p$. It follows that

$$\tilde\sigma^2 = G_p(\tilde\gamma(i,j) : i,j=0,\dots,p), \quad \text{and} \tag{3.17}$$

$$\tilde\sigma_k^2 = G_p(\tilde\gamma_k(i,j) : i,j=0,\dots,p) \tag{3.18}$$

for each $k=1,\dots,m+1$. Since $\hat\gamma(i,j)-\tilde\gamma(i,j) = O(\log\log n/n)$ and $\hat\gamma_k(i,j)-\tilde\gamma_k(i,j) = O(\log\log n/n)$ by the FLIL, we have $\log\hat\sigma^2 - \log\tilde\sigma^2 = O(\log\log n/n)$ and $\log\hat\sigma_k^2 - \log\tilde\sigma_k^2 = O(\log\log n/n)$ for each of the $m+1$ fitted segments. We now show that

$$\sum_{k=1}^{m+1}(\hat\lambda_k - \hat\lambda_{k-1})\log\tilde\sigma_k^2 - \log\tilde\sigma^2 = O\left(\frac{\log\log n}{n}\right), \tag{3.19}$$

which then implies that

$$\sum_{k=1}^{m+1}(\hat\lambda_k - \hat\lambda_{k-1})\log\hat\sigma_k^2 - \log\hat\sigma^2 = O\left(\frac{\log\log n}{n}\right). \tag{3.20}$$

Let $\boldsymbol\gamma = (\gamma(|i-j|) : i,j=0,\dots,p)$ be the vector of true (limiting) ACVFs ranging over lags $0,\dots,p$ (cf. (3.13) and (3.14)). Carrying out a second order Taylor expansion about $\log G_p(\boldsymbol\gamma)$ on each of the $\log\tilde\sigma_k^2$ terms and the $\log\tilde\sigma^2$ term, we obtain

$$\begin{aligned}
&\sum_{k=1}^{m+1}(\hat\lambda_k - \hat\lambda_{k-1})\log\tilde\sigma_k^2 - \log\tilde\sigma^2 \\
&\quad= \sum_{k=1}^{m+1}(\hat\lambda_k - \hat\lambda_{k-1})\log G_p(\tilde{\boldsymbol\gamma}_k) - \log G_p(\tilde{\boldsymbol\gamma}) \\
&\quad= \left[\sum_{k=1}^{m+1}(\hat\lambda_k - \hat\lambda_{k-1})\log G_p(\boldsymbol\gamma) - \log G_p(\boldsymbol\gamma)\right] \\
&\qquad + \left[\sum_{k=1}^{m+1}(\hat\lambda_k - \hat\lambda_{k-1})\nabla\log G_p(\boldsymbol\gamma)(\tilde{\boldsymbol\gamma}_k - \boldsymbol\gamma) - \nabla\log G_p(\boldsymbol\gamma)(\tilde{\boldsymbol\gamma} - \boldsymbol\gamma)\right] \\
&\qquad + \frac{1}{2}\Bigg[\sum_{k=1}^{m+1}(\hat\lambda_k - \hat\lambda_{k-1})(\tilde{\boldsymbol\gamma}_k - \boldsymbol\gamma)^T\nabla^2\log G_p(\boldsymbol\gamma_k^*)(\tilde{\boldsymbol\gamma}_k - \boldsymbol\gamma) \\
&\qquad\qquad - (\tilde{\boldsymbol\gamma} - \boldsymbol\gamma)^T\nabla^2\log G_p(\boldsymbol\gamma^*)(\tilde{\boldsymbol\gamma} - \boldsymbol\gamma)\Bigg],
\end{aligned} \tag{3.21}$$

where $\tilde{\boldsymbol\gamma}_k := (\tilde\gamma_k(i,j) : i,j=0,\dots,p)$ and $\tilde{\boldsymbol\gamma} := (\tilde\gamma(i,j) : i,j=0,\dots,p)$. The variables $\boldsymbol\gamma^*$ and $\boldsymbol\gamma_k^*$ are between $\boldsymbol\gamma$ and $\tilde{\boldsymbol\gamma}$ or between $\boldsymbol\gamma$ and $\tilde{\boldsymbol\gamma}_k$, respectively, for $k = 1, \dots, m+1$, and each variable converges to $\boldsymbol{\gamma}$ almost surely as n goes to infinity.

Both the constant and first-order terms in the Taylor expansion (3.21) are exactly zero due to the form of the conditional maximum likelihood estimates; see (3.15) and (3.16). Within the second order term of the Taylor series expansion, we can apply the FLIL to the μ-centered sample ACVF. It is then readily seen that the second order term in the Taylor series expansion is of order $\log\log n/n$ with probability one (cf. (2.9) and (2.11)), and (3.19) holds. (More precisely, (2.9) applies to the μ-centered sample ACVF for the stationary $\{X'_t\}$. Due to (2.11), it also applies to $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\gamma}}_k$ for $\{X_t\}$.) Thus, by (3.20), (3.2) becomes

$$\begin{aligned}&\frac{2}{n}[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \dots, p) - \mathrm{MDL}(0; p)] \\ &\quad = \quad m(p+4)\frac{\log n}{n} + O\left(\frac{1}{n}\right) + O\left(\frac{\log\log n}{n}\right),\end{aligned}$$

which is greater than zero for large n with probability one. □

Lemma 3.1 can be extended to the case where the autoregressive order is unknown, implying that the estimated AR order is also consistent for the true AR order in this simple case. This result is given in Appendix A.1 as Lemma A.1.

Since Yule-Walker estimates have the same asymptotic distribution as the conditional maximum likelihood estimates (see Section 8.10 in [3]), one would expect that substituting Yule-Walker estimates into the MDL would not change the consistency result. However, due to the delicate argument in the proof of Lemma 3.1 where the sample ACVF terms in the Taylor expansion (3.21) cancel exactly, using Yule-Walker estimates in the MDL does not guarantee consistency of the estimated number of change-points. One example, where the noise has exponential tails, is stated in the next theorem. Yule-Walker variance estimates may provide weakly consistent estimators for the number of change-points if the noise has normal tails, but this needs to be explored further.

THEOREM 3.1. *Assume that the true process $\{X_t\}$ is a stationary mean zero AR(1) given by*

$$X_t = \phi X_{t-1} + \sigma\epsilon_t,$$

where $|\phi| < 1$ and the noise sequence $\{\epsilon_t\}$ is iid with mean zero and variance one. Furthermore suppose that $\{X_t\}$ satisfies assumptions (A1) and (A2) and that the noise ϵ_t has a density function f_ϵ which satisfies

(i) $f_\epsilon(x) > 0$ *for* $-\infty < x < \infty$,
(ii) $f_\epsilon(x) = f_\epsilon(-x)$,
(iii) $\liminf_{u\to\infty} e^{cu}\int_u^\infty f_\epsilon(x)dx > 0$ *for some constant* $c > 0$.

Then, using Yule-Walker estimation in the MDL, for every $\delta > 0$ *and* $C > 0$,

$$\liminf_{n\to\infty} P\left(\mathrm{MDL}(0;1) - \min_{\delta\le\lambda\le 1-\delta}\mathrm{MDL}(1,\lambda;1,1) > C\log n\right) > 0.$$

Proof. See Appendix A.2. □

Though consistency does not hold when Yule-Walker estimates are used in the MDL, weak consistency can still be obtained for the general piecewise AR model (2.1) when using conditional maximum likelihood estimates. The next theorem states that the estimated number of change-points and the estimated AR orders are weakly consistent for their respective true parameters when the true process follows the piecewise AR model (2.1) and meets the assumptions for the FLIL to hold for the sample ACVF within each segment.

An additional point that is worth noting here is that while the strong consistency result of Lemma 3.1 holds conditionally on the initial values $X_{-P+1},\dots,X_0$, the weak consistency version, Theorem 3.2, holds unconditionally as long as the initial values are stochastically bounded (i.e., the $X_i, i = -P+1,\dots,0$ are allowed to depend on the sample size n with $X_i = O_p(1), i = -P+1,\dots,0$). Furthermore, assuming the initial values are stochastically bounded, consider the case that $m_0 > 0$ and $\lambda_k^0 - \lambda_{k-1}^0 \ge \delta, k = 1,\dots,m_0+1$. Then the initial values for the second segment are simply the last P values of the first segment, i.e., $X_{\lambda_1^0 n-i}, i = 0,\dots,P-1$, which are stochastically bounded by the causality assumption (A1). (Indeed, after an initial transient period, the process in the first segment is nearly stationary.) This shows that the initial values for each of the m_0+1 segments are stochastically bounded.

Theorem 3.2. *Assume that the true process* $\{X_t\}$ *follows the piecewise AR model given in (2.1) with* m_0 *change-points and initial values* $X_{-P+1},\dots,X_0$ *and that the relative length of each of the* m_0+1 *segments is at least* δ. *Further suppose assumptions (A1) and (A2) hold for the stationary process determined by the parameter values for each of the* m_0+1 *segmented autoregressions. Then*

$$\hat{m} \xrightarrow{P} m_0$$

and

$$(\hat{p}_1, \dots, \hat{p}_{\hat{m}+1}) \xrightarrow{P} (p_1^0, \dots, p_{m_0+1}^0),$$

where $\hat{m}$ is the estimated number of change-points and $(\hat{p}_1, \dots, \hat{p}_{\hat{m}+1})$ are the estimated AR orders obtained by minimizing the MDL defined in (2.3) using conditional maximum likelihood variance estimates.

Proof. Assume that the true model is the piecewise autoregressive process defined by (2.1). In order to prove weak consistency for the estimator of the number of change-points and the AR order estimates, we need only compare the following two fitted models:

Model **$1'$:** Fit a piecewise autoregressive model to the data set with m_0 relative change-points, $\boldsymbol{\lambda} \in A_{m_0}^{\delta}$, where the AR orders, $p_1^0, \dots, p_{m_0+1}^0$, are known.

