Comparison of Data-Driven Analysis Methods for Identification of Functional Connectivity in fMRI

by

Yongwook Bryce Kim

Sc.B. Mathematics and Physics
Brown University, 2005

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2008

Author .....
Department of Electrical Engineering and Computer Science
January 29, 2008

Certified by .....
Polina Golland
Assistant Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Comparison of Data-Driven Analysis Methods for Identification of Functional Connectivity in fMRI

by

Yongwook Bryce Kim

Submitted to the Department of Electrical Engineering and Computer Science
on January 29, 2008, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Data-driven analysis methods, such as independent component analysis (ICA) and clustering, have found a fruitful application in the analysis of functional magnetic resonance imaging (fMRI) data for identifying functionally connected brain networks. Unlike the traditional regression-based hypothesis-driven analysis methods, the principal advantage of data-driven methods is their applicability to experimental paradigms in the absence of a priori model of brain activity. Although ICA and clustering rely on very different assumptions on the underlying distributions, they produce surprisingly similar results for signals with large variation. The main goal of this thesis is to understand the factors that contribute to the differences in the identification of functional connectivity based on ICA and a more general version of clustering, Gaussian mixture model (GMM), and their relations. We provide a detailed empirical comparison of ICA and clustering based on GMM. We introduce a component-wise matching and comparison scheme of resulting ICA and GMM components based on their correlations. We apply this scheme to the synthetic fMRI data and investigate the influence of noise and length of time course on the performance of ICA and GMM, comparing with ground truth and with each other. For the real fMRI data, we propose a method of choosing a threshold to determine which of resulting components are meaningful to compare using the cumulative distribution function of their empirical correlations. In addition, we present an alternate method to model selection for selecting the optimal total number of components for ICA and GMM using the task-related and contrast functions. For extracting task-related components, we find that GMM outperforms ICA when the total number of components are less then ten and the performance between ICA and GMM is almost identical for larger numbers of the total components. Furthermore, we observe that about a third of the components of each model are meaningful to be compared to the components of the other.

Thesis Supervisor: Polina Golland
Title: Assistant Professor of Computer Science and Engineering

Acknowledgments

My time as a MIT student has been a challenging, but enjoyable one. There were many rough times, but I have grown up personally and intellectually from such experiences. I am grateful that I have had great opportunities to learn the importance of staying healthy, being open to differences, and being happy with myself in this stimulating environment. I can never thank enough the people around me at MIT whom I am greatly indebted.

First of all, I would like to express my sincere gratitude to my research advisor Professor Polina Golland for her unwavering support and encouragement, helping me stay focused, and teaching me how to think critically and carefully. I appreciate her enormous understanding and patience, especially regarding my health conditions. I will always remember her invaluable advice, openness, and passion. I'd also like to thank Professor John Gabrieli for the opportunity to work closely with his group in BCS and Professor Nancy Kanwisher for providing the data used in this thesis.

The members of the vision group have been always great to work with. Thanks to Ray and Thomas for helping my transition into the group. Thanks to Wanmei, Danial, and Sune for heated research discussions, listening to my ideas, and providing helpful feedbacks. I appreciate the members of the EECS KGSA for providing the home away from home. Thanks to Junghoon, Taegsang, and Sejoon for fruitful conversations. My special appreciation goes to Bo. Thanks for sharing with me your valuable time and dreams.

I thank the Boston Red Sox for providing me one of the greatest distractions. It was worth it, again.

And finally, as always, I am deeply grateful to my family for their endless encouragement and love. Thanks mom and dad for unconditional support and trust. Thanks Yongjoo for being the best brother and life-time friend. We often find ourselves in different continents in the world, but you are always in my heart.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since its development in the early 1990s, functional magnetic resonance imaging (fMRI) has played a tremendous role in visualizing human brain activity for the study of mechanisms of human brains and clinical practice. Acquired by a regular magnetic resonance machine with special parameter settings, this non-invasive imaging method measures changes in blood flow, which in turn are an indication of neural activity. fMRI produces four-dimensional time-series images (three-dimensional in space and one-dimensional in time) with relatively low temporal resolution and high spatial resolution.

This new imaging technique has generated a large volume of new high dimensional data and hence, the need for new image analysis methods. In the recent literature, interesting discoveries in human cognitive states have been found using techniques from machine learning, especially multivariate pattern analysis and non-linear pattern classification methods [20, 34, 55]. Much of the work in fMRI data analysis has revolved around the detection of activation at different locations. In addition to localizing activity, we are also interested in how the "activated" areas are related and connected to each other. Functional connectivity, the central theme of this thesis, characterizes these functional interactions and coordinated activations among different parts of the brain.

Traditionally, the regression based, hypothesis-driven approach has been used to detect functional connectivity. Taking this approach, a "seed" region of interest must

be first selected by the user. The network is defined as the areas whose correlation with the seed time course exceeds a pre-defined threshold. This method can work well when the goal is to identify regions that co-activate with a certain part of the brain. Hypothesis-driven methods require prior information on the protocol and hypothesis of an experiment to model the expected hemodynamic response. Moreover, the correlation threshold, directly related to the statistical significance level, must be selected.

Recently, there has been an increasing number of fMRI experiments that investigate the brain activity in a more natural, near protocol-free setting, such as responding to audio-visual input like a movie or rest state scanning [10, 11, 17, 30, 65]. Unlike traditional protocol-based experiments, these new complex experiments do not contain a well-defined onset protocol. Although the traditional seed-based connectivity analysis can be applied to these data, paradigm-free, data-driven exploratory methods such as principal component analysis (PCA) [26], independent component analysis (ICA) [49], and clustering algorithms [30, 32] such as Gaussian mixture model can naturally provide an alternative to comparing each voxel's time course against a hypothesis. They explore the data to find "interesting" components or underlying sources. Structures or patterns in the data, which are difficult to identify a priori, such as unexpected activation and connection, motion related artifacts, and drifts, may be revealed by these components. However, the direct relationship among the data-driven methods is largely unknown and the performance in correctly detecting and classifying functionally connected regions depends on various theoretical and experimental factors.

Recently, several works have compared data-driven analysis methods in the context of functional connectivity. Baumgartner *et al.* [6] use artificially generated activations and show that fuzzy clustering analysis (FCA) outperforms principal component analysis in a noisy data setting by comparing the maximum Pearson correlation coefficient between the simulated activation time course and the representative time courses obtained by FCA and PCA. The superior performance of ICA compared to the regression-based cross-correlation analysis (CCA) in detecting functional connec-

tivity in the resting brain and the effect of seed selection on CCA results are presented in [47]. Meyer-Baese *et al.* [50] present comparative results of several variations of clustering and ICA algorithms by evaluating their performances using task-related activation maps and associated time courses with respect to the experimental protocol in a simple block design fMRI experiment and by conducting receiver operating characteristic analysis. They show a close agreement between clustering and ICA, but also conclude that despite a longer processing time, clustering outperforms ICA in terms of the classification results of the task-related activation. Smolders *et al.* [58] compare results of fuzzy clustering and ICA in terms of within- and between-subject consistency and spatial and temporal correspondence of obtained maps and time courses. They demonstrate a good agreement between FCA and spatial ICA in discriminating the contribution of distinct networks of brain regions to the main cognitive stages of the task (auditory perception, mental imagery and behavioural response). They claim that whereas ICA works optimally on the original time series, averaging with respect to the task onset (and thus introducing some a priori information on the experimental protocol) is essential in the case of FCA leading to a richer decomposition of the spatio-temporal patterns of activation. However, for all of these studies, their comparison scheme was only based on the similarity of the task related component detected by the methods to a predefined reference waveform and disregarded all other components.

Exploratory data analysis methods have also been compared in other areas of medical image analysis and computer vision. Jung *et al.* [41] show the advantages of ICA over PCA in removing electroencephalographic (EEG) artifacts. In the context of face recognition, the literature on the subject is contradictory. Bartlett *et al.* [5], Liu and Wechsler [45] claim that ICA outperforms PCA for face recognition, while Baek *et al.* [3] report a contrary result that PCA outperforms ICA when tested on the FERET database. Delac *et al.* [21] and Draper *et al.* [23] conclude that the performance of methods (PCA, ICA, and Linear Discriminant Analysis) largely depends on a particular task of face recognition such as subject identification and expression recognition and that one method cannot be claimed to perform better

than others in general cases.

Although ICA and clustering rely on very different assumptions on the underlying distributions, they produce surprisingly similar results for signals with large variation. The main goal of this thesis is to understand the factors that contribute to the differences in the identification of functional connectivity based on ICA and a more general version of clustering, Gaussian mixture model (GMM), and their relations. Using the synthetic data with artificial activations and artifacts generated by the generative model of ICA under two experimental conditions (length of the time course and signal-to-noise ratio (SNR) of the data), both spatial maps and their associated time courses estimated by ICA and GMM are compared to each other and to the ground truth. The number of components are chosen via the model selection scheme and all selected components are compared, not just the task-related components. This comparison scheme is verified in a real fMRI study.

This work provides a detailed comparison of ICA and clustering based on Gaussian mixture model, both in terms of generative models and experimental conditions. Contributions of this thesis are as follows.

- We devised a component-wise matching and comparison scheme of resulting ICA and GMM components using their correlations.
- We applied this scheme to the synthetic data and investigated the influence of noise and length of time course on the performance of ICA and GMM, comparing with ground truth and with each other.
- We developed a method of choosing a threshold to determine which of resulting components are meaningful to compare using the cumulative distribution function of their empirical correlations.
- We proposed an alternate method of selecting the optimal total number of components for ICA and GMM using the task-related and contrast functions.
- We applied our methods to real fMRI data in visual recognition experiments.

With ever increasing volume of complex experimental fMRI data, we believe that our work will provide a better understanding of the functional brain networks and a direction for further analysis.

The rest of the thesis is organized as follows. In Chapter 2, we review the basic properties of fMRI, typical fMRI experiment set-ups, pre-processing steps, and the sources of noise in fMRI data. In Chapter 3, we compare three definition (anatomical, functional, effective) of brain connectivity. We also define and explain the notion of functional connectivity. In addition, we discuss previous work on this topic and the standard hypothesis-driven connectivity analysis method. In Chapter 4, we describe the generative models and the algorithms for the three data-driven connectivity models of our interest, PCA, ICA, and GMM and model selection methods. Chapter 5 introduces the component-wise comparison scheme between ICA and GMM. Furthermore, we present the results of investigating the differences of performance of the analysis methods using synthetic and real fMRI data in Chapter 5. We conclude with discussion of future research directions in Chapter 6.

Chapter 2

Functional Magnetic Resonance Imaging

2.1 Overview

Functional magnetic resonance imaging (fMRI) is a recently developed neuroimaging modality that provides an opportunity to study functional human brain activity in a non-invasive way. MRI uses strong magnetic fields to create images of biological tissue. To generate images, an MRI scanner applies a series of changing magnetic gradients and oscillating electromagnetic fields, known as a pulse sequence. By varying this pulse sequence, a particular tissue type of interest (e.g. gray and white matter, tumors, bone damage) can be detected by the scanner. Functional neuroimaging aims to localize different mental processes to different parts of the brain, in effect creating a map of which areas are responsible for which processes. Since the early 1990s, the development of fMRI has catalyzed an explosion of interest in functional neuroimaging and has become a powerful tool in research and clinical applications.

Unlike structural MRI, which measures differences between tissues, fMRI measures signal changes in the brain that are due to changing neural activity. The most popular approach is the fMRI based on blood oxygenation level dependent (BOLD) signal changes, which allows assessment of brain activity via local hemodynamic variations over time [51, 64]. The basic assumption is that increased neural activity

induces an increased demand for oxygen and, in turn, the vascular system increases the amount of oxygenated hemoglobin relative to deoxygenated hemoglobin. Because deoxygenated hemoglobin attenuates the MR signal, which causes a change in the MR decay parameter T_2^*, the vascular response leads to a signal increase that is related to the neural activity. This process is known as hemodynamic response (HDR). In a typical fMRI experiment, external stimuli are presented at intervals of several seconds, causing a change in voxel-signal intensity, delayed and blurred by the hemodynamic response lag. From these changes, researchers can make inferences about the underlying neural activity and how different brain regions may participate in different perceptual, motor, or cognitive processes. However, the precise nature of the relationship between neural activation and the BOLD signal is a subject of current research and is yet to be well understood. Because changes in blood oxygenation occur intrinsically as part of normal brain physiology, fMRI is a non-invasive technique that can be repeated on the same subject as many times as needed.

fMRI provides one of the optimal combined spatial and temporal resolution methods presently available for non-invasive functional brain mapping. Typically, it generates voxels with a spatial resolution of 2 to 5 mm and a temporal resolution of few seconds. However, one of the main drawbacks of fMRI is the relatively low image signal-to-noise ratio (SNR), which is the magnitude of the signal change due to experimental condition divided by the variability in the measurements, depending on both the amount and variability of signal change. Along with other factors such as artifacts, head movement, and undesired physiological sources of variability, this makes detection of the activation-related signal changes a difficult task.

Despite its limitations, fMRI has been widely used in many different application domains in psychology, neurobiology, neurology, radiology, biomedical engineering, electrical engineering, physics, and many others. Especially in cognitive neuroscience, due to its adaptability to many types of experimental paradigms, fMRI has shown great utility in researching object processing and recognition, memory, visual attention, language plasticity, and connectivity between brain regions, to name a few. With a better understanding of the BOLD effect and hemodynamic response and more so-

phisticated data acquisition and analysis techniques, fMRI has a great potential to be used even more widely in research and clinical applications.

2.2 fMRI Experimental Protocols

To functionally associate one or more brain regions with a task that a subject performs, one must first devise an experimental design. Simple tasks in fMRI experiments include presentation of sounds and images, whereas more complex experiments involve watching movies and presentation of instructions for memory and recognition tasks, for example. Experimental design is followed by the image acquisition step, in which the subject lies in a MRI scanner performing a task with his head fixed to avoid movement artifact. These acquired images are used to draw a cognitive interpretation via careful statistical analysis. The experimental design is commonly based on a block-design or an event-related design.

In the case of the block design, each condition is presented for an extended time period, and the different conditions are usually alternated over time. Typically, a block design involves alternations of a task-performing block and a rest block, where no stimuli are presented. A block, also referred as an epoch, contains a sequence of several repetitions of stimuli under the same condition. A single condition may include more than one cognitive task. Block design considers all of them as a single task condition. This is the case of our real visual recognition fMRI data, presented in Chapter 5. Due to the large amount of noise is present in fMRI data, the underlying signal, which should follow the periodic activation pattern, is hardly recognizable even when the voxel is taken from a strongly activated region. This low signal-to-noise ratio of fMRI makes detection of any activation difficult with only one realization of a condition. Thus, the fMRI algorithms are based on averaging over several realizations since averaging increases the signal-to-noise ratio. However, the limiting factor in multiple realizations of experimental conditions is the subject's ability to perform identical tasks without moving or getting tired, which introduces motion artifacts and fatigue effects.

An alternative to the block design is the event-related design, which involves a different stimulus structure. Although the block design has an advantage of excellent detection power, the event-related design has the ability to estimate the shape of the hemodynamic response function. Event-related designs present stimuli one at a time rather than together as a block. Such experimental protocols are characterized by rapid, randomized presentation of stimuli. Time between each trial of stimuli is typically jittered. This has the advantage that the subject does not get used to the experiment, which ensures that the HRF does not change its shape or decrease in amplitude. This is necessary to enable averaging over several realizations. Furthermore, different trial types are intermixed so that each trial is statistically independent from other trials. Since it assumes that the HRFs corresponding to various tasks are different, signals can be analyzed by task category. The possibility of post hoc categorization of an event is another advantage of event-related fMRI. It is in general difficult to draw a conclusion which type of experimental design is better. The design which best suits a specific research hypothesis should be chosen.

2.3 Preprocessing

Preprocessing includes all processes that are performed after image reconstruction and prior to the statistical analysis of the data. The two primary goals of preprocessing are to reduce non-task-related variability in experimental data and to improve validity of statistical analysis [36].

Since almost every fMRI scanner acquires the slices of a volume in succession, each slice is obtained at a different time point. Slice timing correction shifts each voxel's time series within a repetition time (TR) so that all voxels in a given volume appear to have been captured at exactly the same time. This is especially important for long TRs, in which the expected hemodynamic response amplitude can vary significantly. Slice timing correction is typically done using temporal interpolation, which uses information from nearby time points to estimate the amplitude of the signal at the onset of the TR. Interpolation strategies include linear, spline, and sinc interpolations.

Another very important preprocessing step is motion correction. fMRI analysis assumes that each voxel represents a unique part of the brain. In case of head motion, each voxel's time course could be acquired from more than one brain location. The effect of head motion on the signal change is significant, especially near the edge of the brain. A movement of one tenth of a voxel may produce 1-2% signal change, which is not negligible, compared to the very small amount of signal change of fMRI BOLD effects [29]. This requires the use of accurate image registration algorithms to spatially align multiple image volumes. The images are transformed by resampling with respect to a reference image, which is often the first acquired image. In case of the rigid body transformation, the transformation parameters (translation, rotation) for the images are determined by optimizing the goodness of fit to the reference image [28].

In order to facilitate comparisons of the results of analyses aross different subjects, the images in the data are normalized according to a template in the standardized space. This process is called spatial normalization. The most commonly adopted coordinate system is that described by Talairach and Tournoux [60]. Although spatial normalization allows generalization of results to larger population and provides a coordinate space for reporting results, matching between subjects is only possible on a coarse scale, since there is not necessarily a one-to-one mapping of the cortical structures between different brains. No such processing was required in our work, since we did not perform the analysis across subjects.

Spatial filtering with a Gaussian smoothing kernel is often applied to increase signal-to-noise ratio in the data. The increase in SNR is achieved by applying a filter which has the same shape and size as the signal. However, the effectiveness of spatial smoothing diminishes if exact signal properties are not known and the size of the smoothing kernel is larger than the activation area. In the temporal domain, applying a high-pass filter suppresses slow, repetitive physiological signals related to the cardiac cycle or to breathing, as well as the scanner-related drifts.

In some studies, a region of interest (ROI) is selected through segmentation, which classifies voxels within an image into different anatomical divisions. It allows direct,

unbiased measurement of activity within an anatomical region, based on the assumption that functional divisions tend to follow anatomical divisions.

