BAYESIAN CHANGE POINT DETECTION IN SEGMENTED MULTI-GROUP AUTOREGRESSIVE MOVING-AVERAGE DATA FOR THE STUDY OF COVID-19 IN WISCONSIN

by

Russell Latterman

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Mathematics

at

The University of Wisconsin-Milwaukee

May 2024

ABSTRACT
BAYESIAN CHANGE POINT DETECTION
IN SEGMENTED MULTI-GROUP AUTOREGRESSIVE MOVING-AVERAGE DATA
FOR THE STUDY OF COVID-19 IN WISCONSIN

by

Russell Latterman

The University of Wisconsin-Milwaukee, 2024
Under the Supervision of Professor David A. Spade

Changepoint detection involves the discovery of abrupt fluctuations in population dynamics over time. We take a Bayesian approach to estimating points in time at which the parameters of an autoregressive moving average (ARMA) change, applying a Markov chain Monte Carlo method. We specifically assume that data may originate from one of two groups. We provide estimates of all multi-group parameters of a model of this form for both simulated and real-world data sets. We include a provision to resolve the problem of confounding ARMA parameter estimates and variance of segment data. We apply our model to identify points in time at which influential events affecting 2020 and 2021 outbreaks of COVID-19 in Waukesha County, Wisconsin, may have occurred.

Dedication

To my brother, Dr. Ryan Latterman, and his newborn daughter

Eleanora,

my baby niece.

To my Mom for challenging me to

complete my dissertation before she completes her book.

It looks to be a statistical tie.

To my step-Mom for helping me learn from an early age to get along with other kids,

no matter how different we are.

To my brother Nicholas, for always telling me why

his mobile devices and personal computing systems are superior,

but without making me feel inferior (usually).

To my Dad for showing me how expansive the universe is.

To Joaquin Bustoz for inspiring myself and so many of my friends through the ASU Math-Science Honors Program, Carlos Castillo-Chavez for restructuring his entire career path to keep the program running, and Cynthia Romero for the encouragement at every milestone over so many years. I will always remember you all.

To all those who have been there for me,

I am dedicated to being there for you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AR or AR(1) Autoregressive Process, first order

MA or MA(1) Moving Average Process, first order

ARMA or ARMA(1,1) Autoregressive Moving Average process, first order

pdf Probability Density Function

pmf Probability Mass Function

NIH National Institutes of Health

CDC Centers for Disease Control

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. David Spade, and the Department of Mathematical Sciences at UWM for giving me the chance to complete my education at a Research 1 institution. I would like to thank my committee members, Professors Istvan Lauko, Chao Zhu, Vytaras Brazauskas, and Richard Stockbridge for not only being so supportive, but also for taking the time to provide me with insight for future work. Additional thanks to Hans Volkmer and Dexuan Xi for being part of the oral exam committee. Special thanks to authors Farhana Sadia, Sarah Boyd, and Jonathan Keith for their inspirational work in Bayesian change-point detection methods. I am much appreciative of Dr. Elisabeth Bennett of Gonzaga University for advice on time management, along with Dr. Winnie Bennett for so much moral support. As I continue to think of the number of individuals I am grateful for having in my life, it borderlines on the countably infinite. I hope that over time I can manage to show everyone the gratitude they deserve.

Chapter 1

Introduction

The goal of this dissertation is to develop a method for identifying points in time at which sudden changes in the rate of daily reported cases of Covid-19 suddenly change. The methodological focus is on development of a Bayesian segmented autoregressive moving average (ARMA) model. We will build upon some recent work [Sadia et al., 2018], [Fearnhead, 2006], to demonstrate the effectiveness of our model on simulated data and on a commonly studied geological data set referred to as the well log data. We will also test our model on many different simulated data sets. We will provide more descriptive analysis of the properties of a well-known and often studied data set, as well as those of our own simulated data, than are demonstrated by in the literature. We propose a new way to calculate probabilities involved with making decisions as far as whether to accept or reject proposed points at which we believe significant changes in the dynamics of the data occur. We do this to prevent the prediction of unrealistic adjacent change points. We will call this the Adjusted Insertion Rule. We will then analyze daily reported cases of Covid-19 from Waukesha County, Wisconsin. In this chapter we provide a review of the literature on relevant concepts. We then discuss the necessary statistical and epidemiological concepts needed for the formal development of our work. We end this chapter with a brief outline of the dissertation.

1.1 Literature Review

The earliest book on changepoint detection, motivated by optimal control theory, was produced in the 1930s for analyzing quality control data [Shewhart, 1931]. Karen Parshnall gives an historical overview of the development of mathematics in the 1920s through 1950s [Parshnall, 2015], citing a number of early authors who looked at such problems from both Bayesian and Classical perspectives. Girschick and Rubin considered the idea that an unknown location of a changepoint could be thought of as a random variable, taking a Bayesian approach [Girschick and Rubin, 1952], leading to further work by Page [Page (1955, 1957)]]. Page was one of the Pioneers in studying methods for detecting changes in mean of a normal distribution from a classical viewpoint. Barnard built upon Page's work by considering minimum sample size for detecting changes, given how limited numerical methods were [Barnard, 1959]. Shiryaev also contributed toward efficient sampling to reduce computational intensity [Shiryaev, 1963]. Chernoff and Zacks took a Bayesian approach to changepoint detection of the mean of normally distributed data on multiple segments [Chernoff and Zacks, 1964]. They found Bayes' estimators to be too computationally intensive to apply in multi-segment models (more than two segments), but their framework helped build a foundation that would be influential when computational speed improved. Sir Adrian Frederick Smith, the current director of the Alan Turing Institute and president of the Royal Society, built upon many previous works, developing more involved methodological approaches by considering cases in which we consider fewer fixed parameters [Smith, 1975]. Much of this can be likened to a simplification of our work in that segmented normal data is like segmented ARMA data

such that the AR and MA parameters are fixed to be zero. In the late 1970s and early 1980s, work in threshold changes in signal processing began to include provisions for auto-regressive behavior observed in data [Tong and Lim, 1980]. Notable work in changepoint detection for time series data in economics was produced in 1989 [Rapoport and Reichlin 1989], focusing on the idea that sudden events, or shocks can potentially cause a permanent change in observed trends. Markov Chain Monte Carlo Methods (MCMC) began to gain traction in the 1990s. Albert and Chib used segmented AR Gibbs sampling techniques to detect changepoints with the assumption that number of changepoints is fixed [Albert and Chib, 1993]. Further similar work was produced by Stephens for more computationally efficient detection of multiple changepoints [Stephens, 1994]. Innovations in applying MCMC and Gibbs procedures lead to applications in forecasting [Pesaran et al., 2006], and mixture models [Giordani and Kohn, 2008]. Koop and Potter studied changepoint detection in a series of data with an unknown number of changepoints by considering that changepoints occur based on a Poisson processes, using MCMC sampling to estimate changepoints while considering segment means and variances to differ [Koop and Potter, 2007]. Fearnhead developed a Bayesian changepoint detection model using MCMC sampling and applied it to the commonly studied well-log data, showing a computationally efficient method for detecting changepoints and estimating the parameters of an auto-regressive segmented model [Fearnhead, 2006]. Similar work was produced by Fearnhead and Clifford for online data, rather than fixed sized data sets [Fearnhead and Clifford, 2003]. Work that allowed for the use of the Dirichlet distribution in predicting locations of an unknown number of changepoints for segmented normal data produced in 2015 [Stanley et al., 2015]. Benson and Alan applied an adaptive MCMC method for detecting changepoints for AR time series data in which number and location of changepoints are both unknown [Benson and Alan, 2016]. This was very similar to the

work by [Fearnhead, 2006]. Work by [Davis et al., 2006] involved a segmented AR model for detection of changepoints by allowing number of points to be fixed and unknown, while also allowing the AR random noise variance to be considered unknown, as well. The well-log data is analyzed further by a segmented Bayesian ARMA model, with the Gibbs sampler used for parameter estimation. A methodological description of a multi-group segmented ARMA model with Gibbs sampling for parameter estimation has been developed by [Sadia, et al 2018]. When we use the term "group," we are referring to the fact that data on a given segment may belong to a group of data with distinguishable characteristics. This concept will be developed more formally in the background section. Our model largely stems from their work. ARMA parameters are allowed to vary based on group assignment. Means about which ARMA segments are generated differ based on groups. These groups are considered to have the same noise variance term. The authors specifically note that this model is designed for data that have similar variance on all segments, regardless of group. Only the single-group form of this model is apparently applied to real world data. They specifically apply their model to the well-log data. We aim to construct a model that includes a provision to allow noise to differ between segments. We also intend to show that the multi-group version of the model we develop can successfully be applied to the well-log data set. The authors note that non-standard form of many parameters results in the need to use approximation methods, rather than direct sampling from the full posterior distributions. We intend to show that direct sampling from the full-posterior distributions of all parameters is possible. As far as we know, the well log data have not been studied using the multi-group version of the model that they describe, nor using an extension of the model in the form that we describe. Other areas of study for similar models (models that include segmentation with either an AR or MA component) include applications to river flow analysis [Cucina et al., 2019],

climatology [Ruggieri and Antonellis, 2016], and neurocomputing [Liu et al., 2020]. None of the models used by these authors are structured in the same way as that of [Sadia et al., 2018]. There also seems to be an absence of the use of a model of this form for the study of daily reported cases of epidemiological data. This is likely due to the fact that such data sets, in raw form, do not have the characteristics of a stationary ARMA process. Either a transformation of the data would be needed in order to use a segmented ARMA model to analyze such data, or a different kind of segmentation modeling would be needed. For example, we have discovered an application of Bayesian changepoint methods for the study of COVID-19 that focuses on the identification of points at which parameters of an epidemiological model change, [Jiang et al]. A Bayesian Poisson segmented regression model is applied, and a MCMC Metropolis Hastings sampling method is used for acceptance and rejection probabilities. We are estimating parameters of a stationary ARMA model in order to identify changepoints in time at which rates of daily case reporting changes. They are identifying changepoints in data series using a segmented Poisson regression model for the main purpose of estimating parameters of an epidemiological model. We will now discuss the statistical background behind many of the technical terms mentioned in this literature review.

1.2 Statistical Background

We will briefly introduce the background needed for in-depth study of the area of work that our results are built upon. This includes an introduction to ARMA processes, the Gibbs Sampler, a cursory review of the Generalized Gibbs Sampler, the availability of epidemiological data, and why Covid-19 data, in particular are so useful when studying epidemiological trends.

1.2.1 Autoregressive Processes and Moving Average Processes

Autoregressive Processes (AR)

Let $\{X_t\}$ be a sequence of random variables such that:

$X_1 = e_1,$
$X_t = \phi X_{t-1} + e_t \ , \quad \boldsymbol{e} = \{e_1, \dots e_t\} \overset{iid}{\sim} N(0, \sigma^2) \quad$ for $t > 1,$

where $\boldsymbol{e} = \{e_1, \dots e_t\} \overset{iid}{\sim} N(0, \sigma^2)$ represent random noise, and ϕ is referred to as the autoregressive parameter. For our purposes, we will assume normality for random noise, although it should be noted that this is not considered a strict requirement. The above model is known as a first order AR process, which can be abbreviated as AR(1). Our initial observation depends only on random noise because there is no previous observation. After the first observation, the data are linearly dependent on the most recent previous observation. If past data are linearly dependent on themselves p-steps into the past, we have an $AR(p)$ process:

$$
\begin{aligned}
&X_1 = e_1 \\
&X_2 = e_2 + \phi_1 X_1 \\
&\dots \\
&X_t = e_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_{p-1} X_{t-i} \quad \text{for } \ 1 < i < \ p \\
&\dots \\
&X_t = e_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_{p-1} X_{t-p} \quad \text{for } \ t \geq \ p
\end{aligned}
$$

where $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_p\}$ are the autoregressive parameters. The t^{th} observation is linearly related to the previous $t-1$ observations for $t \leq p$. When we get to observations $p+1$ and beyond, the t^{th} observation is linearly related to the previous p observations. An AR(p) process can be expressed in an abbreviated form such as:

$$X_t = e_t + \sum_{i=1}^{k} \phi_i X_{t-1} \qquad k = \min(t-1, p) \quad \{e_t\} \overset{iid}{\sim} N(0, \sigma^2)$$

Moving Average Processes (MA)

Suppose we have a series of random variables $\{X_t\}$ such that:

$$X_t = e_t + \sum_{i=1}^{k} \theta_i e_{t-i} \quad k = \min(t-1, q) \qquad \{e_t\} \overset{iid}{\sim} N(0, \sigma^2)$$

where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_q\}$ are the MA parameters. We call this an MA(q) process. The t^{th} observation is linearly related to the previous $t-1$ error terms for $t \leq q$. When we get to observations $q+1$ and beyond, the t^{th} observation is linearly related to the previous q error terms.

1.2.2 Autoregressive Moving Average Processes (ARMA)

An ARMA process has characteristics of both an autoregressive and a moving average process.

Suppose we have a series of random variables $\{X_t\}$ such that

$$X_t = e_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q},$$
$$e_1, \ldots, e_t \overset{iid}{\sim} N(0, \sigma^2),$$

$\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_p\}$ are the autoregressive (AR) parameters and $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_q\}$ are the moving average (MA) parameters. This is called an autoregressive moving average process of order (p, q), or simply an ARMA(p, q) process. Note that AR and MA process are special cases in which one of the ARMA parameters happens to be zero. Going forward, we will focus on first order processes. We will simply refer to these as ARMA processes, rather than ARMA(1,1) processes, with abbreviated form:

$$X_t = e_t + \phi X_{t-1} + \theta e_{t-1}, \qquad e_1, \ldots, e_t \overset{iid}{\sim} \mathrm{N}(0, \sigma^2).$$

Comment on non-zero mean processes: If we want to discuss an ARMA process centered about a non-zero mean, c, we can simply shift a standard ARMA process, giving us the following more general form:

$$X_t = c + e_t + \phi(X_{t-1} - c) + \theta e_{t-1}$$

Stationarity, Invertibility, and Practical Restrictions on Parameters

A time series $\{X_t\}$ is considered stationary if the statistical properties of the data do not depend on time. The mean and variance must remain constant over time, and autocorrelation between any two terms must also be dependent only on their distance apart, not on time, itself. We will discuss the conditions necessary for stationarity of an ARMA process.

Mean, Variance, and Auto-correlation of AR processes, and restrictions on the AR parameter

Given an AR process centered at c

$$X_t = c + \phi X_{t-1} + e_t \ , \qquad \{e_t\} \overset{iid}{\sim} N(0, \sigma^2).$$

It is known that variance, auto-covariance, and auto-correlation functions are given as follows, given $|\phi| < 1$:

Mean	$E[X_t] = c + \frac{1}{1-\lvert\phi\rvert}$
Variance	$Var[X_t] = \frac{\sigma^2}{1-\phi^2}$
Auto-covariance	$Cov(X_t, X_{t-h}) = \frac{\phi^h \sigma^2}{(1-\phi)^2}$
Auto-correlation	$Corr(X_t, X_{t-h}) = \phi^h$

These of course cannot be valid for $|\phi| \geq 1$, simply from an algebraic perspective. From a practical perspective, let us consider $\phi = 1$. Then the expected value of the current term is equal to the expected value of the previous term plus the expected value of the random noise. The expected value of future terms continues to increase in magnitude. And if $|\phi| > 1$, terms will quickly drift arbitrarily far from their starting position. We thus will not have stationarity.

Mean, Variance, and Auto-correlation of MA Processes, and Restrictions on the MA Parameter

Given an MA process centered at c

$X_t = c + e_t + \theta e_{t-1}$, $\{e_t\} \overset{iid}{\sim} N(0, \sigma^2)$.

It is well known that the mean, variance, auto-covariance, and auto-correlation functions are given as follow

Mean	$E(X_t) = c$
Variance	$Var(X_t) = (1 + \theta^2)\sigma^2$
Auto-covariance	$Cov(X_t, X_{t-h}) = \theta\sigma^2$ for $h = 1$, and 0 for $h > 1$.
Auto-correlation	$Corr(X_t, X_{t-h}) = \frac{\theta}{1+\theta^2}$ for $h = 1$, and 0 for $h > 1$.

Auto-covariance and auto-correlation are zero for any two terms further than one time step apart because the observations are a linear combination of the current error term and the most recent previous error term. These are valid regardless of the value of θ. An MA process is always stationary, but there are some practical reasons to restrict θ to less than 1 in magnitude. For $\theta = 1$, notice that past error is just as influential as current error. For $\theta > 1$, it is as if to say that past error is more influential than current error. This does not make practical sense. Another important property of the first order moving average process is called invertibility. The MA process of the above form is invertible if it can be re-expressed as a stationary AR process of infinite order. It can be shown that we must have $|\theta| < 1$ in magnitude for invertibility. Given that our focus is on the application of ARMA processes, this becomes very important. If we have stationarity of an AR process and invertibility of an MA process, the combination of such processes results in a stationary ARMA process.

Given the ARMA process

$X_t = c + e_t + \phi X_{t-1} + \theta e_{t-1},\ \{e_t\} \overset{iid}{\sim} N(\sigma^2),$

We have the following:

Mean $\quad \mathrm{E}(X_t) = \frac{c}{1-|\phi|}$

Variance $\quad Var(X_t) = \sigma^2 \frac{(1+2\phi\theta+\theta^2)\sigma^2}{1-\phi^2}$

Auto-covariance $\quad Cov(X_t, X_{t-1}) = \sigma^2(\theta+\phi)\Big(1 + \frac{(\theta+\phi)}{1-\phi^2}\Big)$

$$Cov(X_t, X_{t-h}) = \phi^{h-1}\sigma^2(\theta+\phi)\left(1+\frac{(\theta+\phi)}{1-\phi^2}\right) \text{ for } h > 1$$

Auto-correlation $\quad Corr(X_t, X_{t-h}) = \frac{(\phi+\theta)(1+\phi\theta)}{1+2\phi\theta+\theta^2} \text{ for } h = 1$

$$Corr(X_t, X_{t-h}) = \phi^{h-1}\frac{(\phi+\theta)(1+\phi\theta)}{1+2\phi\theta+\theta^2} \text{ for } h > 1$$

Notice that if $|\phi| > 1$, these expressions are not valid.

For our purposes, we will require that for a first order ARMA process of the form given above

$$|\phi| < 1, \text{ and } |\theta| < 1.$$

1.2.3 Changepoints and Segmented ARMA Time Series Data

Suppose we have T observations $X_1, X_2, \ldots, X_T$, on K consecutive segments, or discrete intervals of time steps. Now assume that for each segment, k, the data originate from a first order ARMA processes. We will refer to this as a segmented ARMA time series.

$\boldsymbol{I} = \{I_1, I_2, \ldots, I_K\}$, is the collection of segments, where:

$I_1 = [s_1 = 1, b_1]$,

$I_k = [s_k, b_k = s_{k+1} - 1] \text{ for } 1 < k < K-1$,

$I_K = [s_K, b_k = T]$.

Given data on segment I_k, suppose the data originate from an ARMA process that can be described as follows

$X_{s_k} = c_k + e_{s_k}$ The very first observation does not involve the AR and MA parameters.

$$X_t = c_k + e_t + \phi(x_{t-1} - c_k) + \theta e_{t-1} \text{ for } (s_k + 1) \leq t \leq \ b_k$$

The points $\boldsymbol{s} = (s_1, \ldots, s_k)$ are considered changepoints, or points at which each independent segment begins, and $\boldsymbol{b} = (b_1, \ldots, b_k)$ represent the right endpoints. We call $(s_1, \ldots, s_k)$ change points because the dynamics of the data change, starting at these points. The first point can still be referred to as a change point because we essentially have no data until observing the first point, at which point our collection of data changes from empty to non-empty. In this example, the only differences between segments are the segment means, $\boldsymbol{c} = (c_1, \ldots c_k)$.

Segmentation By Groups

Suppose we have a segmented ARMA process of the form described above, in 1.1.3. Now suppose that there are N distinct families of ARMA process from which data may originate, and that data on any given segment originate from one and only one such family of ARMA processes. We will say that we have N groups from which data originate. Groups must differ in at least one property associated with the ARMA processes. A family of ARMA processes can be described as follows. Suppose on the k^{th} segment, the data originate from a group i family of ARMA processes. Then the data originate from an ARMA process of the form

$$\begin{aligned} &x_t = c_k + e_t \text{ for } t = s_k \quad \text{and} \\ &x_t = c_k + e_t + \phi_i(x_{t-1} - c_k) + \theta_i e_{t-1} \text{ for } s_k \leq t < b_k, \end{aligned} \tag{1.1}$$

where c_k is an arbitrary constant representing the segment mean, ϕ_i and θ_i are the AR and MA parameters, and $e_t \overset{iid}{\sim} N(0, \sigma_i^2)$ where σ_i^2 is the group i error variance. The AR and MA parameters, and the error variance are all specific to the group, but the segment observed mean can be any value. The observed segment mean c_k is assumed to have been drawn at random from a normal distribution with mean μ_i and variance τ_i^2. These are the group i expected segment mean, and the variance of the population of all possible group i segment means. We will often simply refer to these as the group mean and group variance, respectively. The term group variance is not to be confused with the group error variance.

$c_k \sim N(\mu_i, \tau_i^2)$ on segment k, given that segment k data originate from group i.

Some further notation is now described. Let $\boldsymbol{g} = (g_1, \ldots, g_k)$ be the group numbers associated with each segment. For example, if we have 3 groups, and $\boldsymbol{g} = (2,1,3)$, this means that segment I_1 data originated from a group 2 ARMA process, segment I_2 data originate from a group 1 ARMA process, and segment I_3 data originate from a group 3 ARMA process. For a changepoint to occur, either the data must begin to originate from a different group, or the data must originate from the same group, but the data must then become centered about a new segment mean.

Utilizing all of our notation, we can say that given a segment k, the data on I_k originate from a group g_k ARMA process of the form

$$\begin{aligned} &x_t = c_k + e_t \text{ for } t = s_k \quad \text{and} \\ &x_t = c_k + e_t + \phi_{g_k}(x_{t-1} - c_k) + \theta_{g_k} e_{t-1} \text{ for } s_k \leq t < b_k, \end{aligned} \tag{1.2}$$

$e_t \overset{iid}{\sim} N(0, \sigma^2_{g_k})$ Error terms

$c_k \sim N(\mu_{g_k}, \tau^2_{g_k})$ Segment means

It is important to note that, as previously discussed, we are now restricting our focus to first order ARMA process. Therefore, when we apply subscripts to our parameters in the context of group segmentation, the subscripts will specifically refer to groups. For example, if we have two AR parameters ϕ_1 and ϕ_2, we mean to say that ϕ_1 is an autoregressive parameter associated with the group 1 first order ARMA process, and ϕ_2 is the autoregressive parameter associated with the group 2 first order ARMA process. In

the context of a discussion on ARMA groups in this project, we do not imply that these are two autoregressive parameters for one second order ARMA process.

The parameter vectors associated with N groups are given, as follows:

$\boldsymbol{\sigma} = (\sigma_1^2, \sigma_2^2, ..., \sigma_N^2)$ Error variance

$\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_N)$ Expected segment means

$\boldsymbol{\tau} = (\tau_1^2, \tau_2^2, ... \tau_N^2)$ Variances of group segment means

$\boldsymbol{\phi} = (\phi_1, \phi_2, ... \phi_N)$ Autoregressive parameters for groups 1 through N

$\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ Moving Average parameters for groups 1 through N

Example 1.2.1. Visualization of Group Segments.

Here we have generated some data with equal segment lengths to illustrate what segmentation by group might look like. In this scenario we have two groups. We generate seven segments for each group, and we alternate the groups, by segment, for illustrative purposes.

Group 1 Segment Data Generation

Every other segment in this example is generated by first sampling a segment mean from a normal distribution based on the expected segment mean for group 1, μ_1, and the variance of segment means for group 1, τ_1^2.

$c_k \sim N(\mu_1 = -15, \tau_1^2 = 3), \text{given } g_k = 1$

After a segment mean, c_k, is sampled, data are generated based on a group 1 ARMA process centered at c_k. Group 2 segments are generated in a similar manner.

Group 1 (lower segments)

$\phi_1 = -0.8 \qquad \theta_1 = -0.1 \qquad \sigma_1^2 = 1$

$x_{s_k} = c_k + e_t$

$x_t = c_k + e_t - 0.8(x_{t-1} - c_k) - 0.1e_{t-1} \qquad e_t \overset{iid}{\sim} N(0, \sigma_1^2 = 1)$

Group 2 (upper segments)

$\phi_2 = 0.2 \qquad \theta_2 = 0.5 \qquad \sigma_2^2 = 2$

$c_k \sim N(\mu_2 = 15, \tau_2^2 = 10) \text{ given } g_k = 2$

$x_t = c_k + e_t + \phi_i(x_{t-1} - c_k) + \theta_i e_{t-1} \qquad e_t \overset{iid}{\sim} N(0, \sigma_2^2 = 2)$

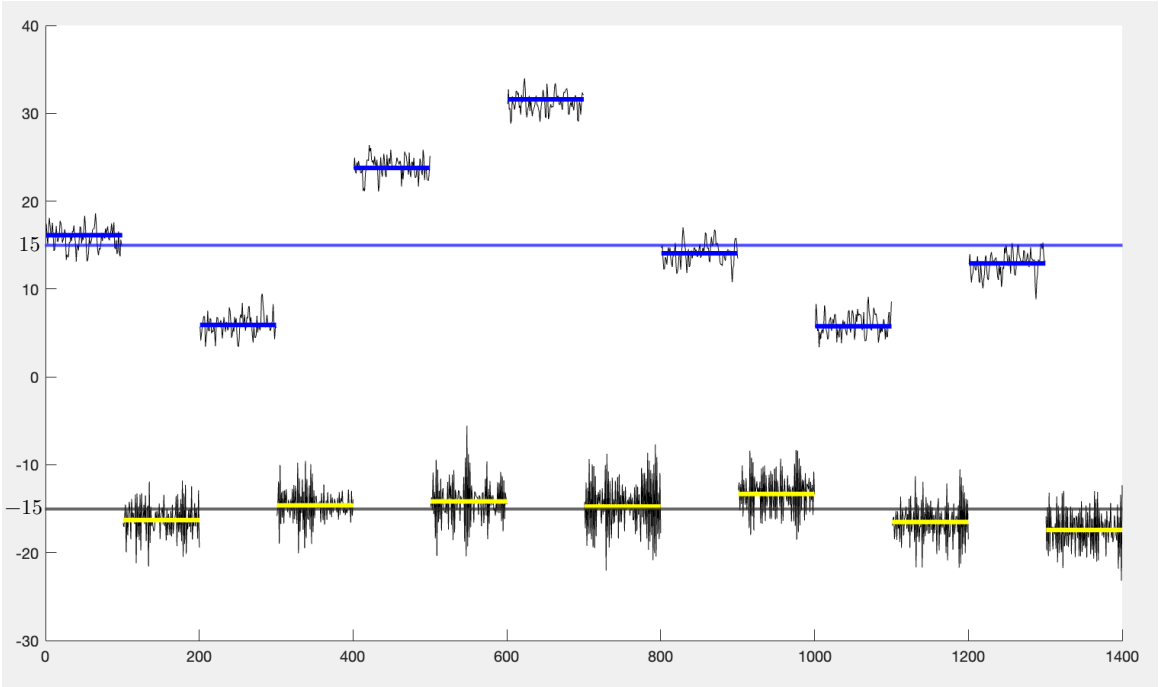


Figure 1.2.1 Hypothetical segmented ARMA data originating from two groups. Group 1 data are in the lower portion of the graph. Group 2 data are in the upper portion.

Fourteen segments of data are generated. Every other segment is a group 1 data segment. The other segments are group 2 data segments. The yellow line segments represent locations of randomly sampled segment means, and the black line represents the expected group 1 segment means. The variance for group 1 segments means is $\tau_1^2 = 3$, which is relatively small compared to the expected group 1 segment mean, $\mu_1 = -15$. This explains why the segments means are somewhat tightly centered about the expected group segment mean. The data within segments appear more erratic for group 1. This is due in large part to the higher magnitude negative AR parameter. Data from group 2 are shown centered about the upper line. The segments are more spread out because the segment mean variance for group 2 is much higher, at $\tau_2^2 = 10$. The spread of the data within the segments is lower because the group 2 error variance is lower, and magnitudes of the AR and MA parameters are much lower.

1.2.4 Markov Chain Monte Carlo Sampling

Suppose we would like to estimate a parameter μ, from a population with density function $p(x|\mu)$. Suppose we set a prior function $p(\mu)$ on the parameter of interest.

By Bayes' Theorem, the posterior distribution is

$$p(\mu|\boldsymbol{x}) = \frac{\textit{joint distribution of the data and } \mu}{\textit{marginal dist of the data}}$$

$$= \frac{(\text{likelihood function of data, given } \mu\ ) \times (\text{prior for } \mu)}{\text{marginal dist of the data}}$$

Symbolically, this is

$$p(\mu|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mu)p(\mu)}{\int p(\boldsymbol{x}|\mu)p(\mu)d\mu}$$

But suppose we either cannot easily integrate the denominator, or suppose that when we do integrate the denominator, we still cannot directly sample from the distribution. We know that $p(\mu|\boldsymbol{x})$ is still proportional to $p(\boldsymbol{x}|\mu)p(\mu)$, as the denominator is simply a proportionality constant. Symbolically, we express this proportionality by

$$p(\mu|\boldsymbol{x}) \propto p(\boldsymbol{x}|\mu)p(\mu).$$

Let us now describe a procedure called the Metropolis Hastings Markov Chain Monte Carlo (MCMC) algorithm. With this method we can still generate a collection of estimated random samples for the parameter μ, based on the overall structure of $p(\boldsymbol{x}|\mu)p(\mu)$. The term Monte Carlo refers to the random sampling that is involved. The value of each estimate will depend on the previous estimate, and only the previous estimate. Thus, we will essentially be generating a Markov Chain. Hence the name Markov Chain Monte Carlo.

1.2.5 Metropolis Hastings Algorithm

Suppose that we know the form of a target distribution $p(\mu)$, up to a normalizing constant: $p(\mu) \propto g(\mu)$.

Choose a proposal distribution $q(\mu)$ that is reasonable and easy to sample from.
Sample an initial estimate for μ_0 from our proposal distribution. Then perform the following iterative process

Metropolis-Hastings Algorithm

For $i = 1,2,..,T$ iterations:

- Draw a candidate sample μ^* from a proposal distribution $q(\mu|\mu_{i-1})$ conditioned on the previous estimate μ_{i-1}.

We then calculate the following ratio:

$$\alpha = \frac{g(\mu^*)/q(\mu^* \,|\mu_{i-1}\,)}{g(\mu_{i-1}\,)/q(\mu_{i-1}|\mu^*)}$$

$$= \frac{g(\mu^*)}{g(\mu_{i-1}\,)} \cdot \frac{q(\mu_{i-1}|\mu^*)}{q(\mu^* \,|\mu_{i-1}\,)} \text{ by rearrangement.}$$

The term on the right, $\frac{q(\mu_{i-1}|\mu^*)}{q(\mu^*|\mu_{i-1}\,)}$, is called the "Hastings Ratio," and it allows us to perform this process without being restricted to symmetric proposal distributions. We have that $q(\mu^*|\mu_{i-1})$ is proportional to the probability that we would choose μ^* if we were to use μ_{i-1} as our estimate for the parameter μ, while $q(\mu_{i-1}|\mu^*)$ is proportional to the probability that we would choose μ_{i-1} as our candidate if our current estimate of μ were to be set as μ^*.

If our proposal distribution is symmetric, this ratio will always be 1 and it can be omitted. In such a case, the algorithm is often simply referred to as the Metropolis algorithm.

- Decision rule:
  Accept μ^* with probability equal to $min\ (1, \alpha)$
  Generate a random number r from a $U(0,1)$ distribution.

If r $< \min(1, \alpha)$, we accept μ^* and set our iterated sample μ_i equal to μ^*.

If $r \geq \min(1, \alpha)$, we reject μ^* and we set $\mu_i = \mu_{i-1}$. This means we do not update our estimate, although we do keep track of this value for each iteration.

After T iterations, we have a collection of random samples, $\{\mu_0, \mu_1, ..., \mu_T\}$ that serves as an estimate for a random sample of T observations from the target distribution. The reason we do not automatically accept our proposal values μ^* on a given iteration i is the fact that we are only sampling from a proposal distribution, which is not the target distribution, itself. The smaller the ratio α happens to be, the less likely it is that μ^* is a

realistic sample from the true target distribution. It is conventional to generate a time series plot of the samples, called a trace plot. The trace-plot will begin to appear stationary, at which point it is as if our samples originate from the target density. An initial segment of time at the beginning of the process will produce what appears to be a nonstationary trend. This is due to the fact that we are sampling from a proposal distribution, not the target distribution, itself. The earlier on we are in the process, the less certain we can be about how well our sampling procedure performs compared to what we would expect if we were able to sample directly from the target distribution. We can discard this as a "burn-in" period. The sample mean of the remaining data can be considered our estimate of the parameter of interest.

We will implement a Metropolis-Hastings decision-making rule algorithm to decide on whether or not to accept or reject proposed change points, and whether or not to delete existing changepoints. For the parameters of our segmented ARMA model, we will use a special case of this Metropolis-Hastings Algorithm called the Gibbs Sampler, explained below.

1.2.6 Gibbs Sampling and Generalized Gibbs Sampling

Gibbs sampling is a special case of the Metropolis Hastings MCMC sampling procedure in which the acceptance probability is always 1. Suppose we are able to sample directly from the target distribution. Then we accept the proposals with probability 1 because we already know that we are sampling from the target density. We can sample over and over

from the target density without needing to calculate acceptance or rejection probabilities. This becomes very convenient when we have multiple parameter that we would like to estimate, as is the case with our ARMA model. We will later show that the full conditional posterior distributions for all of the parameters of our model can be identified, allowing us to apply a Gibbs Sampling procedure.

Gibbs Sampling

Suppose we would like to estimate a number of parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_M)$ from an M-dimensional multivariate distributions. The individual parameters μ_i, may, themselves be multivariate. Suppose we assign priors $\pi_1(\mu_1), \dots \pi_M(\mu_M)$ such that all of the conditional posterior distributions are known. Denote these conditional posterior distributions as.

$p_1(\mu_1|\mu_2, \mu_3, \dots, \mu_M) = p(\mu_1|\cdot)$

…

$p_i(\mu_i|\mu_1, \mu_2, \dots, \mu_{i-1}, \mu_{i+1}, \dots \mu_M) = p(\mu_i|\cdot)$ for $1 < \mathrm{i} < \mathrm{M}$

…

$p_M(\mu_M|\mu_1, \mu_2, \dots, \mu_{i-1}, \mu_{i+1}, \dots \mu_{M-1}) = p(\mu_M|\cdot)$

We will often employ the "bar dot" notation seen above, $p(\mu_i|\cdot)$, as an abbreviation that means "conditioned on all other parameters" or "given all other parameters, when working with conditional distributions.

We start by generating an initial sample of data from prior distributions

Sample $\mu_i^{(0)}$ from each prior distribution $\pi_i(\mu_i)$, for $\mathrm{i} = 1, \dots, \mathrm{M}$.

We now have our prior estimate for the true parameter vector $\boldsymbol{\mu}$.

$\boldsymbol{\mu}^{(0)} = \left(\mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \ldots, \mu_M^{(0)}\right)$

To get our next sample from our posterior distributions, we update our sample vector, one parameter at a time, by directly sampling from the individual posterior distributions, as follows:

For $p_1(\mu_1|\mu_2, \mu_3, \ldots, \mu_M)$, plug in all of our initial sample values, giving us a posterior distribution for μ_1

$$p_1\left(\mu_1 \middle| \mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \ldots, \mu_M^{(0)}\right)$$

Now generate a random sample from the above distribution. This random sample denoted $\mu_1^{(1)}$ is our new estimate of μ_1.

Now we update our vector of estimates by replacing $\mu_1^{(0)}$ with the updated estimate $\mu_1^{(1)}$.

$$\left(\mu_1^{(1)}, \mu_2^{(0)}, \mu_3^{(0)}, \ldots, \mu_M^{(0)}\right)$$

To get our estimate for the second parameter, we plug in $\left(\mu_1^{(1)}, \mu_3^{(0)}, \ldots, \mu_M^{(0)}\right)$ into the conditional distribution for μ_2. This gives us a posterior distribution $p_2\left(\mu_2 \middle| \mu_1^{(1)}, \mu_3^{(0)}, \mu_4^{(0)}, \ldots, \mu_M^{(0)}\right)$. We sample from this distribution to get our updated estimate, $\mu_2^{(1)}$ for μ_2. We do this repeatedly until we get our new set of estimates for our parameters.

Iteration 1

Given the initial sample $\boldsymbol{\mu}^{(0)} = \left(\mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}, \ldots, \mu_M^{(0)}\right)$ from prior distributions,

Plug in $\mu_2^{(0)}, \mu_3^{(0)}, \ldots, \mu_{M-1}^{(0)}, \mu_M^{(0)}$ to $p_1(\mu_1|\,\cdot)$

Sample $\mu_1^{(1)}$ from $p_1\left(\mu_1 \middle| \mu_2^{(0)}, \mu_3^{(0)}, \ldots, \mu_{M-1}^{(0)}, \mu_M^{(0)}\right)$

Plug in $\mu_1^{(1)}$, $\mu_3^{(0)}, \ldots, \mu_{M-1}^{(0)}, \mu_M^{(0)}$ to $p_2(\mu_2|\,\cdot)$

Sample $\mu_2^{(1)}$ from $p_2\big(\mu_2\big|\mu_1^{(1)},\ \mu_3^{(0)},\dots,\mu_{M-1}^{(0)},\mu_M^{(0)}\big)$

Plug in $\mu_1^{(1)},\mu_2^{(1)},\quad \mu_4^{(0)},\dots,\mu_{M-1}^{(0)},\mu_M^{(0)}$ to $p_3(\mu_3|\,\cdot)$

Sample $\mu_3^{(1)}$ from $p_3\big(\mu_2\big|\mu_1^{(1)},\mu_2^{(1)},\ \mu_4^{(0)},\dots,\mu_{M-1}^{(0)},\mu_M^{(0)}\big)$

$\vdots$

Plug in $\mu_1^{(1)},\mu_2^{(1)},\dots,\mu_{m-1}^{(1)}\quad \mu_{m+1}^{(0)},\dots\mu_{M-1}^{(0)},\mu_M^{(0)}$ to $p_m(\mu_m|\,\cdot)$

Sample $\mu_m^{(1)}$ from $p_3\big(\mu_2\big|\mu_1^{(1)},\mu_2^{(1)},\dots,\mu_{m-1}^{(1)},\mu_{m+1}^{(0)},\dots,\mu_{M-1}^{(0)},\mu_M^{(0)}\big)$

$\vdots$

Plug in $\mu_1^{(1)},\mu_2^{(1)},\mu_3^{(1)},\dots,\mu_{M-1}^{(1)}$ to $p_m(\mu_M|\,\cdot)$

Sample $\mu_M^{(1)}$ from $p_M\big(\mu_M\big|\mu_1^{(1)},\mu_2^{(1)},\mu_3^{(1)},\dots\mu_{m-1}^{(1)},\mu_{m+1}^{(1)},\dots,\mu_{M-1}^{(1)}\big)$

We now have our first sample from the posterior conditional distributions

$$\boldsymbol{\mu}^{(1)}=\big(\mu_1^{(1)},\mu_2^{(1)},\mu_3^{(1)},\dots\mu_4^{(1)}\big).$$

We continue to run iterations in this manner.

For iteration k,

Sample $\mu_1^{(k)}$ from $p_1\big(\mu_1\big|\mu_1^{(k-1)},\mu_2^{(k-1)},\mu_3^{(k-1)},\dots,\mu_{M-1}^{(k-1)},\mu_M^{(k-1)}\big)$

Sample $\mu_2^{(k)}$ from $p_2\big(\mu_2\big|\mu_1^{(k)},\ \mu_3^{(k-1)},\dots,\mu_{M-1}^{(k-1)},\mu_M^{(k-1)}\big)$

Sample $\mu_3^{(k)}$ from $p_3\big(\mu_2\big|\mu_1^{(k)},\mu_2^{(k)},\ \mu_4^{(k-1)},\dots,\mu_{M-1}^{(k-1)},\mu_M^{(k-1)}\big)$

$\vdots$

Sample $\mu_m^{(k)}$ from $p_3\big(\mu_2\big|\mu_1^{(k)},\mu_2^{(k)},\dots,\mu_{m-1}^{(k)},\mu_{m+1}^{(k-1)},\dots,\mu_{M-1}^{(k-1)},\mu_M^{(k-1)}\big)$

$\vdots$

Sample $\mu_M^{(k)}$ from $p_M\big(\mu_M\big|\mu_1^{(k)},\mu_2^{(k)},\mu_3^{(k)},\dots\mu_{m-1}^{(k)},\mu_{m+1}^{(k)},\dots,\mu_{M-1}^{k}\big)$

Our k^{th} vector of sample parameters is $\boldsymbol{\mu}^{(k)} = \left(\mu_1^{(k)}, \mu_2^{(k)}, \mu_3^{(k)}, \ldots \mu_M^{(k)}\right)$.

We run this iterative process many times, often many thousands of times. The idea is that after a sufficient number of iterations, our samples behave as though they are coming from the true distribution of the parameter vector $\boldsymbol{\mu}$. Trace plots of the individual parameters can give us an idea of how many iterations we should discard as our burn-in period. The rationale is that, due to uncertainty inherent to our having chosen prior distributions for each parameter, with initial samples originating from these prior distributions, our initial sample vector will not necessarily be a great representative sample of $\boldsymbol{\mu}$. The initial collection of samples, therefore, may not exhibit properties of a random sample from the true population distribution. Burn-in may take longer for some parameters than others. If we see an apparent stationary trend with no sign of strong auto-correlation, we may take the mean of the samples (not including the discarded burn-in samples) as the estimates for the true individual parameters, thus giving us an estimate for the true parameter vector $\boldsymbol{\mu}$.

Gibbs Sampling with Multivariate Parameters

Let us consider the case in which the above parameters $\mu_1, .., \mu_n$ may not all be univariate. This does not necessarily prevent us from applying the Gibbs sampling procedure. For example, suppose $\boldsymbol{\mu} = (\boldsymbol{\mu_1}, \mu_2, \ldots, \mu_M)$ is a vector such that $\boldsymbol{\mu_1} = (\mu_{11}, \mu_{12})$ is a bivariate vector. If we can still sample directly from the full conditional distribution of μ_1, then we can proceed iteratively as was described. If we cannot sample directly from μ_1, then we can still perform the Gibbs sampling procedure as long as we

can sample from the conditional distributions $p(\mu_{11}|\mu_{12})$ and $p(\mu_{12}|\mu_{11})$ separately. Then starting with our initial sample, $\boldsymbol{\mu_1^{(0)}} = \left(\mu_{11}^{(0)}, \mu_{12}^{(0)}\right)$, we will sample $\mu_{11}^{(1)}$ from $p\left(\mu_{11}\middle|\mu_{12}^{(0)}\right)$. Then sample $\mu_{12}^{(1)}$ from $p\left(\mu_{12}\middle|\mu_{11}^{(1)}\right)$, giving us $\boldsymbol{\mu_1^{(1)}} = \left(\mu_{11}^{(1)}, \mu_{12}^{(1)}\right)$.

Gibbs Sampling When Some Full Conditionals are Unknown

We can still perform a similar overall procedure if certain full conditional distributions cannot directly be sampled from. In these cases, we can use the Metropolis Hastings algorithm to simulate samples from certain conditional distributions and draw samples directly for the parameter that have conditional distributions that we can sample directly from.

Generalized Gibbs Sampling

The idea behind the Generalized Gibbs Sampler is that we may end up dealing with a situation in which the dimensionality of the overall parameter space can vary over each iteration. We will try to keep our focus within the scope of what is relevant to what we will be doing in our particular case, given how involved this can become. We will specifically take note of the fact that the number of and location of changepoints will potentially vary after each iteration of our algorithm. For example, let us consider the simple first order segmented moving average model

$$X_t = c + e_t + \theta e_{t-1} \quad \{e_t\} \overset{iid}{\sim} N(0, \sigma^2)$$

Suppose we have a fixed sized data set of T observations, $\{X_t\}_t^T$. When we are trying to determine changepoint locations and estimate how many changepoints we have, $\boldsymbol{s} = (s_1, \ldots s_K)$ and K respectively, we will develop a decision-making process that will be discussed in detail in chapter 2. To introduce the idea of dimensionality change, let us

consider the following idea. Suppose we initially guess that there are only $K^{(0)} = 2$ segments. The first set of changepoint estimates will be denoted $\boldsymbol{s^{(0)}} = \left(s_1, s_2^{(0)}\right)$. Where s_1 is always fixed as the first point in the list. We will also be estimating the parameters of the model, θ and σ^2. To develop this thought process, let us start with an initial guess for θ, which we will call $\theta^{(0)}$. It will be shown in chapter 3 that we can sample directly from conditional distributions for θ and σ^2. Suppose we first sample $\theta^{(1)}$ directly from the conditional distribution of θ, $p\left(\theta|\sigma^{2(0)}, s^{(0)}\right)$. From here we sample $\sigma^{2(1)}$ from $p\big(\sigma^2\big|\mu^{(1)}, s^{(0)}\big)$. This is the typical Gibbs sampling procedure. The sampling procedure for $\boldsymbol{s^{(1)}}$ is much more involved. We will be analyzing the data segment by segment. We will consider inserting a new point, a, in the left segment $\left[s_1, s_2^{(0)}\right]$. Then we will consider the idea of deleting the current point $s_2^{(0)}$. Finally, we will consider inserting a new point, b, in the right segment. Suppose a was inserted. Then the right segment will be $\left[s_2^{(0)} + 1, T\right]$ if $s_2^{(0)}$ was not deleted, or $[a + 1, T]$ if $s_2^{(0)}$ was deleted. Suppose a was not inserted. Then the right segment will still be $\left[s_2^{(0)} + 1, T\right]$ if $s_2^{(0)}$ was not deleted. If a was not inserted and $s_2^{(0)}$ was deleted, then the right segment becomes $[s_1 + 1, T]$. We see that with an initial guess of only two segments, we have six possible outcomes for the dimensionality of the changepoint vector after the first iteration of the algorithm:

$s^{(1)} = (s_1)$	$K^{(1)} = 1$	Old point deleted, no new points inserted.
$s^{(1)} = \left(s_1, s_2^{(0)}\right)$	$K^{(1)} = 2$	No insertions or deletion.
$s^{(1)} = \left(s_1, a, s_2^{(0)}\right)$	$K^{(1)} = 3$	New points, a, appears in the left segment.
$s^{(1)} = \left(s_1, s_2^{(0)}, b\right)$	$K^{(1)} = 3$	New point, b, appears in the right segment.
$s^{(1)} = (s_1, a, b)$	$K^{(1)} = 3$	a and b both inserted, while $s_2^{(0)}$ deleted.
$s^{(1)} = \left(s_1, a, s_2^{(0)}, b\right)$	$K^{(1)} = 4$	a and b both inserted, $s_2^{(0)}$ not deleted.

