On Consistency of Minimum Description Length Model Selection for Piecewise Autoregressions

Richard A. Davis[a,1,*], Stacey A. Hancock[b,2], Yi-Ching Yao[c,3]

[a] *Department of Statistics, 1255 Amsterdam Avenue, MC 4690, Room 1004 SSW, Columbia University, New York, NY, 10027, United States.*
[b] *Department of Statistics, 2204 Donald Bren Hall, University of California Irvine, Irvine, CA, 92697, United States.*
[c] *Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan, ROC.*

Abstract

The Auto-PARM (Automatic Piecewise AutoRegressive Modeling) procedure, developed by Davis, Lee, and Rodriguez-Yam (2006), uses the minimum description length (MDL) principle to estimate the number and locations of structural breaks in a non-stationary time series. Consistency of this model selection procedure has been established when using conditional maximum (Gaussian) likelihood variance estimates. In contrast, the estimate of the number of change-points is inconsistent in general if Yule-Walker variance estimates are used instead. 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. In order to simplify notation and make the arguments more transparent, we only treat in detail the simple case where the time series follows an AR(p) model with no change-points.

*Corresponding author. Tel.: +1 212 851 2131; Fax: +1 212 851 2164.
Email address: rdavis@stat.columbia.edu (Richard A. Davis)

[1]Research supported in part by NSF Grant DMS-1107031.
[2]Research supported in part by the Program of Interdisciplinary Mathematics, Ecology and Statistics (PRIMES) NSF IGERT Grant DGE 0221595, administered by Colorado State University, the NSF East Asia, and Pacific Summer Institute and the National Park Service.
[3]Research supported in part by the Ministry of Science and Technology of Taiwan, ROC.

Keywords: Change-point, structural break, model selection, minimum description length, autoregressive process
JEL: C12, C13, C22, C52

1. Introduction

In recent years, there has been considerable development in non-stationary time series modeling. One prominent subject in non-stationary 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 a small number of 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 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. (2000a, b), Stephens (1994), Yao (1984), and Zhang and Siegmund (2007).

Many of the applications of structural breaks, in fact even the name, come from the econometrics community. Some of the early econometrics papers on this topic include those by Bai and Perron (1998), Bai (1999), and Bai and Perron (2003). As has been well-documented, see, for example, Diebold and Inoue (2001), there is a strong connection between long-memory and structural breaks. Diebold

and Inoue (2001) show that models that incorporate regime shifts can produce data that exhibit long-memory behavior. For financial time series, volatility often appears to exhibit long memory and as such, long-memory GARCH models (e.g., IGARCH) provide reasonable fits to the data. Mikosch and Stărică (2004) provide an explanation for this phenomenon. They show that data from a GARCH model with structural breaks possess long memory characteristics and lead to fitting IGARCH-type models that have long memory.

Though much of the early literature on change-point problems is from a hypothesis testing perspective, the pendulum in recent years has swung to considering this problem in the context of model selection. That is, instead of not including a structural break in the model unless there is strong evidence to do so, model selection approach allows for the inclusion of one or more breaks if the increase in model complexity is offset by a sufficiently *improved fit*. 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 hypothesis of no change-point. 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 Csörgó and Horváth (1997) for a comprehensive review. Recently, there have been some interesting advances on the change-point problem. Here we mention the work of Fryzlewicz (2014), who uses a promising idea based on wild binary segmentation to locate change-points.

This paper examines the procedure Auto-PARM, a method developed by Davis, Lee, and Rodriguez-Yam (2006) for estimating the number and locations of the structural breaks. This method does not assume independence nor a distribution on the data, e.g., normality, 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. Specifically, the time series does not need to follow AR models in each of the segments. The most important ingredient of Auto-PARM is its use of the minimum description length criterion in fitting the model. There are now a number of applications of Auto-PARM in the literature (e.g., see the application to segmentation in climate in Lu, Lund and Lee, 2010) and a general framework can be found in Davis, Lee, and Rodriguez-Yam (2008) and Davis and Yau (2013). A new view on piecewise AR modeling is given in Chan, Yau, and Zhang (2012), who cleverly embed this problem into a LASSO framework.

The estimated (relative) structural break locations are shown to be consistent in Davis et al. (2006) when the number of change-points is known. While the paper leaves open the issue of consistency in the case with an unknown number of change-points, it is shown in Hancock (2008) that the estimated number of change-points and the estimated AR orders are weakly consistent when using conditional maximum (Gaussian) likelihood variance estimates. (See also Davis and Yau, 2013, for related results.)

However, the estimate for the number of change-points may not even be weakly consistent if we use Yule-Walker variance estimates. It seems 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. This unexpected 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. The objective of this paper is to explain this result by contrasting the two approaches and showing the subtle difference between the conditional maximum likelihood and Yule-Walker variance estimates.

The rest of the paper is organized as follows. Section 2 reviews the Auto-PARM procedure. Section 3 discusses the functional law of the iterated logarithm applied to sample autocovariance functions, which plays an important role in Section 4 where a consistency result is established for the Auto-PARM estimate of the number of change-points using conditional maximum likelihood estimation. Section 5 contains an inconsistency result when using Yule-Walker estimation. In order to simplify notation and make the arguments more transparent, we only consider in Sections 4 and 5 the simple case where the time series follows an AR(p) model with no change-points. Section 6 contains a small simulation study on the practical differences between Yule-Walker and conditional maximum likelihood estimation.