Model **$2'$:** Fit a piecewise autoregressive model to the data set with $m_0 + s$ relative change-points, $\boldsymbol{\alpha} \in A_{m_0+s}^{\delta}$, where s is a positive integer. Estimate the autoregressive orders from the data, and denote these orders by $\hat{p}_1, \dots, \hat{p}_{m_0+s+1}$.

If we can show that the MDL for Model $2'$ is larger than the MDL for Model $1'$ for large n in probability, then since for large n, the estimated number of change-points cannot underestimate the true number of change-points with probability one, this implies that the MDL for a model with m change-points, where $m \neq m_0$ and $m < M$, is larger than the MDL for a model with m_0 change-points for large n in probability, and thus, $\hat{m} \xrightarrow{P} m_0$ as n tends to infinity.

We will outline the proof for the simple case where the true number of change-points is $m_0 = 1$, each segment follows an autoregressive model with mean zero and order 1, and we fit AR(1) models to the data. Note that, unlike the case where $m_0 = 0$, the mean zero assumption is no longer a valid simplification. We will address this along with the general case in Appendix A.1. Let λ be the true relative change-point location (we use λ here rather than λ^0 for simplicity), i.e., $\tau = \lambda n$ is the observation at which the change occurs. Denote the true white noise variances in the first and second segments by σ_1^2 and σ_2^2, respectively. For simplicity, we take $s = 1$ in Model $2'$. Models $1'$ and $2'$ then become:

Model **$1'$:** Fit AR(1) models to two segments with relative change-point location λ.

Model **$2'$:** Fit AR(1) models to three segments with relative change-point location estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ obtained by minimizing $\mathrm{MDL}(2, \alpha_1, \alpha_2; 1, 1)$ with respect to $(\alpha_1, \alpha_2) \in A_2^{\delta}$, where A_2^{δ} is defined in (2.2).

We would like to show that

$$\lim_{n\to\infty} P\big(\mathrm{MDL}(2,\hat\alpha_1,\hat\alpha_2;1,1) > \mathrm{MDL}(1,\lambda;1)\big) = 1,$$

where $\mathrm{MDL}(1,\lambda;1)$ is the MDL for Model $1'$, and $\mathrm{MDL}(2,\hat\alpha_1,\hat\alpha_2;1,1)$ is the minimized MDL for Model $2'$. Equivalently, we show that

$$\begin{aligned}&\lim_{n\to\infty} P\big(\mathrm{MDL}(2,\alpha_1,\alpha_2;1,1) > \mathrm{MDL}(1,\lambda;1)\\ &\qquad\forall\ \alpha_1,\alpha_2:\ \delta<\alpha_1<\alpha_1+\delta\le\alpha_2<1-\delta\big) = 1.\end{aligned} \tag{3.22}$$

We will prove (3.22) for the case where $\alpha_1<\lambda<\alpha_2$. If $\lambda<\alpha_1$ or $\alpha_2<\lambda$, the proof is analogous, with the fitted segment $(0,\alpha_1)$ or $(\alpha_2,1)$, respectively, taking the role of the fitted segment (α_1,α_2) in the argument that follows. The argument will depend on how close the fitted change-points, α_1 and α_2, are to the true change-point λ. It will suffice to address the case where only α_1 is allowed to be close to λ. Therefore, (3.22) follows if each of the following statements holds.

(i) $\lim_{n\to\infty} P\big(\mathrm{MDL}(2,\alpha_1,\alpha_2;1,1) > \mathrm{MDL}(1,\lambda;1)\ \forall\ \alpha_1,\alpha_2:$
$\delta<\alpha_1<\lambda-(\log\log n)^2/\log n;\ \lambda+\delta/2<\alpha_2<1-\delta\big)=1.$

(ii) For each finite positive integer N,

$$\begin{aligned}&\lim_{n\to\infty} P\big(\mathrm{MDL}(2,\alpha_1,\alpha_2;1,1) > \mathrm{MDL}(1,\lambda;1)\ \ \forall\ \alpha_1,\alpha_2:\\ &\qquad\lambda-N/n<\alpha_1<\lambda;\ \lambda+\delta/2<\alpha_2<1-\delta\big)=1.\end{aligned}$$

(iii) For every $\epsilon>0$, there exists a positive integer N such that

$$\begin{aligned}&P\big(\mathrm{MDL}(2,\alpha_1,\alpha_2;1,1) > \mathrm{MDL}(1,\lambda;1)\ \ \forall\ \alpha_1,\alpha_2:\\ &\qquad\lambda-(\log\log n)^2/\log n<\alpha_1<\lambda-N/n;\\ &\qquad\lambda+\delta/2<\alpha_2<1-\delta\big)\ >\ 1-\epsilon\end{aligned}$$

for sufficiently large n.

Consider the difference

$$\begin{aligned}&\frac{2}{n}[\mathrm{MDL}(2,\alpha_1,\alpha_2;1,1)-\mathrm{MDL}(1,\lambda;1)]\\ &\quad= O\left(\frac{\log n}{n}\right)+\alpha_1\log\hat\sigma^2_{1,2}+(\alpha_2-\alpha_1)\log\hat\sigma^2_{2,2}\\ &\qquad+(1-\alpha_2)\log\hat\sigma^2_{3,2}-\Big[\lambda\log\hat\sigma^2_{1,1}+(1-\lambda)\log\hat\sigma^2_{2,1}\Big].\end{aligned} \tag{3.23}$$

where $\hat\sigma^2_{k,m}$ is the conditional maximum likelihood variance estimate for the kth segment when fitting a piecewise AR(1) model with m change-points.

Again, the penalty term, $O(\log n/n)$, is positive for large n, so we need only show that the remaining terms are of a smaller order than $O(\log n/n)$.

Partition the interval $(0,1)$ into the intervals $(0,\alpha_1)$, (α_1,λ), (λ,α_2), and $(\alpha_2,1)$. The fundamental idea of the proof is to break the term

$$\begin{aligned}&\alpha_1 \log \hat{\sigma}^2_{1,2} + (\alpha_2-\alpha_1)\log \hat{\sigma}^2_{2,2} + (1-\alpha_2)\log \hat{\sigma}^2_{3,2}\\ &\quad -\left[\lambda \log \hat{\sigma}^2_{1,1} + (1-\lambda)\log \hat{\sigma}^2_{2,1}\right]\end{aligned} \tag{3.24}$$

into a sum over our partition of intervals and examine each true segment, $(0,\lambda)$ or $(\lambda,1)$, separately. Within each true segment, we show that the difference between the terms in the two fitted models is either positive or of smaller order than $\log n/n$. The method we use to compare terms within each true segment will differ depending on if

(i) $(\log\log n)^2/\log n < \lambda - \alpha_1$,
(ii) $\lambda - \alpha_1 < N/n$ for some positive integer N, or
(iii) $N/n < \lambda - \alpha_1 < (\log\log n)^2/\log n$ for some positive integer N.

These three cases correspond to statements (i), (ii), and (iii) made earlier.

For the case where $\alpha_1 < \lambda < \alpha_2$, the only term in (3.24) that we need to partition is $(\alpha_2-\alpha_1)\log \hat{\sigma}^2_{2,2}$. Since

$$\hat{\sigma}^2_{2,2} = \frac{\sum_{t=\alpha_1 n+1}^{\alpha_2 n}\left(X_t - \hat{\phi}X_{t-1}\right)^2}{(\alpha_2-\alpha_1)n},$$

where $\hat{\phi}$ is the AR coefficient estimate when fitting an AR(1) model to the second fitted segment, (α_1,α_2), we can use concavity of the log function and the definition of conditional maximum likelihood estimation to show that for large n,

$$\begin{aligned}&(\alpha_2-\alpha_1)\log \hat{\sigma}^2_{2,2}\\ &\quad\geq (\lambda-\alpha_1)\log\left(\frac{\mathrm{RSS}_{2,1}}{(\lambda-\alpha_1)n}\right) + (\alpha_2-\lambda)\log\left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2-\lambda)n}\right),\end{aligned} \tag{3.25}$$

where

$$\mathrm{RSS}_{2,1} := \sum_{t=\alpha_1 n+1}^{\lambda n}\left(X_t - \hat{\phi}_1 X_{t-1}\right)^2,$$

$$\mathrm{RSS}_{2,2} := \sum_{t=\lambda n+1}^{\alpha_2 n}\left(X_t - \hat{\phi}_2 X_{t-1}\right)^2,$$

$\hat{\phi}_1$ is the AR(1) coefficient estimate fit to the segment $(\alpha_1, \lambda]$, and $\hat{\phi}_2$ is the AR(1) coefficient estimate fit to the segment $(\lambda, \alpha_2]$. Substituting (3.25) into (3.24), we can then group terms corresponding to each true segment and use Taylor series expansions as in Lemma 3.1 to show statement (i).

Cases (ii) and (iii) require further argument since if the fitted change-point α_1 is very close to the true change-point λ, we cannot use $\mathrm{RSS}_{2,1}$ as defined above. In case (ii), the number of observations between $\alpha_1 n$ and $\lambda n - 1$ will be finite in the limit. Therefore, using another Taylor expansion on the log function, it can be shown that

$$(\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} \quad \geq \quad O_p\left(\frac{1}{n}\right) + (\alpha_2 - \lambda) \log\left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n}\right).$$

Again, we can substitute this into (3.24) and use Taylor series expansions to show statement (ii). For case (iii), it can be shown that for any $\epsilon > 0$, there exists a positive integer N such that with probability greater than $1 - \epsilon$,

$$\begin{aligned}&(\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} \\ &\quad \geq \quad (\lambda - \alpha_1) \log\left(\sigma_1^2 + \eta\right) + (\alpha_2 - \lambda) \log\left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n}\right),\end{aligned}$$

for every α_1 such that $(\lambda - \alpha_1)n = N+1, N+2, \ldots, (\log\log n)^2$ and for some $\eta > 0$. Substituting this into (3.24) and using Taylor series expansions as in the proof of Lemma 3.1 leads to statement (iii) and the result follows. □

APPENDIX A: PROOF DETAILS

A.1. Conditional Maximum Likelihood. In this section, we prove the extension of Lemma 3.1 to the case where the AR order is unknown and provide further details for the proof of Theorem 3.2.

