DISSERTATION

ESTIMATION OF STRUCTURAL BREAKS IN NONSTATIONARY TIME SERIES

Submitted by

Stacey Hancock

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2008

UMI Number: 3346426

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3346426

Copyright 2009 by ProQuest LLC.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

COLORADO STATE UNIVERSITY

July 24, 2008

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY STACEY HANCOCK ENTITLED ESTIMATION OF STRUCTURAL BREAKS IN NONSTATIONARY TIME SERIES BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Peter J. Brockwell

N. Thompson Hobbs

Hari K. Iyer (Co-adviser)

Richard A. Davis (Adviser)

Jean D. Opsomer (Acting Department Head)

ABSTRACT OF DISSERTATION

ESTIMATION OF STRUCTURAL BREAKS IN NONSTATIONARY TIME SERIES

Many time series exhibit structural breaks in a variety of ways, the most obvious being a mean level shift. In this case, the mean level of the process is constant over periods of time, jumping to different levels at times called change-points. These jumps may be due to outside influences such as changes in government policy or manufacturing regulations. Structural breaks may also be a result of changes in variability or changes in the spectrum of the process. The goal of this research is to estimate where these structural breaks occur and to provide a model for the data within each stationary segment. The program Auto-PARM (Automatic Piecewise AutoRegressive Modeling procedure), developed by Davis, Lee, and Rodriguez-Yam (2006) , uses the minimum description length principle to estimate the number and locations of change-points in a time series by fitting autoregressive models to each segment.

The research in this dissertation shows that when the true underlying model is segmented autoregressive, the estimates obtained by Auto-PARM are consistent. Under a more general time series model exhibiting structural breaks, Auto-PARM's estimates of the number and locations of change-points are again consistent, and the segmented autoregressive model provides a useful approximation to the true process. Weak consistency proofs are given, as well as simulation results when the true process is not autoregressive.

An example of the application of Auto-PARM as well as a source of inspiration for this research is the analysis of National Park Service sound data. This data was collected by the National Park Service over four years in around twenty of the National Parks by setting recording devices in several sites throughout the parks. The goal of the project is to estimate the amount of manmade sound in the National Parks. Though the project is in its initial stages, Auto-PARM provides a promising method for analyzing sound data by breaking the sound waves into pseudo-stationary pieces. Once the sound data have been broken into pieces, a classification technique can be applied to determine the type of sound in each segment.

Stacey Hancock
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Fall 2008

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the invaluable and continuing support of Dr. Richard Davis. Through his encouragement and patient coaching, he motivated me to become a better scientific researcher. His research, teaching, and professional career are true inspirations. I thank him for all the opportunities he provided and for his friendship.

I am also indebted to Dr. Hari Iyer for his guidance and confidence. Dr. Iyer introduced me to the National Park Service sound data problem, which sparked the research in this dissertation. He is a valued friend, and I thank him for his unending support.

I would like to thank Dr. Yi-Ching Yao and the National Science Foundation's East Asia and Pacific Summer Institute for providing the opportunity to visit Academia Sinica in Taipei, Taiwan. Dr. Yao devoted an immeasurable amount of time to our research, even beyond my visit to Taiwan. The subtle theoretical details of this dissertation would not have been unraveled without Dr. Yao's insightful observations.

I also wish to thank Dr. Peter Brockwell and Dr. N. Thompson Hobbs for serving on my graduate committee. Dr. Brockwell's introduction to time series course gave me the foundation necessary for this dissertation.

The Statistics Department at Colorado State University and the Department of Mathematical Sciences at Montana State University provided the necessary training for this research. I would especially like to thank Dr. Robert Boik and Dr. John Borkowski from Montana State University for cultivating my devotion to Statistics.

Finally, I would like to acknowledge the National Science Foundation's Integrative Graduate Education and Research Traineeship (IGERT) Program, the Program for Interdisciplinary Mathematics, Ecology and Statistics (PRIMES), administered by Colorado State University.

DEDICATION

To my family, for their unending support.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 The Change-Point Problem

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" problem. The change-point problem considers observations ordered by some index, usually time, and postulates that their distribution changes at some unknown point in the sequence. The main objective is to detect the changes and estimate their locations. According to Zacks (1983), "The change-point problem can be considered one of the central problems of statistical inference. . . ." The most common change considered is a mean shift, where the mean of the process is piecewise constant. These jumps may be due to outside influences such as changes in government policy or manufacturing regulations. Structural breaks may also result from changes in variability, changes in the spectrum, or changes in some other feature of the process. Other examples of change-point problems can be found in Jassby and Powell (1990), who consider problems in ecological time series data, and Dias and Embrechts (2004), who discuss finance and insurance applications.

Change-point analysis, as discussed here, is concerned with *a posteriori* detection of change-points with a fixed sample size. Another area of research is the sequential or "on-line" study of changes in a process. Sequential methods are often used in problems of statistical control of industrial processes. For example, when manufacturing a product, data is collected and analyzed in real time, and the goal is to detect a change in the product as soon as possible after it occurs. Cumulative sum procedures

(CUSUM) are popular sequential change-point detection methods. See Zacks (1991) for a review of sequential methods.

Fixed sample change-point analysis can be broken into two main problems: hypothesis testing and estimation. Hypothesis testing is concerned with testing a null of no change-points versus an alternative of one or more change-points. See Bhattacharya (1994), Haccou and Meelis (1988), James, James and Siegmund (1987), Picard (1985), and Yao and Davis (1986) for a more comprehensive review of testing for change-points and for further references. This dissertation is concerned with the estimation side of the change-point problem. Rather than testing if change-points are present, we estimate the number of change-points and the change-point locations, then fit models to each segment. Estimating the change-points not only determines if one or more change-points are present, but gives further information about the number and locations of the change-points.

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. For example, Lee (1997) assumes independent observations from an exponential family, and Hawkins (2001) explores maximum likelihood change-point estimates for a general exponential family. Also, Lee (1996) considers a nonparametric approach for independent data with no distributional assumption. Other papers that explore change-points in independent data include Bhattacharya (1987), Jandhyala and Fotopoulos (1999), Yao and Au (1989), and Yao (1987). 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 is starting to focus more on detecting changes in dependent data, though the majority of this literature concerns hypothesis testing. Research on the estimation of the number and locations of the change-points includes Davis et al. (1995), who consider change-points in autoregressive processes, Kühn (2001), who assumes a weak invariance principle, and Kokoszka and Leipus (2000) on the estimation of change-points in ARCH models. This dissertation examines a method for estimating the number and locations of the change-points that does not assume independence nor a distribution on the data, e.g., normal, and does not assume a specific type of change. The method can detect changes in the mean, variance, spectrum, or other model parameters.

1.2 Automatic Piecewise Autoregressive Modeling (Auto-PARM)

Davis et al. (2006) developed a procedure for modeling a non-stationary time series by segmenting the series into blocks of different autoregressive processes. A random process $\{X_t\}$ is said to follow an autoregressive model of order p (or AR(p)) with mean μ if

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \sigma\epsilon_t$$

where $\{\epsilon_t\}$ is a white noise process with mean zero and unit variance, and $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0$ for all $|z| \leq 1$. The condition on the autoregressive polynomial $\phi(z)$ ensures that the process is causal. Autoregressive models have the unique ability to model any covariance function up to a certain lag. More precisely, for every non-singular covariance matrix $\Gamma_{p+1} = \{\gamma(i-j)\}_{i,j=1}^{p+1}$ of a stationary process, there is a causal AR(p) process whose autocovariances at lags 0, ..., p are exactly $\gamma(0)$, ..., $\gamma(p)$ [8]. This property makes autoregressive models ideal for approximating non-autoregressive processes.

The modeling procedure of Davis et al. (2006), referred to as Auto-PARM (*A*utomatic *P*iecewise *A*uto*R*egressive *M*odeling), uses a model selection criterion

called minimum description length to estimate the number of change-points, the locations of the change-points, the autoregressive model orders, and the autoregressive model parameters. The class of piecewise autoregressive models that Auto-PARM fits to an observed time series $\{y_t\}$ with n observations is as follows. For $j = 1, \ldots, m$, denote the change-point between the jth and $(j+1)$st autoregressive processes as τ_j, where $\tau_0 := 1$ and $\tau_{m+1} := n + 1$. Then the jth piece of the series is modeled as an autoregressive process

$$Y_t = X_{t,j}, \qquad t = \tau_{j-1}, \ldots, \tau_j - 1, \tag{1.1}$$

where $\{X_{t,j}\}$ is the causal AR(p_j) process

$$X_{t,j} = \phi_{j,0} + \phi_{j,1} X_{t-1,j} + \cdots + \phi_{j,p_j} X_{t-p_j,j} + \sigma_j \epsilon_t,$$

$\boldsymbol{\psi}_j := (\phi_{j,0}, \phi_{j,1}, \ldots, \phi_{j,p_j}, \sigma_j)$ is the parameter vector corresponding to this AR(p_j) process with $\phi_{j,p_j} \neq 0$, and the sequence $\{\epsilon_t\}$ is a white noise process with zero mean and unit variance. To ensure identifiability of the parameters, the model assumes that $\boldsymbol{\psi}_j \neq \boldsymbol{\psi}_{j+1}$ for any $j = 1, \ldots, m$. 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. Note that if we denote the mean of the jth segment as μ_j, then $\phi_{j,0} = \mu_j(1 - \phi_{j,1} - \cdots - \phi_{j,p_j})$. Given an observed time series $\{y_t\}_{t=1}^n$, Auto-PARM obtains the "best"-fitting model by finding the "best" combination of the number of change-points m, the change-point locations $\tau_1, \ldots, \tau_m$, and the AR orders $p_1, \ldots, p_{m+1}$. Once these parameters are specified, we can easily compute maximum likelihood estimates of the AR parameters $\boldsymbol{\psi}_j$ for each segment.

The model selection criteria that Auto-PARM uses to obtain the best-fitting model is called minimum description length (MDL). The minimum description length principle is a method for model selection developed by Jorma Rissanen in the 1980's (see, e.g., [40] and [41]). Underlying the concept of MDL is the insight that any regularity or pattern in the data can be used to compress the data [17]. In terms of

coding theory, the "best" model for the data is the one that describes the data with the shortest code length, or the least amount of memory space required to store the data. Rissanen developed a universal coding system in which the code length of a data set using a given model can be expressed as the sum of the code length of the fitted model plus the code length of the residuals given the fitted model. In other words, the data can be described as patterns in the data plus leftover "noise". This interpretation is not unlike model selection criteria which fit a model to the data by minimizing the likelihood plus a penalty term for model complexity, however, MDL prevents overfitting automatically and does not require the often *ad hoc* estimation of a tuning parameter for the penalty term nor the assumption of an underlying "true" model. Using MDL as a model selection criteria has many other advantages. For example, MDL chooses a model that trades off goodness-of-fit with model complexity, adhering to the principle of parsimony. For further review and discussion on minimum description length, see Lee (2001), Hansen and Yu (2001), and Grünwald et al. (2005).

If we denote the whole class of piecewise AR models by $\mathcal{M}$ and any model from this class by $\mathcal{F} \in \mathcal{M}$, then the MDL principle defines the "best"-fitting model from $\mathcal{M}$ as the one that produces the shortest code length that completely describes the observed data $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. Denoting the code length of an object z using model $\mathcal{F}$ by $CL_{\mathcal{F}}(z)$, the code length of the observed data can be expressed as

$$CL_{\mathcal{F}}(\boldsymbol{y}) = CL_{\mathcal{F}}(\hat{\mathcal{F}}) + CL_{\mathcal{F}}(\hat{\boldsymbol{e}}|\hat{\mathcal{F}}),$$

where $CL_{\mathcal{F}}(\hat{\mathcal{F}})$ is the code length of the fitted model $\hat{\mathcal{F}}$ and $CL_{\mathcal{F}}(\hat{\boldsymbol{e}}|\hat{\mathcal{F}})$ is the code length of the corresponding residuals $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ conditional on the fitted model $\hat{\mathcal{F}}$. Thus, deriving expressions for $CL_{\mathcal{F}}(\hat{\mathcal{F}})$ and $CL_{\mathcal{F}}(\hat{\boldsymbol{e}}|\hat{\mathcal{F}})$ results in an expression for $CL_{\mathcal{F}}(\boldsymbol{y})$. The MDL principle states that the best-fitting model is the model $\mathcal{F}$ that minimizes $CL_{\mathcal{F}}(\boldsymbol{y})$. Using Rissanen's results for encoding integers, bounded integers,

and maximum likelihood estimates, we obtain the code length of the fitted model $\hat{\mathcal{F}}$,

$$\begin{aligned} CL_{\mathcal{F}}(\hat{\mathcal{F}}) &= \log_2 m + (m+1)\log_2 n \\ &\quad + \sum_{j=1}^{m+1} \log_2 p_j + \sum_{j=1}^{m+1} \frac{p_j+2}{2} \log_2 n_j, \end{aligned}$$

where n_j is the number of observations in the jth segment. Rissanen demonstrated that the code length for the residuals $\hat{\boldsymbol{e}}$ given the fitted model $\hat{\mathcal{F}}$ is given by the negative of the log-likelihood of the fitted model $\hat{\mathcal{F}}$. Assuming the segments are independent, we can apply quasi-likelihood inference procedures to obtain the approximation

$$CL_{\mathcal{F}}(\hat{\boldsymbol{e}}|\hat{\mathcal{F}}) \approx \sum_{j=1}^{m+1} \left[\frac{n_j}{2} \log(2\pi) - \frac{1}{2} \log|\hat{\boldsymbol{V}}_j| + \frac{1}{2} \boldsymbol{y}_j^T \hat{\boldsymbol{V}}_j \boldsymbol{y}_j \right] \log_2 e.$$

where $\hat{\boldsymbol{V}}_j^{-1}$ is an estimate of the covariance matrix of the vector of observations in the jth segment.

Using logarithm base e rather than base 2, and using the standard approximation to the likelihood for AR models, we define the minimum description length for the piecewise autoregressive process model (1.1) used by Auto-PARM as

$$\begin{aligned} \mathrm{MDL}(m,\tau_1,\ldots,\tau_m;p_1,\ldots,p_{m+1}) &= \log m + (m+1)\log n \\ &\quad + \sum_{j=1}^{m+1} \log p_j + \sum_{j=1}^{m+1} \frac{p_j+2}{2} \log n_j + \sum_{j=1}^{m+1} \frac{n_j}{2} \log(2\pi\hat{\sigma}_j^2), \end{aligned} \tag{1.2}$$

where $\hat{\sigma}_j^2$ is the Yule-Walker estimate of the white noise variance in the jth segment (see [13] for further details on the derivation of the MDL).[1] Auto-PARM selects the best-fitting model for $\boldsymbol{y}$ as the model $\mathcal{F} \in \mathcal{M}$ that minimizes (1.2), which is

[1]The consistency proofs in the following chapters use conditional maximum likelihood estimates rather than Yule-Walker estimates. We can think of using conditional maximum likelihood estimates as being true the to MDL principle since these estimates then result in the exact -2log(likelihood) term (assuming a normal distribution on the errors) rather than the approximation that results from using Yule-Walker estimates. In the end, the Auto-PARM estimates are weakly consistent in both cases.

equivalent to choosing the model with the minimum code length $CL_{\mathcal{F}}(\boldsymbol{y})$. If the number of change-points m is zero, $\log m$ is taken to be zero. Likewise, if an AR order p_j is zero, $\log p_j$ is defined to be zero. Of the five terms in (1.2), the first term represents the code length for m, the second the code length for $n_1, \ldots, n_{m+1}$ (which determine $\tau_1, \ldots, \tau_m$), the third the code length for the AR orders $p_1, \ldots, p_{m+1}$, the fourth the code length for the AR model parameters, and the fifth the code length for the residuals given the fitted model. The best-fitting model is then found by minimizing the MDL with respect to the number of change-points, m, the change-point locations, $\tau_1, \ldots, \tau_m$, and the AR orders, $p_1, \ldots, p_{m+1}$. Note that once these parameters are specified, the AR parameters within each fitted segment, $\boldsymbol{\psi}_j$, can be easily estimated using maximum likelihood. This minimization is carried out using a genetic algorithm which is a numerical minimization technique that mimics natural evolution. See Pasia et al. (2005) for a review of genetic algorithms and Davis et al. (2006) for implementation details in Auto-PARM.

1.3 Statement of the Problem

This dissertation explores the theory behind Auto-PARM. Davis et al. (2006) showed that if the true number of change-points is known, the estimated change-point locations obtained by Auto-PARM are consistent for the true change-point locations if the underlying model is assumed to be piecewise autoregressive. They also demonstrated through simulations that the estimated number of change-points seems to be a consistent estimator for the true number of change-points. This research will prove that the estimated number of change-points and the estimated AR orders are weakly consistent for the true number of change-points and the true AR orders, respectively, and under certain circumstances, are strongly consistent. In addition, we will show that the estimated number and locations of change-points are weakly consistent assuming only that the underlying process is piecewise stationary, strongly mixing and

satisfies a few other minor assumptions. This has implications for Auto-PARM's applicability to real data in which the assumption of a piecewise autoregressive model is overly restrictive.

Previous consistency proofs for the number of change-points in the change-point literature have been demonstrated by, e.g., Yao (1988), Lee (1995), Lee (1996), Lee (1997), and Kuhn (2001). Many of these papers use a modified version of the BIC for model selection, and all of the papers aside from Kuhn (2001) assume independent observations. In general, we will prove consistency by considering the difference between the MDL with the true number of change-points and the MDL with the wrong number of change-points. Strong or weak consistency follows if the MDL for the wrong model is larger than the MDL for the correct model for large n almost surely or in probability, respectively. An outline of the results is as follows.

In Chapter 2, we show that, assuming the true process follows a piecewise autoregressive model, the estimated number of change-points and the estimated AR orders are consistent for the true number of change-points and the true AR orders, respectively. As a tool for the proofs, in Lemmas 2.1 and 2.2 and Theorem 2.1, we use conditional maximum likelihood estimates in the MDL rather than Yule-Walker estimates. We first address the case where the true process has no change-points. Lemma 2.1 assumes that the AR order of the true process is known, and shows that the MDL using $m \geq 1$ change-points is larger than the MDL using no change-points for large n with probability 1. Lemma 2.2 extends Lemma 2.1 to the case where the AR orders are estimated from the data. Lemmas 2.1 and 2.2 imply that the estimated number of change-points and the estimated AR orders are strongly consistent for the true number of change-points (0) and the true AR order when the underlying process has no change-points. Theorem 2.1 uses Lemmas 2.1 and 2.2 to prove that if the true process has $m_0 \geq 0$ change-points, then the estimated number of change-points and the estimated AR orders are weakly consistent for m_0 and the true AR orders. The last section of Chapter 2 shows that the estimated number of change-points and the

estimated AR orders are still weakly consistent for the true number of change-points and the true AR orders when we use Yule-Walker estimates in the MDL. Lemmas 2.3 and 2.4 show weak consistency in the case where the underlying process has no change-points, and Theorem 2.2 shows weak consistency when the underlying process has $m_0 \geq 0$ change-points.

Chapter 3 shows consistency of the Auto-PARM estimates when the underlying process is not necessarily piecewise autoregressive, but piecewise stationary and strongly mixing, and satisfies some other general assumptions. The second section mimics the consistency proofs in Davis et al. (2006) to show that under this general model, when the number of change-points is known, if we choose a large enough AR order for each fitted segment, the estimated change-point locations are again strongly consistent for the true change-point locations. This result is stated as Theorem 3.1. In the third section, we adapt the arguments for Lemma 2.1 and Theorem 2.1 to show that the estimated number of change-points is weakly consistent for the true number of change-points if large enough AR orders are fit to each segment. These results are stated as Lemma 3.1 and Theorem 3.2. In the last section of Chapter 3, we describe some simulations using Auto-PARM on non-autoregressive data.

Chapter 4 addresses a specific application of Auto-PARM to sound data obtained by the National Park Service. In this project, the National Park Service was interested in determining the proportion of man-made sounds in the parks. We used Auto-PARM to break the sound waves into approximately stationary segments, and then applied a classification algorithm to each segment. We also explored a method of windowing the sound wave and using linear discriminant analysis to classify each window. This second method of windowing the data showed promising results, and future work may include examining this problem in more detail possibly using statistical learning methods.

Chapter 2

CONSISTENCY OF AUTO-PARM ESTIMATES FOR A PIECEWISE AR PROCESS

2.1 Introduction

Consider the problem of modeling a nonstationary time series by segmenting the series into blocks of different autoregressive (AR) processes. Let m denote the number of change-points, and for $k = 1, \ldots, m$, denote the change-point between the kth and $(k+1)$st AR processes as τ_k, where $\tau_0 := 1$ and $\tau_{m+1} := n+1$. Let $p_1, \ldots, p_{m+1}$ denote the AR orders of the $m+1$ segments. As described in Chapter 1, the number and locations of change-points plus the autoregressive orders, $(m, \tau_1, \ldots, \tau_m, p_1, \ldots, p_{m+1})$, are estimated by minimizing the minimum description length (MDL),

$$\begin{aligned}\mathrm{MDL}(m,\tau_1,\ldots,\tau_m;p_1,\ldots,p_{m+1}) &= \log m + (m+1)\log n \\ &+ \sum_{k=1}^{m+1} \log p_k + \sum_{k=1}^{m+1} \frac{p_k+2}{2} \log n_k + \sum_{k=1}^{m+1} \frac{n_k}{2} \log(2\pi\hat{\sigma}_k^2), \end{aligned} \tag{2.1}$$

where $\hat{\sigma}_k^2$ is the conditional maximum likelihood estimate of the noise variance when fitting a p_kth order AR model to the kth segment, and n_k is the number of observations in the kth segment, $k = 1, \ldots, m+1$. Note that the dependence of the minimum description length on the autoregressive coefficient parameter estimates is only through the white noise estimates, $\hat{\sigma}_k^2$.

This chapter shows consistency of the estimates obtained by minimizing the MDL. Davis et al. (2006) showed that if the true process follows a piecewise autoregressive model, and the number of change-points is known, then the estimated

change-point locations are strongly consistent. We will prove that the estimate of the number of change-points and the estimated AR orders are weakly consistent. The proof of consistency will rely on detailed behavior of the sample covariances, which we examine through the functional law of the iterated logarithm. In the next section, we will introduce the functional law of the iterated logarithm and show how it applies to the sample covariances of autoregressive processes. In Section 2.3, we will review the piecewise autoregressive model introduced in Chapter 1. Conditional maximum likelihood estimation will be discussed in Section 2.4. Section 2.5 proves consistency of the estimated number of change-points and the AR orders for the case where there are no true change-points. Section 2.6 uses the results from Section 2.5 to show the main consistency result for the estimates of the number of change-points and the AR orders.

The MDL defined in (2.1) uses conditional maximum likelihood estimates for the white noise variances. Alternatively, Davis et al. (2006) use Yule-Walker estimates in their implementation of the MDL. We use conditional maximum likelihood estimates as a tool to simplify the consistency proofs. In practice, Yule-Walker estimates are preferable due to computational speed and stability (numerical), and resulting causal estimates. We will address this difference in Section 2.7, denoting the MDL using Yule-Walker estimates as MDL_Y, and show how the consistency results extend to the case where the MDL is defined with Yule-Walker estimates. The only difference in the consistency results between conditional maximum likelihood estimation and Yule-Walker estimation is in the case where the true process has no change-points. When we use conditional maximum likelihood estimation in this case, we can show that consistency holds almost surely. However, when using Yule-Walker estimation, because Yule-Walker estimates use partial sums of the sample autocovariance function in which the sum index starts at $t = 1 + h$ rather than from $t = 1$, we can only show consistency in probability.

2.2 Functional Law of the Iterated Logarithm

An underlying principle used in this dissertation when proving consistency is the functional law of the iterated logarithm. Some background on this principle will help in understanding the proofs that follow. Suppose we are given a sequence $\{Y_t\}$ of independent and identically distributed random variables with mean 0 and variance 1, and define the partial sum

$$S_n = \sum_{i=1}^{n} Y_i.$$

There are three fundamental asymptotic results about the behavior of S_n as $n \to \infty$: the strong law of large numbers, the central limit theorem, and the law of the iterated logarithm. The *strong law of large numbers* states that

$$\frac{S_n}{n} \to 0 \ a.s.$$

The *central limit theorem* states that

$$\frac{S_n}{\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1).$$

Notice that the strong law of large numbers normalizes S_n by dividing by n, which results in a constant, the mean, as its limit. On the other hand, the central limit theorem provides a normalizing factor for S_n that results in a distributional limit with a random variable as its limit. This difference in limits is due to the size of the normalizing factor. In the strong law of large numbers, the normalizing factor n is too large to determine an exact order of convergence for S_n, but in the central limit theorem, the normalizing factor $\sqrt{n}$ is too small. The *law of the iterated logarithm* addresses the delicacy of the normalizing factor by giving the exact normalizing factor necessary to insure that the supremum of the limit points is 1 and the infimum of the limit points is -1. It states that

$$P\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1\right) = 1,$$

i.e.,

$$S_n = O\left(\sqrt{n \log\log n}\right) \ a.s.$$

Equivalently, the law of the iterated logarithm can be stated in terms of the sample mean:

$$P\left(\limsup_{n\to\infty} \frac{S_n/n}{\sqrt{\frac{2}{n}\log\log n}} = 1\right) = 1.$$

Heyde and Scott (1973) proved a law of the iterated logarithm for stationary processes, and Hannan and Quinn (1979) used this result on sample partial autocorrelations to show consistency of AR order estimates. We will use the law of the iterated logarithm on sample covariances within each stationary segment of the process. However, since the boundaries between fitted segments are not fixed, we need a functional law of the iterated logarithm to determine the order of convergence for sample covariances with random boundaries found by minimizing the MDL.

Strassen (1964) defined a functional law of the iterated logarithm for independent and identically distributed processes. Consider the continuous function η_n on $[0,1]$ obtained by linearly interpolating

$$\eta_n\left(\frac{k}{n}\right) := \frac{S_k}{\sqrt{2n\log\log n}}$$

at $t = k/n$, $k = 1, \ldots, n$.[2] In other words, for $k \le nt \le k+1$,

$$\eta_n(t) := (2n\log\log n)^{-1/2}\left(S_k + (nt-k)Y_{k+1}\right).$$

Consider the set

$$K := \{f \in C[0,1] : f(0) = 0,\ f \text{ absolutely continuous, } \int_0^1 f'^2 \le 1\}.$$

Note that if $f \in K$, then $|f(t)| \le 1$ for all $t \in [0,1]$ and K is uniformly bounded [5]. Let $L(\{\eta_n\})$ be the set of a.s. limit points of the sequence of functions $\{\eta_n\}$.

---

[2]The function η_n need not be a linear interpolation. Any continuous function such that the values of $\eta_n(k/n)$ match the values $S_k/\sqrt{2n\log\log n}$ is sufficient.

Then Strassen's functional law of the iterated logarithm (FLIL) states that $L(\{\eta_n\})$ coincides with K a.s., and for any continuous functional θ on $C[0,1]$, $L(\{\theta(\eta_n)\})$ coincides with $\theta(K)$ a.s. For example, letting

$$\theta(\eta_n) := \eta_n(1) = \frac{S_n}{\sqrt{2n \log\log n}}$$

results in the basic law of the iterated logarithm.

2.2.1 Applying the FLIL to Autoregressive Processes

Throughout the consistency proofs in this dissertation, we use the functional law of the iterated logarithm on the sample covariances and sample means of autoregressive processes. In this subsection, we will describe how to apply the FLIL to AR processes and discuss sufficient conditions in order for the FLIL to hold.

Suppose the process $\{X_t\}$ follows the causal AR(p) model with mean μ

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \sigma\epsilon_t$$

where $\{\epsilon_t\}$ is a white noise process with mean zero and unit variance. It is always assumed that the AR process is causal. That is, we can write

$$X_t - \mu = \sum_{j=0}^{\infty} \psi_j \, \sigma\epsilon_{t-j},$$

for $t = 0, \pm 1, \ldots$, where $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and $\psi_0 = 1$. The coefficients ψ_j are determined by the relation $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = 1/\phi(z)$ where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$.

If we observe a sample of size n from this process, the sample autocovariance function (ACVF) is defined for $0 \le h \le n-1$ as

$$\begin{aligned} \hat{\gamma}(h) &= \frac{1}{n} \sum_{t=1+h}^{n} (X_t - \overline{X})(X_{t-h} - \overline{X}) \qquad (2.2) \\ &= \frac{1}{n} \sum_{t=1+h}^{n} X_t X_{t-h} - \overline{X}^2, \end{aligned}$$

where $\overline{X}$ is the sample mean. Let $\gamma(h) = Cov(X_t, X_{t-h})$ be the true covariance between X_t and X_{t-h}. Then Hannan (1974) showed, under very general conditions, the uniform convergence of sample autocovariances a.s. to the true covariances,

$$\lim_{n\to\infty} \sup_{0\le h<\infty} |\hat{\gamma}(h) - \gamma(h)| = 0, \quad \text{a.s.},$$

where $\hat{\gamma}(h) := 0$ for $h \ge n$. An et al. (1982) improved on the order of this result by showing that if $p(n) = O(\log n)^a$ for some $a < \infty$, then

$$\max_{0\le h\le p(n)} |\hat{\gamma}(h) - \gamma(h)| = O\left(\sqrt{\frac{1}{n}\log\log n}\right). \tag{2.3}$$

For a sequence $\{X_t\}$ of real-valued random variables with finite variance defined on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, one measure of dependence, introduced by Rosenblatt (1956), is the set of strong mixing coefficients. For any two sigma algebras $\mathcal{A}$ and $\mathcal{B}$ in $(\Omega, \mathcal{F}, \mathrm{P})$, let

$$\begin{aligned} \alpha(\mathcal{A}, \mathcal{B}) &= \sup_{A\in\mathcal{A};\, B\in\mathcal{B}} |\mathrm{P}(A\cap B) - \mathrm{P}(A)\mathrm{P}(B)| \\ &= \sup_{A\in\mathcal{A};\, B\in\mathcal{B}} |Cov(\mathrm{I}_A \mathrm{I}_B)| \le 1/4. \end{aligned}$$

Then the strong mixing coefficients $\{\alpha_n\}_{n>0}$ of the sequence $\{X_t\}$ are defined by $\alpha_n = \sup_{k\in\mathbb{Z}} \alpha(\mathcal{F}_k, \mathcal{G}_{k+n})$, where $\mathcal{F}_k = \sigma(X_t : t \le k)$ and $\mathcal{G}_l = \sigma(X_t : t \ge l)$. We make the convention that $\alpha_0 = 1/4$. The sequence $\{X_t\}$ is called a *strong mixing sequence* if $\lim_{n\to\infty} \alpha_n = 0$. One can think of a strong mixing sequence as one in which observations become independent as the lag between them tends to infinity. Rio (1995) showed that the functional law of the iterated logarithm holds for stationary strong mixing sequences under the following condition. Suppose $\{X_i\}_{i\in\mathbb{Z}}$ 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 $\alpha(t) = \alpha_{[t]}$ and denote the quantile function of $|X_0|$ by Q. Then the functional law of the iterated logarithm holds for the sequence $\{X_i\}_{i\in\mathbb{Z}}$ if

$$\int_0^1 \alpha^{-1}(t) Q^2(t) dt < \infty. \tag{2.4}$$

This condition simplifies if the process is strong mixing at a geometric rate. In this case, the functional law of the iterated logarithm holds if

$$E\left(X_0^2 \log^+ |X_0|\right) < \infty \tag{2.5}$$

(see Rio (1995) for proof).

In order to apply the functional law of the iterated logarithm to sample covariances of a stationary strong mixing process, the cross products $\{X_t X_{t-h}\}$ in the sample covariances must also satisfy the assumptions in Rio (1995). Assume for notational convenience that the mean of the process is zero. Under the assumption that the mean is zero, we can use the function

$$\hat{\gamma}^*(h) := \frac{1}{n}\sum_{t=1}^{n} X_t X_{t-h} \tag{2.6}$$

for the sample covariance rather than (2.2) since (2.6) has the same asymptotic properties as (2.2) (see Brockwell and Davis (1991) p. 226). Assume the process $\{X_t\}$ is strong mixing at a geometric rate. Then the processes $\{X_t X_{t-h}\}$ are also strong mixing at a geometric rate, and we can apply the functional law of the iterated logarithm to the sample covariances as long as the cross product processes $\{X_t X_{t-h}\}$ satisfy (2.5). That is, the process $\{X_t\}$ must satisfy the condition

$$E\left(X_0^4 \log^+(X_0^2)\right) < \infty. \tag{2.7}$$

Note that condition (2.7) is automatic if the $(4+\delta)$th moment is finite for some $\delta > 0$. Thus, assuming strong mixing at a geometric rate and (2.7), if we let

$$\theta(\eta_n) := \inf_{0\le t\le 1}\{\eta_n(t)\} = \inf_k \left\{\frac{S_k}{\sqrt{2n\log\log n}}\right\},$$

where

$$S_k = \sum_{t=1}^{k}\left[\frac{X_t X_{t-h} - E(X_t X_{t-h})}{\sqrt{Var(X_t X_{t-h})}}\right],$$

then we can apply the functional law of the iterated logarithm to sample covariances calculated using the change-point locations that minimize the MDL. In other words, e.g., for an AR(p) process $\{X_t\}$ with mean zero, the function

$$f(\lambda) := \inf_{0<\lambda<1} \left\{ \frac{(\hat{\gamma}^*_{0:\lambda}(h) - \gamma(h))/\sqrt{Var(X_t X_{t-h})}}{\sqrt{\frac{2}{n}\log\log n}} \right\}$$

is bounded by one in absolute value a.s., where $\hat{\gamma}^*_{0:\lambda}(h)$ is the sample covariance calculated between observations X_1 and $X_{[\lambda t]}$ and $\gamma(h)$ is the true autocovariance function of the process. This implies that $\inf_{0<\lambda<1}(\hat{\gamma}_{0:\lambda}(h) - \gamma(h))$ is $O(\sqrt{\log\log n/n})$ for each fixed h, and thus, $\sup_{0\le h\le P}\inf_{0<\lambda<1}(\hat{\gamma}_{0:\lambda}(h) - \gamma(h))$ is $O(\sqrt{\log\log n/n})$ for some upper bound $P < \infty$.

Remark 1. When examining sample covariances, we can assume without loss of generality that the process mean μ is zero since the distribution of $X_t - \overline{X}$ is invariant to μ. To show that the FLIL holds for the mean-corrected cross products used in the sample covariance function (2.2), consider, e.g.,

$$\frac{1}{n}\sum_{t=1}^{[sn]}(X_t - \overline{X})^2 = (X_t - \mu)^2 - 2(\overline{X} - \mu)(X_t - \mu) + (\overline{X} - \mu)^2, \tag{2.8}$$

where $s \in (0, 1)$. If we subtract off the variance of the process from the first term of (2.8), this term satisfies the functional law of the iterated logarithm. The second two terms in (2.8) are both of order $\log\log n/n$ since each consists of a product of two random variables which each satisfy the FLIL. By the same argument, the FLIL will hold for cross-covariances $(1/n)\sum_{t=1}^{[sn]}(X_t - \overline{X})(X_{t-h} - \overline{X})$.

In many of the proofs, we use conditional maximum likelihood estimates for the AR parameters. This estimation method will be discussed in more detail in Section 2.4, but we will state here the assumptions necessary for the FLIL to hold in this case. When fitting an AR(p) process, conditional maximum likelihood estimation

uses a definition of the sample covariance that includes initial unobserved values $X_{-p}, X_{-p+1}, \ldots, X_0$. 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 covariance, where $\overline{X}_{a:b} := \sum_{t=a}^{b} X_t/(b-a+1)$. It follows from previous arguments that for a stationary strong mixing process $\{X_t\}$ satisfying (2.7), the FLIL holds for $\hat{\gamma}^*(h)$. However, in order to apply the FLIL to these sample covariances when using conditional maximum likelihood estimation, we must show that the FLIL holds for $\hat{\gamma}^*(h)$ when we condition on any initial values $X_{-p}, X_{-p+1}, \ldots, X_0$.

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

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

where a_{it} is a function, depending on t, of sums and products of $\phi_1, \ldots, \phi_p$ for $i = 0, \ldots, p$. Defining the stationary process

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

it follows that

$$X_t - X_t' = \sum_{j=0}^{\infty} \psi_{t+j} \, \sigma\epsilon_{-j} + a_{0t}(X_0 - \mu) + \cdots + a_{pt}(X_{-p} - \mu).$$

If we assume that $\sum_{j=t}^{\infty} \psi_{t+j} \, \sigma\epsilon_{-j}$ tends to zero at a geometric rate as t goes to infinity, then since each coefficient a_{it} tends to zero at a geometric rate, $X_t - X_t'$ also tends to zero at a geometric rate a.s. as t goes to infinity.