2.4 Signal-to-Noise Ratio of fMRI Signal

Although fMRI has been shown useful and is used extensively in neuroscience research, the level of signal changes in fMRI data still remain low (approximately 1-2%). Signal-to-noise (SNR) ratio is one way to quantify the level of signal changes in the fMRI data. SNR is typically defined as the ratio of the variability in the signal to the variability in the noise. We define SNR in the General Linear Model framework [27], which models the brain as a linear time invariant system with an impulse response function reflecting the hemodynamic properties of the brain. We use design matrix $B = [B_1, B_2]$ for linear regression. GLM assumes the signal is a linear combination of a protocol-dependent component, B_1, a protocol-independent component, B_2, such as physiological noise and drifting, and random noise, ϵ. We construct B_1 by convolving the experimental protocol and the assumed hemodynamic response function modelled as a two gamma function [39], defined as

$$h(t) = \left(\frac{t}{d_1}\right)^{a_1} \exp\left(-\frac{t-d_1}{b_1}\right) - c\left(\frac{t}{d_2}\right)^{a_2} \exp\left(-\frac{t-d_2}{b_2}\right) \tag{2.1}$$

where $d_j = a_j b_j$ is the time to the peak and $a_1 = 6, a_2 = 12, b_1 = b_2 = 0.9$s, and $c = 0.35$. The two gamma function correctly captures the small dip after the HRF has returned to zero. Typically, low order polynomials are used to model B_2. For a given time course $\vec{y_i}$, GLM is often formulated as

$$\vec{y_i} = B_1\vec{\beta_{1i}} + B_2\vec{\beta_{2i}} + \vec{\epsilon_i}, \tag{2.2}$$

where $\vec{\beta_i} = [\hat{\beta_{1i}}, \hat{\beta_{2i}}]$ is a vector of estimated amplitudes of the hemodynamic responses and the protocol independent signals at voxel i. Noise $\vec{\epsilon_i} \sim N(0, \Sigma_i)$. The noise covariance Σ_i is unknown. Assuming the noise is white, $\Sigma_i = \sigma_i^2 I$, we can estimate $\vec{\beta}$

using the least square estimate,

$$\hat{\vec{\beta}}_i = (B^T B)^{-1} B^T \vec{y_i}. \tag{2.3}$$

For a given voxel i, we define our estimated SNR as

$$\widehat{SNR}_i = \frac{|B_1 \hat{\beta}_1|^2}{|\vec{y_i} - B_1 \vec{\beta_{1i}}|^2}. \tag{2.4}$$

We use the average of these estimates over our region of interest. The SNR value is subject to the choice of noise measurement. In the definition above, we define noise in the denominator as anything that is not signal. Each region in the brain contains different components of the noise signal. Data acquired outside the brain region is only subject to the noise of the measurement instrument (e.g., the scanner), whereas data within the brain is related to motion-related noise, thermal and respiratory noise from the body, partial volume effects, flow artifacts, and MR spin history errors [52]. The estimated SNR is an optimistic approximation of the true SNR because the signal and the noise overlap in some frequency bands, and thus part of the noise is treated as signal. We use this estimated SNR as an upper bond of the true SNR. Amount of noise presented in the data largely influences effectiveness of data analysis and modeling algorithms. Therefore, a clear connection between SNR of data and performance of an analysis method is crucial in obtaining accurate interpretation of results.

2.5 Summary

In this chapter, we introduced a brief background on fMRI physics, properties of data, and experimental protocols. Preprocessing steps of fMRI data described in the previous section are applied to the real fMRI data, and signal-to-noise ratio is an important property of the data against which we test performance and robustness of our data driven analysis methods in Chapter 5. We now turn our attention to how fMRI data can be used to reveal connectivity in the brain.

Chapter 3

Functional Connectivity

Many fMRI studies aim to discover patterns of brain activity that are associated with phenomena of interest. The patterns of activity are often called neural correlates, to emphasize that changes in the brain vary with changes in an external phenomenon. Most fMRI analysis methods identify whether a given voxel or a region of interest (ROI) shows significant task-related signal changes. Each voxel or a group of voxels is tested for correlation with the protocol, independently from other voxels or groups in such methods. A collection of voxels whose time courses correlate substantially with the experimental task may implicitly represent coactivation, but do not provide any information about the relations or dependencies among the brain regions that those voxels delineate.

Because fMRI data are collected over time and have a temporal structure, several methods utilize the information about the coherence of activity over time to identify functional connectivity, which represents the pattern of functional relations among brain regions, independent of a particular task-induced activation. This class of methods includes cross-correlation [25], partial least squares [66], and data driven methods such as flat [32] and hierarchical clustering [30], principal component analysis [26], multidimensional scaling [26], and independent component analysis [49].

3.1 Brain Connectivity

The organization of the human brain is based on two complimentary principles, which lead to two corresponding approaches in explaining its function [56]. The first approach is functional segregation. The goal here is to localize function to specific brain areas. This approach is based on the principle of modularity, which is specialization of function within different regions of the brain, where local assemblies of neurons in each area perform their unique operations. The second approach is functional integration, which explains function in terms of information flow between brain areas. This approach is based on the principle that functions are emergent properties of interacting brain areas within networks. Functional segregation has been the dominant approach, but segregation itself does not explain the entire brain function. Recently, more works have been focused on the distributed nature of information processing in neuronal networks in the brain, which attempt to explain "transferred and transformed effects within the segregated regions" [56]. This leads to the study of brain connectivity.

Before we look at the relationships between neuronal networks across the brain, we first need to categorize the different types of brain connectivity. Connectivity refers to several interrelated, yet different aspects of brain organization [35, 44]. The basic distinction is that between structural connectivity, functional connectivity, and effective connectivity. Structural connectivity refers to a network of anatomical connections linking sets of neurons or neuronal elements. On the other hand, functional connectivity is fundamentally a statistical concept. It characterizes deviations from statistical independence between distributed and spatially remote neuronal units. Statistical dependence can be estimated by measuring correlation or covariance [26] or spectral coherence [61]. Functional connectivity often looks for temporal correlations between all neurophysiological events in a system, regardless of the anatomical routes through which such influences are exerted. Furthermore, it does not make any explicit reference to specific causal effects between events. Effective connectivity describes networks of causal effects of one neural element over another in the context

of a particular anatomical model that specifies such routes a priori. Thus, it is often viewed as the intersection of structural and functional connectivity. Such casual effects are inferred through a causal model, which includes structural parameters, and regions and connections of interest are specified by the researcher [54].

3.2 Standard Regression-based Hypothesis-driven Method for Detecting Functional Connectivity

Since fMRI studies rely on the detection of a weak signal in the presence of substantial noise, careful statistical analysis is necessary. As briefly discussed above, the regression based approach has been traditionally applied to detect functional connectivity, especially in early studies of fMRI [4, 10]. Typically, a "seed" region is selected as the first step. It is often a particular area of interest in the brain that we want to find connectivity to, or a group of regions whose time courses exhibit most resemblance to the protocol of an experiment (e.g box-car waveform). Then, functionally connected network is defined as the areas whose correlation with the seed time course exceeds a pre-defined threshold. For a time course $\mathbf{t}$ and a reference waveform $\mathbf{s}$ of the seed region, the correlation coefficient is calculated as

$$r = \frac{\sum(\mathbf{t} - \bar{\mathbf{t}})(\mathbf{s} - \bar{\mathbf{s}})}{\sqrt{(\mathbf{t} - \bar{\mathbf{t}})^2(\mathbf{s} - \bar{\mathbf{s}})^2}}, \tag{3.1}$$

where $\bar{\mathbf{t}}$ and $\bar{\mathbf{s}}$ are the means of the individual time course and the reference waveform, respectively. r has a value of 1 for perfect correlation, a value of zero for no correlation (corresponding to the null hypothesis), and a value of -1 for perfect anti-correlation.

The basic idea is very similar to that of a simple hypothesis testing, where the result is declared as significant if the data sample is unlikely to have occurred under the null hypothesis. An experimental hypothesis represents a prediction about the data or an active voxel, whereas a null hypothesis is based on random chance, corresponding to an assumption that the mean of correlation coefficients between the signals of the seed region and activated areas is same as that with non-activated areas. Therefore,

the standard regression based method is also known as hypothesis-driven analysis.

Hypothesis-driven analysis has two main characteristics. First, this method requires a prior knowledge about the choice of the seed region or an external reference function (not necessarily from within the brain), which often requires information on the protocol of an experiment. Although it is difficult to obtain an exact event timing in more complex experiments, the experimental protocol is pre-defined in a vast majority of fMRI experiments, and hypothesis-driven analysis such as t-test or correlation analysis can be applied. The second characteristic is an choice of the correlation threshold, which is directly related to the significance level for the values of correlation. Obtaining a meaningful correlation coefficient depends on having maximal variability in the signal of interest, compared to experimental noise, and the number of time samples used. The choices of seed regions and threshold values should be carefully compared, especially when group analysis across subjects is performed.

On the other hand, exploratory data analysis methods, such as principal component analysis [26], independent analysis [49], and clustering [32], do not require a pre-determined choice of a seed region. Instead, they discover the interesting seed regions and their associated networks and time courses in an unsupervised way. These methods will be discussed in depth in Chapter 4.

3.3 Prior Work

An increasing amount of attention has been recently paid to the conditions of the human brain at rest and correlations in brain activity during a deactivated state in fMRI studies. Functional connectivity in the motor cortex of resting human brain was demonstrated by Biswal and his group in 1995 [10]. Using echo-planar image pulse sequence with a time resolution of 250ms to rapidly sample a single slice within the brain, they measured fMRI activity in the sensorimotor cortex during a rest condition. Voxels that are "functionally related" were determined by the standard regression-based cross-correlation analysis, which identified voxels whose BOLD activity time courses were significantly correlated with each other despite the subject

not performing any motor task. The seed region, which in this case was the region with time courses of low frequency fluctuations (<0.1 Hz) compared to the MR signal intensity fluctuations of about 2% in the resting brain, and the correlation threshold had to be predetermined before the correlation analysis. Thus, the resulting map presented functional connectivity with that seed region. The authors concluded that correlation of low frequency fluctuations, which may arise from fluctuations in blood oxygenation or flow and are not associated with system noise or cardiac or respiratory peaks, is a demonstration of functional connectivity in the brain. This study was followed by other groups' studies that revealed evidence of connectivity between additional functional areas of the brain, such as the somatosensory and visual cortices [17, 46, 53].

Functional connectivity during the resting state was also measured by independent component analysis in [65]. It was demonstrated that spatial ICA yielded connectivity maps of bilateral auditory, motor and visual cortices, which in part confirmed Biswal's result. In addition, it showed that prefrontal and parietal areas are also functionally connected within and between hemispheres during the resting state. The authors claimed that these connectivity maps obtained by ICA showed an extremely high degree of consistency in spatial, temporal, and frequency parameters within and between subjects. Several other applications of ICA in resting state fMRI data showed similar results that ICA is capable of detecting functional networks beyond the primary (motor, visual, and somatosensory) brain regions [7, 42, 47, 67].

Calhoun *et al.* used ICA to decompose activation patterns into interpretable components during a simulated driving test, which simultaneously engages multiple cognitive elements, such as error monitoring and inhibition and perceiving driving speed [11]. In addition, they also applied ICA to clinical research. In [14], they found that the use of coherent brain networks such as the temporal lobe and default mode networks provides a more reliable measure of disease state than task-correlated fMRI activity, when the goal is to discriminate subjects with bipolar disorder, chronic schizophrenia, and healthy controls.

Another pioneering work in functional connectivity was done by Friston *et al.*

[26]. They defined time-series functional connectivity as temporal correlations between spatially remote neurophysiological events. They modeled a connected brain system as a pattern of activity in terms of correlations or covariance, and used principal component analysis to demonstrate the connectivity during a verbal test. This method is explained in more detail in Section 4.1.

Clustering also has been applied to detect functionally connected networks. For example, in [17], Cordes *et al.* used hierarchical clustering [32] in resting data and found clusters of neighboring voxels whose activity was highly correlated at low frequencies, which suggested functional connectivity similar to that of Biswal [10]. Similarly, Peltier *et al.* [53] classified the low frequency resting state functional connectivity using a self-organizing map (SOP) [43]. In [30], Golland and her colleagues applied a top-down hierarchical clustering approach to the rest-state scan and movie watching data. By incorporating the concept of functional hierarchy and its multi-resolution visualization framework, their results described the co-activation pattern at different scales, which helped the interpretation of the results when compared to the anatomical structure of the brain. They discovered that clustering analysis finds networks consistent with neuroanatomical parcellation of the cortex at the coarse levels of hierarchy, and that the finer levels reveal an interesting, yet unstudied, network structure which exhibits higher variability across subjects and experiments. Various components that lead to the differences in the clustering tree need to be understood to expand this model for use in global analysis.

Besides describing functional relations between brain regions, several approaches have been developed to provide information on the directionality of those relations, called pathway analysis. These include structural equation models and dynamical causal models [54], whose goals are to measure effective connectivity, which is the influence exerted by one neuronal system over another.

3.4 Summary

In this chapter, three types of brain connectivity, namely, anatomical, functional, and effective connectivity were presented and compared to each other. Functional connectivity was more specifically defined as temporal correlations between neurophysiological events. The standard approach of identifying functional connectivity using the hypothesis-driven method was discussed along with the prior work utilizing this method.

Chapter 4

Data-Driven Methods

Data-driven methods provide an alternative to testing each voxel's time course against a hypothesis. Also known as exploratory analysis, data-driven analysis explores the multivariate structure of the data, aiming to identify "interesting" components. These components may reveal structures or patterns in the data, which are difficult to identify a priori, such as unexpected activation and connection, motion related artifacts, and drifts [11]. These unsupervised analysis methods provide generalizations of connectivity analysis in situations where reference seed regions are unknown or difficult to identify reliably. One important motivation and expectation behind the use of these methods is that in many data sets, data points lie in some manifold of much lower dimensionality than that of the original data space [9]. Three most popular methods are clustering, principal component analysis, and independent component analysis, and they will be discussed in the context of functional connectivity in the subsequent sections.

We first define the notations used throughout this chapter:

$\mathbf{X}$: Data, a set of samples/observations.

$\mathbf{x}$: Single sample/observation.

$\mathbf{S}$: Sources.

$\mathbf{s}$: Single source.

K: Number of sources/components.

n: Index for observations.

k: Index for sources.

A: Mixing/projection matrix.

W: Unmixing matrix.

C: Sample covariance matrix.

T: Number of time point in fMRI data.

V: Number of voxels in fMRI data.

N: Number of observations. (For spatial PCA and ICA, $N = T$. For GMM, $N = V$.)

D: Dimension of observation. (For spatial PCA and ICA, $D = V$. For GMM, $D = T$.)

4.1 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique that linearly transforms an original set of variables into a substantially smaller set of uncorrelated variables that captures most of the variance in the original set of variables. It is also known as the Karhunen-Loeve transform [40]. One of the main goals of PCA is to reduce the dimensionality of the original data set. A small set of uncorrelated variables is assumed to represent the underlying sources for observations, and is more computationally efficient in further analysis than a larger set of correlated variables. Thus, PCA is often used as a pre-processing step for other data-driven analysis methods such as clustering and ICA. For investigation of functional connectivity, principal component analysis has been found to be useful. In [26], time-series functional connectivity was investigated by defining it as the temporal correlation between spatially remote neurophysiological events. Besides its use in dimensionality reduction, PCA is widely applied in lossy compression and feature extraction of data and data visualization [40]. In this section, we follow the formulation presented in [9].

The algorithm of principal component analysis is driven by two different ideas, namely maximum variance of transformed data and minimum reconstruction error, which can be shown to be equivalent. In the maximum variance formulation of PCA, it is defined as the orthogonal projection of the original data onto a lower dimensional

linear "principal space," which maximizes the variance of the projected data. Assume we have a data set of observations $\{\mathbf{x}_n\}$ of dimension D, where $n = 1, \ldots, N$ represents the number of samples. We project the data onto a space of dimension $K \leq D$, where the value of K is determined by the user depending on the application. In a simplified case, consider projecting the data onto a one-dimensional space where the direction is defined by a D-dimensional unit vector $\mathbf{u}_1$, such that $\mathbf{u}_1^T\mathbf{u}_1 = 1$. The variance of the projected data is given by

$$\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}}\}^2 = \mathbf{u}_1^T\mathbf{C}\mathbf{u}_1, \tag{4.1}$$

where $\bar{\mathbf{x}}$ is the sample mean of the data

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \tag{4.2}$$

and $\mathbf{C}$ is the sample data covariance matrix defined as

$$\mathbf{C} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T. \tag{4.3}$$

We want to maximize the variance of the projected data with respect to $\mathbf{u}_1$, enforcing the constraint that $\mathbf{u}_1$ is a unit vector. Then by using a Lagrange multiplier, this becomes a maximization problem of

$$\mathbf{u}_1^T\mathbf{C}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^T\mathbf{u}_1), \tag{4.4}$$

where λ_1 is a constant. By taking the derivative with respect to $\mathbf{u}_1$ and setting it to zero, the maximum is achieved when

$$\mathbf{C}\mathbf{u}_1 = \lambda_1\mathbf{u}_1, \tag{4.5}$$

which implies that $\mathbf{u}_1$ is an eigenvector of the covariance matrix $\mathbf{C}$. Also, since $\mathbf{u}_1$ is

a unit vector, the variance of the projected data is given by

$$\mathbf{u}_1^T \mathbf{C} \mathbf{u}_1 = \lambda_1, \tag{4.6}$$

which suggests that $\mathbf{u}_1$ must be the eigenvector with the largest eigenvalue λ_1. This eigenvector defines the first principal component.

The next principal components can be found in an iterative manner by selecting the direction that maximizes the variance of the projected data among all the directions that are orthogonal to the ones that have already been defined as principal components. Therefore, in the general case of projection onto the K-dimensional space, the optimal projection that maximizes the variance of the projected data is achieved when the principal components are defined as the K eigenvectors $(\mathbf{u}_1, \ldots, \mathbf{u}_K)$ of the data covariance matrix $\mathbf{C}$ with the K largest eigenvalues $(\lambda_1, \ldots, \lambda_K)$. Only the first (the mean) and second order (covariance) information of the data governs the principal component analysis.

An alternative formulation of principal component analysis is based on the notion of minimum reconstruction error. In this formulation, it is shown that among all linear projection methods, principal component analysis minimizes the reconstruction error, which is the distance between a data point and its reconstruction from the lower dimensional space,

$$J = \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}_n - \hat{\mathbf{x}}_n||^2, \tag{4.7}$$

where $\hat{\mathbf{x}}_n$ is the projection of point $\mathbf{x}_n$ onto the lower dimensional space.

