Stony Brook University

A Stochastic Segmentation Model for Recurrent Copy Number Alterations in Grouped array-CGH Data

A Dissertation Presented

by

Ying Cai

to

The Graduate School in Partial Fulfillment of the Requirements for the

Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2013

Stony Brook University

The Graduate School

Ying Cai

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Haipeng Xing - Dissertation Advisor
Associate Professor, Applied Mathematics and Statistics

Wei Zhu - Chairperson of Defense
Professor & Deputy Chair, Applied Mathematics and Statistics

Song Wu - Committee Member
Assistant Professor, Applied Mathematics and Statistics

Jinfeng Xu - Committee Member
Assistant Professor, Division of Biostatistics, Department of Population Health, School of Medicine, New York University

This dissertation is accepted by the Graduate School.

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

A Stochastic Segmentation Model for Recurrent Copy Number Alterations in Grouped array-CGH Data

by

Ying Cai

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2013

With the recent advances in high resolution microarrays and next generation sequencing, DNA copy number can now be profiled in a high throughput global manner. This has enabled the systematic study of DNA copy number alterations in tumors, as well as the profiling of inherited population-wide copy number variants. Studies of DNA copy number usually involve many samples that fall into different groups, e.g. tumor subtype or ethnic group. It is often of interest to find recurrent alterations within each group. We develop a stochastic segmentation model for detecting recurrent DNA copy number alterations in grouped array-CGH data. In our model, the parameter in each regime is a random variable following specific regime-specific distribution. Explicit formulas for posterior means can be used to estimate the signal directly without performing segmentation. We give a linear-time algorithm for fitting this model and for estimating its parameters by expectation maximization. Simulation

studies and applications to real grouped array-CGH data illustrate the advantages of the proposed model.

To my parents, my husband Yan with all my love

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor, Prof. Haipeng Xing, for the strict statistical model suggested, for interesting discussions and for his guidance.

I would also thank Prof. Wei Zhu, Prof. Song Wu and Prof. Jinfeng Xu for joining my dissertation committee.

I would like to thank all my group mates and friends, especially Ning Sun, Yifan Mo, Yang Yu, Jie Wu, for all their support and help.

Chapter 1

Introduction

Genome sequencing have been made process at a rapid pace, with the development of high throughput technologies. The high resolution microarrays and next-generation sequencing (NGS) technologies are now widely being applied to a broad range of important topics in biology and medicine, including analyses of transcriptome dynamics, genomic variation, genome structure and etc. These high throughput, high resolution techniques have generated tremendous sequential data, thus there are onerous statistical and computational challenges that are incurred when dealing with such data, like identifying genetic variation, transcriptional network inference, Genome-Wide Association Study (GWAS) or common diseases and complex traits. Obviously, identifying genetic variation and understanding the role in human traits and diseases is an important goal of human genetics.

In this chapter, we will review an important genetic variation, arrayCGH data for copy number alteration (CNA) detection and recurrent copy number alterations (CNAs). Then we shall retrospect the most popular and the latest statistical models to show how they solved the problem of detecting recurrent CNAs. Moreover, the interesting and unsolved questions are shown in the third section as the motivation of our study. The last section gives the outline of this dissertation.

1.1 Recurrent Copy Number Alterations using array-CGH data

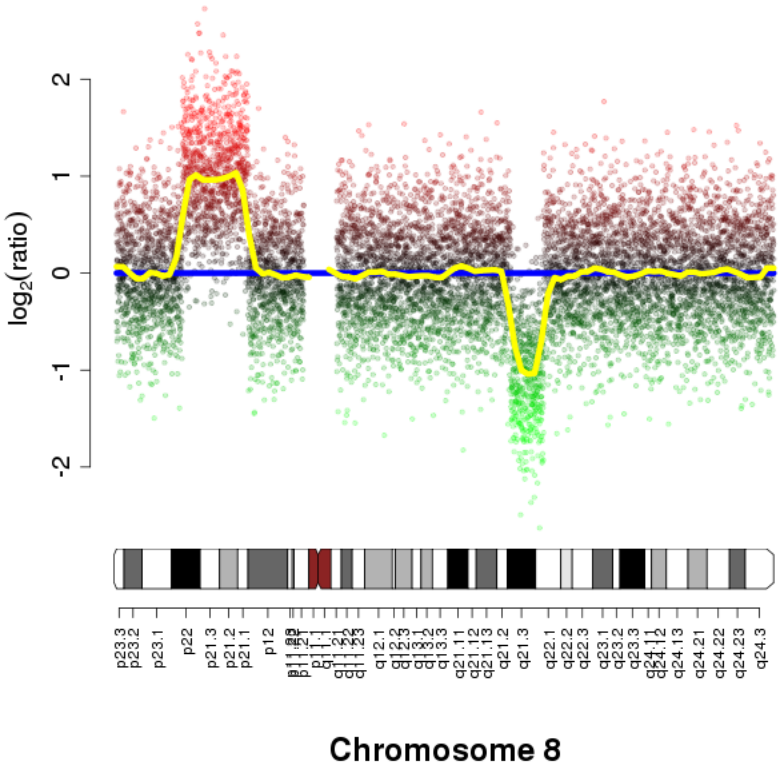
Genetic variation plays a significant role in biological function. It exists in different forms, ranging from gross alterations in the human karyotype to single nucleotide changes. one predominant form of genetic variation is called a Single Nucleotide Polymorphism (SNP). A SNP is defined as a single nucleotide base change in a DNA sequence. The single nucleotide base is replaced by any of the other three bases. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs have been identified in the human genome. If a SNP is frequently found close to a particular gene, it can act as a biological marker for that gene, thereby it may help researcher to locate disease associated gene. An alternative form of genetic variation has gained much interest, it is defined as gains or losses of specific regions of the genome, and varying in size from 1 kb to a complete chromosome arm, not limited to a single nucleotide base (Lee et al., 2007; Scherer et al., 2007; Shah et al., 2008; Hastings et al., 2009), has been estimated to contribute more to genetic diversity than single nucleotide polymorphisms (Feuk et al., 2006) as shown in Figure 1.1.

CNA is a key genetic event in the development and progression of numerous human diseases including cancer, HIV acquisition, autoimmune diseases, and Alzheimer and Parkinsons disease (Pollack et al., 2002; Redon et al. 2006; Beck et al. 2007; Lupski et al. 2007). So far the CNAs reported have been documented in the TCAG database of genomic variants(MacDonald et al., 2013), accounts for roughly 12% of human genomic DNA. Recent advances in high density microarray technologies (Pinkel et al. 1998; Pollack et al. 1999; Snijders et al. 2001; Bignell et al. 2004; Ishkanian et al. 2004; Peiffer et al. 2006; Pinkel et al. 2005) enable high-throughput genome-wide profiling of DNA copy number, hence enabling systematic study of their involvement in diseases. For a given cell sample, these technologies

allow measurement of the average genomic DNA copy number at thousands of locations linearly ordered along the chromosomes. Array-based Comparative Genomic Hybridization (aCGH) is currently the main technology to detect genomic copy number changes(gains or losses) in DNA. For a given cell sample, the average DNA copy number at several thousands of locations along the chromosomes can be quantitatively measured. DNA from a test sample is chopped up into short fragments, then these fragments are labeled with a fluorescent dye of a specific color, while DNA from normal reference sample is labeled with a dye of a different color. The two genomic DNAs, test and reference, are then jointly hybridize to the array of several thousands of short sequences of DNA probes. Hybridization to the probes emits fluorescence, which can be quantified. Because the test and reference are labeled with different colored fluorescent dyes that can be measured independently, the intensity ratio of the test and reference hybridization signals, usually expressed by log 2, gives an indirect measure of copy number of each probe. The aCGH protocol is shown in Figure 1.2. The potential of aCGH has been widely used to identify functional genetic mutations involved in cancer (Beroukhim et al.,2010). SNP array-based technology is another array-based technology, which also enable the detection of CNA (Wang et al., 1998), although it was originally developed for detecting SNP. In SNP array, it is single label, no reference DNA sample is necessary. Instead, CNA is identified via comparing the probe intensities of the test DNA with library of location specific empirical distribution. Besides the array-based technologies, NGS based approaches have also been used in the last few years in detecting CNA in human genomes. This type of method based on the short read data provide an unbiased and comprehensive view of genetic variations, their performance is not fully understood due to the relatively young age of the procedures. It is important to note that, locating CNA in individual samples is only the initial step in the search for interesting genes, but the cancer driver genes are more likely to be found in common or recurrent regions among samples (Korbel et al., 2007; Lai., 2005; Willenbrock et al., 2005; Rueda et al., 2007). Generally,

we can define a recurrent CNAs region as "a set of continuous probes that show a high enough evidence to being altered in at least some samples" (Rueda and Diaz-Uriarte, 2010), as shown in Figure 1.3.

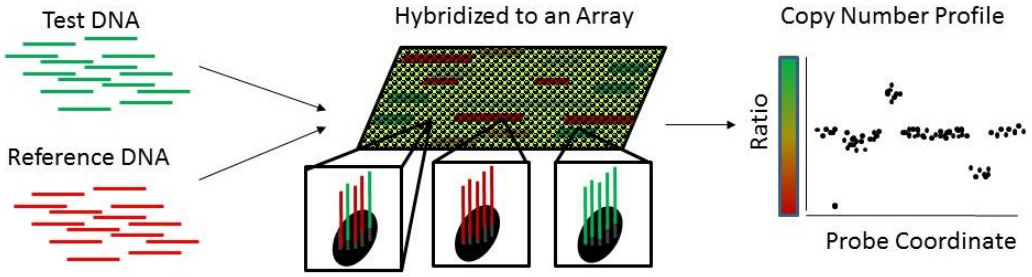
Figure 1.1: Illustration: Copy Number Alteration on Chromosome 8.
(http://www.dnavision.com/biostatistics.php)



1.2 Overview of existing methods

Over the past ten years, a large number of computational and statistical methods have been developed to analyze DNA copy number data. These methods can be broadly divided into two categories: (1) single-sample procedures aimed at accurately identifying regions of gain and loss within an individual sample by various statistical techniques including simple thresholding, change-point models, wavelets, hidden Markov models, and lowess smoothing

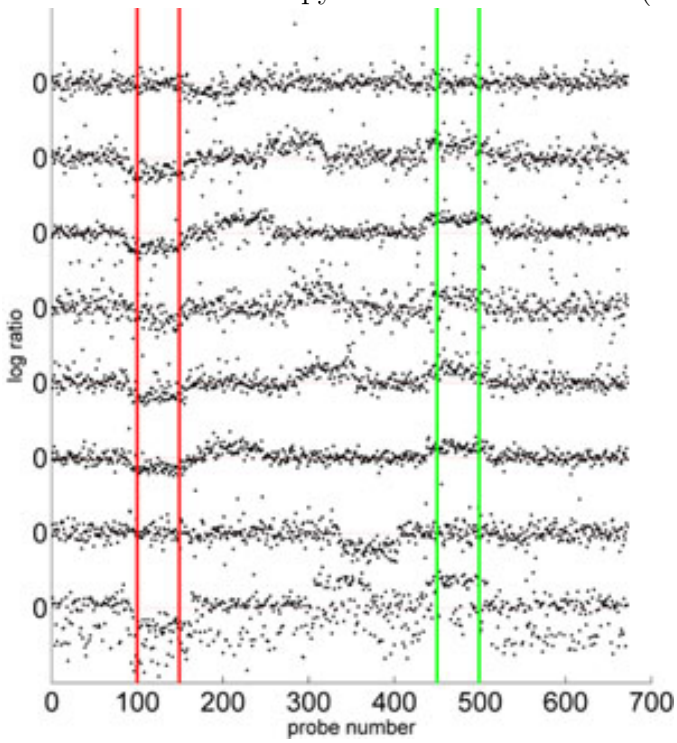
Figure 1.2: Illustration: Array-CGH protocol.
(http://compbio.cs.brown.edu/projects/structvar/)



(for recent reviews and comparisons, see Lai et al., 2005; Willenbrock and Fridlyand, 2005); (2) multi-sample methods aimed at providing an integrative analysis of copy number aberrations across samples (Lipson et al., 2006; Diskin et al., 2006; Rouveirol et al., 2006; Shah et al., 2007; Beroukhim et al., 2007; Guttman et al., 2007; Taylor et al., 2008; Zhang et al., 2009). Two recent reviews (Shah, 2008; Rueda OM and Diaz-Uriarte R, 2010) provide qualitative comparison of of existing methods. In this thesis, we focus on the problem of multi-sample CNAs analysis when the samples have pre-assigned group labels. We develop a method for estimating recurrent aberrations.

The common strategy for identifying recurrent CNAs is to first pre-process each individual sample and make calls of gains and losses, and then to infer recurrent CNAs using a threshold for the frequency of occurrence (Pollack et al., 2002; de Leeuw et al., 2004; Garnis 2006). We describe these methods as alternation frequency-based procedures. For example, Rouveirol et al. (2006) proposed an minimal alteration region (MAR) algorithm that takes as input a set of discretized sequences and outputs a set of minimal recurrent regions. This method works by converting the sequences to a S $*$ T binary matrix (gains and losses are analyzed separately), where S is the number of samples and T is the number of probes on the array. It then tries to find short blocks that are shared by a pre-specified fraction of the samples. Significance testing for aberrant copy number (STAC) proposed by Diskin et

Figure 1.3: Illustration: Recurrent Copy Number Alterations. (Shah et al., 2007)



al. (2006) also take a binary S $*$ T matrix as input, and use a greedy search procedure to find regions ("stacks") that are shared across samples with statistically significant frequency. Guttman et al. (2007) proposed a method, Multiple Sample Analysis (MSA), which can be considered an improvement over STAC. MSA uses the original ratio data, not segmented data, as input data, by searching over a set of possible cutoff values in the segmentation procedure. Genomic Identification of Significant Targets in Cancer (GISTIC) (Beroukhim et al., 2007) is another approach that uses segmentation information. In contrast to STAC and MSA, GISTIC not only considers the location and length of the aberration, but also considers the amplitude of the aberration across samples. In addition, GISTIC accepts continuous segmented input data generated by single sample analysis methods such as Gain and Loss Analysis of DNA (GLAD) algorithm (Hup et al., 2004) and circular binary segmentation (CBS) algorithm (Venkatraman et al., 2007) and define the G-score involving

both the amplitude of the aberration and the frequency of its occurrence across samples. Recently, Walter proposed a novel method Discovering Copy Number Aberrations Manifested In Cancer (DiNAMIC) which employs a novel permutation scheme called cyclic shift to assess the statistical significance of recurrent CNAs in multiple samples (Walter et al., 2011). This method accepts both continuous raw signal and segmented data, either discrete or continuous, since this method makes no distributional assumptions. Morganella et al. (2011) proposed a Genomic Analysis of Important Alterations (GAIA) approach which uses within-sample homogeneity to find recurrent CNAs where a statistical hypothesis model is based on a conservative permutation test. To access statistical significance of CNAs, all these methods, STAC, MSA, GISTIC, DiNAMIC and GIGA, use a permutation approach. GAIA, GISTIC uses multiple hypothesis testing via false discovery rate (FDR) control for the CNAs, while DiNAMIC, STAC and MSA use max-T procedure to control the family-wise error rate (FWER)(Westfall PH and Young SS, 1993). It is important to note that, these above two-step methods require pre-segmentation of each sample, and do not pool information across samples at the raw data stage. While alteration-based procedures may detect some common signals, pre-processing or discretizing the sequences separately may remove information by smoothing over short or low-amplification CNAs.

For finding shared patterns using the raw data to avoid problems with premature thresholding, one stage method which does not require a prior segmentation step will be discussed here. Lipson et al. (2006) proposed a combined scan statistic and provides the statistical significance for each CNA calling. Shah et al. (2007) suggested a multi-layer hierarchical hidden Markov model (HMM) to simultaneously segment all samples.This model assumes that changes must occur in the same direction with the same magnitude. An MCMC algorithm is used to estimate parameters in this model. Klijn et al. (2008) proposed a Kernel Convolution: a Statistical Method for Aberrant Region deTection (KC-SMART) which constructs a

statistic, kernel smoothed estimate (KSE) base on a kernal function. The correlations among copy number data and information from neighboring markers are considered in the statistic, KSE. Zhang et al. (2010) proposed a simultaneous change-point model and a likelihood-based framework for pooling data across multiple samples, and showed by using replicate samples and family trios that pooling data across samples pre-segmentation improves the power of calling short copy number variants. Zhang Q et al. (2010) introduced Correlation Matrix Diagonal Segmentation (CMDS) algorithm, directly using the raw copy number data (intensity ratio) across a set of samples, without the pre-analysis of single sample. This method constructs a Recurrent CNAs (RCNA) score using an easily implemented and fast diagonal transformation technique, the significance of RCNA score can be accessed based on a normal distribution, and then the significance of RCNA regions can be determined. Sergii et al. (2010) presented a CNAnova for identifying recurrent CNAs using unsegmented SNP microarray data .This model can be considered as a one-way analysis of variance and does not require permutation strategy to generate the null distribution. Instead, the distribution can be approximated using probe intensities in normal samples. Recently, Ewald et al. (2013) proposed a Analytical Multi-scale Identification of Recurring Events(ADMIRE), which performs a different kernel smoothing methodology from KC-SMART on the aggregated profile that gains power for detecting recurrent CNAs.

None of the existing methods incorporate grouping information for the samples. Such information is often available and may yield additional insights. For example, in tumor data, we often know the disease stage, lymph node status, and information regarding other biomarkers for each biological sample. In population-wide studies, samples may be categorized by ethnicity or by a particular phenotype. It is often of interest in studies to find regions of the genome that is altered more frequently in a specific group.

In this thesis, we propose for the analysis of recurrent CNAs a new stochastic segmenta-

tion model that incorporate grouping information for multiple samples. We tend to classify the states of copy number as three regimes, amplification, deletion, and baseline. The mean of copy number at different periods have different values, even they belong to the same regime. So in our model, parameter is a random variable instead of a constant in the classic model, following some regime-specific distribution within each regime. The model parameters are represented by a three state irreducible Hidden Markov Chain with transition probability matrix. Our model uses a Bayesian framework, so contains certain hyperparameters. Here I applied Expectation Maximization (EM) algorithm to estimate the hyperparameters. Furthermore, in order to improve the computational efficiency of the estimation, I proposed an approximation to the Bayes estimates, named Bounded Complexity Mixture (BCMIX), which uses only a fixed number of filters, thus reducing the computational complexity.

1.3 A Motivating Question

The classic segmentation model, also called Hidden Markov Model (HMM), in which the parameter is a constant at different periods within each regime, has been comprehensively used for many application, including CNA detection (Fridlyand et al., 2004). However, the mean of copy number data at different periods might not necessarily be same, even they belong to the same regime (amplification, deletion or baseline). To address this issue, Lai and Xing developed a novel Bayesian segmentation model for array-CGH data (Lai et al., 2008). However studies of DNA copy number usually involve many samples. It is often of interest to find recurrent CNAs for grouped samples. So we want to develop our novel stochastic segmentation model under Bayesian framework, which incorporate grouping information for multiple samples. The model is a mixture model, or hierarchical model with the top layer being a finite-state Markov chain and the middle layer being continuous-state Markov chain. Explicit formulas instead of Monte Carlo simulations will be adopted to smooth the

parameters in the model.

1.4 Outline

This dissertation research is based on the motivation mentioned above. This research developed a new model to detect recurrent CNAs in real grouped arrary-CGH data set. It studied the estimation of parameters in a stochastic segmentation model and explored its application to ovarian cancer copy number data. Chapter 2 proposes the stochastic regime switching model, and describes smoothing estimates of parameters and the posterior state probability with explicit formula. Since the proposed model uses a Bayesian framework, and hence contains certain hyperparameters, EM algorithm is used to estimate those hyperparameters. Furthermore, to improve the computational efficiency of the estimation, an approximation to the Bayes estimates is proposed, BCMIX, which uses only a fixed number of filters. Chapter 3 describes the extensive numerical simulation study to check whether our model performs well or not. Three types of simulation studies were designed to evaluates the performance of our method on simulated data. The first type of simulation is to demonstrate its accuracy and robustness compared to the fictitious Bayes (fBayes). The second simulation study exactly follows our model assumption and tests the performance under different simulation settings without using EM algorithm to estimate hyperparameters. In the third simulation study, we examine the effects of different simulation setting on the estimates using EM algorithm. Furthermore, we compared our method to an existing hierarchical HMM method, which shows advantages of our model over current state-of-the-art method. The last part of this section assumes regime switching locations across all samples are non-simultaneous, which is more close to what happens in real data, and tests the performance of our model for non-simultaneous changes. In Chapter 4, our model is applied to analyze ovarian cancer copy number data downloaded from the The Cancer Genome Atlas (TCGA)

to detect recurrent CNAs in fifteen ovarian cancer patients. In the end, some concluding remarks and discussion are given in Chapter 5.

Chapter 2

Estimation in a Novel Stochastic Segmentation Model

2.1 Model Specification

We consider the following model assumptions:

1. The log ratios y_{lt} of CNAs follow the model $y_{lt} = \theta_{lt} + \epsilon_{lt}$ for $l = 1, \ldots, J$ and $t = 1, \ldots, T$, where ϵ_{lt} are independent normal random variables with mean 0 and variance σ_l^2.

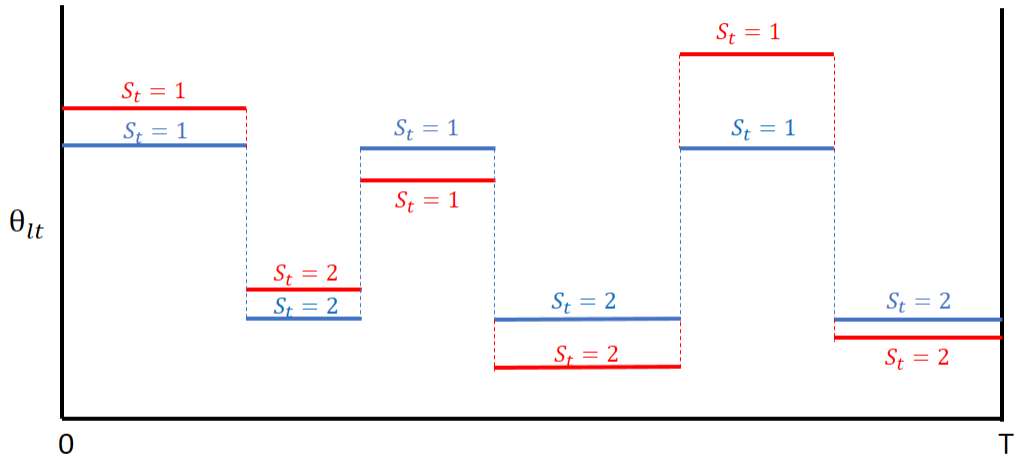
2. The recurrent CNAs across the J's sample are represented by a three-state irreducible hidden Markov chain $\{s_t\}$ with transition probability matrix $P = (p_{ij})$ and stationary distribution π. Here the three states represent the states of copy number "gains", "losses", and "baseline value 0".

3. For each l, the dynamics of $\{\theta_{lt}\}_{t=1,\ldots,T}$ given $\{s_t\}$ is given by $\theta_{lt} = 1_{\{s_t = s_{t-1}\}}\theta_{l,t-1} + 1_{\{s_t \neq s_{t-1}\}}z_{lt}$, in which z_{lt} are independent normal variables with mean $z^{(l,s_t)}$ and covariance $V^{(l,s_t)}$.

Note that the classic segmentation model does not satisfy the third assumption. In the classic model, the parameter is a constant based on the regime. The third assumption adds

the new feature to our novel model by allowing the parameter as a random variable following some regime-specific distribution within each regime. Let's take a two-state scenario as an example. As shown in Figure 2.1, the red lines demonstrates the assumptions of our model, showing an example of possible values of a one-dimensional θ_{lt}. Four transitions occur during the period $0 \leq t \leq T$. Within each state, θ_{lt} take different values instead of a constant in the classic model which is represented by the blue lines. The values are realizations from the state-specific distribution of $\theta_l(s_t)$. The transitions are governed by some hidden Markov chain.

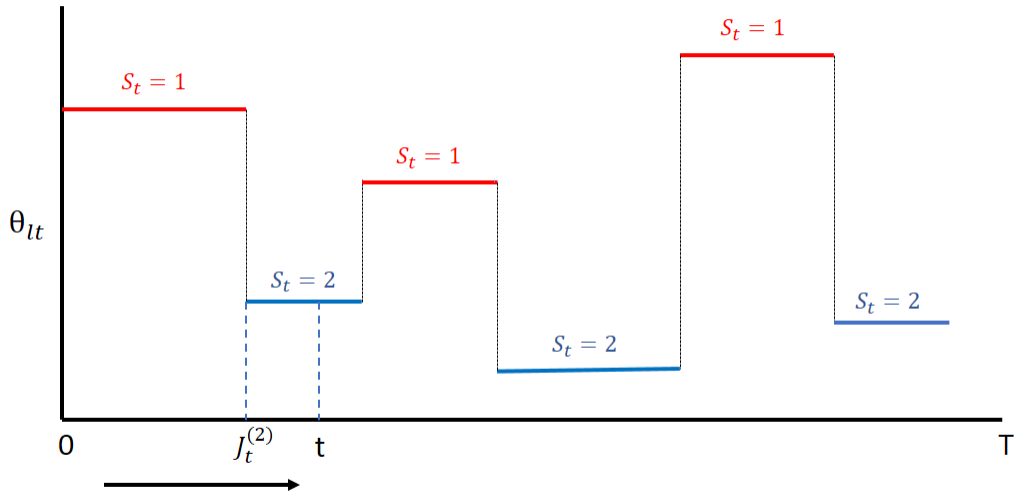
Figure 2.1: Illustration: A comparison between values of $\theta_l(s_t)$ in a classic stochastic segmentation model (blue line) and in our novel stochastic segmentation model (red line).



2.2 The Forward Filtering Estimate of Parameters

Let $J_t^{(k)} = \max\{i \leq t : s_{i-1} \neq s_i = \cdots = s_t = k\}$ be the most recent switching time prior or equal to t and at which s_t switches from a regime other than k to regime k. The definition of $J_t^{(k)}$ is illustrated in Figure 2.2. At time t, $s_t = 2$, and the most recent switching occurs before t is at time $J_t^{(2)}$ as shown in the figure.

Figure 2.2: Illustration: Definition of $J_t^{(k)}$.



Let

$$\xi_t^{(k)} = P(s_t = k|\mathcal{Y}_t), \qquad \xi_{i,t}^{(k)} = P(J_t^{(k)} = i|\mathcal{Y}_t), \tag{2.2.1}$$

for $1 \leq i \leq t$ and $1 \leq k \leq K$, in which $\mathcal{Y}_t = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$ and $\mathcal{Y}_{i,j} = (\boldsymbol{y}_i, \ldots, \boldsymbol{y}_j)$, $\boldsymbol{y} = (y_1, \ldots, y_J)'$, then, by definition, $\xi_t^{(k)} = \sum_{i=1}^{t} \xi_{i,t}^{(k)}$. If we know all the historical information up to time t, $\mathcal{Y}_t$, and that the recent transition occurs at time i from some regime to regime k, we just need to use the information after this transition to estimate the current value of

θ_{lt}.

Since $y_{lt} = \theta_{lt} + \epsilon_{lt}$ for $l = 1, \ldots, J.$ where $\epsilon_{lt} \sim N(0, \sigma_l^2)$, and $\theta_{lt} \sim N(z^{(l,k)}, V^{(l,k)})$. The posterior distribution of θ_{lt} given $\mathcal{Y}_t$ and $J_t^{(k)} = i$ is:

$$\begin{aligned} f(\theta_{lt}|\mathcal{Y}_{it}) &\propto \prod_{j=i}^{t} f(y_{lj}|\theta_{lt}) \cdot f(\theta_{lt}) \\ &\propto \prod_{j=i}^{t} \exp\big(-\frac{(y_{lj}-\theta_{lt})^2}{2\sigma_l^2}\big) \cdot \exp\big(-\frac{(\theta_{lt}-z^{(l,k)})^2}{2V^{(l,k)}}\big) \\ &\propto \exp\big(-\frac{(\theta_{lt}-z_{i,t}^{(l,k)})^2}{2V_{i,t}^{(l,k)}}\big), \end{aligned}$$

where

$$V_{i,j}^{(l,k)} = \Big(\big[V^{(l,k)}\big]^{-1} + \frac{j-i+1}{\sigma_l^2}\Big)^{-1}, z_{i,j}^{(l,k)} = V_{i,j}^{(l,k)}\Big(\big[V^{(l,k)}\big]^{-1} z^{(l,k)} + \frac{1}{\sigma_l^2}\sum_{u=i}^{j} y_{lu}\Big),$$

for $j \geq i$.