We are not explaining this in order to formally develop the model structure in this chapter. That will be carried out in chapter 2. We are explaining this now to illustrate how involved a process with dimensionality change can become. With each change in dimensionality, many other aspects of the model will be affected. On every segment we have a different estimated segment mean. Every time the number of segments change, the error terms must be updated, as the error terms are directly related to the segment means. On the j^{th} iteration, it is only after we complete $K^{(j)}$ insertion steps and $K^{(j)} - 1$ deletion steps, updating error terms and segment means estimates every time an insertion or deletion occurs, that we can then perform our next iteration of samples for θ and σ^2. There is no way of knowing how many total calculations we will ultimately perform, and so we must be very careful when applying a generalized Gibbs sampling procedure.

1.3 Epidemiological Background

Volatility of the Data and Availability of Data

For our purposes, we are interested in studying Covid-19, specifically the rates at which daily reporting of cases change, and the way the rates fluctuate. We are interested in this specific disease for a number of reasons. It has been shown to be far more contagious than any widespread airborne illness we have seen since in decades, which has made epidemic dynamics very volatile. Immunity loss combined with new mutations can cause resurgence of an otherwise subsiding epidemic trend. This can create many ebbs and flows of daily reported cases when reporting rate is high. The more contagious a disease, the more prominent these trends may be, assuming we have enough data available. This is very seldom the cases, and reliable county level data have been almost non-existent for common airborne illnesses until the Covid-19 pandemic, and subsequent annual local epidemics that continue to follow. We now have vastly more publicly available testing

data for the Covid-19 epidemics in the US than we have ever had for any airborne illness. During an annual epidemic of a common illness, such as the flu, we get general outbreak trends, but within an epidemic time-frame, the dynamics and volatility of daily cases vary widely from month to month. Different levels of local social interaction are highly influential. These can include more time spent indoors in the winter in colder areas, national holidays travel in locations in which travel may be more congested than in other locations, and events that unique to many localities can also be influential. County fairs, conventions, local election gatherings, concerts, science fairs, sporting events, and even publicly circulated knowledge of local sales of popular items at nearby shopping centers.

Auto-Regressive and Moving Average Qualities of Epidemiological Data

From a qualitative viewpoint, we can understand that reporting of cases can have auto-regressive and moving average properties. For example, consider that the volume of cases reported on a given day may be affected by volume of cases reported on a previous day relative to the actual number of confirmed positive tests that have occurred. The difference between number of lab-confirmed cases and number of publicly reported confirmed cases on a given day can be thought of as random error in a time series of reported cases. If reporting is behind schedule, the number of cases reported will be lower than the number of lab-confirmed cases. An attempt to catch up might be made by including unreported cases from the previous day in the total cases reported on the current day. This conceptually reminds us of an order 1 moving average process with a negative coefficient. Weekly trends may occur due to higher reporting after a weekend, and an abundance of cases may be reported at the end of a month due to policies on monthly reporting policies. This makes us think of the concept of higher order moving average qualities. Autoregressive qualities can also be considered. The number of actual

infections that occur on a given day are positively affected by infections that occur in prior days. This gives us a sense of some kind of auto-regressive nature of disease spread. If rate of testing throughout the population of interest is high, then daily reporting of cases may have a positive autoregressive quality. Covid-19 is one of very few diseases that have resulted in an abundance of data, as we have just mentioned. Symptoms take up to two weeks to develop, and the infected individual may decide to get tested before having any symptoms, during an active infection, or after symptoms subside. Generally, a lab-reported case is not test-confirmed immediately, although this varies widely. Some testing sites have rapid tests with immediate results, but as far as data collection and reporting procedures go, the rate at which these cases are reported to the public will not be immediate. So, there are factors that conceptually seem to indicate a higher-order auto-regressive quality to reported case data. This general thought process does not imply that a time series of daily reported cases for Covid-19 is going to be mathematically equivalent to what we expect to see from a stationary first order moving average process. In fact, time series of reported cases of diseases are rarely stationary, and their autoregressive and moving average properties are generally not first order. However, the general auto-regressive and moving average qualities of epidemiological case data motivate us to explore this kind of data from a mathematical viewpoint. If data sets can be transformed in such a way as to give us something more tractable to work with (i.e. stationarity and lower-order behavior) then we may be able to study such data sets with mathematical ARMA processes in the form discussed earlier in this chapter. We will illustrate this in the last section of chapter 4.

1.4 Dissertation Outline

Chapter 2 Overview

Model Development and Methodology

Our model stems from the work of [Sadia et al., 2018] and [Fearnhead, 2006]. A Gibbs sampling procedure is applied for estimation of parameters, and a Metropolis-Hastings decision making procedure is used for change point detection. We discuss the original model and present the methodology behind our multi-variable version. We make a modification to the decision rule for accepting or rejecting changepoints in order to improve efficiency. We do not allow an insertion of a changepoint at a location at which a changepoint was most recently deleted. We call this the Adjusted Insertion Rule.

Chapter 3 Overview

Derivation of Full Conditional Posterior Distributions

The full conditional distributions for the ARMA parameters have not been provided in closed form by the authors of the models we have studied [Sadia et al., 2018]. It is stated that in such cases, approximation methods are used. In order to directly apply the Gibbs sampling procedure for all parameters, we will need closed form solutions to the equations that describe these probability distributions. We will prove that the ARMA parameters for the model by [Sadia et al., 2018] exist in closed form, and we will use our results to derive closed form solutions for our multi-variable version of this model.

Chapter 4 Overview

Application to Simulated Data and Geological Data

The authors of the original model present the results of an application to a simulated data set as well as a commonly studied geological data set that is generally referred to as "the well log data." We provided an overview of their results, and we spend the remainder of the chapter demonstrating our model's performance with simulated data and with the well log data set.

Chapter 5 Overview

We apply our model in order estimate points in time at which the rate at which cases of Covid-19 are being reported reach a maximum or minimum. We show that even though raw data on daily reported cases are not characteristic of a stationary process, we can transform the data into a form that can be likened to a segmented ARMA process. This makes our model applicable for the study of Covid-19. We apply our model to a data set of daily reported cases from Waukesha County, Wisconsin, from March 2020 through February 2022, drawing on insight from chapter 4 to do this effectively.

Chapter 6 Overview

We conclude with a summary, a discussion of several limitations, and future directions in which our results may take us.

Chapter 2

Methodology and Model Development

We briefly describe the model produced by [Sadia et al., 2018], presented in the form in which it was implemented.

2.1 Foundational Model

Let $\mathrm{X} = \{x_t\}_{t=1}^{T}$ be a series of T observations on K unknown segments.

Within the k^{th} segment, suppose the data originate from a first order ARMA process of the form:

$$x_t = c_k + e_t + \phi(x_{t-1} - c_k) + \theta e_{t-1} \text{ for } s_k \leq t < s_{k+1} \quad e_t \overset{iid}{\sim} N(0, \sigma^2) \tag{2.1}$$

Let g_k represents the group number that segment k data originate from. If $g_k = 1$, segment data originate from group 1, if $g_k = 2$, segment k data originate from group 2, etc. The ARMA processes from which the segment data originate are classified by two groups. The defining characteristics of groups in this setting are the expected segment mean and the variance of the population of segment means. If data from segment k originate from group 1, then segment k will have mean c_k randomly drawn from a population of group 1 means with group mean μ_1 and variance τ_1^2. In general,
$c_k \sim N(\mu_{g_k}, \tau_{g_k}^2)$.

All other parameters are considered equal by group. The statistical dynamics of the data within any two segments are assumed to differ only based on their respective means. The goal is to identify the locations of the left endpoints of the K segments (the change points), while estimating the AR and MA parameters ϕ and parameters θ, in addition to the noise variance σ^2 and the parameters $c_k, \mu_1, \mu_2, \tau_1^2, \tau_2^2$. This is the form of the model that was implemented by the previous authors [Sadia et al., 2018]. The authors implemented their model in a two-group form on simulated data, apparently while also assuming equal group variances. It appears that they applied a one-group form of the model when analyzing a real-world data set, although it is not explicitly stated. It is important to note that for the purpose of developing the sampling methodology, they gave a general description of their model structure based on N groups. We build our model based on this generalization, in addition to several other generalizations.

2.2 Model Structure

We present the model structure by [Sadia et al., 2018], we then present our model, highlighting the key differences.

Suppose that the data $X_1,\ X_2, \dots,\ X_T$ are observed on K segments or intervals, $\{I_k\}_{k=1}^{K}$

$$I_1 = [1 = s_1, b_1]$$
$$I_k = [s_k, b_k = s_{k+1} - 1] \ \text{ for } 1 < k < K - 1$$
$$I_K = [s_K, b_k = T].$$

such that data between segments are independent, and data within segments are generated by stationary and invertible ARMA processes. The vector $\boldsymbol{s} = (s_1, \dots, s_k)$ represents the changepoints, or points at which each independent segment begins, let $\boldsymbol{b} = (b_1, \dots, b_k)$ represent the right endpoints. The first point, s_1 is always 1. It still makes practical sense to call this a "change" point because once we collect our first data point, we have changed from a state of having no information to having our first observation. Overall, for K segments there are $K - 1$ changepoints that we would like to discover when analyzing data.

Suppose that we have N groups, and that data on any given segment are assumed to originate from an ARMA process associated with one and only one group.

Let $\boldsymbol{g} = (g_1, \dots, g_k)$ be the group numbers associated with each segment. For example, if segment I_1 data originated from a group 3 ARMA process, $g_1 = 3$, etc.

In the original model by [Sadia et al., 2018], the data on segment I_k originate from an ARMA process of the form

$$x_t = c_k + e_t \text{ for } t = s_k \quad \text{and}$$
$$x_t = c_k + e_t + \phi(x_{t-1} - c_k) + \theta e_{t-1} \text{ for } s_k < t < b_k \tag{2.2}$$

$e_t \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ Error terms have equal variance, regardless of group.

$c_k \sim N(\mu_{g_k}, \tau^2_{g_k})$ Segment means may have different expected location, μ_{g_k}, and the variance of the population of sample means, $\tau^2_{g_k}$, may differ based on groups.

The number of segments, K, are unknown, as are the changepoints. The parameters σ^2,ϕ, θ are also unknown and do not differ by group. The model thus implicitly assumes that data on all segments have equal overall variances.

We now develop a changepoint detection algorithm. We start with an initial guess for number of changepoints and their locations. We then move from one segment to the next, choosing a random point within each segment. We decide whether or not to consider this new point a changepoint. If we determine this new point to be a changepoint, we say that a changepoint has been inserted. We then move to the next changepoint and decide whether evidence still suggests that this is a changepoint. If we remove this from our list of estimated changepoints, we say that the changepoint was deleted. We do this by segment, updating our estimate of segment mean and segment group along the way. After moving through all of the segments, we then update all of the remaining parameters.

After this, we summarize the variables referenced, and we will discuss the sampling procedure more formally.

Variables

Fixed/known variables

$\boldsymbol{x} = \{x_t\}_{t=1}^{T}$	Fixed observed data set.
T	Fixed number of data points.
N	Number of groups.

Parameters that we must estimate are:

K	Number of segments. Updated each time a changepoint is inserted or deleted.
$\boldsymbol{s} = (s_1, \dots, s_k)$	Left endpoints of our segments (changepoints), the first of which is always 1. Updated every time a changepoint is inserted or deleted.
$\boldsymbol{b} = (b_1, \dots, b_k)$	Right endpoints of segments. The last of which is always T. Updated every time a changepoint is inserted or deleted.
$\gamma \in (0,1)$	Probability that a changepoint occurs at any given time.
$\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$,	Given any segment, the probability that segment data are from group i is π_i. Updated every time a changepoint is inserted or deleted.

$\boldsymbol{y} = (y_1, y_2, \ldots y_N)$	For each group i, y_i represents the number of segments whose data are said to originate from group i, Updated every time number of segments change.
$\boldsymbol{g} = (g_1, \ldots, g_K)$	Group numbers associated with each segment are assigned each time a changepoint is inserted or deleted. Updated each time a changepoint is inserted or deleted.
$\boldsymbol{c} = (c_1, \ldots, c_K)$	Segment means are updated one at a time each time a changepoint is inserted or deleted.
$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$	Expected segment means by group.
$\boldsymbol{\tau} = (\tau_1^2, \tau_2^2, \ldots \tau_N^2)$	Variances of populations of group segment means.
$\boldsymbol{e} = (e_1, e_2, \ldots, e_T)$	Error terms are updated each time a parameter that can affect the error terms is updated.
σ^2	Error variance. Constant, regardless of group.
ϕ	Autoregressive parameter. Equal for all groups.
θ	Moving average parameter. Equal for all groups.

Our Model Structure

Our model differs in that, instead of single parameters ϕ, θ, and σ^2, we allow these parameters to vary by group, in addition to segment means.

$\boldsymbol{\sigma} = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$ Noise variance by group.

$\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots \phi_N)$ Autoregressive parameters for groups 1 through N.

$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ Moving Average parameters for groups 1 through N.

Variance for an ARMA(1,1) process depends on all three of the above parameters.

$$Var(X_t|g_k = i) = \frac{(1 + 2\phi_i\theta_i + \theta_i^2)\sigma_i^2}{1 - \phi_i^2} \tag{2.3}$$

When we are dealing with a model that involves so many other parameters to estimate, the provision to allow these parameters to vary by group adds relatively little additional computational procedures, making this a practical idea to consider. We replace the $\theta, \phi,$ and σ^2 shown in equation (2.2) with $\theta_{g_k},\ \phi_{g_k},\ \sigma^2_{g_k}$, giving us our segmented multi-group model structure

$$\begin{aligned} &x_t = c_k + e_t \text{ for } t = s_k \quad \text{and} \\ &x_t = c_k + e_t + \phi_{g_k}(x_{t-1} - c_k) + \theta_{g_k} e_{t-1} \text{ for } s_k \leq t < b_k. \end{aligned} \tag{2.4}$$

$e_t \overset{iid}{\sim} N(0, \sigma^2_{g_k})$ Error terms.

$c_k \sim N(\mu_{g_k}, \tau^2_{g_k})$ Segment means.

2.2.1 Bayesian Framework for Our Model

Developing the Likelihood Model

For brevity, we will describe all of the following in terms of our model, which requires that we take variation of ARMA parameters and noise variance into account.

Since we are assuming a fixed probability γ of a changepoint occurring at any given time (except time $t = 1$), the probability of any given segmentation of K segments occurring is

$$p(K, \boldsymbol{s}|\gamma) = \gamma^{K-1}(1-\gamma)^{T-K} \tag{2.5}$$

Given that we have K segments, the probability of assigning these segments to a specific sequence of groups, from N possible groups, is

$$p(\boldsymbol{g}|K, \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_{g_k} = \prod_{n=1}^{N} \pi_n^{y_n} \tag{2.6}$$

We can express the model from equation (2.4) in terms of the error terms, as follows:

$e_t = x_t - c_k$ for $t = s_k$, and

$e_t = x_t - (c_k + e_t + \phi_{g_k}(x_{t-1} - c_k) + \theta_{g_k} e_{t-1}\ )$ for $s_k \leq t < b_k$.

For brevity, we can say that

$e_t = \lambda_t$ where

$\lambda_t = c_k$ for $t = s_k$

$\lambda_t = (c_k + e_t + \phi_{g_k}(x_{t-1} - c_k) + \theta_{g_k} e_{t-1}\ )$ for $s_k < t < b_k$

At times we may also use the notation $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_T)$ to refer to these terms.

Since we are assuming that $\{e_t\} \overset{iid}{\sim} N(0, \sigma^2_{g_k})$ within segments,

$x_t = e_t + \lambda_t \overset{iid}{\sim} N(\lambda_t, \sigma^2_{g_k})$ within segment k when all other parameters are held constant.

Distribution functions for various segment means

We have $p(c_k \big| g_k, \mu_{g_k}, \tau^2_{g_k}) = \frac{1}{\sqrt{2\pi\tau^2_{g_k}}} exp(-\frac{1}{2}(c_k - \mu_{g_k})^2)$, and the data between segments are independent. Thus, the joint distribution of the segment means is given by

$$p(\boldsymbol{c}|\boldsymbol{g}, \boldsymbol{\mu}, \boldsymbol{\tau}^2) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\tau^2_{g_k}}} exp\left(-\frac{1}{2}(c_k - \mu_{g_k})^2\right) \tag{2.7}$$

The probability density of the full data set is given by

$$p(\boldsymbol{x}|K, \boldsymbol{I}, \boldsymbol{s}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{g}, \boldsymbol{\sigma^2}) = \prod_{k=1}^{K} p(x|I_k, \phi_{g_k}, \theta_{g_k}, c_k, \sigma^2_{g_k}),$$

which can be expressed as the double product

$$\mathrm{p}(\boldsymbol{x}|K, \boldsymbol{I}, \boldsymbol{s}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{g}, \boldsymbol{\sigma^2}) = \prod_{k=1}^{K} \prod_{t=s_k}^{b_k} \frac{1}{\sqrt{2\pi\sigma^2_{g_k}}} exp\left(-\frac{1}{2\sigma^2_{g_k}}(x_t - \lambda_t)^2\right) \tag{2.8}$$

Alternatively, we can think in terms of the probability density function of the full set of error terms $e_t = x_t - \lambda_t$. We will use the following quite frequently in chapter 3.

$$p(\boldsymbol{e}|\boldsymbol{x}, K, \boldsymbol{I}, \boldsymbol{s}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{c}, \boldsymbol{\sigma^2}) = \prod_{k=1}^{K} \prod_{t=s_k}^{b_k} \frac{1}{\sqrt{2\pi\sigma^2_{g_k}}} exp\left(-\frac{1}{2\sigma^2_{g_k}}(e_t)^2\right) \tag{2.9}$$

Prior Distributions

We will assign the following prior distributions to our parameters, as follows.

$\phi_{g_k} \sim U(-1,1)$	Group AR parameter. Domain restriction is for stationarity.
$\theta_{g_k} \sim U(-1,1)$	Domain restriction is for invertibility.
$\gamma \sim Beta(1,1) = U(0,1)$	Probability of a changepoint occurring at any given time will be given a non-informative prior.
$\pi_i \sim Dirichlet(1,1,\ldots,1)$	We initially assume equal probability for group assignment.
$\mu_{g_k} \sim N(0,1)$	Expected segment mean for groups.
$c_k \sim N(\mu_{g_k}, \tau^2_{g_k})$	Segment means.
$\tau^2_{g_k} \sim IG(\alpha, \beta)$	Variance of segment means within groups.
$\sigma^2_{g_k} \sim IG(\alpha, \beta)$	Error variances within groups.
$P(g_k = n) = \pi_n$	Group assignment probability.

Table 2.2.1 Prior distributions

We will use the following full conditional posterior distributions to update the parameter Using the Gibbs sampling procedure. These are derived in chapter 3.

$\phi_n \vert \cdot \sim TN(\omega_n,\ \nu_n^2, -1, 1)$	Truncated Normal.
See equations 3.23 and 3.24 for ω_n and ν_n^2.	
$\theta_n \vert \cdot \sim TN(\xi_n,\ \eta_n^2, -1, 1)$	Truncated Normal.
See equations 3.32 and 3.33 for ξ_n and η_n^2.	
$\gamma \vert \cdot \sim Beta(\gamma \vert K, T-K)$	Probability of a changepoint occurring at any given time will be given a non-informative prior.
$\pi_i \vert \cdot \sim Dirichlet(y_1+1,\ \ldots, y_N+1)$	Where $y_1, \ldots, y_n$ represent number of segments associated with groups 1 through n.
$\mu_n \vert \cdot \sim N\left(\frac{1}{m_n}\sum_{k\in G_n} c_k\ , \frac{\tau_n^2}{m_n}\right)$	Expected segment mean for groups.
$c_k \vert \cdot \sim N\left(\mu_{c_k}, \delta_{c_k}^2\right)$	$\mu_{c_k},\ \delta_{c_k}^2$ are given on the next page*.

$\tau_{g_k}^2 \vert \cdot \sim IG\left(\alpha_n + \frac{y_n}{2}, \beta_n + \frac{1}{2}\sum_{k\in G_n}(c_k - \mu_n)^2\right)$

α_n and β_n are the parameters for the group-based priors.

$\sigma_n^2 \vert \cdot \sim IG\left(\alpha_n + \frac{T_n}{2}, \beta_n + \frac{1}{2}\sum_{k\in G_n}\sum_{t=s_k}^{b_k} e_t^2\right)$ Error variances within groups.

α and β are the parameters for the prior. T_n is the total data within group n.

$p(\boldsymbol{g}\vert\cdot)$ The full conditional posterior distributions of each g_n are discrete with parameters proportional to $\prod_{k=1}^K N\left(c_k \vert \mu_{g_k}, \tau_{g_k}^2\right) \prod_{k=1}^K \pi_{g_k}$

We plug all of our parameter values $c_k, \mu_{g_k}, \tau_{g_k}^2,\ \pi_{g_k}$ into $N\left(c_k \vert \mu_{g_k}, \tau_{g_k}^2\right)\pi_{g_k}$, to get the individual probabilities:

$$P(g_k = n \vert \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}) = \frac{N\left(c_k \vert \mu_n, \tau_n^2\right)\pi_n}{\prod_{k=1}^K N\left(c_k \vert \mu_{g_k}, \tau_{g_k}^2\right) \prod_{k=1}^K \pi_{g_k}}$$

Table 2.2.2 Full Conditional Posterior Distributions

*The mean and variance for the full conditional posterior distribution for segment means as listed in table 2.2.2 are denoted μ_{c_k} and $\delta^2_{c_k}$ respectively.

$$\mu_{c_k} = \frac{\tau^2_{g_k}\left((1-\phi)\sum_{t=s_k+1}^{b_k}(x_t - \phi x_{t-1} - \theta e_{t-1}) + x_{s_k}\right) \quad + \sigma^2_{g_k}\mu_{g_k}}{\tau^2_{g_k}(1+(b_k-s_k)(1-\phi)^2) \quad + \sigma^2_{g_k}}$$

$$\delta^2_{c_k} = \frac{\sigma^2_{g_k}\tau^2_{g_k}}{\tau^2_{g_k}(1+(b_k-s_k)(1-\phi)^2) \; + \; \sigma^2_{g_k}}$$

Conditional Dependencies of Parameters
Figure updated to reflect our model

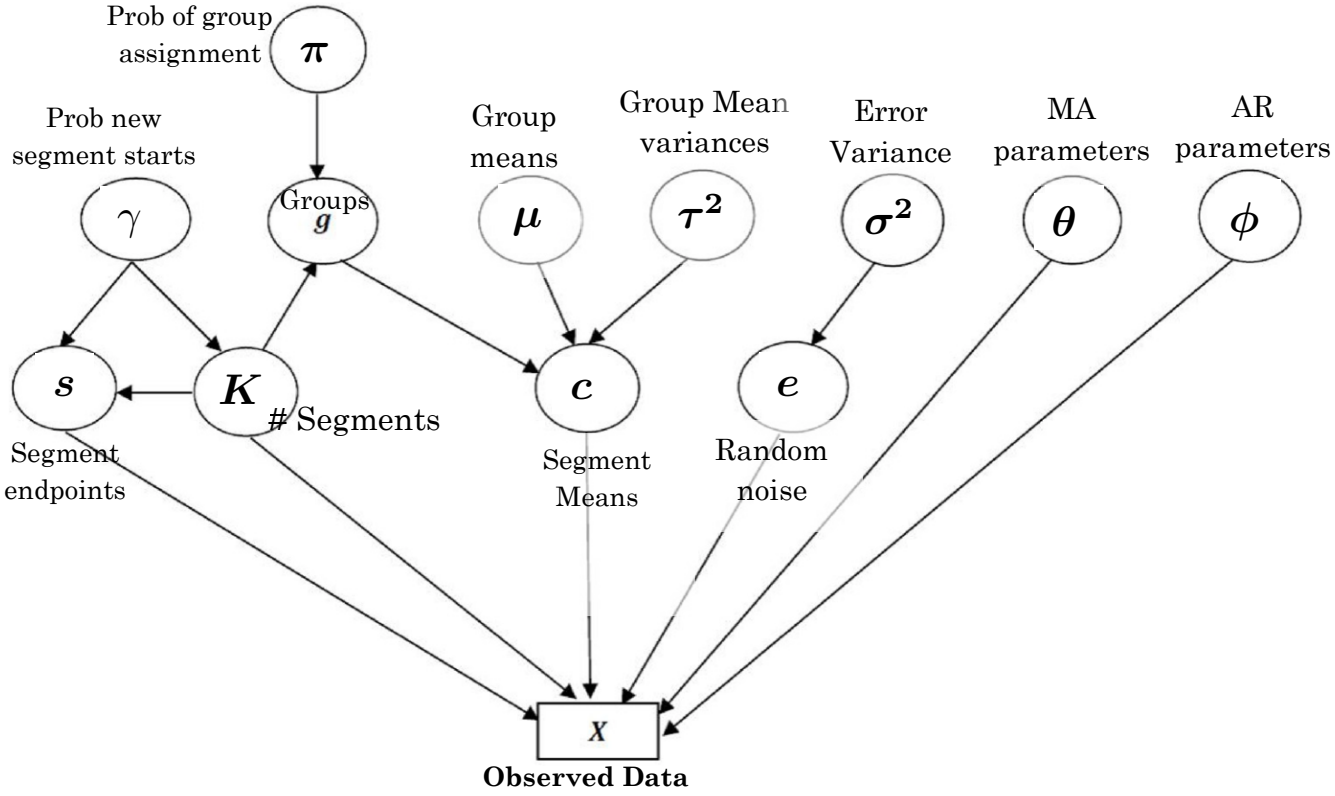


Figure 2.2.1 Adapted from Sadia et al. 2018. Conditional dependencies of parameters.

The above figure shows the dependency of the parameters. We have added descriptions next to each symbol in the figure, and we have changed the symbolic notation for two parameters to reflect our notation. Their original notation for the MA parameter was ψ,

rather than $\boldsymbol{\phi}$. Their original notation for γ was ϕ. The data, of course, depend in some way on all parameters. The conditional posterior distributions for all parameters will be derived in chapter 3. The AR, MA, and error variance are shown in bold to reflect the fact that our model considers these to be multi-variable vectors of parameters based on groups. We will now discuss the procedures for estimating changepoint locations and for updating parameter values.

Choosing Locations to Insert and Delete

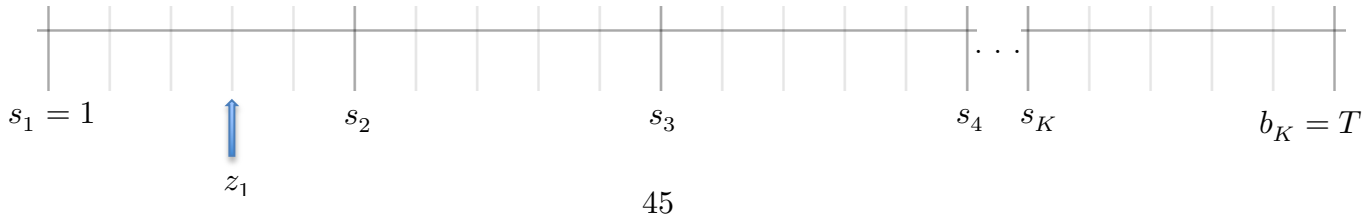
Start off with an initial guess for number of segments. Fix some initial guess γ_0 to be the probability of a changepoint occurring at any given time point. Then generate a list of changepoints as our initial sample. The first point is always 1, signifying the left endpoint of the first interval.

$\boldsymbol{s^{(0)}} = \left\{s_1^{(0)} = 1, s_2^{(0)}, \ldots, s_{K_0}^{(0)}\right\}$, where K_0 is our initial guess for total number of segments.

For brevity, we will omit the superscripts until the end of the insertion-deletion process. We will have a vector of segment endpoints that we will add to as we iterate through the entire list. We will call this S_{new}. The first point will always be 1, the left endpoint. We will start with

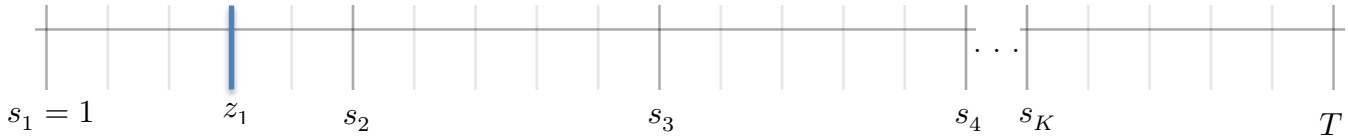
$S_{new} = \{1\}$. This will be a new list of updates. At the end of our insertion/deletion process, S_{new} will become our current list of changepoints.

Insertion step 1: Randomly choose a point z within the first segment as a potential new cutoff point.

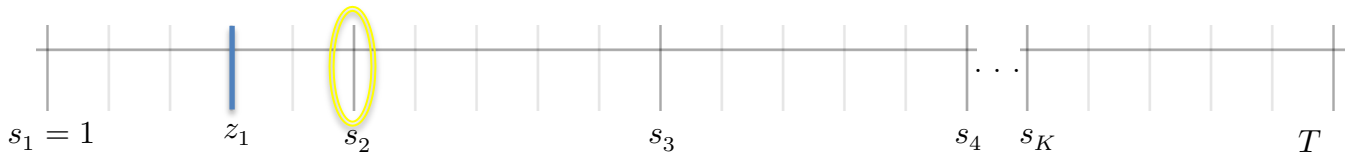


Based on a probability model that we will develop, we will decide whether to accept z as a new changepoint or reject z. If we accept z as a new changepoint, we update our list.

$$S_{new} = \{1, z_1\}$$



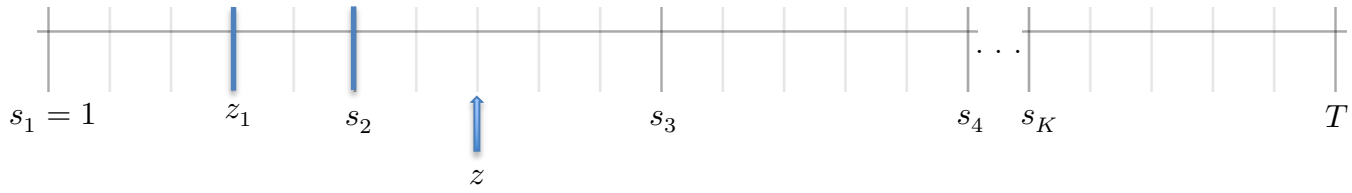
Deletion Step 1: We now move on to the changepoint s_2. We decide whether to delete this point or to keep it.



Suppose we do not end up deleting this point, then we add this to our new list of changepoints.

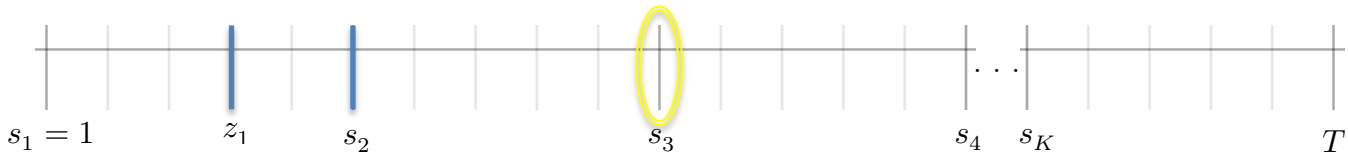
$S_{new} = \{1, z_1, s_2\}$

Insertion Step 2: Since we did not delete s_2, we will perform the same procedure that we performed in the first insertion step. We will randomly propose a new changepoint between s_2 and s_3.



Suppose that we fail to insert a new changepoint at this location. Then after this insertion step, we do not update our S_{new} vector of points.

Deletion step 2: Decide whether to delete s_3.

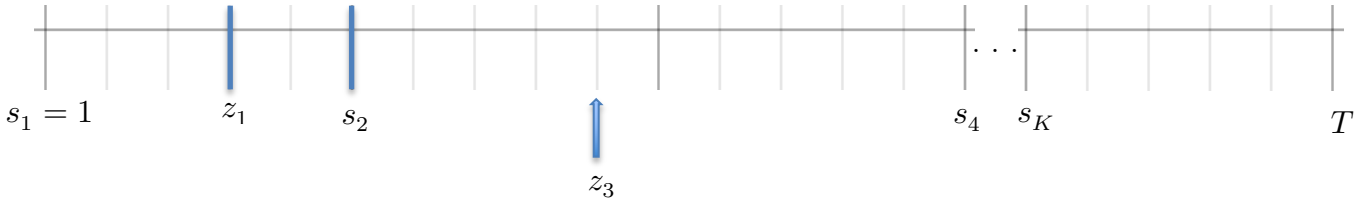


Suppose we do end up deleting s_3.

Then we will not include s_3 in the new vector. The vector remains the same.

Insertion interval spans from s_2 to $s_4 - 1$

Insertion step 3: In this scenario, since we deleted s_3, our insertion interval will span all the way back to the most recent changepoint. We once again randomly choose a point as a potential changepoint. Since we deleted s_3, deciding on whether to insert a point here is, in some sense, equivalent to deciding whether to move s_3 to a new location after having deleted it.



One interesting aspect of this algorithm is that it leaves open the possibility of randomly choosing s_3 again as the potential location for a new changepoint. We could decide to not allow s_3 to be a candidate point during this iteration, although the original algorithm by [Sadia et al., 2018] did not indicate that we do this. Neither did we restrict this in our initial implementation of the code. We later decided to implement our own rule, which will call the Adjusted Insertion Rule. We adjust for a deletion by not allowing that we immediately re-inserted a point that was just deleted. This will be further discussed in later sections. Suppose we decide to insert z_3 as a new changepoint. Then our vector of changepoints becomes

$$S_{new} = \{1, z_1, s_2, z_3\}.$$

Here is a list of a hypothetical continuations of our algorithm

Deletion step 3: Potential deletion point s_4.

Outcome: s_4 not deleted

Update: $S_{new} = \{1, z_1, s_2, z_3, s_4\}$

Insertion Step 4: Randomly choose a point z_4 between s_4 and s_5.

Outcome: Fail to insert z_4

Update: We do not update our vector of changepoints.

Deletion Step 4: Potential deletion point s_5

Outcome: Fail to delete s_5.

Update: s_5 is added to our vector of changepoints

$$S_{new} = \{1, z_1, s_2, z_3, s_4, s_5\}$$

.

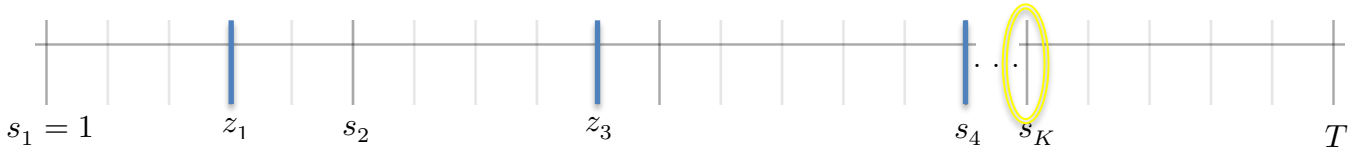
.

Insertion step $K-1$: Randomly choose a point z_{K-1} between s_K and the nearest changepoint to the left of s_K.

Outcome: Insert z_{k-1}

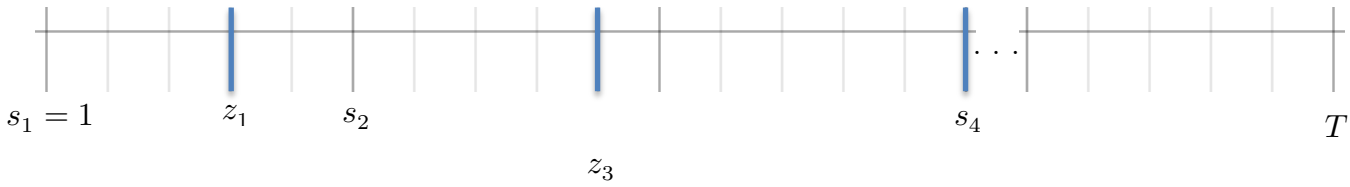
Update: $S_{new} = \{1, z_1, s_2, z_3, s_4, s_5, \dots, z_{k-1}\}$

Deletion step $K-1$: Decide whether to delete s_K. This is the last deletion step.



Outcome: Suppose we delete s_K. Then we do not add s_K to our new vector.

Insertion step K: Last insertion step



Suppose we insert z_K, then our updated vector becomes

$$S_{new} = \{1, z_1, s_2, z_3, s_4, s_5, \dots, z_{k-1}, z_K\}$$

It is important to note that there is no K^{th} deletion step. There are a total of K insertion steps and $K-1$ deletion steps.

After performing K insertion steps and $K-1$ deletion steps, our vector of estimated changepoint locations s becomes S_{new}.

$S_{new} = \{1, z_1, s_2, z_3, s_4, s_5, \ldots, z_{K-1,} z_K\}$.
In our algorithm, the new set of changepoints is re-labelled as:
$\boldsymbol{s}^{(\mathbf{1})} = \left\{1, s_2^{(1)}, s_3^{(1)}, s_4^{(1)}, s_5^{(1)}, s_6^{(1)}, \ldots, s_{K_1-1}^{(1)}, s_{K_1}^{(1)}\right\}$.
Where $s_1^{(1)}$=1 because the first point in our list is always 1.
The new total number of segments is K_1, where K_1 is the length of $\boldsymbol{s}^{(\mathbf{1})}$. This is the estimated number of segments after we perform our first iteration of insertions and deletions.

Issue with singleton set intervals. Can we choose the rightmost endpoint, T, as a potential changepoint? Can we have one changepoint immediately followed by another changepoint? The last insertion step is unique in that we do include the right endpoint in the set of possible points that may be defined as changepoints. From a practical standpoint, if we have an outlier at the very last point, it is possible that this indicates the beginning of a new segment, and that if we were to continue gathering data, we might end up confirming this. Suppose we did end up choosing $z_K = T$, and suppose we did end up keeping this as a changepoint. Then $S_{new} = \{1, z_1, s_2, z_3, s_4, s_5, \ldots, z_{K-1,} z_K = T\}$. We would have an "interval" $I_K = [T, T]$. This implies that we do allow a singleton set to be a legitimate interval of data. In general, the algorithm that we use does allow for adjacent changepoint occurrences, which would imply that we may end up with many singleton

intervals. Whenever we arrive at an interval of length one, we skip the insertion step, as there are no possible points to insert. The deletion step then involves deciding whether to delete this singleton point. Typically, points like this will be deleted, as the probability distribution function we will use to make our insertion/deletion decisions will assign very low probabilities to changepoints that result in very short intervals. However, an issue can arise when a singleton set is deleted. On the subsequent insertion step, the point can immediately be inserted. This is likely to occur on relatively short intervals.

Adjusted Insertion Rule

A proposal for a new decision-making procedure for post-deletion insertion step.

Our decision rule differs slightly compared to what we have seen in other studies. We eliminate the possibility of immediate reversal of a deletion. We simply do this by making an additional important provision. We will call this the Adjusted Insertion Rule. We are not allowed to insert a point at which a deletion occurred in the previous deletion step. In other words, the Adjusted Insertion Rule says that on the k^{th} insertion step, we cannot immediately undo a deletion that occurred on deletion step $k-1$. We apply what we call the adjusted post-deletion procedure to each insertion step that occurs after a deletion.

Adjusted Post-Deletion Procedure:

Suppose a point s_k is deleted. Then we end up with a merged interval $[s_{k-1}, . . , s_k, ... b_k]$

Now s_k no longer denotes a changepoint. Instead s_k simply denotes the location at which we deleted a point. On the subsequent insertion step, we have two cases when applying the Adjusted Post-Deletion Procedure

Case 1: If $[s_{k-1}, . . , s_k, ... b_k]$ is of length 2, then

$$[s_{k-1}, . . , s_k, ... b_k] = [s_{k-1}, s_k]$$

The only possible point in the list at which we could choose to insert a new changepoint is the position s_k. By the Adjusted Insertion Rule, we are not allowed to do this because s_k was just deleted, so we do not do a deletion step.

Case 2: $[s_{k-1}, . . , s_k, ... b_k]$ is of length greater than 2

Then we will randomly choose any point within the adjusted set

$z \in [s_{k-1}, . . s_{k-1}] \cup [\, s_{k+1}, ... b_k]$ as our new potential change point.

All other calculations will be performed in the same way. From this point forward, we will assume that the Adjusted Insertion Rule is being followed unless otherwise noted. For example, in the simulations section and the application to well-log data section, we will consider cases with and without the Adjusted Insertion Rule applied.

Algorithm for Changepoint Decision Rules and Gibbs Sampling of parameters

We need to formally define a decision rule for how to accept or reject these change points. We will describe the method by [Sadia et al., 2018], and we will then describe our method. Theirs resembles a Metropolis Hastings acceptance rule. Ours happens to be based directly on the Metropolis Hastings decision rule described in Chapter 1.

Pseudo-Code Algorithm Structure

Randomly sample all parameters using the prior distributions for each. Produce an initial plot of the data.

For $i = 1{:}\, R$ Where R is number of iterations. Generally, 5 to 10 thousand.

For $k = 1{:}\, K_i$ Where K_i is the current number of estimated segments.

Insertion step

Make decision to insert a proposed changepoint z within segment

*Randomly propose a changepoint point z within the interval $[s_k, b_k]$ to form left and right intervals $[s_k, z]$ *and* $[z, b_k]$.

Using the prior distributions for groups randomly assign groups g_l and g_r to the left and right intervals, given the current probabilities of a randomly chosen segment being from groups 1 through N, respectively.

g_l and g_r are both sampled as discrete uniform with parameters $(\pi_1, \ldots, \pi_N)$.

Using the prior distribution for segment means, conditioned on the randomly assigned groups for the left and right intervals, randomly sample left and right interval means c_l and c_r.

$$c_l \sim N(\mu_{g_l}, \tau^2_{g_l}) \qquad c_l \sim N(\mu_{g_r}, \tau^2_{g_r})$$

*Update the error terms associate with each interval, recursively, based on the structure of the ARMA process.

The original model by [Sadia et al., 2018] requires that we update the error terms, as does our model. Our model differs in that we must also use specific values for ϕ_{g_l} and θ_{g_l} when updating the error terms. Plug in the interval data and the parameters ϕ_{g_l} and θ_{g_l} corresponding to the assigned groups.

Left error terms

$e_t^{(l)} = x_t - c_l$ for $t = s_k$

$e_t^{(l)} = x_t - \left(c_l + \phi_{g_l}(x_{t-1} - c_l) + \theta_{g_l} e_{t-1}^{(l)}\right)$ until $t = z - 1$.

Right error term update procedure is analogous.

*__Calculate the probability of insertion__, given the randomly assigned groups, segments, and updated error terms.

The details on this calculation are more involved and are discussed in 2.2.2.

*Generate a random number from a uniform distribution. If this number is less than the insertion probability, the insertion occurs.

If an insertion occurs, this results in keeping the two new intervals and further updating segment means and groups.

The left and right segment g_l and c_l are further updated by sampling from the full conditional posterior distributions for segment groups and segment means, shown in table 2.2.2. After each update, error terms must be updated recursively to reflect the changes.

Number of existing intervals, K, becomes $K + 1$ if an insertion occurs.

If an insertion does not occur

Do not update any parameters.

Proceed to deletion step.

K does not change.

Deletion step.

We decide whether to remove s_{k+1} by generating a merged interval.

Case scenarios:

If an insertion did previously occur in the insertion step

Current interval is $[z, b_k]$, next interval is $[s_k, s_{k+1}]$

Merged interval is $[z, s_{k+1}]$

If an insertion did not occur

Current interval is $[s_k, b_k]$, next interval is $[s_k, b_{k+1}]$

Merged interval is $[s_k, b_{k+1}]$

*Randomly sample a new group, g_M, for the merged interval by sampling g_M from a uniform distribution with parameters $\pi_1, \ldots, \pi_N$.

*Randomly sample a new mean, c_M, for the merged interval by sampling c_M from the prior distribution for segment means, given the group g_M.

$$c_M \sim N(\mu_{g_M}, \tau^2_{g_M})$$

*Update the error terms recursively.

*Calculate probability of insertion.

This will be discussed in 2.2.2

*Generate a random sample from a $U(0,1)$ distribution. If the sample is less than the deletion probability, the proposed deletion occurs, and we keep the merged interval as our updated new interval.

If a deletion occurs

Further update the segment groups c_M and g_M for the merged segments using the full conditional posterior distributions for segment means and segment groups. Update the error terms each time.

Number of segments K becomes $K-1$ because we have reduced the total number of segments.

If a deletion does not occur, do not update any parameters.

End Insertion/Deletion loop

After performing all insertion and deletion steps

Gibbs Sampling for parameter updates.

Update $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}$, by sampling from the full conditional posterior distributions for each. The order in which the sampling occurs is up to the operator, although it is suggested that a consistent order be used for each iteration.

