ABSTRACT

Title of dissertation: Reversible Jump Hidden Markov Model Analysis of Longitudinal Data with Medical Applications

Jin Yan, Doctor of Philosophy, 2013

Dissertation directed by: Professor Michel Wedel
Robert H. Smith School of Business
Professor Paul J. Smith
Mathematical Statistics Program

Longitudinal datasets that contain the same variables at multiple time occasions from a given subject are frequently observed in current medical studies. Research has been done to develop method to analyze such data and make meaningful inferences. In this dissertation, we use hidden Markov models (HMM) and a modified reversible jump Markov chain Monte Carlo algorithm to analyze the longitudinal medical data .

For an eye tracking data of participants looking at chest X-rays with a potential cancerous nodule, we use the HMM model to find out what areas on the images attract participants attention more, how their eyes jump among these areas, and which scan pattern is related to an effective detection of the nodule. We estimated the total number of areas of interest (AOIs) on each image, as well as their centers, sizes and orientations. We use pixel luminance as prior information, as nodules are often brighter and luminance may thus affect the AOIs. Differences in scan patterns between those who found the nodule and those who didn't, are discussed.

For a HIV clinical trial data, we use the hidden Markov model to estimate the health states each patient at different time points, compare the states with physical phenomena in HIV clinical trials, and predict health development patterns.

KEY WORDS: Eye tracking; Areas of interest; Reversible jump Markov chain Monte Carlo; HIV; Health States;

Reversible Jump Hidden Markov Model Analysis of Longitudinal Data with Medical Applications

by

Jin Yan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee (order by last name):
Dr. Paul J Smith (Co-Chair, Co-Advisor)
Dr. Michel Wedel (Co-Chair, Co-Advisor)
Dr. Shuo Chen
Dr. Benjamin Kedem
Dr. Mei-Ling Ting Lee (Dean's Representative)
Dr. Guoxing Soon
Dr. Grace Yang

Acknowledgments

The dissertation is dedicated to my grandfather who passed away two years ago. He was so happy when knowing that I'm going to pursue a Ph.D. but sadly passed away before he could see the moment. I believe that he is always around to support me during the hard times and I can imagine how happy he is when I get the Ph.D. degree.

I owe my gratitude to all the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my co-advisors, Professor Michel Wedel and Professor Paul J. Smith for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past three years. They have always made themselves available for help and advices, guided me how to become an independent researcher and taught me the skills to write scientific papers. They have also gave me great help and advice in my life. Plus, I have learned how to devote myself into research, concentrate on it and work hard from them. It has been a pleasure to work with and learn from such extraordinary individuals.

My sincere thanks also goes to my mentor in U.S. Food and Drug Administration (FDA), Dr. Guoxing Soon. Without his extraordinary knowledge of HIV clinical trials, this dissertation would lack the compelling section. The great opportunity he provided to me, working in FDA as an intern for two years, is also really appreciated as it gave me great statistical research experience and finical support.

Thanks are also due to Dr. Eliott Siegel, Professor of imaging informatics at the University of Maryland, School of Medicine in Baltimore, to make the chest X-ray images used in this study available, and Alice Li, Ph.D. student in the Smith School of Business for her help in eye-tracking data collection.

Besides, I would like to thank Professor Shuo Chen, Professor Benjamin Kedem, Professor Mei-Ling Ting Lee and Professor Grace Yang for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing the manuscript. Other than that, I would like to thank for Dr. Eric Slud, Dr. Abram M. Kagan and Dr. Koralov Leonid, who gave great statistical courses which helped me build up my strong Statistics background. Ms. Regalado Celeste, Ms. Hidalgo Haydee V. and Ms. Berry, Linette Denise also gave me great help on my graduate life here, which I would love to highly appreciate.

I owe my deepest thanks to my family - my mother, father and my grandmother who have always stood by me and care about me. Their encourages and understanding have helped me fight against impossible odds throughout these five years. Words cannot express the gratitude I owe them.

I would like to acknowledge financial support from the Department of Mathematics, Campus Recreation Services, and U.S. Food and Drug Administration.

It is impossible to remember all, and I apologize to those I've inadvertently left out. Last but not least, thank you all!

Table of Contents

List of Tables vi

List of Figures vii

List of Abbreviations ix

1 Introduction 1

2 Reversible Jump Hidden Markov Model 5
- 2.1 Hidden Markov Model . . . 6
  - 2.1.1 Hidden Markov model . . . 6
  - 2.1.2 Markov Chain Monte Carlo Estimation of HMMs . . . 9
- 2.2 The Forward-Backward Recursions . . . 14
  - 2.2.1 Forward-Backward Recursion . . . 16
  - 2.2.2 Forward-Backward Algorithm . . . 17
- 2.3 Label Switching and Diagnostics . . . 18
  - 2.3.1 Label Switching . . . 19
  - 2.3.2 Diagnostics of MCMC Convergence . . . 21
  - 2.3.3 Determining Model Adequacy . . . 24
- 2.4 Reversible Jump Markov Chain Monte Carlo . . . 25
  - 2.4.1 Calculate the marginal likelihood $m(\mathbf{y}|M_S)$ . . . 26
  - 2.4.2 Reversible Jump Markov Chain Monte Carlo . . . 28
  - 2.4.3 Augmented Birth . . . 32

3 Synthetic Data Analysis 35
- 3.1 Data Generation . . . 35
- 3.2 Model description . . . 36
- 3.3 Summary of Results . . . 41

4 A Reversible Jump Hidden Markov Model Analysis of Search for Cancerous Nodules in X-ray Images 46
- 4.1 Introduction . . . 46
- 4.2 Data Description . . . 49
- 4.3 The Reversible Jump Hidden Markov Model and MCMC . . . 52
  - 4.3.1 Details of the Hidden Markov Model . . . 53
    - 4.3.1.1 Spatial Poisson Process . . . 53
    - 4.3.1.2 Spatial and Luminance Priors . . . 54
    - 4.3.1.3 Partially Hidden Markov Chain . . . 55
  - 4.3.2 Reversible Jump Markov Chain Monte Carlo . . . 56
  - 4.3.3 Posterior Computation and Gibbs Sampler . . . 58
- 4.4 Case Study . . . 62
  - 4.4.1 The Areas of Interest . . . 62
  - 4.4.2 Analysis of Transition Patterns . . . 67

4.4.3 Effects on search performance . . . 71
4.5 Conclusion and Discussion . . . 74

5 A Hidden Markov Model Analysis of HIV Clinical Trial Data for Treatment Naïve Female 79
5.1 Introduction . . . 79
5.2 Data Description . . . 80
5.3 The Hidden Markov Model and MCMC . . . 82
5.3.1 Conditional Distribution for Observed Data . . . 83
5.3.2 Hidden Markov Chain . . . 84
5.3.3 Posterior Computation and Gibbs Sampler . . . 84
5.4 Case Study . . . 86
5.4.1 The number of states . . . 86
5.4.2 Analysis under $S = 4$ . . . 87
5.5 Conclusion and Discussion . . . 93

6 Conclusion and Future Studies 96

A Synthetic Data Study 100
A.1 Calculation for the acceptance rate for the birth and death moves . . 100
A.2 Simulation plots for $\vec{\mu}$ and their autocorrelation . . . 101

B Appendix for Eye Tracking Data Analysis 103
B.1 Prior for $\alpha_i(j^*)$ . . . 103
B.2 Proposal function for $u_i(j)$ . . . 104
B.3 Acceptance Ratio for the Birth Moves . . . 105
B.4 Figures and Tables . . . 107

C Appendix for HIV Clinical Study 114

Bibliography 115

List of Tables

3.1 True parameters for the synthetic datasets. 37
3.2 Estimated $\vec{\mu}$ for both datasets. 45

4.1 Summary of the posterior distribution of $\vec{\mu}$ and Σ for case 14. 66
4.2 Switching rate between AOIs of each participant across images. . . . 71
4.3 Switching rate of each image across respondents. 71
4.4 Effects of search attributes on failures and false positives versus successes in a multinomial logit model. 72

5.1 Total number of patients in each group 81
5.2 Log marginal likelihood under each estimated model 86
5.3 Patients in each major transition pattern and percentage of ever suppressed and rebound 93

C.1 Time windows definition 114

List of Figures

2.1 A simple graphic description of the mechanism of HMM. Here y_t are the outputs and s_t are the hidden states. 8
2.2 Example of label switching for a 5 state hidden Markov model. . . . 20
2.3 Example of graphical diagnostic of convergence for a 4 state HMM. The graph shows the posterior estimates of components of Θ. 23

3.1 Synthetic datasets. The left panel shows the isolated dataset, while the right panel represents the mixed dataset. 36
3.2 Estimated number of states for each dataset again the iterations of the MCMC chain. 42
3.3 Simulation results of $\vec{\mu}$ for dataset 1. From the plot we can see that there is no evidence to say that the MCMC chain did not reach the stationary distribution after 6,000 iterations (burn-in) period. 43
3.4 Autocorrelation plot for each component of $\vec{\mu}$ for dataset 1 after the burn-in period. From the graph we can see there is almost no autocorrelation for the estimation after the burn-in period. 44
3.5 Estimated 80% credible ellipse for each state on both datasets. 45

4.1 One X-ray image with one nodule present (case 14). The circle marks the nodule (not shown on the test images). 50
4.2 Scan-paths for 16 participants (case 14). 52
4.3 Number of states across MCMC iterations 63
4.4 Estimated areas of interest on seven X-ray images. The dashed ellipses are the estimated AOIs, with the state number in center. The circles represent the nodule. Case 14 is used to illustrate the result. . 65
4.5 Boxplot of the posterior distribution of β for each X-ray image. . . . 67
4.6 Posterior distribution of shrinkage parameter A_i for each X-ray, grouped by success-types. 69
4.7 Gray scale plot for the transition probability matrices for the three groups for case 14. The size of the symbol at (s, r) reflects the square root of the posterior mean of $Q_i[s, r]$ across participants in that group, where larger symbols represent greater value. The gray scale of the symbol is proportional to the posterior standard deviation. 70

5.1 Estimated 95% credible ellipse under different number of states. . . . 87
5.2 Estimated 95% credible bands for each state under a 4-state assumption. 88
5.3 Estimated 95% credible bands for each drug at each state for Patients. 90
5.4 Estimated 95% credible bands for each drug at each state. 91
5.5 Estimated 95% credible bands for each drug at each state. 92

A.1 Simulation results of $\vec{\mu}$ for dataset 2. From the plot we can see that there is no evidence to say that the MCMC chain did not reach the stationary distribution after 20,000 iteration (burn-in) period. 101

A.2 Autocorrelation plot for each component of $\vec{\mu}$ for dataset 2 after the burn-in period. From the graph we can see there is almost no autocorrelation for the estimation after the burn-in period. 102

B.1 X-ray images with one nodule present. 108
B.2 X-ray images with one nodule present. 109
B.3 X-ray images with one nodule present. 110
B.4 Scan-paths for 16 participants. 111
B.5 Scan-paths for 16 participants. 112
B.6 Scan-paths for 16 participants. 113

List of Abbreviations

k_1^t	$\{k_1, \ldots, k_t\}$ for any k, k could be either a scaler or a vector
$\mathbf{k}$	$\{k_1, ..., k_T\}$ for any k, k could be either a scaler or a vector
k_{-t}	$\{k'_t, t' \neq t\}$ for any k, k could be either a scaler or a vector
$k^{(j)}$	the jth estimation of k;
ACF	Auto Correlation Function
AOIs	Areas of Interest
ART	Antiretroviral Therapy
FDA	U.S. Food and Drug Administration
HMM	Hidden Markov Model
LOCF	Last Observation Carry Forward (LOCF)
RJMCMC	Reversible Jump Markov Chain Monte Carlo
RJHMM	Reversible Jump Hidden Markov Model
VDMC	Variable-Dimension Monte Carlo

Chapter 1

Introduction

A longitudinal dataset contains the same measurements (variables) at multiple measurement occasions for a given subject or a group. Typically many subjects are involved in the dataset. They are often used in psychology to study individual's developmental trends across time, in sociology to study events throughout generations or individual life times, and in medicine to study how radiologists search for tumors and how patients respond to a drug or a therapy. For example, in eye tracking data, the eye fixation positions for one respondent can be recorded while he/she looks at an image, and each record is a two dimensional vector representing the x and y coordinates of the fixations on the image. The observer may be given a task, for example to identify a possible cancerous nodule in a diagnostic X-ray. Another example occurs for HIV infected patients under antiretroviral therapy (ART), during which their CD4 (a type of white blood cell that fights infection, but could be destroyed by HIV) counts, their HIV 1-RNA viral loads and other health related measurements are recorded in pre-determined time windows to track patients' reactions to the therapy.

One of the biggest advantages of longitudinal datasets is that they allow us to measure changes over time. Unlike those cross-sectional data analyse which compare different individuals with the same characteristics, longitudinal data contains records

of multiple observations for the same people. Therefore the differences observed between subjects can be separated from differences within subjects. Sequences are therefore not confounded with irrelevant between subject variation, and we can better control for unobserved individual heterogeneity. Because of this, using longitudinal data one can make more accurate observations of changes and can better assess causality than using cross-sectional data. This advantage occurs in various other fields. Taking the previous two cases as examples, in the eye tracking data, we can study how respondents' eye fixation position changes as time goes on, while for the HIV clinical trial case, we can study how disease related measures change over time as the patients are under the ART.

The second important advantage of longitudinal data analysis is that the longitudinal data connects the past to the present, and may also be used to predict the future. So we can, for example, evaluate the effect of a drug or a therapy as well as predict how the patients will react to it in the future. This is especially useful in medical research.

The third advantage of longitudinal data analysis is that we can recognize general development patterns for each individual, provide insight into their functional form, and identify individuals whose data do not conform to the general pattern. We can also compare the patterns for different groups of people to reveal differences and study the reasons for the differences. In the eye tracking data, for example, we can compare the eye tracking pattern of individuals under different viewing goals, or we can compare the patterns of those who can quickly find the target with those who cannot. In HIV clinical trial, we can compare the patterns of measurements for

groups of people in different treatment or age group.

There are some commonly used methods for longitudinal data analysis, for example, ANOVA/MANOVA for repeated measures, mixed effect regression models, covariance pattern models (Jennrich and Schluchter, 1986), generalized estimating equation models (Liang and Zeger, 1986; Zeger and Liang, 1986), structural equation models (Sewall, 1921; Haavelmo, 1943; Herbert, 1953) and time series models. However, apart from time series models, common conventional statistical methods usually require independence between observations, but longitudinal data are very likely to violate this assumption. In addition, rather than focusing on inferences on the evolution of the observed data, often inferences focus on the evolution of unobserved states. These unobserved states can be the areas on a image that are attractive to respondents, or potential health states which patients belong to. In this dissertation, we will introduce a hidden Markov model to analyze the longitudinal data, which contains a Markov process to capture the relationship of previous observations to the current one. We will apply the algorithm we developed on two data sets: eye tracking data on chest X-rays, and data from HIV clinical trials.

The structure of the dissertation is as follows: In Chapter 2, we will give a detailed introduction of the reversible jump hidden Markov model we are going to use in this dissertation, together with a literature review. We will define the model using a Bayesian framework, write out the joint likelihood function, and introduce the Markov chain Monte Carlo algorithm to draw estimates of the parameters and missing data that we are interested in. Some possible algorithms that can be used to draw those values will be introduced and reasons for our choices will be given. We

will also introduce ways to check the performance of the algorithm and the quality of the estimates. An idea of how to improve the shortcoming of the previous algorithms will also be described. In Chapter 3, a synthetic data study will be presented to show the algorithm accuracy as well as the correctness of the computer code. We run two different data sets and compare estimated results with the true value that generates the data to validate the model's accuracy. Chapter 4 will contain an application of the algorithm to eye tracking data analysis of search for cancerous nodules in X-ray images. The algorithm will help us define the areas of interest on the X-ray image and will show what features might affect the position of areas of interest. We will also show that with the help of the definition of the areas of interests, we can study respondents' search behavior and how its related to the search efficiency. Chapter 5 will present an application of the algorithm to HIV clinical trials data, which will define the health states of female patients after they start to take some antiretroviral drug. In that chapter we define the health states of patients based on their first 32 weeks of data and predict the development of the patients health states for later times. We will show that we can predict the same health states by using the data till week 32 compared to using the data till week 48. Chapter 6 will provide conclusions as well as some possible model extensions. Finally, the appendices contain graphs and derivations.

Chapter 2

Reversible Jump Hidden Markov Model

This Chapter will give a detailed introduction of the reversible jump hidden Markov model that will be used in the whole dissertation. First of all, in Section 2.1, an detailed introduction to hidden Markov models will be given, combined with literature reviews and model building. After that, the Bayesian structure will be merged into the model and the likelihood function will be written out. From the joint likelihood function, readers can see why a Markov chain Monte Carlo (MCMC) algorithm is needed to get inference from the model. Two frequently used algorithms to estimate the parameters in MCMC, the Metropolis-Hastings algorithm and Gibbs sampler algorithm, will be introduced in details as well. In that section it will also shown that the Gibbs sampler algorithm indeed is a special case of the Metropolis-Hastings algorithm. In Section 2.2, the forward-backward algorithm, which is used to estimate the hidden states, and its original idea: forward-backward recursion, will be introduced. Some other benefits of the forward-backward recursion, such as, the possibility of inference on the parameter and the hidden states without more computational burden, will also be discussed. Section 2.3 mainly focuses on the label switching issue with some suggested solutions, as well as diagnostic methods for the MCMC chain convergence and model adequacy. Finally, in Section 2.4, in the model the number of states S will be treated as unknown, and several ways to

chose the best number of states will be introduced. This section will mainly focus on the reversible jump MCMC (RJMCMC), which we will use in the entire dissertation, and we will propose an augmented birth method to overcome the shortcomings of the RJMCMC algorithm.

2.1 Hidden Markov Model

A hidden Markov model (HMM) is a statistical model in which observed outputs are distributed conditionally given a chain of unobserved (hidden) states which is assumed to be a Markov process. In this section, we will introduce the hidden Markov model (HMM), write out the joint likelihood of the observable data and the hidden states, and introduce Markov Chain Monte Carlo (MCMC) methods to estimate the desired parameters.

2.1.1 Hidden Markov model

Unlike the simple Markov model, the Markov chain which describes hidden states is unobserved, and thus the transition probabilities between the states are difficult to calculate, and need to be estimated from the observed output. The HMM can be treated as a generalization of a mixture model where the hidden states, which represent the mixture component each observation belongs to, are connected by a Markov process rather than independent of each other. In other words, in the mixture model, the state each observation belongs to does not depend on other states. But in the HMM, they are related to the previous states by nature

of a Markov process. Here we can see the advantage of HMM compared to finite mixture models: it allows sharing information over time and can borrow the strength from the history instead of pooling the data without considering the time dimension. It can reflect that occasionally a state with a large output is in fact weaker evidence than a cluster of small outputs in that state.

This Markov process is usually assumed to be first order, which means future states, conditional on the present state, are independent of the past. But it can also be a kth ($k \geq 2$) order Markov process, in which future states, conditional on the present and $k-1$ previous states, is independent of the past. In this dissertation we only consider first order Markov processes. Also notice that there exist time-homogeneous Markov processes, in which the transition probability between states does not change with time, and time-inhomogeneous Markov processes, in which the probability of transitions between states are related to time. In this dissertation, we will use the time-homogeneous Markov processes. In the discussion part we will suggest a potential extension to a time-inhomogeneous Markov process.

The theoretical development of the HMM was first described by Baum et al. (1966). Nowadays it has been widely used in a variety of areas and disciplines, such as signal processing (Juang and Rabiner 1991), health states (Scott 1995), genetics (Churchill 1989), image analysis (Romberg et al. 2000), economics (Hamilton 1989; Albert and Chib 1993) and eye tracking (Van der Lans et al. 2008). Hidden Markov models are especially widely applied in speech, handwriting and gesture recognition, which is called temporal pattern recognition (Starner and Pentland 1995).

The hidden states of the HMM are not required to have a physical meaning

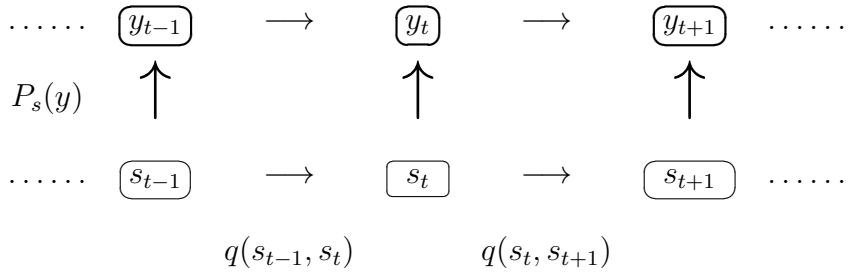


Figure 2.1: A simple graphic description of the mechanism of HMM. Here y_t are the outputs and s_t are the hidden states.

(Scott 2005), but with a reasonable physical meaning the model will be more compelling and easy to understand. Also, even though sometimes the physical meaning is not easy to understand, it might reveal some inner connections between outputs for the same states. This is of special use, for example, for defining health states, which otherwise could be defined only on the basis of several measurements of patients. In our work, we give physical meanings to the hidden states in both applications.

Figure 2.1.1 illustrates a simple example of HMM: assume that for an observed longitudinal chain $\mathbf{y} = \{\vec{y}_1, \ldots, \vec{y}_T\}$, each data point $\vec{y}_t$, $t = 1, \ldots, T$, belongs to an unobserved (hidden) state $s_t \in \{1, \ldots, S\}$, where S is the total number of states. Conditional on the hidden states $\mathbf{s} = \{s_1, \ldots, s_T\}$ and the parameter vector θ, the $\vec{y}_t$ are independently distributed for each time point t with probability

$$P_s(\vec{y}|\theta) = f(\vec{y}_t = \vec{y}|s_t = s, \theta). \tag{2.1}$$

The chain $\mathbf{s} = \{s_1, \ldots, s_T\}$ follows a first order time homogeneous Markov process with the transition matrix $Q_S = (q(r, s))_{S \times S}$, and the initial distribution $\vec{\pi}_0$,

where $r, s \in \mathcal{F}_S$. If we define $k_1^t = \{k_1, \ldots, k_t\}$ for any k (which could be a value or a vector), then formally for any $t \geq 2$

$$\begin{aligned} p(s_t|s_1^{t-1}) &= p(s_t|s_{t-1}) \\ \pi_0(s) &= p(s_1 = s) \\ q(r,s) &= p(s_t = s|s_{t-1} = r). \end{aligned} \tag{2.2}$$

Here we can see that if we set each row of Q_S equal, the HMM will be reduced to a finite mixture model.

2.1.2 Markov Chain Monte Carlo Estimation of HMMs

If we define the whole parameter set $\Theta_S = \{\theta, Q_S\}$, the joint likelihood function of $\{\mathbf{y}, \mathbf{s}\}$ is

$$p(\mathbf{y}, \mathbf{s}|\Theta_S, S) = \pi_0(y_1)P_{s_1}(y_1)\prod_{t=2}^{T} q(s_{t-1}, s_t)P_{s_t}(y_t). \tag{2.3}$$

If we define $\mathcal{F}_S^T$ as the T-dimensional state space, then the full likelihood for the observed data $\mathbf{y}$ is

$$p(\mathbf{y}|\Theta_S, S) = \sum_{s_t \in \mathcal{F}_S} \pi_0(y_1)P_{s_1}(y_1)\prod_{t=2}^{T} q(s_{t-1}, s_t)P_{s_t}(y_t). \tag{2.4}$$

In a Bayesian point of view, we will give prior distributions for the parameter set Θ_S: for θ, usually a conjugate prior $p(\theta)$ is used, while for $\vec{\pi}_0$, and each row $q(r, \cdot)$ of Q_S, an S dimensional Dirichlet distribution $\mathcal{D}_S(\vec{\alpha})$ is used, where $\vec{\alpha}$ is the hyper-parameter. If we use Φ to define all the hyper-parameters including $\vec{\alpha}$, and $p(\Theta_S|\Phi)$ as the prior distribution for the parameter space Θ_S, then the full joint

likelihood will be:

$$\begin{aligned} p(\mathbf{y}, \Theta_S|\Phi, S) &= p(\mathbf{y}|\Theta_S, S)p(\Theta_S|\Phi) \\ &= \sum_{s_t \in \mathcal{F}_S} \left\{ \pi_0(y_1) P_{s_1}(y_1) \prod_{t=2}^{T} q(y_{t-1}, y_t) P_{s_t}(y_t) \right\} \\ &\quad \times p(\theta) \mathcal{D}_S(\vec{\pi}_0|\vec{\alpha}) \prod_{s=1}^{S} \mathcal{D}_S(q(s, \cdot|\vec{\alpha})). \end{aligned} \tag{2.5}$$

Notice that the computation of equation (2.5) is not feasible since it is a sum over S^T elements for a single chain, let alone that there might be multiple chains in a single model. In order to solve the problem, we will use the Markov Chain Monte Carlo (MCMC) algorithm, which involves sampling the parameters and the unobserved values from the desired distribution $\pi(d\Theta)$ (the joint likelihood function here) based on an ergodic (aperiodic and positive recurrent) Markov chain. In practice, we draw the parameters and unobserved values from a proposal distribution in each iteration, and determine whether accept the draw conditional on samples simulated in the immediately preceding step. This gives us a Markov chain that each sample only depends on immediately preceding sample and is independent of any other earlier samples. Under the condition of ergodicity, the Markov chain will also have the desired distribution as its equilibrium distribution. After a large number of iterations (burn in period), the chain is close to convergence and the draws can be used as an approximate sample from the desired distribution. As the number of iterations increase, the quality of the approximation also improves. Getting back to our problems, instead of simulating from equation (2.5), we will simulate both the parameter set Θ_S and the unobserved state $\mathbf{s}$ cyclically from the joint likelihood

function:

$$p(\mathbf{y}, \mathbf{s}, \Theta_S|\Phi, S) = \pi_0(y_1)P_{s_1}(y_1)\prod_{t=2}^{T} q(y_{t-1}, y_t)P_{s_t}(y_t) \times p(\theta)\mathcal{D}_S(\vec{\pi}_0|\vec{\alpha})\prod_{s=1}^{S}\mathcal{D}_S(q(s,\cdot|\vec{\alpha})). \quad (2.6)$$

The history of the MCMC algorithm dates back to the 1940s, and the first reasonable MCMC algorithm was the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). In 1984, Gibbs sampler , which is named after the physicist J. W. Gibbs, is described by Geman and Geman (1984). In summary, the MCMC algorithm is a random walk algorithm, which moves in relatively small steps and thus may take a long time to explore all of the parameter space and might double back. Some other special cases and extensions are *slice sampling* (Neal 2003), *Multiple-try Metropolis* (Liu J. S. 2000) and introducing *Auxiliary variables* into the MCMC chain (Higdon D. M. 1997). In our projects we will only use the Metropolis-Hastings and the auxiliary variable Gibbs sampler algorithms, so we will give a detailed introduction below. For both of these two, the purpose, as we described, is to draw samples η (could be parameters or unobserved states) from the target distribution $\pi(d\eta)$. For notation convenience, we denote $\eta^{(t)}$ as the t^{th} draw of the MCMC.

Metropolis-Hastings algorithm

1. Choose an arbitrary proposal density $q(\eta'|\eta^{(t)})$, which will sample a new value η' based on the current value $\eta^{(t)}$. The proposal density should be a symmetric one, which means $q(\eta'|\eta^{(t)}) = q(\eta^{(t)}|\eta')$.

2. Give an arbitrary point as the first sample $\eta^{(0)}$;

3. For each iteration $t \geq 1$, sample η' from $q(\eta'|\eta^{(t-1)})$;

4. Calculate the acceptance ratio

$$a = \min\{1, \frac{\pi(d\eta')q(\eta^{(t-1)}|\eta')}{\pi(d\eta^{(t-1)})q(\eta'|\eta^{(t-1)})}\};$$

5. Draw a random variable $u \sim \mathcal{U}(0,1)$, if $u < a$, accept $\eta^{(t)} = \eta'$; otherwise keep $\eta^{(t)} = \eta^{(t-1)}$.

A good choice of the proposal density is always very important for Metropolis-Hastings . Bad proposal densities will lower the acceptance ratio and thus slow down the convergence. Also, in multidimensional cases, the Metropolis-Hastings algorithm as described above might requires a proposal density for the whole vector, which is extremely difficult when the dimension is high. An alternative approach that we will describe is the Gibbs sampler algorithm.

Gibbs sampler algorithm

If $\eta = \{\eta_1, \ldots, \eta_K\}$ is K dimensional, the Gibbs sampler will sample η component by component instead of sampling it as a whole. If we denote $h_{-t} = \{h_1, \ldots, h_{t-1}, h_{t+1}, \ldots, h_K\}$ for any h, then the Gibbs sampler can be described as follows:

1. Choose a starting point $\eta^{(0)}$;

2. In each iteration t, let $\eta^{(t)} = \eta^{(t-1)}$, then for each component $r = 1, \ldots, K$, update $\eta_r^{(t)}$ by η_r^* which is sampled from the conditional posterior density $p(\eta_r^*|\eta_{-r}^{(t)})$, until every component is updated.

If in each small step, we define the old vector as $\eta_{old}^{(t)} = \{\eta_1^{(t)}, \ldots, \eta_K^{(t)}\}$ and the newly updated vector as $\eta_{new}^{(t)} = \{\eta_1^{(t)}, \ldots, \eta_{r-1}^{(t)}, \eta_r^*, \eta_{r+1}^{(t)}, \ldots, \eta_K^{(t)}\}$, then

$$\begin{aligned}
\pi(d\eta_{new}^{(t)})q(\eta_r^t|\eta_{-r}^{(t)}) &= p(\eta_{new}^{(t)})q(\eta_r^{(t)}|\eta_{-r}^{(t)}) \\
&= p(\eta_r^*|\eta_{-r}^{(t)})p(\eta_{-r}^{(t)})p(\eta_r|\eta_{-r}^{(t)}) \\
&= p(\eta_r^*)p(\eta_{old}^{(t)}) = \pi(d\eta_{old}^{(t)})q(\eta_r^*|\eta_{-r}^{(t)}),
\end{aligned}$$

thus the Gibbs sampler is indeed a special case of Metropolis-Hastings algorithm, for which the acceptance ratio is always 1. It is only applicable if the full conditional posteriors in 2 are known and can be sampled from.

The individual components can be sampled by different algorithm, such as adaptive rejection sampling (Gilks, W. R. and Wild, P. 1992), a one-dimensional Metropolis-Hastings step, or slice sampling (Neal, Radford M. 2003). The order for updating the rth competent can be either fixed or random, and in our project, we will use a random order rather than a fixed one. Also, instead of estimating the parameter component by component, a modified Gibbs sampler algorithm, blocked Gibbs sampler, can also be used. The blocked Gibbs sampler groups the variables and generates each group conditional on the other groups.

For both the Metropolis-Hastings algorithm and Gibbs sampler , the autocorrelation and the burn-in period are two things we should pay attention to. Because the random walk usually moves in small steps, a set of nearby draws are usually correlated with each other and may not very efficiently represent the independent samples of target distribution. Some therefore prefer to decide a lag l, and only choose every l-th draw to decrease the autocorrelation. Typical l is between $5 - 20$.

Also, even though the Markov chain will converge to the target distribution eventually, the initial draws might follow a quite different distribution function based on the choice of the initial value, so a burn-in period is necessary,and one throws away the draws from this phase. The burn-in period depends on the performance of the MCMC and its autocorrelation, and thus is not fixed. In our project we usually set it to $20,000$ to $30,000$.