Remark 2. We can show that the functional law of the iterated logarithm applies to the sample covariances of the process $\{X_t\}$ conditioned on initial values $X_{-P}, X_{-P+1}, \ldots, X_0$ for some prespecified upper bound P by showing that the difference between the sample covariances for this process and the sample covariances

for the stationary process $\{X_t'\}$ over an interval $[1,[sn]]$ are of order $\sqrt{\log\log n/n}$ uniformly in s where $s \in (0,1)$. This can be shown by considering the difference

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{[sn]}(X_tX_{t-h} - X_t'X_{t-h}') \\
&= \frac{1}{n}\sum_{t=1}^{[sn]}(X_t - X_t')X_{t-h} + \frac{1}{n}\sum_{t=1}^{[sn]}(X_{t-h} - X_{t-h}')X_t' \\
&=: \mathrm{I} + \mathrm{II}.
\end{aligned}$$

Note that $|\psi_j| \le Cr^j$ for some fixed $C > 0$ and $0 < r < 1$. Therefore,

$$\begin{aligned}
|\mathrm{I}| &\le \frac{1}{n}\sum_{t=1}^{[sn]} a_t^* \left(|X_{-p}| + \cdots + |X_0|\right)|X_{t-h}| + \frac{C}{n}\sum_{t=1}^{[sn]}\left(\sum_{j=0}^{\infty} r^{t+j}|\epsilon_{-j}||X_{t-h}|\right) \\
&= \frac{1}{n}\sum_{t=1}^{[sn]}\left[a_t^*(|X_{-p}| + \cdots + |X_0|) + C\left(\sum_{j=0}^{\infty} r^j|\epsilon_{-j}|\right)r^t\right]|X_{t-h}|
\end{aligned}$$

Choose $a > 0$ such that $a\log r < -1/2$ and write

$$\begin{aligned}
|\mathrm{I}| &= \frac{1}{n}\sum_{t=1}^{[a\log n]}(\,\cdot\,) + \frac{1}{n}\sum_{t=[a\log n]+1}^{[sn]}(\,\cdot\,) \\
&=: \mathrm{A} + \mathrm{B}.
\end{aligned}$$

Examining the first sum, we see that

$$\mathrm{A} \le \left[C(|X_{-p}| + \cdots + |X_0|) + C\sum_{j=0}^{\infty} r^j|\epsilon_{-j}|\right]\frac{1}{n}\sum_{t=1}^{[a\log n]}|X_{t-h}|.$$

Now

$$\frac{1}{n}\sum_{t=1}^{[a\log n]}|X_{t-h}| \le \frac{1}{n}[a\log n]M_{[a\log n]},$$

where $M_{[a\log n]} = \max_{1\le t\le[a\log n]}|X_{t-h}|$. By Markov's inequality, we have

$$\begin{aligned}
&\sum_n P\left(\frac{a\log n}{n}\left(\sqrt{\frac{\log\log n}{n}}\right)^{-1} M_{[a\log n]} > \epsilon\right) \\
&\le \sum_n a\log n\; P\left(\frac{a\log n}{n}\left(\sqrt{\frac{\log\log n}{n}}\right)^{-1}|X_{t-h}| > \epsilon\right) \\
&\le \sum_n \epsilon^{-4}\frac{(a\log n)^5}{n^2}\,\frac{E|X_{t-h}|^4}{(\log\log n)^2}.
\end{aligned}$$

Thus, by Borel-Cantelli,

$$\frac{\log n}{n} M_{[a \log n]} = o\left(\sqrt{\frac{\log \log n}{n}}\right) \quad \text{a.s.}$$

Turning to B, we have

$$\begin{aligned} |\mathrm{B}| &\leq Cr^{a \log n}(|X_{-p}| + \cdots + |X_0|)\frac{1}{n}\sum_{t=1}^{n}|X_{t-h}| \\ &\quad + \frac{C}{n}\left(\sum_{j=0}^{\infty} r^j |\epsilon_{-j}|\right) \sum_{t=[a \log n]+1}^{n} r^t |X_{t-h}| \\ &\leq o\left(\sqrt{\frac{\log \log n}{n}}\right) + O(1) r^{a \log n} \frac{1}{n}\sum_{1}^{n}|X_{t-h}| \\ &= o\left(\sqrt{\frac{\log \log n}{n}}\right), \end{aligned}$$

by the choice of a.

2.3 Piecewise Autoregressive Process

In order to show that the estimates obtained by minimizing the MDL are consistent, we need to assume a true model for the data. Throughout this dissertation 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). We assume in this chapter that the true process follows a piecewise autoregressive model with m_0 change-points 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, where $\tau_0^0 := 1$ and $\tau_{m_0+1}^0 := n + 1$. We define $\boldsymbol{\lambda}^0 = (\lambda_1^0, \ldots, \lambda_{m_0}^0)$ such that $0 < \lambda_1^0 < \cdots < \lambda_{m_0}^0 < 1$ and $\tau_k^0 = [\lambda_k^0 n]$ for $k = 1, \ldots, m_0$, where $[x]$ denotes the integer part of x. We set $\lambda_0^0 := 0$ and $\lambda_{m_0+1}^0 := 1$, and define $[\lambda_0^0 n] := 1$ and $[\lambda_{m_0+1}^0 n] := n + 1$. Defined in this way, knowledge of λ_k^0 determines τ_k^0, and both will equivalently be referred to as "change-points", λ_k^0 being a "relative change-point". This is the standard setup when deriving asymptotic results in the change-point problem. As n tends to infinity, the number of observations in each segment, n_k, also tends to infinity, but time is re-scaled to the interval [0,1] (see [13]).

Let $\{\epsilon_{k,t}\}$, $k = 1, \ldots, m_0 + 1$, be independent sequences of independent and identically distributed (iid) random variables with mean zero and unit variance. Then for given initial values X_{-P}, X_{-P+1}, ..., X_0 with P a preassigned upper bound, AR coefficient parameters $\phi_{k,j}$, $k = 1, \ldots, m_0 + 1$, $j = 1, \ldots, p_k^0$, and noise parameters σ_1, ..., σ_{m_0+1}, the *piecewise autoregressive process* $\{X_t\}$ is defined as

$$X_t = \phi_{k,0} + \phi_{k,1} X_{t-1} + \cdots + \phi_{k,p_k^0} X_{t-p_k^0} + \sigma_k \epsilon_{k,t} \qquad \text{for} \qquad t \in [\tau_{k-1}, \tau_k). \tag{2.9}$$

where $\boldsymbol{\psi}_k := (\phi_{k,0}, \phi_{k,1}, \ldots, \phi_{k,p_k^0}, \sigma_k)$ is the parameter vector corresponding to the causal AR(p_k) process in the kth segment and $\boldsymbol{\psi}_k \neq \boldsymbol{\psi}_{k+1}$ for any $k = 1, \ldots, m_0$ (i.e., between consecutive segments, at least one of the AR coefficients, the process mean, the white noise variance, or the AR order must change). Note that if we denote the mean of $\{X_t\}$ in the kth segment by μ_k, then 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

$$X_t - \mu_k = \phi_{k,1}(X_{t-1} - \mu_k) + \cdots + \phi_{k,p_k^0}(X_{t-p_k^0} - \mu_k) + \sigma_k \epsilon_{k,t}.$$

In order for the functional law of the iterated logarithm to apply to the sample covariances of this process, we will need to assume throughout this dissertation that the process

A1. is strong mixing at a geometric rate, and

A2. satisfies the moment condition (2.7) within each segment.

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. Choose $\epsilon > 0$ small such that

$$\epsilon << \min_{i=1,\ldots,m_0+1} \left(\lambda_i^0 - \lambda_{i-1}^0\right), \tag{2.10}$$

and set

$$A_m^\epsilon = \{(\lambda_1, \ldots, \lambda_m) : 0 < \lambda_1 < \cdots < \lambda_m < 1, \lambda_k - \lambda_{k-1} \geq \epsilon, k = 1, \ldots, m+1\}, \tag{2.11}$$

where $\lambda_0 := 0$ and $\lambda_{m+1} := 1$. Setting $\boldsymbol{p}^0 = (p_1^0, \ldots, p_{m_0+1}^0)$, the parameters m_0, $\boldsymbol{\lambda}^0$, and $\boldsymbol{p}^0$ are then estimated by minimizing the MDL over $m \leq M$, $0 \leq \boldsymbol{p} \leq P$, and $\boldsymbol{\lambda} \in A_m^\epsilon$, where M and P are preassigned upper bounds for m and p_k. Denote these estimates by

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

Though not explicit in the notation, it is important to note when studying asymptotic properties that the estimates $\hat{m}$, $\hat{\boldsymbol{\lambda}}$, and $\hat{\boldsymbol{p}}$ all depend on n.

Davis et al. (2006) showed 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, $\hat{\boldsymbol{\lambda}} \to \boldsymbol{\lambda}^0$ a.s. as $n \to \infty$. We would like to 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.

For large n, it is easy to show that the estimated AR orders will not underestimate the true AR order, so in our consistency proofs, we only need to consider the case where the estimated AR orders might overestimate the true AR order. To see this, consider the special case where $m_0 = 0$ and the true AR order is p where the pth AR coefficient, ϕ_p, is not zero. Suppose we fit an AR$(p-1)$ model to the observations. Then the estimated variance converges to the minimum of $E(X_p - a_0 - a_1 X_{p-1} - \cdots - a_{p-1} X_1)^2$ over all $a_0, \ldots, a_{p-1}$, which is greater than the true noise variance. Therefore, underestimating the true AR order results in an increase in the last term of (2.1) by an amount of $O(n)$. Likewise, for large n, the estimated number of change-points, $\hat{m}$, will not underestimate the true number of change-points. Consider the MDL using m change-points where $m < m_0$. Even though the penalty term for the MDL with m change-points is smaller than the penalty term for the MDL with m_0 change-points, when we divide by n, the difference in penalty terms goes to zero. The estimated white noise variance(s) under a model with too few change-points will be larger than the true white noise variance(s) for large n with probability 1, and thus,

the difference $(\mathrm{MDL}(m) - \mathrm{MDL}(m_0))/n$ converges to a positive constant as n tends to infinity.

2.4 Conditional Maximum Likelihood Estimation

Before stating any results, we develop notation for the conditional maximum likelihood estimates of the autoregressive parameters. Consider fitting an AR(p_k) model with mean μ_k to the kth segment, i.e., the segment starting with observation $\tau_{k-1} = [\lambda_{k-1}n]$ and ending with observation $\tau_k - 1 = [\lambda_k n] - 1$, $1 \leq k \leq m+1$. For simplicity, let $a = \tau_{k-1}$, $b = \tau_k - 1$, and $p = p_k$. Assume we are given initial values $X_{-P}, X_{-P+1}, \ldots, X_0$ for the process. Let

$$\mathbf{X}_{a:b} = \begin{pmatrix} X_a \\ X_{a+1} \\ \vdots \\ X_b \end{pmatrix}, \tag{2.12}$$

and

$$\mathbf{M}^p_{a:b} = \begin{pmatrix} 1 & X_{a-1} & X_{a-2} & \ldots & X_{a-p} \\ 1 & X_a & X_{a-1} & \ldots & X_{a-p+1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{b-1} & X_{b-2} & \ldots & X_{b-p} \end{pmatrix}. \tag{2.13}$$

Then the projection of $\mathbf{X}_{a:b}$ onto the linear subspace spanned by the columns of $\mathbf{M}^p_{a:b}$ is

$$\mathrm{P}_{\mathbf{M}^p_{a:b}}(\mathbf{X}_{a:b}) = \mathbf{M}^p_{a:b}\hat{\boldsymbol{\psi}}^p_{a:b},$$

where

$$\begin{aligned} \hat{\boldsymbol{\psi}}^p_{a:b} &= \left({\mathbf{M}^p_{a:b}}^T \mathbf{M}^p_{a:b}\right)^{-1} {\mathbf{M}^p_{a:b}}^T \mathbf{X}_{a:b} \\ &= \left(\hat{\phi}^{p0}_{a:b}, \hat{\phi}^{p1}_{a:b}, \ldots, \hat{\phi}^{pp}_{a:b}\right)^T =: \left(\hat{\phi}^{p0}_{a:b}, \hat{\boldsymbol{\phi}}^p_{a:b}\right)^T \end{aligned}$$

is the conditional maximum likelihood estimate of the intercept $\phi^{p0}_{a:b}$ and autoregressive coefficients, $\boldsymbol{\phi}^p_{a:b} = (\phi^{p1}_{a:b}, \ldots, \phi^{pp}_{a:b})^T$, when an AR($p$) model is fit to the kth segment

$[a, b]$. The conditional maximum likelihood estimate of σ_k^2 is then

$$\begin{aligned}\hat{\sigma}_{k,p}^2 &= \frac{1}{b-a+1}\left\|\mathbf{X}_{a:b} - \mathbf{M}_{a:b}^p \hat{\boldsymbol{\psi}}_{a:b}^p\right\|^2 \\ &= \frac{1}{[(\lambda_k - \lambda_{k-1})n]}\left\|\mathbf{X}_{a:b} - \mathbf{M}_{a:b}^p \hat{\boldsymbol{\psi}}_{a:b}^p\right\|^2. \end{aligned} \tag{2.14}$$

For the remainder of this paper, we will use the divisor $(\lambda_k - \lambda_{k-1})n$ rather than $[(\lambda_k - \lambda_{k-1})n]$ in (2.14). Since

$$\frac{1}{[xn]} - \frac{1}{xn} \leq \frac{1}{[xn]^2} = O\left(\frac{1}{n^2}\right),$$

this substitution will have no effect on the asymptotic results.

2.5 Case of No Change-Points ($m_0 = 0$)

We will prove consistency of the estimates for the number of change-points and the AR orders by first focusing on the case where there are no change-points in the underlying process. Lemmas 2.1 and 2.2 state strong consistency results for this specific case. Lemma 2.1 assumes the AR order is known, and Lemma 2.2 extends Lemma 2.1 to the case where the AR order is unknown and estimated by minimizing the MDL using conditional maximum likelihood estimation.

Assume throughout this section that the true process follows the causal AR(p) model with mean μ and no change-points

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \sigma\epsilon_t, \quad t = 1, \ldots, n, \tag{2.15}$$

where $\phi_0 = \mu(1 - \phi_1 - \cdots - \phi_p)$ and the noise sequence $\{\epsilon_t\}$ is iid with mean zero and unit variance. Initially, we will assume that the AR order p is known, and will fit the following two models to the dataset:

Model 1: Fit an AR(p) model with no change-points.

Model 2: Fit a piecewise AR(p) model with m relative change-points, $\boldsymbol{\lambda} \in A_m^{\epsilon}$.

The MDL for Model 1 is

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

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

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

and $\hat{\sigma}_k^2$ is the conditional maximum likelihood estimate of the AR(p) noise variance in the kth segment, $k = 1,\ldots,m+1$. The next result states that the estimate of the number of change-points is strongly consistent when there are no true change-points and the AR order is known.

Lemma 2.1. *Assume the true process $\{X_t\}$ follows the AR(p) model given in (2.15) with no change-points ($m_0 = 0$) and initial values X_{-P}, $X_{-P+1},\ldots,X_0$, and satisfies assumptions A1 and A2. Then with probability 1, for any $m \geq 1$,*

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

for n large.

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

$$\begin{aligned} & \frac{2}{n} \left[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p) - \mathrm{MDL}(0; p) \right] \\ &= \frac{2 \log m}{n} + m(p+4) \frac{\log n}{n} + \frac{2m \log p}{n} \\ &\quad + \frac{p+2}{n} \sum_{k=1}^{m+1} \log(\hat{\lambda}_k - \hat{\lambda}_{k-1}) \\ &\quad + \sum_{k=1}^{m+1} (\hat{\lambda}_k - \hat{\lambda}_{k-1}) \log \hat{\sigma}_k^2 - \log \hat{\sigma}^2. \end{aligned} \tag{2.16}$$

We will show that (2.16) is strictly positive for n large with probability 1. The first four terms in (2.16) are positive, and each of these terms converges to zero at a rate of either $1/n$ or $\log n/n$. 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

$$\sum_{k=1}^{m+1} (\hat{\lambda}_k - \hat{\lambda}_{k-1}) \log \hat{\sigma}_k^2 - \log \hat{\sigma}^2 \tag{2.17}$$

converges to zero as $n \to \infty$. We will use the functional law of the iterated logarithm on a Taylor series expansion of (2.17) 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.

First consider the kth fitted segment. For simplicity of notation, let $a = \hat{\tau}_{k-1} = [\hat{\lambda}_{k-1} n]$ and $b = \hat{\tau}_k - 1 = [\hat{\lambda}_k n] - 1$. Then the conditional maximum likelihood estimate for the mean and autoregressive coefficients in the kth segment is

$$\hat{\boldsymbol{\psi}}_{a:b} = \left(\mathbf{M}_{a:b}^{p\ T} \mathbf{M}_{a:b}^p \right)^{-1} \mathbf{M}_{a:b}^{p\ T} \mathbf{X}_{a:b}, \tag{2.18}$$

where $\mathbf{X}_{a:b}$ is defined in (2.12), and $\mathbf{M}_{a:b}^p$ is defined in (2.13). The conditional maximum likelihood estimate of the variance in the kth segment is

$$\hat{\sigma}_k^2 = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1}) n} \left\| \mathbf{X}_{a:b} - \mathbf{M}_{a:b}^p \hat{\boldsymbol{\psi}}_{a:b} \right\|^2. \tag{2.19}$$

Substituting (2.18) into (2.19) and taking norms, we obtain

$$\hat{\sigma}_k^2 = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1}) n} \left[\sum_{t=a}^{b} X_t^2 - \mathbf{X}_{a:b}^T \mathbf{M}_{a:b}^p \left(\mathbf{M}_{a:b}^{p\ T} \mathbf{M}_{a:b}^p \right)^{-1} \mathbf{M}_{a:b}^{p\ T} \mathbf{X}_{a:b} \right].$$

If we define the sample means $\overline{X}_{a-i:b-i} := \sum_{t=a}^{b} X_{t-i}/((\hat{\lambda}_k - \hat{\lambda}_{k-1})n)$ for $i = 0, 1, \ldots,$ and define the matrix $\mathbf{V} := \left\{ \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=a}^{b} X_{t-i} X_{t-j} \right\}_{i,j=1}^{p}$, then we can express ${\mathbf{M}_{a:b}^{p}}^T \mathbf{M}_{a:b}^{p}$ as

$$ {\mathbf{M}_{a:b}^{p}}^T \mathbf{M}_{a:b}^{p} = (\hat{\lambda}_k - \hat{\lambda}_{k-1})n \begin{pmatrix} 1 & \overline{\mathbf{X}}^T \\ \overline{\mathbf{X}} & \mathbf{V} \end{pmatrix}, $$

where $\overline{\mathbf{X}}$ is the p x 1 vector $\overline{\mathbf{X}} := \left(\overline{X}_{a-1:b-1}, \ldots, \overline{X}_{a-p:b-p}\right)^T$. Using the formula for the inverse of a block matrix (e.g., p. 51 of [6]), we can express the inverse of ${\mathbf{M}_{a:b}^{p}}^T \mathbf{M}_{a:b}^{p}$ as

$$ \left({\mathbf{M}_{a:b}^{p}}^T \mathbf{M}_{a:b}^{p}\right)^{-1} = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \begin{pmatrix} 1 + \overline{\mathbf{X}}^T (\mathbf{V} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^T)^{-1} \overline{\mathbf{X}} & -\overline{\mathbf{X}}^T (\mathbf{V} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^T)^{-1} \\ -(\mathbf{V} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^T)^{-1} \overline{\mathbf{X}} & (\mathbf{V} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^T)^{-1} \end{pmatrix}. $$

Similarly, we can write

$$ {\mathbf{M}_{a:b}^{p}}^T \mathbf{X}_{a:b} = (\hat{\lambda}_k - \hat{\lambda}_{k-1})n \begin{pmatrix} \overline{X}_{a:b} \\ \mathbf{w} \end{pmatrix}. $$

where $\mathbf{w}$ is defined to be the p x 1 vector

$$ \mathbf{w} := \left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=a}^{b} X_t X_{t-1}, \ldots, \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=a}^{b} X_t X_{t-p} \right)^T. $$

Note that the above expressions imply that the estimated AR parameter vector can be written as

$$ \hat{\boldsymbol{\psi}}_{a:b} = \begin{pmatrix} \overline{X}_{a:b} - \overline{\mathbf{X}}^T \left(\mathbf{V} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}\right)^{-1} \left(\mathbf{w} - \overline{\mathbf{X}}\,\overline{X}_{a:b}\right) \\ \left(\mathbf{V} - \overline{\mathbf{X}}\,\overline{\mathbf{X}}^T\right)^{-1} \left(\mathbf{w} - \overline{\mathbf{X}}\,\overline{X}_{a:b}\right) \end{pmatrix} = \begin{pmatrix} \hat{\phi}_{a:b}^{p0} \\ \hat{\boldsymbol{\phi}}_{a:b}^{p} \end{pmatrix}. $$

Now through tedious but straightforward algebra, by using the shortcut formula for sample covariances

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

we can write $\log \hat{\sigma}_k^2$ as a function of the sample covariances,

$$\log \hat{\sigma}_k^2 = g\left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=a}^{b} (X_{t-i} - \overline{X}_{a-i:b-i})(X_{t-j} - \overline{X}_{a-j:b-j}) : i, j = 0, \ldots, p \right),$$

where

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

Note that $u_{ij} = u_{ji}$ for all $i, j = 0, \ldots, p$, so that g is actually a function of only $p(p+1)/n + p + 1$ independent (or free) variables. However, for ease of notation, we will treat the vector $(u_{ij} : i, j = 0, \ldots, p)$ as $(u_{00}, u_{01}, \ldots, u_{0p}, u_{10}, u_{11}, \ldots, u_{1p}, \ldots, u_{p0}, u_{p1}, \ldots, u_{pp})$.

Since the $\log \hat{\sigma}_k^2$ terms only depend on the sample covariances, and the sample covariances are invariant to shifts in the mean, we can assume without loss of generality that the process mean is zero ($\mu = 0$) for the remainder of the proof. Under this assumption, let $\gamma(i) = E\left(X_t X_{t-i}\right)$ denote the true covariance between X_t and X_{t-i}, and let $\boldsymbol{\gamma} = (\gamma(|i-j|) : i, j = 0, \ldots, p)$ be the vector of covariances ranging over lags $0, \ldots, p$ defined in such a way to match the indices of the vectors of sample covariances,

$$\hat{\boldsymbol{\gamma}}_k := \left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=a_k}^{b_k} (X_{t-i} - \overline{X}_{a_k-i:b_k-i})(X_{t-j} - \overline{X}_{a_k-j:b_k-j}) : i, j = 0, \ldots, p \right),$$

for segments $k = 1, \ldots, m+1$ where $a_k := \hat{\tau}_{k-1}$ and $b_k := \hat{\tau}_k - 1$. Carrying out a second order Taylor expansion about $g(\boldsymbol{\gamma})$ on each of the $\log \hat{\sigma}_k^2$ terms and the $\log \hat{\sigma}^2$

term in (2.17), we obtain

$$\begin{aligned}
&\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\hat{\sigma}_k^2 - \log\hat{\sigma}^2 \\
&= \left[\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})g(\boldsymbol{\gamma}) - g(\boldsymbol{\gamma})\right] \\
&\quad + \left[\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\nabla g(\boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}) - \nabla g(\boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right] \\
&\quad + \frac{1}{2}\left[\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma})^T\nabla^2 g(\boldsymbol{\gamma}_k^*)(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma})\right. \\
&\qquad \left. - (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\nabla^2 g(\boldsymbol{\gamma}^*)(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\right] \qquad (2.20)
\end{aligned}$$

where $\hat{\boldsymbol{\gamma}} := (\frac{1}{n}\sum_{t=1}^{n}(X_{t-i} - \overline{X}_{1-i:n-i})(X_{t-j} - \overline{X}_{1-j:n-j}) : i,j = 0,\ldots,p)$. The term $\nabla g(\boldsymbol{\gamma})$ is the gradient of $g(u_{ij} : i,j = 0,\ldots,p)$ evaluated at $\boldsymbol{\gamma}$. The Hessians of $g(u_{ij} : i,j = 0,\ldots,p)$ evaluated at $\boldsymbol{\gamma}_k^*$ for $k = 1,\ldots,m+1$ and at $\boldsymbol{\gamma}^*$ are denoted by $\nabla^2 g(\boldsymbol{\gamma}_k^*)$ and $\nabla^2 g(\boldsymbol{\gamma}^*)$, respectively. The variables $\boldsymbol{\gamma}^*$ and $\boldsymbol{\gamma}_k^*$ are between $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$ or between $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}_k$, respectively, for $k = 1,\ldots,m+1$, and each variable converges to $\boldsymbol{\gamma}$ almost surely as n goes to infinity.

The first quantity in (2.20) is zero since $\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1}) = 1$. The second quantity is of order $O(\log\log n/n)$ since

$$\begin{aligned}
&\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}) - (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = \sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\hat{\boldsymbol{\gamma}}_k - \hat{\boldsymbol{\gamma}} \\
&= \sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\sum_{t=\hat{\tau}_{k-1}}^{\hat{\tau}_k - 1} X_{t-i}X_{t-j}\right. \\
&\qquad \left. - \overline{X}_{\hat{\tau}_{k-1}-i:\hat{\tau}_k-1-i}\overline{X}_{\hat{\tau}_{k-1}-j:\hat{\tau}_k-1-j} : i,j = 0,\ldots,p\right) \\
&\quad - \left(\frac{1}{n}\sum_{t=1}^{n} X_{t-i}X_{t-j} - \overline{X}_{1-i:n-i}\overline{X}_{1-j:n-j} : i,j = 0,\ldots,p\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\Bigg(\sum_{t=1}^{\hat{\tau}_1-1} X_{t-i}X_{t-j} + \sum_{t=\hat{\tau}_1}^{\hat{\tau}_2-1} X_{t-i}X_{t-j} + \cdots + \sum_{t=\hat{\tau}_m}^{n} X_{t-i}X_{t-j} \\
&\qquad - \sum_{t=1}^{n} X_{t-i}X_{t-j} : i,j = 0,\ldots,p \Bigg) \\
&\quad - \Bigg(\frac{1}{\hat{\lambda}_1} \cdot \frac{\sum_{t=1}^{\hat{\tau}_1-1} X_{t-i}}{n} \cdot \frac{\sum_{t=1}^{\hat{\tau}_1-1} X_{t-j}}{n} + \cdots + \frac{1}{(1-\hat{\lambda}_m)} \cdot \frac{\sum_{t=\hat{\tau}_m}^{n} X_{t-i}}{n} \cdot \frac{\sum_{t=\hat{\tau}_m}^{n} X_{t-j}}{n} \\
&\qquad - \frac{\sum_{t=1}^{n} X_{t-i}}{n} \cdot \frac{\sum_{t=1}^{n} X_{t-j}}{n} : i,j = 0,\ldots,p \Bigg) \\
&= \underline{0} + O\left(\frac{1}{n}\log\log n\right) = O\left(\frac{1}{n}\log\log n\right).
\end{aligned}
$$

The last line follows by the functional law of the iterated logarithm for partial sums of X_{t-i} (since we are assuming $\mu = 0$) and because $\hat{\lambda}_1$ and $(1-\hat{\lambda}_m)$ are bounded away from zero. When necessary, we can apply the functional law of the iterated logarithm by looking at the partial sums as

$$
\sum_{t=\hat{\tau}_{k-1}}^{\hat{\tau}_k-1} (\cdot) = \sum_{t=1}^{\hat{\tau}_k-1} (\cdot) - \sum_{t=1}^{\hat{\tau}_{k-1}-1} (\cdot).
$$

Thus, (2.20) becomes

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

Within the second order term of the Taylor series expansion, we can apply the functional law of the iterated logarithm to the sequences of mean-corrected cross-products as demonstrated in Section 2.2.1. It is then readily seen that the second order term in the Taylor series expansion is of order $\log\log n/n$ with probability 1. Thus, (2.16) becomes

$$
\begin{aligned}
&\frac{2}{n}\left[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p,\ldots,p) - \mathrm{MDL}(0; p)\right] \\
&\quad = m(p+4)\frac{\log n}{n} + O\left(\frac{1}{n}\right) + O\left(\frac{\log\log n}{n}\right), \qquad (2.21)
\end{aligned}
$$

which is greater than zero for large n with probability 1. □

Still under the assumption that the data follow the AR(p) process defined in (2.15), 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 dataset with m relative change-points, $\boldsymbol{\lambda} \in A_m^e$. Estimate the autoregressive orders from the data, and denote these orders by $\hat{p}_1, \ldots, \hat{p}_{m+1}$.

Then the minimum description length for Model 3 is

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

Recall that the minimum description length for Model 2 is

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

and the minimum description length for Model 1 is

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

The next lemma states that with probability 1, 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 2.2. *Assume the true process $\{X_t\}$ follows the AR(p) model given in (2.15) with no change-points ($m_0 = 0$) and initial values X_{-P}, $X_{-P+1}, \ldots, X_0$, and satisfies assumptions A1 and A2. Then with probability 1, for any $m \geq 1$,*

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

for n large.

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

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

We know from Lemma 2.1 that $\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p) - \mathrm{MDL}(0; p) > 0$ for n large with probability 1. Therefore, to prove Lemma 2.2, we need only show that

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

for n large with probability 1.

For Model 3, since the estimated AR orders will not underestimate the true AR order p for n large, it suffices to consider the case of fitting an autoregressive model of order $p+1$ to the kth segment, and autoregressive models of order p to each of the other m segments, so $\hat{p}_k = p + 1$ and $\hat{p}_j = p$ for $j \neq k$, where p is the true order of the process. Then

$$\begin{aligned} &\frac{2}{n}\left[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; \hat{p}_1, \ldots, \hat{p}_{m+1}) - \mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p)\right] \\ &= \frac{2(\log(p+1) - \log p)}{n} + \frac{\log(\hat{\lambda}_k - \hat{\lambda}_{k-1})}{n} \\ &\quad + \frac{\log n}{n} + (\hat{\lambda}_k - \hat{\lambda}_{k-1})\left(\log \hat{\sigma}^2_{k,p+1} - \log \hat{\sigma}^2_{k,p}\right). \end{aligned} \tag{2.22}$$

We will show that $\log \hat{\sigma}^2_{k,p+1} - \log \hat{\sigma}^2_{k,p} = O(\log\log n/n)$, and the result follows.

Again define $a = \hat{\tau}_{k-1} = [\hat{\lambda}_{k-1}n]$ and $b = \hat{\tau}_k - 1 = [\hat{\lambda}_k n] - 1$, $1 \le k \le m+1$. Recall that

$$\hat{\sigma}^2_{k,p+1} = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \left\| \mathbf{X}_{a:b} - \mathrm{P}_{\mathbf{M}^{p+1}_{a:b}} (\mathbf{X}_{a:b}) \right\|^2,$$

where $\mathrm{P}_{\mathbf{M}^{p+1}_{a:b}} (\mathbf{X}_{a:b})$ is the projection of $\mathbf{X}_{a:b}$ onto the $(p+2)$-dimensional column space of

$$\begin{aligned} \mathbf{M}^{p+1}_{a:b} &= \begin{pmatrix} 1 & X_{a-1} & X_{a-2} & \cdots & X_{a-p} & X_{a-(p+1)} \\ 1 & X_a & X_{a-1} & \cdots & X_{a-p+1} & X_p \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & X_{b-1} & X_{b-2} & \cdots & X_{b-p} & X_{b-(p+1)} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{1} & \mathbf{X}_{a-1:b-1} & \mathbf{X}_{a-2:b-2} & \cdots & \mathbf{X}_{a-p:b-p} & \mathbf{X}_{a-(p+1):b-(p+1)} \end{pmatrix}. \end{aligned}$$

This projection is equivalent to the projection of $\mathbf{X}_{a:b}$ onto the $(p+2)$-dimensional column space of

$$\begin{aligned} \tilde{\mathbf{M}}^{p+1}_{a:b} &:= \begin{pmatrix} 1 & X_{a-1} - \overline{X}_{a-1:b-1} & \cdots & X_{a-(p+1)} - \overline{X}_{a-(p+1):b-(p+1)} \\ 1 & X_a - \overline{X}_{a-1:b-1} & \cdots & X_p - \overline{X}_{a-(p+1):b-(p+1)} \\ \vdots & \vdots & & \vdots \\ 1 & X_{b-1} - \overline{X}_{a-1:b-1} & \cdots & X_{b-(p+1)} - \overline{X}_{a-(p+1):b-(p+1)} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{1} & \mathbf{X}_{a-1:b-1} - \overline{X}_{a-1:b-1}\mathbf{1} & \cdots & \mathbf{X}_{a-(p+1):b-(p+1)} - \overline{X}_{a-(p+1):b-(p+1)}\mathbf{1} \end{pmatrix}, \end{aligned}$$

denoted by $\mathrm{P}_{\tilde{\mathbf{M}}^{p+1}_{a:b}} (\mathbf{X}_{a:b})$. Since the second through $(p+2)$nd columns are orthogonal to the first column of $\tilde{\mathbf{M}}^{p+1}_{a:b}$, it is not difficult to show that

$$\mathrm{P}_{\tilde{\mathbf{M}}^{p+1}_{a:b}} (\mathbf{X}_{a:b}) = \overline{X}_{a:b}\mathbf{1} + \mathrm{P}_{\tilde{\mathbf{M}}_{\mathbf{0}}{}^{p+1}_{a:b}} \left(\mathbf{X}_{a:b} - \overline{X}_{a:b}\mathbf{1} \right)$$

where $\tilde{\mathbf{M}}_{\mathbf{0}}{}^{p+1}_{a:b}$ is the matrix $\tilde{\mathbf{M}}^{p+1}_{a:b}$ with the first column removed,

$$\begin{aligned} \tilde{\mathbf{M}}_{\mathbf{0}}{}^{p+1}_{a:b} &:= \begin{pmatrix} X_{a-1} - \overline{X}_{a-1:b-1} & \cdots & X_{a-(p+1)} - \overline{X}_{a-(p+1):b-(p+1)} \\ X_a - \overline{X}_{a-1:b-1} & \cdots & X_p - \overline{X}_{a-(p+1):b-(p+1)} \\ \vdots & & \vdots \\ X_{b-1} - \overline{X}_{a-1:b-1} & \cdots & X_{b-(p+1)} - \overline{X}_{a-(p+1):b-(p+1)} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{X}_{a-1:b-1} - \overline{X}_{a-1:b-1}\mathbf{1} & \cdots & \mathbf{X}_{a-(p+1):b-(p+1)} - \overline{X}_{a-(p+1):b-(p+1)}\mathbf{1} \end{pmatrix}. \end{aligned}$$

Therefore, the estimated white noise variance when fitting an AR$(p+1)$ model to the kth segment becomes

$$\hat{\sigma}^2_{k,p+1} = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \left\| \mathbf{X}_{a:b} - \overline{X}_{a:b}\mathbf{1} - \mathrm{P}_{\tilde{\mathbf{M}}_{\mathbf{0}}{}^{p+1}_{a:b}} \left(\mathbf{X}_{a:b} - \overline{X}_{a:b}\mathbf{1} \right) \right\|^2.$$

Likewise, the estimated white noise variance when fitting an AR(p) model to the kth segment can be expressed as

$$\hat{\sigma}^2_{k,p} = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \left\| \mathbf{X}_{a:b} - \overline{X}_{a:b}\mathbf{1} - \mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^p_{a:b}} \left(\mathbf{X}_{a:b} - \overline{X}_{a:b}\mathbf{1} \right) \right\|^2$$

where $\tilde{\mathbf{M}}\mathbf{0}^p_{a:b}$ is defined accordingly.

