Bayesian Factor Analysis for fMRI Data

by

Lin Sun

(Under the direction of Nicole Lazar)

Abstract

Functional magnetic resonance imaging (fMRI) can give excellent visualization of the activity location in the brain resulting from cognitive function or sensory stimulation . BOLD fMRI takes advantage of the fact that local blood flow increases following an increase in neuronal activity. Among the various fMRI analysis techniques, the Bayesian factor analysis has been widely used for assessment of multivariate dependence and codependence, which solve the problem that the parameters can't be uniquely determined from the likelihood alone in classical factor analysis. In this study, BOLD measurement of the acute effect, a substance shown previously to engage multiple sites within the orbitofrontal cortex, was processed with the Bayesian factor model for comparative group. The flexibility of the Bayesian factor analysis was shown by choosing different modeling strategies to form the prior reference functions, including approximating simultaneously measured behavioral data and the effect of transformed stimulus in semiparametric Bayesian factor model. Bayesian factor analysis provides a powerful approach to understand BOLD response. Several factors have been determined to explain most of variance within the data. It is demonstrated that Bayesian factor analysis successfully associates the activated activity with the Bayesian factors.

Index words: functional Magnetic Resonance Imaging, Bayesian factor analysis, Semiparametric approach, Normalized Cut)

Bayesian Factor Analysis for fMRI Data

by

Lin Sun

B.A., Renmin University of China, 2000

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2013

Bayesian Factor Analysis for fMRI Data

by

Lin Sun

Approved:

Major Professor: Nicole Lazar

Committee: Gauri Datta
Daniel Hall
Jaxk Reeves
Lynne Seymour

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2013

Acknowledgments

I would like to first thank my advisor, Professor Nicole Lazar, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the research. This dissertation would not have taken shape without her invaluable input. Dr. Lazar introduced me to the field of fMRI analysis. Her insight and experience have guided me throughout my research during which time she provided numerous invaluable suggestions. It was a great pleasure for me to conduct this dissertation under her supervision. I would also like to thank Dr.Gauri Datta, Dr. Daniel Hall, Dr. Jaxk Reeves and Dr. Lynne Seymour for their willingness to serve on my committee. I have benefitted by their suggestions of this dissertation.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the dissertation.

And finally, most importantly, I thank my family. I thank my mother and father for everything and my sister, too. And of course I thank my dearest Fei for his understanding and love during the past few years. Their support and encouragement are my source of strength.

Lin Sun

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Functional MRI

This chapter outlines the mechanism of functional Magnetic Resonance Imaging (fMRI) in studying human brain function, and some fundamental features of this imaging modality. The first section briefly introduces some basic facts of Magnetic Resonance Imaging (MRI). In the second section, we discuss in detail the imaging principles of fMRI, in which we focus on the difference between functional MRI and regular MRI. In the last section of this chapter, we'll give a brief overview of the statistical methods that have been proposed to analyze fMRI data.

1.1 What is MRI?

Magnetic resonance tomography (MRT) or magnetic resonance imaging (MRI) is a imaging technique commonly used to visualize the structure and function of the body (Novelline, 1997). MRI offers superior contrast between different soft tissues than computed tomography (CT) does, it is therefore used extensively for cardiovascular, neurological, musculoskeletal, and oncological (cancer) imaging. Physicians use it to diagnosis many kinds of conditions and injuries because of the great ability to customize the exam to the specific medical question being asked. For example, by changing exam parameters of the MRI system, tissues in the body may appear differently in the image. This is quite helpful to the radiologist in determining whether those visible structures are normal or not.

Magnetic resonance imaging is a relatively new technology compared with other popular imaging techniques. For example, the X-ray imaging was first invented in 1895 by a German physicist. It wasn't until the late 1960s, by contrast, that Raymond Damadian discovered

that malignant tissue had different magnetic resonance parameters than normal tissue. He concluded that it should be possible to do tissue characterization based on these differences(Blink, 2004). However, he did not describe a method for generating pictures from such scans. Paul Lauterbur found a way to generate the first MR images, in two and three dimensions, using gradients (Lauterbur, 1973). In 1974, he produced the first ever MR image of a rat tumor. Damadian built the first full body MRI machine, which was used to produce the first image of the human body (Damadian et al., 1977) .

Since its invention, the past several decades have witnessed a rapid increase of information regarding the role of MRI in assessing pathologic conditions of various anatomical structures. The hardware and software became faster, easier to use and more intelligent. More recently, with the development of advanced MRI pulse sequences, people find many more applications for MRI, such as MR angiography, functional imaging and diffusion scanning.

1.1.1 How does MRI Work?

Atoms are the essential building blocks of all matter in this world, and the human body contains billions of them. The atom nucleus spins, or precesses, on an axis. The cylindrical tube of an MRI scanner contains a very powerful electro-magnet (Lauterbur, 1973), and a typical research MRI scanner has a field strength of 3 Tesla (T). Inside the scanner, the magnetic field affects the magnetic nuclei of atoms of body tissues. Atomic nuclei are usually randomly oriented, but they become aligned with the guidance of the magnetic field due to the influence of the electro-magnet (Novelline, 1997). The degree of alignment grows when the magnetic field becomes stronger. The tiny magnetic signals from individual nuclei aggregate when they are all performing in the same direction, which resulted in a signal that is strong enough to measure.

A second radiofrequency (RF) electromagnetic field is then switched on briefly, and it cause the protons to absorb some of its energy (Haacke et al., 1999). Radiofrequency energy, which is generated by electrons traveling through loops of wire, is in the form of fast changing

magnetic and electric fields, and the direction of current flow is rapidly changing back and forth at "radio frequencies", When the field is switched off, the protons unleash the energy at a radiofrequency that can be found by the scanner.

The magnetic field produced by the flow of electrons is an effect of changing across the body (a field gradient). The linearly changing magnetic field is producted by switching gradient coils on and off, so then varied spatial locations change related with varied precession frequencies, and these field gradients are usually pulsed. Field gradients and radiofrequency excitation are then combined to construct the MR image using the two-dimensional Fourier transform (2DFT) with slice selection (Haacke et al., 1999) or by the three-dimensional Fourier transform (3DFT) method.

Since the protons in different tissues revert to their equilibrium state at varied rates (also known as "relaxation rate"), different types of tissues will have different contrast in the MR image. Several variables associated with tissues, e.g. spin density (also known as proton density), T1 and T2 relaxation times (also known as longitudinal relaxation and transverse relaxation respectively), can be used to construct MR images (Hendee and Morgan, 1984). Therefore by modifying the parameters of the scanner, we can use the effect that different tissues have different relaxation rates to generate contrast between different types of tissues, or between other properties (Hendee and Morgan, 1984), such as in fMRI and diffusion MRI. In the next section, we will introduce in the detail the imaging principles of fMRI with a focus on the difference between functional and regular MRI.

1.2 Functional MRI

Functional MRI is a noninvasive medical imaging approach that uses the blood oxygenation level dependent (BOLD) effect to map brain function (Blink, 2004). Functional MRI can give excellent visualization of the activity location in the brain resulting from cognitive function or sensory stimulation (Pekar, 2006). The study of the blood flow or circulation is called "hemodynamics". BOLD fMRI makes use of the fact that local blood flow increases

following an increase in neuronal activity. It has become an important and powerful tool for investigating the blood flow or circulation changes that occur within the working human brain. fMRI therefore enables us to study how the normal brain functions, how the brain is affected by varied diseases, how drugs can regulate brain activities and how brain tries to recover after damage. In recent years, fMRI has played an important role in neuroscience research, and is becoming useful clinically as well, for example, presymptomatic diagnosis and surgical planning (Smith, 2004).

1.2.1 fMRI Principle: The BOLD Effect

The detection of brain areas which are used during task performance (loosely defined) is based on the BOLD effect, which gives a relative measure of oxy-/deoxy-hemoglobin concentration. Oxygen is delivered to neurons in capillary red blood cells by hemoglobin. There is an increased demand for oxygen when neuronal activity increases, which would lead to a growth in cerebral blood flow to areas of increased neural activity. Through the hemodynamic response, blood releases oxygen to the active neurons at a greater rate than to the inactive ones. Hemoglobin is diamagnetic (substance that is repelled by magnetic fields) when oxygenated but paramagnetic (substance that is attracted by magnetic fields) when deoxygenated; this difference leads to a slight difference in the magnetic resonance signal of blood relying on the level of oxygenation. Since blood oxygenation changes according to the levels of neural activity, those differences can be used to detect brain activity. Figure 1.1 Devlin (2005) illustrates blood level change to an activated region; we can see clearly from the Figure that when the neuronal activity increases, so does the concentration of oxygenated hemoglobin. This form of MRI is acknowledged as blood oxygenation level dependent (BOLD) imaging, and is the most common form of functional MRI.

In a typical fMRI study designed to identify brain regions involved in a particular task, which could be language, sensory, visual, auditory and other targeted stimulations, subjects

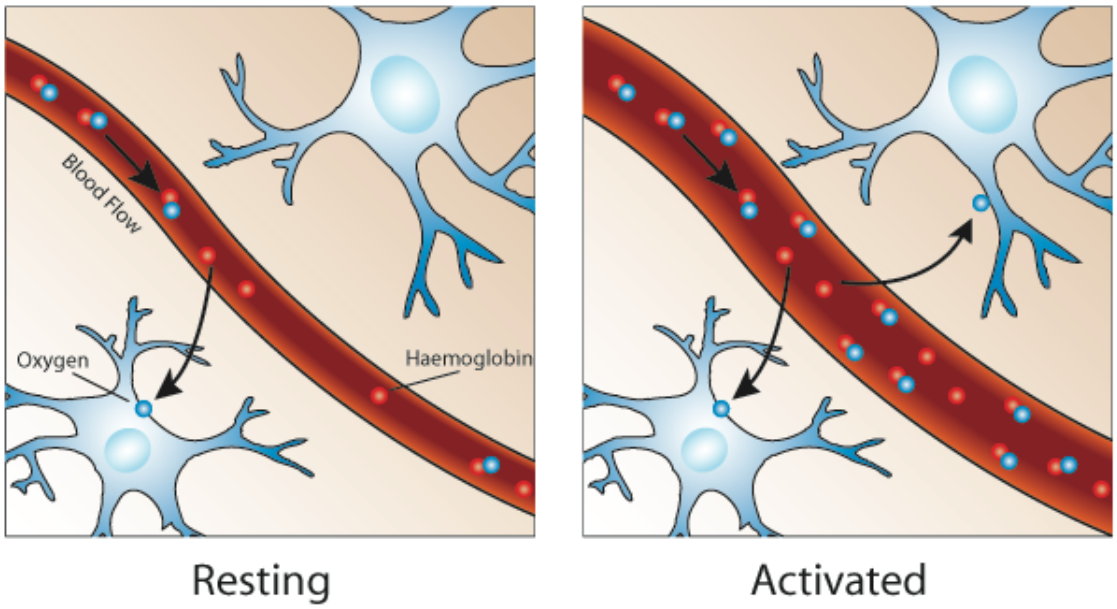

Figure 1.1: Illustration of the blood oxygenation level change during resting (left) and activation (right)(Devlin, 2005).

are located in the scanner to result a conventional scan, and plane lines are set based on conventional imaging methods. The patient performs the task while being imaged. It typically takes between 15 minutes and an hour for an experiment, depending on the goals of the study.

The brain is scanned at a fixed resolution and at a rapid rate, typically once every 2 to 3 seconds. A functional volume is acquired for each image. A single volume is comprised of individual cubical elements called "voxels", and the voxels in the MRI image usually represent cubes of tissue around 2 to 4 mm on every side in humans. More recently, the use of high magnetic fields (van der Zwaaga et al., 2009) and multichannel radiofrequency reception (Griswold et al., 2002), have advanced millimeter scale spatial resolution (1 mm on each side of the cube).

The volume data are typically acquired slice by slice using single-shot echo planar imaging (EPI), which is one of the fastest MRI techniques. As we introduced in the last section,

the magnetic field is induced to vary across the body (a field gradient) during the data acquisition, so that different spatial locations are associated with different precession frequencies. Then, in the presence of a static gradient magnetic field, a radio frequency (RF) pulse is applied to selectively excite nuclear spins of a single slice; After that, the slice-select gradient has been turned off, and the signals resulted from those spins are translated along the dimensions of the slice using fast switching magnetic field gradients (Pekar, 2006). Once the dataset is acquired, Fourier transformation is applied to generate the actual images.

As we explained earlier in the MRI physics section, the relaxation rate is an intrinsic property of the MR signal decay behavior. Its value reflects BOLD effects due to the changes in local magnetic field inhomogeneity that accompany the changes in the oxy-/deoxy-hemoglobin balance, i.e., a magnetic susceptibility effect. Because the images are gotten using an MR sequence, and it is sensitive to variations in local blood oxygenation level, certain region of the images gotten during stimulation can show increased intensity, compared with those gotten while at rest. These higher intensity regions should correspond to the brain regions which are activated by the stimulation. Statistical methods can then be applied to determine reliably these areas of the brain which are associated with this difference as a result, i.e. areas of the brain that are active during the task, process or emotion. Figure 1.2 shows a sample fMRI scan of the brain during auditory exposure.

The activated region resulting from the auditory stimulation is shown in color, which is superposed with axial, coronal and sagittal views of the MRI scan respectively in Figure 1.2a-c. Imagine a human body standing upright, the axial plane is also known as the horizontal plane and it separates the brain into upper and lower regions; the sagittal plane is a up-and-down plane which checks through the body from front to back, and it separates the body into right and left regions; the coronal plane is known as the frontal plane which separates the brain into front and back regions; a 3D view of the activated region on the brain surface is shown in the fourth row.

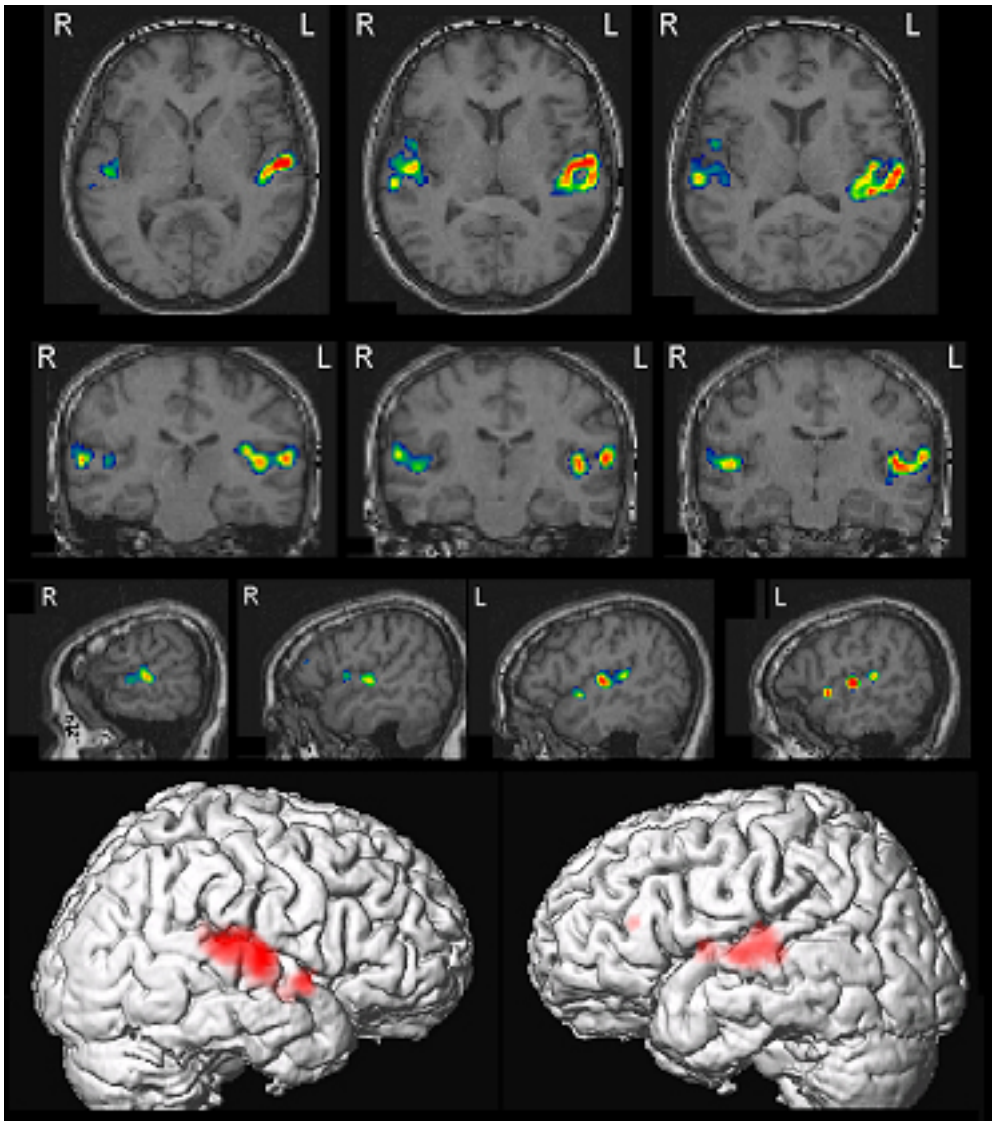

Figure 1.2: A sample fMRI scan of the brain during auditory exposure. The first to third rows show the axial, coronal and sagittal views of the MRI scan respectively. For the axial view, the forehead is at the top, and the back of the head is at the bottom; 3D view of the brain surface is shown in the fourth row. The activated region resulting from auditory exposure is shown in color (oorsuizen.be, 1996).

Typically compared with the noise level, the BOLD response signal is weak. Figure 1.3 shows the BOLD signal changes after the onset of the stimulus, where the subjects are exposed to periodical visual stimulus. In the Figure, the upper line is the visual stimulus presented to the subject (1 : Stimulus turned on, 0 : Stimulus turned off), and the lower line is the magnitude of the BOLD signal, which is measured from the intensity of a single voxel of the fMRI data. From the Figure, we observe that even from a strong visual stimulus, the BOLD signal is quite noisy, which makes brain activation detection a difficult problem. Therefore the goal of fMRI analysis is to robustly and accurately detect those regions of the brain, which are activated by the stimulus, since fMRI data may vary greatly between different human subjects and under different stimulus signals.

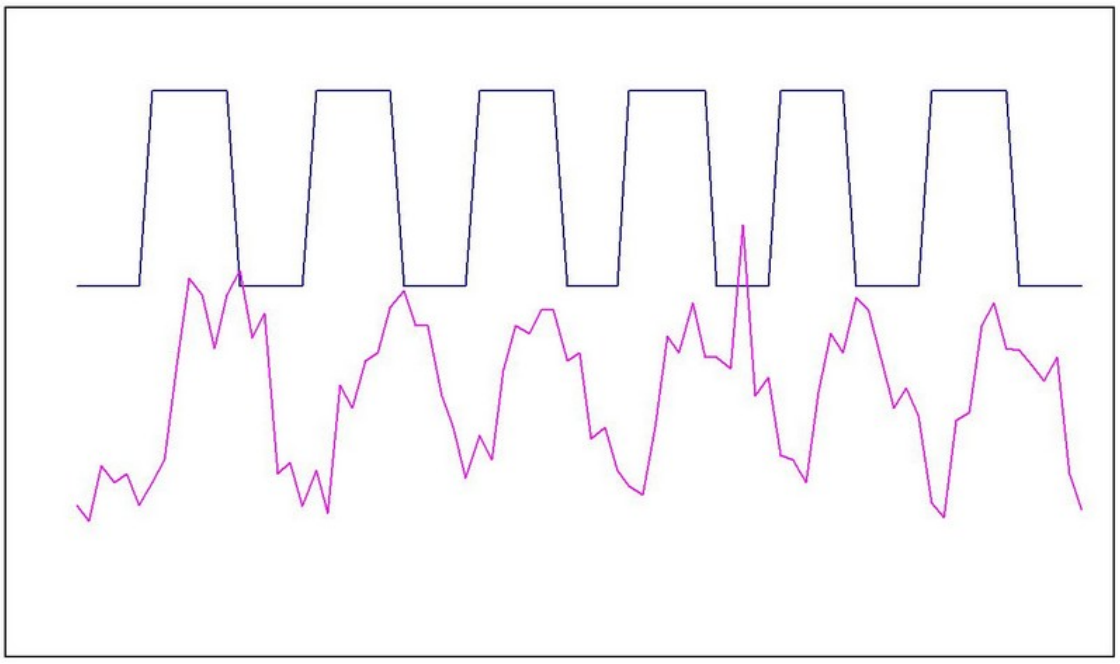

Figure 1.3: Example time series at a strongly activated voxel from fMRI data. Upper line is the stimulus presentation trail, lower line is measured BOLD response. In practice, most voxels will not exhibit such a strong activation pattern.

1.2.2 Statistical Analysis of fMRI Data: An Overview

Statistical methods can not only be applied to reduce the effect of noise in fMRI images, but they can also help scientists to explain the fMRI images, specifically to assist researchers

to robustly detect and localize the activated regions of BOLD responses, thereby enabling us to associate different regions with different cognitive tasks or stimuli. Some statistical methods for fMRI analysis are shown in Figure 1.4.

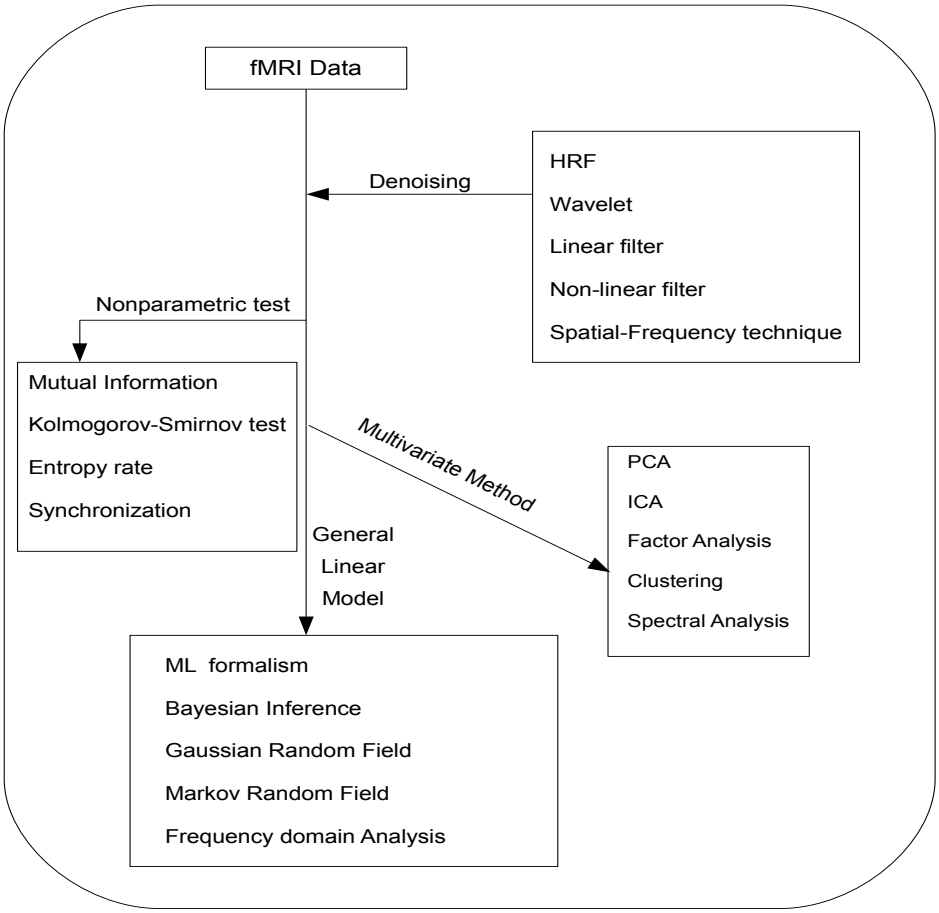

Figure 1.4: Overview of the various methods for fMRI data analysis.

Statistical methods such as the General Linear Model (GLM) have been used to identify functionally specialized brain responses. The GLM is probably the most popular approach to characterizing functional anatomy and disease-related changes (Friston et al., 1995). At each voxel, GLMs are modelled to the fMRI time series, resulting in a group of voxel-specific parameters which are then used to form Posterior Probability Maps (PPMs) or Statistical

Parametric Maps (SPMs). Those maps characterize regionally specific responses to experimental operation. However, the GLM fails to take into consideration the correlations between voxels. To overcome this limitation, parametric and semiparametric spatio-temporal models are used to determine the activity brain correlations at a given actived voxel in adjacent voxels. For instance, spatial correlations by incorporating the time series from neighboring voxels is addressed by Katanoda et al. (2002). Furthermore, Gössl et al. (2001a) use a Bayesian framework conditional autoregressive (CAR) model to determine the correlation between surrounding voxels. Woolrich et al. (2004) consider a Bayesian approach for fMRI data which incorporates spatio-temporal noise modeling and hemodynamic response function (HRF) modeling. Penny et al. (2005) present a Bayesian model with the spatial correlative prior distributions assumption of regression coefficients. The downside of these approaches is that their models usually contain many unknown parameters that need to be estimated, therefore they are quite time consuming for the large datasets which are frequently encountered in fMRI.

The fMRI data typically comprise hundreds of thousands of voxels and hundreds of time points for a single subject, and usually there will be multiple subjects. Dimension reduction techniques are often used to not only alleviate the computational complexity of fMRI processing, but also to potentially help the scientist solve the challenges of finding "good" subspaces in explaining the fMRI data. One prominent example is the so-called "factor models", which explain variability among observed random variables with fewer unobserved random variables called "factors". Factor models can be used to reduce the dimensionality of fMRI data, thereby enabling us to use the smaller number of variables to find meaningful structure in the observations.

The main aim of this dissertation is to explore the issues involved in using factor models, and specifically Bayesian factor models, for analyzing fMRI data. Factor models have been used for the analysis of fMRI data in the past. Langers, for example, compares factor analysis and independent component analysis in fMRI data blind source separation (Langers, 2009).

To the best of my knowledge, however, the Bayesian factor method is a new method in fMRI. Li et al. (2010) apply a variational Bayesian factor partition (VBFP) to resting-state fMRI data. There is no prior work that provides a detailed analysis of fMRI data with the Bayesian factor model. I'm especially interested in finding the answers to the questions: "How many factors are enough to explain the underlying structure of the fMRI data?", and "Which voxels correspond to specific factors in the factor model?". With those questions in mind, we are hoping to achieve the following specific goals.

1. With the factor model, two or more variables can be combined into a single factor, therefore the dimension of the problem is reduced. The datasets acquired for fMRI are typically huge and difficult to interpret; with a relatively small number of factors, we will obtain reduced computational complexity of analysis algorithms, and easier interpretation of the fMRI data.

2. By using a factor model, we can identify groups of variables (factors), to see how they are correlated to each other. This will help us to understand differences among activated regions, as a result of a particular cognitive function. Consequently, we could potentially identify hidden dimensions of the data which may not be apparent from direct analysis. This could lead us to new findings of previously unknown factors associated with different cognitive functions. Therefore, we can understand more about brain activation and brain function.

3. The factor model can help us to identify different brain regions associated with particular cognitive functions, which is a major focus of fMRI analysis.

1.3 Outline of Remainder

In the next chapter, we will introduce factor analysis and then review the existing algorithms for factor analysis; this is followed by an extensive discussion of the convergence issue in Markov Chain Monte Carlo (MCMC) methods, which is critical for the Bayesian factor

models. To facilitate the analysis of the data, we will first partition the image into disjoint regions to ease implementation of the factor models, and to save computation time. Therefore we will introduce suitable image segmentation approaches, i.e. normalized cut segmentation for delineation of anatomical structures in fMRI data.

In Chapter 3, we show the technical details of using the Bayesian factor model to analyze functional MRI data, which is the focus of this dissertation. Priors and posterior estimation are critical components of the Bayesian factor models; these issues will be discussed in the same chapter. We present results from applying Bayesian factor models on real fMRI data.

In Chapter 4, we discuss the details of a semiparametric Bayesian factor model and its application to fMRI data analysis. Specifically, we will discuss whether we should add a hemodynamic response function (HRF) to explain brain activation, and try to get better fitting results with the incorporation of semiparametric Bayesian factor models. The results of using this model are demonstrated and discussed in this chapter.

Lastly, we end with some concluding points and thoughts for future work in Chapter 5.

Chapter 2

Factor Models and Useful Concepts

Factor analysis is a statistical data reduction method. Factor analysis is used to identify hypothetically unobserved variables, or factors (Gorsuch, 1983), which is with the lower number than the number of observed variables. It has been employed to examine a aggregation of data sets, with important experiments design applications, physical science (geochemistry, ecology, and hydrochemistry) (Subbarao et al., 1995) and economics (Berry, 1960). More recently researchers have applied factor analysis to health care data such as DNA microarray, and medical images (fMRI, MRI and CT) (Reyment and Jöreskog, 1996; Machado et al., 1999).

In this chapter, we will introduce the definitions of the factor models and various techniques that have been used for factor analysis. In Section 2.1, we introduce factor models and we also introduce the Markov chain Monte Carlo (MCMC) method and MCMC convergence in Section 2.2 and 2.3. In the context of using factor models to analyze fMRI data, we will partition the image into disjoint regions and apply the procedure on the regions instead of on the individual voxels. This saves computation time without much loss of information. In Section 2.4, we introduce suitable image segmentation approaches that can be applied to partition the image into semantically meaningful regions. In Section 2.5, we introduce Mutual Information to quantify the differences among the images.

2.1 Factor Analysis Model

2.1.1 Introduction

Factor analysis is a class of statistical way used to analyze a large set of interdependent variables and to get a small set of underlying factors to explain the variables. By now, the history of factor analysis spans more than 100 years. Factor analysis began with Charles Spearman (Spearman, 1904). In the first half of the Twentieth Century, psychologists mainly developed factor analysis for the objective of identifying mental abilities by means of psychological testing. Factor analysis took a more general form through Jöreskog's developments in maximum likelihood factor analysis (Jöreskog, 1969).

Principal component analysis (PCA) is another variable-reduction procedure that is related to factor analysis. PCA uses an orthogonal linear transformation to convert a set of possibly related variables into unrelated variables with smaller number called principal components. However, PCA and factor analysis are not identical. One major difference is that PCA produces a variable rotation with the maximum variability, on all variability in the variables, while factor analysis measures how much of the variability is explained by common factors. Another difference is that factor analysis assumes the measured responses are based on underlying and unique factors; in PCA, the principal components are based on unexplainable constructs. PCA is suggested if simple data reduction is of interest. Factor analysis reveals the underlying factors that are responsible for the observed variables. If we assume the variances of the errors in factor analysis are all the same, PCA and factor analysis are roughly equivalent on mathematical definition but different in interpretation.

There are three main reasons why we use factor analysis, specifically,

1. To reduce the large number of variables, we can use a smaller number of factors to find meaningful structure in the observations. For our research, we can use a small set of factors to explain thousands of fMRI voxels.

2. To treat a set of factors as uncorrelated variables for use in multiple regression.

3. To verify the factors that illuminate one variable, or verify one factor that gives insights to a set of variables.

We will next formally define the factor model and various popular techniques that have been used in the past for factor analysis.

2.1.2 Factor Analysis Method

Letting y_i denote the n-vector of response variables ($i = 1 \dots T$), f_i ($i = 1 \dots k$) denote the k-vector factors, where $k < T$, the factor model is expressed as

$$y_i = \beta f_i + \epsilon_i, \tag{2.1.1}$$

or in matrix notation,

$$Y = F\beta' + E \tag{2.1.2}$$

Here, β $(n \times k)$ is the matrix of standardized coefficients - these standardized coefficients denote the correlations between the original variables and the factors. The squared factor loading for a given variable, is the percentage variance in this variable that is explained by a factor, and it is analogous to Pearson's correlation coefficient in regression. f_i is a k-dimensional vector of unobserved "common" factors (factor scores) for the i^{th} variable, and F is a $T \times k$ matrix. Factors are a small number of "latent variables" that explain the observed structures. We impose that factors have zero mean, and variances all equal to 1, and for the orthogonal factor model, they are all orthogonal to each other, i.e. they are uncorrelated. ϵ_i is a n-dimensional error vector with zero mean and finite variance, which is independent of the factors F by assumption.