Error terms should be updated after ϕ is updated, and again after θ is updated

End of iterations loop

Post Iteration Procedures

Perform a final update of all parameters, segment means, groups, and error terms.

After the iterations loop ends, it is highly suggested that the user:

- Produce a graph of the data, along with changepoint locations and identification of segment means.
- Produce trace plots of all parameters to determine whether our parameter estimates seem plausible.

- If an excess number of changepoints are detected, decide on whether to remove superfluous changepoints. This can be done in a number of ways and will later be discussed.

- If too few changepoints are found, or if trace plots indicate a lack of convergence, we may consider running the algorithm for a greater number of iterations.

- If trace plots appear to lack stationarity, we may want to consider changing model assumptions. For example, we may find that a simpler model, like AR or MA, is sufficient or better suited. We may also find it appropriate to have some parameters fixed, rather than vary throughout the simulation. After making changes, the program can be run again.

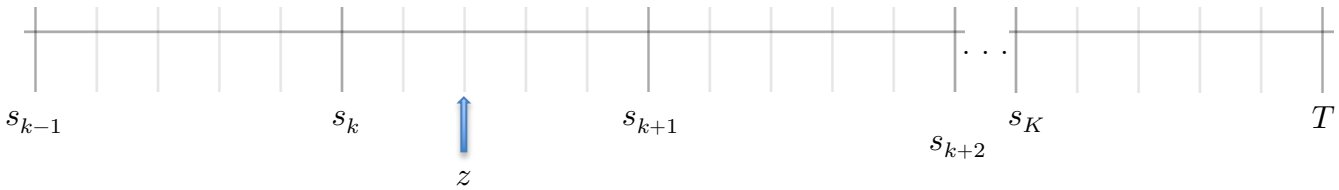
- Intermittent graphical updates. It is suggested that while the loop runs, intermittent graphical updates for the changepoint locations are made. By viewing these, insight may be gained. Similar updates can be made for trace plots.

Calculating Insertion and Deletion Probabilities

We will now develop the method for calculating the acceptance/rejection probabilities for the insertion and deletion steps.

2.2.2 Insertion and Deletion Probabilities

Insertion Step k



If we accept the new changepoint, z, this will result in two new intervals

Use the prior distribution for groups to sample g_l: The left interval is randomly assigned a group using the prior distribution for g. This is a discrete distribution with probability $\pi_1, \ldots, \pi_N$ probabilities of assigning the left segment to groups 1 through N, respectively.

Based on this group assignment, a random segment mean is chosen from the prior distribution for segments means, $c_l \sim N(\mu_{g_l}, \tau_l^2)$.

g_r: The right interval is randomly assigned a group g_r and segment mean c_r in the same way.

In the models by [Sadia et al., 2018] and [Fearnhead, 2006], segments all have equal variance and ARMA parameters., and thus the error variance does not depend on the group. Using these new segment groups and segment means, we iteratively update each error term by plugging in the new segment mean values, along with the current estimates of the ARMA parameters, along with the data.

Left error terms

$$e_t^{(l)} = x_t - c_l \text{ for } t = s_k$$
$$e_t^{(l)} = x_t - (c_l + \phi(x_{t-1} - c_l) + \theta e_{t-1}) \text{ for } s_k < t \leq z - 1$$

Right error terms

$$e_t^{(r)} = x_t - c_r \text{ for } t = s_k$$
$$e_t^{(r)} = x_t - (c_l + \phi(x_{t-1} - c_l) + \theta e_{t-1}) \text{ for } z < t \leq b_k$$

$f\left(e_t^{(l)}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{{e^{(l)}}_t^2}{2\sigma^2}\right) \propto \frac{1}{\sigma} exp\left(-\frac{{e^{(l)}}_t^2}{2\sigma^2}\right)$

and $\quad f\left(e_t^{(r)}\right) \propto \frac{1}{\sigma} exp\left(-\frac{{e^{(l)}}_t^2}{2\sigma^2}\right)$

Probability of accepting the insertion

Based on original model

$$P_{insert} \propto \gamma \prod_{t=s_k}^{z-1} p\left(e_t^{(l)} \middle| 0, \sigma^2\right) \prod_{t=z}^{b_k} p\left(e_t^{(r)} \middle| 0, \sigma^2\right) \frac{1}{H_{K+1}}$$
$$= \frac{\gamma}{H_{K+1}} exp\left(-\frac{1}{2\sigma^2}\left(\sum_{t=s_k}^{z-1}\left(e_t^{(l)}\right)^2 + \sum_{t=z}^{b_k}\left(e_t^{(r)}\right)^2\right)\right)$$

Where H_K and H_{K+1} are correction facts explained below.

Based on our model

The error variance terms differ based on group. The subscripts g_l and g_r denote the groups corresponding to the left and right segments, respectively.

$$P_{insert} \propto \gamma \prod_{t=s_k}^{z-1} p\left(e_t^{(l)} \middle| 0, \sigma_{g_l}^2\right) \prod_{t=z}^{b_k} p\left(e_t^{(r)} \middle| 0, \sigma_{g_r}^2\right) \frac{1}{H_{K+1}}$$
$$= \frac{\gamma}{H_{K+1}} \left(\frac{1}{\sigma_{g_l}}\right)^{z-s_k} \left(\frac{1}{\sigma_{g_r}}\right)^{b_k-z+1} exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_{g_l}^2}\sum_{t=s_k}^{z-1}\left(e_t^{(l)}\right)^2 + \frac{1}{\sigma_{g_r}^2}\sum_{t=z}^{b_k}\left(e_t^{(r)}\right)^2\right)\right)$$

We calculate this product by plugging in the error terms to the pdf of the distribution function for the error terms. Given each t.

We multiply by γ, the probability of a changepoint occurring, and we also scale this based on a correction factor $\frac{1}{H_{K+1}}$ to account for the fact that when we insert a new point, we now have $K+1$ estimated intervals.

Correction Factors H_K and H_{K+1}, and H_{K-1}

These account for number of actions or moves that are possible, given the number of intervals, K. In a full run-through of insertion/deletion steps, we have K insertion steps and $K-1$ deletion steps, K segment group updates and K segment mean updates. This is a total of $4K-1$ actions. There are N possible segment group means to update $\mu_1, \ldots \mu_N$, N segment group mean variances to update $\tau_1^2, \ldots \tau_N^2$, a_1 AR parameters to update (where a_1 is the order of the AR process), a_2 MA parameter to update (where a_2 is the order of the MA process), and three more parameters π, σ^2, and γ to update. It is important to keep a_1 and a_2 general. We are restricting our complete model to either first order ARMA segmentations $a_1 = 1$ and a_2=1, but if we want to compare the performance of such a model to a model that involves only the AR component, only the MA component, or just random noise, one or both of the above may be zero, rather than one. This gives us

$H_K = 4K - 1 + 2N + a_1 + a_2 + 3$

Practical Correction Factors

Given that we will be using two groups, that gives us

For ARMA(1,1)	For MA(1) or AR(1) Models
$H_K = 4K + 8$	$H_K = 4K + 7$
$H_{K+1} = 4K + 12$	$H_{K+1} = 4K + 11$
$H_{K-1} = 4K + 4$	$H_{K-1} = 4K + 3$

For a Brownian Motion process (ARMA(0,0))

$H_K = 4K + 6$
$H_{K+1} = 4K + 10$
$H_{K-1} = 4K + 2$

Probability of Rejecting the Insertion

This is calculated over the entirety of the original interval, without calculating any new error terms.

Original error terms

$$e_t = x_t - c_k \text{ for } t = s_k$$
$$e_t = x_t - (c_k + \phi(x_{t-1} - c_l) + \theta e_{t-1}) \text{ for } s_k < t \leq b_k$$

$$e_t = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{e_t^2}{2\sigma^2}\right) \propto \left(\frac{1}{\sigma}\right) exp\left(-\frac{e_t^2}{2\sigma^2}\right)$$

$$P_{reject} \propto (1-\gamma) \prod_{t=s_k}^{b_k} p(e_t|0,\sigma^2) \cdot \frac{1}{(b_k - s_k)} \cdot \frac{1}{H_K}$$
$$= \frac{(1-\gamma)}{H_K(b_k - s_k)} \left(\frac{1}{\sigma}\right)^{(b_k - s_k)} exp\left(-\frac{1}{2\sigma} \sum_{t=s_k}^{b_k} e_t^2\right)$$

Where $\frac{1}{(b_k - s_k)}$ is 1 over the length of the full interval (not counting the first point) to account for the fact that there are $(b_k - s_k)$ possible points that we could have chosen.

$$e_t \overset{iid}{\sim} \text{N(0,1)}$$

Insertion Decision rule by [Sadia et al., 2018]

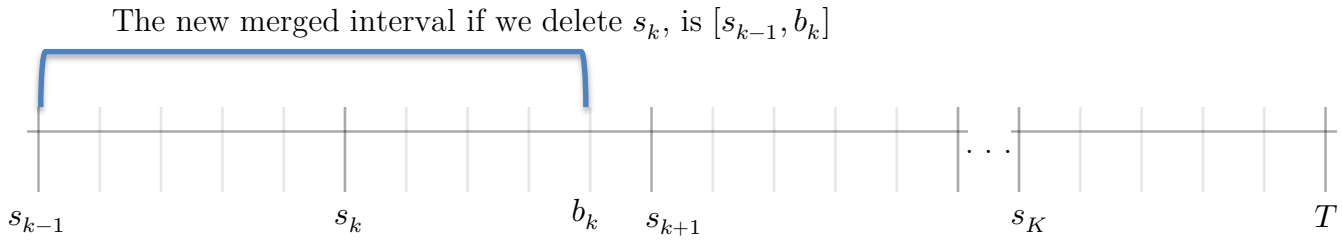
Randomly choose ρ from a uniform(0,1) distribution.

If $\rho < \frac{P_{insert}}{P_{insert} + P_{reject}}$, accept the insertion. Algebraically, this is equivalent to $\frac{1}{1 + \frac{P_{reject}}{P_{insert}}}$.

Note that the Metropolis-Hastings decision rule is somewhat different

Accept if $\frac{P_{inisert}}{P_{reject}} > 1$, otherwise accept with probability $\frac{P_{insert}}{P_{reject}}$. We ended up trying both methods. This is discussed in chapter 4.

Probability of Deletion



We sample a new group g_M, and sample a new segment mean c_M, where the "M" refers to the fact that we have merged two intervals. We update all of the error terms for the new segment.

$e_t^{(M)} = c_M - x_{s_{k-1}}$ for $t = s_{k-1}$

$e_t^{(M)} = c_M - (c_M + \phi(x_{t-1} - c_M) + \theta e_{t-1})$ for $s_{k-1} < t \leq b_k$

The length of the new interval, excluding the leftmost endpoint, is $\frac{1}{(b_k - s_{k-1})}$, which is the sum of the lengths of the individual intervals, excluding the leftmost endpoint of the first interval. The probability of a changepoint not occurring at s_k at any given time t is $(1-\gamma)$. The correction factor is $\frac{1}{H_{K-1}}$, because we have $K-1$ segments if we delete a point.

Original Model: Probability that we delete the proposed point is proportional to

$$P_{delete} \propto (1-\gamma) \prod_{t=s_{k-1}}^{b_k} p\left(\left(e_t^{(M)}\right)^2 \middle| 0, \sigma^2\right) \cdot \frac{1}{(b_k - s_{k-1})} \cdot \frac{1}{H_{K-1}}$$

$$\propto \frac{(1-\gamma)}{H_{K-1}} \frac{1}{(b_k - s_{k-1})} \cdot \left(\frac{1}{\sigma}\right)^{(b_k - s_{k-1})} exp\left(-\frac{1}{2\sigma^2} \sum_{t=s_{k-1}}^{b_k} \left(e_t^{(M)}\right)^2\right)$$

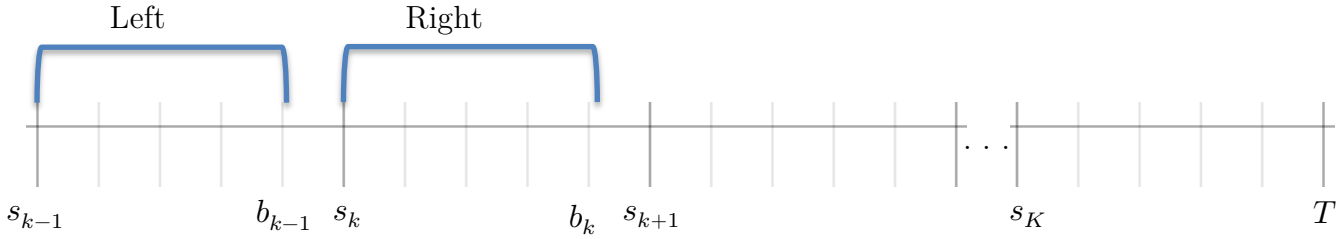
Our Model: Probability that we delete the proposed point is proportional to

$$P_{delete} \propto (1-\gamma) \prod_{t=s_{k-1}}^{b_k} p(e_t^M | 0, \sigma^2_{g_M}) \cdot \frac{1}{(b_k - s_{k-1})} \frac{1}{D_{K-1}}$$

$$\propto \frac{(1-\gamma)}{H_{K-1}} \frac{1}{(b_k - s_{k-1})} \left(\frac{1}{\sigma_{g_l}}\right)^{(b_k - s_{k-1}+1)} \left(\frac{1}{\sigma_{g_r}}\right)^{(b_k - s_k + 1)} exp\left(-\frac{1}{2}\left(\sum_{t=s_{k-1}}^{b_{k-1}} (e_t^M)^2 + \sum_{t=s_k}^{b_k} (e_t^M)^2\right)\right)$$

Probability of No Deletion

We will calculate the rejection probability based on the two existing intervals, one to the left of the proposed deletion point, the other to the right.
Our left and right intervals are $[s_{k-1,}, b_{k-1}], [s_k, b_k]$.



We do not generate new group assignments, segment means, or error terms because we already have those current values. The correction factor is $\frac{1}{H_K}$ because we do not change the number of segments, K. We are currently discussing the original model by [Sadia et al., 2018], in which the error variance is the same for all segments. This allows us to calculate the probability by multiplying over all terms within the left and right intervals.

Original Model: Probability that we do not delete

$$P_{reject} \propto \gamma \prod_{t=s_{k-1}}^{b_{k-1}} p\left(e_t^{(l)} \middle| 0, \sigma^2\right) \prod_{t=s_k}^{b_k} p\left(e_t^{(r)} \middle| 0, \sigma^2\right) \frac{1}{H_K}$$
$$= \gamma \prod_{t=s_{k-1}}^{b_k} p(e_t|0, \sigma^2) \frac{1}{H_K}$$

$$\propto \frac{\gamma}{H_K} \left(\frac{1}{\sigma}\right)^{b_k - s_{k-1}} exp\left(-\frac{1}{2\sigma^2} \sum_{t=s_{k-1}}^{b_k} e_t^2\right)$$

Our model: We account for the error variance corresponding to groups

$$P_{reject} \propto \gamma \prod_{t=s_{k-1}}^{b_{k-1}} p\left(e_t^{(l)} \middle| 0, \sigma_{g_l}^2\right) \prod_{t=s_k}^{b_k} p\left(e_t^{(r)} \middle| 0, \sigma_{g_r}^2\right) \frac{1}{H_K}$$

$$\propto \frac{\gamma}{H_K} \frac{1}{(b_k - s_{k-1})} \left(\frac{1}{\sigma_{g_l}}\right)^{(b_k - s_{k-1}+1)} \left(\frac{1}{\sigma_{g_r}}\right)^{(b_k - s_k+1)} exp\left(-\frac{1}{2}\left(\sum_{t=s_{k-1}}^{b_{k-1}} \frac{1}{\sigma_{g_l}^2} e_t^2 + \sum_{t=s_k}^{b_k} \frac{1}{\sigma_{g_r}^2} e_t^2\right)\right)$$

Summary Comparison of Insertion Probabilities

Original Model	Our Model
The parameters ϕ and θ are constant for both intervals. Left error terms $e_t^{(l)} = x_t - c_l$ for $t = s_k$ $e_t^{(l)} = x_t - \left(c_l + \phi(x_{t-1} - c_l) + \theta e_{t-1}^{(l)}\right)$ for $s_k < t \leq z-1$ Right error terms $e_t^{(r)} = x_t - c_r$ for $t = s_k$ $e_t^{(r)} = x_t - (c_r + \phi(x_{t-1} - c_l) + \theta e_{t-1})$ for $z < t \leq b_k$ Probability of accepting the insertion. The error variance term σ^2 is the same for each segment. $P_{insert} \propto \gamma \prod_{t=s_k}^{z-1} p\left(e_t^{(l)} \middle\| 0, \sigma^2\right) \prod_{t=z}^{b_k} p\left(e_t^{(r)} \middle\| 0, \sigma^2\right) \frac{1}{H_{K+1}}$ $= \frac{\gamma}{H_{K+1}} \left(\frac{1}{\sigma}\right)^{b_k - s_k + 1} \times$ $exp\left(-\frac{1}{2\sigma^2}\left(\sum_{t=s_k}^{z-1}\left(e_t^{(l)}\right)^2 + \sum_{t=z}^{b_k}\left(e_t^{(r)}\right)^2\right)\right)$ Probability Rejecting $P_{reject} \propto$ $\frac{(1-\gamma)}{H_K(b_k - s_k)}\left(\frac{1}{\sigma}\right)^{(b_k - s_k + 1)} exp\left(-\frac{1}{2\sigma}\sum_{s_{k-1}}^{b_k} e_t^2\right)$	The ARMA parameters must correspond to the groups. Left error terms $e_t^{(l)} = x_t - c_l$ for $t = s_k$ $e_t^{(l)} = x_t - \left(c_l + \phi_{g_l}(x_{t-1} - c_l) + \theta_{g_l} e_{t-1}^{(l)}\right)$ for $s_k < t \leq z-1$ Right error terms $e_t^{(r)} = x_t - c_r$ for $t = s_k$ $e_t^{(r)} = x_t - \left(c_r + \phi_{g_r}(x_{t-1} - c_l) + \theta_{g_r} e_{t-1}^{(r)}\right)$ for $z < t \leq b_k$ Probability of accepting the insertion: The error variance terms differ based on the segment groups. $P_{in} \propto \gamma \prod_{t=s_k}^{z-1} p\left(e_t^{(l)} \middle\| 0, \sigma_{g_l}^2\right) \prod_{t=z}^{b_k} p\left(e_t^{(r)} \middle\| 0, \sigma_{g_r}^2\right) \frac{1}{H_{K+1}}$ $= \frac{\gamma}{H_{K+1}} \left(\frac{1}{\sigma_{g_l}}\right)^{z - s_k} \left(\frac{1}{\sigma_{g_r}}\right)^{b_k - z + 1} \times$ $exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_{g_l}^2}\sum_{t=s_k}^{z-1}\left(e_t^{(l)}\right)^2 + \frac{1}{\sigma_{g_r}^2}\sum_{t=z}^{b_k}\left(e_t^{(r)}\right)^2\right)\right)$ Probability of Rejecting Insertion $P_{reject} \propto$ $\frac{(1-\gamma)}{H_K(b_k - s_k)}\left(\frac{1}{\sigma_{g_k}}\right)^{(b_k - s_k + 1)}$ $\times exp\left(-\frac{1}{2}\frac{1}{\sigma_{g_k}^2}\sum_{t=s_k}^{b_k - 1} e_t^2\right)$

Table 2.2.3 Summary of Comparison of Insertion Probabilities

Summary Comparison of Deletion Probabilities

Original Model	Our Model
When we merge the segments, we will use the same values ϕ and θ, regardless of the group the merged segment is assigned to. $e_t^M = c_M - x_{s_{k-1}}$ for $t = s_{k-1}$ $e_t^M = c_M - (x_t + \phi(x_{t-1} - c_M) + \theta e_{t-1})$ for $s_{k-1} < t \le b_k$ The error variance σ^2 does not depend on group. The probability of accepting the deletion is proportional to: $P_1 \propto (1-\gamma) \prod_{t=s_{k-1}}^{b_k} p(e_t^M \mid 0, \sigma^2) \cdot \frac{1}{(b_k - s_{k-1})} \cdot \frac{1}{H_{K-1}}$ $\frac{(1-\gamma)}{D_{K-1}} \frac{1}{(b_k - s_{k-1})} \left(\frac{1}{\sigma}\right)^{(b_k - s_{k-1}+1)}$ $exp\left(-\frac{1}{2\sigma^2} \sum_{t=s_{k-1}}^{b_k} (e_t^M)^2\right)$ **Probability of rejecting the deletion** Taking the product of terms over the left and right segments is the same as taking the product over the merged segment, with no change in σ^2. $P_0 \propto \gamma \prod_{t=s_{k-1}}^{b_k} p(e_t \mid 0, \sigma^2) \frac{1}{H_K}$ $\propto \frac{\gamma}{H_K} \left(\frac{1}{\sigma}\right)^{b_k - s_{k-1}+1} exp\left(-\frac{1}{2\sigma^2} \sum_{t=s_{k-1}}^{b_k} e_t^2\right)$	We use the ARMA parameters associated with the group that the merged segment is randomly assigned to. $e_t^M = c_M - x_{s_{k-1}}$ for $t = s_{k-1}$ $e_t^M = c_M - \left(x_t + \phi_{g_M}(x_{t-1} - c_M) + \theta_{g_M} e_{t-1}\right)$ for $s_{k-1} < t \le b_k$ We must account for the specific error variance for the group of the merged segment. The probability of accepting the deletion is proportional to $P_1 \propto (1-\gamma) \prod_{t=s_{k-1}}^{b_k} p\left(e_t^M \mid 0, \sigma^2_{g_M}\right) \cdot \frac{1}{(b_k - s_{k-1})} \frac{1}{H_{K-1}}$ $\propto \frac{(1-\gamma)}{D_{K-1}} \frac{1}{(b_k - s_{k-1})} \left(\frac{1}{\sigma_{g_l}}\right)^{(b_k - s_{k-1}+1)} \left(\frac{1}{\sigma_{g_r}}\right)^{(b_k - s_k+1)}$ $\exp\left(-\frac{1}{2}\left(\sum_{t=s_{k-1}}^{b_{k-1}} (e_t^M)^2 + \sum_{t=s_k}^{b_k} (e_t^M)^2\right)\right)$ **Probability of rejecting the deletion** Taking the product of terms over the left and right segments requires that we keep track of the difference between the error variance in the left and right segments. $P_0 \propto \gamma \prod_{t=s_{k-1}}^{b_{k-1}} p\left(e_t^{(l)} \mid 0, \sigma^2_{g_l}\right) \prod_{t=s_k}^{b_k} p\left(e_t^{(r)} \mid 0, \sigma^2_{g_r}\right) \frac{1}{H_K}$ $\propto \frac{\gamma}{H_K} \frac{1}{(b_k - s_{k-1})} \left(\frac{1}{\sigma_{g_l}}\right)^{(b_{k-1} - s_{k-1}+1)} \left(\frac{1}{\sigma_{g_r}}\right)^{(b_k - s_k+1)}$ $exp\left(-\frac{1}{2}\left(\sum_{t=s_{k-1}}^{b_{k-1}} \frac{1}{\sigma^2_{g_l}} e_t^2 + \sum_{t=s_k}^{b_k} \frac{1}{\sigma^2_{g_r}} e_t^2\right)\right)$

Table 2.2.4 Summary of Comparison of Deletion Probabilities

The Gibbs Sampler for $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\mu}, \gamma$

The vectors $\boldsymbol{s}$, $\boldsymbol{c}$, and $\boldsymbol{g}$ must be updated many times during the insertion and deletion procedures, and this will vary widely based on iteration step as part of our Generalized Gibbs sampling procedure. However, after we perform the insertion and deletion steps, we use the Gibbs sampling procedure, as described in 1.2.6, to update the remaining parameter vector $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\mu}$, and the parameter γ. The full conditional posterior distributions are shown in table 2.2.2 and mathematically developed in chapter 3. The proofs of our claims of truncated-normality of the full conditional distributions for the ARMA parameters are given in great detail (Theorems 3.1.1 through 3.1.4). The order in which we update the parameters $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\pi}, \boldsymbol{\mu}$ and γ is up to the individual implementing the algorithm. Finally, the error vector $e = (e_1, \ldots, e_T)$ must be updated any time a parameter that affects the estimated error terms is updated. For our version of the multi-group model, the error vector must be updated whenever $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{g}$ and $\boldsymbol{c}$ are updated. The error vector will also be updated every time an insertion or deletion occurs because estimated segment mean is updated in each case.

2.2.3 Exception Handling

Rounding Error

When calculating acceptance and rejection probabilities on relatively long intervals, the products used to produce the acceptance and rejection probabilities tend toward zero, and in some cases go beyond machine rounding capability. This can result in pathological over-acceptance rate for insertion and under-deletion rate. We can avoid this with log transformations. For example, on an interval $I_k = [s_k, b_k]$, recall that the probability of rejecting a proposed insertion point is proportional to:

$$P_{reject} \propto \frac{(1-\gamma)}{H_K(b_k - s_k)} \left(\frac{1}{\sigma_{g_k}}\right)^{(b_k - s_k + 1)} exp\left(-\frac{1}{2}\frac{1}{\sigma^2_{g_k}} \sum_{t=s_k}^{b_k} e_t^2\right)$$

On a relatively long interval, for example, an interval of length over 100 points, which is not uncommon for large data sets, we have over 100 error terms of the form

$$c_k - (x_t + \phi_{g_k}(x_{t-1} - c_M) + \theta_{g_k} e_{t-1})$$

The sum of squares of these terms can vary widely, given how much variation in our estimates for c_k, ϕ, and θ we may have on any given iteration. This can make the exponential term easily tend close enough to zero for machine rounding error to simply interpret this value as zero. This would be fine were this the actual probability of rejecting the insertion, but this is only proportional to the probability of rejection, relative the term that we use for acceptance of probability.

If we consider that the probability of accepting an insertion point is proportional to

$$P_{insert} \propto \gamma \prod_{t=s_k}^{z-1} p\left(e_t^{(l)} \middle| 0, \sigma^2_{g_l}\right) \prod_{t=z}^{b_k} p\left(e_t^{(r)} \middle| 0, \sigma^2_{g_r}\right) \frac{1}{H_{K+1}}$$

A term that could possibly be even smaller than the $P_{rejecet}$ term, we could easily end up with

Probability of Insertion $\frac{P_{insert}}{P_{insert}+P_{reject}}$ being interpreted as zero over zero. This will be interpreted as '*NaN*' or 'Not a number,' by the machine.

Deletion steps can be even more problematic, because a deletion step involves evaluating probabilities over a merged interval, resulting in many more error terms to handle. In several simulations, we discovered that up to one fourth of the time, the machine interpreted our probabilities as '*NaN*.' This can result in over-acceptance of insertions, and over-rejection of proposed deletions. At the end of a simulation, we may have what appear to be far too many detected changepoints. We can prevent this from happening by taking the natural log of these terms, converting a long string of products into a much more manageable collection of summations

For example:

$$\log(P_{reject}) \propto$$
$$\log(1-\gamma) - \log(b_k - s_k) - \log(H_K) - (b_k - s_k)\log(\sigma_{g_k}) - \frac{1}{2\sigma^2_{g_k}} \sum_{t=s_{k-1}}^{b_k} e_t^2$$

Not only does this prevent extreme rounding error, but it also decreases computation time. The calculations of the other terms for the deletion steps are analogous and we will not go into further detail here. But it is an important realization of ours that rounding error must be monitored closely in our setting.

Chapter 3 Full Conditional Distributions

In order to solidify our ability to apply the Gibbs sampler without additional approximation methods, we want to be able to directly utilize all of the full conditional posterior distributions for our parameters. Equations that symbolically described the structure of these distributions are presented by [Sadia et al., 2018]. However, closed-form solutions are not provided for the ϕ and θ, nor c. Without such solutions, a computationally intensive approximation method called the slice-sampler is said to be used. This indirect sampling may be effective, but we would like to be able to sample directly from the full conditional posterior distributions when applying the generalized Gibbs sampling procedure. We resolve this issue by providing closed-form identifiable probability distributions for each parameter, showing ϕ and θ to be Truncated Normal under their model assumptions. We will then use these results to prove that the full conditional posterior distributions for the AR and MA parameters present in our multi-variable model are also Truncated Normal. This allows us to directly utilize the generalized Gibbs sampling procedure to estimate all parameters. Finally, we will provide full conditional posterior distributions for all remaining model parameters, showing that we can directly use the generalized Gibbs sampling for all remaining parameters of our model.

3.1 ARMA Parameters Full Conditional Posterior Distributions

Suppose that the data $\boldsymbol{x} = \{x_1, x_2, ... x_T\}$ where T is the number of observations are observed on K segments such that data between segments are independent, and data within segments are generated by a stationary and invertible ARMA process with autoregressive parameter ϕ, moving average parameter θ, with error terms $e_t \sim^{iid} \mathrm{N}(0, \sigma^2)$. Now assume that for each segment, k, the data are centered about a segment mean denoted by c_k. We will denote the set of K segments, or intervals, on which the data are independent between segments, as follows:

$\boldsymbol{I} = \{I_1, I_2, ..., I_K\}$, where:
$I_1 = [1 = s_1, b_1]$
$I_k = [s_k, b_k = s_{k+1} - 1]$ for $1 < k < K - 1$

$I_K = [s_K, b_K = T]$

The points $\boldsymbol{s} = (s_1, ..., s_k)$ are considered changepoints, or points at which each independent segment begins and $\boldsymbol{b} = (b_1, ..., b_k)$ represent the right endpoints.
Then we can express the ARMA process on segment k by

$X_{s_k} = c_k + e_{s_k}$ The very first observation does not involve the AR or MA parameters.

$X_t = c_k + e_t + \phi(x_{t-1} - c_k) + \theta e_{t-1}$ for $(s_k + 1) \leq t \leq b_k$

Define prior distributions on the model parameters, as follows:

$\phi \sim U(-1,1)$

$\theta \sim U(-1,1)$

$\sigma^2 \sim IG(\alpha, \beta)$ Where α and β are constants greater than 1, chosen based on what is appropriate for a particular application of interest.

Theorem 3.1.1: The full conditional posterior distribution of the moving average parameter, ϕ, is truncated normal on the domain $-1 < \phi < 1$, generated over a normal distribution with mean

$$\omega = \frac{\sum_{k=1}^{K} \sum_{t=s_k-1}^{b_k} (x_{t-1} - c_k)(x_t - c_k - \theta e_{t-1})}{\sum_{k=1}^{K} \sum_{t=s_k+1}^{b_k} (x_{t-1} - c_k)^2} \tag{3.1}$$

and variance

$$\nu^2 = \frac{\sigma^2}{\sum_{k=1}^{K} \sum_{t=s_k+1}^{b_k} (x_{t-1} - c_k)^2} \tag{3.2}$$

In more compact notation:

$$\phi | \cdot \sim TN(\omega,\ \nu^2, -1, 1) \tag{3.3}$$

Proof of Theorem 3.1

Given segment k, the error terms are as follows

$e_{s_k} = x_t - c_k$ for $t = s_k$, because the first term in the sequence does not contain any past information on which ϕ and θ have influence, and

$e_t = x_t - c_k - \phi(x_{t-1} - c_k) - \theta e_{t-1}$ for $t > s_k$.

Denote the vectors of data, segment means, and error terms in bold and italicized form, as follow:

$\boldsymbol{c} = \{c_1,\ c_2,\ \ldots c_K\}$ where K represents the number of independent segments

$\boldsymbol{e} = (e_1, \ldots,\ e_t)$

We will first derive the likelihood function, $f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \phi, \theta, \sigma^2)$.

We know that $e_t \overset{iid}{\sim} N(0, \sigma^2)$. By equation (2.9) we have that

$$f(e_t) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2} e_t^2\right\}$$

Now if we express this in terms of $\boldsymbol{x}, \boldsymbol{c}, \phi, \theta$, and ignoring $\frac{1}{\sqrt{2\pi\sigma^2}}$ as a proportionality constant

$f(e_t|\boldsymbol{x}, c_k, \phi, \theta, \sigma^2) \propto exp\{-\frac{1}{2\sigma^2}(x_t - c_k)^2\}$ at the left endpoints (change points)

$$f(e_t|\boldsymbol{x}, c_k, \phi, \theta, \sigma^2) \propto exp\left\{-\frac{1}{2\sigma^2}(x_t - c_k - \phi(x_{t-1} - c_k) - \theta e_{t-1})^2\right\}\ s_k + 1 < t \leq b_k \tag{3.4}$$

By independence of the error terms, we have

$$\begin{aligned} f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \phi, \theta, \sigma^2\ ) &= \textstyle\prod_{t=1}^{T} f(e_t|x_t, c_k, \phi, \theta, \sigma^2) \\ &= \prod_{k=1}^{K} \prod_{t=s_k}^{b_k} f(e_t|x_t, c_k, \phi, \theta, \sigma^2) \end{aligned}$$

And by (3.4), the above expression can be written as:

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2) \propto \prod_{k-1}^{K} exp\left\{-\frac{1}{2\sigma^2}(x_t - c_k)^2\right\} \prod_{s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma^2}(x_t - c_k - \phi(x_{t-1} - c_k) - \theta e_{t-1})^2\right\}$$

We notice that $exp\{-\frac{1}{2\sigma^2}(x_t - c_k)^2\}$ does not depend on ϕ, and so we can ignore it as a proportionality constant. This essentially allows us to disregard the left endpoints of all segments. This makes sense because if we have a segment with only one point, we do not have any information about the parameter of interest.

Our joint distribution for the error terms can now be expressed as follows:

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2\ ) \ \propto \prod_{k-1}^{K} \prod_{s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma^2}(x_t - c_k - \phi(x_{t-1} - c_k) - \theta e_{t-1})^2\right\} \tag{3.5}$$

The likelihood function will be used to derive the full conditional distribution:

$$p(\phi|\boldsymbol{x},\boldsymbol{e},\boldsymbol{c},\theta,\sigma^2) = \frac{f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2\ )p(\phi)}{\int f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2\ )p(\phi)d\boldsymbol{e}} \tag{3.6}$$

The denominator can be omitted throughout our derivation, as it can be thought of as a proportionality constant.

Now let us consider the exponential term in the product shown in equation 3.5.

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2\ ) \prod_{k=1}^{K} \prod_{t=s_k+1}^{b_k} \underbrace{exp\{-\tfrac{1}{2\sigma^2}(x_t - c_k - \phi(x_{t-1} - c_k) - \theta e_{t-1})^2\}}$$

The term within the exponential function can be rearranged in such a way that allows us to later omit additional terms that do not involve ϕ.

$$exp\left\{-\frac{1}{2\sigma^2}(x_t - c_k - \phi(x_{t-1} - c_k) - \theta e_{t-1})^2\right\}$$

$$= exp\left\{-\frac{1}{2\sigma^2}\big(\phi(x_{t-1} - c_k) - (x_t - c_k - \theta e_{t-1})\big)^2\right\} \tag{3.7}$$

We can expand the inner terms and separate the above expression into the following two exponential factors:

$$exp\left\{-\frac{1}{2\sigma^2}\big(\phi^2(x_{t-1} - c_k)^2 - 2\phi(x_{t-1} - c_k)(x_t - c_k - \theta e_{t-1})\big)\right\} \times exp\left\{-\frac{1}{2\sigma^2}(x_t - c_k - \theta e_{t-1})^2\right\} \tag{3.8}$$

The right-most term does not involve ϕ and can therefore be ignored as a proportionality constant. Then (3.7) is proportional to

$$exp\left\{-\frac{1}{2\sigma^2}\big(\phi^2(x_{t-1} - c_k)^2 - 2\phi(x_{t-1} - c_k)(x_t - c_k - \theta e_{t-1})\big)\right\}$$

Equation (3.5) now becomes

$$f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \phi, \theta, \sigma^2) \propto \prod_{k=1}^{K} \prod_{t=s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma^2}\phi^2(x_{t-1} - c_k)^2 - 2\phi(x_{t-1} - c_k)(x_t - c_k - \theta e_{t-1})\right\} \tag{3.9}$$

This is algebraically equivalent to

$$f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \phi, \theta, \sigma^2) \propto -\frac{1}{2\sigma^2} exp\left\{\sum_{k=1}^{K} \sum_{t=s_k-1}^{b_k} \big((x_{t-1} - c_k)^2 - 2\phi(x_{t-1} - c_k)(x_t - c_k - \theta e_{t-1})\big)\right\} \tag{3.10}$$

By factoring out $-\frac{\phi^2}{2\sigma^2}$, the double summation can be written as two separate summations, as follows

$$exp\left\{-\frac{1}{2\sigma^2}\phi^2\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)^2+\frac{\phi}{\sigma^2}\sum_{k=1}^{K}\sum_{t=s_k-1}^{b_k}(x_{t-1}-c_k)(x_t-c_k-\theta e_{t-1})\right\}.$$

Let $h=\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)^2$ and $z=\sum_{k=1}^{K}\sum_{t=s_k-1}^{b_k}(x_{t-1}-c_k)(x_t-c_k-\theta e_{t-1})$, giving us

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2)\propto exp\left\{-\frac{\phi^2}{2\sigma^2}h+\frac{\phi}{\sigma^2}z\right\} \tag{3.11}$$

With some further algebraic manipulation, we have

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2)\propto exp\left\{-\frac{\phi^2}{2\sigma^2}h+\frac{h}{h}\frac{\phi}{\sigma^2}z\right\}$$

$$=exp\left\{-\frac{\phi^2}{2\left(\frac{\sigma^2}{h}\right)}+\frac{1}{h}\frac{\phi}{\left(\frac{\sigma^2}{h}\right)}z\right\}.$$

We can further abbreviate our expression by $\nu^2=\frac{\sigma^2}{h}$, and now have:

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2)\propto exp\left\{-\frac{1}{2\nu^2}\phi^2+\frac{1}{\nu^2}\left(\frac{1}{h}\right)\phi z\right\}$$

$$=exp\left\{-\frac{1}{2\nu^2}\left(\phi^2-\frac{2z}{h}\phi\right)\right\} \tag{3.12}$$

By completing the square, (3.12) becomes

$$exp\left\{-\frac{1}{2\nu^2}\left(\phi^2-\frac{2z}{h}\phi+\left(\frac{z}{h}\right)^2-\left(\frac{z}{h}\right)^2\right)\right\}$$

$$=exp\left\{-\frac{1}{2\nu^2}\left[\left(\phi-\frac{z}{h}\right)^2-\left(\frac{z}{h}\right)^2\right]\right\}$$

$$=exp\left\{-\frac{1}{2\nu^2}\left(\phi-\frac{z}{h}\right)^2\right\}\ exp\left\{\frac{1}{2\nu^2}\left(\frac{z}{h}\right)^2\right\}.$$

The factor on the right does not involve ϕ and can thus be ignored as a proportionality constant.

We now have that:

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2)\propto exp\left\{-\frac{1}{2\nu^2}(\phi-\omega)^2\right\}\ \text{with}\,\omega=\frac{z}{h}$$

We see that this is the kernel of normal pdf in ϕ with mean of $\omega=\frac{z}{h}$ and variance ν^2.

Recall that ϕ~$U(-1,1)$. Then by equation (3.6), full conditional distribution function can now be expressed as

$$p(\phi|\boldsymbol{x},\boldsymbol{e},\boldsymbol{c},\theta,\sigma^2)\propto exp\left\{-\frac{1}{2\nu^2}(\phi-\omega)^2\right\}I(\phi)_{(-1,1)} \tag{3.13}$$

We have the product of the kernel of the density of a normally distributed random variable and the density of a uniform random variable, giving us the form of a truncated normal distribution. Thus, we have confirmed (3.1), (3.2), and (3.3).

$$\phi|\cdot \sim TN(\omega,\ \nu^2, -1,1)$$

$$\omega = \frac{z}{h} = \frac{\sum_{k=1}^{K}\sum_{t=s_k-1}^{b_k}(x_{t-1}-c_k)(x_t - c_k - \theta e_{t-1})}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)^2}$$

$$\nu^2 = \frac{\sigma^2}{h} = \frac{\sigma^2}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)^2}$$

■

Theorem 3.1.2: The full conditional posterior distribution of the moving average parameter, θ, is truncated normal on the domain $-1 < \theta < 1$, generated over a normal distribution with mean

$$\xi = \frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}\big(x_t - c_k - \phi(x_{t-1}-c_k)\big)}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2} \tag{3.14}$$

and variance

$$\eta^2 = \frac{\sigma^2}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2} \tag{3.15}$$

In more compact notation,

$$\theta|\cdot\ \sim TN(\xi,\ \eta^2, -1,1) \tag{3.16}$$

Proof of Theorem 3.1.2

The process is similar to that of the proof of theorem 3.1.1.

The full conditional distribution for θ is

$$f(\theta|\boldsymbol{x},\boldsymbol{e},\boldsymbol{c},\phi,\sigma^2) = \frac{f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2)p(\theta)}{\int f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2\ )p(\theta)d\boldsymbol{e}}$$

$$\propto f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\sigma^2)p(\theta)$$

$$= \ f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\sigma^2)I(\theta)_{(-1,1)} \tag{3.17}$$

By equation (3.6), the likelihood function is given by

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2\ ) \ \propto \prod_{k=1}^{K}\prod_{t=s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma^2}(x_t - c_k - \phi(x_{t-1}-c_k) - \theta e_{t-1})^2\right\} \tag{3.18}$$

The exponential term can be re-expressed by rearranging terms:

$$exp\left\{-\frac{1}{2\sigma^2}(-\theta e_{t-1} - \phi(x_{t-1}-c_k) + x_t - c_k)^2\right\}$$

$$= exp\left\{-\frac{1}{2\sigma^2}(\theta e_{t-1} + \phi(x_{t-1}-c_k) - x_t + c_k)^2\right\}$$

$$= exp\left\{-\frac{1}{2\sigma^2}(\theta^2 e_{t-1}^2 + 2(\theta e_{t-1})(\phi(x_{t-1}-c_k) - x_t + c_k) + (\phi(x_{t-1}-c_k) - x_t + c_k)^2)\right\}$$

$$= exp\left\{-\frac{1}{2\sigma^2}\left(\theta^2 e_{t-1}^2 + 2e_{t-1}(\phi(x_{t-1}-c_k) - x_t + c_k)\right)\right\} exp\left\{-\frac{1}{2\sigma^2}(\phi(x_{t-1}-c_k) - x_t + c_k)^2\right\}$$

We have separated our expression into two factors. The term on the right does not depend on θ and can thus be discarded as a proportionality constant.

Our likelihood function can now be expressed, as

$$f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \phi, \theta, \sigma^2\ ) \propto \prod_{k=1}^{K}\prod_{t=s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma^2}\left(\theta^2 e_{t-1}^2 + 2(\theta e_{t-1})(\phi(x_{t-1}-c_k) - x_t + c_k)\right)\right\},$$

which is algebraically equivalent to

$$f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \phi, \theta, \sigma^2) \propto exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} \theta^2 e_{t-1}^2 + 2\theta e_{t-1}(\phi(x_{t-1}-c_k) - x_t + c_k)\right\} \tag{3.19}$$

We can separate the inner double sum, as follows

$$\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} \theta^2 e_{t-1}^2 + \sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} 2\theta e_{t-1}(\phi(x_{t-1}-c_k) - x_t + c_k)$$

We see that θ does not depend on t, allowing us to re-express the above as

$$\theta^2 \sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2 \ \ + \ \ 2\theta \sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}(\phi(x_{t-1}-c_k) - x_t + c_k)$$

Now if we factor out $\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2$, we get a quadratic expression in θ

$$\left(\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2\right)\left(\theta^2 + 2\theta\frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}(\phi(x_{t-1}-c_k) - x_t + c_k)}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}\right)$$

By completing the square, the above expression becomes

$$\left(\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2\right)\left\{\left(\theta-\frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}\big(x_t-c_k-\phi(x_{t-1}-c_k)\big)}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}\right)^2\right.$$
$$\left.-\left(\frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}\big(x_t-c_k-\phi(x_{t-1}-c_k)\big)}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}\right)^2\right\}$$

Notice the negative sign that we now have because we have rearranged the terms in the upper sum, as this gives us a more convenient form to work with. The last term in the above summation occurs within the exponential function of equation (3.15). This term does not depend on θ and so it can be excluded as a proportionality constant.

By equation (3.5), we have now shown that the likelihood function can be put into the following form

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2)\propto exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2\right)\left(\theta-\frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}\big(x_t-c_k-\phi(x_{t-1}-c_k)\big)}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}\right)^2\right\} \tag{3.20}$$

Now set

$$\xi=\frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}\big(x_t-c_k-\phi(x_{t-1}-c_k)\big)}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}$$

$$\eta^2=\frac{\sigma^2}{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}$$

Now we have that the likelihood function is $f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi,\theta,\sigma^2)\propto exp\Big\{-\frac{1}{2\eta^2}(\theta-\xi)^2\Big\}$, and so by equation 3.14, the full posterior conditional distribution function can be expressed as follows

$$p(\theta|\boldsymbol{x},\boldsymbol{e},\boldsymbol{c},\theta,\sigma^2)\propto exp\left\{-\frac{1}{2\eta^2}(\theta-\xi)^2\right\}I(\theta)_{(-1,1)} \tag{3.21}$$

This is the kernel of a truncated normal distribution function with parameters ξ and η^2 on the domain $-1<\theta<1$. Thus, we have shown that the full conditional posterior distribution of θ is

$$TN(\xi,\eta^2,-1,1).$$

■

Generalized N-Dimensional Group Case

Suppose we are dealing with segments on which variance of data are unequal. We may be able analyze segments such that the data within segments correspond to an ARMA processes that can be described by groups, as discussed in chapter 1.

Given a T observations $\{x_1, ... x_T\}$ observed on K segments $\boldsymbol{I} = \{I_1, I_2, ..., I_K\}$, where:

$I_1 = [1 = s_1, b_1]$

$I_k = [s_k, b_k = s_{k+1} - 1]$ for $1 < k < K - 1$

$I_K = [s_K, b_k = T]$

$\boldsymbol{s} = (s_1, ..., s_k)$ are the left endpoints and

$\boldsymbol{b} = (b_1, ..., b_k)$ represent the right endpoints.