2. Automatic piecewise autoregressive modeling and the issue of consistency

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 *A*utomatic *P*iecewise *A*uto*R*egressive *M*odeling (Auto-PARM), 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 $\tau_0 := 0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} := n$. 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, $\ldots$, 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. If $\{X_t\}$ is stationary in the first segment with mean denoted by μ_1, then the intercept $\phi_{1,0}$ equals $\mu_1(1 - \phi_{1,1} - \cdots - \phi_{1,p_1})$, and for $t \in (\tau_0, \tau_1]$, we can express the model as

$$X_t - \mu_1 = \phi_{1,1}(X_{t-1} - \mu_1) + \cdots + \phi_{1,p_1}(X_{t-p_1} - \mu_1) + \sigma_1 \epsilon_{1,t}.$$

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 (small) $\delta > 0$ be a preassigned lower bound for the relative length of each of the fitted segments, and

define

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

where $\lambda_0 := 0$ and $\lambda_{m+1} := 1$. (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 (cf. Rissanen (1989)) and standard likelihood approximations, we define the minimum description length for a piecewise autoregressive model (cf. Davis et al. (2006)) as

$$\begin{aligned} &\mathrm{MDL}(m, \lambda_1, \ldots, \lambda_m; p_1, \ldots, p_{m+1}) \\ &= \log^+ m + (m+1) \log n + \sum_{k=1}^{m+1} \log^+ p_k \\ &\quad + \sum_{k=1}^{m+1} \frac{p_k + 2}{2} \log n_k + \sum_{k=1}^{m+1} \frac{n_k}{2} \log(2\pi \hat{\sigma}_k^2), \end{aligned} \tag{2.3}$$

where $\log^+ x = \max\{\log x, 0\}$, $n_k = n(\lambda_k - \lambda_{k-1})$ 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). Two common approaches to estimating the white noise variance are conditional maximum (Gaussian) likelihood estimation (equivalent to conditional least-squares) and Yule-Walker estimation. Since Yule-Walker esti-

mates have the same asymptotic distribution as the conditional maximum likelihood estimates (see Section 8.10 in Brockwell and Davis, 1991), one would expect that substituting Yule-Walker estimates into MDL would not change the consistency result. In Sections 4 and 5, we examine the subtle difference between the two approaches that yields an inconsistency result when using the Yule-Walker variance estimates. In order to simplify notation and make the arguments more transparent, we only consider the simple case where the true process follows an AR(p) model (with $p \geq 1$ and no change-points) and MDL is used to select between

***Model 1*:** The observations follow an AR(p) model,

***Model 2*:** The observations follow a piecewise AR(p) model with $m \geq 1$ (relative) change-points, $\boldsymbol{\lambda} \in A_m^{\delta}$.

We show in Section 4 that with probability 1, Model 1 is selected for large n when using the conditional maximum likelihood estimates, and in Section 5 that as $n \to \infty$, there is a nonnegligible probability that Model 1 is not selected in favor of Model 2 when using Yule-Walker estimates.

3. Functional law of the iterated logarithm

As the consistency proof in Section 4 uses the functional law of the iterated logarithm (FLIL) on the sample ACVF and sample means of autoregressive processes, we 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 $\int_0^1 \alpha^{-1}(v)Q^2(v)dv < \infty$, 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 Rio, 1995, for proof). Therefore, assuming

$$\mathbb{E}\left(X_1^4 (\log^+ |X_1|)^2\right) < \infty \tag{3.1}$$

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 (fixed) $\delta > 0$, we have

$$\sup_{0 \le \lambda < \lambda + \delta \le \lambda' \le 1} \frac{\tilde{\gamma}_{\lambda:\lambda'}(i,j) - \gamma(|i-j|)}{\sqrt{\frac{2}{n}\log\log n}} < C \quad \text{a.s.} \tag{3.2}$$

where $C < \infty$ is a constant,

$$\tilde{\gamma}_{\lambda:\lambda'}(i,j) := \frac{1}{(\lambda'-\lambda)n} \sum_{t=\lambda n+1}^{\lambda' n} (X_{t-i} - \mu)(X_{t-j} - \mu), \tag{3.3}$$

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 (3.2) also holds if $\tilde{\gamma}_{\lambda:\lambda'}(i,j)$ is replaced by

$$\hat{\gamma}_{\lambda:\lambda'}(i,j) := \frac{1}{(\lambda'-\lambda)n} \sum_{t=\lambda n+1}^{\lambda' n} \left(X_{t-i} - \overline{X}_{\lambda n+1-i:\lambda' n-i}\right)\left(X_{t-j} - \overline{X}_{\lambda n+1-j:\lambda' n-j}\right) \tag{3.4}$$

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

$$\sup_{0 \le i,j \le p^*,\ 0 \le \lambda < \lambda + \delta \le \lambda' \le 1} |\tilde{\gamma}_{\lambda:\lambda'}(i,j) - \gamma(|i-j|)| = O(\sqrt{\log\log n/n}) \quad \text{a.s.}, \tag{3.5}$$

$$\sup_{0 \le i,j \le p^*,\ 0 \le \lambda < \lambda + \delta \le \lambda' \le 1} |\hat{\gamma}_{\lambda:\lambda'}(i,j) - \gamma(|i-j|)| = O(\sqrt{\log\log n/n}) \quad \text{a.s.}, \tag{3.6}$$

for any given upper bound $p^* < \infty$.

When applying the FLIL to the sample ACVF of a piecewise autoregressive process, we need to assume 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 (3.1).

As commented in Remark 2.1 of Davis et al. (1995), 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.

Remark 3.1. With more effort, the moment condition (3.1) may be relaxed in order for the results in the following sections to hold. When fitting an AR(p) model to observations $X_1, \ldots, X_n$, conditional maximum likelihood estimation uses a definition of the sample ACVF that includes initial values $X_{-p+1}, \ldots, X_0$. (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, the FLIL must hold for $\hat{\gamma}(h)$ when we condition on any initial values $X_{-p+1}, \ldots, X_0$. It is not difficult to show that for a causal AR(p) process $\{X_t\}$ with initial values $X_{-p+1}, \ldots, X_0$, there exists a stationary AR(p) process $\{X'_t\}$ generated by the same AR coefficients and the same noise sequence such that as $t \to \infty$,

$$X_t - X'_t = O(\rho^t) \text{ a.s.} \tag{3.7}$$

for some constant $0 < \rho < 1$ depending on the AR coefficients $\phi_1, \ldots, \phi_p$. Thus, if $\{X'_t\}$ satisfies (3.5) and (3.6), so does $\{X_t\}$ where in (3.5) and (3.6), $\gamma(h) = \lim_{t\to\infty} \text{Cov}(X_t, X_{t-h}) = \text{Cov}(X'_s, X'_{s-h})$ for all s.