With factor analysis, we can identify the number of underlying factors responsible for covariation in the observed variables. We need to determine whether there are k uncorrelated unobserved variables that describe the relationship among the n observable variables, where $k \ll n$.

We consider the covariance between the observations and the factors,

$$\begin{aligned} \text{cov}(y_i, f_i) &= E(y_i f_i') - E(y_i)E(f_i') \\ &= E[(\beta f_i) f_i'] \\ &= E(\beta f_i f_i') \\ &= \beta E(f_i f_i') \\ &= \beta R \end{aligned} \tag{2.1.3}$$

Under the orthogonal factor model (the factor matrix F is orthogonal), $R = I_k$, the factor loading matrix β is interpreted as a covariance matrix between the n observed variables and the k unobserved factors (i.e. β_{ij} explains the covariance between y_i and f_j.).

As we observed from the model definition, the key problem for factor analysis is to estimate the free parameters β, as well as the factors f_i. In the next sections, we will introduce various popular techniques that have been used for this estimation.

2.1.3 Maximum Likelihood Factor Analysis

There are many approaches in the literature which are used to estimate β and f_i. One of the most popular is the maximum likelihood method (Lawley and Maxwell, 1962). Maximum likelihood factor analysis linearly combines variables to form factors, where the estimators of parameters are those most possibly to have produced in the observed correlation matrix, using maximum likelihood estimation approach and assuming a multivariate normal distribution.

As before, the model assumes that for $i = 1 \dots T$, the f_i are independent with $f_i \sim N(0, I_k)$. We assume independent normal n-vectors errors, $\epsilon_i \sim N(0, \Sigma)$, where $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. In addition, we assume that, for all i and j, f_i and ϵ_j are independent, i.e. $\text{cov}(f_i, \epsilon_j) = 0$, $i \neq j$.

From the assumptions above, the observation given the factor loadings and the factors is normally distributed, $(y_i|\beta, f_i, k) \sim N(\beta f_i, \Sigma)$. The matrices of the factors are also normally

distributed, $(f_i|k) \sim N(0, I_k)$. The joint distribution of each f_i and y_i $(i = 1 \dots T)$ is

$$\begin{aligned} p(f_i, y_i|\beta, \Sigma, k) \propto & \exp(-1/2(f_i - \hat{f}_i)'(I_k + \beta'\Sigma^{-1}\beta)(f_i - \hat{f}_i)) \\ & \exp((-1/2(y_i - \beta f_i)'(\Sigma + \beta\beta')^{-1}(y_i - \beta f_i)) \end{aligned} \tag{2.1.4}$$

where $\hat{f}_i = (I_k + \beta'\Sigma^{-1}\beta)^{-1}\beta'\Sigma^{-1}y_i$, given β, Σ. In this method, the marginal density of the observations is

$$\begin{aligned} p(y_i|\beta, R, \Sigma, k) &= \int p(f_i, y_i|\beta, \Sigma, k)df_i \\ &= (2\pi)^{n/2}|\beta I_k\beta' + \Sigma|^{-1/2} \exp(-1/2(y_i - \beta\hat{f}_i)'(\Sigma + \beta I_k\beta')^{-1}(y_i - \beta\hat{f}_i)) \end{aligned} \tag{2.1.5}$$

For simplicity, we assume the orthogonal factor model, i.e. the factor matrix F is orthogonal and define $\Omega = \beta I_k\beta' + \Sigma$. The likelihood function L is

$$\begin{aligned} L &= p(y_1 \dots y_T|\beta, \Sigma, k) \\ &= (2\pi)^{nT/2}|\Omega|^{-T/2} \exp(-1/2 \sum_{i=1}^{T}(y_i - \beta f_i)'|\Omega|^{-1}(y_i - \beta f_i)) \\ &= (2\pi)^{nT/2}|\Omega|^{-T/2} \exp(-1/2 * \mathrm{tr}(S|\Omega|^{-1}) \end{aligned} \tag{2.1.6}$$

where $S = 1/T \sum_{i=1}^{T}(y_i - \beta f_i)(y_i - \beta f_i)'$.

For the purpose of avoiding indeterminacy, we assume that $\beta'\Sigma^{-1}\beta$ is a diagonal matrix. Then the log likelihood becomes $\log L = nT/2 * \log(2\pi) - T/2 * \log|\Omega| - 1/2 * \mathrm{tr}(S\Omega^{-1})$. The log likelihood estimating equations can then be computed by appealing to the Sherman-Morrison-Woodbury formula (Sherman and Morrison, 1949, 1950; Press et al., 1992).

$$(S - \hat{\Omega})\hat{\Omega}^{-1}\hat{\beta} = 0. \tag{2.1.7}$$

$$\begin{aligned} (S - (\hat{\beta}I_k\hat{\beta}' + \hat{\Sigma}))(\hat{\beta}I_k\hat{\beta}' + \hat{\Sigma})^{-1}\hat{\beta} = 0 \\ \hat{\Sigma} = \mathrm{diag}(S - \hat{\beta}\hat{\beta}') \end{aligned} \tag{2.1.8}$$

Equations (2.1.7) and (2.1.8) for β and Σ yield unique maximum likelihood estimators. Once the factor loadings $\hat{\beta}$ and the error covariance matrix $\hat{\Sigma}$ are computed, we can estimate

the factors $\hat{f}_i$ with

$$\hat{f}_i = (I_k + \hat{\beta}'\hat{\Sigma}^{-1}\hat{\beta})^{-1}\hat{\beta}'\hat{\Sigma}^{-1}y_i \tag{2.1.9}$$

It is well-known that maximum likelihood makes several key assumptions, which should hold true for maximum likelihood factor analysis as well, and those assumptions are: 1. large sample size; 2. continuous measurement of variable (Jöreskog, 1994); 3. independent normal distribution of the n-vectors errors, i.e. $\epsilon_i \sim N(0, \Sigma)$. Although ML estimation is known to be robust to moderate violations, e.g. smaller sample size or non-normal errors, if the following severe violations occur, maximum likelihood factor analysis is not recommended (Harrington, 2008). The critical violations are

1. underestimation of the standard errors, which increases Type I error, especially for leptokurtic data (Hoogland and Boomsma, 1998);

2. the χ^2 tests of overall fit are poorly performed owing to non-normality problem, and there are other underestimated fit problems (the comparative fit index (CFI), Tucker-Lewis index (TLI), etc.) ;

3. the non-normality effects are worse together with smaller sample sizes or missing data (Brown, 2006).

2.1.4 EM Maximum Likelihood Factor Analysis

The Expectation Maximization algorithm for the maximum likelihood method (EM, Rubin and Thayer (1982)) is an iterative method applied to estate the model parameters in factor models (Ghahramani and Hinton, 1996). EM maximum likelihood factor analysis and maximum likelihood factor analysis have the same assumptions (Rubin and Thayer, 1982). The EM algorithm is an alternate way of solving maximum likelihood estimators in a sequence of two shift steps: Expectation(E) step and Maximization(M) step. In the E-step, the log likelihood expectation for the factor F conditional on the observed data Y is calculated. In the M-step, the expected log likelihood which is found in the E-step has been maximized.

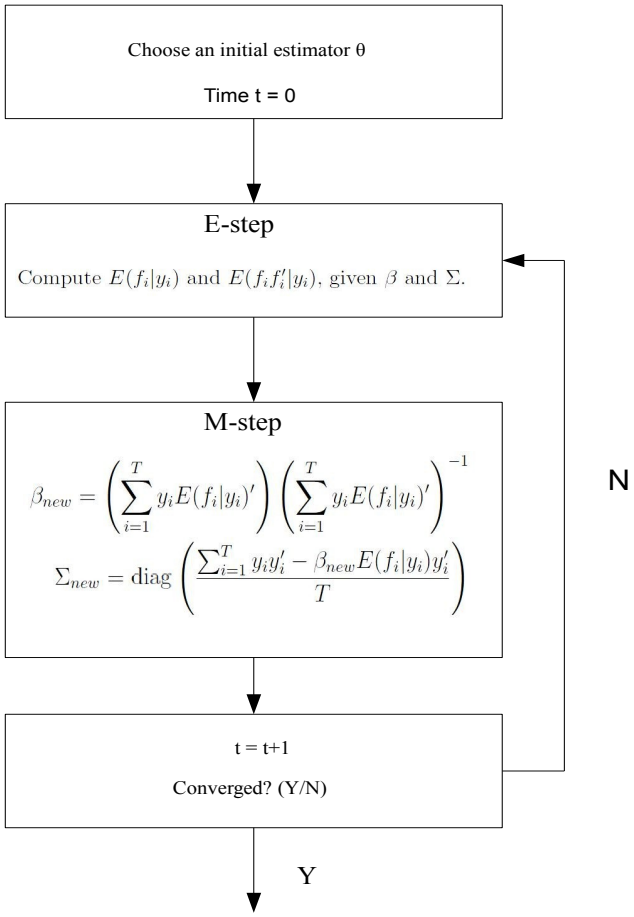

Figure 2.1: An overview of EM algorithm

The conditional mean of the factors is $E(f_i|y_i, \Sigma, \beta, k)$, and the conditional covariance matrix of the factors is $\text{var}(f_i|y_i, \Sigma, \beta, k)$, where

$$\begin{aligned} E(f_i|y_i, \Sigma, \beta, k) &= (I_k^{-1} + \beta'\Sigma^{-1}\beta)^{-1}\beta'\Sigma^{-1}y_i \\ E(f_i f_i'|y_i, \Sigma, \beta, k) &= \text{var}(f_i|y_i, \Sigma, \beta, k) + E(f_i|y_i, \Sigma, \beta, k)E(f_i|y_i, \Sigma, \beta, k)' \end{aligned} \tag{2.1.10}$$

The expected log likelihood which is found in the E-step has been maximized, then we get

$$\begin{aligned} \hat{\beta} &= (F'F)^{-1}F'Y, \\ \hat{\Sigma} &= \text{diag}\left(\frac{(Y - F\beta')'(Y - F\beta')}{T}\right). \end{aligned} \tag{2.1.11}$$

Starting from initial values, we iteratively apply the E-step and M-step with Equations (2.1.10) and (2.1.11) to obtain maximum likelihood estimates. In practice, for large data sets, the EM maximum likelihood algorithm converges slowly. Figure 2.1 shows a schematic of the EM algorithm for this model.

2.2 Markov chain Monte Carlo (MCMC) Method

In this thesis, we use the Markov chain Monte Carlo (MCMC) simulation method to fit the Bayesian factor model (Metropolis and Ulam, 1949; Metropolis et al., 1953). Markov chain Monte Carlo is one of most common computing techniques, which has been very widely used in statistics, computer science, biology, and so on. MCMC is based on probability distribution samples, randomly generating a Markov chain with desired characteristics. Since it is impossible to sample from some particular densities directly, MCMC methods are proposed to simulate Markov chains with these stationary densities (Gilks et al., 1996).

Monte Carlo integration computes numerical integrals using random number generation. Suppose we want to compute an integral $\int_a^b h(x)dx = \int_a^b f(x)p(x)dx$, where $p(x)$ is the probability function of $f(x)$. Suppose that the sample size is n, random variables $x_1, ..., x_n$ are drawn from the density $p(x)$ then we the Monte Carlo estimate of the desired integral as

$$\begin{aligned} \int_a^b h(x) &= \int_a^b f(x)p(x) \\ &\simeq \frac{1}{n}\sum_{i=1}^{n} f(x_i). \end{aligned} \tag{2.2.1}$$

A Markov chain is a sequence of stochastic transitions from one state to another where the next state is independent of all past states, except the current state, i.e.,

$$P(X_{t+1} = s_j | X_0 = s_h, \ldots, X_t = s_i) = P(X_{t+1} = s_j | X_t = s_i) \tag{2.2.2}$$

$P(i,j) = P(X_{t+1} = s_j | X_t = s_i)$ indicates the transition probability in a single step from state s_i to s_j. P is defined as the probability transition matrix, which contains all the transition probabilities $P(i,j)$. $\pi_i(t) = P(X_t = s_i)$ denotes the probability that at time t the chain is in state s_i and $\pi(t)$ denotes the vector of chain probabilities at time t. Then

$$\begin{aligned} \pi_j(t+1) &= P(X_{t+1} = s_j) \\ &= \sum_i P(X_{t+1} = s_j | X_t = s_i) P(X_t = s_i) \\ &= \sum_i P(i,j) \pi_i(t) \end{aligned} \tag{2.2.3}$$

which can be described by the Chapman-Kolmogorov equation as,

$$\pi(t+1) = \pi(t)P = \pi(t-1)P^2 = \pi(0)P^{t+1}. \tag{2.2.4}$$

Let $\pi(\cdot)$ denote a unique stationary distribution of the Markov chain, which means the probability for a given state is independent of the initial distribution of the chain. The distribution of $\pi(\cdot)$ is stationary if,

$$\pi(\cdot) = \pi(\cdot)P. \tag{2.2.5}$$

MCMC can be thought of as drawing random variables from the target distribution $\pi(\cdot)$, then computing the approximate average expectation.

In Bayesian statistics, the Gibbs sampler and the Metropolis-Hastings algorithm (M-H algorithm) are two widely used MCMC algorithms to generate Markov chains converging to a target density $\pi(x)$. The Metropolis-Hastings algorithm was developed by Metropolis et al. (1953), and extended to the general case by Hastings (1970). The M-H algorithm is used for simulating a sequence of random samples, converging to $\pi(x)$, from an equilibrium

proposal distribution $q(x|x^*)$ on which transition matrix P is based. Gibbs sampling was further developed by Geman and Geman (1984) as a special case of the M-H algorithm. Gibbs sampling generates a Markov chain which converges to $\pi(x)$, by sampling from the full conditionals: $X_i \sim \pi(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$.

2.3 Convergence of MCMC Methods

One important and hard problem for MCMC methods is how to decide when to stop the algorithm with some confidence that convergence has been reached. There are several methods that test the convergence of the Markov chain; see Cowles and Carlin (1996) for a more detailed review on this topic. We discuss some of them here, and they will be used in the next chapter for fMRI analysis.

2.3.1 Gelman and Rubin

The method of Gelman and Rubin (Gelman and Rubin, 1992) assumes that m chains can be generated independently. After running $2N$ iterations of the m chains, the sample means and variances of each chain, based on the last N iterations, are compared. Let $\hat{V}$ be the variance estimator that involves mN draws between and within chains, θ_i^t be the value of the t-th draw from chain i. The variance estimator $\hat{V}$ is defined as,

$$\hat{V} = \frac{N-1}{N} W + \left(1 + \frac{1}{m}\right) B, \tag{2.3.1}$$

where the average of the m within chain variances is $W = \frac{1}{m}\sum_{i=1}^{m} s_i^2$, and the within sequence variances is $s_i^2 = \frac{1}{N-1}\sum_{t=N+1}^{2N} \left(\theta_i^t - \bar{\theta}_i\right)^2$. The variance between the means of the m chains is,

$$B = \frac{1}{m-1} \sum_{i=1}^{m} \left(\bar{\theta}_i - \bar{\theta}_\cdot\right)^2, \tag{2.3.2}$$

where $\bar{\theta}_i = \frac{1}{N}\sum_{t=N+1}^{2N} \theta_i^t$, $\bar{\theta}_\cdot = \frac{1}{m}\sum_{t=1}^{m} \bar{\theta}_i$. The convergence ratio is,

$$\hat{R}_c = \frac{d+3}{d+1}\frac{\hat{V}}{W}, \tag{2.3.3}$$

where d is the degrees of freedom, $d \approx 2\hat{V}^2/\text{var}(\hat{V})$ (Brooks and Gelman, 1998). The convergence ratio $\hat{R}_c$ is called the potential scale reduction factor (Brooks and Gelman, 1998). If $\hat{R}_c$ is much larger than 1, it means that the chains should continue to be run. On the other hand, if $\hat{R}_c$ is close to 1, each of the m chains of N iterations is considered to be converged.

2.3.2 Raftery and Lewis

The Raftery-Lewis test (Raftery and Lewis, 1992, 1996) is designed for evaluating accuracy of the estimated simulation percentiles by computing the minimal number of samples, N_{min}, that are needed to reach the desired accuracy. Suppose one is interested in the posterior sample value of θ_q given by $P(\theta \leq \theta_q|y) = q$ and let $\hat{\theta}_q$ denote the estimator with $P(\theta \leq \hat{\theta}_q) = \hat{P}_q$. For some true cumulative probability q, $P(\theta \leq \pm\hat{\theta}_q)$ lies within precision $\pm r$ of the true value with probability s, i.e. $P(\hat{P}_q \in (q-r, q+r)) = s$.

The Raftery-Lewis method can find the length of the burn-in period, M, and the number of iterations needed, N, by estimating $P(\hat{P}_q \in (q-r, q+r)) = s$ accurately. We run the Gibbs sampler for an initial M iterations that we discard. Starting from the $M+1$ iteration, we run the sampler for a further N iterations of which we store every p^{th} iteration, meaning we store $M+p$, $M+2p$, $\ldots$, $M+jp$, s.t. $jp \leq N$. Define $Z_t = I(\theta^t \leq \hat{\theta}_q)$ for all t. The sequence Z_t is not a Markov chain, but the Raftery-Lewis method constructs a $Z_t^{(p)}$ that acts as an approximately Markov model when p is sufficiently large, where $Z_t^{(p)} = Z_{1+(t-1)p}$.

The parameter number p is determined by comparing the first and second-order Markov chain model. p with the smallest value for the first order Markov chain model is preferred. The Bayesian information criterion(BIC) is used to choose p. The p with the smallest BIC value is preferred. The BIC is based on the likelihood ratio test (G_p^2), which is expressed as,

$$\begin{aligned} BIC &= G_p^2 - 2\log(n_p - 2) \\ &= \left(2\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{l=0}^{1} w_{ijl}\log\frac{w_{ijl}}{\hat{w}_{ijl}}\right)^2 - 2\log(n_p - 2), \end{aligned} \tag{2.3.4}$$

where w_{ijl} is the transition probability for t, $t-1$ and $t-2$, Z_t is treated as a Markov chain with transition matrix,

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}, \tag{2.3.5}$$

where α is the probability of changing from state 0 to state 1 and β is the probability of changing from state 1 to state 0. The equilibrium distribution exists, $\pi = (\pi_0, \pi_1) = (\alpha+\beta)^{-1}(\beta, \alpha)$, where $\pi_0 = P(\theta \leq \theta_q|y)$ and $\pi_1 = 1-\pi_0$. The length of the burn-in period, $M = mk$, can be estimated with the following

$$|P(Z_m = i|Z_0 = j) - P(Z_\infty = i)| \leq \epsilon, \tag{2.3.6}$$

where $i, j = 0, 1$, Z_∞ is the stationary distribution of the chain. Equation 2.3.6 holds true with the assumption $1-\alpha-\beta > 0$,

$$m = \frac{\log\left(\frac{(\alpha+\beta)\epsilon}{\max(\alpha,\beta)}\right)}{\log(1-\alpha-\beta)}.$$

If the number of iterations, $N = nk$ and n, are reasonably large, the estimator of $P(\theta \leq \theta_q)$ is defined as $\bar{Z}_n^{(g)}$ $\bar{Z}_n^{(g)} = \frac{1}{n}\sum_{t=1}^{n} Z_t^{(g)}$ is approximately normal $N(q, \frac{1}{n}\frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3})$ (Raftery and Lewis, 1996).

$P(\bar{Z}_n^{(g)} \in (q-r, q+r)) = s$ will be satisfied if

$$n = \frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3}\left(\frac{\Phi^{-1}\left(\frac{1}{2}(s+1)\right)}{r}\right)^2,$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution.

Assuming the samples Z_t are independent, $\alpha = 1-\beta = \pi_1 = 1-q$, in which case $M = 0$, $g = 1$,

$$N_{\min} = \frac{q(1-q)}{r^2}\left(\Phi^{-1}\left(\frac{1}{2}(s+1)\right)\right)^2. \tag{2.3.7}$$

N_{min} is the smallest number of iterations needed to obtain convergence of the chain.

2.3.3 Geweke

The method of Geweke (Geweke, 1992) is based on comparing the two subsequences of the Markov chain in order to detect failure of convergence. Suppose the values of θ^t are computed after each iteration. After the burn-in period, the subsequences of the Markov chain are taken out, with $\theta_1^t : t = 1, \ldots, n_1$ and $\theta_2^t : t = n_a, \ldots, n$, where $1 < n_1 < n_a < n$. Let $\bar{\theta}_1$ and $\bar{\theta}_2$ be the means of the first and second subsequences respectively, i.e.,

$$\bar{\theta}_1 = \frac{1}{n_1} \sum_{t=1}^{n_1} \theta^t \qquad \text{and} \qquad \bar{\theta}_2 = \frac{1}{n - n_a + 1} \sum_{t=n_a}^{n} \theta^t.$$

When the Markov chain has converged, the location values of two subsequences of the chain should be approximately equal, i.e. $\bar{\theta}_1 \approx \bar{\theta}_2$. Let $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ be the estimators of the variance of θ based on the two chains. When the Markov chain is stationary, Geweke's statistic Z_n will have an asymptotically standard normal distribution, therefore we have,

$$Z_n = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{\hat{\Sigma}_1}{n_1} + \frac{\hat{\Sigma}_2}{n - n_a + 1}}} \xrightarrow{n \to \infty} N(0, 1). \tag{2.3.8}$$

The test of the null hypothesis of Geweke's statistic Z_n is a two-sided test. Geweke (1992) suggests using $n_1 = n/10$ and $n_a = 1 + n/2$. This method requires only one single chain and can be applied with any MCMC method.

2.3.4 Yu and Mykland

The graphical method of Yu and Mykland (1998) monitors convergence via a CUSUM path plot (Brooks, 1998), which is based upon the sampler output. After discarding the burn-in period M, the CUSUM plot is constructed in the following six steps using the output $\theta_1, \ldots, \theta_n$.

1. Calculate the mean based on the retained iterates

$$\hat{\mu} = \frac{1}{n - M} \sum_{j=M+1}^{n} \theta_j$$

2. Calculate the observed CUSUM or partial sum

$$\hat{S}_j = \sum_{j=M+1}^{n} (\theta_j - \hat{\mu}) \qquad \text{for } j = M+1, \ldots, n$$

3. Plot $\hat{S}_j$ against j, $j = M+1, \ldots, n$, and connect the successive line segments. In this method, the plot will always begin and end at around 0.

4. Brooks (1998) modified the approach by deriving a statistic to assess convergence, for all $j = M+1, \ldots, n-1$,

$$d_j = \begin{cases} 1 & \text{if } S_{j-1} > S_j \text{ and } S_j < S_{j+1} \\ 1 & \text{if } S_{j-1} < S_j \text{ and } S_j > S_{j+1} \\ 0 & \text{else} \end{cases}$$

5. $D_{M,n}$ is denoted as the index,

$$D_{M,n} = \frac{1}{n-M} \sum_{j=M+1}^{n-1} d_j$$

which could be treated as a Binomial distribution with mean $1/2$ and variance $1/4(n-M)$.

6. Calculate the confidence interval

$$\frac{1}{2} \pm Z_{-\alpha/2} \sqrt{\frac{1}{4(n-M)}}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ upper percentile of the normal distribution.

If $n - n_0$ is large, $D_{n_0,n}$ approximately normal. The situation that $D_{n_0,n}$ belongs to the interval, is not a sufficient but necessary condition for convergence of the chain.

The convergence checks are summarized in Table 2.1.

Table 2.1: Convergence diagnostic tests for MCMC

Method	Quantitative or Graphical	Single or Multiple chain	Test	Applicability
Gelman and Rubin	Quantitative	Multiple	One-sided test based on a variance ratio test statistic	Any MCMC
Raftery and Lewis	Quantitative	Single	Fewer than the necessary chain sample, then rejection	Any MCMC
Geweke	Quantitative	Single	Two-sided test based on a Z-score statistic	Any MCMC
Yu and Mykland with Brooks modification	Graphical and Quantitative	Single	Two sided test based on a Z-score statistic	Any MCMC

2.4 Image Segmentation with Graph Cut

Image segmentation plays a critical role in many medical imaging tasks, which is done by facilitating or automating the delineation of anatomical structures. The segmentation of different structures from two dimensional and three dimensional images is an important first step that you might consider when you analyze medical data. To facilitate the analysis of fMRI data with factor models, we propose the use of a robust and reliable segmentation method that can partition the image into disjoint regions, such that each region can be homogeneous with respect to some properties, e.g. pixel intensity or textures.

Since fMRI data are usually massive, their analysis can be computationally intensive and time-consuming. To deal with this issue, we propose to segment the image into different regions using the normalized cut (Wu and Leahy, 1993), which enables us to reduce computation time and at the same time keep most of the image information. If we consider an image as a graph (Figure 2.2), and each voxel is a node or vertex of the graph (green dots in Figure 2.2), the voxel similarity between neighboring voxels can be conceived as the edges connecting the graph nodes. Each edge can be assigned a weight, which typically is defined

as a function of the intensity difference between voxels, e.g. $e^{\frac{-||I(i)-I(j)||}{\sigma^2}}$, where $I(i) - I(j)$ is the intensity difference, σ is a constant). In Figure 2.2, the yellow line denotes a graph cut, which cuts across the graph edges shown as red lines in the figure, and the sum of the weights corresponding to those red edges are the value of the cut. There are possibly many such cuts, with each corresponding to one candidate segmentation. We can therefore convert the fMRI image partition problem into a graph cut problem by optimizing a certain energy function that is defined based on the graph cut.

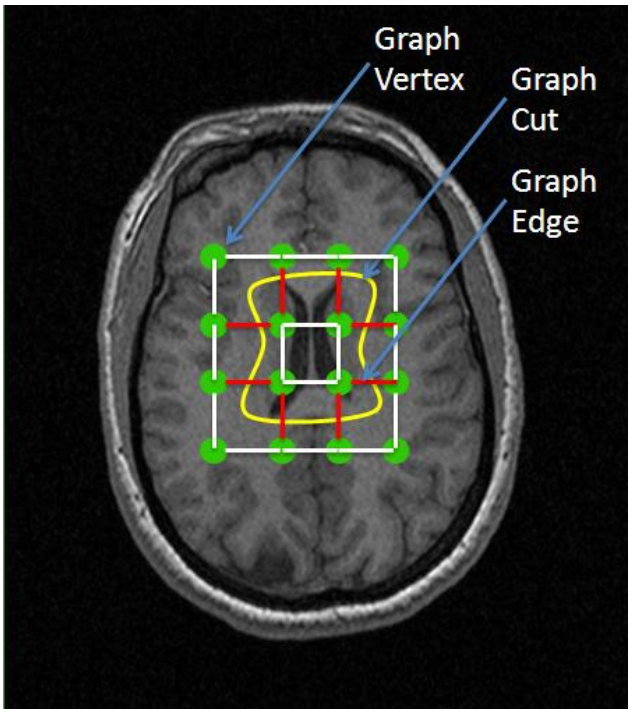

Figure 2.2: Illustration of the graph cuts for fMRI image segmentation problem. The green dots are the image pixels representing the graph nodes (or vertices); the lines (white and red lines) connecting the dots represent graph edges, where each edge is assigned a weight; the yellow line denotes a graph cut, which cuts across the graph edges represented as the red lines, and the sum of the weights corresponding to the red edges is the value of the cut. There are possibly many such cuts; we can convert the fMRI image partition problem into a graph cut problem by optimizing a certain energy function that is defined based on the graph cut.

Next, we will formally define graph cut, and give a brief overview of various graph cut algorithms that have been applied for image segmentation.

2.4.1 Graph Cut

A weighted graph, $G = (V, E)$, has a vertices and b edges, where $V = v_1, \ldots, v_a$ and $E = e_1, \ldots, e_b$. There is a weight w_i associated with each e_i, typically defined as a function of the intensity difference between voxels in the context of image segmentation.

If A, B is a partition of V (such that $A \bigcup B = V$ and $A \bigcap B = \emptyset$), then the cut is the set of all edges of G between the two subsets A and B (dotted red edges in Figure 2.3), which is denoted as $cut(A, B)$. Any edge crossing the cut, $(u, v) \in E$ with $u \in A$ and $v \in B$, is a cut edge (dotted red edges in Figure 2.3). The definition of the size (or value) of the cut is the sum of the weights ($w(u, v)$) of the cut edges ($e = (u, v)$), i.e.

$$\text{cut}(A, B) = \sum_{\substack{u \in A \\ v \in B}} w(u, v) \tag{2.4.1}$$

Figure 2.3 shows two subsets A (blue dots) and B (grey dots) with all the weights, and the cut is defined to be the sum of the weights corresponding to the red edges. For fMRI data, we can think of the image as a graph, where the vertices are the voxels, and edges connecting the graph are represented with voxel similarity, therefore we can use graph cut to partition the fMRI image into disjoint regions.

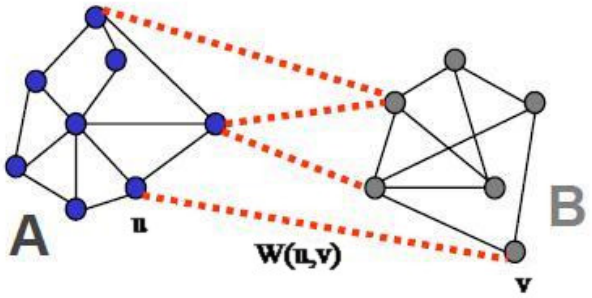

Figure 2.3: Graph cut, with subsets A and B; the red lines are the weights of the edges.

Next, we will introduce some popular energy functions that have been applied for the image segmentation problem, with the focus on the normalized cut algorithm.

2.4.2 Minimum Cut (MinCut)

The minimum cut (MinCut) of a graph is the cut that partitions G into disjoint segments such that the sum of the weights related with edges between the different segments is minimized. However, there is a flaw associated with this type of algorithm, as it favors cutting small sets

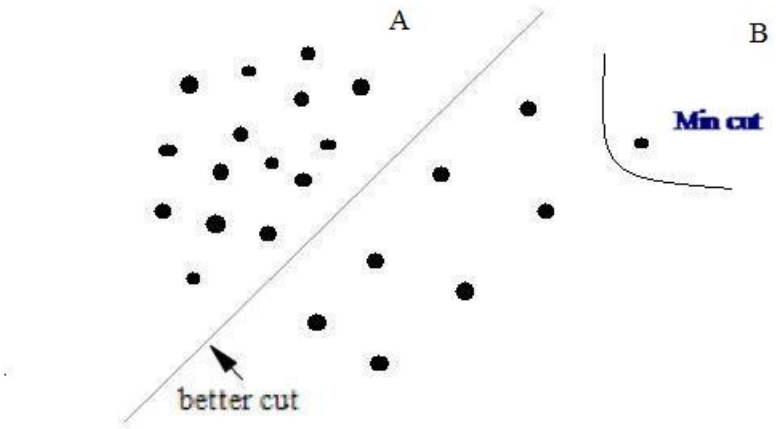

Figure 2.4: MinCut does not choose the better cut, rather prefers the isolated node

of isolated nodes in the graph (Wu and Leahy, 1993) as demonstrated in Figure 2.4, where we are trying to separate subsets A and B. It is obvious that the cut that we are looking for (annotated as "better cut") has a larger cut value than the MinCut, which prefers isolated nodes since they will yield minimal value. This motivates people to look for other alternatives that avoid such unbalanced cuts.