Thus the conditional distribution of θ_{lt} given $\mathcal{Y}_t$ and $J_t^{(k)} = i$, is $g_{i,t}^{(lk)}(\theta_{lt})$ which is defined as

$$\theta_{lt}|\{\mathcal{Y}_t, J_t^{(k)} = i\} \sim N(z_{i,t}^{(lk)}, V_{i,t}^{(lk)}), \tag{2.2.2}$$

If no historical information is given and only the event $s_t = k$ is known, the conditional distribution of θ_{lt} is $g_{0,0}^{(lk)}(\theta_{lt})$ which is defined as

$$\theta_{lt}|\{s_t = k\} \sim N(z^{(l,k)}, V^{(l,k)}). \tag{2.2.3}$$

It follows that the posterior distribution of θ_{lt} given $\mathcal{Y}_t$ is a mixture of normal distributions:

$$\theta_{lt}|\mathcal{Y}_t \sim \sum_{k=1}^{K}\sum_{i=1}^{t} \xi_{i,t}^{(k)} \text{Normal}(z_{i,t}^{(lk)}, V_{i,t}^{(lk)}). \tag{2.2.4}$$

Let us see how to derive the mixture weight $\xi_{i,t}^{(k)}$. First note that

$$f(\theta_{lt}, \boldsymbol{y}_t, s_{t-1} = k|\mathcal{Y}_{t-1}) = \sum_{r=1}^{K} f(\theta_{lt}, \boldsymbol{y}_t, s_{t-1} = k, s_t = l|\mathcal{Y}_{t-1}).$$

When $r \neq k$,

$$\begin{aligned}
&f(\theta_{lt}, \boldsymbol{y}_t, s_{t-1} = k, s_t = l|\mathcal{Y}_{t-1}) \\
=&f(\theta_{lt}, \boldsymbol{y}_t|\mathcal{Y}_{t-1}, s_{t-1} = k, s_t = r)P(s_{t-1} = k, s_t = r|\mathcal{Y}_{t-1}) \\
=&f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(r)} = t)f(\theta_{lt}|\mathcal{Y}_t, J_t^{(r)} = t)P(s_t = r|s_{t-1} = k)P(s_{t-1} = k|\mathcal{Y}_{t-1}) \\
=&f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(r)} = t)g_{t,t}^{(l,r)}(\theta_t)p_{k,r}\xi_{t-1}^{(k)}.
\end{aligned}$$

When $r = k$,

$$\begin{aligned}
&f(\theta_{lt}, \boldsymbol{y}_t, s_{t-1} = k, s_t = k|\mathcal{Y}_{t-1}) = \sum_{i=1}^{t-1} f(J_t^{(k)} = i, \theta_{lt}, \boldsymbol{y}_t|\mathcal{Y}_{t-1}) \\
=&\sum_{i=1}^{t-1} f(\theta_{lt}, \boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(k)} = i)P(s_{t-1} = k, s_t = k|\mathcal{Y}_{t-1}) \\
=&\sum_{i=1}^{t-1} f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(k)} = i)f(\theta_{lt}|\mathcal{Y}_t, J_t^{(k)} = i)P(s_t = k|s_{t-1} = k)P(s_{t-1} = k|\mathcal{Y}_{t-1}) \\
=&\sum_{i=1}^{t-1} f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(k)} = i)g_{i,t}^{(l,k)}(\theta_t)p_{k,k}\xi_{i,t-1}^{(k)}.
\end{aligned}$$

Define

$$\xi_{i,t}^{(k)} \propto \xi_{i,t}^{(k)*} := \begin{cases} \big(\sum_{r\neq k} \xi_{t-1}^{(r)} p_{rk}\big) f(\boldsymbol{y}_t|J_t^{(k)} = t) & i = t, \\ p_{kk}\xi_{i,t-1}^{(k)} f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(k)} = i) & i < t. \end{cases} \tag{2.2.5}$$

So, we will study $f(\boldsymbol{y}_t|J_t^{(k)}=t)$ and $f(\boldsymbol{y}_t|\mathcal{Y}_{t-1},J_t^{(k)}=i)$, when $i=t$.

$$f(\boldsymbol{y}_t|J_t^{(k)}=t)=\int f(\boldsymbol{y}_t|\theta_{lt},J_t^{(k)}=t)f(\theta_{lt}|J_t^{(k)}=t)d\theta_{lt}.$$

$$\begin{aligned}
&f(\boldsymbol{y}_t|\theta_{lt},J_t^{(k)}=t)f(\theta_{lt}|J_t^{(k)}=t)\\
=&\prod_{l=1}^{J}\frac{1}{\sqrt{2\pi\sigma_l^2}}\exp\Big\{-\frac{(y_{lt}-\theta_{lt})^2}{2\sigma_l^2}\Big\}\frac{1}{\sqrt{2\pi V^{(l,k)}}}\exp\Big\{-\frac{(\theta_{lt}-z^{(l,k)})^2}{2V^{(l,k)}}\Big\}\\
=&\prod_{l=1}^{J}\frac{1}{\sqrt{2\pi\sigma_l^2V^{(l,k)}}}\exp\Big\{-\frac{(y_{lt}-\theta_{lt})^2}{2\sigma_l^2}-\frac{(\theta_{lt}-z^{(l,k)})^2}{2V^{(l,k)}}\Big\}\\
=&\prod_{l=1}^{J}\frac{1}{\sqrt{2\pi\sigma_l^2V^{(l,k)}}}\exp\Big\{-\frac{(\theta_{lt}-\widetilde{z})^2}{2\widetilde{V}}-\frac{(z^{(l,k)})^2}{2V^{(l,k)}}-\frac{y_{lt}^2}{2\sigma_l^2}+\frac{\widetilde{z}^2}{2\widetilde{V}}\Big\}.
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{V}=&-\frac{1}{V^{(l,k)}}+\frac{1}{\sigma_l^2}\big)^{-1}=V_{t,t}^{(l,k)},\\
\widetilde{z}=&\widetilde{V}\big(\frac{z^{(l,k)}}{V^{(l,k)}}+\frac{y_{lt}}{\sigma_l^2}\big)=z_{t,t}^{(l,k)}.
\end{aligned}$$

Denote $g_{i,j}^{(l,k)}(u)$ the density function of the Normal$(z_{i,t}^{(l,k)},V_{i,t}^{(l,k)})$ distribution at point u, i.e., $g_{i,j}^{(l,k)}(u)=(2\pi V_{i,j}^{(l,k)})^{-1/2}\exp\{-\frac{(u-z_{i,j}^{(l,k)})^2}{2V_{i,j}^{(l,k)}}\}$,

Thus

$$\begin{aligned}
&f(\boldsymbol{y}_t|\theta_{lt},J_t^{(k)}=t)f(\theta_{lt}|J_t^{(k)}=t)\\
=&\prod_{l=1}^{J}\frac{g_{z_{t,t}^{(l,k)},V_{t,t}^{(l,k)}}(\theta_t)g_{z^{(l,k)},V^{(l,k)}}(o)g_{0,\sigma_l^2}(y_{lt})}{g_{z_{t,t}^{(l,k)},V_{t,t}^{(l,k)}}(o)}.
\end{aligned}$$

Therefore

$$f(\boldsymbol{y}_t|J_t^{(k)}=t)=\prod_{l=1}^{J}\int\frac{g_{z_{t,t}^{(l,k)},V_{t,t}^{(l,k)}}(\theta_t)g_{z^{(l,k)},V^{(l,k)}}(o)g_{0,\sigma_l^2}(y_{lt})}{g_{z_{t,t}^{(l,k)},V_{t,t}^{(l,k)}}(o)}d\theta_{lt}=\frac{\psi_{0,0}^{(k)}}{\psi_{t,t}^{(k)}}\psi_{0,\sigma_l^2}(y_{lt}).$$

where $\psi_{0,0}^{(k)}=\prod_{l=1}^{J}g_{0,0}^{(l,k)}(0)$, and $\psi_{i,j}^{(k)}=\prod_{l=1}^{J}g_{i,j}^{(k)}(0)$.

The second conditional density can be transferred to a similar integral

$$f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(k)}=i) = \int f(\boldsymbol{y}_t|\theta_{lt}, \mathcal{Y}_{t-1}, J_t^{(k)}=i) f(\theta_{lt}|\mathcal{Y}_{t-1}, J_t^{(k)}=i) d\theta_{lt},$$

$$\begin{aligned}
& f(\boldsymbol{y}_t|\theta_{lt}, \mathcal{Y}_{t-1}, J_t^{(k)}=i) f(\theta_{lt}|\mathcal{Y}_{t-1}, J_t^{(k)}=i) \\
=& \prod_{l=1}^{J} \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\Big\{ -\frac{(y_{lt}-\theta_{lt})^2}{2\sigma_l^2} \Big\} \frac{1}{\sqrt{2\pi V^{(l,k)}}} \exp\Big\{ -\frac{(\theta_{lt}-z_{i,t-1}^{(l,k)})^2}{2V_{i,t-1}^{(}l,k)} \Big\} \\
=& \prod_{l=1}^{J} \frac{1}{\sqrt{2\pi\sigma_l^2 V^{(l,k)}}} \exp\Big\{ -\frac{(y_{lt}-\theta_{lt})^2}{2\sigma_l^2} - \frac{(\theta_{lt}-z_{i,t-1}^{(l,k)})^2}{2V_{i,t-1}^{(l,k)}} \Big\} \\
=& \prod_{l=1}^{J} \frac{1}{\sqrt{2\pi\sigma_l^2 V^{(l,k)}}} \exp\Big\{ -\frac{(\theta_{lt}-\widetilde{z})^2}{2\widetilde{V}} - \frac{z_{i,t-1}^{2(l,k)}}{2V_{i,t-1}^{(l,k)}} - \frac{y_{lt}^2}{2\sigma_l^2} + \frac{\widetilde{z}^2}{2\widetilde{V}} \Big\}.
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{V} =& -\frac{1}{V_{i,t-1}^{(l,k)}} + \frac{1}{\sigma_l^2}\big)^{-1} = V_{i,t}^{(l,k)}, \\
\widetilde{z} =& \widetilde{V}\big(\frac{z_{i,t-1}^{(l,k)}}{V_{i,t-1}^{(l,k)}} + \frac{y_{lt}}{\sigma_l^2}\big) = z_{i,t}^{(l,k)}.
\end{aligned}$$

Thus

$$f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(k)}=i) = \frac{\psi_{i,t-1}^{(k)}}{\psi_{i,t}^{(k)}} \psi_{0,\sigma_l^2}(y_{lt}).$$

Then

$$\frac{f(\boldsymbol{y}_t|J_t^{(k)}=t)}{f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}, J_t^{(k)}=i)} = \frac{\psi_{0,0}^{(k)}/\psi_{t,t}^{(k)}}{\psi_{i,t-1}^{(k)}/\psi_{i,t}^{(k)}}. \tag{2.2.6}$$

Plugging (2.2.6) into (2.2.5) yielding $\xi_{i,t}^{(k)} = \frac{\xi_{i,t}^{(k)*}}{\sum_{k=1}^{K}\sum_{i=1}^{t}\xi_{i,t}^{(k)*}}$, where

$$\xi_{i,t}^{(k)*} := \begin{cases} \big(\sum_{r\neq k}\xi_{t-1}^{(r)} p_{rk}\big)\psi_{0,0}^{(k)}/\psi_{t,t}^{(k)} & i=t, \\ p_{kk}\xi_{i,t-1}^{(k)}\psi_{i,t-1}^{(k)}/\psi_{i,t}^{(k)} & i<t, \end{cases} \tag{2.2.7}$$

Then expressions (2.2.1) and (2.2.4) imply

$$P(s_t = k|\mathcal{Y}_t) = \sum_{i=1}^{t} \xi_{i,t}^{(k)}, \qquad E(\theta_{lt}|\mathcal{Y}_t) = \sum_{k=1}^{K}\sum_{i=1}^{t} \xi_{i,t}^{(k)} z_{i,t}^{(l,k)}. \tag{2.2.8}$$

2.3 The Backward Filtering Estimate of Parameters

The model assumption implies that, a stationary distribution of θ_{lt} exists and is given by

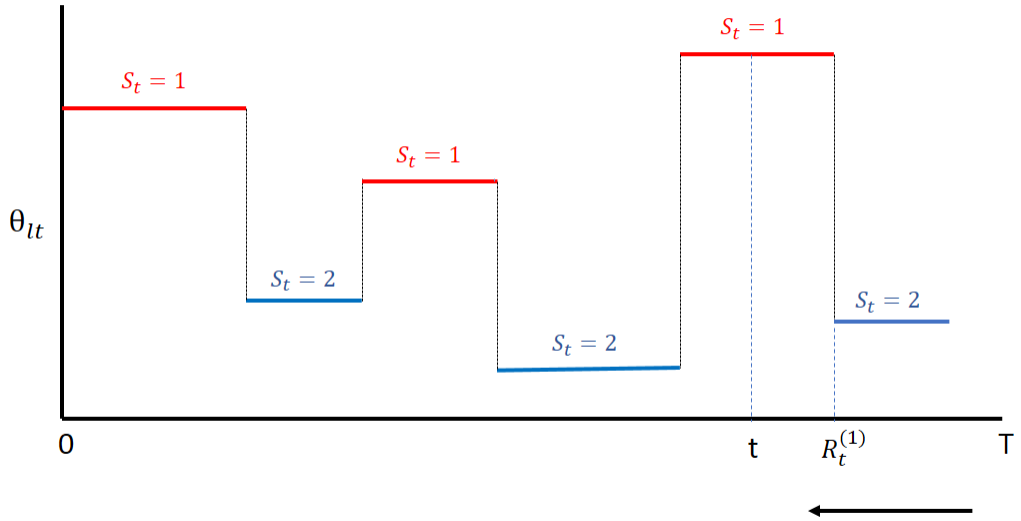
$$\sum_{k=1}^{K} \pi_k \text{Normal}(z^{(l,k)}, V^{(l,k)}). \tag{2.3.1}$$

This indicates that, if θ_{lt} is initialized at the stationary distribution, its time-reversed Markov chain can be defined. This substantially simplifies the smoothing estimates of θ_{lt}. Note that this also imposes stationarity conditions for $\boldsymbol{y}_t$, for instance, if the regression model has an autoregressive component, the stationarity condition for components of θ_{lt} should be imposed. In such case, we shall replace the Normal distribution in (2.3.1) by a truncated Normal distribution that has support in stability region. Such treatment also applies for the smoothing estimates of θ_{lt}. For notational convenience, we still use Normal (instead of truncated Normal) in the sequel.

As indicated at the end of Section 2.1, $\{\theta_{lt}\}$ is a reversible Markov chain. Therefore we can obtain a backward filter that is analogous to (2.2.4). That is, we reverse the time, starting with time T and estimating θ_{lt} for any time t given the "historical" information from time T to t.

Let $R_t^{(k)} = \min\{j \geq t : k = s_t \cdots = s_{j-1} \neq s_j\}$ be the most recent switching time larger than or equal to t when s_t switches from the regime k to another regime. Figure 2.3 illustrates the definition of $R_t^{(k)}$. At time t, the regime is $s_t = 1$, and the most recent transition occurs after t is at $R_t^{(1)}$ as shown in Figure 2.3.

Figure 2.3: Illustration: Definition of $R_t^{(k)}$.



Define

$$\eta_t^{(k)} = P(s_t = k|\mathcal{Y}_{t,T}), \qquad \eta_{j,t}^{(k)} = P(R_t^{(k)} = j|\mathcal{Y}_{t,T})$$

We then use the time-reversed chain of θ_{lt} to obtain a backward analog of (2.2.4),

$$\theta_{l,t+1}|\mathcal{Y}_{t+1,T} \sim \sum_{k=1}^{K}\sum_{j=t+1}^{T} \eta_{t+1,j}^{(k)} \text{Normal}(z_{t+1,j}^{(l,k)}, V_{t+1,j}^{(l,k)}), \tag{2.3.2}$$

in which the weights $\eta_{t+1,j}^{(k)}$ can be obtained by backward induction using the time-reversed counterpart of (2.2.7):

$$\eta_{t+1,j}^{(k)} \propto \eta_{t+1,j}^{(k)*} := \begin{cases} \left(\sum_{r\neq k} \eta_{t+2}^{(r)} \widetilde{p}_{rk}\right) \psi_{0,0}^{(k)} / \psi_{t+1,t+1}^{(k)} & j = t+1, \\ \widetilde{p}_{kk} \eta_{t+2,j}^{(k)} \psi_{t+2,j}^{(k)} / \psi_{t+1,j}^{(k)} & j > t+1, \end{cases} \tag{2.3.3}$$

where $\widetilde{P} = (\widetilde{p}_{rk})$ is the transition matrix of the reversed chain of $\{s_t\}$, and $\widetilde{p}_{rk} = P(s_t = k|s_{t+1} = r)$.

Thus

$$\begin{aligned}\theta_{lt}|\mathcal{Y}_{t+1,T} \sim \sum_{k=1}^{K} \Big\{ \widetilde{p}_{kk} \sum_{j=t+1}^{T} \eta_{t+1,j}^{(k)} \text{Normal}(z_{t+1,j}^{(l,k)}, V_{t+1,j}^{(l,k)}) \\ + \Big(\sum_{l \neq k} \widetilde{p}_{rk} \eta_{t+1}^{(r)} \Big) \text{Normal}(z^{(l,k)}, V^{(l,k)}) \Big\}.\end{aligned} \tag{2.3.4}$$

Then we can go one step further to calculate $f(\theta_{lt}|\mathcal{Y}_{t+1,T})$.

$$\begin{aligned}f(\theta_{lt}|\mathcal{Y}_{t+1,T}) &= \sum_{k=1}^{K} f(\theta_{lt}, s_{t+1} = k|\mathcal{Y}_{t+1,T}) = \sum_{k=1}^{K} P(s_{t+1} = k|\mathcal{Y}_{t+1,T}) f(\theta_{lt}|s_{t+1} = k, \mathcal{Y}_{t+1,T}) \\ &= \sum_{k=1}^{K} \sum_{r=1}^{K} P(s_{t+1} = k|\mathcal{Y}_{t+1,T}) f(\theta_{lt}, s_t = r|s_{t+1} = k, \mathcal{Y}_{t+1,T}) \\ &= \sum_{k=1}^{K} \sum_{r=1}^{K} P(s_{t+1} = k|\mathcal{Y}_{t+1,T}) P(s_t = r|s_{t+1} = k) f(\theta_{lt}|s_{t+1} = k, s_t = r, \mathcal{Y}_{t+1,T}) \\ &= \sum_{k=1}^{K} \sum_{r=1}^{K} P(s_{t+1} = k|\mathcal{Y}_{t+1,T}) \widetilde{p}_{kr} f(\theta_{lt}|s_{t+1} = k, s_t = r, \mathcal{Y}_{t+1,T}).\end{aligned}$$

When $k = r$,

$$\begin{aligned}&\widetilde{p}_{kk} P(s_{t+1} = k|\mathcal{Y}_{t+1,T}) f(\theta_{lt}|s_{t+1} = k, s_t = k, \mathcal{Y}_{t+1,T}) \\ =&\widetilde{p}_{kk} f(\theta, s_{t+1} = k|\mathcal{Y}_{t+1,T})\Big|_{\theta=\theta_{lt}} \\ =&\widetilde{p}_{kk} \sum_{j=t+1}^{T} f(\theta, s_{t+1} = k, R_t^{(k)} = j|\mathcal{Y}_{t+1,T})\Big|_{\theta=\theta_{lt}} \\ =&\widetilde{p}_{kk} \sum_{j=t+1}^{T} P(R_t^{(k)} = j|\mathcal{Y}_{t+1,T}) f(\theta, s_{t+1} = k|R_t^{(k)} = j, \mathcal{Y}_{t+1,T})\Big|_{\theta=\theta_{lt}} \\ =&\widetilde{p}_{kk} \sum_{j=t+1}^{T} \eta_{t+1,j}^{(k)} g_{t+1,j}^{(k)}(\theta)\Big|_{\theta=\theta_{lt}};\end{aligned}$$

when $k \neq r$,

$$\begin{aligned}&\widetilde{p}_{kr} P(s_{t+1} = k|\mathcal{Y}_{t+1,T}) f(\theta_{lt}|s_{t+1} = k, s_t = r, \mathcal{Y}_{t+1,T}) \\ =&\widetilde{p}_{kr} \eta_{t+1}^{(k)} f(\theta|s_t = r)\Big|_{\theta=\theta_{lt}} = \widetilde{p}_{kr} \eta_{t+1}^{(k)} g_{0,0}^{(l,r)}(\theta)\Big|_{\theta=\theta_{lt}}.\end{aligned}$$

Thus

$$\begin{aligned}f(\theta_{lt}|\mathcal{Y}_{t+1,T}) &= \sum_{k=1}^{K} f(\theta_{lt}, s_{t+1} = k|\mathcal{Y}_{t+1,T}) = \sum_{k=1}^{K} P(s_{t+1} = k|\mathcal{Y}_{t+1,T}) f(\theta_{lt}|s_{t+1} = k, \mathcal{Y}_{t+1,T}) \\ &= \sum_{k=1}^{K}\sum_{r=1}^{K} P(s_{t+1} = k|\mathcal{Y}_{t+1,T}) \widetilde{p}_{kr} f(\theta_{lt}|s_{t+1} = k, s_t = r, \mathcal{Y}_{t+1,T}) \\ &= \sum_{k=1}^{K} \widetilde{p}_{kk} \sum_{j=t+1}^{T} \eta_{t+1,j}^{(k)} g_{t+1,j}^{(k)}(\theta_{lt}) + \sum_{k=1}^{K}\sum_{r\neq k} \widetilde{p}_{kr} \eta_{t+1}^{(r)} g_{0,0}^{(k)}(\theta_{lt}).\end{aligned}$$

2.4 Smoothing Estimate of Parameters

Next, we shall use Bayes' theorem to combine the forward filter (2.2.4) with its backward variant (2.3.4) to estimate θ_{lt} given $\mathcal{Y}_T$ $(1 \leq t < T, 1 \leq l < J)$

$$f(\theta_{lt}|\mathcal{Y}_T) = \sum_{k=1}^{K} f(\theta_{lt}, s_t = k|\mathcal{Y}_T) \propto \sum_{k=1}^{K} f(\theta_{lt}, s_t = k|\mathcal{Y}_t) f(\theta_{lt}, s_t = k|\mathcal{Y}_{t+1,T}) \big/ f(\theta_{lt}, s_t = k).$$

From this we can derive the posterior distribution of θ_{lt} given $\mathcal{Y}_T$. Applying Bayes' theorem,

$$g_t^{(l)}(\theta|\mathcal{Y}_T) = \sum_{k=1}^{K} g_t^{(l)}(\theta, s_t = k|\mathcal{Y}_T) \propto \sum_{k=1}^{K} g_t^{(l)}(\theta, s_t = k|\mathcal{Y}_t) g_t^{(l)}(\theta, s_t = k|\mathcal{Y}_{t+1,T}) \big/ f(\theta, s_t = k).$$

where $g_t^{(l)}(\cdot|\mathcal{Y}_T)$, $g_t^{(l)}(\cdot|\mathcal{Y}_t)$, and $g_t^{(l)}(\cdot|\mathcal{Y}_{t+1,T})$ denote the density functions of the absolutely continuous components of θ_{lt} given $\mathcal{Y}_T$, $\mathcal{Y}_t$, and $\mathcal{Y}_{t+1,T}$ respectively.

The right hand side is a mixture of different states.

$$
\begin{aligned}
& g_t^{(l)}(\theta, s_t = k|\mathcal{Y}_t) g_t^{(l)}(\theta, s_t = k|\mathcal{Y}_{t+1,T}) \big/ f(\theta, s_t = k) \\
=& \frac{\Big\{ \sum_{i=1}^{t} P(J_t^{(k)} = i|\mathcal{Y}_t) f(\theta_{lt}|\mathcal{Y}_t, J_t^{(k)} = i) \Big\} \Big\{ \sum_{r=1}^{K} P(s_{t+1} = r|\mathcal{Y}_{t+1,T}) f(\theta_{lt}, s_t = k|s_{t+1} = r, \mathcal{Y}_{t+1,T}) \Big\}}{P(s_t = k) f(\theta_{lt}|s_t = k)} \\
=& \frac{\Big\{ \sum_{i=1}^{t} \xi_{i,t}^{(k)} g_{i,t}^{(l,k)}(\theta) \Big\} \Big\{ \widetilde{p}_{kk} \sum_{j=t+1}^{T} \eta_{t+1,j}^{(k)} g_{t+1,j}^{(l,k)}(\theta) + \Big(\sum_{r \neq k} \widetilde{p}_{rk} \eta_{t+1}^{(r)} \Big) g_{0,0}^{(l,k)}(\theta) \Big\}}{\pi_k g_{0,0}^{(l,k)}(\theta)} \\
=& \sum_{i=1}^{t} \xi_{i,t}^{(k)} \Big(\sum_{r \neq k} \frac{\widetilde{p}_{rk}}{\pi_k} \eta_{t+1}^{(r)} \Big) g_{i,t}^{(l,k)}(\theta) + \frac{\widetilde{p}_{kk}}{\pi_k} \sum_{1 \leq i \leq t < j \leq T} \xi_{i,t}^{(k)} \eta_{t+1,j}^{(k)} \frac{g_{i,t}^{(l,k)}(\theta) g_{t+1,j}^{(l,k)}(\theta)}{g_{0,0}^{(l,k)}(\theta)}.
\end{aligned}
$$

Based on the reversibility of P,

$$
\begin{aligned}
\widetilde{p}_{kk} =& P(s_t = k|s_{t+1} = k) = \frac{P(s_t = k, s_{t+1} = k)}{P(s_{t+1} = k)} \\
=& \frac{P(s_t = k, s_{t+1} = k)}{P(s_t = k)} = P(s_{t+1} = k|s_t = k) = p_{kk}.
\end{aligned}
$$

So the posterior distribution of θ_{lt} given $\mathcal{Y}_T$ is a mixture of normal distributions:

$$
\theta_{lt}|\mathcal{Y}_T \sim \sum_{k=1}^{K} \sum_{1 \leq i \leq t \leq j \leq T} \alpha_{ijt}^{(k)} g_{i,j}^{(l,k)}(\theta_t), \tag{2.4.1}
$$

in which the mixture weights $\alpha_{ij,t}^{(k)}$ are posterior probabilities explained below. Consider the event

$$
C_{ij}^{(k)} = \{ s_i = \cdots = s_j = k, s_i \neq s_{i-1}, s_j \neq s_{j+1} \},
$$

Appendix A shows that, for $i \leq t \leq j$, $\alpha_{ijt}^{(k)} = P(C_{ij}^{(k)}|\mathcal{Y}_t)$ and $\alpha_{ij,t}^{(k)}$ can be calculated

recursively as

$$\alpha_{ijt}^{(k)} = \alpha_{ijt}^{(k)*} \Big/ D_t, \qquad D_t = \sum_{k=1}^{K} \sum_{1\le i\le t\le j\le T} \alpha_{ijt}^{(k)*},$$
$$\alpha_{ijt}^{(k)*} = \begin{cases} \xi_{i,t}^{(k)} \big(\sum_{r\neq k} \eta_{t+1}^{(r)} p_{kr}/\pi_r \big) & i \le t = j, \\ p_{kk} \xi_{i,t}^{(k)} \eta_{t+1,j}^{(k)} \psi_{i,j}^{(k)} \psi_{0,0}^{(k)} \big/ (\pi_k \psi_{i,t}^{(k)} \psi_{t+1,j}^{(k)}) & i \le t < j. \end{cases} \tag{2.4.2}$$

where $\psi_{0,0}^{(k)} = \prod_{l=1}^{J} g_{0,0}^{(l,k)}(0)$, and $\psi_{i,j}^{(k)} = \prod_{l=1}^{J} g_{i,j}^{(k)}(0)$.

Therefore, the smoothing estimates of θ_{lt} and s_t given $\mathcal{Y}_T$ are given by

$$E(\theta_{lt}|\mathcal{Y}_T) = \sum_{k=1}^{K} \sum_{1\le i\le t\le j\le T} \alpha_{ijt}^{(k)} z_{i,j}^{(l,k)}. \tag{2.4.3}$$

$$P(s_t = k|\mathcal{Y}_T) = \sum_{1\le i\le t\le j\le T} \alpha_{ijt}^{(k)}. \tag{2.4.4}$$

One concern here is that, since (2.4.3) are represented as K mixtures of mixtures of normals, it is questionable whether the smoothing formula could differentiate the values of θ_{lt} when K regimes are close to each other. Such identification issue is closed related to the choice of appropriate hyperparameters, and will be discussed in Section 2.6.

2.5 Bounded Complexity Mixture (BCMIX) Approximation

Although the Bayes filter (2.2.4) uses a recursive updating formula (2.2.7) for the weights $\xi_{i,t}^{(k)}$ $(1 \le i \le t, 1 \le k \le K)$, the number of weights increases dramatically with t, resulting in rapidly increasing computational complexity and memory requirements in estimating θ_{lt} as t keeps increasing. To address the issue of computational efficiency, we follow Lai and Xing (2011) and consider a Bounded Complexity Mixture (BCMIX) approximation procedure with much lower computational complexity yet comparable to the Bayes estimates in statistical

efficiency. The idea of BCMIX approximation is to keep only a fixed number M of weights at every stage t, in particular, the most recent m ($1 \leq m < M$) weights $\xi_{i,t}^{(k)}$ (with $t-m < i \leq t$) and the largest $M-m$ of the remaining weights.

Denote $\mathcal{K}_{t-1}^{(k)}$ the set of indices i for which $\xi_{i,t-1}^{(k)}$ in (2.2.7) is kept at stage $t-1$ for regime k. Note that there are at most M indices in $\mathcal{K}_{t-1}^{(k)}$ and $\mathcal{K}_{t-1}^{(k)} \supset \{t-1, \cdots, t-m\}$. When a new observation arrives at time t, we still define $\xi_{i,t}^{(k)*}$ by (2.2.7) for $i \in \{t\} \cup \mathcal{K}_{t-1}^{(k)}$ and denote i_t the index not belonging to the most recent m stages, $\{t, t-1, \ldots, t-m+1\}$, such that

$$\xi_{i_t,t}^{(k)*} = \min\{\xi_{i,t}^{(k)*} : i \in \mathcal{K}_{t-1}^{(k)} \quad \text{and} \quad i \leq t-m\}, \tag{2.5.1}$$

choosing $i_t^{(k)}$ to be the one farthest from t if the minimizing set in (2.5.1) has more than one element. Define $\mathcal{K}_t^{(k)} = \{t\} \cup (\mathcal{K}_{t-1}^{(k)} - \{i_t^{(k)}\})$, and then

$$\xi_{i,t}^{(k)} = \Big(\xi_{i,t}^{(k)*} \Big/ \sum_{j \in \mathcal{K}_t^{(k)}} \xi_{j,t}^{(k)*}\Big), \qquad i \in \mathcal{K}_t^{(k)}, \tag{2.5.2}$$

yields a BCMIX approximation to the forward filter.

Similarly, to obtain a BCMIX approximation to the backward filter (2.3.3), let $\widetilde{\mathcal{K}}_{t+1}^{(k)}$ denote the set of indices j for which $\eta_{j,t+1}^{(k)}$ in (2.3.3) is kept at stage $t+1$ for regime k; thus, $\widetilde{\mathcal{K}}_{t+1}^{(k)} \supset \{t+1, \cdots, t+m\}$. At time t, define $\eta_{j,t}^{(k)}$ by (2.3.3) for $j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}^{(k)}$ and let j_t be the index not belonging to the most recent m stages, $\{t, t+1, \cdots, t+m-1\}$ such that

$$\eta_{j_t,t}^{(k)*} = \min\{\eta_{j,t}^{(k)*} : j \in \widetilde{\mathcal{K}}_{t+1}^{(k)} \quad \text{and} \quad j \geq t+m\}, \tag{2.5.3}$$

choosing $j_t^{(k)}$ to be the one farthest from t if the minimizing set in (2.5.3) has more than one

element. Define $\widetilde{\mathcal{K}}_t^{(k)} = \{t\} \cup (\widetilde{\mathcal{K}}_{t+1}^{(k)} - \{i_t^{(k)}\})$, and then

$$\eta_{j,t}^{(k)} = \Big(\eta_{j,t}^{(k)*} \Big/ \sum_{j \in \widetilde{\mathcal{K}}_t^{(k)}} \eta_{j,t}^{(k)*}\Big), \qquad j \in \widetilde{\mathcal{K}}_t^{(k)}, \tag{2.5.4}$$

yields a BCMIX approximation to the backward filter.

For the smoothing estimate $E(\theta_{lt}|\mathcal{Y}_T)$ and its associated posterior distribution, we construct BCMIX approximations by combining the preceding forward and backward BCMIX filters with index sets $\mathcal{K}_t^{(k)}$ and $\widetilde{\mathcal{K}}_{t+1}^{(k)}$, respectively, at time t. Then the BCMIX approximations to (2.4.2) are given as

$$\widetilde{\alpha}_{ijt} = \alpha_{ijt}^* / \widetilde{D}_t, \qquad \widetilde{D}_t = \sum_{i \in \mathcal{K}_t^{(k)}, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}^{(k)}} \alpha_{ijt}^*,$$

$$\alpha_{ijt}^{(k)*} = \begin{cases} \xi_{i,t}^{(k)} \big(\sum_{r \neq k} \eta_{t+1}^{(r)} p_{kr} / \pi_r \big) & i \in \mathcal{K}_t^{(k)}, \\ p_{kk} \xi_{i,t}^{(k)} \eta_{t+1,j}^{(k)} \psi_{i,j}^{(k)} \psi_{0,0}^{(k)} \big/ (\pi_k \psi_{i,t}^{(k)} \psi_{t+1,j}^{(k)}) & i \in \mathcal{K}_t^{(k)}, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}^{(k)}. \end{cases}$$

Therefore, the BCMIX smoother for θ_{lt} and s_t given $\mathcal{Y}_T$ are expressed as

$$E(\theta_{lt}|\mathcal{Y}_T) \approx \textstyle\sum_{k=1}^K \sum_{i \in \mathcal{K}_t^{(k)}, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}^{(k)}} \widetilde{\alpha}_{ijt}^{(k)} z_{i,j}^{(l,k)}, \tag{2.5.5}$$

$$P(s_t = k|\mathcal{Y}_T) \approx \textstyle\sum_{k=1}^K \sum_{i \in \mathcal{K}_t^{(k)}, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}^{(k)}} \widetilde{\alpha}_{ijt}^{(k)}. \tag{2.5.6}$$

The BCMIX approximation fixes the number of filters as M at each time, and keeps the m closest weights and the other $M - m$ largest weights. This greatly reduces the computational complexity $O(T^2)$ of the filter in Section 2.2, 2.3 and $O(T^3)$ of the smoother in Sections 2.4 to $O(T)$. The specification of M and m are discussed in Section 3.

2.6 Hyperparameter Estimation

The inference procedure in the above sections involve the hyperparameters $\{P, z^{(l,k)}, V^{(l,k)}, \sigma_l^2\}$, $(1 \leq k \leq K,\ 1 \leq l \leq J)$. which can be replaced by their estimates in the empirical Bayes approach. We can show the conditional density function of $\boldsymbol{y}_t$ given $\mathcal{Y}_{t-1}$ is expressed as

$$f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}) = \prod_{l=1}^{J} \left(\sum_{k=1}^{K} \sum_{i=1}^{t} \xi_{i,t}^{(k)*} \right), \tag{2.6.1}$$

where $\xi_{i,t}^{(k)*}$ are given by (2.2.7) and are functions of hyperparameter vector $\Phi = \{P, z(lk), V(lk), \sigma_l^2\}$, $1 \leq k \leq K$, $1 \leq l \leq J$. Given Φ and the observed data $\mathcal{Y}_T$, the log likelihood function is

$$l(\Phi) = \sum_{t=1}^{T} \log f(\boldsymbol{y}_t|\mathcal{Y}_{t-1}) = \sum_{l=1}^{J} \left(\sum_{t=1}^{T} \log \Big\{ \sum_{k=1}^{K} \sum_{i=1}^{t} \xi_{i,t}^{(k)*} \Big\} \right), \tag{2.6.2}$$

in which $f(\cdot|\cdot)$ denotes conditional density function. Maximizing (2.6.2) over Φ yields the maximum likelihood estimate $\widehat{\Phi}$.