Remark 3.2. Like conditional maximum likelihood estimation, Yule-Walker estimation of the parameters is also a function of the *sample autocovariance function*, but now these are defined in a slightly different fashion than above. In fact, Yule-Walker estimates use the more conventional estimates of the ACVF (see Brockwell and Davis, 1991), given by

$$\hat{\gamma}'(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \overline{X}_{1:n})(X_{t+h} - \overline{X}_{1:n}), \quad h \geq 0\,.$$

4. Conditional maximum likelihood

In this section, we assume that the true process follows an AR(p) model with $p \geq 1$

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \sigma\epsilon_t, \;\; t = 1, \ldots, n, \; (\phi_p \neq 0) \tag{4.1}$$

where the noise sequence $\{\epsilon_t\}$ is iid with mean zero and unit variance. (That is, the true process follows Model 1.) When using the MDL together with the conditional maximum likelihood variance estimates, Theorem 4.1 shows that with probability 1, Model 1 is selected over Model 2 for large n.

Theorem 4.1. *Suppose the true process $\{X_t\}$ follows the AR(p) model given in (4.1) with no change-points and initial values $X_{-p+1}, \ldots, X_0$, which satisfies assumptions (A1) and (A2). Then for any $1 \le m \le M_\delta$, with probability one,*

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

for large n, 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 $m \ge 1$ change-points, both of which use conditional maximum likelihood variance estimates.

Proof. By the definition of MDL in (2.3), we have

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

where $\hat{\sigma}^2$ is the conditional maximum likelihood estimate of the AR(p) noise variance over the entire data set, and

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

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

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

$$\begin{aligned} \frac{2}{n} \big[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p) - \mathrm{MDL}(0; p) \big] &= \frac{2 \log m}{n} + m(p+4) \frac{\log n}{n} + \frac{2m \log p}{n} \\ &\quad + \frac{p+2}{n} \sum_{k=1}^{m+1} \log(\hat{\lambda}_k - \hat{\lambda}_{k-1}) + \sum_{k=1}^{m+1} (\hat{\lambda}_k - \hat{\lambda}_{k-1}) \log \hat{\sigma}_k^2 - \log \hat{\sigma}^2. \end{aligned} \tag{4.2}$$

We will show that (4.2) 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 on the right hand side of (4.2) 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 $k = 1, \ldots, m+1$. Thus, since $\hat{\sigma}^2$ also converges to σ^2, the quantity

$$\sum_{k=1}^{m+1} (\hat{\lambda}_k - \hat{\lambda}_{k-1}) \log \hat{\sigma}_k^2 - \log \hat{\sigma}^2 \tag{4.3}$$

converges to zero as $n \to \infty$. We will use the FLIL on a Taylor series expansion of (4.3) to show that this quantity is of order $\log\log n/n$, from which the theorem follows since $\log n/n > \log\log n/n$ for n large.

Defining

$$\mathbf{X}_{a:b} = \left(X_a, X_{a+1}, \cdots, X_b\right)^T \quad \text{and} \tag{4.4}$$

$$\tag{4.5}$$

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

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 $(1, \ldots, 1)^T$ and $\mathbf{X}_{1-i:n-i}$, $i = 1, \ldots, p$, i.e.,

$$n\hat{\sigma}^2 = \left\| \mathbf{X}_{1:n} - \mathrm{P}_{\boldsymbol{N}_{1:n}^p}(\mathbf{X}_{1:n}) \right\|^2, \tag{4.6}$$

where $\mathrm{P}_{\boldsymbol{N}_{1:n}^p}(\mathbf{X}_{1:n})$ is the projection of $\mathbf{X}_{1:n}$ onto the $(p+1)$-dimensional column space of $\boldsymbol{N}_{1:n}^p$. This is the same as the squared norm of the difference between

$\mathbf{X}^*_{1:n}$ and its projection onto the subspace spanned by $\mathbf{X}^*_{1-i:n-i}$, $i = 1, \ldots, p$, where $\mathbf{X}^*_{1:n}$ is the component of $\mathbf{X}_{1:n}$ orthogonal to $(1, \ldots, 1)^T$, i.e.,

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

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

$$\hat{\sigma}^2 = G_p\left(\hat{\gamma}(i,j) : i,j = 0, \ldots, p\right), \tag{4.7}$$

where

$$\hat{\gamma}(i,j) := \hat{\gamma}_{0:1}(i,j) = \frac{1}{n}\sum_{t=1}^{n}(X_{t-i} - \overline{X}_{1-i:n-i})(X_{t-j} - \overline{X}_{1-j:n-j}), \tag{4.8}$$

is the sample ACVF (cf. (3.4)), and

$$G_p\left(u_{ij} : i,j = 0, \ldots, p\right) = u_{00} - (u_{01}, \ldots, u_{0p})\left[\left\{u_{ij}\right\}_{i,j=1}^{p}\right]^{-1}\begin{pmatrix} u_{01} \\ \vdots \\ u_{0p} \end{pmatrix}. \tag{4.9}$$

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

$$(\hat{\lambda}_k - \hat{\lambda}_{k-1})n\hat{\sigma}_k^2 = \left\|\mathbf{X}_{\hat{\lambda}_{k-1}n+1:\hat{\lambda}_k n} - \mathrm{P}_{\boldsymbol{N}^p_{\hat{\lambda}_{k-1}n+1:\hat{\lambda}_k n}}\left(\mathbf{X}_{\hat{\lambda}_{k-1}n+1:\hat{\lambda}_k n}\right)\right\|^2, \tag{4.10}$$

and thus,

$$\hat{\sigma}_k^2 = G_p\left(\hat{\gamma}_k(i,j) : i,j = 0, \ldots, p\right), \tag{4.11}$$

where

$$\begin{aligned}\hat{\gamma}_k(i,j) &:= \hat{\gamma}_{\hat{\lambda}_{k-1}:\hat{\lambda}_k}(i,j) \qquad (4.12)\\ &= \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\sum_{t=\hat{\lambda}_{k-1}n+1}^{\hat{\lambda}_k n}(X_{t-i} - \overline{X}_{\hat{\lambda}_{k-1}n+1-i:\hat{\lambda}_k n-i})(X_{t-j} - \overline{X}_{\hat{\lambda}_{k-1}n+1-j:\hat{\lambda}_k n-j})\end{aligned}$$

is the sample ACVF in the kth segment (cf. (3.4)).