2.4.3 Normalized Cut

To avoid the bias that divides small sets of nodes as in the MinCut procedure, Shi and Malik (2000) give a new measure of disjunction between two subsets. The measure named the normalized cut (Ncut), uses the total edge connection from the region to all the nodes as the cut cost:

$$\text{Ncut}(A,B) = \frac{\text{cut}(A,B)}{\text{assoc}(A,V)} + \frac{\text{cut}(A,B)}{\text{assoc}(B,V)} \tag{2.4.2}$$

where $\text{assoc}(A,V) = \sum_{u\in A, t\in V} w(u,t)$, $\text{assoc}(B,V) = \sum_{v\in B, t\in V} w(v,t)$, is the total connection from nodes in A or B to all nodes in the graph, which can be viewed as "volume" or "size" of the segments. From this definition, a cut with isolated points will no longer have a small Ncut value, since Ncut penalizes unbalanced segments by normalizing the cut by the size of segments. For example, in Figure 2.4, when the cut is unbalanced, we observe that $cut(A,B)$ remains small, and $cut(A,B)$ will be roughly the same as the "volume" of B, i.e. $\text{assoc}(B,V)$. Once $cut(A,B)$ is normalized by segment size $\text{assoc}(B,V)$, the value of normalized cut will be close to 1 therefore it won't be a minimum as in the MinCut case. We could potentially avoid the unbalanced cuts that are encountered in MinCut by the normalization scheme.

Next, we will turn our focus to the normalized cut formulation and how we solve it by posing it as an eigenvalue problem. Let W be an $R \times R$ ($R = |V|$) matrix that includes all edge weights and $W(i,j) = w_{ij}$ is the weight of the edge between voxel i and j, where $W(i,j) = \exp\frac{-||I(i)-I(j)||}{\sigma^2}$, and $I(i)$ and $I(j)$ are image voxel intensities at i and j respectively. Let x be a vector indicating the region that each voxel belongs to, where $x_i = 1$ means region A, $x_i = -1$ means region B. d is a vector of the sum of weight of edges leaving each voxel, d_i is the sum of all edges leaving voxel i ($d_i = \sum_j W(i,j)$). Let D be an $R \times R$ diagonal matrix with d on its diagonal. The $\text{Ncut}(A,B)$ can be rewritten as (Shi and Malik, 2000),

$$\begin{aligned}\text{Ncut}(A,B) &= \frac{\text{cut}(A,B)}{\text{assoc}(A,V)} + \frac{\text{cut}(A,B)}{\text{assoc}(B,V)} \\ &= \frac{\sum_{x_i>0,\, x_j<0} -w_{ij}x_ix_j}{\sum_{x_i>0} d_i} + \frac{\sum_{x_i<0,\, x_j>0} -w_{ij}x_ix_j}{\sum_{x_i<0} d_i}.\end{aligned} \tag{2.4.3}$$

$b = \sum_{x_i>0} d_i / \sum_{x_i<0} d_i$ is set as a constant that provides the ratio of the "proportion" of regions, and let $y = (1+x) - b(1-x)$ where x was 1 or -1 to indicate its region. Therefore y

is a linear transformation of x which encodes the same information, and y is 2 or $-2b$. Then we have,

$$\min_x \text{Ncut}(x) = \min_y \frac{y^T(D-W)y}{y^T D y}, \tag{2.4.4}$$

where $y(i) \in 1, b$ and $y^T D1 = 0$.

The problem in this form can be converted to a Rayleigh-Ritz ratio problem. A Rayleigh-Ritz ratio is expressed as $R(A,x) = \frac{x^*Ax}{x^*x}$, where x is nonzero vector, x^* is the conjugate transpose of x, A is self-joint matrix, i.e. $A = A^*$ (Horn and Johnson, 1985). Then we solve Ncut by posing it as an eigenvalue problem. Let $z = D^{\frac{1}{2}}y$, $\text{Ncut}(x) = \frac{z^T D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}z}{z^T z}$, which is minimized by z that are the smallest eigenvectors, satisfying

$$D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}z = \lambda z, \tag{2.4.5}$$

where $(D-W)$ is positive semidefinite as the Laplacian matrix. Therefore to minimize the Rayleigh quotient, the smallest nonzero eigenvalues should be found and their eigenvectors. With zero eigenvalue, The denominator of the ratio is 0 with zero eigenvalue. $z_0 = D^{-\frac{1}{2}}1$, is the smallest eigenvector of above equation. Let z_1 be the perpendicular eigenvector to z_0.

Since $z = D^{\frac{1}{2}}y$, let $y_0 = 1$ is the eigenvector with 0 eigenvalue, and y_1 is the eigenvector with the smallest nonzero eigenvalue. Then we have $z_1^t z_0 = y_1^t D1 = 0$. The graph is divided by the eigenvector with the smallest nonzero eigenvalue.

2.5 Mutual Information

In fMRI data, there are usually several different groups and multiple subjects within each group. How to compare the results of different groups and subjects is extremely important for our analysis. In recent years mutual information (MI) has received much attention since it has been found to be an effective and robust similarity measure for comparing images (Maes et al., 1997; Viola and Wells, 1997).

MI is a measure of information that one random variable contains about another random variable. For image comparison, MI provides a statistical way to analyze the similarity

between content of two images. The mutual information between two images U and V, $I(U,V)$ can be defined as

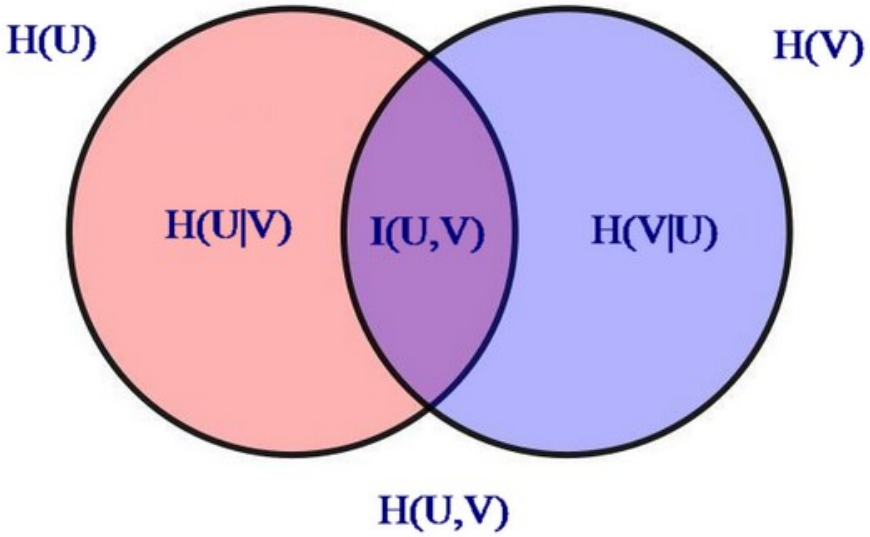

Figure 2.5: The mutual information of random variables U and V. The left circle is $H(U)$, the right is $H(V)$. $H(U,V) = H(U|V) + H(V|U) + I(U,V)$, where $H(U|V)$ is the red part, $H(V|U)$ is the blue part and I(U,V) is the purple part.

$$\begin{aligned} I(U,V) &= H(U) - H(U|V) \\ &= H(V) - H(V|U) \\ &= H(U) + H(V) - H(U,V) \end{aligned} \tag{2.5.1}$$

where $H(U)$ and $H(V)$ are the marginal entropies, $H(U) = -\sum_u p_U(u) \log p_U(u)$ and $H(V) = -\sum_v p_V(v) \log p_V(v)$. The joint entropy of U and V, $H(U,V)$, is

$$H(U,V) = -\sum_{u,v} p_{U,V}(u,v) \log p_{U,V}(u,v) \tag{2.5.2}$$

where $p_U(u)$, $p_V(v)$ and $p_{U,V}(u,v)$ are the probability distribution functions of the variables. For images, $p_U(u)$, $p_V(v)$ and $p_{U,V}(u,v)$ are the mean images' grey value distribution and joint distribution.

From Equation (2.5.1) if U and V are independent, the mutual information $I(U, V)$ will be 0. If image U and image V have equal entropies, then $I(U, V)$ will just be $H(U)$ or $H(V)$. Since $H(U, V) \leq H(U)$, $I(U, V)$ is nonnegative.

In the next chapter, we will discuss the technical details of using the Bayesian factor model and the methods that are introduced in this chapter to analyze functional MRI data. Examples of real fMRI data will also be given.

Chapter 3

Bayesian Factor Analysis for fMRI Data

From Chapter 1, we know that a key problem for fMRI analysis is understanding which parts of the brain are activated during various cognitive tasks. There is also evidence that different fMRI activation patterns are associated with schizophrenia, Alzheimer's disease, and mild traumatic brain injury (MTBI), which suggests that diagnosis of these conditions may eventually be possible on the basis of brain activations.

In this chapter, we will first discuss the technical details of the Bayesian factor model, with a focus on the difference with the regular factor model. Then, we customize the Bayesian factor model for fMRI analysis. We present experimental results on the application of the Bayesian factor model to real fMRI data when eye movement tasks are performed while subjects are in the scanner. Three sets of experimental results are presented to comprehensively evaluate the performance of the Bayesian factor model for fMRI data analysis. The first set of experiments are performed on single subjects, with the focus on the effectiveness of the algorithm; the second experiment set is on a group of fMRI data, which is performed together with the normalized cut segmentation algorithm to improve the efficiency of the model; the third set is on all fMRI data that are collected; here we are primarily interested in the comparison of brain activation between subjects from two different groups, controls against schizophrenia subjects.

3.1 Bayesian Factor Analysis

One of the major differences between the classical factor model and Bayesian factor model is in the estimation of the free parameters. In classical factor analysis, pursuant to many sub-

jective criteria, the loading matrix is produced with an orthogonal rotation. The parameters can't be uniquely determined from the likelihood alone. Bayesian factor analysis is not in this case. The incorporation of prior knowledge in Bayesian factor model eliminates the ambiguity of rotation and obviate the need for model constraints in regular factor model (Press and Shigemasu, 1997). Bayesian approaches solve these problems by treating the parameters θ (β, Σ, F) as unknown random variables, they are then estimated by averaging over the ensemble of models, which can be formulated as:

$$\begin{aligned} P(Y) &= \int P(Y|\theta)P(\theta)d\theta \\ &= \int P(Y|\beta,\Sigma,F)P(\beta|\Sigma)P(F)P(\Sigma)d\Sigma dF d\beta, \end{aligned} \tag{3.1.1}$$

where $P(Y)$ is the evidence for data set $Y = y_1, \ldots, y_T$, β is the matrix of standardized coefficients, known as the factor loading matrix, which represents correlations between the original variables and the factors. F is the factor score matrix which explains the observed structures and Σ is the covariance matrix of the error ϵ. All information about the parameters is acquired by the posterior distribution rather than by maximum likelihood estimation.

The Bayesian approach to factor analysis is widely used for estimations of multivariate dependence and co-dependence. One of the reasons is that the Bayesian approach improves the original model by more informative priors. More specifically, the Bayesian approach has the minimal loss of all conditional models based on the observed data (Howson and Urbach, 2005). Furthermore it is able to generate factor structures as exemplified by the work in time series modeling, for instance, dynamic factor analysis, which is a multivariate time-series analysis technique for estimating underlying common patterns in a set of time series (Lopes and West, 2004). Among all Bayesian approaches used in factor model analysis, the iterative Markov Chain Monte Carlo (MCMC) simulation method is the most popular one, with extensions to dynamic factor components in financial time series modeling (Aguilar and West, 2000; Wang and Wall, 2003).

3.1.1 Methods

Similar to the classical factor analysis, we start with a zero-mean n-variate normal distribution, denoted by $N(0, \Omega)$, where Ω is an $n \times n$ non-singular covariance matrix. Let $Y = (y_1, \ldots, y_i, \ldots, y_T)$ be the T vectors of random observation for $i = 1 \ldots T$. The model assumes that each n-dimensional data vector Y was generated by a k ($k < n$) dimensional factor vector F, which has an unobserved independent zero-mean unit-variance Gaussian distribution (Christensen and Amemiya, 2003), as in Equation (2.1.1) and Equation (2.1.2),

$$y_i = \beta f_i + \epsilon_i \tag{3.1.2}$$

or,

$$Y = F\beta' + E \tag{3.1.3}$$

In the model, let μ be the mean of Y; the marginal density can then be expressed as

$$p(y_t|\beta, \mu, \Sigma) = \int p(f_t)p(y_t|f_t, \beta, \mu, \Sigma)df_t \sim N(y_t|\mu, \beta\beta' + \Sigma). \tag{3.1.4}$$

From Section 2.3 , we know that the likelihood for (β, F, Σ) (Kaufman and Press, 1973) can be computed as

$$p(Y|\beta, F, \Sigma) \propto |\Sigma|^{-nT/2} e^{-1/2(Y-F\beta')'\Sigma^{-1}(Y-F\beta')}. \tag{3.1.5}$$

We will use Equation (3.1.5) in Gibbs sampling MCMC for the parameters of the factor model with k fixed.

3.1.2 Prior and Posterior Distributions for β, Σ and F

Similar to Press (2002), we will use the generalized natural conjugate prior distributions to estimate the parameters. Conjugate prior distributions are a class of prior distributions which possess the property that the posterior are in the same family as the prior.

Assume β is dependent on the covariance matrix Σ, Σ is independent of the factors F, and F is independent of β and Σ; the joint prior distribution for those parameters can then be written as

$$p(\beta, F, \Sigma|k) = p(\beta|\Sigma, k)p(\Sigma)p(F|k). \tag{3.1.6}$$

Since we assume k is fixed in our model, the prior distribution is equivalent to

$$p(\beta, \Sigma, F) \propto p(\beta|\Sigma)P(\Sigma)P(F). \tag{3.1.7}$$

Next, we will discuss how we choose the priors for the model parameters β and Σ. The steps are similar to those introduced in Fokoué (2004) and in Lopes and West (2004). For the factor loadings matrix β, we know $(Y - F\beta')'(Y - F\beta') = (Y - F\hat{\beta}')'(Y - F\hat{\beta}') + (\beta' - \hat{\beta}')F'F(\beta' - \hat{\beta}')$. Since $(Y - F\hat{\beta}')'(Y - F\hat{\beta}')$ is independent of β, we can write the distribution for β as

$$L(\beta) \propto \exp(-1/2 \operatorname{tr} \Sigma^{-1}(\beta' - \hat{\beta}')'F'F(\beta' - \hat{\beta}')). \tag{3.1.8}$$

For ease of exposition, take $\vartheta = \vec{\beta'}$, a vectorized version of the factor loadings matrix, and $\operatorname{tr} \Sigma^{-1}(\beta' - \hat{\beta}')'F'F(\beta' - \hat{\beta}') \equiv (\vartheta - \hat{\vartheta})'(\Sigma^{-1} \otimes F'F)(\vartheta - \hat{\vartheta})$, then we have

$$(\Sigma^{-1} \otimes F'F)^{-1} = \Sigma \otimes (F'F)^{-1}. \tag{3.1.9}$$

Combining Equations (3.1.8) and (3.1.9) together, the likelihood of the loading matrix can be expressed as

$$L(\beta) = L(\vartheta) \propto \exp(-1/2(\vartheta - \hat{\vartheta})'(\Sigma \otimes (F'F)^{-1})^{-1}(\vartheta - \hat{\vartheta})). \tag{3.1.10}$$

This suggests that a normal distribution is a natural conjugate prior for ϑ; it also implies a normal prior for each row β_i of the loading matrix β. The upper-diagonal elements of the positive loading matrix are independent, and can be approximated as $\beta_{ij} \sim N(0, C_0)$, where $i = 1, \ldots, k$, $i \neq j$ and $\beta_{ii} \sim N(0, C_0)1(\beta_{ii} > 0)$ by assuming C_0 is a positive constant. This normal prior restricts the diagonal elements β_{ii} to be strictly positive.

For Σ, from Equation (2.1.6) and Equation (2.1.9), we have the likelihood,

$$L(\Sigma^{-1}) \propto |\Sigma^{-1}|^{T/2} e^{-1/2 \operatorname{tr}(\Sigma^{-1} S)}, \tag{3.1.11}$$

which indicates that the conjugate prior for Σ^{-1} is the Wishart distribution. Since Σ^{-1} is diagonal, we can rewrite Equation (3.1.11) as

$$L(\Sigma^{-1}) \propto \prod_{i=1}^{n} |\sigma_i^{-2}|^{T/2} e^{-1/2 S_{ii} \sigma_i^{-2}}. \tag{3.1.12}$$

We observe that this likelihood is the product of Gamma densities. In other words, a Gamma prior is suggested for each σ_i^{-2}. A prior of common Inverse Gamma distribution is assumed for each σ_i^2. The variances of the prior is independent assumption. With υ and s^2 hyperparameters, the σ_i^2 are token as $\sigma_i^2 \sim IG(\upsilon/2, \upsilon s^2/2)$ independently. υ is the freedom hyperparameter prior degree and s^2 is each σ_i^2 prior mode.

For the factors F, from Equation (2.1.6) and Equation (2.1.9), we know

$$L(F) \propto \exp(-1/2 \operatorname{tr} F'F), \tag{3.1.13}$$

which suggests that the conjugate prior for F is the normal distribution.

In practice, there are a few cases where the number of factors k is known and/or fixed (e.g. the factors under study are the only factors of interest) as we have assumed so far. However, the value of k is very often unknown in real-life applications and hence of interest in and of itself. There are several methods that can be applied to determine the number of factors; we will briefly introduce some of the popular ones next, although we don't address the choice of k explicitly in this dissertation.

The goodness-of-fit test (D'Agostino and Stephens, 1986) is a classical likelihood ratio test, with the null hypothesis stating the covariance matrix of Y has the structure $\Omega = \beta\beta' + \Sigma$, under the normal assumption, $W = nT(\operatorname{tr}(\hat{\Omega}^{-1}) - \log|\hat{\Omega}^{-1}S| - k)$, where S is the sample covariance matrix. If $\Sigma > 0$, then W is asymptotically χ^2 distributed with $1/2[(T+k)^2 - T - k]$ degrees of freedom. If a model with a given number of factors is deemed to be a poor fit, more factors may be added until a good fit is found.

The Akaike Information Criterion (AIC) (Akaike, 1973, 1974) is a quality measure criterion of model selection. By definition, $AIC = -2\log(\text{maximum likelihood}) + 2(\text{number of parameters fitted})$. In general, the goodness of fit is better by increasing the number of free parameters to be measured, the number of free parameters is irrespective. Hence AIC is not only rewarded by goodness of fit, but also might be punished with an function that increase the number of estimate parameters. AIC discourages overfitting to a certain extent.

The Bayesian Information Criterion (BIC) (Schwarz, 1978) is a popular measure of model selection. It is defined as $BIC = -2\log(\text{maximum likelihood}) + \log n$ (number of parameters fitted). AIC has two problems: First, the classic AIC was never proven consistent (Bickel and Zhang, 1992; Zhang, 1993); here consistency means that for the model under consideration and fixed value of the parameter, as the number of observations $n \to \infty$, the choice of k will be asymptotically correct. Second, AIC tends to overestimate the true model order. The penalty term of BIC penalizes complex models more heavily than does that of AIC. So BIC tends to be more conservative, in the sense that it favors to select simpler models; BIC reduces the tendency of AIC to overfit models.

The Reversible Jump Markov Chain Monte Carlo (RJMCMC) is a type of MCMC method that allows for dimension changes in the probability distribution being simulated, i.e. the number of factors is not fixed (Green, 1995; Gamerman and Lopes, 2006), and indeed, is taken as one of the parameters to be estimated.

Finally, the Birth-and-Death MCMC first introduced by Stephens (2000) is similar to RJMCMC, but it is time continuous, and it has a limit on the types of moves permitted in order to simplify implementation (Stephens, 2000).

For simplicity reasons, we start by assuming k is fixed, and we are able to get satisfactory results. Later in this chapter, we will discuss in more detail about the problem of determining the number of factors and how the number of factors affects the overall performance.

Using Bayes' rule, the posterior distribution for the unknown parameters can be written as,

$$\begin{aligned} p(\beta, \Sigma, F|Y) \propto & p(Y|\beta, \Sigma, F)p(\beta, \Sigma, F) \\ \propto & p(Y|\beta, \Sigma, F)sp(\beta|\Sigma)p(\Sigma)P(F) \\ \propto & |\Sigma|^{-nT/2} e^{-1/2(Y-F\beta')'\Sigma^{-1}(Y-F\beta')} \\ & \exp(-1/2(\vartheta - \hat{\vartheta})'(\Sigma(F'F)^{-1})^{-1}(\vartheta - \hat{\vartheta})) \\ & \prod_{i=1}^{n} |\sigma_i^{-2}|^{T/2} e^{-1/2S_{ii}\sigma_i^{-2}} \exp(-1/2 \operatorname{tr} F'F) \end{aligned} \tag{3.1.14}$$

3.2 The Bayesian Factor Model for fMRI Data Analysis

In this section, we will apply the Bayesian factor model (BFM) presented in the last section to analyze fMRI data. First, we will introduce our choice of priors for the free parameters as well as the MCMC sampler. We then describe in detail how the fMRI data are collected for our experiments.

3.2.1 Prior for Parameters and MCMC Sample

Recall from Section 2.1 that the factor loading matrix represents the correlation between a variable and a factor. Hence the loadings range from -1 to 1. Also, as discussed in Section 2.1.2, the square of the loading denotes the percentage variance in the variable that is explained by a factor. With these interpretations in mind, we can consider appropriate choices for the prior.

Another consideration in picking the prior concerns identifiability. We say that there is "weak identifiability" in a model when some of its parameters contain little information. If there is a θ_1 such that $f(\theta_2|\theta_1, y)$ is roughly equal to $f(\theta_2|\theta_1)$ we say that θ_2 is weakly identifiable.

We use fairly non-informative priors similar to Lopes and West (2004) to avoid the identifiability problem. Specifically

$$\beta_{ij} \sim N(0, C_0) \sim N(0, 1) \tag{3.2.1}$$

$$\sigma_i^2 \sim IG(\upsilon/2, \upsilon s^2/2) \sim IG(1.1, 0.05) \tag{3.2.2}$$

where we assume $E(\sigma_i^2) = 0.5$. We have to use a truncated $N(0, 1)$ to accomplish β within $(-1, 1)$ for the lower triangular elements of β individually. These hyperparameter values are made to cover all the data.

We run MCMC simulation with the above prior distribution of β and Σ. Based on MCMC convergence of our data set, 105000 iterations were sampled. Since size of all iterations for one subject will be more than 100 Megabyte, we only keep the last 5000 to avoid the size of the file getting too big.

3.2.2 Data Acquisition

The focus of the study in question is the comparison of activation levels, in two experimental groups, between a task and a rest condition. Specifically, fMRI data are collected for both schizophrenic patients and healthy controls during an eye movement task. Those movement tasks, also known as "saccade" when eyes move in the direction of a visual cue, or anti-saccade if they move in the direction opposite to the cue, are often used in human experimental psychology, since deficits in performing these tasks can be indicative of pathologies such as schizophrenia, or lesions in specific parts of the brain (Pettigrew et al., 1990).

This particular experiment is composed of sixteen participants diagnosed with schizophrenia and fifteen healthy participants. Schizophrenia participants were diagnosed with the Structured Clinical Interview (Patient Edition) for DSM-IV and rated with Scales for the Assessment of Negative Symptoms (SANS), Scales for the Assessment of Positive Symptoms (SAPS), and Global Assessment Functioning (GAF) (Camchong et al., 2008). Brain imaging was performed at the Athens Orthopedic Clinic MRI Center with a GE Signa Horizon LX

1.5T MRI scanner, where a dual mirror box was placed 16 cm above and in front of the participant's eye. The experiment was designed to make stimuli visible to the participant and the participant's eye visible to an eye-tracking camera. Eye movements were then recorded at sampling rate of 60 Hz and displayed on a computer monitor so performance could be monitored continuously (Camchong et al., 2008).

When the imaging process starts, a three-dimensional T1−weighted structural MRI scan within each brain was acquired with spoiled gradient-recall (SPGR) protocol, and here are the imaging parameters: echo time (TE) = 2.8 msec, number of excitations = 2, matrix = 256 × 256, slice thickness of 1.5 mm, in-plane resolution of .97 × .97, sagittal acquisition, 124 contiguous slices, scan time 5 min 41 sec (Camchong et al., 2008).

Then, participants were instructed with task, and two functional runs of anti-saccade trials and ocular motor delayed response trials (see Figure 3.1) were conducted. More specifically, participants performed two blocked runs, which alternated between blocks of fixation and blocks of a single volitional saccade condition, shown in Figure 3.1(a). Task change were signaled by a change in the geometric shape around the fixation cross: a bordering diamond signaled a volitional saccade condition and a bordering square signaled the baseline fixation condition (Camchong et al., 2008). The volitional saccade blocks were either anti-saccade trials ("AS" run), which is shown in Figure 3.1(b), or ocular motor delayed response trials ("ODR" run) shown in Figure 3.1(c). The order of runs was counterbalanced across subjects. We refer to Figure 3.1 for detailed explanation of each of the stimuli. For our experiment, we are only analyzing the data from the anti-saccade runs, i.e. AS vs. fixation.

For each subject, a series of T2-weighted functional images was obtained with spoiled-gradient pulse sequence (SPGR); each slice was made up of 64×64 voxels, in-plane resolution of 3.75×3.75, TE = 40 msec, slice thickness = 4 mm, TR = 1912 msec with two interleaved resulting in an image acquisition time of 3.8 sec and 24 supratentorial contiguous slices. A total of 38 slices were collected. Each subject's brain activity was recorded at 81 different time points. Five images were acquired in each of two blocks; six images were acquired in

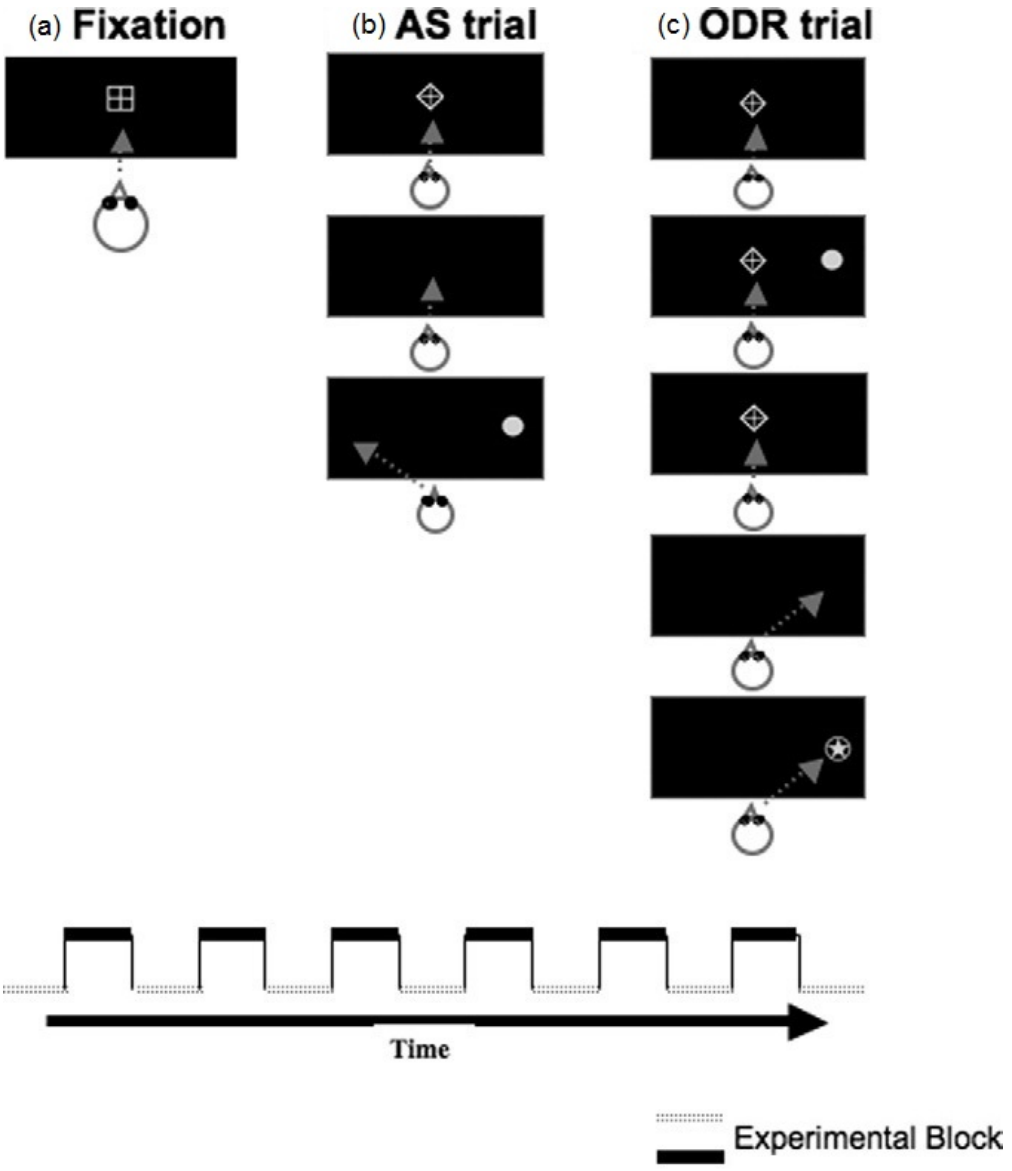

Figure 3.1: From Camchong et al. (2008), antisaccade (AS) and ocular motor delayed response (ODR) task trials. Stimuli presented during AS and ODR runs. During the AS run, participants were presented with 22.5-sec blocks of fixation (a) alternated with blocks of eight AS trials (b). During the ODR run, participants were presented with 22.5-sec blocks of fixation (a) alternated with blocks of six ODR trials (c). Gray arrows show correct eye position.

each of six blocks and seven images were acquired in five blocks for a total of 81. The visual stimulus was presented at alternating intervals (saccade or anti-saccade), which induced a certain type of eye movement, and the associated brain activity was recorded by the fMRI scanner.

To illustrate the data, one sample subject at the 35^{th} time point is shown in Figure 3.2. We only keep the voxels inside of the brain area for our analysis and delete all other irrelevant voxels. Notice that there are 38 slices in total at this particular time point for the sample subject. Slice 28 is chosen by the psychologist who thinks that it covers regions of the brain that are strongly involved in the task; therefore we only analyze this particular slice for all subjects in this dissertation, although ideally other slices close to the 28^{th} should be examined as well.

Consider the factor model in Equations (3.1.2) and (3.1.3) independently at each voxel and fit the model. Let y_{ti} be the observed signal intensity acquired at voxel i, at time t and Y be a matrix of these observations. The data span the period from 1 to 81 repetitions, $t = 1, \ldots, 81$. Observations are considered to have a multivariate normal distribution by previous assumptions. As a first step, we don't consider the correlation among neighbors of each voxel. If neighboring voxels have similar behavior, we expect them to be modeled by common factors.

Next, three sets of experimental results are presented to comprehensively evaluate the performance of the BFM for fMRI data analysis. The first set of experiments is performed on a single subject, with the focus on the effectiveness of the algorithm; the second experiment set is on a group of fMRI data, which is performed together with the normalized cut segmentation algorithm to improve the efficiency of the model; the third set is on all fMRI data that were collected for this study, and we are primarily interested in the comparison between patients from two different groups, controls against schizophrenia patients.