Now using the ideas of MCMC, we can state how we will sample from the equation (2.6), with the use of the blocked Gibbs sampler algorithm. That is, cyclically sample Θ_S from $p(\Theta_S|\mathbf{s}, \mathbf{y}, \Phi, S)$ and $\mathbf{s}$ from $p(\mathbf{s}|\mathbf{y}, \Theta_S, \Phi, S)$. The simulation from $p(\Theta_S|\mathbf{s}, \mathbf{y}, \Phi, S)$ is Gibbs sampler algorithm, while the sampling from $p(\mathbf{s}|\mathbf{y}, \Theta_S, \Phi, S)$ will use what we called the Forward-Backward Algorithm which will be introduced below.

2.2 The Forward-Backward Recursions

In the MCMC procedure, there are two ways to deal with the hidden state $\mathbf{s}$, either integrating it out,or sampling it from $p(\mathbf{s}|\mathbf{y}, \Theta_S, \Phi, S)$ using an augmented variable sampler.. There are several reasons why we choose to sample it instead of integrating it out. First of all, the Metropolis-Hastings algorithms that average over $\mathbf{s}$ will perform poorly when the dimension of the parameter Θ_S is large, as we described before, since it's hard to find a good proposal function that can sample the whole parameter set and most of the proposed draws will be rejected. Second, the highly correlated components of Θ_S under the full distribution $p(\Theta_S|\mathbf{y})$ are often

independent under the conditional distribution $p(\Theta_S|\mathbf{s},\mathbf{y})$ when adding $\mathbf{s}$. Thus it will speed up the mixing of the MCMC (van Dyk and Meng 2000). Thirdly, sometimes the hidden state $\mathbf{s}$ itself, and some functions $g(\mathbf{s})$ might be of interests to the researchers, so sampling $\mathbf{s}$ provide a way to obtain inferences about it. Last but not least, $\mathbf{s}$ and some of its function $h(\mathbf{s})$ might be used to diagnose the model adequacy and convergence of the MCMC procedure (Robert et al. 1999).

There are two ways to sample $\mathbf{s}$ from $p(\mathbf{s}|\mathbf{y},\Theta_S,\Phi,S)$. One of them is the Direct Gibbs (DG) Sampler introduced by Albert and Chib (1993), which is to sample s_t from the full conditional distribution

$$p(s_t = s|s_{-t},\mathbf{y},\Theta_S) \propto q(s_{t-1},s)q(s,s_{t+1})P_s(\vec{y}_t|\Theta_S), \tag{2.7}$$

for $t = 1,\ldots,T$. Notice that the end points need some modification. Since the DG is using the traditional Gibbs Sampler and estimating the hidden states one by one, it will slow down the mixing. It's also mentioned in their paper that the sampled state vector can only be accepted if each state was at least drawn once.

The second algorithm is the Forward-Backward algorithm (Chib 1996; Scott 1999), which is based on the Forward-Backward recursion, a mechanism first developed by Baum et al. (1970) to implement an EM algorithm to maximize the likelihood function (2.4) given the observed value $\mathbf{y}$. In this dissertation we will use the Forward-Backward algorithm instead of DG, so we will give detailed explanation of the Forward-Backward recursion and the Forward-Backward algorithm. We will also discuss why we choose the Forward-Backward rather than DG.

2.2.1 Forward-Backward Recursion

The forward recursion involves calculating the probabilities $p_t(r,s) = p(s_{t-1} = r, s_t = s|y_1^t, \Theta)$, the joint probability of the (s_{t-1}, s_t) conditional on the first t observations, and the matrices $\mathbf{P}_2, \ldots, \mathbf{P}_n$, where $\mathbf{P}_t = (p_t(r,s))$. It is called forward recursion because once $\mathbf{P}_{t-1}$ is calculated, one can calculate $\mathbf{P}_t$ by:

$$\begin{aligned} p_t(r,s) &= p(s_{t-1} = r, s_t = s|y_1^t, \Theta_S) \\ &\propto p(s_{t-1} = r, s_t = s, y_t|y_1^{t-1}, \Theta_S) \\ &= p(s_{t-1} = r|y_1^{t-1}, \Theta_S)p(s_t = s|s_{t-1} = r, y_1^{t-1}, \Theta_S)p(y_t|s_t, s_{t-1}, y_1^{t-1}, \Theta_S) \\ &= \pi_{t-1}(r|\Theta_S)q(r,s)P_s(y_t|\Theta_S). \end{aligned} \tag{2.8}$$

where $\pi_t(r|\Theta_S) = \sum_s p_t(s,r)$ for any t and can be computed once the matrix $\mathbf{P}_t$ is known. Also notice that $\sum_r \sum_s p_t(r,s) = 1$.

The backward recursion involves calculating the probabilities $p'_t(r,s) = p(s_{t-1} = r, s_t = s|\mathbf{y}, \Theta)$, the joint probability of (s_{t-1}, s_t), conditional on the whole observations, and the matrices $\mathbf{P}'_2, \ldots, \mathbf{P}'_T$ where $\mathbf{P}'_t = (p'_t(r,s))$. Notice first that $\mathbf{P}_T = \mathbf{P}'_T$. It is called backward recursion because once we know $\mathbf{P}'_{t+1}$ and $\mathbf{P}_t$, we can compute $\mathbf{P}'_t$ by:

$$\begin{aligned} p'_t(r,s) &= p(s_{t-1} = r, s_t = s|\mathbf{y}, \Theta_S) \\ &= p(s_{t-1} = r|s_t = s, \mathbf{y}, \Theta_S)p(s_t = s|\mathbf{y}, \Theta_S) \\ &= p(s_{t-1} = r|s_t = s, y_1^t, \Theta_S)\pi'_t(s|\Theta_S) \\ &= p_t(r,s)\frac{\pi'_t(s|\Theta_S)}{\pi_t(s|\Theta_S)}, \end{aligned} \tag{2.9}$$

where $\pi'_t(s|\Theta_S) = p(s_t = s|\mathbf{y}, \Theta_S) = \sum_r p'_t(r,s)$.

2.2.2 Forward-Backward Algorithm

The Forward-Backward algorithm (Chib 1996; Scott 1999) uses the idea of the Forward-Backward recursion. It first calculates the forward matrices $\mathbf{P}_2, \ldots, \mathbf{P}_T$ as described in equation 2.8. However, instead of calculating the backward matrices $\mathbf{P}'_2, \ldots, \mathbf{P}'_T$, it runs a *stochastic backward recursion*, which is based on

$$\begin{aligned} p(\mathbf{s}|\mathbf{y}, \Theta_S) &= p(s_T|\mathbf{y}, \Theta_S) \prod_{t=T-1}^{1} p(s_t|s_{t+1}^T, \mathbf{y}, \Theta_S) \\ &= p(s_T|\mathbf{y}, \Theta_S) \prod_{t=T-1}^{1} p(s_t|s_{t+1}, y_1^{t+1}, \Theta_S) \\ &\propto p(s_T|\mathbf{y}, \Theta_S) \prod_{t=T-1}^{1} p_{t+1}(s_t, s_{t+1}), \end{aligned}$$

where the second line is because of the property of the Markov Chain and the third line is Bayes theorem.

Then the Forward-Backward algorithm follows the following steps:

1. Draw s_T from $\pi_T(\cdot|\Theta)$.

2. Sample s_t from $p_{t+1}(\cdot, s_{t+1})$, which is the s_{t+1}th column of $\mathbf{P}_{t+1}$.

There are also two reasons that we prefer Forward-Backward compared to DG, First, instead of adding T additional blocks into the Gibbs sampler algorithm, Forward-Backward will only add one block, which will hugely improve the convergence of the MCMC chain. Second, by sampling $\mathbf{s}$ from $p(\mathbf{s}|\mathbf{y}, \Theta)$ using the Forward-Backward algorithm, the dependence of each state s_t on other previously drawn states is reduced. Scott (2002) showed that the autocovariance of the complete data sufficient statistics drawn by DG is larger than that of the Forward-Backward algorithm and tends to become much larger as the posterior covariance of the hidden

states increases. This faster mixing of $\mathbf{s}$ results in a faster mixing for the parameter set Θ_S. Furthermore, in Chib (1996), it is also mentioned that if the prior is properly assigned to all the parameters, unlike DG, one does not need to reject a particular draw $\mathbf{s}$ that might have some empty states (i.e., no observation was assigned to some states). This also helps reduce the computing load.

The Forward-Backward algorithm also helps us to infer the hidden state $\mathbf{s}$. For example, if we want to estimate the marginal distribution $\pi'_t(s) = p(s_t = s|\mathbf{y})$, one intuitive choice is that

$$\pi'(s) = \frac{1}{G}\sum_{j=1}^{G} I(s_t^{(j)} = s), \tag{2.10}$$

where G is the total number of iteration and $s_t^{(j)}$ is the draw from the jth iteration. However, with the help of the backward recursions, we can also estimate it by the Rao-Blackwellized estimator using the backward recursions (Gelfand and Smith 1990; Casella and Robert 1996),

$$\hat{\pi}_t(s) = \frac{1}{G}\sum_{j=1}^{G} \pi'_t(s|\Theta^{(j)}). \tag{2.11}$$

Once the forward recursions are run, the backward recursions require little effort.

2.3 Label Switching and Diagnostics

There is a common obstacle for HMMs regardless of which sampling strategy is used: *Label Switching*. Also diagnostics for MCMC chain convergence is a major concerns in Bayesian work. In this section we will discuss these two issues.

2.3.1 Label Switching

Figure 2.2 gives us an example of label switching: the graph shows the estimated parameters for a 5 state hidden Markov model (the example is derived from a simplified version of the application in Chapter 5). For each state, there is an estimated mean position $\vec{\mu}$, shown row by row in the figure. We can see that the second row and the fifth row have a paired pattern, where there is a discrete jump in the Markov Chain at around iteration $31,000$, which is suggesting that the labels of state 2 and state 5 switched in the estimation. The reason for label switching is that the HMM joint likelihood function (2.6) is invariant under arbitrary permutations of the state labels, and thus labels cannot be identified in the simulation. If the parameters for two states are exchangeable in their prior distribution, then their marginal posterior distributions are also identical.

Several methods have been developed to solve these problems. The first and most frequently used one is to provide constraints in priors, typically by ordering the means or variances of the mixture components. This method is called *identifiability constraints* (ICs; Jasra et al. 2005). Most constraints involve ordering by the mean or the covariance. Such constraints can be realized by (A) applying the Metropolis-Hastings algorithm and rejecting draws that violate the constraints (Richardson and Green 1997), (B) reparameterizing the model (Robert and Titterington 1998),or (C) permuting the samples to satisfy the constraint in every iteration (Diebolo and Robert 1994). For method (A), the problem is that the acceptance rate of the Metropolis-Hastings might be very low.

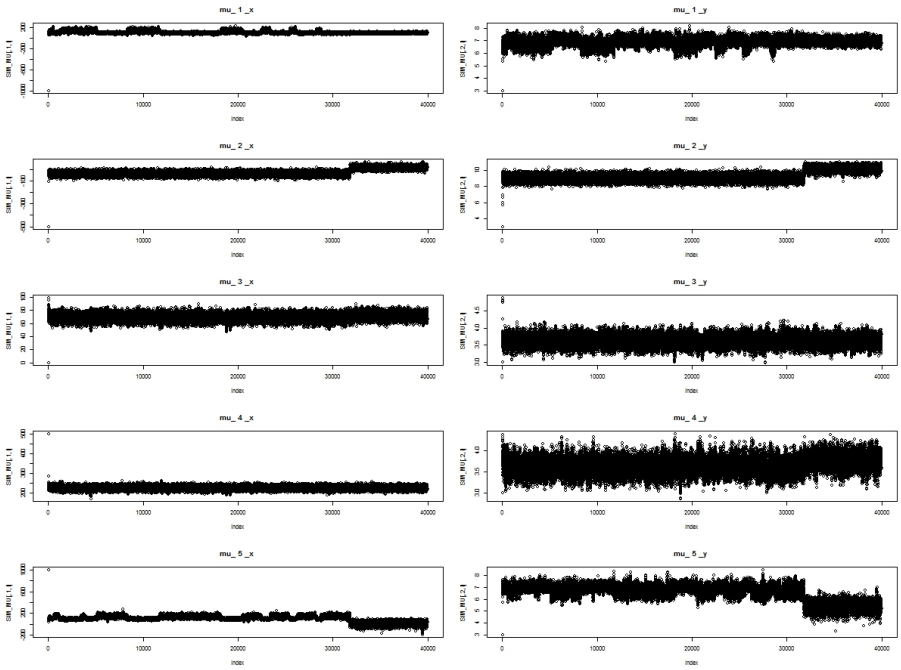


Figure 2.2: Example of label switching for a 5 state hidden Markov model.

However, a weakly informative prior might do little to prevent label switching. Also, for multivariate problems, finding good ICs is quite hard. In this sense, the physical meaning of the hidden state may provide very useful prior information by suggesting an order for Θ_S.

An alternative way is to use a *random permutation sampler* to propose a new permutation of the label at each iteration and to use Metropolis-Hastings to decide whether to accept the new permutation (Frühwirth-Schnatter 2001). This algorithm ensures that the MCMC visits all $S!$ symmetric modes of the parameter space. However, computational burden is an important issue for this algorithm.

The third method to solve label switching is called the *relabelling algorithm* (Stephens 1997a, 1997b, 2000b). It defines a loss function, and then seeks to minimize the loss of performing an action a, which is associated with the parameter

set, by selecting the permutation that minimizes the loss and thus minimize the posterior expected loss. Since the loss function cannot be calculated exactly, the Monte Carlo estimates needs to be used to approximate the loss.

In this dissertation, to make it simple, we will mainly order the mean and use method (A) to avoid label switching. We drew the estimation plots to check label switching and they showed no evidence of label switching.

2.3.2 Diagnostics of MCMC Convergence

Since the estimation from a MCMC analysis should not be treated as reliable until the chain has sufficiently mixed and come close to the stationary distribution, determining the convergence and mixing become important issues that we must address.

There are typically three ways to check the convergence of the MCMC chains: (A) Use the theoretical and mathematical properties of Markov Chains. (B) Diagnose summary statistics and (C) Use the idea of *perfect sampling* (Propp and Wilson 1996). The theoretical diagnostic (A) usually relies on the mathematical properties of the transition kernels of a MCMC chain. The biggest disadvantage is its mathematical complexity. See the important works of Rosenthal (1995a, 1995b), Frieze et al. (1994), and Robert and Polson (1994). Furthermore, even if we know that theoretically the MCMC is convergent, we do not know what the limiting distribution is or when the MCMC chain is close to convergence. Robert et al. (1999) dealt with the MCMC convergence diagnostic issue for HMM based on work of Cowles

and Carloin (1996) and Mengersen et al. (1999), which dealt with the same issue for general MCMC. Here, we will mainly talk about ways to use summary statistics to diagnose the convergence. However, one should be aware that these methods are in fact indicators of non-convergence, instead of evidence for convergence.

1. Choose some functions $c(\Theta_S)$ and determine whether they have reached a stable distribution under the MCMC estimation. The functions $c(\Theta_S)$ are usually not specified and we can choose them for convenience. We can either use a diagnostic for each component of Θ_S or we can find a function c that deals with the whole parameter set. A natural function in the latter case is the posterior distribution $p(\Theta_S|\mathbf{y})$, which can be easily obtained from the MCMC process with the help of the Forward-Backward recursion. (See details in Section 2.4.1.)

2. To judge whether the MCMC chain has stabilized to its limit distribution, another intuitive way is to judge it graphically by the traceplots of the path of the MCMC runs. For example Figure 2.3 shows a diagnostic for the components of a four state HMM, from which we can see that the MCMC chain has stabilized.

3. Autocorrelation: If the correlation within a single parameter chain is high, it indicates slow mixing and even may mess up the individual convergence. This is because the chain will explore less space in a limited time and thus the results might be problematic. So plotting the autocorrelation can help to identify the lag k by which the chain should be thinned, meaning that only

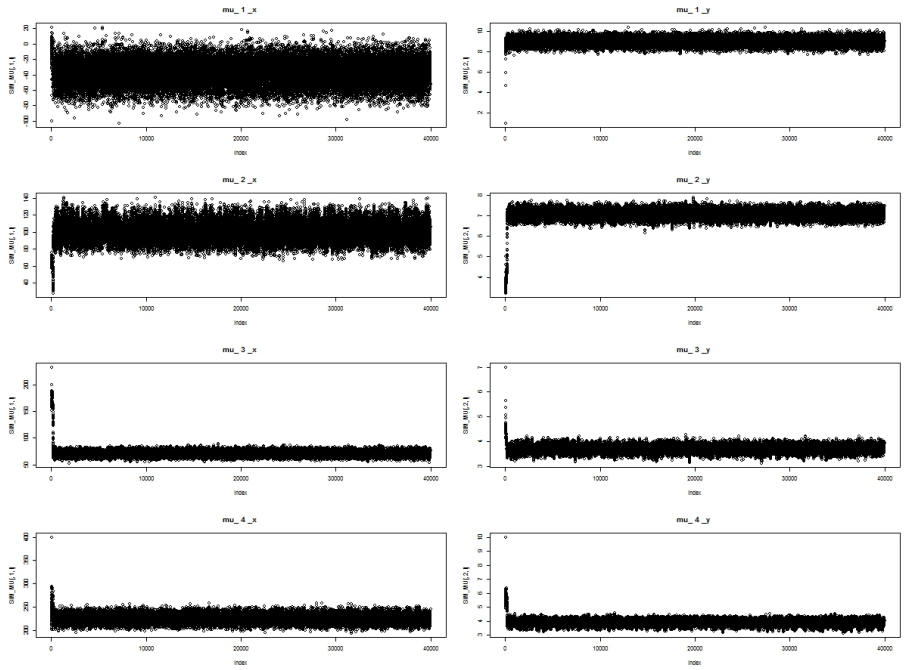


Figure 2.3: Example of graphical diagnostic of convergence for a 4 state HMM. The graph shows the posterior estimates of components of Θ.

every kth realization should be retained for analysis.

4. Some other standard empirical diagnostics;

   - *The Geweke Time-Series Diagnostic* (Geweke 1992): Compare some early window of the chain after the burn-in period with a non-overlapping late window of the chain by a difference of means test. The test uses an asymptotic approximation for the standard error of the difference. The proposed statistic is asymptotically standard normal. A significant difference will be evidence that the two selected portions of the chain differ substantially and also evidence that the chain is not converged. This is a time-series way of checking the convergence of the MCMC chain.

However, its shortcoming is that the test statistics maybe affected by the choice of the windows.

- *The Gelman and Rubin Multiple Sequence Diagnostic* (Gelman and Rubin, 1992a): Run the MCMC with different starting points that are overdispersed, then compare these chains using an ANOVA-based test: calculate the within chain variance V_w, between chain variance V_b and estimated variance $\widehat{Var} = (1 - 1/n)V_w + (1/n)V_b$, then calculate the *estimated scale reduction*

$$\hat{R} = \sqrt{\frac{\widehat{Var}}{W}}$$

to judge whether the chains have reached the same distribution. Values of $\hat{R}$ less than roughly 1.1 or 1.2 are acceptable evidence that the chains are following the same distribution according to Gelman (1996).

2.3.3 Determining Model Adequacy

Model adequacy is another important issue that we should pay attention to. Three useful diagnostics are listed here:

1. If the hidden states are interpretable, the posterior distribution of $\mathbf{s}$ can be an great diagnostic to test the model adequacy.

2. The autocorrelation function (ACF) (MacDonald and Zucchini 1997) is another useful diagnostic to determine whether the model is adequate especially for regression model. For example, if the model involves a regression, then the statistical tools can calculate the posterior distribution of the ACF for the

residuals. If all the marginal distributions of the ACF cover 0, this suggests that the dependence in the residuals is small enough to be ignored, and thus the model is reasonably adequate.

3. The predictive distribution of data $p(y|\mathbf{y})$ can also be used as a diagnostic for model adequacy. The summary of $p(y|\mathbf{y})$ can be achieved by posterior predictive checks (Rubin 1984; Meng 1994; Gelman et al. 1995). Roughly speaking, for each $\Theta_S^{(h)}$ which is the hth draw of the MCMC chain, simulate a dataset $\mathbf{y}^{(j)}$ from $p(\mathbf{y}|\Theta_S^{(j)})$ and then compute the summary for $\mathbf{y}^{(j)}$.

2.4 Reversible Jump Markov Chain Monte Carlo

In the previous sections describing the HMM and its estimation, we assume that the number of states in the model is S. However, S could also be unknown and thus need estimation. Assume that the probability for the total number of states equal S is $p(S)$ and the maximum possible number is $\mathcal{S}$. Then the full likelihood is

$$p(\mathbf{y}, \mathbf{s}, \Theta_S, S|\Phi) = \left(\sum_{s_t \in \mathcal{F}_S} \pi_0(y_1) P_{s_1}(y_1) \prod_{t=2}^{T} q(y_{t-1}, y_t) P_{s_t}(y_t) \right) p(S) p(\Theta_S|\Phi). \quad (2.12)$$

There are various ways to choose S, such as to calculate the marginal likelihood $m(\mathbf{y}|M_S)$ from the MCMC output (Newton and Raftery 1994; Gelfand and Dey 1993; Chib 1995), where M_S is the estimated model when the number of states is S, and Variable-Dimension Monte Carlo (VDMC). We will first give a brief introduction of calculating $p(S|\mathbf{y})$, and then we introduce VDMC and especially focus on one algorithm: Reversible Jump Markov Chain Monte Carlo (RJMCMC).

2.4.1 Calculate the marginal likelihood $m(\mathbf{y}|M_S)$

If we define $M_S, S \in \{1, \ldots, \mathcal{S}\}$ as the model with the number of state S, the marginal likelihood under model M_S is

$$m(\mathbf{y})_{M_S} = \int p_S(\mathbf{y}|\Theta_S)\pi(\Theta_S)d\Theta_S, \tag{2.13}$$

where $p_S(\mathbf{y}|\Theta_S)$ is the density function of the observed data $\mathbf{y}$ under model M_S, and $\pi(\Theta_S)$ is the prior density. There are several estimators of the marginal likelihood. The Newton and Raftery (1994) estimator is

$$\hat{m}_{NR}(\mathbf{y})_{M_S} = \left\{ \frac{1}{G} \sum_{j=1}^{G} \left(p_S(\mathbf{y}|\Theta_S^{(j)}) \right)^{-1} \right\}^{-1},$$

where G is the total number of draws and $\Theta_S^{(j)}$ is the jth draw. The Newton and Raftery estimator is the harmonic mean of the likelihood values, which converges to the true marginal likelihood as G goes to infinity, but is not stable (Chib 1995).

Another possible estimator is proposed by Gelfand and Dey (1993):

$$\hat{m}_{GD}(\mathbf{y})_{M_S} = \left\{ \frac{1}{G} \sum_{j=1}^{G} \left(\frac{p(\Theta_S^{(j)})}{p_S(\mathbf{y}|\Theta_S^{(j)})\pi(\Theta_S^{(j)})} \right) \right\}^{-1},$$

where $p(\Theta_S^{(j)})$ is another density function whose tails are thinner than the denominator. The Gelfand and Dey estimator also converges to the true marginal likelihood as G goes to infinity. However, it requires a tuning function which might be hard to find.

The third method is proposed by Chib (1995), which calculates the log marginal likelihood by:

$$\begin{aligned} \ln m(\mathbf{y})_{M_S} &\approx \ln \hat{m}(\mathbf{y})_{M_S} \\ &= \ln p_S(\mathbf{y}|\Theta_S^*) + \ln \pi(\Theta_S^*) - \ln \pi_S(\Theta_S^*|\mathbf{y}), \end{aligned} \tag{2.14}$$

in which Θ_S^* is an arbitrary parameter set for model M_S and $\pi_S(\Theta_S^*|\mathbf{y})$ is the posterior density estimate at Θ_S^* under model M_S. However, for estimation efficiency, it is better to choose a Θ_S^* that has high density under the posterior, and we usually choose the mean of the estimated posterior distribution.

More generally, if the estimated parameter set Θ_S can be decomposed into blocks $\Theta_{S,1}, \Theta_{S,2}, \ldots, \Theta_{S,K}$, the estimator can be written as

$$\begin{aligned} \ln \hat{m}(\mathbf{y})_{M_S} &= \ln p_S(\mathbf{y}|\Theta_S^*) + \ln \pi(\Theta_S^*) \\ &\quad - \sum_{r=1}^{K} \ln \pi_S(\Theta_{S,r}^*|\mathbf{y}, \Theta_{S,h}^*, h = 1, \ldots, r-1), \end{aligned} \tag{2.15}$$

where $\pi_S(\Theta_{S,r}^*|\mathbf{y}, \Theta_{S,h}^*, h = 1, \ldots, r-1)$ is the posterior density of $\Theta_{S,r}^*$ given $\Theta_{S,h}^*, h = 1, \ldots, r-1$.

Notice that under a hidden Markov model, the posterior density is

$$\pi_S(\Theta_S^*|\mathbf{y}, M_S) = \int \pi(\Theta_S^*|\mathbf{y}, \mathbf{s}) p(\mathbf{s}|\mathbf{y}) d\mathbf{s},$$

where $\mathbf{s}$ is the vector of estimated hidden states. Chib (1995) proposed that the posterior density can be estimated by the following expression, which converges to the true value as G goes to infinity.

$$\begin{aligned} &\pi_S(\Theta_{S,r}^*|\mathbf{y}, \Theta_{S,h}^*(h < r)) \\ &\quad \approx \hat{\pi}_S(\Theta_{S,r}^*|\mathbf{y}, \Theta_{S,h}^*(h < r)) \\ &\quad = \frac{1}{G} \sum_{j=1}^{G} \pi_S(\Theta_{S,r}^*|\mathbf{y}, \Theta_{S,1}^*, \ldots, \Theta_{S,r-1}^*, \mathbf{s}^{(j)}, \Theta_{S,r+1}^{(j)}, \ldots, \Theta_{S,K}^{(j)}). \end{aligned} \tag{2.16}$$

However, notice that the $\mathbf{s}^{(j)}, \Theta_{S,r+1}^{(j)}, \ldots, \Theta_{S,K}^{(j)}$ here are drawn from the sub-model fixing $\Theta_{S,1}^*, \ldots, \Theta_{S,r-1}^*$, instead of the original full model.

The calculation of the sample density $p_S(\mathbf{y}|\Theta_S^*)$ will use the forward step which calculates $\pi_{t-1}(s) = p(s_{t-1} = s|y_1^{t-1}, \Theta_S^*, M_S)$ mentioned in Section 2.2 as follows:

$$\begin{aligned} p_S(\mathbf{y}|\Theta_S^*) &= \prod_{t=1}^{T} p(y_t|y_1^{t-1}, \Theta_S^*, M_S) \\ &= \prod_{t=1}^{T} \left[\prod_{r=1}^{S} p(s_t = r|y_1^{t-1}, M_S) p(y_t|s_t = r, \Theta_S^*, M_S) \right] \\ &= \prod_{t=1}^{T} \left\{ \prod_{r=1}^{S} \left[\left(\sum_{s=1}^{S} q(s,r)\pi_{t-1}(s) \right) p(y_t|s_t = r, \Theta_S^*, M_S) \right] \right\}. \end{aligned} \tag{2.17}$$

Chib (1995) showed that this estimator numerically has small standard error for HMM, which can be regarded as a accurate estimator for the marginal likelihood.

2.4.2 Reversible Jump Markov Chain Monte Carlo

Another method of choosing S is what is called a variable-dimension Monte Carlo (VDMC). The idea is to sample S from its conditional distribution $p(S|\Theta^{(t)}, \mathbf{y}, \mathbf{s}^{(t)})$ in each iteration of the MCMC so that the chain will reach the stationary distribution $p(S|\mathbf{y})$. The number of VDMC methods has recently grown, to include jump diffusion (Grenander and Miller 1994), reversible jump Markov chain Monte Carlo (RJMCMC; Green 1995) and birth-death sampler (Stephens 2000a; Cappé et al. 2001).

We will use the RJMCMC algorithm in this dissertation. Briefly speaking, the RJMCMC is a Metropolis-Hastings method: conditional on all the other estimated parameters, the algorithm proposes a way to change the currently state space Θ_S to $\tilde{\Theta}_S$ by an arbitrary proposal density $q(\Theta_S, \tilde{\Theta_S})$. The move is then accepted with

probability

$$\min\left\{1, R = \frac{\pi(d\tilde{\Theta}_S)q(\tilde{\Theta}_S, d\Theta_S)}{\pi(d\Theta_S)q(\Theta_S, d\tilde{\Theta}_S)}\right\}$$

where $\pi(d\Theta_S)$ is the target probability to sample the parameter set Θ_S. The move from Θ_S to $\tilde{\Theta}_S$ either increases or decreases the dimension.

The reason that the algorithm is called reversible jump MCMC is that the moves are formed as reversible pairs: if there is a move that increases the parameter space, there is another one that decreases it back accordingly. To choose the movement for decreasing and increasing the parameter space, one should consider not only dimension matching, but also the practicability of the movement as well as the acceptance rate (efficiency). If the efficiency of the reversible jump is very low, the Markov chain will converge slowly and have a high autocorrelation. However, if one could find a proposal movement that ensures the new state space, including the new states, has a similar posterior support as the current one, then both the movement and its reverse will have a reasonable acceptance rate (Green 1997). There are some common move types such as the birth-death and the split-merge (Green 1995).

- **Birth-Death Moves**

  In this dissertation, we will use birth and death moves, which involve the birth or death of an empty state. In each step of the MCMC, a birth move or its reverse, a death move, is randomly determined with a birth rate b_S and a death rate $d_S = 1 - b_S$ if the current number of states is S. Notice that $d_1 = b_{\mathcal{S}} = 0$ and death can only happen when there is at least one empty state.

  1. If a birth is proposed and the current number of states is S, a new state j^*

is generated, and the parameter vector is changed from a S-dimensional Θ_S to a $S+1$-dimension $\tilde{\Theta}_{S+1}$ by inserting the new element Θ^{j^*} from proposal function $\Theta^{j^*} \sim p(\Theta^*)$. The value of j^* is determined by a restriction of the parameter space and other state indices are reassigned accordingly. The transition matrix Q is changed to $\tilde{Q}$ by inserting a j^*th row $\tilde{q}(j*,.)$ from the Dirichlet distribution $\tilde{q}(j*,.) \sim \mathcal{D}_{S+1}(\vec{\alpha})$ and by changing other elements according to equations (2.18) below, for any $j \neq j^*$ and $j = 1, \ldots, S$,

$$\begin{aligned} \tilde{q}(j, j^*) &= u(j) \sim p(u(j)); \\ \tilde{q}(j, r) &= [1 - u(j)]q(j, r),\ r \neq j^*; \end{aligned} \tag{2.18}$$

The labelling of the hidden state chain will change according to the following prescription: for those time occasions whose state in rank is before the new state, the state label of that point will stay the same, while for those points whose rank is after the new state, the state label of that point will increase by one.

2. If a death is proposed, one empty state j^* in the image is randomly chosen to die and the state indices are reordered accordingly. The whole j^*th row and column of Q are deleted and each row is normalized to sum to 1. The hidden state labels are also changed accordingly: For those time occasions whose state ranked is before the death state, the state number will stay the same, while for those after the death state, the state number will decrease by one.

3. Finally, the acceptance ratio for this birth move is

$$\min\left\{1, R = \frac{\pi(d\tilde{\Theta}_S)q(\tilde{\Theta}_S, d\Theta_S)}{\pi(d\Theta_S)q(\Theta_S, d\tilde{\Theta}_S)}\right\},$$

while for a death move it is $\min\{1, R^{-1}\}$ (Green 1995).