Suppose that we have N groups, as described in chapter 1. On any given segment, the data are assumed to originate from an ARMA process associated with one and only one group.

Define the sets $\boldsymbol{G_n}$ to represent indices of all segments associated with group n. For example, if the data from segments I_1, I_5, I_{11} originate from group n, and no other segment is associated with group n, then $\boldsymbol{G_n} = \{1,5,11\}$.

Let $\boldsymbol{g} = (g_1, ..., g_k)$ be the set of group numbers associated with each segment. For example, if segment I_1 data originated from a group 3 ARMA process, $g_1 = 3$, etc.

We can express the ARMA process on any given segment I_k by

$X_{s_k} = c_k + e_{s_k}$ The very first observation does not involve the AR or MA parameters.

$X_t = c_k + e_t + \phi_{g_k}(x_{t-1} - c_k) + \theta_{g_k} e_{t-1}$ for $(s_k + 1) \leq t \leq b_k$.

Let $\boldsymbol{c} = \{c_1,\ c_2,\ \ldots c_K\}$ be the segment means and $\boldsymbol{e} = (e_1, \ldots,\ e_t) \overset{iid}{\sim} N(0, \sigma^2_{g_k})$ the vector of error terms.

Suppose the prior distributions on all AR and MA

$\phi_{g_k} \sim U(-1,1)$ $\theta_{g_k} \sim U(-1,1)$

Theorem 3.1.3: Under the above conditions, given K segments and N possible groups, the full posterior distribution of the parameter ϕ_n associated with any group n is truncated normal of the form

$$\phi_n | \cdot \sim TN(\omega_n,\ \nu^2_n, -1,1) \tag{3.22}$$

$$\omega_n = \frac{\sum_{k \in G_n} \sum_{t=s_k-1}^{b_k} (x_{t-1} - c_k)(x_t - c_k - \theta_n e_{t-1})}{\sum_{k \in G_n} \sum_{t=s_k+1}^{b_k} (x_{t-1} - c_k)^2} \tag{3.23}$$

$$\nu^2_n = \frac{\sigma^2_n}{\sum_{k \in G_n} \sum_{t=s_k+1}^{b_k} (x_{t-1} - c_k)^2} \tag{3.24}$$

Proof of Theorem 3.1.3

Consider an arbitrary group n from N possible groups.

By equation (3.4)

$$f(e_t|\boldsymbol{x}, c_k, \phi_n, \theta_n, \sigma_n^2, g_k = n) \propto exp\left\{-\frac{1}{2\sigma_n^2}(x_t - c_k - \phi_n(x_{t-1} - c_k) - \theta_n e_{t-1})^2\right\},$$

and by independence of the error terms, we get the same form for the joint distribution as the form we see in equation (3.5).

$$f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}) \propto \prod_{k-1}^{K} \prod_{s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma_n^2}(x_t - c_k - \phi_{g_k}(x_{t-1} - c_k) - \theta_{g_k} e_{t-1})^2\right\} \tag{3.25}$$

By rearranging the order in which we multiply terms, we can express the above in a form that has one product of terms associated with group n, and another product of all terms not associated with group n.

$$f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}) \propto \prod_{k \in G_n} \prod_{t=s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma_n^2}(x_t - c_k - \phi_{g_k}(x_{t-1} - c_k) - \theta_n e_{t-1})^2\right\} \times$$
$$\prod_{k \notin G_n} \prod_{t=s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma_{g_k}^2}(x_t - c_k - \phi_{g_k}(x_{t-1} - c_k) - \theta_{g_k} e_{t-1})^2\right\}$$

Now the product of all terms not involving group n can be considered a proportionality constant, allowing us to now express the joint distribution of error terms, as follows:

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi_n,\theta_n,\sigma_n^2) \propto \prod_{k\in G_n}\prod_{t=s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma_n^2}(x_t-c_k-\phi_n(x_{t-1}-c_k)-\theta_n e_{t-1})^2\right\}$$

By equation (3.8), this is algebraically equivalent to

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi_n,\theta_n,\sigma_n^2) \propto$$

$$exp\left(\sum_{k\in G_n}\sum_{t=s_k-1}^{b_k} -\frac{1}{2\sigma_n^2}\Big((x_{t-1}-c_k)^2-2\phi_n(x_{t-1}-c_k)(x_t-c_k-\theta_n e_{t-1})\Big)\right) \tag{3.26}$$

We can factor out $-\frac{\phi_n}{2\sigma_n^2}$ and re-express the exponential term as

$$exp\left\{-\frac{1}{2\sigma^2}\phi_n^2\left(\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)^2+\frac{\phi_n}{\sigma_n^2}\sum_{k\in G_n}\sum_{t=s_k-1}^{b_k}(x_{t-1}-c_k)(x_t-c_k-\theta_n e_{t-1})\right)\right\}$$

Let $h_n=\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)^2$ and $z_n=\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)(x_t-c_k-\theta\epsilon_{t-1})$

Then

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi_n,\theta_n,\sigma_n^2) \propto exp\left\{-\frac{\phi_n^2}{2\sigma_n^2}h_n+\frac{\phi_n}{\sigma_n^2}z_n\right\} \tag{3.27}$$

$$= exp\left\{-\frac{\phi_n^2}{2\left(\frac{\sigma_n^2}{h_n}\right)}+\frac{1}{h}\frac{\phi_n}{\left(\frac{\sigma_n^2}{h_n}\right)}z_n\right\}$$

$$= exp\left\{-\frac{1}{2\nu_n^2}\phi_n^2+\frac{1}{\nu_n^2}\left(\frac{z_n}{h_n}\right)\phi_n\right\} \text{ where } \nu_n^2=\frac{\sigma_n^2}{h_n}$$

$$= exp\left\{-\frac{1}{2\nu_n^2}(\phi_n^2-2\omega_n\phi_n)\right\} \text{ where } \omega_n=\frac{z_n}{h_n}$$

$$= exp\left\{-\frac{1}{2\nu_n^2}[(\phi_n-\omega_n)^2-(\omega_n)^2]\right\} \text{ by completing the square}$$

$$= exp\left\{-\frac{1}{2\nu_n^2}[(\phi_n-\omega_n)^2]\right\} exp\{-(\omega_n)^2\}$$

Now the term on the right can be ignored as a proportionality constant. As a result, we see that the likelihood function has the norm of a Truncated Normal pdf.

$$p(\phi_n|\boldsymbol{x},\boldsymbol{e},\boldsymbol{c},\theta_n,\sigma_n^2) \propto f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi_n,\theta_n,\sigma_n^2)I(\phi_n)_{(-1,1)}$$
$$\propto exp\left\{-\frac{1}{2\nu_n^2}(\phi_n-\omega_n)^2\right\}I(\phi_n)_{(-1,1)}$$

That is

$$\phi_n|\cdot \sim TN(\omega_n,\ \nu_n^2,-1,1) \tag{3.28}$$

$$\omega_n = \frac{\sum_{k\in G_n}\sum_{t=s_k-1}^{b_k}(x_{t-1}-c_k)(x_t-c_k-\theta_n e_{t-1})}{\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)^2} \tag{3.29}$$

$$\nu_n^2 = \frac{\sigma_n^2}{\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k}(x_{t-1}-c_k)^2} \tag{3.30}$$

Since group n was chosen arbitrarily, we have shown that this is true for all groups.

■

Theorem 3.1.4: Given K segments and N possible groups, the full conditional posterior distribution of the parameter θ_n associated with any group n is truncated normal of the form

$$\theta_n|\cdot \sim TN(\xi_n,\ \eta_n^2,-1,1) \tag{3.31}$$

$$\xi_n = \frac{\sum_{k\in G_n}\sum_{t=s_k-1}^{b_k} e_{t-1}\Big(x_t-c_k-\phi_n(x_{t-1}-c_k)\Big)}{\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2} \tag{3.32}$$

$$\eta_n^2 = \frac{\sigma_n^2}{\sum_{k \in G_n} \sum_{t=s_k+1}^{b_k} e_{t-1}^2} \tag{3.33}$$

Proof of Theorem 3.1.4

The full conditional distribution for θ is

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{c}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}) = \frac{f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}) p(\boldsymbol{\theta})}{\int f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}\ ) p(\boldsymbol{\theta}) d\boldsymbol{e}} \tag{3.34}$$

As we saw in the proof of Theorem 3.3, we can express the likelihood function as a product of individual distribution functions such that the terms on the left all have to do with group n, and the terms on the right are not associated with group n.

$$f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\sigma}) \propto \prod_{k \in G_n} \prod_{t=s_k+1}^{b_k} f(e_t|\boldsymbol{x}, c_k, \phi_n, \theta_n, \sigma_n^2) \prod_{k \notin G_n} \prod_{t=s_k+1}^{b_k} f(e_t|\boldsymbol{x}, c_k, \phi_{g_k}, \theta_{g_k}, \sigma_{g_k}^2)$$

The terms on the right do not involve θ_n and can thus be ignored as proportionality constants. This gives us

$$f(\boldsymbol{e}|\boldsymbol{x}, \boldsymbol{c}, \phi_n, \theta_n, \sigma_n^2) \propto \prod_{k \in G_n} \prod_{t=s_k+1}^{b_k} f(e_t|\boldsymbol{x}, c_k, \phi_n, \theta_n, \sigma_n^2) \tag{3.35}$$

In terms of the parameters, this is

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi_n,\theta_n,\sigma_n^2) \propto \prod_{k\in G_n}\prod_{t=s_k+1}^{b_k} exp\left\{-\frac{1}{2\sigma_n^2}\left(\theta_n^2 e_{t-1}^2 + 2(\theta_n e_{t-1})(\phi_n(x_{t-1}-c_k)-x_t+c_k)\right)\right\}$$

which is algebraically equivalent to

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi_n,\theta_n,\sigma_n^2) \propto exp\left\{-\frac{1}{2\sigma^2}\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k}\theta_n^2 e_{t-1}^2 + 2\theta_n e_{t-1}(\phi_n(x_{t-1}-c_k)-x_t+c_k)\right\} \tag{3.36}$$

After performing a bit of algebra just as we did in theorem 3.2, the double sum in the exponential function can be expressed as follows:

$$-\frac{\theta_n^2}{2\sigma_n^2}\left(\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2 + 2\theta_n\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}(\phi_n(x_{t-1}-c_k)-x_t+c_k)\right)$$

Now we can factor out $\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2$, giving us a quadratic form in θ_n on the right side.

$$-\frac{\theta_n^2}{2\sigma_n^2}\left(\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2\left(\theta_n^2 + 2\theta_n\frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}(\phi_n(x_{t-1}-c_k)-x_t+c_k)}{\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}\right)\right)$$

By completing the square, the above expression becomes

$$-\frac{\theta_n^2}{2\sigma_n^2}\left(\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2\left(\theta_n - \frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}(\phi_n(x_{t-1}-c_k)-x_t+c_k)}{\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}\right)^2 - \left(\frac{\sum_{k=1}^{K}\sum_{t=s_k+1}^{b_k} e_{t-1}(\phi_n(x_{t-1}-c_k)-x_t+c_k)}{\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}\right)^2\right)$$

The term being subtracted occurs in the exponential function and does not depend on θ_n and can therefore be ignored as a proportionality constant.

Now set

$$\xi_n = \frac{\sum_{k\in G_n}\sum_{t=s_k-1}^{b_k} e_{t-1}(x_t - c_k - \phi_n(x_{t-1}-c_k))}{\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}$$

and

$$\eta_n^2 = \frac{\sigma_n^2}{\sum_{k\in G_n}\sum_{t=s_k+1}^{b_k} e_{t-1}^2}$$

Then the likelihood function can be expressed in the form similar to what we see in equation (3.17).

$$f(\boldsymbol{e}|\boldsymbol{x},\boldsymbol{c},\phi_n,\theta_n,\sigma_n^2) \propto exp\left\{-\frac{1}{2\eta_n^2}(\theta_n-\xi_n)^2\right\} \tag{3.37}$$

The full conditional distribution function of θ_n is now

$$p(\theta_n|\boldsymbol{x},\boldsymbol{e},\boldsymbol{c},\theta_n,\sigma_n^2) \propto exp\left\{-\frac{1}{2\eta_n^2}(\theta_n-\xi_n)^2\right\} I(\theta_n)_{(-1,1)} \tag{3.38}$$

This is the kernel of a truncated normal distribution function parameters ξ and η^2 on the domain $-1 < \theta < 1$. Thus, we have shown

$$\theta_n | \cdot \sim TN(\xi_n, \eta_n^2, -1,1).$$

Since n was arbitrarily chosen, we have shown that this is true for all groups.

■

3.2 Segment Means and Group Sampling Distributions

Conditional Posterior Distribution of c paired with error terms e

The most time-intensive sampling is that of the segment means and group assignments. This is because every time a point is inserted or deleted, we must sample from the full conditional distributions of the segment groups and means. There can be up to K insertions and $K\text{-}1$ deletions per iteration. Given that we will run the code for at least 5000 iterations, and that we generally have well over 20 estimated segments, we will be sampling from the posterior distribution for segment means well over 100,000 times during most simulations. We will show that segment means can be directly sampled from a normal distribution.

Theorem 3.2.1

Given an arbitrary k^{th} segment $I_k = [s_k, b_k]$, the full conditional posterior distribution of the segment mean, c_k, is normal with mean

$$\mu_{c_k} = \frac{\tau^2_{g_k}\left((1-\phi_{g_k})\sum_{t=s_k+1}^{b_k}(x_t - \phi_{g_k}x_{t-1} - \theta_{g_k}e_{t-1}) + x_{s_k}\right) \quad + \sigma^2_{g_k}\mu_{g_k}}{\tau^2_{g_k}(1+(b_k-s_k)(1-\phi_{g_k})^2) \quad + \sigma^2_{g_k}}$$

and variance

$$\delta^2_{c_k} = \frac{\sigma^2_{g_k}\tau^2_{g_k}}{\tau^2_{g_k}(1+(b_k-s_k)(1-\phi_{g_k})^2) \; + \; \sigma^2_{g_k}}$$

Proof of Theorem 3.2.1

The prior on the segment mean is $c_k \sim N(\mu_k, \; \tau^2_{g_k})$. Also note that $e_t \overset{iid}{\sim} N(0, \sigma^2_{g_k})$.

Recall that

$x_t = c_k + e_t$ for $t = s_k$

$x_t = c_k + e_t + \phi_{g_k}(x_{t-1} - c_k) + \theta_{g_k}e_{t-1} \quad s_k < t \leq b_k$

And so

$e_t = c_t - c_k, \;$ for $t = s_k$

$e_t = x_t - c_k - \phi_{g_k}(x_{t-1} - c_k) - \theta_{g_k}e_{t-1} \quad s_k < t \leq b_k$

The full conditional posterior distribution is proportional to the following product:

$$p(c_k|\boldsymbol{e}, \boldsymbol{x}, g_k) \propto p(\boldsymbol{e}|c_k, g_k)p(c_k) = \prod_{t=s_k}^{b_k} p(e_t|x_t, g_{\boldsymbol{k}})p(c_k|\mu_{g_k}, \tau^2_{g_k})$$

$$\propto \left(\frac{1}{\sigma_{g_k}}\right)^{(b_k - s_k + 1)} \frac{1}{\tau^2_{g_k}} exp\left(-\frac{1}{2\sigma^2_{g_k}}\sum_{t=s_k}^{b_k}(e_t)^2\right) exp\left(-\frac{1}{2\tau^2_{g_k}}(c_k - \mu_{g_k})^2\right)$$

$$\propto exp\left(\left(-\frac{1}{2\sigma^2_{g_k}}\sum_{t=s_k}^{b_k}(e_t)^2\right) - \frac{1}{2\tau^2_{g_k}}(c_k - \mu_{g_k})^2\right)$$

We can re-express the above by noticing that the first term of the summation is just

$$e_t = x_t - c_k \quad \text{for } t = s_k$$

Our exponential expression becomes

$$exp\left(-\frac{1}{2\sigma_{g_k}^2}\left((x_{s_k} - c_k)^2 + \sum_{t=s_k+1}^{b_k} \left(x_t - c_k - \phi_{g_k}(x_{t-1} - c_k) - \theta_{g_k} e_{t-1} \right)^2 \right) - \frac{1}{2\tau_{g_k}^2}(c_k - \mu_{g_k})^2 \right) \quad (3.39)$$

The term within the sum can be re-expressed as follows

$$\left(x_t - c_k - \phi_{g_k}(x_{t-1} - c_k) - \theta_{g_k} e_{t-1} \right)^2$$

$$\left(x_t - \phi_{g_k} x_{t-1} - \theta_{g_k} e_{t-1} - c_k \left(1 - \phi_{g_k} \right) \right)^2$$

And expanded

$$c_k^2 \left(1 - \phi_{g_k} \right)^2 - 2c_k \left(1 - \phi_{g_k} \right) \left(x_t - \phi_{g_k} x_{t-1} - \theta_{g_k} e_{t-1} \right) + \left(x_t - \phi_{g_k} x_{t-1} - \theta_{g_k} e_{t-1} \right)^2$$

We now have a quadratic form in c_k.

The term to the left of the sum can be re-expressed as

$$(x_{s_k} - c_k)^2 = x_{s_k}^2 - 2x_{s_k} c_k + c_k^2$$

The term on the far right can be expanded to

$$-\frac{1}{2\tau_{g_k}^2}(c_k^2 - 2c_k \mu_{g_k}) - \frac{1}{2\tau_{g_k}^2}(\mu_{g_k}^2)$$

We can discard the factors that do not depend on c_k. We now have that our original expression (3.39) is proportional to

$$exp\left(-\frac{1}{2\sigma^2_{g_k}}\left(-2x_{s_k}c_k + c_k^2 + \sum_{t=s_k+1}^{b_k}\left(c_k^2(1-\phi_{g_k})^2 - 2c_k(1-\phi_{g_k})(x_t - \phi x_{t-1} - \theta_{g_k}e_{t-1})\right)\right) - \frac{c_k^2 - 2c_k\mu_{g_k}}{2\tau^2_{g_k}}\right)$$

$$\times exp\left(\ -\frac{1}{2\sigma^2_{g_k}}\left(x^2_{s_k} + \left(x_t - \phi_{g_k}x_{t-1} - \theta_{g_k}e_{t-1}\right)^2\right) - \frac{1}{2\tau^2_{g_k}}(\mu^2_{g_k})\right)$$

The factor on the right can be ignored as a proportionality constant. We will now focus on the expression within the exponential factor on the left.

$$\text{Let } A_t = (x_t - \phi x_{t-1} - \theta_{g_k}e_{t-1})$$

Now we have

$$-\frac{1}{2\sigma^2_{g_k}}\left(-2x_{s_k}c_k + c_k^2 + \sum_{t=s_k+1}^{b_k} c_k^2\left(1-\phi_{g_k}\right)^2 - 2c_k\sum_{t=s_k+1}^{b_k}(1-\phi_{g_k})A_t\right) - \frac{c_k^2}{2\tau^2_{g_k}} + \frac{c_k\mu_{g_k}}{\tau^2_{g_k}} \tag{3.40}$$

We can re-arrange the order of the terms in (3.40) to get the following form

$$\frac{-1}{2\sigma^2_{g_k}}\left(c_k^2 + c_k^2(b_k - s_k)\left(1-\phi_{g_k}\right)^2 - 2c_k\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t - 2x_{s_k}c_k\right) - \frac{c_k^2}{2\tau^2_{g_k}} + \frac{c_k\mu_{g_k}}{\tau^2_{g_k}}$$

We can re-express this as the sum of terms involve c_k^2 and the terms that involve c_k

$$\frac{-1}{2\sigma^2_{g_k}}\left(c_k^2 + c_k^2(b_k - s_k)\left(1-\phi_{g_k}\right)^2\right) - \frac{c_k^2}{2\tau^2_{g_k}} \quad + \quad \frac{-1}{2\sigma^2_{g_k}}\left(-2c_k\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t - 2x_{s_k}c_k\right) + \frac{c_k\mu_{g_k}}{\tau^2_{g_k}}$$

$$\left[\frac{-1}{2\sigma^2_{g_k}}\left(1 + (b_k - s_k)\left(1-\phi_{g_k}\right)^2\right) - \frac{1}{2\tau^2_{g_k}}\right]c_k^2 + \frac{-1}{2\sigma^2_{g_k}}\left(-2c_k\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t - 2x_{s_k}c_k\right) + \frac{c_k\mu_{g_k}}{\tau^2_{g_k}}$$

We can factor out $-\frac{1}{2}$, from the entire expression to get

$$-\frac{1}{2}\left[\frac{1}{\sigma^2_{g_k}}\left(1+\left(b_k-s_k\right)\left(1-\phi_{g_k}\right)^2\right)-\frac{1}{2\tau^2_{g_k}}\right]c_k^2+\frac{1}{\sigma^2_{g_k}}\left(-2c_k\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t-2x_{s_k}c_k\right)-2\frac{c_k\mu_{g_k}}{\tau^2_{g_k}}$$

We can factor out $-2c_k$ from the term on the right to get

$$-\frac{1}{2}\left(\left[\frac{1}{\sigma^2_{g_k}}\left(1+\left(b_k-s_k\right)\left(1-\phi_{g_k}\right)^2\right)+\frac{1}{\tau^2_{g_k}}\right]c_k^2+\left[\frac{-2c_k}{\sigma^2_{g_k}}\left(\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t+x_{s_k}\right)+\frac{\mu_{g_k}}{\tau^2_{g_k}}\right]\right) \tag{3.41}$$

$$\text{Let } B=\frac{1}{\sigma^2_{g_k}}\left(1+\left(b_k-s_k\right)\left(1-\phi_{g_k}\right)^2\right)+\frac{1}{\tau^2_{g_k}}$$

Then (3.41) becomes

$$-\frac{1}{2}\left(Bc_k^2+\left[\frac{-2c_k}{\sigma^2_{g_k}}\left(\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t+x_{s_k}\right)+\frac{\mu_{g_k}}{\tau^2_{g_k}}\right]\right)$$

$$=-\frac{1}{2}B\left(c_k^2-2\frac{\left[\frac{1}{\sigma^2_{g_k}}\left(\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t+x_{s_k}\right)+\frac{\mu_{g_k}}{\tau^2_{g_k}}\right]}{B}c_k\right) \tag{3.42}$$

By completing the square, (3.42) is proportional to

$$-\frac{1}{2(B^{-1})}\left(c_k-\frac{\left[\frac{1}{\sigma^2_{g_k}}\left(\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t+x_{s_k}\right)+\frac{\mu_{g_k}}{\tau^2_{g_k}}\right]}{B}\right)^2-\frac{1}{2(B^{-1})}\left(\frac{\left[\frac{1}{\sigma^2_{g_k}}\left(\sum_{t=s_k+1}^{b_k}\left(1-\phi_{g_k}\right)A_t+x_{s_k}\right)+\frac{\mu_{g_k}}{\tau^2_{g_k}}\right]}{B}\right)^2$$

The term on the right occurs in the exponential function and does not depend on c_k. Thus, it can be discarded as a proportionality constant. We now have that our full conditional posterior distribution is proportional to

$$exp\left(-\frac{1}{2(B^{-1})}\left(c_k-\frac{\left[\frac{1}{\sigma^2_{g_k}}\left(\sum_{t=s_k+1}^{b_k}(1-\phi_{g_k})A_t+x_{s_k}\right)+\frac{\mu_{g_k}}{\tau^2_{g_k}}\right]}{B}\right)^2\right)$$

This is the kernel of a normal distribution with mean

$$\mu_{c_k}=\frac{\left[\frac{1}{\sigma^2_{g_k}}\left((1-\phi_{g_k})\sum_{t=s_k+1}^{b_k}A_t+x_{s_k}\right)+\frac{\mu_{g_k}}{\tau^2_{g_k}}\right]}{B}$$

And variance

$$\delta^2_{c_k}=\frac{1}{B}$$

where

$$A_t=(x_t-\phi_{g_k}x_{t-1}-\theta_{g_k}e_{t-1}).$$

and

$$B=\frac{1}{\sigma^2_{g_k}}(1+(b_k-s_k)(1-\phi_{g_k})^2)+\frac{1}{\tau^2_{g_k}}$$

These are algebraically equivalent to

$$\mu_{c_k}=\frac{\tau^2_{g_k}\left((1-\phi_{g_k})\sum_{t=s_k+1}^{b_k}(x_t-\phi_{g_k}x_{t-1}-\theta_{g_k}e_{t-1})+x_{s_k}\right)\quad+\sigma^2_{g_k}\mu_{g_k}}{\tau^2_{g_k}(1+(b_k-s_k)(1-\phi_{g_k})^2)\quad+\sigma^2_{g_k}}$$

$$\delta^2_{c_k}=\frac{\sigma^2_{g_k}\tau^2_{g_k}}{\tau^2_{g_k}(1+(b_k-s_k)(1-\phi_{g_k})^2)\ +\ \sigma^2_{g_k}}$$

■

Note that if we are working with a single point, that being a segment of length 1, we have the given expressions below. This happens when we have adjacent changepoints.

$$\text{mean}\ \frac{\left[\frac{x_{s_k}}{\sigma^2_{g_k}}+\frac{\mu_{g_k}}{\tau^2_{g_k}}\right]}{\left[\frac{1}{\sigma^2_{g_k}}+\frac{1}{\tau^2_{g_k}}\right]}=\frac{x_{s_k}\tau^2_{g_k}+\mu_{g_k}\sigma^2_{g_k}}{\sigma^2_{g_k}+\tau^2_{g_k}}\quad\text{and}\quad\text{variance}\ \frac{\sigma^2_{g_k}\tau^2_{g_k}}{\sigma^2_{g_k}+\tau^2_{g_k}}.$$

Conditional Posterior Distribution of for groups.

After every insertion or deletion, we must sample a group from the new segments from the full conditional posterior distribution for groups. This derivation is quickly arrived at. We can generate a weighted discrete distribution from which to sample in order to assign a segment to a group. Given the prior $c_k \sim N(\mu_{g_k}, \tau^2_{g_k})$, and given that $\pi_1, \ldots, \pi_N$ represent the probabilities that a randomly occurring segment will have data originating from groups 1 through N, respectively, we have

$$p(\boldsymbol{g}|\boldsymbol{c}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\tau}) \propto p(\boldsymbol{c}|\boldsymbol{g}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(g)$$

$$\propto \prod_{k=1}^{K} N(c_k|\mu_{g_k}, \tau^2_{g_k}) \prod_{k=1}^{K} \pi_{g_k}$$

Groups are independent, and segment groups can be updated one segment at a time. A given segment k be assigned to group m with probability:

$$P(g_k = m|c_k, \tau^2_m, \pi_m) = \frac{\frac{1}{\tau_m} exp\left(-\frac{1}{2\tau^2_m}(c_k - \mu_m)^2\right)\pi_m}{\sum_{n=1}^{N} \frac{1}{\tau_n} exp\left(-\frac{1}{2}(c_{g_n} - \mu_{g_n})^2\right)\pi_n}$$

In the two-group case that we implement, this simplifies to:

$$P(g_k = 1|c_k, \tau_m^2, \pi_m) = \frac{\frac{1}{\tau_1} exp\left(-\frac{1}{2\tau_n^2}(c_k - \mu_1)^2\right)\pi_1}{\frac{1}{\tau_1} exp\left(-\frac{1}{2\tau_n^2}(c_k - \mu_1)^2\right)\pi_1 + \frac{1}{\tau_2} exp\left(-\frac{1}{2\tau_n^2}(c_k - \mu_2)^2\right)\pi_2}$$

$$P(g_k = 2|c_k, \tau_m^2, \pi_m) = 1 - P(g_k = 1|c_k, \tau_m^2, \pi_m).$$

3.3 Other Parameters

Conditional Posterior Distribution of Group Assignment Probabilities $\boldsymbol{\pi}$

Theorem 3.3.1

The prior for group probability assignments is just a trivial Dirichlet distribution $\boldsymbol{\pi} \sim Dirichlet(1, \ldots, 1)$ a symmetric Dirichlet prior, which assigns equal probability to all possible groups. Suppose we have K segments and N groups. Let $y_1,\ y_2, \ldots,\ y_N$ represent the number of segments with data originating from group 1, group 2, … etc. The joint conditional posterior distribution for the group probabilities is

$$\boldsymbol{\pi} \sim Dirichlet(y_1 + 1, \ldots, y_N + 1)$$

Proof of Theorem 3.3.1

The full conditional distribution of $\boldsymbol{\pi}$ is

$$p(\boldsymbol{\pi}|\boldsymbol{K}, \boldsymbol{g}) \propto p(\boldsymbol{g}|K, \boldsymbol{\pi})p(\pi)$$

By section 3.2.2 we know that $p(\boldsymbol{g}|K, \boldsymbol{\pi}) = \prod_{n=1}^{N} \pi_n^{y_n} = \prod_{n=1}^{N} \pi_n^{y_n+1-1}$

$$\propto D\text{irichlet}(y_1 + 1, y_2 + 1, \ldots, y_N + 1).$$

Since $\boldsymbol{\pi}$~$Dirichlet(1,1,...1)$,

$$p(\pi) \propto \prod_{\boldsymbol{n=1}}^{\boldsymbol{N}} \pi_n^{1-1} = 1$$

giving us

$$p(\boldsymbol{\pi}|K,g) \propto \prod_{\boldsymbol{n}=1}^{N} \pi_n^{(y_n+1)-1} \prod_{n=1}^{N} 1 = \prod_{n=1}^{N} \pi_n^{(y_n+1)-1} \tag{3.43}$$

This is proportional to the probability density function of a Dirichlet$(y_1+1,\ \ldots, y_N+1)$ distribution function. This shows that that the conditional posterior distribution of $\boldsymbol{\pi}$ is

$$Dirichlet(y_1+1,\ \ldots, y_N+1).$$

■

Given that we will be applying a two-group model for the applications portion of the project, a simplified sampling procedure may be of interest. We will only need to sample π_1 from a Beta distribution when dealing with 2 groups. This is computationally much faster than using a Dirichlet sampling function, and it can be shown to be nearly three times faster when implemented specifically in Matlab.

Corollary 3.3.2: When $N = 2$, $\pi_1 \sim Beta(y_1, K - y_1)$ and $\pi_2 = 1 - \pi_1$.

Proof of Corollary 3.3.2.

By the Fundamental Theorem of Probability. $\pi_2 = 1 - (\pi_1 + \pi_3 + \cdots + \pi_N)$, which is simply $1 - \pi_1$ when $N = 2$. Clearly it is only necessary to sample π_1 in order to implicitly have a sample for π_2.

By theorem 3.2.1, the joint distribution function for (π_1, π_2) is the pdf of a two parameter Dirichlet distribution:

$$f(\pi_1, \pi_2 | y_1 + 1, y_2 + 1) = \frac{\prod_{i=1}^{2} \Gamma(y_i + 1)}{\Gamma\big(\sum_{i=1}^{2} (y_i + 1)\big)} \prod_{i=1}^{2} \pi_n^{y_i} \tag{3.44}$$

And given that we only have two groups, the number of segments associated with group 2 is just K minus the number of segments associated with group 1. That is,

$$y_2 = K - y_1 \text{ and } \pi_2 = 1 - \pi_1.$$

Our density function (3.44) then becomes a function of only π_1 with parameter y_1.

$$f(\pi_1, \pi_2 | y_1 + 1, y_2 + 1) = f(\pi_1, (1 - \pi_1) | y_1 + 1, K - y_1 + 1)$$

$$= \frac{(y_1!)(y_2!)}{(y_1 + y_2 + 1)!} \pi_1^{y_1} \pi_2^{y2} = \frac{(y_1!)\big((K - y_1)!\big)}{\big(y_1 + (K - y_1 + 1)\big)!} \pi_1^{y_1} (1 - \pi_1)^{K - y_1}$$

We now only have to consider the distribution that only depends on π_1 and y_1.

The above is the pdf of a Beta distribution. Thus

$$\pi_1|\cdot \sim Beta(\pi_1|\alpha = y_1, \beta = K - y_1).$$

■

Conditional Posterior Distribution of γ

Probability of a changepoint occurring at any given time will be given a non-informative prior, $\gamma \sim$ Beta(1,1) = U(0,1). The conditional posterior distribution for γ is thus distributed Beta($\gamma|K, T-K$).

We see that

$$\begin{aligned} p(\gamma|K, \boldsymbol{s}) &\propto p(K, \boldsymbol{s}|\gamma)p(\gamma) \\ &= \lambda^K(1-\lambda)^{T-K} I_{(0,1)}. \end{aligned}$$

Which is the kernel of a $Beta(\gamma|K, T-K)$ pdf.

■

Conditional Posterior Distribution of $\boldsymbol{\mu}$

The prior for expected mean of the population of group means is weakly informative $\mu_{g_k} \sim N(0,1)$.

$$p(\boldsymbol{\mu}|\boldsymbol{c},\boldsymbol{g},\boldsymbol{\tau}) = \frac{p(\boldsymbol{c}|\boldsymbol{g},\boldsymbol{\mu},\boldsymbol{\tau})p(\boldsymbol{\mu})}{\int p(c|\boldsymbol{g},\boldsymbol{\mu},\boldsymbol{\tau})p(\boldsymbol{\mu})d\boldsymbol{\mu}}$$

$$\propto p(\boldsymbol{c}|\boldsymbol{g},\boldsymbol{\mu},\boldsymbol{\tau})p(\boldsymbol{\mu})$$

$$\propto \prod_{k=1}^{K} N(c_k|g_{k,}\mu_k,\tau_k^2)\prod_{k=1}^{K} p(\mu_k)$$

$$\propto \prod_{k=1}^{K} N(c_k|g_{k,}\mu_k,\tau_k^2)\prod_{k=1}^{K} N(0,1)$$

The non-informative prior does not depend on c_k, and so we can ignore $\prod_{k=1}^{K} N(0,1)$ as a proportionality constant, giving us

$$p(\boldsymbol{\mu}|\boldsymbol{c},\boldsymbol{g},\boldsymbol{\mu},\boldsymbol{\tau}) \propto \prod_{k=1}^{K} N(c_k|\mu_k,\tau_k^2)$$

Let n be an arbitrary group. Let G_n be the set of all segments associated with group n. Then we can separate the above product into

$$\prod_{k\in G_n} N(c_k|\mu_n,\tau_n^2)\prod_{k\notin G_n} N(c_k|\mu_{g_k},\tau_{g_k}^2)$$

The product on the right is a constant with respect to group n. We can now see that for group n, the full conditional posterior distribution for the mean of the population of segment means, μ_n, has a pdf proportional to

$$\prod_{k\in G_n} N(c_k|\mu_n,\tau_n^2) = exp\left(\sum_{k\in G_n} -\frac{1}{2\tau_n^2}(c_k-\mu_n)^2\right)$$

If we swap the sign within the term $(c_k-\mu_n)^2$, we end up with

$$exp\left(\sum_{k\in G_n} -\frac{1}{2\tau_n^2}(\mu_n-c_k)^2\right)$$

$$= exp\left(-\frac{1}{2\tau_n^2}\left(\sum_{k\in G_n}\mu_n^2 - 2\mu_n\sum_{k\in G_n} c_k + \sum_{k\in G_n} c_k^2\right)\right)$$

$$exp\left(-\frac{1}{2\tau_n^2}\left(\sum_{k\in G_n}\mu_n^2 - 2\mu_n\sum_{k\in G_n}c_k\right)\right)exp\left(-\frac{1}{2\tau_n^2}\left(\sum_{k\in G_n}c_k^2\right)\right)$$

The term on the right does not depend on μ_n and so it can be discarded as a proportionality constant, giving us

$$exp\left(-\frac{1}{2\tau_n^2}\left(\sum_{k\in G_n}\mu_n^2 - 2\mu_n\sum_{k\in G_n}c_k\right)\right)$$

Let $m > 0$ be the number of segments associated with group n. (If $m = 0$, we will not update the mean for group m). Then the above expression becomes

$$exp\left(-\frac{1}{2\tau_n^2}\left(m\mu_n^2 - 2\mu_n\sum_{k\in G_n}c_k\right)\right)$$

$$= exp\left(-\frac{m}{2\tau_n^2}\left(\mu_n^2 - 2\frac{\mu_n}{m}\sum_{k\in G_n}c_k\right)\right)$$

By completing the square, we get

$$exp\left(\frac{m}{2\tau_n^2} - \left[\left(\mu_n - \frac{1}{m}\sum_{k\in G_n}c_k\right)^2 - \left(\frac{1}{m}\sum_{k\in G_n}c_k\right)^2\right]\right)$$

$$= exp\left(\frac{m}{2\tau_n^2} - \left[\left(\mu_n - \frac{1}{m}\sum_{k\in G_n}c_k\right)^2\right]\right)exp\left(-\frac{m}{2\tau_n^2}\left[-\left(\frac{1}{m}\sum_{k\in G_n}c_k\right)^2\right]\right)$$

The term on the right does not depend on μ_n and so it can be ignored as a proportionality constant.

$$\propto exp\left(-\frac{m}{2\tau_n^2}\left(\mu_n - \frac{1}{m}\sum_{k\in G_n}c_k\right)^2\right)$$

$$\propto exp\left(-\frac{1}{2\frac{\tau_n^2}{m}}\left(\mu_n - \frac{1}{m}\sum_{k\in G_n}c_k\right)^2\right) \tag{3.45}$$

This is the kernel of a normal pdf in μ_n with mean $\frac{1}{m}\sum_{k\in G_n} c_k$ and variance $\frac{\tau_n^2}{m}$.

Let us label m more specifically as m_n to state more explicitly that m_n refers to the number of segments associated with group n.

Since n was chosen arbitrarily, we have shown that for all groups, n, the full conditional posterior distribution for μ_n is

$$N\left(\frac{1}{m_n}\sum_{k\in G_n} c_k\,, \frac{\tau_n^2}{m_n}\right)$$

■

Conditional Posterior Distribution of σ_n^2

The original model by [Sadia et al., 2018], it is assumed that error variance, σ^2, is equal for all groups. Given an inverted gamma prior on σ^2 with shape parameter α, scale parameter β, they state that the full posterior conditional distribution for σ^2 is also inverted gamma with parameters

$$\text{Shape } \ a = \alpha + \frac{T}{2}$$

$$\text{Scale } \ \ b = \beta + \frac{1}{2}\sum_{t=1}^{T} e_t^2$$

With our model, we assign an inverse gamma prior to each σ_n^2 with shape and scale parameters denoted α_n and β_n. We will show that the full conditional posterior conditional distribution for each σ_n^2 is inverted gamma with

$$\text{Shape } \ a_n = \alpha_n + \frac{T_n}{2}$$

$$\text{Scale } \ \ b_n = \beta_n + \frac{1}{2}\sum_{k \in G_n}\sum_{t=s_k}^{b_k} e_t^2$$

Where T_n is the total number of data points over all intervals associated with group n.

$$T_n = \sum_{k \in G_n} (b_k - s_k + 1).$$

Let G_n be the set of all segments, k, associated with group n.

Then

$$p(\sigma_n^2 | \boldsymbol{e}, G_n) \propto p(\boldsymbol{e} | G_n, \sigma_n^2) p(\sigma_n^2).$$

The prior for each σ_n^2 is inverted gamma,

$$p(\sigma_n^2|\alpha_n,\beta_n) = \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)}(\sigma_n^2)^{-\alpha_n-1}\, exp\left\{-\frac{\beta_n}{\sigma_n^2}\right\}$$

and the error terms are iid $\mathrm{N}(0,\sigma_n^2)$, giving us

$$p(\boldsymbol{e}|G_n,\sigma_n^2)p(\sigma_{g_n}^2) \propto \left(\prod_{k\in G_n}\prod_{t=s_k}^{b_k} N(e_t|\lambda_t,\sigma_{g_k}^2)\right)\frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)}(\sigma_n^2)^{-\alpha_n-1}\, exp\left\{-\frac{\beta_n}{\sigma_n^2}\right\}$$

$$= \left(\prod_{k\in G_n}\prod_{t=s_k}^{b_k}\frac{1}{\sigma_n}\right) exp\left(-\frac{1}{2\sigma_n^2}\sum_{k\in G_n}\sum_{t=s_k}^{b_k} e_t^2\right)\frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)}(\sigma_n^2)^{-\alpha_n-1}\, exp\left\{-\frac{\beta_n}{\sigma_n^2}\right\} \tag{3.46}$$

The term $\frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)}$ does not depend on σ_n^2 and so it can be ignored as a proportionality constant.

The product $\prod_{k\in G_n}\prod_{t=s_k}^{b_k}\frac{1}{\sigma_n}$ can be re-expressed as $\left(\frac{1}{\sigma_n}\right)^{\sum_{k\in G_n}(b_k-s_k+1)} = (\sigma_n^2)^{-\frac{T_n}{2}}$.

The above expression (3.46) is thus proportional to

$$(\sigma_n^2)^{-\frac{T_n}{2}}(\sigma_n^2)^{-\alpha_n-1}\, exp\left(-\frac{1}{2\sigma_n^2}\sum_{k\in G_n}\sum_{t=s_k}^{b_k} e_t^2 - \frac{\beta_n}{\sigma_n^2}\right)$$

$$= (\sigma_n)^{-\left(\frac{T_n}{2}+\alpha_n\right)-1}\, exp\left(-\frac{1}{\sigma_n^2}\left(\frac{1}{2}\sum_{k\in G_n}\sum_{t=s_k}^{b_k} e_t^2 + \beta_n\right)\right)$$

This is the kernel of an inverted gamma distribution function in σ_n^2 with shape parameter

$$a_n = \frac{T_n}{2} + \alpha_n$$

and scale parameter

$$b_n = \frac{1}{2}\sum_{k\in G_n}\sum_{t=s_k}^{b_k} e_t^2 + \beta_n$$

■

Conditional Posterior Distribution of $\boldsymbol{\tau}$

Given an inverted gamma prior on τ_n^2 with shape parameter α_n, scale parameter β_n, it can be shown that the posterior conditional distribution is also inverted gamma with parameters

$$\text{Shape } a_n = \alpha_n + \frac{y_n}{2}$$

$$\text{Scale } b_n = \beta_n + \frac{1}{2}\sum_{k \in G_n} \left(c_k - \mu_n\right)^2$$

Where y_n is the number of segments associated with group n. The algebra is analogous to that of the derivation of the full conditional posterior for σ_n^2.

$$p(\tau_n^2|\boldsymbol{c}, G_n) \propto p(\boldsymbol{c}|G_n)p(\tau_n^2)$$

The prior for each n is

$$p(\tau_n^2|\alpha_n, \beta_n) = \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)}(\tau_n^2)^{-\alpha_n - 1}\, exp\left\{-\frac{\beta_n}{\tau_n^2}\right\}.$$

Then

$$p(\tau_n^2|\boldsymbol{c}, G_n) \propto p(\boldsymbol{c}|G_n)p(\tau_n^2)$$

$$\propto \left(\prod_{k \in G_n} N(c_k|\mu_n, \tau_n^2)\right) \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)}(\tau_n^2)^{-\alpha_n - 1}\, exp\left\{-\frac{\beta_n}{\tau_n^2}\right\}$$

$$\propto \left(\frac{1}{\tau_n^2}\right)^{\frac{y_n}{2}} (\tau_n^2)^{-\alpha_n - 1}\, exp\left(-\frac{1}{2\tau_n^2}\sum_{k \in G_n}(c_k - \mu_n)^2 \;\; - \frac{\beta_n}{\tau_n^2}\right)$$

$$\propto (\tau_n^2)^{-(\frac{y_n}{2}+\alpha_n)-1}\, exp\left(-\frac{1}{\tau_n^2}\left(\frac{1}{2}\sum_{k\in G_n}(c_k-\mu_n)^2+\beta_n\right)\right)$$

This is the kernel of an inverted gamma distribution with parameters

$$\text{Shape } a_n = \alpha_n + \frac{y_n}{2}$$

$$\text{Scale } b_n = \beta_n + \frac{1}{2}\sum_{k\in G_n}(c_k-\mu_n)^2$$

■

Chapter 4

Simulations and Application to Real World Data

4.1 Existing Simulated Data Results

One simulation by [Sadia et al., 2018] is observed in which a data set was generated from an ARMA(1,1) process with specific parameter values of $\phi = 0.22$ and $\theta = 0.6$ (See figure 4.1.1). The data vary in that the means of the segments differ. The data are generated in such a way as to provide segments of equal length. Their model performs well on this data set, accurately detecting 17 of out of 20 change points, as shown in figure 4.1.2. They actually detect 18 points total points, but one of these points is adjacent to the second changepoint, resulting in an interval of length one. This is subtly apparent by the thickness of the vertical blue line near the second changepoint in figure 4.1.2. This indicates that the model has the potential to keep points that happen to be one unit of time apart. This is an issue we later address when analyzing our own simulated data in section 2 of this chapter. Trace plots of the AR and MA parameters are given, showing a good model fit.