In extending Lemma 3.1, still under the assumption that the data follow the AR(p) process defined in (3.1), now we will fit a model to the data which does not assume that the AR order of the true process is known:

***Model 3*:** Fit a piecewise autoregressive model to the data set with $m \geq 1$ relative change-points, $\boldsymbol{\lambda} \in A_m^\delta$. Estimate the autoregressive orders from the data by minimizing $\mathrm{MDL}(m, \boldsymbol{\lambda}; p_1, \ldots, p_{m+1})$ over $0 \leq p_j \leq P, j = 1, \ldots, m+1$, and denote these estimated orders by $\hat{p}_1, \ldots, \hat{p}_{m+1}$.

Then the minimum description length for Model 3 is

$$\begin{aligned}
&\mathrm{MDL}(m, \boldsymbol{\lambda}; \hat{p}_1, \ldots, \hat{p}_{m+1}) \\
&\quad = \log m + (m+1)\log n + \sum_{k=1}^{m+1} \log^+ \hat{p}_k \\
&\qquad + \sum_{k=1}^{m+1} \frac{\hat{p}_k + 2}{2} \log\left((\lambda_k - \lambda_{k-1})n\right) \\
&\qquad + \sum_{k=1}^{m+1} \frac{(\lambda_k - \lambda_{k-1})n}{2} \log\left(2\pi\hat{\sigma}_k^2(\hat{p}_k)\right),
\end{aligned}$$

where $\hat{\sigma}_k^2(p')$ denotes the conditional maximum likelihood variance estimate when fitting an AR(p') model to the data $X_t, \lambda_{k-1}n < t \le \lambda_k n$. The next lemma states that with probability one, the minimum description length for Model 1 is strictly smaller than the minimized MDL for Model 3 for n large, implying that the estimate of the number of change-points and the AR order estimates are strongly consistent when there are no true change-points.

LEMMA A.1. *Assume that the true process $\{X_t\}$ follows the AR(p) model given in (3.1) with no change-points ($m_0 = 0$) and initial values $X_{-P+1}, \ldots,$ X_0, and satisfies assumptions (A1) and (A2). Then (i) for any $m \ge 1$, with probability one, for n large,*

$$\mathrm{MDL}(0; p) < \min_{\boldsymbol{\lambda} \in A_m^\delta} \mathrm{MDL}(m, \boldsymbol{\lambda}; \hat{p}_1, \ldots, \hat{p}_{m+1});$$

and (ii) for any $p' \ne p$, with probability one, for n large,

$$\mathrm{MDL}(0; p) < \mathrm{MDL}(0; p').$$

PROOF. Let $\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in A_m^\delta} \{\mathrm{MDL}(m, \boldsymbol{\lambda}; \hat{p}_1, \ldots, \hat{p}_{m+1})\}$. Note that

$$\begin{aligned}
&\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; \hat{p}_1, \ldots, \hat{p}_{m+1}) - \mathrm{MDL}(0; p) \\
&\quad = \left[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; \hat{p}_1, \ldots, \hat{p}_{m+1}) - \mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p)\right] \\
&\qquad + \left[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p) - \mathrm{MDL}(0; p)\right],
\end{aligned}$$

where $\hat{p}_k = \hat{p}_k(\hat{\boldsymbol{\lambda}}), k = 1, \ldots, m+1$, depend on $\hat{\boldsymbol{\lambda}}$. We know from Lemma 3.1 that $\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p) - \mathrm{MDL}(0; p) > 0$ for n large with probability one. Therefore, to prove Lemma A.1, we need only show that

$$\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; \hat{p}_1, \ldots, \hat{p}_{m+1}) - \mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p) \ge 0$$

for n large with probability one.

It suffices to prove that for any $p_1, \dots, p_{m+1}$ not all equal to p, with probability one, for n large,

$$\min_{\boldsymbol{\lambda} \in A_m^\delta} \{\mathrm{MDL}(m, \boldsymbol{\lambda}; p_1, \dots, p_{m+1}) - \mathrm{MDL}(m, \boldsymbol{\lambda}; p, \dots, p)\} > 0. \tag{A.1}$$

Note by (2.3) that

$$\begin{aligned} &\mathrm{MDL}(m, \boldsymbol{\lambda}; p_1, \dots, p_{m+1}) - \mathrm{MDL}(m, \boldsymbol{\lambda}; p, \dots, p) \\ &\quad = \sum_{k=1}^{m+1} \biggl\{ (\log^+ p_k - \log^+ p) + \frac{p_k - p}{2} \log(\lambda_k n) \\ &\qquad + \frac{(\lambda_k - \lambda_{k-1})n}{2} \bigl(\log \hat{\sigma}_k^2(p_k) - \log \hat{\sigma}_k^2(p) \bigr) \biggr\}. \end{aligned} \tag{A.2}$$

We claim that for $p_k \neq p$, with probability one, for n large,

$$\begin{aligned} &\min_{\boldsymbol{\lambda} \in A_m^\delta} \biggl\{ (\log^+ p_k - \log^+ p) + \frac{p_k - p}{2} \log(\lambda_k n) + \\ &\qquad \frac{(\lambda_k - \lambda_{k-1})n}{2} \bigl(\log \hat{\sigma}_k^2(p_k) - \log \hat{\sigma}_k^2(p) \bigr) \biggr\} > 0, \end{aligned} \tag{A.3}$$

which together with (A.2) implies (A.1).

If $p_k < p$, we have by Lemma A.2(i) below that

$$\begin{aligned} v(p_k) &= \lim_{n \to \infty} \min_{0 \le \lambda < \lambda + \delta \le \lambda' \le 1} \hat{\sigma}_{\lambda:\lambda'}^2(p_k) \\ &\le \lim_{n \to \infty} \min_{\boldsymbol{\lambda} \in A_m^\delta} \hat{\sigma}_k^2(p_k) \\ &\le \lim_{n \to \infty} \max_{\boldsymbol{\lambda} \in A_m^\delta} \hat{\sigma}_k^2(p_k) \\ &\le \lim_{n \to \infty} \max_{0 \le \lambda < \lambda + \delta \le \lambda' \le 1} \hat{\sigma}_{\lambda:\lambda'}^2(p_k) = v(p_k) \text{ a.s.} \end{aligned}$$

where $v(p_k)(> \sigma^2)$ is given in Lemma A.2 and $\hat{\sigma}_{\lambda:\lambda'}^2(p')$ denotes the conditional maximum likelihood variance estimate when fitting an AR(p') model to the data $X_t, \lambda n < t \le \lambda' n$. But $\hat{\sigma}_k^2(p)$ converges a.s. to σ^2 uniformly in $\boldsymbol{\lambda} \in A_m^\delta$, which implies that (A.3) holds for $p_k < p$.

For $p_k > p$, by Lemma A.2(ii),

$$\begin{aligned} 0 &\le \max_{\boldsymbol{\lambda} \in A_m^\delta} \{\log \hat{\sigma}_k^2(p) - \log \hat{\sigma}_k^2(p_k)\} \\ &\le \max_{0 \le \lambda < \lambda + \delta \le \lambda' \le 1} \{\log \hat{\sigma}_{\lambda:\lambda'}^2(p) - \log \hat{\sigma}_{\lambda:\lambda'}^2(p_k)\} = O\left(\frac{\log \log n}{n}\right) \text{ a.s.,} \end{aligned}$$

implying that (A.3) holds for $p_k > p$. This proves part (i).

To prove part (ii), when fitting an AR(p') model with $p' < p$ to the entire data set, by Lemma A.2(i), the conditional maximum likelihood variance estimate $\hat{\sigma}^2_{0:1}(p')$ converges a.s. to $v(p') > \sigma^2$, so that $\mathrm{MDL}(0; p') > \mathrm{MDL}(0; p)$ for n large with probability one. For $p' > p$, we have by Lemma A.2(ii) that

$$\log \hat{\sigma}^2_{0:1}(p) - \log \hat{\sigma}^2_{0:1}(p') = O\left(\frac{\log\log n}{n}\right) \text{ a.s.,}$$

implying $\mathrm{MDL}(0; p') > \mathrm{MDL}(0; p)$ for n large with probability one. This proves part (ii).

□

Lemma A.2. *Under the same assumptions as in Lemma A.1, let*

$$v(0) := \lim_{t\to\infty} \min_{a_0} \mathbb{E}(X_t - a_0)^2$$

and for $p' = 1, 2, \ldots,$

$$v(p') := \lim_{t\to\infty} \min_{a_0,\ldots,a_{p'}} \mathbb{E}[X_t - (a_0 + a_1 X_{t-1} + \cdots + a_{p'} X_{t-p'})]^2.$$

For $0 \le \lambda < \lambda' \le 1$, *let* $\hat{\sigma}^2_{\lambda:\lambda'}(p')$ *denote the conditional maximum likelihood variance estimate when fitting an* AR(p') *model to the data* $X_t, \lambda n < t \le \lambda' n$.