While the causal AR(p) process $\{X_t\}$ is not assumed to be stationary, it 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 (3.7) 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

$$\mu := \lim_{t\to\infty} \mathbb{E}(X_t) \text{ and } \gamma(h) := \lim_{t\to\infty} \mathrm{Cov}(X_t, X_{t-h}). \tag{4.13}$$

Note that

$$\mu = \mathbb{E}(X'_t) \text{ and } \gamma(h) = \mathrm{Cov}(X'_t, X'_{t-h}) \text{ for all } t. \tag{4.14}$$

Without loss of generality, we take $\mu = 0$ since the estimates $\hat{\sigma}_k^2$ are location invariant, and consider the μ-centered sample ACVF (cf. (3.3)),

$$\tilde{\gamma}(i,j) := \tilde{\gamma}_{0:1}(i,j) = \frac{1}{n}\sum_{t=1}^{n}(X_{t-i}-\mu)(X_{t-j}-\mu) = \frac{1}{n}\sum_{t=1}^{n} X_{t-i}X_{t-j}, \text{ and} \tag{4.15}$$

$$\tilde{\gamma}_k(i,j) := \tilde{\gamma}_{\hat{\lambda}_{k-1}:\hat{\lambda}_k}(i,j) = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=\hat{\lambda}_{k-1}n+1}^{\hat{\lambda}_k n} X_{t-i}X_{t-j}. \tag{4.16}$$

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 $\mathbf{X}_{\hat{\lambda}_{k-1}n+1-i:\hat{\lambda}_k n-i}$, $i = 1, \ldots, p$. It follows that

$$\tilde{\sigma}^2 = G_p(\tilde{\gamma}(i,j) : i,j = 0, \ldots, p), \text{ and} \tag{4.17}$$

$$\tilde{\sigma}_k^2 = G_p(\tilde{\gamma}_k(i,j) : i,j = 0, \ldots, p) \tag{4.18}$$

for each $k = 1, \ldots, m+1$. 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

$$\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\tilde{\sigma}_k^2 - \log\tilde{\sigma}^2 = O\left(\frac{\log\log n}{n}\right), \tag{4.19}$$

which then implies that

$$\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\hat{\sigma}_k^2 - \log\hat{\sigma}^2 = O\left(\frac{\log\log n}{n}\right). \tag{4.20}$$

Let $\boldsymbol{\gamma} = (\gamma(|i-j|) : i,j = 0,\ldots,p)$ be the vector of true (limiting) ACVFs ranging over lags $0,\ldots,p$ (cf. (4.13) and (4.14)). Carrying out a second order Taylor expansion on each of the $\log\tilde{\sigma}_k^2$ terms and the $\log\tilde{\sigma}^2$ term, we obtain

$$\begin{aligned}
&\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\tilde{\sigma}_k^2 - \log\tilde{\sigma}^2 \\
&= \sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log G_p(\tilde{\boldsymbol{\gamma}}_k) - \log G_p(\tilde{\boldsymbol{\gamma}}) \\
&= \left[\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log G_p(\boldsymbol{\gamma}) - \log G_p(\boldsymbol{\gamma})\right] \\
&\quad + \left[\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\nabla\log G_p(\boldsymbol{\gamma})(\tilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}) - \nabla\log G_p(\boldsymbol{\gamma})(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right] \\
&\quad + \frac{1}{2}\left[\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})(\tilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma})^T\nabla^2\log G_p(\boldsymbol{\gamma}_k^*)(\tilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma})\right. \\
&\qquad \left. - (\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\nabla^2\log G_p(\boldsymbol{\gamma}^*)(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right],
\end{aligned} \tag{4.21}$$

where $\tilde{\boldsymbol{\gamma}}_k := (\tilde{\gamma}_k(i,j) : i,j = 0,\ldots,p)$ and $\tilde{\boldsymbol{\gamma}} := (\tilde{\gamma}(i,j) : i,j = 0,\ldots,p)$. 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 $k = 1,\ldots,m+1$, 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 (4.21) are exactly zero due to the form of the conditional maximum likelihood estimates; see

(4.15) and (4.16). 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. (3.5) and (3.7)), and (4.19) holds. (More precisely, (3.5) applies to the μ-centered sample ACVF for the stationary $\{X'_t\}$. Due to (3.7), it also applies to $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\gamma}}_k$ for $\{X_t\}$.) Thus, by (4.20), (4.2) becomes

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

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

5. Yule-Walker

As in the last section, we assume that the true process follows the AR(p) model in (4.1) (i.e. Model 1). To show that the consistency result breaks down when using the MDL together with the Yule-Walker variance estimates, it suffices to prove that there is a nonnegligible probability that Model 1 is not selected in favor of Model 2 with $m = 1$ for large n, which is stated formally in the following theorem.

Theorem 5.1. *Assume that the true process $\{X_t\}$ follows the AR(p) model in (4.1) where 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) $\liminf_{u\to\infty} e^{cu}\int_u^\infty f_\epsilon(x)dx > 0$ *for some constant* $c > 0$.