Since Φ is a $[(K-1)K + d(d+1)K + 1]$-dimensional vector and the functions $\xi_{i,t}^{(k)}$ have to be computed recursively for $1 \leq t \leq T$, direct maximization of (2.6.2) is computationally expensive due to the curse of dimensionality. We now use the Expectation Maximization (EM) algorithm to exploit the much simpler structure of the log likelihood $l_c(\Phi)$ of the

complete data $\{(\boldsymbol{y}_t, s_t, \theta_{lt}), 1 \le t \le T, 1 \le l \le J\}$:

$$
\begin{aligned}
l_c(\Phi|s_0) &= \sum_{l=1}^{J}\sum_{t=1}^{T} f(y_{lt}, \theta_{lt}, s_t | \{y_{li}, \theta_{li}, s_i; i = 0, \ldots, t-1, l = 0, \ldots, J\}) \\
&= \sum_{l=1}^{J}\sum_{t=1}^{T} \Big\{ \log f(y_{lt}|\theta_{lt}) + \sum_{k=1}^{K} f(\theta_{lt}|s_t = k) 1_{\{s_t = k\}} + \sum_{k,r=1}^{K} \log(p_{kr}) 1_{\{s_{t-1}=k, s_t = r\}} \Big\} \\
&= -\frac{1}{2}\sum_{l=1}^{J}\sum_{t=1}^{T} \Big\{ \frac{(y_{lt} - \theta_{lt})^2}{\sigma_l^2} + \log(2\pi\sigma_l^2) \Big\} + \sum_{l=1}^{J}\sum_{t=1}^{T}\sum_{k,r=1}^{K} \log(p_{kr}) 1_{\{s_{t-1}=k, s_t=r\}} \\
&- \frac{1}{2}\sum_{l=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} \Big\{ \frac{(\theta_{lt} - z^{(l,k)})^2}{V^{(l,k)}} + \log((2\pi V^{(l,k)}) \Big\} 1_{\{s_t = k, s_t \ne s_{t-1}\}}
\end{aligned}
\tag{2.6.3}
$$

The E-step of the EM algorithm calculates $E(l_c(\Phi)|\mathcal{Y}_T)$ which is

$$
\begin{aligned}
E(l_c(\Phi)|\mathcal{Y}_T) = &-\frac{1}{2}\sum_{l=1}^{J}\sum_{t=1}^{T} \frac{1}{\sigma_l^2} E[(y_{lt} - \theta_{lt})^2 | \mathcal{Y}_T] - \frac{T}{2}\log(2\pi\sigma_l^2) \\
&- \frac{1}{2}\sum_{l=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} E[\frac{(\theta_{lt} - z^{(l,k)})^2}{V^{(l,k)}} 1_{\{s_t=k\}} | \mathcal{Y}_T] \\
&- \frac{T}{2}\sum_{l=1}^{J}\sum_{k=1}^{K} \log((2\pi V^{(l,k)}) E(1_{\{s_t=k\}}|\mathcal{Y}_T) + J\sum_{t=1}^{T}\sum_{k,r=1}^{K} \log(p_{kr}) E(1_{\{s_{t-1}=k, s_t=r\}}|\mathcal{Y}_T).
\end{aligned}
\tag{2.6.4}
$$

It involves $E[(y_{lt} - \theta_{lt})^2|\mathcal{Y}_T]$, $E[\frac{(\theta_{lt} - z(lk))^2}{V(lk)} 1_{\{s_t=k\}}|\mathcal{Y}_T]$, and the conditional probabilities $E(1_{\{s_t=k\}}|\mathcal{Y}_T) = P(s_t = k|\mathcal{Y}_T)$ and $E(1_{\{s_{t-1}=k, s_t=r\}}|\mathcal{Y}_T) = P(s_{t-1} = k, s_t = r|\mathcal{Y}_T)$. For the first conditional probability,

$$
\begin{aligned}
P(s_t = k|\mathcal{Y}_T) &= \sum_{i=1}^{t} P(J_t^{(k)} = i|\mathcal{Y}_T) = \sum_{i=1}^{t}\sum_{j=t}^{T} P(J_t^{(k)} = i, R_t^{(k)} = j|\mathcal{Y}_T) \\
&= \sum_{1 \le i \le t \le j \le T} P(C_{ij}^{(k)}|\mathcal{Y}_T) = \sum_{1 \le i \le t \le j \le T} \alpha_{ijt}^{(k)}.
\end{aligned}
$$

For the second conditional probability,

$$P(s_{t-1}=k, s_t=r|\mathcal{Y}_T) = P(s_t=r|s_{t-1}=k,\mathcal{Y}_T)P(s_{t-1}=k|\mathcal{Y}_T). \tag{2.6.5}$$

From the above derivation, we know that

$$P(s_{t-1}=k|\mathcal{Y}_T) = \sum_{1\le i\le t-1\le j\le T} \alpha^{(k)}_{i,j,t-1}.$$

Furthermore,

$$\begin{aligned}
&P(s_t=j|s_{t-1}=i,\mathcal{Y}_T) = \frac{P(s_t=j, s_{t-1}=i,\mathcal{Y}_T)}{P(s_{t-1}=i,\mathcal{Y}_T)}\\
=&\frac{P(s_t=j, s_{t-1}=i,\mathcal{Y}_t|\mathcal{Y}_{t+1,T})}{P(s_{t-1}=i,\mathcal{Y}_t|\mathcal{Y}_{t+1,T})}\\
=&\frac{P(s_{t-1}=i,\mathcal{Y}_t|s_t=j)P(s_t=j|\mathcal{Y}_{t+1,T})}{P(s_{t-1}=i,\mathcal{Y}_t|\mathcal{Y}_{t+1,T})}\\
=&\frac{P(s_{t-1}=i,\mathcal{Y}_t)P(s_t=j|s_{t-1}=i,\mathcal{Y}_t)}{P(s_t=j)}\frac{P(s_t=j|\mathcal{Y}_{t+1,T})}{P(s_{t-1}=i,\mathcal{Y}_t|\mathcal{Y}_{t+1,T})}\\
=&\frac{P(s_{t-1}=i,\mathcal{Y}_t)}{P(s_{t-1}=i,\mathcal{Y}_t|\mathcal{Y}_{t+1,T})}\frac{P(s_t=j|s_{t-1}=i,\boldsymbol{y}_t)P(s_t=j|\mathcal{Y}_{t+1,T})}{P(s_t=j)}\\
\propto&\frac{P(s_t=j,\boldsymbol{y}_t|s_{t-1}=i)P(s_t=j|\mathcal{Y}_{t+1,T})}{P(s_t=j)}\\
=&\frac{f(\boldsymbol{y}_t|s_t=j,s_{t-1}=i)P(s_t=j|s_{t-1}=i)\sum_{k=1}^K P(s_t=j,s_{t+1}=k|\mathcal{Y}_{t+1,T})}{P(s_t=j)}\\
=&\frac{f(\boldsymbol{y}_t|s_t=j,s_{t-1}=i)P(s_t=j|s_{t-1}=i)\sum_{k=1}^K P(s_t=j|s_{t+1}=k,\mathcal{Y}_{t+1,T})P(s_{t+1}=k|\mathcal{Y}_{t+1,T})}{P(s_t=j)}\\
=&\frac{f(\boldsymbol{y}_t|s_t=j)P(s_t=j|s_{t-1}=i)\sum_{k=1}^K P(s_t=j|s_{t+1}=k)P(s_{t+1}=k|\mathcal{Y}_{t+1,T})}{P(s_t=j)}\\
=&\frac{\psi^{(j)}_{0,0}/\psi^{(j)}_{t,t}p_{ij}\sum_{k=1}^K \widetilde{p}_{kj}\eta^k_{t+1}}{\pi_j}.
\end{aligned}$$

Thus

$$P(s_t = r|s_{t-1} = k, \mathcal{Y}_T) = \frac{\psi_{0,0}^{(r)}/\psi_{t,t}^{(r)} p_{kr} \widetilde{P}_r' \eta_{t+1}/\pi_r}{\sum_{i=1}^{K} \left[\psi_{0,0}^{(i)}/\psi_{t,t}^{(i)} p_{ki} \widetilde{P}_i' \eta_{t+1}/\pi_i\right]}. \tag{2.6.6}$$

Plugging (2.6.6) into (2.6.5), we have

$$P(s_t = r, s_{t-1} = k|\mathcal{Y}_T) = \frac{\psi_{0,0}^{(r)}/\psi_{t,t}^{(r)} p_{kr} \widetilde{P}_r' \eta_{t+1}/\pi_r}{\sum_{i=1}^{K} \left[\psi_{0,0}^{(i)}/\psi_{t,t}^{(i)} p_{ki} \widetilde{P}_i' \eta_{t+1}/\pi_i\right]} \sum_{1\le i\le t-1\le j\le T} \alpha_{i,j,t-1}^{(k)}.$$

Then the conditional probabilities are:

$$\begin{aligned} E(1_{\{s_t=k\}}|\mathcal{Y}_T) =& \sum_{1\le i\le t\le j\le T} \alpha_{ijt}^{(k)}, \\ E(1_{\{s_{t-1}=k, s_t=r\}}|\mathcal{Y}_T) =& \frac{\psi_{0,0}^{(r)}/\psi_{t,t}^{(r)} p_{kr} \widetilde{P}_r' \eta_{t+1}/\pi_r}{\sum_{i=1}^{K} \left[\psi_{0,0}^{(i)}/\psi_{t,t}^{(i)} p_{ki} \widetilde{P}_i' \eta_{t+1}/\pi_i\right]} \sum_{1\le i\le t-1\le j\le T} \alpha_{i,j,t-1}^{(k)}. \end{aligned} \tag{2.6.7}$$

The M-step of the EM algorithm involves calculating the partial derivatives of (2.6.4) with respect to Φ. Simple algebra yields the following updating formulas for Φ.

$$\begin{aligned} \widehat{p}_{kr,\text{new}} &= \frac{\sum_{t=2}^{T} P(s_{t-1} = k, s_t = r|\mathcal{Y}_T, \widehat{\Phi}_{\text{old}})}{\sum_{t=2}^{T} P(s_{t-1} = k|\mathcal{Y}_T, \widehat{\Phi}_{\text{old}})}, \\ \widehat{z}(lk)_{\text{new}} &= \frac{\sum_{t=1}^{T} E(\theta_{lt} 1_{\{s_t=k\}}|\mathcal{Y}_T, \widehat{\Phi}_{\text{old}})}{\sum_{t=1}^{T} P(s_t = k|\mathcal{Y}_T, \widehat{\Phi}_{\text{old}})}, \\ \widehat{V}(lk)_{\text{new}} &= \frac{\sum_{t=1}^{T} E[(\theta_{lt} - \widehat{z}(lk)_{\text{old}})^2 1_{\{s_t=k\}}|\mathcal{Y}_T, \widehat{\Phi}_{\text{old}}]}{\sum_{t=1}^{T} P(s_t = k|\mathcal{Y}_T, \widehat{\Phi}_{\text{old}})}, \\ \widehat{\sigma}_{l,\text{new}}^2 &= \frac{1}{T} \sum_{t=1}^{T} E[(y_{lt} - \theta_{lt})^2|\mathcal{Y}_T, \widehat{\Phi}_{\text{old}}]. \end{aligned} \tag{2.6.8}$$

In (2.6.8), $P(s_t = k|\mathcal{Y}_T)$ can be computed by (2.4.4), and other items are given as follows,

$$P(s_{t-1} = k, s_t = r|\mathcal{Y}_T) = \frac{\psi_{0,0}^{(r)}/\psi_{t,t}^{(r)} p_{kr} \widetilde{P}_r' \eta_{t+1}/\pi_r}{\sum_{i=1}^{K} \left[\psi_{0,0}^{(i)}/\psi_{t,t}^{(i)} p_{ki} \widetilde{P}_i' \eta_{t+1}/\pi_i\right]} \sum_{1\le i\le t-1\le j\le T} \alpha_{i,j,t-1}^{(k)}.$$

$$E(\theta_{lt}1_{\{s_t=k\}}|\mathcal{Y}_T) = \sum_{1\le i\le t\le j\le T} E(\theta_{lt}|P(J_t^{(k)}=i, R_t^{(k)}=j, \mathcal{Y}_T)P(J_t^{(k)}=i, R_t^{(k)}=j|\mathcal{Y}_T),$$

in which $P(J_t^{(k)}=i, R_t^{(k)}=j|\mathcal{Y}_T) = P(C_{ij}^{(k)}|\mathcal{Y}_T) = \alpha_{ijt}^{(k)}$. Given $C_{ij}^{(k)}$ and $\mathcal{Y}_T$, the conditional density of θ_{lt} is $g_{i,j}^{(k)}(\theta_{lt})$, which is a normal distribution as given (2.2.2) with mean of $z_{i,j}^{(l,k)}$ and variance of $V_{i,j}^{(l,k)}$. Thus

$$E(\theta_{lt}1_{\{s_t=k\}}|\mathcal{Y}_T) = \sum_{1\le i\le t\le j\le T} \alpha_{ijt}^{(k)} z_{i,j}^{(l,k)},$$

Similarly, for the second posterior expectation,

$$\begin{aligned}
&E[(\theta_{lt}-z^{(l,k)})^2 1_{\{s_t=k\}}|\mathcal{Y}_T] \\
=&\sum_{1\le i\le t\le j\le T} E[(\theta_{lt}-z^{(l,k)})^2|C_{ij}^{(k)}, \mathcal{Y}_T]P(C_{ij}^{(k)}|\mathcal{Y}_T) \\
=&\sum_{1\le i\le t\le j\le T} \alpha_{ijt}^{(k)} E[\theta_{lt}^2 - 2\theta_{lt}z^{(l,k)} + (z^{(l,k)})^2|C_{ij}^{(k)}, \mathcal{Y}_T] \\
2=&\sum_{1\le i\le t\le j\le T} \alpha_{ijt}^{(k)} \Big\{ E(\theta_{lt}^2|C_{ij}^{(k)}, \mathcal{Y}_T) - 2z^{(l,k)}E(\theta_{lt}|C_{ij}^{(k)}, \mathcal{Y}_T) + (z^{(l,k)})^2 \Big\}
\end{aligned}$$

Here $E(\theta_{lt}|C_{ij}^{(k)}, \mathcal{Y}_T) = z_{i,j}^{(l,k)}$ and

$$\begin{aligned}
&E(\theta_{lt}^2|C_{ij}^{(k)}, \mathcal{Y}_T) = var(\theta_{lt}|C_{ij}^{(k)}, \mathcal{Y}_T) + (E(\theta_{lt}|C_{ij}^{(k)}, \mathcal{Y}_T))^2 \\
=&\Big(V_{i,j}^{(l,k)} + (z_{i,j}^{(l,k)})^2\Big).
\end{aligned}$$

So

$$\begin{aligned}
&E\big[(\theta_{lt}-z^{(l,k)})^2 1_{\{s_t=k\}}|\mathcal{Y}_T\big] \\
=&\sum_{1\le i\le t\le j\le T} \alpha_{ijt}^{(k)} \Big(z_{i,j}^{(l,k)^2} + V_{i,j}^{(l,k)} - 2z^{(l,k)}\cdot z_{i,j}^{(l,k)} + (z^{(l,k)})^2 \Big)
\end{aligned}$$

For the last posterior expectation, according to (2.4.1) and the above proof,

$$\begin{aligned}
&E[(y_{lt}-\theta_{lt})^2|\mathcal{Y}_T]\\
=&\sum_{k=1}^{K}\sum_{1\le i\le t\le j\le T}\alpha_{ijt}^{(k)}E[(y_{lt}-\theta_{lt})^2|C_{ij}^{(k)},\mathcal{Y}_T]\\
=&\sum_{k=1}^{K}\sum_{1\le i\le t\le j\le T}\alpha_{ijt}^{(k)}\Big\{y_{lt}^2-2E(\theta_{lt}|C_{ij}^{(k)},\mathcal{Y}_T)y_{lt}+E(\theta_{lt}^2|C_{ij}^{(k)},\mathcal{Y}_T)\Big\}\\
=&\sum_{k=1}^{K}\sum_{1\le i\le t\le j\le T}\alpha_{ijt}^{(k)}\Big\{y_{lt}^2-2z_{i,j}^{(l,k)}y_{lt}+V_{i,j}^{(l,k)}+(z_{i,j}^{(l,k)})^2\Big\}
\end{aligned}$$

In summary, the posterior expectations necessary for the updating formulas can be calculated as

$$P(s_{t-1}=k,s_t=r|\mathcal{Y}_T)=\frac{\psi_{0,0}^{(r)}/\psi_{t,t}^{(r)}p_{kr}\widetilde{P}_r'\eta_{t+1}/\pi_r}{\sum_{i=1}^{K}\Big[\psi_{0,0}^{(i)}/\psi_{t,t}^{(i)}p_{ki}\widetilde{P}_i'\eta_{t+1}/\pi_i\Big]}\sum_{1\le i\le t-1\le j\le T}\alpha_{i,j,t-1}^{(k)}. \tag{2.6.9}$$

$$E(\theta_{lt}1_{\{s_t=k\}}|\mathcal{Y}_T)=\sum_{1\le i\le t\le j\le T}\alpha_{ijt}^{(k)}z_{i,j}^{(k)}, \tag{2.6.10}$$

$$\begin{aligned}
&E\big[(\theta_{lt}-z(lk)^2 1_{\{s_t=k\}}|\mathcal{Y}_T\big]\\
=&\sum_{1\le i\le t\le j\le T}\alpha_{ijt}^{(k)}\Big(z_{i,j}^{(l,k)^2}+V_{i,j}^{(l,k)}-2z^{(l,k)}\cdot z_{i,j}^{(l,k)}+(z^{(l,k)})^2\Big)
\end{aligned} \tag{2.6.11}$$

$$E[(y_{lt}-\theta_{lt})^2|\mathcal{Y}_T]=\sum_{k=1}^{K}\sum_{1\le i\le t\le j\le T}\alpha_{ijt}^{(k)}\Big\{y_{lt}^2-2z_{i,j}^{(l,k)}y_{lt}+V_{i,j}^{(l,k)}+(z_{i,j}^{(l,k)})^2\Big\} \tag{2.6.12}$$

The iteration scheme (2.6.8) is carried out until convergence or until some prescribed upper bound on the number of iterations is reached.

To speed up the computations involved in the EM algorithm, one can use the BCMIX approximations in Section 2.5 instead of the full recursions to determine the items (2.6.9)-(2.6.12). Our simulation studies shows that the EM procedure converge very fast.

2.6.1 Implementation

We have shown the posterior distribution of parameter θ_{lt} is mixture of distributions. In this section, we describe in detail how to implement the algorithms, presenting explicit formulas. Let us start with a description of Bayes algorithm.

Step 1 Calculating $V_{i,j}^{(l,k)}$ and $z_{i,j}^{(l,k)}$. Similar to (2.2.2), given $\mathcal{Y}_T$ and $C_{ij}^{(k)}, i < j$ we use

$$V_{i,j}^{(l,k)} = \big([V^{(l,k)}]^{-1} + \frac{j-i+1}{\sigma_l^2}\big)^{-1},$$
$$z_{i,j}^{(l,k)} = V_{i,j}^{(l,k)}\big([V^{(l,k)}]^{-1} z^{(l,k)} + \frac{\sum_{u=i}^{j} y_{lu}}{\sigma_l^2}\big).$$

The results can be saved in two four-dimensional matrices for future calculation. More specifically, $g_{i,j}^{(l,k)}(\theta_t)$ is calculated by (2.2.2). If there is no information other than $s_t = k$ is given, the conditional distribution is $g_{0,0}^{(l,k)}(\theta_t)$ as in (2.2.3). Using $V_{i,j}^{(l,k)}$, $z_{i,j}^{(l,k)}$, we can also calculate the conditional densities $\psi_{0,0}^{(k)}$ and $\psi_{i,j}^{(k)}$. They are also used to calculate the smoothing estimate of θ_{lt} by (2.4.3).

Step 2 Calculating the forward filter (2.2.7) in a recursive manner.

(A) Start with $t = 1$. According to (2.2.7), we have

$$\xi_{1,1}^{(k)} \propto \xi_{1,1}^{(k)*} = \big(\sum_{r \neq k} \xi_0^{(r)} p_{rk}\big)\psi_{0,0}^{(k)}/\psi_{1,1}^{(k)}.$$

Substitute $\xi_0^{(r)}$ for $r \neq k$ by the stationary distribution π_r, use $\psi_{0,0}^{(k)}$ and $\psi_{1,1}^{(k)}$ to calculate $\big(\sum_{r \neq k} \pi_r p_{rk}\big)\psi_{0,0}^{(k)}/\psi_{1,1}^{(k)}$, which gives the value of $\xi_{1,1}^{(k)*}$, and therefore $\xi_{1,1}^{(k)} = \frac{\xi_{1,1}^{(k)*}}{\sum_{k=1}^{K} \xi_{1,1}^{(k)*}}$.

(B) At $t > 1$, calculate $\xi_{t,t}^{(k)*} = \big(\sum_{r \neq k} \xi_{t-1}^{(r)} p_{rk}\big)\psi_{0,0}^{(k)}/\psi_{t,t}^{(k)}$ directly. Use $\xi_{i,t-1}^{(k)}$ to calculate $\xi_{i,t}^{(k)*} = p_{kk}\xi_{i,t-1}^{(k)}\psi_{i,t-1}^{(k)}/\psi_{i,t}^{(k)}$ for $i < t$. Normalize $\xi_{i,t}^{(k)*}$ by dividing $\sum_{1 \leq i \leq t} \xi_{i,t}^{(k)*}$ to get $\xi_{i,t}^{(k)}$. Keep doing (B) until $t = T$.

Step 3 Calculating the backward filter (2.3.3) in a recursive manner. The backward filter $\eta_{j,t+1}^{(k)}$ can be calculated similarly by starting with $t = T$.

Step 4 Calculating the smoothing mixture weight (2.4.2) and the smoothing estimate (2.4.3).

One main challenge for this procedure is the computational complexity due to the space needed to save the matrices and number of weights which is increasing with t. There are two ways to increase the computational efficiency of this procedure.

The first modification is to implement the BCMIX approximation so that number of weights will be a fixed number M. The cost associated with the method is to keep the index set $\mathcal{K}_t^{(k)}$ for forward filter $\xi_{i,t}^{(k)}$ and $\widetilde{\mathcal{K}}_{t+1}^{(k)}$ for backward filter $\eta_{j,t+1}^{(k)}$. The basic procedure is similar to the preceding one with calculation of up to $M + 1$ weights for each stage t. The detailed procedure is as follows.

Step 1 Calculating $V_{i,j}^{(l,k)}$ and $z_{i,j}^{(l,k)}$.

Step 2 Calculating the BCMIX forward filter (2.5.2) in a recursive manner.

(A) For $1 \leq t \leq M$, use the Bayes procedure to calculate $\xi_{i,t}^{(k)*}$, $\xi_{i,t}^{(k)}$. The index set $\mathcal{K}_t^{(k)}$ at stage t is $\{1, \cdots, t\}$.

(B) At $t > M$, use new information at stage t to calculate $\psi_{t,t}^{(k)}$ and therefore $\xi_{t,t}^{(k)*} = \big(\sum_{r \neq k} \xi_{t-1}^{(r)} p_{rk}\big)\psi_{0,0}^{(k)}/\psi_{t,t}^{(k)}$. Use $\xi_{i,t-1}^{(k)}$ to calculate $\xi_{i,t}^{(k)*} = p_{kk}\xi_{i,t-1}^{(k)}\psi_{i,t-1}^{(k)}/f_{i,t-1}^{(k)}$ for $i \in \mathcal{K}_{t-1}^{(k)}$. Compare the weights in $\mathcal{K}_{t-1}^{(k)} - \{i_t^{(k)}\}$ and drop the smallest one. The remaining M weights form the new index set $\mathcal{K}_t^{(k)}$, and $\xi_{i,t}^{(k)} = \frac{\xi_{i,t}^{(k)*}}{\sum_{j \in \mathcal{K}_t^{(k)}} \xi_{j,t}^{(k)*}}$. Keep doing (B) until $t = T$, saving both the index sets and the BCMIX forward filters for future calculation.

Step 3 Calculating the BCMIX backward filter (2.5.4) in a recursive manner starting with $t = T$.

Step 4 Calculating the BCMIX smoothing mixture weight (2.4.2) and the smoothing

estimate (2.4.3).

When one takes a second look at the BCMIX procedure, it is easy to find that only a small portion of the huge precalculated matrices $z_{i,j}^{(l,k)}$ and $V_{i,j}^{(l,k)}$ have been used. So it wastes a lot of space and time to calculate all the elements. However, we do not know which elements to use before calculating the index sets. A better idea is to calculate $z_{i,j}^{(l,k)}$ and $V_{i,j}^{(l,k)}$ when we need them. One more challenge is that the formulas to calculate $z_{i,j}^{(l,k)}$ and $V_{i,j}^{(l,k)}$ involve matrix inversion, which will take a long time to implement. Instead of calculating $z_{i,j}^{(l,k)}$ and $V_{i,j}^{(l,k)}$ directly, we can calculate $VI_{i,j}^{(l,k)} := (V_{i,j}^{(l,k)})^{-1}$ and $VIZ_{i,j}^{(l,k)} := (V_{i,j}^{(l,k)})^{-1} z_{i,j}^{(l,k)}$ by the following simple recursive formulas if we know $VI_{i,j-1}^{(l,k)}$ and $VIZ_{i,j-1}^{(l,k)}$

$$\begin{aligned} VI_{i,j}^{(l,k)} &= [V^{(l,k)}]^{-1} + \frac{j-i+1}{\sigma_l^2} = VI_{i,j-1}^{(l,k)} + \frac{1}{\sigma_l^2}, \\ VIZ_{i,j}^{(l,k)} &= [V^{(l,k)}]^{-1} z^{(l,k)} + \frac{\sum_{t=i}^{j} y_{lt}}{\sigma_l^2} = VIZ_{i,j-1}^{(l,k)} + \frac{\sum_{t=i}^{j} y_{lt}}{\sigma_l^2}. \end{aligned} \tag{2.6.13}$$

So the BCMIX algorithm can be further simplified by adding this recursive updating feature. The detailed procedure is as follows.

Step 1 Calculating the BCMIX forward filter (2.5.2) in a recursive manner from $t = 1$. Follow Step 2 in the above BCMIX algorithm. Assume at stage $t-1$ we have finished calculating $\xi_{i,t-1}^{(k)}$ and $\mathcal{K}_{t-1}^{(k)}$, and saved all the $VI_{i,t-1}^{(l,k)}$ and $VIZ_{i,t-1}^{(l,k)}$ for $i \in \mathcal{K}_{t-1}^{(k)}$. At stage t, $VI_{i,t}^{(l,k)}$ and $VIZ_{i,t}^{(l,k)}$ for $i \in \mathcal{K}_{t-1}^{(k)}$ can be calculated by (2.6.13). $VI_{t,t}^{(l,k)} = [V^{(l,k)}]^{-1} + \frac{1}{\sigma_l^2}$, and $VIZ_{t,t}^{(l,k)} = [V^{(l,k)}]^{-1} z^{(l,k)} + \frac{y_{lt}}{\sigma_l^2}$. They are used to calculate $\psi_{i,t}^{(k)}$, $\psi_{t,t}^{(k)}$ by

$$\begin{aligned} \psi_{i,t}^{(k)} = \prod_{l=1}^{J} g_{i,t}^{(l,k)} &\propto \prod_{l=1}^{J} (V_{i,t}^{(l,k)})^{-1/2} \exp\{-\frac{(z_{i,t}^{(l,k)})^2}{2V_{i,t}^{(l,k)}}\} \\ &\propto \prod_{l=1}^{J} (VI_{i,t}^{(l,k)})^{1/2} \exp\{-\frac{(VIZ_{i,t}^{(l,k)})^2}{2VI_{i,t}^{(l,k)}}\}, \end{aligned}$$

and therefore $\xi_{i,t}^{(k)*}$ are calculated for all $i \in \{t\} \cup \mathcal{K}_{t-1}^{(k)}$. A small weight is dropped by the BCMIX rule and the remaining index set $\mathcal{K}_t^{(k)}$, $\xi_{i,t}^{(k)}$, $VI_{i,t}^{(l,k)}$ and $VIZ_{i,t}^{(l,k)}$ are saved.

Step 2 Calculating the BCMIX backward filter (2.5.4) in a recursive manner starting with $t = T$. If we know $VI_{i-1,j}^{(l,k)}$ and $VIZ_{i-1,j}^{(l,k)}$, and want to calculate $VI_{i,j}^{(l,k)}$ and $VIZ_{i,j}^{(l,k)}$ by the recursive formulas

$$VI_{i,j}^{(l,k)} = VI_{i-1,j}^{(l,k)} + \frac{1}{\sigma_l^2}, \quad VIZ_{i,j}^{(l,k)} \ = VIZ_{i-1,j}^{(l,k)} + \frac{\sum_{t=i}^{j} y_{lt}}{\sigma_l^2}.$$

Using these updating formulas, we can recursively calculate $VI_{t+1,j}^{(l,k)}$ and $VIZ_{t+1,j}^{(l,k)}$ for $j \in \widetilde{\mathcal{K}}_{t+1}^{(k)}$ and conduct Step 3 in the above BCMIX algorithm.

Step 3 Calculating the BCMIX smoothing mixture weight $\widetilde{\alpha}_{ijt}^{(k)}$ and the smoothing estimate $\hat{\theta}_{l,t|T}$. We can evaluate $VI_{i,j}^{(l,k)}$ and $VIZ_{i,j}^{(l,k)}$ for $i \in \mathcal{K}_t^{(k)}, j \in \widetilde{\mathcal{K}}_{t+1}^{(k)}$ by

$$\begin{aligned}
VI_{i,j}^{(l,k)} &= [V^{(l,k)}]^{-1} + \frac{j-i+1}{\sigma_l^2} \\
&= ([V^{(l,k)}]^{-1} + \frac{t-i+1}{\sigma_l^2}) + ([V^{(l,k)}]^{-1} + \frac{j-t}{\sigma_l^2}) - [V^{(l,k)}]^{-1} \\
&= VI_{i,t}^{(l,k)} + VI_{t+1,j}^{(l,k)} - [V^{(l,k)}]^{-1}, \\
VIZ_{i,j}^{(l,k)} &= [V^{(l,k)}]^{-1} z^{(l,k)} + \frac{\sum_{u=i}^{j} y_{lu}}{\sigma_l^2} \\
&= ([V^{(l,k)}]^{-1} z^{(l,k)} + \frac{\sum_{u=i}^{t} y_{lu}}{\sigma_l^2}) + ([V^{(l,k)}]^{-1} z^{(l,k)} + \frac{\sum_{u=t+1}^{j} y_{lu}}{\sigma_l^2}) - [V^{(l,k)}]^{-1} z^{(l,k)} \\
&= VIZ_{i,t}^{(l,k)} + VIZ_{t+1,j}^{(l,k)} - [V^{(l,k)}]^{-1} z^{(l,k)}.
\end{aligned} \tag{2.6.14}$$

The smoothing estimate of θ_{lt} can be calculated as $\hat{\theta}_{l,t|T}$ defined in (2.5.5). Furthermore, the inference on regimes can be conducted using (2.5.6) by substituting α_{ijt} by $\widetilde{\alpha}_{ijt}$ calculated in Step 3.

Chapter 3

Simulation Studies

In this chapter, implementation of intensive simulation experiments is described. Firstly, some general criterion are introduced, including sum of squared error (SSE),and the identification ratio (IR) of true state calling. Then the performances of the fBayes and BCMIX estimates are compared, through Monte Carlo simulations. The result shows the BCMIX is statistically and computational efficient. Afterwards, we examine the relationship between the BCMIX performance and simulation settings. Additionally, we compare our model to an existing hierarchical HMM model. The choice of hyperparameters is also discussed.