- **Split-Merge Moves**

The split and merge moves randomly choose to split a state or merge two adjacent existing ones with probability sp_S and $co_S = 1 - sp_S$ if the current number of states is S. Naturally $sp_{\mathcal{S}} = co_1 = 0$. Since we will not use the split and merge moves in our project, we will only give a brief introduction here.

1. If a merge is proposed, a randomly chosen adjacent pair of states (j_1, j_2) will be merged into a single new state j^*, thus decreasing the current state space by 1. The parameters and the hidden state chain are changed accordingly to keep the move reversible.

2. If a split is proposed, a randomly chosen state j^* will be split into two new components j_1, j_2 according to the algorithm described in Robert et al. (2000). The hidden chain is also changed using a restricted backward algorithm described in the appendix of Robert et al. (2000).

3. Then the acceptance rate is calculated to determine whether accept this move or not.

2.4.3 Augmented Birth

The RJMCMC algorithm is known to have low acceptance rates (Green 1995), since it's usually hard to find a good move to increase or decrease the dimension of the parameter space while ensuring that the newly proposed state has similar posterior support as the currently existing states. Researchers have been working to find possible modification to increase the acceptance rates. There are some interesting and important developments, such as *order methods* (Brooks et al. 2003), which constructs a move that is based on roughly knowing where is a good place to jump to and then imposing various constraints on the acceptance rate; the *saturated space approach* (Brooks et al. 2003), which augments the state space with auxiliary variables to ensure that all models share the same largest dimension so that when making a move, the information can be retained and a jump is proposed to a sensible place; the *adaptive MCMC* (Roberts and Rosenthal 2006), which uses the idea that under certain suitable conditions, the proposal function or mechanisms can depend on past realisations of the chain instead of only depending on the current state. In other words, the proposal mechanisms can be determined online during the run of the MCMC and the history of the chain may be used to make inferences about the targeting distribution. Finally, the *delayed rejection* (Tierney and Mira 1999; Green and Mira 2001), proposes a second attempt if the first move is rejected while still maintaining the balance of the compound transition. Some other methods are introduced by Al-Awadhi et al. (2004), Tjelmeland and Hegstad (2001) and (Brooks et al. 2003).

In this dissertation we will use the idea of the augmented birth, since for our synthetic data and in the applications. We found that in our MCMC chain, the death move works well, but not the birth move, which is hardly ever accepted. An extreme example is: if we start from a very large number of states, the MCMC chain quickly decreases to the true number of states, and the value S stays there, but if we start from $S_0 = 2$, the chain does not move up for a long time,if at all. The reason is that for the death move, an empty state, which is an unlikely state, is removed. Thus the move is in itself moving to a more likely state space and thus the acceptance rate is high. For the birth move, however, the new state is randomly generated from the uninformative prior, so there is a very high possibility that the new state space is less likely and the move will be rejected.

However, since we know that for a MCMC chain, the burn-in period should not influence the final stationary distribution, one can propose more moves in the burn-in period: apart from the regular birth and death moves. We propose an augmented birth according to a certain probability in each iteration of the MCMC chain to induce more mixing (especially more births) in the burn-in period. The augmented birth is always accepted, which will make the MCMC chain non-reversible. However, the algorithm needs to approach the true MCMC chain and thus have the same stationary distribution (target distribution) as the original RJMCMC chain. Thus, the probability of proposing an augmented birth should decrease to 0 as the number of iterations increases. In practice, we choose the probability for the augmented birth as follows:

$$P_b(n) \to 0 \text{ in probability 1 as } n \to \infty,$$

where n is the iteration number. An extreme example of possible density is a step function $P_b(n) = c$ when $n \leq 0$ and $P_b(n) = 0$ for $n > n_0$, where $c < 1$ is a small number and n_0 is an integer. Other examples can be types of functions that decrease at rate of n^{-2} or higher.

Chapter 3

Synthetic Data Analysis

In order to check the accuracy of the reversible jump hidden Markov algorithm and the computer code, we simulated synthetic datasets based on the model assumptions and ran the algorithm to check the results. Section 3.1 describes how the two datasets were generated, Section 3.2 shows the model calculation and simulation, and Section 3.3 summarizes the estimation results and compares then with the true values.

3.1 Data Generation

The synthetic data is generated according to the following steps, based on the model assumptions.

1. Generate a 4 state $1,000$ time-point hidden Markov chain based on the properties of a time-homogeneous first order Markov process. That is, randomly draw the stating state s_1 from the starting distribution $\vec{\pi}$ and sequentially draw the following states s_t conditional on the last immediate state s_{t-1} according to the transition matrix Q.

2. Conditional on the generated hidden states s_t, each observed value $\vec{y}_t$ is generated from a bivariate normal distribution. That is, $\vec{y}_t$ is generated from $\mathcal{N}_2(\vec{\mu}_s, \Sigma_s)$ if $s_t = s$.

Here we simulate two datasets. For the first dataset, the observed outcomes $\mathbf{y}$ is isolated from each other by assigning small variance-covariance matrices Σ_s, while for the second dataset the observed outcomes $\mathbf{y}$ are mixed by assigning large variance-covariance matrices Σ_s. Table 3.1 gives the true parameters of the two datasets and Figure 3.1 plots the two datasets.

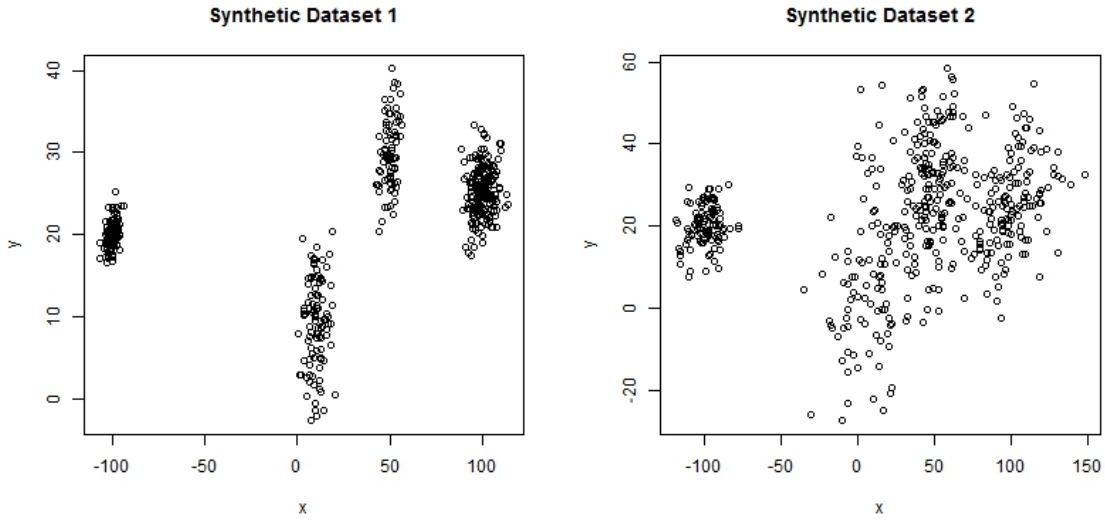


Figure 3.1: Synthetic datasets. The left panel shows the isolated dataset, while the right panel represents the mixed dataset.

3.2 Model description

Let $\mathbf{g}_{obs} = \mathbf{y}$, $\mathbf{g}_{mis} = \mathbf{s}$, $\Theta_S = \{\vec{\mu}_s, \Sigma_s, Q, \vec{\pi} : s = 1, \ldots, S\}$ be the parameter set. Then the joint likelihood function, similar to Equation (2.3) can be written as

$$p(\mathbf{g}_{obs}, \mathbf{g}_{mis}, S|\Theta_S) = p(S)\vec{\pi}(s_1)\mathcal{N}_2(\vec{y}_1|\vec{\mu}_{s_1}, \Sigma_{s_1}) \prod_{t=2}^{1000} \{q(s_{t-1}, s_t)\mathcal{N}_2(\vec{y}_t|\vec{\mu}_{s_t}, \Sigma_{s_t})\} . \quad (3.1)$$

Here we give independent conjugate priors to each component of the parameter

Table 3.1: True parameters for the synthetic datasets.

	Dataset 1	Dataset 2
$\vec{\pi}$	$(0.25, 0.25, 0.25, 0.25)'$	$(0.25, 0.25, 0.25, 0.25)'$
Q	$\begin{pmatrix} 0.6 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.5 & 0.1 & 0.3 \\ 0.2 & 0.3 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{pmatrix}$	$\begin{pmatrix} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{pmatrix}$
$\vec{\mu}$	$\vec{\mu}_1$ $\vec{\mu}_2$ $(-100, 20)'$ $(10, 10)'$ $\vec{\mu}_3$ $\vec{\mu}_4$ $(50, 30)'$ $(100, 25)'$	$\vec{\mu}_1$ $\vec{\mu}_2$ $(-100, 20)'$ $(10, 10)'$ $\vec{\mu}_3$ $\vec{\mu}_4$ $(50, 30)'$ $(100, 25)'$
Σ	Σ_1 Σ_2 $\begin{pmatrix} 6 & 1 \\ 1 & 3 \end{pmatrix}$ $\begin{pmatrix} 20 & 4 \\ 4 & 30 \end{pmatrix}$ Σ_3 Σ_4 $\begin{pmatrix} 10 & 5 \\ 5 & 20 \end{pmatrix}$ $\begin{pmatrix} 20 & 0 \\ 0 & 10 \end{pmatrix}$	Σ_1 Σ_2 $\begin{pmatrix} 60 & 10 \\ 10 & 30 \end{pmatrix}$ $\begin{pmatrix} 200 & 40 \\ 40 & 300 \end{pmatrix}$ Σ_3 Σ_4 $\begin{pmatrix} 100 & 50 \\ 50 & 200 \end{pmatrix}$ $\begin{pmatrix} 200 & 50 \\ 50 & 100 \end{pmatrix}$

set Θ_S. That is, for any $s = 1, \ldots, S$, the prior for $\{\vec{\mu}_s, \Sigma_s\}$ is the conjugate Normal-Inverse-Wishart distribution:

$$p(\vec{\mu}_s, \Sigma_s) = p(\vec{\mu}_s|\Sigma_s)p(\Sigma_s) = \mathcal{N}_2(\vec{\mu}_s|\vec{m}_0, \frac{\Sigma_0}{\lambda_0}) \times \mathcal{W}_2^{-1}(\Sigma_s|V_0, \kappa_0),$$

where $\mathcal{W}_2^{-1}$ is the two-dimensional inverse Wishart distribution, and $\vec{m}_0, \lambda_0, V_0$ and κ_0 are hyper-parameters. For $\vec{\pi}$ and each row $q(s, \cdot)$ of Q, the priors are the $S-$dimensional Dirichlet distribution $\mathcal{D}_S(\mathbf{1})$. The prior for the number of states S is uniform on $\{1, \ldots, \mathcal{S}\}$. Therefore the joint likelihood can be written as

$$\begin{aligned}
& p(\mathbf{g}_{obs}, \mathbf{g}_{mis}, S, \Theta_S|\Phi) \\
= \; & \frac{1}{\mathcal{S}}\vec{\pi}(s_1)\mathcal{N}_2(\vec{y}_1|\vec{\mu}_{s_1}, \Sigma_{s_1}) \prod_{t=2}^{1000} \{q(s_{t-1}, s_t)\mathcal{N}_2(\vec{y}_t|\vec{\mu}_{s_t}, \Sigma_{s_t})\} \\
& \times \left\{\mathcal{D}_S(\vec{\pi}|\mathbf{1}) \prod_{s=1}^{S} \mathcal{D}_S(q(s, \cdot)|\mathbf{1})\right\} \prod_{s=1}^{S} \left\{\mathcal{N}_2(\vec{\mu}_s|\vec{m}_0, \frac{\Sigma_0}{\lambda_0})\mathcal{W}_2^{-1}(\Sigma_s|V_0, \kappa_0)\right\} \\
= \; & \frac{1}{\mathcal{S}} \times \vec{\pi}(s_1)\mathcal{D}_S(\vec{\pi}|\mathbf{1}) \\
& \times \prod_{s=1}^{S} \left\{ \prod_{\{t:t-1=s\}} q(s_{t-1}, s_t)\mathcal{D}_S(q(s, \cdot)|\mathbf{1}) \right\} \\
& \times \prod_{s=1}^{S} \left\{ \prod_{\{t:s_t=s\}} \mathcal{N}_2(\vec{y}_t|\vec{\mu}_s, \Sigma_s)\mathcal{N}_2(\vec{\mu}_s|\vec{m}_0, \frac{\Sigma_0}{\lambda_0})\mathcal{W}_2^{-1}(\Sigma_s|V_0, \kappa_0) \right\} \\
\propto \; & \frac{1}{\mathcal{S}} \left\{ \left(\prod_{s=1}^{S} \pi(s)^{n_0(s)} \right) \mathcal{D}_S(\pi|\mathbf{1}) \right\} \\
& \times \prod_{s=1}^{S} \left\{ \left(\prod_{r=1}^{S} q(s, r)^{n(s,r)} \right) \mathcal{D}_S(q(s, \cdot)|\mathbf{1}) \right\} \\
& \times \prod_{s=1}^{S} \left[\mathcal{N}_2\left(\vec{\mu}_s|M_s, G_s\right) \mathcal{W}_2\left(\Sigma_s^{-1}|D_s\right) \right], \qquad (3.2)
\end{aligned}$$

with the help of following notations:

$$\begin{aligned}
\Phi &= \{\vec{m}_0, \lambda_0, V_0, \kappa_0, \mathcal{S}\}, \\
n_0(s) &= I(s_1 = s),
\end{aligned}$$

$$
\begin{aligned}
n(s,r) &= \sum_{t=2}^{1000} I(s_{t-1}=s, s_t=r), \\
c_s &= \sum_{t=1}^{1000} I(s_t=s), \\
\vec{a}_s &= \frac{1}{c_s}(\sum_{t:s_t=s} \vec{y}_t), \\
B_s^2 &= \sum_{t:s_t=s} (\vec{y}_t - \vec{a}_s)(\vec{y}_t - \vec{a}_s)', \\
M_s &= (\lambda_0 + c_s)^{-1}(\lambda_0 \vec{m}_0 + c_s \vec{a}_s), \\
G_s &= (\lambda_0 + c_s)^{-1}\Sigma_s, \\
D_s &= \left\{[V_0^{-1} + B_s^2 + \lambda_0 c_s(\lambda_0 + c_s)^{-1}(\vec{a}_s - \vec{m}_0)(\vec{a}_s - \vec{m}_0)']^{-1}, \kappa_0 + c_s\right\}.
\end{aligned}
$$

Based on Equation 3.2, we can use the MCMC algorithm described before to iteratively simulate parameters and missing values as follows.

1. Generate Θ_S from the following independent posterior distribution:

$$
\begin{aligned}
\vec{\pi} &\sim \mathcal{D}_S(\vec{n}_0 + \mathbf{1}), \\
q(s,\cdot) &\sim \mathcal{D}_S(n(s,\cdot) + \mathbf{1}), \\
\{\vec{\mu}_s, \Sigma_s\} &\sim \mathcal{N}_2\left(\vec{\mu}_s | M_s, G_s\right) \mathcal{W}_2\left(\Sigma_s^{-1} | D_s\right), s = 1, \ldots, S
\end{aligned}
$$

2. Generate the hidden state $\mathbf{s}$ from by the Forward-Backward algorithm;

3. Increase or decrease the number of states S using the birth and death moves:

   (a) The birth rate b_s and death rate d_s all equals 0.5 for $s = 2, \ldots, 5$ and $b_1 = d_6 = 1$.

   (b) If we propose to generate a birth and the current number of states is S, the new state j^* is generated from the prior

$$
\{\vec{\mu}_{j^*}, \Sigma_{j^*}\} \sim \mathcal{N}_2(\vec{\mu} | \vec{m}_0, \frac{\Sigma_0}{\lambda_0}) \mathcal{W}_2^{-1}(\Sigma | V_0, \kappa_0),
$$

where j^* is determined by the x-coordinates of $\vec{\mu}$. The new row $q(j^*, \cdot)$ is generated from $\mathcal{D}_{S+1}(\mathbf{1})$, and for each $j \neq j^*$, $u(j)$ is simulated from the beta distribution $\mathcal{B}(1, cS)$. The acceptance rate is equal to

$$\min\left\{1, R = \frac{(S+1)d_{S+1}}{(k_0+1)b_S c^{S+1}} \prod_{j \neq j^*} (1-u(j))^{d_j-(c-1)S}\right\},$$

where k_0 is the number of empty states in the current iteration. The detailed calculation is shown in Appendix A.1.

(c) If we propose to generate a death, one of the empty states j^* is randomly chosen to die and other states are relabelled. $\{\vec{\mu}_{j^*}, \Sigma_{j^*}\}$ will be deleted, the j^*th row and column of Q will also be deleted and the left components of Q are normalized to satisfy a transition matrix property. The acceptance rate is $\min\{1, R^{-1}\}$.

4. In each step, we artificially add a birth following the exact birth moves described in 3(b). The augmented births are proposed according to the probability

$$P_{1b}(n) = \begin{cases} 0.001 & n <= 4000 \\ 16000/n^2 & n > 4000. \end{cases}$$

and

$$P_{2b}(n) = \begin{cases} 0.005 & n <= 4000 \\ 80000/n^2 & n > 4000. \end{cases}$$

3.3 Summary of Results

In both of the MCMC chains, we use the following hyper-parameters: $m_0 = (20, 10)'$, $\lambda_0 = 0.01$, $V_0 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$, $\kappa_0 = 5$, $c = 15$ and $\mathcal{S} = 6$. In the mixed data, the parameters are more difficult to identify than in the isolated data. Therefore we run $10,000$ iterations for dataset 1 with a burn-in of $6,000$ iterations, and run $28,000$ iterations for dataset 2 with a burn-in $20,000$ iterations. For both chains we keep every 1 out of 5 iterations after the burn-in period. Both chains have converged after the burn-in.

Figure 3.2 shows the number of states for the two datasets. Both of them reach the true number of states, 4, after the burn-in period, which indicates that our algorithm can identify the correct number of states.

The convergence of the MCMC chain can also be shown by the estimation plots for $\vec{\mu}$ as well as the autocorrelation plot after the burn-in period. Here we will only show the estimation plots of $\vec{\mu}$ for dataset 1 (Figure 3.3) and the auto-correlation after the burn-in period (Figure 3.4). The same plots for dataset 2 showed that the MCMC chain are noisy in the burn-in period, but eventually settles done to the true values. The plots are included in the Appendices (Figure A.1 and Figure A.2). It shows that

Table 3.2 shows the estimated $\vec{\mu}$ under both data sets. We can see that both of the two estimates are quite close to the true value, which suggests good performance of the model. In Figure 3.5 we show the estimated 80% credible regions for each state based on the estimated variance-covariance matrices. We can see that for both

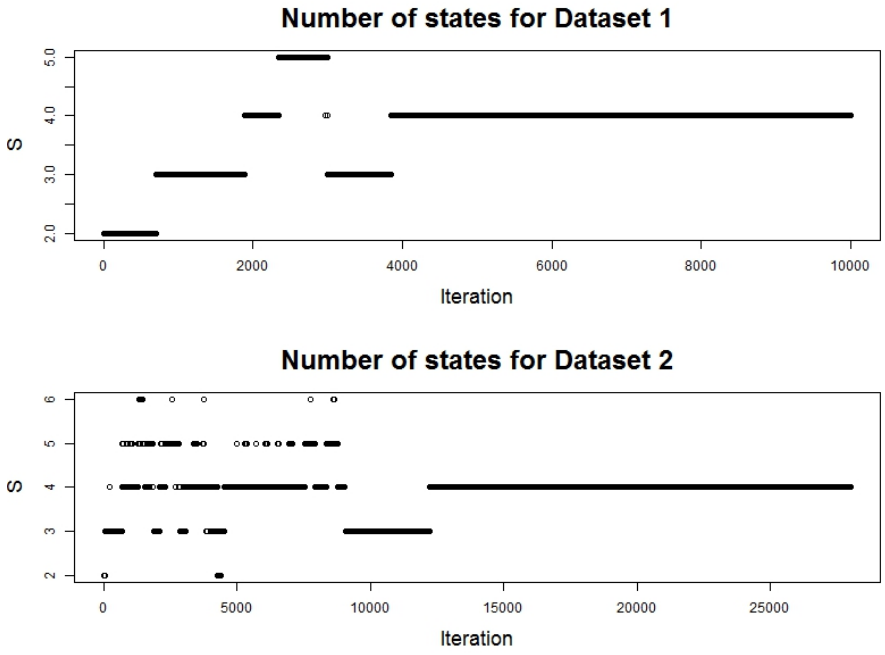


Figure 3.2: Estimated number of states for each dataset again the iterations of the MCMC chain.

datasets, the points are clearly assigned to different states.

Finally, for the isolated dataset (dataset 1), the estimated hidden states are 100% identical to the true hidden states generated from the Markov chain, while for the mixed dataset (dataset 2), 98.2% of the estimated hidden states are exactly equal the true state $\mathbf{s}$, which suggests accurate estimation of the hidden states.

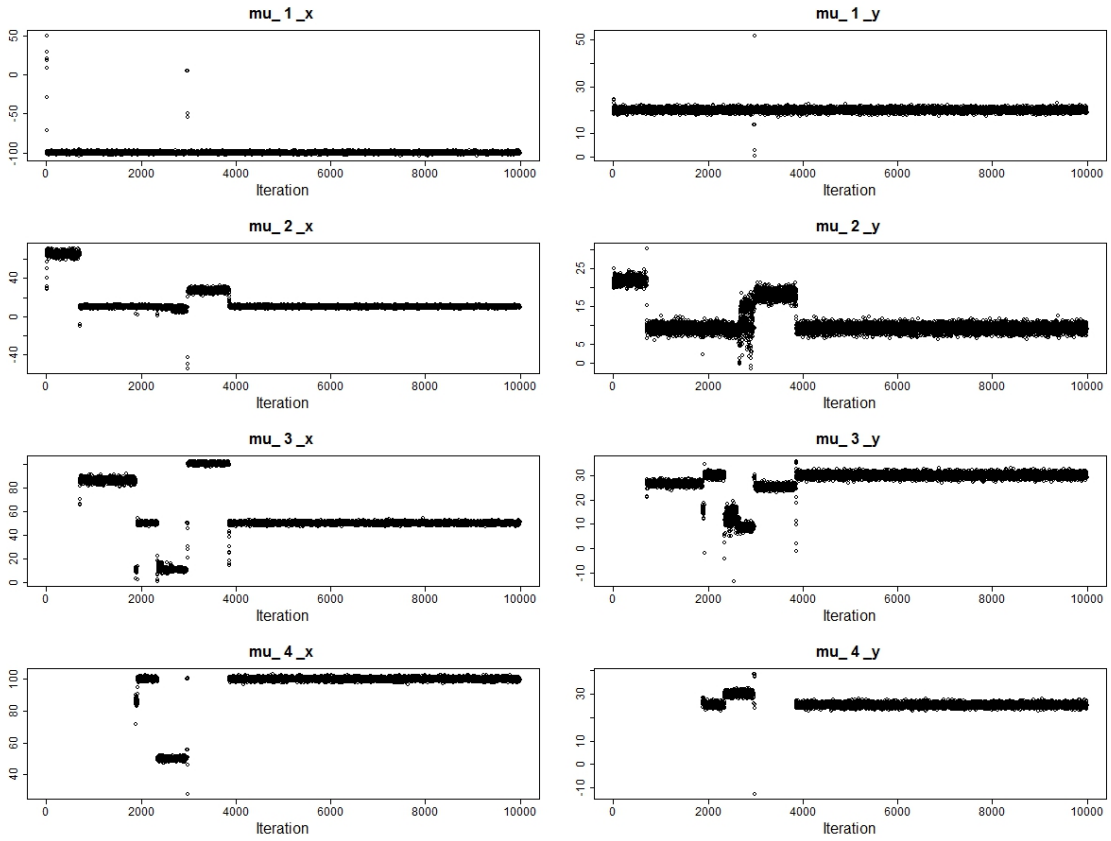


Figure 3.3: Simulation results of $\vec{\mu}$ for dataset 1. From the plot we can see that there is no evidence to say that the MCMC chain did not reach the stationary distribution after 6,000 iterations (burn-in) period.

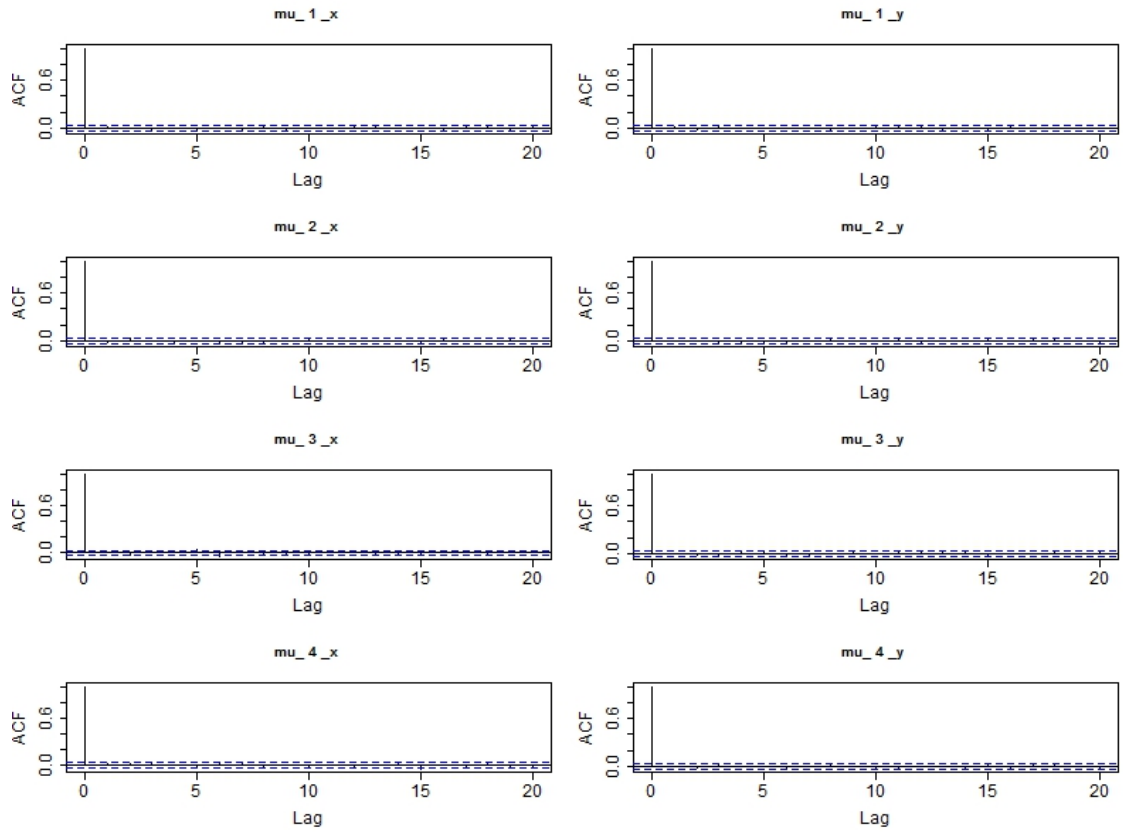


Figure 3.4: Autocorrelation plot for each component of $\vec{\mu}$ for dataset 1 after the burn-in period. From the graph we can see there is almost no autocorrelation for the estimation after the burn-in period.

Table 3.2: Estimated $\vec{\mu}$ for both datasets.

True $\vec{\mu}$	$\vec{\mu}_1$	$\vec{\mu}_2$	$\vec{\mu}_3$	$\vec{\mu}_4$
	$\begin{pmatrix} -100 \\ 20 \end{pmatrix}$	$\begin{pmatrix} 10 \\ 10 \end{pmatrix}$	$\begin{pmatrix} 50 \\ 30 \end{pmatrix}$	$\begin{pmatrix} 100 \\ 25 \end{pmatrix}$
Dataset 1	$\vec{\mu}_1$	$\vec{\mu}_2$	$\vec{\mu}_3$	$\vec{\mu}_4$
	$\begin{pmatrix} -100.05 \\ 19.99 \end{pmatrix}$	$\begin{pmatrix} 10.18 \\ 9.10 \end{pmatrix}$	$\begin{pmatrix} 49.91 \\ 29.97 \end{pmatrix}$	$\begin{pmatrix} 100.05 \\ 25.23 \end{pmatrix}$
Dataset 2	$\vec{\mu}_1$	$\vec{\mu}_2$	$\vec{\mu}_3$	$\vec{\mu}_4$
	$\begin{pmatrix} -99.39 \\ 20.29 \end{pmatrix}$	$\begin{pmatrix} 9.82 \\ 8.44 \end{pmatrix}$	$\begin{pmatrix} 48.62 \\ 30.46 \end{pmatrix}$	$\begin{pmatrix} 99.37 \\ 25.10 \end{pmatrix}$

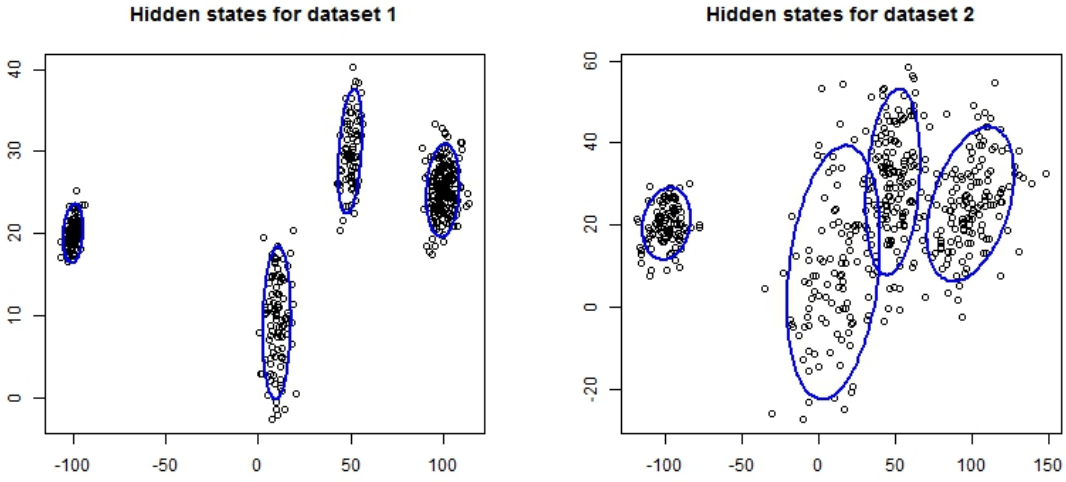


Figure 3.5: Estimated 80% credible ellipse for each state on both datasets.

Chapter 4

A Reversible Jump Hidden Markov Model Analysis of Search for Cancerous Nodules in X-ray Images

4.1 Introduction

Radiology has witnessed revolutionary changes in the last two decades, due to the development of medical imaging and medical image processing technologies. With these developments, skills in searching for anomalies in medical images have become an ever more critical component of the expertise of radiologists. An increasing amount of research therefore attempts to better understand these skills, with the aim of improving the computerized processing and display of medical images, and the effectiveness with which radiologists search them for abnormalities.