(i) If $p' < p$, *then* $\sigma^2 < v(p') = G_{p'}(\gamma(|i-j|) : 0 \le i, j \le p')$, *and* $\hat{\sigma}^2_{\lambda:\lambda'}(p')$ *converges a.s. to* $v(p')$ *uniformly in* $0 \le \lambda < \lambda + \delta \le \lambda' \le 1$, *i.e.*

$$\lim_{n\to\infty} \min_{0\le\lambda<\lambda+\delta\le\lambda'\le1} \hat{\sigma}^2_{\lambda:\lambda'}(p') = v(p') \textit{ a.s.,}$$
$$\lim_{n\to\infty} \max_{0\le\lambda<\lambda+\delta\le\lambda'\le1} \hat{\sigma}^2_{\lambda:\lambda'}(p') = v(p') \textit{ a.s.,}$$

where $G_{p'}$ *is defined as in (3.9) and* $\gamma(h) := \lim_{t\to\infty} \mathrm{Cov}(X_t, X_{t-h})$.

(ii) If $p' > p$, *then*

$$\max_{0\le\lambda<\lambda+\delta\le\lambda'\le1} \{\log \hat{\sigma}^2_{\lambda:\lambda'}(p) - \log \hat{\sigma}^2_{\lambda:\lambda'}(p')\} = O\left(\frac{\log\log n}{n}\right) \textit{ a.s.}$$

Proof. Recall the argument in Section 2.3 showing that there exists a causal stationary AR(p) process $\{X'_t\}$ generated by the same AR coefficients and the same noise sequence such that (2.11) holds. So it suffices to prove the lemma under the additional assumption that $\{X_t\}$ is stationary. Then

$\gamma(h) = \mathrm{Cov}(X_t, X_{t-h})$ for all t. Let μ denote the common mean of X_t. Letting $a = \lambda n + 1$ and $b = \lambda' n$, we have (cf. (3.6) and (3.7))

$$\begin{aligned}\hat{\sigma}^2_{\lambda:\lambda'}(p') &= \frac{1}{(\lambda' - \lambda)n} \|\mathbf{X}_{a:b} - P_{\boldsymbol{N}^{p'}_{a:b}}(X_{a:b})\|^2 \\ &= G_{p'}(\hat{\gamma}_{\lambda:\lambda'}(i,j) : i,j = 0,\ldots,p'),\end{aligned}$$

where $N^{p'}_{a:b}$, $G_{p'}$ and $\hat{\gamma}_{\lambda:\lambda'}(i,j)$ are defined as in (3.5), (3.9) and (2.8), respectively. Part (i) follows by observing that for $p' < p$,

$$\begin{aligned}\sigma^2 < v(p') &= \min_{a_0,\ldots,a_{p'}} \mathbb{E}[X_t - (a_0 + a_1 X_{t-1} + \cdots + a_{p'} X_{t-p'})]^2 \text{ (by stationarity)} \\ &= \min_{a_1,\ldots,a_{p'}} \mathbb{E}[(X_t - \mu) - \{a_1(X_{t-1} - \mu) + \cdots + a_{p'}(X_{t-p'} - \mu)\}]^2 \\ &= G_{p'}(\gamma(|i-j|) : i,j = 0,1,\ldots,p')\end{aligned}$$

and that uniformly in $0 \le \lambda < \lambda + \delta \le \lambda' \le 1$,

$$\lim_{n\to\infty} \hat{\gamma}_{\lambda:\lambda'}(i,j) = \gamma(|i-j|) \text{ a.s.}, i,j = 0,\ldots,p'.$$

For part (ii), we will only prove that

(A.4)

$$\max_{0\le\lambda<\lambda+\delta\le\lambda'\le1} \{\log \hat{\sigma}^2_{\lambda:\,\lambda'}(p) - \log \hat{\sigma}^2_{\lambda:\,\lambda'}(p+1)\} = O\left(\frac{\log\log n}{n}\right) \text{ a.s.},$$

or equivalently,

$$\max_{0\le\lambda<\lambda+\delta\le\lambda'\le1} \{\hat{\sigma}^2_{\lambda:\,\lambda'}(p) - \hat{\sigma}^2_{\lambda:\,\lambda'}(p+1)\} = O\left(\frac{\log\log n}{n}\right) \text{ a.s.}$$

A similar argument can be used to show that for $l = 1,2,\ldots,$

$$\max_{0\le\lambda<\lambda+\delta\le\lambda'\le1} \{\log \hat{\sigma}^2_{\lambda:\,\lambda'}(p+l) - \log \hat{\sigma}^2_{\lambda:\,\lambda'}(p+l+1)\} = O\left(\frac{\log\log n}{n}\right) \text{ a.s.}$$

Without loss of generality, assume $\mu = 0$ in what follows. Letting $a = \lambda n + 1, b = \lambda' n$, and $p* = p$ or $p+1$, denote by $\boldsymbol{M}^{p*}_{a:b}$ the matrix $\boldsymbol{N}^{p*}_{a:b}$ with the first column of 1's removed, i.e.

$$\boldsymbol{M}^{p*}_{a:b} = (\boldsymbol{X}_{a-1:b-1} \cdots \boldsymbol{X}_{a-p*:b-p*}). \tag{A.5}$$

Define

$$\tilde{\sigma}^2_{\lambda:\lambda'}(p*) := \frac{1}{(\lambda'-\lambda)n} \|\boldsymbol{X}_{a:b} - P_{\boldsymbol{M}^{p*}_{a:b}} \boldsymbol{X}_{a:b}\|^2, \tag{A.6}$$

which can be readily shown (cf. (3.17) and (3.18)) to equal $G_{p*}(\tilde{\gamma}_{\lambda:\lambda'}(i,j) : i,j = 0,\dots,p*)$ where $\tilde{\gamma}_{\lambda:\lambda'}(i,j)$ is defined as in (2.7) with $\mu = 0$, i.e.

$$\tilde{\gamma}_{\lambda:\lambda'}(i,j) = \frac{1}{(\lambda' - \lambda)n} \sum_{t=a}^{b} X_{t-i} X_{t-j}.$$

Since uniformly in $0 \le \lambda < \lambda + \delta \le \lambda' \le 1$,

$$\hat{\gamma}_{\lambda:\lambda'}(i,j) - \tilde{\gamma}_{\lambda:\lambda'}(i,j) = O\left(\frac{\log\log n}{n}\right) \text{ a.s.},$$

we have uniformly in $0 \le \lambda < \lambda + \delta \le \lambda' \le 1$,

$$\begin{aligned}
&\hat{\sigma}^2_{\lambda:\lambda'}(p*) - \tilde{\sigma}^2_{\lambda:\lambda'}(p*) \\
&= G_{p*}(\hat{\gamma}_{\lambda:\lambda'}(i,j) : i,j = 0,\dots,p*) - G_{p*}(\tilde{\gamma}_{\lambda:\lambda'}(i,j) : i,j = 0,\dots,p*) \\
&= O\left(\frac{\log\log n}{n}\right) \text{ a.s.}
\end{aligned}$$

Thus, (A.5) holds if we can show that

$$\text{(A.7)} \quad \max_{0\le\lambda<\lambda+\delta\le\lambda'\le 1} \{\tilde{\sigma}^2_{\lambda:\lambda'}(p) - \tilde{\sigma}^2_{\lambda:\lambda'}(p+1)\} = O\left(\frac{\log\log n}{n}\right) \text{ a.s.}$$

We can express the projection of $\mathbf{X}_{a:b}$ onto the column space of $\mathbf{M}^{p+1}_{a:b}$ as

(A.8)

$$\begin{aligned}
\mathrm{P}_{\mathbf{M}^{p+1}_{a:b}} \mathbf{X}_{a:b} &= \mathbf{M}^{p+1}_{a:b} \hat{\boldsymbol{\phi}}^{p+1}_{a:b} \\
&= \hat{\phi}^{p+1,1}_{a:b} \mathbf{X}_{a-1:b-1} + \cdots + \hat{\phi}^{p+1,p+1}_{a:b} \mathbf{X}_{a-(p+1):b-(p+1)} \\
&= \hat{\phi}^{p+1,1}_{a:b} \mathbf{X}_{a-1:b-1} + \cdots + \hat{\phi}^{p+1,p}_{a:b} \mathbf{X}_{a-p:b-p} \\
&\quad + \hat{\phi}^{p+1,p+1}_{a:b} \Big[\mathrm{P}_{\mathbf{M}^{p}_{a:b}} (\mathbf{X}_{a-(p+1):b-(p+1)}) \\
&\quad + \mathrm{P}^{\perp}_{\mathbf{M}^{p}_{a:b}} (\mathbf{X}_{a-(p+1):b-(p+1)}) \Big],
\end{aligned}$$

where

$$\text{(A.9)} \quad \begin{aligned}
\hat{\boldsymbol{\phi}}^{p+1}_{a:b} &= \left(\hat{\phi}^{p+1,1}_{a:b}, \dots, \hat{\phi}^{p+1,p+1}_{a:b} \right)^T \\
&= \left(\mathbf{M}^{p+1\,T}_{a:b} \mathbf{M}^{p+1}_{a:b} \right)^{-1} \mathbf{M}^{p+1\,T}_{a:b} \mathbf{X}_{a:b},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{P}^{\perp}_{\mathbf{M}^{p}_{a:b}} (\mathbf{X}_{a-(p+1):b-(p+1)}) &= \mathbf{X}_{a-(p+1):b-(p+1)} \\
&\quad - \mathrm{P}_{\mathbf{M}^{p}_{a:b}} (\mathbf{X}_{a-(p+1):b-(p+1)}).
\end{aligned}$$