3.1 Comparison Criterion

The sum of squared errors between the true and estimated parameters is used to assess the performance of the estimation of parameter θ_{lt}. In our model, if a time series of T observations is generated, with a series of $\hat{\theta}_{lt}$ estimated, the SSE is defined by

$$SSE = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J}\sum_{l=1}^{J}(\theta_{lt} - \hat{\theta}_{lt})^2$$

We also need to evaluate the performance of the smoothed probability $\hat{r}_{t|T}^{(k)} = P(s_t =$

$k|\mathcal{Y}_t)$ as discussed in previous chapter. We use this probability to provide assessment of the hidden state s_t belonging to regime k. However, this is not a logical variable only taking a value of 1 or 0, but a probability theoretically close to 1 or 0. When there is a transition from some regime to another one, the probability might show some fuzziness. An intuitive and simple way to make the inference on s_t is to compare the smoothed probability $\hat{r}_{t|T}^{(k)}$ with 0.5. If for any $1 \leq k \leq K$, $\hat{r}_{t|T}^{(k)} > 0.5$, we identity $s_t = k$. More specifically, to evaluate the performance of this procedure, we define an identification ratio as

$$IR := \frac{\sum_{t=1}^{T} \sum_{k=1}^{K} \mathbf{1}_{(\hat{r}_{t|T}^{(k)} > 0.5) \cap (s_t = k)}}{T},$$

where $\mathbf{1}$ denotes an indicator function, and T is the length of the sequence. If the true regime is k, and a probability reasonably close to 1, $\hat{r}_{t|T}^{(k)} > 0.5$, is obtained from the procedure, then $(\hat{r}_{t|T}^{(k)} > 0.5) \cap (s_t = k)$ is true, and the indicator function returns 1 for stage t.

3.2 Simulation 1: Comparison between Bayes and BCMIX Estimates

As mentioned in Section 2.5, the Bayes method is accurate but computationally inefficient since the number of weights increases with t, resulting in rapidly increasing computational complexity and memory requirement in estimating θ_{lt} as t keeps increasing. The BCMIX approximation is much faster and does not need to save so many variables. This section is to compare the performances of the fBayes method described later and the BCMIX approximation described in Section 2.5.

We use three states in our model, $K = 3$, and the values of the parameter θ_{lt} depend on the hidden state s_t. The hyperparameters consist of $\{P, z^{(l,k)}, V^{(l,k)}, \sigma_l^2\}$, $1 \leq k \leq K$, $1 \leq l \leq J$. In all the examples shown in this section, data are generated according to

hyperparameter values: $(z^{(l,1)}, V^{(l,1)}) = (1, 0.22)$, $(z^{(l,2)}, V^{(l,2)}) = (0, 0.22)$, $(z^{(l,3)}, V^{(l,3)}) = (-1, 0.22)$ $P = \begin{pmatrix} 0.998 & 0.001 & 0.001 \\ 0.001 & 0.998 & 0.001 \\ 0.001 & 0.001 & 0.998 \end{pmatrix}$, and $\sigma_l^2 = 1$, $1 \leq l \leq 10$. Furthermore, given s_t, θ_{lt} is a realization from a truncated Normal$(z^{(l,s_t)}, V^{(l,s_t)})$ distribution such that $|\theta_{lt}| < 2$ to make the series stationary. We generate $N = 500$ series, each of length $T = 1000$, and consider s_t changing over time in four scenarios:

Scenario 1. There is one transition from regime 1 to regime 2, one transition from regime 2 to regime 3. $s_t = 1$ for $1 \leq t \leq 200$; $s_t = 2$ for $201 \leq t \leq 400$; $s_t = 3$ for $401 \leq t \leq 1000$.

Scenario 2. There is one transition from regime 1 to regime 2, one transition from regime 2 to regime 3. $s_t = 1$ for $1 \leq t \leq 500$; $s_t = 2$ for $501 \leq t \leq 700$; $s_t = 3$ for $701 \leq t \leq 1000$.

Scenario 3. There are four transitions among regime 1, regime 2 and regime 3. $s_t = 2$ for $1 \leq t \leq 200$; $s_t = 1$ for $201 \leq t \leq 400$; $s_t = 2$ for $401 \leq t \leq 600$, $s_t = 3$ for $601 \leq t \leq 800$, $s_t = 2$ for $801 \leq t \leq 1000$.

Scenario 4. There are six transitions among regime 1, regime 2 and regime 3. $s_t = 2$ for $1 \leq t \leq 200$; $s_t = 3$ for $201 \leq t \leq 300$; $s_t = 1$ for $301 \leq t \leq 400$, $s_t = 2$ for $401 \leq t \leq 600$, $s_t = 3$ for $601 \leq t \leq 700$, $s_t = 1$ for $701 \leq t \leq 800$,$s_t = 2$ for $801 \leq t \leq 1000$.

In each scenario, we assume the true hyperparameters are given, and compute both the BCMIX estimate. As mentioned in Section 2.5, the performance of the BCMIX procedure depends on the specification of M and m. This dependence is examined here, choosing $M = 2m$ and M =10, 20, 30 and 40. Furthermore, to access the performance of the method, we consider a simple benchmark in which the hidden state is known so that the Bayes estimates of θ_{lt} among three transitions are given by the standard Bayesian formulas for normal populations (Section 2.7 of Box and Tiao (1973)). This is called a "fictitious Bayes" estimate. Tables 3.1 compare fictitious Bayes estimate (fBayes) and the BCMIX estimate

Table 3.1: Performance of sum of squared errors (SSE) for fBayes and BCMIX estimates. Standard errors are given in parentheses below the estimates.

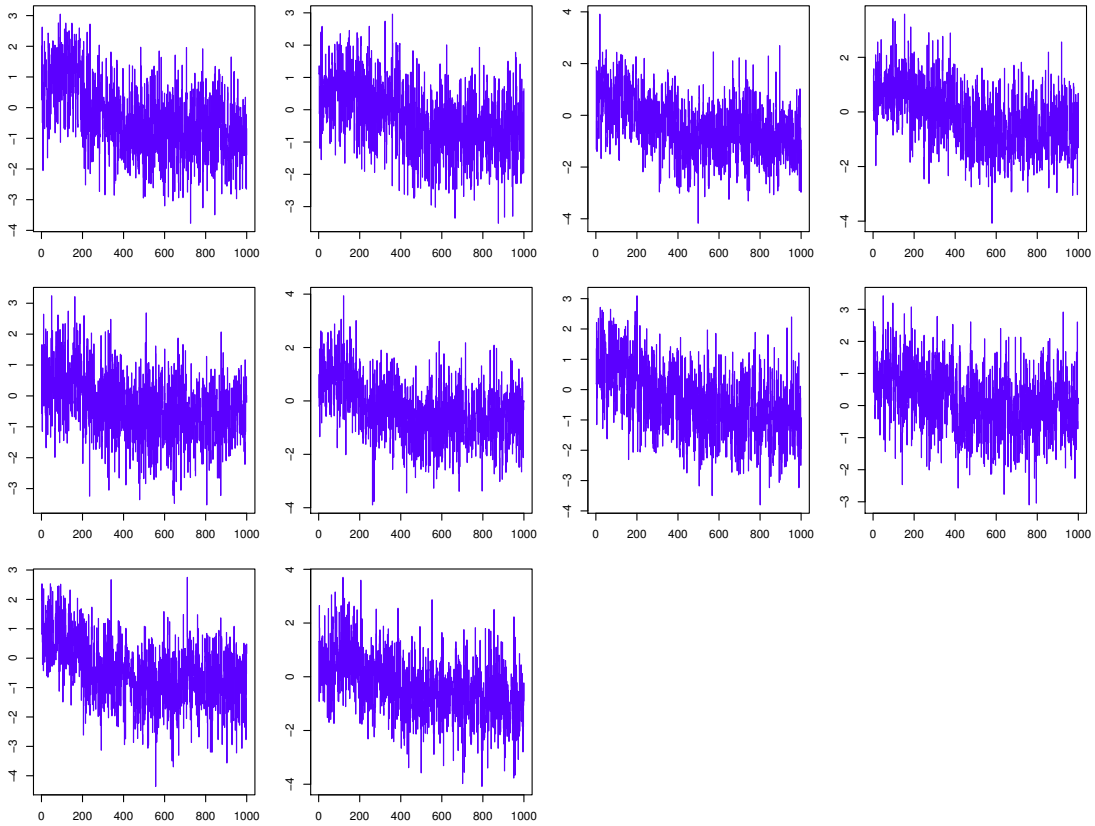
Scenarios	*fBayes*	*BCMIX* (10,5)	(20,10)	(30,15)	(40,20)
Scenario 1	0.00287	0.00320	0.00319	0.00319	0.00319
	(3.26E-05)	(3.77E-05)	(3.77E-05)	(3.77E-05)	(3.77E-05)
Scenario 2	0.00294	0.00329	0.00328	0.00328	0.00328
	(3.35E-05)	(3.93E-05)	(3.91E-05)	(3.91E-05)	(3.91E-05)
Scenario 3	0.00487	0.00540	0.00540	0.00540	0.00540
	(4.34E-05)	(4.75E-05)	(4.75E-05)	(4.75E-05)	(4.75E-05)
Scenario 4	0.00676	0.00758	0.00758	0.00758	0.00758
	(5.18E-05)	(6.17E-05)	(6.13E-05)	(6.12E-05)	(6.12E-05)

(BCMIX) in terms of the SSE.

The first two columns in Table 3.1 show that the fictitious Bayes estimate show smaller SSE than BCMIX(10,5) estimate. As discussed in Section 3.1, SSE is not an appropriate criterion for evaluating the performance of different estimation procedures. But this comparison illustrates the effectiveness of BCMIX estimate. Furthermore, the relative difference between BCMIX(10,5) estimate and fBayes estimate in SSE is less than 2% in all scenarios, which demonstrates BCMIX has very promising results.The last four columns in Table 3.1 show that the average SSE over 500 sequences changes with respect to the different values of M and m. As mentioned in Section 2.5, the approximation should improve as M and m become larger since more filters are kept at each stage. However, based on Table 3.1, we cannot see the trend clearly although in each scenario all the BCMIX estimates have similar SSE. This observation shows the BCMIX procedure is very robust for this model. The estimation results do not change dramatically when M and m are getting larger.

Let us take a second look at Table 3.1 to compare the results of different scenarios. Scenarios 1 and 2 both experience two transitions, but at different times. Both fBayes and

Figure 3.1: A selected series y_{lt} for 10 samples in Scenarios 1 (from left to right and top to bottom).



BCMIX estimates have similar SSE in Scenario 1 and 2. The estimating errors become larger in Scenarios 3 and 4 when there are more transitions and the distances between two successive transitions become smaller. The result is consistent with the prediction that the more change points, the larger SSE.

To visualize the simulation results, here we show some figures. Figure 3.1, 3.2, 3.3, 3.4 show a randomly selected simulation path y_{lt} in four scenarios, respectively. From the four figures we find some changing patterns in each series, but cannot tell the number and locations of the transitions by observing the series. Figure 3.5, 3.6, 3.7, 3.8, show the true θ_{lt} and estimated $\hat{\theta}_{lt|T}$ of the corresponding series in four scenarios, respectively. Before we

Figure 3.2: A selected series y_{lt} for 10 samples in Scenarios 2 (from left to right and top to bottom).

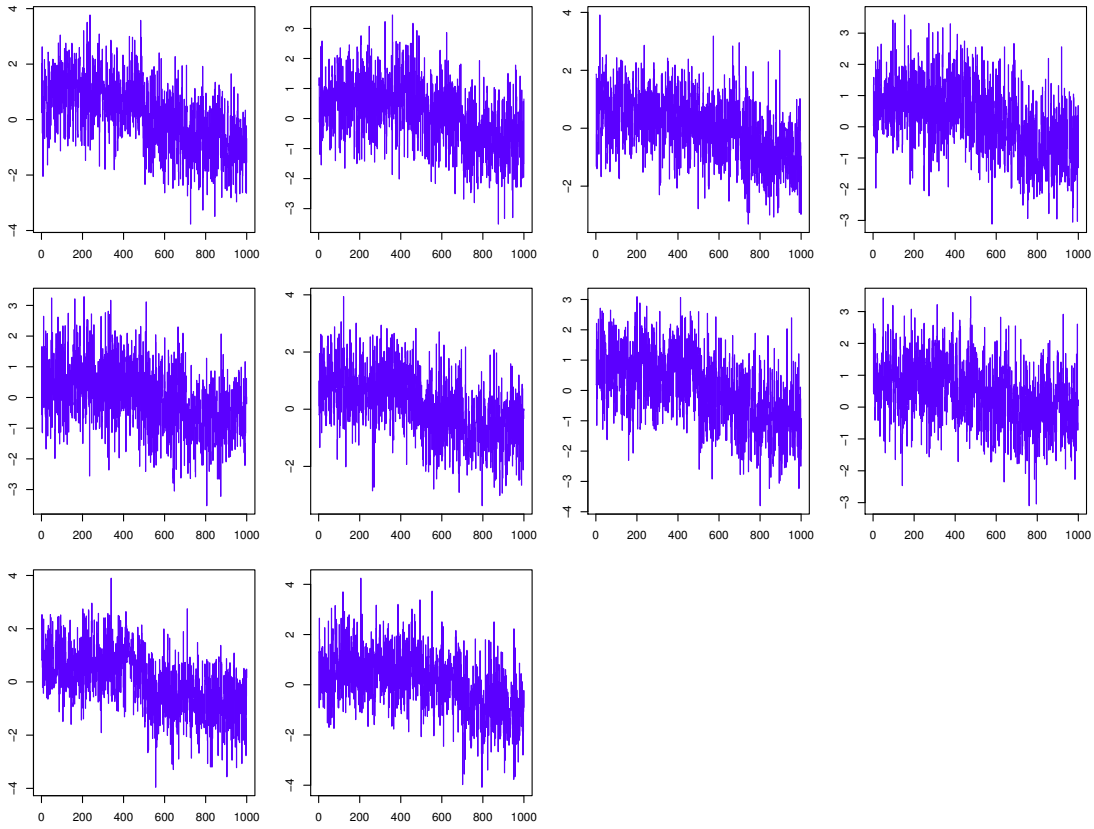


Figure 3.3: A selected series y_{lt} for 10 samples in Scenarios 3 (from left to right and top to bottom).

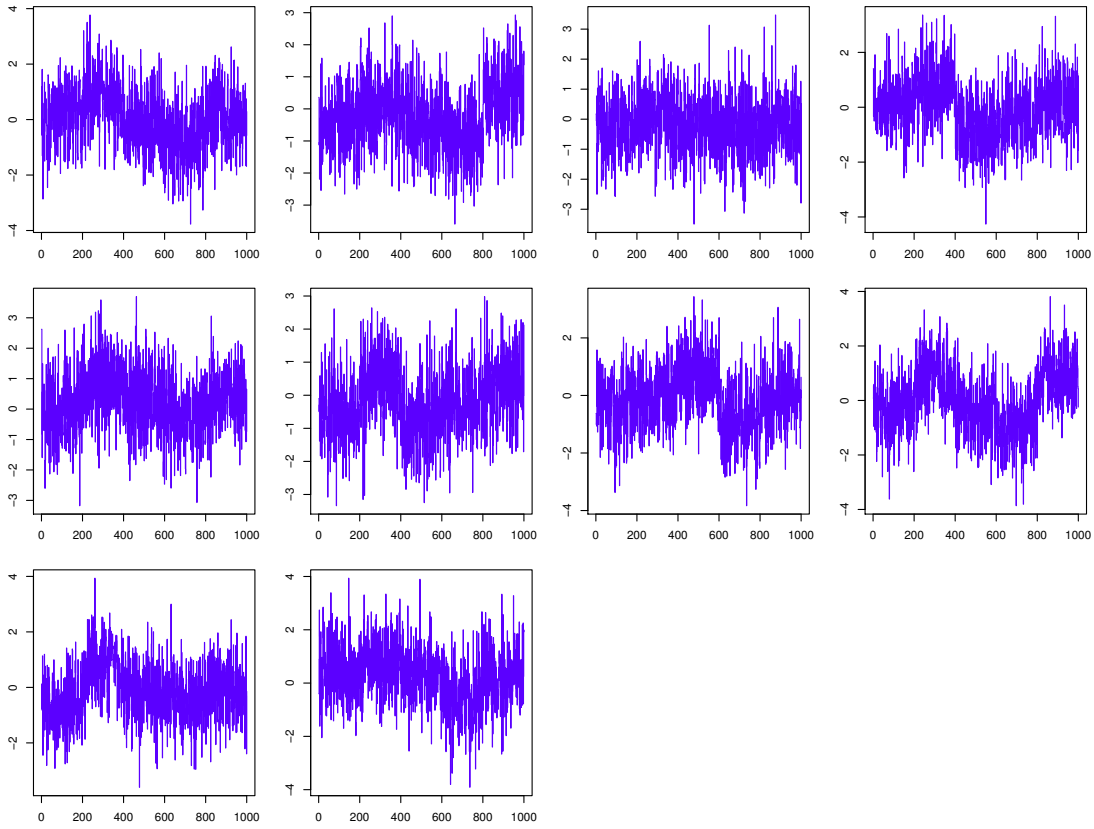
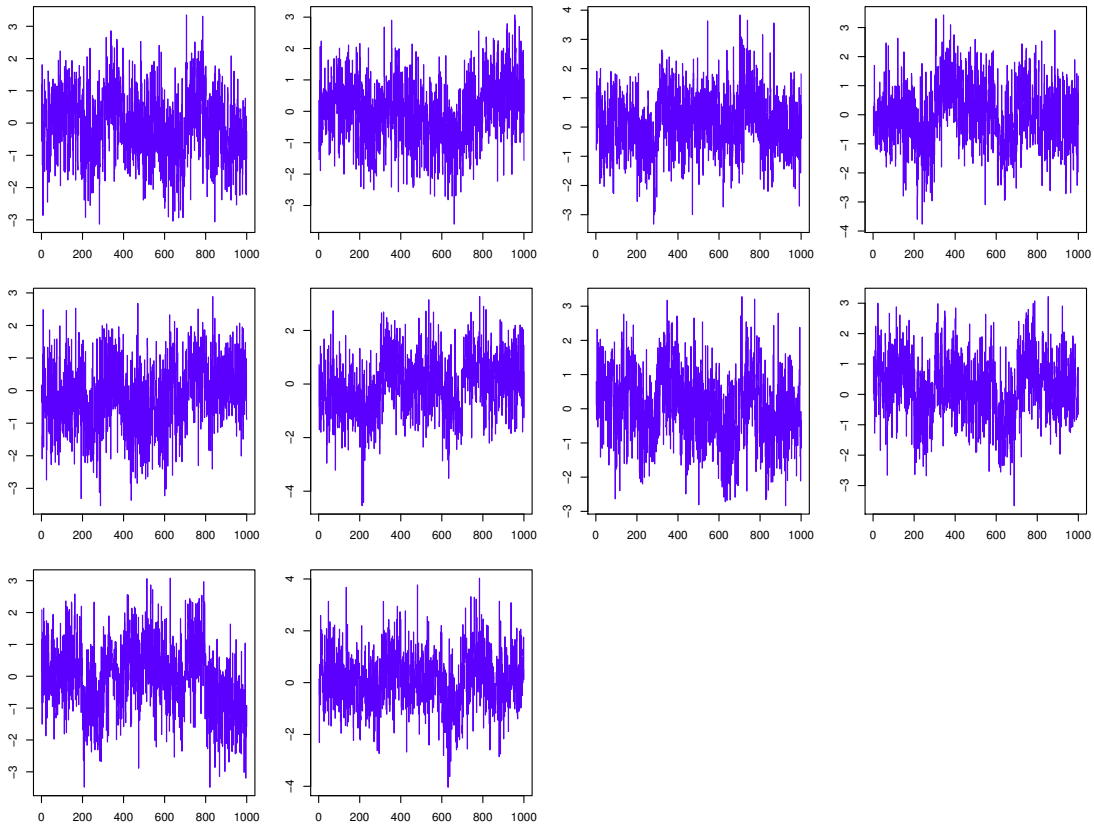


Figure 3.4: A selected series y_{lt} for 10 samples in Scenarios 4 (from left to right and top to bottom).



analyze the estimates, let us observe the true parameters in different regimes to have a better understanding of the model. In the last plot (Figure 3.8), there are three regimes and six transitions from regime 2 to 3, then to regime 1, and then back to regime 2, then to regime 3, then to regime 1, then back to regime 2 again. However, values of θ_{lt} within each regime are not the same. This is the new feature of our model as specified in assumption (A3). Different from the classic segmentation model in which θ_{lt} is a constant within each regime, in our model θ_{lt} is a random variable following some distribution within each regime. Now let us look at the estimation results. In the first two scenarios (Figure 3.5, 3.6), the estimated parameters are very close to the true θ_{lt}. In the last two scenarios (Figure 3.7,3.8) there are significant deviations between θ_{lt} and $\hat{\theta}_{lt|T}$.

Figure 3.9, 3.10, 3.11, 3.12 show the true and estimated $P(s_t = 1)$, $P(s_t = 2)$, $P(s_t = 3)$ of each series in four scenarios, respectively. Specifically, if the true regime is 1, the true probability of $P(s_t = 1) = 1$; if the true regime is 2 or 3, the true probability of $P(s_t = 1) = 0$. There are three regimes in our simulation setup. In all four scenarios, the estimated probabilities in BCMIX procedure (red line) are very close to the true probabilities (blue line). So the BCMIX procedure is robust and efficient to make inference on regimes.

Figure 3.5: BCMIX estimates (red line) of $\hat{\theta}_{lt|T}$ and true θ_{lt} (blue line) of the selected series for 10 samples in Scenarios 1 (from left to right and top to bottom).

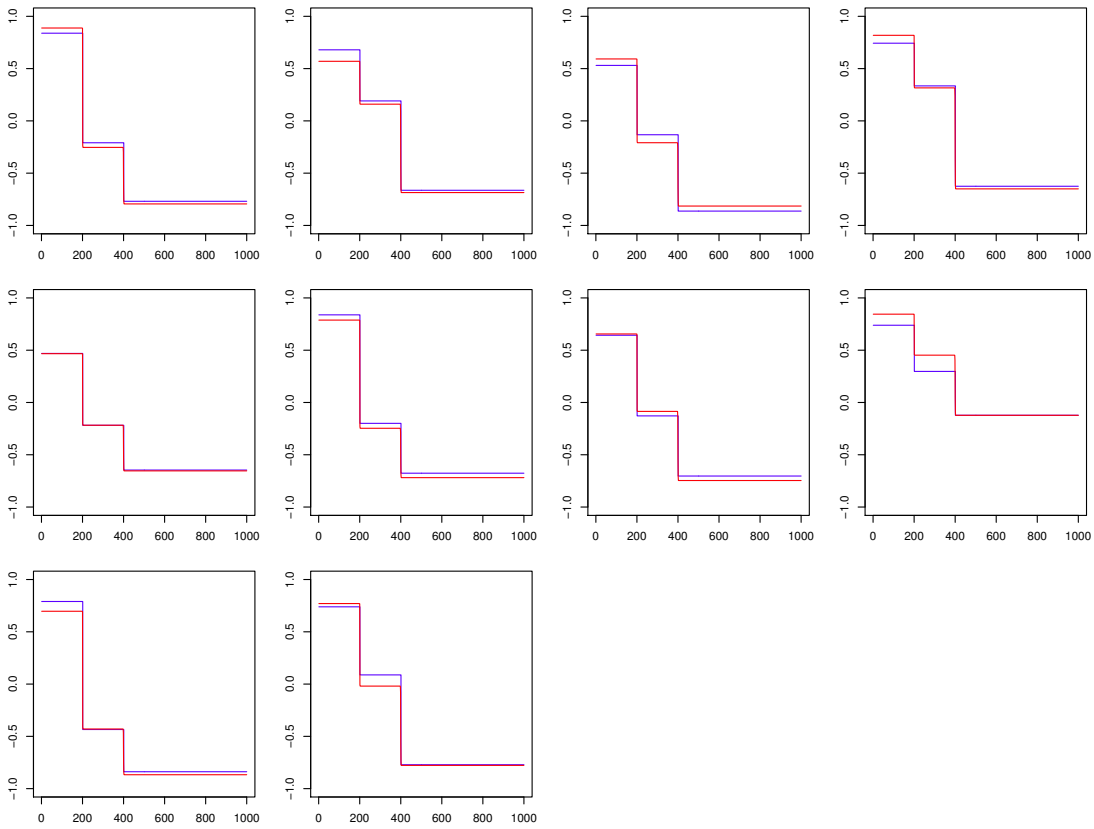


Figure 3.6: BCMIX estimates (red line) of $\hat{\theta}_{lt|T}$ and true θ_{lt} (blue line) of the selected series for 10 samples in Scenarios 2 (from left to right and top to bottom).

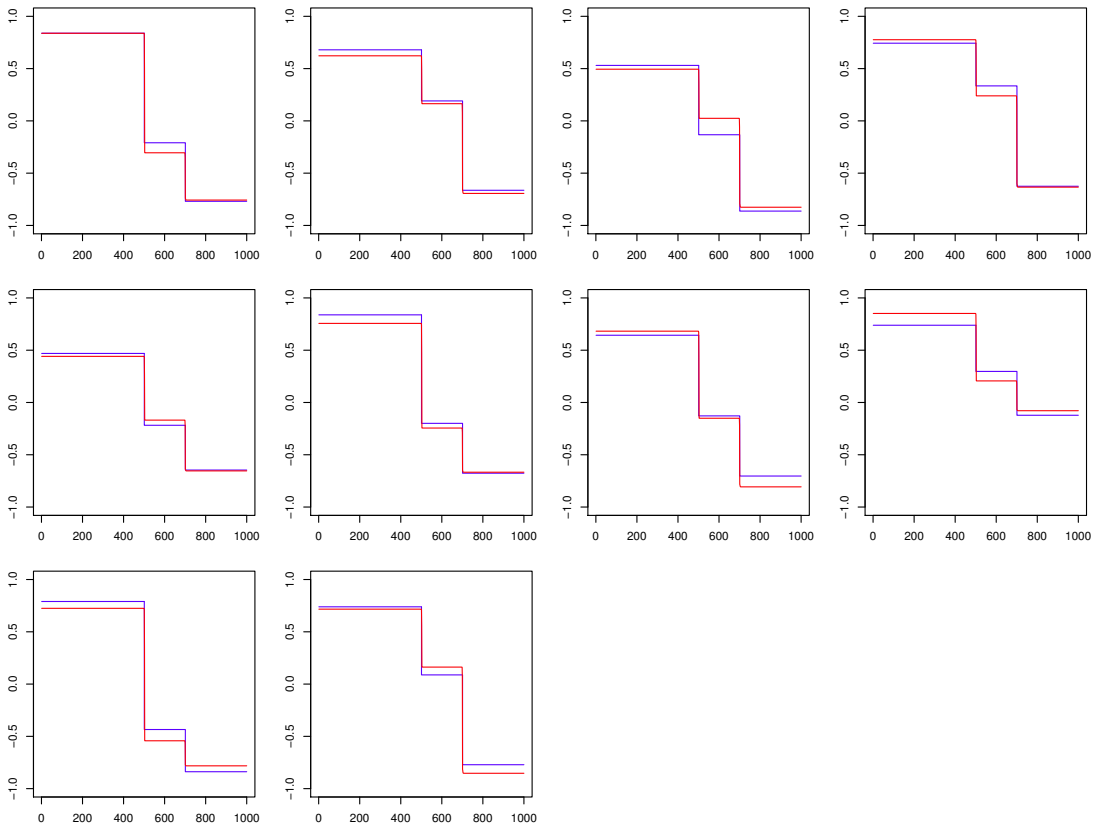


Figure 3.7: BCMIX estimates (red line) of $\hat{\theta}_{lt|T}$ and true θ_{lt} (blue line) of the selected series for 10 samples in Scenarios 3 (from left to right and top to bottom).

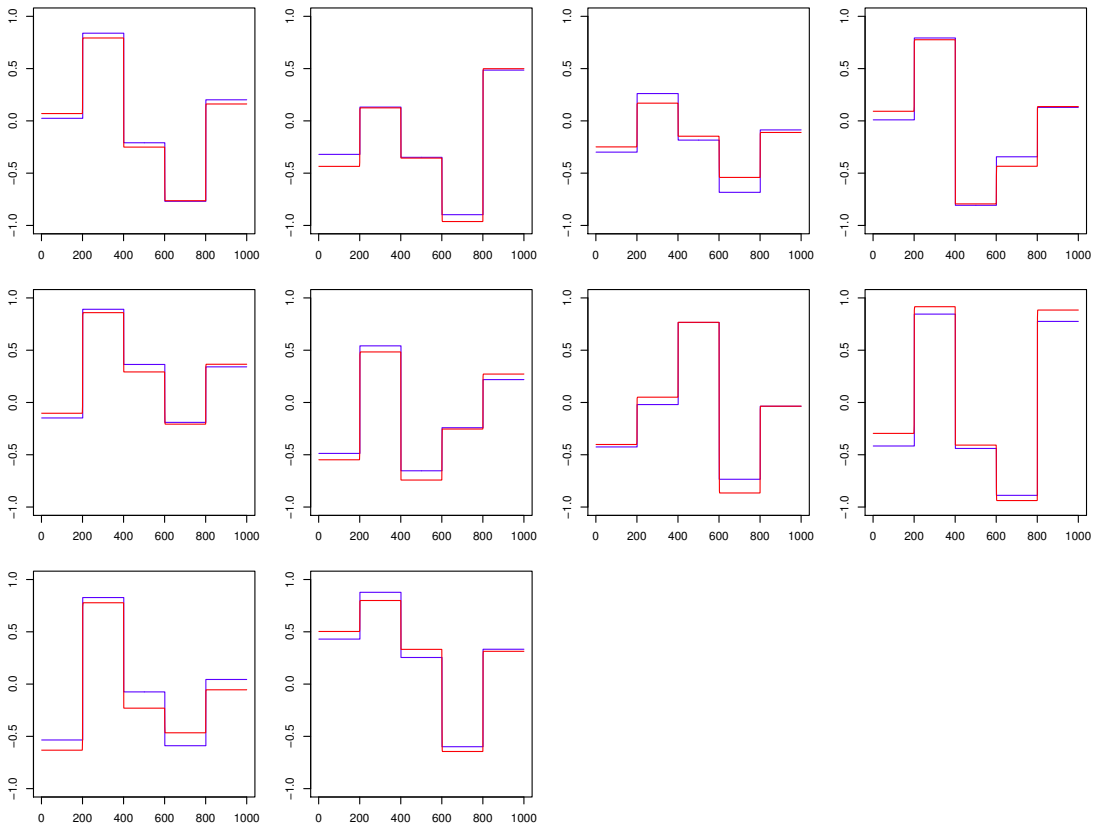


Figure 3.8: BCMIX estimates (red line) of $\hat{\theta}_{lt|T}$ and true θ_{lt} (blue line) of the selected series for 10 samples in Scenarios 4 (from left to right and top to bottom).