The search for potentially cancerous nodules on chest X-ray images is an important application area. General estimates suggest that in lung screening programs 76% of pulmonary nodules, which can represent lung cancer, may be overlooked in chest X-rays (Nodine and Kundell 1987). Lewellyn-Thomas (1963) was one of the first to record radiologists' eye movements to study the problem of errors in the detection of nodules in X-ray images. The eye movements that humans make while trying to detect these nodules provide insights into the search process that cannot be otherwise easily obtained (Zelinsky 2008), and they can be unobtrusively recorded

through modern eye tracking equipment. Several studies have used eye movement data to investigate which image and nodule characteristics contribute to correct and incorrect detections (DeValk and Eilkman 1984), whether visual search patterns can reveal the nature of errors (Nodine and Kundel 1987), and how radiologists find nodules as well as how this is affected by experience and training (Krupinski 2000). These studies show that features such as nodule size, location and conspicuity have little or no effect on time to detection, but size and conspicuity of the nodule have been found to affect times until a nodule is first detected (Krupinski et al. 2003). These studies have also shown that there is high variability between radiologists in their ability to detect nodules (Krupinski 2000), but the reasons are not yet well understood. Further, the results of these studies have documented that search is more random than systematic, and that prolonged search tends to lead to more errors (Oestmann et al. 1988; Gur et al. 1991). Nonetheless, optimal strategies focus attention on areas where nodules are most likely to be located (DeValk and Eilkman 1984; Kundel 1974).

Much of eye tracking research relies on the definition of so called areas of interest (AOIs) on the image, used to classify the eye movements. There are two main types of eye movements: fixations and saccades. During an eye fixation the eye is relatively still and visual information around the point of regard on the image is extracted and projected onto the fovea, which is a small area of the retina with high acuity, covering about two degrees of visual angle. The duration of fixations ranges from 100-500 milliseconds (Rayner 1978), and is mostly beyond our conscious control. During a saccade the eye is rapidly redirected to other potentially infor-

mative regions on the image. The duration of saccades is typically in the range of 20-50 milliseconds (Rayner 1978), and they can be both involuntary and voluntary. During a saccade vision is suppressed. A sequence of fixations and saccades on an image is called a scan-path. The definition of AOIs allows eye tracking researchers to investigate what aspects of an image observers inspect, by observing how long they dwell within an AOI and how they move from one AOI to another. The location and shape of AOIs are most often a priori defined based on hypotheses of the researcher, but in some studies they are based on image segmentation and manual coding or clustering to find regions with a high density of fixation points (see for a review, Holmqvist et al. 2011). For chest X-rays, the nodule typically engenders one such AOI. Despite their potential to provide a better understanding and more accurate inferences on the search process, statistical models of eye movements that identify areas of interest that are important from the perspective of the observer have not been developed. Here, we propose such a model and apply it to describe the search for nodules in chest X-rays. We build on the prior work by Van der Lans, Pieters and Wedel (2008), who developed a hidden Markov model that describes fixation-patterns at the level of the pixels in an image. Similar to their model, we treat the observed eye movements as a two-dimensional spatial Poisson process, but we allow eye fixations to cluster in small areas on the image that might potentially represent a nodule. These AOIs are unknown in number, position, size and orientation. To analyze data from an eye tracking experiment in which sixteen participants search for nodules on seven images that contained a nodule, we use a reversible jump hidden Markov model to estimate not only the characteristics of the AOIs, but also

their total number. After identifying the AOIs with our model, we study the effect of scan patterns and image characteristics on the ability of participants to detect a nodule.

4.2 Data Description

The data were collected on participants in the eye tracking lab of a major US University [name withheld for reasons of confidentiality]. The data used in this paper are part of a larger study. The participants are undergraduate students who volunteered to participate for course credit. They were instructed about the purpose and duration of the study. The instructions and images were presented on 21-inch LCD monitors. Images were shown full screen at a resolution of $1,024 \times 768$ pixels (approximately $15.5^\circ \times 19.4^\circ$ of visual angle) in black and white. Participants seated in front of the monitor were first exposed to an annotated X-ray image of a chest to familiarize themselves with various structures (heart, trachea, hilum, diaphragm, spine, ribs and lungs). Then, to provide them with some training in finding nodules, they were exposed to 14 X-ray images of which seven did and seven did not contain a potentially cancerous nodule. They were instructed to search for a possible nodule, present as a bright round shape in the image, and take as much time as needed for each image. After their search of an image was complete, they were provided with feedback on the presence and location of the nodule.

Next, participants were exposed to 14 testing X-ray images and were asked to search for a nodule, possibly present on the lungs. Among the 14 X-ray images,

seven contained a single nodule and seven did not. However, the participants do not know this information. Figure 4.1 shows an example of an X-ray image with a nodule present. Other X-ray images are contained in Appendix B, Figures B.1-B.3. Notice that the nodule is difficult to detect, which is even the case for trained radiologists. The images were presented in random order and participants were asked to search for a single nodule on each of them. During their search of each image, participants' eye movements were recorded with infrared corneal reflection eye tracking methodology. The eye-tracker registers the point of regard every 50 milliseconds with a spatial resolution better than 0.5° and measures the reflection of infrared rays emitted by the eye-tracker on the cornea (the hard outer layer) of both eyes. The eye tracking equipment allows participants to freely move their heads within a virtual box of about 20 inches, while miniature cameras track the position of the head and eyes.

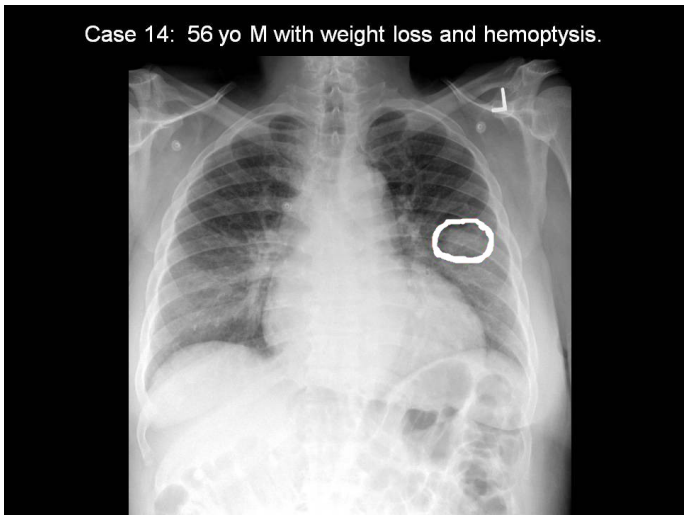


Figure 4.1: One X-ray image with one nodule present (case 14). The circle marks the nodule (not shown on the test images).

As introduced before, eye movements on stationary scenes consist primarily of fixations and saccades (Rayner 1978). In the present study, fixations of individual participants, measured in pixel coordinates, are the unit of analysis. From the raw data produced by the eye tracking equipment, eye-fixation positions were computed by the algorithm of Van der Lans et al. (2011). This algorithm uses information on the point of regard of both eyes to derive fixation locations in pixel coordinates and filters out blinks and other anomalies from the eye movement traces. Participants who had less than 40 points-of-regard recorded during exposure were removed from the study. This resulted in a final set of 16 participants with usable data (one participant missing case 11). When the participants believed they had reached a decision on the presence of a nodule in an image, they were asked to fixate the location in question for about 2 seconds and click on a button on the bottom right of the screen to proceed to the next image. They were also asked to record whether they believe that there is a nodule. In the present study, to illustrate the model, the fixation sequences of 16 participants on seven X-ray images which have a nodule present will be used for analysis (Cases 1, 3, 9, 10, 11, 12, 14). Figure 4.2 provides the scan-path pattern for all 16 participants on case 14, which is shown in Figure 4.1. Figure 4.2 illustrates the heterogeneity of the scan-path across participants, and shows that they search between 40 seconds and almost 4 minutes for a nodule. It highlights the need to use model-based inference to draw generalizable conclusions about search behavior for nodules from these data. The same scan-path plots are included in Appendix B, Figure B.4-B.6

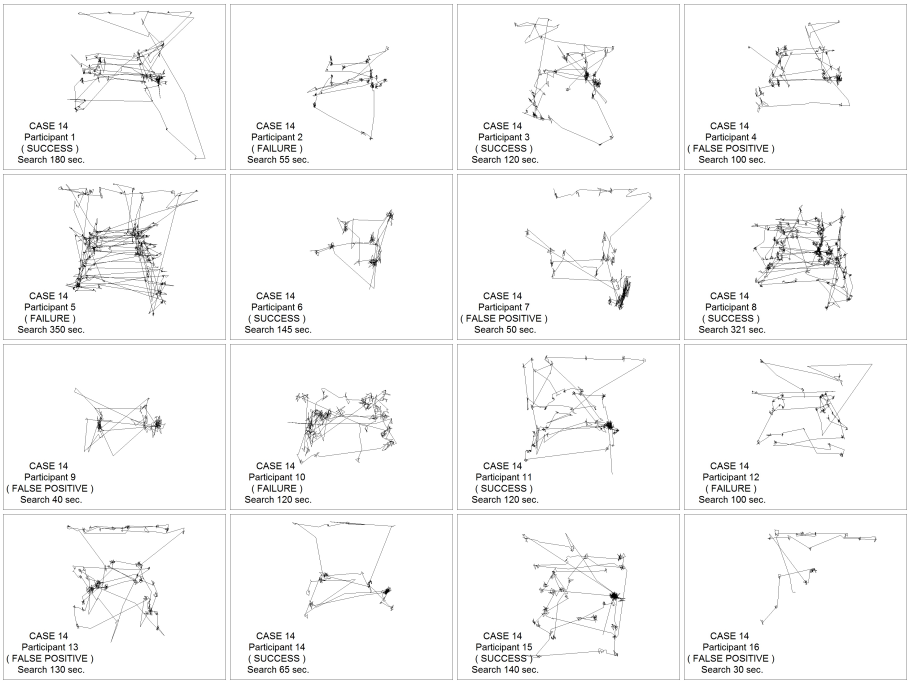


Figure 4.2: Scan-paths for 16 participants (case 14).

4.3 The Reversible Jump Hidden Markov Model and MCMC

In this section we describe the Reversible Jump Hidden Markov Model (RJHMM) used to identify AOIs on a given X-ray image from participants' eye tracking data. The AOIs are assumed to be common to all participants. The number of AOIs is unknown, and they have unknown centers, sizes and orientations. Participants' eyes jump among these AOIs when searching for a nodule on the X-ray image in question. This process is modeled by a time homogeneous Markov chain, where different participants may have a different transition matrix. Subsection 3.1 defines the Hidden Markov Model (HMM)W, Subsection 3.2 describes the reversible Jump Markov Chain Monte Carlo (MCMC) algorithm for its estimation, and Subsection 3.3 presents the details of the full conditional distributions and the Gibbs sampler.

4.3.1 Details of the Hidden Markov Model

We assume that, regardless of participants and time, a given image contains S AOIs, which may partly overlap. The data of participant i is a sequence of pairs $(\vec{y}_{i,t}, s_{i,t})$ with length T_i, where $\vec{y}_{i,t}$ is a two-dimensional vector representing the fixation point at time t and $s_{i,t}$ is the corresponding AOI. We denote $\mathbf{y}_i = \{\vec{y}_{i,1}, \ldots, \vec{y}_{i,T_i}\}$ to be the fixation sequence for participant i and $\mathbf{s}_i = \{s_{i,1}, \ldots, s_{i,T_i}\}$ to be the sequence of AOIs, or states, visited. For participant i at time t, $(\vec{y}_{i,t}, s_{i,t})$ is uniquely defined. However, only $\vec{y}_{i,t}$ is observable. Our approach will estimate the hidden states $\mathbf{s}_i$ from the observed $\mathbf{y}_i$ as explained below.

4.3.1.1 Spatial Poisson Process

The model is based on the assumption that conditional on AOI $s_{i,t}$, the fixation positions $\vec{y}_{i,t}$ are independently distributed as a spatial Poisson process, with intensity function

$$\lambda_s(\vec{y}) = \lambda_{i,t}(\vec{y}_{i,t} = \vec{y} | s_{i,t} = s) = \mathcal{N}_2(\vec{y}|\vec{\mu}_s, \Sigma_s) I(\vec{y} \in \Omega),\ s = i, \ldots, S$$

where $\mathcal{N}_2$ denotes a bivariate normal density, Ω denotes the image, $\vec{\mu}_s = (\mu_s^x, \mu_s^y)'$ is the spatial center of the s^{th} AOI and $\Sigma_s = \begin{pmatrix} \sigma_{s,x}^2 & \sigma_{s,xy} \\ \sigma_{s,xy} & \sigma_{s,y}^2 \end{pmatrix}$ is its spatial covariance matrix. Thus, the unobserved AOIs are assumed to have the shape of ellipses constrained to the region of the image.

By the property of spatial Poisson process, the probability of fixating at point

$\vec{y}_{i,t}$ conditional on $s_{i,t}$ is:

$$P_s(\vec{y}) = P(\vec{y}_{i,t} = \vec{y} | s_{i,t} = s) = \frac{\lambda_s(\vec{y})}{\int_{\vec{u}\in\Omega} \lambda_s(\vec{u})} = \frac{\mathcal{N}_2(\vec{y}|\vec{\mu}_s, \Sigma_s) I(\vec{y} \in \Omega)}{\sum_{\vec{u}\in\Omega} \mathcal{N}_2(\vec{u}|\vec{\mu}_s, \Sigma_s)},$$

where $\vec{u} = (u_x, u_y) \in \Omega$ and both $\vec{y}_{i,t}$ and $\vec{u}$ are measured in pixel-units.

4.3.1.2 Spatial and Luminance Priors

On X-ray images, independent of ambient and monitor lighting conditions, nodules consist of roundish regions that are brighter than the background. A region of relatively more luminous pixels is thus a priori more likely to contain a nodule. Thus, apart from spatial information, pixel luminance should a priori influence the likely location of a nodule. So we assign both spatial and luminance based priors to the locations of the centers of the AOIs (in other applications, other image features may influence the prior). The luminance of a pixel value, which lie in $[0, 1]$,is computed as a linear function of RGB, following the NTSC and JPEG standards (Gevers 2001): $X_{luminance} = 0.229R + 0.587G + 0.114B$. The higher the value, the brighter the point is.

For any $s = 1, \ldots, S$, the spatial prior for $\{\vec{\mu}_s, \Sigma_s\}$ is the Normal-Inverse-Wishart distribution:

$$P_{sp}(\vec{\mu}_s, \Sigma_s) = P(\vec{\mu}_s|\Sigma_s)P(\Sigma_s) = \mathcal{N}_2(\vec{\mu}_s|\vec{m}_s, \lambda_s^{-1}\Sigma_s) I(\vec{\mu}_s \in \Omega) \times \mathcal{W}_2^{-1}(\Sigma_s|V_s, \kappa_s),$$

where $\mathcal{W}_2^{-1}$ is the two-dimensional inverse Wishart distribution, and $\vec{m}_s, \lambda_s, V_s$ and κ_s are hyper-parameters. The luminance prior is: $P_{lu}(\vec{\mu}_s) \propto (\beta + 1)(x_{\vec{\mu}_s})^\beta$, where $x_{\vec{\mu}_s} \in [0, 1]$ is the luminance value at $\vec{\mu}_s$, and β has a non-informative uniform hyper-prior on $(-1, 1)$. If the posterior of β has a substantial part of its mass above 0, it

means that the brighter, more luminous points are more likely to be the center of the AOIs. Assuming the spatial and luminance priors are a priori independent, the prior for $\{\vec{\mu}_s, \Sigma_s\}$ is

$$P(\vec{\mu}_s, \Sigma_s) = P_{sp}(\vec{\mu}_s, \Sigma_s) \times P_{lu}(\vec{\mu}_s).$$

4.3.1.3 Partially Hidden Markov Chain

Instead of using a hidden Markov chain in which all states are unobserved, for this particular application, we assume that the first state represents the headline or image caption. Such a headline was always present above each X-ray image, so that for any fixations $\vec{y}_{i,t}$ on the headline, we have $s_{i,t} = 1$. There are several advantages of doing this. First, the headline can never contain the nodule, and therefore can be treated as an a priori known and predefined AOI. Second, it is of interest to investigate how participants jump between the headline and the other AOIs in the image, for example to confirm a final judgement, and specifying the headline as an a priori defined AOI makes this possible.

Thus the sequence $\mathbf{s}_i = \{\vec{s}_{i,1}, \ldots, \vec{s}_{i,T_i}\}$ is a partially hidden chain for participant i, and each $\mathbf{s}_i$ is assumed to follow a first order time homogeneous Markov chain with transition matrix $Q_i = (q_i(r,s))$ and initial distribution $\vec{\pi}_{0i} = \{\pi_{0i}(1), \ldots, \pi_{0i}(S)\}$, where for $s, r \in \{1, 2, \ldots, S\}$, $t = 2, \ldots, T_i$, $q_i(r,s) = p(s_{i,t} = s | s_{i,t-1} = r)$ and $\pi_{0i}(s) = p(s_{i,1} = s)$. For any $r = 1, \ldots, S$, $q(r, \cdot)$, the r^{th} row of Q_i, together with $\vec{\pi}_{0i}$, are a priori independently distributed according to a Dirichlet distribution $\mathcal{D}_S(\vec{\alpha}_i)$. In order to include unobserved differences among participants, we adopt

the following methods (Scott 2005). Assume $\vec{\alpha}_i = A_i\vec{\phi}_i$ where $A_i = \sum_{j=1}^{S} \alpha_i(j)$ is the shrinkage parameter and $\vec{\phi}_i \sim \mathcal{D}(\mathbf{1})$ is the location parameter. The A_i represent the number of prior observations in the distribution of Q_i and is given a uniform shrinkage prior. The transformed prior for $\vec{\alpha}_i$ is

$$p(\vec{\alpha}_i) = \frac{z_0\Gamma(S)}{(z_0 + A_i)^2 A_i^{S-1}},\ i = 1, \ldots, I.$$

The number of states S is unknown, and we assume that a priori it is equally likely to take on any value between 1 to $\mathcal{S}$: $P(S = k) = 1/\mathcal{S}$. Recall that state 1 is the headline. In order to avoid label switching, we order the remaining AOIs by their x-coordinates from left to right.

4.3.2 Reversible Jump Markov Chain Monte Carlo

Previous eye tracking analysis has used a fixed number of predefined AOIs. In our model, however, we assume that the number of AOIs is unknown and thus it must be estimated. We will use a special case of Reversible Jump Markov Chain Monte Carlo (RJMCMC) for HMM (Green 1995; Robert et al. 2000). The idea of RJMCMC is that at each iteration of the Gibbs sampler, after updating all the hidden states and variables, S is randomly increased or decreased by 1 and the hidden states rearranged accordingly. Then a Metropolis-Hastings step is used to determine whether to accept this change or not. The key is that generating or eliminating states should be a reversible pair.

Richardson and Green (1997) describe two reversible jump algorithms. One involves the birth or death of an empty state, while the second one involves a split

or merge of existing states. Both of these steps, however, are known to have low acceptance rates (Green 1995). Our algorithm involves modified birth and death steps. That is, in each step of the MCMC, a birth or a death is randomly proposed with a birth rate b_S and a death rate $d_S = 1 - b_S$ if the current number of states is S. Notice that $d_1 = b_{\mathcal{S}} = 0$ and death can only happen when there is at least one empty state.

If the current number of state is S and a birth is proposed, a new state j^* on the image is generated with center and variance covariance matrix $\{\vec{\mu}_{j^*}, \Sigma_{j^*}\}$ simulated from the prior distribution $\mathcal{N}_2(\vec{\mu}|\vec{m}_0, \lambda_0^{-1}\Sigma_0)\mathcal{W}^{-1}(\Sigma|V_0, \kappa_0)$. The value of j^* is determined by the rank of its x-coordinate and other state indices are reassigned accordingly. The parameter vector $\vec{\alpha}_i$ is also changed into a $S+1$ dimensional vector $\tilde{\alpha}_i$ by inserting $\alpha_i(j^*) \sim (S-1)[A_i + \alpha_i(j^*)]^{1-S} A_i^{S-1}$ (Appendix B.1). The transition matrix Q_i is changed to $\tilde{Q}_i$. The new row of $\tilde{Q}_i$, $\tilde{q}_i(j^*, \cdot)$ is generated from a $k+1$ dimensional Dirichlet distribution $\mathcal{D}_{k+1}(\tilde{\alpha}_i)$. The new column is changed according to equations (4.1) below, for any $j \neq j^*$ and $j = 1, \ldots, S$. In order to simplify computation, we use a beta distribution $\mathcal{B}(\alpha_i(j^*), A_i + d_{ij} - S + 1)$ as the proposal function to generate $u_i(j)$ (Appendix B.2). Here $d_{ij} = \sum_{t=1}^{T_i - 1} I(s_{i,t} = j)$.

$$\begin{aligned} \tilde{q}_i(j, j^*) &= u_i(j); \\ \tilde{q}_i(j, r) &= [1 - u_i(j)] q_i(j, r), \ r \neq j^*; \end{aligned} \tag{4.1}$$

At this step β does not change. Finally, this birth is accepted with a rate of $\min\left\{1, R = \frac{\pi(d\tilde{\Theta}) q(\tilde{\Theta}, d\Theta)}{\pi(d\Theta) q(\Theta, d\tilde{\Theta})}\right\}$ (Green 1995; Appendix B.3).

If a death is proposed, one empty state j^* on the image is randomly chosen

to die and the indices are reordered accordingly. The whole j^*th row and column of Q_i are deleted and each row is normalized to sum to 1. The acceptance rate is $\min\{1, R^{-1}\}$ (Green 1995).

To overcome the low acceptance rate of births, we will apply an additional birth step at each iteration with the probability of birth converges in probability to 0 as the number of iterations goes to infinity. That is, at the n^{th} iteration of the Gibbs sampler, the additional birth occurs with probability $P_b(n)$, and if a birth is determined, the exact same birth process described above is followed but will be accepted with probability 1. $P_b(n) \to 0$ in probability as $n \to \infty$. Thus, asymptotically this Markov chain converges to the desired Markov chain but mixes substantially better during the burn-in period.

4.3.3 Posterior Computation and Gibbs Sampler

In summary, the parameters in the model are S and $\Theta_S = \{\vec{\mu}_s, \Sigma_s, Q_i, \vec{\pi}_{0i}, \vec{\alpha}_i, \beta : i = 1, \ldots, I, s = 1, \ldots, S\}$, the observed data are $\mathbf{g}_{obs} = \{\mathbf{y}_1, \ldots, \mathbf{y}_I\}$, the latent data are $\mathbf{g}_{mis} = \{\mathbf{s}_1, \ldots, \mathbf{s}_I\}$, the hyper-parameters are $\Phi = \{\vec{m}_s, \lambda_s, V_s, \kappa_s, z_0, \mathcal{S} : s = 1, \ldots, \mathcal{S}\}$ and $X = \{x_u, u \in \Omega\}$ are the luminances at each pixel point on image. The joint likelihood function is:

$$
\begin{aligned}
& p(\mathbf{g}_{obs}, \mathbf{g}_{mis}, S, \Theta_S | \Phi, X) \\
&= \frac{1}{\mathcal{S}} \prod_{i=1}^{I} \left\{ \pi_{0i}(s_{i,1}) \prod_{t=2}^{T_i} q_i(s_{i,t-1}, s_{i,t}) \prod_{t=1}^{T_i} \frac{\mathcal{N}_2(\vec{y}_{i,t} | \vec{\mu}_{s_{i,t}}, \Sigma_{s_{i,t}}) I(\vec{y}_{i,t} \in \Omega)}{\sum_{\vec{u} \in \Omega} \mathcal{N}_2(\vec{u} | \vec{\mu}_{s_{i,t}}, \Sigma_{s_{i,t}})} \right\} \\
&\quad \times \prod_{i=1}^{I} \left\{ \mathcal{D}_S(\vec{\pi}_{0i} | \vec{\alpha}_i) \cdot \prod_{s=1}^{S} \mathcal{D}_S(q_i(s, \cdot) | \vec{\alpha}_i) \cdot \frac{z_0 \Gamma(S)}{[z_0 + A_i]^2 [A_i]^{s-1}} \right\} \\
&\quad \times \prod_{s=1}^{S} \left\{ \mathcal{N}_2(\vec{\mu}_s | \vec{m}_s, \frac{\Sigma_s}{\lambda_s}) \cdot I(\vec{\mu}_s \in \Omega) \cdot \mathcal{W}_2^{-1}(\Sigma_s | V_s, \kappa_s) \cdot (\beta + 1)(x_{\vec{\mu}_s})^{\beta} \right\} p(\beta)
\end{aligned}
\tag{4.2}
$$

To simplify notation, for $r, s \in \{1, \ldots, \mathcal{S}\}$, $i = 1, \ldots, I$, we define the following quantities: $n_{0i}(s) = I(s_{i,1} = s)$, $n_i(s, r) = \sum_{t=2}^{T_i} I(s_{i,t-1} = s, s_{i,t} = r)$, $c_s = \sum_{i=1}^{I} \sum_{t=1}^{T_i} I(s_{i,t} = s)$, $\vec{a}_s = c_s^{-1} (\sum_{i=1}^{I} \sum_{t : s_{i,t} = s} \vec{y}_{i,t})$, $B_s^2 = \sum_{i=1}^{I} \sum_{t : s_{i,t} = s} (\vec{y}_{i,t} - \vec{a}_s)(\vec{y}_{i,t} - \vec{a}_s)'$, $M_s = (\lambda_s + c_s)^{-1} (\lambda_s \vec{m}_s + c_s \vec{a}_s)$, $G_s = (\lambda_s + c_s)^{-1} \Sigma_s$, and

$$
D_s = \left\{ [V_s^{-1} + B_s^2 + \lambda_s c_s (\lambda_s + c_s)^{-1} (\vec{a}_s - \vec{m}_s)(\vec{a}_s - \vec{m}_s)']^{-1}, \kappa_s + c_s \right\}.
$$

Then equation (4.2) can be written as:

$$
\begin{aligned}
& p(\mathbf{g}_{obs}, \mathbf{g}_{mis}, S, \Theta_S | \Phi, X) \\
&\propto \frac{1}{2\mathcal{S}} \prod_{i=1}^{I} \left\{ \left(\prod_{s=1}^{S} \pi_{0i}(s)^{n_{0i}(s)} \right) \mathcal{D}_S(\pi_{0i} | \vec{\alpha}_i) \times \prod_{s=1}^{S} \left[\left(\prod_{r=1}^{S} q_i(s, r)^{n_i(s,r)} \right) \mathcal{D}_S(q_i(s, \cdot) | \vec{\alpha}_i) \right] \right\} \\
&\quad \times \prod_{s=1}^{S} \left[\frac{\mathcal{N}_2(\vec{\mu}_s | M_s, G_s) \, \mathcal{W}_2(\Sigma_s^{-1} | D_s) \, (x_{\vec{\mu}_s})^{\beta}}{(\sum_{\vec{u} \in \Omega} \mathcal{N}_2(\vec{u} | \vec{\mu}_s, \Sigma_s))^{c_s}} \right] \prod_{i=1}^{I} \left[\frac{z_0 \Gamma(S)}{(z_0 + A_i)^2 (A_i)^{s-1}} \right] (\beta + 1)^S
\end{aligned}
\tag{4.3}
$$

The following prior settings are used: $\vec{m}_1 = (460, 60)'$ for the headline state, the approximate image center $\vec{m}_s = (460, 300)'$ for all other s, $V_s = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}$, $\lambda_s = 0.005$, and $\kappa_s = 5$ for all s. We did not find that small perturbations of z_0

had an impact on the posterior distribution (Scott 2005), so we use $z_0 = 1$. To a priori not favor a particular number of states, we use a birth rate $b_k = 0.5$ for $s = 2, \ldots, \mathcal{S} - 1$, $b_1 = 1$ and $b_{\mathcal{S}} = 0$. We also use $\mathcal{S} = 14$ (the number of states on the X-ray images never reached 13 in the estimation).

The parameters Θ are sampled from their posterior distributions using the MCMC algorithm, which samples parameters from $p(\Theta_S|S, \mathbf{g}_{mis}, \mathbf{g}_{obs})$, hidden states from $p(\mathbf{g}_{mis}|S, \Theta_S, \mathbf{g}_{obs})$, and the number of states from $p(S|\mathbf{g}_{obs}, \mathbf{g}_{mis}, \Theta_S)$ circularly. The simulation of $\mathbf{g}_{mis}$ utilizes the forward-backward (FB) algorithm (Scott 2002). Conditioning on $\mathbf{g}_{mis}$ and S, the sampling of Θ breaks into the following four components for which Gibbs sampling or Metroplis-Hasting sampling is used:

- For $i = 1, \ldots, I$, given $\{\vec{\mu}_s, \Sigma_s : s = 1, \ldots, S\}$, $\vec{\pi}_{0i}$, Q_i and β, the $\vec{\alpha}_i$ are independent. If denote $\vec{\pi}_{0i}$ as $q_i(0, \cdot)$, the posterior distribution of $\vec{\alpha}_i$ can be written as:

$$p(\vec{\alpha}_i|S, \mathbf{g}_{mis}, \mathbf{g}_{obs}, \vec{\pi}_{0i}, Q_i, \vec{\mu}_s, \Sigma_s, \beta) = \frac{\Gamma(A_i)}{(z_0 + A_i)^2 (A_i)^{S-1}} \times \prod_{s=0}^{S} \prod_{r=1}^{S} \frac{q_i(s,r)^{\alpha_i(r)-1}}{\Gamma(\alpha_i(r))}. \tag{4.4}$$

- For $i = 1, \ldots, I$, given $\{\vec{\mu}_s, \Sigma_s : s = 1, \ldots, S\}$, $\vec{\alpha}_i$ and β, $q_i(s, \cdot)$ are independent for s with:

$$p(\vec{\pi}_{0i}|S, \mathbf{g}_{mis}, \mathbf{g}_{obs}, \vec{\alpha}_i, \vec{\mu}_s, \Sigma_s, \beta) = \mathcal{D}_S(\vec{\pi}_{0i}|\vec{\alpha}_i + \vec{n}_{0i}), \tag{4.5}$$

$$p(q_i(s, \cdot)|S, \mathbf{g}_{mis}, \mathbf{g}_{obs}, \vec{\alpha}_i, \vec{\mu}_s, \Sigma_s, \beta) = \mathcal{D}_S(q_i(s, \cdot)|\vec{\alpha}_i + \vec{n}_i(s, \cdot)). \tag{4.6}$$

- For $s = 1, \ldots, S$, given $\vec{\alpha}_i$, $\vec{\pi}_{0i}$, Q_i and β, the pairs $\{\vec{\mu}_s, \Sigma_s\}$ are independent

with:

$$p(\vec{\mu}_s, \Sigma_s | S, \mathbf{g}_{mis}, \mathbf{g}_{obs}, \vec{\alpha}_i, \vec{\pi}_{0i}, Q_i, \beta) \propto \frac{\mathcal{N}_2(\vec{\mu}_s | M_s, G_s)\mathcal{W}_2(\Sigma_s^{-1} | D_s)(x_{\vec{\mu}_s})^\beta}{[\sum_{\vec{u}\in\Omega}\mathcal{N}_2(\vec{u}|\vec{\mu}_s, \Sigma_s)]^{c_s}}. \tag{4.7}$$

- Given $\{\vec{\mu}_s, \Sigma_s : s = 1, \ldots, S\}$, $\vec{\alpha}_i$, $\vec{\pi}_{0i}$ and Q_i for $i = 1, \ldots, I$,

$$p(\beta | S, \mathbf{g}_{mis}, \mathbf{g}_{obs}, \vec{\mu}_s, \vec{\alpha}_i, \vec{\pi}_{0i}, Q_i) \propto (\beta + 1)^S \prod_{s=1}^{S} (x_{\vec{\mu}_s})^\beta. \tag{4.8}$$

To sample S from $p(S|\mathbf{g}_{mis}, \mathbf{g}_{obs}, \Theta_S)$, the previously described birth and death step is added to each iteration (Green 1995), followed by the additional birth step described above, with probability

$$P_b(n) = \begin{cases} 0.005 & n <= 2000 \\ 20000/n^2 & n > 2000. \end{cases}$$

We use the last 10,000 draws with a burn-in of 30,000 and thinning the draws by 1 in 10. Convergence of the MCMC was checked by standard methods and using different start points. These results showed no signs of lack of convergence after the burn-in. The autocorrelation plot showed no sign of significant autocorrelation at lags larger than 10.