The projection of $\mathbf{X}_{a-(p+1):b-(p+1)}$ onto the column space of $\mathrm{P}_{\mathbf{M}^p_{a:b}}$ will be denoted as

$$\mathrm{P}_{\mathbf{M}^p_{a:b}}(\mathbf{X}_{a-(p+1):b-(p+1)}) = \mathbf{M}^p_{a:b}\tilde{\boldsymbol{\phi}}^p_{a:b} = \sum_{j=1}^{p} \tilde{\phi}^{p,j}_{a:b}\mathbf{X}_{a-j:b-j}, \tag{A.10}$$

where we use $\tilde{\boldsymbol{\phi}}$ rather than $\hat{\boldsymbol{\phi}}$ to distinguish the estimated coefficients

$$\tilde{\boldsymbol{\phi}}^p_{a:b} = \left(\mathbf{M}^{p\,T}_{a:b}\mathbf{M}^p_{a:b}\right)^{-1}\mathbf{M}^{p\,T}_{a:b}\mathbf{X}_{a-(p+1):b-(p+1)}$$

from the estimated coefficients

$$\hat{\boldsymbol{\phi}}^p_{a:b} = \left(\mathbf{M}^{p\,T}_{a:b}\mathbf{M}^p_{a:b}\right)^{-1}\mathbf{M}^{p\,T}_{a:b}\mathbf{X}_{a:b}.$$

Since $\mathrm{P}_{\mathbf{M}^p_{a:b}}\mathbf{X}_{a-(p+1):b-(p+1)}$ is in the column space of $\mathbf{M}^p_{a:b}$,

$$\begin{aligned}\hat{\phi}^{p+1,1}_{a:b}\mathbf{X}_{a-1:b-1}+ \cdots +\hat{\phi}^{p+1,p}_{a:b}\mathbf{X}_{a-p:b-p} + \hat{\phi}^{p+1,p+1}_{a:b}\mathrm{P}_{\mathbf{M}^p_{a:b}}\mathbf{X}_{a-(p+1):b-(p+1)}& \\ = \quad \hat{\phi}^{p,1}_{a:b}\mathbf{X}_{a-1:b-1} + \cdots + \hat{\phi}^{p,p}_{a:b}\mathbf{X}_{a-p:b-p},&\end{aligned}$$

and thus

$$\begin{aligned}\tilde{\sigma}^2_{\lambda:\lambda'}(p) - \tilde{\sigma}^2_{\lambda:\lambda'}(p+1) \quad &= \quad \frac{1}{(\lambda'-\lambda)n}\left\|\mathbf{X}_{a:b} - \mathbf{M}^p_{a:b}\hat{\boldsymbol{\phi}}^p_{a:b}\right\|^2 \\ &\quad - \frac{1}{(\lambda'-\lambda)n}\left\|\mathbf{X}_{a:b} - \mathbf{M}^{p+1}_{a:b}\hat{\boldsymbol{\phi}}^{p+1}_{a:b}\right\|^2 \\ &= \quad \frac{1}{(\lambda'-\lambda)n}\left(\hat{\phi}^{p+1,p+1}_{a:b}\right)^2\left\|\mathrm{P}^{\perp}_{\mathbf{M}^p_{a:b}}(\mathbf{X}_{a-(p+1):b-(p+1)})\right\|^2.\end{aligned}$$

Note that

$$\begin{aligned}&\frac{1}{(\lambda'-\lambda)n}\left\|\mathrm{P}^{\perp}_{\mathbf{M}^p_{a:b}}\mathbf{X}_{a-(p+1):b-(p+1)}\right\|^2 \\ &\quad = \quad \frac{1}{(\lambda'-\lambda)n}\left\|\mathbf{X}_{a-(p+1):b-(p+1)}\right\|^2 - \mathbf{u}^T\mathbf{V}^{-1}\mathbf{u},\end{aligned}$$

where the components of $\boldsymbol{u} = (u_1, \ldots, u_p)$ and $\boldsymbol{V} = \{v_{ij}\}^p_{i,j=1}$ are given by

$$\begin{aligned}u_i \quad &= \quad \frac{1}{(\lambda'-\lambda)n}\sum_{t=a}^{b} X_{t-p-1}X_{t-i} \text{ and} \\ v_{ij} \quad &= \quad \frac{1}{(\lambda'-\lambda)n}\sum_{t=a}^{b} X_{t-i}X_{t-j}.\end{aligned}$$

From this observation, it is readily seen that uniformly in $0 \le \lambda < \lambda + \delta \le \lambda' \le 1$,

$$\lim_{n\to\infty} \frac{1}{(\lambda'-\lambda)n} \left\| \mathrm{P}^{\perp}_{\mathbf{M}^p_{a:b}} (\mathbf{X}_{a-(p+1):b-(p+1)}) \right\|^2 = c_p > 0 \text{ a.s.},$$

where

$$c_p = \min_{a_1,\ldots,a_p} \mathbb{E}[X_{t-p-1} - (a_1 X_{t-1} + \cdots + a_p X_{t-p})]^2.$$

Therefore, if we can show that $\left(\hat{\phi}^{p+1,p+1}_{a:b}\right)^2 = O(\log\log n/n)$ uniformly in $0 \le \lambda < \lambda+\delta \le \lambda' \le 1$, it follows that $\tilde{\sigma}^2_{\lambda:\lambda'}(p) - \tilde{\sigma}^2_{\lambda:\lambda'}(p+1) = O(\log\log n/n)$ uniformly in $0 \le \lambda < \lambda + \delta \le \lambda' \le 1$ and (A.7) holds.

We would like to apply the FLIL on the $(p+1)$st component of $\hat{\boldsymbol{\phi}}^{p+1}_{a:b}$, denoted by $\hat{\phi}^{p+1,p+1}_{a:b}$ (cf. (A.8) and (A.9)). A standard argument yields

$$\begin{aligned} \hat{\phi}^{p+1,p+1}_{a:b} &= \frac{\left\langle \mathbf{X}_{a:b}, \mathbf{X}_{a-(p+1):b-(p+1)} - \mathrm{P}_{\mathbf{M}^p_{a:b}} \mathbf{X}_{a-(p+1):b-(p+1)} \right\rangle}{\left\| \mathbf{X}_{a-(p+1):b-(p+1)} - \mathrm{P}_{\mathbf{M}^p_{a:b}} \mathbf{X}_{a-(p+1):b-(p+1)} \right\|^2} \\ \text{(A.11)} \qquad &= \frac{\left\langle \mathbf{X}_{a:b}, \mathbf{X}_{a-(p+1):b-(p+1)} \right\rangle - \sum_{j=1}^p \tilde{\phi}^{p,j}_{a:b} \left\langle \mathbf{X}_{a:b}, \mathbf{X}_{a-j,b-j} \right\rangle}{\left\| \mathbf{X}_{a-(p+1):b-(p+1)} - \sum_{j=1}^p \tilde{\phi}^{p,j}_{a:b} \mathbf{X}_{a-j,b-j} \right\|^2}, \end{aligned}$$

where $\mathrm{P}_{\mathbf{M}^p_{a:b}} \mathbf{X}_{a-(p+1):b-(p+1)} = \mathbf{M}^p_{a:b} \tilde{\boldsymbol{\phi}}^p_{a:b} = \sum_{j=1}^p \tilde{\phi}^{p,j}_{a:b} \mathbf{X}_{a-j,b-j}$ (cf. (A.10)).

Define $s_{ij}(r) = \sum_{t=1}^r X_{t-i} X_{t-j}$ for $r = 1, 2, \ldots$. Using the projection theorem, we can show that (A.11) is a function of the $s_{ij}(.)$'s, and thus we can apply the FLIL. That is,

$$\text{(A.12)} \qquad \hat{\phi}^{p+1,p+1}_{a:b} = h\left(\frac{1}{b-a+1} (s_{ij}(b) - s_{ij}(a-1)) : i,j = 0,1,\ldots,p+1 \right).$$

Using a first order Taylor expansion about $\boldsymbol{\gamma} = (\gamma(|i-j|) : i,j = 0,1,\ldots,p+1)$ on $\hat{\phi}^{p+1,p+1}_{a:b}$,

$$\text{(A.13)} \qquad \hat{\phi}^{p+1,p+1}_{a:b} = h(\boldsymbol{\gamma}) + \nabla h(\boldsymbol{\gamma}*) \left(\frac{1}{b-a+1} \boldsymbol{s}_{a:b} - \boldsymbol{\gamma} \right),$$

where $\boldsymbol{s}_{a:b} = (s_{ij}(b) - s_{ij}(a-1) : i,j = 0,1,\ldots,p+1)$ and $\boldsymbol{\gamma}*$ is between $\boldsymbol{\gamma}$ and $\frac{1}{b-a+1}\boldsymbol{s}_{a:b}$. By the FLIL, for $i,j = 0,1,\ldots,p+1$, uniformly in $0 \le \lambda <$

$\lambda + \delta \le \lambda' \le 1$,

$$\begin{aligned}
&\frac{1}{b-a+1}\left[s_{ij}(b) - s_{ij}(a-1)\right] - \gamma(|i-j|) \\
&\quad = \frac{1}{\lambda' - \lambda}\left(\frac{s_{ij}b - b\gamma(|i-j|)}{n} - \frac{s_{ij}(a-1) - (a-1)\gamma(|i-j|)}{n}\right) \\
\text{(A.14)}\quad &\quad = O\left(\sqrt{\frac{1}{n}\log\log n}\right) \text{ a.s.}
\end{aligned}$$

Finally, to see $h(\boldsymbol{\gamma}) = 0$ (which implies by (A.13) and (A.14) that $(\hat{\phi}_{a:b}^{p+1,p+1})^2 = O(\log\log n/n)$ a.s. uniformly in $0 \le \lambda < \lambda + \delta \le \lambda' \le 1$), let

(A.15)

$$\hat{\boldsymbol{\phi}}^* = (\phi_1^*, \dots, \phi_{p+1}^*) := \underset{a_1,\dots,a_{p+1}}{\arg\min}\ \mathbb{E}[X_t - (a_1 X_{t-1} + \cdots + a_{p+1}X_{t-p-1})]^2.$$

Since the stationary AR(p) process $\{X_t\}$ satisfies (3.1) with $\mu = 0$, we have $\phi_i^* = \phi_i, i = 1, \dots, p$ and $\phi_{p+1}^* = 0$. By (A.8), (A.9) and (A.15), $\hat{\boldsymbol{\phi}}_{a:b}^{p+1}$ converges a.s. to $\boldsymbol{\phi}^*$. In particular, $\lim_{n\to\infty} \hat{\phi}_{a:b}^{p+1,p+1} = \phi_{p+1}^* = 0$ a.s. By(A.12), $h(\boldsymbol{\gamma}) = 0$ since $(s_{ij}(b) - s_{ij}(a-1))/(b-a+1) \to \gamma(|i-j|)$ a.s. for $i, j = 0, 1, \dots, p+1$. This completes the proof.