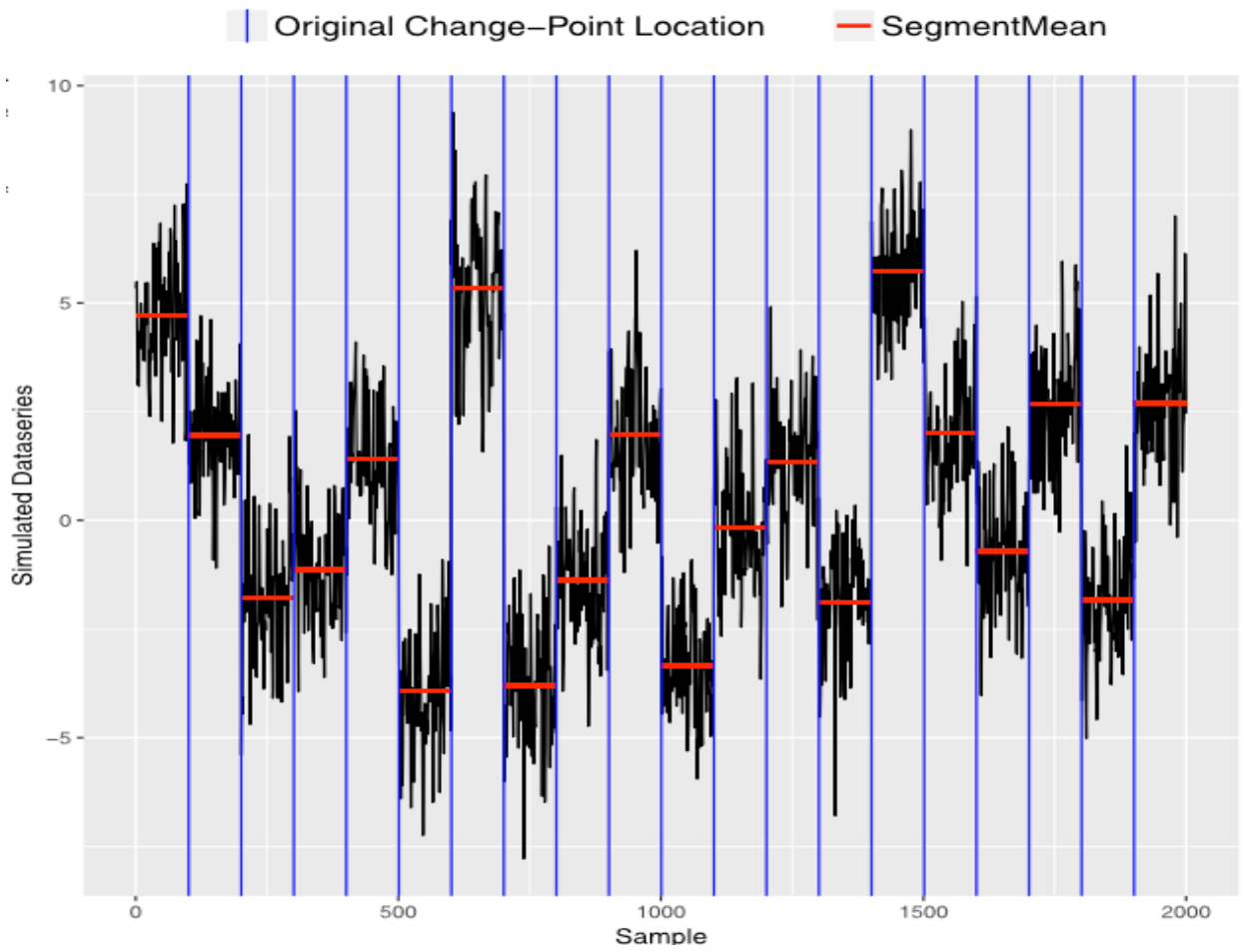


Figure 4.1.1 [Sadia et al., 2018] Page 11. True changepoint locations are illustrated by the vertical lines. Horizontal lines indicate segment means.

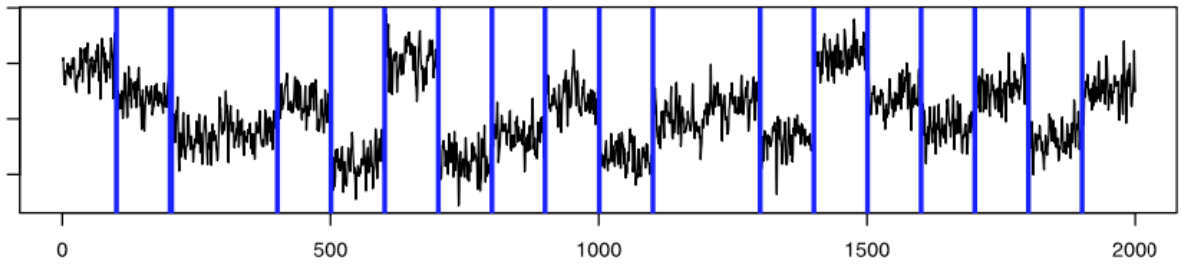


Figure 4.1.2 [Sadia et al., 2018] Page 13. This indicates the changepoints detected by their algorithm after 5000 iterations. Adjacent changepoint detection occurs next to the second change point, indicated by the double thickness of the vertical blue line.

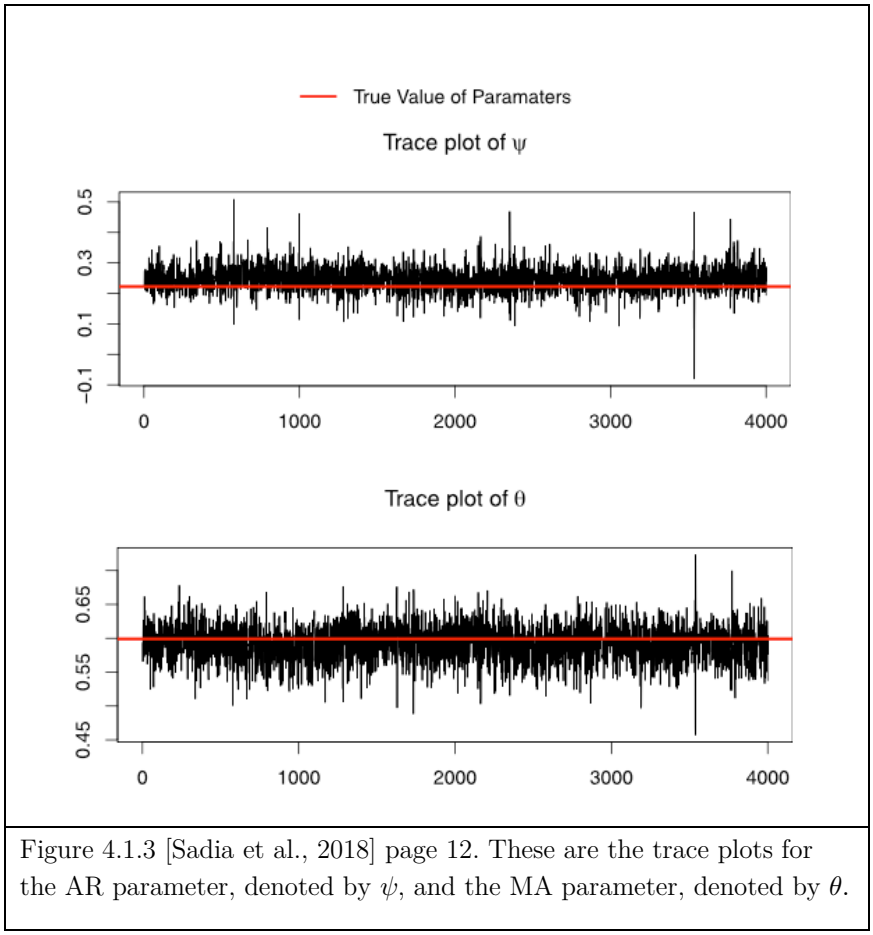


Figure 4.1.3 [Sadia et al., 2018] page 12. These are the trace plots for the AR parameter, denoted by ψ, and the MA parameter, denoted by θ.

Their model performs well on the data set provided. They state that other parameters were also estimated accurately, although we do not have further detail on this. Note that in their paper, ψ represents the AR parameter, rather than ϕ.

4.2 Application of Our Model to Simulated Data

We apply our model as described in equation (2.4) to several simulated data sets. The framework for a model by [Sadia et al., 2018] assumes that groups can be separated based on segment mean, and it is therefore explicitly stated that the model is not appropriate for cases in which data between groups have differing variance. As we can see, even in individual segments of data from MA, AR, and ARMA processes with varying parameters, the spread of data may vary widely. If data originate from different groups, it may be difficult to accurately estimate the ARMA parameters, and it may result in difficulty in detecting changepoints. AR parameters toward the extremes are the most influential, as their effect on variance is greatest. The noise variance is held constant for all of these figures. Were error variance also to differ, we would see an even greater range of variation in the data.

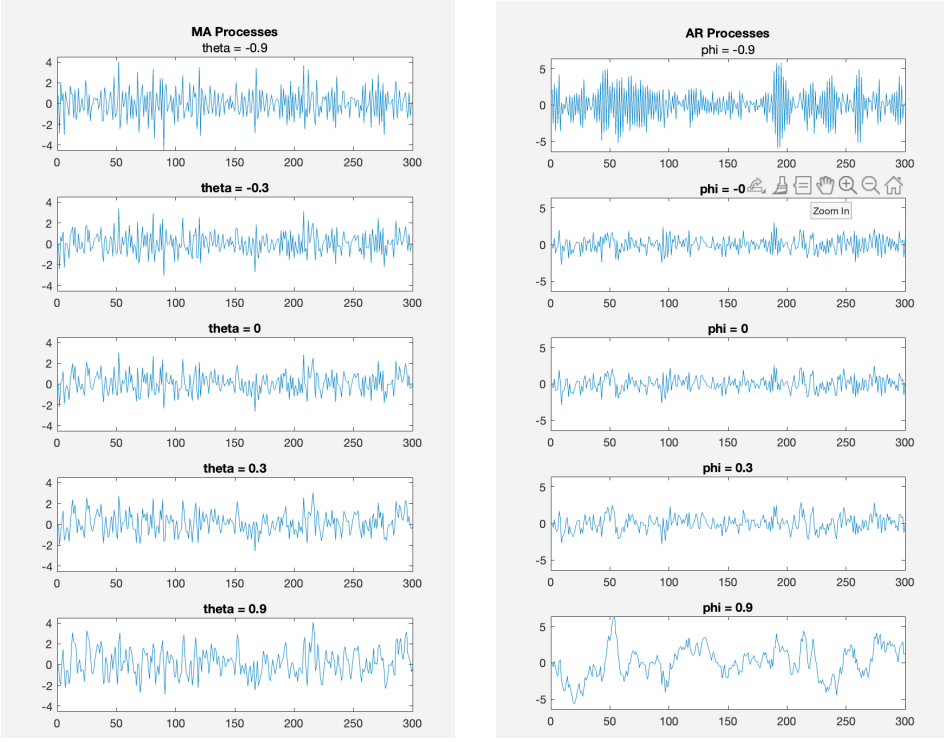


Figure 4.2.1 Examples of MA(1) processes (left) and AR(1) processes (right).

We generate a number of different data sets. Our tests differ in that segment lengths and locations are randomized. Here is an example of a typical simulation. Two groups are defined with relatively disparate expected segment means. Group 1 has noise variance 0.7 while group 2 has noise variance 0.4. Group means are -10 and 10, respectively. The trace plots show convergence to the true values of the ARMA parameters. We also provide trace plots for all other parameters. A notable aspect is that we see a lot of variation in the estimates for group mean variances. This is due to a difficulty in estimating a parameter like variance with relatively small sample sizes. Each group has ten segments. As the simulation runs, segment groups assignments change. At times, we may have more segments defined as group 1 or group 2 than the other, increasing the uncertainty involved with estimating group variance.

4.2.1 Groups with Different Random Noise Variance

Example 4.2.1 Data with Differing Random Noise Variance

We generate 20 segments of randomly varying length, with minimum length 50, are generated over a collection of 2000 data points. Segments have equal AR and MA parameters and different error variances.

Group 1 Data on Segment k. Higher noise variance.

$$\begin{aligned} x_t &= c_k + e_t + \phi_1(x_{t-1} - c_k) + \theta_1 e_{t-1} \\ &= c_k + e_t + 0.6(x_{t-1} - c_k) + 0.2e_{t-1} \end{aligned}$$

$$\begin{aligned} c_k &\sim N(\mu_k = -10, \tau_k^2 = 16) \\ e_t &\overset{iid}{\sim} N(0, \sigma^2 = 0.7) \end{aligned}$$

Group 2 Data on Segment k.

$$\begin{aligned} x_t &= c_k + e_t + \phi_1(x_{t-1} - c_k) + \theta_1 e_{t-1} \\ &= c_k + e_t + 0.6(x_{t-1} - c_k) + 0.2e_{t-1} \\ c_k &\sim N(\mu_k = 10, \tau_k^2 = 16) \\ e_t &\overset{iid}{\sim} N(0, \sigma^2 = 0.4) \end{aligned}$$

All changepoints are detected and sorted into distinct groups. Random noise variance is 0.7 for group 1 and only 0.4 for group 2. Other model parameters are the same. Vertical dashed lines indicate segment endpoints, with the interior vertical dashed lines representing changepoints. Blue horizontal line segments represent segments means of group 1 data (lower segments) while yellow represent group 2 segments (upper).

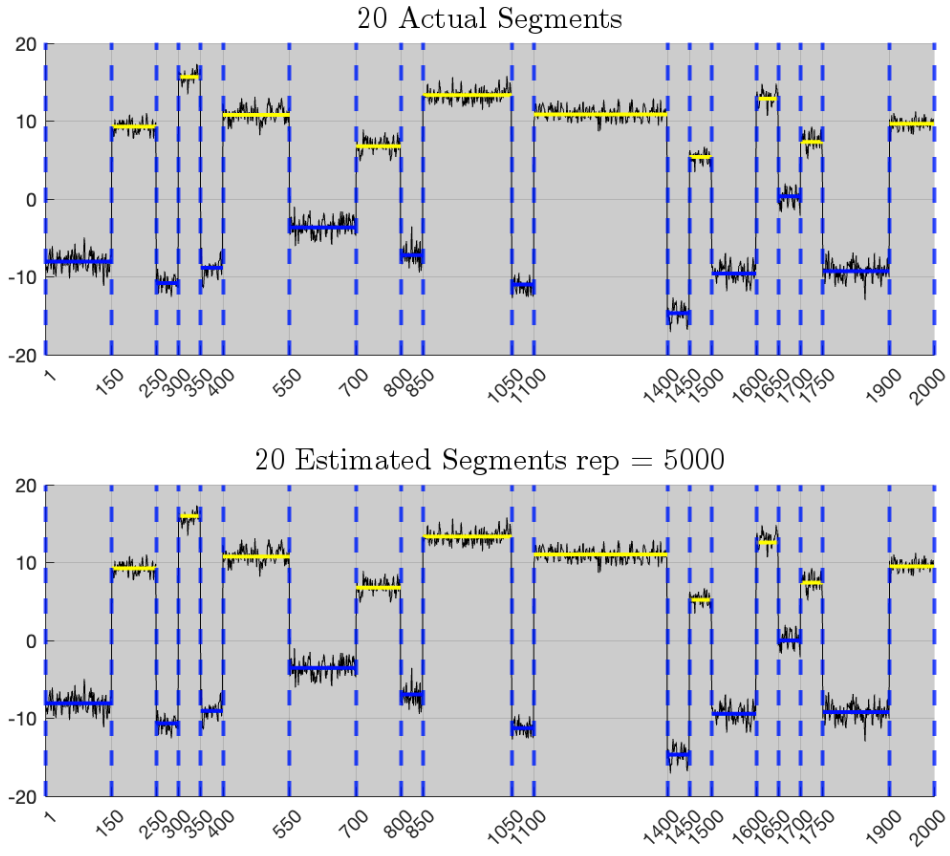


Figure 4.2.2 Changepoint detection results for example 4.2.1 for data with similar ARMA parameters and differing random noise.

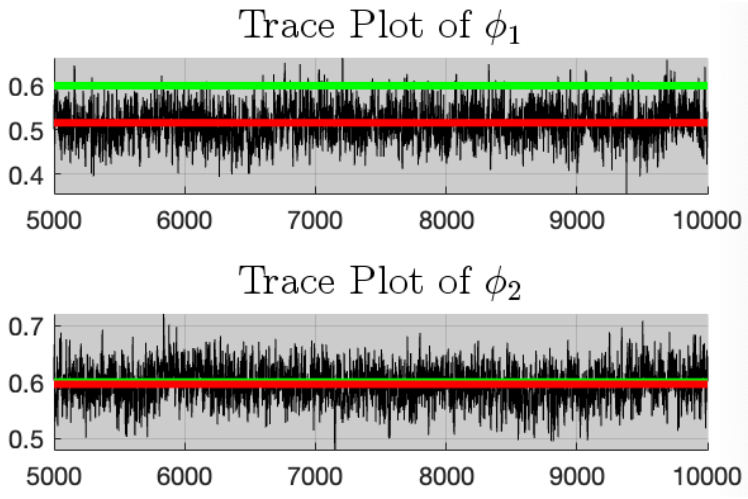


Figure 4.2.3 Trace plots of the AR parameters. The actual value is set to 0.6 for both groups. Red indicates estimate. Green represents true parameter value.

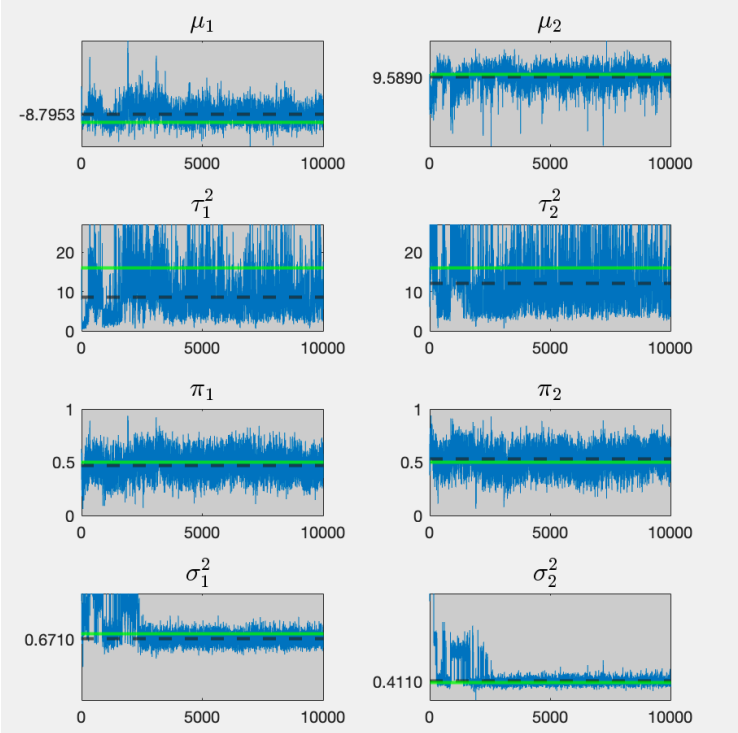


Figure 4.2.4 Trace plots of all parameters. The algorithm does well in determining the difference between random noise from groups 1 and 2, as shown in the bottom row.

Least Squares Estimator of Parameters for Example 4.2.1

Since we know the true locations of the changepoints, we can compare the results of our Gibbs sampler to those of a least-squares best-fit ARMA modeling procedure. We take each of the group 1 and group 2 segments and combine them to form two series. Notice that group 2 has more data than group 1. This is an artifact of having randomly generated 10 segments for each group, with each segment having a random length. Thus, a difference in total observations is expected with our simulated data. Series of combined segment data are shown in figure 4.2.5, all centered at the true segment mean locations, separated by group. All off the group 2 segments in figure 4.2.8 (upper segments with yellow indicating segment means) were combined to form the upper series in this image. The group one segments combine to form the lower series in this image.

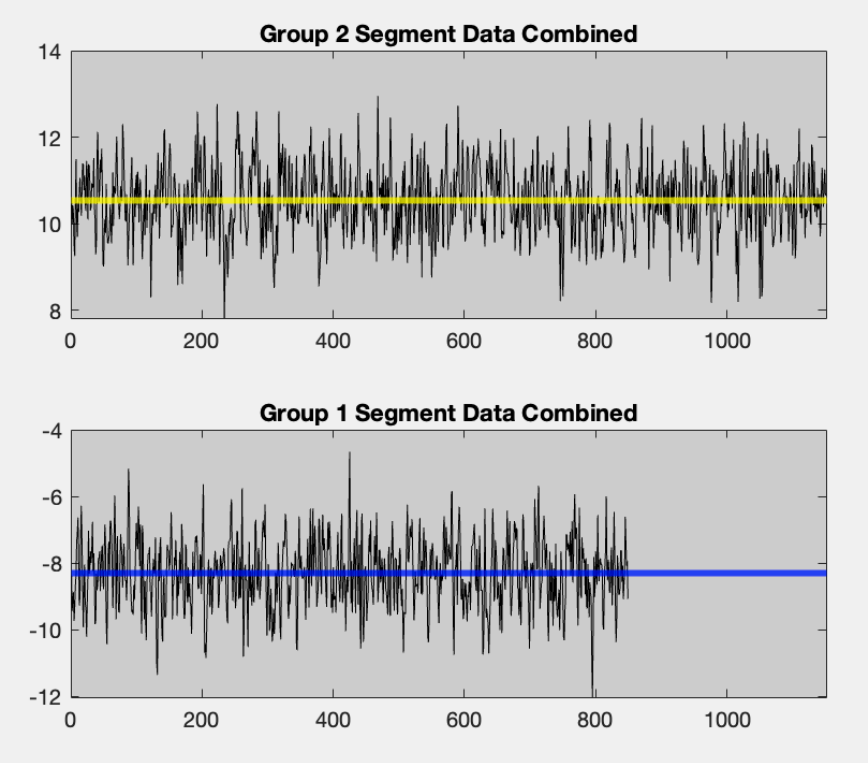


Figure 4.2.5 Combined segment data to form two series from groups 1 and 2.

MATLAB's least squares estimator was not used to find the changepoints. Table 4.2.1 gives a comparison of what Matlab estimates would be when given all of the group 1 data and all of the group 2 data combined, as if these were two full sequences of ARMA process data.

	Gibbs Sampler after 5000 iterations	**Least Squares Estimator in Matlab**	**Actual Value**
ϕ_1	0.122	0.134	0.200
ϕ_2	0.185	0.141	0.200
θ_1	0.653	0.618	0.600
θ_2	0.614	0.616	0.600
σ_1^2	0.683	0.671	0.700
σ_2^2	0.415	0.402	0.400
μ_1	-8.736	-8.312	-10.00
μ_2	9.241	10.252	10.00

Table 4.2.1 Comparison of Gibbs Sampler, Matlab least squares estimator, and true parameter values for example 4.2.1.

We manually calculate the MSE for the Gibbs sampler, Matlab least squares estimator, and the true values of the ARMA parameters that were used to generate the original data sets. Let c_1 and c_2 be the sample means of all data when group 1 segments and group 2 segments are combined and centered, respectively.

Given the combined left data, let ϕ_i and θ_i represent the estimates for the ARMA parameters for group i. We calculate the MSE, given the means of the combined segment data recursively, as follows

$$e_1 = x_1 - c_1$$
$$e_t = x_t - c_2 - \phi_i(x_{t-1} - c_2) - \theta_i e_{t-1} \qquad 1 < t \leq T_i \ ,$$

where T_i is the total number of data points in group 1.

We perform this process using the ARMA parameter estimates from the Gibbs Sampler, the Matlab least squares estimator, and finally using the actual fixed values used to generate the data from the beginning.

We see that the estimated mean squared error for the Gibbs sampler and least squares estimators are very similar, with the Gibbs sampler seeming to outperform Matlab in in the group 2 case. This is likely due to the fact that we have generated group data on random segments and later combined segments, which is not identical to generating two consecutive series of ARMA data. We of course expect the MSE to be lowest when using the actual parameters that the data were generated from.

	Gibbs Sampler	**Least Squares Estimator in Matlab**	**Using true ARMA parameters**
MSE Group 1	0.523	.513	0.439
MSE Group 2	0.273	.303	0.264

Table 4.2.2 Comparison of the performance of the Gibbs Sampler vs a Least Squares estimation procedure for fitting an ARMA model to the data.

Example 4.2.2 Differing AR and MA Parameters

In our first example, we had fixed AR and MA parameters, with different variances for random noise. We will now demonstrate the performance of our model with groups that have vary widely in terms of the AR and MA parameters. We use the same segment endpoints as those from example 4.2.1 for visual comparison. We generated data based on the following group structure:

ARMA Parameters

Group 1: $\phi_1 = 0.8$ $\theta_1 = 0.8$
Group 2: $\phi_2 = -0.8$ $\theta_2 = -0.8$

Group 1 Data

$$\begin{aligned} x_t &= c_k + e_t + \phi_1(x_{t-1} - c_k) + \theta_1 e_{t-1} \\ &= c_k + e_t + 0.8(x_{t-1} - c_k) + 0.8e_{t-1} \end{aligned}$$

$$c_k \sim N(\mu_k = -10, \tau_k^2 = 16)$$
$$e_t \overset{iid}{\sim} N(0, \sigma^2 = 0.9)$$

Group 2 Data

$$\begin{aligned} x_t &= c_k + e_t + \phi_1(x_{t-1} - c_k) + \theta_1 e_{t-1} \\ &= c_k + e_t - 0.8(x_{t-1} - c_k) - 0.8e_{t-1} \end{aligned}$$
$$c_k \sim N(\mu_k = 10, \tau_k^2 = 16)$$
$$e_t \overset{iid}{\sim} N(0, \sigma^2 = 0.9)$$

Both group segments have equal noise variance 0.9. Group expected segment mean and segment mean variances are identical and the same as those from example 4.2.1. We still allow the algorithm to assume variation in noise between groups, as this allows our overall estimates to converge based on the overall behavior within the combined data from each group.

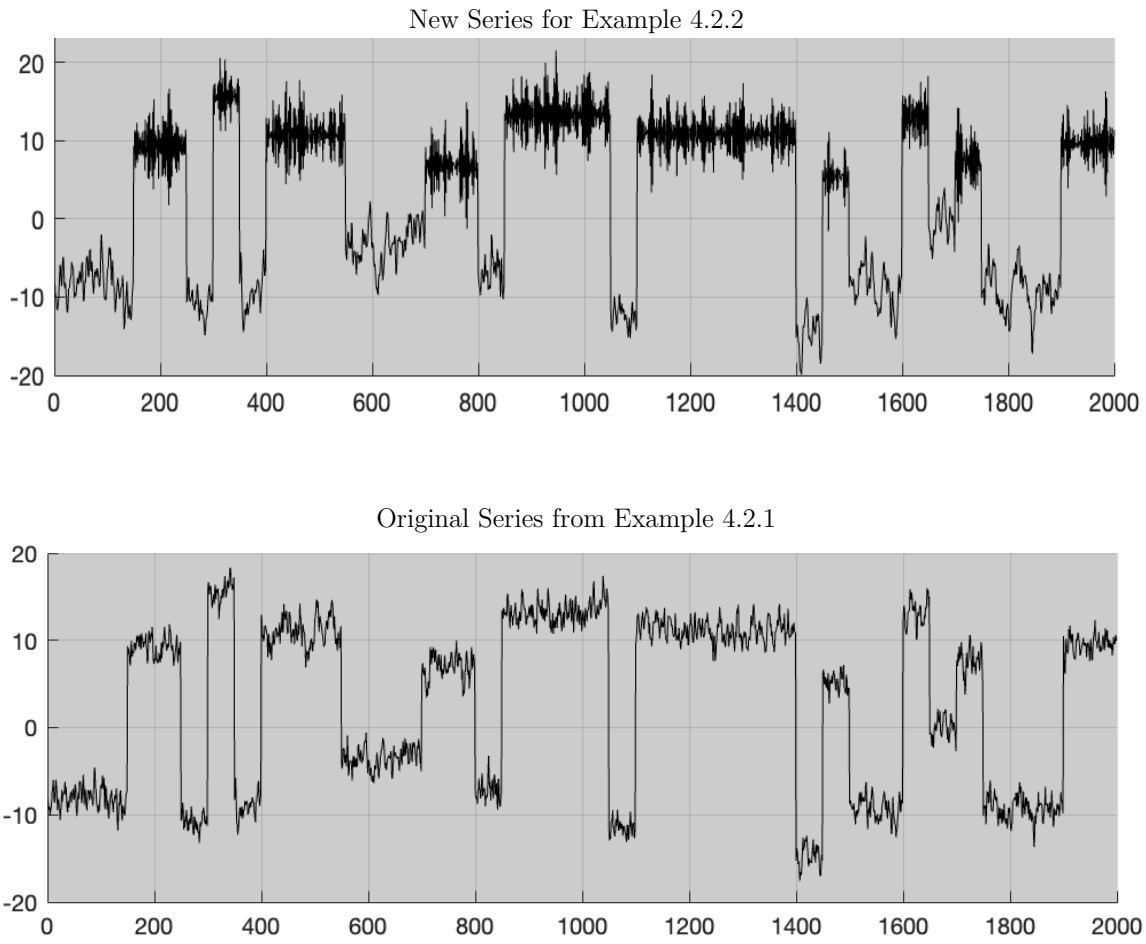


Figure 4.2.6 Data series for Example 4.2.2 compared to that of 4.2.1.

Group 1 data, (lower data), have large magnitude positive AR and MA parameters. This causes terms to be more positively correlated with one another, producing what appears to be non-stationary linear or even periodic trends, despite the data coming from a stationary distribution.

Group 2 data, (upper data), oscillate erratically. Both the AR and MA parameters are large in magnitude and negative. Terms tend to rapidly deviate away from each other. This creates what visually appears to be a non-stationary feature, that being unequal

variance over time. Were it unknown that the data do originate from two stationary processes by group, we might suspect non-stationarity. This does not affect the outcome of the simulation.

After 5,000 iterations, we get reliable identification of all changepoints, with only one misidentification of an additional changepoint at $t = 27$. It takes over 20,000 iterations for the algorithm to accurately estimate the model parameters, however.

Changepoint Detection for Example 4.2.2 Data

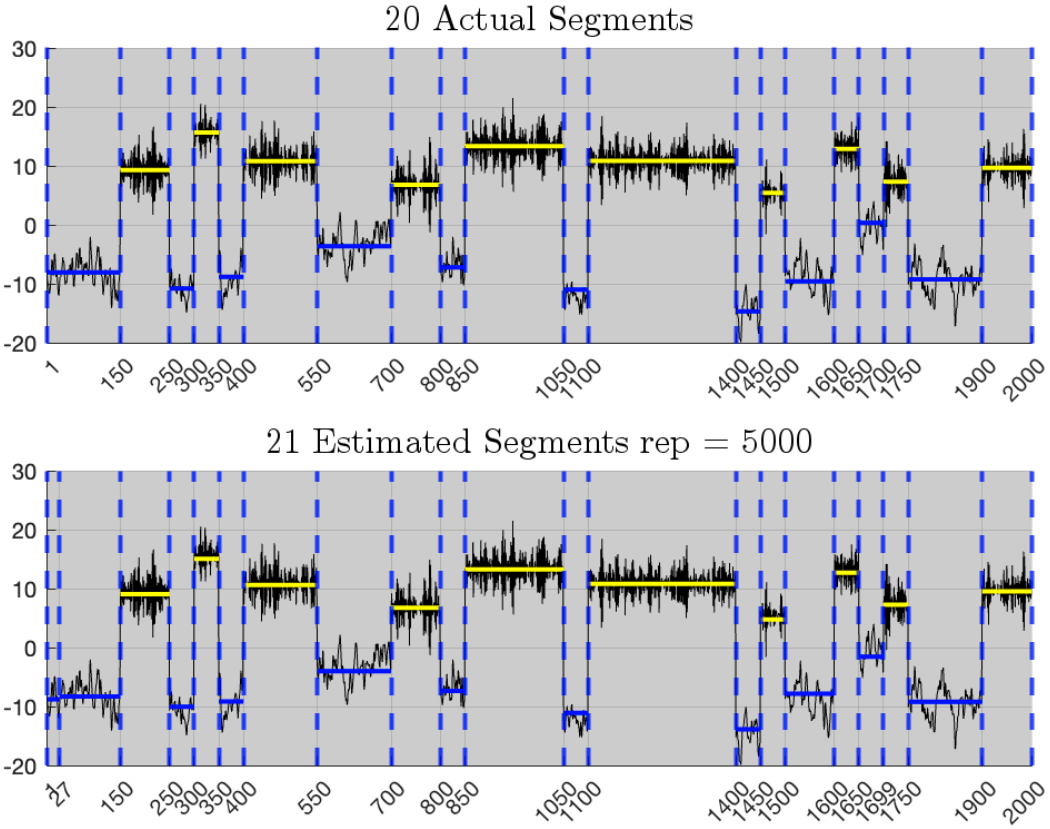


Figure 4.2.7 All changepoints are detected and segments are sorted into groups.

Trace Plots for Example 4.2.2

As we can see, 5000 iterations is not sufficient to accurately estimate our model parameters. Trace plots of ARMA parameters are not stable during a simulation with only 5000 iterations. Despite the data being visibly separated into segments and distinct groups, the volatility and variability of data within segments makes parameter estimation difficult without many more iterations.

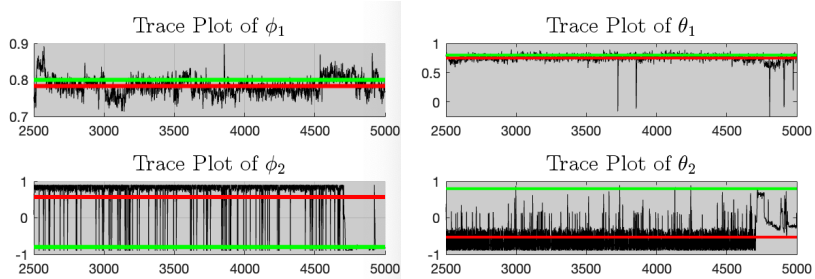


Figure 4.2.8 Trace plots of ARMA parameters for example 4.2.2 at 5,000 iterations.

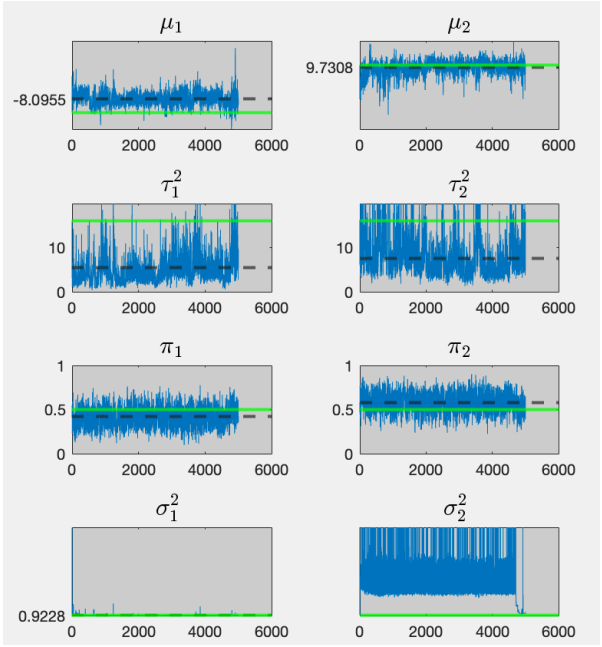


Figure 4.2.9 Trace plots of other parameters.

It takes almost 30,000 iterations to get convergence of the second moving average parameter to the true value. Even with this many iterations, it appears that it may take even longer to get convergence for θ in group 2. This likely has to do with how fast the data are oscillating when the data are highly negatively correlated with each other through both the AR parameter and MA parameter.

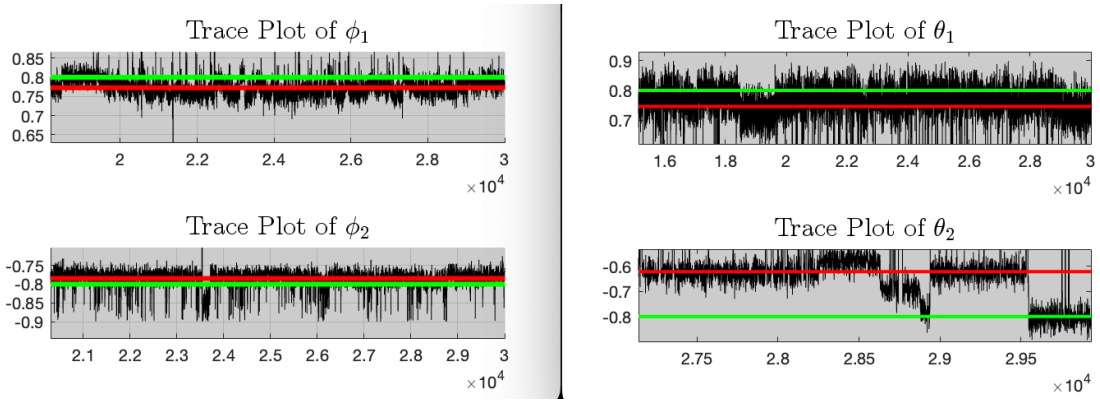


Figure 4.2.10 AR and MA trace plots for Example 4.2.2 at 30,000 iterations.

The parameters converge, both close to the true value of 0.9, but with a high degree of skew for error variance, as seen in the lower row. This does make sense, as the inverted gamma distribution is skewed, and the data vary widely from one segment to the next.

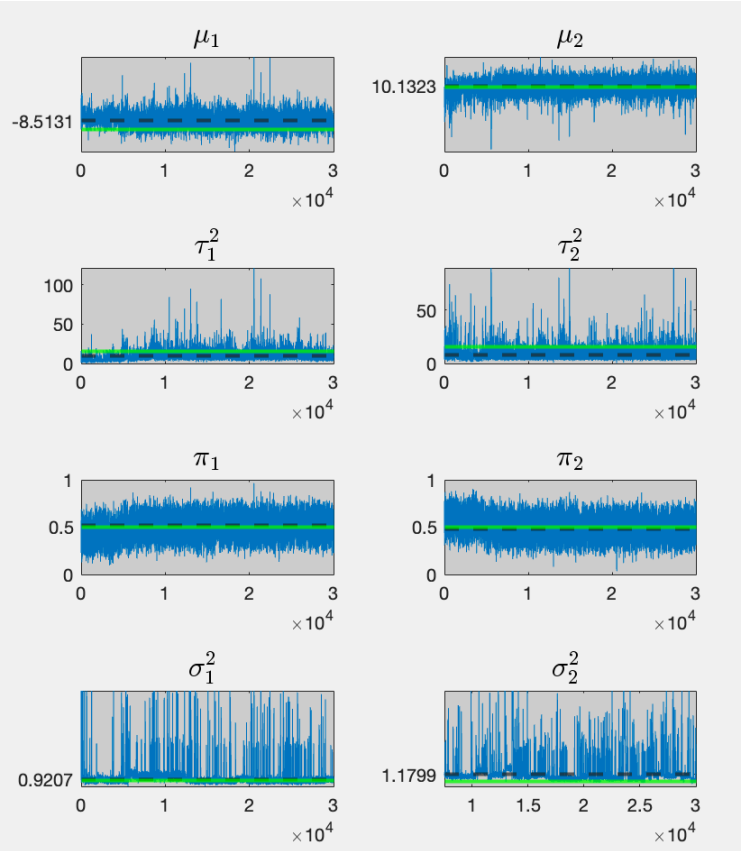


Figure 4.2.11 Underlying model parameter trace plots for example 4.2.2 at 30,000 iterations

Comparison of Gibbs Sampler with Least Squares Estimator for Example 4.2.2 Data

	Gibbs Sampler after 30,000 iterations	**Least Squares Estimator in Matlab**	**Actual Value**
ϕ_1	0.772	0.636	0.800
ϕ_2	-0.785	-0.780	-0.800
θ_1	0.746	0.537	0.800
θ_2	-0.621	-0.639	-0.800
σ_1^2	0.921	1.064	0.900
σ_2^2	1.180	1.085	0.900
μ_1	-8.513	-7.211	-10.00
μ_2	10.132	9.625	10.00

Table 4.2.3 Comparison of Gibbs Sampler, Matlab least squares estimator, and true parameter values when AR and MA parameters differ.

Note that we see potential convergence to nearly the exact value of θ_2 in the trace plot as we approach 30,000 iterations, but we have used a more conservative figure here, taking the mean after discarding the burn-in period. The MSE for groups 1 and 2 are given in table 4.2.4. The Gibbs sampler actually out-performs the least squares estimator for group 1 and performs nearly as well for group 2.

Combined segments centered at their respective sample means for groups 1 and two are shown in figure 4.2.12. Not that we have less data for group 1 because the segments were randomly generated.

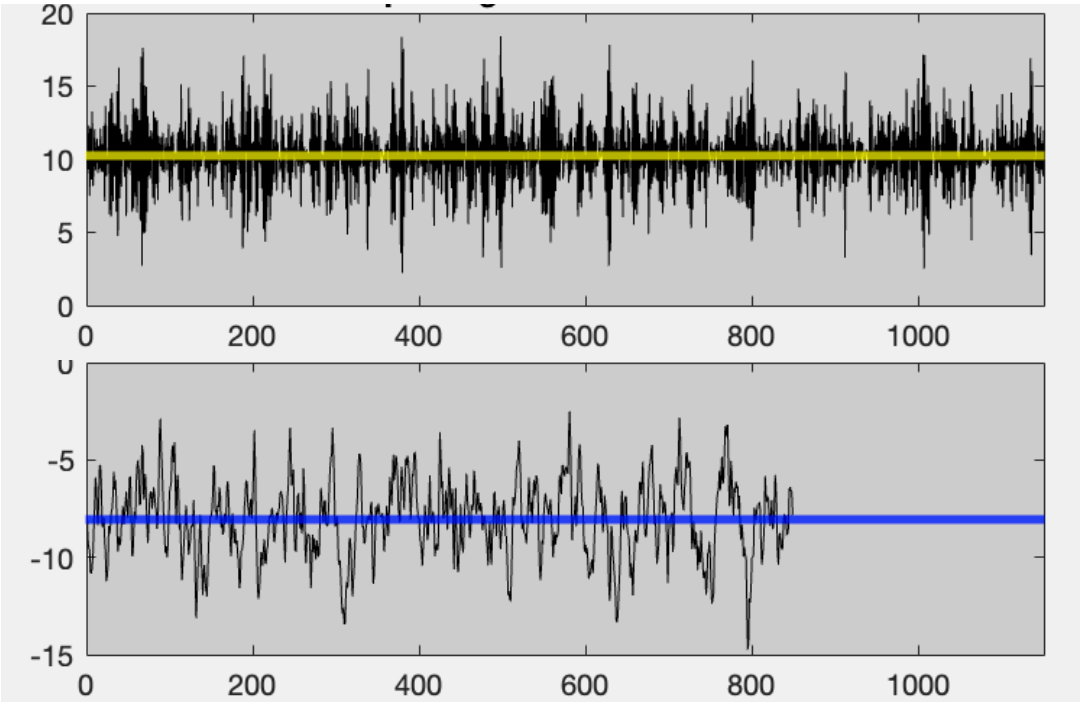


Figure 4.2.12 Group 2 segment data combined (top) and group 1 data combined, (bottom).

	Gibbs Sampler	**Least Squares Estimator in Matlab**	**Using true ARMA parameters**
MSE Group 1	0.089	0.242	0.072
MSE Group 2	8.958	8.839	7.707

Table 4.2.4 Comparison of the performance of the Gibbs Sampler vs a Least Squares estimation procedure for fitting an ARMA model to the data with widely varying AR and MA parameters.

4.2.2 Overlapping Groups

If data originate from two similar groups such that the main difference is expected segment mean, and if the segment mean variance is relatively high, our model will assign segments to groups in a somewhat random way. In cases like this, it may be appropriate to re-run the simulation under a one-group assumption.

Example 4.2.3 Relatively High Segment Mean Variance

We have the AR and MA parameters all set to 0.2, the group means are -5 and 5 for groups 1 and 2, respectively. We see, at first, from the yellow and blue line segments in figure 4.2.13, that more segments are considered to be part of group 1 than group 2. By the 9000^{th} iteration, almost all segments are classified as being group 2. But by the 10000^{th} iteration, we are back to a group assignment somewhat similar to what we see early on in the search algorithm, although the changepoint locations are more accurately determined after 10000 iterations than 3000.

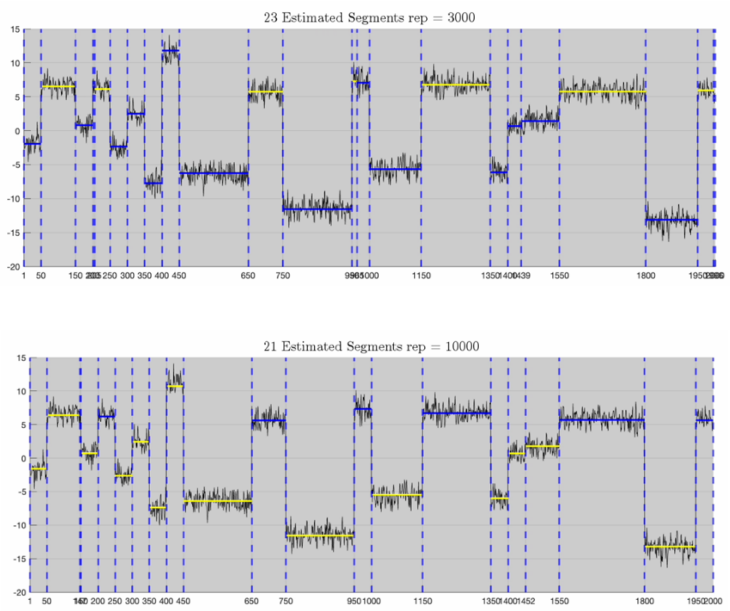


Figure 4.2.13 Data series with high segment mean variance by group. Group assignments are flipped going from iteration 3,000 to iteration 10,000.

Trace plots of AR and MA parameters for example 4.2.3

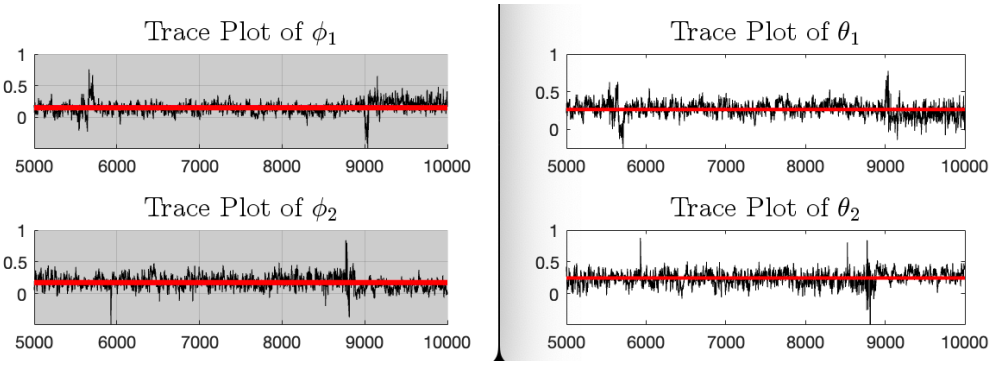


Figure 4.2.14 Trace plots of the AR and MA for example 4.2.3.