To obtain a solution that minimizes the reconstruction error, we assume we have a complete orthonormal set of D-dimensional basis vectors $\{\mathbf{u}_i\}$, where $i = 1, \ldots, D$. We want to approximate each data point using a set of $K \leq D$ basis vectors of the lower-dimensional space of the projected data. Then the approximation of each data point $\mathbf{x}_n$ can be expressed by

$$\hat{\mathbf{x}}_n = \sum_{i=1}^{K} a_{ni} \mathbf{u}_i + \sum_{i=K+1}^{D} b_i \mathbf{u}_i, \tag{4.8}$$

where a_{ni}'s are the coefficients of the basis vectors for each point and b_i's are constants for all data points. We seek $\{\mathbf{u}_i\}$, $\{a_{ni}\}$, and $\{b_i\}$ that minimize the reconstruction error J. Taking the first derivative of J with respect to $\{a_{ni}\}$ and $\{b_i\}$, and making use of the orthonormality condition of basis vectors, we obtain $a_{nj} = \mathbf{x}_n^T\mathbf{u}_j$ for $j = 1, \ldots, K$ and $b_j = \bar{\mathbf{x}}^T\mathbf{u}_j$ for $j = K+1, \ldots, D$. Substituting $\{a_{ni}\}$ and $\{b_i\}$ and making use of the relation $\mathbf{x}_n = \sum_{i=1}^{D}(\mathbf{x}_n^T\mathbf{u}_i)\mathbf{u}_i$ gives

$$\mathbf{x}_n - \hat{\mathbf{x}}_n = \sum_{i=K+1}^{D}\{(\mathbf{x}_n - \bar{\mathbf{x}})^T\mathbf{u}_i\}\mathbf{u}_i. \tag{4.9}$$

Therefore, we obtain the reconstruction error J in terms of the basis vectors in the form of

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=K+1}^{D}(\mathbf{x}_n^T\mathbf{u}_i - \bar{\mathbf{x}}^T\mathbf{u}_i)^2 = \sum_{i=K+1}^{D}\mathbf{u}_i^T\mathbf{C}\mathbf{u}_i. \tag{4.10}$$

Then, the solution for the minimization of J with the constraint that $\{\mathbf{u}_i\}$ are orthonormal is given by choosing $\{\mathbf{u}_i\}$ as the eigenvectors of the covariance matrix $\mathbf{C}$, namely,

$$\mathbf{C}\mathbf{u}_i = \lambda_i\mathbf{u}_i, \tag{4.11}$$

where $i = 1, \ldots, D$, and the corresponding reconstruction error J is given by

$$J = \sum_{i=K+1}^{D}\lambda_i \tag{4.12}$$

which is the sum of the eigenvalues of eigenvectors that are normal to the principal subspace. The minimum reconstruction error is achieved by choosing such eigenvectors with the $D - K$ smallest eigenvalues. Equivalently, the eigenvectors with the K largest eigenvalues define the basis vectors of the principal subspace. Therefore, we have shown that the maximum variance and minimum reconstruction error formulations of PCA give identical solutions.

There remains the problem of choosing the dimensionality of the principal space where we project the original data onto, or equivalently, the number of principal components, K. One can choose K based on a priori knowledge or use automatic

procedures. Several measures have been adopted in choosing the number of principal components. One popular way, which examines the proportion of variance, is to select K such that the top K principal components explains 90 per cent of the total variance in the data. Since the variance in the data is explained in terms of the eigenvalues λ_i's of the data covariance matrix $\mathbf{C}$, we pick the optimal K such that

$$\frac{\lambda_1 + \cdots + \lambda_K}{\lambda_1 + \lambda_2 + \cdots + \lambda_D} = 0.9 \tag{4.13}$$

holds. Adding another principal component beyond K would not substantially increase the variance explained. As in the case of many time series of images, such as fMRI experiments, where inputs are highly correlated in space and time, there will be a small number of eigenvectors of the data covariance matrix with large eigenvalues. Therefore, a large amount of dimensionality reduction can be achieved via principal component analysis. Another approach in selecting the number of principal components is to adopt a model selection technique. This approach is discussed in depth in Section 4.5.

In the setting of fMRI time-series data, let the data be represented as an T x V matrix $\mathbf{X}$, where each row represents a time point and column presents a voxel. In terms of a generative model, we assume that the observed time course $\mathbf{x}$ comes from a multivariate Gaussian distribution with mean $E[\mathbf{x}] = \mu$ and covariance $\text{Cov}[\mathbf{x}] = \Sigma$, i.e., $\mathbf{x} \sim N(\mu, \Sigma)$. Then, the computation of the principal components is reduced to the solution of an eigenvalue-eigenvector decomposition of a correlation/covariance matrix. In this work, we follow the convention that principal components are the normalized eigenvectors from the decomposition. Following the formulation in [26], a connected brain system is represented as a pattern of activity in terms of correlations or covariance, $\mathbf{X}^T\mathbf{X}$, depending on the normalization of the data. The subtracting the mean from the data is necessary in order to force the first principal component to represent the direction that captures the most variance *within* the data, rather than with respect to the origin of the coordinate system. $\mathbf{X}^T\mathbf{X}$ expressed as correlation is preferred to covariance when the variables are in different units or their variances differ

widely. Then, applying Singular Value Decomposition (SVD) on $\mathbf{X}$, $\mathbf{X} = U\Lambda\Psi^T$, the normalized time-series matrix $\mathbf{X}$ is decomposed into two sets of orthonormal vectors U and Ψ, which represent patterns in space and in time, respectively, and Λ, which is a diagonal matrix of singular values in a decreasing order. Since $\mathbf{X}^T\mathbf{X}$ defines the functional connectivity matrix, rearranging the above equation into $\mathbf{X}^T\mathbf{X}\Psi = \Lambda\ ^2\Psi$ implies that the columns of Ψ are the eigenvectors of the functional connectivity matrix. Thus, the first eigenvector represents a spatial pattern that embodies the most variance. Other eigenvectors are sorted in terms of the amount of variance they explain. Since these eigenvectors or spatial modes can be represented as an image, they are often called eigenimages [63], each of which can be seen as a template for important features. In addition, each column of U depicts the time dependent profile of each eigenimage and reflects the level at which an eigenimage is expressed over time or under each experimental condition. By comparing the temporal expression of the first few eigenimages with the variation in experimental factors over time, we can determine a distributed functional system associated with these various factors.

4.2 Probabilistic Principal Component Analysis

In [62], Tipping and Bishop developed a more precise probabilistic formulation of PCA using a Gaussian latent variable model, similar to factor analysis. This probabilistic formulation of PCA provides a way to find a low-dimensional representation of higher dimensional data with a well-defined probability distribution, and enables comparison to other generative models within a density estimation framework.

Let us consider a latent variable model which fits data $\mathbf{x}$ of dimension D to its corresponding lower-dimensional representation $\mathbf{z}$ of dimension K. This continuous latent variable $\mathbf{z}$ corresponds to the principal subspace. Assuming that this lower-dimensional representation of $\mathbf{x}$ is linear, we aim to find a projection matrix $\mathbf{A}$, which spans a linear space within the data subspace corresponding to the principal subspace, and offset μ such that $\mathbf{x} = \mathbf{A}\mathbf{z} + \mu$, where μ is the mean offset of the data permitting the model to have non-zero mean. We evaluate the estimates of the parameters with

an objective function, which in this case is the squared-error in representation,

$$\mathbf{A}^* = \text{argmin}_{\mathbf{A}} ||\mathbf{x} - \mathbf{A}\mathbf{z} - \mu||^2. \tag{4.14}$$

Extending the model to explicitly represent the noise present in the observations by additive isotropic noise $\epsilon \sim N(0, \sigma^2\mathbf{I})$, an observation $\mathbf{x}$ is generated by

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mu + \epsilon. \tag{4.15}$$

Assuming the prior distribution over $\mathbf{z}$ is a standard normal distribution,

$$p(\mathbf{z}) = N(\mathbf{z}|0, \mathbf{I}), \tag{4.16}$$

with the conditional distribution of the observed variable $\mathbf{x}$, which is also Gaussian,

$$p(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|\mathbf{A}\mathbf{z} + \mu, \sigma^2\mathbf{I}) \tag{4.17}$$

we can compute the marginal distribution of $\mathbf{x}$,

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \sim N(\mu, \mathbf{B}), \tag{4.18}$$

where the covariance matrix $\mathbf{B}$ is defined by

$$\mathbf{B} = \mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{I}. \tag{4.19}$$

Equation 4.18 defines the probability model of the high dimensional observations. Given the model and $\mathbf{B}$, it implies that the likelihood of any observation $\mathbf{x}$ can be directly evaluated.

For a set of data observations $\mathbf{X} = \{\mathbf{x}_n\}$, the corresponding log-likelihood function

is given by

$$\ln p(\mathbf{X}|\mu, \mathbf{A}, \sigma^2) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mu, \mathbf{A}, \sigma^2) = -\frac{N}{2}\{D\ln(2\pi) + \ln|\mathbf{B}| + \mathrm{Tr}(\mathbf{B}^{-1}\mathbf{C})\}, \tag{4.20}$$

where the maximum likelihood estimator for μ is given by the mean of the data $\bar{\mathbf{x}}$, and $\mathbf{C}$ is the sample covariance matrix defined in Equation 4.3. Finding the maximum likelihood estimator for other parameters is non-travial, but Tipping and Bishop [62] show that the closed-form solutions of the maximum likelihood estimates of the parameters $\mathbf{A}$ and σ^2 are obtained when

$$\widehat{\mathbf{A}}_{ML} = \mathbf{U}(\mathbf{L} - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{4.21}$$

where $\mathbf{R}$ is an arbitrary rotation matrix, $\mathbf{U}$ is the matrix of the eigenvectors of the observation covariance matrix $\mathbf{C}$, and $\mathbf{L}$ is the matrix whose diagonal contains the corresponding eigenvalues. At the stationary point of the likelihood function for $\mathbf{A} = \widehat{\mathbf{A}}_{ML}$, the corresponding maximum likelihood estimate for σ^2 is

$$\hat{\sigma}^2_{ML} = \frac{1}{D-K}\sum_{i=K+1}^{D} \lambda_i, \tag{4.22}$$

where λ_i is the ith eigenvalue of the observation covariance matrix $\mathbf{C}$, assuming the eigenvalues are arranged in order of descending magnitude. It was shown by Tipping and Bishop that the maximum of the likelihood function is obtained when the chosen K eigenvectors correspond to the K largest eigenvalues.

There are two main advantages of probabilistic principal component. First, it provides an explicit probability model of the data, $p(\mathbf{X})$, in the density estimate framework, which allows us to compute the likelihood of any observation and to compare the result of probabilistic principal component to other exploratory data analysis methods. Second, in a generative viewpoint, this probability model can be used to provide samples from the distribution of PCA. We will not directly model our fMRI data as probabilistic PCA, but instead use it to compute the likelihood in

model selection.

4.3 Independent Component Analysis

Originally developed by Bell and Sejnowski [8], independent component analysis (ICA) is a powerful explorative analysis technique used in many applications that tackle blind source separation (BSS) problems [37]. ICA assumes that the original data variable $\mathbf{X}$ is a linear weighted sum of a set of unknown latent source variables $\mathbf{S}$; i.e. $\mathbf{X} = \mathbf{AS}$, where $\mathbf{A}$ is the matrix of mixing coefficients. The latent variables, known as the independent components of the observed data, are assumed to be non-gaussian and mutually independent. As the linear mixing system is unknown, both the source variables $\mathbf{S}$ and the weights $\mathbf{A}$ are iteratively estimated in ICA.

ICA was first introduced for fMRI analysis by McKeown *et al.* [49]. They used the ICA algorithm to investigate task-related human brain activity in fMRI data and showed that ICA can be used to reliably partition fMRI data sets into meaningful basic components, including task and function related physiological changes, non-task related signal changes, and artifactual components. Despite its strict linearity assumption and debate on the choice of spatial or temporal independency, ICA has been widely applied, especially in analysis of complex fMRI data. For instance, Calhoun *et al.* [11] used ICA to decompose activation patterns into interpretable components during a simulated driving test, which is an example of a near-protocol-free fMRI experiment.

Since we are interested in the functional connectivity of the brain, we want to obtain spatial maps which represent independent functional networks and their associate time courses. Thus, for our analysis of fMRI data, we use spatial independent component analysis (sICA), rather than time independent component analysis which produces a set of time courses that are as independent to each other and their associated spatial maps. Esposito *et al.* [24] compared two ICA algorithms that have been used so far for spatial ICA (sICA) of fMRI time-series in the literature: Infomax [8] and Fixed-Point [38]. They found that whereas both algorithms produced highly

accurate results, because of its adaptive nature, they concluded that the Infomax approach appears to be better suited to investigate activation phenomena that are not predictable or adequately modelled by inferential techniques. We chose to use the Infomax approach for our fMRI analysis.

4.3.1 Generative Model

As in the case of PCA, the generative model used in ICA is a linear mixture of latent random variables. Assuming we have N observations of such mixtures, each observation of mixture $\mathbf{x}_n$ is expressed as a linearly weighted sum of K independent sources, $\mathbf{s}_k$,

$$\mathbf{x}_n = a_{n1}\mathbf{s}_1 + \cdots + a_{nK}\mathbf{s}_K \tag{4.23}$$

for $n = 1, \ldots, N$ and a_{nk}'s represent the mixing coefficients. ICA assumes that each mixture $\mathbf{x}$ and each source $\mathbf{s}$ are random variables. In the matrix form, where the mixed signals are represented as a data matrix $\mathbf{X}$, the generative model can be expressed as

$$\mathbf{X} = \mathbf{AS}. \tag{4.24}$$

In other words, for a total of K sources, each row $\mathbf{s}_k^T$ of the source matrix $\mathbf{S}$ contains a single independent component, whereas each column $\mathbf{a}_k$ of the mixing matrix $\mathbf{A}$ comprises the corresponding mixing weights.

Unlike principal component analysis where we want our sources to be uncorrelated with each other, independent component analysis imposes a stricter condition that sources are statistically independent of each other. Given $\mathbf{X} = \mathbf{AS}$, we want to find the optimal unmixing matrix $\mathbf{W}^* \sim \mathbf{A}^{-1}$ in $\mathbf{S} = \mathbf{WX}$ such that the components of $\mathbf{S}$ are statistically independent.

Since both $\mathbf{A}$ and $\mathbf{S}$ are unknown, the ICA model has the following ambiguities. First, the scaling and signs of the sources can not be determined. For example, any scalar multiplier in the mixing vector $\mathbf{a}_k$ can be cancelled by dividing the source vector $\mathbf{s}_k$ by the same scalar. The second ambiguity is that the order of the independent components is not fixed since we can freely change the order of the linear terms

in Equation 4.23. The sign and scaling of components are often normalized after performing ICA to deal with these ambiguities. In our case, we sorted the components by their energy and adjusted the components so that each component has a positive mean.

4.3.2 Independence of Sources

What differentiates ICA from PCA and other exploratory data analysis methods is the assumption that the underlying sources, or equivalently, the components of $\mathbf{S}$ are statistically independent. Statistical independence means that the sources do not contain any information about each other. In other words, the joint probability density function (pdf) of the sources is the product of its marginal probability densities for all sources,

$$p(\mathbf{s}_1, ..., \mathbf{s}_K) = \prod_{k=1}^{K} p(\mathbf{s}_k). \tag{4.25}$$

Since the exact determination of the pdfs is generally not feasible, it is difficult to obtain a closed form solution of ICA. Instead, we estimate the sources by approximating independence with an objective function. This objective function, measuring the non-Gaussianity of the estimated sources, is often based on mutual information (infomax) [8] or negentropy (fixed-point algorithm) [38]. In practice, we use iterative methods, such as gradient descent, to optimize the objective function of ICA.

The use of non-Gaussianity as a measure of independence is justified by the central limit theorem. Central limit theorem states that the distribution of the sum of independent, identically distributed random variables tends to be more Gaussian than the original ones. In other words, the more non-Gaussian the sources are, the more independent they have to be. This builds the link between independence and non-Gaussianity.

One important measure of non-Gaussianity is given by negentropy, which is based on the information-theoretic quantity of entropy. Treating each source $\mathbf{s}$ as a discrete random variable, the entropy H of the discrete random variable $\mathbf{s}$ with the probability

distribution $p(\cdot)$ is defined as

$$H(\mathbf{s}) = -\sum_{\mathbf{a}} p(\mathbf{s} = a_i) \log p(\mathbf{s} = a_i), \tag{4.26}$$

where the a_i are all the possible values of $\mathbf{s}$. By definition, entropy measures the amount of information contained in the observation of the random variable $\mathbf{s}$. The more random (unstructured and unpredictable) the random variable is, the larger its entropy is. This definition can be generalized to differential entropy for continuous random variables or vectors, where the summation in the entropy equation is replaced by an integral.

In the differential entropy setting, Gaussian variables have the largest entropy among all random variables of equal variance, implying that the Gaussian distribution is the least structured of all distributions [19]. This allows entropy to be used as a measure of non-Gaussianity. Negentropy J is defined as

$$J(\mathbf{s}) = H(\mathbf{s}_{gauss}) - H(\mathbf{s}), \tag{4.27}$$

where $\mathbf{s}_{gauss}$ is a Gaussian random variable with the same covariance as $\mathbf{s}$. Negentropy is always non-negative and is equal to zero if and only if $\mathbf{s}$ is Gaussianly distributed. In other words, negentropy measures the difference between the Gaussian distribution and that of the independent variables, and shows how non-Gaussian the independent variables are. In the case of unit variance $\mathbf{s}$, entropy and negentropy differ only by a constant. The above definition of negentropy requires an exact pdf of the random variable. To make estimation feasible in practice, negentropy can be approximated without knowing exact pdfs by using other measures of non-Gaussianity, such as skewness and kurtosis, which are, the third and fourth order cumulants, respectively [38].

Mutual information measures how much dependence is shared among random variables. The mutual information I between K random variables is defined using

entropy as

$$I(\mathbf{s}_1, \ldots, \mathbf{s}_K) = \sum_{k=1}^{K} H(\mathbf{s}_k) - H(\mathbf{s}). \tag{4.28}$$

This is equivalent to the Kullback-Leibler divergence (relative entropy) between the joint density $p(\mathbf{s})$ and the product of its marginal densities [19]. Mutual information is always non-negative and zero if and only if the variables are statistically independent. Since mutual information measures the amount of information shared between random variables and captures the whole dependence structure of the variables beyond the simple covariance, it can be used as a natural measure of independence. Thus, estimating the independent components is possible by minimizing the mutual information between them. However, in practice, minimizing mutual information can be highly computationally expensive.

From the definition of negentropy, we observe that negentropy differs from mutual information only by a constant C; i.e.

$$I(\mathbf{s}_1, \ldots, \mathbf{s}_K) = C - \sum_{k=1}^{K} J(\mathbf{s}_k). \tag{4.29}$$

This shows the fundamental relation between negnetropy and mutual information. Therefore, maximizing negentropy is equivalent to minimizing mutual information when estimating independence.