Let $\mathbf{Y}_{a:b} = \mathbf{X}_{a:b} - \overline{X}_{a:b}\mathbf{1}$. Then

$$\tilde{\mathbf{M}}\mathbf{0}^{p+1}_{a:b} = \left(\mathbf{Y}_{a-1:b-1} \cdots \mathbf{Y}_{a-(p+1):b-(p+1)} \right),$$

and we can express the projection of $\mathbf{Y}_{a:b}$ onto the column space of $\tilde{\mathbf{M}}\mathbf{0}^{p+1}_{a:b}$ as

$$\begin{aligned}
\mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^{p+1}_{a:b}} \mathbf{Y}_{a:b} &= \tilde{\mathbf{M}}\mathbf{0}^{p+1}_{a:b} \hat{\boldsymbol{\phi}}^{p+1}_{a:b} \\
&= \hat{\phi}^{p+1,1}_{a:b} \mathbf{Y}_{a-1:b-1} + \cdots + \hat{\phi}^{p+1,p+1}_{a:b} \mathbf{Y}_{a-(p+1):b-(p+1)} \\
&= \hat{\phi}^{p+1,1}_{a:b} \mathbf{Y}_{a-1:b-1} + \cdots + \hat{\phi}^{p+1,p}_{a:b} \mathbf{Y}_{a-p:b-p} \\
&\quad + \hat{\phi}^{p+1,p+1}_{a:b} \left(\mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^p_{a:b}} (\mathbf{Y}_{a-(p+1):b-(p+1)}) + \mathrm{P}^{\perp}_{\tilde{\mathbf{M}}\mathbf{0}^p_{a:b}} (\mathbf{Y}_{a-(p+1):b-(p+1)}) \right),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\phi}}^{p+1}_{a:b} &= \left(\tilde{\mathbf{M}}\mathbf{0}^{p+1\,T}_{a:b} \tilde{\mathbf{M}}\mathbf{0}^{p+1}_{a:b} \right)^{-1} \tilde{\mathbf{M}}\mathbf{0}^{p+1\,T}_{a:b} \mathbf{Y}_{a:b} \\
&= \left(\hat{\phi}^{p+1,1}_{a:b}, \ldots, \hat{\phi}^{p+1,p+1}_{a:b} \right)^T,
\end{aligned}$$

and

$$\mathrm{P}^{\perp}_{\tilde{\mathbf{M}}\mathbf{0}^p_{a:b}} (\mathbf{Y}_{a-(p+1):b-(p+1)}) = \mathbf{Y}_{a-(p+1):b-(p+1)} - \mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^p_{a:b}} (\mathbf{Y}_{a-(p+1):b-(p+1)}).$$

The projection of $\mathbf{Y}_{a-(p+1):b-(p+1)}$ onto the column space of $\mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^p_{a:b}}$ will be denoted as

$$\mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^p_{a:b}} (\mathbf{Y}_{a-(p+1):b-(p+1)}) = \tilde{\mathbf{M}}\mathbf{0}^p_{a:b} \tilde{\boldsymbol{\phi}}^p_{a:b} = \sum_{j=1}^{p} \tilde{\phi}^{p,j}_{a:b} \mathbf{Y}_{a-j,b-j},$$

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

$$\tilde{\boldsymbol{\phi}}^p_{a:b} = \left(\tilde{\mathbf{M}}\mathbf{0}^{p\,T}_{a:b} \tilde{\mathbf{M}}\mathbf{0}^p_{a:b} \right)^{-1} \tilde{\mathbf{M}}\mathbf{0}^{p\,T}_{a:b} \mathbf{Y}_{a-(p+1):b-(p+1)}$$

from the estimated coefficients

$$\hat{\boldsymbol{\phi}}^{p}_{a:b} = \left(\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}{}^{T}\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}\right)^{-1}\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}{}^{T}\mathbf{Y}_{a:b}.$$

Since $\mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}}\mathbf{Y}_{a-(p+1):b-(p+1)}$ is in the span of $\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}$,

$$\begin{aligned}\hat{\phi}^{p+1,1}_{a:b}\mathbf{Y}_{a-1:b-1} + \cdots + \hat{\phi}^{p+1,p}_{a:b}\mathbf{Y}_{a-p:b-p} + \hat{\phi}^{p+1,p+1}_{a:b}\mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}}\mathbf{Y}_{a-(p+1):b-(p+1)} \\ = \quad \hat{\phi}^{p,1}_{a:b}\mathbf{Y}_{a-1:b-1} + \cdots + \hat{\phi}^{p,p}_{a:b}\mathbf{Y}_{a-p:b-p},\end{aligned}$$

and thus the difference between the conditional maximum likelihood variance estimates in the kth segment is

$$\begin{aligned}\hat{\sigma}^2_{k,p} - \hat{\sigma}^2_{k,p+1} &= \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\left\|\mathbf{Y}_{a:b} - \tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}\hat{\boldsymbol{\phi}}^{p}_{a:b}\right\|^2 \\ &\quad - \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\left\|\mathbf{Y}_{a:b} - \tilde{\mathbf{M}}\mathbf{0}^{p+1}_{a:b}\hat{\boldsymbol{\phi}}^{p+1}_{a:b}\right\|^2 \\ &= \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\left(\hat{\phi}^{p+1,p+1}_{a:b}\right)^2\left\|\mathrm{P}^{\perp}_{\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}}(\mathbf{Y}_{a-(p+1):b-(p+1)})\right\|^2.\end{aligned}$$

Note that

$$\begin{aligned}&\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\left\|\mathrm{P}^{\perp}_{\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}}\mathbf{Y}_{a-(p+1):b-(p+1)}\right\|^2 \\ &= \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\left\|\mathbf{Y}_{a-(p+1):b-(p+1)}\right\|^2 - \mathbf{u}^T\mathbf{V}^{-1}\mathbf{u},\end{aligned}$$

where

$$\begin{aligned}\mathbf{u} &= \begin{pmatrix} \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\sum_{t=a}^{b} Y_{t-p-1}Y_{t-1} \\ \vdots \\ \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\sum_{t=a}^{b} Y_{t-p-1}Y_{t-p}\end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\sum_{t=a}^{b}(X_{t-p-1} - \overline{X}_{a-p-1:b-p-1})(X_{t-1} - \overline{X}_{a-1:b-1}) \\ \vdots \\ \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\sum_{t=a}^{b}(X_{t-p-1} - \overline{X}_{a-p-1:b-p-1})(X_{t-p} - \overline{X}_{a-p:b-p})\end{pmatrix},\end{aligned}$$

and $\mathbf{V} = \left\{v_{ij}\right\}_{i,j=1}^{p}$ where

$$\begin{aligned}v_{ij} &= \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\sum_{t=a}^{b} Y_{t-i}Y_{t-j} \\ &= \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\sum_{t=a}^{b}(X_{t-i} - \overline{X}_{a-i:b-i})(X_{t-j} - \overline{X}_{a-j:b-j}).\end{aligned}$$

From this observation, it is readily seen that

$$\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n}\left\|\mathrm{P}^{\perp}_{\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}}(\mathbf{Y}_{a-(p+1):b-(p+1)})\right\|^2 \to c_p > 0,$$

with probability 1, as n goes to infinity, where

$$c_p = \min_{a_0,a_1,\ldots,a_p} E[X_{t-p-1} - (a_0 + a_1 X_{t-1} + \cdots + a_p X_{t-p})]^2.$$

Therefore, if we can show that $\left(\hat{\phi}^{p+1,p+1}_{a:b}\right)^2 = O(\log\log n/n)$, it follows that $\log\hat{\sigma}^2_{k,p} - \log\hat{\sigma}^2_{k,p+1} = O(\log\log n/n)$.

Recall that $\hat{\boldsymbol{\phi}}^{p+1}_{a:b} = \left(\tilde{\mathbf{M}}\mathbf{0}^{p+1\,T}_{a:b}\tilde{\mathbf{M}}\mathbf{0}^{p+1}_{a:b}\right)^{-1}\tilde{\mathbf{M}}\mathbf{0}^{p+1\,T}_{a:b}\mathbf{Y}_{a:b}$. We would like to apply the functional law of the iterated logarithm on the $(p+1)$st component of $\hat{\boldsymbol{\phi}}^{p+1}_{a:b}$, denoted by $\hat{\phi}^{p+1,p+1}_{a:b}$. As discussed in Section 2.2.1, we can apply the functional law of the iterated logarithm to the sample covariances of the process $\{X_t\}$. Assume the true order of the kth segment is p. Then adapting the argument given on p. 171 of Brockwell and Davis (1991), where we assume inner products are dot products in Euclidean space, we have,

$$\begin{aligned}\hat{\phi}^{p+1,p+1}_{a:b} &= \frac{\left\langle \mathbf{Y}_{a:b},\, \mathbf{Y}_{a-(p+1):b-(p+1)} - \mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}}\mathbf{Y}_{a-(p+1):b-(p+1)}\right\rangle}{\left\|\mathbf{Y}_{a-(p+1):b-(p+1)} - \mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}}\mathbf{Y}_{a-(p+1):b-(p+1)}\right\|^2} \\ &= \frac{\left\langle \mathbf{Y}_{a:b},\, \mathbf{Y}_{a-(p+1):b-(p+1)}\right\rangle - \sum_{j=1}^{p}\tilde{\phi}^{p,j}_{a:b}\left\langle \mathbf{Y}_{a:b},\, \mathbf{Y}_{a-j,b-j}\right\rangle}{\left\|\mathbf{Y}_{a-(p+1):b-(p+1)} - \sum_{j=1}^{p}\tilde{\phi}^{p,j}_{a:b}\mathbf{Y}_{a-j,b-j}\right\|^2}, \end{aligned} \tag{2.23}$$

where $\mathrm{P}_{\tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}}\mathbf{Y}_{a-(p+1):b-(p+1)} = \tilde{\mathbf{M}}\mathbf{0}^{p}_{a:b}\tilde{\boldsymbol{\phi}}^{p}_{a:b} = \sum_{j=1}^{p}\tilde{\phi}^{p,j}_{a:b}\mathbf{Y}_{a-j,b-j}$.

Define $s_{ij}(r) = \sum_{t=1}^{r} Y_{t-i}Y_{t-j}$ for $r = 1, 2, \ldots$, and $\gamma(|i-j|) = \mathrm{Cov}(X_t, X_{t+|i-j|})$. To apply the functional law of the iterated logarithm on $\hat{\phi}^{p+1,p+1}_{a:b}$, we need to show that it is a function of the $s_{ij}(.)$'s. Consider the numerator of (2.23)

$$\begin{aligned}&\left\langle \mathbf{Y}_{a:b},\, \mathbf{Y}_{a-(p+1):b-(p+1)}\right\rangle - \sum_{j=1}^{p}\tilde{\phi}^{p,j}_{a:b}\left\langle \mathbf{Y}_{a:b},\, \mathbf{Y}_{a-j,b-j}\right\rangle \\ &= \sum_{t=a}^{b} Y_t Y_{t-(p+1)} - \sum_{j=1}^{p}\left(\tilde{\phi}^{p,j}_{a:b}\sum_{t=a}^{b} Y_t Y_{t-j}\right),\end{aligned}$$

where $\sum_{t=a}^{b} Y_t Y_{t-(p+1)} = s_{0,p+1}(b) - s_{0,p+1}(a-1)$ and $\sum_{t=a}^{b} Y_t Y_{t-j} = s_{0,j}(b) - s_{0,j}(a-1)$. By the projection theorem, the denominator of (2.23) becomes

$$\begin{aligned}
&\left\| \mathbf{Y}_{a-(p+1):b-(p+1)} - \sum_{j=1}^{p} \tilde{\phi}_{a:b}^{p,j} \mathbf{Y}_{a-j,b-j} \right\|^2 \\
&= \left\| \mathbf{Y}_{a-(p+1):b-(p+1)} \right\|^2 - \left\| \sum_{j=1}^{p} \tilde{\phi}_{a:b}^{p,j} \mathbf{Y}_{a-j,b-j} \right\|^2 \\
&= \left\| \mathbf{Y}_{a-(p+1):b-(p+1)} \right\|^2 - \left\| \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p} \tilde{\boldsymbol{\phi}}_{a:b}^{p} \right\|^2 \\
&= \sum_{t=a}^{b} Y_{t-(p+1)}^2 - \tilde{\boldsymbol{\phi}}_{a:b}^{p\ T} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p\ T} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p} \tilde{\boldsymbol{\phi}}_{a:b}^{p}.
\end{aligned}$$

As in the numerator, $\sum_{t=a}^{b} Y_{t-(p+1)}^2 = s_{p+1,p+1}(b) - s_{p+1,p+1}(a-1)$, and

$$\begin{aligned}
\tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p\ T} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p} &= \left[\sum_{t=a}^{b} Y_{t-i} Y_{t-j} \right]_{i,j=1}^{p} \\
&= \left[s_{ij}(b) - s_{ij}(a-1) \right]_{i,j=1}^{p}.
\end{aligned}$$

In both the numerator and the denominator of (2.23), $\tilde{\boldsymbol{\phi}}_{a:b}^{p}$ is a function of the $s_{ij}(.)$'s since

$$\begin{aligned}
\tilde{\boldsymbol{\phi}}_{a:b}^{p} &= \left(\tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p\ T} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p} \right)^{-1} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p\ T} \mathbf{Y}_{a-(p+1):b-(p+1)} \\
&= \left[\left\{ s_{ij}(b) - s_{ij}(a-1) \right\}_{i,j=1}^{p} \right]^{-1} \begin{pmatrix} s_{1,p+1}(b) - s_{1,p+1}(a-1) \\ \vdots \\ s_{p,p+1}(b) - s_{p,p+1}(a-1) \end{pmatrix}.
\end{aligned}$$

Therefore, (2.23) becomes

$$\begin{aligned}
\hat{\phi}_{a:b}^{p+1,p+1} &= \left[\frac{1}{b-a+1} \sum_{t=a}^{b} Y_t Y_{t-(p+1)} - \frac{1}{b-a+1} \mathbf{Z} \times \right. \\
&\quad \left. \left(\frac{1}{b-a+1} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p\ T} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p} \right)^{-1} \left(\frac{1}{b-a+1} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p\ T} \mathbf{Y}_{a:b} \right) \right] \div \\
&\quad \left[\frac{1}{b-a+1} \sum_{t=a}^{b} Y_{t-(p+1)}^2 - \left(\frac{1}{b-a+1} \mathbf{Y}_{a:b}^{T} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p} \right) \times \right. \\
&\quad \left. \left(\frac{1}{b-a+1} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p\ T} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p} \right)^{-1} \left(\frac{1}{b-a+1} \tilde{\mathbf{M}}_{\mathbf{0}a:b}^{\ p\ T} \mathbf{Y}_{a:b} \right) \right] \\
&= h\left(\frac{1}{b-a+1} \left(s_{ij}(b) - s_{ij}(a-1) \right); i,j = 0,1,\ldots,p+1 \right),
\end{aligned}$$

where $\mathbf{Z} = \left(\sum_{t=a}^{b} Y_t Y_{t-1}, \ldots, \sum_{t=a}^{b} Y_t Y_{t-p}\right)$. Let $\gamma(|i-j|) = \mathrm{Cov}(X_t, X_{t+|i-j|})$. Using a first order Taylor expansion about $\boldsymbol{\gamma} = (\gamma(|i-j|); i, j = 0, 1, \ldots, p+1)$ on $\hat{\phi}_{a:b}^{p+1,p+1}$,

$$\begin{aligned}\hat{\phi}_{a:b}^{p+1,p+1} &= h\left(\gamma(|i-j|); i, j = 0, 1, \ldots, p+1\right) \\ &\quad + \nabla h\left(\gamma(|i-j|); i, j = 0, 1, \ldots, p+1\right)\left(\frac{1}{b-a+1}\boldsymbol{s} - \boldsymbol{\gamma}\right),\end{aligned}$$

where $\boldsymbol{s} = (s_{ij}(b) - s_{ij}(a-1); i, j = 0, 1, \ldots, p+1)$. Note that $\frac{s_{ij}(r)}{r}$ goes to $\gamma(|i-j|)$ as r goes to infinity. Since $a = [\hat{\lambda}_{k-1} n]$ and $b = [\hat{\lambda}_k n] - 1$, for any $i, j = 1, \ldots, p$,

$$\begin{aligned}&\frac{1}{b-a+1}\left(s_{ij}(b) - s_{ij}(a-1)\right) - \gamma(|i-j|) \\ &= \frac{1}{\hat{\lambda}_k - \hat{\lambda}_{k-1}}\left(\frac{s_{ij}(\hat{\lambda}_k n - 1) - s_{ij}(\hat{\lambda}_{k-1} n - 1)}{n} - (\hat{\lambda}_k - \hat{\lambda}_{k-1})\gamma(|i-j|)\right) \\ &= \frac{1}{\hat{\lambda}_k - \hat{\lambda}_{k-1}}\left(\frac{s_{ij}(\hat{\lambda}_k n - 1) - (\hat{\lambda}_k n - 1)\gamma(|i-j|)}{n}\right. \\ &\quad \left. + \frac{s_{ij}(\hat{\lambda}_{k-1} n - 1) - (\hat{\lambda}_{k-1} n - 1)\gamma(|i-j|)}{n} - \frac{2\gamma(|i-j|)}{n}\right) \\ &= \frac{1}{\hat{\lambda}_k - \hat{\lambda}_{k-1}}\left(O\left(\sqrt{\frac{1}{n}\log\log n}\right) + O\left(\sqrt{\frac{1}{n}\log\log n}\right) - O\left(\frac{1}{n}\right)\right), \quad (2.24)\end{aligned}$$

where (2.24) follows by the functional law of the iterated logarithm. Since $\hat{\lambda}_k - \hat{\lambda}_{k-1} \geq \epsilon > 0$ and $h\left(\gamma(|i-j|); i, j = 0, 1, \ldots, p+1\right) = \phi^{p+1,p+1} = 0$, then $\left(\hat{\phi}_{a:b}^{p+1,p+1}\right)^2 = O\left(\log\log n/n\right)$. By induction, $\log\hat{\sigma}_{k,p}^2 - \log\hat{\sigma}_{k,p+s}^2 = O\left(\log\log n/n\right)$ for any positive integer s. Since the segment k was arbitrary, the result holds for any segment or segments. □

2.6 Consistency of the Number of Change-points and AR Orders Estimates

Lemmas 2.1 and 2.2 imply that if the true model is an autoregressive process with no change-points, then for any $m \geq 1$, the minimum description length for a fitted model with m change-points will be larger than the minimum description

length for a fitted model with no change-points for large n with probability 1. In other words, when the true model has no change-points, $\hat{m}$ is a strongly consistent estimator of 0, and $\hat{p}_1, \ldots, \hat{p}_{\hat{m}+1}$ are all strongly consistent estimators of the true AR order, p. Now we would like to show that if the true number of change-points is $m_0 \geq 1$ and the true AR orders are $p_1^0, \ldots, p_{m_0+1}^0$, then $\hat{m}$ is a consistent estimator for m_0 and $(\hat{p}_1, \ldots, \hat{p}_{\hat{m}+1})$ is consistent for $(p_1^0, \ldots, p_{m_0+1}^0)$. Since we can only show that the difference between minimum description lengths for a model with the true number of change-points and a model with more than the true number of change-points includes terms $O_p(1/n)$ rather than $O(1/n)$, we only obtain weak consistency of $\hat{m}$ and $(\hat{p}_1, \ldots, \hat{p}_{\hat{m}+1})$.

Assume throughout this section that the true model is the piecewise autoregressive process defined by (2.9). 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 dataset with m_0 relative change-points, $\boldsymbol{\lambda} \in A_{m_0}^{\epsilon}$, where the AR orders, $p_1^0, \ldots, p_{m_0+1}^0$, are known.
>
> *Model* 2′: Fit a piecewise autoregressive model to the dataset with $m_0 + 1$ relative change-points, $\boldsymbol{\alpha} \in A_{m_0+1}^{\epsilon}$. Estimate the autoregressive orders from the data, and denote these orders by $\hat{p}_1, \ldots, \hat{p}_{m_0+2}$.

If we can show that the MDL for Model 2′ is larger than the MDL for Model 1′ for large n in probability, then it follows that the MDL for a model with $m_0 + s$ change-points for any $1 \leq s \leq M - m_0$, where M is a prespecified upper bound for the fitted number of change-points, is larger than the MDL for Model 1′ (the true model) for large n in probability. Since for large n, the estimated number of change-points cannot underestimate the true number of change-points with probability 1, this implies that

the MDL for a model with m changepoints, 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.

The minimum description length for Model 1′ is

$$\inf_{\boldsymbol{\lambda} \in A^{\epsilon}_{m_0}} \text{MDL}(m_0, \boldsymbol{\lambda}; p^0_1, \ldots, p^0_{m_0+1}),$$

where

$$\begin{aligned}\text{MDL}(m_0, \boldsymbol{\lambda}; p^0_1, \ldots, p^0_{m_0+1}) &= \log m_0 + (m_0+1)\log n \\ &\quad + \sum_{k=1}^{m_0+1} \log p^0_k + \sum_{k=1}^{m_0+1} \frac{p^0_k + 2}{2} \log((\lambda_k - \lambda_{k-1})n) \\ &\quad + \sum_{k=1}^{m_0+1} \frac{(\lambda_k - \lambda_{k-1})n}{2} \log(2\pi\hat{\sigma}^2_{k,m_0}),\end{aligned}$$

$\hat{\sigma}^2_{k,m_0}$ is the conditional maximum likelihood estimate of the process variance when fitting an autoregressive model of order p^0_k to the kth of the $m_0 + 1$ segments, and $A^{\epsilon}_{m_0}$ is defined in (2.11). The minimum description length for Model 2′ is

$$\inf_{\boldsymbol{\alpha} \in A^{\epsilon}_{m_0+1}} \text{MDL}(m_0 + 1, \boldsymbol{\alpha}; \hat{p}_1, \ldots, \hat{p}_{m_0+2}),$$

where

$$\begin{aligned}\text{MDL}(m_0 + 1, \boldsymbol{\alpha}; \hat{p}_1, \ldots, \hat{p}_{m_0+2}) &= \log(m_0+1) + (m_0+2)\log n \\ &\quad + \sum_{j=1}^{m_0+2} \log \hat{p}_j + \sum_{j=1}^{m_0+2} \frac{\hat{p}_j + 2}{2} \log((\alpha_j - \alpha_{j-1})n) \\ &\quad + \sum_{j=1}^{m_0+2} \frac{(\alpha_j - \alpha_{j-1})n}{2} \log(2\pi\hat{\sigma}^2_{j,m_0+1}),\end{aligned}$$

$\hat{\sigma}^2_{j,m_0+1}$ is the conditional maximum likelihood estimate of the process variance when fitting an autoregressive model of order $\hat{p}_j$ to the jth of the $m_0 + 2$ segments, and $A^{\epsilon}_{m_0+1}$ is defined in (2.11). Let

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in A^{\epsilon}_{m_0}} \left\{ \frac{2}{n} \text{MDL}(m_0, \boldsymbol{\lambda}; p^0_1, \ldots, p^0_{m_0+1}) \right\},$$

and

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in A^{\epsilon}_{m_0+1}}{\arg\min} \left\{ \frac{2}{n} \mathrm{MDL}(m_0 + 1, \boldsymbol{\alpha}; \hat{p}_1, \ldots, \hat{p}_{m_0+2}) \right\}.$$

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 a piecewise AR model and meets the assumptions for the functional law of the iterated logarithm to hold for sample covariances within each segment.

Theorem 2.1. *Assume the true process $\{X_t\}$ follows the AR(p) model given in (2.15) with m_0 change-points and initial values X_{-P}, $X_{-P+1}, \ldots, X_0$, and satisfies assumptions A1 and A2. Then $\hat{m} \xrightarrow{P} m_0$, where $\hat{m}$ is the estimated number of change-points obtained by minimizing the MDL defined using conditional maximum likelihood white noise estimates.*

Note that the statement $\hat{m} \xrightarrow{P} m_0$ is equivalent to the statement

$$\lim_{n\to\infty} P\left(\inf_{\substack{m \neq m_0 \\ m < M}} \{\mathrm{MDL}(m, \hat{\boldsymbol{\alpha}}; \hat{p}_1, \ldots, \hat{p}_{m+1})\} > \mathrm{MDL}(m_0, \hat{\boldsymbol{\lambda}}; p^0_1, \ldots, p^0_{m_0+1}) \right) = 1$$

for a fixed upper bound M. We will show that this result follows if

$$\lim_{n\to\infty} P\left(\mathrm{MDL}(m_0 + 1, \hat{\boldsymbol{\alpha}}; \hat{p}_1, \ldots, \hat{p}_{m_0+2}) > \mathrm{MDL}(m_0, \boldsymbol{\lambda}^0; p^0_1, \ldots, p^0_{m_0+1})\right) = 1.$$

Before we prove Theorem 2.1 in the general case, 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. Let λ be the true relative change-point location[3], i.e., $\tau = [\lambda n]$ is the observation at which the change occurs. Denote the autoregressive coefficient in the first segment by ϕ_1 and the autoregressive coefficient in the second segment by ϕ_2.

---

[3]For this simple case, we drop the zero superscript on the true parameters for notational convenience.

Likewise, denote the white noise variances in the first and second segments by σ_1^2 and σ_2^2, respectively. Thus, $\phi_1 \neq \phi_2$ and/or $\sigma_1^2 \neq \sigma_2^2$. Compare two fitted models:

1. Fit AR(1) models to two segments with relative change-point location λ.

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 $(\alpha_1, \alpha_2) \in A_2^\epsilon$, where A_2^ϵ is defined in (2.11).

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 the first fitted model, and $\mathrm{MDL}(2, \hat{\alpha}_1, \hat{\alpha}_2; 1, 1)$ is the minimized MDL for the second fitted model. Equivalently, we show that

$$\lim_{n\to\infty} P\big(\mathrm{MDL}(2, \alpha_1, \alpha_2; 1, 1) > \mathrm{MDL}(1, \lambda; 1) \quad \forall\, \alpha_1, \alpha_2 :\ \epsilon < \alpha_1 < \alpha_1 + \epsilon < \alpha_2 < 1 - \epsilon\big) = 1. \tag{2.25}$$

In proving (2.25), we can assume without loss of generality that $\alpha_1 < \lambda < \alpha_2$ since if $\lambda < \alpha_1$ or $\alpha_2 < \lambda$, the same argument can be applied to the fitted segments $(0, \alpha_1)$ or $(\alpha_2, 1)$, repectively, as the argument we will use for the fitted segment (α_1, α_2) when $\alpha_1 < \lambda < \alpha_2$. Therefore, (2.25) follows if each of the following statements hold.

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

(ii) For each finite positive integer N,

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

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

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

for sufficiently large n.

In addition to these three statements, three corresponding statements where α_2 (rather than α_1) is allowed to be close to λ must hold. These corresponding statements can be proven in a similar manner as the previous three statements, and thus, it suffices to show only that the above three statements hold.

Consider the difference

$$\begin{aligned}
&\frac{2}{n}[\mathrm{MDL}(2,\alpha_1,\alpha_2;1,1) - \mathrm{MDL}(1,\lambda;1)] \\
&= \left(\frac{1}{n}\right)\Big[2\log 2 + 3(\log\alpha_1 + \log(\alpha_2-\alpha_1) + \log(1-\alpha_2) \\
&\quad - \log\lambda - \log(1-\lambda))\Big] + \frac{5\log n}{n} \\
&\quad + \alpha_1\log\hat{\sigma}^2_{1,2} + (\alpha_2-\alpha_1)\log\hat{\sigma}^2_{2,2} + (1-\alpha_2)\log\hat{\sigma}^2_{3,2} \\
&\quad - \left[\lambda\log\hat{\sigma}^2_{1,1} + (1-\lambda)\log\hat{\sigma}^2_{2,1}\right] \\
&= O\left(\frac{1}{n}\right) + O\left(\frac{\log n}{n}\right) + \alpha_1\log\hat{\sigma}^2_{1,2} + (\alpha_2-\alpha_1)\log\hat{\sigma}^2_{2,2} \\
&\quad + (1-\alpha_2)\log\hat{\sigma}^2_{3,2} - \left[\lambda\log\hat{\sigma}^2_{1,1} + (1-\lambda)\log\hat{\sigma}^2_{2,1}\right].
\end{aligned}$$

We will show that the probability this difference is positive for any α_1 and α_2 with $\epsilon < \alpha_1 < \lambda < \lambda+\epsilon/2 < \alpha_2 < 1-\epsilon$ converges to one as n tends to infinity. The penalty terms, $O(1/n)$ and $O(\log n/n)$, are positive, so we need only show that the remaining terms are of a smaller order than the first two terms in probability. 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

$$\alpha_1\log\hat{\sigma}^2_{1,2} + (\alpha_2-\alpha_1)\log\hat{\sigma}^2_{2,2} + (1-\alpha_2)\log\hat{\sigma}^2_{3,2} \\ - \left[\lambda\log\hat{\sigma}^2_{1,1} + (1-\lambda)\log\hat{\sigma}^2_{2,1}\right] \qquad (2.26)$$

into a sum over our partition of intervals. Then we can look at each true segment individually, and show that the difference between the terms in the two fitted models within each true segment 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.

In order to look at (2.26) as a sum over the true segments, we must determine how to break up the terms from the fitted segments. The segments $(0, \alpha_1)$ and $(\alpha_2, 1)$ do not contain any true change-points, so we can leave the terms $\alpha_1 \log \hat{\sigma}^2_{1,2}$ and $(1-\alpha_2)\log\hat{\sigma}^2_{3,2}$ as is. However, the interval (α_1, α_2) contains one true change-point, λ, so we will break up the term $(\alpha_2 - \alpha_1)\log\hat{\sigma}^2_{2,2}$ as follows. Let $\hat{\phi}$ be the autoregressive coefficient estimate when fitting an AR(1) model to the second fitted segment, (α_1, α_2). Then

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

$$= (\alpha_2-\alpha_1)\log\left[\frac{1}{(\alpha_2-\alpha_1)n}\left(\sum_{t=[\alpha_1 n]}^{[\lambda n]-1}\left(X_t - \hat{\phi}X_{t-1}\right)^2 + \sum_{t=[\lambda n]}^{[\alpha_2 n]-1}\left(X_t - \hat{\phi}X_{t-1}\right)^2\right)\right].$$

Define

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

and

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

Then we can write

$$(\alpha_2-\alpha_1)\log\hat{\sigma}^2_{2,2} = (\alpha_2-\alpha_1)\log\left[\frac{\mathrm{RSS}_{2,1}}{(\alpha_2-\alpha_1)n} + \frac{\mathrm{RSS}_{2,2}}{(\alpha_2-\alpha_1)n}\right].$$

Keep in mind that since $\hat{\phi}$ is calculated using observations $[\alpha_1 n], \ldots, [\alpha_2 n]$, even though $\mathrm{RSS}_{2,1}$ is a sum over observations $[\alpha_1 n], \ldots, [\lambda n]$, it also depends on the observations $[\lambda n], \ldots, [\alpha_2 n]$. Likewise, $\mathrm{RSS}_{2,2}$ depends on all of the observations between $[\alpha_1 n]$ and $[\alpha_2 n]$ even though it is a sum only over observations $[\lambda n], \ldots, [\alpha_2 n]$.

We will first show that statement (i) holds. If $\lambda - \alpha_1 > (\log\log n)^2/\log n$, or equivalently, if $\lambda - \alpha_1 >> \log\log n/\log n$, and if $\lambda + \epsilon/2 < \alpha_2$, then for large n,

$$\begin{aligned}
(\alpha_2 - \alpha_1)\log\hat{\sigma}^2_{2,2} &= (\alpha_2 - \alpha_1)\log\left[\frac{\mathrm{RSS}_{2,1}}{(\alpha_2 - \alpha_1)n} + \frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \alpha_1)n}\right] \\
&= (\alpha_2 - \alpha_1)\log\left[\frac{\lambda - \alpha_1}{\alpha_2 - \alpha_1}\cdot\frac{\mathrm{RSS}_{2,1}}{(\lambda - \alpha_1)n} + \frac{\alpha_2 - \lambda}{\alpha_2 - \alpha_1}\cdot\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n}\right] \\
&> (\lambda - \alpha_1)\log\left(\frac{\mathrm{RSS}_{2,1}}{(\lambda - \alpha_1)n}\right) + (\alpha_2 - \lambda)\log\left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n}\right) \\
&\geq (\lambda - \alpha_1)\log\left(\frac{\mathrm{RSS}^*_{2,1}}{(\lambda - \alpha_1)n}\right) + (\alpha_2 - \lambda)\log\left(\frac{\mathrm{RSS}^*_{2,2}}{(\alpha_2 - \lambda)n}\right),
\end{aligned}$$

where

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

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

$\hat{\phi}_1$ is the AR(1) coefficient estimate fit to the segment (α_1, λ), and $\hat{\phi}_2$ is the AR(1) coefficient estimate fit to the segment (λ, α_2). The first inequality follows by concavity of the log function. The second inequality holds since, by definition, the conditional maximum likelihood AR coefficient estimates used in $\mathrm{RSS}^*_{2,1}$ are fit to the segment (α_1, λ), by minimizing the quantity

$$\sum_{t=[\alpha_1 n]}^{[\lambda n]-1}(X_t - aX_{t-1})^2$$

with respect to a. Likewise,

$$\mathrm{RSS}^*_{2,2} = \arg\min_a\left\{\sum_{t=[\lambda n]}^{[\alpha_2 n]-1}(X_t - aX_{t-1})^2\right\}.$$

The term (2.26) now becomes for large n,

$$\begin{aligned}
&\alpha_1 \log \hat{\sigma}^2_{1,2} + (\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} + (1 - \alpha_2) \log \hat{\sigma}^2_{3,2} \\
&\qquad - \left[\lambda \log \hat{\sigma}^2_{1,1} + (1 - \lambda) \log \hat{\sigma}^2_{2,1}\right] \\
> \quad &\left[\alpha_1 \log \hat{\sigma}^2_{1,2} + (\lambda - \alpha_1) \log \left(\frac{\mathrm{RSS}^*_{2,1}}{(\lambda - \alpha_1)n}\right) - \lambda \log \hat{\sigma}^2_{1,1}\right] \\
&+ \left[(\alpha_2 - \lambda) \log \left(\frac{\mathrm{RSS}^*_{2,2}}{(\alpha_2 - \lambda)n}\right)\right. \\
&\qquad \left. + (1 - \alpha_2) \log \hat{\sigma}^2_{3,2} - (1 - \lambda) \log \hat{\sigma}^2_{2,1}\right]. \qquad (2.27)
\end{aligned}$$

If we expand the first term in brackets using a second order Taylor series expansion on each of the log functions, as in (2.20) in the proof of Lemma 2.1, then we can show that the constant and first order terms cancel, and we are left with the second order terms. The first order terms are exactly zero rather than of order $\log\log n/n$ as in the proof of Lemma 2.1 since in this case, we are assuming the mean of the process is zero, and thus can study asymptotics without estimating the mean. If we were to estimate the mean, i.e., minimize the quantity $(X_t - a_0 - a_1 X_{t-1})^2$ for the residual sum of squares, then the first order terms would again be of order $\log\log n/n$ a.s. In Lemma 2.1, by applying the functional law of the iterated logarithm, we were able to show that the second order terms were $o(\log n/n)$ with probability 1, but since α_1 may now be within ϵ of λ, it is not clear if the functional law of the iterated logarithm still holds. This is because the second order terms involve sums of the type

$$\frac{\sum_{t=[\alpha_1 n]}^{[\lambda n]-1} X_{t-i} X_{t-j}}{(\lambda - \alpha_1)n},$$

and the functional law of the iterated logarithm may not apply to these sums if α_1 is too close to λ. We will show that if $\lambda - \alpha_1 > (\log\log n)^2/\log n$, the functional law of the iterated logarithm between observations $[\alpha_1 n]$ and $[\lambda n]$ still holds.

Let $\hat{\gamma}$ denote the vector of estimated covariances for the segment (α_1, λ),

$$\hat{\gamma} := \left(\frac{\sum_{t=[\alpha_1 n]}^{[\lambda n]-1} X_t^2}{(\lambda - \alpha_1)n}, \frac{\sum_{t=[\alpha_1 n]}^{[\lambda n]-1} X_t X_{t-1}}{(\lambda - \alpha_1)n}, \frac{\sum_{t=[\alpha_1 n]}^{[\lambda n]-1} X_{t-1} X_t}{(\lambda - \alpha_1)n}, \frac{\sum_{t=[\alpha_1 n]}^{[\lambda n]-1} X_{t-1}^2}{(\lambda - \alpha_1)n}\right).$$

Then the "(i,j)th" component of $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ for $i,j = 0,1$, where $\boldsymbol{\gamma}$ is the vector of true covariances for the segment $(0,\lambda)$, is

$$\begin{aligned}
&\frac{\sum_{t=[\alpha_1 n]}^{[\lambda n]-1}\left(X_{t-i}X_{t-j} - \gamma(|i-j|)\right)}{(\lambda-\alpha_1)n} \\
&= \frac{1}{\lambda-\alpha_1}\left(\frac{\sum_{t=1}^{[\lambda n]-1}\left(X_{t-i}X_{t-j} - \gamma(|i-j|)\right)}{n}\right. \\
&\qquad \left. - \frac{\sum_{t=1}^{[\alpha_1 n]-1}\left(X_{t-i}X_{t-j} - \gamma(|i-j|)\right)}{n}\right),
\end{aligned}$$

which, by the law of the iterated logarithm, is bounded by

$$\frac{O\left(\sqrt{\frac{1}{n}\log\log n}\right)}{\lambda-\alpha_1}. \tag{2.28}$$

Since $\lambda - \alpha_1 >> \log\log n/\log n$, (2.28) is $o(1)$, implying that $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}$, and hence the term $\boldsymbol{\gamma}^*$ in the second order term of the Taylor series expansion (which is between $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$) converges to $\boldsymbol{\gamma}$. Thus, the corresponding second order term is bounded by

$$(\lambda-\alpha_1)\frac{\log\log n/n}{(\lambda-\alpha_1)^2} = \frac{\log\log n/n}{\lambda-\alpha_1} = o\left(\frac{\log n}{n}\right).$$

Since this is of smaller order than the penalty terms when we subtract

$$(2/n)\mathrm{MDL}(1,\lambda;1)$$

from

$$(2/n)\mathrm{MDL}(2,\alpha_1,\alpha_2;1,1),$$

and since the second term in brackets of (2.27) is of order $\log\log n/n$ by Lemma 2.1[4], we conclude that

$$\begin{aligned}
\lim_{n\to\infty} P\Big(&\frac{2}{n}[\mathrm{MDL}(2,\alpha_1,\alpha_2;1,1) - \mathrm{MDL}(1,\lambda;1)] > 0 \ \ \forall\, \alpha_1,\alpha_2 : \\
&\lambda-\alpha_1 > (\log\log n)^2/\log n;\ \lambda+\epsilon/2 < \alpha_2 < 1-\epsilon\Big) = 1.
\end{aligned}$$

---

[4]We can apply Lemma 2.1 to the segment $(\lambda,1)$ since $\alpha_2 - \lambda > \epsilon/2$, treating $\epsilon/2$ as the "ϵ" in Lemma 2.1.