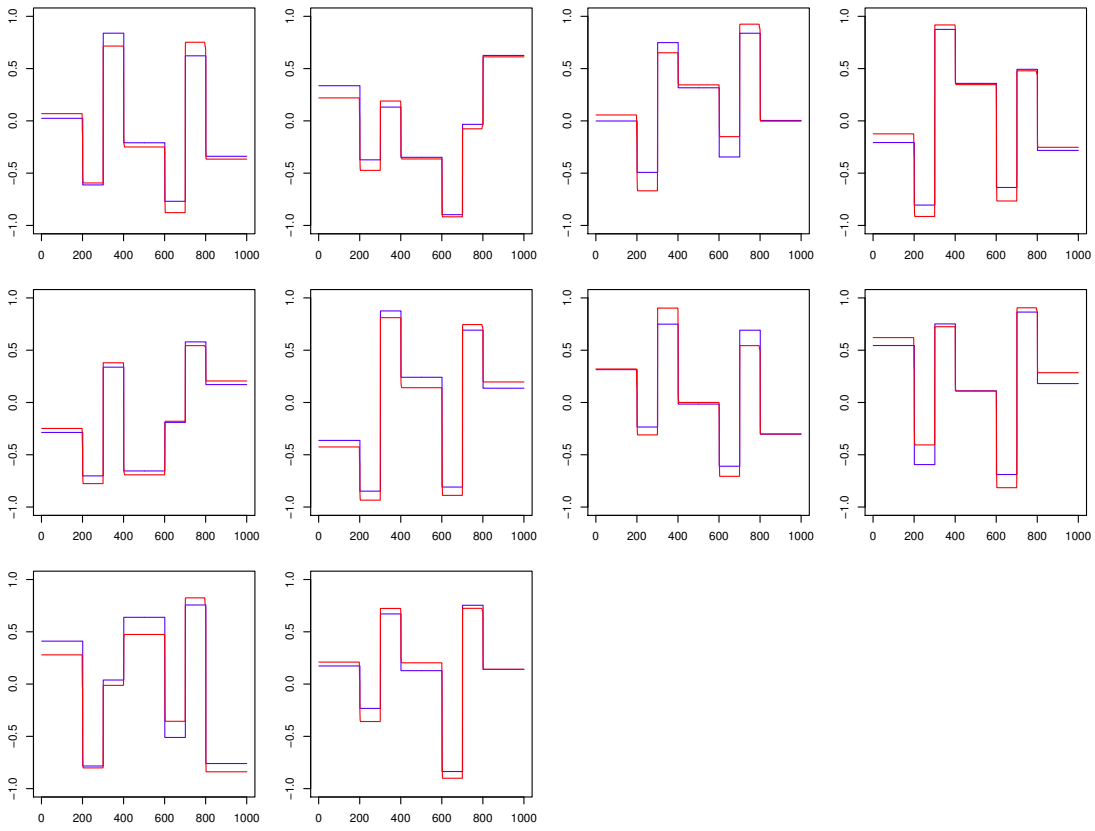


Figure 3.9: BCMIX estimates (red line) of $\hat{r}_{t|T}^{(1)}$ and true $P(s_t = 1)$ (blue line) (top), $P(s_t = 2)$ (blue line) (middle), $P(s_t = 3)$ (blue line) (bottom) of the selected series for 10 samples in Scenarios 1.

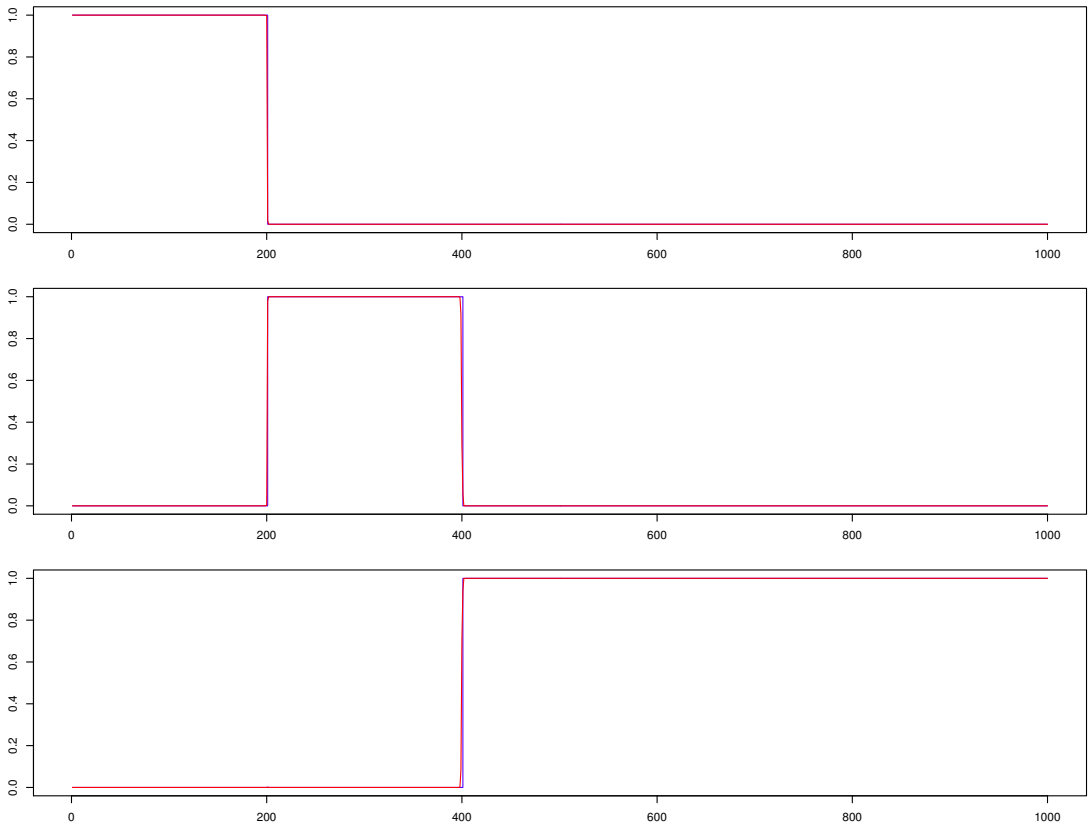


Figure 3.10: BCMIX estimates (red line) of $\hat{r}_{t|T}^{(1)}$ and true $P(s_t = 1)$ (blue line) (top), $P(s_t = 2)$ (blue line) (middle), $P(s_t = 3)$ (blue line) (bottom) of the selected series for 10 samples in Scenarios 2.

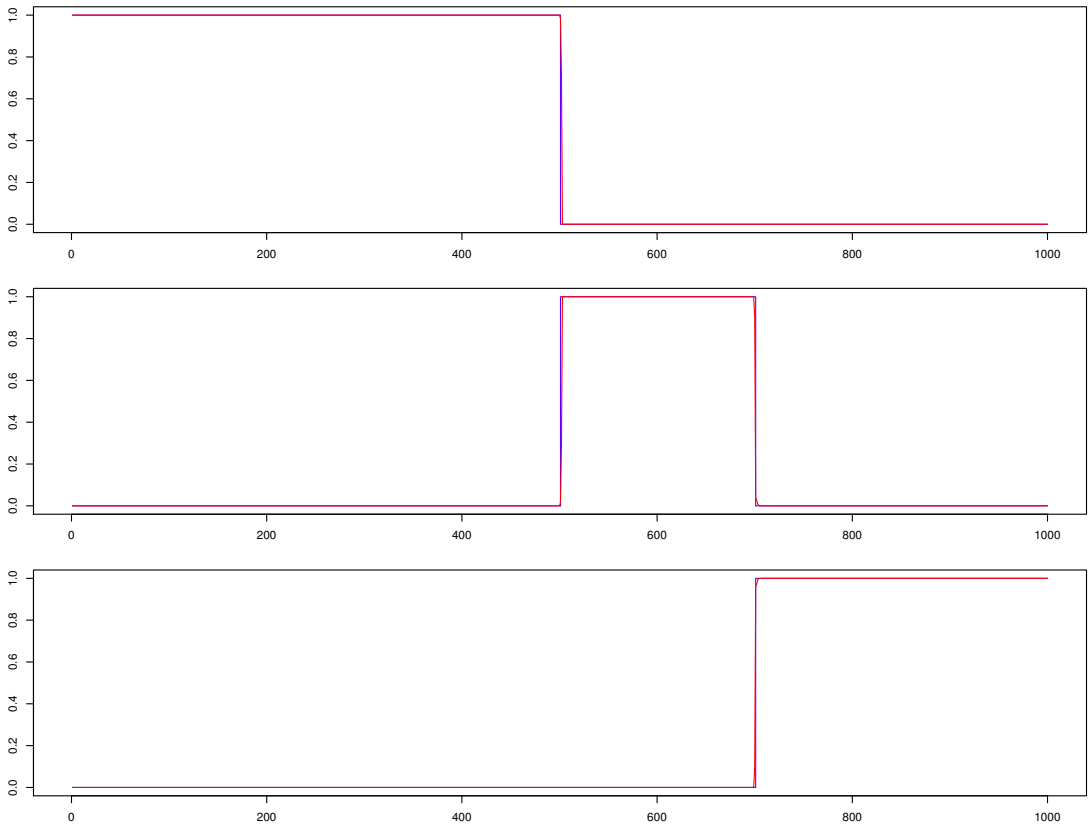


Figure 3.11: BCMIX estimates (red line) of $\hat{r}_{t|T}^{(1)}$ and true $P(s_t = 1)$ (blue line) (top), $P(s_t = 2)$ (blue line) (middle), $P(s_t = 3)$ (blue line) (bottom) of the selected series for 10 samples in Scenarios 3.

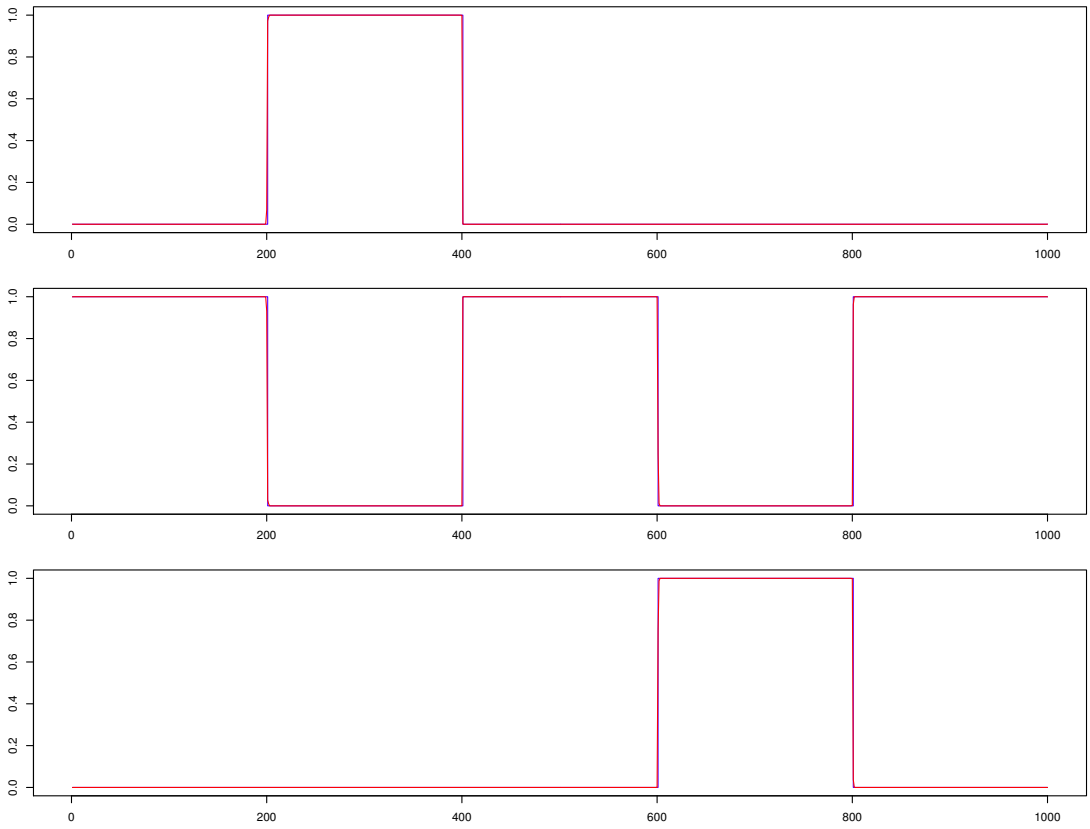
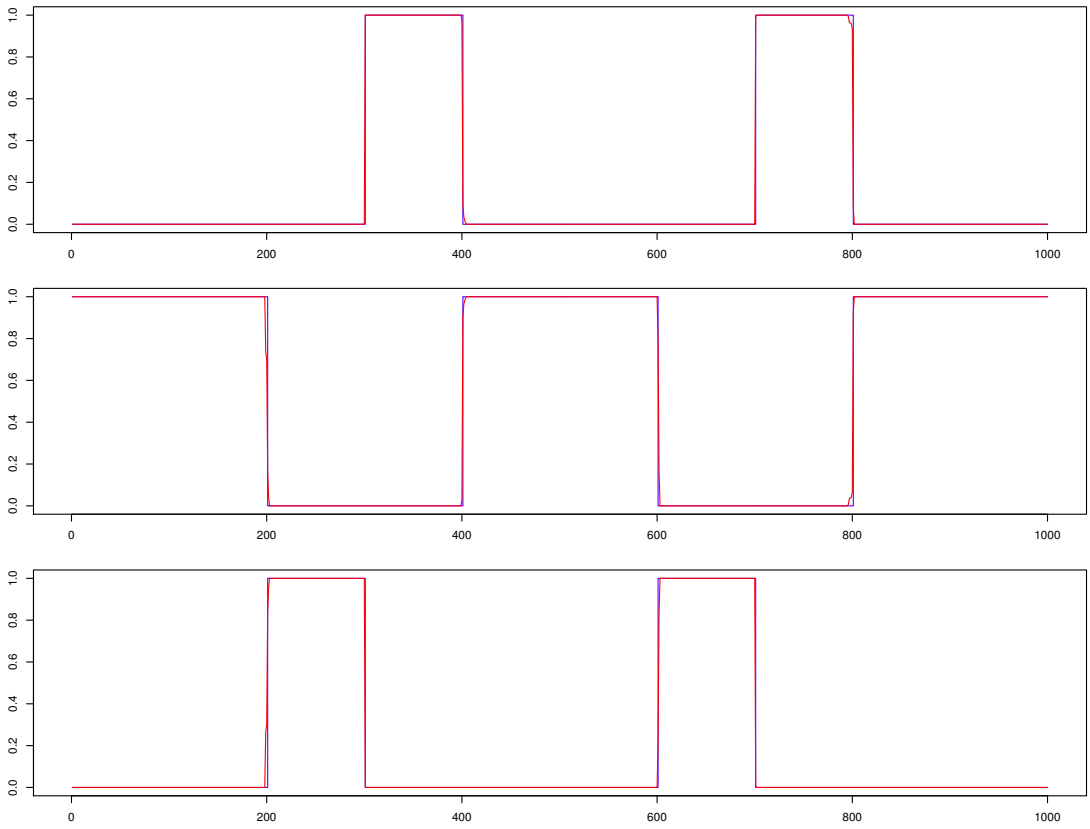


Figure 3.12: BCMIX estimates (red line) of $\hat{r}_{t|T}^{(1)}$ and true $P(s_t = 1)$ (blue line) (top), $P(s_t = 2)$ (blue line) (middle), $P(s_t = 3)$ (blue line) (bottom) of the selected series for 10 samples in Scenarios 4.



3.3 Simulation 2: Large Simulation with Different Simulation Setting

In this section, we will examine the effects of different simulation settings on the estimates. As mentioned in the last section, we will use the BCMIX procedure for large scale simulation studies with specific M and m. We assume we have the exact same model described in chapter 2. There are three states, $K = 3$. We still use the same parameter settings with the previous simulation study: $(z(l,1), V(l,1)) = (1, 0.22)$, $(z(l,2), V(l,2)) = (0, 0.22)$, $(z(l,3), V(l,3)) = (-1, 0.22)$, $\sigma_l^2 = 1$, $1 \leq l \leq 10$. The transition matrix is $P = \begin{pmatrix} 1-p_1-q_1 & p_1 & q_1 \\ p_2 & 1-p_2-q_2 & q_2 \\ p_3 & q_3 & 1-p_3-q_3 \end{pmatrix}$, which has the following settings:

Scenario 1. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.001, 0.001, 0.001, 0.001, 0.001, 0.001)$.

Scenario 2. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.002, 0.002, 0.002, 0.002, 0.002, 0.002)$.

Scenario 3. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.004, 0.004, 0.004, 0.004, 0.004, 0.004)$.

Scenario 4. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.008, 0.008, 0.008, 0.008, 0.008, 0.008)$.

Scenario 5. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.016, 0.016, 0.016, 0.016, 0.016, 0.016)$.

Scenario 6. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.002, 0.001, 0.002, 0.002, 0.001, 0.002)$.

Scenario 7. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.004, 0.001, 0.004, 0.004, 0.001, 0.004)$.

Scenario 8. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.001, 0.002, 0.001, 0.001, 0.001, 0.001)$.

Scenario 9. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.001, 0.004, 0.001, 0.001, 0.001, 0.001)$.

We generate 500 series, $N = 500$, and each of length T takes the values of 3000, 4000, 5000, 6000 and 7000 for each scenario. The BCMIX procedure with $M = 20$ and $m = 10$ is adopted to estimate the smoothing parameters and give inference on the states. Tables 3.2 compare the estimates in different scenarios in terms of the SSE. Each table has 5 columns

and 9 rows, in which every "cell" is the result of 500 times simulation for that specific scenario.

Let us look at Tables 3.2 column by column. Within each column, the sample size T is the same, but the value of p_1,q_1,p_2,q_2,p_3,q_3 is different. This infers that the transition matrix $P = \begin{pmatrix} 1-p_1-q_1 & p_1 & q_1 \\ p_2 & 1-p_2-q_2 & q_2 \\ p_3 & q_3 & 1-p_3-q_3 \end{pmatrix}$ is different and thus, the mean number of the change point is different for each row. Although we cannot guarantee the number of transitions are the same for each scenario since the they are generated by the Markov chain, we knew the more transitions should be expected once the p_1,q_1,p_2,q_2,p_3,q_3 become larger. From scenario 1 to scenario 5, p_1,q_1,p_2,q_2,p_3,q_3 become larger, so more transitions should be expected. Presumably the errors are getting larger when experience more transitions. For example, when $T = 7000$, $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.008$, SSE is 0.00531. The quantity of SSE decreases to 0.00411 when p_1,q_1,p_2,q_2,p_3,q_3 changes to 0.004, and decreases to 0.00284 when p_1,q_1,p_2,q_2,p_3,q_3 change to 0.002. The quantity increases to 0.00625 when p_1,q_1,p_2,q_2,p_3,q_3 become 0.016. For scenario 1, scenario 6 and scenario 8 , $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.001$, SSE is 0.00179, and increases to 0.00254 when q_1,p_3 remain at 0.001 and p_1,p_2,q_2,q_3 change to 0.002. The value of SSE slightly increase to 0.00193, when only q_1 change to 0.002.

From Tables 3.2, we can examine the effects of sample size on the performance. When T changes from 3000 to 4000, there is a slightly decrease in SSE. After that, the differences between SSE are quite small. For example, in scenario 5, $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.016$, SSE is 0.00639, when $T = 3000$, then there are almost no distinction among $T = 4000$, $T = 5000$, $T = 6000$ and $T = 7000$. In short, we can tell BCMIX has very good performance on all of these p_1,q_1,p_2,q_2,p_3,q_3 settings, because the largest SSE is only about 0.639%.

Table 3.3 summarizes the identification ratio (IR) in each scenario. Most ratios are

Table 3.2: Performance of Sum of squared errors (SSE) for BCMIX estimates for K=3. Standard errors are given in parentheses.

Scenarios	$T = 3000$	$T = 4000$	$T = 5000$	$T = 6000$	$T = 7000$
Scenario 1	0.00196	0.00187	0.00187	0.00179	0.00179
	(3.27E-05)	(2.70E-05)	(2.44E-05)	(2.16E-05)	(2.14E-05)
Scenario 2	0.00301	0.00292	0.00290	0.00286	0.00284
	(3.64E-05)	(3.09E-05)	(2.69E-05)	(2.41E-05)	(2.25E-05)
Scenario 3	0.00421	0.00414	0.00411	0.00414	0.00411
	(3.62E-05)	(3.02E-05)	(2.62E-05)	(2.50E-05)	(2.36E-05)
Scenario 4	0.00541	0.00529	0.00531	0.00527	0.00531
	(3.52E-05)	(3.00E-05)	(2.59E-05)	(2.35E-05)	(2.35E-05)
Scenario 5	0.00639	0.00625	0.00625	0.00626	0.00625
	(3.54E-05)	(2.99E-05)	(2.78E-05)	(2.36E-05)	(2.24E-05)
Scenario 6	0.00272	0.00264	0.00264	0.00258	0.00254
	(3.61E-05)	(3.08E-05)	(2.59E-05)	(2.50E-05)	(2.27E-05)
Scenario 7	0.00375	0.00363	0.00367	0.00365	0.00360
	(3.59E-05)	(3.11E-05)	(2.80E-05)	(2.53E-05)	(2.28E-05)
Scenario 8	0.00231	0.00201	0.00203	0.00195	0.00193
	(3.46E-05)	(2.89E-05)	(2.43E-05)	(2.24E-05)	(2.08E-05)
Scenario 9	0.00227	0.00218	0.00220	0.00212	0.00211
	(3.54E-05)	(3.03E-05)	(2.69E-05)	(2.43E-05)	(2.29E-05)

Table 3.3: Performance of identification ratio (IR) for BCMIX estimates for K=3. Standard errors are given in parentheses.

Scenarios	$T = 3000$	$T = 4000$	$T = 5000$	$T = 6000$	$T = 7000$
Scenario 1	0.956 (4.91E-03)	0.962 (3.66E-03)	0.954 (4.57E-03)	0.959 (3.44E-03)	0.962 (3.32E-03)
Scenario 2	0.956 (4.20E-03)	0.968 (2.86E-03)	0.960 (2.93E-03)	0.958 (2.63E-03)	0.958 (2.39E-03)
Scenario 3	0.955 (2.87E-03)	0.954 (2.66E-03)	0.957 (2.39E-03)	0.960 (2.09E-03)	0.962 (1.82E-03)
Scenario 4	0.951 (2.78E-03)	0.956 (2.20E-03)	0.960 (2.03E-03)	0.953 (1.88E-03)	0.956 (1.76E-03)
Scenario 5	0.948 (2.46E-03)	0.952 (2.00E-03)	0.956 (1.79E-03)	0.949 (1.65E-03)	0.950 (1.64E-03)
Scenario 6	0.964 (3.82E-03)	0.974 (2.68E-03)	0.971 (2.61E-03)	0.971 (2.66E-03)	0.970 (2.32E-03)
Scenario 7	0.981 (2.16E-03)	0.980 (2.00E-03)	0.977 (2.08E-03)	0.984 (1.49E-03)	0.983 (1.40E-03)
Scenario 8	0.950 (5.32E-03)	0.956 (3.89E-03)	0.947 (4.48E-03)	0.958 (3.24E-03)	0.959 (3.24E-03)
Scenario 9	0.952 (5.10E-03)	0.953 (3.75E-03)	0.955 (3.87E-03)	0.956 (3.12E-03)	0.957 (2.85E-03)

greater than 95%, showing that the BCMIX procedure is very effective in identifying the transitions.

3.4 Simulation 3: Large Simulation with Different Simulation Setting using EM Algorithm to Estimate Hyperparameters

In this section, the effects of different simulation settings on the estimates using EM algorithm are examined. Similar with previous sections, the BCMIX procedure for large scale simulation studies with specific M and m is used. It assumes that we have the exact same model described in chapter 2. Again there are three states, $K = 3$, true parameters for simulation are, $(z^{(l,1)}, V^{(l,1)}) = (1, 0.16)$, $(z^{(l,2)}, V^{(l,2)}) = (0, 0.16)$, $(z^{(l,3)}, V^{(l,3)}) = (-1, 0.16)$ and $\sigma_l^2 = 1$, $1 \leq l \leq 10$. $P = \begin{pmatrix} 1-p_1-q_1 & p_1 & q_1 \\ p_2 & 1-p_2-q_2 & q_2 \\ p_3 & q_3 & 1-p_3-q_3 \end{pmatrix}$ which has the following settings:

Scenario 1. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.001, 0.001, 0.001, 0.001, 0.001, 0.001)$.

Scenario 2. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.002, 0.002, 0.002, 0.002, 0.002, 0.002)$.

Scenario 3. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.004, 0.004, 0.004, 0.004, 0.004, 0.004)$.

Scenario 4. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.008, 0.008, 0.008, 0.008, 0.008, 0.008)$.

Scenario 5. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.016, 0.016, 0.016, 0.016, 0.016, 0.016)$.

Scenario 6. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.002, 0.001, 0.002, 0.002, 0.001, 0.002)$.

Scenario 7. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.004, 0.001, 0.004, 0.004, 0.001, 0.004)$.

Scenario 8. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.001, 0.002, 0.001, 0.001, 0.001, 0.001)$.

Scenario 9. $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.001, 0.004, 0.001, 0.001, 0.001, 0.001)$.

Let $N = 500$ and T take the values of 3000, 4000, 5000, 6000 and 7000 for each scenario. In each scenario, hyperparameter estimate uses EM algorithm until convergence. Then the estimates are computed. We give the hyperparameters some initial values as below: $(z^{(l,1)}, V^{(l,1)}) = (0.9, 0.12)$, $(z^{(l,2)}, V^{(l,2)}) = (0.1, 0.12)$, $(z^{(l,3)}, V^{(l,3)}) = (-1.1, 0.12)$ and $\sigma_l^2 =$

Table 3.4: Performance of sum of squared errors (SSE) using EM for K=3. Standard errors are given in parentheses.

Scenarios	$T = 3000$	$T = 4000$	$T = 5000$	$T = 6000$	$T = 7000$
Scenario 1	0.00228 (3.72E-05)	0.00213 (3.01E-05)	0.00208 (2.64E-05)	0.00201 (2.36E-05)	0.00198 (2.22E-05)
Scenario 2	0.00339 (4.10E-05)	0.00322 (3.41E-05)	0.00317 (2.88E-05)	0.00313 (2.58E-05)	0.00304 (2.38E-05)
Scenario 3	0.00462 (4.00E-05)	0.00447 (3.32E-05)	0.00443 (3.09E-05)	0.00442 (2.73E-05)	0.00431 (2.40E-05)
Scenario 4	0.00585 (3.89E-05)	0.00562 (3.18E-05)	0.00560 (2.76E-05)	0.00557 (2.54E-05)	0.00555 (2.43E-05)
Scenario 5	0.00675 (3.83E-05)	0.00659 (3.14E-05)	0.00653 (2.93E-05)	0.00652 (2.47E-05)	0.00645 (2.34E-05)
Scenario 6	0.00303 (3.86E-05)	0.00289 (3.35E-05)	0.00282 (2.84E-05)	0.00276 (2.55E-05)	0.00273 (2.36E-05)
Scenario 7	0.00399 (3.83E-05)	0.00387 (3.42E-05)	0.00382 (2.96E-05)	0.00379 (2.70E-05)	0.00375 (2.41E-05)
Scenario 8	0.00248 (3.97E-05)	0.00228 (3.18E-05)	0.00225 (2.61E-05)	0.00219 (2.47E-05)	0.00213 (2.24E-05)
Scenario 9	0.00266 (3.86E-05)	0.00247 (3.33E-05)	0.00247 (2.99E-05)	0.00237 (2.62E-05)	0.00233 (2.45E-05)

1.1, $1 \leq l \leq 10$. and $(p_1, q_1, p_2, q_2, p_3, q_3) = (0.01, 0.01, 0.01, 0.01, 0.01, 0.01)$. The BCMIX procedure with $M = 20$ and $m = 10$ is adopted to estimate the smoothing parameters and produce inference on the states. Table 3.4 compares the estimates in different scenarios in terms of the SSE. Each table has 5 columns and 9 rows, in which every "cell" contains the result of 500 times simulation for that specific scenario.

In Table 3.4, within each column, the sample size T is the same, but the values of p_1,q_1,p_2,q_2,p_3,q_3 are different. From scenario 1 to scenario 5, p_1,q_1,p_2,q_2,p_3,q_3 become larger, so more transitions should be expected. Presumably the errors are getting larger when experience more transitions. For example, when $T = 7000$, $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.008$, SSE is 0.00555. The quantity of SSE decreases to 0.00431 when p_1,q_1,p_2,q_2,p_3,q_3

Table 3.5: Performance of identification ratio (IR) using EM for K=3. Standard errors are given in parentheses.

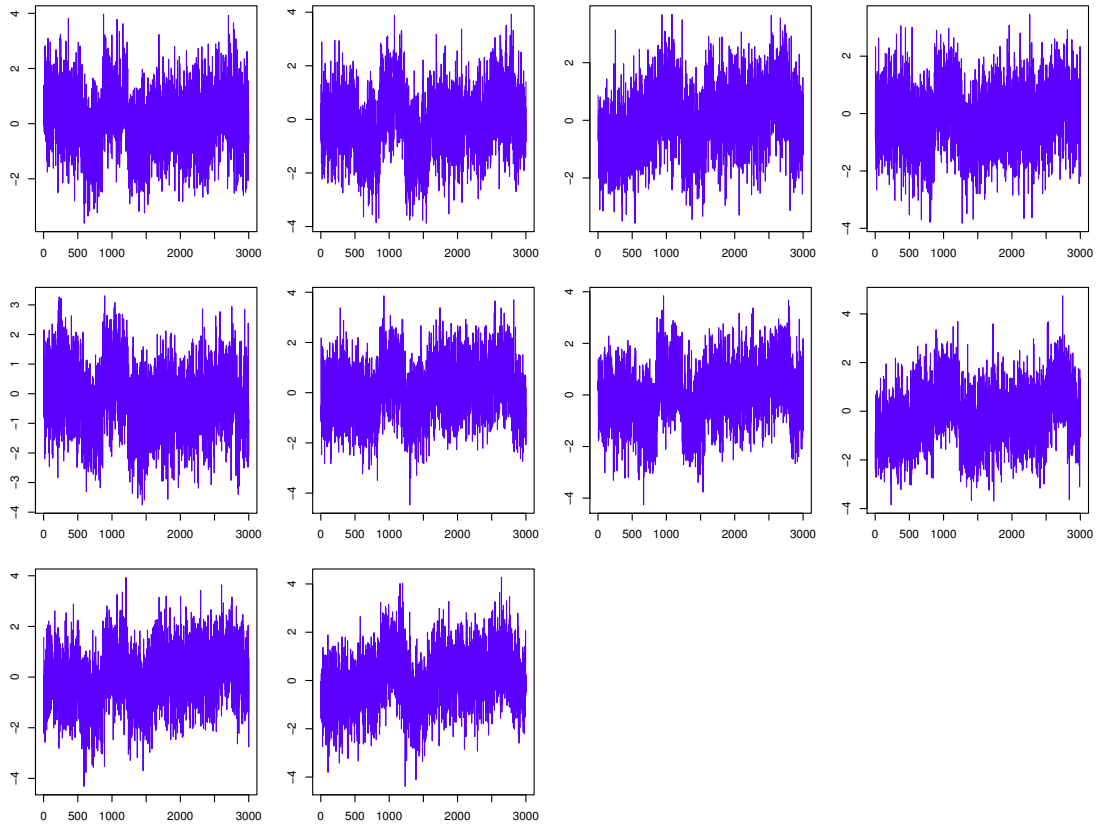
Scenarios	$T = 3000$	$T = 4000$	$T = 5000$	$T = 6000$	$T = 7000$
Scenario 1	0.926	0.919	0.880	0.870	0.874
	(6.18E-03)	(5.66E-03)	(6.72E-03)	(6.23E-03)	(6.19E-03)
Scenario 2	0.890	0.863	0.842	0.831	0.827
	(6.21E-03)	(5.59E-03)	(5.39E-03)	(5.51E-03)	(5.21E-03)
Scenario 3	0.857	0.824	0.816	0.808	0.802
	(5.14E-03)	(5.51E-03)	(4.86E-03)	(4.55E-03)	(4.47E-03)
Scenario 4	0.829	0.819	0.795	0.781	0.777
	(4.85E-03)	(4.61E-03)	(4.15E-03)	(4.25E-03)	(4.01E-03)
Scenario 5	0.826	0.802	0.787	0.776	0.766
	(4.52E-03)	(4.36E-03)	(4.00E-03)	(3.95E-03)	(3.78E-03)
Scenario 6	0.921	0.912	0.885	0.872	0.870
	(5.58E-03)	(5.09E-03)	(5.72E-03)	(5.82E-03)	(5.24E-03)
Scenario 7	0.935	0.918	0.919	0.908	0.904
	(4.28E-03)	(4.79E-03)	(3.99E-03)	(4.19E-03)	(4.17E-03)
Scenario 8	0.904	0.895	0.857	0.857	0.845
	(7.46E-03)	(6.53E-03)	(7.41E-03)	(6.62E-03)	(6.34E-03)
Scenario 9	0.905	0.876	0.858	0.841	0.831
	(7.13E-03)	(7.15E-03)	(6.98E-03)	(7.02E-03)	(7.03E-03)

changes to 0.004, and decreases to 0.00304 when p_1,q_1,p_2,q_2,p_3,q_3 change to 0.002. The quantity increases to 0.00645 when p_1,q_1,p_2,q_2,p_3,q_3 become 0.016. For scenario 1, scenario 6 and scenario 8 , $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.001$, SSE is 0.00198, and increases to 0.00273 when q_1, p_3 remain at 0.001 and p_1, p_2, q_2, q_3 change to 0.002. The value of SSE slightly increases to 0.00213, when only q_1 changes to 0.002.