4.3.3 The Infomax Algorithm

Infomax is an implementation of ICA from a neural network viewpoint, based on minimization of mutual information between independent components [8]. In the Infomax framework, a self-organizing learning algorithm is chosen to maximize the output entropy, or the information flow, of a neural network of non-linear units. The network has N input and output neurons, and an N x N weight matrix $\mathbf{W}$ connecting the input layer neurons with the output layer neurons. $\mathbf{X}$ is an input the to neural

network. Assuming sigmoidal units, the neuron's outputs are given by

$$\mathbf{s} = g(\mathbf{D}) \text{ with } \mathbf{D} = \mathbf{W}\mathbf{X} \tag{4.30}$$

where is $g(\cdot)$ a specified non-linear function. This non-linear function, which provides necessary higher-order statistical information, is chosen to be a logistic function

$$g(D_i) = \frac{1}{1 + e^{-D_i}}, \tag{4.31}$$

where D_i represents a row in the matrix $\mathbf{D}$ for $i = 1, \cdots, N$.

The main idea of this algorithm is to find an optimal weight matrix $\mathbf{W}$ iteratively such that the output joint entropy $H(\mathbf{s})$ is maximized. In the simplified case of only two outputs, where $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2)$, $I(\mathbf{s}) = H(\mathbf{s}_1) + H(\mathbf{s}_2) - H(\mathbf{s})$ holds by the definition of mutual information. Hence, we can minimize the mutual information by maximizing the joint entropy. Then, by another equivalent definition of mutual information, $I(\mathbf{X}, \mathbf{s}) = H(\mathbf{s}) - H(\mathbf{s}|\mathbf{X})$, the information flow between the input and the output is maximized by maximizing the joint entropy $H(\mathbf{s})$ since the last term vanishes due to the deterministic nature of $\mathbf{s}$ given $\mathbf{X}$ and $g(\cdot)$.

To find an optimal weight matrix $\mathbf{W}$, the algorithm first initializes $\mathbf{W}$ to the identity matrix $\mathbf{I}$. Using small batches of data drawn randomly from $\mathbf{X}$ without substitution, the elements of $\mathbf{W}$ are updated based on the following rule:

$$\Delta\mathbf{W} = -\epsilon\left(\frac{\partial H(\mathbf{s})}{\partial \mathbf{W}}\right)\mathbf{W}^T\mathbf{W} = -\epsilon(\mathbf{I} + f(\mathbf{D})\mathbf{D}^T)\mathbf{W}, \tag{4.32}$$

where ϵ is the learning rate (typically near 0.01) and the vector function g has elements

$$f_i(D_i) = \frac{\partial}{\partial D_i} \ln\left(\frac{\partial g_i}{\partial D_i}\right) = (1 - 2s_i). \tag{4.33}$$

Equation 4.32 is known as the Infomax algorithm. The $\mathbf{W}^T\mathbf{W}$ term in Equation 4.32, first proposed by Amari *et al.* [2], avoids matrix inversions and speeds up convergence. During training, the learning rate is reduced gradually until the weight matrix stops

changing appreciably. The choice of nonlinearity depends on the application type. In the context of fMRI, where relatively few highly active voxels are usually expected in a large volume, the distribution of the estimated components is assumed to be super-Gaussian. Therefore, a sigmoidal function is appropriate for such an application [49].

4.3.4 Maximum Likelihood Formulation

ICA can also be formulated in the maximum likelihood framework [38, 59]. For a mixture vector variable $\mathbf{x}$ with the joint pdf $p(\mathbf{x})$ and a source vector variable $\mathbf{s}$ with the joint pdf $p(\mathbf{s})$, such that $\mathbf{s} = \mathbf{W}^*\mathbf{x}$, where $\mathbf{W}^*$ is the optimal unmixing matrix, the density of $\mathbf{x}$ given $\mathbf{W}^*$ is

$$p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{s}}(\mathbf{s})|\mathbf{W}^*|. \tag{4.34}$$

$|\mathbf{W}^*| = |\partial\mathbf{s}/\partial\mathbf{x}|$ is the Jacobian of $\mathbf{s}$ with respect to $\mathbf{x}$. Equation 4.34 defines the likelihood of the observed mixtures $\mathbf{x}$. We assume that we can set the density of $\mathbf{s}$ a priori.

For any non-optimal unmixing matrix $\mathbf{W}$, the resulting signals are given by $\mathbf{y} = \mathbf{W}\mathbf{x}$. Explicitly stating the dependence on $\mathbf{W}$, the likelihood $p_{\mathbf{x}}(\mathbf{x}|\mathbf{W})$ of the mixture $\mathbf{x}$ given $\mathbf{W}$ is

$$p_{\mathbf{x}}(\mathbf{x}|\mathbf{W}) = p_{\mathbf{s}}(\mathbf{W}\mathbf{x})|\mathbf{W}|. \tag{4.35}$$

The maximum of this likelihood is achieved when $\mathbf{W}$ is the optimal unmixing matrix $\mathbf{W}^*$. Therefore, the quality of any presumed unmixing matrix $\mathbf{W}$ can be evaluated by the above equation, and we can optimize Equation 4.35 to find the particular $\mathbf{W}$ that maximizes the likelihood of the mixture.

Since $\mathbf{W}$ is the parameter that needs to be estimated to calculate the maximum likelihood, the joint pdf $p_{\mathbf{x}}(\mathbf{x}|\mathbf{W})$ for $\mathbf{x}$ can be treated as if it was a function of the parameter $\mathbf{W}$. We denote this joint pdf as the likelihood function $L(\mathbf{W})$. Assuming the K source signals are statistically independent, such that the joint pdf $p_{\mathbf{s}}$ is the product of its marginal pdfs, it allows the logarithm of Equation 4.35 to be written

as

$$\ln L(\mathbf{W}) = \ln p_{\mathbf{x}}(\mathbf{X}|\mathbf{W}) = \sum_{k=1}^{K}\sum_{n=1}^{N} \ln p_{\mathbf{s}}(\mathbf{w}_k^T \mathbf{x}_n) + N \ln |\mathbf{W}|, \tag{4.36}$$

where $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\}^T$. The matrix $\mathbf{W}$ that maximizes this function is the maximum likelihood estimate of the optimal unmixing matrix $\mathbf{W}^*$. This maximum likelihood formulation of ICA enables comparison to other exploratory methods, such as probabilistic PCA and Gaussian mixture model.

Furthermore, it can be shown that the maximum likelihood estimation formulation is equivalent to the Infomax approach [16]. To see this connection, we consider the expectation of the log-likelihood,

$$\frac{1}{N} E[\ln L(\mathbf{W})] = \sum_{k=1}^{K} E[\ln p_{\mathbf{s}}(\mathbf{w}_k^T \mathbf{x})] + \ln |\mathbf{W}|. \tag{4.37}$$

If the true distribution of $\mathbf{w}_k^T \mathbf{x}$ were equal to the pre-defined $p_{\mathbf{s}}(\cdot)$, then the first term on the right hand side would be equal to $-\sum_{k=1}^{K} H(\mathbf{w}_k^T \mathbf{x})$, by the definition of entropy. For an invertible linear transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$, the mutual information is expressed as

$$I(\mathbf{y}_1, \ldots, \mathbf{y}_K) = \sum_{k=1}^{K} H(\mathbf{y}_k) - H(\mathbf{x}) - \ln |\mathbf{W}|. \tag{4.38}$$

Therefore, combining Equation 4.37 and 4.38 shows that the likelihood would be equal to the negative of the mutual information up to an additive constant. More precisely, exact equivalence arises when the non-linearities $g_i(\cdot)$ used in the neural network are chosen as the cumulative distribution functions corresponding to the densities $p_{\mathbf{s}}(\cdot)$, i.e., $g_i'(\cdot) = p_{\mathbf{s}}(\cdot)$.

4.3.5 Spatial ICA for fMRI

In the case of spatial independent component analysis (sICA) [49], we assume that spatio-temporal fMRI data come from the linear mixing of different brain processes whose spatial distributions are invariant over time and statistically independent. The fMRI data is represented by a $T \times V$ data matrix $\mathbf{X}$, where T is the length of the time

course and V is the total number of voxels in the volume. In case of sICA, each image is treated as a sample. sICA provides an unsupervised composition so that each row of $\mathbf{S}$ of dimension $K \times V$ contains an independent spatial pattern and the column of $\mathbf{A}$ holds its corresponding activation time-course, $\mathbf{X} = \mathbf{AS}$. Typically, spatial maps, treated as a segmentation of the volume, are sparse and non-overlapping, but exact interpretation of the components is difficult. Within each independent spatial component, we declare voxels with a large magnitude of coefficients as functionally connected. This is similar to the definition of connectivity maps in PCA [26]. One problem of ICA is that it is difficult to assign a statistical significance to a value in the spatial maps since the amplitude of a separated signal is determined up to sign and scale. In practice, a z-map conversion is adopted to convert a spatial map with a non-Gaussian distribution into a z-map with a Gaussian distribution. For each voxel within a spatial map, we first subtract the mean of the spatial map from the voxel value, and then divide it by the standard deviation of the spatial map. This enables assignment of significance levels based on the transformed z-map values [49].

There is another subtle step for dealing with fMRI data in ICA. Before estimating the independent components, the observed data $\mathbf{X}$ is whitened, that is, the samples made uncorrelated and their variances one. Whitening is a linear transformation that can be constructed using principal component analysis (PCA). Since whitening reduces the number of free parameters, it makes the estimation of independent components computationally easier. Specifically, the mixing matrix $\mathbf{A}$ becomes orthonormal, making its inverse $\mathbf{W}$ easy to calculate. In addition, by excluding the weakest principal components, the dimension of the data can be reduced in a way that optimally preserves the total variance, which improves the signal-to-noise ratio of the data.

Calhoun *et al.* investigated many properties of ICA when applied to fMRI data. They provided a generative model for validating and comparing results when different choices of algorithms and preprocessing stages were performed [15]. They generated artificial fMRI data using the synthesis model, performed analysis of the data using ICA, and evaluated the performance using the Kullback-Leibler divergence between

the true source and the estimated component. In their work, Infomax outperformed Fixed-Point for the choice of ICA algorithm, and PCA outperformed clustering for the choice preprosessing. They concluded that the best combination is Infomax with PCA. Based on this result, we chose Infomax as our choice of algorithm and PCA as our preprossing step for the analysis of fMRI data in the next chapter. Furthermore, Calhoun *et al.* compared the spatial and temporal ICA using the Fixed-Point algorithm [13]. With synthetic activations, they found a good correspondence between the resulting components of sICA and tICA for an activation study with a single activation, but also observed some divergence for a visual paradigm in which two closely related regions were active. For further perspectives of ICA on fMRI data, such as validation, group analysis, and applications to clinical research, one can refer to the review articles in this topic [12, 48].

4.4 Clustering: Gaussian Mixture Model

Clustering, or data segmentation, algorithms aim to group a collection of data points into subsets such that the points in each subset are more closely related to each other than those in other subsets, while each cluster itself is as different as possible from other clusters. In many real data cases where multiple clusters are present, a simple probability distribution is insufficient to capture the structure of the data. A linear combination of more basic distributions, known as mixture distribution, gives a better characterization by providing a framework upon which to build a more complex, richer class of density models. In this section, we follow the formulations presented in [9].

In terms of a generative model, we assume that the data sample $\mathbf{x}$ is generated from a mixture density,

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|\Theta_k)P(\Theta_k) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\Theta_k), \tag{4.39}$$

where Θ_k are clusters and $p(\mathbf{x}|\Theta_k)$ and $p(\Theta_k) = \pi_k$ represent component densities and mixture proportions, respectively. The number of clusters K must be pre-defined.

Gaussian mixture model (GMM) is a mixture distribution where each distribution in the mixture is assumed to be a single multivariate Gaussian, $p(\mathbf{x}|\Theta_k) = N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$. With GMM, any continuous probability density can be approximated to some arbitrary accuracy by using a sufficiently large number of single Gaussians and by adjusting their means and covariances as well as the mixture weights of the linear combination.

If a joint distribution is defined over observed and latent variables, we can obtain the distribution of observed variables alone by marginalizing the joint distribution over the latent variables. This makes enables relatively complex marginal distributions over the observed variables to be defined in terms of more tractable joint distributions over the expanded space of observed and latent variables [9]. In other words, the latent variables allow complicated distributions to be formed from simpler distributions.

To better understand this model, we introduce a discrete K-dimensional binary latent variable $\mathbf{z}$. An element z_k of $\mathbf{z}$ can have a value of either 0 or 1, and only one element of $\mathbf{z}$ is equal to one. There exists only K possible states of the vector $\mathbf{z}$. The marginal distribution of $\mathbf{z}$ can then be specified in terms of the mixing coefficients,

$$p(z_k = 1) = \pi_k, \tag{4.40}$$

where the mixing coefficients $\{\pi_i\}$ must satisfy

$$0 \leq \pi_k \leq 1 \tag{4.41}$$

and

$$\sum_{k=1}^{K} \pi_k = 1 \tag{4.42}$$

in order to be a valid probability distribution. Then the density of $\mathbf{z}$ can be written as

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{4.43}$$

Since the conditional distribution of $\mathbf{x}$ given a particular $\mathbf{z}$ is defined as a Gaussian

distribution in GMM,

$$p(\mathbf{x}|z_k) = N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k), \tag{4.44}$$

the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ can be written in the form

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)^{z_k}. \tag{4.45}$$

Now, with the introduction of the latent variable $\mathbf{z}$ and marginalizing the joint distribution of $\mathbf{x}$ and $\mathbf{z}$ over all possible states of $\mathbf{z}$, the distribution of $\mathbf{x}$ is obtained in the form of

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k), \tag{4.46}$$

which is equivalent to Equation 4.39. In generative viewpoint, we first generate a value of $\mathbf{z}$ according to the mixture coefficients (Equation 4.43). Then, a data point is generated from the Gaussian distribution which corresponds to the outcome of $\mathbf{z}$ (Equation 4.45). Therefore, for every observation $\mathbf{x}_n$, there exists a corresponding latent variable $\mathbf{z}_n$. This leads to another important quantity, the conditional distribution of $\mathbf{z}$ given $\mathbf{x}$, which can be viewed as the responsibility of component k for explaining the observation of data $\mathbf{x}$. We let $\gamma(z_k)$ denote this conditional distribution and use Bayes' theorem to obtain

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{k'=1}^{K} p(z_{k'} = 1)p(\mathbf{x}|z_{k'} = 1)} \tag{4.47}$$

$$= \frac{\pi_k N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} N(\mathbf{x}|\mu_{k'}, \mathbf{\Sigma}_{k'})}. \tag{4.48}$$

Note that π_k can be interpreted as the prior probability of $z_k = 1$, and $\gamma(z_k)$ as the corresponding posterior probability after observing $\mathbf{x}$.

Representing a data set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ as an $N \times D$ matrix $\mathbf{X}$ (for fMRI data, N = V and D = T), the log-likelihood of the data is given by

$$\ln p(\mathbf{X}|\pi, \mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k). \tag{4.49}$$

We want to estimate the parameters $\{\pi_i\}, \{\mu_i\}, \text{and} \{\Sigma_i\}$ from the sample for $i = 1, ..., K$ such that the log-likelihood function is maximized. The usual way of estimating the parameters, which is to set the first derivative of the log-likelihood with respect to each parameter to zero and solve for the parameter, does not lead to a closed form solution in this case. Such parameters are expressed in terms of the responsibility term $\gamma(z_k)$, which in turn also involves the parameters we want to estimate, as shown in Equation 4.48. Therefore, we use the Expectation-Maximization (EM) algorithm [22] to estimate the parameters in an iterative scheme.

The basic heuristic of the EM algorithm for estimating the parameters of GMM is as follows [9]:

1. Initialize the means $\{\mu_i\}$, covariances $\{\Sigma_i\}$, and mixing coefficients $\{\pi_i\}$, and compute the initial value of the log likelihood.

2. Expectation step: Evaluate the responsibilities using the current parameter values according to the relative density of each data point under each Gaussian component,
$$\gamma(z_k) = \frac{\pi_k N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} N(\mathbf{x}|\mu_{k'}, \mathbf{\Sigma}_{k'})}. \tag{4.50}$$

3. Maximization step: Update the estimates of the parameters using the new responsibilities,
$$N_k = \sum_{n=1}^{N} \gamma(z_k) \tag{4.51}$$
$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_k) \mathbf{x}_n \tag{4.52}$$
$$\mathbf{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_k)(\mathbf{x}_n - \mu_k^{new})(\mathbf{x}_n - \mu_k^{new})^T \tag{4.53}$$
$$\pi_k^{new} = \frac{N_k}{N}. \tag{4.54}$$

4. Compute the log-likelihood

$$\ln p(\mathbf{X}|\pi, \mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \tag{4.55}$$

and evaluate the convergence of the parameters and the log-likelihood. Iterate steps 2 and 3 until convergence.

Each update of the parameters that alternates between the expectation and maximization steps is guaranteed to increase the log-likelihood function [22]. Thus, we repeat this procedure until the change in the likelihood or the parameters falls below some set threshold. This is usually done in a finite number of iterations. However, as in all gradient ascent procedures, there will generally be multiple local maxima of the likelihood function. It is important to note that the EM algorithm is not guaranteed to reach the global maximum. In addition, the results of the EM algorithm depend on the initialization of the parameters. Therefore, multiple runs using several random assignments should be run and we choose the run which has the highest likelihood.

Despite its ability to model complex densities, GMM has several additional weaknesses. First, the number of clusters must be pre-defined. Also, similar to ICA, there is a problem of identifiability, since for K clusters, there exist $K!$ equivalent solutions. The ordering of components is arbitrary. As the number of clusters gets larger, it becomes more difficult to interpret each cluster and compare across subjects. One way to improve the interpretability is to incorporate a hierarchical structure among clusters [30].

Computing full covariance matrices for all K Gaussians is computationally inefficient and can easily introduce singularities. For high dimensional fMRI data, modeling of the full covariance is impractical. Most methods obtain reasonable results modeling only the variance elements in each Gaussian.

In the fMRI setting, if we treat the time course of a voxel as a sample in the data $\mathbf{X}$, GMM simultaneously estimates an optimal partition of the volume into a set of disjoint networks and the representative time courses associated with these networks. After the algorithm converges, the responsibilities $\gamma(\mathbf{z})$ for each representative time

course can be treated as a spatial map, which represents probabilistic segmentation of the volume with respect to a particular time course representative. The exponential form of each Gaussian, combined with high dimensionality of the input space, generates essentially binary posterior probabilities.

4.5 Model Selection

Model selection is a task of selecting a statistical model, which has the best generalization of the given data, from a set of potential models [9]. As mentioned in the previous sections, we saw that the total number of components, K, is pre-specified in principal component analysis, independent component analysis, and Gaussian mixture model. The number of components in those models also determines the degree of freedom of the model and controls the model complexity. Therefore, we need to determine such parameters for our model with a goal in mind that we want the model to achieve the best predictability on new data sets. In addition, we also consider a range of different types of models in order to find out the one that best describes our data.