Next consider statements (ii) and (iii), where $(\log\log n)^2/\log n > \lambda - \alpha_1$. In statement (ii), since $\lambda - \alpha_1 < N/n$ for some positive integer N, we cannot consider the quantity $\mathrm{RSS}^*_{2,1}/[(\lambda - \alpha_1)n]$ since this quantity does not necessarily converge to the true variance in the first segment. Instead, since $\mathrm{RSS}_{2,1}$ only depends on a finite number of terms and hence is $O_p(1)$, we can use the fact that

$$\frac{\mathrm{RSS}_{2,1}}{(\alpha_2 - \alpha_1)n} = O_p\left(\frac{1}{n}\right).$$

In this case, letting $y := \mathrm{RSS}_{2,2}/((\alpha_2 - \lambda)n)$ for simplicity of notation,

$$\begin{aligned}
(\alpha_2 - \alpha_1)\log\hat{\sigma}^2_{2,2} &= (\alpha_2 - \alpha_1)\log\left[\frac{\mathrm{RSS}_{2,1}}{(\alpha_2 - \alpha_1)n} + \frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \alpha_1)n}\right] \\
&= (\alpha_2 - \alpha_1)\log\left[O_p\left(\frac{1}{n}\right) + \frac{\alpha_2 - \lambda}{\alpha_2 - \alpha_1}\cdot y \right. \\
&\qquad \left. + \frac{\alpha_1}{\alpha_2 - \alpha_1}\cdot y - \frac{\alpha_1}{\alpha_2 - \alpha_1}\cdot y\right] \\
&= (\alpha_2 - \alpha_1)\log\left[O_p\left(\frac{1}{n}\right) - \frac{\lambda - \alpha_1}{\alpha_2 - \alpha_1}\cdot y + \frac{\alpha_2 - \alpha_1}{\alpha_2 - \alpha_1}\cdot y\right] \\
&= (\alpha_2 - \alpha_1)\log\left[O_p\left(\frac{1}{n}\right) + y\right] \qquad (2.29)
\end{aligned}$$

since $\lambda - \alpha_1 = O(1/n)$. If we perform a Taylor expansion on the log function about y, then (2.29) becomes

$$(\alpha_2 - \alpha_1)\left(\log y + \frac{1}{y}\cdot O_p\left(\frac{1}{n}\right) + R\right),$$

where R is a remainder term with order $O_p(1/n^2)$. Therefore,

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

which equals

$$(\alpha_2 - \lambda)\log\left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n}\right) + O_p\left(\frac{1}{n}\right)$$

since $(\lambda - \alpha_1)\log\left(\mathrm{RSS}_{2,2}/[(\alpha_2 - \lambda)n]\right) = O_p(1/n)$. By the definition of conditional maximum likelihood estimates, for large n,

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

where $\mathrm{RSS}^*_{2,2}$ is defined as before. This implies that for large n,

$$\begin{aligned}
&\alpha_1 \log \hat{\sigma}^2_{1,2} + (\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} + (1 - \alpha_2) \log \hat{\sigma}^2_{3,2} \\
&\geq \alpha_1 \log \hat{\sigma}^2_{1,2} + (\alpha_2 - \lambda) \log \left(\frac{\mathrm{RSS}^*_{2,2}}{(\alpha_2 - \lambda)n} \right) + O_p\left(\frac{1}{n}\right) \\
&\quad + (1 - \alpha_2) \log \hat{\sigma}^2_{3,2},
\end{aligned}$$

and, therefore, (2.26) becomes

$$\begin{aligned}
&\alpha_1 \log \hat{\sigma}^2_{1,2} + (\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} + (1 - \alpha_2) \log \hat{\sigma}^2_{3,2} - \left[\lambda \log \hat{\sigma}^2_{1,1} + (1 - \lambda) \log \hat{\sigma}^2_{2,1}\right] \\
&\geq \left[\alpha_1 \log \hat{\sigma}^2_{1,2} + O_p\left(\frac{1}{n}\right) - \lambda \log \hat{\sigma}^2_{1,1}\right] \\
&\quad + \left[(\alpha_2 - \lambda) \log \left(\frac{\mathrm{RSS}^*_{2,2}}{(\alpha_2 - \lambda)n} \right) + (1 - \alpha_2) \log \hat{\sigma}^2_{3,2} - (1 - \lambda) \log \hat{\sigma}^2_{2,1}\right].
\end{aligned}$$

The first term in brackets is $O_p(\log \log n/n)$ by a Taylor series expansion argument. The second term in brackets is again of order $\log \log n/n$ a.s. by Lemma 2.1. Thus, (2.26) is greater than or equal to something of order $\log \log n/n$ in probability, and therefore, for each finite positive integer N,

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

Now consider statement (iii): $N/n < \lambda - \alpha_1 < (\log \log n)^2/\log n$ for some positive integer N. We showed previously that

$$\begin{aligned}
(\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} \ &> \ (\lambda - \alpha_1) \log \left(\frac{\mathrm{RSS}_{2,1}}{(\lambda - \alpha_1)n} \right) \\
&\quad + (\alpha_2 - \lambda) \log \left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n} \right) \qquad (2.30)
\end{aligned}$$

by concavity of the log function. In this case, we will show that with high probability, $\mathrm{RSS}_{2,1}/((\lambda - \alpha_1)n)$ is greater by a fixed (small) constant $\eta > 0$ than the true variance in the segment $(0, \lambda)$, σ^2_1. This will allow us to replace $\mathrm{RSS}_{2,1}$ in the inequality (2.30) with $(\lambda - \alpha_1)n(\sigma^2_1 + \eta)$ for some small $\eta > 0$.

Suppose that $\phi_1 \neq \phi_2$. Intuitively, since $\alpha_1 \to \lambda$ as n tends towards infinity, for large n, the fitted AR coefficient in the segment (α_1, α_2), $\hat{\phi}$, should be close to the true coefficient of the segment $(\lambda, 1)$, ϕ_2, rather than the true coefficient of the segment $(0, \lambda)$, ϕ_1. Thus, if we define

$$c := E\left(X_{t+1} - \phi_2 X_t\right)^2,$$

where $\{X_t\}$ is a stationary process following an AR(1) model with AR coefficient ϕ_1 and white noise variance σ_1^2, then if $N \to \infty$, $\mathrm{RSS}_{2,1}/[(\lambda - \alpha_1)n]$ converges to c in probability as $n \to \infty$, and

$$\begin{aligned} c &= E\left(X_{t+1} - \phi_2 X_t\right)^2 \\ &= E\left[(\phi_1 X_t - \phi_2 X_t) + (X_{t+1} - \phi_1 X_t)\right]^2 \\ &= E((\phi_1 - \phi_2)X_t)^2 + E[2(\phi_1 - \phi_2)X_t(X_{t+1} - \phi_1 X_t)] \\ &\quad + E(X_{t+1} - \phi_1 X_t)^2 \\ &= E((\phi_1 - \phi_2)X_t)^2 + \sigma_1^2 \\ &> \sigma_1^2. \end{aligned}$$

The cross-product term is zero since $X_{t+1} = \phi_1 X_t + \sigma_1 \epsilon_{t+1}$ where $\{\epsilon_t\}$ is white noise with mean 0 and unit variance, X_t is independent of $\{\epsilon_j\}_{j>t}$, and thus,

$$E(X_t(X_{t+1} - \phi_1 X_t)) = \sigma_1 E(X_t \epsilon_{t+1}) = 0.$$

We would like to show that for every $\delta > 0$, there exists a positive integer N such that with probability greater than $1 - \delta$,

$$\frac{\mathrm{RSS}_{2,1}}{(\lambda - \alpha_1)n} > \frac{c + \sigma_1^2}{2} \tag{2.31}$$

for every α_1 such that $[(\lambda - \alpha_1)n] = N + 1, N + 2, \ldots, (\log\log n)^2$. Recall that $\mathrm{RSS}_{2,1}$ depends not only on α_1, but also on α_2, since the fitted AR coefficient in the residual sum of squares is calculated between observations $[\alpha_1 n]$ and $[\alpha_2 n]$.

To prove (2.31), first assume that the fitted coefficient between α_1 and α_2 is exactly the true AR(1) coefficient of the second true segment, ϕ_2. Then

$$\mathrm{RSS}_{2,1} = \sum_{t=[\alpha_1 n]}^{[\lambda n]-1} (X_t - \phi_2 X_{t-1})^2 .$$

Since the process $\{X_t\}$ is stationary in reverse time, we have, with probability 1,

$$\sum_{t=-1}^{-K} \frac{(X_t - \phi_2 X_{t-1})^2}{K} \to c$$

as K goes to infinity. This implies that for every $\delta > 0$, there exists an $N > 0$ such that with probability greater than $1 - \delta$,

$$\sum_{t=-1}^{-K} \frac{(X_t - \phi_2 X_{t-1})^2}{K} > \frac{c + \sigma_1^2}{2}$$

for all $K > N$. Now, stationarity of the process implies that with probability greater than $1 - \delta$,

$$\frac{\mathrm{RSS}_{2,1}}{(\lambda - \alpha_1)n} > \frac{c + \sigma_1^2}{2}$$

for all α_1 such that $[(\lambda - \alpha_1)n] = N + 1, N + 2, \ldots, (\log\log n)^2$.

To remove the assumption that the fitted coefficient is the true second coefficient, note that for any ϕ',

$$\begin{aligned} &\frac{\sum_{t=-1}^{-K} [(X_t - \phi_2 X_{t-1})^2 - (X_t - \phi' X_{t-1})^2]}{K} \\ &= 2(\phi' - \phi_2)\frac{\sum_{t=-1}^{-K} X_t X_{t-1}}{K} + (\phi_2^2 - (\phi')^2)\frac{\sum_{t=-1}^{-K} X_{t-1}^2}{K}. \end{aligned}$$

It follows that for any $\delta > 0$, there exists $N > 0$ and (small) $r > 0$ such that with probability greater than $1 - \delta$,

$$\frac{\sum_{t=-1}^{-K} (X_t - \phi' X_{t-1})^2}{K} > \frac{c + 2\sigma_1^2}{3}$$

for all $K > N$ and all ϕ' satisfying $|\phi_2 - \phi'| < r$. Finally, note that the fitted coefficient in the segment (α_1, α_2) with $(\log\log n)^2 < (\lambda - \alpha_1)n$ and $(\alpha_2 - \lambda)n > \epsilon n/2$ should be close to ϕ_2 with high probability.

Therefore, since (2.31) holds, in the case where $\phi_1 \neq \phi_2$, for any $\delta > 0$, there exists an integer $N > 0$ such that with probability greater than $1 - \delta$,

$$\frac{\mathrm{RSS}_{2,1}}{(\lambda - \alpha_1)n} > \sigma_1^2 + \eta \tag{2.32}$$

for every α_1 such that $[(\lambda - \alpha_1)n] = N+1, N+2, \ldots, (\log\log n)^2$ and for some $\eta > 0$. Thus, for $N/n < \lambda - \alpha_1 < (\log\log n)^2/\log n$ and for $\lambda + \epsilon/2 < \alpha_2 < 1 - \epsilon$,

$$\begin{aligned}(\alpha_2 - \alpha_1)\log\hat{\sigma}_{2,2}^2 &> (\lambda - \alpha_1)\log\left(\frac{\mathrm{RSS}_{2,1}}{(\lambda - \alpha_1)n}\right) + (\alpha_2 - \lambda)\log\left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n}\right) \\ &> (\lambda - \alpha_1)\log\left(\sigma_1^2 + \eta\right) + (\alpha_2 - \lambda)\log\left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n}\right)\end{aligned}$$

for large n with high probability, so (2.26) becomes

$$\begin{aligned}&\alpha_1\log\hat{\sigma}_{1,2}^2 + (\alpha_2 - \alpha_1)\log\hat{\sigma}_{2,2}^2 + (1 - \alpha_2)\log\hat{\sigma}_{3,2}^2 \\ &\quad - \left[\lambda\log\hat{\sigma}_{1,1}^2 + (1 - \lambda)\log\hat{\sigma}_{2,1}^2\right] \\ &\geq \left[\alpha_1\log\hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1)\log\left(\sigma_1^2 + \eta\right) - \lambda\log\hat{\sigma}_{1,1}^2\right] \\ &\quad + \Bigg[(\alpha_2 - \lambda)\log\left(\frac{\mathrm{RSS}_{2,2}^*}{(\alpha_2 - \lambda)n}\right) \\ &\quad + (1 - \alpha_2)\log\hat{\sigma}_{3,2}^2 - (1 - \lambda)\log\hat{\sigma}_{2,1}^2\Bigg].\end{aligned} \tag{2.33}$$

Consider the first term in brackets in (2.33). Let $\hat{\phi}_{1:\alpha_1}$ denote the fitted AR coefficient between observations 1 and $[n\alpha_1] - 1$, and let $\hat{\phi}_{1:\lambda}$ denote the fitted AR coefficient between observations 1 and $[n\lambda] - 1$. Then

$$\hat{\sigma}_{1,1}^2 = \frac{\sum_{t=1}^{[\lambda n]-1}(X_{t+1} - \hat{\phi}_{1:\lambda}X_t)^2}{\lambda n}$$

and

$$\hat{\sigma}_{1,2}^2 = \frac{\sum_{t=1}^{[\alpha_1 n]-1}(X_{t+1} - \hat{\phi}_{1:\alpha_1}X_t)^2}{\alpha_1 n}.$$

Define

$$\tilde{\sigma}_1^2 = \frac{\sum_{t=[\alpha_1 n]}^{[\lambda n]-1}(X_{t+1} - \hat{\phi}_{1:\alpha_1}X_t)^2}{(\lambda - \alpha_1)n}.$$

Then since

$$\sum_{t=1}^{[\lambda n]-1} (X_{t+1} - \hat{\phi}_{1:\lambda} X_t)^2 \leq \sum_{t=1}^{[\lambda n]-1} (X_{t+1} - \hat{\phi}_{1:\alpha_1} X_t)^2,$$

we have

$$\lambda n \hat{\sigma}_{1,1}^2 \leq \alpha_1 n \hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1) n \tilde{\sigma}_1^2,$$

which implies

$$\hat{\sigma}_{1,1}^2 \leq \frac{\alpha_1}{\lambda}\, \hat{\sigma}_{1,2}^2 + \frac{\lambda - \alpha_1}{\lambda}\, \tilde{\sigma}_1^2.$$

Also, since $\hat{\sigma}_{1,1}^2$, $\hat{\sigma}_{1,2}^2$, and $\tilde{\sigma}_1^2$ all converge to σ_1^2 as n goes to infinity, by choosing N large enough, we have $|\hat{\sigma}_{1,1}^2 - \sigma_1^2| < \eta'$, $|\hat{\sigma}_{1,2}^2 - \sigma_1^2| < \eta'$, and $|\tilde{\sigma}_1^2 - \sigma_1^2| < \eta'$ with high probability, where η' is a small positive number depending on η to be determined later. If $|\hat{\sigma}_{1,2}^2 - \sigma_1^2| < \eta'$ and $|\tilde{\sigma}_1^2 - \sigma_1^2| < \eta'$, then the first term in brackets in (2.33) becomes

$$\begin{aligned}
&\alpha_1 \log \hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1) \log\left(\sigma_1^2 + \eta\right) - \lambda \log \hat{\sigma}_{1,1}^2 \\
&> \alpha_1 \log \hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1) \log\left(\sigma_1^2 + \eta\right) - \lambda \log \left[\frac{\alpha_1}{\lambda}\, \hat{\sigma}_{1,2}^2 + \frac{\lambda - \alpha_1}{\lambda}\, \tilde{\sigma}_1^2\right] \\
&> \alpha_1 \log \hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1) \log\left(\sigma_1^2 + \eta\right) - \lambda \log \left[\frac{\alpha_1}{\lambda}\, \hat{\sigma}_{1,2}^2 + \frac{\lambda - \alpha_1}{\lambda}\, (\sigma_1^2 + \eta')\right] \\
&= \alpha_1 \log \hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1) \log\left(\sigma_1^2 + \eta\right) - \lambda \log \left[\hat{\sigma}_{1,2}^2 + \left(1 - \frac{\alpha_1}{\lambda}\right)(\sigma_1^2 + \eta' - \hat{\sigma}_{1,2}^2)\right] \\
&= \alpha_1 \log \hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1) \log\left(\sigma_1^2 + \eta\right) \\
&\quad - \lambda \log \hat{\sigma}_{1,2}^2 - \lambda \log \left[1 + \left(1 - \frac{\alpha_1}{\lambda}\right)\left(\frac{\sigma_1^2 + \eta'}{\hat{\sigma}_{1,2}^2} - 1\right)\right] \\
&\geq (\lambda - \alpha_1)[\log(\sigma_1^2 + \eta) - \log \hat{\sigma}_{1,2}^2] - \lambda \left(1 - \frac{\alpha_1}{\lambda}\right)\left(\frac{\sigma_1^2 + \eta'}{\hat{\sigma}_{1,2}^2} - 1\right)
\end{aligned}$$

since $\log(1 + x) \leq x$, and thus,

$$\begin{aligned}
&\alpha_1 \log \hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1) \log\left(\sigma_1^2 + \eta\right) - \lambda \log \hat{\sigma}_{1,1}^2 \\
&> (\lambda - \alpha_1)\left[\log\left(\frac{\sigma_1^2 + \eta}{\hat{\sigma}_{1,2}^2}\right) - \frac{\sigma_1^2 + \eta'}{\hat{\sigma}_{1,2}^2} + 1\right] \\
&> (\lambda - \alpha_1)\left[\log\left(\frac{\sigma_1^2 + \eta}{\sigma_1^2 + \eta'}\right) - \frac{\sigma_1^2 + \eta'}{\sigma_1^2 - \eta'} + 1\right] \qquad (2.34)
\end{aligned}$$

since $\sigma_1^2 - \eta' < \hat{\sigma}_{1,2}^2 < \sigma_1^2 + \eta'$. Now we choose $\eta' > 0$ so small that the term inside the brackets in (2.34) is positive. This implies that for any given $\delta > 0$, there exists a positive integer N such that

$$P\big(\alpha_1 \log \hat{\sigma}_{1,2}^2 + (\lambda - \alpha_1) \log\left(\sigma_1^2 + \eta\right) - \lambda \log \hat{\sigma}_{1,1}^2 \geq 0$$
$$\forall\ \alpha_1 = \lambda - N/n, \lambda - (N+1)/n, \ldots, \lambda - \log\log n/\log n\big) > 1 - \delta$$

for large n. Again, the second term in brackets in (2.33) is of order $\log\log n/n$ by Lemma 2.1. Thus, (2.26) is greater than or equal to something nonnegative plus something of order $\log\log n/n$ in probability, and therefore, for every $\delta > 0$, there exists a positive integer N such that

$$\begin{aligned} P\big(&\mathrm{MDL}(2, \alpha_1, \alpha_2; 1, 1) > \mathrm{MDL}(1, \lambda; 1)\ \ \forall\, \alpha_1, \alpha_2 : \\ &\lambda - (\log\log n)^2/\log n < \alpha_1 < \lambda - N/n; \\ &\lambda + \epsilon/2 < \alpha_2 < 1 - \epsilon\big) \quad > \quad 1 - \delta \end{aligned} \tag{2.35}$$

for sufficiently large n.

In the case where $\phi_1 = \phi_2 = \phi$, but $\sigma_1^2 \neq \sigma_2^2$, the argument is slightly modified. Note that since the three fitted autoregressive coefficients for the segments $(0, \alpha_1)$, (α_1, α_2), and $(\alpha_2, 1)$ should all be close to the true autoregressive coefficient, ϕ, we have $\mathrm{RSS}_{2,1}/((\lambda - \alpha_1)n) \to \sigma_1^2$ and $\mathrm{RSS}_{2,2}/((\alpha_2 - \lambda)n) \to \sigma_2^2$. Define

$$R := \frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right).$$

Since $x > \log(1+x)$ for all $x \neq 0$, we know that $R > 0$. By expanding $\log(u + w(v-u))$ about u using a Taylor series expansion, we can show that there exists a small $c > 0$ such that for every u, v, w satisfying $|u - \sigma_2^2| < c$, $|v - \sigma_1^2| < c$ and $0 < w < c$,

$$\log(u + w(v-u)) \geq \log u + \frac{1}{u} w(v-u) - \frac{wR}{3}.$$

Since $\lim_{c \to 0}(v/u - 1 - \log(v/u)) = R$, by continuity, $R > 0$ implies that

$$\frac{v}{u} - 1 - \log\left(\frac{v}{u}\right) > \frac{2R}{3}$$

for u and v satisfying $|u-\sigma_2^2|<c$ and $|v-\sigma_1^2|<c$ (if necessary, making c smaller). Then, since

$$\begin{aligned}(1-w)\log u + w\log v &= \log u + w\log\left(\frac{v}{u}\right)\\ &\le \log u + w\left(\frac{v}{u}-1-\frac{2R}{3}\right),\end{aligned}$$

it follows that

$$\begin{aligned}\log[(1-w)u+wv] &= \log[u+w(v-u)]\\ &\ge \log u + \frac{1}{u}w(v-u) - \frac{wR}{3}\\ &\ge (1-w)\log u + w\log v + \frac{wR}{3} \qquad (2.36)\end{aligned}$$

for every u, v, w satisfying $|u-\sigma_2^2|<c$, $|v-\sigma_1^2|<c$ and $0<w<c$. Letting $(\lambda-\alpha_1)/(\alpha_2-\alpha_1)=w$, $\mathrm{RSS}_{2,1}/((\lambda-\alpha_1)n)=v$, and $\mathrm{RSS}_{2,2}/((\alpha_2-\lambda)n)=u$, we can now apply (2.36) to $(\alpha_2-\alpha_1)\log\hat{\sigma}_{2,2}^2$ as follows:

$$\begin{aligned}(\alpha_2-\alpha_1)\log\hat{\sigma}_{2,2}^2 &= (\alpha_2-\alpha_1)\log[wv+(1-w)u]\\ &> (\alpha_2-\alpha_1)[(1-w)\log u + w\log v + wR/3]\\ &= (\alpha_2-\lambda)\log u + (\lambda-\alpha_1)(\log v + R/3)\\ &> (\alpha_2-\lambda)\log u + (\lambda-\alpha_1)\log Q,\end{aligned}$$

where Q is any number satisfying $Q>\sigma_1^2$ and $\log Q<\log\sigma_1^2+R/3$. (Note that Q plays the role of $\sigma_1^2+\eta$ in (2.32).) The first inequality above holds if $|u-\sigma_2^2|<c$, $|v-\sigma_1^2|<c$ and $0<w<c$. For n large, $w<c$ since $\lambda-\alpha_1<(\log\log n)^2/\log n$. Also, for large n and large N, $\mathrm{RSS}_{2,1}/((\lambda-\alpha)n)$ and $\mathrm{RSS}_{2,2}/((\alpha_2-\lambda)n)$ are close to σ_1^2 and σ_2^2, respectively, with high probability. The second inequality holds when v is close to σ_1^2. It follows that the above inequality holds with probability greater than $1-\delta$ if N is large enough.

Therefore, for $N/n<\lambda-\alpha_1<(\log\log n)^2/\log n$ and for $\lambda+\epsilon/2<\alpha_2<1-\epsilon$,

$$(\alpha_2-\alpha_1)\log\hat{\sigma}_{2,2}^2 > (\alpha_2-\lambda)\log\left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2-\lambda)n}\right)+(\lambda-\alpha_1)\log Q$$

for large n and large N with high probability, so (2.26) becomes

$$\begin{aligned}
&\alpha_1 \log \hat{\sigma}^2_{1,2} + (\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} + (1 - \hat{\alpha}_2) \log \hat{\sigma}^2_{3,2} \\
&\quad - \left[\lambda \log \hat{\sigma}^2_{1,1} + (1 - \lambda) \log \hat{\sigma}^2_{2,1}\right] \\
&\geq \left[\alpha_1 \log \hat{\sigma}^2_{1,2} + (\lambda - \alpha_1) \log Q - \lambda \log \hat{\sigma}^2_{1,1}\right] \\
&\quad + \left[(\alpha_2 - \lambda) \log \left(\frac{\mathrm{RSS}^*_{2,2}}{(\alpha_2 - \lambda)n}\right)\right. \\
&\quad \left. + (1 - \alpha_2) \log \hat{\sigma}^2_{3,2} - (1 - \lambda) \log \hat{\sigma}^2_{2,1}\right].
\end{aligned}$$

As in the case where $\phi_1 \neq \phi_2$, since $\hat{\sigma}^2_{1,2} \to \sigma^2_1$ and $\hat{\sigma}^2_{1,1} \to \sigma^2_1$ as n goes to infinity, the first term in brackets is nonnegative with probability approaching 1. Again, the second term in brackets is of order $\log\log n/n$ by Lemma 2.1. Thus, (2.26) is greater than or equal to something nonnegative plus something of order $\log\log n/n$ in probability, and therefore, for every $\delta > 0$, there exists a positive integer N such that (2.35) again holds for sufficiently large n.

Suppose now that the true process still has one true change-point where each segment follows an AR(1) model, but now the means of the two true segments, μ_1 and μ_2, are not necessarily zero. Then when examining residual sums of squares, rather than minimizing the quantity $\sum(X_t - aX_{t-1})^2$ with respect to a, we minimize the quantity $\sum(X_t - a_0 - a_1 X_{t-1})^2$ with respect to a_0 and a_1. In this case, we denote the true AR parameters in the first true segment by $\phi_{1,0}$ and $\phi_{1,1}$, and the true AR parameters in the second true segment by $\phi_{2,0}$ and $\phi_{2,1}$, where $\phi_{1,0} = \mu_1(1 - \phi_{1,1})$ and $\phi_{2,0} = \mu_2(1 - \phi_{2,1})$. For cases (i) $(\log\log n)^2/\log n < \lambda - \alpha_1$, and (ii) $\lambda - \alpha_1 < N/n$ for some positive integer N, this difference has no effect on the previous arguments. For case (iii) $N/n < \lambda - \alpha_1 < (\log\log n)^2/\log n$ for some positive integer N, we need only show again that with high probability, $\mathrm{RSS}_{2,1}/((\lambda - \alpha_1)n)$ is greater by $\eta > 0$ than the true variance in the segment $(0, \lambda)$.

Under the assumption of case (iii), if we assume $\phi_{1,1} \neq \phi_{2,1}$ and define

$$c := E\left(X_{t+1} - \phi_{2,0} - \phi_{2,1}X_t\right)^2,$$

where $\{X_t\}$ is a stationary process following an AR(1) model with mean μ_1, AR coefficients $\phi_{1,0} = \mu_1(1 - \phi_{1,1})$ and $\phi_{1,1}$, and white noise variance σ_1^2, then if $N \to \infty$, $\mathrm{RSS}_{2,1}/((\lambda - \alpha_1)n)$ converges to c in probability as $n \to \infty$. To see that $c > \sigma_1^2$, we have

$$\begin{aligned} c &= E\left(X_{t+1} - \phi_{2,0} - \phi_{2,1}X_t\right)^2 \\ &= E\left((X_{t+1} - \mu_2) - \phi_{2,1}(X_t - \mu_2)\right)^2 \\ &= E\left[(\phi_{1,1} - \phi_{2,1})(X_t - \mu_1) + (X_{t+1} - \mu_1) - \phi_{1,1}(X_t - \mu_1)\right]^2 \\ &= E((\phi_{1,1} - \phi_{2,1})(X_t - \mu_1))^2 + \sigma_1^2 \\ &> \sigma_1^2. \end{aligned}$$

By replacing X_t by $X_t - \mu_1$ in the mean zero argument, (2.35) follows accordingly. If $\phi_{1,1} = \phi_{2,1} = \phi$ and $\sigma_1^2 \neq \sigma_2^2$, the proof of (2.35) is analogous to the proof in the mean zero case.

In the mean zero case, $\mu_1 = \mu_2 = 0$. Now, we may have the case where $\mu_1 \neq \mu_2$. For case (iii) $N/n < \lambda - \alpha_1 < (\log\log n)^2/\log n$ for some positive integer N, we have shown how to prove that with high probability, $\mathrm{RSS}_{2,1}/((\lambda - \alpha_1)n)$ is greater than the true variance in the segment $(0, \lambda)$ when $\phi_{1,1} \neq \phi_{2,1}$ or when $\phi_{1,1} = \phi_{2,1} = \phi$ and $\sigma_1^2 \neq \sigma_2^2$. We must now address the possibility that $\phi_{1,1} = \phi_{2,1} = \phi$, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $\mu_1 \neq \mu_2$. In this case, $\phi_{1,0} = \mu_1(1 - \phi)$ and $\phi_{2,0} = \mu_2(1 - \phi)$. As before, if $N \to \infty$, $\mathrm{RSS}_{2,1}/((\lambda - \alpha_1)n)$ converges to c in probability as $n \to \infty$ where c is defined as

$$c := E\left(X_{t+1} - \phi_{2,0} - \phi X_t\right)^2$$

with $\{X_t\}$ a stationary process following an AR(1) model with mean μ_1, AR coefficients $\phi_{1,0}$ and ϕ, and white noise variance σ^2. We can show that $c > \sigma^2$ as

follows:

$$\begin{aligned} c &= E\left(X_{t+1} - \phi_{2,0} - \phi X_t\right)^2 \\ &= E\left((X_{t+1} - \mu_2) - \phi(X_t - \mu_2)\right)^2 \\ &= E\left[(X_{t+1} - \mu_1) - \phi(X_t - \mu_1) + (\mu_1 - \mu_2)(1-\phi)\right]^2 \\ &= E(\sigma\epsilon_{t+1})^2 + 2(\mu_1 - \mu_2)(1-\phi)\sigma E(\epsilon_{t+1}) + (\mu_1 - \mu_2)^2(1-\phi)^2 \\ &= \sigma^2 + (\mu_1 - \mu_2)^2(1-\phi)^2 \\ &> \sigma^2. \end{aligned}$$

The rest of the argument to show (2.35) follows as in the case where $\phi_{1,1} \neq \phi_{2,1}$.

Keeping the outline of the proof for this special case in mind, we will now prove Theorem 2.1.

Proof of Theorem 2.1. It suffices to show that

$$\lim_{n\to\infty} P\left(\mathrm{MDL}(m_0+1, \hat{\boldsymbol{\alpha}}; \hat{p}_1, \ldots, \hat{p}_{m_0+2}) > \mathrm{MDL}(m_0, \hat{\boldsymbol{\lambda}}; p_1^0, \ldots, p_{m_0+1}^0)\right) = 1$$

since this implies

$$\lim_{n\to\infty} P\left(\inf_{\substack{m\neq m_0 \\ m<M}} \{\mathrm{MDL}(m, \hat{\boldsymbol{\alpha}}; \hat{p}_1, \ldots, \hat{p}_{m+1})\} > \mathrm{MDL}(m_0, \hat{\boldsymbol{\lambda}}; p_1^0, \ldots, p_{m_0+1}^0)\right) = 1,$$

where M is a prespecified upper bound, and the result follows.

By definition of $\hat{\boldsymbol{\lambda}}$,

$$\mathrm{MDL}(m_0, \hat{\boldsymbol{\lambda}}; p_1^0, \ldots, p_{m_0+1}^0) \leq \mathrm{MDL}(m_0, \boldsymbol{\lambda}^0; p_1^0, \ldots, p_{m_0+1}^0),$$

where $\boldsymbol{\lambda}^0$ is a vector of the true change-point locations. Therefore, we need only show that

$$\lim_{n\to\infty} P\left(\mathrm{MDL}(m_0+1, \hat{\boldsymbol{\alpha}}; \hat{p}_1, \ldots, \hat{p}_{m_0+2}) > \mathrm{MDL}(m_0, \boldsymbol{\lambda}^0; p_1^0, \ldots, p_{m_0+1}^0)\right) = 1. \quad (2.37)$$

Equivalently, as the simple case outlined above, we can show that

$$\lim_{n\to\infty} P\Big(\mathrm{MDL}(m_0+1,\boldsymbol{\alpha};\hat{p}_1,\ldots,\hat{p}_{m_0+2}) > \mathrm{MDL}(m_0,\boldsymbol{\lambda}^0;p_1^0,\ldots,p_{m_0+1}^0)\ \forall\ \boldsymbol{\alpha}\in A_{m_0+1}^{\epsilon}\Big)=1 \tag{2.38}$$

where $A_{m_0+1}^{\epsilon}$ is defined as in (2.11).

Consider the difference

$$\begin{aligned}
&\frac{2}{n}[\mathrm{MDL}(m_0+1,\boldsymbol{\alpha};\hat{p}_1,\ldots,\hat{p}_{m_0+2})-\mathrm{MDL}(m_0,\boldsymbol{\lambda}^0;p_1^0,\ldots,p_{m_0+1}^0)]\\
&= \left(\frac{1}{n}\right)\Big[2\log(m_0+1)-2\log m_0+\sum_{j=1}^{m_0+2}(2\log\hat{p}_j+(\hat{p}_j+2)\log(\alpha_j-\alpha_{j-1}))\\
&\quad -\sum_{k=1}^{m_0+1}\left(2\log p_k^0+(p_k^0+2)\log(\lambda_k^0-\lambda_{k-1}^0)\right)\Big]\\
&\quad +\left(\frac{\log n}{n}\right)\Big[4+\sum_{j=1}^{m_0+2}\hat{p}_j-\sum_{k=1}^{m_0+1}p_k^0\Big]\\
&\quad +\sum_{j=1}^{m_0+2}(\alpha_j-\alpha_{j-1})\log\hat{\sigma}^2_{j,m_0+1}-\sum_{k=1}^{m_0+1}(\lambda_k^0-\lambda_{k-1}^0)\log\hat{\sigma}^2_{k,m_0}\\
&= O\left(\frac{1}{n}\right)+O\left(\frac{\log n}{n}\right)\\
&\quad +\sum_{j=1}^{m_0+2}(\alpha_j-\alpha_{j-1})\log\hat{\sigma}^2_{j,m_0+1}-\sum_{k=1}^{m_0+1}(\lambda_k^0-\lambda_{k-1}^0)\log\hat{\sigma}^2_{k,m_0}.
\end{aligned}\tag{2.39}$$

Since for large n, the estimated AR orders are greater than or equal to the true AR orders, the sum of the $O(1/n)$ and $O(\log n/n)$ penalty terms in (2.39) is strictly positive for large n. The $O(1/n)$ term itself may not be positive, but since it goes to zero faster than the $O(\log n/n)$ term and since the $O(\log n/n)$ term is strictly positive, the sum of the two terms is strictly positive for large n. Therefore, it suffices to show that for all $\boldsymbol{\alpha}\in A_{m_0+1}^{\epsilon}$,

$$\sum_{j=1}^{m_0+2}(\alpha_j-\alpha_{j-1})\log\hat{\sigma}^2_{j,m_0+1}-\sum_{k=1}^{m_0+1}(\lambda_k^0-\lambda_{k-1}^0)\log\hat{\sigma}^2_{k,m_0}\ \geq\ o_p\left(\frac{\log n}{n}\right). \tag{2.40}$$

This will imply (2.38), and the theorem follows.