Rerunning the Gibbs sampler for the example 4.2.3 data under a one group assumption gives us identification of all changepoints while also providing more stable and accurate estimates of the ARMA parameters. Relatively stable results are seen when we apply the

model under a one-group assumption when data come from groups with similar properties and segment means vary widely.

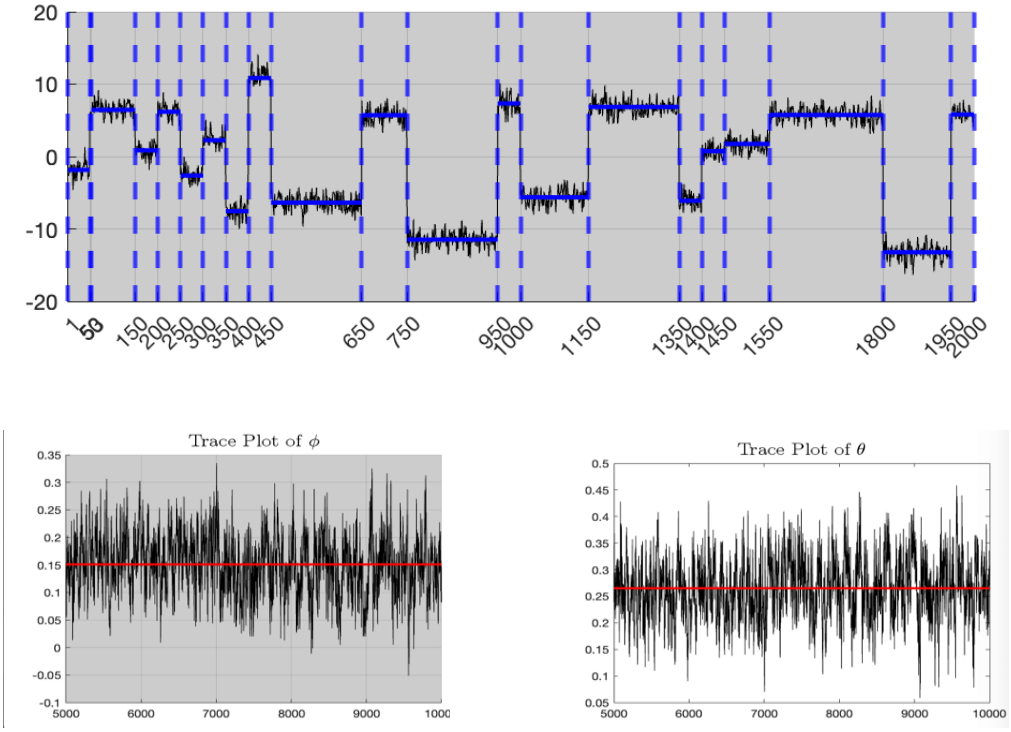


Figure 4.2.15 Changepoints and trace plots of single group AR and MA parameters for example 4.2.3

Example 4.2.4 Heavy Overlap with No Apparent Groups

When data come from two groups with the same segment mean, it may be difficult to even see a clear view of where the actual changepoints are located. For example, suppose data technically come from two different groups, but the group means and variances are so similar that it may be difficult to tell the difference between groups. Here the data originate from two groups with identical ARMA parameters set at 0.2 for both the MA and AR parameters, and random noise variance is set at 0.9. If groups were to differ only in expected segment mean, for example a mean of -0.5 for group 1 and expected segment mean of 0.5 for group 2, and moderate segment mean variance of 1 for both groups, we

may end up with a series that looks something like we see in figure 4.2.15 on the following page. The true changepoint locations and their respective groups are shown in the upper left graph. The lower left graph shows the results of 10,000 iterations of the Gibbs sampler using a one-group model. We get relatively stable convergence of the ARMA parameters, although we do miss some of the changepoints. For example, in the left segment, we do not detect any significant changepoints. This does still seem reasonable, as our model performs best at detecting changes in segment means. The sampler performs well when estimating the other parameters of this data, as well. For brevity we omit these plots.

Changepoint Detection Using One Group Approach. Results for Example 4.2.4

Data that originate from two groups with nearly identical properties and relatively little fluctuation in segment means. A one group model does fairly well. When segment means are too close together, changepoints are not detected, as seen in the left part of the graph in the lower left corner.

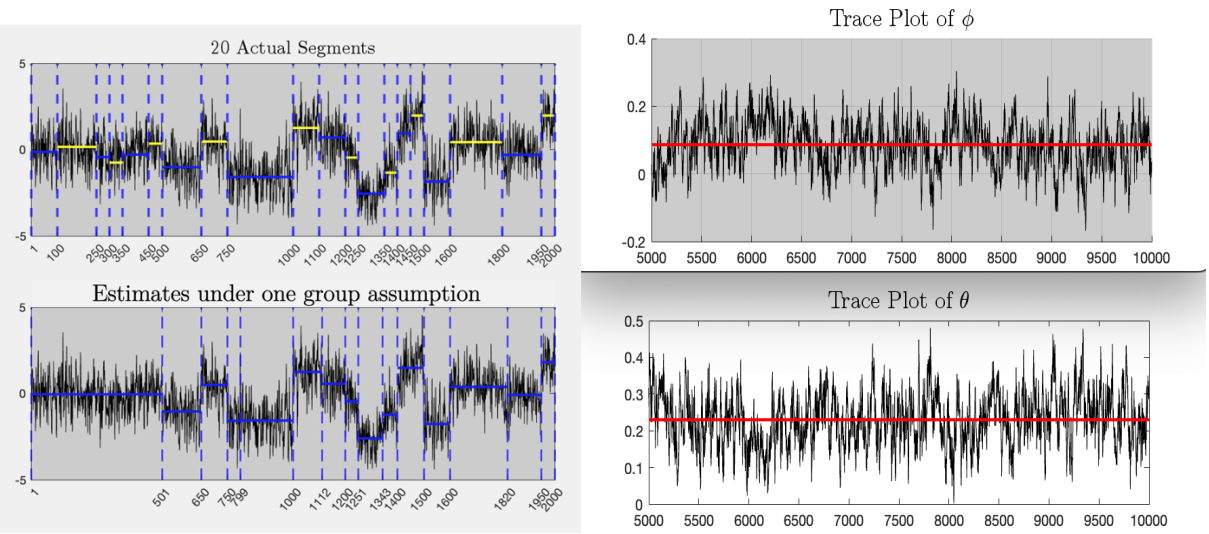


Figure 4.2.16 Results of one-group model applied to finding change points in a data set that has groups that are difficult to identify.

Discrepancies Due to Erroneous Equal Noise Variance Assumption

If we re-run the simulation from example 4.2.1 using the model structure of [Sadia et al., 2018] or [Fearnhead, 2006] we over-identify changepoints with this data series due to inability to tell the difference between group error variances. We end up with a random noise variance estimate that appears close to the mean of the two true variances. The top graph in figure 4.2.17 shows overidentification of changepoints in the data from example 4.2.1 when the model is set to assume that random noise is equal by group. The bottom row shows trace plots of the estimate random noise error variance from the original simulation from example 4.2.1 (bottom left two plots), and the trace plot of random noise variance when the model is set to assume that noise does not vary by group. The green line in the bottom right plot indicates the location of the true variance for group two.

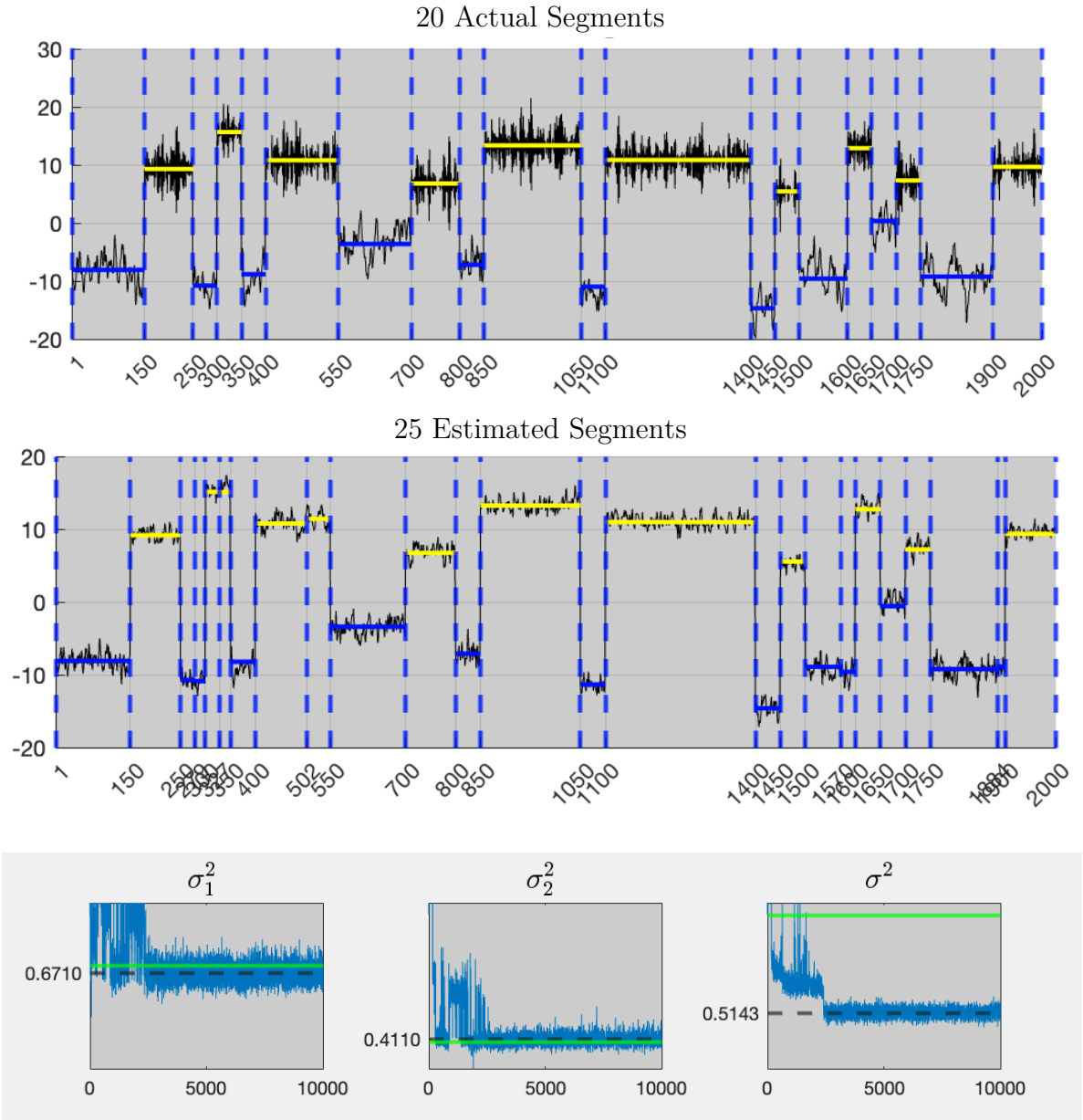


Figure 4.2.17 Changepoint detections and trace plots of true error variance vs estimated error variance when we assume that groups have equal variance.

Addressing the Double Insertion Issue Using the Adjusted Insertion Rule

In chapter 2, we discussed the concept of the double insertion of points inherent to the original model. With our model, we control for this by applying what we call the Adjusted Insertion Rule. If a changepoint is deleted, we do not allow the changepoint to be immediately re-inserted on the subsequent insertion step, as is the case with previous models. If we were to allow points to be re-inserted immediately after deletion, it would increase the chance of adjacent changepoints occurring, and it may interfere with our ability to accurately estimate parameters. We will re-visit the data set from example 4.2.2. It was already difficult to get convergence to the correct parameter values in that example. We now present the results of an attempt to estimate parameters without correcting for double-insertions. Not only are the trace plots less stable, even after 30,000 iterations, but we also erroneously detect changepoints that are too close together, including an adjacent changepoint.

On the left are the results of the trace plots of the ARMA parameters from example 2, in which the Adjusted Insertion Rule was applied. Figures are zoomed further in. On the right are the trace plots when we do not control for double insertion.

We notice that in both cases it is difficult to get convergence in this example, given the high magnitude of the parameters. But on the right column, we see that we have even less precision, even after 30,000 iterations. Additionally, we have a higher chance of ending up with adjacent changepoints, as seen in figure 4.2.19.

We see several changepoints that are fairly close to each other, and one case in which two change points are adjacent to each other at around 1650. This is more likely to occur when we do not control for double insertion, especially when dealing with data sets that have large magnitude ARMA parameters.

Original trace plots for Example 4.2.2 simulation are shown in the left column of figure 4.2.18. The analogous plots when we do not adjust for double-insertion are shown in the right column. The trace plots are less stable. We also see evidence of near exact convergence of θ_2 to the correct value when we approach 30,000 iterations, although it also appears that we may need even more iterations than this.

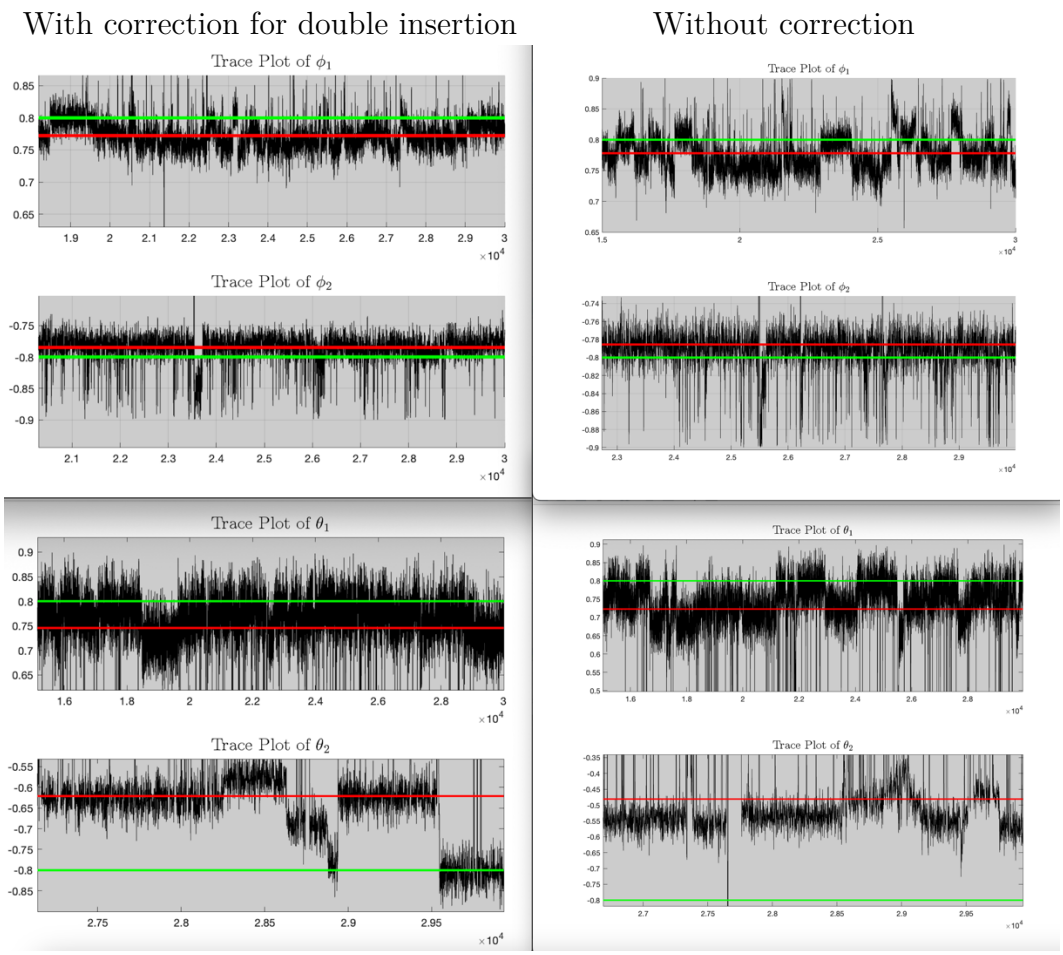


Figure 4.2.18 Trace plots of AR and MA parameters after 30,000 iterations for example 4.2.2 data with Adjusted Insertion Rule applied (left column) and without Adjusted Insertion Rule applied (right column).

Figure 4.2.19 illustrates what can happen when we do not adjust for inadvertent reversal of deleted points following a deletion step. The bottom graph shows that we end up with adjacent changepoints at 1650 and 1651, which falsely indicate that a segment with unique properties of length 1 exists at this location. We also get a very unexpectedly short segment nearly 1450. In neither case does this occur due to outliers.

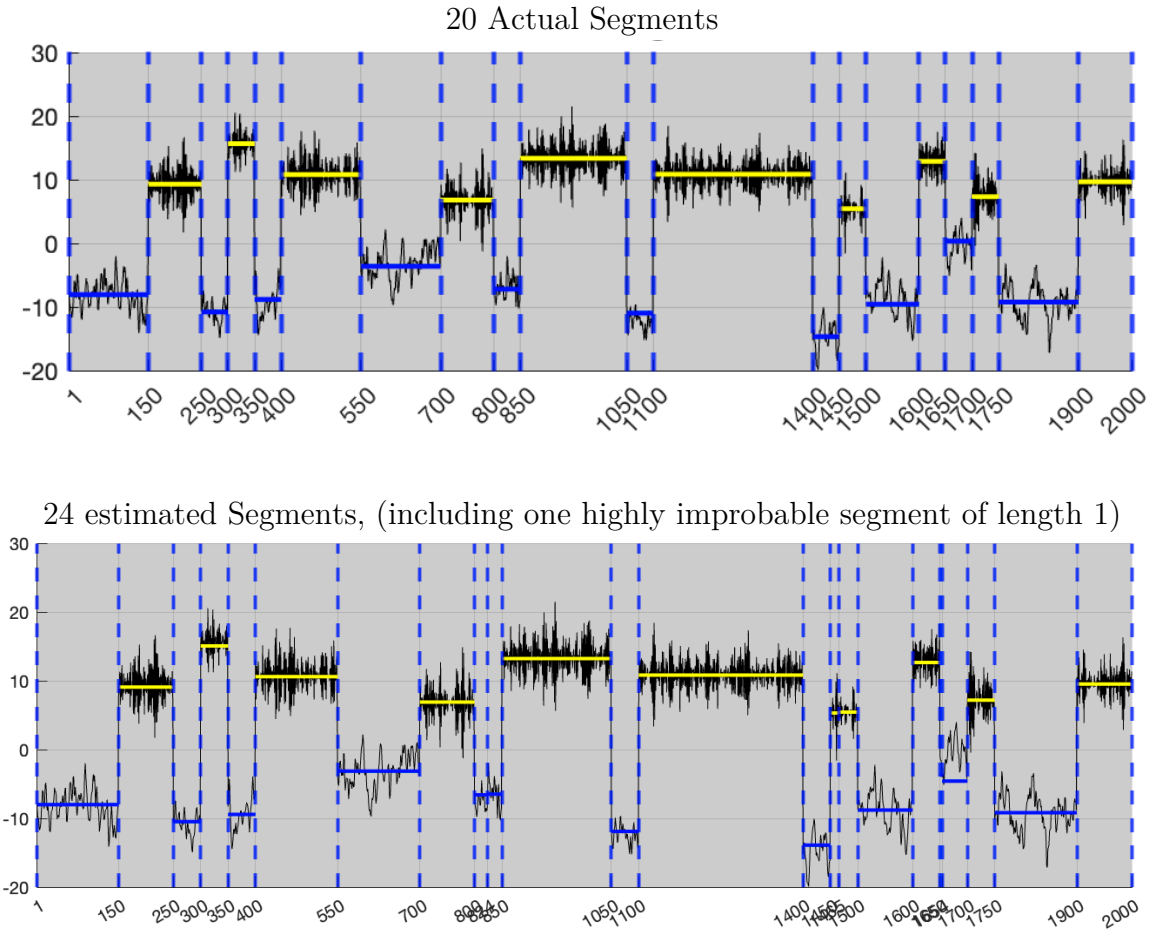


Figure 4.2.19 Illustration of adjacent changepoint issue due to a potential reversal of the deletion step when the Adjusted Insertion Rule is not applied. The data used are from example 4.2.2.

4.2.3 Fixing Parameters

Tighter Bounds for ϕ and θ

We may want to include a provision in the code to prevent unrealistically high or low parameter estimates while running the code. We can bound ϕ and θ to a tighter interval, for example $(-0.9,0.9)$. This is reasonable when post-simulation data indicate that neither parameter is high in magnitude. Allowing these parameter to vary too much could result in instability, with random estimates of our parameters reaching the very edge of the boundaries early on in simulations. After re-running the simulation on the same data set from example 4.2.1, we get convergence to the true parameters in only about 500 iterations when bounding ϕ_2 to $(-0.9,0.9)$.

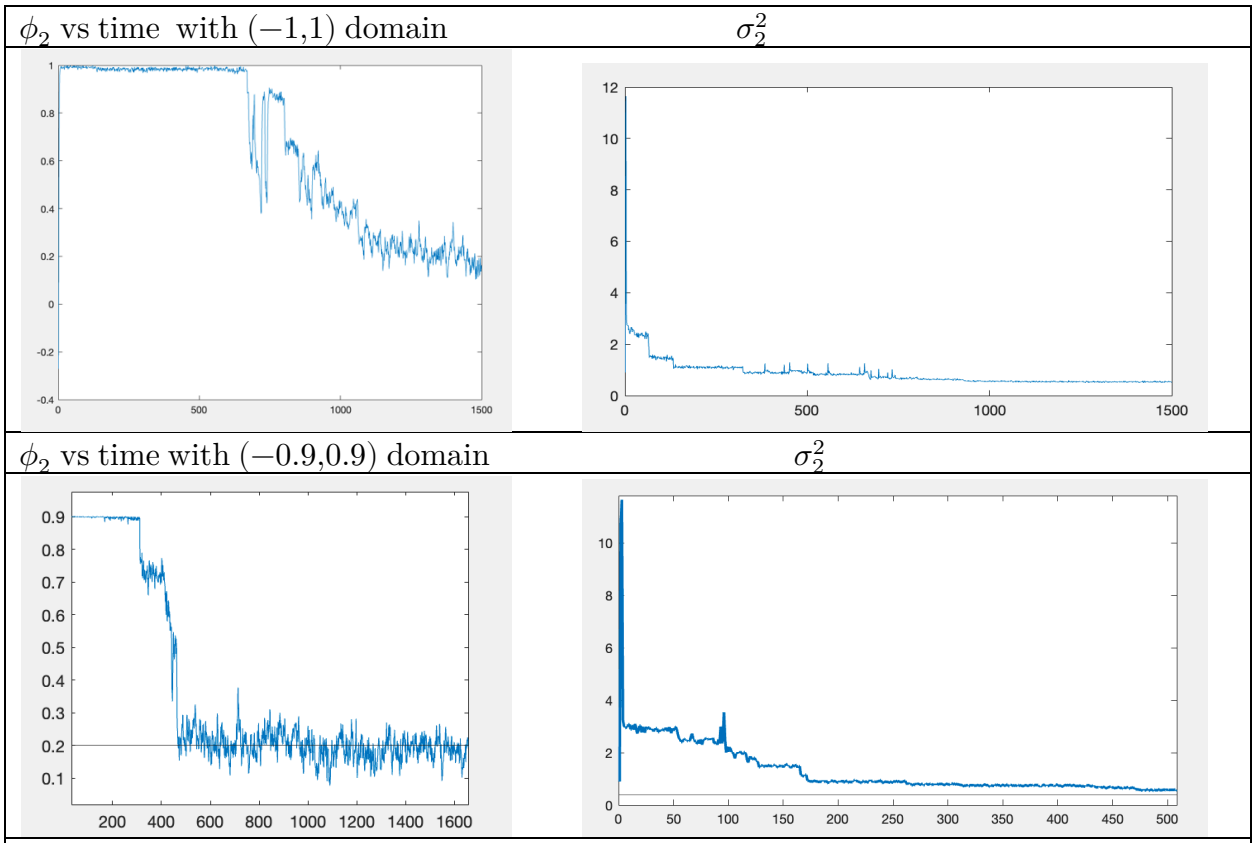


Figure 4.2.20 Trace plots showing faster convergence when bounding the AR parameter.

We may also want to fix π_1 and π_2 if we believe that we likely have a similar number of groups. During the Gibbs sampling procedure, it makes sense to simply fix these to be equal, reducing the required computation without compromising the ability of the algorithm to sample equally from both group 1 and group 2. The results of a simulation with the same data set from example 4.2.1 are shown. We did mis-identify one change-point, but this is very subtle, and it occurs at a location at which the data appear to have a non-linear trend. This is not an unexpected occurrence.

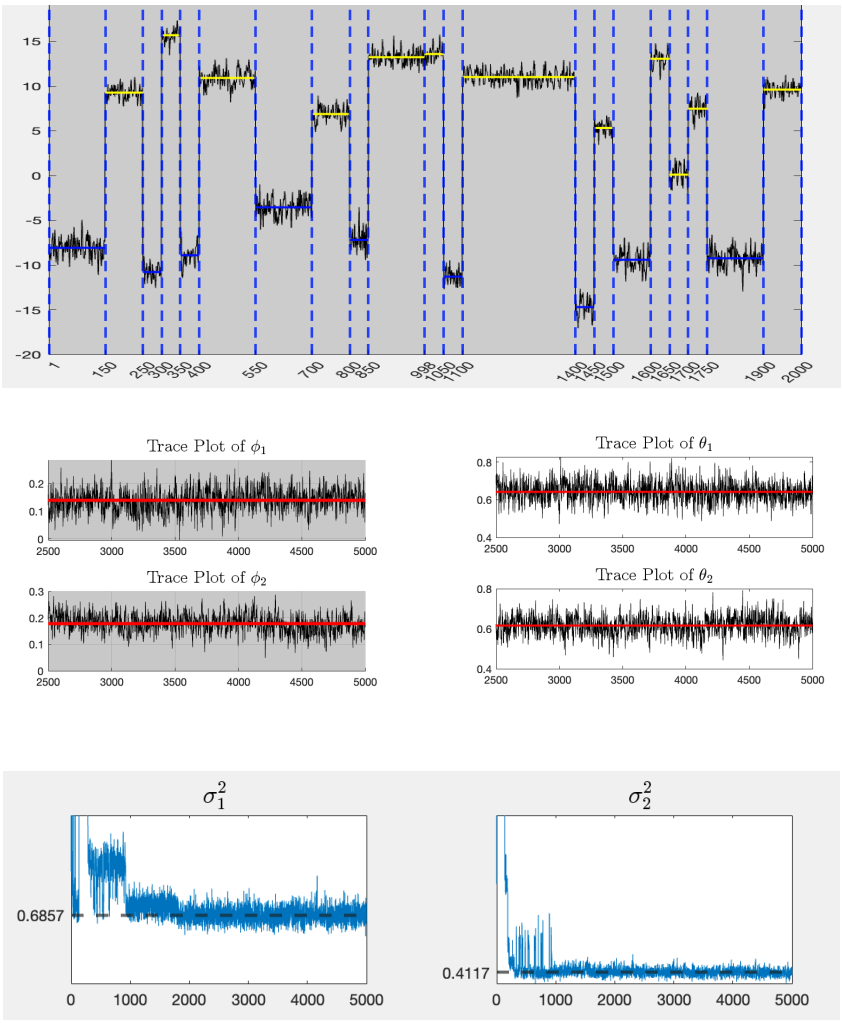


Figure 4.2.21 Results of re-running the simulation from example 4.2.1 with group probabilities fixed to be 0.5.

4.3 Well-log Data – Existing Model Results

The term "well-logging" broadly refers to the continual collection of data during a drilling process, often well drilling, hence the name. The pioneering of well-logging is attributed to the Schlumberger brothers in the 1920s [Bowker, 1973]. Their offshore drilling company has grown to be the world's largest in terms of annual revenue, as of 2022 [Joshi, 2022], in part due to the advantages gained through continual data collection. One such modern method of well logging involves nuclear magnetic resonance response to drilling as a means of measuring porosity of rock strata and identifying points at which the composition of the strata change. Abrupt changes in measurement can sometimes yield data that exhibit characteristics of a segmented ARMA process. The specific data set of interest was originally gathered by the Schlumberger corporation and was first analyzed from a Bayesian perspective by [Ó Ruanaidh and Fitzgerald, 1994]. We simply refer to this as the well log data. We chose this data set in part because it appears so frequently in Bayesian changepoint detection studies, including that of [Sadia et al., 2018]. The results of the application of the model produced by [Sadia et al., 2018] are presented here.

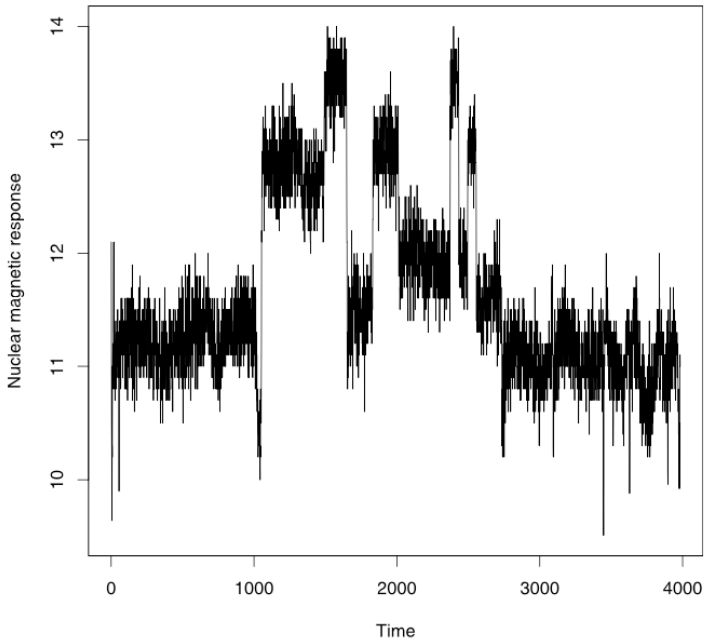


Fig 8. Well-log data. This data provides information about the rock structure of the well. Some change-points are present in the data reflecting the presence of a new rock.

https://doi.org/10.1371/journal.pone.0208927.g008

Figure 4.3.1 Well log data as presented by Sadia et al., 2018, page 15.

Figure 4.3.1 is the image of the data set they present, with about 50 outliers removed. They also compare the changepoints detected using their model, as highlighted in green in the image, as well as with simplified versions of the model. By holding the MA parameter fixed at zero, an AR model is produced. An MA model is produced by holding the AR parameter fixed at zero. Figure 4.3.2 shows the results of these three model applications. We made a composite image of the three outputs from figure 4.3.2 in order to more-directly view the overlapping and non-overlapping results. The composite image is shown in figure 4.3.3. There is quite a bit of agreement at many of the more prominent locations, while none of the models agree much in terms of changepoints detected below time 1000 or above time 3000. Interestingly, the MA model interprets just as many locations to be changepoints as the ARMA model, although there is little agreement in the extreme ends of the data set. Overall, their model appears to perform well, especially when identifying

changepoints toward within the range of 1000 to 3000. Apparently this was analyzed under a one-group modeling assumption. Trace plots of parameter estimations are not provided. Results for AR(1) model are similar to those of [Fearnhead, 2006].

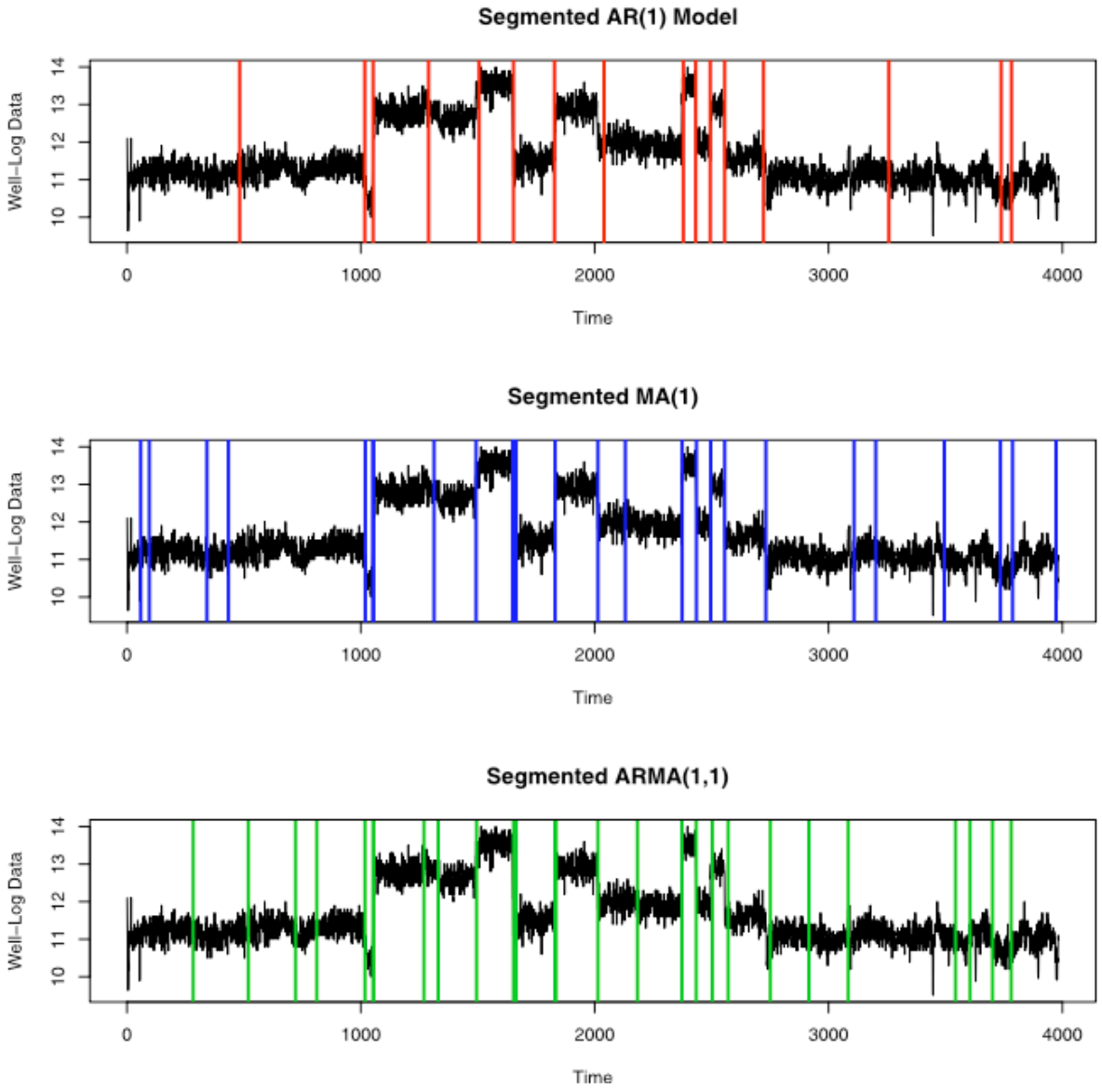


Figure 4.3.2 Changepoints detected by AR, MA, and ARMA versions of the model by Sadia et al., 2018, page 19.

Our Composite Image of the Results of [Sadia et al., 2018]

The light pink represent AR detected changepoints that do not align with those of other models. The green represent the ARMA changepoints that do not overlap with those of other models, and the blue represent the MA model changepoints. Purple represent overlap between AR and MA, bluish-green represent overlap of ARMA and MA, dark green represents the overlap between ARMA and AR models, and the darkest lines indicate an overlap of all three. For the readers who may be viewing a greyscale version of this document, the darkest lines represent overlap, and the lighter lines represent changepoints that are unique to a particular model.

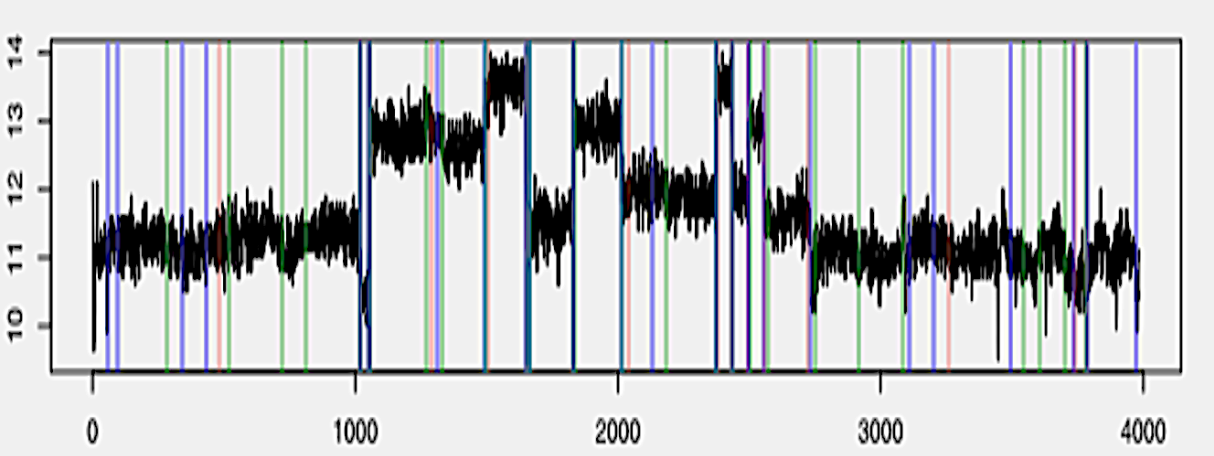


Figure 4.3.3 Composite image of the three images from [Sadia et al., 2018].

4.4 Well Log Data with Our Model Utilized

We apply our multi-group model as described in section to of chapter 2 to analyze the well-log data set. Our identification of changepoints is similar what is seen in 4.3. We do not detect nearly as many points in the extremes of the data set. The estimation of the parameters is not discussed by [Sadia at al, 2018], or by [Fearnhead, 2006]. Their results

are also not described in terms of a group-based model. We provide our results, along with trace plots of parameter estimates for the model parameters. We discover subtle differences in the parameters, by group, suggesting that groups may differ in terms of random noise. Vertical dashed lines in figure 4.4.1 represent segment endpoints. Horizontal blue segments represent group 1 segment means. Horizontal yellow indicate group 2. Note that dashed lines at the far right and far left indicate the end of the data set, while the interior dashed lines represent changepoints. Figure 4.4.2 includes trace plots of the group means, group variances, and error variances by group. The sudden shift that we see in the means is due to an algorithmic decision to re-define group 1 as group 2 and vice versa, which can sometimes happen before burn-in is reached.

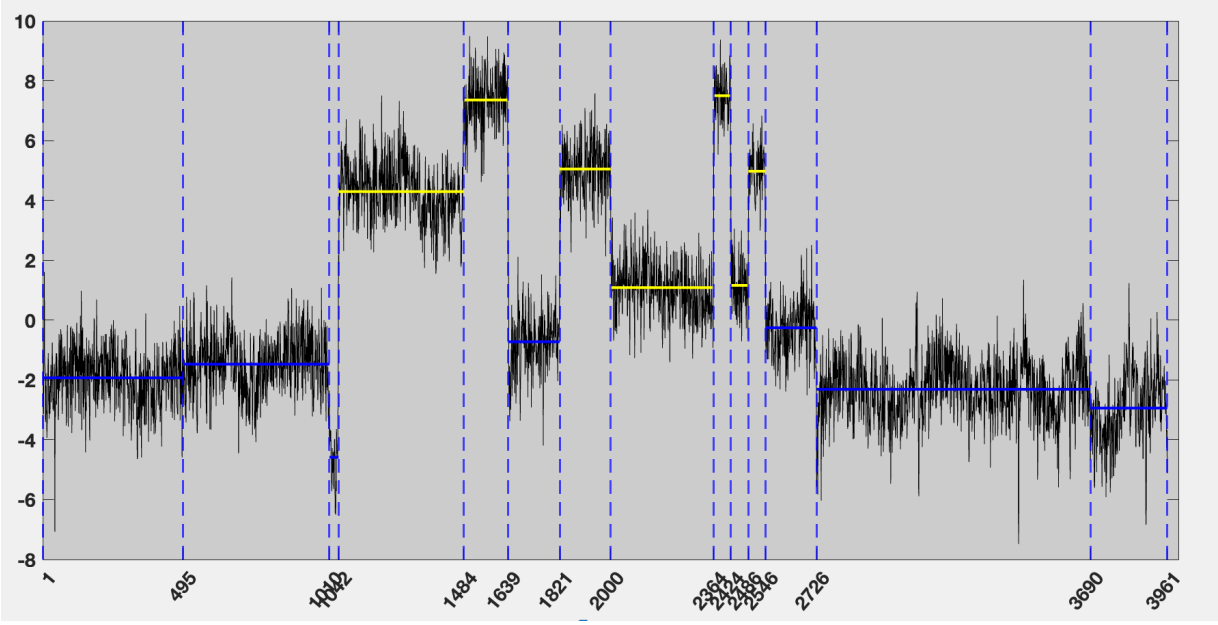


Figure 4.4.1 Initial changepoints detected from the well-log data using the two-group model with 10,000 iterations.

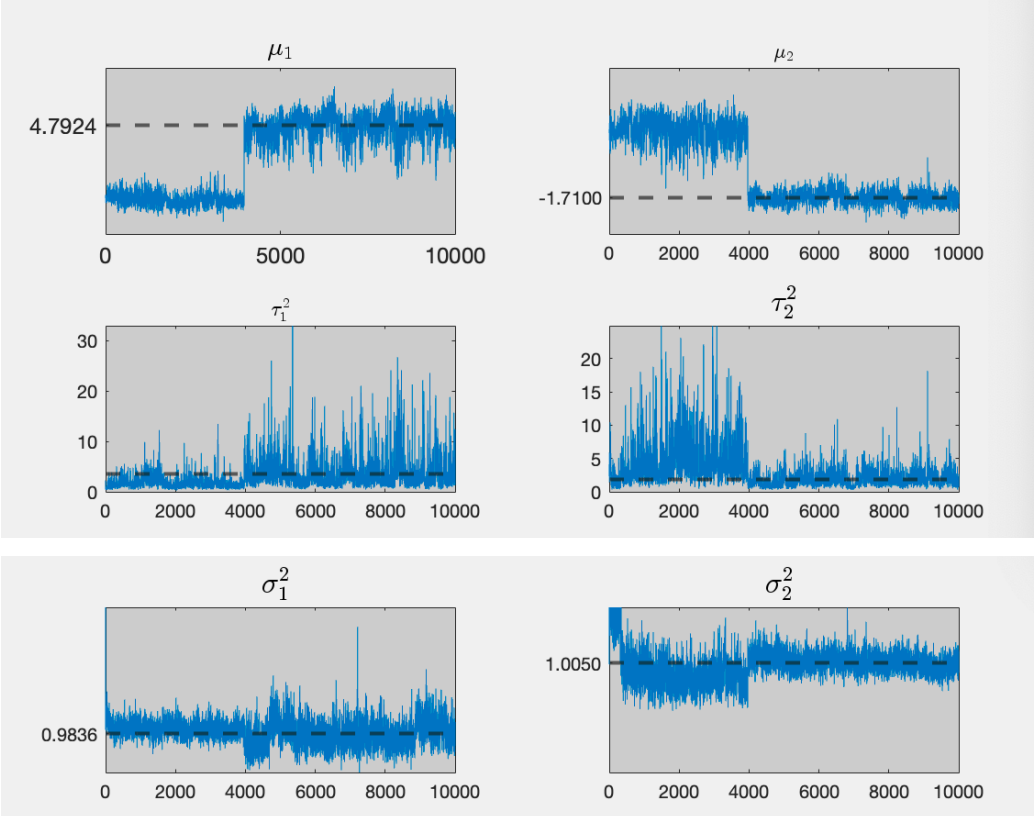


Figure 4.4.2 Trace plots of group parameters for our initial 10,000 iterations using the two-group model.

Figures 4.4.3 and 4.4.4 show trace plots for the MA and AR parameter estimates after running our two-group version of the model on the well-log data for 10,000 iterations. We get somewhat stable mean for the MA parameters, although variance in the trace plot seems to increase after 8000 iterations. We notice a similar increase in the variance of the trace plot of the group two AR parameter in figure 4.4.4. This is to be expected, as the full conditional posterior distributions of the AR parameters are influenced by the MA parameters and vice versa, as shown in Theorems 3.1 and 3.2.