From Table 3.4, we can observe the effects of sample size on the performance. When T changes from 3000 to 4000, there is a slightly decrease in SSE. After that, the differences between SSE are quite small. For example, in scenario 5, $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.016$, SSE is 0.00675, when $T = 3000$, then there are almost no difference among $T = 4000$, $T = 5000$, $T = 6000$ and $T = 7000$. In summary, we can say BCMIX has very good performance on all of these p_1,q_1,p_2,q_2,p_3,q_3 settings since the largest SSE is still quite small, only about 0.675%.

Table 3.5 summarizes the identification ratio (IR) in each scenario. Most ratios are greater than 80%, showing that the BCMIX procedure is very efficient in identifying the transitions.

Figure 3.13: A selected series y_{lt} for 10 samples in Scenario 1 (from left to right and top to bottom).



As in the first section, we will show some figures of a randomly selected simulation path in each scenario to visualize the simulation results. Considering the limitations of space, I only show the results of the first two scenarios. Figures 3.13, 3.16, show the series y_{lt} in two scenarios with $T = 3000$ for 10 samples. We find more fluctuations in magnitude in each series when p_1,q_1,p_2,q_2,p_3,q_3 become larger. Figures 3.14, 3.17 compare the true θ_{lt} with $\hat{\theta}_{lt|T}$ of the same series for 10 samples in these two scenarios. From Figures 3.14, 3.17, it is clear that when p_1,q_1,p_2,q_2,p_3,q_3 become larger, the series experiences more frequent transitions. For example, in the Figure 3.14 with $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.001$, there are 6 transitions in total for each sample, while in the Figure 3.17 with $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.02$,

Figure 3.14: BCMIX estimates $\hat{\theta}_{lt|T}$ (dashed line) and true θ_{lt} (solid line) of the selected series for 10 samples in Scenario 1 (from left to right and top to bottom).

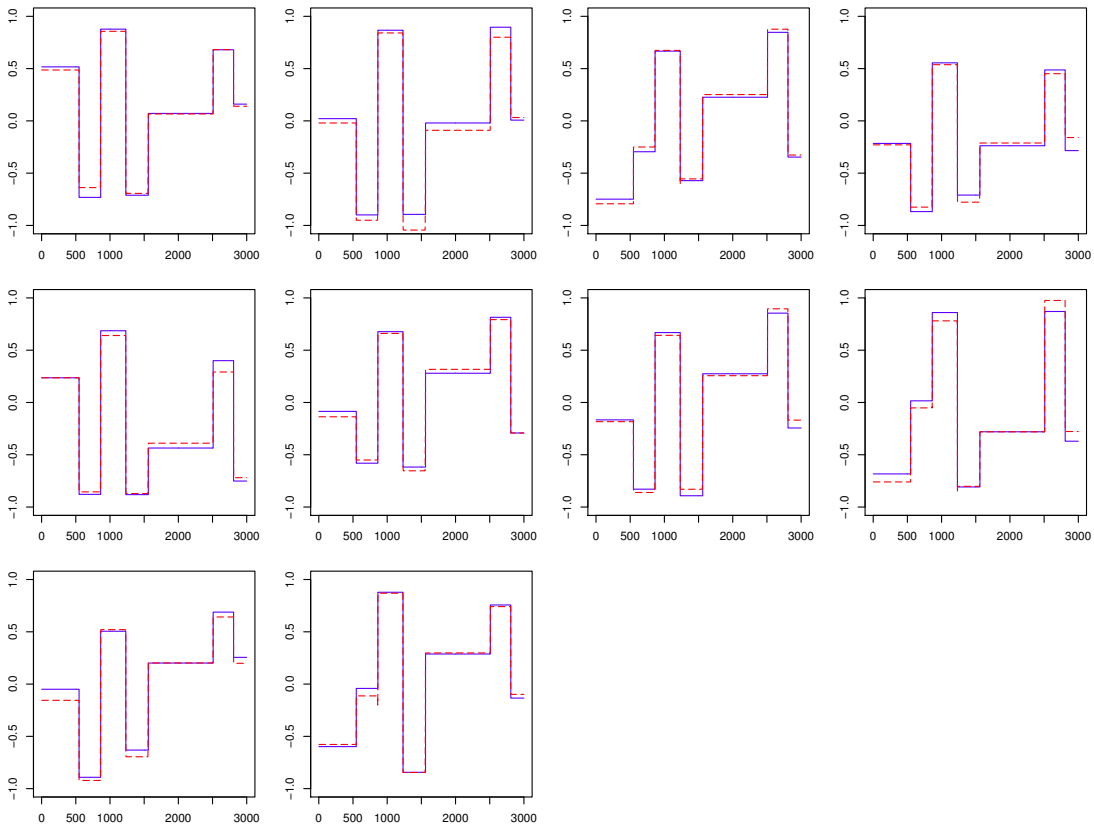


Figure 3.15: BCMIX estimates $\hat{r}_{t|T}^{(1)}$ (red points) and true $P(s_t = 1)$ (solid line) (top), $P(s_t = 2)$ (solid line) (middle), $P(s_t = 3)$ (solid line) (bottom) of the selected series for 10 samples in Scenario 1.

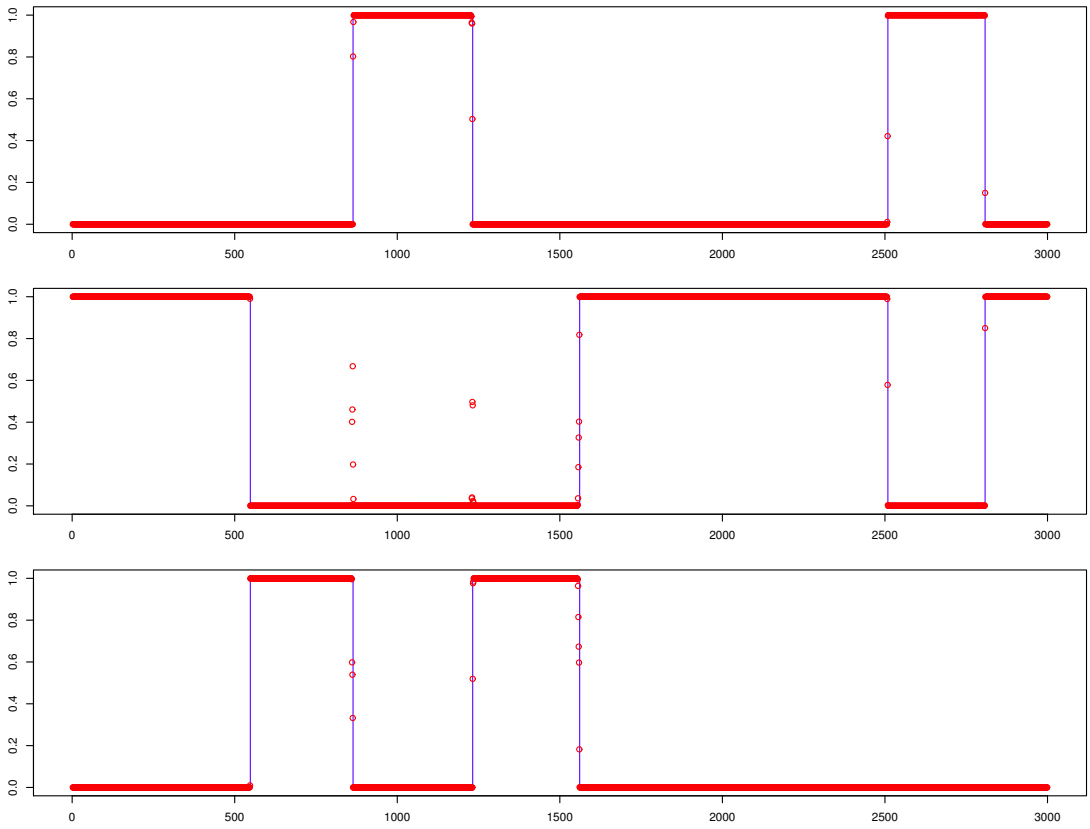


Figure 3.16: A selected series y_{lt} for 10 samples in Scenario 2 (from left to right and top to bottom).

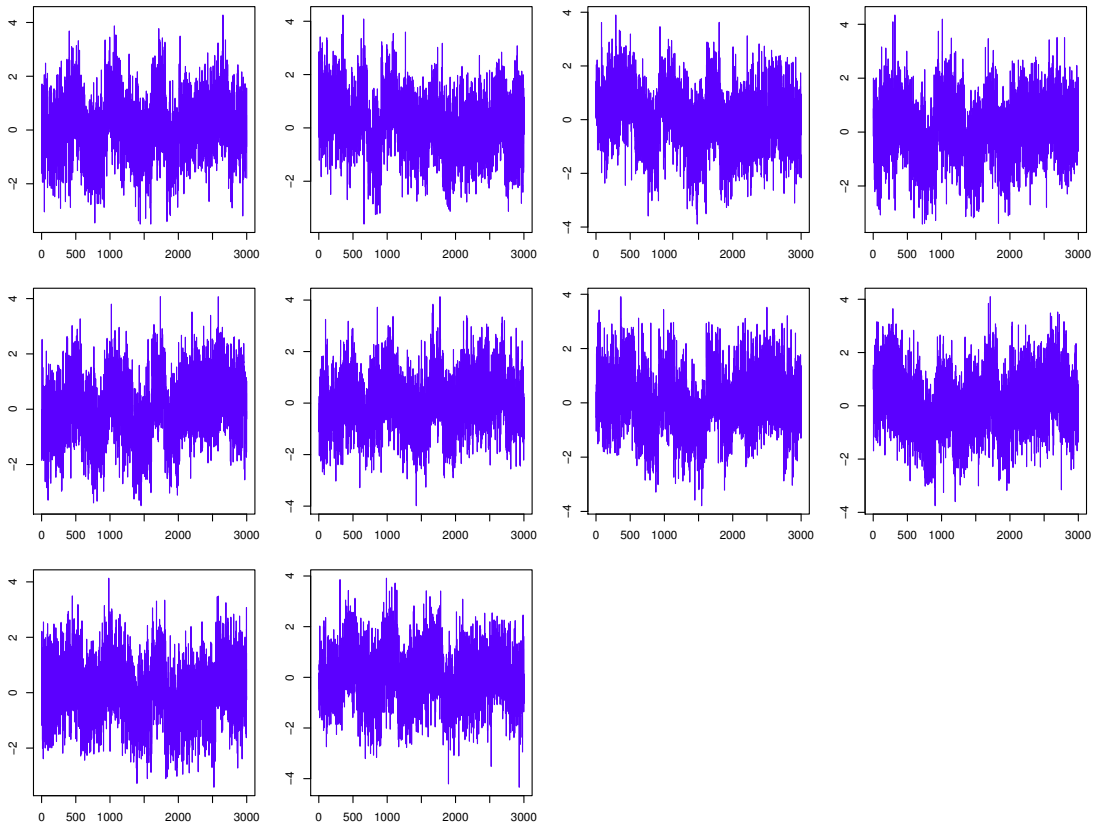


Figure 3.17: BCMIX estimates $\hat{\theta}_{lt|T}$ (dashed line) and true θ_{lt} (solid line) of the selected series for 10 samples in Scenario 2 (from left to right and top to bottom).

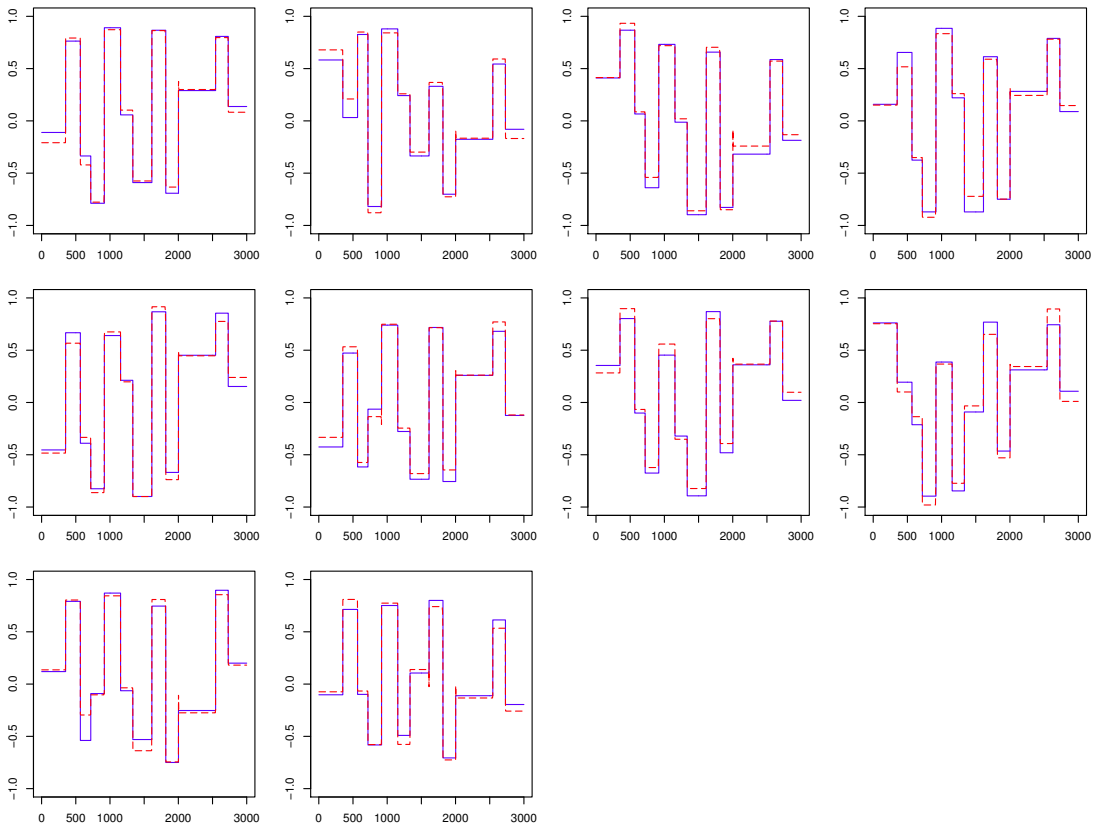
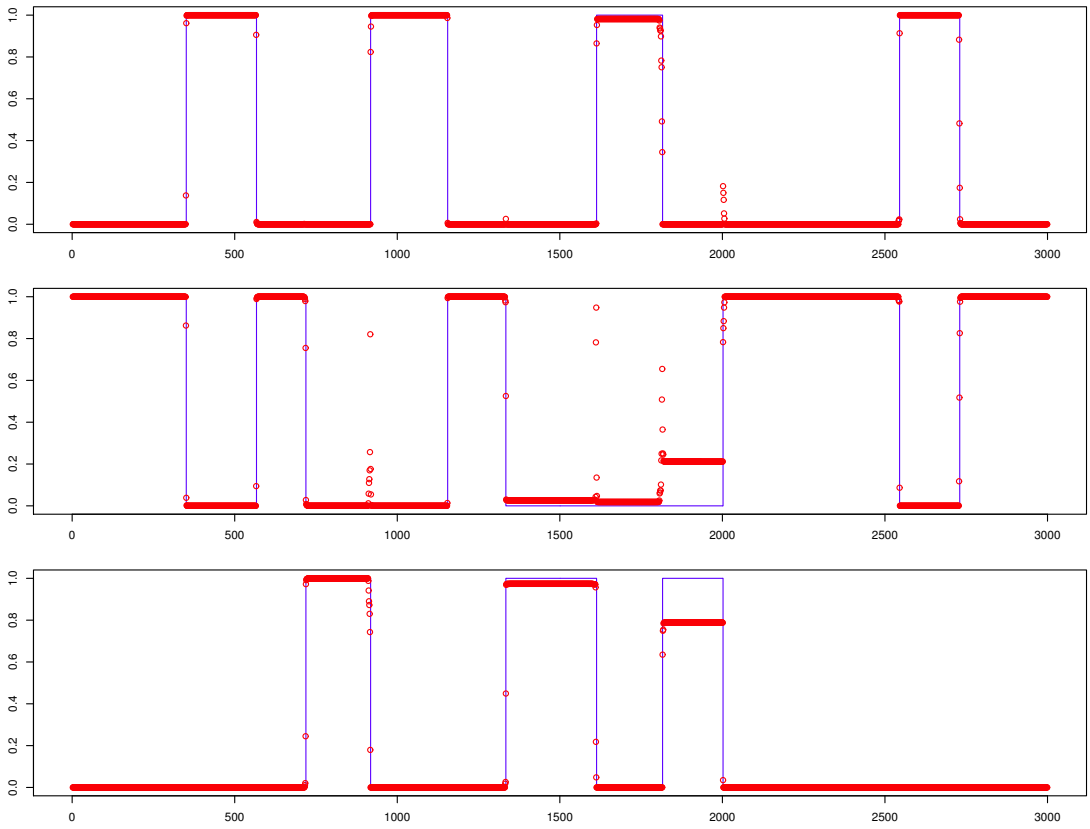


Figure 3.18: BCMIX estimates $\hat{r}_{t|T}^{(1)}$ (red points) and true $P(s_t = 1)$ (solid line) (top), $P(s_t = 2)$ (solid line) (middle), $P(s_t = 3)$ (solid line) (bottom) of the selected series for 10 samples in Scenario 2.



there are 11 transitions. In each plot of each figure there are three regimes, but the values of θ_{lt} within each regime are not constant. In each plot, the estimated parameter is close to the true θ_{lt} with some errors, which become more significant when there are more transitions.

Figure 3.15, 3.18 shows the true and estimated $P(s_t = 1)$, $P(s_t = 2)$, $P(s_t = 3)$ of the same series for 10 samples in each scenario. To clearly show the fuzziness around each transition, I use points to denote the estimated probabilities. It is clear that when there is a transition, it takes a while to recognize it. So the probability of $P(s_t = 1)$ does not jump directly from 1 to 0 or 0 to 1. Instead it adjusts step by step and takes some values in between. These "middle" points may affect the identification ratio. Moreover, there are more middle points when there are more frequent transitions, although the IR is higher. In our simulation, the three states are close with large variance, thus make the model difficult to correctly estimate the posterior mean. in Figure 3.16 with $p_1 = q_1 = p_2 = q_2 = p_3 = q_3 = 0.002$, we can discover some transitions with "blur" boundaries and we cannot know the number of transitions and the magnitude of each states. The Figure 3.17 demonstrates that the "shape" of the means are much different with the previous ones. In this scenario, the mean of state 1 sometimes are smaller than the mean of state 2 or 3 in some plots, since the large variation of the hidden variable. Moreover, the Figure 3.18 has clearly shows the fuzziness around each transitions. Even though there are some obvious differences between true state probability and estimated probability, their value still indicate the correct state calling.

Next, in order to access the advantage of our model, we compare our model to an existing hierarchical hidden Markov model (HMM) proposed by Shah et al., (2007) in terms of the identification ratio of true state calling. Shah assumed that the hidden parameter is conditionally independent given state, and follows a normal distribution. While in our model, we assume the dynamics of parameter is a Markov process, based solely on the state of the most recent time point, which is more close to what happens in real world.

Table 3.6: Performance of identification ratio (IR) using hierarchical HMM for K=3. Standard errors are given in parentheses.

Scenarios	$T = 3000$	$T = 4000$
Scenario 1	0.733 (2.43E-02)	0.729 (2.63E-02)
Scenario 2	0.726 (1.54E-02)	0.710 (1.36E-02)
Scenario 3	0.699 (2.53E-02)	0.695 (3.62E-02)

Furthermore, instead of using Markov Chain Monte Carlo (MCMC) algorithm to estimate hyperparameters in Shah's model, the hyperparameters in our model are estimated by an EM algorithm. To speed up the computations involved in the EM algorithm, we use the BCMIX approximations. So the running time of our model is quite fast.

Table 3.6 summarizes the identification ratio (IR) using hierarchical HMM in some scenarios chosen from the above simulation settings. Table has 2 columns and 3 rows, in which every "cell" is the result of 100 times simulation for that specific scenario instead of 500 times simulation, because 500 times simulation takes too much time.

As mentioned before, in this simulation, the three states are very close to each other with large variance, thus it is difficult to make a correct state calling. We can see all the ratios are about 70%, which is typically smaller than the corresponding ratios in Table 3.5. This comparison illustrates that our model is more accurate and effective than Shah's hierarchical HMM model in identifying the transitions, in the case that the regimes are very close to each other.

At the end of this section, we will display a special simulation which is not satisfied with the assumptions of our model. Instead of considering the change points are the same across all 10 samples, we will set that the change points are different across 10 samples which is corresponding with the real condition. There are two states, $K = 2$, true parameter for simulation, $(z^{(l,1)}, V^{(l,1)}) = (1, 0.01)$, $(z^{(l,2)}, V^{(l,2)}) = (-1, 0.01)$, and $\sigma_l^2 = 0.04$, $1 \leq l \leq 10$. The transition matrix is $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$, which has the following settings:

Scenario 1. $(p, q) = (0.001, 0.001)$.

Scenario 2. $(p, q) = (0.002, 0.001)$.

Scenario 3. $(p, q) = (0.002, 0.002)$.

Scenario 4. $(p, q) = (0.004, 0.001)$.

Scenario 5. $(p, q) = (0.004, 0.002)$.

Scenario 6. $(p, q) = (0.008, 0.004)$.

Scenario 7. $(p, q) = (0.008, 0.008)$.

Scenario 8. $(p, q) = (0.016, 0.008)$.

Scenario 9. $(p, q) = (0.016, 0.016)$.

$N = 500$, and T takes the values of 3000, 4000, 5000, 6000, 7000 and 8000 for each scenario. In each scenario, hyperparameter estimate uses EM algorithm until convergence. Then the estimates are computed. We give the hyperparameters some initial values as below: $(z^{(l,1)}, V^{(l,1)}) = (0.6, 0.1)$, $(z^{(l,2)}, V^{(l,2)}) = (-1.4, 0.1)$ and $\sigma_l^2 = 0.1$, $1 \leq l \leq 10$. and $(p, q) = (0.01, 0.01)$. The BCMIX procedure with $M = 20$ and $m = 10$ is adopted to estimate the smoothing parameters and give inference on the states. Tables 3.7 compares the estimates in different scenarios in terms of the SSE. Each table has 5 columns and 9 rows, in which every "cell" is the result of 500 times simulation for that specific scenario.

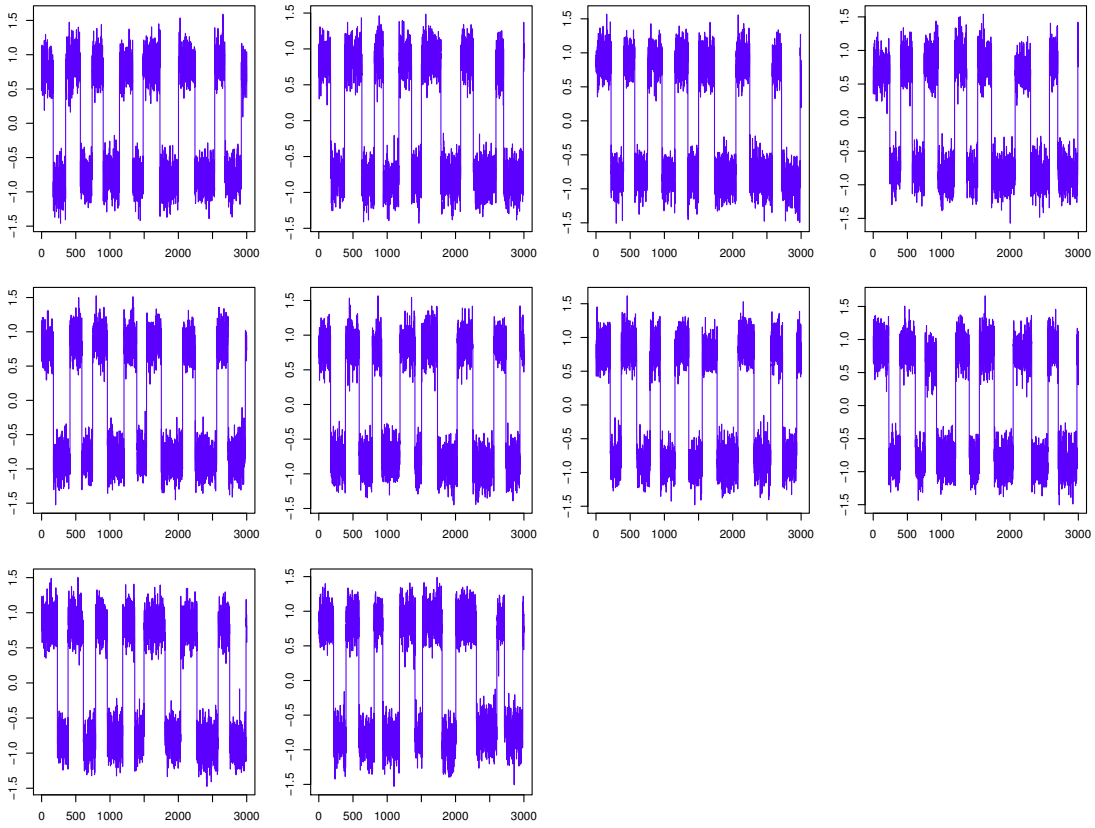
Let us Look at Tables 3.7 column by column. In each column, the sample size T is

Table 3.7: Performance of Sum of squared errors (SSE) using EM for K=2. Standard errors are given in parentheses.

Scenarios	$T = 3000$	$T = 4000$	$T = 5000$	$T = 6000$	$T = 7000$	$T = 8000$
$p = 0.001$	0.00141	0.00139	0.00134	0.00128	0.00130	0.00130
$q = 0.001$	(3.17E-05)	(2.77E-05)	(2.59E-05)	(2.28E-05)	(2.22E-05)	(2.01E-05)
$p = 0.002$	0.00168	0.00166	0.00165	0.00158	0.00157	0.00158
$q = 0.001$	(3.66E-05)	(3.25E-05)	(3.08E-05)	(2.63E-05)	(2.49E-05)	(2.19E-05)
$p = 0.002$	0.00227	0.00228	0.00228	0.00224	0.00222	0.00222
$q = 0.002$	(4.45E-05)	(3.94E-05)	(3.09E-05)	(2.94E-05)	(2.64E-05)	(2.57E-05)
$p = 0.004$	0.00192	0.00184	0.00181	0.00177	0.00174	0.00176
$q = 0.001$	(4.06E-05)	(3.56E-05)	(3.28E-05)	(2.92E-05)	(2.73E-05)	(2.50E-05)
$p = 0.004$	0.00276	0.00266	0.00271	0.00267	0.00267	0.00263
$q = 0.002$	(4.86E-05)	(4.30E-05)	(3.50E-05)	(3.19E-05)	(2.88E-05)	(2.69E-05)
$p = 0.008$	0.00450	0.00435	0.00439	0.00438	0.00438	0.00436
$q = 0.004$	(5.79E-05)	(4.72E-05)	(4.17E-05)	(4.06E-05)	(3.49E-05)	(3.03E-05)
$p = 0.008$	0.00585	0.00571	0.00575	0.00570	0.00571	0.00571
$q = 0.008$	(5.34E-05)	(4.80E-05)	(4.24E-05)	(3.99E-05)	(3.50E-05)	(3.35E-05)
$p = 0.016$	0.00664	0.00658	0.00659	0.00662	0.00655	0.00658
$q = 0.008$	(5.69E-05)	(4.92E-05)	(4.52E-05)	(4.14E-05)	(3.86E-05)	(3.47E-05)
$p = 0.016$	0.00803	0.00801	0.00799	0.00793	0.00798	0.00799
$q = 0.016$	(5.38E-05)	(4.65E-05)	(4.17E-05)	(4.00E-05)	(3.60E-05)	(3.28E-05)

fixed, but p and q are changing. Therefore the transition matrix $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ is different for each row. From top to bottom p and q become larger, so more transitions should be expected and the larger are SSE. For example, when $T = 8000$, $p = 0.008$, and $q = 0.008$, SSE is 0.00571. The quantity SSE decreases to 0.00436 when p remains at 0.008 and q changes to 0.004, and decreases to 0.00263 when p and q change to 0.004 and 0.002 respectively. The quantity increases to 0.00799 when both p and q become 0.016. When T changes from 3000 to 8000, the differences between SSE are small. For example, in scenario 9, $p = q = 0.016$, SSE is around 0.008, from $T = 3000$ to $T = 8000$. In short, we can tell BCMIX has very good performance on all of these p,q settings since the largest SSE is only about 0.803%.

Figure 3.19: A selected series y_{lt} for 10 samples in Scenario 9 (from left to right and top to bottom).



Because of the limitation of report's length. We choose the largest p, q value (in scenario 9) in order to keep many transitions in the sequence. Considering the visualization, we set $T = 3000$ and do 500 simulation with different seeds. To visualize the results, we display the Figure 3.19 and find many fluctuations in magnitude in each series corresponding to each sample. Figure 3.20 compares the true θ_{lt} with $\hat{\theta}_{lt|T}$ of the same series for each sample. Even though most of the estimated parameters are very close to the true θ_{lt}, there are more significant difference between the true θ_{lt} and estimated $\hat{\theta}_{lt|T}$, since the change points are different across 10 samples and many transitions in this scenario.

Figure 3.21 shows the true and estimated $P(s_t = 1)$ of each series for 10 samples.

Figure 3.20: BCMIX estimates $\hat{\theta}_{lt|T}$ (dashed line) and true θ_{lt} (solid line) of the selected series for 10 samples in Scenarios 9 (from left to right and top to bottom).