Typically, when the size of the data set is large, we select some of the data as the training set to train a given model with a range of values for its complexity parameters. Then we compare the parameters on independent data, called validation set, and choose the one that gives the best prediction. In case of limited data where the given model is fitted iteratively in the validation set, another test set of data is necessary to evaluate the final performance of the selected model.

When we use maximum likelihood as the measure of performance, the performance on the training set is not the best indicator of predictability due to the problem of over-fitting. In other words, as we increase the number of the complexity parameters, the likelihood of the training set will increase, but due to this close fitting to the training data, the model loses the power to accurately predict the new data.

Various information criteria have been proposed as a measure of the goodness of a fit of an estimated statistical model, which rely only on the training data and

overcomes the bias due to over-fitting. They overcome the problem of bias in the likelihood approach by introducing the penalty term for complexity which regularizes and offsets the over-fitting for more complex models. Information criteria also make use of only the training data so they can compare the complexity parameters and models in a single training run. In the case of fMRI experiments, where the amount of available data is limited due to the small size of experimental subjects and insubstantial number of repeated runs for each subject, the use of information criteria can help to determine the number of components in a data-driven analysis model.

One of the classical and basic information criteria is Akaike Information Criterion (AIC) developed by Hirotsugu Akaike in [1]. It uses information theoretic criterion and selects the model for which the score

$$\ln p(\mathbf{X}|\mathbf{\Omega}) - K' \tag{4.56}$$

is largest, where $\ln p(\mathbf{X}|\mathbf{\Omega})$ is the best-fit log-likelihood for the data $\mathbf{X}$. $\mathbf{\Omega}$ is the set of parameters, and K' represents the number of free parameters in the model. Regardless of the number of free parameters in generating the data, the goodness of fit is improved by increasing the number of free parameters to be estimated. Hence, AIC not only rewards goodness of fit, but also discourages over-fitting by including a penalty, which is an increasing function of the number of estimated parameters.

Bayesian Information Criterion (BIC) [57], on the other hand, selects the model for which the score

$$\ln p(\mathbf{X}) \sim \ln p(\mathbf{X}|\mathbf{\Omega}) - \frac{1}{2}K' \ln N \tag{4.57}$$

is the largest, where N presents the total number of data samples. Compared to AIC, BIC penalizes the number of parameters in the model more severely, and favors a simpler model. It is shown in [57] that, in the asymptotic case where N approaches infinity, corresponding to having infinite number of samples, BIC always outputs the correct model.

While we use GMM on the full data, sICA is a two step process. We pre-process our $T \times V$ data (whitening and reducing dimensions to $K \leq T$) using PCA prior

to applying sICA on the new retained $K \times V$ data to have the K spatial sources represented by a $K \times V$ matrix [15, 49]. Thus, for the likelihood term in AIC and BIC, in order to determine the optimal number of total sources for the full data using sICA, we subtract the likelihood function of probabilistic PCA (Equation 4.20) on the $(T-K) \times V$ disregarded data from the likelihood function of ICA (Equation 4.36) on the $K \times V$ retained data. The likelihood of probabilistic PCA on the disregarded components penalizes the dimensionality reduction for throwing out important information in the original data.

4.6 Summary

In this chapter, we described the main ideas, the generative models, and the algorithms for three data-driven connectivity models of our interest, PCA, ICA, and GMM. We also reviewed model selection, which is used to determine the total number of components used in the models. In addition, we discussed the applications of these methods in the context of fMRI analysis. We apply these methods in identifying functional connectivity in the next chapter and discuss their similarities and differences.

Chapter 5

Empirical Study

In this chapter, we present our comparison scheme on the performance of ICA and GMM, and necessary preprocessing steps using the methods presented in Chapter 4. In Section 5.2, the synthetic data examples used for our studies, and the analysis of the results are presented. In Section 5.3, we extend our comparison scheme to a real object recognition fMRI study.

5.1 Comparison Scheme

In this section, we discuss our approach for comparing the performance of ICA and GMM, and necessary preprocessing steps using the methods presented in Chapter 4.

5.1.1 Preprocessing and Component Selection

In fMRI studies, experimental raw data is usually preprocessed in many ways to enhance the quality of analysis, as discussed in Chapter 2. In our study, we do not emphasize the effects of the standard preprocessing techniques. However, we discuss the effect of normalization to eliminate mean on the performance of ICA and GMM. GMM, along with other clustering analysis methods, is based on grouping image voxels together by the similarity of their profile in time. Thus, when using GMM, we subtract the mean of a time course from that time course to make sure that we

cluster based on the signal shape rather than signal amplitude. For example, without subtracting the mean, two time signals with an identical shape, but with significantly different mean amplitudes are unlikely to be grouped in a same cluster. On the other hand, sICA considers the intensity profiles of each image in the data. Thus, normalizing the data across space for each image enhances the result of sICA analysis.

We evaluated the effect of different ways of normalizing the data in the experiments with synthetic data described in Section 5.3. Time-averaging the data for GMM gave a much better result than performing GMM on the raw or space-averaged data, and was similar to that of space-time averaging (averaged across both space and time). For sICA, space-avaraging the data gave a better result than using raw or time-averaged data, and again was similar to that of space-time averaging. The overall dependance of performance on the type of normalization was more significant for GMM than sICA. To make our comparison of sICA and GMM on identical data, we normalize our data both across space and time prior to our analysis.

Another major issue when performing sICA and GMM is that we need to specify the number of sources a priori. In order to approximate such number, K, we ran AIC and BIC on our normalized data with sICA and GMM over a range of values of K from 2 to 105. From this range, we obtained an estimate of the number of sources that is most likely, based on AIC and BIC. We repeated this process 30 times and selected the average as our K. Due to the known problem of AIC and BIC that they underestimate the true number of sources with a finite number of samples, we only interpreted the outcomes as "suggested" number of sources and used them to approximate the real number of sources. We ran ICA and GMM with the number of sources suggested by AIC and BIC when analyzing their performances, described in the following sections.

5.1.2 Comparison between ICA and GMM

With the preprocessed data and the estimated optimal number of sources K for ICA and GMM suggested by the model selection methods, we performed the comparison of the performance of ICA and GMM on classifying functional connectivity and dis-

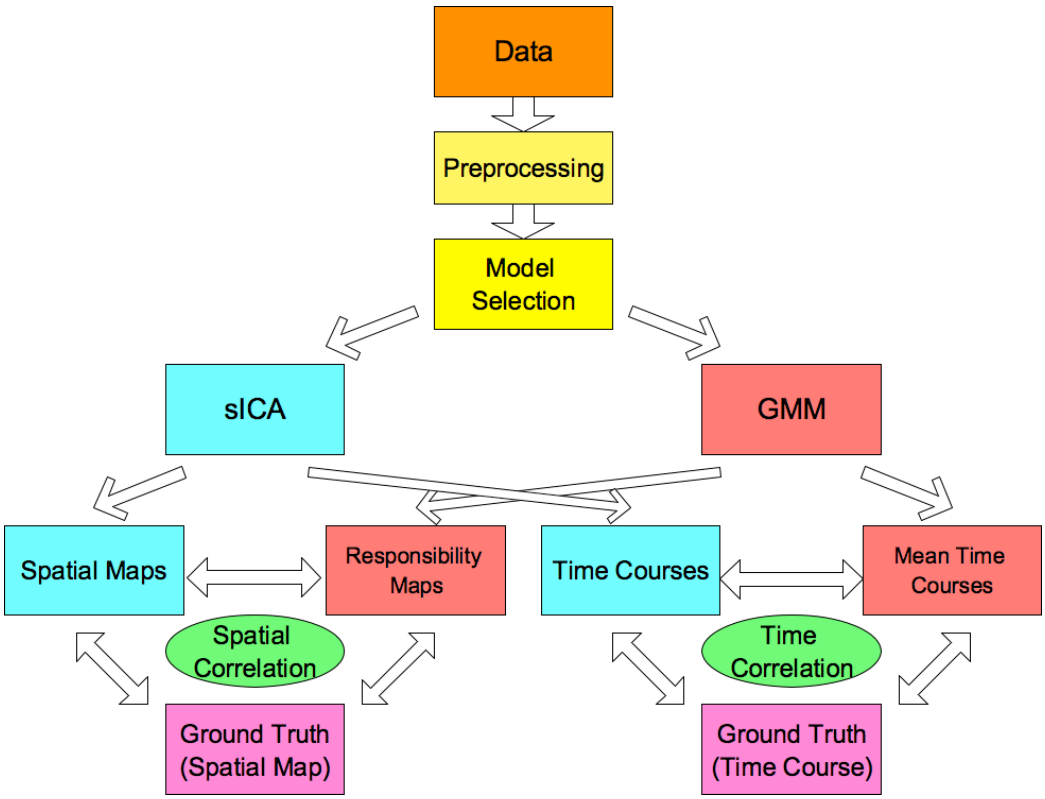


Figure 5-1: *Comparison scheme of ICA, GMM, and the ground truth. The ground truth is not available for real fMRI studies for components other than the ones corresponding to the experimental protocol.*

tinguishing different sources. The complete comparison scheme is shown in Fig. 5-1.

We first perform spatial independent component analysis (sICA) on a T x V dimensional data with the prespecified number of sources, K. As a result, we obtain K V-dimensional spatial maps and their associated time courses of length T. The objective of sICA is to have the sources represented as spatial maps that are as independent as possible. On the other hand, applying the data to Gaussian mixture model (GMM) with K components groups similar data time courses together into K clusters. As discussed in Chapter 4, GMM produces the T-dimensional mean time courses of each cluster and their responsibility maps of dimension V, which explain the probability that each voxel belongs to a particular cluster. The EM-based results of GMM depends on the initialization condition. We perform 10 runs of the algorithm using different random initializations and select the outcomes of the run which gives

the maximum likelihood of the data.

In this work, we propose a component-based comparison, which directly compares all components of sICA to their corresponding components in GMM. Since the ordering of components in each model is arbitrary, we need to match each component of one method to that of another. We investigate two methods for comparing the resulting components. In the first method, we first match each component of sICA to its corresponding Gaussian cluster by selecting the cluster whose mean time course is the most correlated with the sICA time course. This can be done by using correlation across time, defined as

$$c_t(i,j) = \frac{\mathbf{t}_i^T \mathbf{t}_j}{\sqrt{\mathbf{t}_i^T \mathbf{t}_i}\sqrt{\mathbf{t}_j^T \mathbf{t}_j}} \tag{5.1}$$

where i, j denote the index of components in sICA and GMM, respectively, and $\mathbf{t}$ is a mean-corrected time course. The matching between sICA and GMM is one-to-one. In case of a conflict, for example, where one component of sICA is claimed by multiple Gaussians, the time course of Gaussian which has the highest energy gets the priority.

After the matching is done, we compare the spatial maps of sICA to the responsibility maps of GMM. When interpreting the results, many studies first transform the spatial maps of sICA to z-scores to give pseudo-statistical interpretation to voxel values. However, it is also well-known that sICA spatial maps acquired from fMRI studies have a property that they are very sparse and non-overlapping. Only small portion of voxels have significantly high absolute coefficients whereas others have values near zero, similar to the structure of a binary map. This property enables a direct comparison of sICA spatial maps to responsibility maps of GMM, which also contains only small portion of voxels with probability near 1. The proximity of two components obtained by sICA and GMM is evaluated using the spatial correlation, defined as

$$c_s(i,j) = \frac{\mathbf{s}_i^T \mathbf{s}_j}{\sqrt{\mathbf{s}_i^T \mathbf{s}_i}\sqrt{\mathbf{s}_j^T \mathbf{s}_j}} \tag{5.2}$$

where i, j denote the index of components in sICA and GMM, respectively, and $\mathbf{s}$ is a mean-corrected spatial or responsibility map. The significance of this spatial correlation coefficient is described by its corresponding p-value. In some studies, sICA spatial maps are thresholded with a prespecified value, when they are compared to a GMM responsibility map [58].

This evaluation process can also swap the roles of time-correlation and space-correlation using this alternative definition. We first match each component based on the correlation of spatial maps and then compare the components using the time correlation of their time courses.

In the presence of ground truth about our data, we can easily extend this method to incorporate it, and perform a three-way comparison: between sICA and ground truth, between GMM and ground truth, and between sICA and GMM. The comparison results are examined to check which analysis method performs better under particular conditions of data and experiments. We examine each component and pay careful attention to non-task related "lower" components, which is where we expect the differences between ICA and GMM to arise. Task-related components estimated from ICA and GMM are usually very similar to each other.

5.2 Synthetic Data

This section contains a description of synthetic data and the results of the comparison scheme discussed in the previous section on influences of noise level and length of experiment on the resulting components.

5.2.1 Data Generation

Simulated synthetic data were generated to investigate the influence of noise level and length of a time course on the performance of ICA and GMM. Each data had the size of V = 5000 voxels and T = 300 time points. The simulated signals and noise used in this section are presented in Figure 5-2. Two types of signals (Signals A and B) were constructed to represent a consistently task-related (CTR) hemodynamic

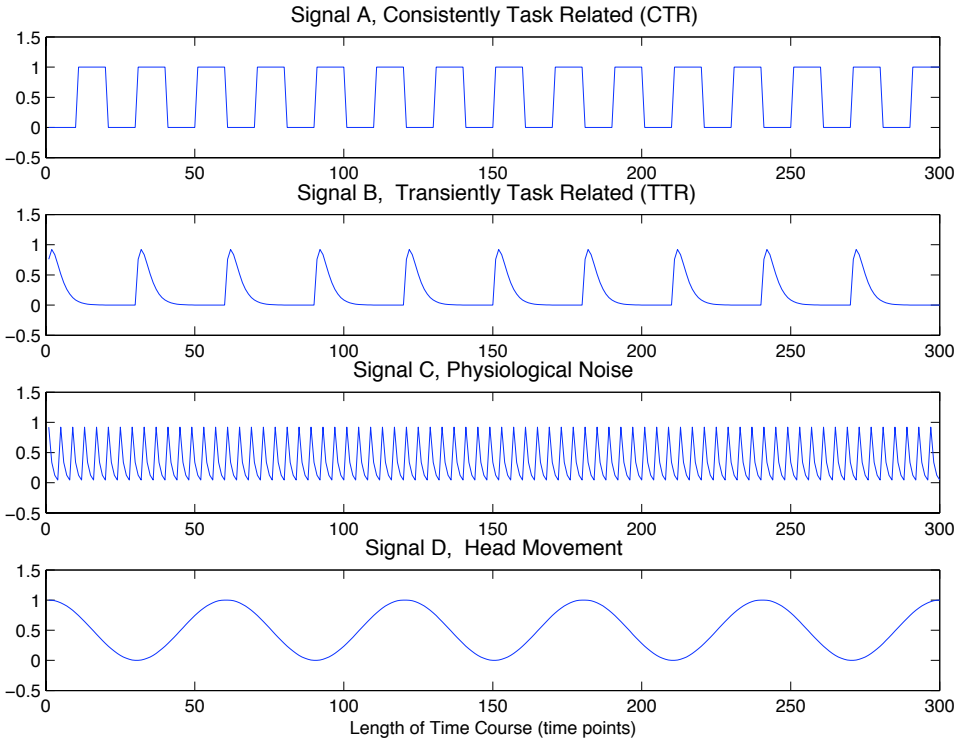


Figure 5-2: *Synthetic Data Signals. Signal A is consistently task related with block-design waveform. Signal B is generated as a gamma function to represent transiently task related component. Signal C is also a gamma function modeling physiological noise. Signal D is a sine wave simulating head motion.*

response and a transiently task-related (TTR) hemodynamic response of the brain, respectively. The CTR signal has a property that it is periodic and slowly varying in sync with the box-car experimental waveform. The experimental protocol was assumed to have an alternating pattern of ON and OFF, where each of them lasts over 10 time points. The TTR signal is also periodic, but is transient compared to the CTR signal. Here, it was constructed using the Gamma function and has the period of 30 time points. Note that the periods of CTR and TTR signals are not equal. If they were to be equal, then the rise of the signals will happen at the same time, which leads to the observation that two signals are not independent to each other while their corresponding spatial maps are still independent. Although independence in time is not required for sICA, we kept the sources independent both in space and

time in our synthetic data by assigning a different period to TTR compared to that of CTR, so that TTR is not in sync with the experimental protocol. Therefore, TTR in this case is not fully task related in time, but has the shape of a typical TTR signal. In section 5.2.2, we discussed the results when the independence in time no longer holds.

Non-tasked related signal sources were also generated. Signal C models the physiology-related noise such as heart beats. This periodic signal was constructed using the Gamma function with the period of 4 time points. Signal D simulates the motion-related signal such as the slow head movement. This type of signal usually varies very slowly with large transient. In order to preserve these properties, we chose a sine wave with the period of 60 time points to generate Signal D.

In typical fMRI data, the task related signals (Signals A and B) are corrupted by the presence of non-task related components such as Signals C and D and random noise. To make the comparison of these signals on the equal level, we set the maximum and minimum amplitudes of all signals to be 1 and 0, respectively, as shown in Figure 5-2. Another confounding source in fMRI data is the variation in the baseline magnetization of the scanner. However, we did not include it in our synthetic data set since its effect can easily be removed by detrending the data by fitting a low degree polynomial. Signals A, B, C, and D were added to 125 voxels each, where each type of signal source constitutes 2.5 percent of the total number of voxels in a volume. These regions were made non-overlapping in order to follow the non-overlapping and sparseness properties of fMRI spatial maps.

We added a Gaussian random noise to our data set to simulate a noisy environment. In our real fMRI studies, the estimated SNR is about 0.5. We generated a set of 14 synthetic data to investigate the effect of noise by controlling the variance of the Gaussian random noise, over a range of real SNR from 0.1 to 2.0, which corresponds to the estimated SNR of 0.18 to 3.7. More data were generated around the estimated SNR of our real fMRI data. In order to analyze the effect of the length of the time courses on the performance of ICA and GMM, we generated a set of 8 synthetic data with above signals over a range of time length from 50 to 300 time points for each

SNR value, with more data around 105 time points, which is the length of the time courses in our real fMRI study.

5.2.2 Effects of Noise on the Identified Components

In this section, we study the effects of noise on the performance of sICA and GMM. We applied sICA and GMM on 14 data sets of SNR from 0.1 to 2.0, which is a typical SNR range for fMRI data. Data were more finely sampled around SNR of 0.3, where a significant change of performances occurred, and it is near the estimated SNR of our real fMRI data. Aikake information criterion and Bayesian information criterion correctly approximated the number of sources as 5 (four types of signals plus random noise) for both sICA and GMM for all cases of SNRs.