□

We proved Theorem 3.2 for the simple case where the true number of change-points is $m_0 = 1$, each segment follows an autoregressive model with mean zero and order 1, and we fit AR(1) models to the data. The following extends the proof to the general case where the true process follows the piecewise AR model (2.1) and we compare the MDL for the following two models:

***Model* 1′:** Fit a piecewise autoregressive model to the data set with m_0 relative change-points, $\boldsymbol{\lambda} \in A_{m_0}^{\delta}$, where the AR orders, $p_1^0, \dots, p_{m_0+1}^0$, are known.

***Model* 2′:** Fit a piecewise autoregressive model to the data set with $m_0 + s$ relative change-points, $\boldsymbol{\alpha} \in A_{m_0+s}^{\delta}$, where s is a positive integer. Estimate the autoregressive orders from the data, and denote these orders by $\hat{p}_1, \dots, \hat{p}_{m_0+s+1}$.

Proof of Theorem 3.2. It suffices to show that

$$\begin{aligned}
\lim_{n\to\infty} P\Big(&\mathrm{MDL}(m_0 + s, \hat{\boldsymbol{\alpha}}; \hat{p}_1, \dots, \hat{p}_{m_0+s+1}) \\
&> \mathrm{MDL}(m_0, \hat{\boldsymbol{\lambda}}; p_1^0, \dots, p_{m_0+1}^0)\Big) = 1
\end{aligned}$$

since this implies

$$\lim_{n\to\infty} P\biggl(\min_{\substack{m\neq m_0\\ 0\le m\le M}}\{\mathrm{MDL}(m,\hat{\boldsymbol{\alpha}};\hat{p}_1,\dots,\hat{p}_{m+1})\} \\ > \mathrm{MDL}(m_0,\hat{\boldsymbol{\lambda}};p_1^0,\dots,p_{m_0+1}^0)\biggr) = 1,$$

where M is a prespecified upper bound, and the result follows. Equivalently, as the simple case outlined previously, we can show that

$$\lim_{n\to\infty} P\Bigl(\mathrm{MDL}(m_0+s,\boldsymbol{\alpha};\hat{p}_1,\dots,\hat{p}_{m_0+s+1}) \\ > \mathrm{MDL}(m_0,\boldsymbol{\lambda}^0;p_1^0,\dots,p_{m_0+1}^0)\ \forall\ \boldsymbol{\alpha}\in A_{m_0+s}^{\delta}\Bigr) = 1 \tag{A.16}$$

where $A_{m_0+s}^{\delta}$ is defined in (2.2).

Consider the difference

$$\begin{aligned} &\frac{2}{n}[\mathrm{MDL}(m_0+s,\boldsymbol{\alpha};\hat{p}_1,\dots,\hat{p}_{m_0+s+1}) - \mathrm{MDL}(m_0,\boldsymbol{\lambda}^0;p_1^0,\dots,p_{m_0+1}^0)] \\ &\quad = O\left(\frac{\log n}{n}\right) + \sum_{j=1}^{m_0+s+1}(\alpha_j-\alpha_{j-1})\log\hat{\sigma}_{j,m_0+s}^2 \\ &\qquad - \sum_{k=1}^{m_0+1}(\lambda_k^0-\lambda_{k-1}^0)\log\hat{\sigma}_{k,m_0}^2. \end{aligned} \tag{A.17}$$

where $\hat{\sigma}_{j,m_0+s}^2$ is the conditional maximum likelihood variance estimate for the jth fitted segment when fitting a piecewise AR model with m_0+s change-points and estimated AR orders, and $\hat{\sigma}_{k,m_0}^2$ is the conditional maximum likelihood variance estimate when fitting an AR(p_k^0) model to the kth of m_0+1 fitted segments. Since for large n, the estimated AR orders are greater than or equal to the true AR orders, the $O(\log n/n)$ penalty term in (A.17) is strictly positive for large n. Therefore, it suffices to show that for all $\boldsymbol{\alpha}\in A_{m_0+s}^{\delta}$, the term

$$\sum_{j=1}^{m_0+s+1}(\alpha_j-\alpha_{j-1})\log\hat{\sigma}_{j,m_0+s}^2 - \sum_{k=1}^{m_0+1}(\lambda_k^0-\lambda_{k-1}^0)\log\hat{\sigma}_{k,m_0}^2 \tag{A.18}$$

is either positive or is of smaller order than $\log n/n$. This will imply (A.16), and the theorem follows.

We will show (A.18) by combining the two summations into one sum over the true segments, and applying the arguments demonstrated previously to each term within the sum. In other words, rather than summing the terms $(\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+s}$ over the indices of the fitted change-point locations, $\alpha_1, \ldots, \alpha_{m_0+s}$, we will break up each term and sum over the indices of the true change-point locations, $\lambda^0_1, \ldots, \lambda^0_{m_0}$. Then we will examine each summand individually.

We first give the argument for the case where the true process has mean zero in every segment, and then describe extensions to the non-zero mean case at the end. First focus on the jth fitted segment, (α_{j-1}, α_j). If this segment does not contain any true change-points, there is no need to partition the interval further. Suppose, however, the segment contains 1 true change-point, denoted by $\lambda^0_{k(j)}$, where $k(j)$ denotes the index of the true change-point contained in the jth fitted segment. In the case where the fitted segment (α_{j-1}, α_j) contains one true change-point, this segment corresponds to (α_1, α_2) in the simple case demonstrated previously. In other words, we need only consider the cases

(i) $(\log \log n)^2 / \log n < \lambda^0_{k(j)} - \alpha_{j-1}$,
(ii) $\lambda^0_{k(j)} - \alpha_{j-1} < N/n$ for some positive integer N, or
(iii) $N/n < \lambda^0_{k(j)} - \alpha_{j-1} < (\log \log n)^2 / \log n$ for some positive integer N.

Then, using the same arguments as in the simple case, but applying Lemma A.1 rather than Lemma 3.1 to account for the estimated AR orders, for case (i),

$$\begin{aligned}(\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+1} \geq (\lambda^0_{k(j)} - \alpha_{j-1}) \log \left(\frac{\mathrm{RSS}_{j,1}}{(\lambda^0_{k(j)} - \alpha_{j-1})n} \right) \\ + (\alpha_j - \lambda^0_{k(j)}) \log \left(\frac{\mathrm{RSS}_{j,2}}{(\alpha_j - \lambda^0_{k(j)})n} \right),\end{aligned}$$

where

$$\mathrm{RSS}_{j,1} := \sum_{t=\alpha_{j-1}n+1}^{\lambda^0_{k(j)}n} \left(X_t - \hat{\phi}^{\hat{p}_j 1}_{\alpha_{j-1}n+1:\lambda^0_{k(j)}n} X_{t-1} - \cdots - \hat{\phi}^{\hat{p}_j \hat{p}_j}_{\alpha_{j-1}n+1:\lambda^0_{k(j)}n} X_{t-\hat{p}_j} \right)^2$$

and

$$\mathrm{RSS}_{j,2} := \sum_{t=\lambda^0_{k(j)}n+1}^{\alpha_j n} \left(X_t - \hat{\phi}^{\hat{p}_j 1}_{\lambda^0_{k(j)}n+1:\alpha_j n} X_{t-1} - \cdots - \hat{\phi}^{\hat{p}_j \hat{p}_j}_{\lambda^0_{k(j)}n+1:\alpha_j n} X_{t-\hat{p}_j} \right)^2$$

are the residual sum of squares over the ith sub-segment of the jth fitted segment, but using AR coefficients estimated only within that sub-segment rather than using the entire jth fitted segment. For case (ii),

$$(\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+1} \geq O_p\left(\frac{1}{n}\right) + (\alpha_j - \lambda^0_{k(j)}) \log \left(\frac{\mathrm{RSS}_{j,2}}{(\alpha_j - \lambda^0_{k(j)})n}\right),$$

and for case (iii),

$$\begin{aligned}(\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+1} \geq (\lambda^0_{k(j)} - \alpha_{j-1}) \log \left(\sigma^2_{k(j)} + \eta\right)& \\ + (\alpha_j - \lambda^0_{k(j)}) \log \left(\frac{\mathrm{RSS}_{j,2}}{(\alpha_j - \lambda^0_{k(j)})n}\right)&,\end{aligned}$$

for some small $\eta > 0$ where $\sigma^2_{k(j)}$ is the true variance in the $k(j)$th true segment.

Suppose now that (α_{j-1}, α_j) contains more than one true change-point. In the simple case, since there was only one true change-point, we only needed to consider when either α_1 was close to λ, or when α_2 was close to λ, but both α_1 and α_2 couldn't be close to λ simultaneously. Now, both α_{j-1} and α_j could potentially be close to a true fitted changepoint. We will address this case by adding a fictitious fitted change-point at the center of each true segment completely contained within the jth fitted segment. This can only reduce the log-likelihood term of the fitted MDL,

$$\sum_{j=1}^{m_0+s+1} (\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+s},$$

but we can show that even with this reduction, the MDL of the fitted model is still greater than the MDL corresponding to the true model. With the addition of the fictitious fitted change-points, each fitted segment will then contain either no true change-points or one true change-point.