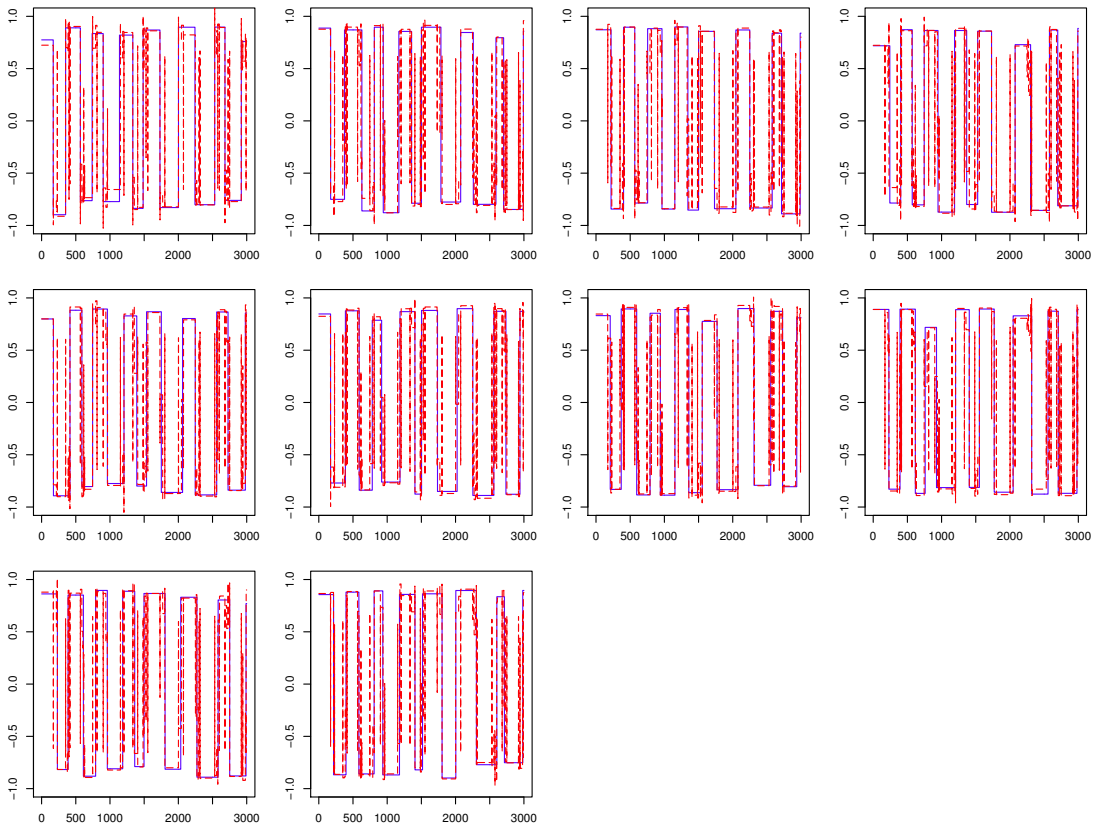
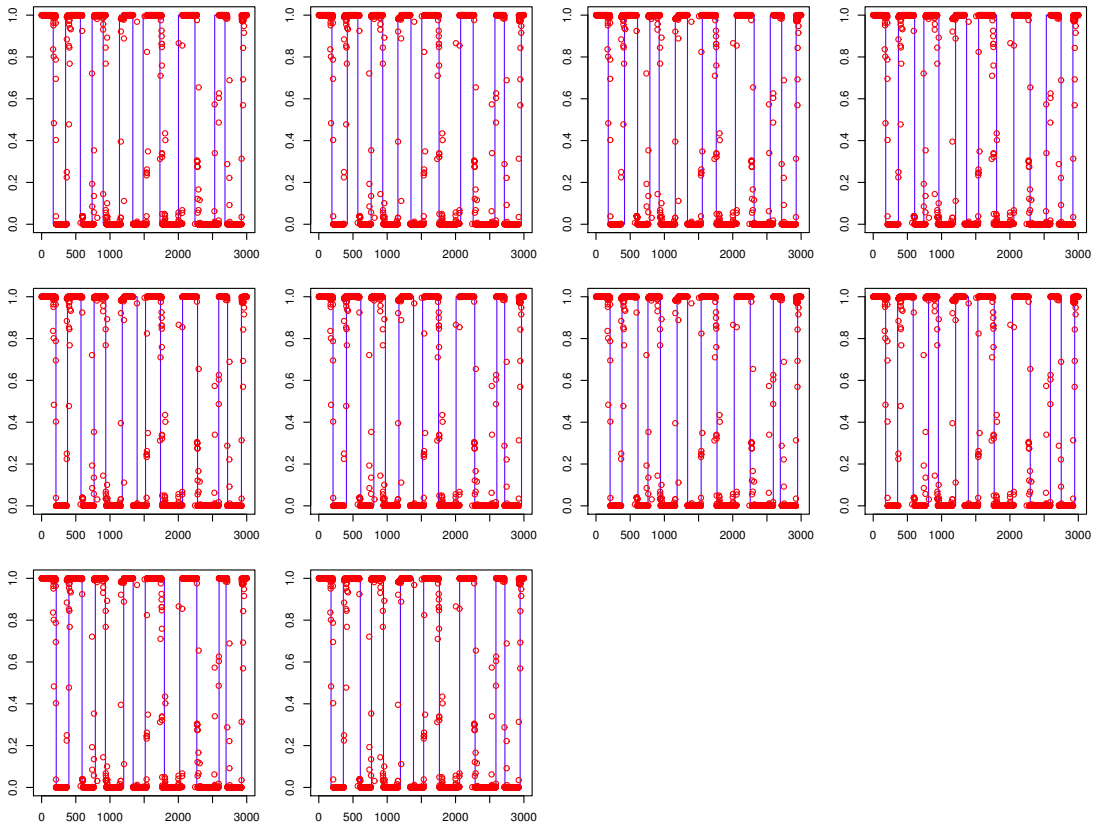


Figure 3.21: BCMIX estimates $\hat{r}_{t|T}^{(1)}$ (red points) and true $P(s_t = 1)$ (solid line) (top), $P(s_t = 2)$ (solid line) (middle), $P(s_t = 3)$ (solid line) (bottom) of the selected series for 10 samples in Scenario 9.



Specifically, if the true regime is 1, the true probability of $P(s_t = 1) = 1$; if the true regime is 2, the true probability of $P(s_t = 1) = 0$. There are two regimes in our simulation setup, hence $P(s_t = 2) = 1 - P(s_t = 1)$ for $1 \leq t \leq T$. So we only show the probability of regime 1. Slight differences between the estimated state probabilities and true state probabilities. Since there are more frequent transitions, as shown in Figure 3.21, the estimated probabilities show some fuzziness around transitions. We can see there are some red dots which represent the posterior state probability appears in the middle of 0 and 1. That is why we use $P(s_t = 1) > 0.5$ to make inference on the unknown regime. Therefore, even though this simulation is not satisfied with the assumptions of our model, it still works well for estimating parameter and state calling.

Chapter 4

Real Data Analysis

4.1 Data/Ovarian Cancer

In this section, we will apply the stochastic segmentation model to one real data set: copy number results for ovarian serous cystadenocarcinoma (OV) using Array based-CGH technology, CGH-1x1M_G4447A platform, archive type Level_2 (normalized signals for copy number alterations of aggregated regions, per probe or probe set), version 11.2.0 from Memorial Sloan-Kettering Cancer Center (MSKCC). The data files available for public use, can be downloaded from The Cancer Genome Atlas (TCGA). We will display some relative results of our model such as posterior mean and state probability.

This data is published on April 1st, 2010 in TCGA database. Fifteen OV cancer patients were selected. Three states model were used. State 1 represents amplification, state 2 is baseline, state 3 represents deletion.

Ovarian cancer is the fifth leading cause of cancer death in women. It is cancer that starts in the ovaries. The ovaries are the female reproductive organs that produce eggs. This cancer mainly develops in older women. About half of the women who are diagnosed with ovarian cancer are 63 years or older. It is more common in white women than African-

American women. The American Cancer Society estimates for ovarian cancer in the United States for 2013 are: About 22,240 women will receive a new diagnosis of ovarian cancer. About 14,230 women will die from ovarian cancer. Ovarian cancers display a high degree of complex genetic variations. The previous literature results show that the most frequently affected chromosomes in ovarian cancer are chromosome 1, chromosome 8 and chromosome 17. So we chose these three chromosomes to display the results. Analysis on chromosome 17 is more important and introduced first, then analysis on chromosome 1 and 8, since chromosome 17 includes more copy number alterations and well known mutant genes related to OV cancer.

4.2 Analysis on Chromosome 17

We chose 15 series from 15 OV cancer patients to visualize the estimation of posterior mean and state probability. There are 20009 probes on Chromosome 17. With $K = 3$ and hyperparameter estimate using EM algorithm, we give the hyperparameters some initial values as below: $(z^{(l,1)}, V^{(l,1)}) = (0.5, 0.01)$, $(z^{(l,2)}, V^{(l,2)}) = (0, 0.01)$, $(z^{(l,3)}, V^{(l,3)}) = (-0.6, 0.01)$ and $\sigma_l^2 = 0.1$, $1 \leq l \leq 15$. and $P = \begin{pmatrix} 0.98 & 0.01 & 0.01 \\ 0.01 & 0.98 & 0.01 \\ 0.01 & 0.01 & 0.98 \end{pmatrix}$.

Table 4.1 shows the estimated hyperparameters by the EM algorithm . The corresponding estimated transition probability matrices are showed in Table 4.2.

Figure 4.1 displays the observations along the probes for 15 patients from left to right and top to bottom shown by grey lines. Red line displays the posterior estimation of mean for each patient. Figure 4.2 displays the posterior state probability (Red: Amplification; Blue: Baseline; Green: Deletion). We can find that deletions occur in the first half, and amplifications occur in the second half. Most of 15 observations from Fig 4.1 show the similar

Table 4.1: Hyperparameter estimate using EM algorithm for 15 samples.

	z_1	z_2	z_3	V_1	V_2	V_3	σ^2
$sample1$	0.1072	0.0037	-0.6472	0.1601	0.0484	0.2239	0.0394
$sample2$	0.1145	0.0334	-0.5832	0.1517	0.0246	0.1679	0.0613
$sample3$	0.3780	0.2098	-0.6881	0.1094	0.0821	0.3018	0.1006
$sample4$	0.6321	0.4072	-0.5638	0.0184	0.1868	0.2079	0.0678
$sample5$	0.2339	0.2018	-0.6702	0.1022	0.1238	0.2113	0.0727
$sample6$	0.3039	0.0796	-0.5852	0.0910	0.0442	0.1412	0.0762
$sample7$	0.3201	0.0189	-0.6737	0.0821	0.0595	0.2490	0.0559
$sample8$	0.1726	0.0063	-0.4995	0.1309	0.0537	0.1090	0.0497
$sample9$	0.1297	0.0130	-0.5551	0.1414	0.0252	0.0479	0.0424
$sample10$	0.8393	0.2749	-0.6858	0.1278	0.2673	0.1875	0.0648
$sample11$	0.0783	-0.0399	-0.5553	0.1844	0.0519	0.0497	0.0459
$sample12$	0.4209	-0.0013	-0.5415	0.0431	0.0412	0.0639	0.1452
$sample13$	0.1233	-0.0095	-0.5610	0.1439	0.0687	0.0882	0.0669
$sample14$	0.1179	-0.0257	-0.4885	0.1654	0.0611	0.0864	0.0616
$sample15$	0.2935	-0.2219	-0.7257	0.0950	0.1244	0.2608	0.0671

Table 4.2: Estimated transition probabilities.

	State 1	State 2	State 3
State 1	0.7433	0.2539	0.0028
State 2	0.0893	0.8874	0.0233
State 3	0.1198	0.2374	0.6428

pattern. From these two figures, we can conclude that our model has good performance on detecting transitions and generate reasonable state call. Although the posterior state probabilities here have many fuzziness, it can improve state call using a cut line $p = 0.5$. So, the above results show that our stochastic segmentation model can successfully capture recurrent aberrations across 15 OV cancer patients.

The most common aberrations for serous histology of OV cancer are deletions of 17p (Dimova et al., 2009; Mankoo et al., 2011; Zhang et al., 2013). Engler utilized TCGA dataset and found that most of deletions of chr17 map to 17p11.2 (Chr17:17646236 - 21720090) and 17p12(Chr17:10689461 - 16833125)(Engler et al., 2012). The recent research (Zhang et al., 2013) found that deletions in OV cancer was in the regions containing BRCA1, TP53 mutations. In general, a deletion indicates the presence of a tumor suppressor gene, and an amplification indicates the presence of at least one oncogenes. BRCA1 and TP53 are well-known tumor suppressor genes. Mutation in BRCA1 has been associated with higher risks of OV cancer, and they are hereditary, so the mutation and the cancer risk can occur in families. Mutations of the TP53 gene are the most common and most frequently studied molecular alterations in human cancer. Several earlier studies have suggested that TP53 plays a role in serous OV cancer.

Our result as shown in Figure 4.3 indicates that the recurrent copy number deletions involved 17p11.2, 17p12 which is consistent with the above studies using existing statistical models. For example, Engler used GISTIC analysis to identify amplifications and deletions based on segmented copy number data generated by CBS algorithm (Engler et at., 2012). We also detected well known tumor suppressor genes TP53(17p13.1), BRCA1(17q21), and transcription factors RAI1 (17p11.2),SREBF1 (17p11.2). There are totally 136 unique known genes involving in recurrent copy number deletion regions.

Pathway enrichment analysis using these 136 known genes based on IPA database was

Figure 4.1: The observationbs (grey line) and the posterior estimation of mean for 15 samples (from left to right and top to bottom).

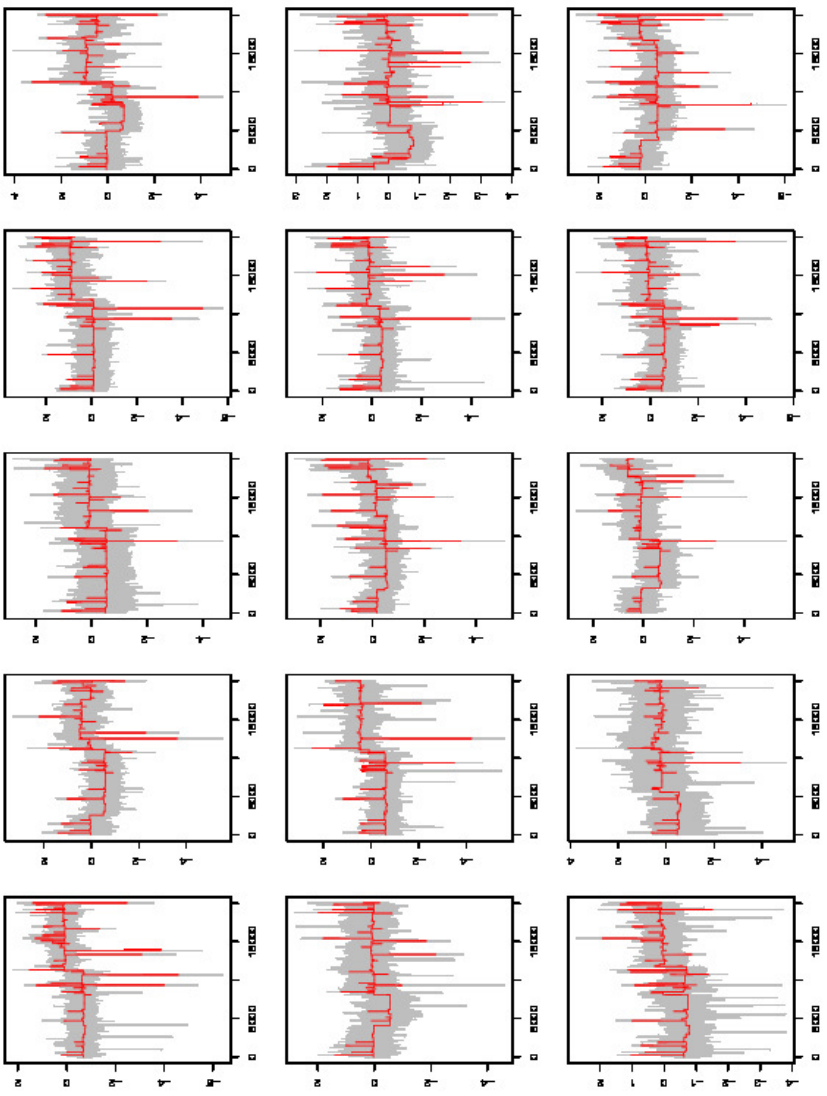


Figure 4.2: The posterior estimation of state probability for 15 samples $P(amplification)$ (top), $P(baseline)$ (middle), $P(deletion)$ (bottom).

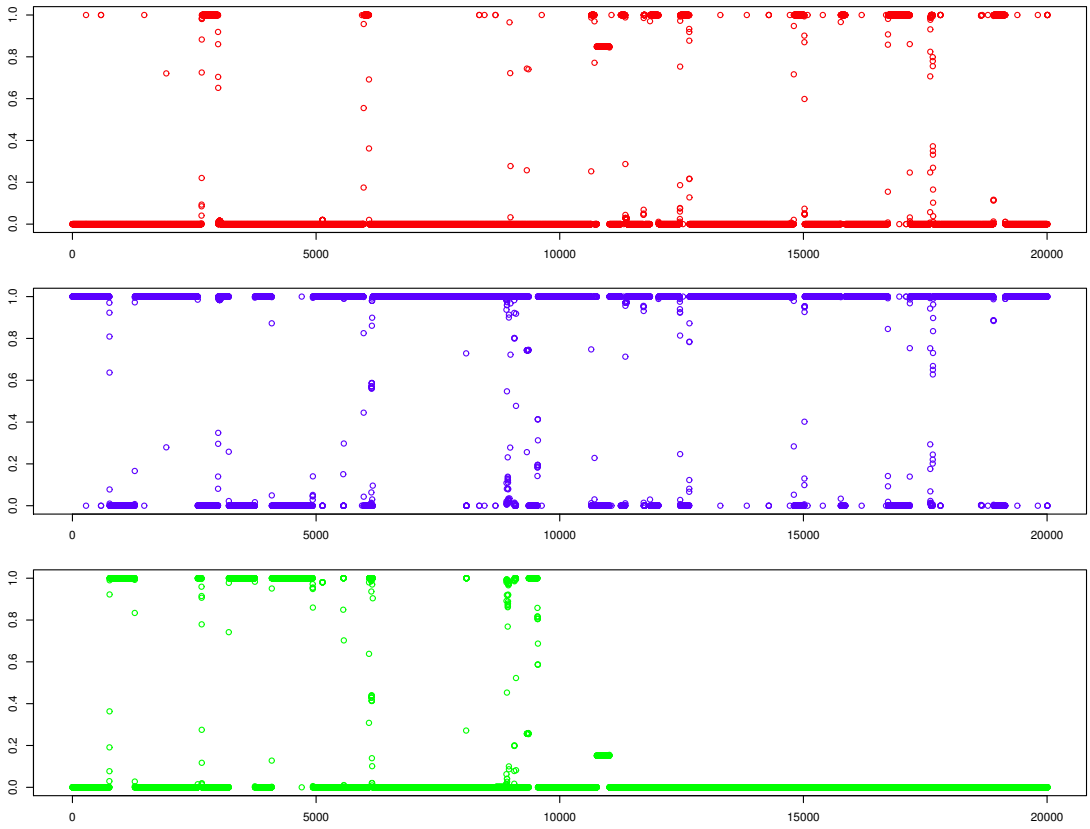


Figure 4.3: The posterior estimation of deletion probability for 15 samples with marked recurrent deletions.

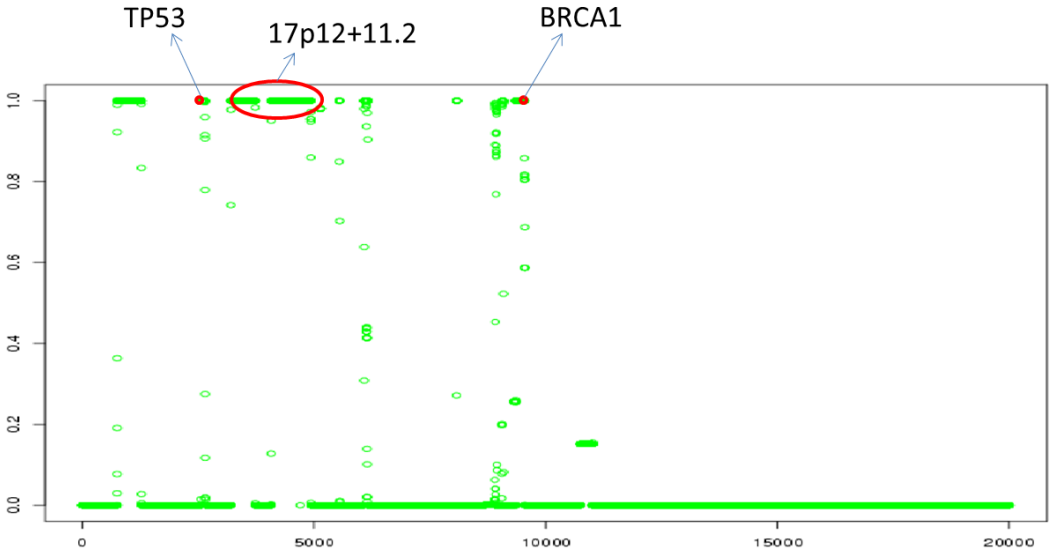
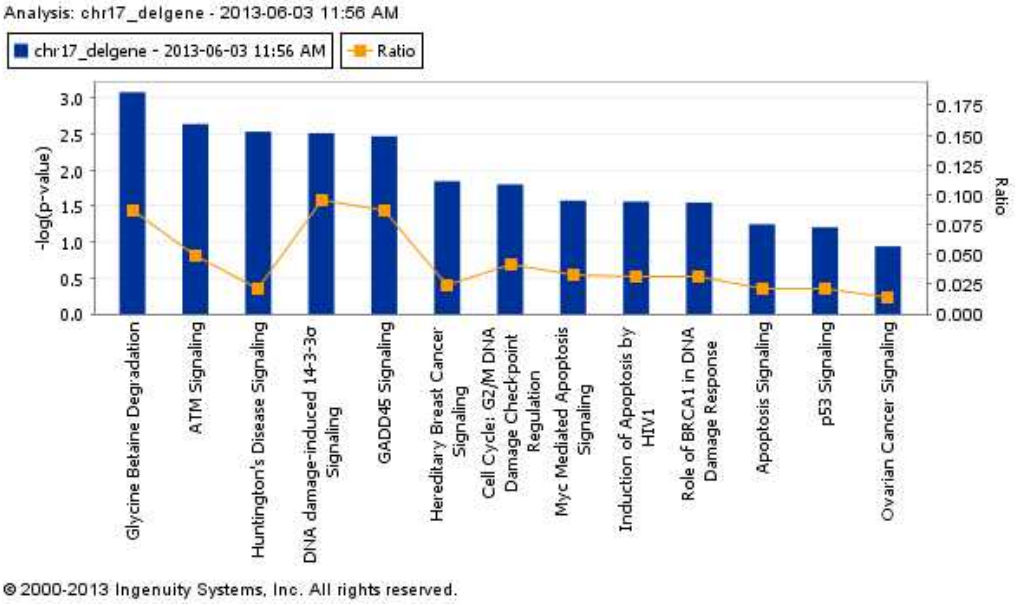


Figure 4.4: Canonical pathway analysis of genes from recurrent regions of copy number variants.



carried out in this study. Significantly enriched pathways with Fishers exact p-value less than 0.05 are listed in this bar plot as shown in Figure 4.4. Although only one of the chromosomes was analyzed, this already reveals some biological mechanisms and pathway changes involved in ovarian cancer. First obviously, the ovarian cancer signaling pathway was found enriched. Particularly, the GADD45 and p53 signaling pathways are enriched. Both these two factors, especially p53, are well established tumor suppressor proteins. More importantly, almost half of the pathways are basic and critical cellular processes such as DNA repair, cell cycle regulation and apoptosis. Changes in these pathways indicate severe disruptions of normal cellular functions. This could either be causing the cancer or be the result of cancer.

4.3 Analysis on Chromosome 1

We choose 15 series from the same 15 OV cancer patients and apply the same method from previous section to estimate posterior mean and state probability. There are 55274 probes on Chromosome 1. We set $K = 3$ and use EM to estimate hyperparameter estimate.Some initial values of hyperparameters are given as below: $(z^{(l,1)}, V^{(l,1)}) = (0.55, 0.01)$, $(z^{(l,2)}, V^{(l,2)}) = (0, 0.01)$, $(z^{(l,3)}, V^{(l,3)}) = (-0.55, 0.01)$ and $\sigma_l^2 = 0.1$, $1 \leq l \leq 15$. and $P = \begin{pmatrix} 0.98 & 0.01 & 0.01 \\ 0.01 & 0.98 & 0.01 \\ 0.01 & 0.01 & 0.98 \end{pmatrix}$.

Table 4.3 shows the estimated hyperparameters by the EM algorithm. Table 4.4 shows the corresponding estimated transition probability matrices.

Figure 4.5 displays the observations along the probes for 15 patients shown by grey lines. The posterior estimation of mean for each patient is represented by red line. Figure 4.6 displays the posterior state probability. From these two figures, we can say that our model has good performance on smoothing the signal and can successfully detect recurrent

Table 4.3: Hyperparameter estimate using EM algorithm for 15 samples.

	z_1	z_2	z_3	V_1	V_2	V_3	σ^2
$sample1$	0.1308	0.0187	-0.6950	0.1835	0.0389	0.0568	0.0393
$sample2$	0.1340	0.0431	-0.4722	0.1776	0.0205	0.0748	0.0612
$sample3$	0.4457	0.2161	-0.0650	0.0549	0.0957	0.3443	0.1008
$sample4$	0.6304	0.4280	0.1577	0.0073	0.2063	0.6220	0.0678
$sample5$	0.2843	0.2027	-0.2837	0.1067	0.1172	0.1261	0.0726
$sample6$	0.2550	0.1059	-0.0310	0.1189	0.0599	0.3434	0.0760
$sample7$	0.3652	0.0399	-0.2272	0.0791	0.0629	0.2406	0.0559
$sample8$	0.2193	0.0199	-0.4590	0.1284	0.0500	0.0233	0.0496
$sample9$	0.1310	0.0252	-0.4042	0.1827	0.0229	0.0314	0.0423
$sample10$	0.8470	0.3270	-0.1480	0.1029	0.3071	0.2362	0.0649
$sample11$	0.0873	-0.0235	-0.6472	0.2244	0.0423	0.0216	0.0457
$sample12$	0.4159	0.0384	-0.2024	0.0565	0.0576	0.1507	0.1449
$sample13$	0.1379	0.0087	-0.6267	0.1732	0.0589	0.0216	0.0669
$sample14$	0.1574	-0.0202	-0.0385	0.1797	0.0569	0.3181	0.0615
$sample15$	0.3355	-0.1802	-0.4393	0.1136	0.1200	0.0618	0.0671

Table 4.4: Estimated transition probabilities.

	State 1	State 2	State 3
State 1	0.6639	0.3332	0.0029
State 2	0.0907	0.8859	0.0234
State 3	0.0158	0.3477	0.6365

aberrations cross 15 OV cancer patients.

Dimova found that amplifications of 1p12 (Chr1:117350000 - 118700000) and 1q23.2 (Chr1:157700000 - 158500000), the deletions of 1p36.21 and 1p36.1, were associated with ovarian cancer (Dimova et al., 2009). Engler found that most of amplifications of chr1 map to 1p34.2 (Chr1:39685801 - 40370914) including the oncogene MYCL1 and 1q42.3(Chr1:232669917 - 234247146), most of deletions map to 1p36.33 (Chr1:823965 - 2511264) containing TP73 mutations (Engler et al., 2012).

As a result, Figure 4.7 shows the recurrent copy number amplifications involved 1p34.2,

1p12, 1q23.2 and 1q42.3, deletions involved 1p36.33, 1p36.21 and 1p36.13 which are consistent with these above earlier studies. We also detected well known oncogene MYCL1 (1p34.2) and tumor suppressor genes TP73 (1p36.33). There are totally 178 unique known genes involving in recurrent regions of copy number variants.

Pathway enrichment analysis using these 178 known genes based on IPA database was used to identify biological themes. Significantly enriched pathways with Fishers exact p-value less than 0.05 are listed in this bar plot as shown in Figure 4.8. Most of the pathways are related to cancer. First obviously, the breast cancer signaling and apoptosis signaling pathways were found enriched. Defects in apoptosis signaling contribute to resistance of tumors (Schulze-Bergkamen et al., 2004). NRF2-mediated oxidative stress response is the most significantly regulated pathway, which has been related to breast cancer (Seng et al., 2007; Liu et al., 2010). Ubiquitination regulates degradation of cellular proteins by the ubiquitin proteasome system, controlling a proteins half-life and expression levels. A change of ubiquitination activity is associated with ovarian tumorigenesis, so the protein ubiquitination pathway might be involved in breast ovarian progression. Notch signaling plays the paradoxical role in numerous human cancers including ovarian cancer (Rose, 2009). Notch appears to act as both an oncogene and a tumor suppressor gene depending on the cellular context.

Figure 4.5: The observationbs (grey line) and the posterior estimation of mean for 15 samples (from left to right and top to bottom).

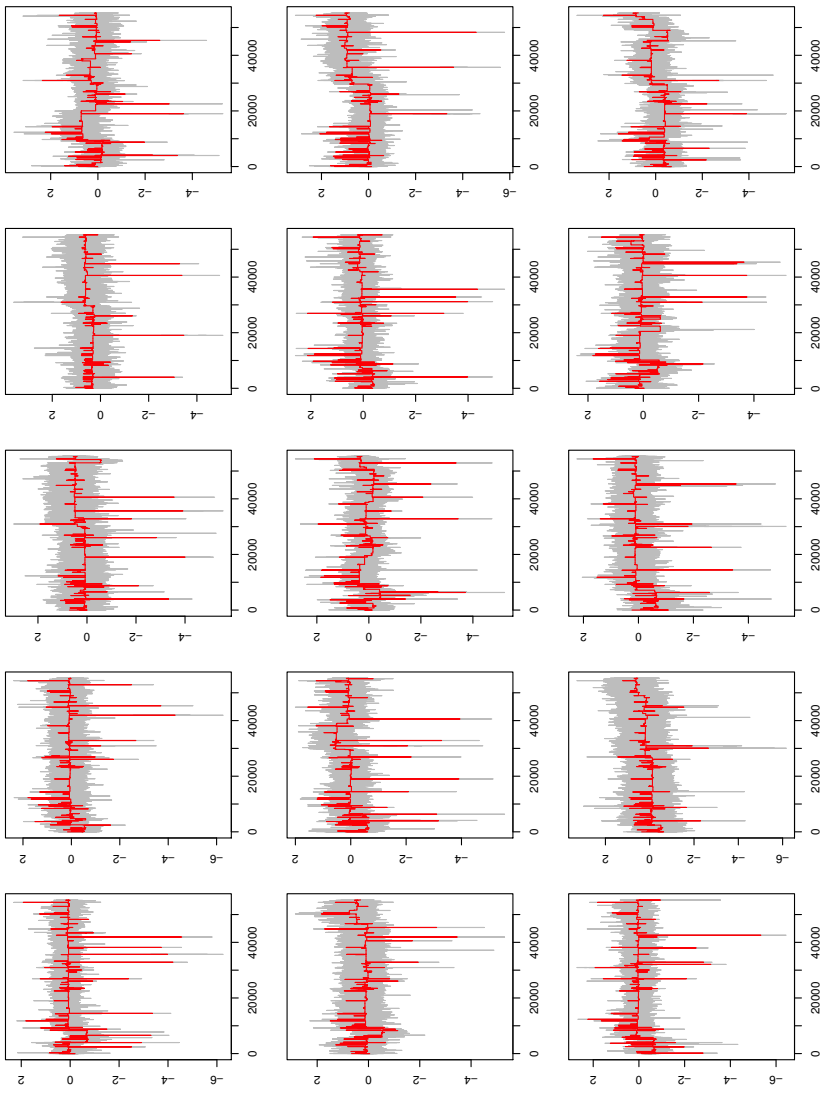


Figure 4.6: The posterior estimation of state probability for 15 samples $P(amplification)$ (top), $P(baseline)$ (middle), $P(deletion)$ (bottom).

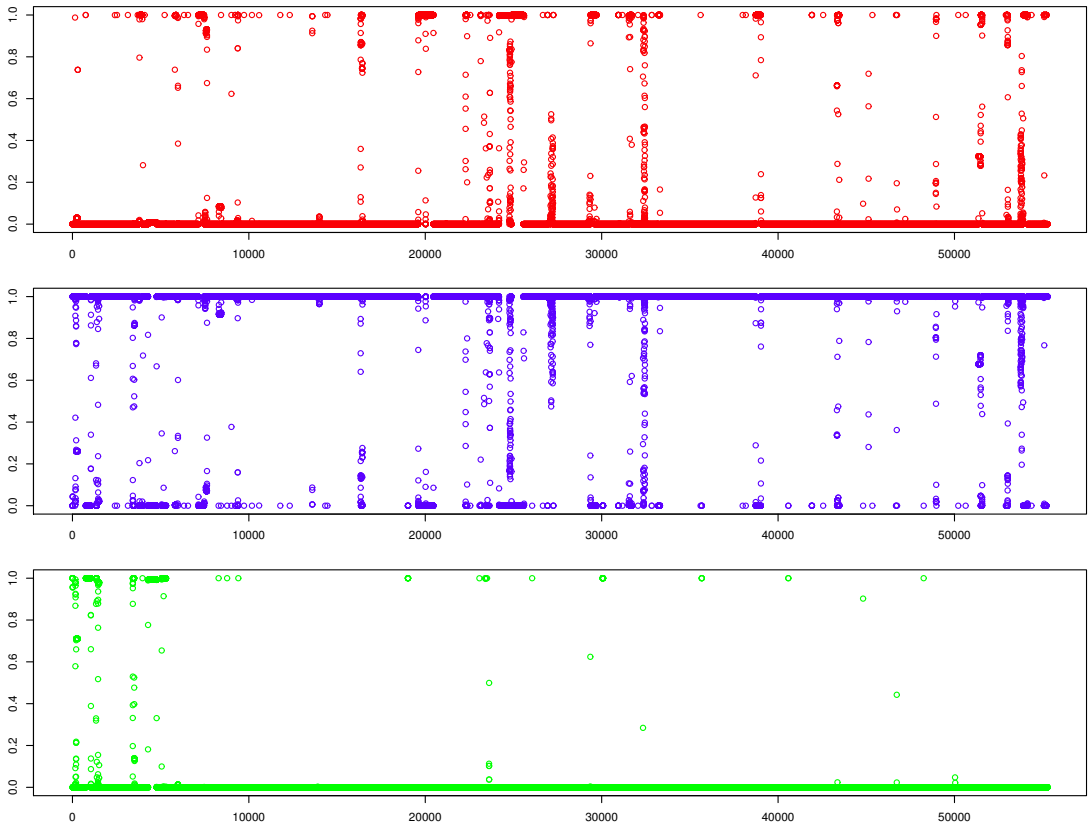


Figure 4.7: The posterior estimation of state probability for 15 samples with marked recurrent regions of copy number variants.