To investigate the performance of our model and reversible jump algorithm, we analyzed two simulated data sets. Each data set was generated assuming four states, while in each state $\vec{y}_{i,t}$ was generated from a spatial Poisson process with a different bivariate log-normal intensity. In the first data set, the states were well separated because the elements of the variance-covariance matrices were small, while in second data set the elements of the variance-covariance matrix were large, which

resulted in substantial overlap between the states. The transitions between the four states were generated according to a Markov chain. The results of the estimation showed that, for each data-set, the MCMC chains mixed well during the burn-in, and the algorithm quickly converged to the true number of states after that. The recovery of the parameters, $\vec{\mu}_s$, Σ_s and Q was also quite accurate.

4.4 Case Study

4.4.1 The Areas of Interest

In this section we describe the analysis of the data described in Section 2. The number of states S is the first issue we address. The initial number of states is uniformly generated from 1 to $\mathcal{S}$, the maximum $\mathcal{S}$ is set to 14 and the birth and death rates are set to be equal. The MCMC output shows that for each of the seven X-ray images, in the last $10,000$ iterations the number of states remains constant (Figure 4.3), and these are the most likely number of states based on the data. The average of the number of states in the seven X-ray images is 7.14 and the median is 7 (SD=1.07).

Figure 4.4 shows the results for all seven X-ray images with a nodule on it, each from 16 participants' eye-fixation data. Each panel in the figure plots the posterior means of $\vec{\mu}_s$ and the 80% probability ellipse, representing Σ_s for each state. The nodule on each image is indicated by a small circle. As pre-defined, the headline is state 1, and the AOIs on the image are sorted in the order of their x coordinates. Figure 4.4 reveals that each of the images not only has a different

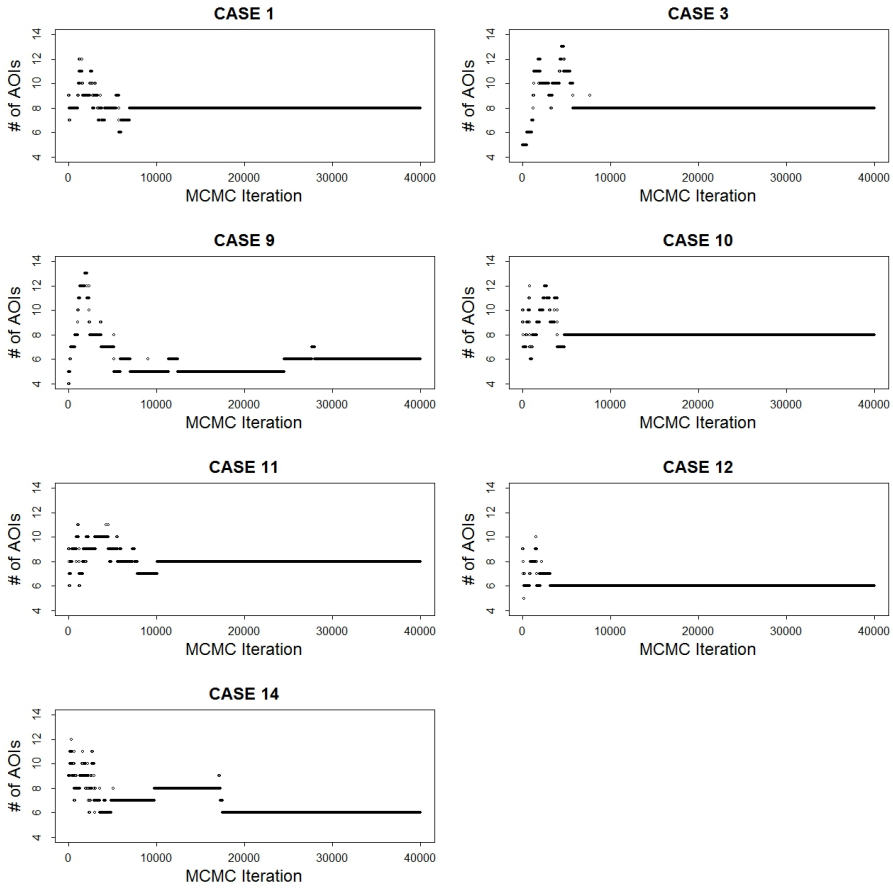


Figure 4.3: Number of states across MCMC iterations

number of AOIs, but also that the sizes and shapes of these AOIs are substantially different. This indicates that the classical approach of identifying AOIs a priori will yield a suboptimal partitioning of the images, and makes it more difficult to represent individuals' search behavior across informative regions in the image adequately. The reason that the number and shapes of the AOIs are various in different images is that the structures visible in chest X-rays vary considerably from patient to patient. Importantly, we can see from the figure that 6 out of the 7 images have one well defined state (AOI) surrounding the nodule (Case 3 has a larger nodule that is covered by two states). The size of the nodule state is significantly smaller than

other AOIs ($9,559$ pixels on average for nodule states and $54,114$ pixels for others). The number of participants that correctly find the nodule is negatively related to the size of nodule state ($p = 0.06$). Our model thus can detect the nodule in X-ray images quite well, based on eye-movement data.

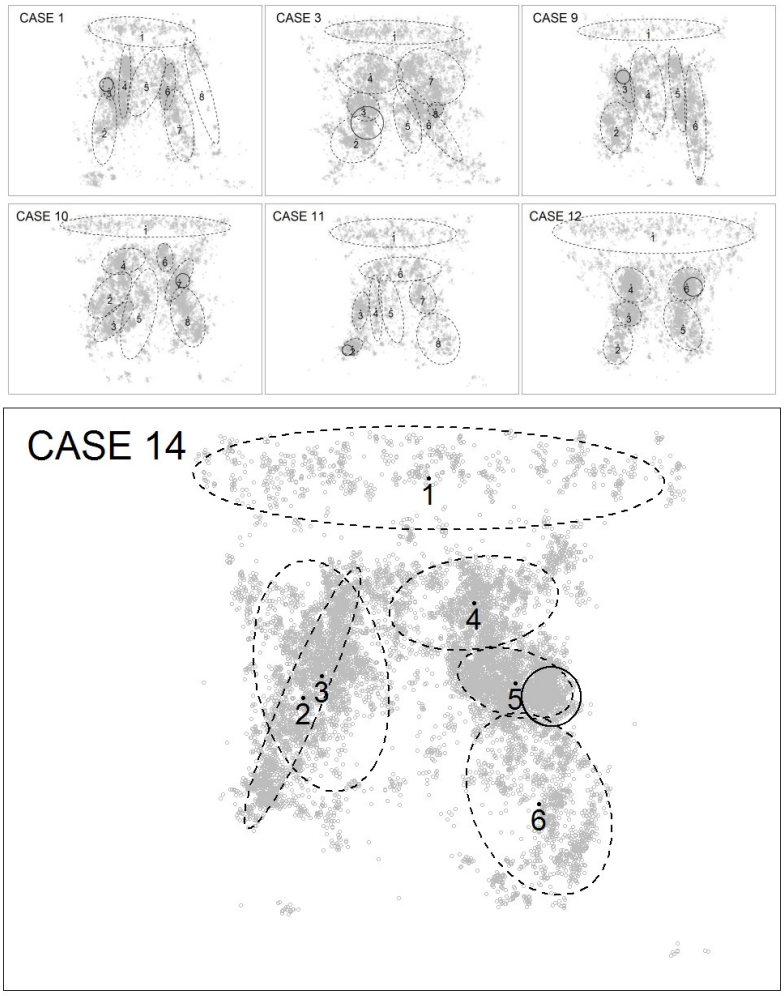


Figure 4.4: Estimated areas of interest on seven X-ray images. The dashed ellipses are the estimated AOIs, with the state number in center. The circles represent the nodule. Case 14 is used to illustrate the result.

Using case 14 as an example, we provide the posterior mean and standard deviation of $\vec{\mu}_s$ in the upper part of Table 4.1. The table shows that each AOI is clearly separated from the others (see Figure 4.4), and the small standard deviations suggest that the eye movement data provide a substantial amount of information on the location of the AOIs. Notice that the y-axis is labeled from top to bottom. The lower part of Table 4.1 displays the posterior mean and stand deviation of the elements of Σ_s. It shows that there are substantial differences in the shape and orientation of the AOIs, and that the nodule state 5 is most concentrated.

Table 4.1: Summary of the posterior distribution of $\vec{\mu}$ and Σ for case 14.

$\vec{\mu}$	$\mu_1(x)$	$\mu_1(y)$	$\mu_2(x)$	$\mu_2(y)$	$\mu_3(x)$	$\mu_3(y)$	$\mu_4(x)$	$\mu_4(y)$	$\mu_5(x)$	$\mu_5(y)$	$\mu_6(x)$	$\mu_6(y)$
Mean	529.3	65.4	361.3	358.5	386.0	329.0	589.5	232.2	644.5	339.0	676.4	499.8
SD	5.7	1.3	1.2	2.3	1.6	2.4	1.6	0.9	0.7	0.5	1.6	2.2

Σ	Σ_1		Σ_2		Σ_3		Σ_4		Σ_5		Σ_6	
	30783.2	595.0	1872.4	-3893.8	2547.3	912.8	3933.7	-382.3	1827.9	337.2	2938.4	1020.1
Mean	(1507.3)	(221.9)	(65.1)	(128.6)	(96.7)	(123.9)	(138.4)	(58.2)	(41.6)	(20.3)	(122.3)	(118.8)
(SD)	595.0	1477.5	-3893.8	9372.3	912.8	7357.5	-382.3	1222.0	337.2	697.3	1020.1	4595.5
	(221.9)	(115.1)	(128.6)	(279.6)	(123.9)	(254.2)	(58.2)	(42.1)	(20.3)	(21.3)	(118.8)	(196.2)

The parameter β, reflecting the influence of the luminance, may reveal why certain regions in the image are visually salient. Figure 4.5 presents a box-plot of the posterior distribution of β for each image. We can see that the estimates of β are very similar across the images; the average of β is typically around 0.8 and the

95% CIs do not cover 0. This shows that the brighter a point is, the more likely it will pop out in the X-ray image and attract eye movements to it. This is important because a nodule is usually brighter than its surroundings, and as a consequence, it will have a larger probability to be found by viewers: if the nodule is 10% brighter, the probability that it pops out will increase by 8%. This underlines the potential of image enhancing methods to increase the intensity of a nodule relative to its surroundings.

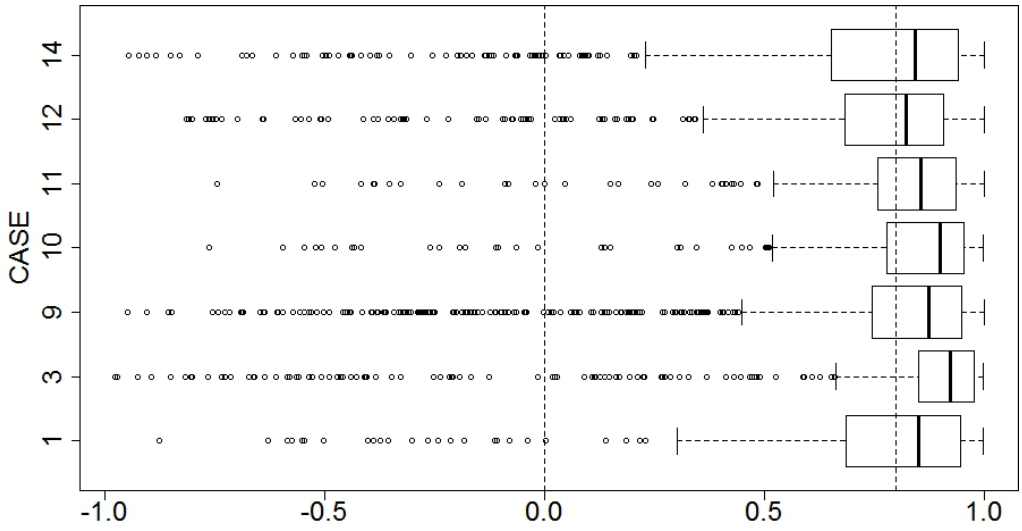


Figure 4.5: Boxplot of the posterior distribution of β for each X-ray image.

4.4.2 Analysis of Transition Patterns

Other important results of the model are the transition patterns of participants' eye fixations among AOIs. Based on participants' answers as well as their eye fixation data, we classify them into three groups: "Success": those cases in which a participant really found the nodule, "False Positive": when a participant

claims to have found but in reality didn't find the nodule, and "Failure": when participants did not find a nodule that was present. To study the effects of different transition patterns on search effectiveness, we will compare several statistics among the three groups.

For a given image, each participant has a different transition matrix Q_i, where the Q_i are connected by the shrinkage parameter A_i. Still using case 14 as an example, Figure 4.6 displays a box-plot of the simulated of A_i for each i, which describes its marginal posterior distribution, grouped by success-failure type. From the graph we can see that the shrinkage parameter for some of the false positives is much larger than those for the two other groups, suggesting that there is less information on their transition between AOIs. To study this in more depth, we calculate the Bayesian shrinkage factor $K_i^r = A_i/(A_i + \sum_{s=1}^{S} n_i(r,s))$ (Scott 2005) for each state r and plot its posterior median for each participant in the second panel of Figure 4.6. It shows that the participants who believe there is no nodule on the image have a relative smaller shrinkage factor than the others. This indicates that there is much information on their transition pattern between AOIs, likely because their eyes jump around the AOIs more extensively than others. Many of the false positives indeed have large shrinkage factors in Figure 4.6, which indicates few transitions between AOIs. Besides, we can see that those with short search times usually have a larger shrinkage factor, which is more evidence of lack transition information. Notice that the nodule state (state 5) generally has a smaller shrinkage factor, which suggests that participants' eyes move through the nodule state frequently trying to find it.

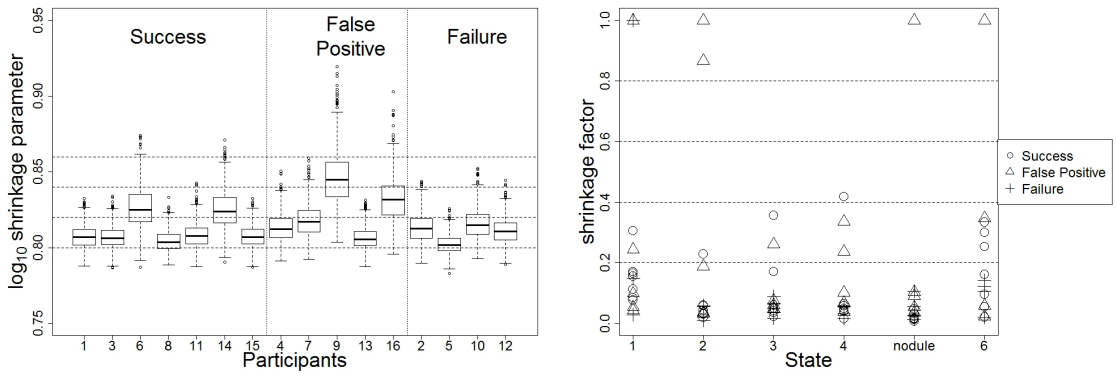


Figure 4.6: Posterior distribution of shrinkage parameter A_i for each X-ray, grouped by success-types.

We present the transition probabilities Q_i by using the gray scale plot in Figure 4.7. Notice that for case 14, the nodule is in state 5. The left panel shows that the value of $q(5,5)$ is the largest, which suggests that once the participants are in the AOI that contains the nodule, their eyes tend to stay fixated in that region, indicating that the nodule is correctly recognized. On the other hand, the center and right panels show smaller values of $q(5,5)$ (success (0.97), false positives (0.75), failure (0.93)). This suggests that the participants have a higher probability to move to other AOIs even if they reach the AOI that contains the nodule, and indicates recognition errors (Krupinski 2000), as the nodule is observed, but apparently not recognized as such. Further, the right panel shows that participants that fail to find the nodule jump from the headline to each of the AOIs much more frequently. This may indicate that they are looking for cues in the headline about whether or not the patient in question may have a nodule.

In addition, we report the switching rate: the percentage of state changes

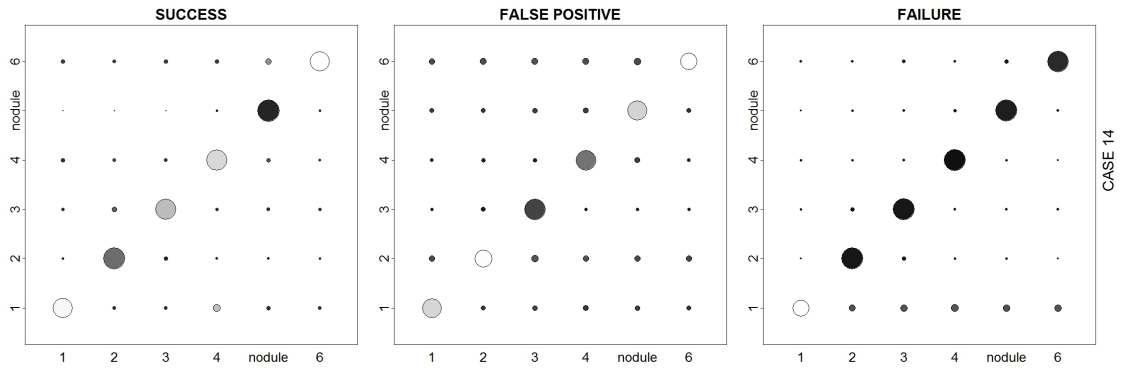


Figure 4.7: Gray scale plot for the transition probability matrices for the three groups for case 14. The size of the symbol at (s, r) reflects the square root of the posterior mean of $Q_i[s, r]$ across participants in that group, where larger symbols represent greater value. The gray scale of the symbol is proportional to the posterior standard deviation.

throughout the search for the nodule. Table 4.2 shows the posterior switching rate for each participant across images, while Table 4.3 shows the posterior switching rate for each image across participants. Table 4.2 shows that the average switching rates are low, roughly between 2.7% and 5.7%. Participants thus mostly search within the AOIs, and occasionally switch to another AOI. There are relatively large differences between participants in these switching rates, up to a factor 2 or more. Differences between the images are smaller, but still sizeable. The question is whether these differences lead to different search results. This question will be answered in the next section.

Table 4.2: Switching rate between AOIs of each participant across images.

Participant	1	2	3	4	5	6	7	8
Ave. (%)	3.7	4.4	3.1	3.7	3.1	5.3	5.0	3.0
SD	(1.4)	(2.0)	(0.4)	(0.9)	(0.5)	(1.6)	(3.7)	(0.6)
Total Success	4	1	2	1	3	2	0	3
Participant	9	10	11	12	13	14	15	16
Ave. (%)	4.3	3.9	3.3	4.4	5.0	5.7	2.7	4.5
SD	(1.3)	(1.1)	(0.5)	(1.3)	(2.9)	(4.3)	(0.7)	(2.4)
Total Success	3	1	4	2	3	2	3	1

Table 4.3: Switching rate of each image across respondents.

Image	CASE1	CASE3	CASE9	CASE10	CASE11	CASE12	CASE14
Ave. (%)	4.6	4.9	2.9	4.1	5.0	3.5	3.4
SD	(2.2)	(2.5)	(1.3)	(1.2)	(3.0)	(1.1)	(1.4)
Total Success	7	1	7	1	8	4	7

4.4.3 Effects on search performance

We want to find the differences in eye movement patterns between the successes, the false positives and the failures. For this purpose, we pool the data across images and participants, and study the effect of features that might lead to a difference in search performance, using a (Bayesian) multinomial logistic regression. Let $Y_i = 0$ indicates success, $Y_i = 1$ false positive and $Y_i = 2$ failure, we use a model in which the successes are the baseline category, and the explanatory vari-

ables are the following features of the visual search process: $X_{i,1}$ {the total length of search in seconds}, $X_{i,2}$ {the percentage of time spend in the headline AOI}, $X_{i,3}$ {the transitions rate between AOIs}, $X_{i,4}$ {systematic search rate (the percentage of transitions that are left-right or up-down between two adjacent states)}, and $X_{i,5}$ {$M_2 = 1 - \lambda_2$, where λ_2 is the second largest eigenvalue of the transition matrix, which measures the time for the chain to reach the stationary distribution}. Here $i \in \{1, \ldots, 16\}$ indicates the respondents and $X_{i,1}$ to $X_{i,5}$ are standardized.

Table 4.4: Effects of search attributes on failures and false positives versus successes in a multinomial logit model.

		Total Length	Headline Rate	Transition Rate	Systematic Scan	M_2
False Positives	Mean	2.397	0.648	-2.780	1,377	4.382
vs.	SD	0.858	0.544	1.129	0.785	1.465
Successes	Median	2.419	0.644	-2.640	1.397	4.189
	95% CI	(0.716, 4.078)	(-0.419, 1.714)	(-4.993, -0.566)	(-0.162, 2.916)	(1.511, 7.252)
Failures	Mean	3.374	0.659	-1.978	2.14	5.512
vs.	SD	0.995	0.693	1.249	0.913	1.637
Successes	Median	3.333	0.633	-1.903	2.104	5.335
	95% CI	(1.423, 5.324)	(-0.700, 2.017)	(-4.426, 0.470)	(0.350, 3.930)	(2.303, 8.721)

Table 4.4 shows the posterior mean and standard deviation for each parameter. It tells us that: (A) The total scan length is associated with more false positives and failures, which indicates that participants who can't find a nodule tend to search longer, or alternatively that if the nodule does not pop out quickly and attract attention, a longer search for it is not associated with a higher success rate. (B) The time

one spends on the headline does not have a significant effect on success rate. (C) More frequent transitions among AOIs are associated with more successes relative to false positives. This suggests that to accurately find a nodule on an image, one should scan the whole image widely and visit all AOIs instead of searching one area and move to the next area after fully searching the previous one. (D) Furthermore, systematic search is associated with a higher number of failures relative to successes (but is unrelated to false positives). This may suggests that to successfully find a nodule on an image, one should scan the image completely and compare all AOIs, rather than systematically scan the image left-right or top-bottom. The last two findings are at odds with common recommendations to radiology residents to systematically search X-ray images. (E) Last but not least, a longer time to reach the stationary distribution is associated with fewer nodules being found and increases both failures and false positives. This implies that successes are faster to reach the stationary distribution, which has a higher probability for the nodule state. Indeed, taking case 14 as an example, the stationary probability of the nodule state for the successes is higher (mean=0.43, SD=0.06) than the false positives (mean=0.19, SD=0.17) or failures (mean=0.24, SD=0.14). Interestingly, however, for some of the false positives and failures, the stationary distribution has a large probability for the nodule state as well (0.44 for participant 9 who is false positive, and 0.44 for participant 2 who is failure). This suggests a decision error rather than a search error (Krupinski 2000), as these participants seem to have detected the nodule, recognize it and spend most of their time looking at it, but make the wrong decision in the end. Training for these types of participants should focus on decision making,

rather than search.

4.5 Conclusion and Discussion

In this article we have proposed a model to identify critical regions in images from the eye movements people make during visual search. Visual search is one of the tasks humans face most frequently, for example when searching for lost objects at home, searching for packages on the shelves of supermarkets or on shopping web sites, searching for traffic signs while driving, searching for suspicious objects in X-ray images at security checkpoints, and searching for anomalies in medical images. Eye movement recordings provide detailed moment-to-moment insights into the visual search process in those situations (Zelinsky 2008). But eye movement recordings produce a large amount of data, and statistical models are needed to make inferences on the search process (Van der Lans et al. 2008).

One key step in the analysis of eye movements is the definition of Areas of Interest: regions on the image hypothesized to be relevant by the researcher a priori. In this research, we have developed a Hidden Markov Model to identify such areas of interest from the eye movement patterns themselves, and characterized search behavior on these areas in a way that allows participants to be heterogeneous. The effect of specific image features (luminance) was included in the model. As the number of areas is a priori unknown, we used a reversible jump algorithm to estimate it (Robert et al. 2000), along with the parameters of the model that characterize location and shape of the AOIs and observers' moment-to-moment eye movements

across them. We used a modified birth and death algorithm in which we added a random birth step that decays in probability across the draws of the MCMC chain to overcome the low acceptance rate of the standard birth and death reversible jump algorithm. The resulting MCMC chain mixes well during the burn-in period, converges quickly to the stationary distribution of the desired chain, and recovers the number of hidden states very well.

We applied the model to eye movement data collected during the search for potentially cancerous nodules on chest X-rays. The search for pulmonary nodules is an extremely difficult task, even for experienced radiologists who may miss them up to 30 percent of the time when the nodule is in retrospect visible, but even higher otherwise (Krupinski 2000). Describing search for such nodules is the focus of a rapidly growing area of inquiry, which aims to improve the search effectiveness of radiologists by showing which search strategies work and which do not. Applying the proposed model to eye movement data collected from a sample of participants, we found an average of about seven well separated areas of interest in chest X-ray images. One of these areas, which was substantially smaller than the others, contained the pulmonary nodule. Participants mostly search within the critical areas, and only occasionally (up to 5 percent of the time) switch from one area to another. How they switch matters. Participants that switch systematically between these regions, in a left-right and top-down fashion, are associate with a lower rate of finding nodules. In addition and in line with earlier research, we found that longer search tends to lead to lower accuracy (Nodine et al. 2002; Oestmann et al. 1988; Gur et al. 1991). If the nodule is not recognized quickly, longer search tends to

result in more errors. Obviously, this finding has important consequences for the training of radiologists, but because it may also reflect lower accuracy and longer searches in difficult cases, we recommend that future research addresses this issue.

Prior research has typically subjectively defined one area of interest that covers the nodule (Krupinski et al., 2003), but recovering the nodule from only the X-ray image without human judgment proves to be very difficult, as current systems come up with as much as 2 or 3 false positives per image. Using our model, we found one area that covers the nodule precisely for all images except for one (in which it was covered by two areas). It is the combination of the features of the digital image and the eye movements of participants that allows us to identify the nodule with such high precision. The nodule area was substantially smaller than the others, indicating that participants focus their attention on a small region around the nodule. The smaller the size of that area was, the higher the accuracy of search on the image in question. In addition, participants that are effective in finding the nodule revisit this area more frequently throughout their search. This knowledge may help radiologists in utilizing more effective search strategies, and may facilitate the development of systems that facilitate the detection of nodules by using information from both images and participants. The scan-paths of successful searches more quickly reached their stationary distributions, which are characterized by a higher probability of the nodule state. Interestingly, for some participants the stationary distribution of the scan-path for false positives and failures also shows a high probability for the nodule state, which suggests decision (but not search) errors that can be alleviated by training or feedback during inspection. Non-nodule states

that have a large stationary probability can be used for gaze-contingent feedback in which they are highlighted to improve search performance (Krupinski et al. 1993). Further, the results showed that areas on the images that are brighter than their surroundings pop out and attract more eye movements, which increase the chance that the nodule is detected and recognized. Thus, gaze contingent digital image processing methods through which the luminance and contrast of candidate locations can be enhanced based on what the observer looks at from moment-to-moment, and zoom and resolution enhancement driven real time by eye tracking, are tools that may improve search. Thus, our analysis suggests a variety feedback, training and image enhancement strategies to improve the search for nodules in chest X-ray images by radiologists. However, our study has some limitations. First, our participants were not trained radiologists. Second, we focused on the detection of only a single nodule, whereas in a real life scenario a radiologist is looking for all types of pathology, including abdominal disease, rib fractures, neck disease, disease of the spine and heart and blood vessels and so on. Third, in real life situations multiple nodules may be present. It would be of importance to apply the method developed in this study to eye-tracking data collected in those settings.

Our model has a much wider potential for application in eye tracking research, in fields such as web-usability, engineering, marketing and psychology. The most compelling advance over current methods for eye-tracking data analysis is the flexibility in identifying areas of interest, both in total number and features such as position, size and shape. In those fields of research other image features beyond luminance, for example edges, contrast, colors and shapes can be included as prior

information. There are several possible extensions of the model to better accommodate the specific applications in those fields of research. For example, instead of assuming that the participants share the same areas of interest, they could be different for different participants, in relation to their knowledge of the stimulus or their viewing purpose. Second, instead of assuming a time homogeneous Markov chain for jumps of the eyes between the regions, an inhomogeneous chain would capture changing transition patterns throughout the task. Last but not least, for dynamic stimuli such as video clips or sequences of X-ray images, the number of areas of interest could be allowed to change throughout the exposure. We hope that the research presented in this paper provides a useful starting point for these developments.

Chapter 5

A Hidden Markov Model Analysis of HIV Clinical Trial Data for Treatment Naïve Female

5.1 Introduction

In the US, approximately 27% of the HIV infected persons and 30% of the new AIDS cases are women (Trends in HIV/AIDS diagnoses 2005). Kojic el al. (2007) pointed out that women who are HIV infected might undergo an earlier menopause (46.5 years old) than those who are not infected ('49 years old for African American and 51 years old for Caucasian women). Patterson et al. (2009) evaluated association of menopause with primary antiretroviral therapy (ART) response by comparing the CD4 cell counts and HIV viral load levels of treatment-naïve female from 2 ACTG primary treatment studies at Week 24, Week 48 and Week 96. Their study shows no significant difference in CD4 cell counts or HIV viral load response to ART between menopausal and premenopausal females. However, this is the largest study so far to address ART response in HIV-infected menopausal female and the effect of age and/or menopause on the treatment results of HIV disease are not clear till now. In this study, we would like to use a hidden Markov model to analyze the health states to which patients belong during their treatment, find the physical meaning for each health state, and compare the transition pattern of patients among

these states to see whether the pre-menopausal patients act differently from post-menopausal patients in terms of their HIV viral load and CD4 cell counts change. The second important issue in the analysis of HIV treatment clinical trial data is to quickly and accurately predict the patients health status in order to make decisions like changes of treatment. With the help of the hidden Markov chain, we will able to predict the status development of the patients based on history data.

5.2 Data Description

In our analysis, we studied females who had never taken any HIV treatment (treatment naïve) to learn how their health status changes during the treatment, in terms of their virologic responses and immunologic responses, through the treatment. The virologic response is defined as the natural log of the HIV viral load and the immunologic response is defined as the CD4 cell count changes from baseline. The data are from registrational HIV randomized clinical trials (RCT) submitted to FDA from 2000-2010. Among all the treatment naïve studies, we choose the studies for analysis based on two criteria: firstly, at least 90% of initial patients stay on study till Week 48; secondly, there are at least 10 female patients greater than 50 years old. From the dataset, we used viral load and CD4 count for each patient at Week 4, Week 8, Week 12, Week 16, Week 24, Week 32, Week 40, and Week 48, which are typical visit times used in clinical trials (time window definitions for each visit time are in Appendix C.1). The viral load is left censored by either 50 copies/ml or 400 copies/ml due to viral load assay limitation (indicated in the dataset). For

missing data, last observation carry forward (LOCF) imputation is used. Patients who have CD4 cell count change larger than 1000 cells/mL are deleted as outliers (less than 1%). In the dataset we use, each subject belongs to one clinical study and is assigned to one treatment arm, and these treatment arms can be classified into the following drug class combinations: NRTI/PI, NRTI/NNRTI and NRTI/CCR5. To study the menopausal effects for patients, we only choose female patients who are less than 35 years old or who are more than 50 years old, since the former group is very likely pre-menopausal group while the latter one is post-menopausal. To study treatment difference as well as age difference, we classified the patients into six pre-defined groups by their age group (≤ 35 years old and ≥ 50 years old) and their drug classes. In the dataset we use, there are six groups. The total number of patients in each group are shown in Table 5.1.