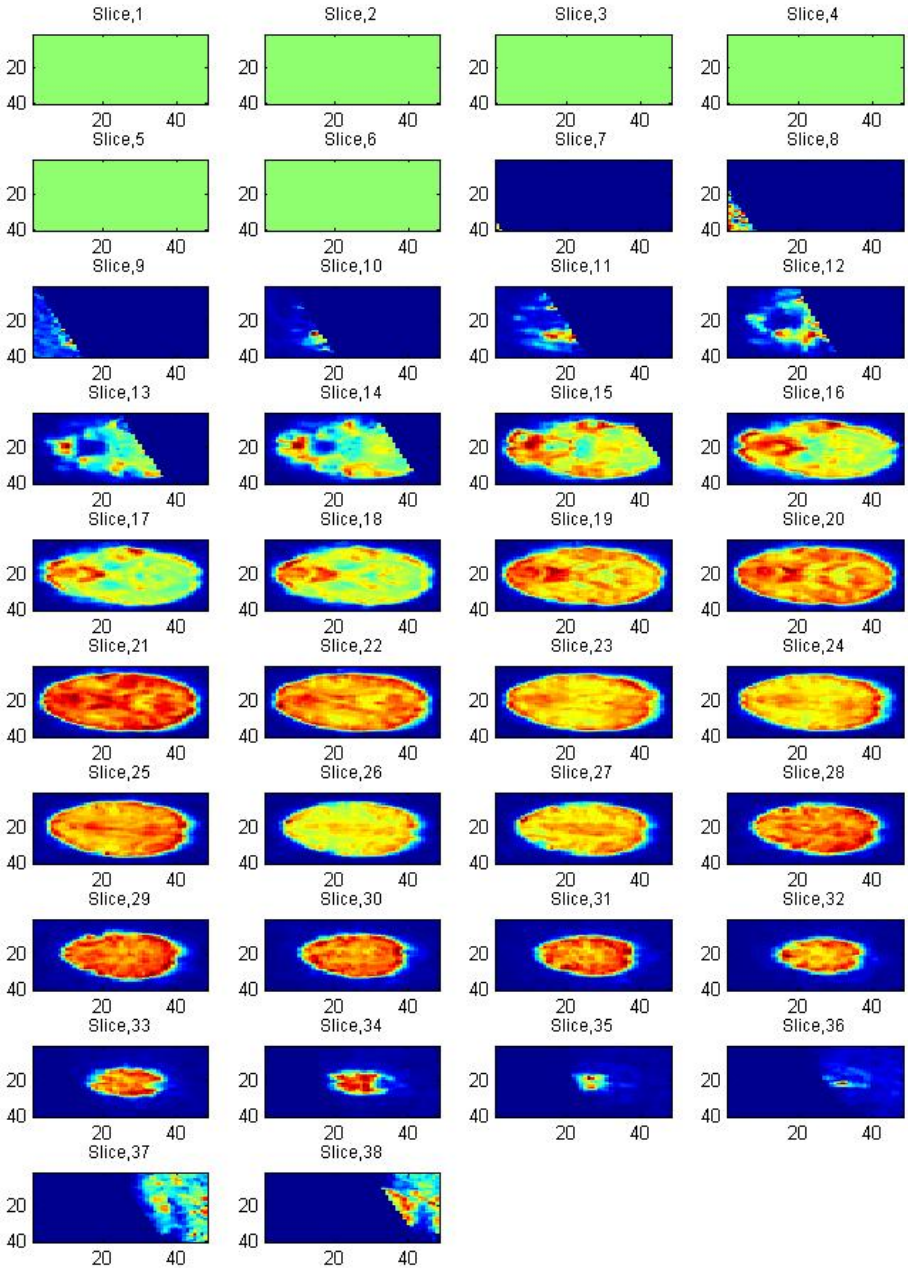

Figure 3.2: Sample fMRI data subject slice image at time ($T = 35$)

3.3 Experimental Results with Single Subject fMRI Data

In this section, we analyze data from a single subject to illustrate the effectiveness of the BFM when it is applied to fMRI analysis. The choice of the number of factors is discussed in the first part of the section, followed by MCMC convergence analysis for the subject. Finally, we present the BFM results for the same subject.

3.3.1 Sample Data Result without Image Segmentation

One randomly chosen control subject is used for this particular experiment. We assume the number of factors to be fixed to simplify the computation. As described in the previous section, the number of factors implies the cumulative percent of variance explained. We chose the smallest number of factors for which the cumulative percent of variance explained was no less than 95%, as shown in Figure 3.3. In this case, we need 56 factors to explain 95% of the variation. The total number of brain voxels in this slice for this subject is 833, so we have achieved a large reduction. With the reduced dimension, it is easier to understand the data structure and explore activation patterns.

For this subject, we summarize the results of the BFM on a single slice of fMRI data. The algorithm was written in Matlab. It takes approximately 38 hours to run on 32 8−core Power4 server with a total of 16GB of RAM for the single subject. This is too time consuming to be useful in real applications. In Section 3.4, we will try to reduce the computation time by segmenting the image with the techniques we introduced in the last chapter before we apply the BFM.

3.3.2 Result of MCMC Convergence for Single Subject fMRI

Since we use MCMC to fit the BFM, we have to diagnose convergence. In Chapter 2, we discussed several popular methods to determine whether convergence is indeed achieved. Among them, the Gelman-Rubin diagnostic is quite time consuming as it requires comparison of results from multiple chains, therefore we choose to not to use it in this dissertation.

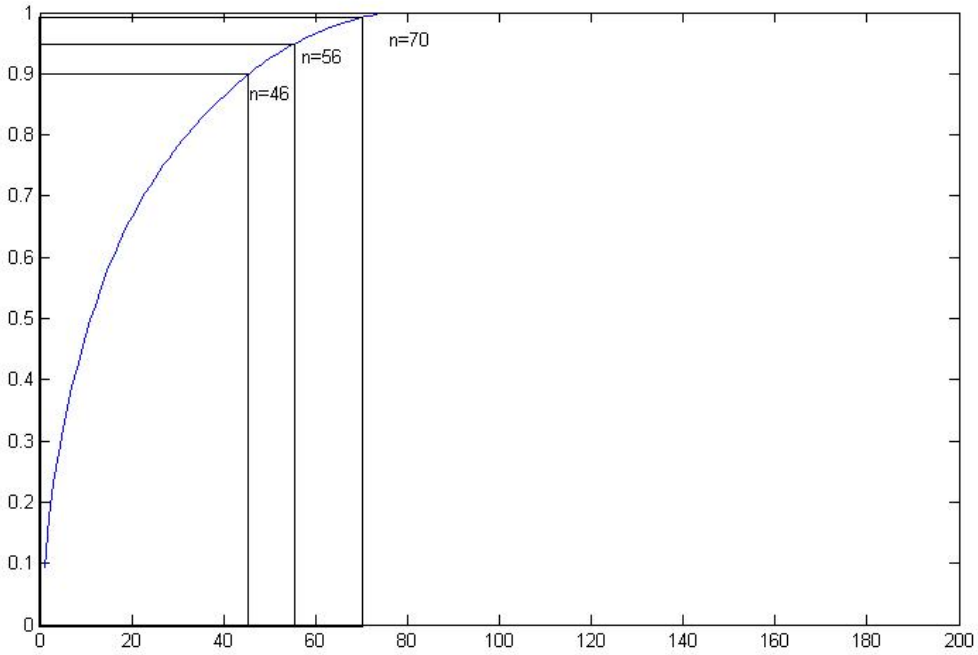

Figure 3.3: Selection of the number of factors. In traditional factor analysis, the number of factors is as above. The first 46 factors explain 90% of the variation in the data, the first 56 factors explain 95% of the variation. To explain 70% of the variation, we need to include 23 factors in the model.

Instead, we use the Raftery and Lewis, Geweke, and Brooks methods for our convergence analysis in this section. The analysis is based on the results from a chain of 105000 iterations, from which we discard the first 100000 iterations to save the physical space for the intermediate results; this won't affect the convergence analysis results, as typically the convergence is derived from the tail-end of the iterations.

Table 3.1: Raftery and Lewis convergence diagnostic test

	Burn-in	N	Nmin	I-stat		Burn-in	N	Nmin	I-stat
$\beta_{1,1}$	2	907	939	0.9659	$\beta_{16,10}$	2	932	937	0.9947
$\beta_{2,1}$	2	937	939	0.9979	$\beta_{17,10}$	2	946	937	1.0010
$\beta_{3,1}$	2	967	937	1.0298	$\beta_{18,10}$	3	927	939	0.9893
$\beta_{4,1}$	2	907	939	0.9659	$\beta_{19,10}$	2	930	937	0.9925
$\beta_{5,1}$	3	905	939	0.9638	$\beta_{20,10}$	2	942	939	1.0053
$\beta_{6,1}$	2	963	939	1.0255	$\beta_{21,10}$	2	958	937	1.0224
$\beta_{7,1}$	2	885	939	0.9424	$\beta_{22,10}$	2	943	937	1.0064
$\beta_{8,1}$	2	920	939	0.9766	$\beta_{23,10}$	2	964	937	1.0288
$\beta_{9,1}$	2	958	937	1.0202	$\beta_{24,10}$	3	932	939	0.9947
$\beta_{10,1}$	2	917	939	0.9766	$\beta_{25,10}$	3	956	937	1.0203
$\beta_{11,1}$	2	932	939	0.9925	$\beta_{26,10}$	2	912	937	0.9733
$\beta_{12,1}$	2	934	939	0.9946	$\beta_{27,10}$	3	945	937	1.0085
$\beta_{13,1}$	3	921	937	0.9808	$\beta_{28,10}$	3	894	937	0.9541
$\beta_{14,1}$	2	978	939	1.0415	$\beta_{29,10}$	2	907	939	0.9680
$\beta_{15,1}$	2	932	939	0.9925	$\beta_{30,10}$	2	958	937	1.0224
$\beta_{101,20}$	2	934	939	0.9970	$\beta_{116,30}$	3	943	939	1.0064
$\beta_{102,20}$	2	958	937	1.0224	$\beta_{117,30}$	2	912	939	0.9733
$\beta_{103,20}$	3	960	937	1.0245	$\beta_{118,30}$	2	958	937	1.0224
$\beta_{104,20}$	2	967	939	1.0320	$\beta_{119,30}$	2	932	939	0.9947
$\beta_{105,20}$	2	932	937	0.9947	$\beta_{120,30}$	2	927	937	0.9893
$\beta_{106,20}$	2	906	939	0.9669	$\beta_{121,30}$	3	943	937	1.0064
$\beta_{107,20}$	3	896	939	0.9562	$\beta_{122,30}$	3	958	939	1.0224
$\beta_{108,20}$	2	927	937	0.9893	$\beta_{123,30}$	2	912	937	0.9733
$\beta_{109,20}$	3	946	937	1.0096	$\beta_{124,30}$	2	964	937	1.0288
$\beta_{110,20}$	2	912	939	0.9733	$\beta_{125,30}$	3	927	939	0.9893
$\beta_{111,20}$	2	934	939	0.9968	$\beta_{126,30}$	2	946	937	1.0096
$\beta_{112,20}$	2	917	937	0.9787	$\beta_{127,30}$	2	907	937	0.9680
$\beta_{113,20}$	2	932	939	0.9947	$\beta_{128,30}$	2	958	939	1.0224
$\beta_{114,20}$	2	958	937	1.0224	$\beta_{129,30}$	3	964	937	1.0288
$\beta_{115,20}$	2	927	939	0.9893	$\beta_{130,30}$	2	930	939	0.9925

Table 3.1 shows the results of the Raftery and Lewis method on part of the parameter β vector after 100000 iterations of the MCMC sampling algorithm (since β is a 833×56 matrix, we can not show the complete results). From Equation (2.3.6), q is typically set to be

0.025 to provide the basis for a 95% posterior credible interval estimate. We empirically set the desired accuracy $r = 0.01$ and the required probability $s = 0.9$, as suggested in Raftery and Lewis (1996).

The results presented in Table 3.1 report that only 2 or 3 draws should be discarded after the the first 100000 iterations. The "N" column represents the total number of draws (including burn-in) recommended to achieve the desired level of accuracy. Those are roughly in the range of 890 to 1050. The "Nmin" column shows the number of draws if the data indicate an independent and identically distributed (iid) chain. The I-statistic is the ratio of the "N" and "Nmin" columns. The I-statistic measures the amount by which autocorrelation inflates the sample size.

From Figure 3.4, I-statistics of the chain are quite close to each other, all around 1. Raftery and Lewis (1992) suggest that I-statistic values above 5 indicate high correlation between coefficients, suggesting that the model needs to be reparameterized, while I-statistic values near 1 indicate good mixing. We can conclude that our chain is mixing well.

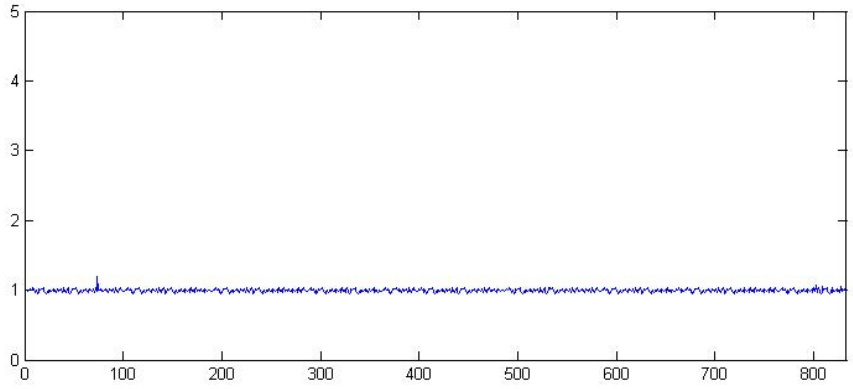

Figure 3.4: I-statistic for the β. x-axis is $\beta(., 1)$, y-ax the I-statistic values for β.is is The part values for β are shown in Table 3.1.

Geweke's method on the other hand uses two parts of one MCMC chain to test convergence. Once a simulation has converged, the means and variances of a parameter's posterior distributions from the first half and the second half of the MCMC chain will be roughly equal.

For our data, we use the first 10% and the last 50% to calculate the Geweke's Z-scores and p values as suggested by Geweke (1992).

Table 3.2: Geweke convergence diagnostic test

	z-score	$Pr > \|z\|$		z-score	$Pr > \|z\|$		z-score	$Pr > \|z\|$
$\beta_{1,1}$	0.99	0.32	$\beta_{21,1}$	-1.15	0.25	$\beta_{41,1}$	0.47	0.64
$\beta_{2,1}$	-0.65	0.52	$\beta_{22,1}$	0.14	0.89	$\beta_{42,1}$	0.4	0.69
$\beta_{3,1}$	0.01	0.99	$\beta_{23,1}$	0.5	0.62	$\beta_{43,1}$	-1.92	0.05
$\beta_{4,1}$	-1.41	0.16	$\beta_{24,1}$	1.06	0.29	$\beta_{44,1}$	0.82	0.41
$\beta_{5,1}$	-0.41	0.68	$\beta_{25,1}$	1.8	0.07	$\beta_{45,1}$	0.61	0.54
$\beta_{6,1}$	-1.87	0.06	$\beta_{26,1}$	-0.18	0.86	$\beta_{46,1}$	0.2	0.84
$\beta_{7,1}$	0.11	0.91	$\beta_{27,1}$	1.47	0.14	$\beta_{47,1}$	0	1
$\beta_{8,1}$	-1.78	0.08	$\beta_{28,1}$	-0.42	0.67	$\beta_{48,1}$	-0.95	0.34
$\beta_{9,1}$	-1.9	0.06	$\beta_{29,1}$	1.37	0.17	$\beta_{49,1}$	0.4	0.69
$\beta_{10,1}$	-0.15	0.88	$\beta_{30,1}$	-1.54	0.12	$\beta_{50,1}$	0.51	0.61
$\beta_{11,1}$	1.7	0.09	$\beta_{31,1}$	-1.11	0.27	$\beta_{51,1}$	-2.05	0.04
$\beta_{12,1}$	0.2	0.84	$\beta_{32,1}$	-0.14	0.89	$\beta_{52,1}$	1.53	0.13
$\beta_{13,1}$	-1.08	0.28	$\beta_{33,1}$	-0.35	0.73	$\beta_{53,1}$	-0.42	0.67
$\beta_{14,1}$	-0.48	0.63	$\beta_{34,1}$	-1.25	0.21	$\beta_{54,1}$	-1.13	0.26
$\beta_{15,1}$	1.26	0.21	$\beta_{35,1}$	-0.07	0.94	$\beta_{55,1}$	-0.36	0.72
$\beta_{16,1}$	-0.32	0.75	$\beta_{36,1}$	0.32	0.75	$\beta_{56,1}$	0.18	0.86
$\beta_{17,1}$	0.72	0.47	$\beta_{37,1}$	-0.24	0.81	$\beta_{57,1}$	-1	0.32
$\beta_{18,1}$	0.59	0.56	$\beta_{38,1}$	0.95	0.34	$\beta_{58,1}$	-1.18	0.24
$\beta_{19,1}$	0.93	0.35	$\beta_{39,1}$	-0.68	0.5	$\beta_{59,1}$	-0.67	0.5
$\beta_{20,1}$	0.46	0.65	$\beta_{40,1}$	0.7	0.48	$\beta_{60,1}$	0.01	0.99

The results from Geweke's method in Table 3.2 indicate that at $\alpha = 0.05$ level, we can not reject the null hypothesis, i.e. the means and variances of the two parts of the chain are equal. The first 10% and the last 50% of the MCMC chain are not significantly different, therefore the chain indeed has converged according to Geweke's method.

Finally, we use the Brooks convergence diagnostic, which assumes the index D_T has a Binomial distribution with mean $\frac{1}{2}$ and variance $\frac{1}{4(n-n_0)}$, to test the chain. The confidence interval is $\frac{1}{2} \pm Z_{-\alpha/2}\sqrt{\frac{1}{4(n-n_0)}}$ for D_T at level α; in our case, $n = 105000$, $n_0 = 100000$, $\alpha = 0.05$, and the interval is $(0.4861, 0.5139)$. Brooks (1998) suggests that if the data fall in the bound, we may consider the data are from a stationary chain.

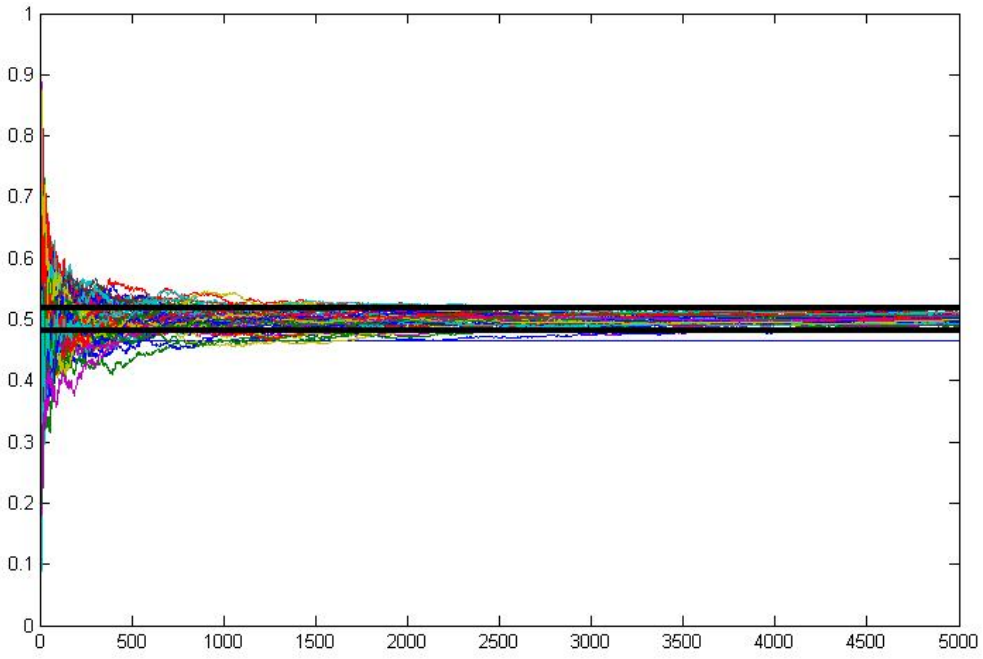

Figure 3.5: Brooks confidence interval for β. The solid dark lines are the limits of the 95% confidence interval. The colored lines come from the different beta parameters.

Figure 3.5, which plots the Brooks confidence interval for all the β, shows at the end of chain, most data fall into the confidence interval. Again, this suggests convergence of chain.

3.3.3 Model Result

In this experiment, we will use the prior distribution for β,Σ and F as discussed in Section 3.1.2. The results (β,Σ and F) of the oblimin rotation factor model (Equation (2.1.2)) are used as the initial values for the BFM. The conditional posterior distribution for f_t can be expressed as independent normal distributions, where

$$f_t \sim N((I_k + \beta'\Sigma^{-1}\beta)^{-1}\beta'\Sigma^{-1}y_t, (I_k + \beta'\Sigma^{-1}\beta)^{-1})$$

For β and Σ, we use Equations (3.1.10) and (3.1.11) as our conditional posterior distributions.

After 100000 iterations, we saved MCMC sampling results for the last 5000 iterations. Simulation results on those 5000 images can not be shown at the same time due to the size. We took the posterior mean to summarize the posterior distribution from the BFM; this is shown in Figure 3.6(b) , along with the original data slice in Figure 3.6(a). The mutual information between the posterior mean and the original image is 0.9674. The closer the MI value is to 1, the more similar the image pairs are, so this result indicates that the posterior distribution is very similar to the original Slice 28 image. Therefore, we consider the BFM simulation to be effective.

We also use communality to evaluate the effectiveness of the BFM. The communality, denoted by h_i^2, is the proportion of a variable's variance explained by a factor structure. One of the goals of factor analysis is to determine factors that can explain as much of the variables' communalities as possible (Walker and Maddan, 2008). The communality h_i^2 for the i^{th} variable explained by all factors can be estimated using

$$h_i^2 = \sum_{j=1}^{k} \hat{\beta}_{ij}^2. \tag{3.3.1}$$

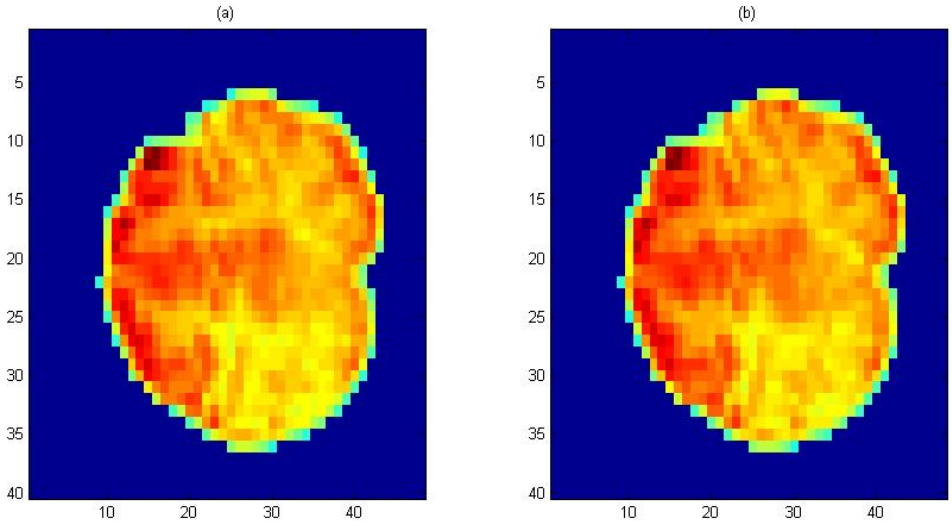

Figure 3.6: (a)Sample original data at T = 35; (b)BFM average posterior result

We compute the communality for each variable, which in our case is all the voxels within the brain slice under investigation. Figure 3.7(a) shows the histogram of the communalities. Notice that most communalities are greater than 0.90, with an average communality over all voxels in the slice of 0.96, indicating good explainability of the factors in our model. To better understand the contribution of individual factors to the communalities of each voxel, in Figure 3.7(b), we also plot the percentage of the voxel's variance that is explained by the single biggest factor with the greatest explanatory power, i.e. $\max_j \hat{\beta}_{ij}^2$ in the brain image. We observe that the majority of the $\max_j \hat{\beta}_{ij}^2$ are in the range of 0.3 to 0.6. The average of $\max_j \hat{\beta}_{ij}^2$ across all voxels is 0.39. From Figure 3.7(b), $\max_j \hat{\beta}_{ij}^2$ values for the voxels on the right side of the slice are higher. This means that the voxels in the posterior part of the brain are more explainable by a single factor. Some of the brain voxels with a higher $\max_j \hat{\beta}_{ij}^2$ value might be eye-movement task relevant since the posterior result is involved with the eye movement task (time).

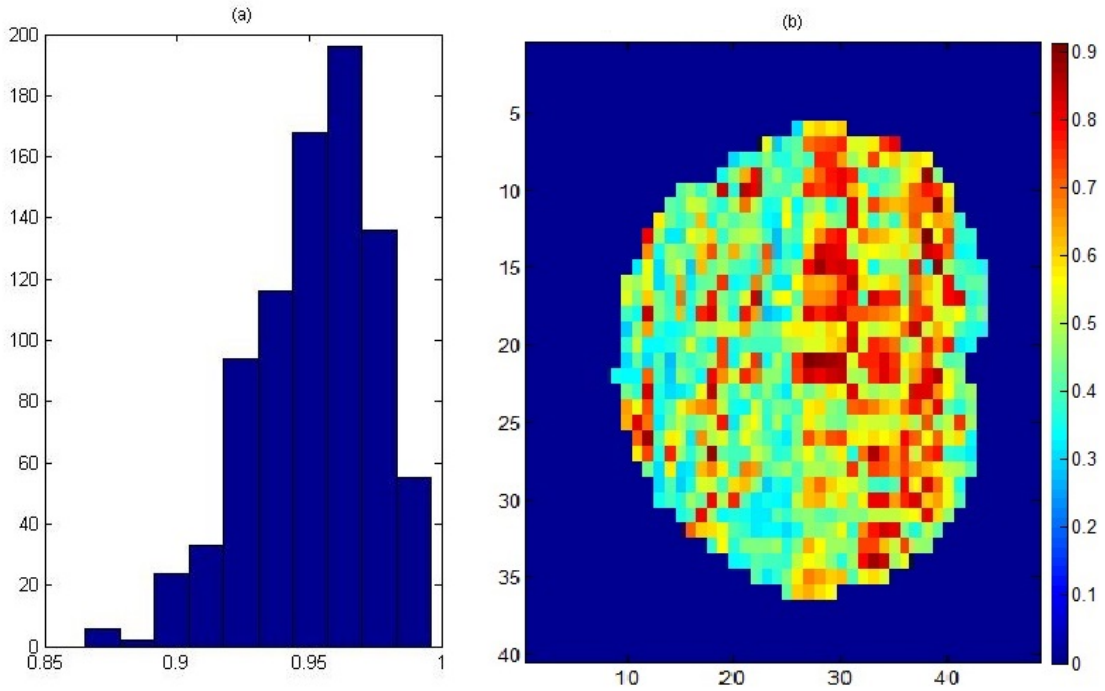

Figure 3.7: (a) Histogram of the communalities for all voxels in the brain region; (b) Proportion of voxel's variance that is explained by the single biggest factor within the brain.

To summarize the findings based on the results with a single fMRI subject: 1. MCMC simulation of the BFM is effective, but not efficient. Posterior distributions of the BFM based on MCMC simulation have good fits compared to the original data; 2. most voxels have a communality of more than 0.9, which indicates good explanatory power of the BFM; 3. computation time of simulations is unacceptably long (38 hours for one slice). We are going to explore other practical approaches in Section 3.4 to reduce the computation time, for which we will use normalized cut (Ncut) to pre-segment the images prior to application of the factor model..

3.4 Experimental Results with Multiple fMRI Subjects

In this section, we present the experimental results using normalized cut with a group of subjects. Instead of dealing with individual fMRI voxels, we group voxels into segments using Ncut and perform the BFM on the segment level, thereby reducing the computational complexity. Further, Ncut is able to preserve inter-voxel connectivity and homogeneity within each segment and therefore should not negatively affect the performance of the BFM.

We randomly choose five controls and five schizophrenic subjects to evaluate the performance of the normalized cut. First, we use AIC and BIC to determine the optimal number of segments as in Tepper et al. (2011); we then discuss in detail how to select the right number of factors for the BFM. Finally, we present the results of the BFM for the ten subjects that we randomly selected.

3.4.1 Result of Normalized Cut

An important issue for all image segmentation algorithms is how to select the optimal number of segments. If there are too many segments, it won't save computation time; if there are too few segments, the variation of the voxels within each segment may be too big to use for the factor analysis. In this dissertation, we propose to use AIC and BIC to determine the number of segments. The number of segments should not be as small as the number

of factors and also not be as large as the number of voxels. We consider that the range of $(150, 300)$ probably might be the number of segments that we are looking for. Figure 3.8 and Figure 3.9 show AIC and BIC values for the different numbers of segments for the five control and five schizophrenia subjects, respectively. We choose the minimal values of BIC for the number of segments, giving $211, 287, 295, 291, 194$ for the control subjects (their numbers with minimal AIC value are $246, 287, 272, 291, 268$) and $261, 297, 298, 123, 283$ for the schizophrenia subjects (their numbers with minimal AIC value are $261, 293, 282, 182, 283$).

Figure 3.10 and Figure 3.11 show the results of the Ncut segmentation algorithm for the control subjects and schizophrenic subjects respectively. From the Figures, we can see that Ncut tends to group neighboring voxels with similar intensities together. Different brain regions with different voxel intensities are clearly delineated into different groups by the normalized cut algorithm (Equation (2.4.2)). Further, the Ncut algorithm with smallest BIC value is able to preserve major components of the image. Within each segment, we assign the average voxel intensity as the segment value for further analysis. The areas outlined in black in Figures 3.10 and 3.11 are regions that obviously differ from their neighboring voxels. Ncut assigns these to distinct segments.

3.4.2 Number of factors

The other important issue in the BFM is how to select the optimum number of factors. This is a tricky problem, as on one hand, we need a smaller dimensional subspace to accurately reflect the observed data, and on the other hand, we have to consider time-saving and the penalty for over-parameterizing.

Similar to the single fMRI experiment of Section 3.3, we choose to use the smallest number of factors such that the cumulative percentage of variance explained is no less than 95%. Figures 3.12 and 3.13 demonstrate this process for control subjects and schizophrenic

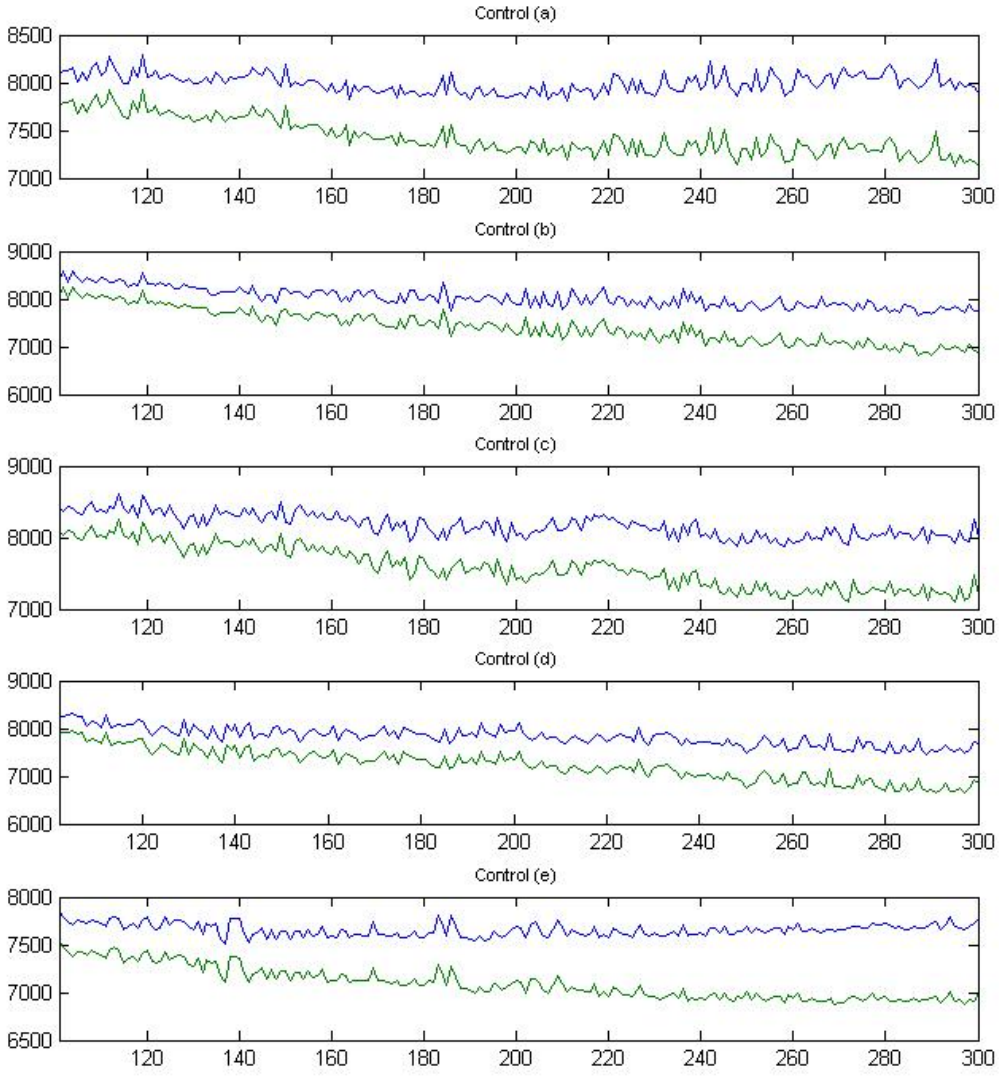

Figure 3.8: AIC and BIC values for the control subjects' segmentation. The upper line is BIC, and the lower line is AIC. y-axis is the value of AIC or BIC, x-axis is the number of segments for Ncut.