Then, using Yule-Walker estimation in the MDL, for every $0 < \delta < 1/2$ *and* $C > 0$,

$$\liminf_{n\to\infty} P\left(\mathrm{MDL}(0;p) - \min_{\delta\le\lambda\le 1-\delta}\mathrm{MDL}(1,\lambda;p,p) > C\log n\right) > 0. \tag{5.1}$$

PROOF. As in the proof of Theorem 4.1, note that the causal process $\{X_t\}$ can be approximated by a stationary AR(p) process $\{X'_t\}$ with the same AR coefficients and the same noise sequence in such a way that (3.7) holds. So it suffices to prove the theorem under the additional assumption that $\{X_t\}$ is stationary. By (2.3), the MDL for an AR(p) process with no change-points equals

$$\mathrm{MDL}(0;p) = \frac{p+4}{2}\log n + \log p + \frac{n}{2}\log(2\pi\hat{\sigma}'^2), \tag{5.2}$$

while the MDL when calculated under one change-point for an AR(p) in each of the two segments is

$$\begin{aligned}\mathrm{MDL}(1,\hat{\lambda};p,p) &= \min_{\delta\le\lambda\le 1-\delta}\mathrm{MDL}(1,\lambda;p,p)\\ &= \min_{\delta\le\lambda\le 1-\delta}\Big\{(p+4)\log n + \frac{p+2}{2}\big(\log(\lambda)+\log(1-\lambda)\big)\\ &\quad + 2\log p + \frac{n}{2}\left(\lambda\log(2\pi\hat{\sigma}_1'^2(\lambda)) + (1-\lambda)\log(2\pi\hat{\sigma}_2'^2(\lambda))\right)\Big\},\end{aligned} \tag{5.3}$$

where now $\hat{\sigma}'^2$ is the Yule-Walker variance estimate of the entire sequence, $\hat{\sigma}_1'^2(\lambda)$ is the Yule-Walker variance estimate from observations $1,\ldots,\lambda n$, and $\hat{\sigma}_2'^2(\lambda)$ is the Yule-Walker variance estimate from observations $\lambda n+1,\ldots,n$. The change-point estimate $\hat{\lambda}$ is obtained by minimizing $\mathrm{MDL}(1,\lambda;p,p)$ with respect to $\lambda \in A_1^\delta = [\delta, 1-\delta]$. By (5.2) and (5.3), to establish (5.1) for all $0 < \delta < 1/2$ and $C > 0$, it suffices to show that for all $0 < \delta < 1/2$ and $C > 0$,

$$\liminf_{n\to\infty} P\left(\log\hat{\sigma}'^2 - \min_{\delta\le\lambda\le 1-\delta}\left\{\lambda\log\hat{\sigma}_1'^2(\lambda) + (1-\lambda)\log\hat{\sigma}_2'^2(\lambda)\right\} > Cn^{-1}\log n\right) > 0. \tag{5.4}$$

For $\lambda \in [\delta, 1-\delta], k = 1, 2$ and $h = 0, 1, \ldots,$ define

$$\hat{\gamma}'(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \overline{X}_{1:n})(X_{t+h} - \overline{X}_{1:n}),$$

$$\hat{\gamma}'_{k,\lambda}(h) = \frac{1}{(\lambda_k - \lambda_{k-1})n}\sum_{t=\lambda_{k-1}n+1}^{\lambda_k n-h}(X_t - \overline{X}_{\lambda_{k-1}n+1:\lambda_k n})(X_{t+h} - \overline{X}_{\lambda_{k-1}n+1:\lambda_k n}),$$

where $\lambda_0 := 0, \lambda_1 := \lambda$ and $\lambda_2 := 1$. Let $\hat{\gamma}'(i,j) := \hat{\gamma}'(|i-j|)$ and $\hat{\gamma}'_{k,\lambda}(i,j) := \hat{\gamma}'_{k,\lambda}(|i-j|)$. Then we have $\hat{\sigma}'^2 = G_p(\hat{\gamma}'(i,j) : i,j = 0,\ldots,p)$ and $\hat{\sigma}'^2_k(\lambda) = G_p(\hat{\gamma}'_{k,\lambda}(i,j) : i,j = 0,\ldots,p)$ where G_p is given in (4.9). Since the above variance estimates are all location invariant, we assume without loss of generality that the stationary process $\{X_t\}$ has mean $\mu = 0$, and consider the following μ-centered version of $\hat{\gamma}'$ and $\hat{\gamma}'_{k,\lambda}$

$$\tilde{\gamma}'(h) := \frac{1}{n}\sum_{t=1}^{n-h} X_t X_{t+h},$$

$$\tilde{\gamma}'_{k,\lambda}(h) := \frac{1}{(\lambda_k - \lambda_{k-1})n}\sum_{t=\lambda_{k-1}n+1}^{\lambda_k n-h} X_t X_{t+h},$$

$$\tilde{\gamma}'(i,j) := \tilde{\gamma}'(|i-j|), \text{ and } \tilde{\gamma}'_{k,\lambda}(i,j) := \tilde{\gamma}'_{k,\lambda}(|i-j|).$$

By FLIL, $\hat{\gamma}'(i,j) - \tilde{\gamma}'(i,j) = O(n^{-1}\log\log n)$ a.s., and

$$\sup_{\delta\le\lambda\le 1-\delta}|\hat{\gamma}'_{k,\lambda}(i,j) - \tilde{\gamma}'_{k,\lambda}(i,j)| = O(n^{-1}\log\log n) \text{ a.s.},$$

implying that $\hat{\sigma}'^2 - \tilde{\sigma}'^2 = (n^{-1}\log\log n)$ a.s. and

$$\sup_{\delta\le\lambda\le 1-\delta}|\hat{\sigma}'^2_k(\lambda) - \tilde{\sigma}'^2_k(\lambda)| = O(n^{-1}\log\log n) \text{ a.s., } k = 1, 2,$$

where $\tilde{\sigma}'^2 := G_p(\tilde{\gamma}'(i,j) : i,j = 0,\ldots,p)$ and $\tilde{\sigma}'^2_k(\lambda) := G_p(\tilde{\gamma}'^2_{k,\lambda}(i,j) : i,j = 0,\ldots,p), k = 1, 2$. To prove (5.4), it suffices to show that for all $0 < \delta < 1/2$ and $C > 0$,

$$\liminf_{n\to\infty} P\left(\log\tilde{\sigma}'^2 - \min_{\delta\le\lambda\le 1-\delta}\left\{\lambda\log\tilde{\sigma}'^2_1(\lambda) + (1-\lambda)\log\tilde{\sigma}'^2_2(\lambda)\right\} > Cn^{-1}\log n\right) > 0. \tag{5.5}$$