Table 5.1: Total number of patients in each group

Group	Patient Number
≤ 35 and NRTI/CCR5	77
≤ 35 and NRTI/NNRTI	480
≤ 35 and NRTI/PI	96
≥ 50 and NRTI/CCR5	23
≥ 50 and NRTI/NNRTI	81
≥ 50 and NRTI/PI	8

5.3 The Hidden Markov Model and MCMC

In this section we will describe the hidden Markov model (HMM) we are going to use to identify the states of patients. For patient i at time occasion t, the observed data are the two dimensional vector $\vec{y}_{i,t}$, where $\vec{y}_{i,t}(1)$ is the log viral load and $\vec{y}_{i,t}(2)$ is the CD4 cell count change from baseline, and the indicator $C_{i,t}$ for censoring of the viral load. The variable $C_{i,t}$ could take three values. If $C_{i,t} = 0$, it means the log viral load is not censored; if $C_{i,t} = 1$, it means the log viral load is censored by $\log(400)$; and if $C_{i,t} = 2$, it means the log viral load is censored by $\log(50)$. Denote $\mathbf{y}_i = \{\vec{y}_{i,1}, \ldots, \vec{y}_{i,T}\}$ and $\mathbf{c}_i = \{c_{i,1}, \ldots, c_{i,T}\}$ where T is the total number of time occasions, and $T = 8$ if we use the data till week 48. We assume that, regardless of ART, patients, age group or time, there are S health states, and that for each state $s \in \{1, \ldots, S\}$, the log viral load and CD4 cell count changes follow a bivariate truncated normal distribution with the parameter $\{\vec{\mu}_s, \Sigma_s\}$. The mean $\vec{\mu}_i$ can suggest the average amount of log viral load and CD4 change in that state, while the variance-covariance matrix represents how the two responses are correlated. Then each record $y_{i,t}$ belongs to one state, which can be denoted as $s_{i,t}$, and we use $\mathbf{s}_i = \{s_{i,1}, \ldots, s_{i,T}\}$ to denote the state chain for patient i. Notice that $\mathbf{s}_i$ is the hidden chain that we cannot observe. Also notice that for those $\vec{y}_{i,t}$ that is censored, there should be an unobserved true data $\vec{y}^*_{i,t} < \vec{y}_{i,t}$, and we can use $\mathbf{y}^*_i$ to denote the unobserved true data.

5.3.1 Conditional Distribution for Observed Data

If we define the observed data as $\mathbf{g}_{obs} = \{\{\mathbf{y}_1, \mathbf{c}_1\}, \ldots, \{\mathbf{y}_I, \mathbf{c}_I\}\}$, the unobserved chain $\mathbf{g}_{chain} = \{\mathbf{s}_1, \ldots, \mathbf{s}_I\}$, and the unobserved true value $\mathbf{g}_{cen} = \{\mathbf{y}_1^*, \ldots, \mathbf{y}_I^*\}$, where I is the total amount of patients, and also define Θ_S as the parameter $\{\{\vec{\mu}_s, \Sigma_s\}, s = 1, \ldots, S\}$, then the conditional density of $\mathbf{g}_{obs}$ based on the hidden chain $\mathbf{g}_{chain}$, $\mathbf{g}_{cen}$ and Θ_S is

$$p(\mathbf{g}_{obs}|\mathbf{g}_{chain}, \mathbf{g}_{cen}, \Theta_S) = \prod_{i=1}^{I}\prod_{t=1}^{T} P(\vec{y}_{i,t}|s_{i,t} = s, \Theta_S) = \prod_{i=1}^{I}\prod_{t=1}^{T} P_S(\vec{y}_{i,t}). \tag{5.1}$$

If the observed data $\vec{y}_{i,t}$ is not censored, then

$$P_S(\vec{y}_{i,t}) = \mathcal{N}_2(\vec{y}_{i,t}|\vec{\mu}_{s_{i,t}}, \Sigma_{s_{i,t}}),$$

where $\mathcal{N}_2$ is the bivariate normal distribution. If $\vec{y}_{i,t}(1)$ is censored and it corresponds to the unobserved data $\vec{y}^*_{i,t}(1)$ ($\vec{y}^*_{i,t}(2) = \vec{y}_{i,t}(2)$ since CD4 cell count is not censored), then

$$P_S(\vec{y}_{i,t}) = p(\vec{y}_{i,t}(1) > \vec{y}^*_{i,t}(1)|\vec{y}_{i,t}(2), \vec{\mu}_{s_{i,t}}, \Sigma_{s_{i,t}}).$$

and $P_S(\vec{y}^*_{i,t}) = \mathcal{N}_2(\vec{y}^*_{i,t}|\vec{\mu}_{s_{i,t}}, \Sigma_{s_{i,t}})$.

If we use $\tilde{y}_{i,t}$ to denote all the true values, that is, if $c_{i,t} = 0$, $\tilde{y}_{i,t} = \vec{y}_{i,t}$, and if $c_{i,t} \neq 0$, $\tilde{y}_{i,t} = \vec{y}^*_{i,t}$, then the joint likelihood function of $\mathbf{g}_{obs}$ and $\mathbf{g}_{cen}$ given the hidden chain $\mathbf{g}_{chain}$ and parameter set Θ_S is

$$p(\mathbf{g}_{obs}, \mathbf{g}_{cen}|\mathbf{g}_{chain}, \Theta_S) = \prod_{i=1}^{I}\prod_{t=1}^{T} \mathcal{N}_2(\tilde{y}_{i,t}|\vec{\mu}_{s_{i,t}}, \Sigma_{s_{i,t}}) \tag{5.2}$$

The prior distributions for $\{\vec{\mu}_s, \Sigma_s\}$, $s = 1, \ldots, S$ are all non-conjugate Normal-Inverse-Wishart distributions

$$p_(\vec{\mu}_s, \Sigma_s) = p(\vec{\mu}_s|\Sigma_s)p(\Sigma_s) = \mathcal{N}_2(\vec{\mu}_s|\vec{m}_s, \frac{\Sigma_s}{\lambda_s})\mathcal{W}_2^{-1}(\Sigma_s|V_s, \kappa_s).$$

5.3.2 Hidden Markov Chain

We assume that each hidden chain $\mathbf{s}_i$ follows a first order time-homogeneous Markov process. As described above, the patients can be classified into six different classes by the drug class they took as well as their age groups, and we assume that each class has its own transition matrix which is shared by all patients in that group. That is, if patient i belongs to group $W_i = w$, $w \in \{1, \ldots, W\}$, then the hidden chain $\mathbf{s}_i$ for this patient follows a first order time-homogeneous Markov process with transition matrix $Q_w = (q_w(r, s))$, $r, s \in \{1, \ldots, S\}$ and first state probability $\vec{\pi}_w = \{\pi_w(1), \ldots, \pi_w(S)\}$. Each row of Q_w, together with $\vec{\pi}_c$ is given a conjugate Dirichlet distribution $\mathcal{D}_S(\vec{\alpha}_w)$ with the hyper-parameter $\vec{\alpha}_w$. To capture the class heterogeneity, the hyper parameter $\vec{\alpha}_w$ is given a shrinkage prior

$$p(\vec{\alpha}_w) = \frac{z_0 \Gamma(S)}{(z_0 + A_w)^2 A_w^{S-1}}, \ w = 1, \ldots, W.$$

where z_0 is a constant.

5.3.3 Posterior Computation and Gibbs Sampler

In summary, the observed data are $\mathbf{g}_{obs} = \{\{\mathbf{y}_1, \mathbf{c}_1\}, \ldots, \{\mathbf{y}_I, \mathbf{c}_I\}\}$. The unobserved data are $\mathbf{g}_{chain} = \{\mathbf{s}_1, \ldots, \mathbf{s}_I\}$ and the unobserved true value $\mathbf{g}_{cen} = \{\mathbf{y}_1^*, \ldots, \mathbf{y}_I^*\}$. If the number of states is S, we expand the parameter notation $\Theta_S = \{\vec{\mu}_s, \Sigma_s, Q_c, \vec{\pi}_c, \vec{\alpha}_c : c = 1, \ldots, C, s = 1, \ldots, S\}$ to include all the parameters appearing in the model, and use $\Phi = \{\vec{m}_s, \lambda_s, V_s, \kappa_s, z_0 : s = 1, \ldots, \mathcal{S}\}$ to denote all the hyper-parameters. Then the joint likelihood of the observed data and the

missing data is

$$
\begin{aligned}
& p(\mathbf{g}_{obs}, \mathbf{g}_{chain}, \mathbf{g}_{cen}, \Theta_S | \Phi) \\
= & \prod_{i=1}^{I} \left\{ \pi_{w_i}(s_{i,1}) \prod_{t=2}^{T} q_{w_i}(s_{i,t-1}, s_{i,t}) \prod_{t=1}^{T} \mathcal{N}_2(\tilde{y}_{i,t} | \vec{\mu}_{s_{i,t}}, \Sigma_{s_{i,t}}) \prod_{t: C_{i,t} \neq 0} I(\vec{y}_{i,t}(1) > \vec{y}^{*}_{i,t}(1)) \right\} \\
& \times \prod_{w=1}^{W} \left\{ \mathcal{D}_S(\vec{\pi}_w | \vec{\alpha}_w) \cdot \prod_{s=1}^{S} \mathcal{D}_S(q_w(s, \cdot) | \vec{\alpha}_w) \cdot \frac{z_0 \Gamma(S)}{[z_0 + A_w]^2 [A_w]^{s-1}} \right\} \\
& \times \prod_{s=1}^{S} \left\{ \mathcal{N}_2(\vec{\mu}_s | \vec{m}_s, \frac{\Sigma_s}{\lambda_s}) \cdot \mathcal{W}_2^{-1}(\Sigma_s | V_s, \kappa_s) \right\}. \qquad (5.3)
\end{aligned}
$$

The settings for prior distributions are as follows: $\vec{m}_s = (4, 100)'$, $\lambda_s = 0.05$, $V_s = 5\mathbf{I}$ and $\kappa_s = 5$ for all s, where $\mathbf{I}$ is the identity matrix with diagonal equal to one. The value for z_0 is still set to 1 as we described in Chapter 4.

We will use the MCMC algorithm to sample the parameters and the hidden values. Generally speaking, in each step h, we will draw the unobserved $\mathbf{g}_{cen}$ from $p(\mathbf{g}_{cen} | \mathbf{g}_{obs}, \mathbf{g}_{chain}, \Theta_S, \Phi)$, the hidden chain $\mathbf{g}_{chain}$ from $p(\mathbf{g}_{chain} | \mathbf{g}_{obs}, \mathbf{g}_{cen}, \Theta_S, \Phi)$, and the parameter Θ_S from $p(\Theta_S | \mathbf{g}_{obs}, \mathbf{g}_{cen}, \mathbf{g}_{chain}, \Phi)$. The genital of the unobserved true value is simply sample a value from the bivariate normal distribution truncated by the censoring value. The sampling of the hidden chain uses the Forward-Backward algorithm we described in Section 2.2.2. Finally the draws of the parameter set are similar to what we described in both Chapter 3 and Chapter 4 which are sampling from the regular conjugate posterior distributions.

We ran the fixed state HMM analysis for each number of states separately, then we compared the log marginal likelihood $m_S(\mathbf{g}_{obs})$, as we described in Section 2.4.1, Equation (2.15), to decide the best number of states.

5.4 Case Study

We ran estimations for fixed number of states: $S = 2$, $S = 3$, $S = 4$ and $S = 5$ separately, each for $30,000$ iterations and the first $20,000$ iterations are dropped as the burn in period. We then kept 1 out of every 20 draws to decrease autocorrelation. To check the convergence of the MCMC, we ran each model from different starting points. The results showed perfectly matching parameter estimates, which suggests convergence of the MCMC. To avoid label switching, we labeled the states by the order of mean for the log viral value, from large to small.

5.4.1 The number of states

To determine the number of states, we first calculated the marginal likelihood under each estimated model (Chib 1995), which is shown in Table 5.2. and then draw the state segmentation under each fixed number of states, which are shown in Figure 5.1. Combining these two, we see that $S = 4$ is the best choice as the log marginal likelihood is the smallest and the graph also shows that 4 states is a suitable choice.

Table 5.2: Log marginal likelihood under each estimated model

S	2	3	4	5
$m_S(\mathbf{g}_{obs})$	-46106.08	-44615.14	-40975.26	-43277.74

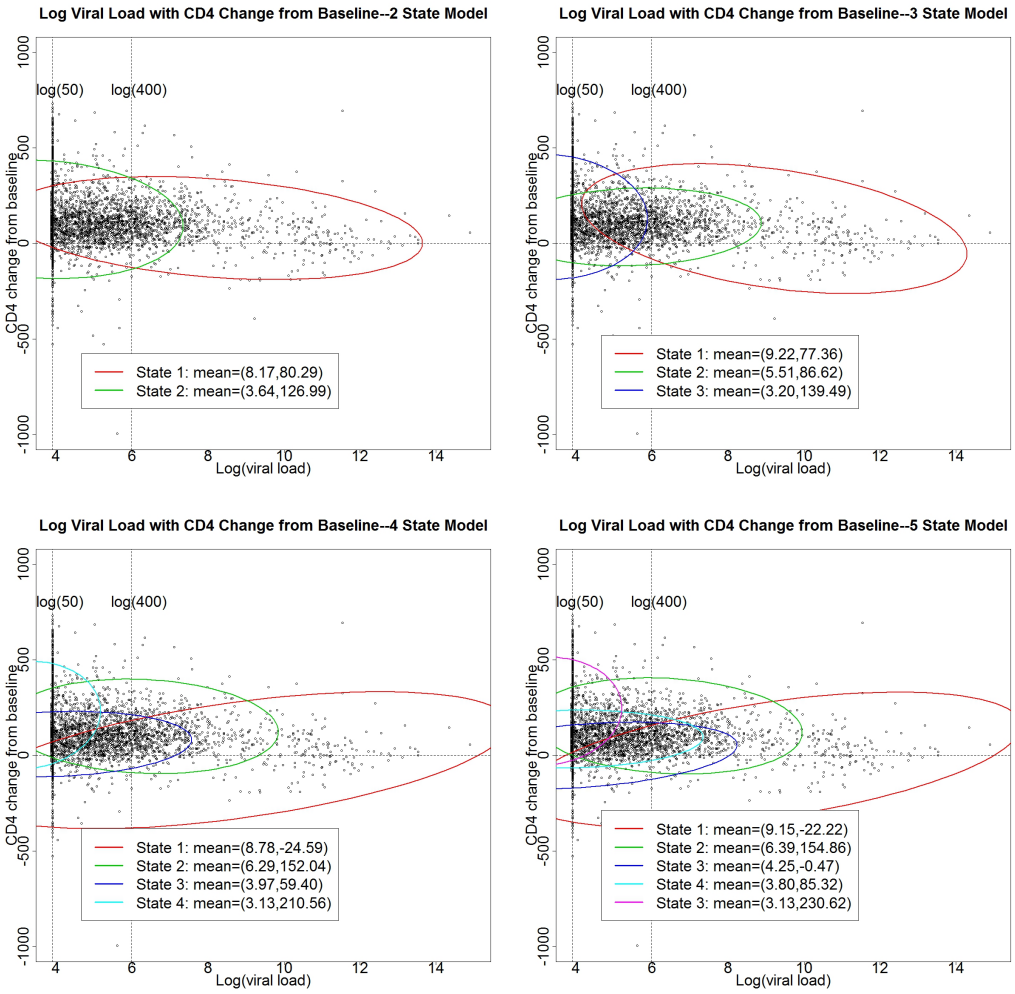


Figure 5.1: Estimated 95% credible ellipse under different number of states.

5.4.2 Analysis under $S = 4$

Figure 5.2 shows the pooled data for each patient in each time point, with the estimated states and their 95% credible ellipses. From the graph we can see that: (A) the first state is the "severe" state, which includes records that are either viral load failures (viral load greater than 400 copies/ml) or CD4 decrease (CD4 change negative), which suggests that the treatment is not effective or drug resistance is developing. (B) the second state is the "progress" state, where patients' viral load

are mostly below 2000 copies/ml and some even below the 400 copies/ml assay limit with most of them having CD4 cell count increase. (C) the third state is a "moderate" state, where most patients are below the 400 copies/ml assay limit, with a large portion reaching the 50 copies/ml assay limit. (D) the fourth state is the "effective" state, where all viral load are below 400 copies/ml with most patients reach the 50 copies/ml assay limit. This is the state where the treatment controls the virus best.

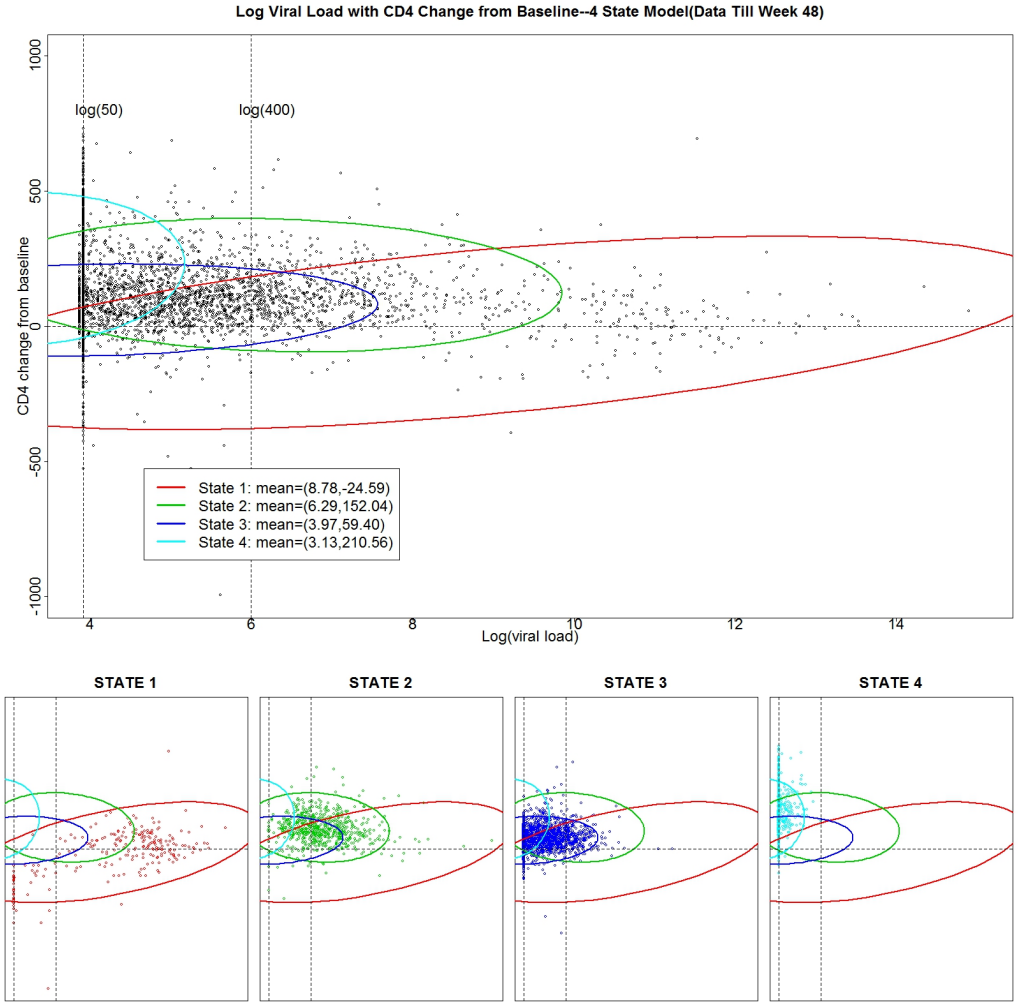


Figure 5.2: Estimated 95% credible bands for each state under a 4-state assumption.

To study how each group is different, we calculated the percentage of patients in each state during each time window group by group. Figure 5.3 shows the trend of these percentages. First of all, we can see that the percentage of patients in state one, which is the "severe" state with a large viral load or with a CD4 decrease, is slightly increasing as time goes on, but is generally lower than 20%. This suggests that most drugs become effective in the first 4 Weeks, but later on some viruses became resistant to the drugs and the viral load rebounds. Second, the percentage of patients in state 2 and state 3 gradually decrease while at the same time the percentage of patients in state 4 increases. Notice that state 4 is the "effective" state which has a low viral load and a big CD4 increase. It is suggesting that with the help of the treatment, a large group of patients gets better. However, also notice that the slopes in state 4 are decreasing as time goes on, typically from Week 16 or Week 24, which suggests that the most gains are at the beginning, and gradually slow down as time goes on, which is also reasonable from the perspective of medical theory.

To evaluate the drug difference, we calculate the percentage of patients in each state at each time windows for different drugs, as well as the 95% credible bands. The plots are shown in Figure 5.4. From the plot we can see that in state 1 and state 3, these three drugs are not significantly different in terms of the proportion of patients in that state. However, from the trend in state 2 and state 4 we can see that NRTI/PI are significantly different from NRTI/CCR5 and NRTI/NNRTI, as fewer patients with NRTI/PI get into the "effective" states.

To evaluate the age effect, we also calculated the percentage of patients in

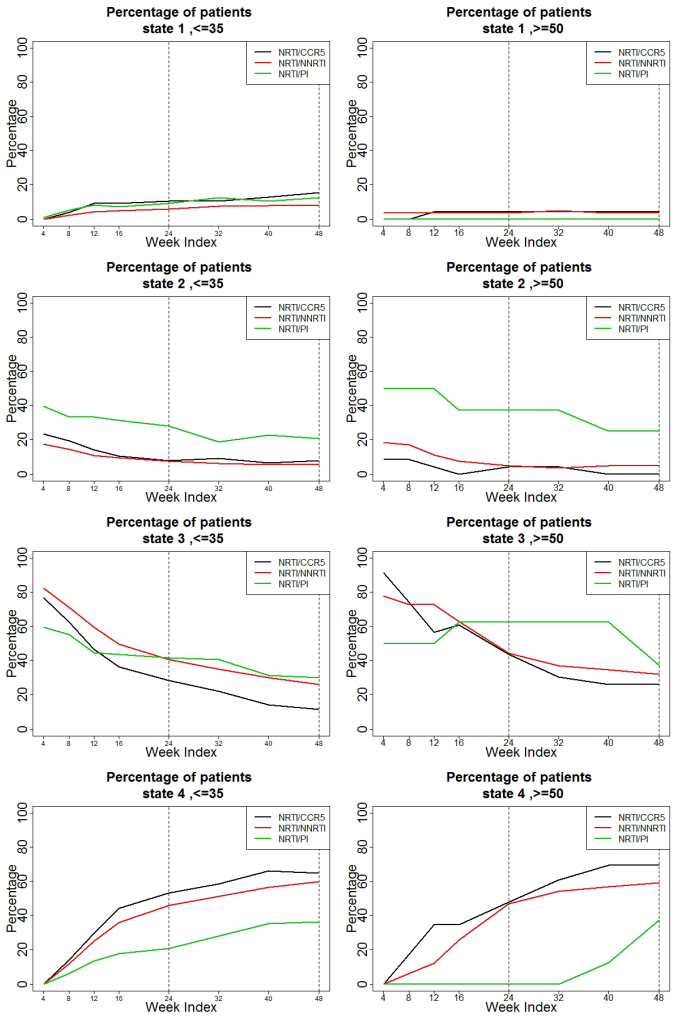


Figure 5.3: Estimated 95% credible bands for each drug at each state for Patients.

each state at each time windows for the age groups ≤ 35 and ≥ 50. Notice that this percentage never changes in the draws, so the 95% credible bands are exactly the same as the estimated mean, which is shown in Figure 5.5. From the figure we can see that there are no significant difference between naïve females younger than 35 years old and older than 50 years old.

Based on how patients transition among these states, 98% of the patients follows one of the six major transition patterns:

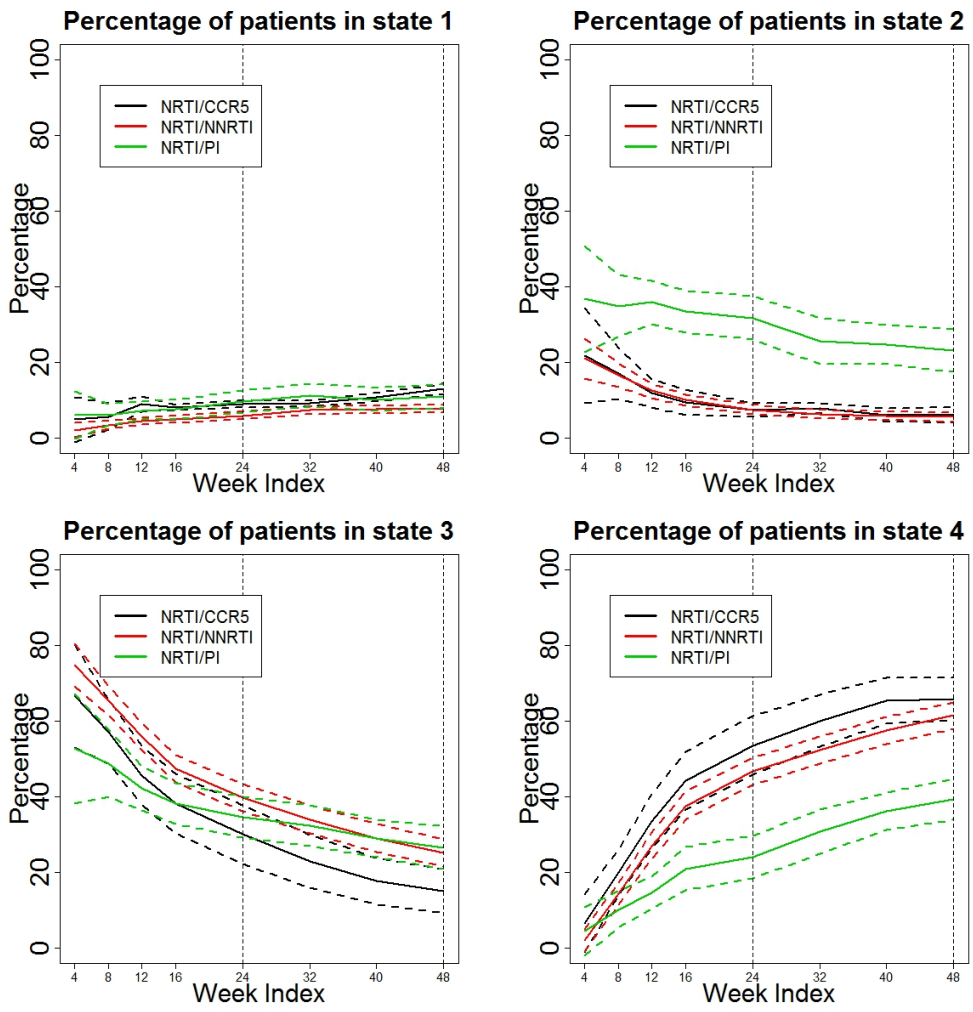


Figure 5.4: Estimated 95% credible bands for each drug at each state.

A: stay at state 2;

B: jump from state 2 or state 3 to state 1;

C: jump between state 2 and state 3;

D: stay at state 3;

E: jump from state 2 gradually to state 4;

F: jump from state 3 to state 4;

These transitions should be related to two common phenomena in HIV disease:

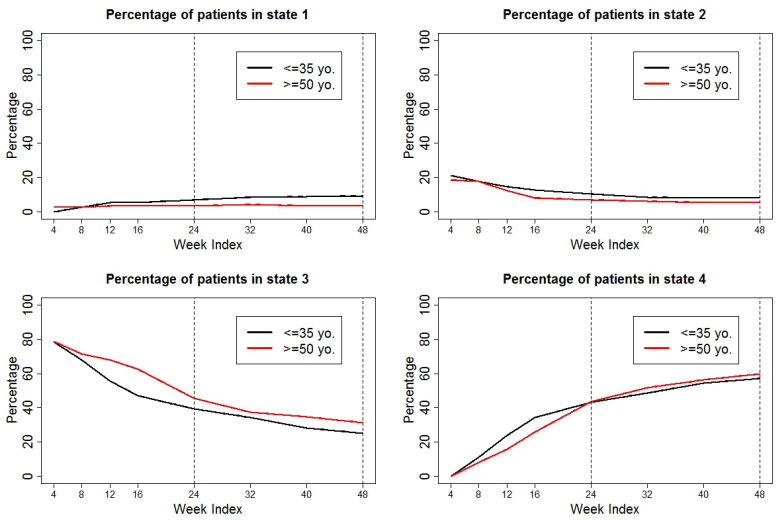


Figure 5.5: Estimated 95% credible bands for each drug at each state.

ever suppressed (viral load lower than the assay limit) and rebound (viral load go beyond the assay limit again after being suppressed). We conjecture that pattern (C)-(F) should be related to ever suppressed, and most rebound patients should have patten (A),(B) or (C). Table 5.3 shows the number of patients that have each of these major transition patterns and the percentage of patients ever suppressed by 400 copies/ml and 50 copies/ml under these pattern, as well as the percentage of patients rebound by the same assay limit. The last row also shows the percentage of patients suppressed below 50 copies/ml assay limit and rebound to 400 copies/ml. We can see from the table that our state perfectly picks up ever suppressed below 400 copies/ml as well as suppressed below 50 but rebound to 400 copies/ml.

To see whether we can predict the transition pattern of patients, we used the data till week 32 to compare with the results from data till week 48. The result shows that 90% of patients share the same transition pattern category, while for the remaining 10%, 2% change from A to F, 3% changes from C to F, and 4%

Table 5.3: Patients in each major transition pattern and percentage of ever suppressed and rebound

	A	B	C	D	E	F
Total	31	55	26	180	66	354
Ever Suppressed 400 (%)	64.5	65.5	100	100	100	100
Ever Suppressed 50(%)	19.4	34.5	88.5	95	95.4	100
Rebound 400 (%)	70	88.9	69.2	10.6	30.3	3.7
Rebound 50 (%)	100	94.7	91.3	35.7	28.6	18.1
Suppressed 50 rebound 400 (%)	100	89.5	60.9	7.6	1.6	2.5

change from D to F. All these patients are gradually moving to the "effective" group between Week 32 and Week 48. However, for those who rebound or stay in the "severe" group, the two results are identical. This suggests that we can use the HMM to filter out those patients who do not respond well to the assigned drug or treatment, and doctors can change the treatment at an early stage.