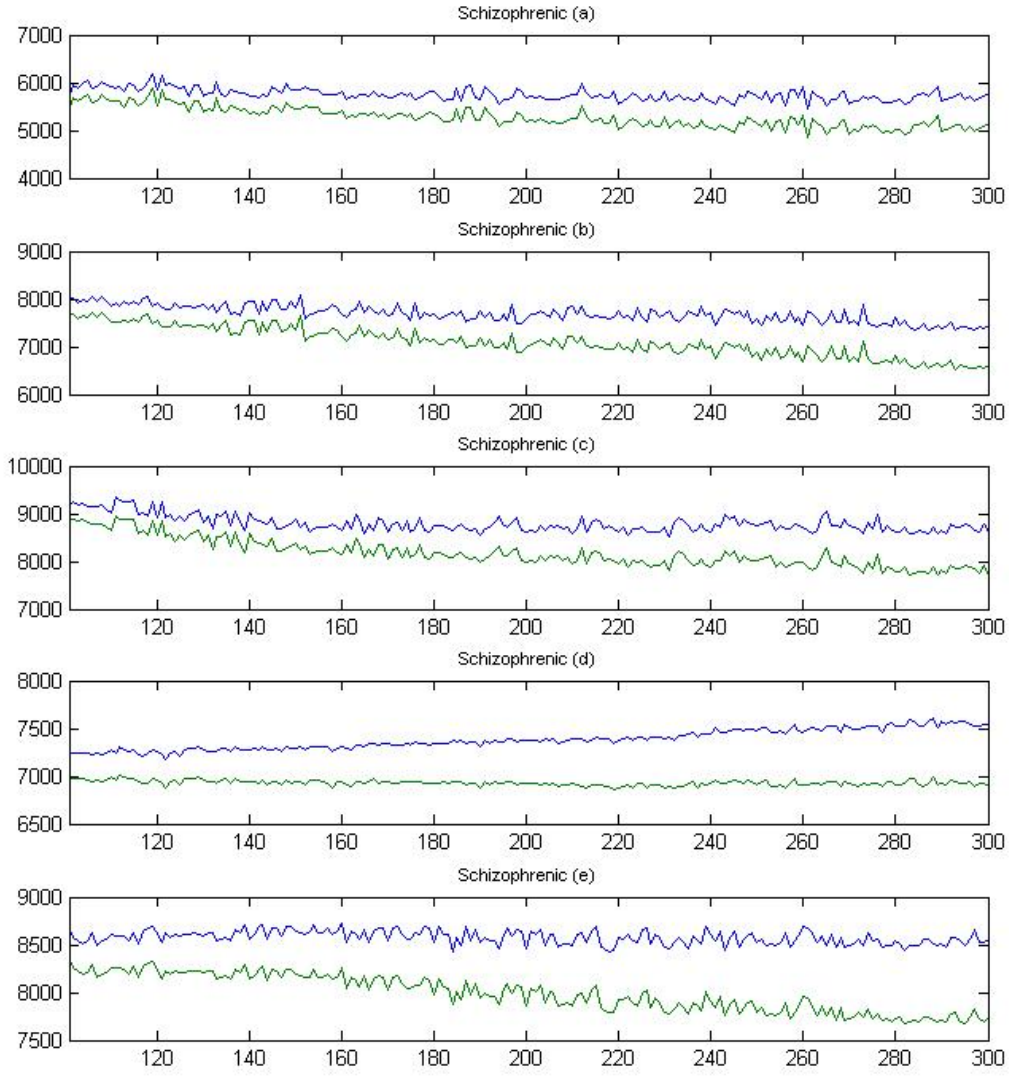

Figure 3.9: AIC and BIC values for the schizophrenic subjects' segmentation. The upper line is BIC, and the lower line is AIC. y-axis is the value of AIC or BIC, x-axis is the number of segments for Ncut.

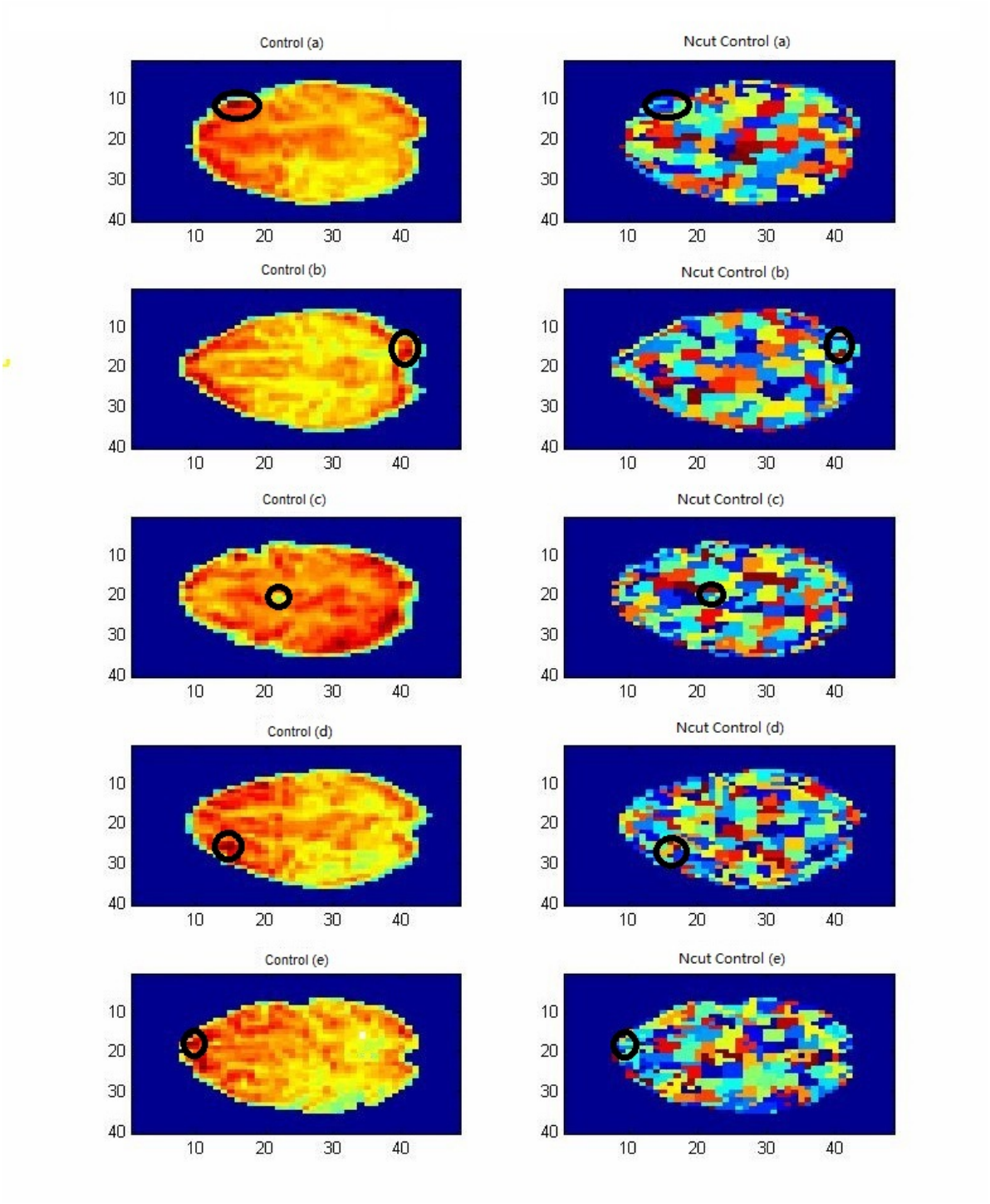

Figure 3.10: Subplots (a)-(e) show the original images and Ncut images of the smallest BIC value in sample control data.

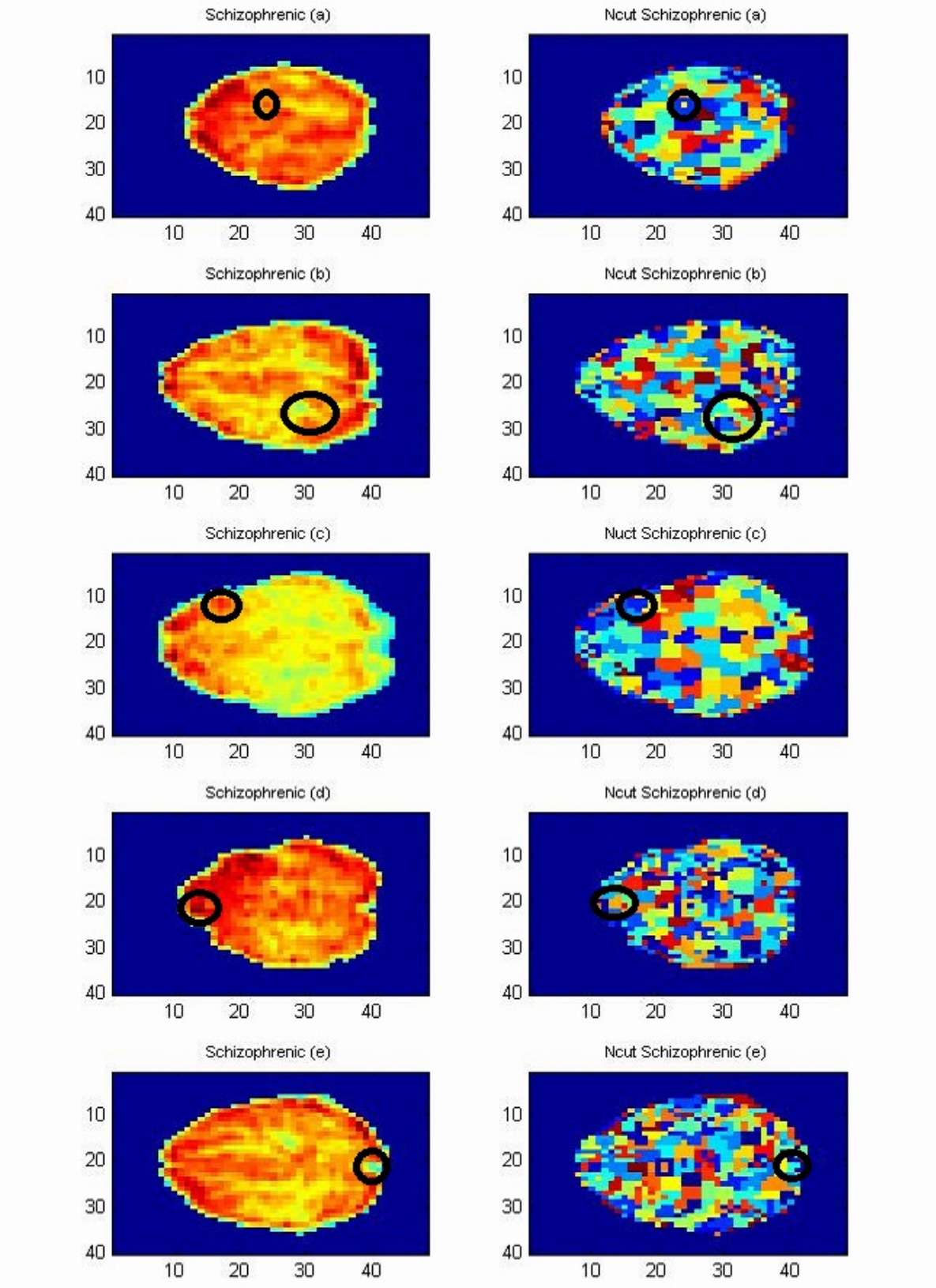

Figure 3.11: Subplots (a)-(e) show the original images and Ncut images of the smallest BIC value in sample schizophrenic data.

subjects respectively.

Figure 3.12 shows the numbers of factors and the cumulative percent of variance that they explain in the five control subjects. The number of factors that we select for those five subjects ranges from 26 to 42 using the 95% variation rule. Further, the first factors for those subjects explain $31 - 52\%$ of the total variation, and the first seven factors together in different control subjects explain more than 70% of the variation. From about the tenth factor onward, the growth trend weakens.

Figure 3.13 shows the numbers of factors and the cumulative percent of variance that they explain in the five schizophrenic subjects. The number of factors that we select for those five subjects ranges from 22 to 33. The first factors for those subjects explain $26 - 47\%$ of the total variation, and the first seven factors together explain around 80% of the variance. From the eighth factor onward, the growth trend weakens.

3.4.3 Result of the Bayesian factor model

We summarize the results of the BFM on slice 28 of the 10 subjects that are examined in this section. The algorithm was written in Matlab. It takes approximately 4 hours to run on 32 8-CPU Power4 nodes with 16GB of RAM for each subject, compared to 38 hours for a single subject previously.

Table 3.3: Mutual information for control and schizophrenic subjects

Subject	Mutual Information				
	1	2	3	4	5
Control	0.9132	0.9397	0.8843	0.9078	0.8902
Schizophrenic	0.9198	0.8801	0.9131	0.9076	0.9056

Similar to the experiment setup in Section 3.3, after 100000 iterations, we keep the last 5000 iterations of the MCMC sampling results. We use the average of the simulation results to summarize the posterior draws; these are presented in Figure 3.14 and Figure 3.15. By

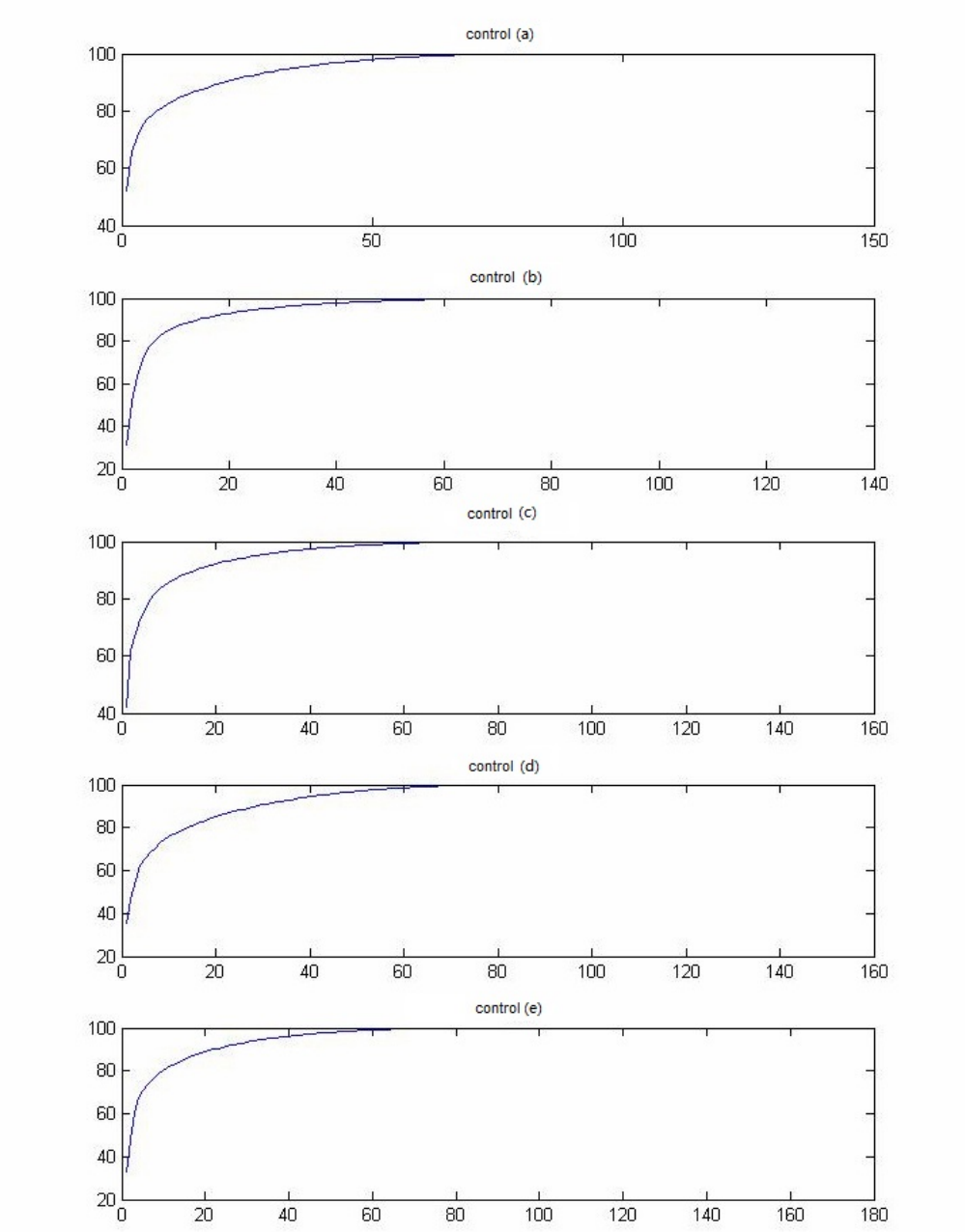

Figure 3.12: Selection of the number of factors for the five randomly chosen control subjects. x-axis is the number of factors, and y-axis is the percentage of variance explained. (a). The first 35 factors explain 95% of the variation for control subject 1; (b) the first 26 factors explain 95% of the variation in control subject 2; (c) the first 29 factors explain 95% of the variation in control subject 3; (d) the first 42 factors explain 95% of the variation in control subject 4; (e) the first 35 factors explain 95% of the variation in control subject 5.

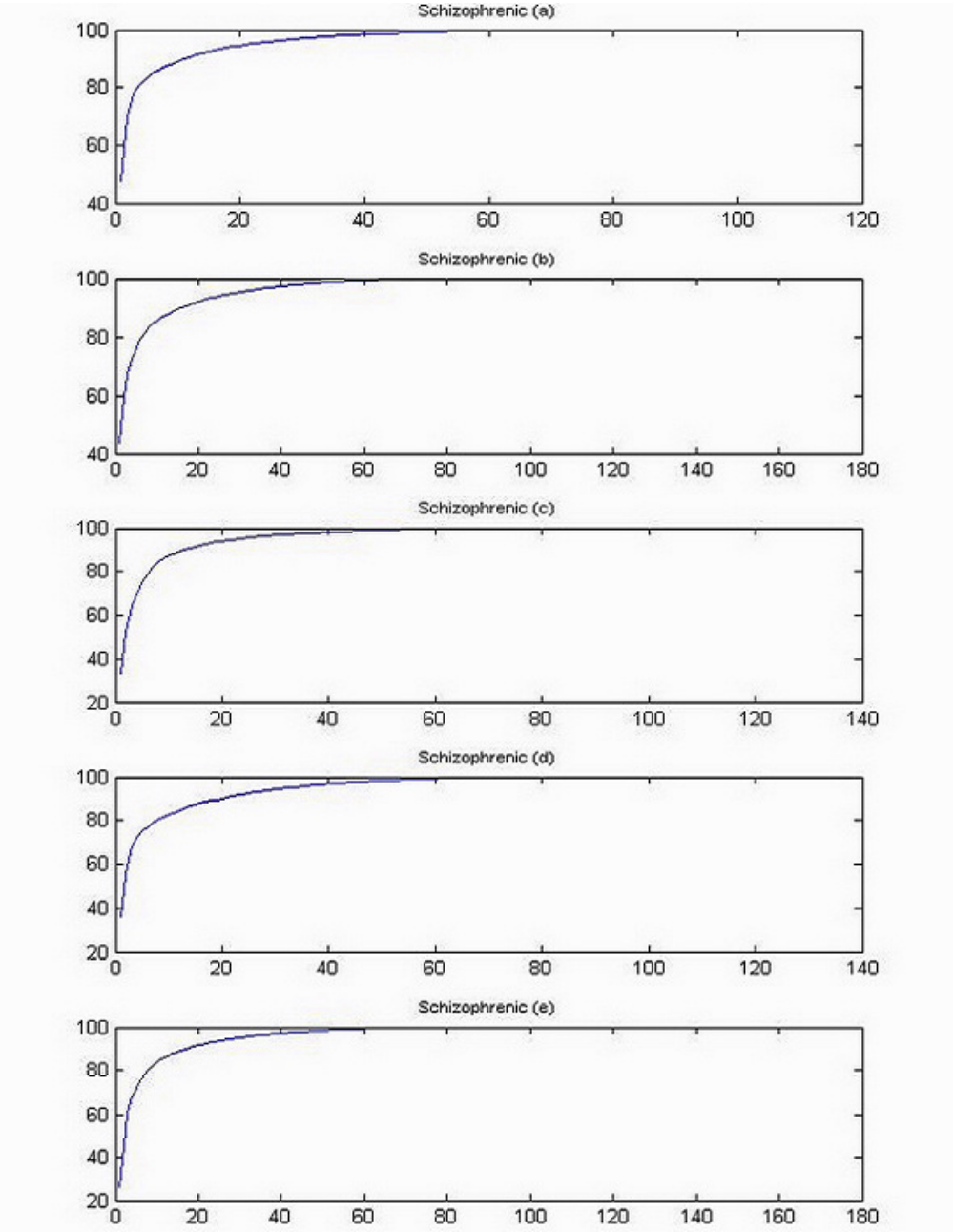

Figure 3.13: Selection of the numbers of factors for the five randomly chosen schizophrenic subjects. x-axis is the number of factors, and y-axis is the percentage of variance explained. (a) The first 22 factors explain 95% of the variation in schizophrenic subject 1; (b) the first 29 factors explain 95% of the variation in schizophrenic subject 2; (3) the first 23 factors explain 95% of the variation in Schizophrenic subject 3; (d) the first 33 factors explain 95% of the variation in schizophrenic subject 4; (e) the first 30 factors explain 95% of the variation in schizophrenic subject 5.

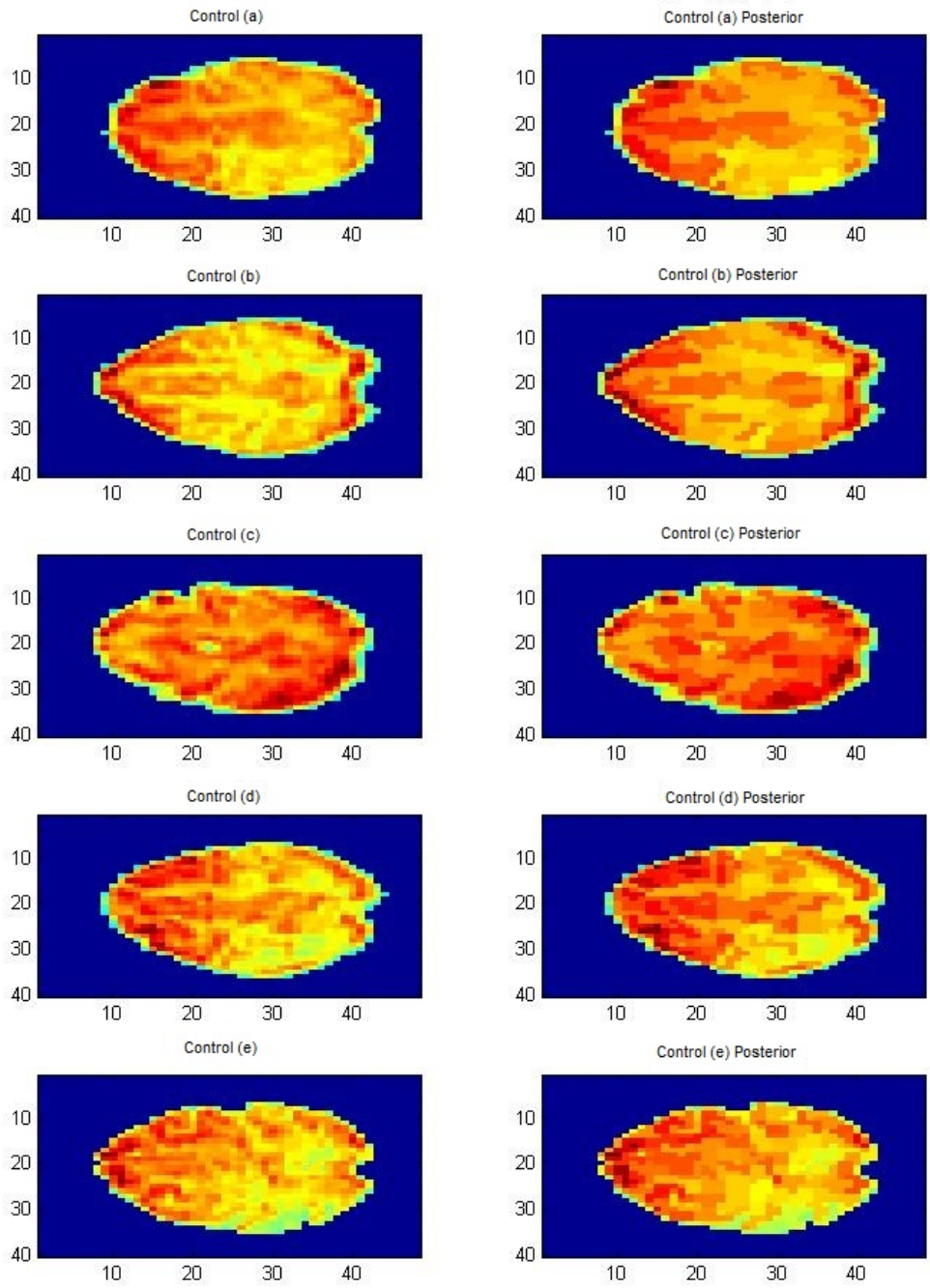

Figure 3.14: Original image and posterior distribution of slice 28 ($\hat{Y}$) in sample control subjects. From 5000 iterations of the MCMC sampling algorithm result after 100000 iterations, we display average of the posterior.

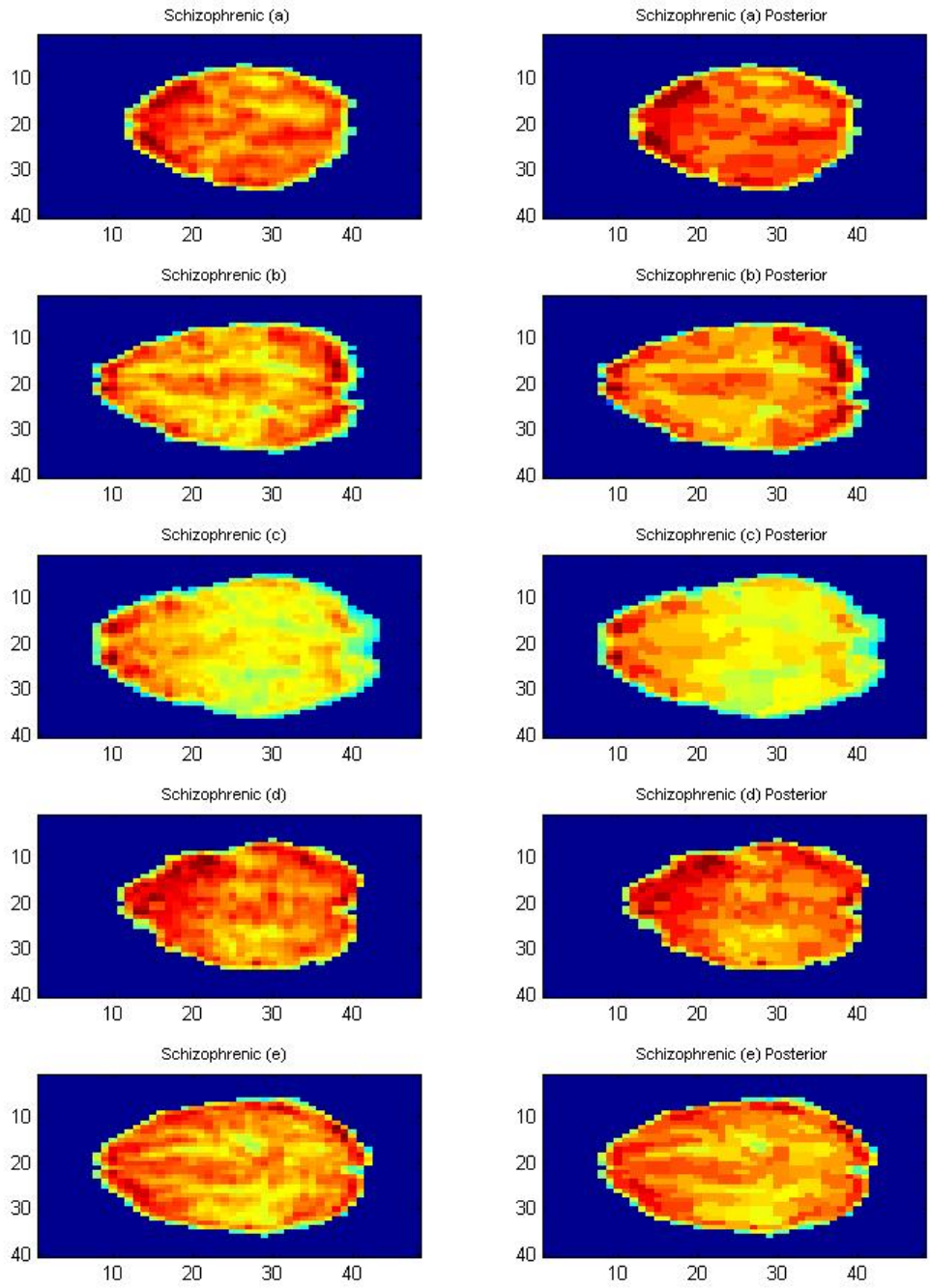

Figure 3.15: Original image and posterior distribution of slice 28 ($\hat{Y}$) in sample schizophrenic subjects. From 5000 iterations of the MCMC sampling algorithm result after 100000 iterations, we display average of the posterior.

visually comparing the original images and posterior distributions in Figures 3.14 and 3.15, they indeed look quite similar. To measure exactly how similar the two images are, we compute the MI between the original and simulated images; the results are presented in Table 3.3. The MI values confirm that results of applying the BFM to images segmented by Ncut are very similar to the original images.

From Figures 3.14 and 3.15, the high levels of activation are also exhibited in both original and posterior images. We are able to preserve most of the information present in the image data by Ncut, and at the same time, the computational complexity is successfully reduced.