For each true segment that does not contain a fitted change-point, add a fictitious fitted change-point at the midpoint of this segment. Once we have added the necessary fictitious fitted change-points, re-label the fitted change-points as α'_1, α'_2,..., α'_{m_0+b} where $b \geq 1$. It follows that

$$\begin{aligned}&\sum_{j=1}^{m_0+s+1} (\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+s} - \sum_{k=1}^{m_0+1} (\lambda^0_k - \lambda^0_{k-1}) \log \hat{\sigma}^2_{k,m_0} \\ \geq\ &\sum_{j=1}^{m_0+b} (\alpha'_j - \alpha'_{j-1}) \log \hat{\sigma}^2_{j,m_0+b} - \sum_{k=1}^{m_0+1} (\lambda^0_k - \lambda^0_{k-1}) \log \hat{\sigma}^2_{k,m_0}\end{aligned}$$

$$\geq \sum_{j=1}^{m_0+b}\left[A_j + \left(\alpha'_j - \lambda^0_{k(j)}\right)\log\left(\frac{\mathrm{RSS}_{j,2}}{(\alpha'_j - \lambda^0_{k(j)})n}\right)\right] - \sum_{k=1}^{m_0+1}(\lambda^0_k - \lambda^0_{k-1})\log\hat{\sigma}^2_{k,m_0}, \tag{A.19}$$

where $\hat{\sigma}^2_{j,m_0+b}$ is the estimated variance within the re-labeled jth fitted segment and A_j is

(i) $(\lambda^0_{k(j)} - \alpha'_{j-1})\log(\mathrm{RSS}_{j,1}/((\lambda^0_{k(j)} - \alpha'_{j-1})n))$,
(ii) $O_p(1/n)$, or
(iii) $(\lambda^0_{k(j)} - \alpha'_{j-1})\log(\sigma^2_{k(j)} + \eta)$ for some $\eta > 0$,

depending on how close α'_{j-1} is to $\lambda^0_{k(j)}$. Note that since $\alpha_0 := 0$, if the first fitted segment contains one true change-point, then A_1 must equal $\lambda^0_1\log(\mathrm{RSS}_{1,1}/(\lambda^0_1 n))$. Note also that if the jth fitted segment does not contain any true change-points, then the term in brackets in (A.19) is simply

$$(\alpha'_j - \alpha'_{j-1})\log\hat{\sigma}^2_{j,m_0+b}.$$

The next step is to combine the two sums in (A.19) into one sum, indexed over the true change-points. In order to make this step, we need some further notation for the re-labeled fitted change-points contained within the kth true segment. Consider the kth true segment, $(\lambda^0_{k-1}, \lambda^0_k)$. This segment must contain at least one re-labeled fitted change-point, so we can break the segment into sub-segments, with the partition being determined by the re-labeled fitted change-points contained in the kth true segment. Let r_k+1, $0 \leq r_k \leq m_0$, be the number of fitted change-points contained in $(\lambda^0_{k-1}, \lambda^0_k)$. Then we can partition $(\lambda^0_{k-1}, \lambda^0_k]$ into the r_k+2 intervals $(\lambda^0_{k-1}, \alpha'_{j(k)})$, $(\alpha'_{j(k)}, \alpha'_{j(k)+1})$, $\ldots$, $(\alpha'_{j(k)+r_k}, \lambda^0_k]$, where $j(k)$ denotes the index of the first re-labeled fitted change-point contained in the kth true segment. Then we can re-index

$$\sum_{j=1}^{m_0+b}(\alpha'_j - \alpha'_{j-1})\log\hat{\sigma}^2_{j,m_0+b}$$

according to the true change-points as follows:

$$\begin{aligned}
&\sum_{j=1}^{m_0+b} (\alpha'_j - \alpha'_{j-1}) \log \hat{\sigma}^2_{j,m_0+b} \\
&\quad \ge \sum_{j=1}^{m_0+b} \left[A_j + \left(\alpha'_j - \lambda^0_{k(j)}\right) \log\left(\frac{\mathrm{RSS}_{j,2}}{(\alpha'_j - \lambda^0_{k(j)})n} \right) \right] \\
&\quad = \sum_{k=1}^{m_0+1} \left[(\alpha'_{j(k)} - \lambda^0_{k-1}) \log\left(\frac{\mathrm{RSS}_{j(k),2}}{(\alpha'_{j(k)} - \lambda^0_{k-1})n} \right) \right. \\
&\qquad\qquad \left. + \sum_{i=1}^{r_k} (\alpha'_{j(k)+i} - \alpha'_{j(k)+i-1}) \log \hat{\sigma}^2_{j(k)+i,m_0+b} + A_{j(k)+r_k+1} \right],
\end{aligned}$$

where for $k = 1, \dots, m_0 + 1$ and $i = 1, \dots, r_k$, $\hat{\sigma}^2_{j(k)+i,m_0+b}$ is the conditional maximum likelihood estimate of the variance when fitting an $\mathrm{AR}(\hat{p}_{j(k)+i})$ model to the $(j(k)+i)$th re-labeled fitted segment, and $A_{j(k)+r_k+1}$ is defined as before for the $(j(k) + r_k + 1)$st re-labeled fitted segment. Thus, (A.19) becomes

$$\begin{aligned}
&\sum_{j=1}^{m_0+b} (\alpha'_j - \alpha'_{j-1}) \log \hat{\sigma}^2_{j,m_0+b} - \sum_{k=1}^{m_0+1} (\lambda^0_k - \lambda^0_{k-1}) \log \hat{\sigma}^2_{k,m_0} \\
&\quad \ge \sum_{k=1}^{m_0+1} \left[(\alpha'_{j(k)} - \lambda^0_{k-1}) \log\left(\frac{\mathrm{RSS}_{j(k),2}}{(\alpha'_{j(k)} - \lambda^0_{k-1})n} \right) \right. \\
&\qquad\qquad + \sum_{i=1}^{r_k} (\alpha'_{j(k)+i} - \alpha'_{j(k)+i-1}) \log \hat{\sigma}^2_{j(k)+i,m_0+b} \\
&\qquad\qquad \left. + A_{j(k)+r_k+1} - (\lambda^0_k - \lambda^0_{k-1}) \log \hat{\sigma}^2_{k,m_0} \right].
\end{aligned} \tag{A.20}$$

Now, within each summand of (A.20), we can apply the same arguments as in the simple case using Lemma A.1 rather than Lemma 3.1, and (A.18) follows.

For the case where the means of each segment are unknown, we can follow an argument similar to that used in the simple case of one true change-point as demonstrated previously. When calculating the estimated white noise variances, rather than minimizing the quantity $\sum (X_t - a_1 X_{t-1} - \dots - a_{\hat{p}} X_{t-\hat{p}})^2$, the estimates minimize the quantity $\sum (X_t - a_0 - a_1 X_{t-1} - \dots - a_{\hat{p}} X_{t-\hat{p}})^2$. Since Lemmas 3.1 and A.1 hold for non-zero means, the result follows. □

A.2. Yule-Walker. In this section, we provide further details for the proof of Theorem 3.1.

Proof of Theorem 3.1. The MDL for an AR(1) process with no change-point is given by

$$\mathrm{MDL}(0;1) = \frac{5}{2}\log n + \frac{n}{2}\log(2\pi\hat{\sigma}^2),$$

while the MDL when calculated under one change-point for an AR(1) is

$$\begin{aligned}\mathrm{MDL}(1,\hat{\lambda};1,1) &= \min_{\delta\le\lambda\le 1-\delta}\mathrm{MDL}(1,\lambda;1,1)\\ &= \min_{\delta\le\lambda\le 1-\delta}\Big\{5\log n + \frac{3}{2}\big(\log(\lambda)+\log(1-\lambda)\big)\\ &+ \frac{n}{2}\Big(\lambda\log(2\pi\hat{\sigma}_1^2(\lambda n)) + (1-\lambda)\log(2\pi\hat{\sigma}_2^2(\lambda n))\Big)\Big\},\end{aligned}\tag{A.21}$$

where now $\hat{\sigma}^2$ is the Yule-Walker variance estimate of the entire sequence, $\hat{\sigma}_1^2(\lambda n)$ is the Yule-Walker variance estimate from observations $1,\dots,\lambda n$, and $\hat{\sigma}_2^2(\lambda n)$ is the Yule-Walker variance estimate from observations $\lambda n+1,\dots,n$. The change-point estimate $\hat{\lambda}$ is obtained by minimizing $\mathrm{MDL}(1,\lambda;1,1)$ with respect to $\lambda\in A_1^\delta$, A_1^δ defined in (2.2).