5.5 Conclusion and Discussion

In this chapter we use the HMM model to analyse the data from HIV clinical trials to study how naïve female patients' virologic responses and immunologic responses change with time. We used both graphical diagnosis and Chib's (1995) estimation of the log marginal likelihood to choose the best number of states. We

found four states that potentially represent the health states of patients, which we define as "severe", "progressing", "moderate" and "effective". We compared the proportion of patients in each health state during each time window to compare the treatment differences and age differences. We find that generally speaking patients taken NRTI/PI act differently from patients take NRTI/CCR5 or NRTI/NNRTI based on the 95% credible intervals. We also find no significant difference for female patients younger than 35 years old verse patients elder than 50 years old, which matches the results from a Meta-analysis. We also classified patients into six transition patterns based on their health state changes, and point out that these patterns are correlated to two frequent phenomena in HIV treatment: ever suppressed and rebound. In addition, we ran the same analysis with the data till week 32 to check whether we can predict patients' jumping pattern at an early time. The results shows that 90% of patients have the same transition pattern using the Week 32 data and the Week 48 data, which suggests that we can predict the transition pattern using the algorithm at an early stage. In particular, all those patients who are not responding well on the treatment stay the same during both Week 32 and Week 48 analysis, which suggests that we can identify them at an early stage and change the drug or treatment.

There are several potential extensions that might be interesting. Firstly, we can add treatment naïve males to compare gender differences for the same drugs. Secondly, as the visit time in our dataset is not equally distributed (4 week intervals in the first 16 weeks but 8 week intervals afterwards), we could assume that there is missing data within the 8 week intervals to make the time intervals equally dis-

tributed. With the help of MCMC, we could estimate the missing data and draw the transition matrix Q and other parameters. Thirdly, instead of assuming that patients in each group share the same transition matrix Q, we could assume that each patient i has a particular Q_i, with the help of which we might able to even better predict the health transition pattern in an early stage and help doctors to make decisions early.

Chapter 6

Conclusion and Future Studies

In this dissertation we introduced a reversible jump hidden Markov model to analyze longitudinal data. We proposed a way to add feature information into the model in order to improve the analysis. We also discussed a shortcoming of the reversible jump Markov chain Monte Carlo and proposed a simple but effective way to overcome it. Firstly we applied the algorithm to two synthetic datasets to check the algorithm accuracy and correctness of the code. Then we applied the algorithm to eye movement data collected during the search for cancerous nodules on X-ray images. With the algorithm we are able to identify the positions, size and shapes of the areas of interest (AOIs) on a given image, together with their total number. The AOIs are modelled as ellipses whose centers are represented by bivariate normal means and whose sizes and orientations are interpreted as variance-covariance matrices. We also studied how the luminance of each pixel point on the image would effect the position of AOIs, and the result turns out to be that the brighter points are more likely possible nodules than the dark points on the X-ray images. Furthermore, we classified the respondents into three groups based on their judgements of the nodule on the image, and defined their search pattern and summarized some search features with the help of the estimated AOIs. We compared searching features that might be related to respondents' answer accuracy. Analysis

through Bayesian multivariate logistic regression showed that the longer searches are related to more failures or false positives, while more jumps between AOIs and unsystematic scan related to more success in finding the nodule.

We also analysed HIV clinical trial data and studied how HIV infected female patients respond to antiretroviral treatment over time, in terms of their immunologic response, the CD4 count and virologic response, the HIV viral load. Pooling the data of all patients at each time occasion together, we found that patients can be classified into three health states by their CD4 count change and log viral load. During their treatment, the patients might move from one state to another. The trend could be reliably estimated by the data of the first 24 weeks. We can use the trend to predict how the patient will act in later weeks, such as Week 40 and Week 48. This might help doctors to make early decision about whether to switch patients' therapy.

Our model could be applied in many other areas, apart from these two applications. One possible area is in visual marketing, such as advertisements, websites or supermarket shelf layouts. With the eye tracking data, we are able to figure out which areas are more attractive to consumers or viewers and why. We could study how other image features, such as contrast, edges, colors and shapes will affect the choice of the AOIs. We could also learn how customers' attention jumps between text and pictures. With these answers, we could make suggestions on how to produce attractive advertisements or websites to attract consumers' attention and as a result increase revenue. There are also many other open questions on eye-tracking data analysis in visual marketing. Currently we are able to figure out the AOIs on

printed ads, and how consumers' attention jumps among these AOIs. However, we are not able to tell how consumers process the image contents in their minds and as a result, know how image contents influences consumers perceptions, judgements and purchasing behavior. If the mechanism of eye tracking can be understood more deeply using our model and further developed ones, it will benefit disciplines such as marketing, psychology, radiology and engineering.

There are also possible extensions of the model. The first extension is the that instead of assuming that the AOIs are shared across subjects, it can assumed that the position of AOIs for each individual or each class can shift from the expected center. This assumption includes subject heterogeneity which might relate to subjects' innate differences, their knowledge of the stimulus or their viewing purpose. The shift may not only occur for centers of AOIs, but also for the variance covariance matrices, reflecting the shapes of the AOIs as well. Either way means that subjects ideally share the same state but in reality have difference.

The second possible extension is that instead of assuming a time homogeneous Markov chain for jumps between states, an inhomogeneous chain could be used to capture changes of the transition patterns throughout the task. This could potentially provide understanding of how respondents process the image information as time goes on. If the transition matrix changes at a latter time, it may suggests that as respondents accumulate knowledge of the image, their way of searching will be effected. This potentially provide a way to study how respondents process image information.

Last but not least, we could also assume that the number of states also varies

throughout the process. One possibility for this assumption is that for the eye tracking data, as respondents accumulate the information from the image, some areas originally of interest will no longer be of interest any more, as respondents refine the image in their minds into fewer regions. This may be especially useful for dynamic stimuli such as video clips or sequences of X-ray images, where the number of AOIs could be allowed to change throughout the exposure. We hope that the research presented in this dissertation provides a useful starting point for the developments mentioned above.

Appendix A

Synthetic Data Study

A.1 Calculation of the acceptance rate for the birth and death moves

If denote $\tilde{A}$ as the newly expended value corresponding to the old one A for any A, and $q(0,\cdot)$ as $\vec{\pi}$, then if the current number of state is S, for a birth move,

$$\begin{aligned}
R &= \frac{\pi(d\tilde{\Theta}_S)q(\tilde{\Theta}_S, d\Theta_S)}{\pi(d\Theta_S)q(\Theta_S, d\tilde{\Theta}_S)} = \frac{L(\mathbf{y},\mathbf{s},\tilde{\Theta}_{S+1}, S+1|\Phi)}{L(\mathbf{y},\mathbf{s},\Theta_S,S|\Phi)} \times \frac{q(\tilde{\Theta}_S, d\Theta_S)}{q(\Theta_S, d\tilde{\Theta}_S)} \\
&= \frac{\frac{1}{S+1}\prod_{t=1}^{1000}\tilde{q}(s_{t-1},s_t)\prod_{t=1}^{1000}\mathcal{N}_2(\vec{y}_t|\tilde{\mu}_{s_t},\tilde{\Sigma}_{s_t})}{\frac{1}{S}\prod_{t=1}^{1000}q(s_{t-1},s_t)\prod_{t=1}^{1000}\mathcal{N}_2(\vec{y}_t|\vec{\mu}_{s_t},\Sigma_{s_t})} \\
&\times \frac{\left\{\mathcal{D}_{S+1}(\tilde{\pi}|\mathbf{1})\prod_{s=1}^{S+1}\mathcal{D}_{S+1}(\tilde{q}(s,\cdot)|\mathbf{1})\right\}\prod_{s=1}^{S+1}\left\{\mathcal{N}_2(\tilde{\mu}_s|\vec{m}_0,\frac{\Sigma_0}{\lambda_0})\mathcal{W}_2^{-1}(\tilde{\Sigma}_s|V_0,\kappa_0)\right\}}{\left\{\mathcal{D}_S(\pi|\mathbf{1})\prod_{s=1}^{S}\mathcal{D}_S(q(s,\cdot)|\mathbf{1})\right\}\prod_{s=1}^{S}\left\{\mathcal{N}_2(\vec{\mu}_s|\vec{m}_0,\frac{\Sigma_0}{\lambda_0})\mathcal{W}_2^{-1}(\Sigma_s|V_0,\kappa_0)\right\}} \\
&\times \frac{d_{S+1}\times J}{b_S(k_0+1)\mathcal{N}_2(\mu_{j^*}|\vec{m}_0,\frac{\Sigma_0}{\lambda_0})\mathcal{W}_2^{-1}(\Sigma_{j^*}|V_0,\kappa_0)\mathcal{D}_{S+1}(\tilde{q}(j^*,\cdot))\prod\limits_{j\neq j^*}p(u(j))}
\end{aligned} \tag{A.1}$$

where $J = \prod\limits_{j\neq j^*}(1-u(j))^{S-1}$ is the Jacobin of transform from $\{\Theta_S, \tilde{q}(j^*,\cdot), u(j), j\neq j^*\}$ to $\tilde{\Theta}_{S+1}$, and

$$p(u(j)) \sim \mathcal{B}(1,cS) = \frac{\Gamma(cS+1)(1-u(j))^{cS-1}}{\Gamma(cS)\Gamma(1)} = cS(1-u(j))^{cS-1}.$$

Notice that for a birth move, (A) the real state a points belongs to does not really change (only the label changes), so all the expressions involve $\vec{y}_t$ are cancelled in Equation A.1; (B) Since the prior for each $\{\vec{\mu}_s, \Sigma_s\}$ as well as $\{\vec{\mu}_{j^*}, \Sigma_{j^*}\}$ are the same, their prior distribution also cancelled in Equation A.1; (C) For any vector $\vec{q}$, $\mathcal{D}_S(\vec{q}|\mathbf{1}) = \Gamma(S)$; (D) for any t, $\tilde{q}(s_{t-1},s_t) = (1-u(j))q(s_{t-1},s_t)$ if $s_{t-1}=j$, so

$$\frac{\prod_{t=1}^{1000}\tilde{q}(s_{t-1},s_t)}{\prod_{t=1}^{1000}q(s_{t-1},s_t)} = \prod_{j\neq j^*}(1-u(j))^{d_j},$$

where $d_j = \sum_{t=1}^{999} I(s_t = j)$ for $j \neq j^*$ and $d_0 = 1$. Now Equation A.1 becomes

$$\begin{aligned} R &= \frac{Sd_{S+1}}{(S+1)b_S(k_0+1)} \times \frac{\Gamma(S+1)^{S+2}}{\Gamma(S)^{S+1}} \\ &\times \frac{(1-u(s_i)) \prod\limits_{j\neq j^*} (1-u(j))^{S-1} \prod_{j\neq j^*}(1-u(j))^{d_j}}{\prod_{j\neq j^*} cS(1-u(j))^{cS-1}\Gamma(S+1)} \\ &= \frac{Sd_{S+1}}{(S+1)b_S(k_0+1)c^{S+1}} \prod_{j\neq j^*} (1-u(j))^{d_j+S-cS}. \end{aligned} \tag{A.2}$$

A.2 Simulation plots for $\vec{\mu}$ and their autocorrelation

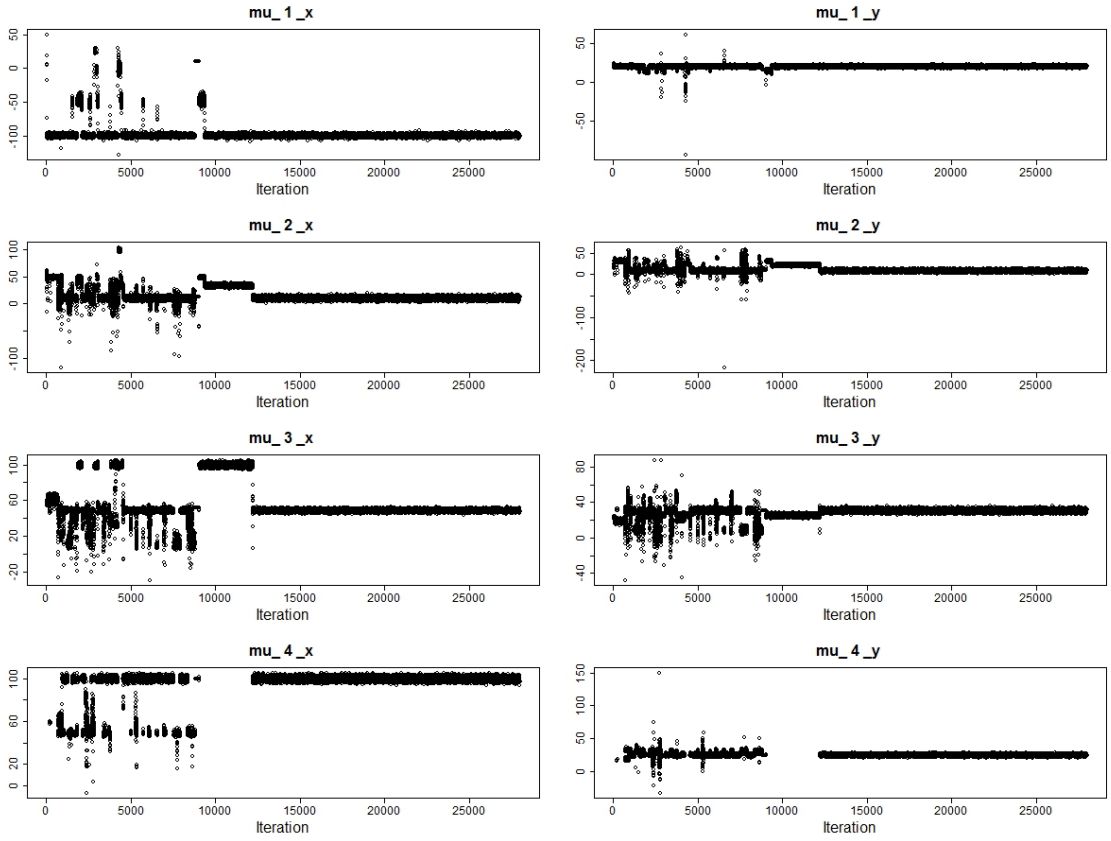


Figure A.1: Simulation results of $\vec{\mu}$ for dataset 2. From the plot we can see that there is no evidence to say that the MCMC chain did not reach the stationary distribution after 20,000 iteration (burn-in) period.

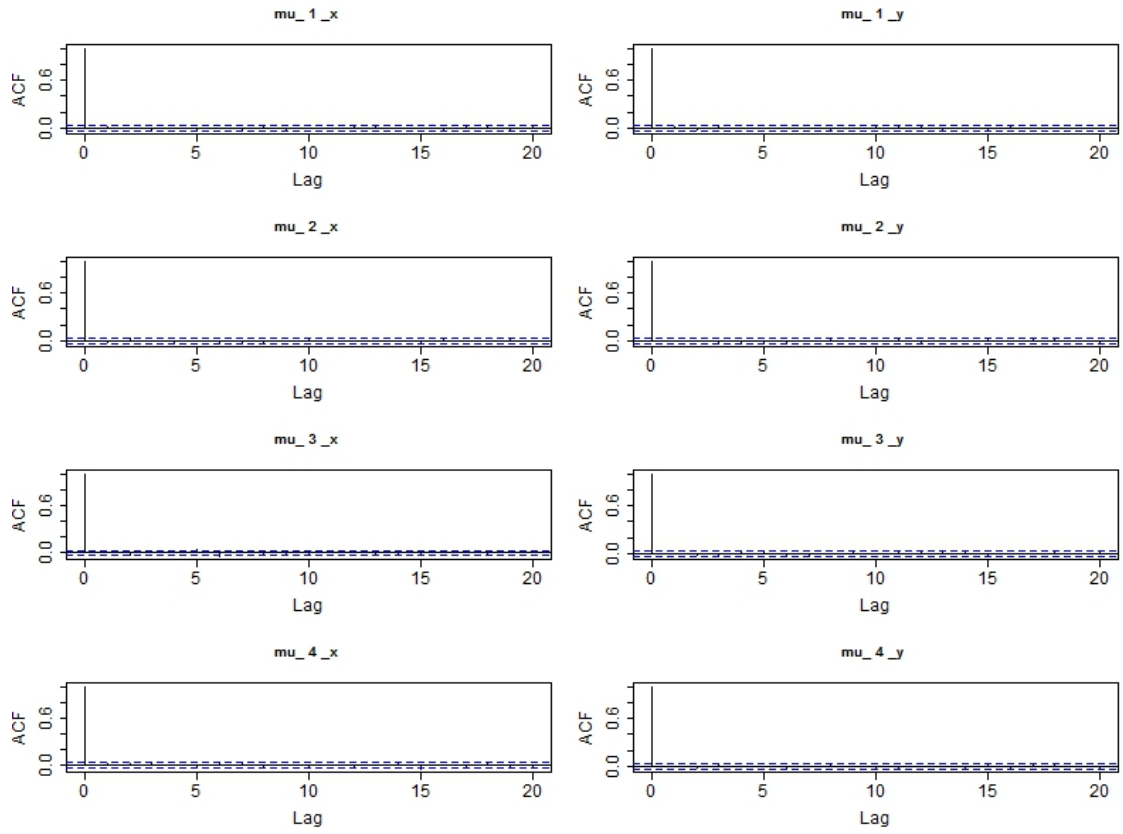


Figure A.2: Autocorrelation plot for each component of $\vec{\mu}$ for dataset 2 after the burn-in period. From the graph we can see there is almost no autocorrelation for the estimation after the burn-in period.

Appendix B

Appendix for Eye Tracking Data Analysis

B.1 Prior for $\alpha_i(j^*)$

According to the prior of $\vec{\alpha}_i$, which is S dimensional, for $i = 1, \ldots, I$,

$$p(\tilde{\alpha}_i) = p(\vec{\alpha}_i, \alpha_i(j^*)) = \frac{z_0\Gamma(S+1)}{(z_0+\tilde{A}_i)^2(\tilde{A}_i)^{S+1-1}} = \frac{z_0\Gamma(S+1)}{[z_0+A_i+\alpha_i(j^*)]^2[A_i+\alpha_i(j^*)]^S},$$

So,

$$\begin{aligned} p(\alpha_i(j^*)|\ \vec{\alpha}_i) &= \frac{p(\vec{\alpha}_i, \alpha_i(j^*))}{p(\vec{\alpha}_i)} \\ &= \frac{z_0\Gamma(S+1)}{[z_0+A_i+\alpha_i(j^*)]^2[A_i+\alpha_i(j^*)]^S} \div \frac{z_0\Gamma(S)}{(z_0+A_i)^2(A_i)^{S-1}} \\ &= \frac{S(z_0+A_i)^2A_i^{S-1}}{[z_0+A_i+\alpha_i(j^*)]^2[A_i+\alpha_i(j^*)]^S} \end{aligned} \tag{B.1}$$

If we can simulate $\alpha(j^*)$ directly from equation (B.1), the term $p(\tilde{\alpha}_i) \setminus p(\vec{\alpha}_i)$ in the acceptance rate will cancel, which will simplify the calculation and increase the acceptance rate to 1. However, sampling directly from equation (B.1) is difficult since it's not a common distribution and there is no closed form of the integral. One way to handle this is to choose another proposal function $p(\alpha_i(j^*)))$ that can partly simplify the computation. To do this, we can see that from equation (B.1) is comprised of two parts, one part is $\{z_0 + A_i + \alpha_i(j^*)\}^{-2}$, the other one is $[A_i + \alpha_i(j^*)]^{-S}$. Thus, we can use one of the following two distributions as proposals:

$$g_1(\alpha_i(j^*)) = \frac{z_0+A_i}{[z_0+A_i+\alpha_i(j^*)]^2}$$

$$g_2(\alpha_i(j^*)) \quad = \quad \frac{(S-1)A_i^{S-1}}{[A_i+\alpha_i(j^*)]^S}$$

Graphical inspection of the density functions reveals that $g_2(\alpha_i(j^*))$ is very close to the target distribution and it's easy to simulate from. So we will use equation (B.2) as the proposal function.

$$p(\alpha_i(j^*)) = g_2(\alpha_i(j^*)) = \frac{(S-1)A_i^{S-1}}{[A_i+\alpha_i(j^*)]^S} \tag{B.2}$$

B.2 Proposal function for $u_i(j)$

According to the equations (4.1) and (4.6), for $j = 0, \ldots, S+1$, $i = 1, \ldots, I$

$$p(\tilde{q}_i(j,)|S, \mathbf{d}_{mis}, \mathbf{d}_{obs}, \tilde{\alpha}_i, \Phi_S) = p(q_i(j,), u_i(j)) = \mathcal{D}_{S+1}(\tilde{q}_i(j,)|\ \tilde{\alpha}_i + \tilde{n}_i(j,)),$$

where $\tilde{n}_i(j,)$ is inserting $n_i(j, j^*) = 0$ to $\vec{n}_i(j,)$ and j^* is the newly generated state. If let Be stands for a beta function, then

$$\begin{aligned}
& p(\tilde{q}_i(j,)|S, \mathbf{d}_{mis}, \mathbf{d}_{obs}, \tilde{\alpha}_i, \Phi_S) \\
&= \frac{1}{Be(\tilde{\alpha}_i)} \left\{ \prod_{r=1, r\neq j^*}^{S+1} \tilde{q}_i(j,r)^{\tilde{\alpha}_i(r)+n_i(j,r)-1} \right\} \tilde{q}_i(j,j^*)^{\alpha_i(j^*)-1} \\
&= \frac{1}{Be(\tilde{\alpha}_i)} \prod_{r=1, r\neq j^*}^{S+1} \{[1-u_i(j)]q_i(j,r)\}^{\tilde{\alpha}_i(r)+n_i(j,r)-1} \times u_i(j)^{\alpha_i(j^*)-1} \\
&= \frac{1}{Be(\tilde{\alpha}_i)} [1-u_i(j)]^{\sum_{r=1, r\neq j^*}^{S+1} [\alpha_i(r)+n_i(j,r)-1]} \times u_i(j)^{\alpha_i(j^*)-1} \\
&= \frac{1}{Be(\tilde{\alpha}_i)} u_i(j)^{\alpha_i(j^*)-1} [1-u_i(j)]^{A_i+d_{ij}-S}
\end{aligned}$$

So we can use the proposal function:

$$p(u_i(j)) = \mathcal{B}(\alpha_i(j^*), A_i + d_{ij} - S + 1) \tag{B.3}$$

B.3 Acceptance Ratio for the Birth Moves

According to Green (1997), the acceptance ratio for a birth is

$$R = \min\left\{1, \frac{\pi(d\tilde{\Theta})q(\tilde{\Theta}, d\Theta)}{\pi(d\Theta)q(\Theta, d\tilde{\Theta})}\right\}, \tag{B.4}$$

If current number of state is S, then by equations (4.1)

$$\begin{aligned}\frac{\pi(d\tilde{\Theta})q(\tilde{\Theta}, d\Theta)}{\pi(d\Theta)q(\Theta, d\tilde{\Theta})} &= \frac{p(\mathbf{d}_{obs}, \ldots, \mathbf{d}_{mis}, S+1, \tilde{\Theta}_{S+1}|\tilde{\Phi}, X)}{p(\mathbf{d}_{obs}, \ldots, \mathbf{d}_{mis}, S, \Theta_S|\Phi, X)} \times \frac{d_{S+1}}{b_S(k_0+1)} \\ &\times \quad J\left[p(\vec{\mu}_{j^*}, \Sigma_{j^*})\prod_{i=1}^{I}\left\{p(\alpha_i(j^*))p(\tilde{q}_i(j^*, \cdot))\prod_{j=0, j\neq j^*}^{S+1} p(u_i(j))\right\}\right]^{-1}\end{aligned}$$

Here k_0 is the number of empty states at the current iteration. J is the Jacobian of the transformation from (Q_i, u_i) to $\tilde{Q}_i$ without the j^*th row, and we know

$$J = \prod_{i=1}^{I}\prod_{j\neq j^*}(1 - u_i(j))^{S-1},$$

(Notice in Robert(2000) the $\prod$ in equation (15) on page 64, incorrectly appear as $\sum$.)

If we use the full likelihood function for $p(\mathbf{d}_{obs}, \ldots, \mathbf{d}_{mis}, S, \Theta_S|\Phi, X)$ as in equation (4.2), any parts concerning $\mathcal{N}_2(\vec{y}_{i,t}|\vec{\mu}_{s_i,t}, \Sigma_{s_i,t})$ and $\mathcal{N}_2(\vec{\mu}_s|\ \vec{m}_s, \Sigma_s(\lambda_s)^{-1})\mathcal{W}_2(\Sigma_s^{-1}|\ V_s, \kappa_s)$ for $s = 1, \ldots, S+1$ cancel since we add an empty state and each fixation didn't really change the state it belongs to, only changed the label. So now, for a iteration that currently have S states, by equation (4.2)

$$\begin{aligned}\frac{\pi(d\tilde{\Theta})q(\tilde{\Theta}, d\Theta)}{\pi(d\Theta)q(\Theta, d\tilde{\Theta})} &= \prod_{i=1}^{I}\left\{\frac{\tilde{\pi}_{0i}(\tilde{s}_{i,1})\prod_{t=2}^{T_i}\tilde{q}_i(\tilde{s}_{i,t-1}, \tilde{s}_{i,t})\mathcal{D}_{S+1}(\tilde{\pi}_{0i}|\tilde{\alpha}_i)\prod_{j=1}^{S+1}\mathcal{D}_{S+1}(\tilde{q}_i(j, \cdot)|\tilde{\alpha}_i)}{\vec{\pi}_{0i}(s_{i,1})\prod_{t=2}^{T_i}q_i(s_{i,t-1}, s_{i,t})\mathcal{D}_S(\vec{\pi}_{0i}|\vec{\alpha}_i)\prod_{j=1}^{S}\mathcal{D}_S(q_i(j, \cdot)|\vec{\alpha}_i)}\right\} \\ &\times \quad \prod_{i=1}^{I}\left\{\frac{z_0\Gamma(S+1)}{z_0 + \tilde{A}_i]^2(\tilde{A}_i)^{S+1-1}} \times \frac{(z_0 + A_i)^2(A_i)^{S-1}}{z_0\Gamma(S)}\right\}(\beta+1)(x_{\vec{\mu}_{j^*}})^{\beta}S\end{aligned}$$

$$\times \quad \frac{d_{S+1}}{b_S(k_0+1)} \times \frac{\prod_{i=1}^{I}\prod_{j\neq j^*}(1-u_i(j))^{S-1}}{\prod_{i=1}^{I}\left\{p(\tilde{q}_i(j^*,))\prod_{j\neq j^*}p(u_i(j))p(\alpha_i(j^*))\right\}}, \tag{B.5}$$

where $\tilde{A}_i = A_i + \alpha_i(j^*)$.

Notice that for $\tilde{\pi}_{0i}(\tilde{s}_{i,1})\prod_{t=2}^{T_i}\tilde{q}_i(\tilde{s}_{i,t-1},\tilde{s}_{i,t})$ and $\vec{\pi}_{0i}(s_{i,1})\prod_{t=2}^{T_i}q_i(s_{i,t-1},s_{i,t})$, the only difference is the term $1-u_i(j)$ where $s_{i,t-1}=j$, because the states all correspond to the old states. The power of $1-u_i(j)$ equals the total number of fixations falls in state j in the chain for subject i except the last observation, which is equal to d_{ij}, $j\in\{0,1,\ldots,S+1, j\neq j^*\}$. So this simplifies to $\prod_{j\neq j^*}[1-u_i(j)]^{d_{ij}}$. Now we have:

$$\begin{aligned}
\frac{\pi(\tilde{\Theta})q(\tilde{\Theta},d\Theta)}{\pi(d\Theta)q(\Theta,d\tilde{\Theta})} &= \prod_{i=1}^{I}\left\{\prod_{j\neq j^*}[1-u_i(j)]^{d_{ij}}\frac{\mathcal{D}_{S+1}(\tilde{\pi}_{0i}|\tilde{\alpha}_i)\prod_{j=1}^{S+1}\mathcal{D}_{S+1}(\tilde{q}_i(j,)|\tilde{\alpha}_i)}{\mathcal{D}_S(\vec{\pi}_{0i}|\vec{\alpha}_i)\prod_{j=1}^{S}\mathcal{D}_{S+1}(q_i(j,)|\vec{\alpha}_i)}\right\} \\
&\times \prod_{i=1}^{I}\frac{S(z_0+A_i)^2(A_i)^{S-1}}{(z_0+\tilde{A}_i)^2(\tilde{A}_i)^S}\times(\beta+1)(x_{\vec{\mu}_{j^*}})^{\beta}S \\
&\times \frac{d_{S+1}}{b_S(k_0+1)}\times\frac{\prod_{i=1}^{I}\prod_{j\neq j^*}(1-u_i(j))^{S-1}}{\prod_{i=1}^{I}\left\{p(\tilde{q}_i(j^*,))\prod_{j\neq j^*}p(u_i(j))p(\alpha_i(j^*))\right\}}
\end{aligned} \tag{B.6}$$

Then notice that $\mathcal{D}_{S+1}(\tilde{q}_i(j^*,)|\tilde{\alpha}_i)$ cancels with $p(\tilde{q}_i(j^*,))$, and if we use $\vec{\pi}_{0i}$ as $q_i(0,)$, then for $j\neq j^*$,

$$\begin{aligned}
\frac{\mathcal{D}_{S+1}(\tilde{q}_i(j,)|\ \tilde{\alpha}_i)}{\mathcal{D}_S(q_i(j,)|\alpha_i)p(u_i(j))} &= \frac{Be(\vec{\alpha}_i)}{Be(\tilde{\alpha}_i)}\left\{\prod_{r\neq j^*}\frac{\tilde{q}_i(j,r)^{\tilde{\alpha}_i(r)-1}}{q_i(j,r)^{\vec{\alpha}_i(r)-1}}\right\}q_i(j,j^*)^{\tilde{\alpha}_i(j^*)-1} \\
&\times \frac{Be(\alpha_i(j^*),A_i+d_{ij}-S+1)}{u_i(j)^{\alpha_i(j^*)-1}[1-u_i(j)]^{A_i+d_{ij}-S}} \\
&= \frac{\Gamma(\tilde{A}_i)}{\Gamma(A_i)\Gamma(\alpha_i(j^*))}\left\{\prod_{r\neq j^*}\frac{[(1-u_i(j))q_i(j,r)]^{\alpha_i(r)-1}}{q_i(j,r)^{\alpha_i(r)-1}}\right\}u_i(j^*)^{\alpha_i(j^*)-1} \\
&\times \frac{\Gamma(\alpha_i(j^*))\Gamma(A_i+d_{ij}-S+1)}{\Gamma(\tilde{A}_i+d_{ij}-S+1)u_i(j)^{\alpha_i(j^*)-1}[1-u_i(j)]^{A_i+d_{ij}-S}} \\
&= \frac{\Gamma(\tilde{A}_i)\Gamma(A_i+d_{ij}-S+1)}{\Gamma(A_i)\Gamma(\tilde{A}_i+d_{ij}-S+1)}[1-u_i(j)]^{-d_{ij}} \\
&= \frac{Be(\tilde{A}_i,A_i+d_{ij}-S(S)+1)}{Be(A_i,\tilde{A}_i+d_{ij}-S(S)+1)}[1-u_i(j)]^{-d_{ij}}
\end{aligned}$$

So now the equation becomes:

$$\begin{aligned}\frac{\pi(d\tilde{\Theta})q(\tilde{\Theta},d\Theta)}{\pi(d\Theta)q(\Theta,d\tilde{\Theta})} &= \prod_{i=1}^{I}\prod_{j\neq j^*}[1-u_i(j)]^{d_{ij}}\frac{Be(\tilde{A}_i,A_i+d_{ij}-S+1)}{Be(A_i,\tilde{A}_i+d_{ij}-S+1)}[1-u_i(j)]^{-d_{ij}} \\ &\times \prod_{i=1}^{I}\frac{S(z_0+A_i)^2(A_i)^{S-1}}{(z_0+\tilde{A}_i)^2(\tilde{A}_i)^S}\times(\beta+1)(x_{\vec{\mu}_{j^*}})^{\beta}S \\ &\times \frac{d_{S+1}}{b_S(k_0+1)}\times\prod_{i=1}^{I}\frac{\prod_{j\neq j^*}(1-u_i(j))^{S-1}}{p(\alpha_i(j^*))}\end{aligned} \tag{B.7}$$