We will show (2.40) by combining the two summations on the left side of the equality 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 $(\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+1}$ over the indices of the fitted change-point locations, $j = 1, \ldots, m_0 + 2$, we will break up each term and sum over the indices of the true change-point locations, $k = 1, \ldots, m_0 + 1$. Then within each true segment, we can look at the difference between the Model 2′ term and the Model 1′ term. This will require some fairly complicated notation, so please bear with us.

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 and $k(j)$ denotes the index of the true change-point contained in the jth fitted segment. We can partition the interval (α_{j-1}, α_j) into the 2 sub-segments $(\alpha_{j-1}, \lambda^0_{k(j)})$, $(\lambda^0_{k(j)}, \alpha_j)$, and write

$$
\begin{aligned}
&(\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+1} \\
&= (\alpha_j - \alpha_{j-1}) \log \left[\frac{\sum_{t=[\alpha_{j-1}n]}^{[\alpha_j n]-1} \left(X_t - \hat{\phi}^{\hat{p}_j 1}_{\alpha_{j-1}:\alpha_j} X_{t-1} - \cdots - \hat{\phi}^{\hat{p}_j \hat{p}_j}_{\alpha_{j-1}:\alpha_j} X_{t-\hat{p}_j} \right)^2}{(\alpha_j - \alpha_{j-1})n} \right] \\
&= (\alpha_j - \alpha_{j-1}) \log \left[\frac{1}{(\alpha_j - \alpha_{j-1})n} \left(\sum_{t=[\alpha_{j-1}n]}^{[\lambda^0_{k(j)}n]-1} (\,\cdot\,)^2 + \sum_{t=[\lambda^0_{k(j)}n]}^{[\alpha_j n]-1} (\,\cdot\,)^2 \right) \right] \\
&=: (\alpha_j - \alpha_{j-1}) \log \left[\frac{1}{(\alpha_j - \alpha_{j-1})n} \left(\mathrm{RSS}_{j,1} + \mathrm{RSS}_{j,2} \right) \right], \qquad (2.41)
\end{aligned}
$$

where $(\hat{\phi}^{\hat{p}_j 1}_{\alpha_{j-1}:\alpha_j}, \ldots, \hat{\phi}^{\hat{p}_j \hat{p}_j}_{\alpha_{j-1}:\alpha_j})$ is the vector of conditional maximum likelihood AR coefficient estimates when fitting an AR($\hat{p}_j$) model to the jth fitted segment, (α_{j-1}, α_j), and $\mathrm{RSS}_{j,i}$ is defined to be the residual sum of squares over the ith sub-segment of the jth fitted segment. Notice that both of the residual sums of squares, $\mathrm{RSS}_{j,i}$, are

based on the fitted AR coefficients between observations $[\alpha_{j-1}n]$ and $[\alpha_j n]-1$, so the index i only specifies the limits of the sum.

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 2.2 rather than Lemma 2.1 to account for the estimated AR orders, for case (i),

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

where

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

and

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

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),

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

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,

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

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 center of this segment. For instance, if $(\lambda^0_k - \lambda^0_{k-1})$ does not contain any fitted change-points, add a fitted change-point at $(\lambda^0_k + \lambda^0_{k-1})/2$. Once we have added the necessary fictitious fitted change-points, re-label the fitted change-points as $\alpha'_1, \alpha'_2, \ldots, \alpha'_{m_0+b}$ where $b \geq 1$. It follows that

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

$$\geq \sum_{j=1}^{m_0+b} \left[A_j + (\alpha'_j - \lambda^0_{k(j)}) \log \left(\frac{\mathrm{RSS}^*_{j,2}}{(\alpha'_j - \lambda^0_{k(j)})n} \right) \right] - \sum_{k=1}^{m_0+1} (\lambda^0_k - \lambda^0_{k-1}) \log \hat{\sigma}^2_{k,m_0}, \tag{2.42}$$

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 (2.42) 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 (2.42) 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:

$$\begin{aligned}
&\sum_{j=1}^{m_0+b} (\alpha'_j - \alpha'_{j-1}) \log \hat{\sigma}^2_{j,m_0+b} \\
&\geq \sum_{j=1}^{m_0+b} \left[A_j + (\alpha'_j - \lambda^0_{k(j)}) \log \left(\frac{\mathrm{RSS}^*_{j,2}}{(\alpha'_j - \lambda^0_{k(j)})n} \right) \right] \\
&= \sum_{k=1}^{m_0+1} \left[(\alpha'_{j(k)} - \lambda^0_{k-1}) \log \left(\frac{\mathrm{RSS}^*_{j(k),2}}{(\alpha'_{j(k)} - \lambda^0_{k-1})n} \right) \right. \\
&\qquad \left. + \sum_{i=1}^{r_k} (\alpha'_{j(k)+i} - \alpha'_{j(k)+i-1}) \log \hat{\sigma}^2_{j(k)+i,m_0+b} + A_{j(k)+r_k+1} \right], \qquad (2.43)
\end{aligned}$$

where for $k = 1, \ldots, m_0+1$ and $i = 1, \ldots, r_k$, $\hat{\sigma}^2_{j(k)+i,m_0+b}$ 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, the difference (2.42) becomes

$$\begin{aligned}
&\sum_{j=1}^{m_0+b} (\alpha'_j - \alpha'_{j-1}) \log \hat{\sigma}^2_{j,m_0+b} - \sum_{k=1}^{m_0+1} (\lambda^0_k - \lambda^0_{k-1}) \log \hat{\sigma}^2_{k,m_0} \\
&\geq \sum_{k=1}^{m_0+1} \left[(\alpha'_{j(k)} - \lambda^0_{k-1}) \log \left(\frac{\mathrm{RSS}^*_{j(k),2}}{(\alpha'_{j(k)} - \lambda^0_{k-1})n} \right) \right. \\
&\qquad + \sum_{i=1}^{r_k} (\alpha'_{j(k)+i} - \alpha'_{j(k)+i-1}) \log \hat{\sigma}^2_{j(k)+i,m_0+b} \\
&\qquad \left. + A_{j(k)+r_k+1} - (\lambda^0_k - \lambda^0_{k-1}) \log \hat{\sigma}^2_{k,m_0} \right]. \qquad (2.44)
\end{aligned}$$

Now, within each summand of (2.44), we can apply the same arguments as in the simple case using Lemma 2.2 rather than Lemma 2.1, and (2.40) follows. Note that if we compare the MDL of Model 1′ to the MDL of a model with $m_0 + s$ change-points, $1 \leq s \leq M - m_0$, rather than to the MDL of a model with $m_0 + 1$ change-points, the proof is identical.

For the case where the means of each segment are not necessarily zero, 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 $\sum(X_t - a_1X_{t-1} - \ldots - a_{\hat{p}}X_{t-\hat{p}})^2$, the estimates minimize the quantity $\sum(X_t - a_0 - a_1X_{t-1} - \ldots - a_{\hat{p}}X_{t-\hat{p}})^2$. Since Lemmas 2.1 and 2.2 hold for non-zero means, the result follows. □

2.7 Consistency of Auto-PARM Estimates Using Yule-Walker Estimation

In the previous section, we used conditional maximum likelihood white noise variance estimates in the definition of the minimum description length. However, Auto-PARM uses Yule-Walker white noise estimates. Yule-Walker estimation is specific to an autoregressive model of order p where the estimates are obtained by equating the sample and theoretical autocovariances at lags 0, 1,..., p. Due to its computational simplicity via the Durbin-Levinson algorithm, Yule-Walker estimation is the most common method of estimating autoregressive parameters. Also, Yule-Walker estimates always produce a causal estimated model and have the same asymptotic distribution as the conditional maximum likelihood estimates (see Section 8.10 in [8]). In this section, we will show that the estimates of the number of change-points and the AR orders are still weakly consistent when using Yule-Walker estimates.

First, we will define the Yule-Walker estimates of the autoregressive coefficients and white noise variance. We then show that the difference between Yule-Walker estimates and conditional maximum likelihood estimates is $O_p(1/n)$. This result allows us to extend Lemmas 2.1 and 2.2 and Theorem 2.1 to prove that the estimate of the number of change-points is weakly consistent when using Yule-Walker estimates in the MDL. Lemmas 2.3 and 2.4 consider the case where the true process has no change-points ($m_0 = 0$). Theorem 2.2 extends the weak consistency results of Lemmas 2.3 and 2.4 to the case where $m_0 \geq 0$.

For an autoregressive model of order p with mean μ,

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \cdots + \phi_p(X_{t-p} - \mu) + \sigma\epsilon_t,$$

where ϵ_t is a white noise process with mean zero and unit variance, the Yule-Walker estimate of the AR(p) coefficient vector is defined as

$$\hat{\boldsymbol{\phi}} = \hat{\Gamma}_p^{-1}\hat{\boldsymbol{\gamma}}_p,$$

where $\hat{\Gamma}_p$ is the sample covariance matrix $\{\hat{\gamma}(i-j)\}_{i,j=1}^{p}$ and $\hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}(1), \ldots, \hat{\gamma}(p))^T$. The Yule-Walker estimate of the white noise variance is then

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{\gamma}}_p,$$

and the Yule-Walker estimator of the mean μ is the sample mean, $\overline{X} = \sum_{t=1}^{n} X_t/n$.

In a sample of size n, Yule-Walker estimates use the following definition of the sample covariance:

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

In contrast, conditional maximum likelihood estimation (in our formulation) uses the sample covariance

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

where $\overline{X}_{a:b} := \sum_{t=a}^{b} X_t/(b-a+1)$. This is not quite equivalent to

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

where $\overline{X} = \overline{X}_{1:n}$, but the two quantities have the same asymptotic properties.

Assume again that the true model is piecewise autoregressive, as defined in (2.9), but we now estimate the number of change-points, change-point locations, and AR orders, $(m_0, \tau_1^0, \ldots, \tau_{m_0}^0, p_1^0, \ldots, p_{m_0+1}^0)$, by minimizing the minimum description length,

$$\begin{aligned}\mathrm{MDL}_Y(m, \boldsymbol{\lambda}; \boldsymbol{p}) &= \log m + (m+1)\log n + \sum_{k=1}^{m+1}\log p_k \\ &\quad + \sum_{k=1}^{m+1}\frac{p_k+2}{2}\log\left((\lambda_k - \lambda_{k-1})n\right) \\ &\quad + \sum_{k=1}^{m+1}\frac{(\lambda_k - \lambda_{k-1})n}{2}\log(2\pi\hat{\sigma}_k^2), \end{aligned} \tag{2.45}$$

with respect to $m \leq M$, $0 \leq \boldsymbol{p} \leq P$, and $\boldsymbol{\lambda} \in A_m^{\epsilon}$, where $\hat{\sigma}_k^2$ is the Yule-Walker estimate of the noise variance when fitting a p_kth order AR model to the kth segment. Here we adapt the proofs of Lemmas 2.1 and 2.2 and Theorem 2.1 for Yule-Walker estimates. The fundamental difference between the proofs using conditional maximum likelihood estimation and the proofs using Yule-Walker estimation is in the definition of the sample covariances. Therefore, we would first like to quantify the order of the difference between the two sample covariance definitions.

For notational simplicity, assume that the mean of the process is known to be zero. Consider the case where there are no change-points in the true process, i.e., $m_0 = 0$. Suppose we fit one change-point to this data at observation $\tau = [\lambda n]$. Using conditional maximum likelihood estimation, the sample covariances in the last fitted segment involve terms like

$$\frac{1}{(1-\lambda)n} \sum_{t=[\lambda n]}^{n} X_t X_{t-h}.$$

However, if we use Yule-Walker estimation, the sample covariances in the last fitted segment will involve terms like

$$\frac{1}{(1-\lambda)n} \sum_{t=[\lambda n]+h}^{n} X_t X_{t-h}.$$

We can examine the difference between the two sample covariances in the last fitted segment as follows:

$$\begin{aligned} & n\left|\frac{1}{n}\sum_{t=[\lambda n]}^{n} X_t X_{t-h} - \frac{1}{n}\sum_{t=[\lambda n]+h}^{n} X_t X_{t-h}\right| \\ &= \left|X_{[\lambda n]} X_{[\lambda n]-h} + \cdots + X_{[\lambda n]+h-1} X_{[\lambda n]-1}\right| \\ &\leq \left|X_{[\lambda n]} X_{[\lambda n]-h}\right| + \cdots + \left|X_{[\lambda n]+h-1} X_{[\lambda n]-1}\right|. \end{aligned} \tag{2.46}$$

Note that λ is not treated as being fixed in the expression above. Define (2.46) as

$$W_\lambda := \left|X_{[\lambda n]} X_{[\lambda n]-h}\right| + \cdots + \left|X_{[\lambda n]+h-1} X_{[\lambda n]-1}\right|.$$

Using stationarity of the process, then, for any $M > 0$,

$$\begin{aligned} P\left(\max_{0<\lambda<1} \left| \frac{1}{n} \sum_{t=[\lambda n]}^{n} X_t X_{t-h} - \frac{1}{n} \sum_{t=[\lambda n]+h}^{n} X_t X_{t-h} \right| > M \right) \\ \leq \; P\left(\frac{1}{n} \max_{0<\lambda<1} W_\lambda > M \right) \\ \leq \; nP\left(W_0 > nM\right) \\ \leq \; \frac{n}{nM} E(W_0) \\ \rightarrow \; 0 \quad \text{as} \quad M \rightarrow \infty, \end{aligned}$$

where the last inequality follows by Markov's inequality. This implies that

$$\sup_{0<\lambda<1} \left| \frac{1}{n} \sum_{t=[\lambda n]}^{n} X_t X_{t-h} - \frac{1}{n} \sum_{t=[\lambda n]+h}^{n} X_t X_{t-h} \right| = O_p\left(\frac{1}{n}\right),$$

and in general, when fitting m change-points,

$$\hat{\gamma}_k^*(h) - \hat{\gamma}_k(h) = O_p\left(\frac{1}{n}\right), \tag{2.47}$$

where $\hat{\gamma}_k^*(h)$ is the conditional maximum likelihood sample covariance in the kth fitted segment, and $\hat{\gamma}_k(h)$ is the Yule-Walker sample covariance in the kth fitted segment.

In the case where there are one or more change-points in the true process, still assuming that the mean of each segment is zero, we can again show (2.47) by the following argument. Assume there are $m_0 \geq 1$ true change-points and we fit $m \geq 1$ change-points to the data. Denote the true relative change-point locations by $\lambda_1^0, \ldots, \lambda_{m_0}^0$ and the fitted relative change-point locations by $\alpha_1, \ldots, \alpha_m$. Then the difference between the two sample covariance definitions in the kth fitted segment, $k = 1, \ldots, m+1$, can be examined through the expression

$$\begin{aligned} & n \left| \frac{1}{n} \sum_{t=[\alpha_{k-1} n]}^{[\alpha_k n]-1} X_t X_{t-h} - \frac{1}{n} \sum_{t=[\alpha_{k-1} n]+h}^{[\alpha_k n]-1} X_t X_{t-h} \right| \\ &= \left| X_{[\alpha_{k-1} n]} X_{[\alpha_{k-1} n]-h} + \cdots + X_{[\alpha_{k-1} n]+h-1} X_{[\alpha_{k-1} n]-1} \right| \\ &\leq \left| X_{[\alpha_{k-1} n]} X_{[\alpha_{k-1} n]-h} \right| + \cdots + \left| X_{[\alpha_{k-1} n]+h-1} X_{[\alpha_{k-1} n]-1} \right|. \end{aligned} \tag{2.48}$$

For large enough n, the indices of the h terms in (2.48) can be split over at most 2 true segments. Suppose, for example, that the sum (2.48) straddles the true relative change-point λ^0_{k-1}. Then for $t = [\alpha_{k-1}n], \ldots, [\alpha_{k-1}n] + h - 1$, the distribution of the cross-product $X_t X_{t-h}$ corresponds to one of three possibilities: (1) $t - h < [\lambda^0_{k-1}n]$ and $t \geq [\lambda^0_{k-1}n]$, (2) t, $t - h < [\lambda^0_{k-1}n]$, or (3) $t, t - h \geq [\lambda^0_{k-1}n]$.

Let n be large enough so that $[\lambda^0_j n] - [\lambda^0_{j-1}n] > h$ for any $j = 1, \ldots, m_0 + 1$, and define (2.48) as

$$W_{\alpha_{k-1}} := \left|X_{[\alpha_{k-1}n]}X_{[\alpha_{k-1}n]-h}\right| + \cdots + \left|X_{[\alpha_{k-1}n]+h-1}X_{[\alpha_{k-1}n]-1}\right|.$$

Then $W_{\alpha_{k-1}}$ may take on n distinct sets of indices, where each set of indices corresponds to one of $2hm_0 + 1$ distributions ($m_0 + 1$ possibilities if $W_{\alpha_{k-1}}$ lies entirely in one true segment plus $(2h-1)m_0$ possibilities if $W_{\alpha_{k-1}}$ straddles two true segments). Let $V_1, \ldots, V_{2hm_0+1}$ be random variables corresponding to the $2hm_0 + 1$ possible distributions of $W_{\alpha_{k-1}}$, and let κ_i, $i = 1, \ldots, 2hm_0 + 1$, be the number of $W_{\alpha_{k-1}}$'s that have distribution V_i, which implies $\sum_{i=1}^{2hm_0+1} \kappa_i = n$. Then, for any $M > 0$,

$$\begin{aligned}
P\left(\max_{\substack{0<\alpha_{k-1}<1 \\ \alpha_{k-1}+\epsilon<\alpha_k<1}} \left|\frac{1}{n}\sum_{t=[\alpha_{k-1}n]}^{[\alpha_k n]-1} X_t X_{t-h} - \frac{1}{n}\sum_{t=[\alpha_{k-1}n]+h}^{[\alpha_k n]-1} X_t X_{t-h}\right| > M\right) & \\
\leq \; P\left(\frac{1}{n}\max_{0<\alpha_{k-1}<1} W_{\alpha_{k-1}} > M\right) & \\
\leq \; \sum_{i=1}^{2hm_0+1} \kappa_i P\left(V_i > nM\right) & \\
\leq \; \sum_{i=1}^{2hm_0+1} \kappa_i \frac{E(V_i)}{nM} & \\
\to \; 0 \quad \text{as} \quad M \to \infty, &
\end{aligned}$$

where ϵ is defined in (2.10), and thus, (2.47) again holds.

If the mean of each segment is not necessarily zero, the previous argument becomes slightly more complicated. We can use the shortcut formulas

$$\hat{\gamma}(h) \;=\; \frac{1}{(\alpha_k - \alpha_{k-1})n} \sum_{t=[\alpha_{k-1}n]+h}^{[\alpha_k n]-1} X_t X_{t-h} - \overline{X}^2_{[\alpha_{k-1}n]:[\alpha_k n]-1},$$

and

$$\hat{\gamma}^*(h) = \frac{1}{(\alpha_k - \alpha_{k-1})n} \sum_{t=[\alpha_{k-1}n]}^{[\alpha_k n]-1} X_t X_{t-h} - \overline{X}_{[\alpha_{k-1}n]:[\alpha_k n]-1} \overline{X}_{[\alpha_{k-1}n]-h:[\alpha_k n]-1-h},$$

then look at the differences

$$\left| \frac{1}{(\alpha_k - \alpha_{k-1})n} \sum_{t=[\alpha_{k-1}n]}^{[\alpha_k n]-1} X_t X_{t-h} - \frac{1}{(\alpha_k - \alpha_{k-1})n} \sum_{t=[\alpha_{k-1}n]+h}^{[\alpha_k n]-1} X_t X_{t-h} \right|$$

and

$$\left| \overline{X}_{[\alpha_{k-1}n]:[\alpha_k n]-1} \overline{X}_{[\alpha_{k-1}n]-h:[\alpha_k n]-1-h} - \overline{X}^2_{[\alpha_{k-1}n]:[\alpha_k n]-1} \right|$$

separately. The supremum over $0 < \alpha_{k-1} < 1$ of the first difference is again $O_p(1/n)$, and it is straightforward to show that the supremum over $0 < \alpha_{k-1} < 1$ of the second difference is also of order $1/n$ in probability.

In extending Lemmas 2.1 and 2.2 and Theorem 2.1, we can use (2.47) to show that for the kth fitted segment found by minimizing the MDL calculated using Yule-Walker estimates, the difference between the conditional maximum likelihood white noise estimate, denoted by $\tilde{\sigma}^{*2}_k$, and the Yule-Walker white noise estimate, denoted by $\hat{\sigma}^2_k$, is $O_p(1/n)$.[5] Since each of these estimates are functions of sample covariances, we use the mean value theorem for multivariable functions and (2.47) to show that

$$\log \tilde{\sigma}^{*2}_k - \log \hat{\sigma}^2_k = O_p\left(\frac{1}{n}\right). \tag{2.49}$$

The extensions of Lemmas 2.1 and 2.2 and Theorem 2.1 will then follow immediately from this result.

Assume that the true process follows the piecewise autoregressive model defined in (2.9). Though (2.49) holds for the non-zero mean case, for notational convenience, we will only show the proof for the case where the true process mean is zero. The

---

[5]We use a "$\sim$" rather than "$\wedge$" on the conditional likelihood estimates since the estimates are obtained using estimated change-point locations found by minimizing the MDL calculated with Yule-Walker white noise estimates.

extension to the non-zero mean case follows accordingly. Suppose we fit a piecewise AR model to the data with m change-points and estimated change-point locations $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$, where the estimated locations are found by minimizing the MDL calculated with Yule-Walker estimates with respect to $\boldsymbol{\lambda} \in A_m^{\epsilon}$. Then, by definition, the Yule-Walker estimate of the white noise variance for an AR(p) model in the kth fitted segment is a function of the sample covariances:

$$\hat{\sigma}_k^2 = \hat{\gamma}_k(0) - \hat{\boldsymbol{\gamma}}_{k,p}^T \hat{\Gamma}_{k,p}^{-1} \hat{\boldsymbol{\gamma}}_{k,p},$$

where $\hat{\Gamma}_{k,p}$ is the sample covariance matrix of the kth fitted segment, $\{\hat{\gamma}_k(i-j)\}_{i,j=1}^p$, $\hat{\boldsymbol{\gamma}}_{k,p} = (\hat{\gamma}_k(1), \ldots, \hat{\gamma}_k(p))^T$, and the sample covariance function in the kth segment is defined as

$$\hat{\gamma}_k(h) = \frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=\hat{\tau}_{k-1}+h}^{\hat{\tau}_k - 1} X_t X_{t-h},$$

where $\hat{\tau}_{k-1} = [\hat{\lambda}_{k-1} n]$ and $\hat{\tau}_k = [\hat{\lambda}_k n]$. Therefore, for $k = 1, \ldots, m+1$, we can write

$$\log \hat{\sigma}_k^2 = g\left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=\hat{\tau}_{k-1}+|i-j|}^{\hat{\tau}_k - 1} X_{t-i} X_{t-j} : i, j = 0, \ldots, p \right),$$

where, as in the proof of Lemma 2.1,

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

Note that the definition of the function $g(\cdot)$ is the same for both the conditional maximum likelihood and Yule-Walker estimates; only the argument of the function changes. If we replace $\hat{\gamma}_k(h)$ by the conditional maximum likelihood estimate $\tilde{\gamma}_k^*(h)$ using the estimated change-point locations found by minimizing the MDL calculated with Yule-Walker estimates, then

$$\log \tilde{\sigma}_k^{*2} = g\left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=\hat{\tau}_{k-1}}^{\hat{\tau}_k - 1} X_{t-i} X_{t-j} : i, j = 0 \ldots, p \right).$$

Note that the conditional maximum likelihood sample covariances $\tilde{\gamma}_k^*(h)$ obtained using the Yule-Walker estimated change-point locations are different than the conditional maximum likelihood sample covariances $\hat{\gamma}_k^*(h)$, which use the estimated change-point locations found by minimizing the MDL calculated with conditional maximum likelihood variance estimates. This will be employed in the proofs later in the section.

First assume that the true process has no change-points ($m_0 = 0$). Denote the true covariance between X_t and X_{t-h} by $\gamma(h) = E[X_t X_{t-h}]$, and let $\boldsymbol{\gamma} = (\gamma(|i-j|) : i, j = 0, \ldots, p)$ be the vector of covariances ranging over lags $0, \ldots, p$ defined in such a way to match the indices of the vectors of sample covariances,

$$\hat{\boldsymbol{\gamma}}_k := \left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=\hat{\tau}_{k-1}+|i-j|}^{\hat{\tau}_k - 1} X_{t-i} X_{t-j} : i, j = 0, \ldots, p \right)$$

and

$$\tilde{\boldsymbol{\gamma}}_k^* := \left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=\hat{\tau}_{k-1}}^{\hat{\tau}_k - 1} X_{t-i} X_{t-j} : i, j = 0, \ldots, p \right)$$

for $k = 1, \ldots, m+1$. By the mean value theorem for multivariable functions, the difference in white noise variance estimates becomes

$$\begin{aligned} \log \tilde{\sigma}_k^{*2} - \log \hat{\sigma}_k^2 &= g(\tilde{\boldsymbol{\gamma}}_k^*) - g(\hat{\boldsymbol{\gamma}}_k) \\ &= \nabla g(\tilde{\boldsymbol{\gamma}})(\tilde{\boldsymbol{\gamma}}_k^* - \hat{\boldsymbol{\gamma}}_k) \\ &= O_p\left(\frac{1}{n}\right) \end{aligned} \tag{2.50}$$

where the variable $\tilde{\boldsymbol{\gamma}}$ is between $\tilde{\boldsymbol{\gamma}}^*$ and $\hat{\boldsymbol{\gamma}}$, and converges to $\boldsymbol{\gamma}$ almost surely as n tends to infinity. The gradient of $g(u_{ij} : i, j = 0, \ldots, p)$ evaluated at $\tilde{\boldsymbol{\gamma}}$ is denoted by $\nabla g(\boldsymbol{\gamma})$. It is important to note that in (2.50), $\tilde{\sigma}_k^{*2}$ is calculated using the estimated change-point locations found by minimizing the MDL which uses Yule-Walker estimates. Again, this is not the same quantity as the estimated white noise variances obtained by minimizing the MDL using conditional maximum likelihood estimates. This will be apparent in the proofs that follow.

It is straightforward to show (2.49) for the case where $m_0 \geq 1$ by defining $\boldsymbol{\gamma}$ as the linear combination of true covariances to which the sample covariances converge. For the case where the true segment means are not necessarily zero, we calculate sample covariances using the mean corrected observations. Again, the previous arguments will ensue if we define $\boldsymbol{\gamma}$ as the linear combination of true covariances to which the mean corrected sample covariances converge.

We can now extend Lemmas 2.1 and 2.2 and Theorem 2.1 to the case where Yule-Walker estimation is used. Note that when Yule-Walker estimates are used, the results corresponding to Lemmas 2.1 and 2.2 show consistency in probability rather than almost surely.

Lemma 2.3. *Assume the true process $\{X_t\}$ follows the AR(p) model given in (2.15) with no change-points ($m_0 = 0$) and initial values X_{-P}, $X_{-P+1}, \ldots, X_0$, and satisfies assumptions A1 and A2. Then for any $m \geq 1$,*

$$\lim_{n\to\infty} P\left(\mathrm{MDL}_Y(0;p) < \inf_{\boldsymbol{\lambda}\in A_m^\epsilon} \mathrm{MDL}_Y(m,\boldsymbol{\lambda};p,\ldots,p)\right) = 1.$$

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

$$\begin{aligned}
&\frac{2}{n}\left[\mathrm{MDL}_Y(m,\hat{\boldsymbol{\lambda}};p,\ldots,p) - \mathrm{MDL}_Y(0;p)\right] \\
&= \frac{2\log m}{n} + m(p+4)\frac{\log n}{n} + \frac{2m\log p}{n} \\
&\quad + \frac{p+2}{n}\sum_{k=1}^{m+1}\log(\hat{\lambda}_k - \hat{\lambda}_{k-1}) \\
&\quad + \sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\hat{\sigma}_k^2 - \log\hat{\sigma}^2 \\
&= m(p+4)\frac{\log n}{n} + O\left(\frac{1}{n}\right) + \sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\hat{\sigma}_k^2 - \log\hat{\sigma}^2 \qquad (2.51)
\end{aligned}$$

We will show that (2.51) is strictly positive for n large in probability by showing that the quantity

$$\sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\hat{\sigma}_k^2 - \log\hat{\sigma}^2 \qquad (2.52)$$

is of order $\log\log n/n$ in probability.

Let $\tilde{\sigma}_k^{*2}$ denote the conditional maximum likelihood estimate of the white noise variance in the kth fitted segment where the change-point locations are obtained by minimizing $\mathrm{MDL}_Y(m, \boldsymbol{\lambda}; p, \ldots, p)$, and $\hat{\sigma}^{*2}$, the conditional maximum likelihood estimate of the white noise variance for the entire data set. Then by (2.49), (2.52) can be expressed as

$$\begin{aligned}
&\sum_{k=1}^{m+1}\left[(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\tilde{\sigma}_k^{*2} + (\hat{\lambda}_k - \hat{\lambda}_{k-1})(\log\hat{\sigma}_k^2 - \log\tilde{\sigma}_k^{*2})\right] \\
&\qquad - \log\hat{\sigma}^{*2} + (\log\hat{\sigma}^{*2} - \log\hat{\sigma}^2) \\
&= \sum_{k=1}^{m+1}(\hat{\lambda}_k - \hat{\lambda}_{k-1})\log\tilde{\sigma}_k^{*2} - \log\hat{\sigma}^{*2} + O_p\left(\frac{1}{n}\right). \qquad (2.53)
\end{aligned}$$

Denote the fitted change-point locations obtained by minimizing the MDL calculated using conditional maximum likelihood estimation by $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in A_m^\epsilon}\left\{\frac{2}{n}\mathrm{MDL}(m, \boldsymbol{\alpha}; p, \ldots, p)\right\}$, and the conditional maximum likelihood white noise variance estimates using these fitted locations by $\hat{\sigma}_k^{*2}$. Then (2.53) is greater than or equal to

$$\sum_{k=1}^{m+1}(\hat{\alpha}_k - \hat{\alpha}_{k-1})\log\hat{\sigma}_k^{*2} - \log\hat{\sigma}^{*2} + O_p\left(\frac{1}{n}\right),$$

which, by Lemma 2.1, is $O(\log\log n/n)$. Therefore, (2.51) becomes

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

which implies

$$\begin{aligned}
&\frac{n}{\log n}\cdot\frac{2}{n}\left[\mathrm{MDL}_Y(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p) - \mathrm{MDL}_Y(0; p)\right] \\
&= m(p+4) + O\left(\frac{1}{\log n}\right) + O_p\left(\frac{\log\log n}{\log n}\right) \\
&= m(p+4) + O_p\left(\frac{\log\log n}{\log n}\right). \qquad (2.54)
\end{aligned}$$

Since $m \geq 1$, for any $\epsilon > 0$, there exists an integer N such that

$$\left| O_p\left(\frac{\log\log n}{\log n}\right)\right| < \frac{m(p+4)}{2}$$

for all $n > N$. It follows that

$$P\left(\mathrm{MDL}_Y(0;p) < \inf_{\boldsymbol{\lambda}\in A_m^\epsilon} \mathrm{MDL}_Y(m,\boldsymbol{\lambda};p,\ldots,p)\right) > 1-\epsilon$$

for n large, and the result follows. □

The extension of Lemma 2.2 to the Yule-Walker case again follows from (2.49).

Lemma 2.4. *Assume the true process $\{X_t\}$ follows the AR(p) model given in (2.15) with no change-points ($m_0 = 0$) and initial values X_{-P}, $X_{-P+1},\ldots,X_0$, and satisfies assumptions A1 and A2. Then for any $m \geq 1$,*

$$\lim_{n\to\infty} P\left(\mathrm{MDL}_Y(0;p) < \inf_{\boldsymbol{\lambda}\in A_m^\epsilon} \mathrm{MDL}_Y(m,\boldsymbol{\lambda};\hat{p}_1,\ldots,\hat{p}_{m+1})\right) = 1,$$

where $\hat{p}_1,\ldots,\hat{p}_{m+1}$ are estimated from the data by minimizing the MDL calculated with Yule-Walker estimates.

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

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

We know from Lemma 2.3 that $\mathrm{MDL}_Y(m,\hat{\boldsymbol{\lambda}};p,\ldots,p) - \mathrm{MDL}_Y(0;p) > 0$ for n large in probability. Therefore, to prove Lemma 2.4, we need only show that

$$\lim_{n\to\infty} P\left(\mathrm{MDL}_Y(m,\hat{\boldsymbol{\lambda}};\hat{p}_1,\ldots,\hat{p}_{m+1}) - \mathrm{MDL}_Y(m,\hat{\boldsymbol{\lambda}};p,\ldots,p) \geq 0\right) = 1 \qquad (2.55)$$

As in the proof of Lemma 2.2, it suffices to consider the case of fitting an autoregressive model of order $p+1$ to the kth segment, and autoregressive models of order

p to each of the other m segments, so $\hat{p}_k = p+1$ and $\hat{p}_j = p$ for $j \neq k$, where p is the true order of the process. Denote the conditional maximum likelihood white noise variance estimate when fitting an AR(p) model to the segment $(\hat{\lambda}_{k-1}, \hat{\lambda}_k)$ by $\tilde{\sigma}^{*2}_{k,p}$. Then by (2.49),

$$\begin{aligned}
&\frac{2}{n}\left[\mathrm{MDL}_Y(m, \hat{\boldsymbol{\lambda}}; \hat{p}_1, \ldots, \hat{p}_{m+1}) - \mathrm{MDL}_Y(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p)\right] \\
&= \frac{2(\log(p+1) - \log p)}{n} + \frac{\log(\hat{\lambda}_k - \hat{\lambda}_{k-1})}{n} \\
&\quad + \frac{\log n}{n} + (\hat{\lambda}_k - \hat{\lambda}_{k-1})\left(\log \hat{\sigma}^2_{k,p+1} - \log \hat{\sigma}^2_{k,p}\right) \\
&= O\left(\frac{1}{n}\right) + \frac{\log n}{n} + (\hat{\lambda}_k - \hat{\lambda}_{k-1})\Big[\left(\log \hat{\sigma}^2_{k,p+1} - \log \tilde{\sigma}^{*2}_{k,p+1}\right) \\
&\quad + \left(\log \tilde{\sigma}^{*2}_{k,p} - \log \hat{\sigma}^2_{k,p}\right) + \left(\log \tilde{\sigma}^{*2}_{k,p+1} - \log \tilde{\sigma}^{*2}_{k,p}\right)\Big] \\
&= O\left(\frac{1}{n}\right) + \frac{\log n}{n} + O_p\left(\frac{1}{n}\right) + (\hat{\lambda}_k - \hat{\lambda}_{k-1})\left(\log \tilde{\sigma}^{*2}_{k,p+1} - \log \tilde{\sigma}^{*2}_{k,p}\right).
\end{aligned}$$

Since the specific change-point locations had no effect on the argument in the proof of Lemma 2.2, $\log \tilde{\sigma}^{*2}_{k,p+1} - \log \tilde{\sigma}^{*2}_{k,p} = O(\log\log n/n)$, and thus (2.55) follows. □

We can also use (2.49) to easily extend Theorem 2.1 to the Yule-Walker estimation case.