Let $\tau=\lambda n$. Since the mean μ of the process is zero, define

$$\tilde{\sigma}^2 := \frac{1}{n}\sum_{t=1}^n X_t^2 - \frac{\left(\frac{1}{n}\sum_{t=1}^{n-1}X_tX_{t+1}\right)^2}{\frac{1}{n}\sum_{t=1}^n X_t^2},\tag{A.22}$$

$$\tilde{\sigma}_1^2(\tau) := \frac{1}{\tau}\sum_{t=1}^\tau X_t^2 - \frac{\left(\frac{1}{\tau}\sum_{t=1}^{\tau-1}X_tX_{t+1}\right)^2}{\frac{1}{\tau}\sum_{t=1}^\tau X_t^2},\tag{A.23}$$

$$\tilde{\sigma}_2^2(\tau) := \frac{1}{n-\tau}\sum_{t=\tau+1}^n X_t^2 - \frac{\left(\frac{1}{n-\tau}\sum_{t=\tau+1}^{n-1}X_tX_{t+1}\right)^2}{\frac{1}{n-\tau}\sum_{t=\tau+1}^n X_t^2},\tag{A.24}$$

which are μ-centered versions of $\hat{\sigma}^2$, $\hat{\sigma}_1^2(\tau)$ and $\hat{\sigma}_2^2(\tau)$, respectively. Since by the FLIL, we have

$$\begin{aligned}\hat{\sigma}^2 - \tilde{\sigma}^2 &= O(\log\log n/n)\text{ a.s.},\\ \max_{\delta\le\tau/n\le 1-\delta}\left|\hat{\sigma}_i^2(\tau)-\tilde{\sigma}_i^2(\tau)\right| &= O(\log\log n/n)\text{ a.s.},\ i=1,2,\end{aligned}$$

it suffices to prove that for every $C > 0$,

$$\liminf_{n\to\infty} P\left(\log\tilde{\sigma}^2 - \min_{\delta\le\frac{\tau}{n}\le 1-\delta}\left\{\frac{\tau}{n}\log\tilde{\sigma}_1^2(\tau) + \frac{n-\tau}{n}\log\tilde{\sigma}_2^2(\tau)\right\} > \frac{C\log n}{n}\right) > 0. \tag{A.25}$$

Performing a Taylor expansion about σ^2 on the log of (A.22), we obtain

$$\begin{aligned}\log\tilde{\sigma}^2 &= \log\left[\frac{\sigma^2}{1-\phi^2} + \frac{1}{n}\sum_{t=1}^{n}\left(X_t^2 - \frac{\sigma^2}{1-\phi^2}\right)\right. \\ &\qquad \left. - \frac{\left(\frac{\phi\sigma^2}{1-\phi^2} + \frac{1}{n}\sum_{t=0}^{n-1}\left(X_tX_{t+1} - \frac{\phi\sigma^2}{1-\phi^2}\right) - \frac{X_0X_1}{n}\right)^2}{\frac{\sigma^2}{1-\phi^2} + \frac{1}{n}\sum_{t=1}^{n}\left(X_t^2 - \frac{\sigma^2}{1-\phi^2}\right)}\right] \\ &= \log\sigma^2 + \frac{1+\phi^2}{\sigma^2}\frac{1}{n}\sum_{t=1}^{n}\left(X_t^2 - \frac{\sigma^2}{1-\phi^2}\right) \\ &\qquad - \frac{2\phi}{\sigma^2}\left[\frac{1}{n}\sum_{t=0}^{n-1}\left(X_tX_{t+1} - \frac{\phi\sigma^2}{1-\phi^2}\right) - \frac{X_0X_1}{n}\right] + O\left(\frac{\log\log n}{n}\right).\end{aligned}$$

Let

$$S_n := \frac{1+\phi^2}{\sigma^2}\sum_{t=1}^{n}\left(X_t^2 - \frac{\sigma^2}{1-\phi^2}\right) - \frac{2\phi}{\sigma^2}\sum_{t=0}^{n-1}\left(X_tX_{t+1} - \frac{\phi\sigma^2}{1-\phi^2}\right),$$

so that

$$\log\tilde{\sigma}^2 = \log\sigma^2 + \frac{1}{n}S_n + \frac{2\phi}{\sigma^2}\frac{X_0X_1}{n} + O\left(\frac{\log\log n}{n}\right).$$

Similarly, for (A.23) and (A.24),

$$\max_{\delta\le\frac{\tau}{n}\le 1-\delta}\left|\log\tilde{\sigma}_1^2(\tau) \ - \ \left(\log\sigma^2 + \frac{1}{\tau}S_\tau + \frac{2\phi}{\sigma^2}\frac{X_0X_1}{\tau}\right)\right|$$

and

$$\max_{\delta\le\frac{\tau}{n}\le 1-\delta}\left|\log\tilde{\sigma}_2^2(\tau) \ - \ \left(\log\sigma^2 + \frac{1}{n-\tau}(S_n - S_\tau) + \frac{2\phi}{\sigma^2}\frac{X_\tau X_{\tau+1}}{n-\tau}\right)\right|$$

are both $O\left(\frac{\log\log n}{n}\right)$. It follows that

$$\begin{aligned}\max_{\delta\le\frac{\tau}{n}\le 1-\delta}&\left|\log\tilde{\sigma}^2 - \left(\frac{\tau}{n}\log\tilde{\sigma}_1^2(\tau) + \frac{n-\tau}{n}\log\tilde{\sigma}_2^2(\tau)\right)\right. \\ &\left. + \frac{2\phi}{\sigma^2}\frac{X_\tau X_{\tau+1}}{n}\right| = O\left(\frac{\log\log n}{n}\right).\end{aligned} \tag{A.26}$$

Under the assumptions for the density f_ϵ of the noise, it is clear that

$$\lim_{n\to\infty} P\left(\max\left\{\epsilon_t : \delta \le \frac{t}{n} \le 1-\delta\right\} > \frac{1}{2c}\log n\right) = 1. \tag{A.27}$$

Let $T := \max\left\{[n\delta]+1 \le t \le n(1-\delta) : \epsilon_t > \frac{1}{2c}\log n\right\}$ if the specified set is non-empty, and define $T := [n\delta]+1$ otherwise. Note by (A.27) that $P(\epsilon_T > \frac{1}{2c}\log n) \to 1$ as $n \to \infty$. Moreover, X_{T-1} and ϵ_T are independent, and X_{T-1} has the same (stationary) distribution as X_0 since T is a stopping time in reverse time, which has an everywhere-positive density function by condition (i). Therefore, from (A.26), we have

$$\begin{aligned}
&\log\tilde\sigma^2 - \left(\frac{T-1}{n}\log\tilde\sigma_1^2(T-1) + \frac{n-T+1}{n}\log\tilde\sigma_2^2(T-1)\right)\\
&= \frac{-2\phi}{\sigma^2}\frac{X_{T-1}X_T}{n} + O\left(\frac{\log\log n}{n}\right)\\
&= \frac{-2\phi}{\sigma^2}\frac{(\phi X_{T-1}^2 + \sigma X_{T-1}\epsilon_T)}{n} + O\left(\frac{\log\log n}{n}\right)\\
&= \left(\frac{-2\phi}{\sigma}X_{T-1}\frac{\epsilon_T}{\log n}\right)\frac{\log n}{n} - \frac{2\phi^2}{\sigma^2}\frac{X_{T-1}^2}{n} + O\left(\frac{\log\log n}{n}\right)\\
&= \left(\frac{-2\phi}{\sigma}X_{T-1}\frac{\epsilon_T}{\log n}\right)\frac{\log n}{n} - O_p(\frac{1}{n}) + O\left(\frac{\log\log n}{n}\right). \qquad \text{(A.28)}
\end{aligned}$$

For every $C > 0$, we have $P\left(\frac{-2\phi}{\sigma}X_{T-1} > C\right) > 0$, implying (A.25) for every $C > 0$. This completes the proof.

□

REFERENCES

[1] Athreya, K. B. and Pantula, S. G. (1986). Mixing properties of Harris chains and autoregressive processes. *Journal of Applied Probability*, **23** 880–892.

[2] Athreya, K. B. and Pantula, S. G. (1986). A note on strong mixing of ARMA processes. *Statistics & Probability Letters*, **4** 187–190.

[3] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* 2nd ed. Springer-Verlag.

[4] Chen, J. and Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, **92** 739–747.

[5] Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, **35** 999–1018.

[6] Csörgó, M. and Horvth, L. (1997). *Limit Theorems in Change-Point Analysis.* Wiley.

[7] Davis, R. A., Huang, D. and Yao, Y. C. (1995). Testing for a change in the parameter values and order of an autoregressive model. *The Annals of Statistics*, **23** 282–304.

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

[9] Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, **16** 203–213.

[10] Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, **37** 323–341.

[11] Kokoszka, P. and Leipus, R. (2000). Change-point estimation in ARCH models. *Bernoulli*, **6** 513–539.

[12] Kühn, C. (2001). An estimator of the number of change points based on weak invariance principle. *Statistics & Probability Letters*, **51** 189–196.

[13] Lee, C. B. (1996). Nonparametric multiple change-point estimators. *Statistics & Probability Letters*, **27** 295–304.

[14] Lee, C. B. (1997). Estimating the number of change points in exponential families distributions. *Scandinavian Journal of Statistics, Theory and Applications*, **24** 201–210.

[15] Ling, S. (2007). Testing for change-points in time series models and limiting theorems for ned sequences. *Annals of Statistics*, **35** 1213–1237.

[16] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer.

[17] Perreault, L., Bernier, J., Bobée, B. and Parent, E. (2000). Bayesian changepoint analysis in hydrometeorological time series. part 1. The normal model revisited. *Journal of Hydrology*, **235** 221–241.

[18] Perreault, L., Bernier, J., Bobée, B. and Parent, E. (2000). Bayesian changepoint analysis in hydrometeorological time series. part 2. Comparison of change-point models and forecasting. *Journal of Hydrology*, **235** 242–263.

[19] Rio, E. (1995). The functional law of the iterated logarithm for stationary strongly mixing sequences. *The Annals of Probability*, **23** 1188–1203.

[20] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company.

[21] Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, **43** 159–178.

[22] Sullivan, J. H. (2002). Estimating the locations of multiple change points in the mean. *Computational Statistics*, **17** 289–296.

[23] Yao, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics*, **12** 1434–1447.

[24] Yao, Y. C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, **6** 181–189.

[25] Yao, Y. C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhya: The Indian Journal of Statistics*, **51** 370–381.

[26] Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, **63** 22–32.

R. A. Davis
Department of Statistics
Columbia University
1255 Amsterdam Avenue, MC 4690
Room 1026 SSW
New York, NY 10027
E-mail: rdavis@stat.columbia.edu

S. Hancock
Department of Mathematics
Reed College
3203 SE Woodstock Blvd
Portland, OR 97202
E-mail: shancock@reed.edu

Y.-C. Yao
Institute of Statistical Science
Academia Sinica
128 Academia Rd. Sec. 2
Taipei 115, Taiwan
E-mail: yao@stat.sinica.edu.tw