Spatial Domain

We performed a component-wise comparison between sICA and GMM on the effect of the noise level in the spatial domain. Here, the estimated time courses from sICA and GMM were matched to each other by the similarity of their time courses, and the matched spatial maps of sICA were compared to their corresponding GMM responsibility maps using spatial correlation (ICA-GMM). Moreover, we conducted the same analysis between sICA and the ground truth (ICA-GT), and GMM and the ground truth (GMM-GT). Table 5.1 shows the average of the absolute values of correlation coefficient between resulting ICA and GMM components for their spatial maps and time courses for SNR = 0.1. The component numbers 1, 2, 3, 4, and 5 correspond to the consistently task-related component, head movement, physiological noise, transiently task-related component, and the noise component, respectively. As shown in the table, we observe that for a given component of one model, there is only one obviously corresponding component of the other model. Our matching scheme correctly selected the right pairs. The noise component of GMM (5) is highly correlated with all ICA components except the noise component, whereas the noise component of ICA (5) is uncorrelated with all GMM components. The matching

		GMM				
	#	1	2	3	4	5
ICA	1	0.6591	0.0253	0.0995	0.036	0.2757
	2	0.0146	0.4945	0.0238	0.0662	0.1991
	3	0.0317	0.0283	0.3753	0.135	0.1967
	4	0.0111	0.0382	0.0843	0.2507	0.1706
	5	0.0103	0.0257	0.0272	0.0716	0.0362

(a) Correlation Coefficients of Spatial Maps

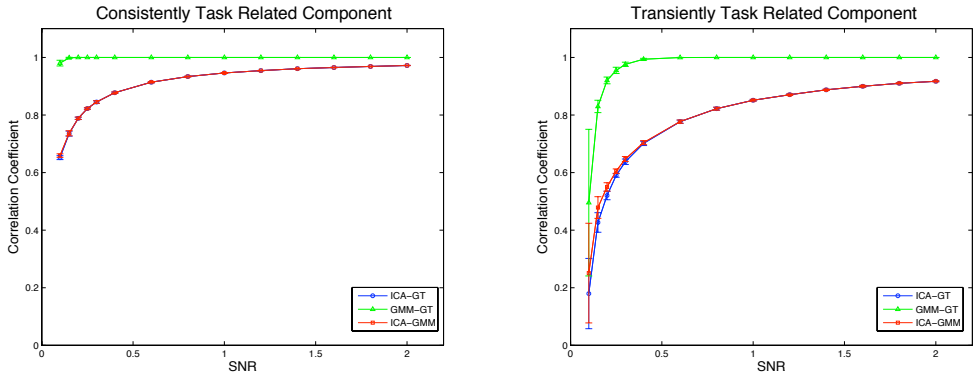
		GMM				
	#	1	2	3	4	5
ICA	1	0.9581	0.0494	0.2018	0.0674	0.5065
	2	0.0249	0.8364	0.0448	0.1109	0.335
	3	0.0585	0.048	0.6298	0.2213	0.3187
	4	0.0169	0.0626	0.1442	0.4113	0.2806
	5	0.0159	0.0411	0.0466	0.1211	0.0619

(b) Correlation Coefficients of Time Courses

Table 5.1: Average of the absolute values of correlations coefficient between ICA and GMM components for SNR = 0.1.

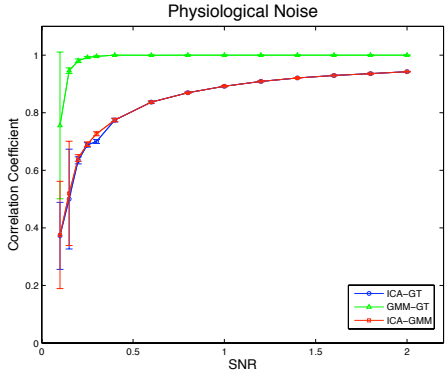
between components were even more obviously for higher SNRs.

The results of comparison in the spatial domain are shown in Figure 5-3. Each of the plots shows the performance of ICA and GMM on the consistently task related, transiently task related, physiological noise, and head movement components, respectively, when compared to the ground truth (ICA-GT, GMM-GT) and to each other (ICA-GMM). GMM outperformed ICA when compared to the ground truth for the entire range of SNR. For this set of synthetic data examples, the spatial maps of the ground truth were almost perfectly retrieved by the responsibility maps of GMM. Although they are not as good as the results of GMM, almost all of the spatial correlation coefficients obtained by sICA had p-values less than 0.005. In addition, for each method, the accuracy of estimates for each component was ordered in the following manner (from the best to the worst): the consistently task-related component, head movement, physiological noise, and transiently task-related component. This order is identical to the reversed order of the estimated kurtosis of the time courses of the

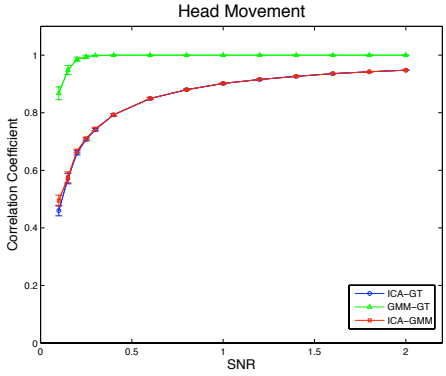


(a) Consistently task-related component

(b) Transiently task-related component



(c) Physiological Noise

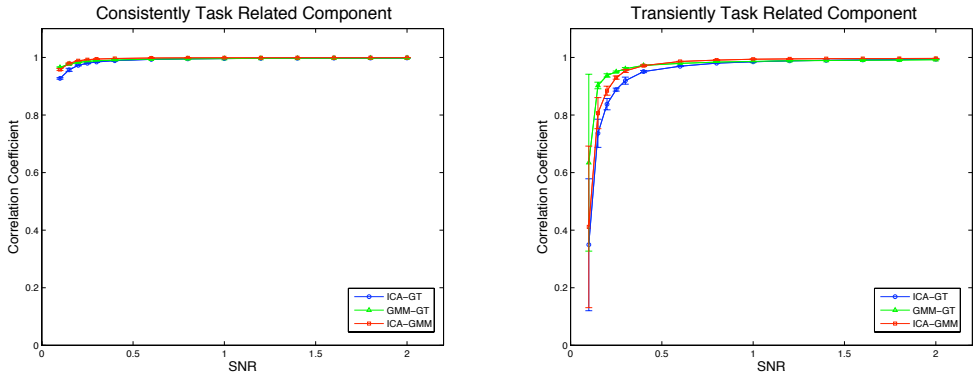


(d) Head Movement

Figure 5-3: Component-wise comparison of the effect of noise level on the spatial correlation of the estimated sICA and GMM maps with ground truth and with each other. T = 300 time points. Error bars come from ten independent repeats.

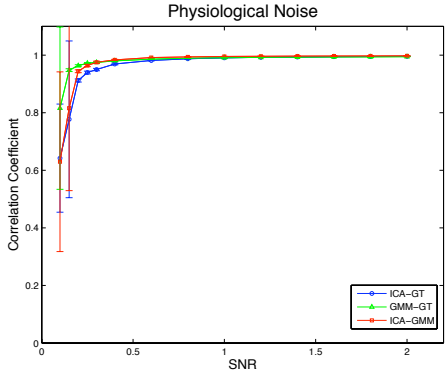
ground truth, where CTR has the lowest (sub-Gaussian) and TTR has the highest (super-Gaussian) kurtosis value among all signal types.

The superior performance of GMM over sICA also persisted under a slightly different setting of our synthetic data. In a new data set, we simulated the transiently task-related component to have the same frequency with CTR. sICA no longer separated the two corresponding regions of voxels and their associated time courses with five components although sICA managed to separate them when we increase the number of components. On the other hand, GMM, which only considers the shape of time signals, still separated CTR from TTR, with only five clusters as before.

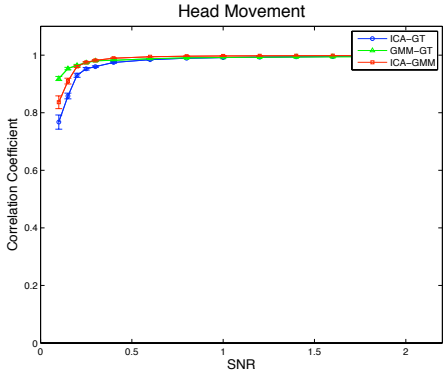


(a) Consistently task-related component

(b) Transiently task-related component



(c) Physiological Noise



(d) Head Movement

Figure 5-4: Component-wise comparison of the effect of noise level on the time correlation of the estimated sICA and GMM time courses with ground truth and with each other. T = 300 time points. Error bars come from ten independent repeats.

Time Domain

Similarly, we performed a component-wise comparison between sICA and GMM in the time domain. Here, the estimated spatial maps of sICA and GMM responsibility maps were matched to each other by their spatial correlations, and the matched time courses of sICA and GMM were compared to each other using time correlation (ICA-GMM). Moreover, we conducted the same analysis between sICA and the ground truth (ICA-GT), and GMM and the ground truth (GMM-GT).

The results of comparison in the time domain are shown in Figure 5-4. The

plots each show the performance of ICA and GMM on consistently task related, transiently task related, physiological noise, head movement components, respectively, when compared to the ground truth (ICA-GT, GMM-GT) and to each other (ICA-GMM). Identical to the results obtained in the spatial domain, for both ICA and GMM, the performance of each component was ordered in the following manner (from the best to the worst): consistently task-related component, head movement, physiological noise, and transiently task-related component.

The average of the time correlation coefficients was higher than that of the spatial correlation coefficients. This leads to a conclusion that the estimated time courses from sICA and GMM were closer to the ground truth time courses than the estimated spatial maps were to the ground truth in space. Furthermore, the difference of the time correlation coefficients between sICA and GMM time courses was much smaller than that of the spatial correlation coefficients of their spatial maps. This implies that sICA and GMM generated very similar time courses, but different spatial maps.

Figure 5-5 shows a zoomed plot of consistently task related component in time over a range SNR from 0.3 to 2. When compared to the ground truth, both methods performed extremely well. Here, we observe an interesting behavior. For SNR values below 1, GMM clearly outperforms ICA. However, for SNR above 1, we observe that ICA outperforms GMM. For SNR greater than 1.4, the range of time correlation coefficients of ICA was even outside the margin of error of the correlation coefficients of GMM. This behavior was also observed in all the other components of the simulated data.

5.2.3 Effect of the Length of Experiment on the Identified Components

In this section, we study the effects of length of the time courses (T) of data on the performance of sICA and GMM. We applied sICA and GMM on 8 data sets of T from 50 to 300 time points, which is a typical range for fMRI data. Data were more finely sampled around T = 105 time points, where significant change of performances

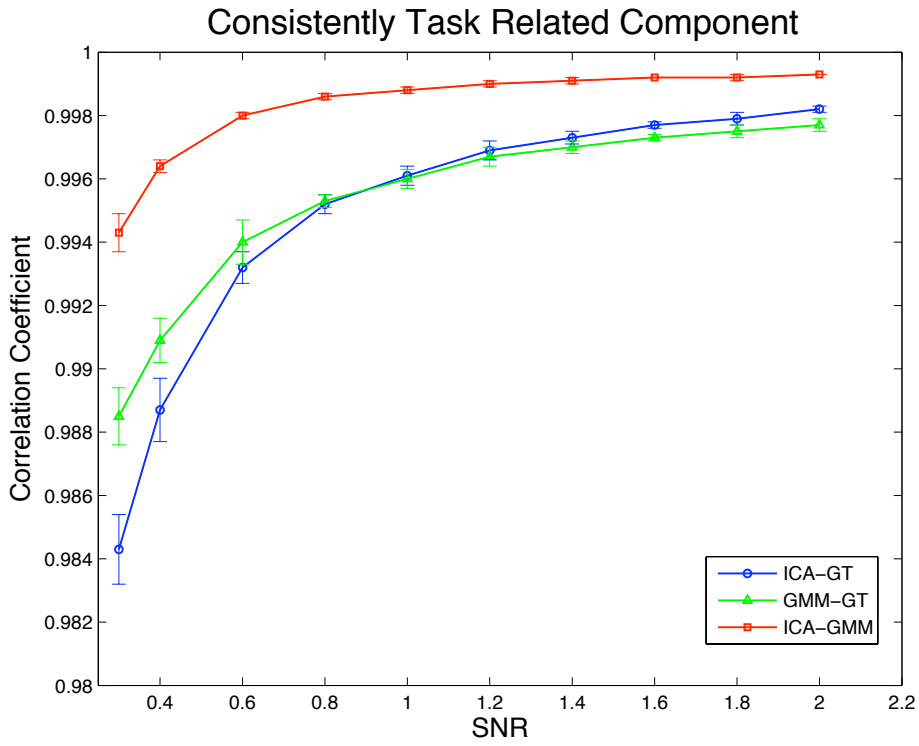


Figure 5-5: *Zoomed plot of consistently task related component in time. Below SNR = 1, GMM outperforms ICA. Above SNR = 1, ICA outperforms GMM. Error bars come from ten independent repeats.*

occurred, and it is also the length of the time course of our real data. The length of time courses corresponds to the number of samples for sICA and the dimension of each Gaussian for GMM. For a data with SNR = 0.3, Aikake information criterion and Bayesian information criterion correctly approximated the number of sources as 5 (four types of signals plus random noise) for both sICA and GMM.

Spatial Domain

We performed a component-wise comparison between sICA and GMM on the effect of the length of time courses in the spatial domain. The estimated time courses from sICA and GMM were matched to each other by the similarity of their time intensity profiles, and we compared the matched spatial maps of sICA to their corresponding GMM responsibility maps using spatial correlation (ICA-GMM). Moreover, we con-

		GMM				
	#	1	2	3	4	5
ICA	1	0.5797	0.1021	0.1314	0.0823	0.261
	2	0.033	0.2948	0.1397	0.1521	0.2184
	3	0.0351	0.1785	0.3468	0.0961	0.2353
	4	0.022	0.1001	0.1119	0.3691	0.2339
	5	0.0155	0.0462	0.1034	0.0665	0.072

(a) Correlation Coefficients of Spatial Maps

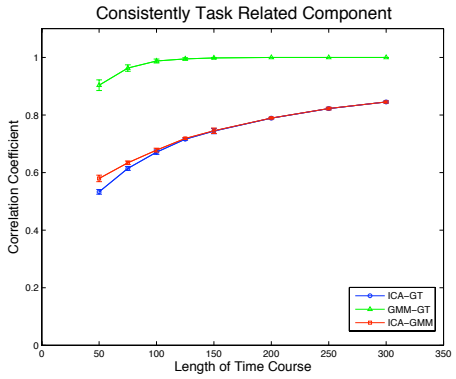
		GMM				
	#	1	2	3	4	5
ICA	1	0.9694	0.2361	0.2786	0.1628	0.4628
	2	0.0848	0.5332	0.2498	0.268	0.3718
	3	0.0828	0.3267	0.5753	0.1622	0.4008
	4	0.0469	0.1788	0.1989	0.5781	0.4162
	5	0.0338	0.0887	0.1822	0.1062	0.1292

(b) Correlation Coefficients of Time Courses

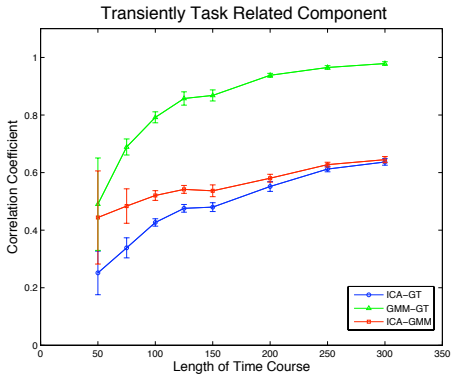
Table 5.2: Average of the absolute values of correlations coefficient between ICA and GMM components for T = 50.

ducted the same analysis between sICA and the ground truth (ICA-GT), and GMM and the ground truth (GMM-GT). Table 5.2 shows the average of the absolute values of correlation coefficient between resulting ICA and GMM components for their spatial maps and time courses for T = 50. The component numbers 1, 2, 3, 4, and 5 correspond to the consistently task-related component, head movement, physiological noise, transiently task-related component, and the noise component, respectively. As in the case with SNRs, we observe that for a given component of one model, there is only one obviously corresponding component of the other model. Our matching scheme correctly selected the right pairs. The noise component of GMM (5) is highly correlated with all ICA components except the noise component, whereas the noise component of ICA (5) is uncorrelated with all GMM components. The matching between components were even more obviously for longer time courses.

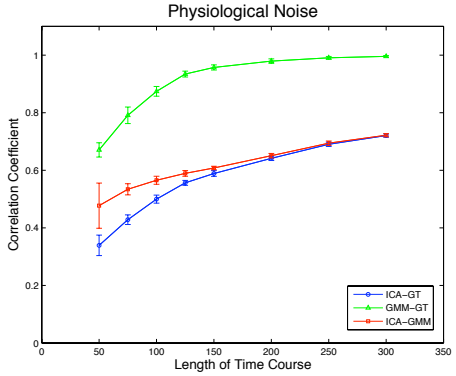
The results of comparison in the spatial domain are shown in Figure 5-6 for the fixed value of SNR = 0.3. Each of the plots shows the performance of ICA



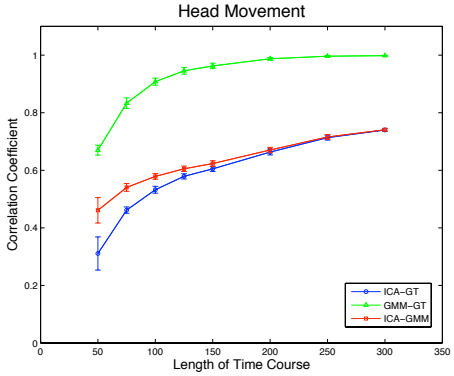
(a) Consistently task-related component



(b) Transiently task-related component



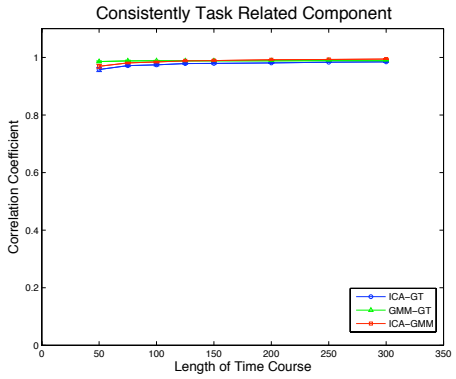
(c) Physiological Noise



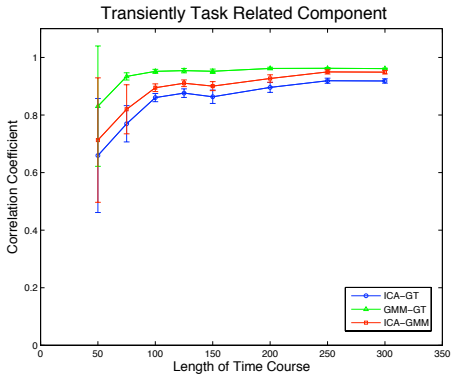
(d) Head Movement

Figure 5-6: Component-wise comparison of the effect of length of time courses on the spatial correlation of the estimated sICA and GMM spatial maps with ground truth and with each other. SNR = 0.3. Error bars come from ten independent repeats.