Theorem 2.2. *Assume the true process $\{X_t\}$ follows the AR(p) model given in (2.15) with m_0 change-points and initial values X_{-P}, $X_{-P+1}, \ldots, X_0$, and satisfies assumptions A1 and A2. Then $\hat{m} \xrightarrow{P} m_0$, where $\hat{m}$ is the estimated number of change-points obtained by minimizing the MDL defined using Yule-Walker white noise variance estimates.*

Proof. As in the proof of Theorem 2.1, the statement $\hat{m} \xrightarrow{P} m_0$ is equivalent to the statement

$$\lim_{n\to\infty} P\left(\inf_{\substack{m\neq m_0 \\ m<M}} \{\mathrm{MDL}_Y(m, \hat{\boldsymbol{\alpha}}; \hat{p}_1, \ldots, \hat{p}_{m+1})\} > \mathrm{MDL}_Y(m_0, \hat{\boldsymbol{\lambda}}; p^0_1, \ldots, p^0_{m_0+1})\right) = 1$$

for a fixed upper bound M. By previous arguments, this result holds if

$$\lim_{n\to\infty} P\left(\mathrm{MDL}_Y(m_0+1, \hat{\boldsymbol{\alpha}}; \hat{p}_1, \ldots, \hat{p}_{m_0+2}) > \mathrm{MDL}_Y(m_0, \hat{\boldsymbol{\lambda}}; p_1^0, \ldots, p_{m_0+1}^0)\right) = 1.$$

For any vector $\boldsymbol{\alpha} \in A_{m_0+1}^{\epsilon}$, consider the difference

$$\begin{aligned}
&\frac{2}{n}[\mathrm{MDL}_Y(m_0+1, \boldsymbol{\alpha}; \hat{p}_1, \ldots, \hat{p}_{m_0+2}) - \mathrm{MDL}_Y(m_0, \boldsymbol{\lambda}^0; p_1^0, \ldots, p_{m_0+1}^0)] \\
&\quad = O\left(\frac{\log n}{n}\right) + \sum_{j=1}^{m_0+2} (\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^2_{j,m_0+1} - \sum_{k=1}^{m_0+1} (\lambda_k^0 - \lambda_{k-1}^0) \log \hat{\sigma}^2_{k,m_0}, \quad (2.56)
\end{aligned}$$

where $\hat{\sigma}^2_{j,m_0+1}$ is the Yule-Walker estimate of the white noise variance when fitting an $\mathrm{AR}(\hat{p}_j)$ model to the segment (α_{j-1}, α_j), $\hat{\sigma}^2_{k,m_0}$ is the Yule-Walker estimate of the white noise variance when fitting an $\mathrm{AR}(p_k^0)$ model to the segment $(\lambda_{k-1}^0, \lambda_k^0)$, $\boldsymbol{\lambda}^0$ denotes the true change-point locations, and the $O(\log n/n)$ term is strictly positive. Denote the conditional maximum likelihood estimate of the white noise variance when fitting an $\mathrm{AR}(\hat{p}_j)$ model to the segment (α_{j-1}, α_j) by $\hat{\sigma}^{*2}_{j,m_0+1}$, and the conditional maximum likelihood estimate of the white noise variance when fitting an $\mathrm{AR}(p_k^0)$ model to the segment $(\lambda_{k-1}^0, \lambda_k^0)$ by $\hat{\sigma}^{*2}_{k,m_0}$. Then we can write the difference between the two sums in (2.56) as

$$\begin{aligned}
&\sum_{j=1}^{m_0+2} \left[(\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^{*2}_{j,m_0+1} + (\alpha_j - \alpha_{j-1}) \left(\log \hat{\sigma}^2_{j,m_0+1} - \log \hat{\sigma}^{*2}_{j,m_0+1}\right)\right] \\
&\qquad - \sum_{k=1}^{m_0+1} \left[(\lambda_k^0 - \lambda_{k-1}^0) \log \hat{\sigma}^{*2}_{k,m_0} + (\lambda_k^0 - \lambda_{k-1}^0) \left(\log \hat{\sigma}^2_{k,m_0} - \log \hat{\sigma}^{*2}_{k,m_0}\right)\right] \\
&\quad = O_p\left(\frac{1}{n}\right) + \sum_{j=1}^{m_0+2} (\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^{*2}_{j,m_0+1} - \sum_{k=1}^{m_0+1} (\lambda_k^0 - \lambda_{k-1}^0) \log \hat{\sigma}^{*2}_{k,m_0}, \quad (2.57)
\end{aligned}$$

where the last equation follows by (2.49). By Theorem 2.1,

$$\sum_{j=1}^{m_0+2} (\alpha_j - \alpha_{j-1}) \log \hat{\sigma}^{*2}_{j,m_0+1} - \sum_{k=1}^{m_0+1} (\lambda_k^0 - \lambda_{k-1}^0) \log \hat{\sigma}^{*2}_{k,m_0} \geq o_p\left(\frac{\log n}{n}\right),$$

and the result follows. □

Chapter 3

CONSISTENCY OF AUTO-PARM ESTIMATES FOR A PIECEWISE STATIONARY PROCESS

3.1 Introduction

In the previous chapter, we showed that the estimates of the number of change-points and the AR orders obtained by Auto-PARM are weakly consistent when the underlying model is piecewise autoregressive. Davis et al. (2006) showed that the estimated change-point locations are strongly consistent under this model when the true number of change-points is known. In this chapter, we will relax the assumption that the underlying model is piecewise autoregressive, and examine the behavior of the estimated change-point locations and number of change-points under a more general model. In fact, we only need to assume the underlying process is stationary and strong mixing plus a few easily met assumptions to show that the estimates of the change-point locations and the number of change-points are consistent.

The first section shows that when the true number of change-points is known, the estimated change-point locations are strongly consistent under a general stationary strong mixing model. In the second section, we show weak consistency of the estimated number of change-points assuming a stationary strong mixing model plus conditions to ensure that the sample covariances satisfy the functional law of the iterated logarithm. The third section demonstrates some simulation results when the underlying process is not piecewise autoregressive.

3.2 Estimating the Change-Point Locations

Assume $\{Y_t\}_{t=1}^n$ is a segmented stationary process such that the jth segment is modeled as

$$Y_t = X_{t-\tau_{j-1}+1,j}, \qquad \tau_{j-1} \leq t < \tau_j, \tag{3.1}$$

where $\tau_0 := 1$, $\tau_{m_0+1} := n+1$, and for each $j = 1, \ldots, m_0+1$, $\{X_{t,j}\}_{t=1}^{\infty}$ is a stationary ergodic process with mean $\mu_j := E(X_{t,j})$ and, for $h = 0, 1, \ldots$, autocovariance function $\gamma_j(h) := Cov(X_{t,j}, X_{t+h,j})$. We assume that the pieces $\{X_{t,j}\}$, $j = 1, \ldots, m_0+1$, are independent. The next section will require further assumptions on the underlying process, but the results in this section only require that the underlying process is segmented stationary ergodic with finite second moment where changes between segments are in the mean or autocovariance function. Define the relative change-points $0 < \lambda_1^0 < \cdots < \lambda_{m_0}^0 < 1$ such that $\tau_j = [\lambda_j^0 n]$ ($[x]$ is the integer part of x), where $\lambda_0^0 = 0$, $\lambda_{m_0+1}^0 = 1$, and $[0] := 1$.

We will emulate the proofs in the Appendix of Davis et al. (2006) which show that the estimated relative change-points $\hat{\tau}_j/n$ are strongly consistent for the true relative change-point locations λ_j^0, $j = 1, \ldots, m_0$ under the assumption that m_0, the number of change-points, is known. When the true model is piecewise autoregressive, the change between segments may be in the mean, variance, or autoregressive parameters. Changes in the variance or autoregressive parameters are equivalent to changes in the autocovariance function. Under the weaker assumption that the process is stationary, we will assume that for $j = 2, \ldots, m_0+1$, either $\mu_j \neq \mu_{j-1}$ or there exists an h such that $\gamma_j(h) \neq \gamma_{j-1}(h)$.

Throughout this section, we will approach the problem by assuming that each stationary segment of $\{Y_t\}$ has mean zero, as in [13], though the results can easily be extended to the non-zero mean case. Thus, changes between segments will be in the covariance functions. That is, for every $j = 2, \ldots, m_0+1$, there exists an h such that $\gamma_j(h) \neq \gamma_{j-1}(h)$. First, we need the following result from Davis et al. (2006).

Proposition 3.1. *[Proposition A.1 of Davis et al. (2006)] Suppose that $\{X_t\}$ is a stationary ergodic process with $E|X_1| < \infty$. Then, with probability 1, the process*

$$S_n(s) = \frac{1}{n}\sum_{t=1}^{[sn]} X_t$$

converges to the process sEX_1 on the space $D[0,1]$, the space of functions on the interval $[0,1]$ that are right-continuous and have left-hand limits.

Proof. See Davis et al. (2006).

Recall that the zero mean autoregressive model of order p is

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t,$$

where Z_t is a white noise process with mean zero and variance σ^2. The Yule-Walker estimate of the AR(p) coefficient vector is defined as

$$\hat{\boldsymbol{\phi}} = \hat{\Gamma}_p^{-1}\hat{\boldsymbol{\gamma}}_p,$$

where $\hat{\Gamma}_p$ is the sample covariance matrix $\{\hat{\gamma}(i-j)\}_{i,j=1}^{p}$ and $\hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}(1), \ldots, \hat{\gamma}(p))^T$. The Yule-Walker estimate of the white noise variance is then

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}^T\hat{\boldsymbol{\gamma}}_p.$$

The following proposition states that if we fit an autoregressive model of order p to a stationary ergodic process, then the estimates of the autoregressive coefficients and the white noise variance converge to quantities determined by the covariance function of the process. The proof of this proposition follows the proof of Proposition A.2 in Davis et al. (2006).

Proposition 3.2. *Suppose that $\{X_t\}$ is a stationary ergodic process with $E(X_1) = 0$, $E|X_1X_{1+h}| < \infty$ for all $h = 0, 1, \ldots, P$, and covariance function $\gamma(h) := E(X_1X_{1+h})$.*

For $r, s \in [0,1]$ $(r < s)$ and $p = 0, 1, \ldots, P$, let $\hat{\boldsymbol{\phi}}(r,s,p)$ and $\hat{\sigma}^2(r,s,p)$ be the Yule-Walker estimates of the AR(p) coefficient vector and process variance, respectively, based on fitting an AR(p) model to the data $X_{[rn]+1}, \ldots, X_{[sn]}$. Then, with probability 1,

$$\hat{\boldsymbol{\phi}}(r,s,p) \to \boldsymbol{\phi}(p) \qquad \text{and} \qquad \hat{\sigma}^2(r,s,p) \to \sigma^2(p),$$

where $\boldsymbol{\phi}(p)$ and $\sigma^2(p)$ are defined in the proof.

Proof. Assume $1 \le p \le P$. Define the sample covariance function

$$\hat{\gamma}(h) := \frac{1}{[sn] - [rn]} \sum_{t=[rn]+1}^{[sn]-h} X_{t+h} X_t,$$

and let $\hat{\Gamma}_p$ be the sample covariance matrix $\{\hat{\gamma}(i-j)\}_{i,j=1}^{p}$ and $\hat{\boldsymbol{\gamma}}_p = (\hat{\gamma}(1), \ldots, \hat{\gamma}(p))^T$. Then $\hat{\boldsymbol{\phi}}(r,s,p)$ and $\hat{\sigma}^2(r,s,p)$ are defined as

$$\hat{\boldsymbol{\phi}}(r,s,p) = \hat{\Gamma}_p^{-1} \hat{\boldsymbol{\gamma}}_p \qquad \text{and} \qquad \hat{\sigma}^2(r,s,p) = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}(r,s,p)^T \hat{\boldsymbol{\gamma}}_p.$$

Since $\hat{\boldsymbol{\phi}}(r,s,p)$ and $\hat{\sigma}^2(r,s,p)$ are continuous functions of the sample covariances, their asymptotic properties will follow immediately from the asymptotic properties of $\hat{\gamma}(h)$. We can examine the asymptotic properties of $\hat{\gamma}(h)$ by considering

$$\hat{\gamma}^*(h) := \frac{1}{[sn] - [rn]} \sum_{t=[rn]+1}^{[sn]} X_{t+h} X_t,$$

since $\hat{\gamma}(h)$ and $\hat{\gamma}^*(h)$ have the same limiting behavior and, using the same argument as in the proof of (2.47), $\hat{\gamma}^*(h) - \hat{\gamma}(h) = O_p(1/n)$. Because $\{X_t\}$ is a stationary ergodic process, $\{X_{t+h} X_t\}$ is also a stationary ergodic process. Applying Proposition 3.1, let B_h be the probability one set on which the partial sum process for $\{X_{t+h} X_t\}$ converges, and set

$$B = \bigcap_{h=0}^{p} B_h.$$

Since $P(B) = 1$, for $h = 0, \ldots, p$,

$$\hat{\gamma}^*(h) \quad = \quad \frac{1}{[sn] - [rn]} \sum_{t=[rn]+1}^{[sn]} X_{t+h} X_t$$

$$
\begin{aligned}
&= \left(\frac{(s-r)n}{[(s-r)n]}\right)\frac{1}{(s-r)n}\sum_{t=[rn]+1}^{[sn]} X_{t+h}X_t \\
&= \left(\frac{(s-r)n}{[(s-r)n]}\right)\frac{1}{s-r}\left[\frac{1}{n}\sum_{t=1}^{[sn]} X_{t+h}X_t - \frac{1}{n}\sum_{t=1}^{[rn]} X_{t+h}X_t\right] \\
&\to \frac{1}{s-r}\left[sE(X_{1+h}X_1) - rE(X_{1+h}X_1)\right] = \gamma(h)
\end{aligned}
$$

as n goes to infinity with probability 1. Thus, for each $h = 0, \ldots, p$, $\hat{\gamma}(h) \to \gamma(h)$ as $n \to \infty$ with probability 1, and therefore,

$$\hat{\boldsymbol{\phi}}(r,s,p) = \hat{\Gamma}_p^{-1}\hat{\boldsymbol{\gamma}}_p \quad \to \quad \Gamma_p^{-1}\boldsymbol{\gamma}_p =: \boldsymbol{\phi}(p),$$

and

$$\hat{\sigma}^2(r,s,p) = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}(r,s,p)^T\hat{\boldsymbol{\gamma}}_p \quad \to \quad \gamma(0) - \boldsymbol{\phi}(p)^T\boldsymbol{\gamma}_p =: \sigma^2(p),$$

as n goes to infinity with probability 1, where $\Gamma_p = \{\gamma(i-j)\}_{i,j=1}^p$ and $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))^T$.

In the case where $p = 0$, we model X_t as a white noise sequence with mean zero and variance σ^2. The estimate of σ^2 is $\hat{\gamma}(0)$, which converges to $\gamma(0) = \sigma^2 =: \sigma^2(0)$ with probability 1. □

The next proposition extends Proposition 3.2 to a piecewise stationary ergodic process with mean zero. The proof of this proposition follows the proof of Proposition A.3 in Davis et al. (2006).

Proposition 3.3. *Suppose that $\{Y_t\}$ is the segmented stationary ergodic process defined in (3.1) where $\mu_j = 0$ for each $j = 1, \ldots, m_0 + 1$. For r, $s \in [0,1]$ $(r < s)$ and $p = 0, \ldots, P$, let $\hat{\boldsymbol{\phi}}_Y(r,s,p)$ and $\hat{\sigma}_Y^2(r,s,p)$ be the Yule-Walker estimates of the AR(p) coefficient vector and process variance, respectively, based on fitting an AR(p) model to the data $Y_{[rn]+1}, \ldots, Y_{[sn]}$. Then, with probability 1,*

$$\hat{\boldsymbol{\phi}}_Y(r,s,p) \to \boldsymbol{\phi}_Y^*(p) \qquad \text{and} \qquad \hat{\sigma}_Y^2(r,s,p) \to \sigma_Y^{*2}(p),$$

where $\boldsymbol{\phi}_Y^*(p)$ and $\sigma_Y^{*2}(p)$ are defined in the proof.

Proof. For the kth true segment, $k = 1, \ldots, m_0 + 1$, let B_k^* be the probability one set on which

$$\frac{1}{n}\sum_{t=1}^{[ns]} X_{t,k}, \quad \frac{1}{n}\sum_{t=1}^{[ns]} |X_{t,k}|, \quad \frac{1}{n}\sum_{t=1}^{[ns]} X_{t-i,k}X_{t-j,k}, \quad \text{and} \quad \frac{1}{n}\sum_{t=1}^{[ns]} |X_{t-i,k}X_{t-j,k}|$$

converge as $n \to \infty$ for all $i, j = 1, \ldots, p$. Set

$$B^* = \bigcap_{k=1}^{m_0+1} B_k^*,$$

and note that $P(B^*) = 1$. Since r, $s \in [0,1]$, $r < s$, then $r \in [\lambda_{i-1}^0, \lambda_i^0)$ and $s \in (\lambda_{i-1+k}^0, \lambda_{i+k}^0]$, for some $i = 1, \ldots, m_0 + 1$ and $k = 0, \ldots, m_0 + 1 - i$. That is, r is in the ith segment, and s is in the $(i+k)$th segment. Then for $\omega \in B^*$, we have

$$\begin{aligned}
\hat{\gamma}_Y(h) &= \frac{1}{[sn]-[rn]} \sum_{t=[rn]+1}^{[sn]-h} Y_{t+h}Y_t \\
&= \left(\frac{(s-r)n}{[(s-r)n]}\right)\frac{1}{s-r}\left[\frac{1}{n}\sum_{t=[rn]+1}^{[\lambda_i^0 n]-h} X_{t+h-\tau_{i-1}+1,i}X_{t-\tau_{i-1}+1,i}\right. \\
&\quad + \frac{1}{n}\sum_{t=[\lambda_i^0 n]+1}^{[\lambda_{i+1}^0 n]-h} X_{t+h-\tau_i+1,i+1}X_{t-\tau_i+1,i+1} \\
&\quad \left. + \cdots + \frac{1}{n}\sum_{t=[\lambda_{i-1+k}^0 n]+1}^{[sn]-h} X_{t-\tau_{i+k-1}+1,i+k}X_{t-\tau_{i+k-1}+1,i+k} + o(1)\right] \\
&\to \frac{1}{s-r}\left[(\lambda_i^0 - r)\gamma_i(h) + (\lambda_{i+1}^0 - \lambda_i^0)\gamma_{i+1}(h) + \cdots + (s - \lambda_{i-1+k}^0)\gamma_{i+k}(h)\right] \\
&=: a_0\gamma_i(h) + \ldots + a_k\gamma_{i+k}(h)
\end{aligned}$$

by Proposition 3.2. For $1 \le p \le P$, let $\hat{\Gamma}_{Y,p}$ be the sample covariance matrix $\{\hat{\gamma}_Y(i_1 - i_2)\}_{i_1,i_2=1}^p$ and $\hat{\boldsymbol{\gamma}}_{Y,p} = (\hat{\gamma}_Y(1), \ldots, \hat{\gamma}_Y(p))^T$. Then

$$\begin{aligned}
\hat{\boldsymbol{\phi}}_Y(r,s,p) = \hat{\Gamma}_{Y,p}^{-1}\hat{\boldsymbol{\gamma}}_{Y,p} &\to \left(\sum_{l=0}^{k} a_l\Gamma_{i+l,p}\right)^{-1}\sum_{l=0}^{k} a_l\boldsymbol{\gamma}_{i+l,p} \\
&=: \boldsymbol{\phi}_Y^*(r,s,p) \qquad (3.2)
\end{aligned}$$

and

$$\begin{aligned}
\hat{\sigma}_Y^2(r,s,p) = \hat{\gamma}_Y(0) - \hat{\boldsymbol{\phi}}_Y(r,s,p)^T\hat{\boldsymbol{\gamma}}_{Y,p} &\to \sum_{l=0}^{k} a_l\gamma_{i+l}(0) - \boldsymbol{\phi}_Y^*(r,s,p)^T\sum_{l=0}^{k} a_l\boldsymbol{\gamma}_{i+l,p} \\
&=: \sigma_Y^{*2}(r,s,p) \qquad (3.3)
\end{aligned}$$

where $\Gamma_{i+l,p} = \{\gamma_{i+l}(i_1 - i_2)\}_{i_1,i_2=1}^{p}$ and $\boldsymbol{\gamma}_{i+l,p} = (\gamma_{i+l}(1), \ldots, \gamma_{i+l}(p))^T$. Note that if $k = 0$,

$$\boldsymbol{\phi}_Y^*(r, s, p) = \Gamma_{i,p}^{-1}\boldsymbol{\gamma}_{i,p}$$

and

$$\sigma_Y^{*2} = \gamma_i(0) - \boldsymbol{\gamma}_{i,p}^T\Gamma_{i,p}^{-1}\boldsymbol{\gamma}_{i,p}.$$

If $p = 0$, no AR coefficients are estimated, and

$$\hat{\sigma}_Y^2(r, s, 0) = \hat{\gamma}_Y(0) \quad \rightarrow \quad \sum_{l=0}^{k} a_l\gamma_{i+l}(0) =: \sigma_Y^{*2}(r, s, 0).$$

□

We are now ready to prove that the estimated change-point locations converge to the true change-point locations a.s. when the true number of change-points m_0 is known. If we fix the fitted autoregressive order large enough such that covariances between true segments differ at some lag between zero and the fixed autoregressive order, and fit autoregressive models of this order to the data, then the estimated change-point locations will converge to the true change-point locations. Note that we cannot choose the autoregressive order too large since we only have a finite number of observations in each segment. The proof of this result follows the proof of Proposition A.4 in Davis et al. (2006).

Theorem 3.1. *Suppose that $\{Y_t\}$ is the segmented stationary ergodic process defined in (3.1) where $\mu_j = 0$ for each $j = 1, \ldots, m_0 + 1$. Choose p^* such that for each $j = 2, \ldots, m_0 + 1$, there exists an $h \in \{0, \ldots, p^*\}$ such that $\gamma_j(h) \neq \gamma_{j-1}(h)$. Let A_m^ϵ be defined as in (2.11). If*

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in A_{m_0}^\epsilon} \left\{ \frac{2}{n}\mathrm{MDL}_Y(m_0, \boldsymbol{\lambda}; \boldsymbol{p}^*) \right\},$$

where $\boldsymbol{p}^* = p^*\mathbf{1}$ and $\mathbf{1}$ is an $(m_0 + 1)$ x 1 vector of ones, then $\hat{\boldsymbol{\lambda}} \rightarrow \boldsymbol{\lambda}^0$ a.s.

Proof. Let B^* be the probability one event described in the proof of Proposition 3.3. For $\omega \in B^*$, suppose that $\hat{\boldsymbol{\lambda}} \not\to \boldsymbol{\lambda}^0$. Then, since the sequence of $\hat{\boldsymbol{\lambda}}$ is bounded, there exists a subsequence $\{n'_k\}$ such that $\hat{\boldsymbol{\lambda}} \to \boldsymbol{\lambda}^*$ on the subsequence for some $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_{m_0}^*)$. It follows that

$$\frac{2}{n}\mathrm{MDL}_Y(m_0, \hat{\boldsymbol{\lambda}}; \boldsymbol{p}^*) \to \sum_{j=1}^{m_0+1} (\lambda_j^* - \lambda_{j-1}^*) \log \sigma_Y^{*2}(\lambda_{j-1}^*, \lambda_j^*, p^*),$$

where σ_Y^{*2} is defined in (3.3), within the proof of Proposition 3.3.

Consider the jth limiting segment, $(\lambda_{j-1}^*, \lambda_j^*)$. If $\lambda_{i-1}^0 \le \lambda_{j-1}^* < \lambda_j^* \le \lambda_i^0$ for some $i = 1, \ldots, m_0 + 1$, then

$$\sigma_Y^{*2}(\lambda_{j-1}^*, \lambda_j^*, p^*) = \gamma_i(0) - \boldsymbol{\gamma}_{i,p^*}^T \Gamma_{i,p^*}^{-1} \boldsymbol{\gamma}_{i,p^*} \ .$$

Note that this quantity is the one-step mean squared prediction error based on the previous p^* observations. In other words, $\sigma_Y^{*2}(\lambda_{j-1}^*, \lambda_j^*, p^*) = E(X_{t,i} - \hat{X}_{t,i})^2$, where

$$\hat{X}_{t,i} = \phi_{p^*1,i} X_{t-1,i} + \cdots + \phi_{p^*p^*,i} X_{t-p^*,i},$$

and

$$\boldsymbol{\phi}_{p^*,i} = (\phi_{p^*1,i}, \ldots, \phi_{p^*p^*,i})^T = \Gamma_{i,p^*}^{-1} \boldsymbol{\gamma}_{i,p^*} \ .$$

Since Γ_{i,p^*} is nonsingular, the one-step prediction coefficients are uniquely determined, and, by the projection theorem, the one-step mean squared prediction error is the minimum error over all possible coefficients.

If $\lambda_{i-1}^0 \le \lambda_{j-1}^* < \lambda_i^0 < \cdots < \lambda_{i-1+k}^0 < \lambda_j^* \le \lambda_{i+k}^0$ for some $i = 1, \ldots, m_0 + 1$ and $k = 1, \ldots, m_0 + 1 - i$, then

$$\sigma_Y^{*2}(\lambda_{j-1}^*, \lambda_j^*, p^*) = \sum_{l=0}^{k} a_l \gamma_{i+l}(0) - \left(\sum_{l=0}^{k} a_l \boldsymbol{\gamma}_{i+l,p^*} \right)^T \left(\sum_{l=0}^{k} a_l \Gamma_{i+l,p^*} \right)^{-1} \sum_{l=0}^{k} a_l \boldsymbol{\gamma}_{i+l,p^*},$$

where

$$\begin{aligned} &(a_0, a_1, \ldots, a_{k-1}, a_k) \\ &:= \frac{1}{\lambda_j^* - \lambda_{j-1}^*} \left(\lambda_i^0 - \lambda_{j-1}^*,\ \lambda_{i+1}^0 - \lambda_i^0, \ldots, \lambda_{i+k-1}^0 - \lambda_{i+k-2}^0,\ \lambda_j^* - \lambda_{i+k-1}^0 \right). \end{aligned}$$

We can again think of $\sigma_Y^{*2}(\lambda_{j-1}^*, \lambda_j^*, p^*)$ as a one-step mean squared prediction error, but since the prediction coefficients are calculated over more than one true segment, the prediction coefficients do not minimize the prediction error. Therefore,

$$\begin{aligned}
&\sigma_Y^{*2}(\lambda_{j-1}^*, \lambda_j^*, p^*) \\
&\geq\ a_0\sigma_Y^{*2}(\lambda_{i-1}^0, \lambda_i^0, p^*) + a_1\sigma_Y^{*2}(\lambda_i^0, \lambda_{i+1}^0, p^*) + \cdots \\
&\quad + a_{k-1}\sigma_Y^{*2}(\lambda_{i+k-2}^0, \lambda_{i+k-1}^0, p^*) + a_k\sigma_Y^{*2}(\lambda_{i+k-1}^0, \lambda_{i+k}^0, p^*).
\end{aligned}$$

By strict concavity of the log function,

$$\begin{aligned}
&(\lambda_j^* - \lambda_{j-1}^*)\log\sigma_Y^{*2}(\lambda_{j-1}^*, \lambda_j^*, p^*) \\
&>\ (\lambda_j^* - \lambda_{j-1}^*)\Big[a_0\log\sigma_Y^{*2}(\lambda_{i-1}^0, \lambda_i^0, p^*) + a_1\log\sigma_Y^{*2}(\lambda_i^0, \lambda_{i+1}^0, p^*) + \cdots \\
&\quad + a_{k-1}\log\sigma_Y^{*2}(\lambda_{i+k-2}^0, \lambda_{i+k-1}^0, p^*) + a_k\log\sigma_Y^{*2}(\lambda_{i+k-1}^0, \lambda_{i+k}^0, p^*)\Big] \\
&=\ (\lambda_i^0 - \lambda_{j-1}^*)\log\sigma_Y^{*2}(\lambda_{i-1}^0, \lambda_i^0, p^*) + (\lambda_{i+1}^0 - \lambda_i^0)\log\sigma_Y^{*2}(\lambda_i^0, \lambda_{i+1}^0, p^*) + \cdots \\
&\quad + (\lambda_{i+k-1}^0 - \lambda_{i+k-2}^0)\log\sigma_Y^{*2}(\lambda_{i+k-2}^0, \lambda_{i+k-1}^0, p^*) \\
&\quad + (\lambda_j^* - \lambda_{i+k-1}^0)\log\sigma_Y^{*2}(\lambda_{i+k-1}^0, \lambda_{i+k}^0, p^*).
\end{aligned}$$

This implies that

$$\begin{aligned}
\lim_{n\to\infty}\frac{2}{n}\mathrm{MDL}_Y(m_0, \hat{\boldsymbol{\lambda}}; \boldsymbol{p}^*) &> \sum_{i=1}^{m_0+1}(\lambda_i^0 - \lambda_{i-1}^0)\log\sigma_Y^{*2}(\lambda_{i-1}^0, \lambda_i^0, p^*) \\
&= \lim_{n\to\infty}\frac{2}{n}\mathrm{MDL}_Y(m_0, \hat{\boldsymbol{\lambda}}; \boldsymbol{p}^*) \\
&\geq \lim_{n\to\infty}\frac{2}{n}\mathrm{MDL}_Y(m_0, \hat{\boldsymbol{\lambda}}; \boldsymbol{p}^*),
\end{aligned}$$

where the last inequality follows from the definition of $\hat{\boldsymbol{\lambda}}$. Since this is a contradiction, we conclude that $\hat{\boldsymbol{\lambda}} \to \boldsymbol{\lambda}^0$ a.s. □

3.3 Estimating the Number of Change-Points

As in the previous section, we will assume the underlying process $\{Y_t\}_{t=1}^n$ is a segmented stationary process such that the jth segment is modeled as

$$Y_t = X_{t-\tau_{j-1}+1,j}, \qquad \tau_{j-1} \leq t < \tau_j, \tag{3.4}$$

where $\tau_0 := 1$, $\tau_{m_0+1} := n+1$, and for each $j = 1, \ldots, m_0+1$, $\{X_{t,j}\}_{t=1}^{\infty}$ is a stationary process with mean $\mu_j := E(X_{t,j})$ and, for $h = 0, 1, \ldots$, autocovariance function $\gamma_j(h)$ $:= Cov(X_{t,j}, X_{t+h,j})$. Recall from Chapter 2 that Rio (1995) showed the functional law of the iterated logarithm holds for stationary strong mixing sequences under condition (2.4). Therefore, if we assume the underlying process

A1. is strong mixing at a geometric rate, and

A2. satisfies the moment condition (2.7) within each segment,

then we can again apply the functional law of the iterated logarithm on the sample covariances to prove consistency of the estimate of the number of change-points.

Many general processes satisfy these conditions, for example, linear processes with geometric coefficient decay and generalized autoregressive conditional heteroscedastic (GARCH) models under certain conditions. Examples of linear processes with geometric coefficient decay include all autoregressive moving average (ARMA) models. In particular, Athreya and Pantula (1986) showed that an ARMA(p, q) process $\{X_t\}$ where

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

is strong mixing if

(i) $E(\log^+ |\epsilon_1|) < \infty$,

(ii) the distribution of ϵ_1 has a nontrivial absolutely continuous component,

(iii) $\mathbf{X}_0 = (X_0, X_{-1}, \ldots, X_{1-p})$ is independent of $\{\epsilon_t\}$,

(iv) $\{\epsilon_t\}$ are independent and identically distributed random variables, and

(v) $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0$ for all $|z| \leq 1$.

Another interesting example is the GARCH (generalized autoregressive conditionally heteroscedastic) process, which plays a central role in modeling financial time series with volatility. For example, it is well known that the GARCH(1,1) process is geometrically strong mixing (see, e.g., Proposition 5 in [9]). We will demonstrate the application of Theorem 3.2 to a piecewise moving average process (ARMA(0,q)) and a piecewise GARCH process through simulations in the next section.

Lemma 2.1 stated that if we assume the process is AR(p) with no change-points, then the estimate of the number of change-points is strongly consistent. We can now relax the assumption that the process is AR(p) and generalize Lemma 2.1 as the following result.

Lemma 3.1. *Assume the true process $\{X_t\}$ follows the model given in (3.4) with no change-points ($m_0 = 0$) and satisfies assumptions A1 and A2. Then with probability 1, for any order $p = 0, 1, \ldots,$*

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

for n large, where

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

and

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

As before, $\hat{\sigma}^2$ is the conditional maximum likelihood estimate of the AR(p) noise variance over the entire dataset, and $\hat{\sigma}_k^2$ is the conditional maximum likelihood estimate of the AR(p) noise variance in the kth segment, $k = 1, \ldots, m+1$. The set A_m^ϵ is defined in (2.11).

Proof. The proof of Lemma 3.1 follows the same argument as the proof of Lemma 2.1. For simplicity of notation, assume the mean of the process is zero. The argument for a non-zero mean follows accordingly. Let $\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda} \in A_m^\epsilon}{\arg\min} \left\{ \frac{2}{n} \mathrm{MDL}(m, \boldsymbol{\lambda}; p, \ldots, p) \right\}$, and consider the quantity

$$\begin{aligned} & \frac{2}{n}\left[\mathrm{MDL}(m, \hat{\boldsymbol{\lambda}}; p, \ldots, p) - \mathrm{MDL}(0; p)\right] \\ &= \frac{2\log m}{n} + m(p+4)\frac{\log n}{n} + \frac{2m \log p}{n} \\ &\quad + \frac{p+2}{n} \sum_{k=1}^{m+1} \log(\hat{\lambda}_k - \hat{\lambda}_{k-1}) \\ &\quad + \sum_{k=1}^{m+1} (\hat{\lambda}_k - \hat{\lambda}_{k-1}) \log \hat{\sigma}_k^2 - \log \hat{\sigma}^2. \end{aligned} \tag{3.5}$$

Recall that we can write $\log \hat{\sigma}_k^2$ as a function of the sample covariances,

$$\log \hat{\sigma}_k^2 = g\left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=a}^{b} X_{t-i} X_{t-j} : i, j = 0, \ldots, p \right),$$

where

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

Let $\gamma(i) = E\left[X_t X_{t-i}\right]$ denote the true covariance between X_t and X_{t-i}, and let $\boldsymbol{\gamma} = (\gamma(|i-j|) : i, j = 0, \ldots, p)$ be the vector of covariances ranging over lags $0, \ldots, p$ defined in such a way to match the indices of the vectors of sample covariances,

$$\hat{\boldsymbol{\gamma}}_k := \left(\frac{1}{(\hat{\lambda}_k - \hat{\lambda}_{k-1})n} \sum_{t=a}^{b} X_{t-i} X_{t-j} : i, j = 0, \ldots, p \right),$$

$k = 1, \ldots, m+1$. Then since $g(\cdot)$ is a continuous function, by the strong law of large numbers, $g(\hat{\boldsymbol{\gamma}}_k)$ converges to $g(\boldsymbol{\gamma})$ with probability 1. The equation (2.20) follows, and the rest of the proof matches that of Lemma 2.1. □

Lemma 3.1 states that for any fixed autoregressive order p, if the true process has no change-points, then the estimated number of change-points will converge to

zero a.s. The next result generalizes Theorem 2.1 for a fixed fitted AR order. If we fix the fitted AR order such that covariances between true segments differ at some lag between zero and the fixed AR order, then the estimated number of change-points will converge to the true number of change-points in probability.

Theorem 3.2. *Assume the true process* $\{Y_t\}$ *follows the model given in (3.4) with* m_0 *change-points and satisfies assumptions A1 and A2. Choose* p^* *such that for each* $j = 2, \ldots, m_0 + 1$, *there exists an* $h \in \{0, \ldots, p^*\}$ *such that* $\gamma_j(h) \neq \gamma_{j-1}(h)$, *and fit AR*($p^*$) *models using the minimum description length. Then* $\hat{m} \xrightarrow{P} m_0$, *where* $\hat{m}$ *is the estimated number of change-points obtained by minimizing the MDL defined using conditional maximum likelihood white noise estimates with fixed AR order* p^*.

Let

$$\hat{\boldsymbol{\lambda}} = \arg \min_{\boldsymbol{\lambda} \in A^{\epsilon}_{m_0}} \left\{ \frac{2}{n} \mathrm{MDL}(m_0, \boldsymbol{\lambda}; \boldsymbol{p}^*_{m_0+1}) \right\},$$

and

$$\hat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha} \in A^{\epsilon}_{m_0+1}} \left\{ \frac{2}{n} \mathrm{MDL}(m_0 + 1, \boldsymbol{\alpha}; \boldsymbol{p}^*_{m_0+2}) \right\},$$

where $\boldsymbol{p}^*_{m_0+2} = p^* \mathbf{1}_{m_0+2}$, $\boldsymbol{p}^*_{m_0+1} = p^* \mathbf{1}_{m_0+1}$, $\mathbf{1}_m$ is an m x 1 vector of ones, and A^{ϵ}_m is defined in (2.11). Then, as in Theorem 2.1, Theorem 3.2 follows if

$$\lim_{n \to \infty} P\left(\mathrm{MDL}(m_0 + 1, \hat{\boldsymbol{\alpha}}; \boldsymbol{p}^*_{m_0+2}) - \mathrm{MDL}(m_0, \hat{\boldsymbol{\lambda}}; \boldsymbol{p}^*_{m_0+1}) > 0 \right) = 1.$$

As in the special case preceding the proof of Theorem 2.1, we will first examine the case where the process mean is zero and $m_0 = 1$ with true relative change-point location λ. Assume we fit AR(p^*) models to the data and compare two fitted models:

1. Fit AR(p^*) models to two segments with relative change-point location λ.

2. Fit AR(p^*) models to three segments with relative change-point location estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ obtained by minimizing $\mathrm{MDL}(2, \alpha_1, \alpha_2; p^*, p^*)$ with respect to $(\alpha_1, \alpha_2) \in A^{\epsilon}_2$, where A^{ϵ}_2 is defined in (2.11).