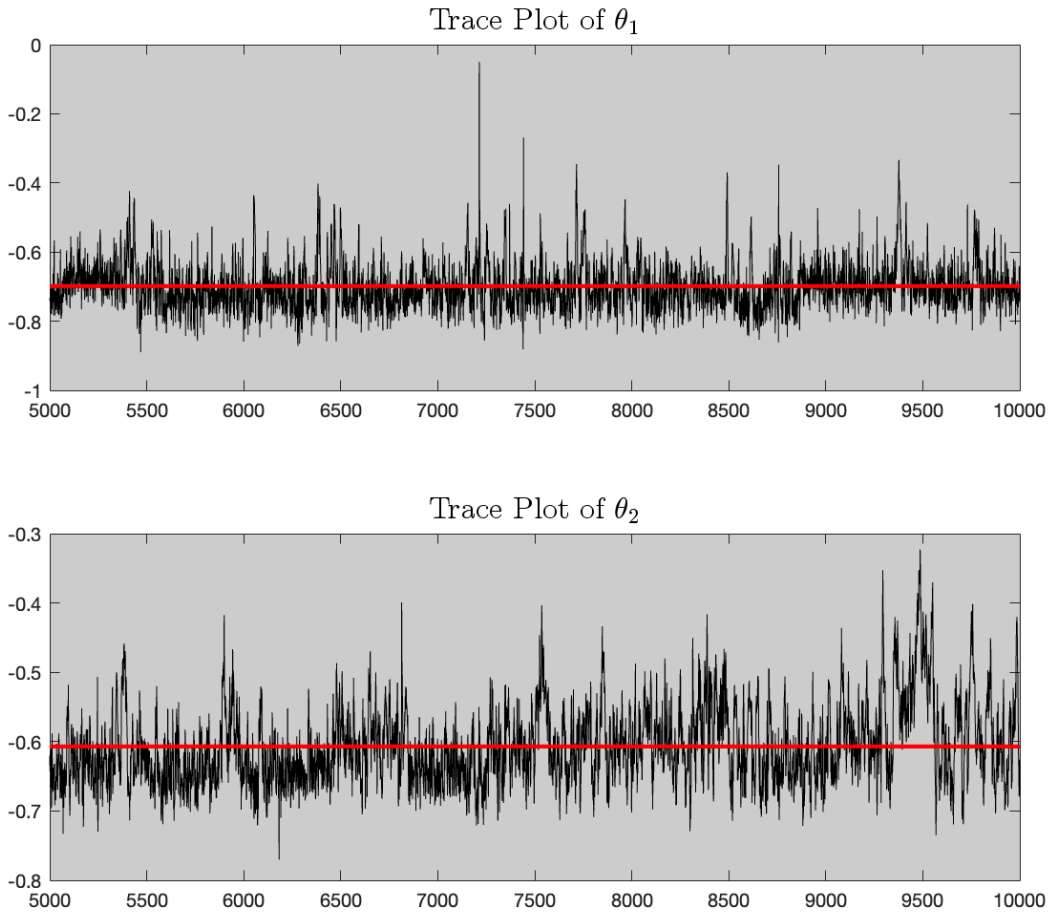


Figure 4.4.3 Trace plots for the MA parameters for the two-group model after 10,000 iterations.

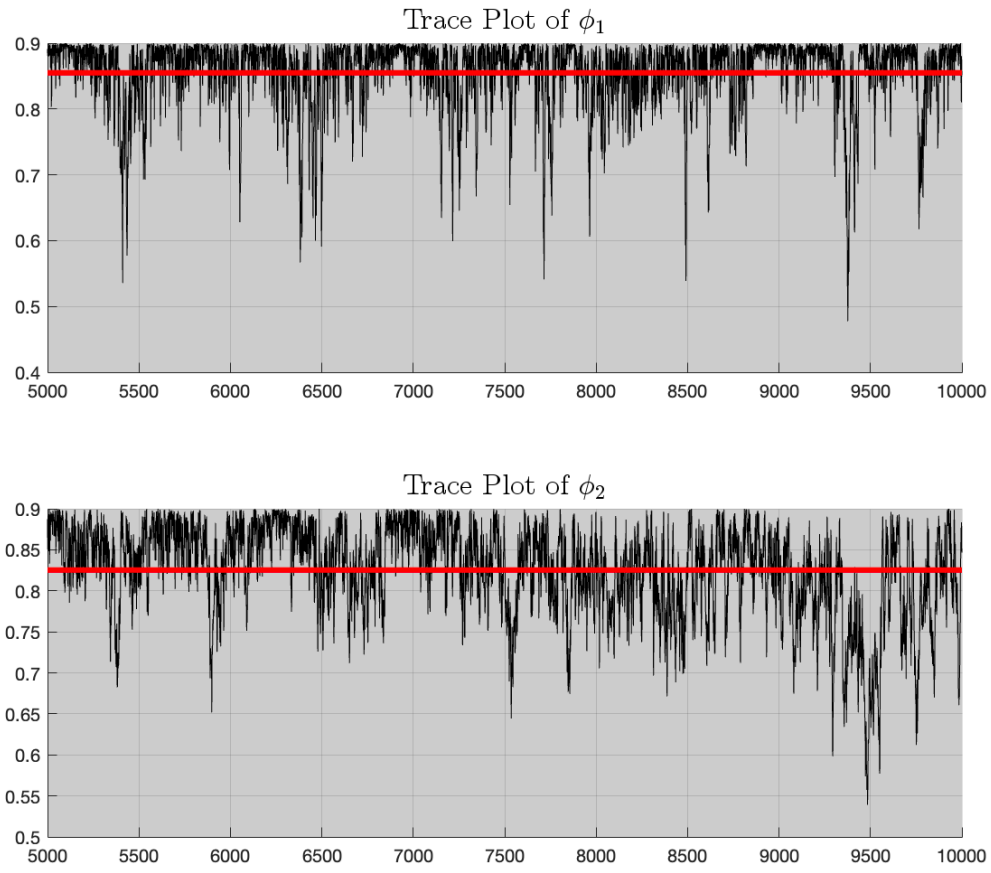


Figure 4.4.4 Trace plots for AR parameters after 10,000 iterations of the two-group model on the well-log data.

Parameter Estimates for Well Log Data

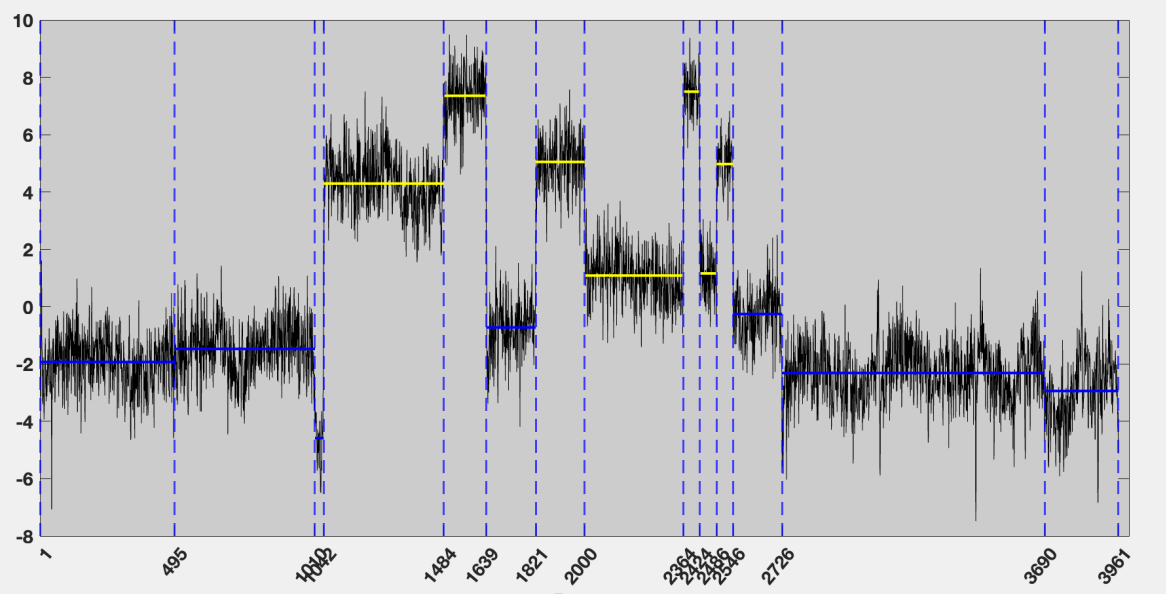
	Gibbs Sampler after 1000 iterations	**Least Squares Estimator in Matlab**
ϕ_1	0.855	0.934
ϕ_2	0.825	0.948
θ_1	-0.697	-0.807
θ_2	-0.606	-0.935
σ_1^2	0.683	0.671
σ_2^2	0.415	0.402
μ_1	4.792	4.529
μ_2	-1.710	-2.229

Table 4.4.1 Comparison of Gibbs Sampler, Matlab least squares estimator, and true parameter values.

Given the instability of the parameter estimates, we might want to consider performing a greater number of iterations. We will now demonstrate the results using 30,000 iterations. Changepoint detection is similar, but ARMA parameter estimates and underlying model parameter estimation produces more stable trace plots.

We get nearly identical changepoint detection with 30,000 iterations vs 10,000 iterations

10,000 Iterations



30,000 Iterations

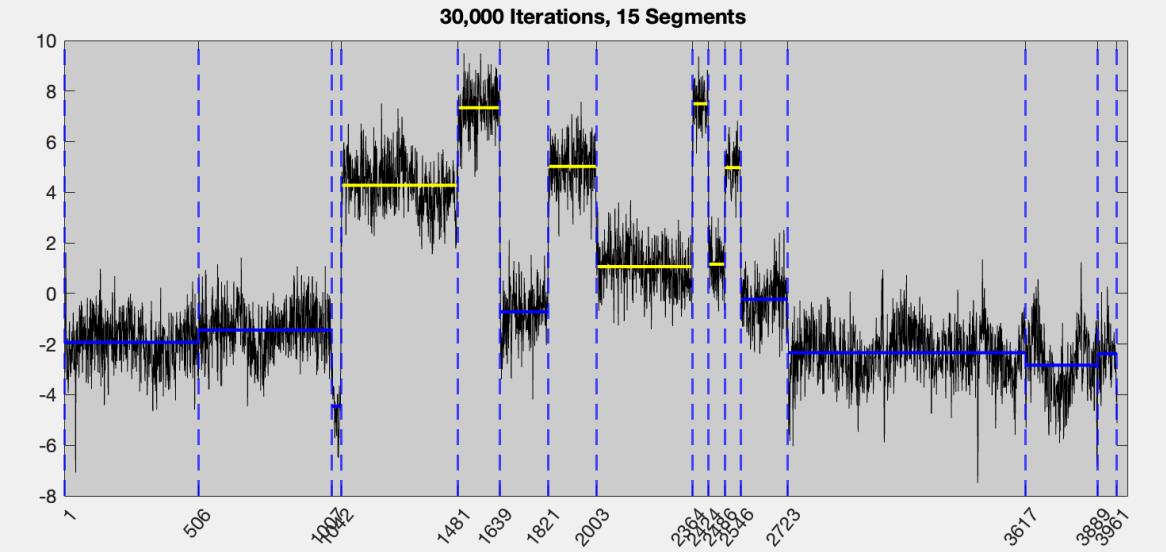


Figure 4.4.5 Results of 30,000 iterations comparted to 10,000 for the application of the two-group model to the well log data.

Trace Plots for two-group 30,000 iteration trial with Well-log Data

Much more stable estimates are found for the parameters, although the estimates for group variance are highly skewed. It is possible that this is causing us difficulty in identifying changepoints in group1, as the data do not vary much in group 1.

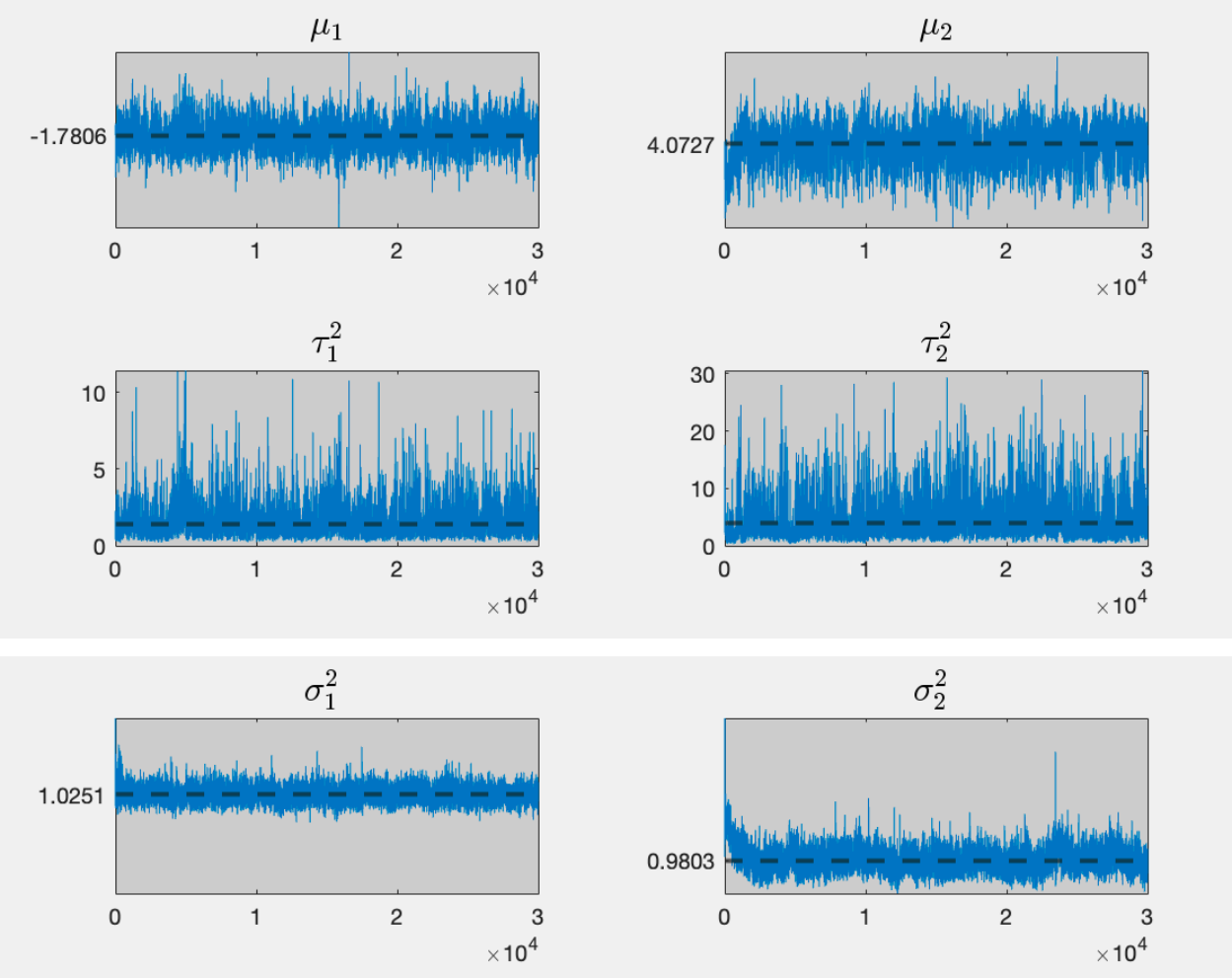


Figure 4.4.6 Trace plots of parameters after 30,000 iterations.

We see better stability of the trace-plots for the moving average parameters after 30,000 iterations.

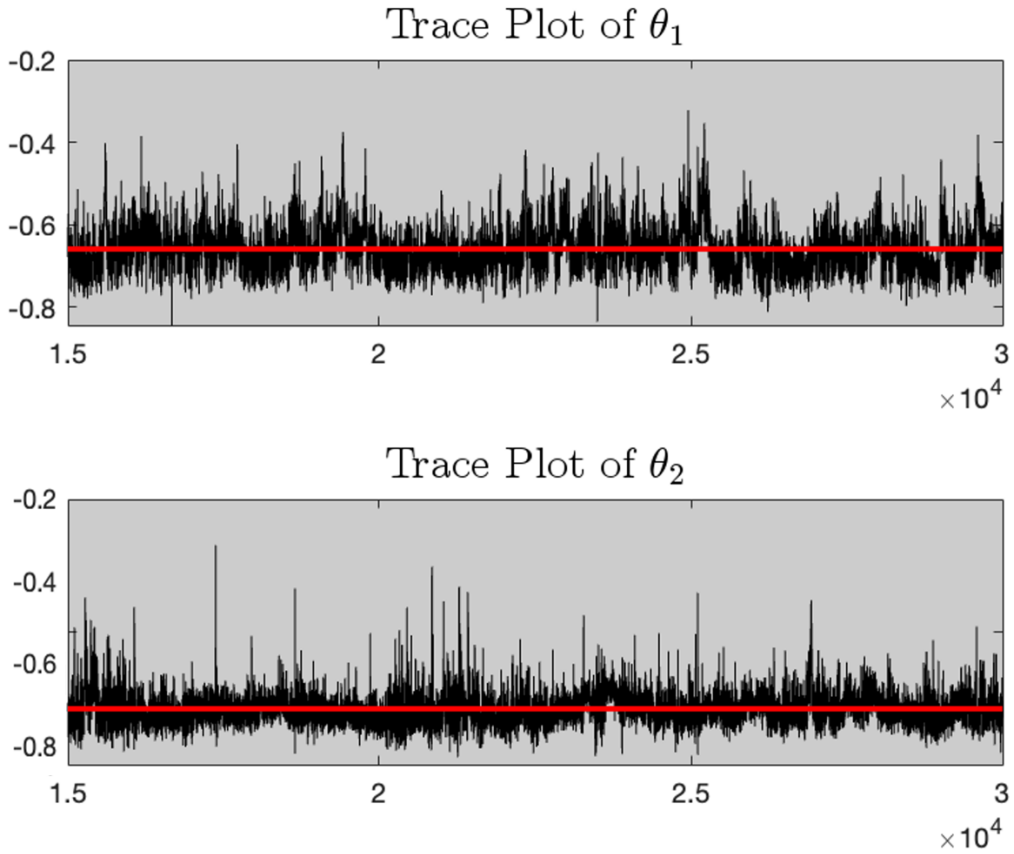


Figure 4.4.7 Trace plots of moving average parameters for well-log data after 30,000 iterations.

The AR parameters estimates are still not stable after 30,000 iterations, although they are more tightly bound upward compared to the 10,000 iteration case. We have bounded the parameter above by 0.98. This is essentially a folded image of what otherwise would look like a segment of data with a high AR parameter. The sequence of estimates for the AR parameter, is, itself, similar in shape to data from an AR process with very high magnitude negative parameter.

Two Group Case

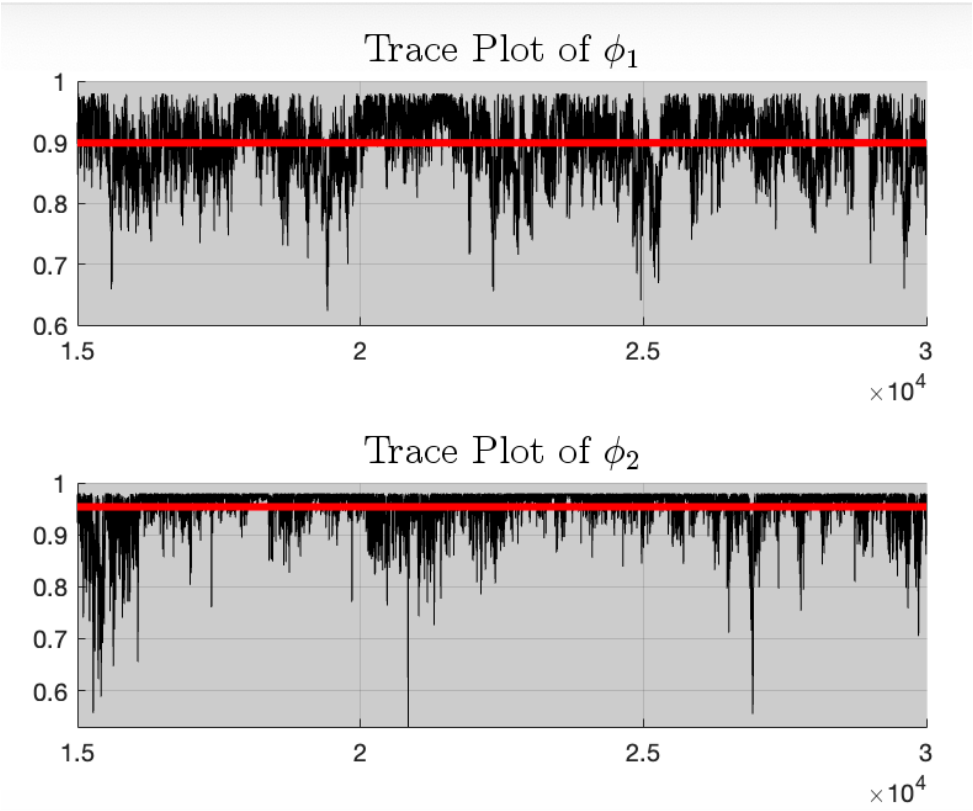


Figure 4.4.8 Trace plots for the AR parameters after 30,000 iterations on the well-log data.

Results after 30,000 Iterations

	Gibbs Sampler after 30,000 iterations	**Least Squares Estimator in Matlab**
ϕ_1	0.899	0.945
ϕ_2	0.954	0.982
θ_1	-0.659	-0.807
θ_2	-0.790	-0.823
σ_1^2	1.025	0.671
σ_2^2	0.980	0.402
μ_1	-1.781	-2.223
μ_2	4.073	3.861

Table 4.4.2 Comparison of Gibbs Sampler, Matlab least squares estimator, and true parameter values after 30,000 iterations.

Changepoint Detection using MA Models

Given the instability of the AR parameter trace plots, and given that the results of [Sadia et al., 2018] seem to indicate good performance with a one-group MA version of the model, we tested our model multi-group model from chapter 2, section 2, in one and two-group form while holding the AR parameter fixed at zero. We get nearly identical results with the one and two group models. The most elusive area is around 1320. Surprisingly, this point is not detected with high probability with any of our models involving the AR parameter. This is in contrast to [Sadia et al., 2018]. All of their models detect a point in this general vicinity.

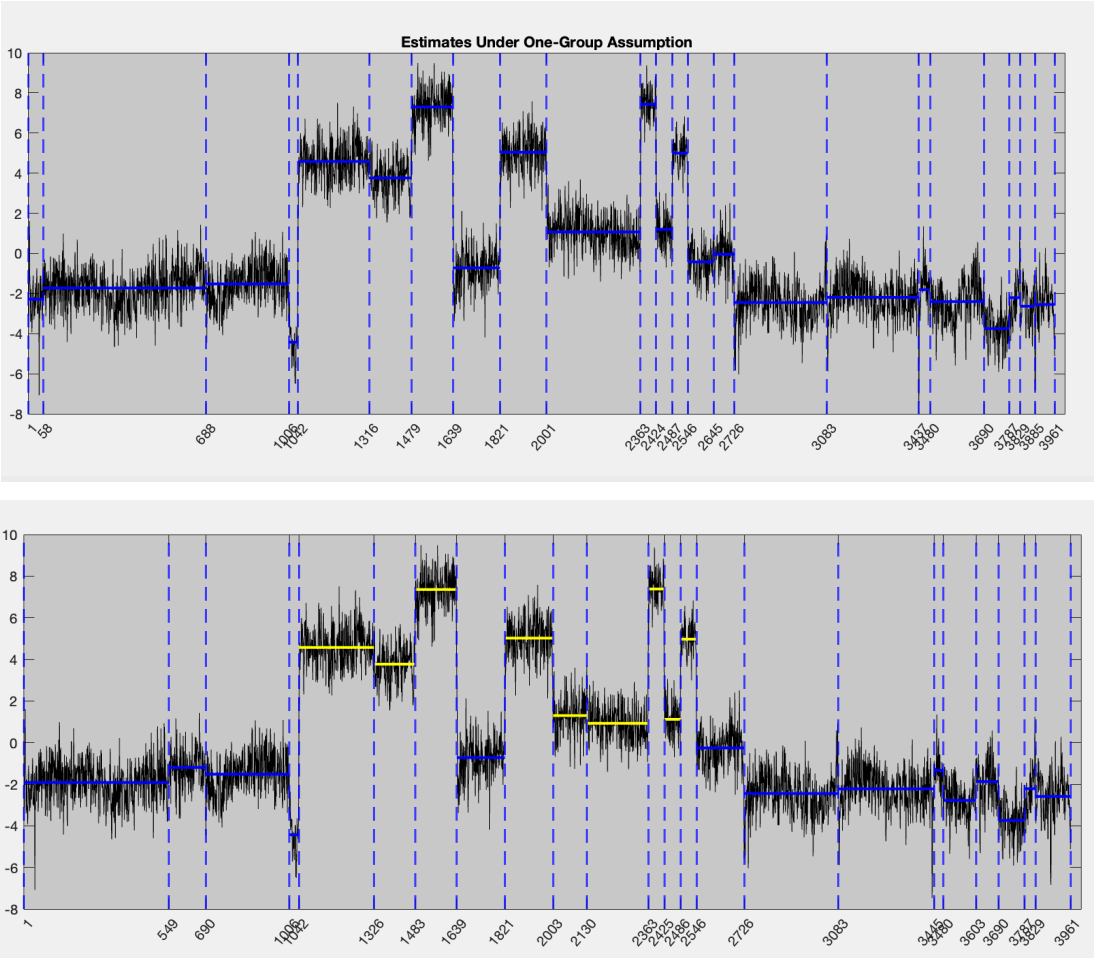


Figure 4.4.9 Changepoints detected after 10,000 iterations of the one and two group MA-only models, shown in the top and bottom graphs respectively.

We have a somewhat stable convergence to a MA parameter estimate when we consider the entire set to be one group, although we have some apparent periodicity, as shown in the first plot in figure 4.4.10. The convergence is very stable in the two group cases with 10,000 iterations. Interestingly, the magnitude of the MA parameters is much lower when we exclude the AR component of the model. From figures 4.4.6 on page 145 and 4.4.6 on page 149, we see that the MA parameters are estimated to be at least 5 times greater in magnitude. This is likely due to the fact that the full posterior conditional distribution of the MA parameter no longer depends on the AR parameter when the AR parameter is fixed to zero.

Trace plots for One and Two Group MA Models for Well-Log Data

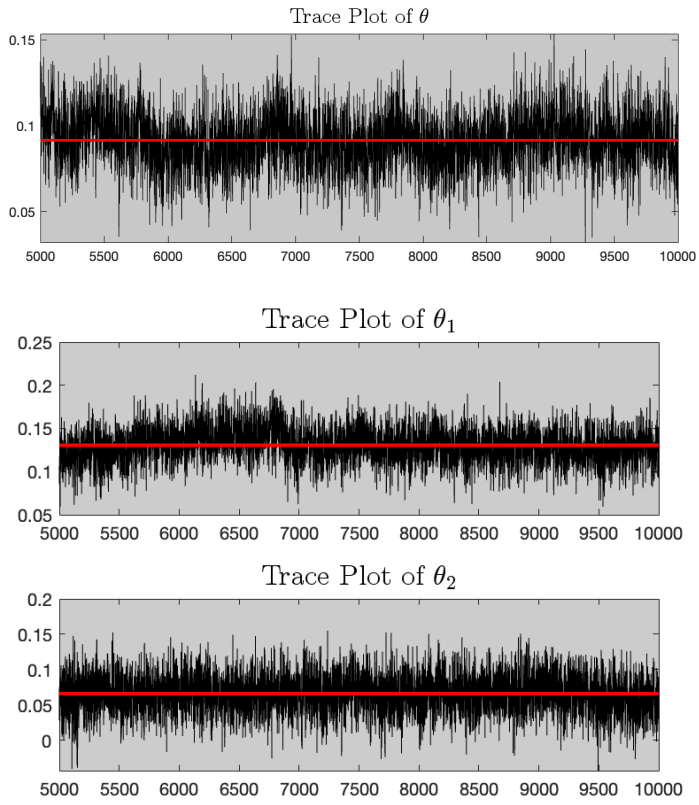


Figure 4.4.10 Trace plots of the MA parameter for the one-group MA only model (top row) and the MA parameters for the two-group MA-only model (bottom two rows).

The parameter estimates for the one and two group MA parameters are compared to those of the least squares estimates (tables 4.4.3 and 4.4.4, respectively). Notice that the group means are in much closer agreement in the two-group model, as shown in table 4.4.3. This is likely because we have detected many more changepoints in the right end of the data set than we did with the multi-group ARMA model, all of which seem to have means lower than the data in the left side of the data set. This could indicate that this model would appropriately be modeled as a three-group model, with the third group represented by the data from about 2726 onward, although this is beyond the scope of this project.

One and Two Group MA Process vs Least Squares Estimator

	Gibbs Sampler after 30,000 iterations	**Least Squares Estimator in Matlab**
θ	0.089	0.1274
σ^2	.918	0.9311
θ_1	0.130	0.159
θ_2	0.066	0.0475
σ_1^2	.918	0.880
σ_2^2	0.880	0.402
μ_1	-2.036	-2.223
μ_2	3.703	3.861

Table 4.4.3 Gibbs Sampler vs Matlab least squares estimator.

One Group MA Process vs Least Squares Estimator

	Gibbs Sampler after 30,000 iterations	**Least Squares Estimator in Matlab**
θ	0.089	0.128
σ^2	0.918	0.931

Table 4.4.4 Comparison of Gibbs Sampler, Matlab least squares estimator of the MA parameter and random noise with the one-group only MA model.

4.5 Final remarks on simulations and application to well-log data

Our model performs in a manner similar to models we have studied by [Sadia et al., 2018] and [Fearnhead, 2006] on simple data sets. Additionally, we have the ability to accurately estimate the parameters of a two-group ARMA model using the Gibbs sampling procedures we have discussed. We have demonstrated this with several examples. Our parameter estimates are similar to what we would get were we to know the correct locations of the segments ahead of time, using a least squares estimator for ARMA parameters. We have also resolved an issue that previously arose due to the possibility of double-insertion and immediate reversal of deletion within the sampling loop by applying the Adjusted Insertion Rule. This has resulted in far fewer single-point segments due to lower acceptance of adjacent changepoints. This may actually be a more useful contribution than the production of the multi-group model itself, in some ways, because the possibility of adjacent changepoints is problematic, whether our model includes provisions for multiple groups or not. Our model does not seem to be as sensitive to changepoints in the tails of the well-log data as that of [Sadia et al., 2018] when we use a multi-group ARMA model. It takes many iterations to get stable trace plots for the AR

parameters, and the estimates for the AR parameters are very high. None of the authors we reviewed produced trace plots, so we do not have a comparison. Our model does effectively identify two distinct groups of data, and group 2 seems to consistently have higher detected random noise variance when analyzed using various forms of the model. The model performs better from a two-group approach. This seems to indicate that the data can be described effectively as groups with distinct properties. Our results also indicate that a two group ARMA model does not perform as well as two group MA model for the well log date. Overall, our results indicate that our model performs sufficiently well when searching for changepoints for us confidently proceed to applying it to new data sets.

Key Finding on the Data Structure of the Well-log Data

We seem to have discovered that our model finds the most change-points and matches up closest to post-simulation least squares estimation of parameters when we utilize the multi-group MA-only model with an allowance for error variance to differ by group. We have not observed this specific suggestion in any of the literature that we reviewed, and it may prove to be an insightful finding. We also seem to identify distinct differences in random noise when analyzing the data using a two group MA model. We cannot yet draw any formal conclusions from this, however, as more analysis would be needed. This is more of a key finding of the structure of the data set relative to how it is interpreted by our model. Not necessarily a true key finding regarding the physical implications of the data. More study into well-logging and geology are needed, which is beyond the scope of this project.

Chapter 5 Application for the Study of Covid-19

We are interested in studying outbreaks at a more local level than most studies present. Most mathematical literature involves state or national level data sets. We decided on a data set in a mid-sized county in the Milwaukee area. The goal is to identify points in time at which the rate of change of daily cases reaches zero. These represent points in time at which daily cases are being reported at either a locally maximum rate, locally minimum rate, or when case reporting begins to level off and briefly becomes somewhat constant. This can be likened to an inflection point. At a maximum, we are at a point at which an epidemic may start to subside in subsequent days. At a minimum, we may experience an extended time period of little disease activity, as in July of 2021, or we may be at a point at which an outbreak is about to occur. It may be possible to identify factors that are related to these local extrema, making it useful to keep note of the dates corresponding to these points in the series. As a last note, if we find many local max and min values within several days of each other, this indicates a sequence of days during which daily reporting of cases may be erratic, inconsistently high or low every few days, or this may indicate a sustained period during which an outbreak is occurring, but the test results are not being reported consistently enough for us to know whether we are dealing with a significant outbreak or not. For these many practical purposes, identifying these points in time can be useful. Our model is not necessarily intended to be superior to other methods used to find max and min rates of cases reported, but we will demonstrate that it is relatively easy to import a data series and quickly identify many useful dates.

5.1 Outbreaks in Waukesha County, Wisconsin 2020 and 2021

Waukesha county is likely to have a greater proportion of reported cases from local residents and smaller adjacent counties. This gives us the local-scale quality that we seek. We could have chosen a largest data set, from Milwaukee County, but these data encompass so much of the state population, with many reported cases from individuals living in other counties, as to make such a data set more akin to a state-level set. The data set was provided by the Wisconsin Department of Health Services. We observe two major outbreaks in each of two years, 2020 and 2021, and many within-year smaller outbreaks. In the series of raw data, we also observe what appear to be weekly spikes in daily reported cases. This is likely due to collection and reporting procedures. We also see what appear to be monthly spikes. And of course, data vary widely over time. The raw data clearly do not follow a stationary ARMA process. But as discussed in section 3 of chapter 1, there are some inherent auto-regressive and moving average properties. In order for our model to be applicable, we will need to develop a method to transform the data into segments of stationary data. We will first reduce some of the periodicity by taking a 14-day rolling average in order to reduce the effect of these periodic spikes on model smoothness, illustrated by the yellow curve. Our model is designed to analyze data that are dependent on the previous data by only one time step. If we have spikes in the data every week, this would be characteristic of a 7^{th} order process. Taking this rolling average helps balance this out. The blue plot represents the daily reported cases. The overlapping yellow curve represents the rolling average daily case estimates. We see this in figure 5.1.1.

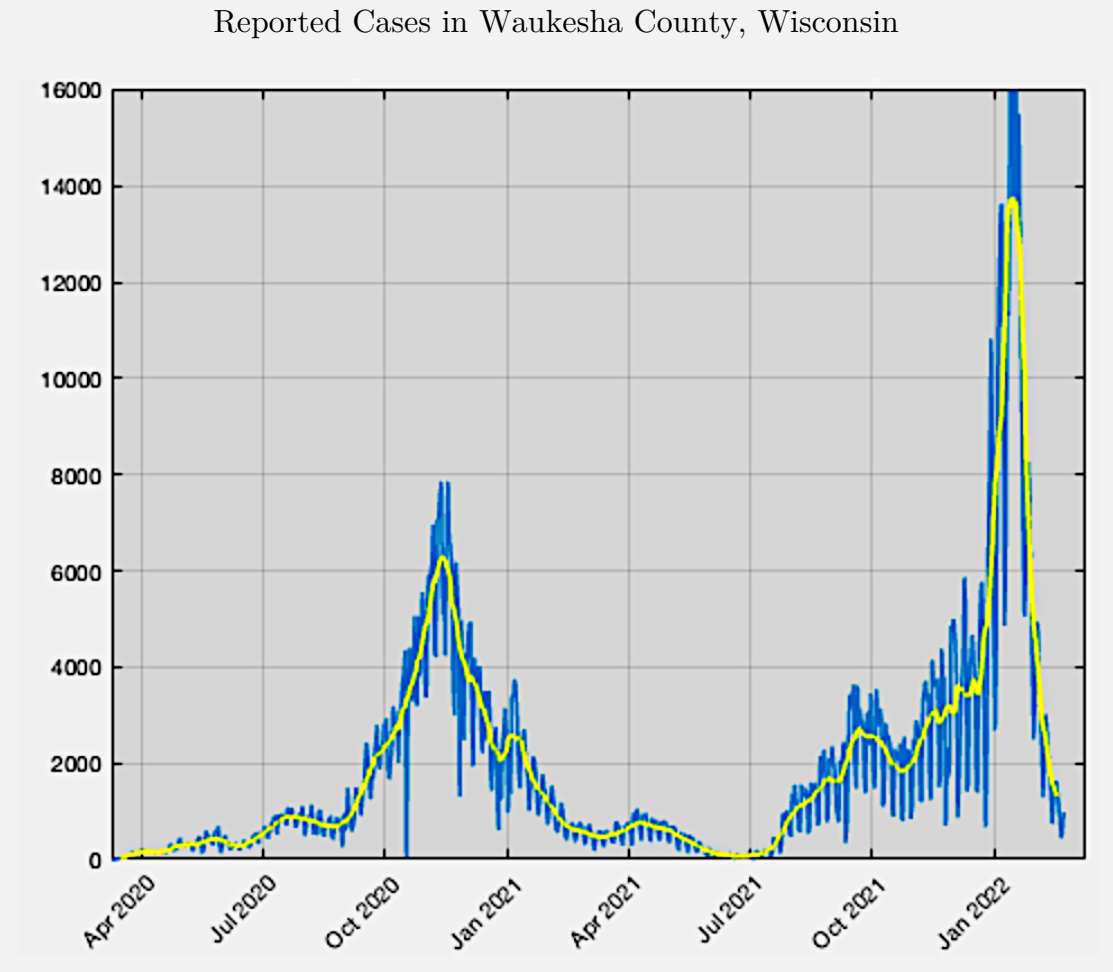


Figure 5.1.1 Series of reported cases (blue) with 14-day rolling average of data overlayed in yellow.

Of course, after we construct the 14-day rolling average, we still do not see a series that is characteristic of a segmented ARMA process described by our model. But our goal is to find the local maxima and minima, and locations at which rate of change suddenly changes. Interestingly, we do begin to see more of a segmented series when we generate a series of estimates of rates of change of daily reported cases. This can be likened to a series of derivatives of a smooth curve 14-day rolling average with respect to time. We do this by fitting a simple linear regression line to collections of seven successive data points

from the series of 14-day rolling averages of reported cases. The slopes of these regression lines are our estimates of rate of change of reported cases per day.

Our series of rate of change data still do not the exhibit the qualities of a stationary ARMA process, as can be seen in the upper left image in figure 5.1.2. But root and log transformations of the absolute value of the data can produce an apparent segmented series. The original sign can be re-assigned each data point, resulting in a series of roots of rates of change in cases per day, or log-rates-of-change.

Appearance of segments occur due to the fact that the data do not follow a smooth path, but may exhibit abrupt changes in how quickly the rates of change of reported case data, themselves, change. These jumps in reporting rates are the changepoints we will be looking for. These jumps will be described by differences in segment means. The most prominent changes occur when daily case rates go from positive to negative. These are the local max, local min, and inflection points. Change in sign of rates of change are especially prominent in the root transformation of data, but changes in magnitude are not. The higher order the root, the more prominent the segmentation, making it easy to discover local extrema. But the overall shape of the data seems to be lost when using root transformations of the data. A natural log transformation, however, seems to allow us to retain the shape of the data while still producing tractable segmentation. There is also the issue of a lack of random noise in the data. We reduced the issue of having high order autoregressive and moving average qualities when we generated a 14-day rolling average series from the original data set. But this smoothing out of the data also removed most of the random noise we observed in the original data set. When we perform a root or log transformation, we still end up with a series of data with no additional random noise.

And, in fact, we may be reducing random noise even further. We can apply a commonly used technique called a jitter, whereby we add a small amount of random noise to a data series to give us a more realistic view of what the data would look like if some of the original random noise had been retained.

Notice that the estimated group 1 and group 2 random noise differ, and that the random noise is not estimated to be 1, even though we added random noise of variance 1. This is due to the fact that some of the variation should be attributed to the AR and MA parameters, as this data is simulated based on our constructed data set. This difference does make sense, by group, as we expect a more precipitous drop in reported cases, post epidemic, due to sudden over-testing relative to cases. The results do indicate a collection of data with multi-group ARMA qualities.

The segment of oscillations indicates a potential third theoretical group that has mean zero and very little variation due to anything other than random noise. This could explain the overabundance of changepoints later in the year during year two. Even if we were to adapt our code to implement a three-group model, we would not have a large sample size of data from this hypothetical third group.

Natural Log Data Transformation

Based on the fact that outbreaks grow exponentially at the outset, we decided to attempt a natural log transformation of the rate of change series from the 14-day rolling average data. We once again add some random noise back to the data to simulate the random noise present in the original data set. The dots represent the transformed data without the noise. The orange scatter in the background represents the data with random noise added. The segments appear less prominent than those generated with the high-order exponential transformation, but interestingly the Gibbs sampler performs very well. We are able to avoid an over-abundance of changepoints and our ARMA model parameters converge to at least somewhat intuitive values. The MA parameter estimates make especially good sense. When disease data reporting tends to increase exponentially, we can see a slower or more gradual shape to the curve. We expect error terms to be negatively correlated, and so we should expect negative MA parameters. We get convergence close to -0.4 for both MA parameters. We expect a strong positive correlation due to auto-regressive behavior, as discussed. We see this in the trace plots for the AR parameters. However, the AR parameters apparently exhibit properties similar to those of the well log data. Being so high, this indicates that an AR component may not result in a good overall ARMA model fit. Other than this, the Gibbs sampler performs well with the transformed data in this form. We provide the original series of data with the 14-day rolling averaged overlayed in yellow. Additionally, we show the locations of the detected changepoints and their corresponding dates, indicated by the vertical black lines. We ran the simulation many times, given that the random noise we are adding will change each time. The dates we find are all very similar, usually within a day or two of each other, and we have no problem with adjacent change points.

Data Transformation Process

Data transformed by taking the natural log of the absolute value of the data. The sign of the original data is then re-introduced, producing the blue curved data points we see in this graph. With random noise added, we get the orange scattered line graph.

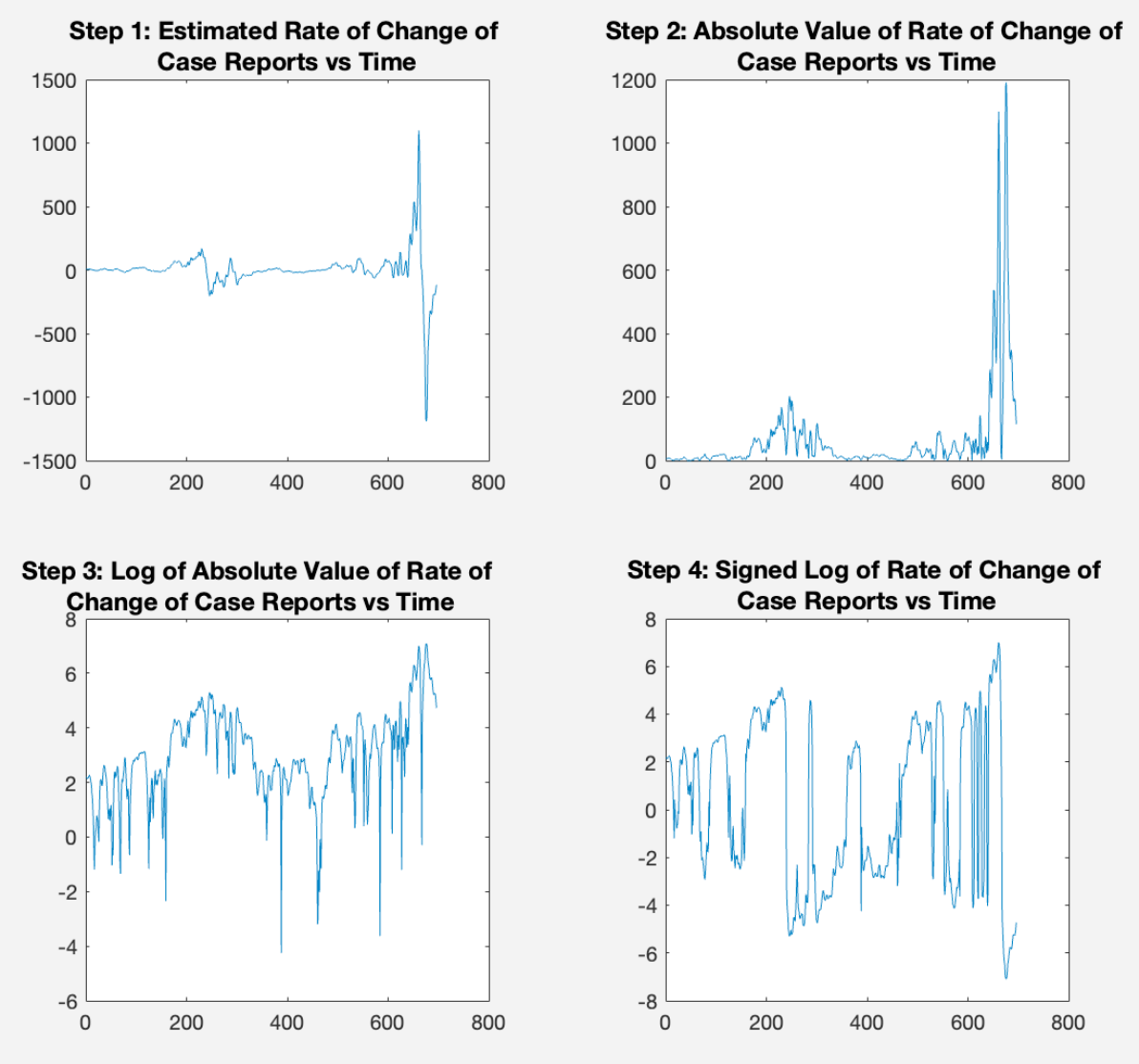


Figure 5.1.2 Data transformation from raw data to segmented data of signed log rates of change of daily reported cases.

Step 6: Add Jitter to Transformed Data

Random noise is added, normally distributed with variance 0.25. This relatively low variance allows for the retention of the shape of the data series while re-introducing randomness that was lost during the data transformation process, making these data compatible with an ARMA modeling procedure.

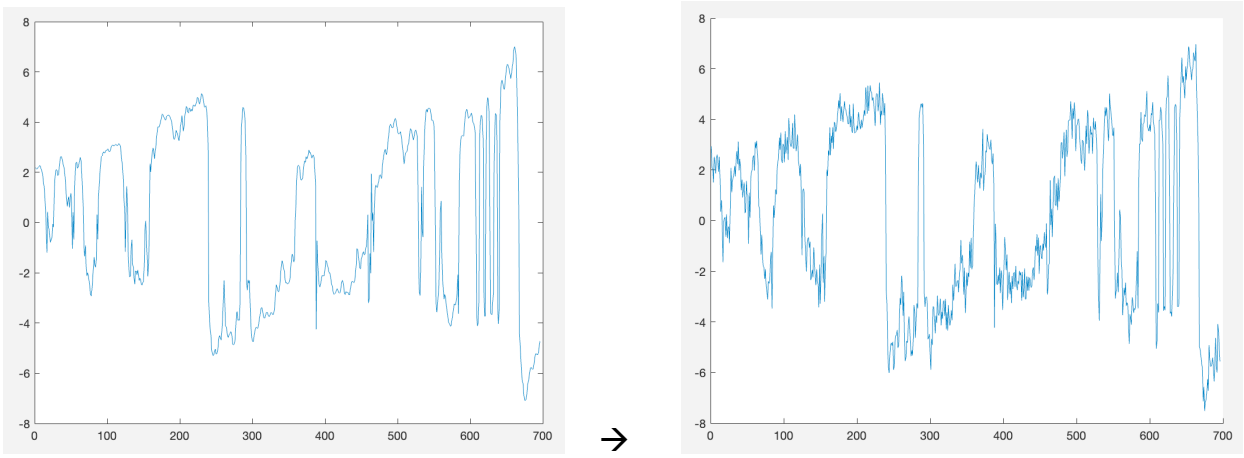


Figure 5.1.3 Transformed data before adding random noise (left) and after adding random noise (right).