We now prove (5.5) for the case $p = 1$. (The general case $p > 1$ can be treated similarly with more complicated notation and extensive calculations.) Writing $\phi = \phi_1 (\neq 0)$, the AR(1) process $\{X_t\}$ satisfies $X_t = \phi X_{t-1} + \sigma\epsilon_t$ (since μ is assumed to be 0). Let $\tau = \lambda n$, an integer between δn and $(1-\delta)n$. It is readily seen that

$$\tilde{\sigma}'^2 = \frac{1}{n}\sum_{t=1}^{n} X_t^2 - \frac{\left(\frac{1}{n}\sum_{t=1}^{n-1} X_t X_{t+1}\right)^2}{\frac{1}{n}\sum_{t=1}^{n} X_t^2}, \tag{5.6}$$

$$\tilde{\sigma}_1'^2(\lambda) = \tilde{\sigma}_1'^2(\tau/n) = \frac{1}{\tau}\sum_{t=1}^{\tau} X_t^2 - \frac{\left(\frac{1}{\tau}\sum_{t=1}^{\tau-1} X_t X_{t+1}\right)^2}{\frac{1}{\tau}\sum_{t=1}^{\tau} X_t^2}, \tag{5.7}$$

$$\tilde{\sigma}_2'^2(\lambda) = \tilde{\sigma}_2'^2(\tau/n) = \frac{1}{n-\tau}\sum_{t=\tau+1}^{n} X_t^2 - \frac{\left(\frac{1}{n-\tau}\sum_{t=\tau+1}^{n-1} X_t X_{t+1}\right)^2}{\frac{1}{n-\tau}\sum_{t=\tau+1}^{n} X_t^2}. \tag{5.8}$$

Performing a Taylor expansion on the log of (5.6), we obtain

$$\begin{aligned}
\log\tilde{\sigma}'^2 &= \log\left[\frac{\sigma^2}{1-\phi^2} + \frac{1}{n}\sum_{t=1}^{n}\left(X_t^2 - \frac{\sigma^2}{1-\phi^2}\right)\right. \\
&\quad \left. - \frac{\left(\frac{\phi\sigma^2}{1-\phi^2} + \frac{1}{n}\sum_{t=0}^{n-1}\left(X_t X_{t+1} - \frac{\phi\sigma^2}{1-\phi^2}\right) - \frac{X_0 X_1}{n}\right)^2}{\frac{\sigma^2}{1-\phi^2} + \frac{1}{n}\sum_{t=1}^{n}\left(X_t^2 - \frac{\sigma^2}{1-\phi^2}\right)}\right] \\
&= \log\sigma^2 + \frac{1+\phi^2}{\sigma^2}\frac{1}{n}\sum_{t=1}^{n}\left(X_t^2 - \frac{\sigma^2}{1-\phi^2}\right) \\
&\quad - \frac{2\phi}{\sigma^2}\left[\frac{1}{n}\sum_{t=0}^{n-1}\left(X_t X_{t+1} - \frac{\phi\sigma^2}{1-\phi^2}\right) - \frac{X_0 X_1}{n}\right] + O\left(\frac{\log\log n}{n}\right) \quad \text{a.s. ,}
\end{aligned}$$

where the second equality follows from the facts that

$$\Big[\frac{\phi\sigma^2}{1-\phi^2}+\frac{1}{n}\sum_{t=0}^{n-1}\left(X_tX_{t+1}-\frac{\phi\sigma^2}{1-\phi^2}\right)-\frac{X_0X_1}{n}\Big]^2$$
$$=\left(\frac{\phi\sigma^2}{1-\phi^2}\right)^2+\left(\frac{2\phi\sigma^2}{1-\phi^2}\right)\left[\frac{1}{n}\sum_{t=0}^{n-1}\left(X_tX_{t+1}-\frac{\phi\sigma^2}{1-\phi^2}\right)-\frac{X_0X_1}{n}\right]$$
$$+O\left(n^{-1}\log\log n\right)\quad\text{a.s.},$$

and that

$$\Big[\frac{\sigma^2}{1-\phi^2}+\frac{1}{n}\sum_{t=1}^{n}\left(X_t^2-\frac{\sigma^2}{1-\phi^2}\right)\Big]^{-1}$$
$$=\left(\frac{\sigma^2}{1-\phi^2}\right)^{-1}\Big[1-\left(\frac{\sigma^2}{1-\phi^2}\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\left(X_t^2-\frac{\sigma^2}{1-\phi^2}\right)+O\left(\frac{\log\log n}{n}\right)\Big]\quad\text{a.s.}$$

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

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

Similarly, for (5.7) and (5.8),

$$\max_{\delta\le\frac{\tau}{n}\le1-\delta}\left|\log\tilde{\sigma}_1'^2(\tau/n)\;-\;\left(\log\sigma^2+\frac{1}{\tau}S_\tau+\frac{2\phi}{\sigma^2}\frac{X_0X_1}{\tau}\right)\right|$$

and

$$\max_{\delta\le\frac{\tau}{n}\le1-\delta}\left|\log\tilde{\sigma}_2'^2(\tau/n)\;-\;\left(\log\sigma^2+\frac{1}{n-\tau}(S_n-S_\tau)+\frac{2\phi}{\sigma^2}\frac{X_\tau X_{\tau+1}}{n-\tau}\right)\right|$$

are both $O\left(n^{-1}\log\log n\right)$. It follows that

$$\max_{\delta\le\frac{\tau}{n}\le1-\delta}\left|\log\tilde{\sigma}'^2-\left(\frac{\tau}{n}\log\tilde{\sigma}_1'^2(\tau/n)+\frac{n-\tau}{n}\log\tilde{\sigma}_2'^2(\tau/n)\right)\right.$$
$$\left.+\frac{2\phi}{\sigma^2}\frac{X_\tau X_{\tau+1}}{n}\right|=O\left(n^{-1}\log\log n\right).\qquad(5.9)$$

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

$$\lim_{n\to\infty} P\left(\max\left\{\epsilon_t : \delta \le \frac{t}{n} \le 1-\delta\right\} > \frac{1}{2c}\log n\right) = 1. \tag{5.10}$$