Figure 3.16 shows the cumulative effect of adding factors to the model in one of the control subjects. As expected, the estimates become more accurate as the number of factors increases. This is illustrated in Figure 3.16, which summarizes the difference between adding successive factors. The first panel gives $\hat{Y} = \hat{\beta}_1 \hat{f}_1$ and the panel labeled i gives $\hat{Y} = \hat{\beta}_1 \hat{f}_1 + \hat{\beta}_2 \hat{f}_2 + \cdots + \hat{\beta}_i \hat{f}_i$. Even though the first factor explains 52% of the variance in the traditional factor model, the effect of this factor isn't that strong. When the second factor is added in, the contour outline of the brain is revealed. As higher factors have ever-weakening influence, the differences among the last several panels are understandably small. In Section 3.5.5, we will discuss correlation between the factors and activation.

3.5 Experimental Results with All fMRI Subjects

Analyzing individual subjects is not the only goal that we have, we are also quite interested in comparing data from different subject groups, i.e. controls against patients with schizophrenia. Various methods have been proposed in the literature to compare data from different groups (see for example Zhang and Shen, 2012). One popular approach is to bring all the subjects into alignment, and then take the mean or median images among those in

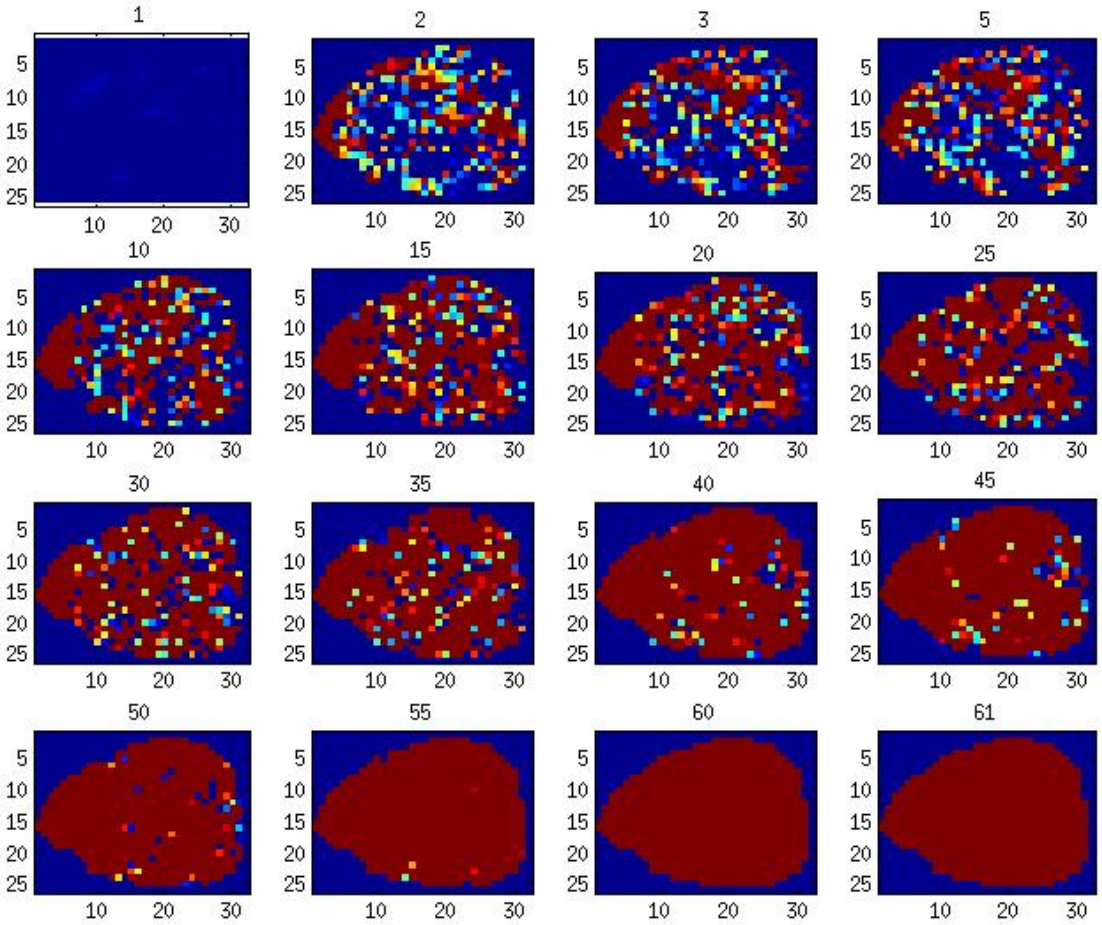

Figure 3.16: Panel 1 shows $\hat{Y}$ given the effect of the first factor. The plot in panel 2 shows $\hat{Y}$ by the combined effect of the first and second factors and so on. Finally is the full 81 factors effect. The color scales in all panels are the same.

the same group, which are then used as atlases for comparison (for example, Argall et al., 2006). In this section, we consider the use of such an approach. After artifact detection (ART) SPM registration, we fit the BFM to the representative mean/median images from the two groups separately. We then explore whether the results of the factor model can be used to differentiate between the groups. If so, this will potentially be another application of the BFM for fMRI analysis.

First, we present BFM results using the data mean, followed by results of the data median. We then compare the results of the two approaches with mutual information. Finally, we try to associate the factors with task-relevant activation..

3.5.1 the Bayesian Factor Model Result of Data Mean

One known issue with the data mean is that it is affected by outliers, a problem not shared by the median. In our experiments, we however hypothesize that if the brain images within a group are well-aligned, e.g. groupwise nonrigid registration algorithm is a popular approach for such alignment (Liao et al., 2012), the outlier issue can be alleviated or avoided, i.e. the effects of outlier are reduced. We observe that it is indeed the case for the subjects that we are analyzing in this thesis, which is evident from the plot of the mean/median of the image (Figures 3.20, 3.24). We average each voxel value over the subjects in the two groups as data mean value.

Result of Normalized Cut

We use normalized cut to reduce the mean data as shown in Figure 3.17. We choose 268 and 287 segments for control and schizophrenic mean data respectively, corresponding to roughly 1/4 of the number of original voxels based on the smallest BIC and AIC values in Figure 3.17.

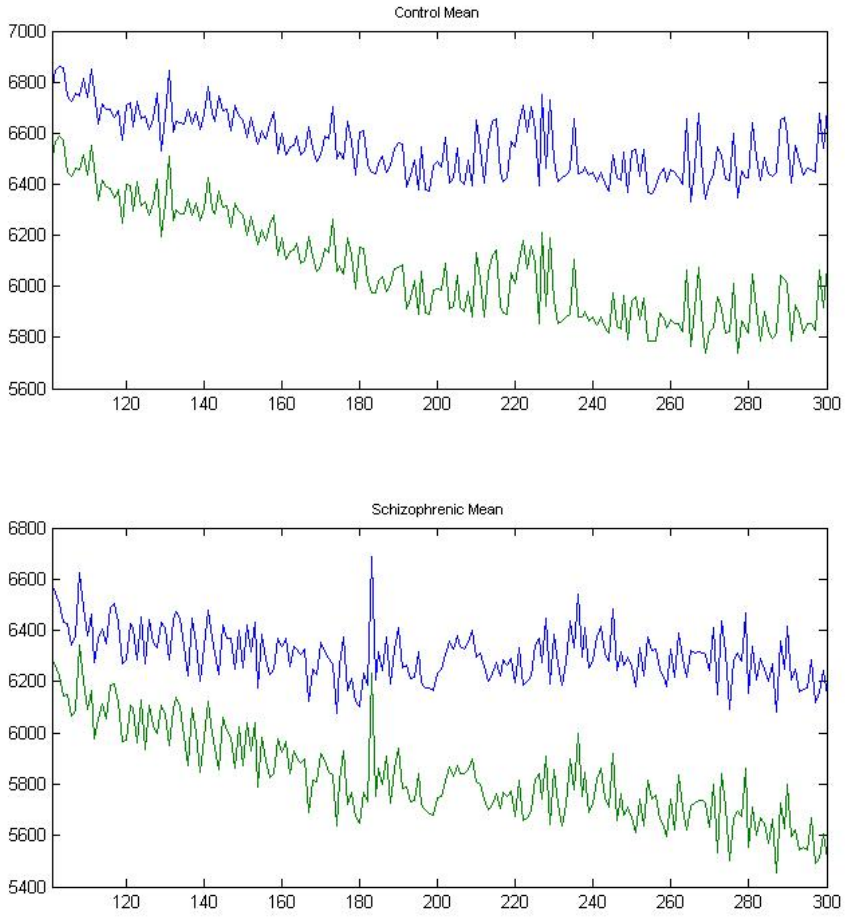

Figure 3.17: AIC and BIC values for the mean of control and schizophrenic subjects segmentation. The upper line is BIC, and the lower line is AIC values. The y-axis shows the value of the selection criteria, x-axis is the number of segmentations for Ncut.

Number of factors for data mean

Picking the appropriate number of factors is an important issue for factor analysis, which may greatly affect the overall result. Similar to the experiments in previous sections, we use the smallest number of factors for which the cumulative percent of variance explained is no less than 95%.

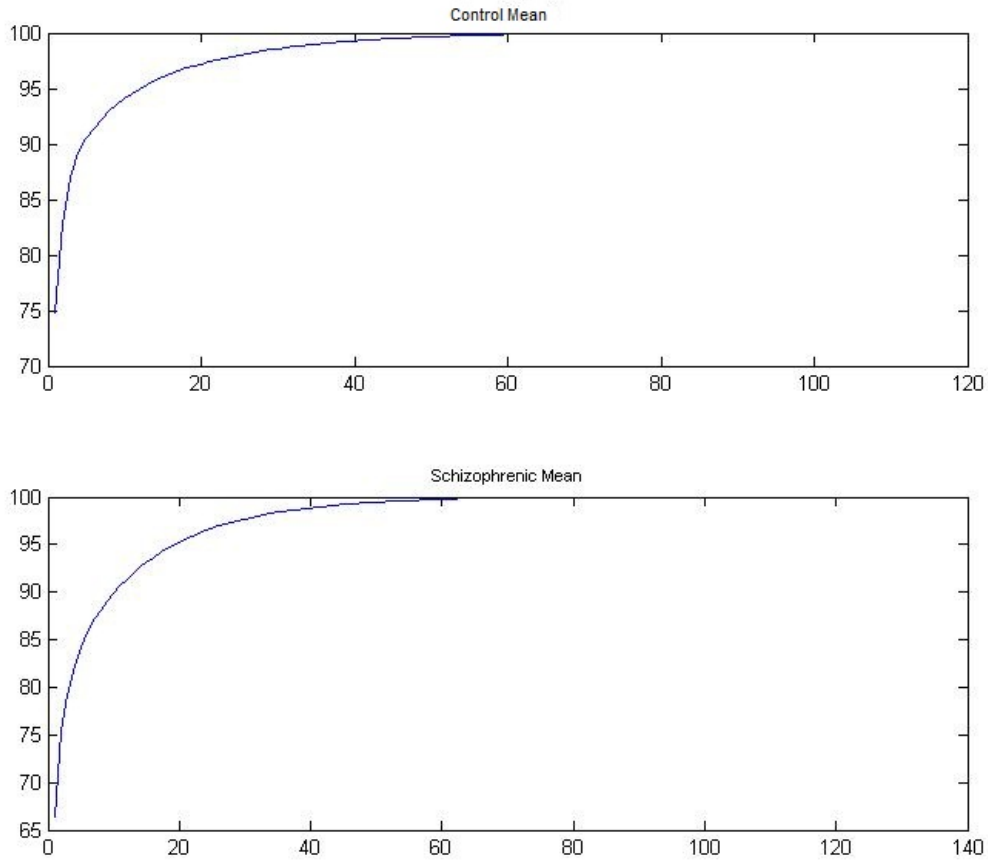

Figure 3.18: Selection of the numbers of factors in control subjects factor models. In traditional factor analysis, the numbers of factors are as above. The first 13 factors explain 95% of the variation in control average data; the first 20 factors explain 95% of the variation in average schizophrenic data.

Figure 3.18 shows the numbers of factors and the cumulative percent of variation that they explain in the average factor models. The first five factors for the control group and eleven factors for the schizophrenic group explain 90% of the variation in the data; 95% of the variation is explained by the first 13 factors of the control group and by the first 20 factors of the schizophrenic group. To explain 98% of the variation, we need to include 25 factors for the control group and 32 factors for the schizophrenic group. In the control group, the first factor explains 75% of the variation; in the schizophrenic group, the first factor explains 66% of the variation. The first five factors together in the two groups explain about 85% of the variation. Starting from about the seventh factor, the growth trend weakens. For the schizophrenic group, more factors are needed than in control group at the same variation level. In the orthogonal rotation BFM, more factors means that the data need more independent structures to be explained. The results on the necessary numbers of factors might indicate that the data of the schizophrenic group as a whole could be more complicated than the data of the control group.

3.5.2 Result of the Bayesian factor model for data mean

We summarize the results of BFM on Slice 28 of all subjects in each group. It takes approximately 4 hours to run on 32 8-CPU Power4 nodes with 16GB of RAM for the mean of each group. The histograms of communalities for the control and schizophrenic groups are shown in Figure 3.19.

From the Figure, we can see that most of the communalities are greater than 0.75 for both groups' mean Ncut images; the average of the communalities for the control group is 0.8719, and for the schizophrenic group, it is 0.8825; both are considered high in terms of explainability. This demonstrates that much of the variability has been explained by the BFM.

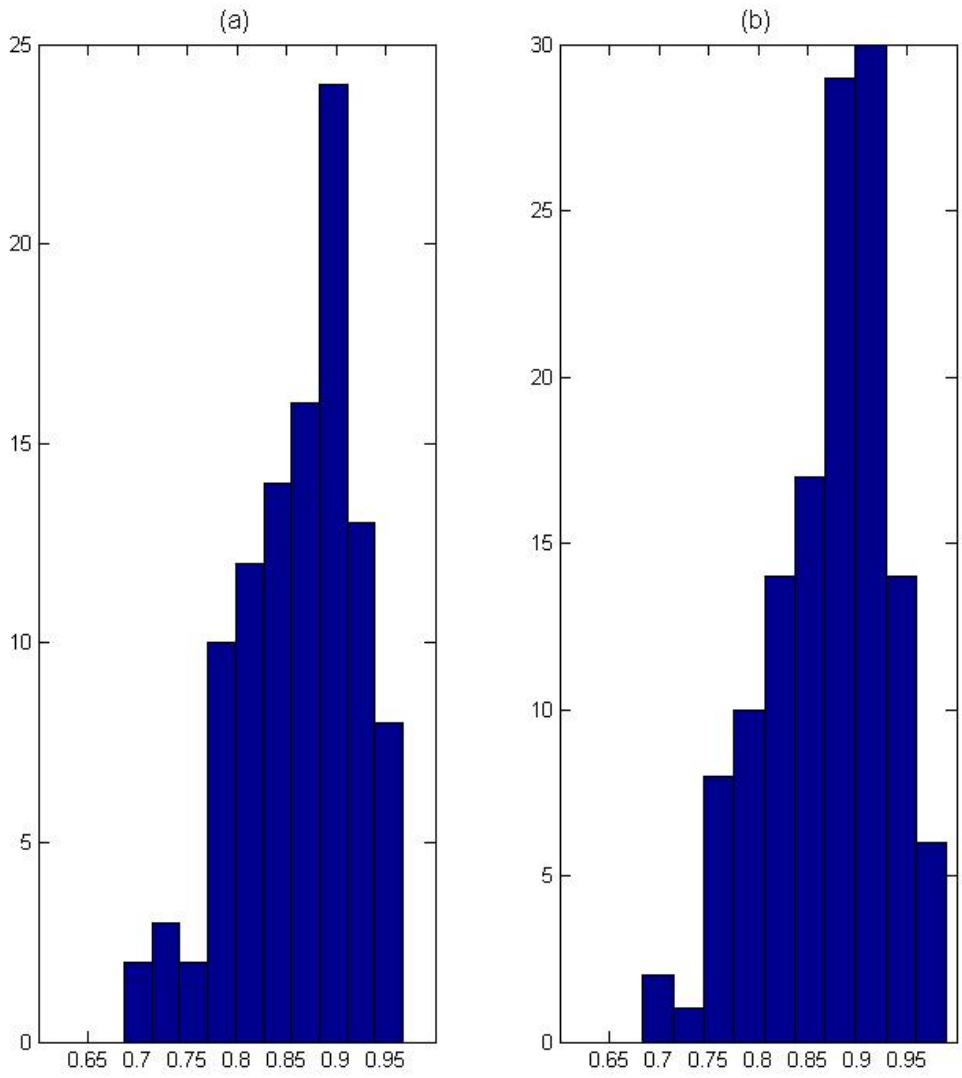

Figure 3.19: (a). Communality histogram of the mean Ncut control group; (b). communality histogram of the mean Ncut schizophrenic group

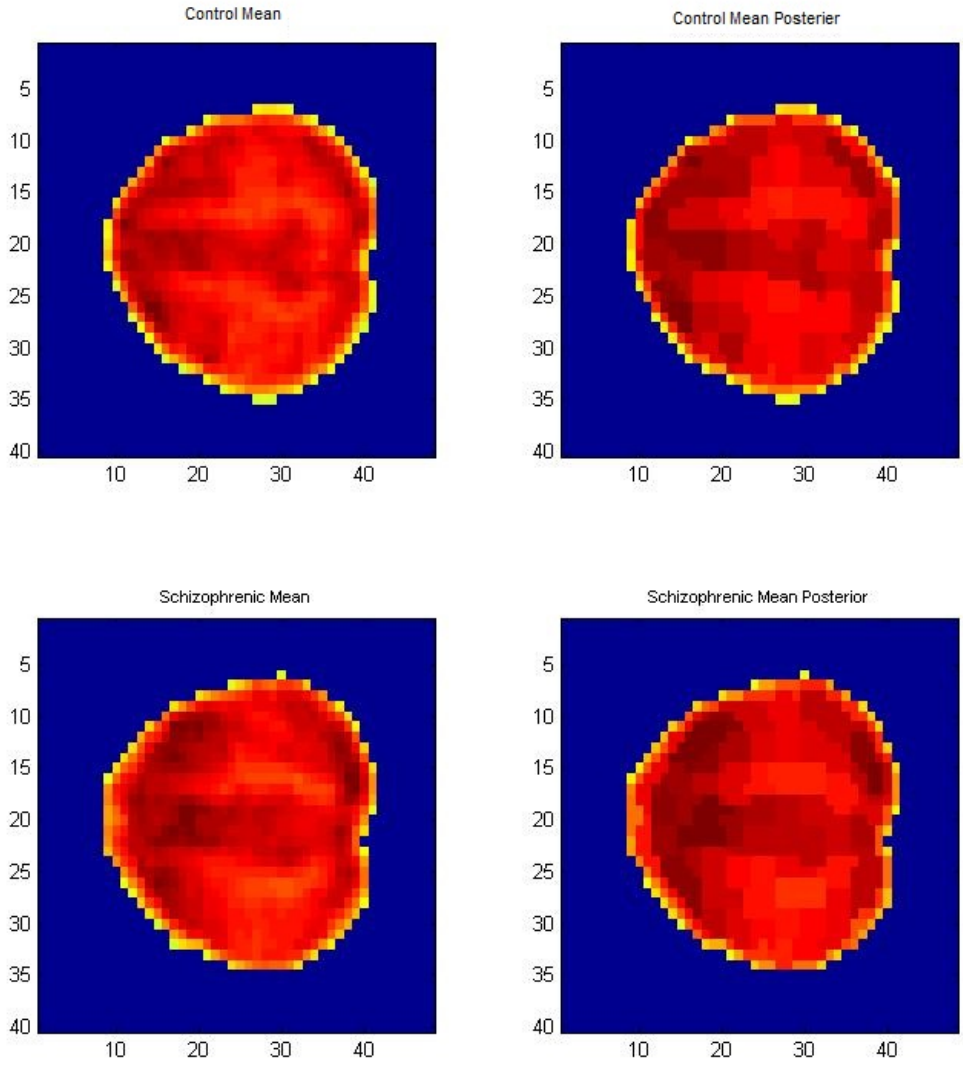

Figure 3.20: Original average image and posterior distribution of slice 28 ($\hat{Y}$) in average data. From 5000 iterations of the MCMC sampling algorithm result after 100000 iterations, we display the average of posterior.

The mutual informations for the control group and the schizophrenic group between mean Ncut images without the BFM and posterior mean images are 0.9011 and 0.8812, which indicates that the mean images and posterior images in Figure 3.20 are indeed quite similar.

3.5.3 the Bayesian factor model Result of Data Median

In this subsection, we use the medians of the two groups instead of the means. We take the median of each voxel value over the subjects in the two groups as the data median value.

Result of Normalized Cut

We use Ncut to reduce the median data as shown in Figure 3.21. We choose 286 and 298 segmentations for the control and schizophrenic median images, respectively, based on the smallest BIC and AIC values.

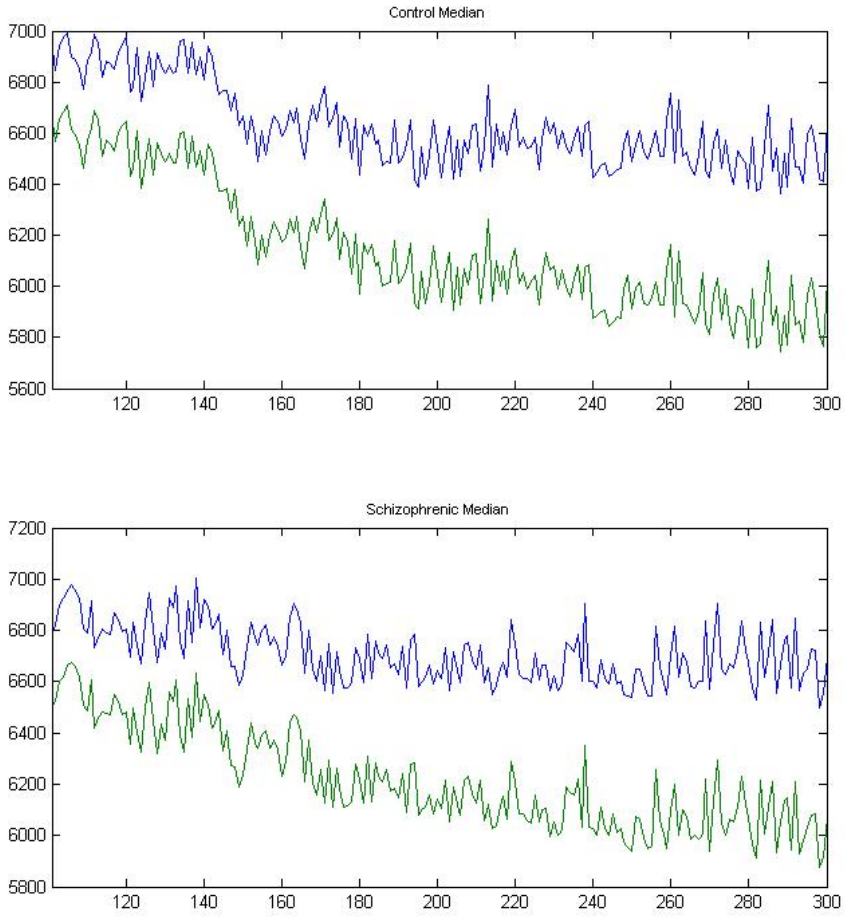

Figure 3.21: AIC and BIC values for the median of control and schizophrenic subjects segmentation. The upper line is BIC, and the lower line is AIC values. The y-axis shows the value of the selection criteria, x-axis is the number of segmentations for Ncut.

Number of factors for data median

As in Section 3.5.1, we choose the smallest number of factors for which the cumulative percent of variance explained was no less than 95%.

Figure 3.22 shows the numbers of factors and the cumulative percent of variation that they explain in the median factor models. The first 20 factors for the control group and 23 factors for the schizophrenic group explain 90% of the variation in the data; 95% of the variation is explained by the first 32 factors for the control group and by the first 35 factors for the schizophrenic group. To explain 98% of the variation, we need to include 46 factors for the control group and 49 factors of the schizophrenic group. In the control group, the first factor explains 52% of the variation; in the schizophrenic group, the first factor explains 45% of the variation. The first six factors together in the two groups explain about 75% of the variation. Starting from about the tenth factor, the growth trend weakens. At the same variation level, the factor numbers for the two groups are similar, in contrast with the findings based on the mean image.

Result of the Bayesian factor model for data median

We summarize the results of the BFM on Slice 28 of all subjects in each group in this subsection. It takes approximately 4 hours to run on 32 8-CPU Power4 nodes with 16GB of RAM for the median of each group. The histograms of communalities for the control and schizophrenic groups are shown in Figure 3.23.

From the Figure, we can see that most of the communalities are greater than 0.80 for both groups median Ncut image data; the average of the communalities for the control group is 0.9041, and for the schizophrenic group, it is 0.9146. Much of the variance has been explained by the BFM. We notice further that the communalities from the median are larger than those from the group mean.

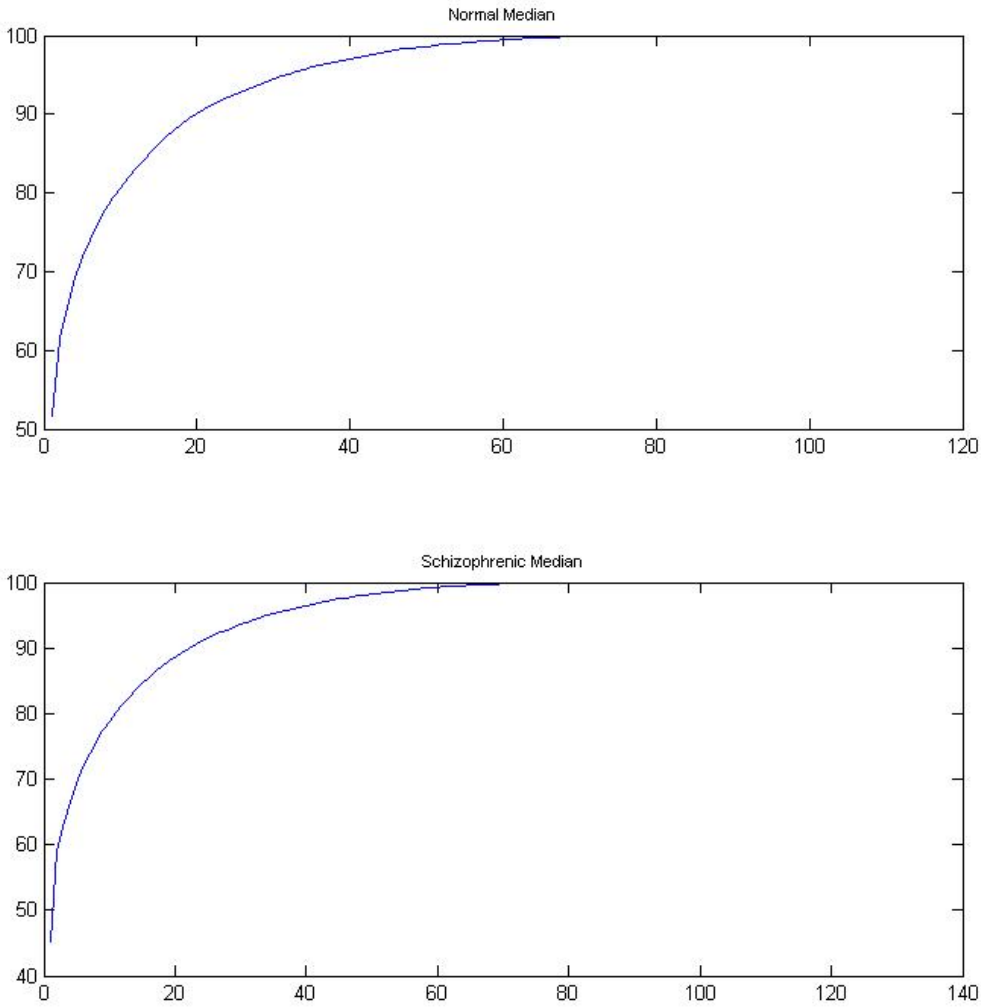

Figure 3.22: Selection of the numbers of factors in median data factor models. In traditional factor analysis, the numbers of factors are as above. The first 32 factors explain 95% of the variation in the median control data; the first 35 factors explain 95% of the variation in the median schizophrenic data.

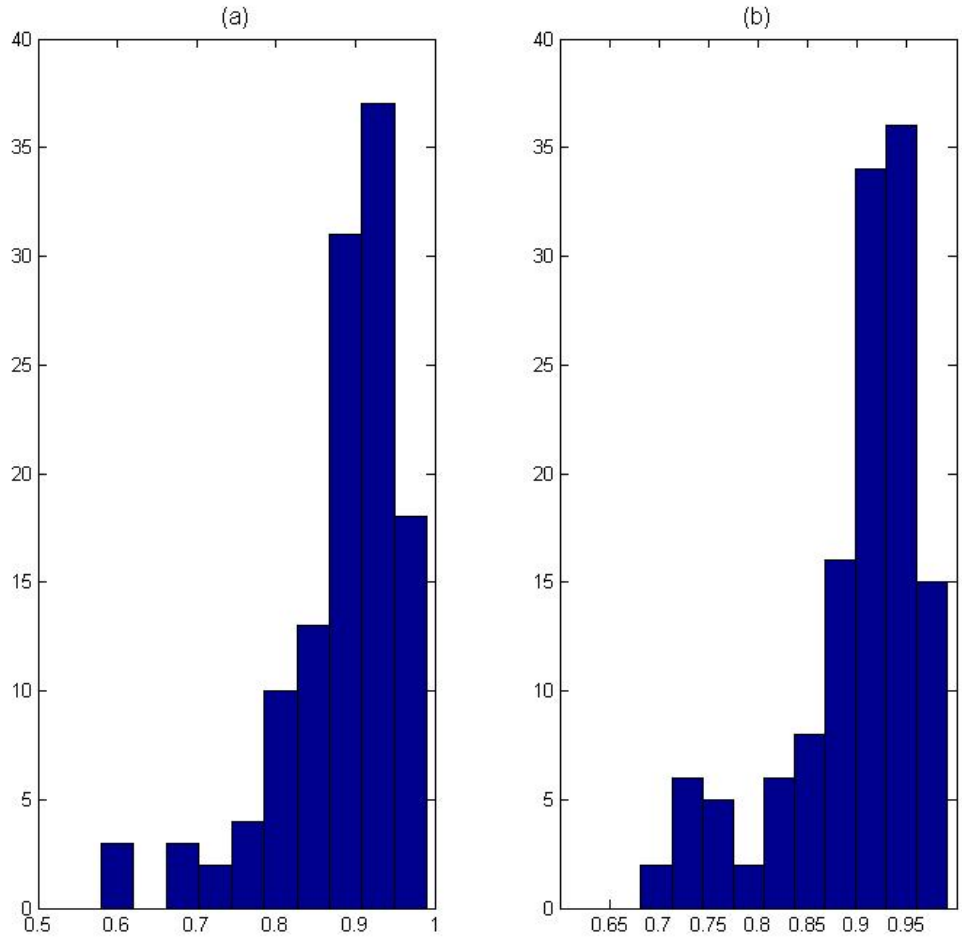

Figure 3.23: (a). Communality distribution of the median Ncut control group; (b). communality distribution of the median Ncut schizophrenic group

The mutual information between median Ncut images, two group posterior images with median images are 0.9165 and 0.8901. The posterior result is close to the median images as can be seen in Figure 3.24. Next, we compare the results of the median and mean data with mutual information.