We first show that

$$\lim_{n\to\infty} P\big(\mathrm{MDL}(2,\hat{\alpha}_1,\hat{\alpha}_2;p^*,p^*) > \mathrm{MDL}(1,\lambda;p^*)\big) = 1.$$

where $\mathrm{MDL}(1,\lambda;p^*)$ is the MDL for the first fitted model, and $\mathrm{MDL}(2,\hat{\alpha}_1,\hat{\alpha}_2;p^*,p^*)$ is the minimized MDL for the second fitted model. Equivalently, we can show that

$$\lim_{n\to\infty} P\big(\mathrm{MDL}(2,\alpha_1,\alpha_2;p^*,p^*) > \mathrm{MDL}(1,\lambda;p^*) \\ \forall\, \alpha_1,\alpha_2 :\ \epsilon < \alpha_1 < \alpha_1+\epsilon < \alpha_2 < 1-\epsilon\big) = 1. \tag{3.6}$$

The main difference between the proof of Theorem 2.1 and the proof of Theorem 3.2 is in the treatment of the white noise variance estimates. In Theorem 2.1, we assumed that the underlying process followed a piecewise autoregressive model. We no longer make that assumption for the proof of Theorem 3.2, and thus must consider the behavior of the white noise variance estimates under less restrictive assumptions. Note that when the process is not piecewise autoregressive, these estimates are not necessarily estimating a white noise variance, but rather, are estimating the one-step prediction error in their respective fitted segments.

Assume, without loss of generality, that $\alpha_1 < \lambda < \alpha_2$, and consider the white noise variance estimate within the segment (α_1,α_2),

$$\hat{\sigma}^2_{2,2} = \frac{\sum_{t=[\alpha_1 n]}^{[\alpha_2 n]-1}\left(Y_t - \hat{\phi}_1 Y_{t-1} - \cdots - \hat{\phi}_{p^*} Y_{t-p^*}\right)^2}{(\alpha_2-\alpha_1)n},$$

where the autoregressive coefficient estimates $\hat{\phi}_1,\ldots,\hat{\phi}_{p^*}$ are calculated by minimizing the quantity

$$\sum_{t=[\alpha_1 n]}^{[\alpha_2 n]-1}\left(Y_t - a_1 Y_{t-1} - \cdots - a_{p^*} Y_{t-p^*}\right)^2,$$

with respect to $a_1,\ldots,a_{p^*}$. We can again break $\hat{\sigma}^2_{2,2}$ into two sums,

$$\hat{\sigma}^2_{2,2} = \frac{1}{(\alpha_2-\alpha_1)n}\left[\sum_{t=[\alpha_1 n]}^{[\lambda n]-1}\left(Y_t - \hat{\phi}_1 Y_{t-1} - \cdots - \hat{\phi}_{p^*} Y_{t-p^*}\right)^2 \right. \\ \left. + \sum_{t=[\lambda n]}^{[\alpha_2 n]-1}\left(Y_t - \hat{\phi}_1 Y_{t-1} - \cdots - \hat{\phi}_{p^*} Y_{t-p^*}\right)^2\right].$$

If α_1 and α_2 are not too close to λ, then,

$$\hat{\sigma}^2_{2,2} \geq \frac{1}{(\alpha_2-\alpha_1)n}\left[\sum_{t=[\alpha_1 n]}^{[\lambda n]-1}\left(Y_t-\hat{\phi}_{1,1}Y_{t-1}-\cdots-\hat{\phi}_{p^*,1}Y_{t-p^*}\right)^2 + \sum_{t=[\lambda n]}^{[\alpha_2 n]-1}\left(Y_t-\hat{\phi}_{1,2}Y_{t-1}-\cdots-\hat{\phi}_{p^*,2}Y_{t-p^*}\right)^2\right],$$

where $\hat{\phi}_{1,1},\ldots,\hat{\phi}_{p^*,1}$ is the minimizer of

$$\sum_{t=[\alpha_1 n]}^{[\lambda n]-1}\left(Y_t-a_1Y_{t-1}-\cdots-a_{p^*}Y_{t-p^*}\right)^2,$$

and $\hat{\phi}_{1,2},\ldots,\hat{\phi}_{p^*,2}$ is the minimizer of

$$\sum_{t=[\lambda n]}^{[\alpha_2 n]-1}\left(Y_t-a_1Y_{t-1}-\cdots-a_{p^*}Y_{t-p^*}\right)^2,$$

both with respect to $a_1,\ldots,a_{p^*}$. This inequality holds by definition of conditional maximum likelihood estimation and does not rely on any assumption about the true process. Define the residual sum of squares quantities as follows:

$$\begin{aligned}
\mathrm{RSS}_{2,1} &:= \sum_{t=[\alpha_1 n]}^{[\lambda n]-1}\left(Y_t-\hat{\phi}_1Y_{t-1}-\cdots-\hat{\phi}_{p^*}Y_{t-p^*}\right)^2\\
\mathrm{RSS}_{2,2} &:= \sum_{t=[\lambda n]}^{[\alpha_2 n]-1}\left(Y_t-\hat{\phi}_1Y_{t-1}-\cdots-\hat{\phi}_{p^*}Y_{t-p^*}\right)^2\\
\mathrm{RSS}^*_{2,1} &:= \sum_{t=[\alpha_1 n]}^{[\lambda n]-1}\left(Y_t-\hat{\phi}_{1,1}Y_{t-1}-\cdots-\hat{\phi}_{p^*,1}Y_{t-p^*}\right)^2\\
\mathrm{RSS}^*_{2,2} &:= \sum_{t=[\lambda n]}^{[\alpha_2 n]-1}\left(Y_t-\hat{\phi}_{1,2}Y_{t-1}-\cdots-\hat{\phi}_{p^*,2}Y_{t-p^*}\right)^2.
\end{aligned}$$

As in the special case before the proof of Theorem 2.1, without loss of generality, assume $\epsilon<\alpha_1<\lambda<\lambda+\epsilon/2<\alpha_2<1-\epsilon$, and consider the following three cases:

(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.

If (i) holds, we can use the above argument to break the term

$$\alpha_1 \log \hat{\sigma}^2_{1,2} + (\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} + (1 - \hat{\alpha}_2) \log \hat{\sigma}^2_{3,2} - \left[\lambda \log \hat{\sigma}^2_{1,1} + (1 - \lambda) \log \hat{\sigma}^2_{2,1}\right] \tag{3.7}$$

in the difference between the two minimum description lengths, $\frac{2}{n}[\mathrm{MDL}(2, \alpha_1, \alpha_2; p^*, p^*) - \mathrm{MDL}(1, \lambda; p^*)]$, into a sum over the intervals $(0, \alpha_1)$, (α_1, λ), (λ, α_2), and $(\alpha_2, 1)$ as in (2.27). Then, by the functional law of the iterated logarithm and Lemma 3.1, (3.7) is of order $\log\log n/n$.

If (ii) holds, the argument used in the proof of Theorem 2.1 when $\lambda - \alpha_1 < N/n$ for some positive integer N carries over directly to this proof since the argument does not use the assumption that the underlying process is autoregressive. Thus, (3.7) is again $O_p(\log\log n/n)$ in this case.

Now assume (iii) holds, $N/n < \lambda - \alpha_1 < (\log\log n)^2/\log n$ for some positive integer N. We know that

$$\begin{aligned}(\alpha_2 - \alpha_1) \log \hat{\sigma}^2_{2,2} \;>\; & (\lambda - \alpha_1) \log \left(\frac{\mathrm{RSS}_{2,1}}{(\lambda - \alpha_1)n} \right) \\ & + (\alpha_2 - \lambda) \log \left(\frac{\mathrm{RSS}_{2,2}}{(\alpha_2 - \lambda)n} \right)\end{aligned} \tag{3.8}$$

by concavity of the log function. In the proof of Theorem 2.1, we showed that with high probability, $\mathrm{RSS}_{2,1}/((\lambda - \alpha_1)n)$ is greater by a fixed constant $\eta > 0$ than the true variance in the segment $(0, \lambda)$. When we only assume stationarity, we can look at $\mathrm{RSS}_{2,1}$ as prediction error, as in the proof of Theorem 3.1, and show that with high probability, $\mathrm{RSS}_{2,1}/((\lambda - \alpha_1)n)$ is greater than the minimum prediction error for the process in the first true segment.

We will first define notation for the true one-step prediction coefficients in the kth true segment, $k = 1, 2$. Let the one-step predictor in terms of the past p^* observations for the kth true segment be denoted as

$$\hat{X}_{t,k} = \phi_{p^*1,k} X_{t-1,k} + \cdots + \phi_{p^*p^*,k} X_{t-p^*,k},$$

where X_t has mean zero and autocovariance function $\gamma_k(h)$. The one-step prediction coefficients are calculated by

$$\boldsymbol{\phi}_{p^*,k} = \Gamma_{k,p^*}^{-1}\boldsymbol{\gamma}_{k,p^*},$$

where $\Gamma_{k,p^*} := \{\gamma_k(|i-j|)\}_{i,j=1}^{p^*}$ and $\boldsymbol{\gamma}_{k,p^*} := (\gamma_k(1),\ldots,\gamma_k(p^*))^T$.

Intuitively, since $\alpha_1 \to \lambda$ as $n \to \infty$, the estimated AR coefficients used to calculate $\mathrm{RSS}_{2,1}$ will converge to the true one-step prediction coefficients for the segment $(\lambda, 1)$. This is because the coefficients in $\mathrm{RSS}_{2,1}$ are calculated within the interval (α_1, α_2) which converges to (λ, α_2). Let

$$c := E\left(X_{t,1} - \phi_{p^*1,2}X_{t-1,1} - \cdots - \phi_{p^*p^*,2}X_{t-p^*,1}\right)^2$$

Then $\mathrm{RSS}_{2,1}/((\lambda-\alpha_1)n)$ converges to c, where, assuming $\phi_{p^*i,1} \neq \phi_{p^*i,2}$ for at least one $i = 1,\ldots,p^*$,

$$c \;>\; E\left(X_{t,1} - \phi_{p^*1,1}X_{t-1,1} - \cdots - \phi_{p^*p^*,1}X_{t-p^*,1}\right)^2$$

since $\boldsymbol{\phi}_{p^*,1}$ minimizes the one-step prediction error for the process in segment $(0,\lambda)$. If we replace σ_1^2 by $E\left(X_{t,1} - \phi_{p^*1,1}X_{t-1,1} - \cdots - \phi_{p^*p^*,1}X_{t-p^*,1}\right)^2$ in (2.33), then (2.35) follows.

Proof of Theorem 3.2. The proof of Theorem 3.2 follows the proof of Theorem 2.1 directly by treating the residual sums of squares as one-step estimated prediction errors rather than estimated variances. If we replace $\sigma_{k(j)}^2$ by the true prediction error in the $k(j)$th true segment,

$$E\left(X_{t,k(j)} - \phi_{p^*1,k(j)}X_{t-1,k(j)} - \cdots - \phi_{p^*p^*,k(j)}X_{t-p^*,k(j)}\right)^2,$$

in case (iii), where $k(j)$ denotes the index of the true change-point contained in the jth fitted segment in the case where the jth fitted segment contains exactly one true change-point, then the rest of the proof follows accordingly. □

In Lemma 3.1 and Theorem 3.2, we used conditional maximum likelihood estimation in the definition of the MDL. However, the results again extend to Yule-Walker estimation as in the last section of Chapter 2.

3.4 Simulation Results

Davis et al. (2006) conducted five simulation experiments evaluating the practical performance of Auto-PARM. In this section, we add to these simulation experiments by conducting simulations when the true process is not piecewise autoregressive.

3.4.1 Piecewise Moving Average Process

For this experiment, we simulated 1000 observations from the model

$$Y_t = \begin{cases} Z_t + 0.9Z_{t-1} & \text{if } 1 \leq t \leq 400 \\ Z_t - 0.9Z_{t-1} & \text{if } 401 \leq t \leq 1000, \end{cases} \tag{3.9}$$

where $Z_t \sim \text{IID}N(0,1)$. A realization from this model is shown in Figure 3.1.

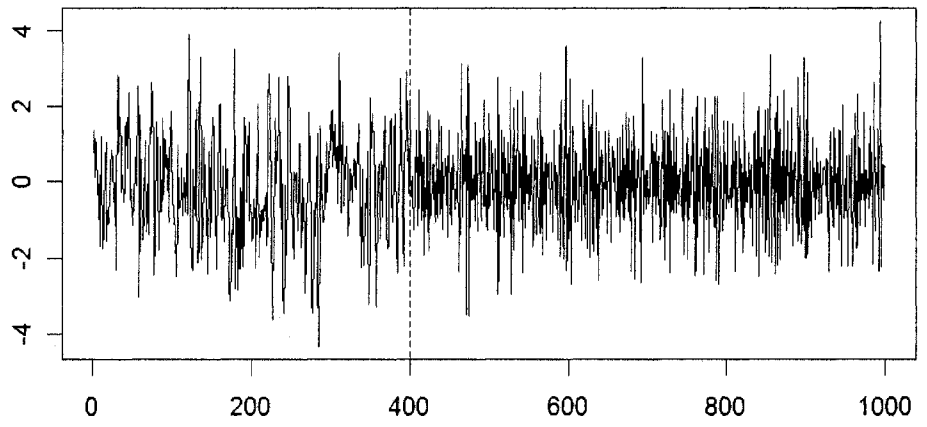

Figure 3.1: Realization from the process in (3.9).

The spectral density functions of the two moving average processes are shown in Figures 3.2 and 3.3, and the autocorrelation functions and partial autocorrelation functions are shown in Figures 3.4 and 3.5.

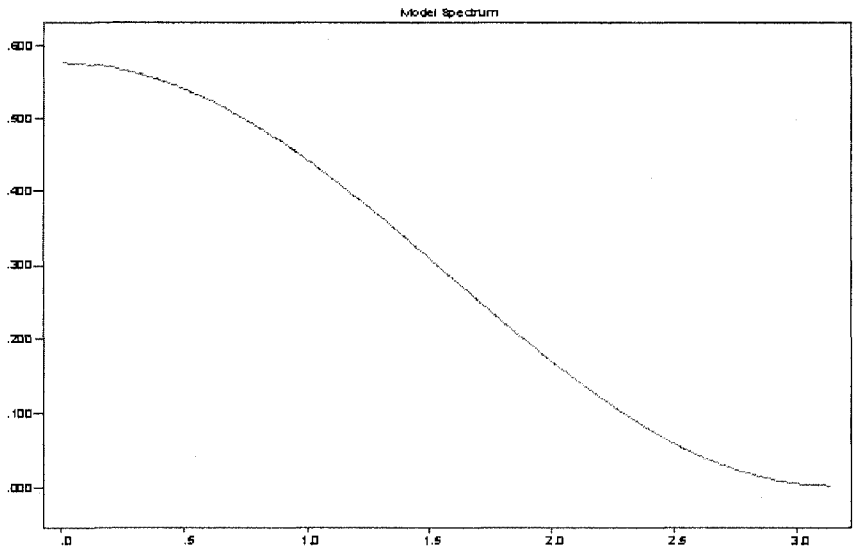

Figure 3.2: Spectral density function for first segment in (3.9).

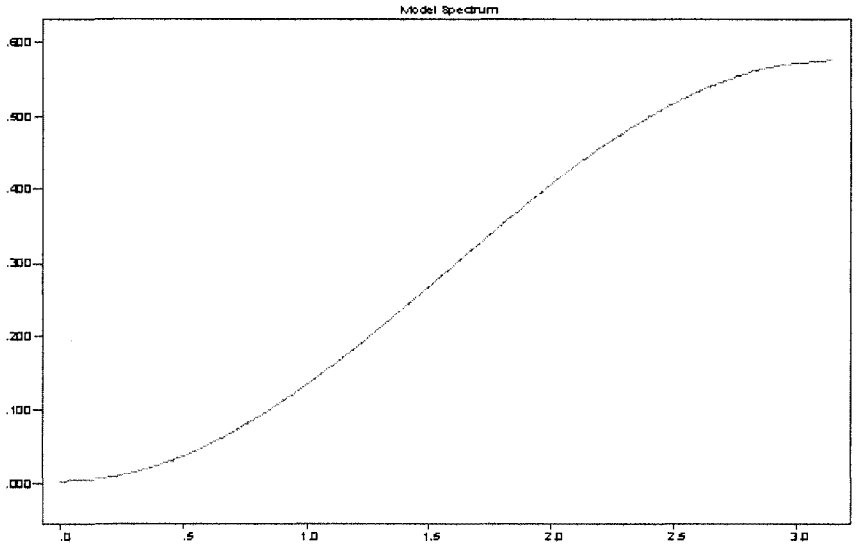

Figure 3.3: Spectral density function for second segment in (3.9).

We simulated 500 realizations of the process in (3.9) and applied Auto-PARM to each realization. All 500 applications of Auto-PARM detected two segments, with a mean change-point location of 401.0 and standard deviation 3.0. A histogram of the change-point location estimates is shown in Figure 3.6. The average estimated AR

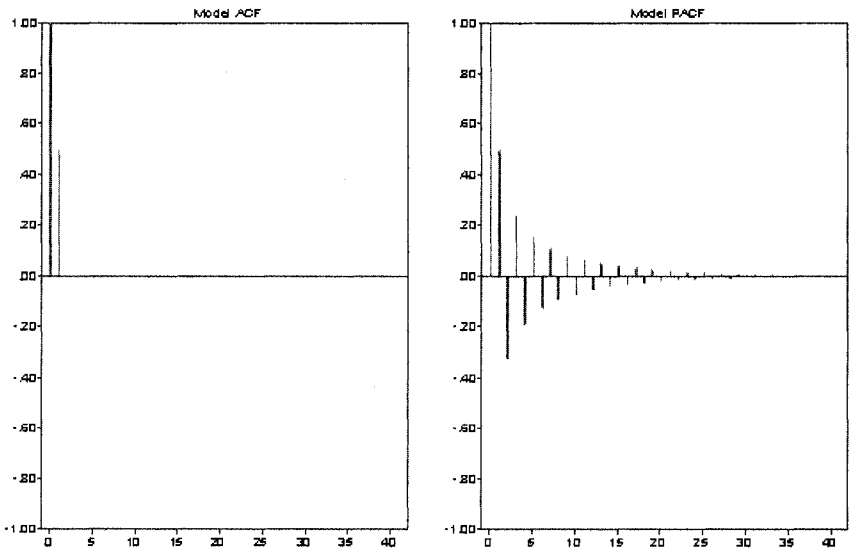

Figure 3.4: Autocorrelation and partial autocorrelation functions for first segment in (3.9).

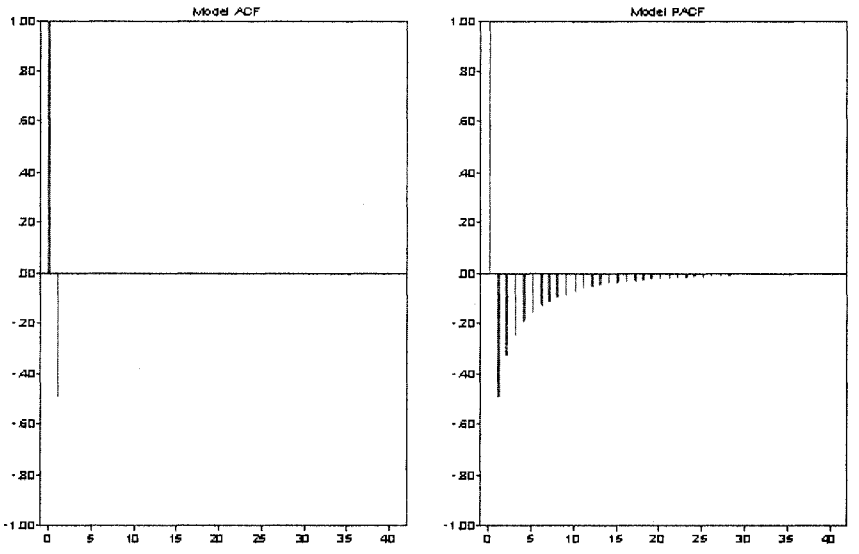

Figure 3.5: Autocorrelation and partial autocorrelation functions for second segment in (3.9).

order in the first segment was around 6, and around 7 in the second segment. Table 3.1 lists the relative frequencies of the AR order estimates.

We also applied Auto-PARM to 500 different realizations with the additional constraint that the fitted autoregressive order was 5. Again, every application of

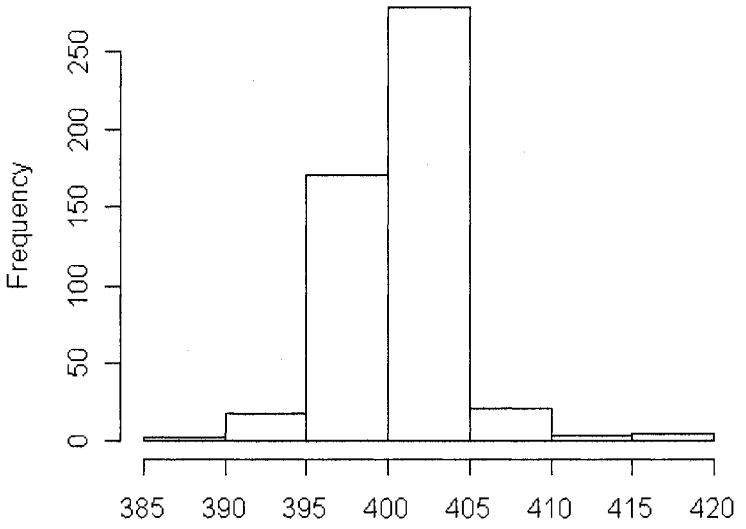

Figure 3.6: Change-point location estimates.

Table 3.1: Relative Frequencies of Auto-PARM AR order estimates.

Order	0	1	2	3	4	5	6	7	8	9	10	11	$\geq$12
p_1	0	0	0	0.8	13.8	21.8	29.2	19.6	9.4	3.0	1.6	0.6	0.2
p_2	0	0	0	0	1.0	10.0	19.6	27.4	20.2	11.2	6.8	2.0	1.8

Auto-PARM detected two segments. The average change-point location was 401.0 and standard deviation 2.9, which was very similar to the results when the AR order was fit from the data. Table 3.2 summarizes the AR parameter estimates obtained by Auto-PARM in the 500 realizations. Note that the true process is not autoregressive, but the autoregressive model provides a useful approximation to the true model. The change between segments is reflected in the first, third, and fifth AR coefficient estimates. The magnitudes of these coefficients are nearly the same between segments, but in the first segment, the coefficients are positive, whereas in the second segment, they are negative. For the simulation experiment where the AR order was fit to the data, in the case where the fitted AR orders were 5 for both segments, the average

AR parameter estimates are nearly the same as the average estimates for the fixed AR order simulation. The standard deviations of the parameter estimates are slightly smaller when the AR order is estimated from the data, but this is most likely due to the small number of realizations which resulted in an AR(5) to AR(5) fit. Only ten of the 500 realizations fit AR orders of 5 to both segments. A summary of these estimates is shown in Table 3.3.

Table 3.2: Summary of AR parameter estimates with AR order fixed at 5.

		Parameter					
Segment		ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	σ^2
1	Mean	0.81	-0.64	0.47	-0.32	0.15	1.06
	SD	0.05	0.06	0.06	0.06	0.05	0.08
2	Mean	-0.82	-0.65	-0.48	-0.32	-0.16	1.06
	SD	0.04	0.05	0.05	0.05	0.04	0.06

Table 3.3: Summary of AR parameter estimates with AR order fit to data where the estimated AR orders were both 5.

		Parameter					
Segment		ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	σ^2
1	Mean	0.84	-0.65	0.50	-0.35	0.17	1.04
	SD	0.03	0.05	0.06	0.04	0.03	0.06
2	Mean	-0.84	-0.67	-0.51	-0.33	-0.17	1.05
	SD	0.03	0.03	0.03	0.04	0.03	0.06

3.4.2 Piecewise GARCH Process

Generalized Autoregressive Conditional Heteroscedastic (GARCH) models are commonly used for analyzing financial time series. The GARCH model takes into account excess kurtosis (heavy tails) and changes in volatility, two common stylized facts about financial data. It can provide accurate forecasts of variances of asset returns through its ability to model time-varying conditional variances, and has wide applications in financial time series data, which include risk management, portfolio

management and asset allocation, option pricing, foreign exchange, and the term structure of interest rates [34].

We say $\{Z_t\}$ follows a GARCH(p, q) model if it is a causal strictly and weakly stationary solution of

$$\begin{aligned} Z_t &= \sqrt{h_t} e_t, \qquad \{e_t\} \sim \text{IID}(0,1), \\ h_t &= \alpha_0 + \sum_{i=1}^{p} \alpha_i Z_{t-i}^2 + \sum_{i=1}^{q} \beta_i h_{t-i}, \end{aligned} \tag{3.10}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$, and $\beta_i \geq 0$ for each i [7]. The equations (3.10) have a causal weakly stationary solution if and only if

$$\sum_{i=1}^{p} \alpha_i + \sum_{i=1}^{q} \beta_i < 1,$$

in which case there is exactly one such solution. The random variable h_t is the conditional variance of Z_t given Z_s, $s < t$. The GARCH(p, q) model is a generalization of the ARCH(p) model [15], where the ARCH(p) model is equivalent to a GARCH(p, q) model with $q = 0$.

Though there is an obvious dependence structure in the variables $\{Z_t\}$ of a GARCH(p, q) process, the sequence is uncorrelated, that is, $E(Z_s Z_t) = 0$ for $s \neq t$. If we apply Auto-PARM to a zero-mean segmented GARCH process, one may ask, "what is changing?" The key idea when using Auto-PARM in this case is the variance of the process in each segment, i.e., $\gamma_k(0)$, is changing between consecutive segments. Therefore, in principle, one could take $p^* = 0$ when applying Theorem (3.2) to a segmented GARCH process. However, in the simulation that follows, we allowed Auto-PARM to fit AR orders to demonstrate how Auto-PARM detects the lack of autocorrelation in the process. The majority of estimated AR orders were zero, indicating that the variables in each segment are uncorrelated.

If we assume that the variance changes between segments, then we can apply Theorem 3.2 to estimate changes in a segmented GARCH process. We must also

assume that the processes in each segment are strong mixing with a geometric rate, have finite fourth moments and satisfy (2.7). For the simulated process that follows, this condition can be verified using Proposition 2 in Lindner (2008) .

In this simulation experiment, we simulated 500 realizations from the piecewise GARCH process $Z_t = \sqrt{h_t}e_t$, where

$$h_t = \begin{cases} 1.20 + 0.06Z_{t-1} + 0.84h_{t-1} & \text{if } 1 \leq t \leq 300 \\ 1.20 + 0.06Z_{t-1} + 0.13Z_{t-2} & \\ \quad +0.33h_{t-1} + 0.18h_{t-2} & \text{if } 301 \leq t \leq 800 \\ 0.50 + 0.06Z_{t-1} + 0.13Z_{t-2} & \\ \quad +0.33h_{t-1} + 0.18h_{t-2} & \text{if } 801 \leq t \leq 900, \end{cases} \tag{3.11}$$

where $\{e_t\} \sim \text{IID}N(0,1)$. The theoretical variances in each segment are 12, 4, and 1.67, respectively. We applied Auto-PARM to each realization. A realization from this model is shown in Figure 3.7.

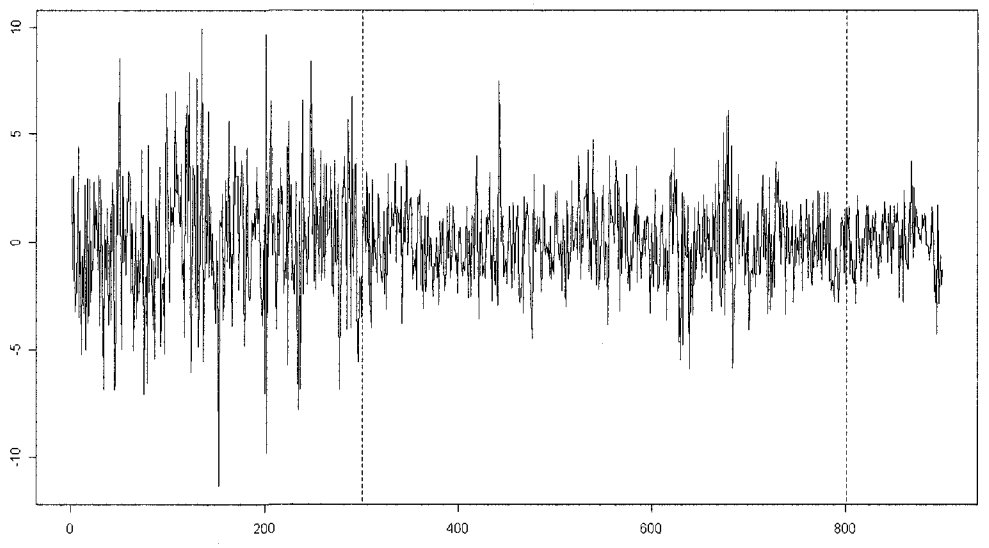

Figure 3.7: Realization from the process in (3.11).

Auto-PARM detected two segments in 34.8% of the 500 realizations, and three segments in 61.2% (see Table 3.4). The mean change-point location estimates for the realizations where Auto-PARM detected three segments are 299 and 782 with

standard deviations 33.1 and 80.6, respectively. The mean change-point location estimates are very close to the true change-point locations of 301 and 801. When Auto-PARM detected only two segments, the mean change-point location estimate was 308 with a standard deviation of 31.3. This indicates that Auto-PARM only detected the first true change-point in these realizations. Thus, Auto-PARM detected the first true change-point in 96% of the realizations.

Table 3.4: Summary of Estimated Change-points for the Process (3.11).

Number of Segments	Percent	Location Mean	Location Standard Deviation
2	34.8	308.4	31.3
3	61.2	298.5	33.1
		782.3	80.6
4	3.6		
5	0.4		

Since the observations from a GARCH model are uncorrelated over time, we would expect that the best autoregressive model fit to the data would have order zero. That is, the fitted model would be an estimated mean plus a white noise term with mean zero and an estimated variance. Table 3.5 shows the relative frequencies of the estimated autoregressive orders in the 500 realizations when Auto-PARM fit three segments. The majority of the realizations fit AR(0) models to each segment.

Table 3.5: Relative Frequencies of Estimated AR Orders for the Process (3.11).

Order		0	1	2	3
Segment	1	97.7	2.0	0.3	0
	2	93.8	4.9	0.7	0.7
	3	81.4	15.7	2.9	0

From these results, we can see that Auto-PARM performs well even on processes that are not autoregressive. Even though GARCH models have a higher order dependence structure than second moments, Auto-PARM can detect changes in the

process by fitting AR(0) models to the process. These results show that Auto-PARM has applications in financial data and other time series data which do not follow autoregressive models.

Chapter 4

APPLYING AUTO-PARM TO NATIONAL PARK SERVICE DATA

4.1 Introduction

Due to a growing concern about manmade noise in the National Parks, the National Park Service has been collecting natural sound data in about 20 of the 388 National Parks. At several sites throughout a park, tripods with microphones attached are used to record the surrounding sounds. The purpose of collecting this data is to measure and monitor noise pollution. In trying to estimate the proportion of sound that is unnatural (manmade), several challenges emerge. First, due to the large amount of recordings, the data cannot be analyzed without sampling or automatic procedures. Second, when listening to the recordings, humans tend to misclassify sounds fairly often. One way to address these challenges is to develop an automatic algorithm that can take a large amount of audio data, partition it into homogeneous sound segments, and classify each segment as a known type of sound with a low misclassification rate.

In general, this problem can be described as follows. Suppose we have a training data set of realizations from a finite collection of known stochastic processes $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K\}$. We observe a time series $X = \{x_1, \ldots, x_n\}$ which is a concatenation of realizations from these processes. Thus, the interval $[1, n]$ may be partitioned into subintervals $[\tau_0 := 1, \tau_1)$, $[\tau_1, \tau_2)$, $\ldots$, $[\tau_m, \tau_{m+1} := n + 1)$ such that each subset $\{x_t : t \in [\tau_{i-1}, \tau_i)\}$ is a realization from process $\mathcal{P}_{k_i}$ with $k_{i-1} \neq k_i$ for all $i = 1, \ldots, m + 1$, i.e., neighboring segments of X come from different processes. The observed data set X in our case is a 1-dimensional audio time series. The observations

x_t are amplitudes, and the training data set of known stochastic processes is a set of realizations from different sounds, such as a clap of thunder, the call of a squirrel, or the drone of snowmobiles. The goal is to use the training data set to estimate the unknown number of segments, $m+1$, the change-points, $\tau_1, \ldots, \tau_m$, and the process types $k_1, \ldots, k_{m+1}$ in the observed series X. Estimating m and $\tau_1, \ldots, \tau_m$ is referred to as a segmentation problem, and once we have estimates of m and $\tau_1, \ldots, \tau_m$, the estimation of $k_1, \ldots, k_{m+1}$ can be viewed as a classification problem.

Auto-PARM provides a straight-forward method of segmenting the sound wave into approximately stationary pieces. Once the sound wave is segmented, we can apply a classification algorithm to categorize each piece as a specific sound type. If consecutive pieces are of the same sound type, they are merged to form one piece.

4.2 Methods and Results

4.2.1 Data Preprocessing

The data consist of 15 recordings of separate sounds commonly heard in the parks and two 1-hour recordings of the surrounding environment in Yellowstone National Park. We will refer to the recordings of 15 common sounds as the index data set and the two 1-hour recordings as the real data. The 15 types of sounds in the index set are elk bugling, coyotes, people talking, H-D motorcycle, snow groomer, rotary snowplow, 2 stroke snowmobiles, jet, propeller plane, helicopter, LE ranger siren, red squirrel, thunder, raven, and mud pots/thermal activity. Originally, we planned to use the index set as the training data set and the real data as the observed data to be segmented and classified. However, since many of the sounds in the real data were not in the index set, we treated the index set and the real data separately, as if each data set was an observed data set, but one in which I knew the sound types and break points. Thus, each data set served as both a training data set and an observed data set.

Before analyzing the data, the data needed to be normalized so that sounds could be compared without volume influencing the classification process. The most common way to normalize sound data is to use peak normalization. Peak normalization multiplies the sound wave by a positive constant so that the maximum absolute value of the amplitudes becomes 1. This type of normalization may not be appropriate in this case because large outliers can strongly affect the normalization. For example, during a thunder storm, the loud cracks of thunder will have the most influence rather than the longer rumbling sounds afterwards. Thus, we decided to use root mean square (RMS) normalization rather than peak normalization. RMS normalization differs from peak normalization by using the average RMS power (average squared amplitude) as its reference rather than the maximum absolute value of the amplitudes. This produces an overall change in loudness to a specified decibel (dB) level. Due to the fact that the reference is an average of the loudness, outliers have less of an effect than in peak normalization.

For each index sound, we applied RMS normalization in Adobe®Audition®1.5 to each sound type by choosing the smallest dB level such that overclipping was 0%. The decibel levels chosen for each sound are shown in Table 4.1. The raven sound was not included in the preliminary analysis since it was difficult to isolate the part of the sound wave where the raven call was present.

For the real data, we did not analyze the entire hour at once since the data set was so large. Instead, we took a small section of the data, classified the sounds present by listening to the data and noting where the sounds changed, and used this section for analysis. The sound categories for the real data are jet, jet and snowmobile, people (man), people (woman), raven, snowmobile, cross-country skiers, cross-country skiers and jet, background noise, other bird, and other. Background noise is when no sound is present except for wind or water which is present at all times. Other bird is a bird other than a raven. Other is a sound that didn't fall into any of the other

Table 4.1: RMS Normalization Levels for Index Set

Sound Type	Decibel Level (dB)
Coyotes	-15
Elk	-16
H-D Motorcycle	-10
Helicopter	-11
Jet	-11
Mudpots	-10
People	-5
Propeller Plane	-8
Siren	-3
Snowgroomer	-8
Snowmobile	-7
Snowplow	-16
Squirrel	-13
Thunder	-8

categories. RMS normalization was applied to the entire section of data that was under consideration.

Some unresolved issues in data preprocessing need to be considered. RMS normalization requires you to specify the size of each window used in calculating the RMS value. The default in Adobe®Audition®1.5 is 50 ms. Also, we normalized each index set separately, but one may want to normalize the sound waves simultaneously, such as in the real data. In addition, sometimes volume of a sound wave may be helpful in classification, such as a siren getting louder as it gets closer. These are topics for further study.

4.2.2 Segmentation and Classification Using Auto-PARM and Spectral Densities

The first approach to the problem used the program Auto-PARM to segment the data, then used spectral density estimates to classify each segment. Auto-PARM estimated change-points and fit an AR model to each segment. The fitted spectral densities of each segment were used for classification. We plotted the log spectral density for each break using the function `spec.ar` in the statistical software package

R. Since there were multiple segments from each sound, comparisons of plots could be made both within each sound type and between different sounds.