Let $T := \max\left\{[n\delta]+2 \le t \le n(1-\delta) : \epsilon_t > \frac{1}{2c}\log n\right\}$ if the specified set is non-empty, and define $T := [n\delta]+2$ otherwise. Note by (5.10) 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 (5.9), we have

$$\begin{aligned}
&\log\tilde{\sigma}'^2 - \left(\frac{T-1}{n}\log\tilde{\sigma}_1'^2((T-1)/n) + \frac{n-T+1}{n}\log\tilde{\sigma}_2'^2((T-1)/n)\right) \\
&= \frac{-2\phi}{\sigma^2}\frac{X_{T-1}X_T}{n} + O\left(n^{-1}\log\log n\right) \\
&= \frac{-2\phi}{\sigma^2}\frac{(\phi X_{T-1}^2 + \sigma X_{T-1}\epsilon_T)}{n} + O\left(n^{-1}\log\log n\right) \\
&= \left(\frac{-\phi}{c\sigma}X_{T-1}\frac{2c\epsilon_T}{\log n}\right)\frac{\log n}{n} - \frac{2\phi^2}{\sigma^2}\frac{X_{T-1}^2}{n} + O\left(n^{-1}\log\log n\right) \\
&= \left(\frac{-\phi}{c\sigma}X_{T-1}\frac{2c\epsilon_T}{\log n}\right)\frac{\log n}{n} - O_p(n^{-1}) + O\left(n^{-1}\log\log n\right). \qquad (5.11)
\end{aligned}$$

For every $C > 0$, we have $P\left(\frac{-\phi}{c\sigma}X_{T-1} > C\right) > 0$, implying (5.5) for every $C > 0$ (and $0 < \delta < 1/2$). This completes the proof. □

Remark 5.1. Theorem 4.1 requires only the moment condition (3.1) in order for the consistency result to hold when using the conditional maximum likelihood estimates. On the other hand, for the consistency result to break down in the Yule-Walker case, condition (ii) in Theorem 5.1 essentially requires the noise distribution to have exponential (or heavier) tails. It would be of theoretical interest to see if the consistency result can be saved in the Yule-Walker case when the noise distribution has very light tails (such as the normal distribution). Furthermore, we may extend Theorem 5.1 to the case where the true process follows a

piecewise autoregressive model with $m \geq 1$ change-points and show that there is a nonnegligible probability that more than m change-points will be selected by using Yule-Walker estimation in the MDL.

6. Simulation Results

In order to investigate the practical differences between conditional maximum likelihood and Yule-Walker estimation when applied to structural break detection, we conducted a small simulation study. We were also curious if the lack of consistency specified in Theorem 5.1 could be discerned in simulated data. To explore these issues, we simulated a first-order mean-zero autoregressive process (with no change-points) given by

$$X_t = .8X_{t-1} + \epsilon_t, \quad \{\epsilon_t\} \sim IID(0, \sigma^2),$$

using several different noise distributions. For each simulated process, we calculated the MDL assuming no change-points (and AR order 1), MDL(0;1), as well as the MDL assuming one change-point (and AR order 1 in each segment), MDL$(1, \lambda; 1, 1)$, with change-point location λn for $\delta \leq \lambda \leq 1 - \delta$. The MDL was calculated for both conditional maximum likelihood (CML) and Yule-Walker (YW) estimates, where the mean in each segment was assumed known equal to zero and not estimated (so that the term $\sum_{k=1}^{m+1} \frac{p_k+2}{2} \log n_k$ in (2.3) was replaced by $\sum_{k=1}^{m+1} \frac{p_k+1}{2} \log n_k = \sum_{k=1}^{m+1} \log n_k, m = 0, 1$). We ran the simulations for a variety of noise distributions, sample sizes (n), and δ values. Based on 1000 replications for each case, the proportion of replications with $\min_{\delta \leq \lambda \leq 1-\delta}$ MDL$(1, \lambda; 1, 1) <$ MDL(0; 1) (indicating evidence for a change-point) is reported in Tables 1, 2 and 3.

The most definitive results concern the case of a t-distribution with 5 degrees of freedom. This distribution satisfies the conditions of Theorem 5.1 and as one

can see in Table 1, with $\delta = .005$, the rejection rate increases from 21.2% to 35.8% as the sample size grows from 1000 to 50,000. In contrast, the rejection rates are getting smaller (from 17.1% to 8.7%) in the CML case as n increases. Similar patterns persist for the four other choices of δ.

As seen in the normal case (Table 2), the rejection rates for both CML and YW are both converging to zero (no more than one out of 1000 replications rejects the true model for sample sizes greater than 1000 and all choices of δ). This does not contradict Theorem 5.1 since the normal distribution does not satisfy condition (ii) of the theorem. These results suggest that both CML and YW produce a consistent procedure for structural break detection for normal noise.

For Laplace noise (see Table 3), both estimates YW and CML have quite small rejection rates (3.3% and .6%, respectively, for $n = 50,000$ and $\delta = .1$). By Theorem 5.1 the asymptotic rate for YW is positive so that an enormous sample size may be required in order to confirm this result. So from a practical perspective, there may be little consequence in using YW estimates in this context.

In other simulation cases with noise distributions that have infinite fourth moment, it appears that the rejection rates (not reported here) for both CML and YW are rather large. In this case, neither the conditions of Theorem 4.1 or Theorem 5.1 are met. The heavy-tailed case will be the subject of future research and is beyond the scope of the current paper.

Based on this limited study, the use of YW for structural break estimation could have an impact for structural break detection if the tails of the noise distribution are heavy. It would certainly be of practical interest to get a better sense about the range of distributions for which the use of Yule-Walker estimation may be problematic in structural break detection. An advantage of Yule-Walker estimates is that they remain numerically stable and can be computed quickly over various subsets of the time series, so one does not want to eliminate this estimation

procedure from consideration.

References

Athreya, K. B., Pantula, S. G., 1986a. Mixing properties of Harris chains and autoregressive processes. Journal of Applied Probability 23, 880–892.

Athreya, K. B., Pantula, S. G., 1986b. A note on strong mixing of ARMA processes. Statistics & Probability Letters 4, 187–190.

Bai, J., 1999. Likelihood ratio tests for multiple structural changes. Journal of Econometrics 91 (2), 299–323.

Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica, 47–78.

Bai, J., Perron, P., 2003. Computation and analysis of multiple structural change models. Journal of Applied Econometrics 18 (1), 1–22.

Brockwell, P. J., Davis, R. A., 1991. Time Series: Theory and Methods, 2nd Edition. Springer-Verlag.

Chan, N. H., Yau, C. Y., Zhang, R.-M., 2014. Group LASSO for structural break time series. Journal of the American Statistical Association 109, 590–599.

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

Chernoff, H., 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.

Csörgó, M., Horváth, L., 1997. Limit Theorems in Change-Point Analysis. Wiley.

Davis, R. A., Huang, D., 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.

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

Davis, R. A., Lee, T. C. M., Rodriguez-Yam, G. A., 2008. Break detection for a class of nonlinear time series models. Journal of Time Series Analysis 29, 834–867.

Davis, R. A., Yau, C. Y., 2013. Consistency of minimum description length model selection for piecewise stationary time series models. Electronic Journal of Statistics 7, 381–411.

Diebold, F. X., Inoue, A., 2001. Long memory and regime switching. Journal of Econometrics 105 (1), 131–159.

Fearnhead, P., 2006. Exact and efficient Bayesian inference for multiple change-point problems. Statistics and Computing 16, 203–213.

Fryzlewicz, P., 2014. Wild binary segmentation for multiple change-point detection. Annals of Statistics 42, 2243–2281.

Hancock, S., 2008. Estimation of structural breaks in nonstationary time series. Ph.D. thesis, Colorado State University.

Hawkins, D. M., 2001. Fitting multiple change-point models to data. Computational Statistics & Data Analysis 37, 323–341.

Kokoszka, P., Leipus, R., 2000. Change-point estimation in ARCH models. Bernoulli 6, 513–539.

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

Lee, C. B., 1996. Nonparametric multiple change-point estimators. Statistics & Probability Letters 27, 295–304.

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

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

Lu, Q., Lund, R., Lee, T. C. M., 2010. An MDL approach to the climiate segmentation problem. Annals of Applied Statistics 4, 299–319.

Meyn, S. P., Tweedie, R. L., 1993. Markov Chains and Stochastic Stability. Springer.

Mikosch, T., Stărică, C., 2004. Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. Review of Economics and Statistics 86 (1), 378–390.

Perreault, L., Bernier, J., Bobée, B., Parent, E., 2000a. Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited. Journal of Hydrology 235, 221–241.

Perreault, L., Bernier, J., Bobée, B., Parent, E., 2000b. Bayesian change-point

analysis in hydrometeorological time series. Part 2. Comparison of change-point models and forecasting. Journal of Hydrology 235, 242–263.

Rio, E., 1995. The functional law of the iterated logarithm for stationary strongly mixing sequences. The Annals of Probability 23, 1188–1203.

Rissanen, J., 1989. Stochastic Complexity in Statistical Inquiry. World Scientific Publishing Company.

Stephens, D. A., 1994. Bayesian retrospective multiple-changepoint identification. Applied Statistics 43, 159–178.

Sullivan, J. H., 2002. Estimating the locations of multiple change points in the mean. Computational Statistics 17, 289–296.

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

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

Yao, Y. C., Au, S. T., 1989. Least-squares estimation of a step function. Sankhya: The Indian Journal of Statistics 51, 370–381.

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

Table 1: Rejection rate with t noise distribution with 5 degrees of freedom.

Method	n	$\delta = 0.005$	0.01	0.03	0.05	0.1
YW	1000	0.212	0.204	0.190	0.170	0.150
CML	1000	0.171	0.152	0.137	0.125	0.105
YW	2000	0.183	0.174	0.156	0.146	0.131
CML	2000	0.127	0.117	0.105	0.097	0.079
YW	5000	0.204	0.193	0.169	0.161	0.140
CML	5000	0.112	0.102	0.081	0.072	0.059
YW	10000	0.248	0.238	0.212	0.195	0.164
CML	10000	0.120	0.112	0.090	0.073	0.052
YW	20000	0.286	0.275	0.254	0.237	0.206
CML	20000	0.104	0.089	0.072	0.059	0.045
YW	50000	0.358	0.341	0.318	0.303	0.270
CML	50000	0.087	0.074	0.053	0.047	0.036

Table 2: Rejection rate with Gaussian noise distribution.

Method	n	$\delta = 0.005$	0.01	0.03	0.05	0.1
YW	1000	0.008	0.004	0.003	0.002	0.001
CML	1000	0.006	0.002	0.002	0.002	0.001
YW	2000	0.001	0.000	0.000	0.000	0.000
CML	2000	0.000	0.000	0.000	0.000	0.000
YW	5000	0.001	0.000	0.000	0.000	0.000
CML	5000	0.000	0.000	0.000	0.000	0.000
YW	10000	0.000	0.000	0.000	0.000	0.000
CML	10000	0.000	0.000	0.000	0.000	0.000
YW	20000	0.000	0.000	0.000	0.000	0.000
CML	20000	0.000	0.000	0.000	0.000	0.000
YW	50000	0.000	0.000	0.000	0.000	0.000
CML	50000	0.000	0.000	0.000	0.000	0.000

Table 3: Rejection rate with Laplace noise distribution.

Method	n	$\delta = 0.005$	0.01	0.03	0.05	0.1
YW	1000	0.125	0.115	0.090	0.081	0.057
CML	1000	0.128	0.097	0.070	0.061	0.044
YW	2000	0.082	0.076	0.056	0.052	0.040
CML	2000	0.068	0.062	0.040	0.037	0.022
YW	5000	0.079	0.070	0.055	0.050	0.036
CML	5000	0.041	0.033	0.026	0.022	0.017
YW	10000	0.066	0.062	0.049	0.044	0.034
CML	10000	0.029	0.024	0.016	0.014	0.007
YW	20000	0.050	0.040	0.028	0.023	0.020
CML	20000	0.027	0.016	0.008	0.006	0.005
YW	50000	0.055	0.051	0.043	0.039	0.033
CML	50000	0.014	0.013	0.008	0.007	0.006