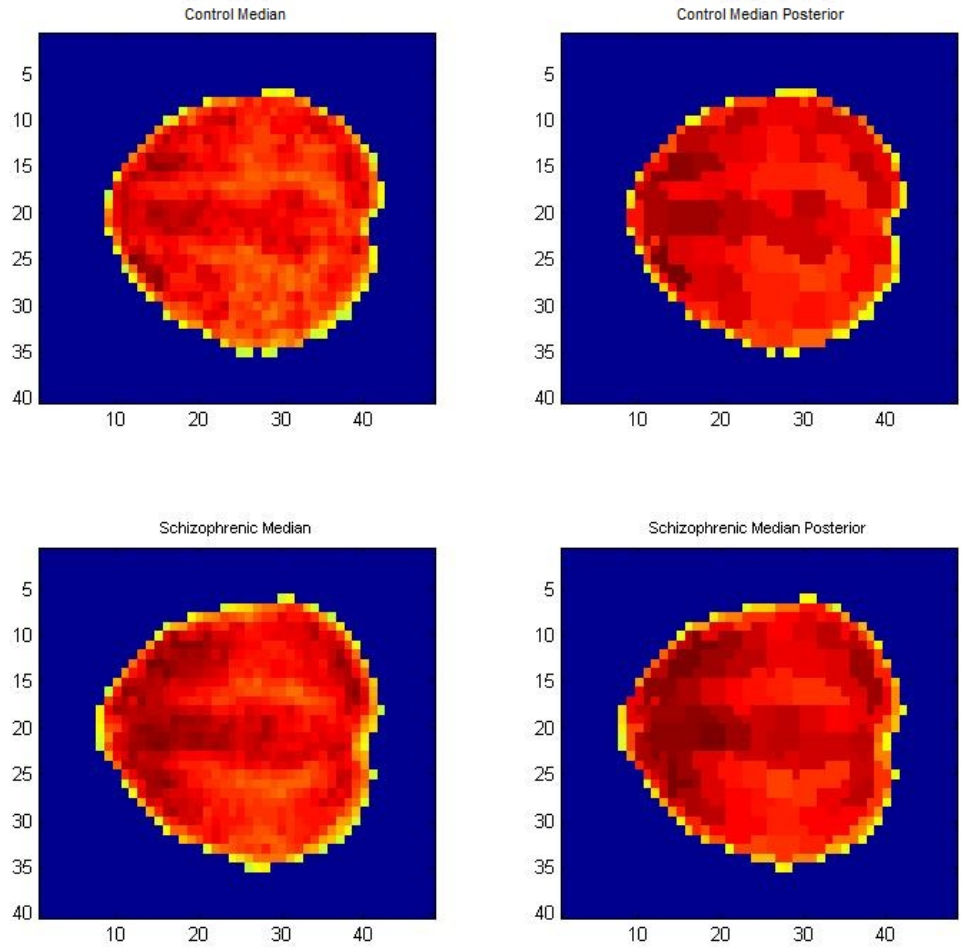

Figure 3.24: Median image and posterior median of slice 28 ($\hat{Y}$). From 5000 iterations of the MCMC sampling algorithm result after 100000 iterations, we chose the average of posterior simulation result.

3.5.4 Comparing mean and median results with mutual information

In order to determine which average method, mean or median should be used, we use mutual information to compare the original individual images with the mean and median simulation results respectively; results are summarized in Table 3.4.

Table 3.4: Mutual information for mean and median

Subject	Control Subjects		Schizophrenic Subjects	
	Mean	Median	Mean	Median
1	0.4727	0.5253	0.3317	0.3495
2	0.4481	0.4896	0.5050	0.5369
3	0.3521	0.3388	0.3007	0.3931
4	0.4861	0.5263	0.3799	0.4224
5	0.5188	0.5922	0.6736	0.7269
6	0.5268	0.5626	0.4581	0.5625
7	0.5092	0.4828	0.5176	0.5138
8	0.5674	0.6222	0.5022	0.5411
9	0.4724	0.4393	0.2338	0.3284
10	0.3074	0.3662	0.4955	0.5559
11	0.5239	0.5148	0.5185	0.6542
12	0.5483	0.5792	0.3703	0.4033
13	0.4646	0.4901	0.4494	0.5086
14	0.5173	0.5258	0.2941	0.3499
15	0.4630	0.4455	0.4582	0.5676
16			0.4257	0.5201
$\bar{x}$	0.4785	0.5000	0.4321	0.4959
Σ	0.0049	0.0062	0.0121	0.1278

Since mean and median are the measures that trend central, average images reduce the variability. The magnitudes of the mutual information values between the original individual images and group average images (mean or median) are relatively small.

The MIs of the original and median images are a little better than the values between the original and mean images in most subjects. We notice that the variance of MI values in the control group is smaller than that in the schizophrenic group. From these group data, the control group appears to be more homogeneous. We prefer the median image as it keeps more integral information. Also the better overall communalities of median images indicate that more variation has been explained than for the mean images. And the median image

is more robust, less liable to be affected by anomalies in the data. In the next subsection, therefore, we will use the median images to analyze the data.

3.5.5 Bayesian factors and fMRI activated regions

fMRI has revolutionized our ability to understand brain function. One of the key problems for fMRI analysis is how to determine the neuronal activated regions. In this subsection, we propose to use the BFM for the task.

fMRI activation quantification

We conduct two sample t-tests separately for each group, to compare voxelwise levels of activation during task and rest, as in Ward (2000). If the t-test result is significant at a level of $\alpha = 0.05$, we consider the voxel to be activated; segments containing at least 20% activated voxels are considered to be activated, as shown in Figure 3.25.

We think there are two problems with the way that we use Ncut to reduce the number of regions. One is that it includes the inactivated voxels. The other is that inactivated voxel values will weaken the level of activation of the activated voxels in the same Ncut regions. If the voxel of one individual subject is tested as activation, then this voxel is thought be a activation as the blue diamonds in the average image, Figure 3.26.

As discussed in the last paragraph, since median measure reduce the variability, and voxels in activated Ncut regions are not all themselves activated, the activated regions of Median images and individual subjects are not coincident. In the schizophrenia group, subjects are more diversified from Figure 3.26. Median image and Ncut keep most of the activations of individual subjects.

Association between Bayesian factors and activated regions

From the rotated factor loading in the Bayesian factor result, we can figure out that each region can be explained mainly by one factor.

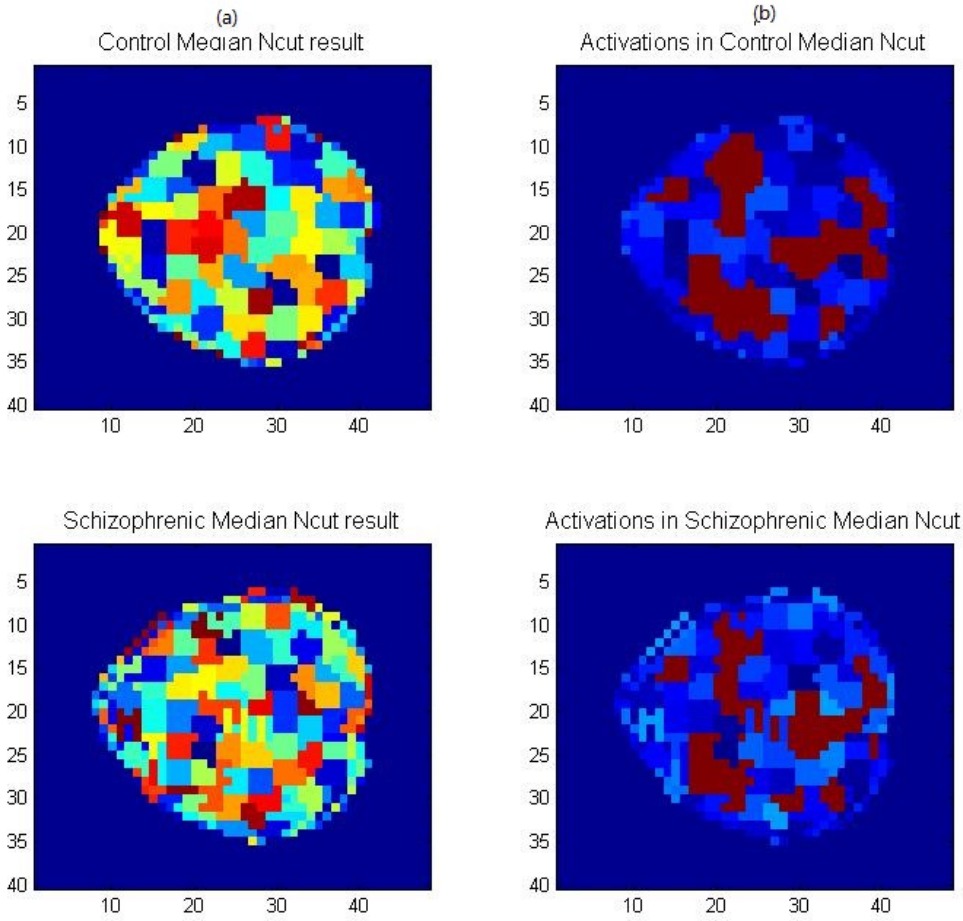

Figure 3.25: (a) Ncut result images. (b), the red cuts are the active cuts in control and schizophrenia median images.

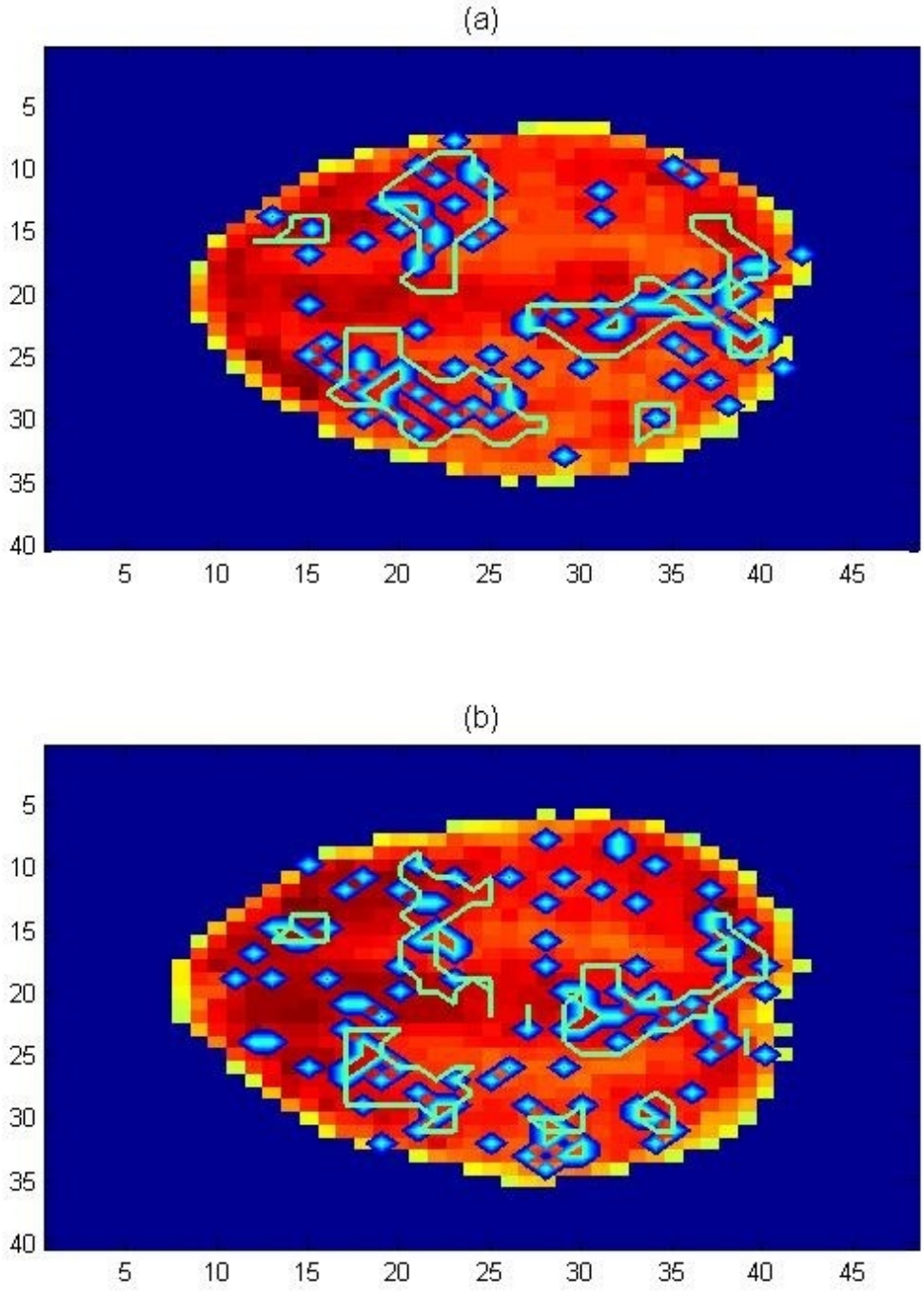

Figure 3.26: (a): the green lines are control Ncut active cuts, the blue diamonds are the active voxels in the individual subjects on the control median image. (b): the green lines are schizophrenia Ncut active cuts, the blue diamonds are the active voxels in the individual schizophrenia subjects on the schizophrenia median images.

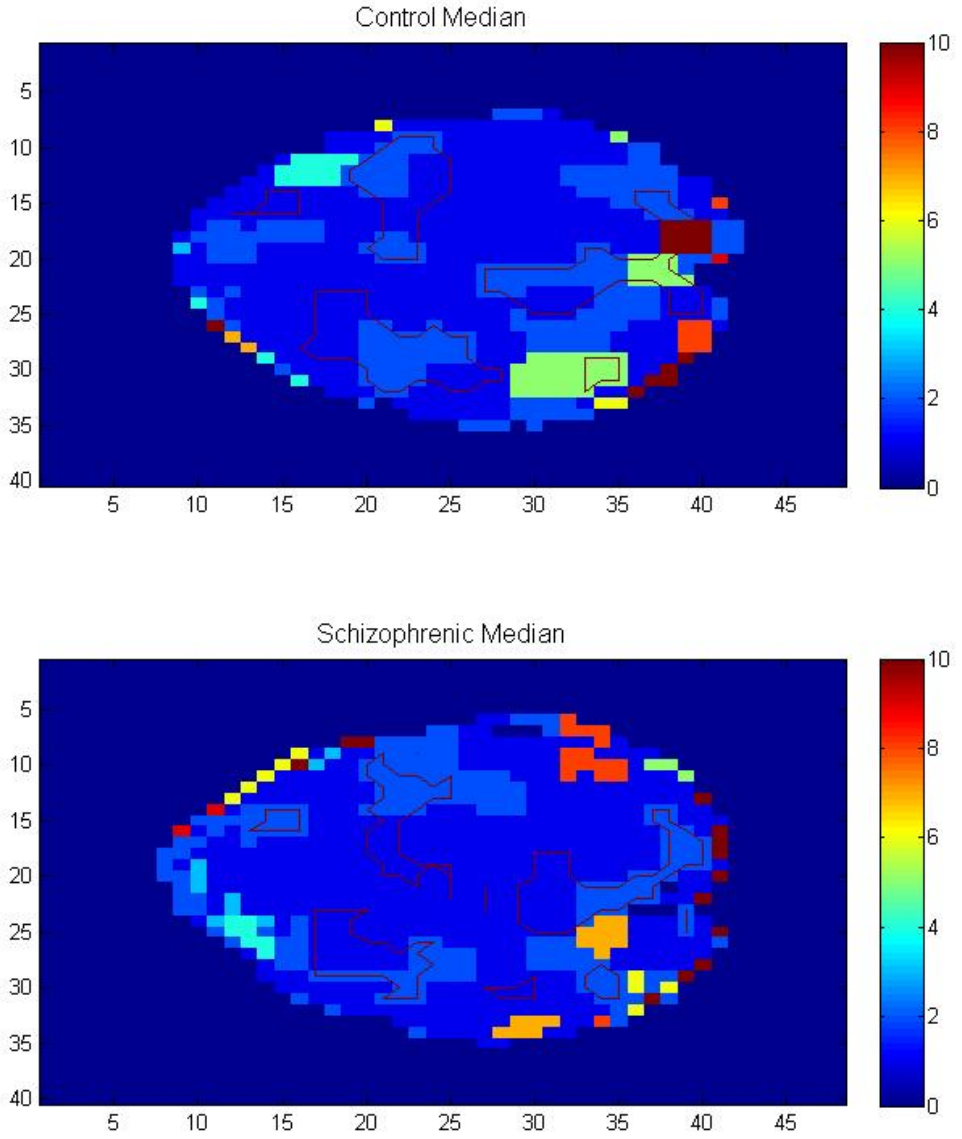

Figure 3.27: Illustration of association between factors and fMRI activated cut regions. All the factors which are larger than Factor 10 are shown as Factor 10. The red outlined segments are the areas of activation.

In Figure 3.27, Factor 1 contains the maximum variance among all factors in both control and schizophrenic groups, therefore most of the voxels are mainly correlated with Factor 1. Factor 2 has the next largest factor scores for most of the images. Both Factor 1 and 2 explain part of the activated area for both control and schizophrenic groups. Other minor factors aside from those two tend to explain the edge of the image.

3.6 Discussion

In this chapter, we introduce the BFM and its application to fMRI data analysis. Through a series of experiments, we observe that the BFM indeed provides us an effective and efficient method to understand brain activity. Listed here are some general observations and discussion of the results.

1. Effectivity of the BFM

From the MI results in Sections 3 and 4, it is evident that the BFM can be used to successfully explain the fMRI data. In our experiments, we choose to use a fixed number of factors which explain 95% of the variance.

2. Number of factors

The number of factors is fixed throughout the experiment; the question remains how to select the optimum number of factors if we treat it as another parameter of interest.

3. Efficacy of the BFM in terms of explaining activated voxels

From the experimental results, it is evident that the majority of the activated segments score high on the first two factors. However the first two factors also account for certain inactivated segments. Therefore, we cannot associate activated segments (or subsegments) exclusively with one specific factor. Future work will involve finding models to separate those inactivated segments from the top factors.

4. Median or mean for comparison between two groups?

There is more than one subject in each of the experimental groups. We use both the mean and median images to compare the differences between the two groups. From our experimental results, the median image obtains slightly better model fitting results than the mean image, which can be explained from the fact that the median image is more robust, hence less liable to be affected by anomalies in the data.

5. Grouping neighboring voxels for the BFM analysis

We use normalized cut to group similar neighboring voxels. Are there any better alternatives available to cluster fMRI voxels into different groups? Further study is needed to evaluate different alternative approaches. We will discuss this issue in greater detail in the future work chapter of this dissertation.

6. How to deal with stimulus part in fMRI data?

Time series of fMRI data are inherently influenced by the stimulus process. How to incorporate the effect of those signals into the model is an active field of research, and is the topic of the next chapter.

Chapter 4

Semiparametric Bayesian Analysis for fMRI Data

The existing fMRI analysis literature offers many options to handle noisy brain signals, such as highpass filters (Mather, 2004), hemodynamic response function (HRF) convolution and correlation estimation. Among them, HRF estimation, which models the brain as a linear "black box" system, is one popular approach to analyzing BOLD fMRI data (Friston et al., 1994; Marrelec et al., 2003). In this chapter, we will discuss whether adding a transformed stimulus part, i.e. HRF to explain evoked activation (for example, Ciuciu et al., 2003; Mumford and Poldrack, 2007; Gössl et al., 2001a; Metropolis et al., 1953), to the original BFM will help us to get better fitting results. We will also demonstrate how the added HRF influences the fMRI signal. The proposed method therefore becomes a semiparametric model, which we will explain in detail in the following sections.

4.1 Semiparametric Bayesian Analysis for fMRI Analysis

Semiparametric models combine a parametric component with a nonparametric component, where the parametric component might be some mixed linear model (Teh et al., 2005; Ruppert et al., 2003; Lin and Carroll, 2001). The normal linear random effects model by Dey et al. (1998) can be expressed as

$$Y = X\beta + Zb + \epsilon, \tag{4.1.1}$$

where Y is the outcome matrix. The first term is the parametric term, X is a fixed covariate matrix, β is a parametric matrix of regression coefficients. For the nonparametric part, Z is

a covariate matrix for the random effects, b, and ϵ is the error matrix. Assume b and ϵ are independent and $b \sim N(0, D)$, $\epsilon \sim N(0, \Sigma)$. Under these assumptions,

$$E(Y|\beta, b) \sim N(X\beta + Zb, \Sigma). \tag{4.1.2}$$

In Chapter 3, we introduced the BFM for fMRI data analysis. In this chapter, we extend the Bayesian factor analysis by including a nonparametric component (Carota and Parmigiani, 1996; Xu, 2007). The semiparametric model for fMRI can be expressed as,

$$y_{ti} = f_t\beta_i + z_t b_i + \epsilon_{ti} \qquad \text{for } i = 1, \ldots, n \tag{4.1.3}$$

This model describes the signal Y_i at voxel i as a combination of the factor loading matrix β and a nonparametric matrix Z superposed by a white-noise error term ϵ_i.

In this model, Z is a function of the presented on-off stimulus or task onset. One way to formulate Z is by first applying a temporal shift to the original stimulus by a time-delay d, which is followed by a convolution with a parametric hemodynamic response function (HRF) h, so that:

$$z_t = \sum_{s=0}^{t-d_i} h(s, \theta) x_{t-d-s} \tag{4.1.4}$$

where x is on-off stimulus, e.g. fixation and anti-saccade, in the example we have been considering in this work; this can be considered as a binary onset stimulus (1 or 0). The HRF h, through θ, carries the temporal delay and dispersion of that signal. A typical HRF is shown in Figure 4.1. After several seconds, the BOLD level reaches a peak, usually followed by an undershoot before returning to baseline.

The Poisson function (Friston et al., 1994) was the first distribution used in the literature to represent the HRF. A major disadvantage of the Poisson model is that it is not flexible enough to capture all of the nuances of the HRF, since it has just a single parameter and

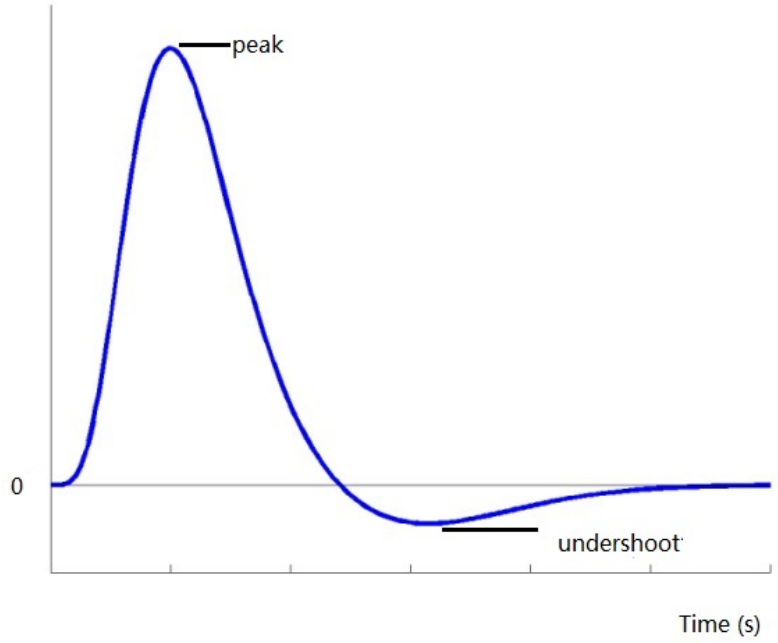

Figure 4.1: A typical HRF of neural activity.

models discrete phenomena. Nowadays the HRF is often modeled as a Gamma(α, γ) (Gössl et al., 2000) or double Gamma function (Friston et al., 1998). The Gamma HRF can be expressed as

$$z_{it} = \sum_{s=0}^{t-d_i} s^{\alpha-1} \frac{\gamma^{\alpha} e^{-\gamma s}}{\Gamma(\alpha)} x_{t-d_i-s} \tag{4.1.5}$$

and the double Gamma function as

$$z_{it} = \sum_{s=0}^{t-d_i} \left(\frac{s^{\alpha_1-1} \gamma_1^{\alpha_1} e^{-\gamma_1 s}}{\Gamma(\alpha_1)} - c \frac{s^{\alpha_2-1} \gamma_2^{\alpha_2} e^{-\gamma_2 s}}{\Gamma(\alpha_2)} \right) x_{t-d_i-s} \tag{4.1.6}$$

where α_1 and α_2 control the shape, γ_1 and γ_2 control the scale. c is the ratio of the response to post stimulus undershoot. Post stimulus undershoot indicates the transient signal fall below baseline signal level after cessation of stimulus.

4.2 Experimental Results

In this section, we first examine the choice of using the Gamma and double Gamma distributions to represent the HRF. We then validate the prosposed model by comparing the results from the semiparametric Bayesian factor model (SBFM) with the BFM.

4.2.1 Gamma HRF and Double Gamma HRF

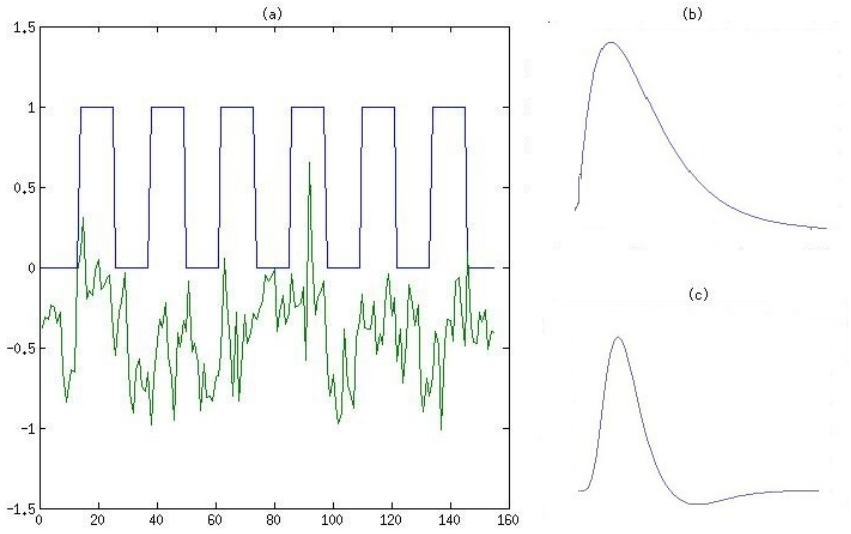

Figure 4.2: (a)Upper line is the stimulus presentation trail, lower line is BOLD response measured at a single voxel from our data. There are strong active stimulus time series. (b)sample z_t by the Gamma HRF and (c) sample z_t by the double Gamma HRF with undershoot.

For comparison purposes, we experimented with both the Gamma HRF and double Gamma HRF. The parameters of the Gamma HRF function in Equation (4.1.5) are set to $\alpha = 5$ and $\gamma = 0.2$. Prior variances for α and γ are set to 1000, prior parameters of Gamma are fixed at 1 as in Gössl et al. (2001b).

The parameters of the double Gamma HRF function in Equation (4.1.6) are set to $\alpha_1 = 6$, $\alpha_2 = 16$, $\gamma_1 = \gamma_2 = 1$ and $c = 1/6$ as in Friston et al. (1998). Prior variances for α and γ are set to 1000, prior parameters of Gamma are fixed at 1 as in Gamma HRF.

The stimulus and the BOLD response at a random voxel are shown in Figure 4.2(a). There is a strong correlation between the stimulus and the BOLD response. Figure 4.2(b) is the sample z_t by the Gamma HRF, in which $d = 0$, $\alpha = 5$ and $\gamma = 0.2$. Figure 4.2(c) is the sample z_t by the double Gamma HRF, in which $d_1 = 6$ (delay for response), $d_2 = 16$ (delay for undershoot), $\alpha_1 = 6$, $\alpha_2 = 16$, $\gamma_1 = \gamma_2 = 1$ and $c = 1/6$ as in Friston et al. (1998).

Figure 4.3 shows the BFM and the SBFM with Gamma and double Gamma HRF simulation posterior results in the two groups Slice 28 median Ncut data through all 81 time points. From the Figure, the two groups have different regions of brain activation. But better estimation of activated brain regions are not observed between the BFM and the SBFM from this figure.

4.2.2 Comparison with the Bayesian factor model

To compare the results between the BFM and the SBFM, we also compute the mutual information between posterior images with the original median image for both the BFM and the SBFM. The values of the mutual information are summarized in Table 4.1. The MIs of the SBFM are slightly larger than those of the Bayesian factor model; furthermore the SSE is smaller for the SBFM than the BFM, the SBFM is better equipped to explain our data .

Table 4.1: Mutual information for the BFM and the SBFM

Method	Control	Schizophrenic
SBFM with Gamma HRF	0.9201	0.8936
SBFM with double Gamma HRF	0.9233	0.8941
BFM	0.9165	0.8901

Figure 4.4 shows the SSEs of the SBFM with Gamma and double Gamma HRF and SSE of the BFM. SSE of the SBFM in both groups are smaller than those of the BFM.

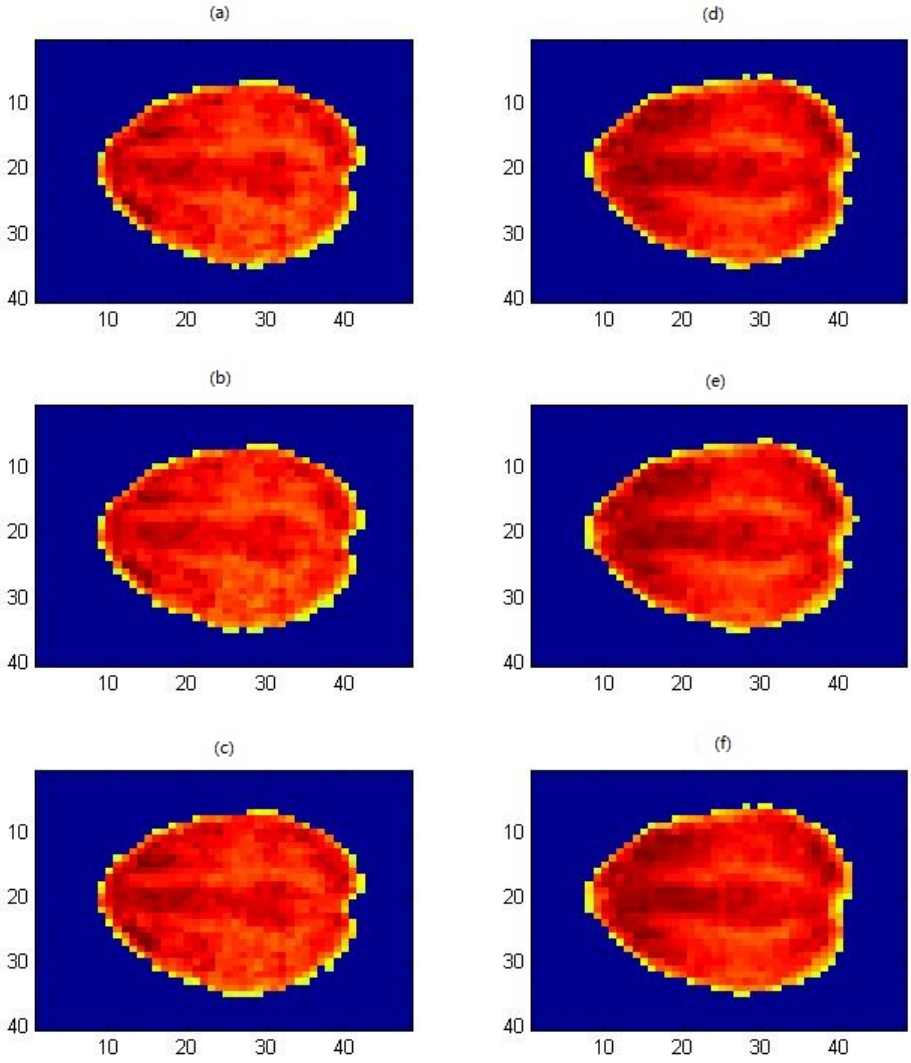

Figure 4.3: (a) the BFM posterior result for the control group; (b) the SBFM with Gamma HRF for the control group; (c) the SBFM with the double Gamma HRF for the control group; (d) the BFM posterior result for the schizophrenic group; (e) the SBFM with the Gamma HRF for the schizophrenic group; (f) the SBFM with the double Gamma HRF for the schizophrenic group.