We first tried this method on the index data set. For each recording, we applied Auto-PARM to detect change-points and to estimate AR parameters. We calculated the fitted spectral density for each segment. Looking at the plots, it seemed as though there was a lot of variability in spectral densities within each sound indicating that it may not be a good classifier. A sample of the log spectral density plots are shown in figures 4.1−4.12. It is apparent that for some sounds such as coyote, elk, squirrel and thunder, the variability in spectral density fits is very large. Other sounds such as motorcycle, helicopter, siren, and snow groomer have low variability between spectral density plots. However, many of these sounds have spectral densities that are similar to other sounds' spectral densities. For example, the spectral densities of motorcycle, helicopter, jet and propeller plane are all very similar.

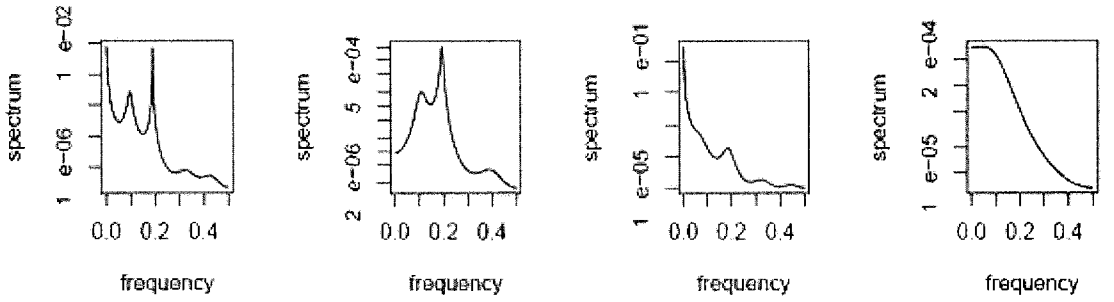

Figure 4.1: Coyote.

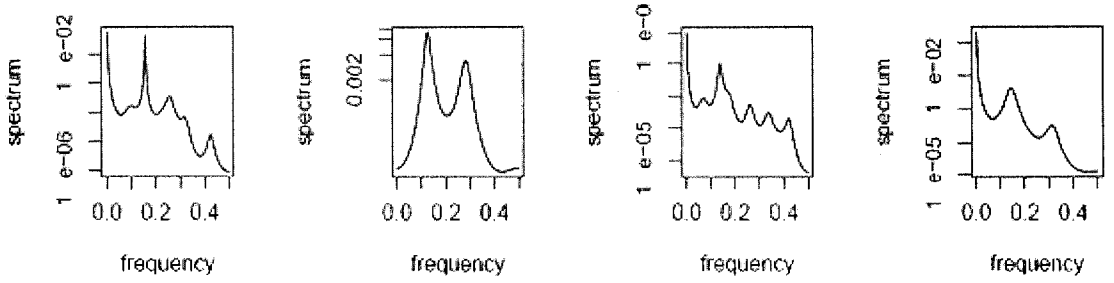

Figure 4.2: Elk.

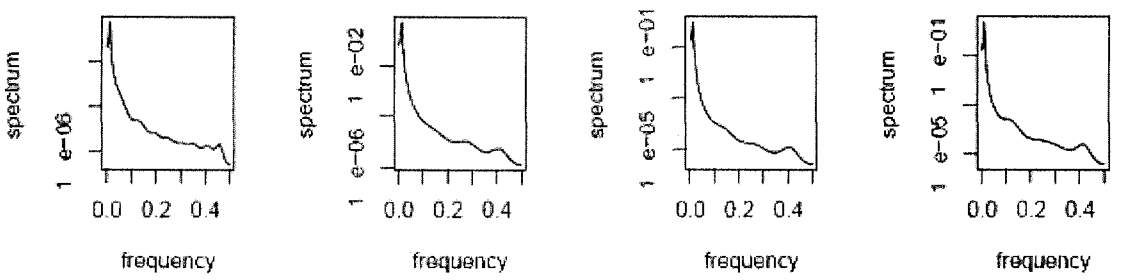

Figure 4.3: H-D Motorcycle.

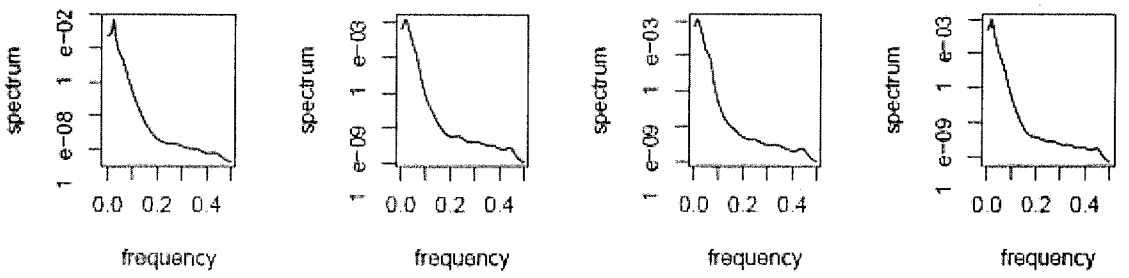

Figure 4.4: Helicopter.

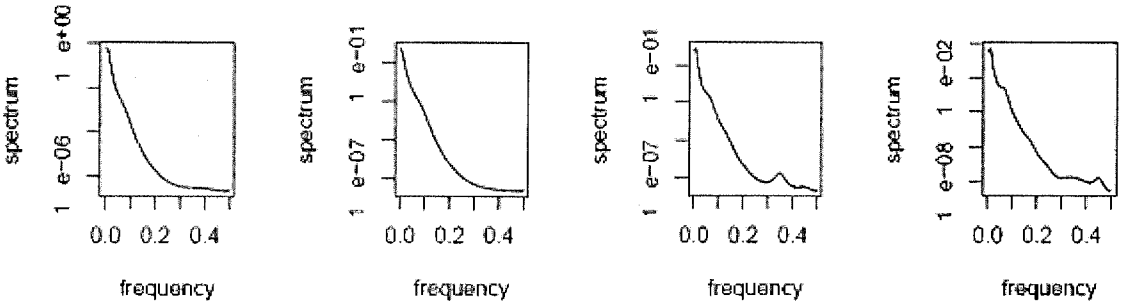

Figure 4.5: Jet.

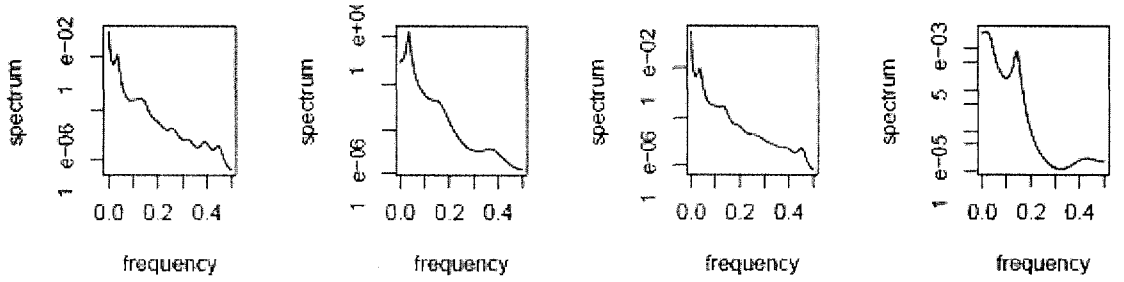

Figure 4.6: Mudpots.

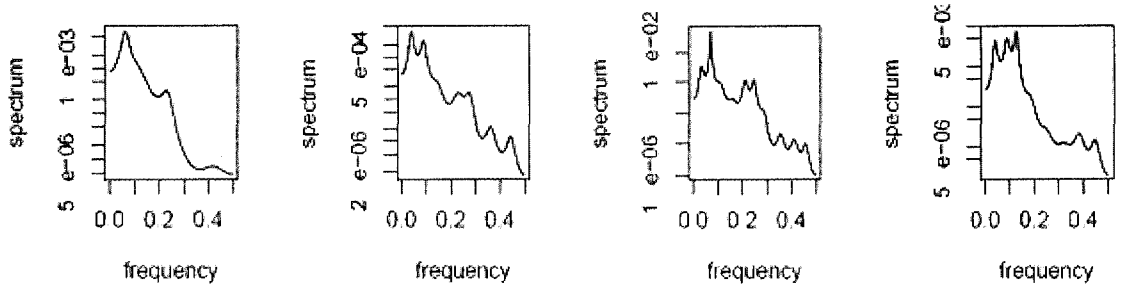

Figure 4.7: People.

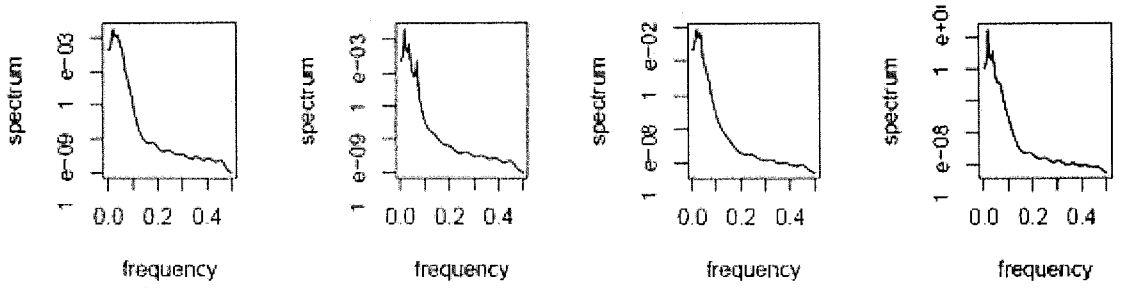

Figure 4.8: Propeller Plane.

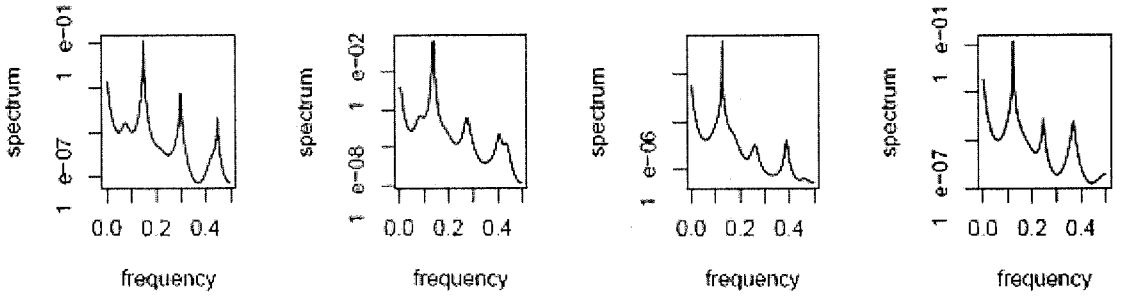

Figure 4.9: Siren.

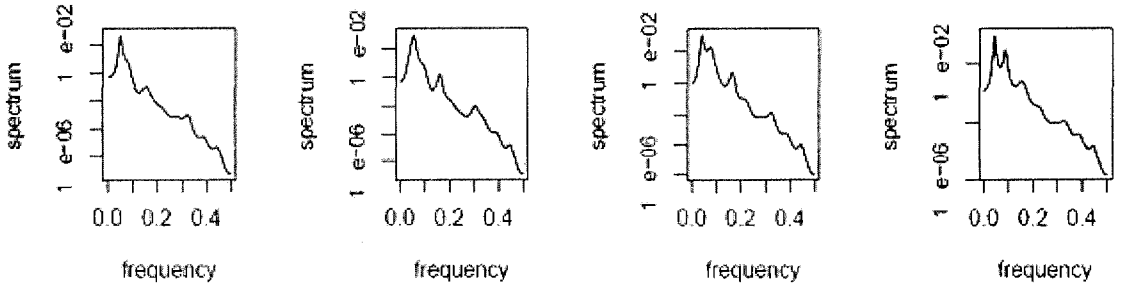

Figure 4.10: Snow Groomer.

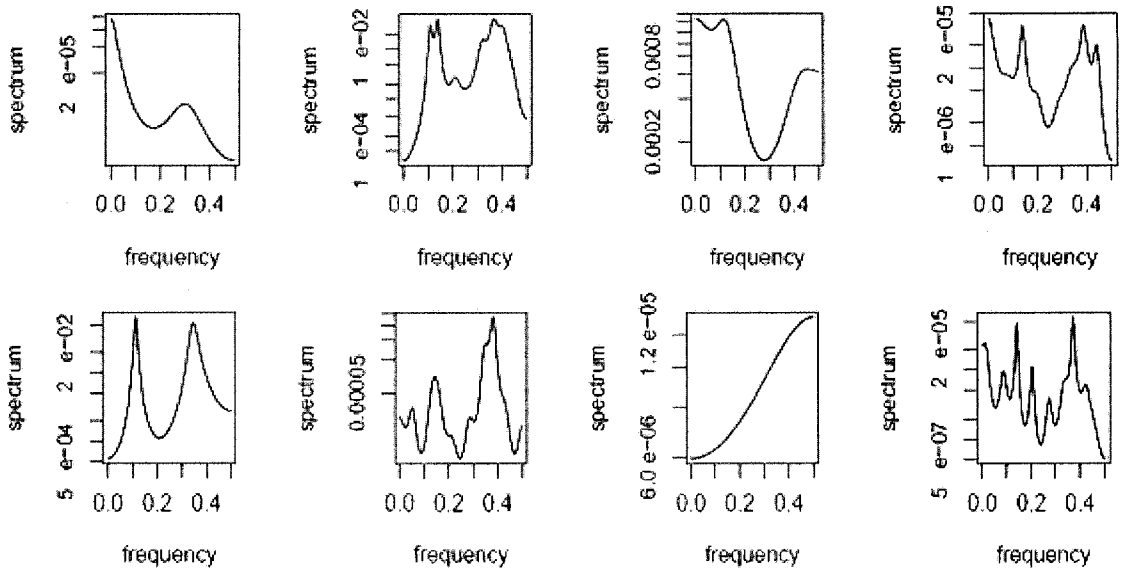

Figure 4.11: Squirrel.

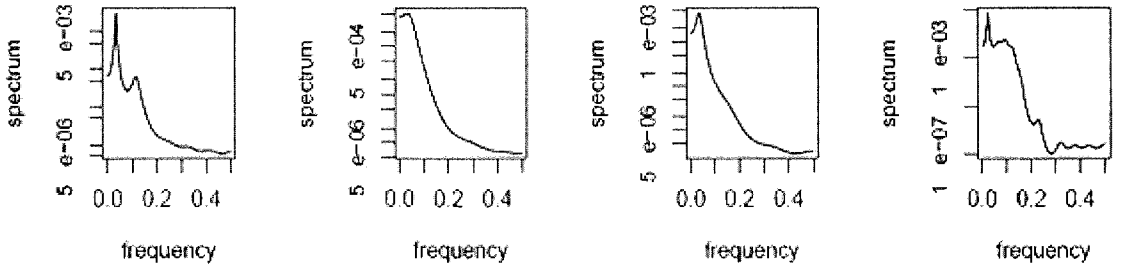

Figure 4.12: Thunder.

In order to quantitatively compare two spectral densities, we used the area under the squared difference between the two densities as a measure of dissimilarity. In other words, given two spectral densities $f(\lambda)$ and $g(\lambda)$, we calculated the area under the curve $(f(\lambda)-g(\lambda))^2$. For each segment, the `spec.ar` function in R outputs (frequency, density) pairs rather than an explicit function. The default number of pairs is 500, and the range of frequencies is between 0 and 0.5. Thus, in order to estimate the area under the squared difference function, we used Simpson's rule on these pairs.

This dissimilarity measure was used to classify sounds in the following manner. Auto-PARM was applied to each sound in the index data set. This broke each sound

into multiple segments and fit autoregressive models to each segment. A randomly chosen segment from the entire index data set was used as the sound we planned to classify. We calculated the dissimilarity in the spectral densities between this randomly chosen sound segment and every other segment in the index set. For each type of sound, we took the minimum dissimilarity as a measure of how close the new segment was to that type of sound. Then we took the minimum over all sounds, and classified the new segment as the type of sound with the smallest dissimilarity. This measure of dissimilarity was not successful in correctly classifying the sounds in the index set. Error rates were very high. Since it did not work on the index set, we did not try it on the real data.

4.2.3 Feature Extraction and Linear Discriminant Analysis

The second approach to this problem eliminated the segmentation procedure and instead windowed the data. Using the statistical learning paradigm, we treated each window as one observation and concentrated on variables (features) measured on each window that could be used in a classification algorithm. The spectral density was no longer used as a feature. Thus, we did not need to assume that within each window, the sound wave could be modeled as an autoregressive process. With this approach, change-points can be estimated using the classification results for each window by combining neighboring windows if a specified majority of the windows are classified as the same sound.

Speech recognition literature suggests several features that have the potential to successfully discriminate between different sounds. The most successful feature in the current literature is the set of mel-frequency cepstral coefficients, and these coefficients were used as features in our analysis. See Appendix I for further description of mel-frequency cepstral coefficients.

The sound files were read into R and a vector of sound type labels indicating the sound type at each sample time was created. We used Matlab to calculate the first 13

mel-frequency cepstral coefficients for each window using the `mfcc.m` function from Malcolm Slaney's Auditory Toolbox [43]. The sound type for each window was taken to be the sound type of the first sample point in the window. A random sample of windows was taken to be the training set for linear discriminant analysis. The other windows were used as a test set. The `discrim` procedure in SAS was used to run the linear discriminant analysis, treating windows as observations, sound type as the class type, and the 13 mel-frequency cepstral coefficients as 13 explanatory variables.

For the index set, this method worked well. We used the default window size of 256 samples with an overlap of 80 samples between consecutive windows. The proportions misclassified in the training and test data sets are shown in Table 4.2. The rates for the training data set were obtained by cross-validation. None of the squirrel windows were misclassified, whereas with the first approach, the fitted spectral densities for the squirrel sound varied greatly. The sounds with the highest misclassification rate were the H-D motorcycle, helicopter, propeller plane, and siren.

Table 4.2: Misclassification Rates for Index Data

Sound Type	Training Data	Test Data
Coyotes	0.1214	0.1438
Elk	0.1288	0.0993
H-D Motorcycle	0.1813	0.1700
Helicopter	0.1847	0.2011
Jet	0.0266	0.0225
Mudpots	0.0485	0.0289
People	0.0135	0.0083
Propeller Plane	0.1852	0.1806
Siren	0.1503	0.1959
Snowgroomer	0.0184	0.0120
Snowplow	0.0425	0.0563
Squirrel	0.0000	0.0000
Thunder	0.1711	0.1712
Total	0.0979	0.0992

The pairwise generalized squared distances between sound types are shown in Table 4.3.

Table 4.3: Generalized Squared Distance between Sound Types for Index Data

	Coyote	Elk	Groomer	Heli	Jet	Motor	Mudpots	People	Plow	Prop	Siren	Squirrel	Thunder
Coyote	0	18.9	71.2	101.6	89.1	24.8	39.1	31.8	84.0	102.5	12.5	115.1	54.3
Elk	18.9	0	59.8	144.7	123.6	32.7	36.1	25.9	87.0	139.4	26.7	74.7	83.8
Groomer	71.2	59.8	0	81.2	61.9	45.5	20.6	19.7	7.7	70.2	94.6	160.7	49.8
Heli	101.6	144.7	81.2	0	20.0	72.0	76.4	89.3	54.1	2.9	147.1	332.7	33.2
Jet	89.1	123.6	61.9	20.0	0	74.8	67.6	67.2	38.9	24.4	107.4	302.5	7.2
Motor	24.8	32.7	45.5	72.0	74.8	0	9.8	21.0	55.8	66.6	57.2	118.8	56.4
Mudpots	39.1	36.1	20.6	76.4	67.6	9.8	0	11.3	31.3	66.3	66.8	124.5	49.4
People	31.8	25.9	19.7	89.3	67.2	21.0	11.3	0	31.2	83.1	52.8	106.7	43.9
Plow	84.0	87.0	7.7	54.1	38.9	55.8	31.3	31.2	0	45.9	112.2	224.0	36.9
Prop	102.5	139.4	70.2	2.9	24.4	66.6	66.7	83.1	45.9	0	151.2	324.8	37.8
Siren	12.5	26.7	94.6	147.1	107.4	57.2	66.8	52.8	112.2	151.2	0	124.4	65.3
Squirrel	115.1	74.7	160.7	332.7	302.5	118.8	124.5	106.7	224.0	324.8	124.4	0	240.2
Thunder	54.3	83.8	49.8	33.2	7.2	56.4	49.4	43.9	36.9	37.8	65.3	240.2	0

In terms of these distances, jet and thunder are similar, plow and groomer are similar, and propeller plane and helicopter are very similar. These close distances account for the large misclassification error rates in these sound types.

The next step was to try this method on the real data. We tried windowing the data both with and without overlap between consecutive windows. The misclassification rates using a window size of 256 and an overlap of 80 samples are shown in Table 4.4.

Table 4.4: Misclassification Rates for Real Data With Overlap

Sound Type	Training Data	Test Data
Background	0.4777	0.4776
Jet	0.1696	0.2053
Jet + Snowmobile	0.1538	0.1495
Other	0.5000	0.6000
Other Bird	0.1975	0.2423
People (Man)	0.6250	0.6310
People (Woman)	0.5821	0.5742
Raven	0.6000	0.7576
Snowmobile	0.1374	0.1064
Xcountry Skis	0.5719	0.5720
Xcountry Skis + Jet	0.2391	0.2481
Total	0.3867	0.4149

These rates are much higher than those in the index data set, but the rates for the training set are fairly similar for the test set, indicating that the model was not overfit. The highest misclassification rates were for background noise, people, raven, cross-country skis, and other.

Pairwise generalized squared distances between sound types in the real data are shown in Table 4.5. Misclassification rates using the same window size but without overlap are shown in Table 4.6. The test data misclassification rates using no overlap are fairly similar to those using an overlap of 80 samples. For the training data, some of the misclassification rates increased immensely when the windows did not overlap.

Table 4.5: Generalized Squared Distance between Sound Types for Real Data

	Bkgrnd	Jet	Jet+Snwmble	Other	Other Bird	Man	Woman	Raven	Snwmble	Skis	Skis+Jet
Bkgrnd	0	9.7	50.8	5.5	6.1	4.1	8.9	3.2	9.9	1.4	25.7
Jet	9.7	0	20.9	7.3	15.0	9.3	7.9	13.3	16.0	5.9	8.3
Jet+Snwmble	50.8	20.9	0	40.3	55.9	47.6	41.2	57.5	45.8	38.4	7.9
Other	5.5	7.3	40.3	0	15.3	2.9	3.5	8.7	17.3	5.0	19.8
Other Bird	6.1	15.0	55.9	15.3	0	10.3	18.4	9.8	15.9	6.4	30.0
Man	4.1	9.3	47.6	2.9	10.3	0	3.8	7.6	17.7	4.3	23.5
Woman	8.9	7.9	41.2	3.5	18.4	3.8	0	10.3	21.8	8.1	20.0
Raven	3.2	13.3	57.5	8.7	9.8	7.6	10.3	0	15.0	5.2	30.4
Snwmble	9.9	16.0	45.8	17.3	15.9	17.7	21.8	15.0	0	11.4	31.2
Skis	1.4	5.8	38.4	5.0	6.4	4.3	8.1	5.2	11.4	0	16.5
Skis+Jet	25.7	8.3	7.9	19.8	30.0	23.5	20.0	30.4	31.2	16.5	0

Table 4.6: Misclassification Rates for Real Data Without Overlap

Sound Type	Training Data	Test Data
Background	0.5057	0.5068
Jet	0.1164	0.1532
Jet + Snowmobile	0.1623	0.1642
Other	1.0000	0.6250
Other Bird	0.2162	0.2143
People (Man)	0.9375	0.5294
People (Woman)	0.5000	0.6349
Raven	1.0000	0.7778
Snowmobile	0.1087	0.1283
Xcountry Skis	0.6547	0.6151
Xcountry Skis + Jet	0.2290	0.2380
Total	0.4937	0.4170

When using crossvalidation to determine error rates in the training data, both raven and other had 100% error, i.e., none of the windows were classified correctly. However, there were only two raven sound windows in the data set and 7 other sound windows. In fact, using resubstitution on the training data set gave 0% misclassification error for raven; both raven windows were correctly classified as raven. Resubstitution gave an error rate of 71.43% for other. Therefore, misclassification rates for these sound types should not be trusted in this case due to the small sample size for these sounds. The pairwise generalized distances between sound types when there was no overlap between windows were very similar to those when there was an overlap of 80 samples and thus are not shown here.

4.3 Future Directions

The second approach, using mel-cepstral frequency coefficients and linear discriminant analysis, produced encouraging results and has the potential to be a successful method to solving this problem. However, much research must be done with other data sets to determine if the approach is applicable to a wide range of sound data. In addition, it may be beneficial to investigate more features in addition to the

mel-cepstral coefficients, and to try a different classifier, e.g., support vector machines, neural networks, or tree-based methods.

There are other approaches to the problem that are worth investigating. One such approach involves hidden Markov models (HMMs). Bayesian approaches using hidden Markov models and Markov chain Monte Carlo methods have recently been shown to work fairly well on change-point problems. In future work, it will be of interest to consider approaching this problem using these methods.

Chapter 5

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation explores the asymptotic properties of the Auto-PARM estimates, in particular, consistency. We show that not only the estimated change-point locations, but also the estimated number of change-points and estimated autoregressive orders are consistent for the true values when the underlying model is piecewise autoregressive. When the underlying model is not piecewise autoregressive, but is piecewise stationary and strong mixing plus satisfies some other small assumptions, the estimated number and locations of change-points are still consistent for the true values. These results demonstrate the advantages of Auto-PARM and provide a foundation for the theory behind the method.

Auto-PARM can be applied to many different problems in a variety of fields. One such application is the segmentation of sound data for the National Park Service. This is not a standard application of a change-point problem, but instead combines segmentation and classification. Further research is needed to explore classification methods, including those from statistical learning theory. Other applications include problems in seismology, psychology, economics, finance, and ecology.

The consistency results proven in this dissertation provide a large piece of the asymptotic theory for Auto-PARM. We also plan to examine the asymptotics of the estimated autoregressive orders in Auto-PARM when the underlying process is not autoregressive. Future research will study the connection between the theory behind Auto-PARM using the MDL principle and Bayesian methods. Zhang and Siegmund (2007) develop a "modified BIC" model selection criterion for independent normal

data with a change in the mean, which estimates the Bayes factor. Their criterion appears to be very similar to the MDL criterion. We will explore this connection and attempt to derive the MDL criterion for Auto-PARM from a Bayesian perspective. Sample size calculations and confidence intervals for the Auto-PARM estimates are also of interest.

References

[1] H.-Z. An, Z.-G. Chen, and E. J. Hannan. Autocorrelation, autoregression and autoregressive approximation. *The Annals of Statistics*, 10:926–936, 1982.

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

[3] P. K. Bhattacharya. Maximum likelihood estimation of a change-point in the distribution of independent random variables: general multiparameter case. *Journal of Multivariate Analysis*, 23:183–208, 1987.

[4] P. K. Bhattacharya. Some aspects of change-point analysis. In E. Carlstein, H.-G. Mller, and D. Siegmund, editors, *Change-Point Problems*, pages 28–56. Institute of Mathematical Statistics, 1994.

[5] N. H. Bingham. Variants on the law of the iterated logarithm. *Bulletin of the London Mathematical Society*, 18:433–467, 1986.

[6] R. J. Boik. *A Pair of Primers: Primer on Matrix Analysis and Primer on Linear Statistical Models.* Department of Mathematical Sciences, Montana State University - Bozeman, December 2006.

[7] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[8] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer-Verlag, 2nd edition, 1991.

[9] M. Carrasco and X. Chen. Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, 18:17–39, 2002.

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

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

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

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

[14] A. Dias and P. Embrechts. Change-point analysis for dependence structures in finance and insurance. In G. P. Szeg, editor, *Risk measures for the 21st century*, pages 321–336. Wiley, 2004.

[15] R. F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50:987–1007, 1982.

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

[17] P. D. Grünwald, I. J. Myung, and M. A. Pitt, editors. *Advances in Minimum Description Length: Theory and Applications*. The MIT Press, 2005.

[18] P. Haccou and E. Meelis. The likelihood ratio test for the change point problem for exponentially distributed random variables. *Stochastic processes and their applications*, 27:121–139, 1988.

[19] E. J. Hannan. The uniform convergence of autocovariances. *The Annals of Statistics*, 2:803–806, 1974.

[20] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B (Methodological)*, 41:190–195, 1979.

[21] M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96:746–774, 2001.

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

[23] C. C. Heyde and D. J. Scott. Invariance principles for the law of the iterated logarithm for martingales and processes with stationary increments. *The Annals of Probability*, 1:428–436, 1973.

[24] B. James, K. L. James, and D. Siegmund. Tests for a change-point. *Biometrika*, 74:71–83, 1987.

[25] V. K. Jandhyala and S. B. Fotopoulos. Capturing the distributional behaviour of the maximum likelihood estimator of a changepoint. *Biometrika*, 86:129–140, 1999.

[26] A. D. Jassby and T. M. Powell. Detecting changes in ecological time series. *Ecology*, 71:2044–2052, 1990.

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

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

[29] C. B. Lee. Estimating the number of change points in a sequence of independent normal random variables. *Statistics & Probability Letters*, 25:241–248, 1995.

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

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

[32] T. C. M. Lee. An introduction to coding theory and the two-part minimum description length principle. *International Statistical Review*, 69:169–183, 2001.

[33] A. M. Lindner. Stationarity, mixing, distributional properties and moments of GARCH(p,q). In T. G. Anderson, R. A. Davis, J.-P. Kreiß, and T. Mikosch, editors, *Handbook of Financial Time Series*. Springer, 2008.

[34] The MathWorks, Inc. *GARCH Toolbox User's Guide*, 2 edition, 1999-2002.

[35] J. M. Pasia, A. Y. Hermosilla, and H. Ombao. A useful tool for statistical estimation: genetic algorithms. *Journal of Statistical Computation and Simulation*, 75:237–251, 2005.

[36] L. Perreault, J. Bernier, B. Bobée, and E. Parent. Bayesian change-point analysis in hydrometeorological time series. part 1. The normal model revisited. *Journal of Hydrology*, 235:221–241, 2000.

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

[38] D. Picard. Testing and estimating change-points in time series. *Advances in Applied Probability*, 17:841–867, 1985.

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

[40] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society, series B*, 49:223–239, 1987.

[41] J. Rissanen. *Information and complexity in statistical modeling*. Springer, 2007.

[42] M. Rosenblatt. A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences U.S.A.*, 42:43–47, 1956.

[43] M. Slaney. Auditory toolbox version 2. Technical Report 1998-010, Interval Research Corporation, 1998.

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

[45] V. Strassen. An invariance principle for the law of the iterated logarithm. *Z. Wahrscheinlichkeitstheorie*, 3:211–226, 1964.

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

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

[48] Y. C. Yao. Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *The Annals of Statistics*, 15:1321–1328, 1987.

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

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

[51] Y. C. Yao and R. A. Davis. The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *Sankhya: The Indian Journal of Statistics*, 48:339–353, 1986.

[52] S. Zacks. Survey of classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures of testing and estimation. In *Recent Advances in Statistics: Papers in honor of Herman Chernoff on his Sixtieth Birthday*, pages 245–269. Academic Press, Inc., 1983.

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

Appendix I: Mel-Cepstral Frequency Coefficients

The Cepstrum

Suppose we have a time series $x(t)$ sampled at discrete points in time. In terms of sound data, the time series will be the amplitudes of the sound wave. The *cepstrum* of the series, first defined by Bogart, Healy and Tukey [] in a 1963 paper on echoes resulting from earthquakes and bomb explosions, is obtained by taking the Fourier transform of the log of the power spectral density of the series (Fourier transform of its autocovariance function). In other words,

$$\text{cepstrum of signal} = \text{DFT}(\ \log(\ \text{power spectral density of } x(t)\ )\ ).$$

There are variants on the definition of cepstrum presented by Bogart, Healy and Tukey. It is common to use the inverse Fourier transform of the log of the power spectral density rather than the Fourier transform, i.e.,

$$\text{cepstrum of signal} = \text{IDFT}(\ \log(\ \text{power spectral density of } x(t)\ )\ ).$$

Also, the discrete cosine transform is sometimes used:

$$\text{cepstrum of signal} = \text{DCT}(\ \log(\ \text{power spectral density of } x(t)\ )\ ).$$

The cepstrum can be seen as information about rate of change in the different spectrum bands. Since taking the spectrum of the original time series converts the information from time to frequency, doing a Fourier transform again converts the frequency

information into how fast the spectrum "crosses zero". The low order cepstral coefficients are sensitive to overall spectral slope and the high-order cepstral coeffecients are susceptible to noise.

The word cepstrum comes from mixing the first four letters in the word spectrum. This is because the cepstrum is like the spectrum of the spectrum. We can think of the spectrum as a "frequency series", which if estimated digitally, will be discrete. Analogous to the terms "magnitude" (or amplitude), "frequency", and "phase" in a time series, we can use the terms "gamnitude", "quefrency", and "saphe" to describe these features in the spectrum.

The independent variable of a cepstral graph is called the *quefrency.* The quefrency is a measure of time, though not in the sense of a signal in the time domain. For example, if the sampling rate of an audio signal is 44100 Hz and there is a large peak in the cepstrum whose quefrency is 100 samples, the peak indicates the presence of a pitch that is $44100/100 = 441$ Hz. This peak occurs in the cepstrum because the harmonics in the spectrum are periodic, and the period corresponds to the pitch.

Mel Scale

The mel scale, proposed by Stevens, Volkman and Newman in 1937 [] (J. Acoust. Soc. Am 8(3) 185–190) is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mels. Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments. As a result, four octaves on the hertz scale above 500 Hz are judged to comprise about two octaves on the mel scale. The name mel comes from the word melody to indicate that the scale is based on pitch comparisons.

To convert f hertz into m mel use:

$$m = 1127.01048 \log_e(1 + f/700),$$

and the inverse:

$$f = 700 \left(e^{m/1127.01048} - 1\right).$$

A graph of the relationship between the mel scale and the Hertz scale is shown in Figure 6.1

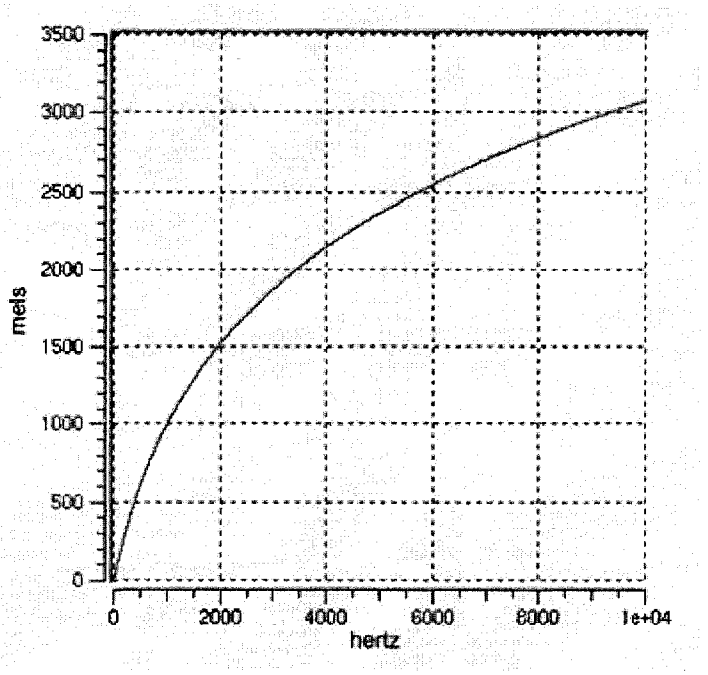

Figure 6.1: Relationship between mel scale and Hertz scale.

Mel Frequency Cepstral Coefficients (MFCCs)

Mel frequency cepstral coefficients are calculated by adding one step to the cepstrum calculation - converting the log spectral density or log Fourier transform to mel scale before taking the discrete cosine transform. The steps to calculate the mel frequency cepstral coefficients for a window of an audio signal are as follows:

1. Calculate the DFT of the frame, $X(k)$.

2. Filter the squared magnitudes by a mel-filter bank, $H_m(k)$:

$$S(m) = \log \sum_{k=0}^{N-1} |X(k)|^2 H_m(k)$$

for $1 \leq m \leq M$.

3. The MFCC is the DCT of the log-energies $S(m)$:

$$c(n) = \sum_{m=1}^{M} S(m) \cos\left(\frac{\pi n(m - 3/2)}{M}\right)$$

for $0 \leq n \leq M - 1$. M is 24-40, but only the first 13 coefficients are used.

The basic difference between the FFT/DCT and the MFCC is that in the MFCC, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT. This allows for better processing of data, for example, in audio compression.