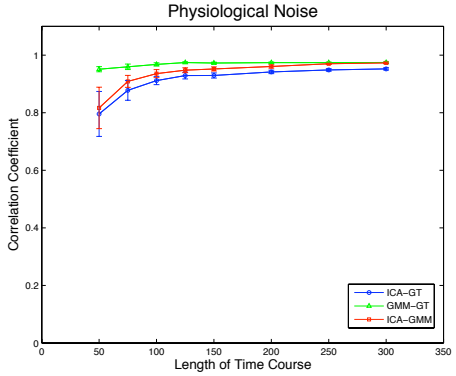
and GMM on consistently task related, transiently task related, physiological, and motion-related components, respectively, when compared to the ground truth (ICA-GT, GMM-GT) and to each other (ICA-GMM). We observe that GMM outperformed ICA when compared to the ground truth over the entire range of values of T. Although they were not as good as the results of GMM, almost all of the spatial correlation coefficients obtained by sICA had p-values less than 0.005. Both sICA and GMM performed well with a small margin of error for Ts longer than 100 time points, when compared to the ground truth. Consistent with the results of variable SNR from the previous section, for each method, the performance of each component was ordered



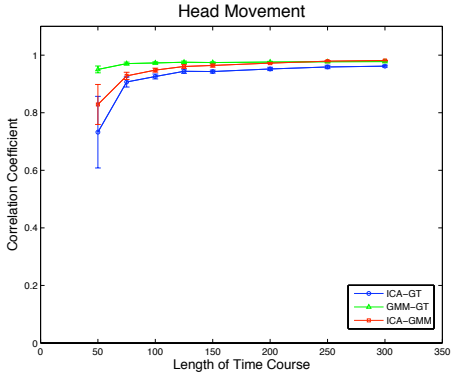
(a) Consistently task-related component



(b) Transiently task-related component



(c) Physiological Noise



(d) Head Movement

Figure 5-7: Component-wise comparison of the effect of length of time courses on the time correlation of the estimated sICA and GMM time courses with ground truth and with each other. SNR = 0.3. Error bars come from ten independent repeats.

in the following manner (from the best to the worst): consistently task-related component, head movement, physiological noise, and transiently task-related component.

Time Domain

Similar to the analysis in the previous section, we performed a component-wise component comparison between sICA and GMM in the time domain. The results of comparison in the time domain are shown in Figure 5-7 for the fixed value of SNR = 0.3. The plots each show the performance of ICA and GMM on the consistently task related, transiently task related, physiological noise, and motion-related components,

respectively, when compared to the ground truth (ICA-GT, GMM-GT) and to each other (ICA-GMM). Again, we observed that GMM outperformed ICA when compared to the ground truth over the entire range of values for T. This is different from the analysis in the time domain with the variable SNRs where ICA outperformed GMM for SNRs over a certain threshold. Although they are not as good as the results of GMM, almost all of the time correlation coefficients obtained by sICA had p-values less than 0.005. Both sICA and GMM performed well with a small margin of error for T longer than 100 time points, when compared to the ground truth.

Similar to the analysis in Section 5.2.2, the average of the time correlation coefficients was higher than that of the spatial correlation coefficients. This again leads to a conclusion that the estimated time courses from sICA and GMM were closer to the ground truth time courses than the estimated spatial maps were to the ground truth in space. Furthermore, the difference of the time correlation coefficients between sICA and GMM time courses was much smaller than that of the spatial correlation coefficients of their spatial maps. Again, it implies that different spatial maps were estimated by sICA and GMM with very similar time courses.

It is a well-known notion in fMRI analysis that better analysis results can be achieved by averaging over multiple runs of experiments as it improves the SNR of the data. To test this, we divided our data of SNR = 0.3 and T = 300 time points into 3 pieces over time and averaged them over. With this new averaged data of T = 100 time points, the performance of the consistently task related component was better (2 standard deviations above) than that of T = 300 time points. However, by averaging the data, we lose our ability to analyze all of the other simulated components in the data. Since the performance of sICA and GMM on the non-averaged data with T = 100 time points was already outstanding and not too distant from that of T = 300 time points, we conclude that it is better to use the non-averaged data for our purpose and subsequent analysis with our real fMRI studies.

5.3 Real fMRI Experiments

This section contains a description of the real fMRI study, which we used to compare the performance of ICA and GMM on the resulting components. With the real fMRI data, we present another way of choosing the optimal number of total components based on the comparison between the results of ICA and GMM to the ground truth. Furthermore, in the absence of ground truth, we propose a way of selecting a threshold to determine which pairs of ICA and GMM components are meaningful to compare using their correlation coefficient matrices.

5.3.1 Description of Data

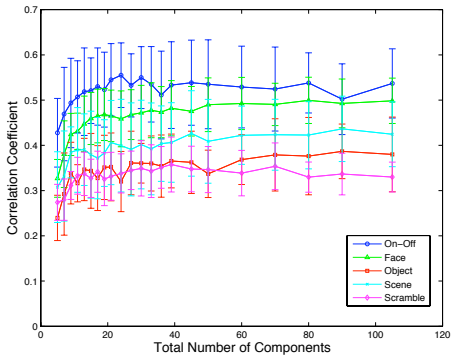
To compare the performance of ICA and GMM, we used a set of fMRI scans obtained during a visual object recognition task for high level vision conducted by Professor Nancy Kanwisher's group in the Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology. Each experiment consisted of five rest epochs and four task epochs. With the TR of 3 seconds, each rest epoch contained five time points. In the rest condition, the subjects were instructed not to move and to concentrate on the noise of the scanner. In the task condition, the subjects were presented with a series of pre-recorded visual images. Each task epoch consisted of a series of four different categories of visual stimuli: faces, objects, scenes, and scrambles. Each task category lasted over five time points, making each task epoch contain twenty time points. Within each task epoch, the ordering of the four categories were randomized to minimize the effect of ordering in the analysis. Experiments were repeated eight times for each subject (eight runs per subject). The original study contained eight subjects, but for the purpose of component-wise comparison of ICA and GMM, we present the results for one subject. Furthermore, we only included the voxels in the brain with the mask of the union of fusiform face area (FFA), parahippocampal place area (PPA), lateral occipital complex (LO), and areas which activated significantly compared to the fixation. FFA, PPA, and LO are known to be responsible for face processing, place processing (scenes, houses), and object/shape processing,

respectively. Our data had a size of T = 105 time points and V = 9703 voxels. Other preprocessing steps included motion and time-correction and Gaussian smoothing. We further subtracted the mean both across time and space prior to our analysis. Within the areas contained in the mask, the estimated SNR was 0.4.

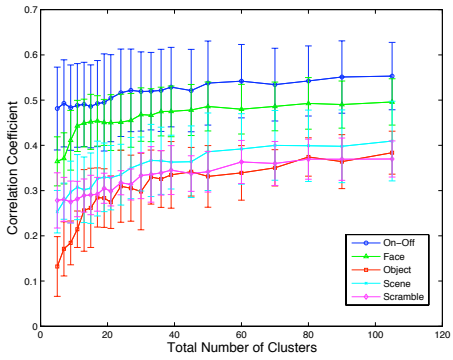
5.3.2 Comparison on Task-related Components

With the presence of the experimental protocol, we devised a set of pseudo ground truth for our fMRI data, which consist of eleven types of box-car waveforms. The first type is the simple contrast between the rest condition (0) and the task condition (1). We also made four types of pseudo ground truth, which are category-specific task-related functions in the form of an image category (faces, objects, scenes, or scrambled images) (1) versus all of the other conditions (0), including the rest epoch and the other image categories. In addition, we designed six contrast functions between the image categories: face vs. object, face vs. scene, face vs. scramble, object vs. scene, object vs. scramble, and scene vs. scramble. One of the contrasting categories was assigned a value of 1, whereas the other was given -1. All of the other conditions had a value of 0. We did not make additional opposite contrast (for example, object vs. face) functions because the resulting correlation coefficient would merely have the opposite sign of the same magnitude (compared to, for example, face vs. object) when the contrast function is correlated with a GMM or ICA time course. All of the box-car waveforms here were convolved with the estimated hemodynamic function presented in Chapter 2 and were 105 time points long.

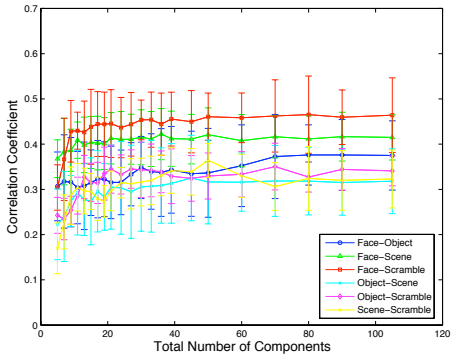
Over a range of the total number of components, K, from 5 to 105, we performed ICA and GMM. For a given K, we correlated our ground truth to all of the resulting time courses of ICA and GMM. Then, for each type of our ground truth, we selected the corresponding ICA and GMM components which had the highest correlation coefficient. We repeated this procedure over the range of K over eight runs to track each model's ability to identify the task-related and category-contrasting components, and to find an optimal K^* for this purpose. The results are shown in Figure 5-8. Figures (a) and (b) each describes the correlation coefficients of the best matched



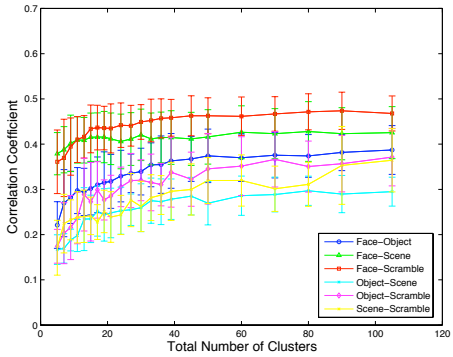
(a) ICA, Task-related.



(b) GMM, Task-related.



(c) ICA, Contrast-related.



(d) GMM, Contrast-related.

Figure 5-8: Comparison of the ground truth to the best matched ICA and GMM components in time for different total number of components.

task-related components for ICA and GMM, respectively. Similarly, figures (c) and (d) show the correlation coefficients of the best matched contrast-related components for ICA and GMM, respectively.

For each model, the resulting components were highly correlated with our ground truth in the order of the face, scene, object, and scramble task-related ground truth components. Furthermore, the contrast functions which involve the face category were most correlated with the estimated ICA and GMM time courses. Comparing the figures (a) to (b) and (c) to (d), we found that there is no big difference in the identifiability of task-related and contrast-related components between ICA and

GMM for a large K. However, there exist several differences between the models. For K less than 10, GMM time courses had better correspondences with the task-related and contrast-related ground truth, except the ones involving the object category. This is largely in agreement with our results on the synthetic data in the previous section where K was 5.

The "elbow" values of K, after which point the correlation coefficients do not increase significantly for larger values of K, are about 15 for ICA and 30 for GMM. This implies that ICA can capture the category specific task and contrast-related components with a smaller number of the total components than GMM. This is contrary to the findings based on the Bayesian information criterion, which suggested a larger required number of total components (79) for ICA than that of GMM (51) to explain the entire data the best. This contrast indicates that ICA requires a smaller number of K to extract the task and contrast-related components, but also needs a large K to describe the entire data, whereas the difference between those values is small for GMM. In other words, in order to best describe the data, ICA needs to dedicate a large number of components to model non-task-related components. This can in part be explained by the non-Gaussianity assumption of ICA components that the noise information in the data, which are in many cases assumed as Gaussian, is unable to be modeled by a single or a small number of ICA components and thus is broken into many ICA components. On the other hand, GMM is able to model Gaussian noise with one or a few more of its Gaussian components. The "elbow" values are used in the component-wise comparison between ICA and GMM in the next section.

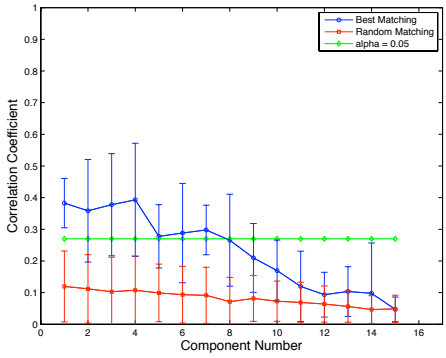
5.3.3 Component-wise Comparison between ICA and GMM

We conducted the component-wise comparison between ICA and GMM on our real fMRI data for five values of the pre-specified total number of components (K = 15, 30, 50, 80, 105). K = 15 and 30 were approximately the optimal total number of components suggested by the "elbow" information for extracting the task-related and contrast-related components by ICA and GMM, respectively, in the previous section.

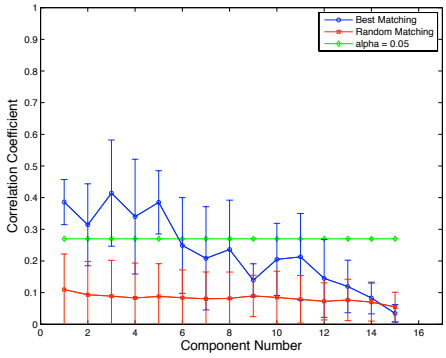
$K = 50$ and 80 were approximately the suggested number of total components for GMM and ICA based on the Bayesian information criterion. $K = 105$ is equivalent to performing the full spatial ICA without any data reduction. The comparisons were done following the scheme presented in Section 5.1.2.

Table A.1 is the matrix of the average of the absolute values of correlation coefficients between the ICA spatial maps and the GMM responsibility maps for $K = 15$ over eight runs. Similarly, Table A.2 shows the matrix of the average of the absolute values of correlation coefficients between the ICA and GMM time courses. Unlike the synthetic data case, the matching between the components of ICA and GMM is more difficult with the real fMRI data, because, for some components of ICA, there exist more than one corresponding GMM components that are highly correlated with, and vice versa. However, comparing the matched time courses with our ground truth presented in the previous section, the matches based on spatial maps and time courses are largely in agreement. This verification becomes significantly more difficult for larger number of components.

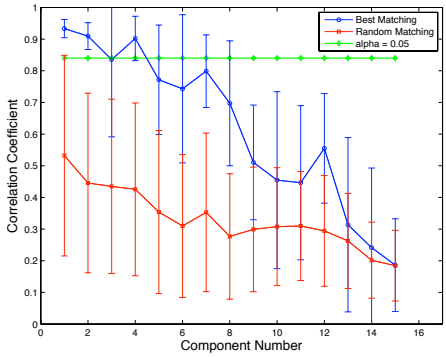
Figure 5-9 shows the comparison results for $K = 15$. Figure (a) shows the correlation coefficients between the spatial maps of ICA and the responsibility maps of GMM (blue line). We first sorted the ICA components by energy, and then matched GMM components to those of ICA by the correlation between their time courses with respect to the order of ICA. In other words, we found the best matching component of GMM for the first component of ICA, excluded that GMM component from the selection pool, and then repeated the matching procedure for the next component of ICA. While the blue line shows a series of correlation coefficients of spatial maps of the best matched components of ICA and GMM, the red line shows the correlation coefficients when we randomly matched the ICA components to those of GMM, still on one-to-one basis. For instance, the first component of ICA was randomly matched to a GMM component, the second ICA component was also randomly assigned a GMM component (excluding the one that already has been picked by the previous ICA components), and so forth. Essentially, each value of the red line shows the average of the correlation coefficients to all GMM components for a given ICA com-



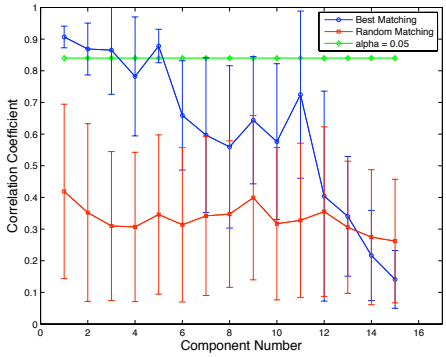
(a) ICA vs. GMM, Spatial Maps.



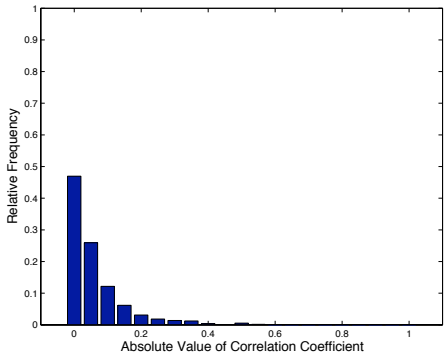
(b) GMM vs. ICA, Spatial Maps.



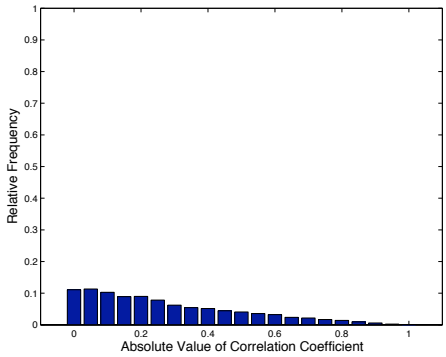
(c) ICA vs. GMM, Time Courses.



(d) GMM vs. ICA, Time Courses.



(e) Histogram of correlation coefficients. Spatial Maps.



(f) Histogram of correlation coefficients. Time Courses.

Figure 5-9: Component-wise comparison of ICA and GMM with $K = 15$ components.

ponent. The bigger the gap between the blue and red lines is, the more reliable the match between the components of ICA and GMM is. For "lower" components of ICA, we observe that two lines overlap. This implies that the matched GMM component was not optimal for a given ICA component due to the fact the the best matches for that ICA component were already taken by the "higher" order ICA components. However, we also observe that the error bars of the blue and red lines do not overlap for the top half components. Based on the correlation matrices of spatial maps of size K x K over eight runs ($K = 15$, in this case), we built a histogram of all correlation coefficient values, shown in Figure (e). The green line in Figure (a) and (b) is the value of the correlation coefficient in space where the empirical cumulative distribution function of the histogram reaches 0.95 ($\alpha = 0.05$).