The SBFM moderately outperforms the BFM. The models are improved by involving the stimulus effect in fMRI BOLD data.

Similarly as Section 3.5.5, from the rotated factor loading in our semiparametric Bayesian factor result, we can figure out that one specified cut could be mainly explained by one specified factor.

In Figures 4.5 and 4.6, Factor 1 explains the maximum variance among all factors in both the control and schizophrenic groups, therefore most of the voxels are mainly correlated with the first factor. Factor 2 has the next largest factor scores for most images. Both Factors 1 and 2, in particular, Factor 2 explain part of the activated area for both controls and schizophrenic group. For both HRF, the first two factors are the main factors that explain around 78% of the activated area. Other minor factors aside from those two tend to explain the edge of the image.

As previously discussed in Section 3.4, for the different subjects, the reactions may not be at exactly the same voxels. Activated Ncut regions will show more generalization for the group data. For both the Gamma and double Gamma HRFs, the posterior results of the activation are similar. Comparing with the result of the Bayesian factor model, there are slight improvements in the SBFM from the MI and SSE results.

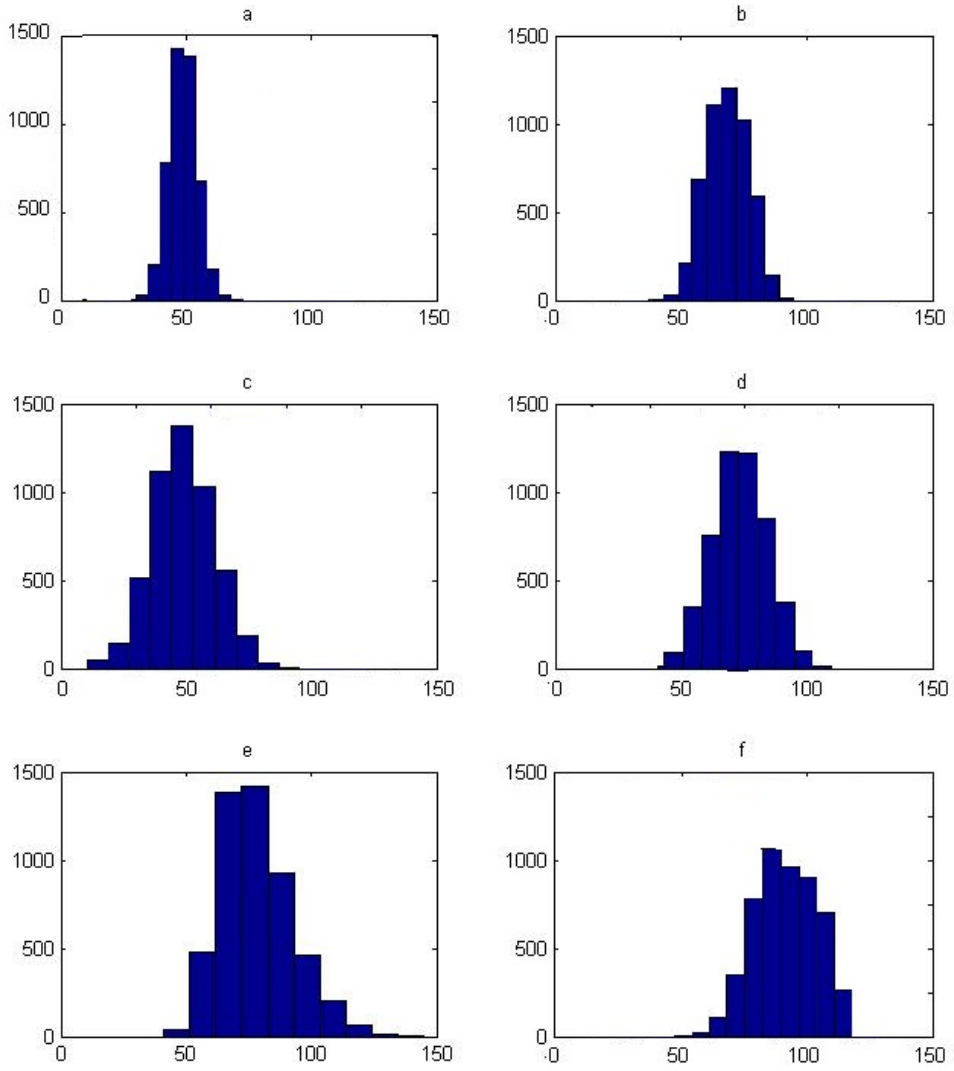

Figure 4.4: (a) SSE of the SBFM with Gamma HRF for the median control subject group; (b) SSE of the SBFM with Gamma HRF for the median schizophrenia subject group; (c) SSE of the SBFM with double Gamma HRF for the median control subject group; (d) SSE of the SBFM with double Gamma for the median schizophrenia subject group; (e) SSE of the BFM for the median control subject group; (f) SSE of the BFM for the median schizophrenia subject group. The distributions of the SBFM SSE look roughly normal. With normality assumption, SSE of the SBFM are better than that of the BFM.

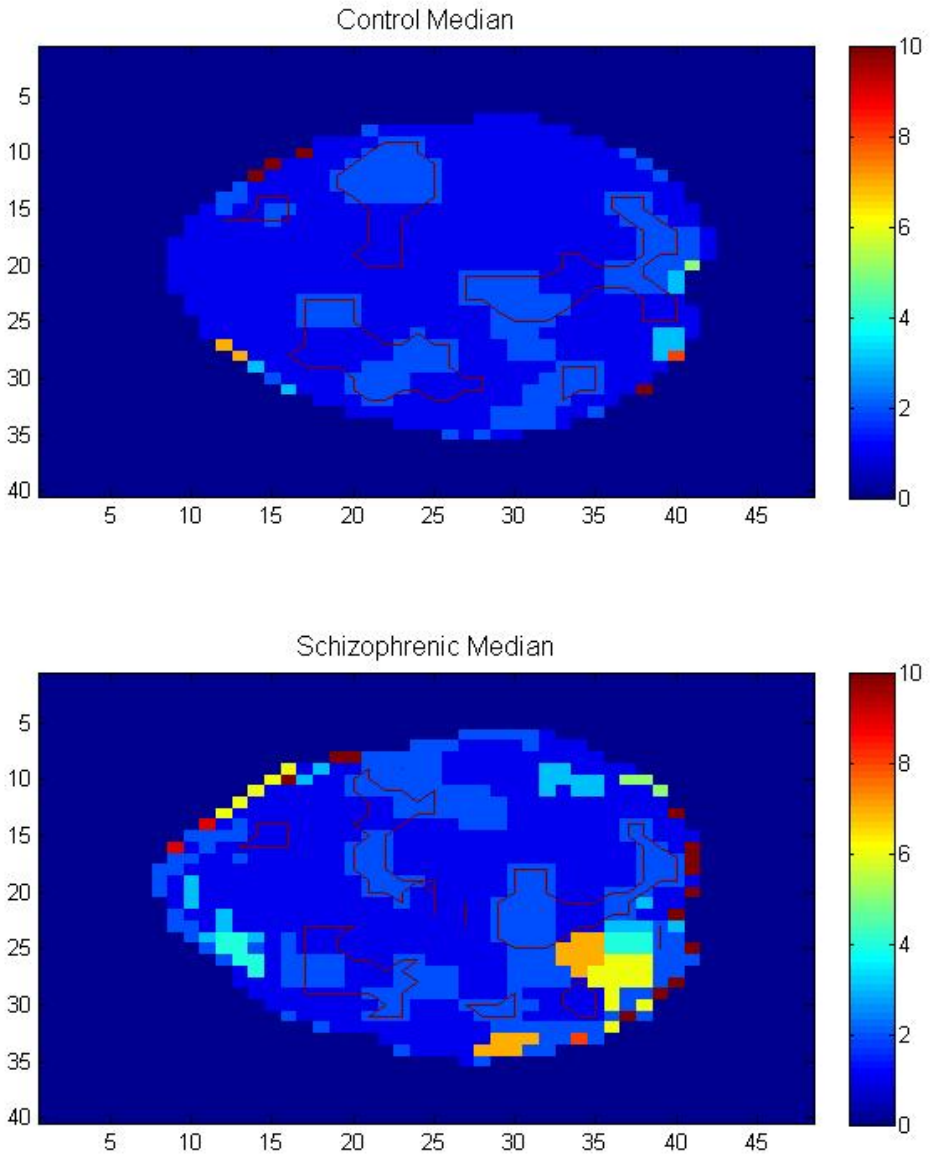

Figure 4.5: Illustration of association between factors and fMRI activated cut regions with Gamma HRF. All the factors which are larger than Factor 10 are shown as Factor 10. The red outlined regions are the areas of activation.

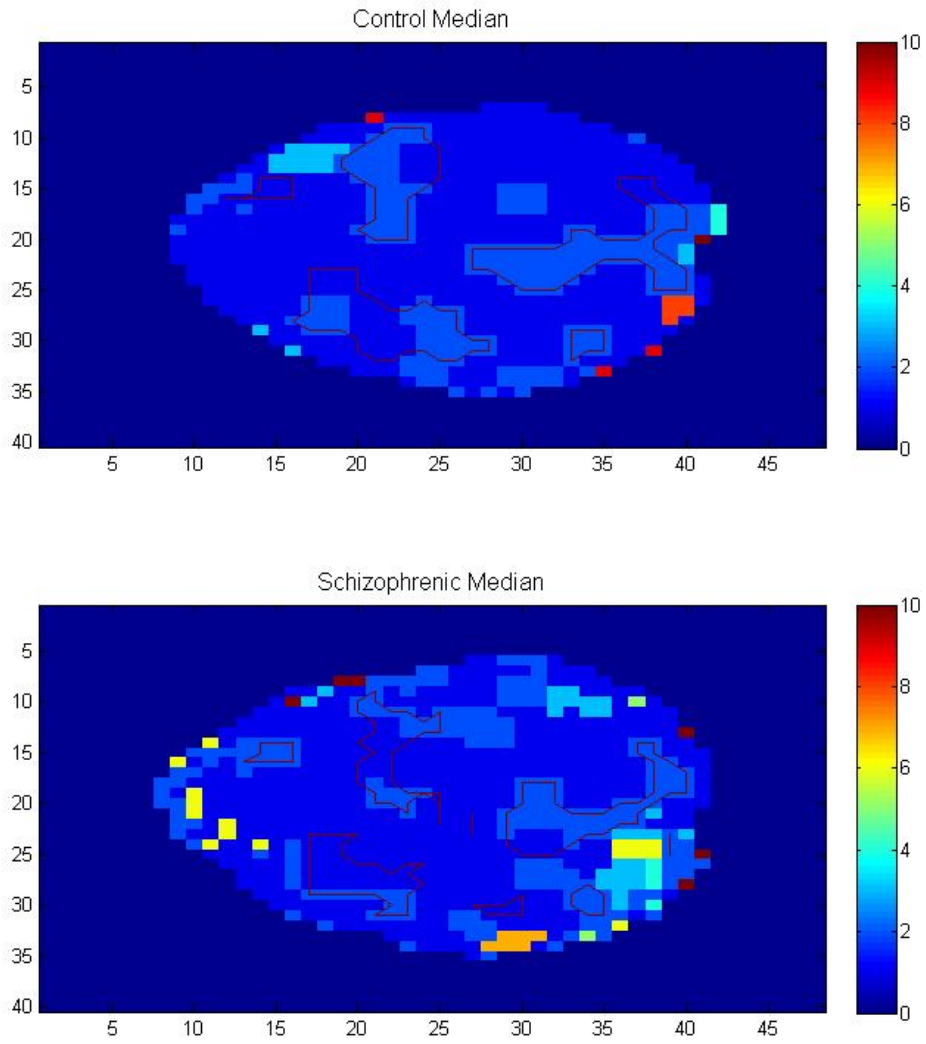

Figure 4.6: Illustration of association between factors and fMRI activated cut regions with double Gamma HRF. All the factors which are larger than Factor 10 are shown as Factor 10. The red outlined regions are the areas of activation.

4.3 Discussion

In this chapter, we analyze fMRI data with the SBFM to account directly for the effect of the stimulus. We experiment with both the Gamma HRF and the double Gamma HRF in the SBFM, and compare with the results that we obtained from the regular Bayesian factor model. It is evident from the results that the SBFM has smaller SSE and larger MI values compared to the Bayesian factor models. As for the two SBFMs with the Gamma HRF and the double Gamma HRF, the double Gamma HRF performs slightly better with control subjects while the Gamma HRF does better with schizophrenic subjects, although the differences are not great.

There are still certain aspects of the SBFM that we have yet to explore::

First of all, can we get still better results by using a different HRF? We experimented with the Gamma and the double Gamma HRFs in this dissertation. More recently, people have proposed to use finite impulse response (FIR) to simulate fMRI data (Glover (1999), Goutte et al. (2000)). Further, FIR can be used to estimate HRF of arbitrary shape for each stimulus type at each voxel of the brain. If we are going to use FIR combined with HRF, similar issues remain as to how to determine the prior distributions of HRF parameters.

Secondly, we choose the time delay to be 6 seconds in our double Gamma HRF experiment. Based on practical data, the time delay indeed exists and is very important to us. Are the delay times the same for different stimulus types? Exploring the effect of a different time delay will be another future topic.

Last but not least, how can we further reduce the computation time of the SBFM? In our experiment, the computation for the SBFM takes approximately 11 hours to run on 8-CPU Power4 nodes with 16GB of RAM for each group median data. Therefore it is important to seek all possible alternatives to shorten the computation complexity.

Chapter 5

Conclusions and Future Work

In this chapter, we will discuss several possible future directions of research that might provide the next steps to advance the Bayesian factor model in fMRI data analysis.

The goal of this dissertation was to develop a BFM for fMRI data. In Chapter 1, I introduced the mechanism of fMRI in studying human brain function. Statistical analysis can help researchers to robustly and accurately detect activated regions in the fMRI data. We think the BFM is an efficient approach to better understand the BOLD response. In Chapter 2, I described a general factor model and all concepts which are used in the following chapters. In Chapter 3, the main focus of this dissertation has been performed for analyzing and estimating fMRI data by the BFM. Experiments explored BOLD effect differences in different study groups using the BFM. Finally, in Chapter 4, the influence of the stimulus in BOLD response was measured by the SBFM.

5.1 Future Work

5.1.1 Number of Factors

We could fit the BFM with the number of factors that explains more (e.g. 98%) of the variance of the data, then fit a SBFM to explain the random effect part. The resultant number of factors is still noticeably smaller than the original dimension of the data. There might be several issues with this approach. First, the bigger number of factors for the BFM, the better fit result will be achieved for the model. However, the issue of over-fitting might occur, we still need to choose the proper number of factors by model selection. I prefer AIC and BIC over RJMCMC or Birth-and-Death MCMC because of computation time.

Secondly, not all voxels significantly respond to the stimulus, i.e. b_i is not always significant. We have to decide how to fit the SBFM for only a subset of voxels instead of all the voxels.

5.1.2 Appropriate Time-Delay

Up to now it is assumed that the hemodynamic response function parametric form is fixed. It is no easy task to estimate these parameters, and they are not typically known to us given the empirical data, therefore it is worth the effort to try various methods to estimate these parameters (Mumford and Poldrack, 2007). For example, the fMRI responses are modeled as a temporal change of HRF. The change is estimated at every voxel. We can use this to compare delays for different stimuli at the same voxel, or for the same stimulus at different voxels.

Liao et al. (2002) treat the time-delay as a variational value (compared to a fixed value, the time-delay would vary under certain conditions). First, create a parametric family of the hemodynamic response function by starting with a indication HRF $h_0(t)$, which should be similar to the true HRF, for instance the Gamma difference HRF (Sarty, 2006). Then propose varying the delay by changing the origin HRF by δ (Friston et al., 1998),

$$h(t;\delta) = h_0(t-\delta) \tag{5.1.1}$$

and how to estimate δ is the problem. For practical goals, the delay d should be estimated by the time to the first peak of the HRF, that is to say, $d = d_0 + \delta$, where d_0 is the time to the first peak of the indication HRF. Assume that the varied HRF $h(t;\delta)$ is established by linear combinations of two functions $h_0(t)$ and

$$h_1(t) = \frac{\partial h(t;\delta)}{\partial \delta}|_{\delta=0} = -\dot{h}_0(t) \tag{5.1.2}$$

where dot denotes derivative with reference to time. If the time scale change δ is small, that is, the true HRF is similar to the indication HRF, then since a Taylor series of the first two

terms, i.e. $\delta = 0$, $h(t;\delta)$ is estimated: $h(t;\delta) \approx h_0(t) + h_1(t)\delta$. The stimulus with $h_0(t)$ and $h_1(t)$ could be convolved and estimated by least squares.

In our model, the flexible d can solve the problem that different stimuli might have different delays. We can think of the HRF as a linear model: Set the range of δ, and change the time-delay d. The remaining question is how to choose the range of δ and whether linear model is the best model to estimate HRF.

5.1.3 Neighbor Correlation

The correlations between neighboring voxels can be strong in fMRI data. We could cluster neighbors to reduce the dimension and save computation time (Baune et al., 1999). In this work, we choose to use the normalized cut to group the neighboring voxels into homogeneous segments. However, there are many alternative approaches in the literature, e.g. clustering-based image segmentation methods, for example, K-means by MacQueen (1967), fuzzy C-means algorithm by Bezdek et al. (1987) and watershed transformation algorithm by Couprie and Bertrand (1997).

Proper clustering algorithms can reveal structures in the fMRI data based on similarities defined by the chosen distance measure, and the voxels in the same cluster exhibit similar signal strength. The set of cluster centers should be representative of the various structures in the fMRI data. The process is to choose the center voxel, calculate the distance of neighbor voxels to choose the qualified neighbor voxels, then detect the correlation between the center voxel and neighboring voxels. If the correlation is significant, cluster them, if not, go to the next one. After clustering, we can refit the BFM.

5.1.4 Others

In the semiparametric Bayesian factor analysis, we assume the activation stays constant over the task time. That means, without considering the circumstances of the time-delay, if on-off stimulus $x = 1$ for 12 repetitions, then observations will stay at a high activation level for 12 repetitions. But in actually, the stimulus isn't exactly synchronized with the activation, which may lag behind a bit. Furthermore, activations are strong at the beginning of the stimulus, then typically decrease because of saturation of response. We can build a model to account for these effects.

The computation time is too long, especially for the semiparametric model. Due to the huge amount of data, MCMC is very time-consuming. Perhaps we should consider other ways to replace MCMC method. One alternative method is the Kalman filter and smoother (Mardia et al., 1998), which products an effective computational model to appraise the process state, one way to look at it is that minimizes the mean squared error. Under the assumptions that the error terms are normal, it also estimates the posterior mean together with the standard deviations and it is computationally more effective. The model has to be transformed into state space form,

$$y_{ti} = \upsilon_t \alpha_{ti} + \epsilon_{ti}, \epsilon_{ti} \sim N(0, \sigma_i^2) \tag{5.1.3}$$

$$\alpha_{t,i} = M\alpha_{t-1,i} + \zeta_{ti}, \zeta_{ti} \sim N(0, Q) \tag{5.1.4}$$

where $\upsilon_t = (\beta, 0, z_{t,i}, 0)$, $\alpha_{ti} = (f_t, f_{t-1}, b_i, b_{i-1})'$. For the task, we need to define M and Q and estimate the posterior of α_{ti} and υ_t.

5.2 Conclusions

fMRI is a powerful technique to measure brain activity by blood oxygen level changes. Due to the complicated data and various noise interference, the analysis of fMRI data is a very challenging problem.

In this dissertation, we applied Bayesian Factor Model as well as semiparametric Bayesian Factor Model to analyze fMRI data, which has never been done previously. By using Bayesian factor model, we can identify groups of variables (factors), to see how they are related to each other. This will help us to understand differences among activated regions, as a result of a particular cognitive function. Consequently, we could potentially identify hidden dimensions of the data, which may not be apparent from direct analysis. From various experiments that we performed on a single fMRI subject as well as a group of fMRI subjects, it is evident that BFM indeed provides a reliable solution that can help people to explore and better understand the relationship between various factors and activation region of BOLD signal. BFM and SBFM could also potentially help people to better understand various brain diseases by comparing those different factors for control and diseased group. Furthermore, we believe that the BFM could be applied to other types of medical images other than fMRI, and it can help to discover hidden patterns contained in the data.

Another contribution of the dissertation is that we proposed to incorporate a Normalized-cut based segmentation scheme into the BFM to improve the efficiency of the algorithm while preserving most of the information contents within the image data.

Experimental results confirm the following findings that we set out to explore.

1. BFM provides a powerful approach to understand BOLD response. Several factors have been determined to explain most of variance within the data. It is demonstrated that BFM successfully associates the activated activity with the Bayesian factors.

2. SBFM has been used to measure stimulus partial information and improve the inference performance. Compared with BFM, the added hemodynamic response function (HRF) indeed improves the model performance.

3. Normalized Cut is an efficient way to group image voxels together without losing any image information. The significant time saving of using Ncut was demonstrated in this dissertation.

Bibliography

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics*, 18:338–357.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Argall, B., Saad, Z., and Beauchamp, M. (2006). Simplified intersubject averaging on the cortical surface using SUMA. *Human Brain Mapping*, 27:14–27.

Baune, A., Sommer, F. T., Erb, M., Wildgruber, D., Kardatzki, B., Palm, G., and Grodd, W. (1999). Dynamical cluster analysis of cortical fMRI activation. *NeuroImage*, 9:477–489.

Berry, B. J. L. (1960). An inductive approach to the regionalization of economic development. *Regional Science*, 6:83–84.

Bezdek, J., Hathaway, R., Sobin, M., and Tucker, W. (1987). Convergence theory for fuzzy c-means: Counterexamples and repairs. *IEEE Transactions on Systems, Man and Cybernetics*, 17:873–877.

Bickel, P. and Zhang, p. (1992). Variable selection in nonparametric regression with categorical covariates. *Journal of the American Statistical Association*, 87:90–97.

Blink, E. J. (2004). An easy introduction to basic MRI physics. `http://mri-physics.net/textuk.html`.

Brooks, S. (1998). Quantitative convergence diagnosis for MCMC via CUSUMS. *Statistics and Computing*, 8:267–274.

Brooks, S. and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455.

Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research*. Chapman and Hall.

Camchong, J., Dyckman, K., Austin, B., Clementz, B., and McDowell, J. (2008). Common neural circuitry supporting volitional saccades and its disruption in schizophrenia patients and relatives. *Biological Psychiatry*, 64:1042–1050.

Carota, C. and Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. *Bayesian Statistics*, 5:507–511.

Christensen, W. and Amemiya, Y. (2003). Modeling and prediction for multivariate spatial factor analysis. *Journal of Statistical Planning and Inference*, 115:543–546.

Ciuciu, P., Poline, J., Marrelec, J., Idier, J., Pallier, C., and Benali, H. (2003). Unsupervised robust nonparametric estimation of the hemodynamic response function for any fMRI experiment. *IEEE Transactions on Medical Imaging*, 22:1235– 1251.

Couprie, M. and Bertrand, G. (1997). Topological grayscale watershed transformation. In *in SPIE Vision Geometry V Proceedings, 3168*, pages 136–146.

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.

D'Agostino, R. and Stephens, M. (1986). *Goodness-of-Fit Techniques*. CRC Press.

Damadian, R., Goldsmith, M., and Minkoff, L. (1977). NMR in cancer: XVI. FONAR image of the live human body. *Physiological Chemistry and Physics*, 9:97–100.

Devlin, H. (2005). Introduction to fMRI. `http://www.fmrib.ox.ac.uk/research/introduction-to-fmri`.

Dey, D., MüIler, P., and Sinha, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics (Lecture Notes in Statistics)*. Springer.

Fokoué, E. (2004). Stochastic determination of the intrinsic structure in Bayesian factor analysis. Technical Report No.17, Statistical and Applied Mathematical Sciences Institute.

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., and Frackowiak, R. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210.

Friston, K., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping*, 1:153–171.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Ruggb, M. D., and Turner, R. (1998). Event-related fMRI: Characterizing differential responses. *NeuroImage*, 7:30–40.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC Interdisciplinary Statistics.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*, pages 169–193. Oxford University Press.

Ghahramani, Z. and Hinton, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical report. Department of Computer Science, University of Toronto, Toronto, Canada.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC Interdisciplinary Statistics.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9:416–429.

Gorsuch, R. L. (1983). *Factor Analysis*. Lawrence Erlbaum Associates.

Gössl, C., Auer, D., and Fahrmeir, L. (2000). Dynamic models in fMRI. *Magnetic Resonance in Medicine*, 43:72–81.

Gössl, C., Auer, D., and Fahrmeir, L. (2001a). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57:554–562.

Gössl, C., Fahrmeir, L., and Auer, D. (2001b). Bayesian modeling of the hemodynamic response function in BOLD fMRI. *NeuroImage*, 14:140–148.

Goutte, C., Nielsen, F., and Hansen, L. K. (2000). Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE Transactions on Medical Imaging*, 19:1188–1188.

Green, P. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., and Haase, A. (2002). Generalized autocalibrating partially parallelf acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, 47:1202–1210.

Haacke, E. M., Brown, R. W., Thompson, M. R., and Venkatesan, R. (1999). *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Wiley, John and Sons, Incorporated.

Harrington, D. (2008). *Confirmatory Factor Analysis*. Oxford University Press.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57:97–109.

Hendee, W. and Morgan, C. (1984). Magnetic resonance imaging Part I-Physical principles. *The Western Journal of Medicine*, 141:491–500.

Hoogland, J. J. and Boomsma, A. (1998). Robustness studies in in covariance structure modeling: An overview and meta-analysis. *Sociological Methods and Research*, 26:329–367.

Horn, R. A. and Johnson, C. A. (1985). *Matrix Analysis*. Cambridge University Press.

Howson, C. and Urbach, P. (2005). *Scientific Reasoning: The Bayesian Approach*. Open Court, 3rd edition.

Jöreskog, K. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 59:381–389.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34:183–202.

Katanoda, K., Matsuda, Y., and Sugishita, M. (2002). A spatio-temporal regression model for the analysis of functional MRI data. *NeuroImage*, 17:1415–1428.

Kaufman, G. and Press, S. (1973). Bayesian factor analysis. `http://dspace.mit.edu/bitstream/1721.1/1870/1/SWP-0662-14514648.pdf`.

Langers, D. (2009). Blind source separation of fMRI data by means of factor analytic transformations. *NeuroImage*, 47:77–87.

Lauterbur, P. C. (1973). Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*, 242:190–191.

Lawley, D. N. and Maxwell, A. E. (1962). Factor analysis as a statistical method. *The Statistician*, 12:209–229.

Li, Y.-O., Mukherjee, P., Nagarajan, S., and Attias, H. (2010). A novel variational Bayesian method for spatiotemporal decomposition of resting-state fMRI. Technical report.

Liao, C. H., Worsley, K. J., Poline, J. B., Aston, J. A. D., Duncan, G. H., and Evans, A. C. (2002). Estimating the delay of the fMRI response. *NeuroImage*, 16:593–606.

Liao, S., Jia, H., Wu, G., and Shen, D. (2012). A novel framework for longitudinal atlas construction with groupwise registration of subject image sequences. *NeuroImage*, 59:1275–1289.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96:1045–1056.

Lopes, H. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67.

Machado, A., Gee, J., and Campos, M. (1999). *Exploratory Factor Analysis in Morphometry*. Lecture Notes in Computer Science. Springer.

MacQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Maes, F., Collignon, A., V, D., Marchal, G., and Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16:187–198.

Mardia, K. V., Goodall, C., Redfern, E. J., and Alonso, F. J. (1998). The kriged Kalman filter. *Test*, 7:217–282.

Marrelec, G., Ciuciu, P., Pelegrini-Issac, M., and Benali, H. (2003). Estimation of the hemodynamic response function in event-related functional MRI: Directed acyclic graphs for a general Bayesian inference framework. *Information Processing in Medical Imaging*, LNCS 2732:635–646.

Mather, P. M. (2004). *Computer Processing of Remotely Sensed Images: An Introduction*. John Wiley and Sons.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341.

Mumford, J. A. and Poldrack, R. A. (2007). Modeling group fMRI data. *Social Cognitive and Affective Neuroscience*, 2:251–257.

Novelline, R. (1997). *Squire's Fundamentals of Radiology*. Harvard University Press.

oorsuizen.be (1996). Neurosurgically treatable causes of tinnitus. Types of tinnitus and their treatment.

Pekar, J. J. (2006). A brief introduction to functional MRI. *IEEE Engineering in Medicine and Biology Magazine*, 6:24–25.

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24:350–362.

Pettigrew, J. D., Wallman, J., and Wildsoet, C. F. (1990). Saccadic oscillations facilitate ocular perfusion from the avian pecten. *Nature*, 343:362–363.

Press, S. and Shigemasu, K. (1997). Bayesian inference in factor analysis-revised. Technical Report No. 243, Department of Statistics, University of California, Riverside.

Press, S. J. (2002). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. Wiley-Interscience.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Raftery, A. E. and Lewis, S. (1992). How many iterations in the Gibbs sampler. *In Bayesian Statistics*, 4:763–773.

Raftery, A. E. and Lewis, S. M. (1996). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In *In Practical Markov Chain Monte Carlo*, pages 115–130. Chapman and Hall.

Reyment, R. and Jöreskog, K. (1996). *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, 47:69–76.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

Sarty, G. E. (2006). *Computing Brain Activity Maps from fMRI Time-Series Images*. Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sherman, J. and Morrison, W. J. (1949). Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix (abstract). *Annals of Mathematical Statistics*, 20:621.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 21:124–127.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905.

Smith, S. M. (2004). Overview of fMRI analysis. *The British Journal of Radiology*, 77:167–175.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components, an alternative to reversible jump methods. *Annals of Statistics*, 28:40–74.

Subbarao, C., Subbarao, N., and Chandu, S. (1995). Characterisation of groundwater contamination using factor analysis. *Environmental Geology*, 28:175–180.

Teh, Y., Seeger, M., and Jordan, M. (2005). Semiparametric Latent Factor Models. In *Workshop on Artificial Intelligence and Statistics 10*.

Tepper, M., Musé, P., Almansa, A., and Mejail, M. (2011). Automatically finding clusters in normalized cuts. *Pattern Recognition*, 44:1372–1386.

van der Zwaaga, W., Francisa, S., Heada, K., Petersa, A., Gowlanda, P., Morrisa, P., and Bowtell, R. (2009). fMRI at 1.5, 3 and 7 T: Characterising BOLD signal changes. *NeuroImage*, 47:1425–1434.

Viola, P. and Wells, III, W. M. (1997). Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24:137–154.

Walker, J. and Maddan, S. (2008). *Statistics in Criminology and Criminal Justice: Analysis and Interpretation*. Jones and Bartlett Learning.

Wang, F. and Wall, M. M. (2003). Generalized common spatial factor model. *Biostatistics*, 4:569–582.

Ward, B. D. (2000). Simultaneous inference for fMRI data. `http://afni.nimh.nih.gov/pub/dist/doc/manual/AlphaSim.pdf`. AFNI 3dDeconvolve Documentation, Medical College of Wisconsin, Milwaukee, WI.

Woolrich, M. W., Behrens, T. E. J., and Smith, S. M. (2004). Constrained linear basis sets for HRF modelling using variational Bayes. *NeuroImage*, 21:1748–1761.

Wu, Z. and Leahy, R. (1993). An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:1101–1113.

Xu, L. (2007). *Bayesian Spatial Modeling of fMRI Data: A Multiple Subject Analysis*. PhD thesis, Department of Biostatistics, the University of Michigan.

Yu, B. and Mykland, P. (1998). Looking at Markov samplers through CUSUM path plots: A simple diagnostic idea. *Statistics and Computing*, 8:275–286.

Zhang, D. and Shen, D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, 59:895–907.

Zhang, P. (1993). On the convergence rate of model selection criteria. *Communications in Statistics: Theory and Methods*, 22:2765–2775.