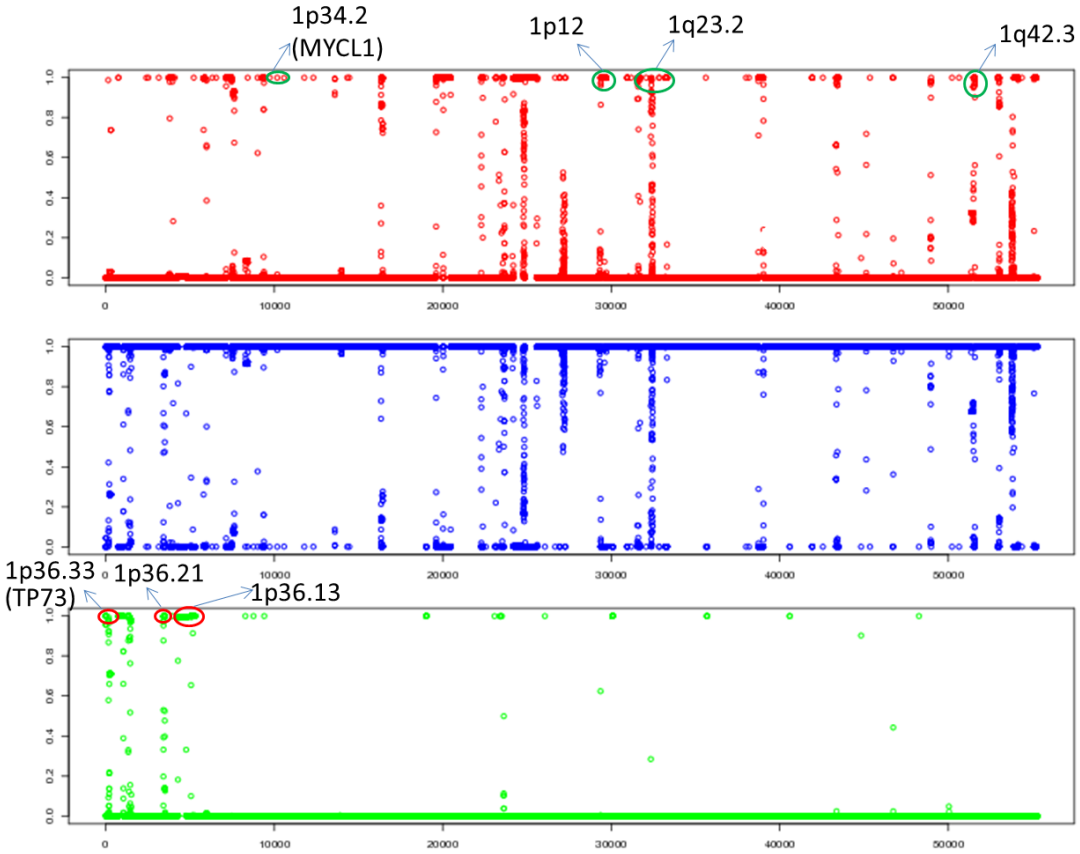
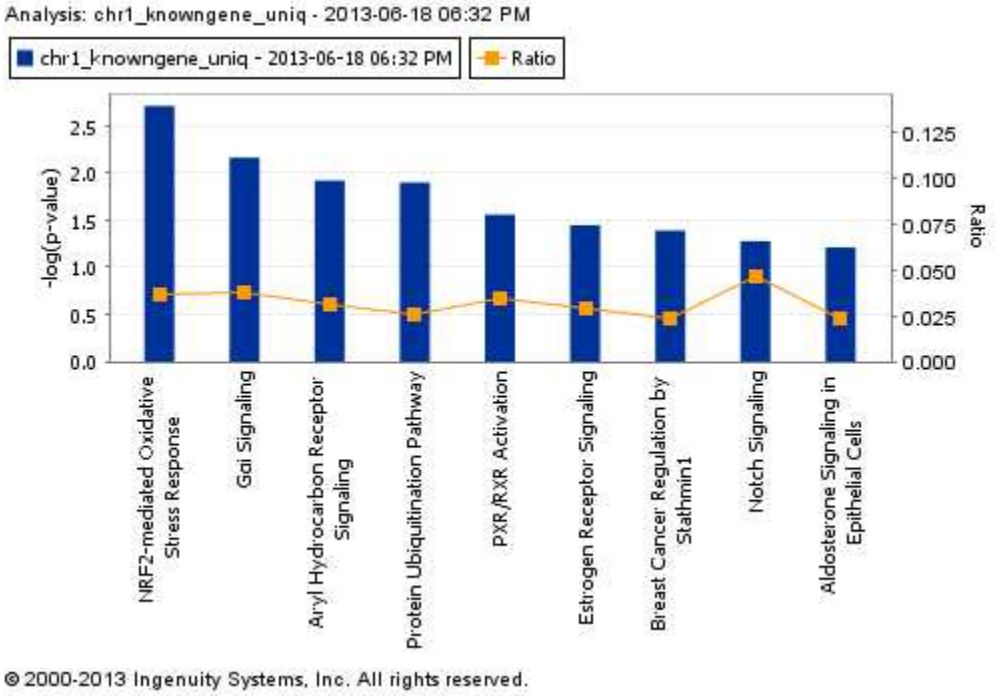


Figure 4.8: Canonical pathway analysis of genes from recurrent regions of copy number variants.



4.4 Analysis on Chromosome 8

On chromosome 8, Same subjects and methods are applied to estimate posterior mean and state probability. There are 30473 probes on Chromosome 8. We use three states in our model, $K = 3$, and hyperparameter estimate using EM algorithm, we give the hyperparameters some initial values as below: $(z^{(l,1)}, V^{(l,1)}) = (0.75, 0.0265)$, $(z^{(l,2)}, V^{(l,2)}) = (0, 0.02)$, $(z^{(l,3)}, V^{(l,3)}) = (-0.7, 0.02)$ and $\sigma_l^2 = 0.1$, $1 \leq l \leq 15$. and $P = \begin{pmatrix} 0.98 & 0.01 & 0.01 \\ 0.01 & 0.98 & 0.01 \\ 0.01 & 0.01 & 0.98 \end{pmatrix}$.

Table 4.5 shows the estimated hyperparameters by the EM algorithm . The corresponding estimated transition probability matrices are showed in Table 4.6. Figure 4.9 displays the observations along the probes for 15 patients (grey lines) and the posterior estimation of mean for each patients (red lines). Figure 4.10 displays the posterior state probability.

Table 4.5: Hyperparameter estimate using EM algorithm for 15 samples.

	z_1	z_2	z_3	V_1	V_2	V_3	σ^2
$sample1$	0.2603	-0.1956	-0.7526	0.3020	0.2455	0.0053	0.0380
$sample2$	0.0980	0.0907	0.0394	0.4276	0.0092	0.5640	0.0606
$sample3$	0.0922	0.0669	0.0053	0.4353	0.0055	0.5150	0.0973
$sample4$	0.5129	0.4127	0.2519	0.2017	0.2061	0.9331	0.0690
$sample5$	0.2324	-0.1269	-0.3572	0.3009	0.1308	0.1967	0.0719
$sample6$	0.3069	-0.1085	-0.6065	0.2480	0.1472	0.0134	0.0769
$sample7$	0.9516	0.1449	-0.6286	0.1210	0.4612	0.0103	0.0580
$sample8$	0.3308	0.0752	-0.3229	0.2015	0.0845	0.1466	0.0460
$sample9$	0.1058	-0.0123	-0.3466	0.4234	0.0465	0.1630	0.0408
$sample10$	0.4248	0.2157	-0.1009	0.1100	0.0922	0.3737	0.0595
$sample11$	0.0508	0.0160	-0.0178	0.4957	0.0013	0.4793	0.0422
$sample12$	0.9665	0.4874	-0.5248	0.0520	0.6475	0.0362	0.1568
$sample13$	0.5788	-0.5090	-2.2307	0.9309	0.9882	3.3140	0.0823
$sample14$	1.0638	0.4795	0.0162	0.1624	0.4680	0.5264	0.0643
$sample15$	0.4003	-0.0427	-0.9629	0.1639	0.1556	1.0574	0.0671

Dimova found that amplifications of 8q13.2 (Chr8:69490000 - 70200000), the deletions of 8p23.1 and 8p21.2, were associated with ovarian cancer (Dimova et al., 2009). Engler utilized TCGA dataset and found that most of amplifications of chr8 map to 8q24.21 (Chr8:128870582 - 129868380) containing oncogene MYC, most of deletions map to 8p23.2 (Chr8: 1422246- 3652163) (Engler et al., 2012).

From Figure 4.11, we can see our results show that the recurrent copy number amplifications involved 8q13.2 and 8q24.21, deletions involved 8p23.2, 8p23.1 and 8p21.2, which are consistent with the above studies. We also detected well known oncogene MYC (8q24.21). There are totally 289 unique known genes involving in recurrent copy number amplification and deletion regions.

We implemented pathway enrichment analysis using these 289 known genes based on IPA database. Figure 4.12 shows significantly enriched pathways with Fishers exact p-value less than 0.05. As we know, these three G-protein coupled receptor related signaling pathways:

Table 4.6: Estimated transition probabilities.

	State 1	State 2	State 3
State 1	0.6761	0.3230	0.0009
State 2	0.1139	0.8060	0.0801
State 3	0.0057	0.2181	0.7762

Figure 4.9: The observationbs (grey line) and the posterior estimation of mean for 15 samples (from left to right and top to bottom).

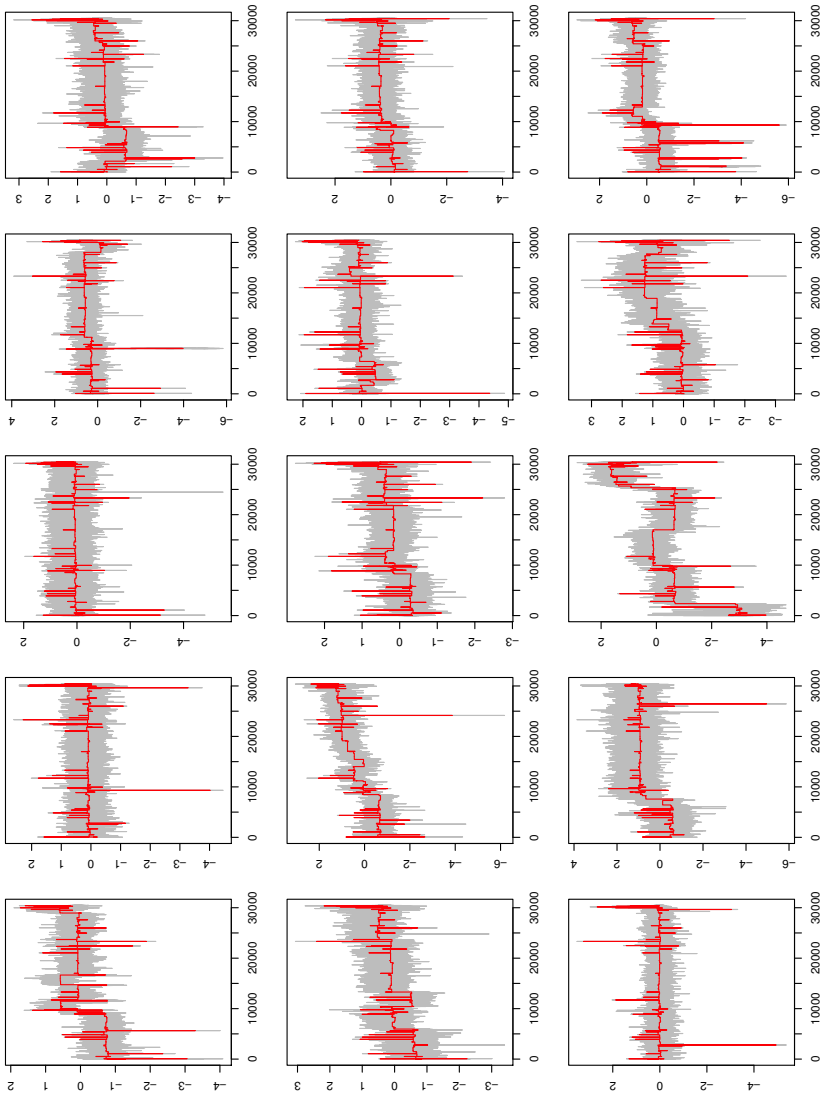


Figure 4.10: The posterior estimation of state probability for 15 samples $P(amplification)$ (top), $P(baseline)$ (middle), $P(deletion)$ (bottom).

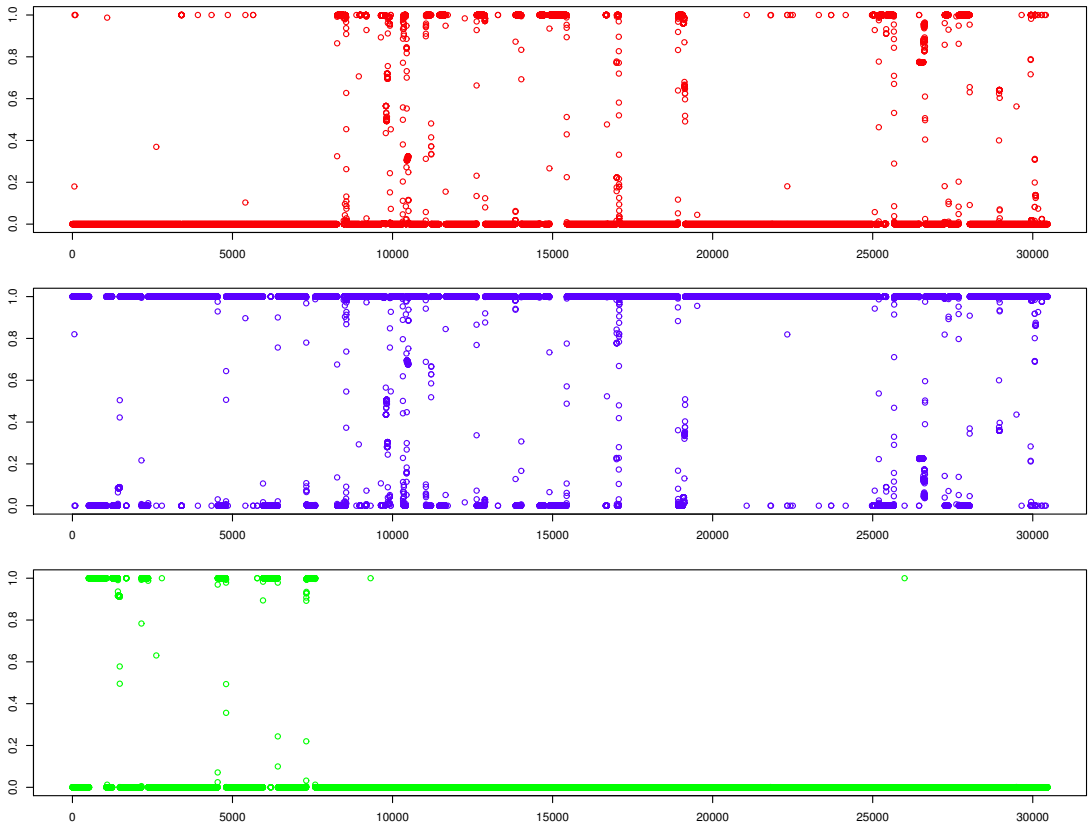


Figure 4.11: The posterior estimation of state probability for 15 samples with marked recurrent regions of copy number variants

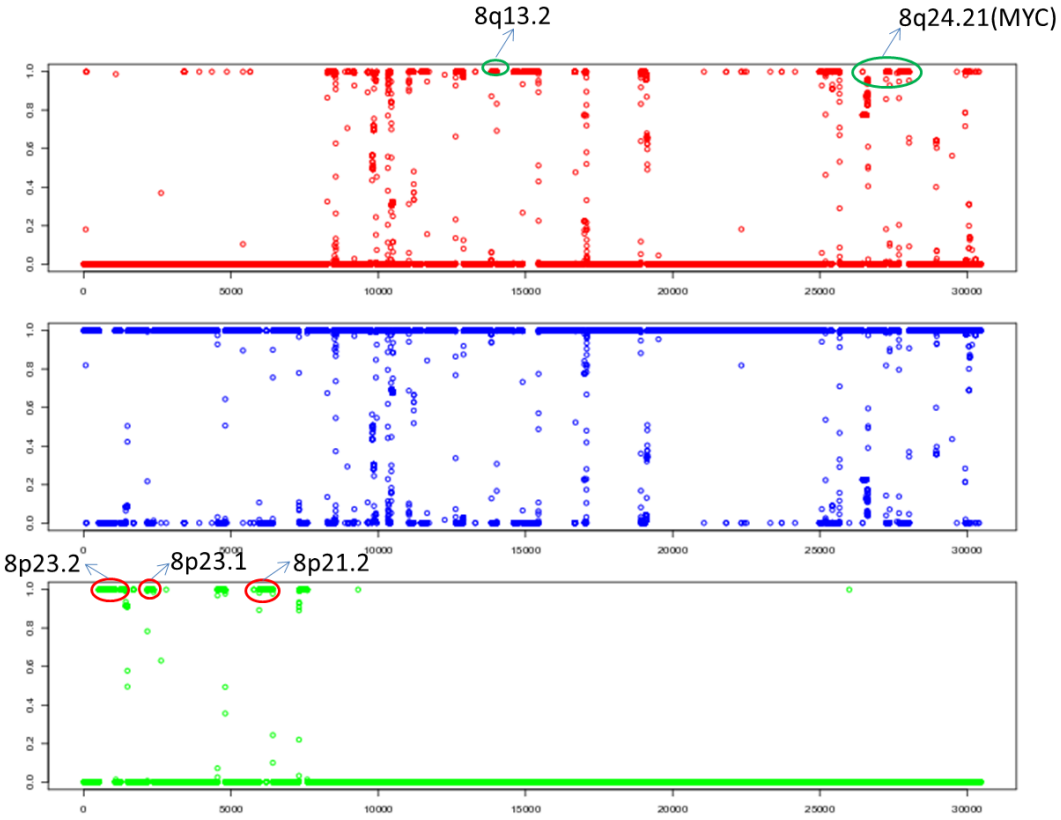
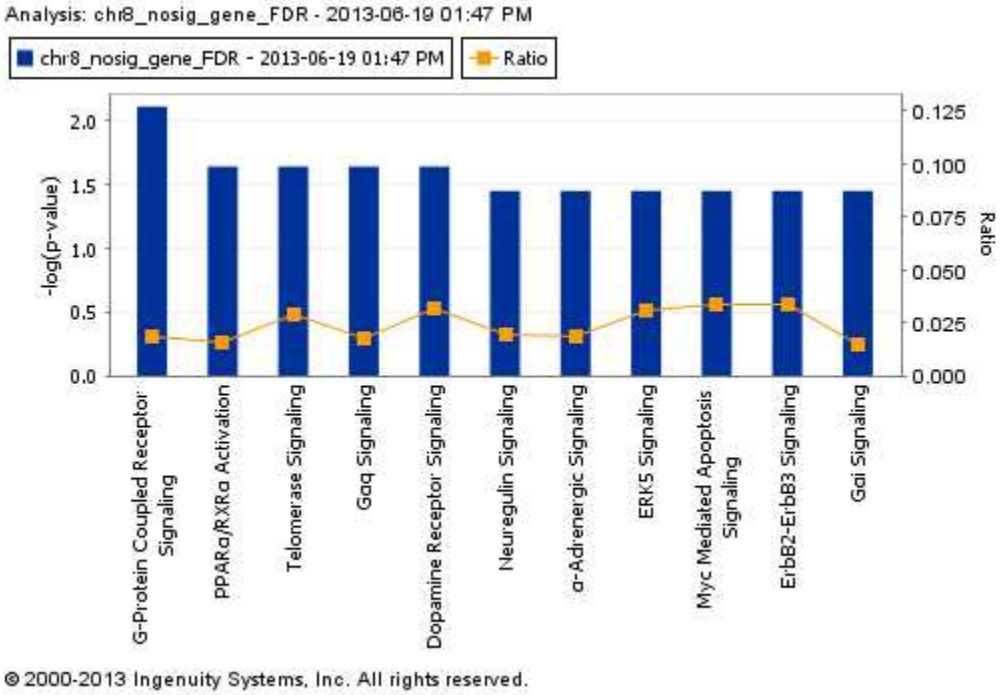


Figure 4.12: Canonical pathway analysis of genes from recurrent regions of copy number variants.



the G-protein coupled receptor signaling, Gaq signaling and Gai signaling pathways are very important molecular pathways. Aberrant signaling through G-protein coupled receptors promotes metastasis which is the major cause of breast cancer death. Particularly, the MYC signaling pathway was found enriched. MYC is a key regulator of cell growth, proliferation, metabolism, differentiation, and apoptosis. MYC deregulation contributes to many cancers development and progression. Additionally, previous studies demonstrate that ERK5 signaling is involved in breast and prostate cancer proliferation and tumorigenesis (Zhou et al., 2008), and the ErbB2/ErbB3 dimer functions as an oncogenic unit to drive tumor cell proliferation (Holbro et al., 2003).

Chapter 5

Conclusions and Discussions

For the analysis of grouped array-CGH data, we proposed a class of stochastic segmentation models and an associated inference framework that has attractive statistical and computational properties. The stochastic segmentation model in Chapter 2 assumes that $y_{lt} = \theta_{lt} + \epsilon_{lt}$ for $l = 1, \ldots, J$ and $t = 1, \ldots, T$, where ϵ_{lt} are independent normal random variables with mean 0 and variance σ_l^2, and θ_{lt} is an unknown step function whose prior distribution depends on a finite state hidden Markov chain s_t. After the hidden state shifts from one regime to another regime, the model parameters jump to another set of values, which are generated by regime-dependent prior distributions and hence are not necessarily same as those within the same regime during the past.

A forward filtering procedure shows the posterior distribution of the parameter as a mixture distribution with explicit weights which can be calculated recursively. Furthermore, based on the reversibility of the hidden Markov chain, a backward filtering procedure can be conducted in a similar way. Based on Bayes' theorem, both the smoothing estimate of parameter and probability of regimes can be calculated explicitly to save a time-consuming numerical filtering procedure. The hyperparameters in the model can be estimated by the Expectation-Maximum (EM) algorithm. Furthermore, a Bounded Complexity Mix-

ture (BCMIX) Approximation is shown to have much lower computational complexity yet comparable to the Bayes estimates in statistical efficiency. Simulation studies evaluate the fBayes and BCMIX estimates in terms of the sum of squared errors (SSE). Moreover, the accuracy of identifying the transitions is evaluated by an Identification Ratio (IR). In order to access the advantage of our model, we compare our model to an existing statistical model in terms of IR. The result of comparison illustrates that our model is more accurate and effective than that model in identifying transitions. At the end of this thesis, we apply this model to the real grouped array-CGH data set to detect recurrent copy number alterations.

An important benefit of our Bayesian model is that we can derive analytical filtering and explicit smoothing formulas for the posterior distributions of model parameters and make inference on regimes. The BCMIX estimate has much lower computational complexity yet comparable to the Bayes estimate in statistical efficiency. Furthermore, our model not only can handle the simultaneous change model, but also work well with non-simultaneous model, which is more close to what happens in real world.

As mentioned in the introduction section, for detecting recurrent CNAs, many methods based on two-step procedure with the first step being segmentation for each sample, the second step being finding recurrent across multiple samples, have been applied. It is important to note that, the two-step procedure may miss information across the samples. This motivates us to consider a joint model which analyzes the recurrent events in one sweep. Besides, the noise information for recurrent events across the samples can be averaged out in our joint model. Analyzing each sample separately may strengthen or weaken some information that might be very important for recurrent event.

As for improving our model, it means to improve the power of the model that identifies genes from recurrent events. This may depend on different types of cancers. so we need incorporate the detailed structure in different cancers.

Reference

Beck JB, Schmuths H, Schaal BA . Global population genetics of Arabidopsis thaliana. *Molec Ecol* in press (2007).

Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel PS, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences* 0710052104+ (2007).

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899-905 (2010).

Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, Futreal PA, Weber B, Shapero MH, Wooster R. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Research* 14, 287295 (2004).

Box GE. and Tiao GC. *Bayesian Inference in Statistical Analysis*. Addison-Wesley (1973).

de Leeuw RJ, Davies JJ, Rosenwald A, Bebb G, Gascoyne RD, Dyer MJ, Staudt LM, Martinez-Climent JA, Lam WL. Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Human Molecular Genetics* 13, 1827-1837 (2004).

Dimova I, Orsetti B, Negre V, Rouge C, Ursule L, Lasorsa L, Dimitrov R, Doganov N, Toncheva D, Theillet C. Genomic markers for ovarian cancer at chromosomes 1, 8 and 17 revealed by array CGH analysis. *Tumori* 95, 357-366 (2009).

Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T, Stoeckert CJ Jr, Weber BL, Maris JM, Grant GR. STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Research* 16, 11491158 (2006).

Engler DA, Gupta S, Growdon WB, Drapkin RI, Nitta M, Sergent PA, Allred SF, Gross J, Deavers MT, Kuo WL. Genome wide DNA copy number analysis of serous type ovarian carcinomas identifies genetic markers predictive of clinical outcome. *PLoS One* 7, e30996 (2012).

Ewald van Dyk, Marcel J.T. Reinders and Lodewyk F.A. Wessels. A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. *Nucleic Acids Research* 41,e100 (2013).

Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nature Reviews Genetics* 7, 8597 (2006).

Fridlyand J, Snijders A, Pinkel D, Albertson DG, Jain AN. Application of hidden Markov models to the analysis of the array-CGH data. *Journal of Multivariate Analysis* 90, 132-153 (2004).

Garnis C, Lockwood WW, Vucic E, Ge Y, Girard L, Minna JD, Gazdar AF, Lam S, MacAulay C, Lam WL. High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH.*International Journal of Cancer* 118, 1556-1564 (2006).

Guttman M, Mies C, Dudycz-Sulicz K, Diskin SJ, Baldwin DA, Stoeckert CJ Jr, Grant GR. Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics* 3, e143+ (2007).

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nature reviews. Genetics* 10, 551564.(2009).

Goecks J, Nekrutenko A, Taylor J; Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11, R86 (2010).

Holbro T, Beerli RR, Maurer F, Koziczak M, Barbas CF 3rd, Hynes NE. The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation. *Proceeding of the National Academy of Sciences of the United States of America* 100, 89338938 (2003).

Hup P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. regions. *Bioinformatics* 20, 34133422 (2004).

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam WL. A tiling resolution DNA microarray with complete coverage of the human genome. *Nature Genetics* 36, 299303 (2004).

Klijn C, Holstege H, de Ridder J, Liu X, Reinders M, Jonkers J, Wessels L. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Research* 36, e13 (2008).

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. PairedEnd Mapping Reveals Extensive Structural Variation in the Human Genome. *Science* 318, 420426 (2007).

Lai TL, Xing H, Zhang N. Stochastic segmentation models for array-based comparative genomic hybridization data analysis.*Biostatistics* 9, 290-307 (2008).

Lai TL and Xing H. A simple Bayesian approach to multiple change-points. *Statistica Sinica*, 21, 539-569 (2011).

Lai WRR, Johnson MDD, Kucherlapati R, Park PJJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21, 37633770 (2005).

Lee C, Iafrate AJ, Brothman AR. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nature Genetics* 39, S48S54 (2007).

Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhini Z. Efficient calculation of interval scores for dna copy number data analysis. *Journal of Computational Biology* 13, 215228 (2006).

Liu Q, Zhang H, Smeester L, Zou F, Kesic M, Jaspers I, Pi J, Fry RC. The NRF2-mediated oxidative stress response pathway is associated with tumor cell resistance to arsenic trioxide across the NCI-60 panel. *BMC Medical Genomics* 3:37, (2010).

Lupski JR. An evolution revolution provides further revelation. *Bioessays* 29, 1182-1184 (2007).

MacDonald JR, Ziman R, Yuen RKC, Feuk L, and Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research* 17 (2013).

Mankoo P, Shen R, Schultz N, Levine D, Sander C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One* 6, e24709 (2011).

Morganella S, Pagnotta SM, Ceccarelli M. Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics* 27, 2949-2956 (2011).

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL. High- resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research* 16, 11361148 (2006).

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High reso- lution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* 20, 207211 (1998).

Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* 37, 1117 (2005).

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO. Genome-wide analysis of DNA copy-number changes using cDNA microarrays.. *Nature Genetics* 23, 4146 (1999).

Pollack JR, Srlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Brresen-Dale AL, Brown PO. Microarray analysis reveals a major direct role of DNA copy number alternation in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences* 99, 1296312968 (2002).

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzlez JR, Gratacs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy num- ber in the human genome. *Nature* 444, 444454 (2006).

Rouveirol C, Stransky N, Hup P, Rosa PL, Viara E, Barillot E, Radvanyi F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics* 22, 849856 (2006).

Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 11, 3146 (2010).

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie

X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553560 (2007).

Rose SL. Notch signaling pathway in ovarian cancer. *International Journal of Gynecological Cancer* 19, 564-566 (2009)

Rueda OM, DiazUriarte R. Flexible and Accurate Detection of Genomic CopyNumber Changes from aCGH. *PLoS Computional Biology* 3, e122 (2007).

Rueda OM, Diaz-Uriarte R. Finding Recurrent Copy Number Alteration Regions: A Review of Methods. *Current Bioinformatics* 5, 1-17 (2010).

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L. Challenges and standards in integrating surveys of structural variation. *Nature Genetics* 39, S7S15 (2007).

Schulze-Bergkamen H, Krammer PH. Apoptosis in cancer–implications for therapy. *Seminars in Oncology* 31, 90119 (2004).

Seng S, Avraham HK, Jiang S, Yang S, Sekine M, Kimelman N, Li H, Avraham S. The nuclear matrix protein, NRP/B, enhances Nrf2-mediated oxidative stress responses in breast cancer cells. *Cancer Research* 67, 85968604 (2007)

Shah SP, Lam WL, Ng RT, Murphy KP. Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics* 23, 450458 (2007).

Shah SP. Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenetic and Genome Research* 123, 343351 (2008).

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG. Assembly of microarrays for genome-wide measurement of DNA copy number.. *Nature Genetics* 29, 263264 (2001).

Taylor BS, Barretina J, Socci ND, Decarolis P, Ladanyi M, Meyerson M, Singer S, Sander C. (2008). Functional copy-number alterations in cancer. *PLoS ONE* 3, e3179+ (2008).

Vakhno S, Tavar S. CNAnova: a new approach for finding recurrent copy number abnormalities in cancer SNP microarray data *Bioinformatics* 26, 1395-1402 (2010).

Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23, 657-663 (2007).

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077-1082 (1998).

Walter V, Nobel AB, Wright FA. DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors. *Bioinformatics* 27, 678-685 (2011).

Westfall PH, Young SS. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. *New York* (1993).

Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyse. *Bioinformatics* 21, 40844091 (2005).

Zhang J, Shi YH, Lalonde E, Li LL, Cavallone L, Ferenczy A, Gotlieb WH, Foulkes WD and Majewski J. Exome profiling of primary, metastatic and recurrent ovarian carcinomas in a BRCA1-positive patient. *BMC Cancer* 13, 146 (2013).

Zhang N, Siegmund D, Ji H, Li JZ. Detecting simultaneous change-points in multiple sequences. *Biometrika* 97, 631-645 (2010).

Zhang Q, Ding L, Larson DE, Koboldt DC, McLellan MD, Chen K, Shi X, Kraja A, Mardis ER, Wilson RK, Borecki IB, Province MA. CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics* 26, 464-469 (2010).

Zhou C, Nitschke AM, Xiong W, Zhang Q, Tang Y, Bloch M, Elliott S, Zhu Y, Bazzone L, Yu D, Weldon CB, Schiff R, McLachlan JA, Beckman BS, Wiese TE, Nephew KP, Shan B, Burow ME, Wang G. Proteomic analysis of tumor necrosis factor-alpha resistant human breast cancer cells reveals a MEK5/Erk5-mediated epithelial-mesenchymal transition phenotype. *Breast Cancer Research* 10, R105 (2008).