And according to equation (B.2):$\frac{(S-1)A_i^{S-1}}{[A_i+\alpha_i(j^*)]^{S-1}}$

$$\begin{aligned}\frac{\pi(d\tilde{\Theta})q(\tilde{\Theta},d\Theta)}{\pi(d\Theta)q(\Theta,d\tilde{\Theta})} &= \prod_{i=1}^{I}\prod_{j\neq j^*}\frac{Be(\tilde{A}_i,A_i+d_{ij}-S+1)}{Be(A_i,\tilde{A}_i+d_{ij}-S+1)} \\ &\times \prod_{i=1}^{I}\frac{S(z_0+A_i)^2(A_i)^{S-1}}{(z_0+\tilde{A}_i)^2(\tilde{A}_i)^S}\times(\beta+1)(x_{\vec{\mu}_{j^*}})^{\beta}S \\ &\times \frac{d_{S+1}}{b_S(k_0+1)}\times\prod_{i=1}^{I}\frac{\prod_{j\neq j^*}(1-u_i(j))^{S-1}[A_i+\alpha_i(j^*)]^S}{(S-1)A_i^{S-1}} \\ &= \frac{S(\beta+1)(x_{\vec{\mu}_{j^*}})^{\beta}d_{S+1}}{b_S(k_0+1)}\prod_{i=1}^{I}\left\{\frac{S(z_0+A_i)^2}{(S-1)(z_0+\tilde{A}_i)^2}\right\} \\ &\times \prod_{i=1}^{I}\prod_{j\neq j^*}\left\{\frac{Be(\tilde{A}_i,A_i+d_{ij}-S+1)}{Be(A_i,\tilde{A}_i+d_{ij}-S+1)}(1-u_i(j))^{S-1}\right\}\end{aligned} \tag{B.8}$$

B.4 Figures and Tables

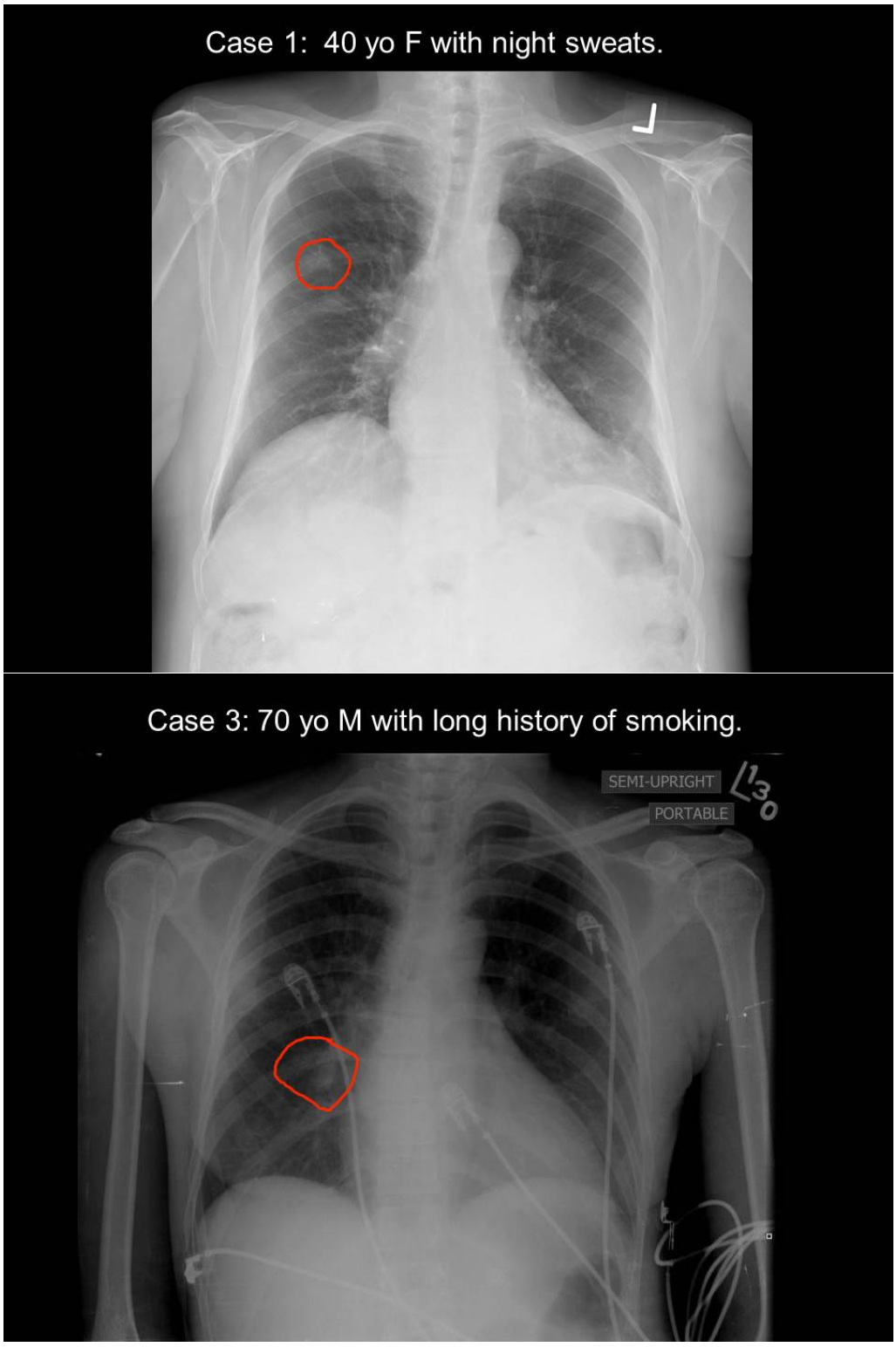


Figure B.1: X-ray images with one nodule present.

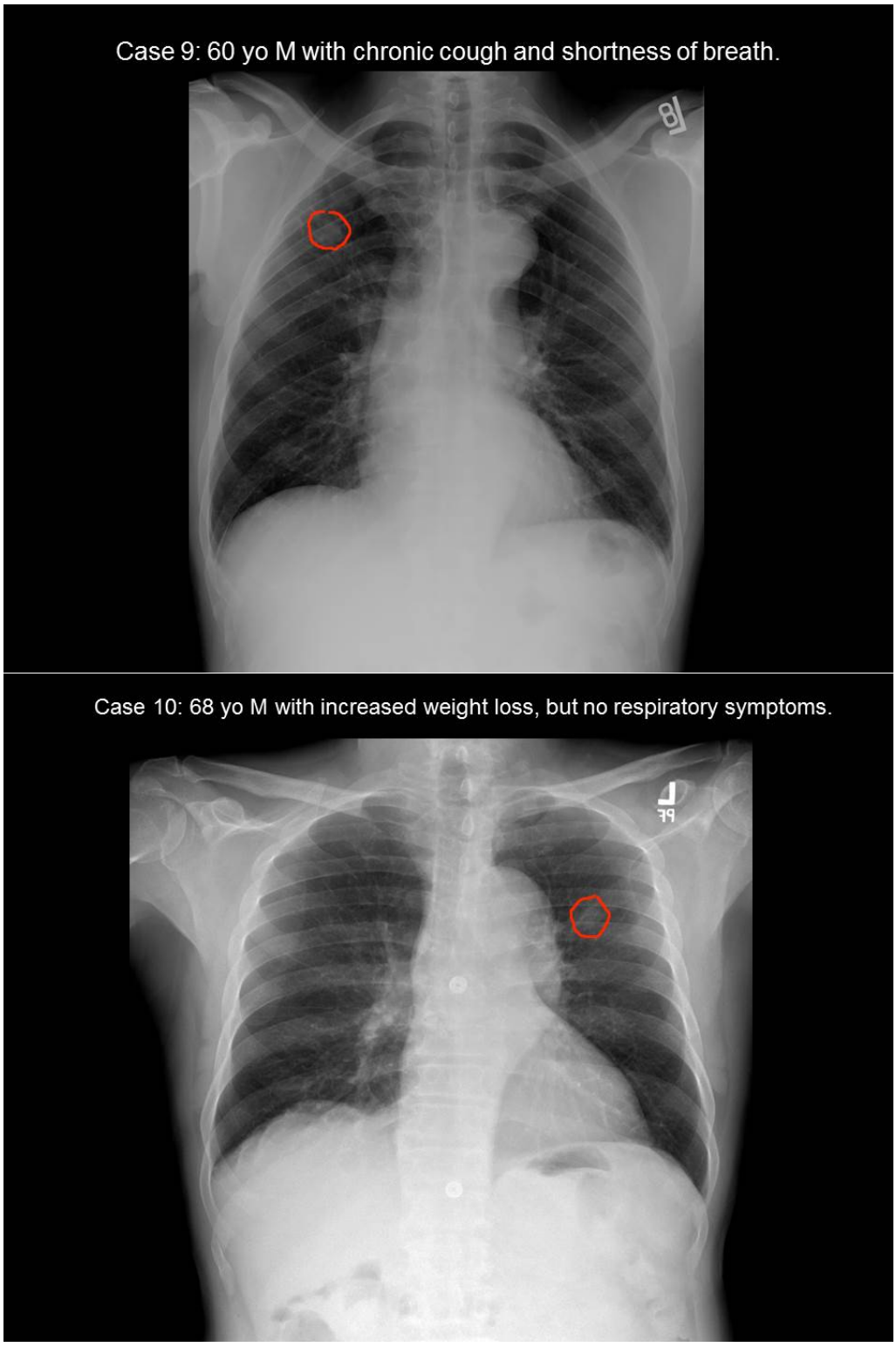


Figure B.2: X-ray images with one nodule present.

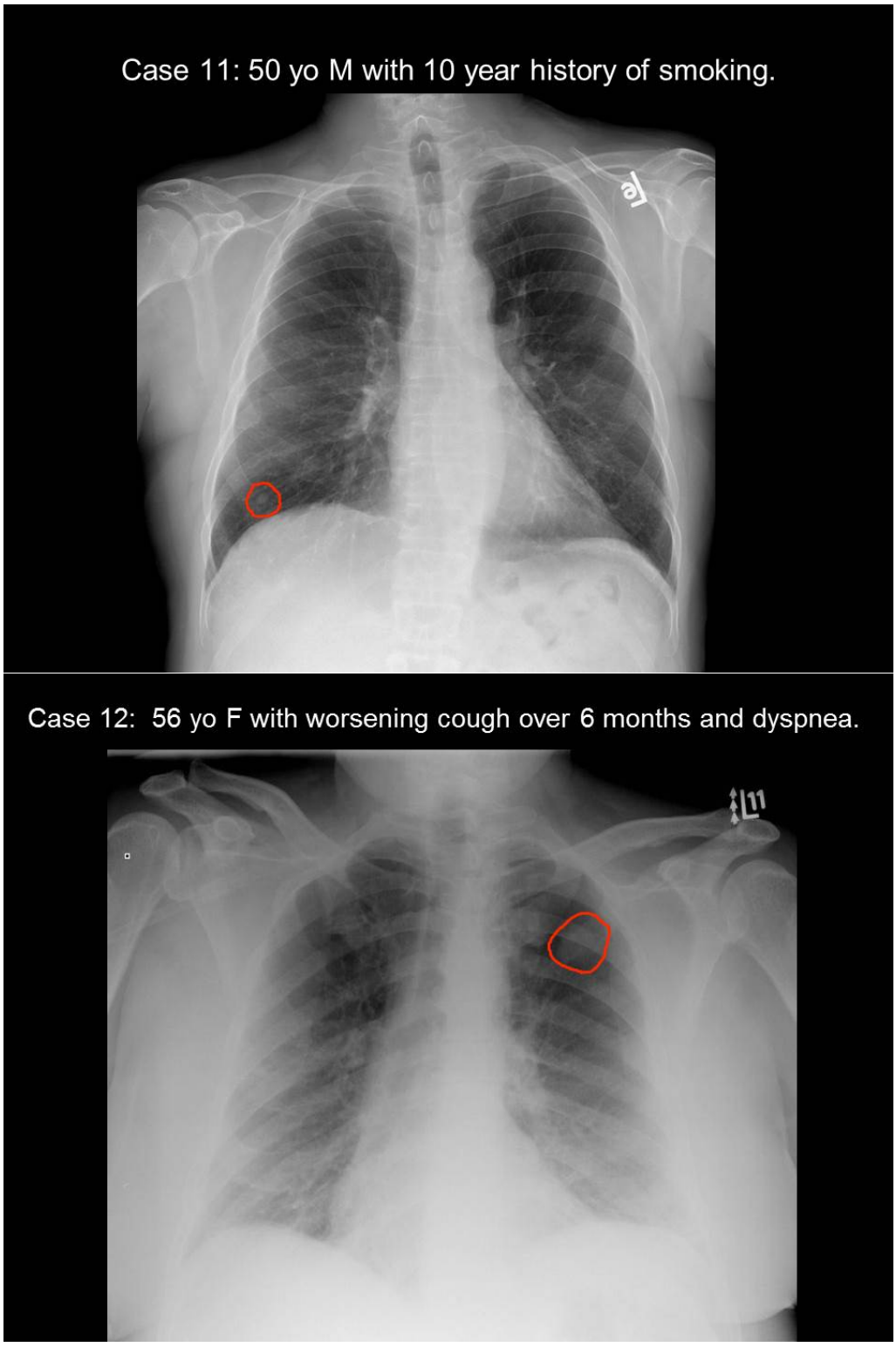


Figure B.3: X-ray images with one nodule present.

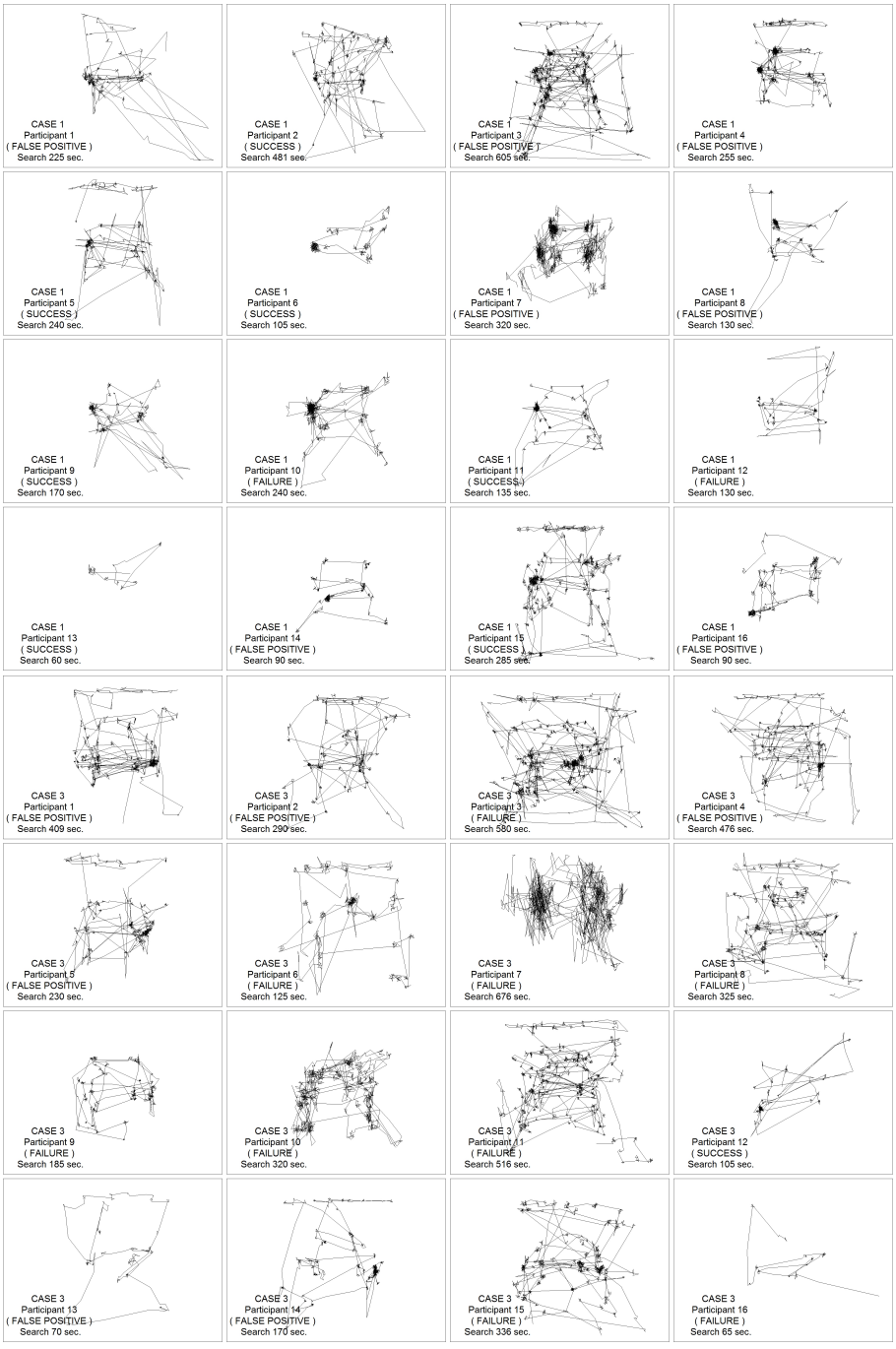


Figure B.4: Scan-paths for 16 participants.



Figure B.5: Scan-paths for 16 participants.



Figure B.6: Scan-paths for 16 participants.

Appendix C

Appendix for HIV Clinical Study

Table C.1: Time windows definition

Week	4	8	12	16
Window (Days)	(1,42]	(42,70]	(70,98]	(98,140]
Week	24	32	49	48
Window (Days)	(140,196]	(196,252]	(252,308]	(308,378]

Bibliography

[1] Al-Awadhi, F., Hurn, M., and Jennison, C. (2004), "Improving the acceptance rate of reversible jump MCMC proposals", *Statistics and Probability Letter*, 69, 189-198.

[2] Albert, J. H., and Chib, S. (1993), "Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts", *Journal of Business and Economic Statistics*, 11, 1-15.

[3] Baum, L. E., and Petrie, T. (1966), "Statistical inference for probabilistic functions of finite state Markov chains", *The Annals of Mathematical Statistics*, 37, 1554-1563.

[4] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics,* 41, 164-171.

[5] Brooks, S.P., Giudici, P., and Roberts, G. O. (2003), "Efficient construction of reversible jump Markov chain Monte Carlo proposal distribution", *Journal of the Royal Statistical Society: Series B,*, 65, 3-55.

[6] Capp, O., Robert, C. P., and Rydn, T. (2001), "Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers," *Technical report, CREST*

[7] Cappé, O., Robert, C. P. and Rydén, T. (2003), "Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers," *Journal of the Royal Statistical Society: Series B* 65, 679-700.

[8] Casella, G., and Robert, C. P. (1996), "RaoBlackwellisation of sampling schemes," *Biometrika,* 83, 81-94.

[9] Celeux, G., Hurn, M., and Robert, C. P. (2000), "Computational and inferential difficulties with mixture prior distributions," *Journal of the American Statistical Association,* 95, 957-970.

[10] Chib, S. (1995), "Marginal likelihood from the Gibbs output", *Journal of American Statistical Association,* 90, 1313-1321.

[11] Cowles, M. K., and Carlin, B. P. (1996), "Markov chain Monte Carlo convergence diagnostics: a comparative review," *Journal of the American Statistical Association,* 91, 883-904.

[12] DeValk, J.P.J., and Eijkman, E.G.J. (1984), "Analysis of eye fixations during the diagnostic interpretation of chest radiographs," *Medical, Biological Engineering and Computing*, 22, 353-360.

[13] Diebolt, J. and Robert, C.P. (1994), "Estimation of finite mixture distributions through Bayesian sampling," *Journal of Royal Statistics Society. Serial B,* 56, 363-375.

[14] Duchowski, A. T. (2003), *Eye Tracking Methodology: Theory and Practice*, New York: SpringerVerlag.

[15] Frieze , A., Kanna, R., and Polson, N.G. (1994), "Sampling from log-concave distributions," *Annals of Applied Probability,* 4, 812-837.

[16] Fruhwirth-Schnatter, S. (2001), "Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models," *Journal of American Statistics Association,* 96, 194-209.

[17] Fruhwirth-Schnatter, S. (ed.) (2006), *Finite Mixture and Markov Switching Models,* New York: SpringerVerlag.

[18] Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association,* 85, 398-409.

[19] Gelfand, A. E., and Dey, D. K. (1994), Bayesian model choice: asymptotics and exact calculations," *Journal of the Royal Statistical Society,Ser. B,* 56, 501-514.

[20] Geman, S.; Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721-741.

[21] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), "Bayesian data analysis," New York: Chapman and Hall.

[22] Geweke, J. (1992), "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in Bayesian Statistics, 4, J.M. Bernardo, A.F.M. Smith, A. P. Dawid, and J. O. Berger(eds.) Oxford: Oxford University Press, pp. 169-193.

[23] Gevers, T. (2001), "Color-based retrieval," in *Principles of Visual Information Retrieval,* ed. M.S. Lew, London: Springer-Verlag, pp. 11-50.

[24] Gilks, W. R. and Wild, P. (1992), "Adaptive rejection sampling for Gibbs sampling", *Applied Statistics* 41, 337-348.

[25] Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika,* 82, 711-732.

[26] Gur, D., Rockette, H.E., Sumkinm, J.H., Hoy, R.J., Feist, J.H., Thaete, F.L., King, J.L., Slasky, Miketic, L.M., and Straub W.H. (1991), "Correlations between time required for radiological diagnoses, readers performance, display environments, and difficulty of cases," *Medical Imaging V: PACS Design and Evaluation,* 1446, 284-288.

[27] Haavelmo, T. (1943), "The Statistical implications of a system of simultaneous equations," *Econometrica,* 11, 1-12.

[28] Herbert, S. (1953), "Causal ordering and identifiability," in Studies in Econometric Method, Hood, W.C., Koopmans, T.C. (eds) New York: Wiley, pp. 49-74.

[29] Higdon, D. M. (1997), "Auxiliary variable methods for Markov chain Monte Carlo with applications", *Journal of the American Statistical Association*, 93, 585-595.

[30] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Halszka J., and Van de Weijer, J. (2011), "*Eye tracking: A Comprehensive Guide to Methods and Measures*," Oxford: Oxford University Press.

[31] Jennrich, R., and Schluchter, M. (1986), "Unbalanced repeated measures models with structured covariance matrices," *Biometrics,* 42, 805-820.

[32] Kojic, E.M., Wang, C. C., and Cu-Uvin, S. (2007), "HIV and menopause: a review," *Jounral of Womens Health,* 16, 1402-11.

[33] Krupinski, E.A., Nodine, C. F., Kundel, H. L (1993), "Perceptual enhancement of tumor targets in chest x-ray images," *Perception and Performance*, 53, 519-526.

[34] Krupinski, E.A., and Roehrig, H. (2000), "The influence of a perceptually linearized display on observer performance and visual search," *Academic Radiology,* 7, 8-13.

[35] Krupinski, E.A., Roehrig, H., Dallas, W.J., and Roehrig, H.(2003), "Searching for nodules: what features attract attention and influence detection?" *Academic Radiology,* 10, 861-868.

[36] Kundel, H.L., Nodine, C.F., and Toto, L.(1974), "Eye movements and the detection of lung tumors in chest images," in Theoretical and Applied Aspects of Eye Movement Research, eds. A.G. Gale AND F. Johnson, Amsterdam: Elsevier Science, pp. 297-304.

[37] Leroux, B. G., and Puterman, M. L. (1992), "Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models," *Biometrics,* 48, 545-558.

[38] Lewellyn-Thomas E., and Lansdown E.L. (1963), "Visual search patterns of radiologists in training," *Radiology*, 81, 288-292.

[39] Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika,* 73, 13-22.

[40] Liu, J. S., Liang, F. and Wong, W. H. (2000), "The multiple-try method and local optimization in Metropolis sampling", *Journal of the American Statistical Association*, 95, 121-134.

[41] Meng, X.-L. (1994), "Posterior Predictive p-Values," *The Annals of Statistics,* 22, 1142-1160.

[42] Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999), "MCMC convergence diagnostics: a review, in Bayesian Statistics 6, eds. J. Berger, J. Bernardo, A. Dawid, D. Lindley, and A. Smith, New York: Oxford University Press, pp. 415-440.

[43] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equations of state calculations by fast computing machines", *Journal of Chemical Physics*, 21, 1087-1092.

[44] Neal, R. M. (2003), "Slice sampling", *Annals of Statistics,* 31, 705-767.

[45] Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian inference by the weighted likelihood bootstrap," *Journal of the Royal Statistical Society, Ser. B,* 56, 3-48.

[46] Nodine, C.F., and Kundel, H.L. (1987), "The cognitive side of visual search in radiology," in Eye movements: from physiology to cognition, eds. J.K. O'Regan and A. Levy-Schoen, Elsevier Science Publishers B.V., pp. 573-582.

[47] Nodine, C.F., Mello-Thoms, C., Kundel, H.L., and Weinstein, S.P. (2002), "Time course of perception and decision marking during mammographic interpretation," *American journal of roentgenology,* 179, 917-23.

[48] Oestmann, J.W., Greene, R., Kushner, D.C., Bourgouin, P.M., Linetsky, L., Llewellyn, H.J.(1988), "Lung lesions: correlation between viewing time and detection," *Radiology,* 166, 451-453.

[49] Patterson, K. B., Cohn, S., Uyanik. J. et al. (2009), "Treatment responses in antiretroviral treatmnet-naïve premenopausal and postmenopausal HIV-1-infected women: an analysis from AIDS clinical trials group studies," CID, 49, 473-476.

[50] Pieters, R. and Wedel, M. (2007), "Goal Control of Attention to Advertising: The Yarbus Implication," *Journal of participant Research,* 34, 224-233.

[51] Propp , J. G. and Wilson, D. B. (1996), "Exact sampling with coupled Markov chains and applications to statistical mechanics," *Random Structures and Algorithms,* 9, 223-252.

[52] Rayner, K.(1977), "Visual attention in reading: Eye movements reflect cognitive processes," *Memory and Cognition,* 5, 443-448.

[53] Rayner, K. (1978), "Eye movement latencies for parafoveally presented words," *Bulletin of the Psychonomic Society,* 11, 13-16.

[54] Richardson, S., and Green, P. (1997), "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society, Ser. B,* 59, 731-792.

[55] Robert, C. P., and Titterington, D. M. (1998), "Reparametrization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation," *Statistics and Computing,* 8, 145-158.

[56] Robert, C. P., Rydn, T., and Titterington, D. M. (1999), "Convergence controls for MCMC algorithms, with applications to hidden Markov chains," *Journal of Statistical Computation and Simulation,* 64, 327-355.

[57] Robert, C. P., Rydn, T., and Titterington, D. M. (2000), "Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method," *Journal of the Royal Statistical Society, Ser. B,* 62, 57-76.

[58] Robert , G. O. and Polson, N. G. (1994), "Methods for estimating L^2 convergence of Markov chain Monte Carlo,"I in Bayesian Analysis in Statistics and Econometrices: Essays in Honor of Arnold Zellner, D. Berry, K. Chaloner, and J. Geweke (eds.). New York: John Wiley and Sons, pp. 373-384.

[59] Robert , G. O. and Rosenthal, J. S. (2006), "Examples of adaptive MCMC," in Harris recurrence of Metropolis-within-Gibbs and Trans-dimensional Markov.

[60] Rubin, D. B. (1984), "Bayesianly justiciable and relevant frequency calculations for the applied statistician," *The Annals of Statistics,* 12, 1151-1172.

[61] Rosenthal, J.S. (1995a), "Minorization conditions and convergence rates for Markov chain Monte Carlo," *Journal of American Statistical Association,* 90, 558-566.

[62] Rosenthal, J.S. (1995b), "Rates of convergence for Gibbs sampling for variance components models," *Annals of Statistics,* 23, 740-61.

[63] Scott, S. L. (1999), "Bayesian Analysis of a Two-State Markov Modulated Poisson Process," *Journal of Computational and Graphical Statistics,* 8, 662-670.

[64] Scott, S. L. (2002), "Bayesian methods for hidden Markov models: recursive computing in the 21st Century," *Journal of the American Statistical Association,* 97, 337-351.

[65] Scott, S. L., James, G.M., and Sugar, C. A. (2005), "Hidden Markov Models for longitudinal comparisons," *Journal of the American Statistical Association,* 100, 359-369.

[66] Sommers, P.M., and Conlisk, J. (1979), "Eigenvalue immobility measures for Markov chain," *Journal of Mathematical Sociology,* 6, 253-276.

[67] Starner, T., and Pentland, A. (1995), "Real-time American sign language visual recognition from video using hidden Markov models", Master's Thesis, MIT, Program in Media Arts.

[68] Stephens, M. (1997a), "Bayesian methods for mixtures of normal distribution," D.Phil. dissertation, Deptment of Statistics, Univerity Oxford.

[69] Stephens, M. (1997b), "Discussion of "On Bayesian analysis of mixtures with an unknown number of components"," by S. Richardson and P. J. Green, *Journal of Royal Statistics Society. Serial B,* 59, 768-769.

[70] Stephens, M. (2000a), "Bayesian Analysis of Mixtures With an Unknown Number of Components: An Alternative to Reversible Jump Methods," *The Annals of Statistics,* 28, 40-74.

[71] Stephens, M.(2000b), "Dealing with label switching in mixture models," *Journal of Royal Statistics Society. Serial B,* 62, 795-809.

[72] Stratonovich, R.L. (1960), "Conditional Markov processes", *Theory of Probability and its Applications*, 5, 156-178.

[73] Tjemeland, H. and Hegstad, B. K. (2001), "Model jumping proposals in MCMC", *Scandinavian Journal of Statistics*, 28, 205-223.

[74] Trends in HIV/AIDS diagnoses-33 states (2005), "MMWR", 54, 1149.

[75] Van der Lans, R., Pieters, R., and Wedel, M. (2008), "Eye movement analysis of search effectiveness," *Journal of the American Statistical Association,* 103, 452-461.

[76] Van der Lans, R., Wedel, M., and Pieters, R. (2011), "Defining eye-fixation sequences across individuals and tasks: the Binocular-Individual Threshold (BIT) algorithm," *Behavior Research Methods,* 43, 239-257.

[77] Van Dyk, D. A., and Meng, X.-L. (2001), "The art of data augmentation," *Journal of Computational and Graphical Statistics,* 10, 1-50.

[78] Wedel, M. and Pieters, R. (2000), "Eye Fixations on Advertisements and Memory for Brands: a Model and Findings," *Marketing Science,* 19, 297-312.

[79] Wedel, M. and Pieters, R. (2006), "Eye Tracking for Visual Marketing," *Foundations and Trends in Marketing,* 1, 231-320.

[80] Wedel, M., and Pieters, R. (eds.)(2008), *Visual Marketing,* New York: Lawrence Erlbaum Associates.

[81] Wright, S. S. (1921), "Correlation and causation," *Journal of Agricultural Research,* 20, 557-85.

[82] Zeger, S. L. and Liang, K.-Y. (1986), "Longitudinal data analysis for discrete and continuous outcomes," *Biometries,* 42, 121-130.

[83] Zelinsky, G.J. (2008), "A theory of eye movements during target acquisition," *Psychological Review,* 115(4), 787-835.