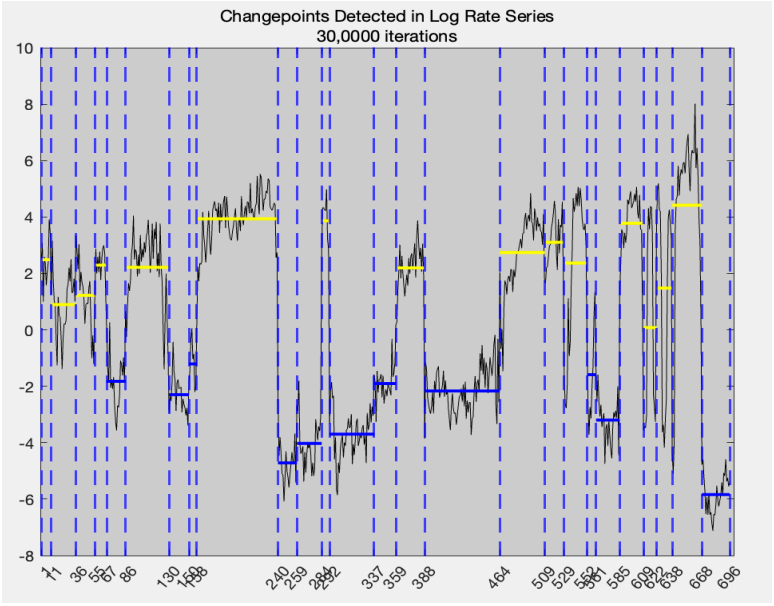


Figure 5.1.4 Changepoints detected after 30,000 iterations.

Time Series of Daily Cases from March 2020 to February 2022

All of the detected change points other than two (8-15-2020) and (2-18-2021) visibly correspond to local maxima, minima, inflection points. However, some points occur in locations of greater uncertainty as far as how the disease is progressing, as seem from October through December 2021. These points can be likened more to detection of inflection points, as reporting is relatively constant from the middle of November through the middle of December of 2021.

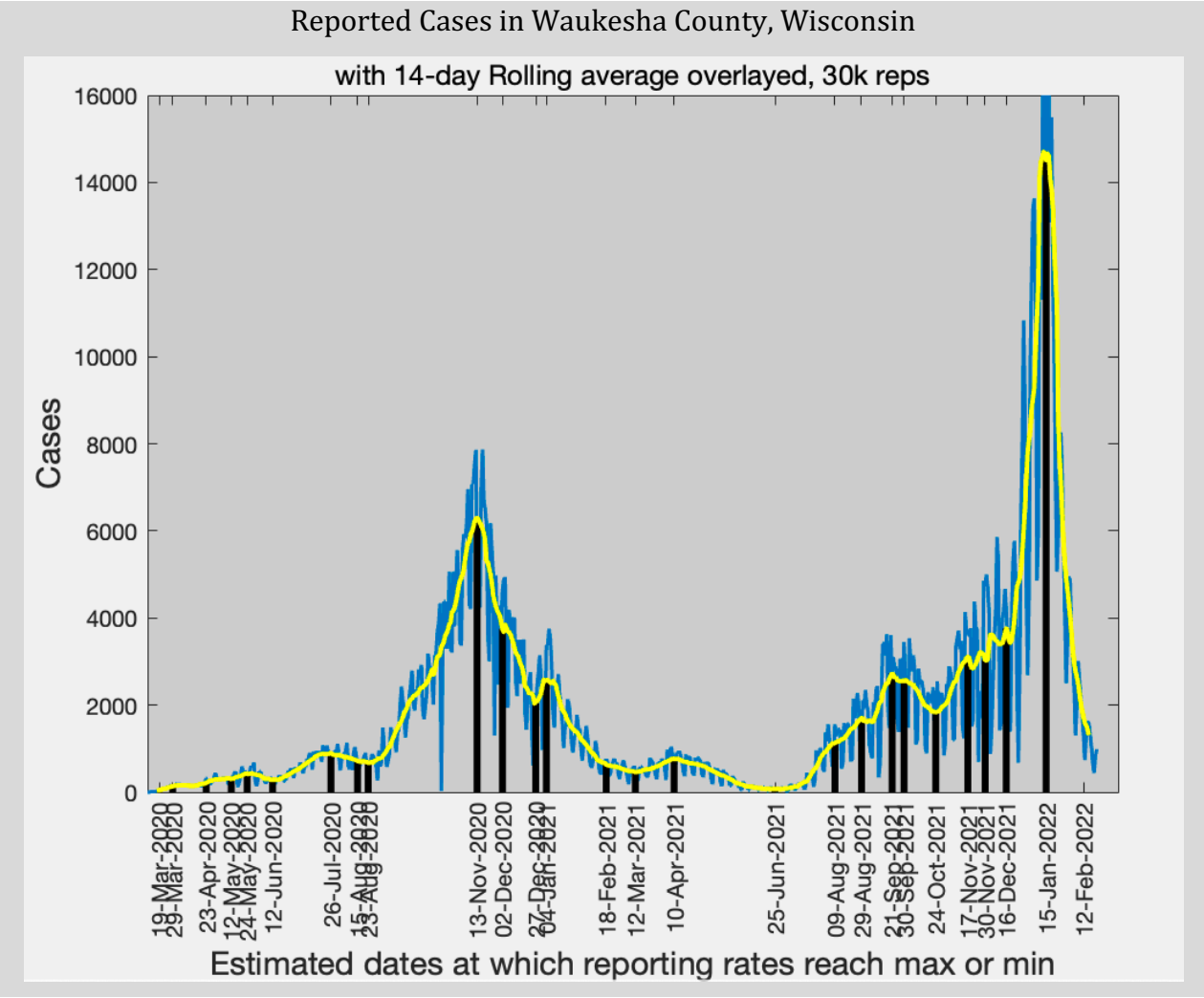


Figure 5.1.5 Changepoints detected are indicated by vertical black line segments.

Time Series with Local Max, Local Min, and Inflection Points for Year 1

Here is a closer look at the data with the detected change points during the first year. We appear to discover either maxima, minima, or relatively flat locations toward the beginning of the epidemic.

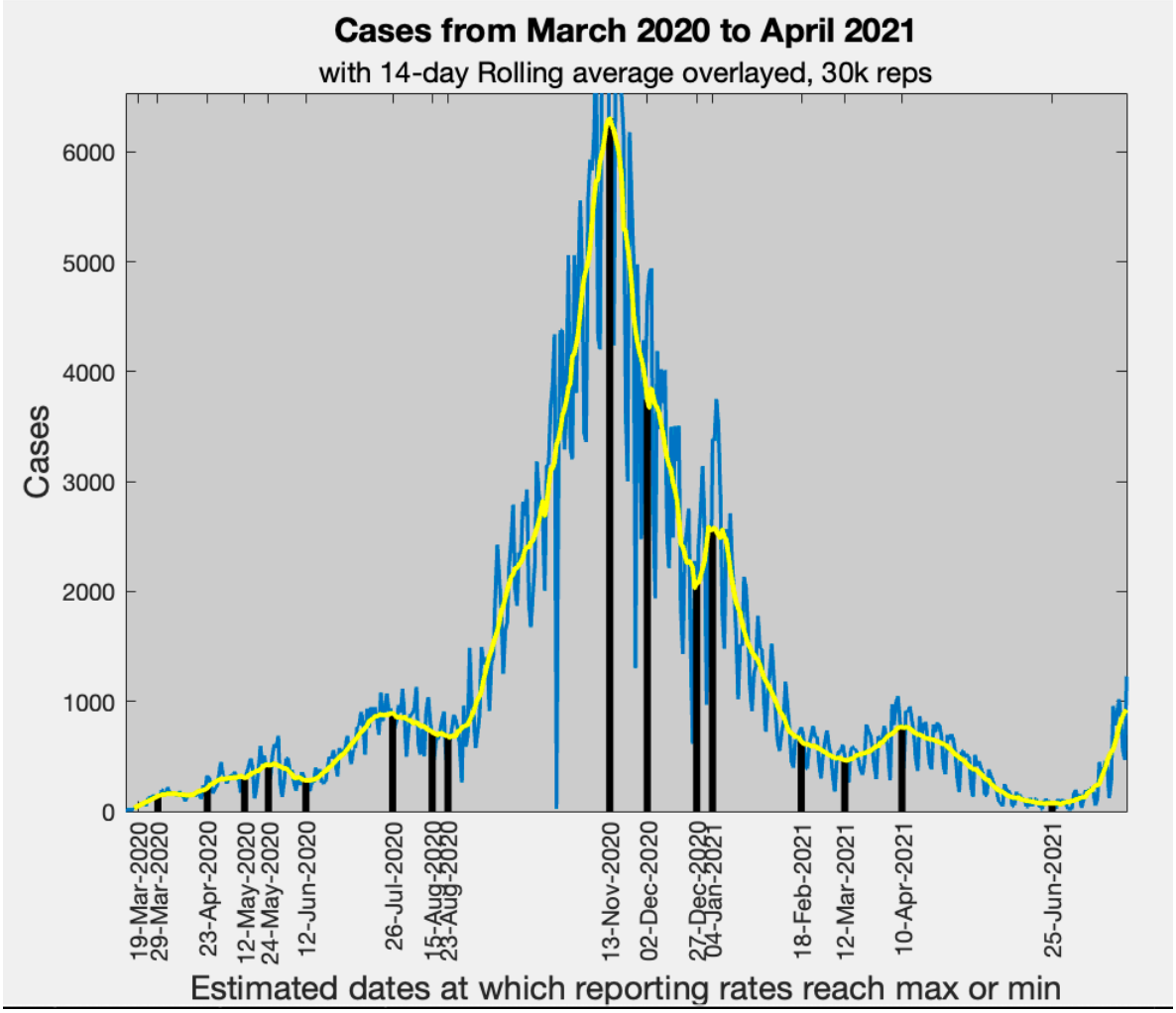


Figure 5.1.6 Time series of cases with changepoints for outbreak in year 1.

Time Series with Local Max, Local Min, and Inflection Points for Year 2

We have a close look at the local extrema detected for the second year. The points identified from November through December seem to be more characteristic of inflection points than extrema, given that the rate of change is relatively close to zero over this time period.

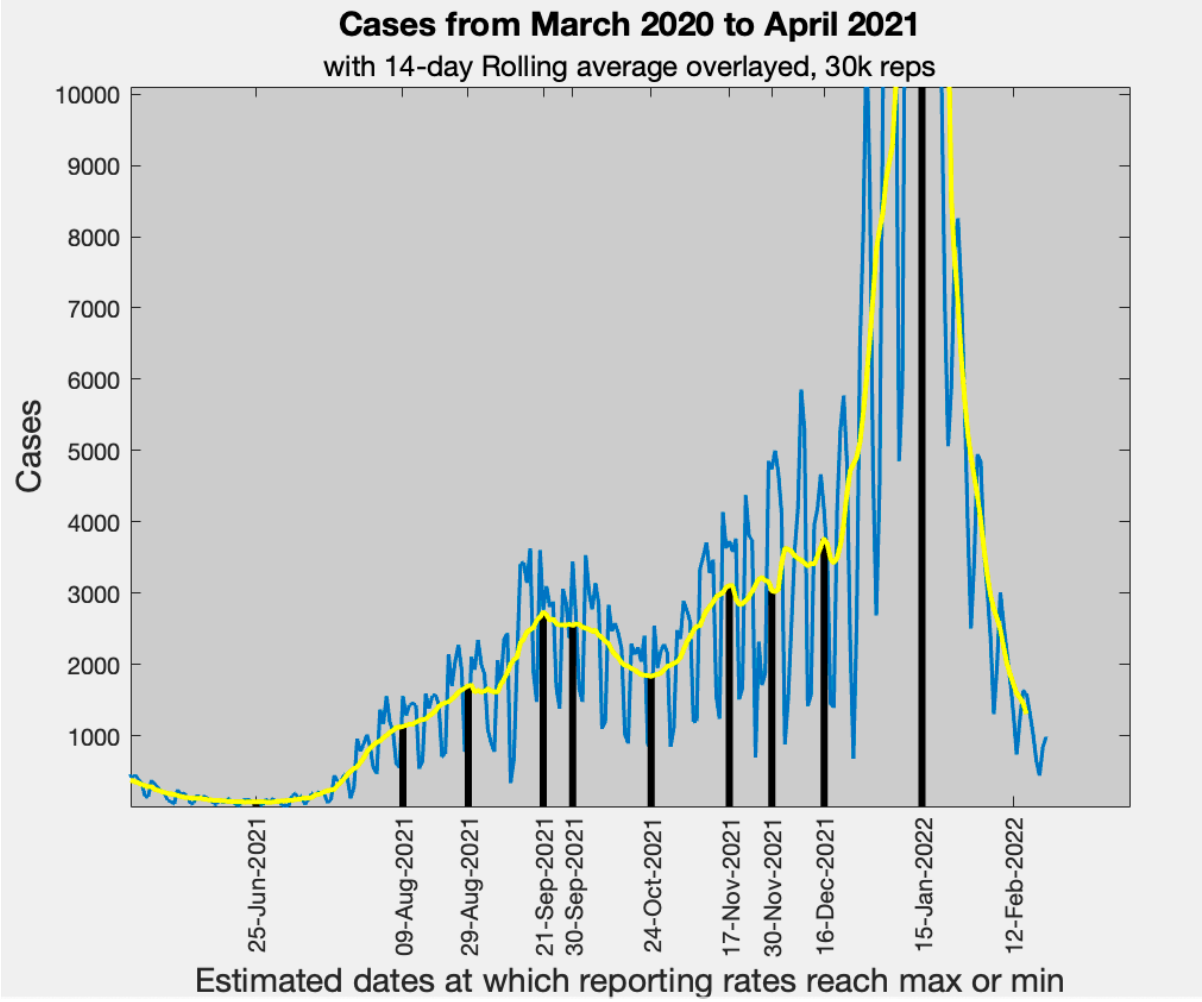


Figure 5.1.7 Time series of cases with changepoints for outbreak in year 2.

Trace Plots for Underlying Model Parameters

We see a contrast between the estimated random noise for group 1 vs group 2, shown in the third row. This may indicate that when rate of change of daily infections is positive (group 2) we may be detecting more volatility in the reporting of data. Upper row is expected group means of log-rates of change. Second row is estimated variance of segment means.

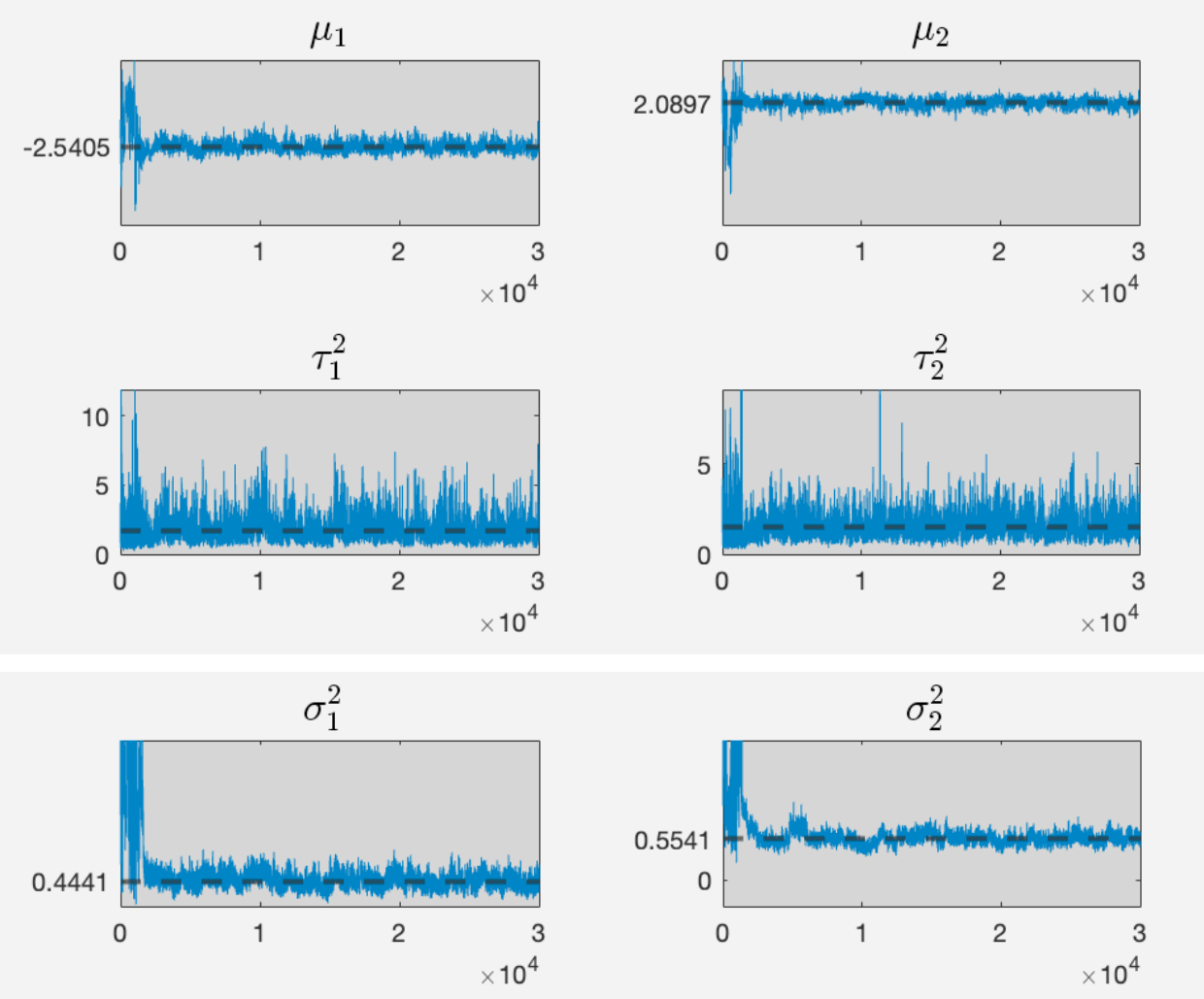


Figure 5.1.8 Trace plots for underlying model parameters for Covid-19 study

Trace Plots for ARMA Parameters Under Natural Logarithmic Transformation

We expected a negative correlation between errors in case reporting, which we observe through the MA parameters in the left column of figure 5.1.9. We get similar magnitude for both groups, indicating that reporting error is similarly influential during times at which daily cases are increasing vs times at which daily cases are decreasing. The high magnitude for the AR parameters indicates that the data are strongly correlated. This is likely due to the fact that we tend to get an exponential increase in daily cases as an epidemic worsens.

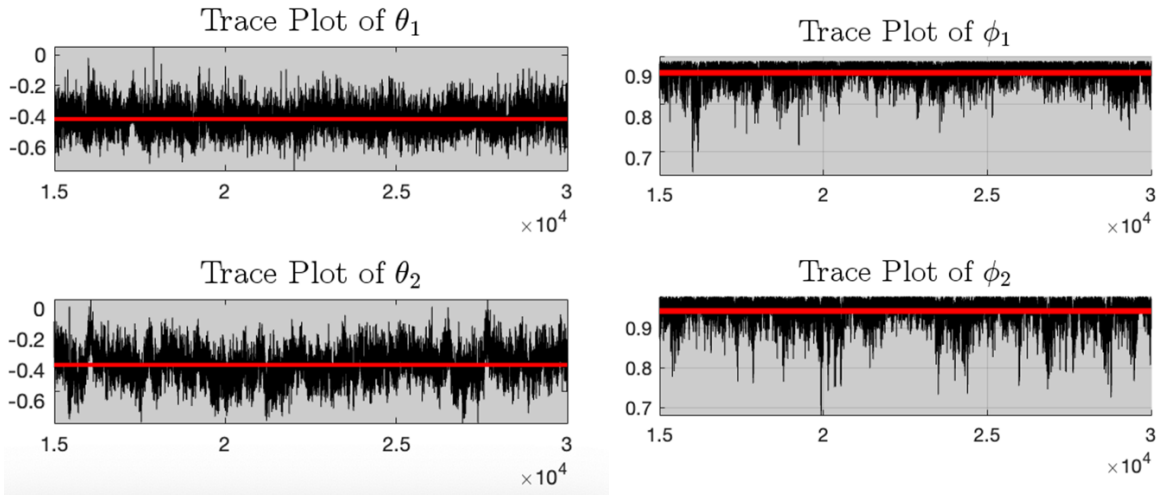


Figure 5.1.9 Trace plots of moving average parameters (left column) and auto-regressive parameters (right column).

Parameter Estimates for Log Transformation of Rate of Change of Reported Cases of Reported Cases of Covid-19

A comparison of Gibbs Sampler, Matlab least squares estimator, and true parameter values is given in table 5.1.1. The AR parameters estimates are lower at 10,000 iterations due to a wider spread in the trace-plot shown in figure 5.1.11. The trace plots for the first 10,000 iterations are more erratic, pulling the mean lower. The least squares estimations are done, post-simulation. We do not get results that are incredibly different. The most notable result is actually the group variance estimation. Our disparate estimates for error variance actually make intuitive sense, as reporting data are more volatile as daily cases begin to increase.

	Gibbs Sampler 10k iterations	**Gibbs Sampler 30k iterations**	**Least Squares Estimator in Matlab**
ϕ_1	0.826	0.930	0.722
ϕ_2	0.798	0.943	0.824
θ_1	-0.370	-0.420	-0.318
θ_2	-0.253	-0.357	-0.428
σ_1^2	0.496	0.444	0.434
σ_2^2	0.397	0.554	0.434
μ_1	-2.444	-2.540	-2.284
μ_2	2.340	2.090	2.274

Table 5.1.1 Comparison of Gibbs Sampler at 10k and 30k iterations vs Least Squares Estimator for Covid-19 data.

Comment on inaccurate ARMA parameter estimation with 10,000 iterations

We had fairly good results when running 30,000 iterations. Figures 4.5.11 and 4.5.12 show less stable trace plots at only 10,000 iterations especially prominent for group 2 AR and MA parameters. This gives a false impression that group 2 errors are less impactful.

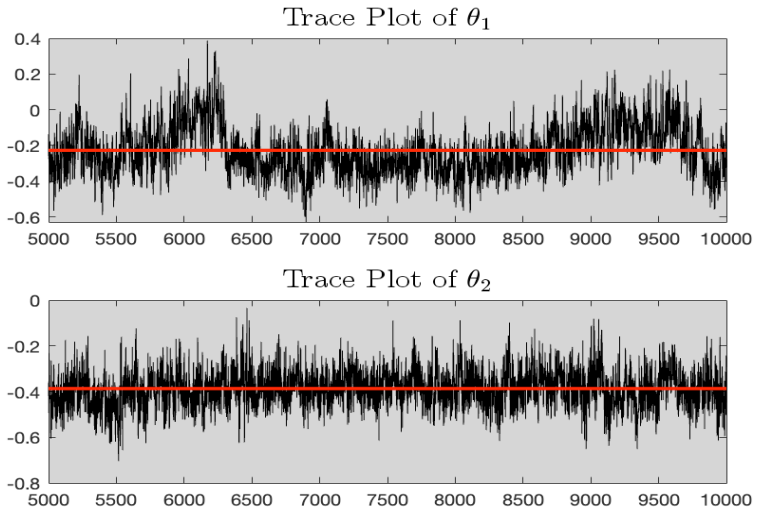


Figure 5.1.10 Moving average trace plots for log rate of change of daily reported cases after 5000 iterations.

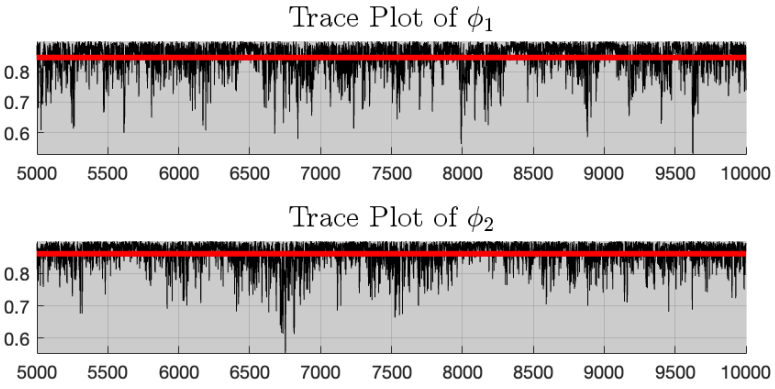


Figure 5.1.11 AR trace plots for log rate of change of daily reported cases after 5000 iterations.

5.2 Comment on Early Outbreaks and Associated Events

At the beginnings of outbreaks, we not only find areas of local maxima and minima, but locations at which rates may change in magnitude slightly but suddenly. This may not be apparent from the graphs. We see zoomed-in images from the beginning of the original outbreak in figure 4.5.13 and the beginning of the second outbreak in 4.5.14. The overall results may prove insightful from a broad perspective, but it is difficult to assign any causality to any individual events local to Waukesha County. We can still discuss some general insight we gain.

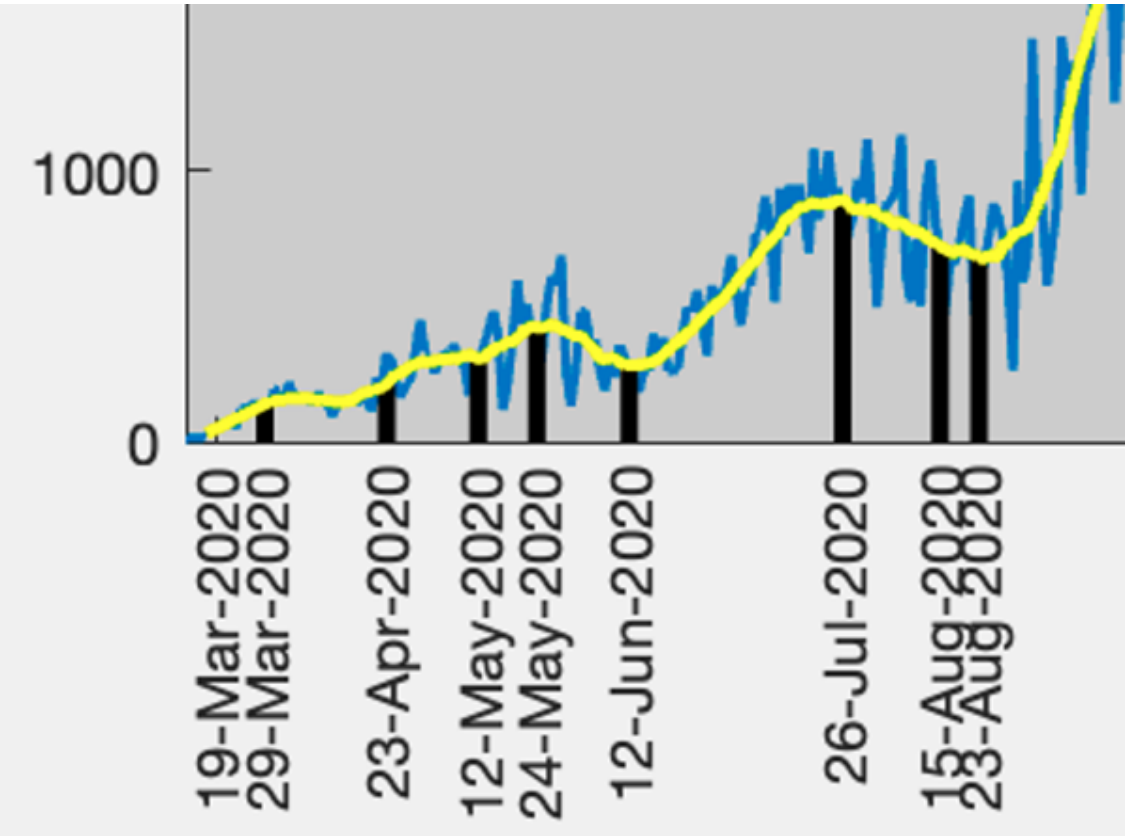


Figure 5.2.1 Zoom in of daily cases at the beginning of the first-year outbreak with 14-day rolling average. Vertical lines indicating points at which rate of change of reported daily cases has changed in magnitude.

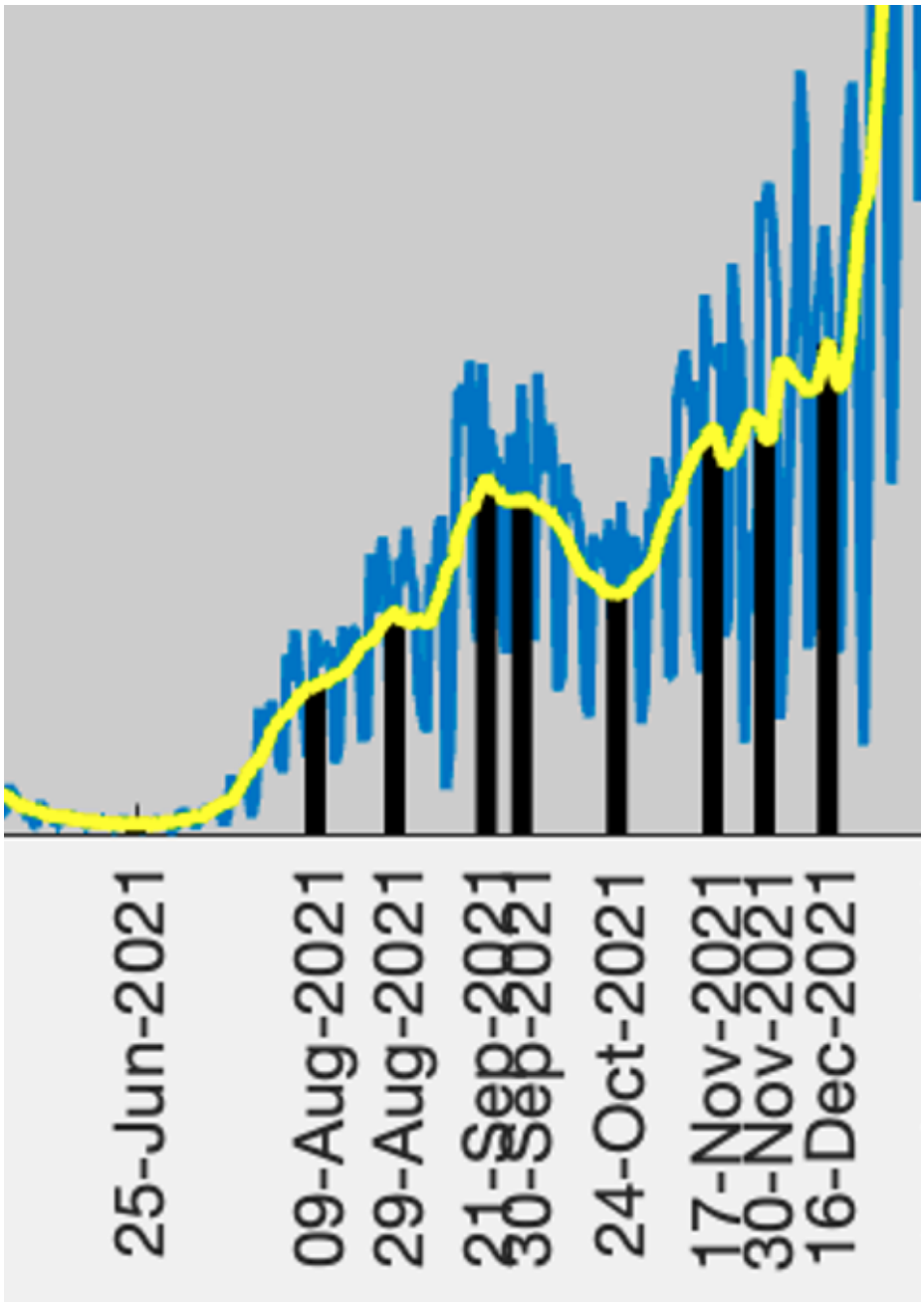


Figure 5.2.2 Zoom in of daily cases at the beginning of the second-year outbreak with 14-day rolling average. Vertical lines indicating points at which rate of change of reported daily cases has changed in magnitude.

The overall volatility seen in the second outbreak may have to do with the simultaneous increase in overall summer activities in the area, as well as the returning of students to in-person classes in the fall.

Our model estimates that reporting rates were relatively minimal on August 29th of 2020, after which a sharp increase finally begins. This may indicate the effectiveness of social distancing and mask-wearing during the summer of this year in temporarily mitigating the spread. Four weeks later, daily case reporting is rapid, as is still to be expected to how contagious Covid is known to be. During the second-year outbreak, we see a relative minimum in reporting estimated on October 24th, exactly eight weeks later than the date seen during the first year. Rather than seeing a sharp increase after the students returned to in-person education, we see the opposite of what we saw during the first year in which students did not attend classes in person. We might have expected to find a sharp increase in reported cases not only due to a return to in-person education, but also due to local and national efforts to increase testing rates at schools. This decline reaching a minimum halfway through the academic fall semester may indicate the effectiveness of vaccination efforts and mask-wearing.

As we said, we cannot assign causality to local events, but we can get a general idea of what might be going on, and what we do find is not unexpected when we do take into consideration what is known about the dynamics of Covid-19. If anything, we have verified that our model does detect points in time that we can reasonably consider to be associated with local minimum, local maximum, points of inflection, and points at which case reporting rate suddenly increase or decrease.

Chapter 6 Discussion and Future Work

By developing a multivariate version of the model by [Sadia et al., 2018], while also developing a new method to calculate acceptance and rejection probabilities (Adjusted Insertion Rule), we seem to have resolved the potential issue that can arise when error variance differs between group segments and eliminate false identification of adjacent changepoints (segments of length one defined as unique series of data). Our algorithm performs surprisingly well in accurately identifying the random noise variances for individual groups. This is something that is in no way apparent by visually inspecting the data. This can make the model less susceptible to outliers, while also potentially lowering overidentification of changepoints in groups with lower error variance. Our sampling procedure differs from others we have observed in that we include a provision to prevent an immediate reversal of the algorithm's decision to delete a proposed changepoint (the Adjusted Insertion Rule), lowering the chance of having more estimated segments than we should. We also addressed the issue that can arise with machine rounding error when attempting to calculate acceptance and rejection probabilities on relatively long intervals. We discovered that without accounting for this, up to one third of our deletion attempts had inaccurate or undefined acceptance probability, resulting in an overly high rate of rejection of proposed deletions. This prevents over-detection of changepoints by preventing the machine from automatically skipping deletion steps. We provided additional analysis of the well-log data set, including estimates of all ARMA parameters for a multi-group model for this data. We have shown that the data evidently originate from two distinct groups with differing estimated ARMA parameters, and subtle but distinct differences in random noise variance. We have also shown that it may be

important to analyze the data using many more iterations of the Gibbs sampler than we have seen in the literature by showing that trace plots do not begin to stabilize until well after 10,000 iterations.

We also proved that closed form conditional posterior distribution functions for the AR and MA parameters do exist, and we derived these distributions, showing them to be truncated normal. This allows us to directly sample from the distributions of these parameters in order to apply our Gibbs sampling procedure as we intended. We explicitly derived a special case posterior distribution for group probabilities when we are dealing with exactly two groups, which simplified our sampling procedure by requiring that we only calculate a probability for one group using a less involved density function. Our model was demonstrated on a wide variation of simulated data sets for segmented ARMA models with varying ARMA parameters and segment error variances than we have observed with earlier models. Finally, we successfully applied our model to a local data set for daily cases of Covid-19. We were able to transform the data set into a form that is characteristic of a two-group segmented ARMA process. We did not observe such use of a segmented ARMA model for this purpose for time series data in the literature we studied. We used our insight from analyzing the well-log data when deciding on an appropriate number of iterations, giving us more stable trace plots of our parameters. We identified many notable locations in the data series which seem to line up with points in time at which rates of change of daily reported cases of Covid-19 in Waukesha County, Wisconsin reached local maxima and minima. These appear to generally match up with what we would expect to observe at the times of year with which they correspond, and we demonstrate this visually by overlaying line segments indicating the detected changepoints with a graph of the time series of the original data.

Limitations and Future Work

Our model works well with data sets that have groups such that the expected means of segments and variance of segment means result in groups segments that are relatively far apart. When data originate from groups that have similar expected segment means, our model has a greater difficulty in distinguishing the difference between groups. This poses no issue to us when analyzing our epidemiological data, as we have groups that are very distinct, with relatively low variance in segment means. This is likely due to the fact that our insertion and deletion probabilities do not depend on the ARMA parameters of the model. Were we to include a provision for this, we would be better equipped to handle segmented data that do not differ greatly by segment mean. Our model is also less sensitive to outliers insomuch as when outliers occur in one group, it does not affect the estimated error variance of the other group. Thus, outliers only affect the groups they occur in, rather than all group parameters. However, our model may not be less sensitive to outliers in terms of their effect on within-group changepoints. We may want to include a provision to flag outliers during simulations to avoid the need for removing them. We would like to implement a similar method and compare our results. Our model is also not designed to detect whether data originate from stationary distributions or not. This may be why it takes so many iterations of our Gibbs sampler to estimate parameters of the well-log data.

Addressing Higher-order ARMA Data and Non-Stationary Trends

A natural progression of this model would be the extension to a general finite dimensional ARMA(p, q) model in order to deal with data that do not seem to originate from first order processes. Another avenue would be the addition of a linear trend component. This

could allow us to detect changepoints by considering segmented data to be centered about a line with a constant slope, rather than centered about a fixed mean. We have already tested computational time needed for such estimates by calculating estimated slopes of regression lines fit through every segment of our data during a typical run of our current model and have found that relatively little additional computational time is needed.

Verification of Effectiveness in Preventing Over-detection of Changepoints

We claim that our Adjusted Insertion Rule sampling procedure prevents over-detection of changepoints, and numerical simulations seem to support this. It seems obvious, as it is easy to show that on individual iterative steps, we lower the probability of immediately rejecting the decision to delete improbable points. Our numerical results also seem to clearly suggest that this reduces the number of false identifications of changepoints over the course of the entire iterative process. Mathematical proof of this is lacking, however, and it may prove very insightful to provide such a proof.

Application to Additional Epidemiological Data sets and Simulated Data

We may want to develop a simulated data generation method for epidemiological data so that we can test our current model and other models on data with a wide range of characteristics, much like what we did with our general ARMA data simulations. We may also want to apply our model to other epidemiological data sets that have already been studied in different ways at the state and national level, as well as to additional county-level data sets that have not been studied. Such data may include data on Covid-19 or other illnesses.

Changepoint Detection for Online Data

Our model is designed for retrospective study. A natural extension of this would be an adaptation for the analysis of online data.

Final Comments

From the vantagepoint of what originally motivated our study, our model has served us quite well. We wanted to build a tool that we could use to study the dynamics of Covid-19. We can read in raw data of reported daily cases and get an immediate transformation of non-stationary data into a form that is compatible with a segmented ARMA model. We do not need to resort to more advanced methods, such as adding a linear regression component or using segmented exponential regression. We get results that include visualization with specifically identified dates, rather than having to manually produce the graphs and manually look for these points. Additionally, we get further insight as far as what our trace plots tell us about the auto-regressive and moving average properties of the transformed version of the data, as discussed in section 5.1. This is something that we cannot see at all by looking at a scatter plot of the data. Our model is useful for what we originally set out to study, and it may be insightful to others who would like to study the dynamics of Covid-19 using segmented ARMA processes from a Bayesian perspective. Given what we have shown with simulated data and commonly studied well log data, along with our method for transforming non-stationary data into a form compatible with our model, we are confident in our ability to utilize this model for the study of many other types of data, as well.

References

Albert, J. and S. Chib. "Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts." *Journal of Business & Economic Statistics*, vol. 11 (1993): 1-15.

Allaud, L. A. "Schlumberger: The History of a Technique." *John Wiley & Sons Inc.* (1978).

Barnard, G.A. "Control Charts and Stochastic Processes." *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, vol. 21 (1959): 239-257.

Bowker, G. "A Well Ordered Reality: Aspects of the Development of Schlumberger, 1920-39." *Social Studies of Science*, vol. 17, no. 4 (1987): 611–55.

Benson, A. and N. Friel. "Adaptive MCMC for multiple changepoint analysis with applications to large datasets." *Electronic Journal of Statistics*, vol. 12, no. 2 (2018): 3365-3396.

Chernoff, H. and S. Zacks. "Estimating the Current Mean of a Normal Distribution Which Is Subjected to Changes in Time." *The Annals of Mathematical Statistics*, vol. 35, no. 3 (1964): 999–1018.

Cucina, D., Rizzo, M. and E., Ursu. "Multiple changepoint detection for periodic autoregressive models with an application to river flow analysis." *Stochastic Environmental Research and Risk Assessment,* vol. 33, no. 4-6 (2019): 1137–1157.

Davis, R.A., Lee, C.M., and G. A. Rodríguez-Yam. "Structural Break Estimation for Nonstationary Time Series Models." Journal of the American Statistical Association, vol. 101 (2006): 223-239.

Fearnhead, P. and P. Clifford. "On-line inference for hidden Markov models via particle filters." *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, vol. 65, no. 4 (2003): 887–899.

Fearnhead, P. "Exact and efficient Bayesian inference for multiple changepoint problems." *Journal of Statistical Computing*, vol. 16 (2006): 203–213.

Fearnhead, P. and R. Guillem. "Changepoint Detection in the Presence of Outliers." *Journal of the American Statistical Association.* vol. 114, no. 525 (2016): 169-193.

Giordani, P. and R. Kohn. "Efficient Bayesian Inference for Multiple Change Point and Mixture Innovation Models." *Journal of Business and Economic Statistics*, vol. 26, no. 1 (2008): 66–77.

Girshick. M.A. and H. Rubin. "A Bayes Approach to a Quality Control Model." *Annals of Mathematics and Statistics*, vol. 23, no. 1, (1952): 114 – 125.

Jiang, S., Zhou, Q., Zhan, X., and Q. Li. "BayesSMILES: Bayesian Segmentation Modeling for Longitudinal Epidemiological Studies." *National Library of Medicine medRxiv,* preprint, (2021).

Koop, G. and Potter, S. M. (2007). "Estimation and Forecasting in Models with Multiple Breaks." *The Review of Economic Studies*, vol. 74, no. 3: 763–789.

Ó Ruanaidh, J. J. K., and W. J. Fitzgerald. "Retrospective Changepoint Detection. In Numerical Bayesian Methods Applied to Signal Processing." *Springer Science & Business Media*, (2012): 96-121.

Page, E.S. "A test for a change in a parameter occurring at an unknown point." *Biometrika*, vol. 42 (1955): 523-527.

Page, E. S. "On Problems in Which a Change in a Parameter Occurs at an Unknown Point." *Biometrika*, vol. 44, no. 1-2 (1957): 248–252.

Parshall, K. "A New Era in the Development of Our Science The American Mathematical Research Community, 1920-1950." In: Rowe, D., Horng, WS. (eds) A Delicate Balance: Global Perspectives on Innovation and Tradition in the History of Mathematics. Trends in the History of Science, (2015): 275-308

Pesaran, M. H., Pettenuzzo, D., and A. Timmermann. "Forecasting Time Series Subject to Multiple Structural Breaks," *Review of Economic Studies*, vol. 73, no. 4 (2006): 1057– 1084.

Punskaya, E., C. Andrieu, A. Doucet and W. J. Fitzgerald. "Bayesian curve fitting using MCMC with applications to signal segmentation." IEEE Trans*actions on* Signal Process*ing,* vol. 50, no. 3 (2002): 747-758.

Rosen, O. and R. Kohn. "Bayesian mixtures of autoregressive models." *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, (2011): 174–195.

Ruggieri, E. "A Bayesian Approach to Detecting Change Points in Climatic Records." *International 901 Journal of Climatology*, vol. 33 (2013): 520-528.

Sadia F., Boyd S., and J.M. Keith. "Bayesian change-point modeling with segmented ARMA model." *PLoS One, vol 13. no. 12* (2018): 1-23.

Shewhart, W.A. "Economic Control of Quality of Manufactured Product." *D. Van Nostrand Company, Inc.* (1931).

Shiryaev, A. N. "On Optimum Methods in Quickest Detection Problems." *Theory of Probability and Its Applications,* vol. 8 (1963): 22-46.

Smith, A. F. M. "A Bayesian approach to inference about a change-point in a sequence of random variables." *Biometrika*, vol. 62 (1975): 407–416.

Stanley I., Ko, M., Terence, T., Chong, L. and P. Ghosh. "Dirichlet Process Hidden Markov Multiple Change-point Model." *Bayesian Analysis,* vol. 10, no. 2 (2015): 275-296.

Stephens, D. A. "Bayesian Retrospective Multiple-Changepoint Identification." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 43, no. 1 (1994): 159-178.

Sugasawa, S. and S. Hashimoto. "Robust Bayesian Changepoint Analysis in the Presence of Outliers." *Intelligent Decision Technologies: Proceedings of the 13th KES-IDT 2021 Conference (Smart Innovation, Systems and Technologies, 238) 1st ed. 2021 Edition*, (2021): 469-478.

Tong, H. and K.S. Lim. "Threshold Autoregression, Limit Cycles and Data." *Journal of the Royal Statistical Society, Series B (Statistical Methodology),* vol. 42, (1980): 245-292.