From Table A.1, we notice that the correlation matrix is not symmetrical as the correlation coefficient between the i'th component of ICA and the j'th component of GMM is different from that between the j'th component of ICA and the i'th component of GMM. Similar to Figure (a), Figure (b) shows the results when the matching was done with respect to the order of GMM components. Figure (c) presents the results when we matched the components using their spatial maps and compared the models using their associated time courses. As in Figure (a), the matching was based on the ICA components. Figure (d) is similar to Figure (c) except that the matching was done with respect to the order of GMM. Figure (f) shows the histogram of all correlation coefficients between ICA and GMM time courses. The green line in Figure (c) and (d) is the value of the correlation coefficient in time where the empirical cumulative distribution function of the histogram reaches 0.95 ($\alpha = 0.05$).

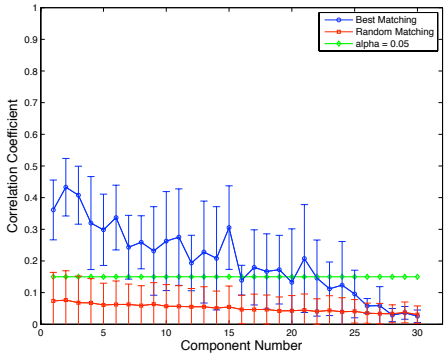
The green line, i.e. the correlation coefficient value which corresponds to $\alpha = 0.05$, is used as a threshold to claim that the component pairs of ICA and GMM, which have higher correlation coefficients than the threshold, are meaningful matches between the models and also that it may not be meaningful to compare other pairs of components (where the blue line is lower than the green line). Based on this threshold, we can claim that only about top six out of the fifteen components are meaningful to compare between ICA and GMM. As mentioned previously, for "lower" components of ICA,

we observe that the blue and red lines overlap implying that the matching was not optimal for those components. However, this overlap is only observed only for the last few components, which we claimed that the match between those components were not relevant.

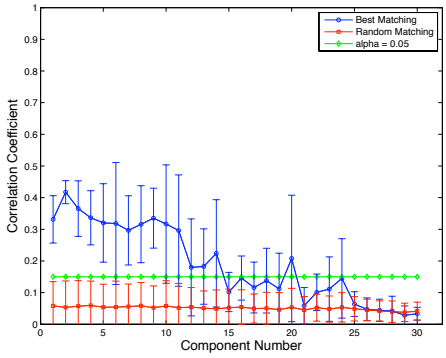
Figure 5-10, Figure 5-11, Figure 5-12, and Figure 5-13 show the results of our component-wise comparison between ICA and GMM for $K = 30$, 50, 80, and 105, respectively. In general, the threshold values become smaller as the total number of components becomes larger. From Figure 5-10, we observe that approximately the top 17 out the 30 spatial components and 10 of the 30 time components were meaningful to compare. Similarly, from Figure 5-11, approximately the top 27 out of the 50 spatial components and the top 17 time components were selected to be relevant.

On the other hand, we notice a slightly different behaviors for larger Ks. The error bars of the red line begin including the blue line overlap for components approximately after the 30th component. This implies that our one-to-one matching scheme may not be effective when we conduct ICA and GMM for a large number of the total components ($K > 30$). Based on Figure 5-12, we observe that approximately the top 45 out the 80 spatial components and 25 of the 80 time components were meaningful to compare. Furthermore, from Figure 5-13, we selected approximately the top 45 out the 105 spatial components and 32 of the 105 time components as relevant comparisons.

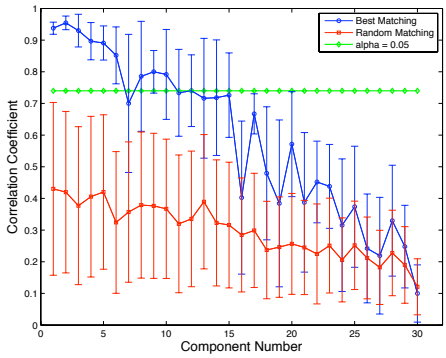
All in all, based on our results and taking the lower value of the spatial and time components to be conservative, we conclude, in general, that approximately the top third of the total components are meaningful to compare with our matching and comparison scheme.



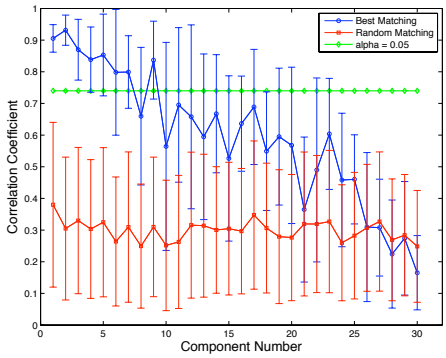
(a) ICA vs. GMM, Spatial Maps.



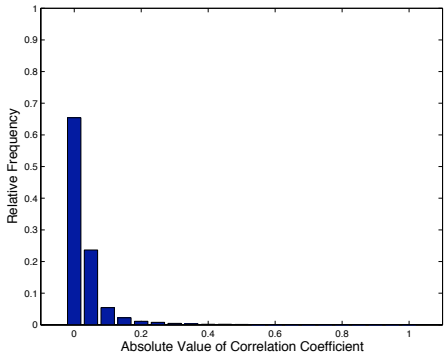
(b) GMM vs. ICA, Spatial Maps.



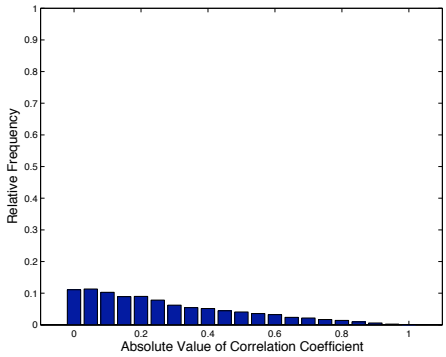
(c) ICA vs. GMM, Time Courses.



(d) GMM vs. ICA, Time Courses.

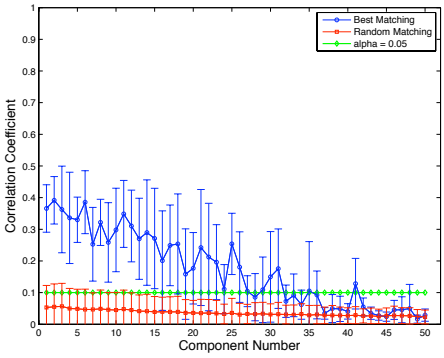


(e) Histogram of correlation coefficients. Spatial Maps.

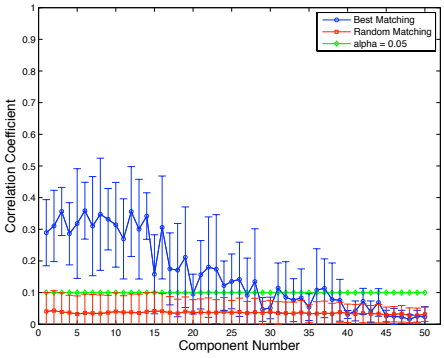


(f) Histogram of correlation coefficients. Time Courses.

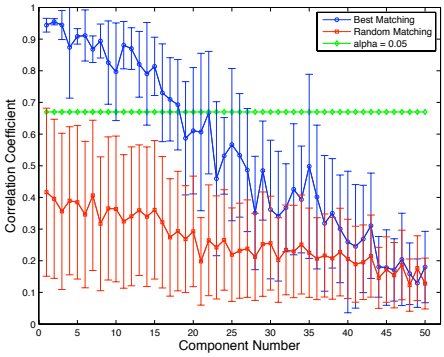
Figure 5-10: Component-wise comparison of ICA and GMM with $K = 30$ components.



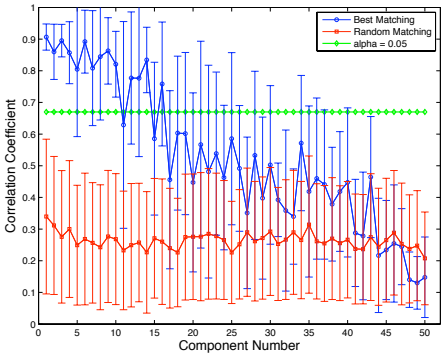
(a) ICA vs. GMM, Spatial Maps.



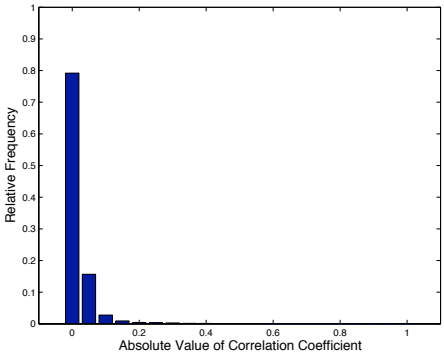
(b) GMM vs. ICA, Spatial Maps.



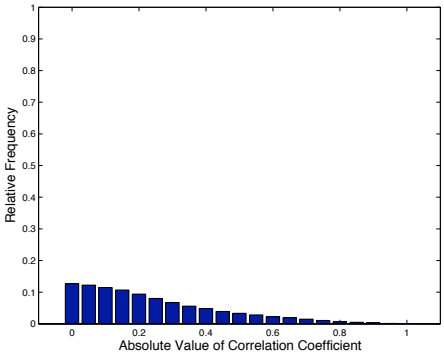
(c) ICA vs. GMM, Time Courses.



(d) GMM vs. ICA, Time Courses.

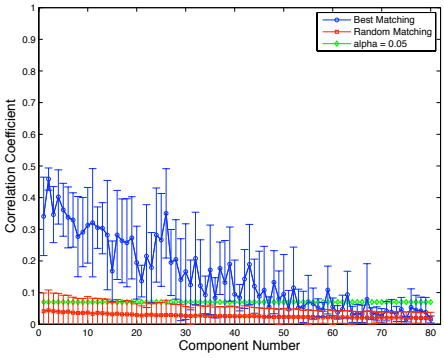


(e) Histogram of correlation coefficients. Spatial Maps.

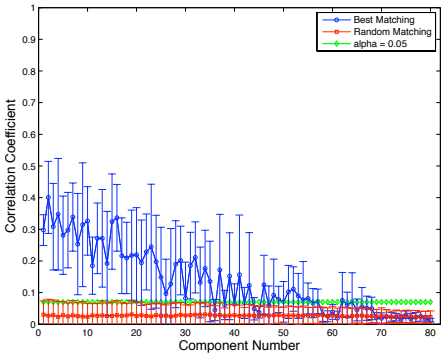


(f) Histogram of correlation coefficients. Time Courses.

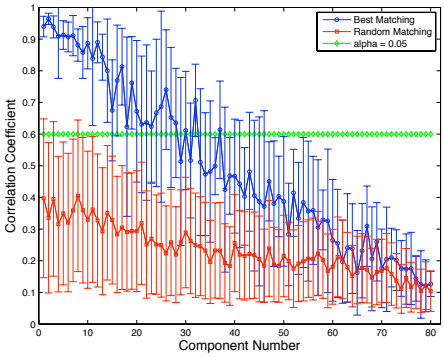
Figure 5-11: Component-wise comparison of ICA and GMM with $K = 50$ components.



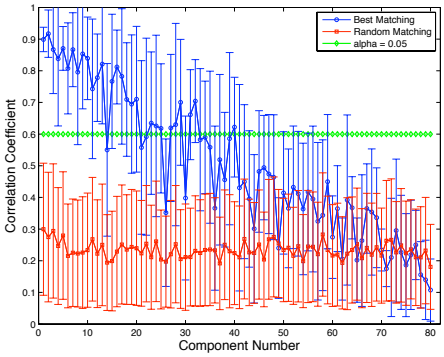
(a) ICA vs. GMM, Spatial Maps.



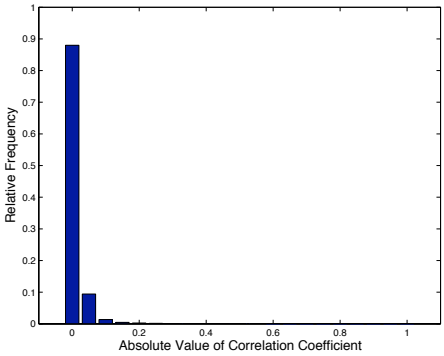
(b) GMM vs. ICA, Spatial Maps.



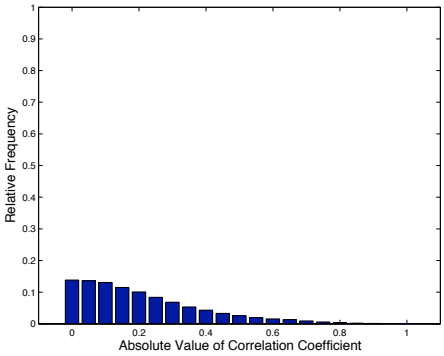
(c) ICA vs. GMM, Time Courses.



(d) GMM vs. ICA, Time Courses.

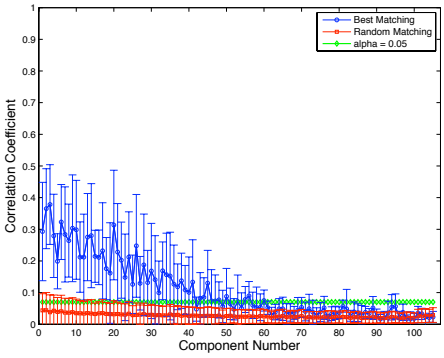


(e) Histogram of correlation coefficients. Spatial Maps.

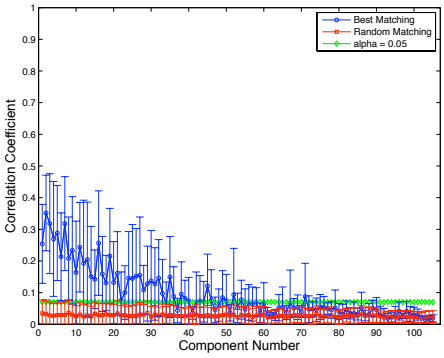


(f) Histogram of correlation coefficients. Time Courses.

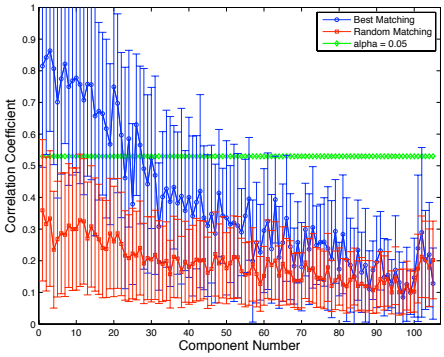
Figure 5-12: Component-wise comparison of ICA and GMM with $K = 80$ components.



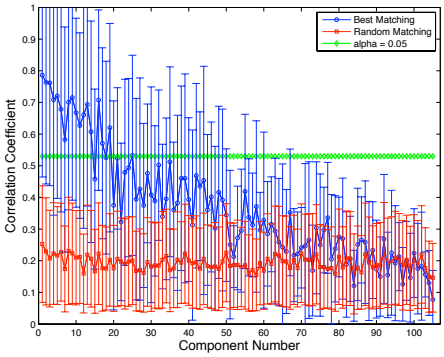
(a) ICA vs. GMM, Spatial Maps.



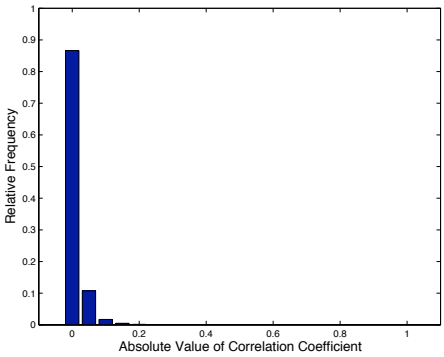
(b) GMM vs. ICA, Spatial Maps.



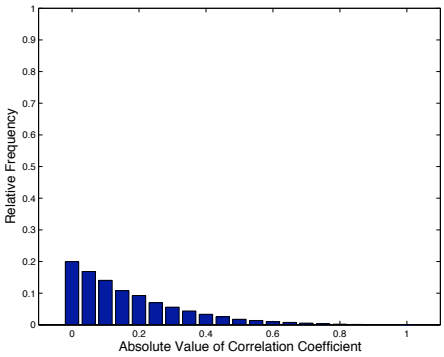
(c) ICA vs. GMM, Time Courses.



(d) GMM vs. ICA, Time Courses.



(e) Histogram of correlation coefficients. Spatial Maps.



(f) Histogram of correlation coefficients. Time Courses.

Figure 5-13: Component-wise comparison of ICA and GMM with $K = 105$ components.

5.4 Summary

In this chapter, we introduced our component-wise comparison scheme to compare the performance of ICA and GMM on identifying the functional connectivity. We applied this scheme on the synthetic data and investigate the influence of noise and length of time course on the performance of ICA and GMM. We further extended the analysis of our comparison scheme to a visual recognition real fMRI data. In addition, we proposed an alternate method of selecting the optimal total number of components for ICA and GMM when the goal was to extract the task and contrast-related components. In the next chapter, we discuss the pros and cons of our comparison scheme based on the results from the synthetic and real fMRI data.

Chapter 6

Discussion and Conclusions

This thesis reviewed several representative data-driven analysis techniques for identifying functional connectivity in fMRI and presented a component-wise matching and comparison scheme of resulting ICA and GMM components using their correlation. We investigated the effectiveness of this comparison scheme using synthetic and real fMRI studies. We found in both synthetic and real data that GMM outperforms ICA when the pre-specified total number of components in each model was less than 10.

There remain several points where our comparison scheme can improve and needs to be further examined. First of all, prior to matching the components of ICA and GMM, the components in each model were ordered in terms of their energy. Since it is not necessary that the most significant components have the highest energy, we could also incorporate kurtosis and the size of the activations in each component when ordering the components. With the current scheme, when we match the components of one model with respect to those of the other model, we start the one-to-one matching from the component of the other model which has the highest energy. Another way of matching which should be tried in the future to relax this strict order of matching is to use the bipartite graph matching algorithm [18], such as used in the marriage problem [33], using the correlation matrix of the components.

Furthermore, in experiments where the pre-specified total number of components in each model was large, we observed that a component in one model can be highly correlated with multiple components of the other model. This arises partly due

to the phenomenon that what used to be explained by one component when the total number of components was small is now more finely separated into multiple components. There is no theoretical analysis of when this breakdown occurs in each model. When multiple components in one model correspond to a single component in the other model, our component-wise comparison scheme might not be turn out to be optimal.

On a visual recognition real fMRI data, we proposed a method of choosing a threshold to determine which of resulting components of ICA and GMM are meaningful to compare using the cumulative distribution function of their empirical correlations. To complement this approach for selecting components that are valid to compare, we could incorporate permutation testing [31]. By randomly permuting the orders of images and voxels in the data and then performing our comparison scheme on this new data with many iterations, we could obtain another set of measure in which we can test the effectiveness of our comparison scheme. In addition, we plan to further apply and examine our comparison scheme to other variations of ICA algorithms such as the Fixed-Point algorithm and other clustering methods in the near future.

With ever increasing volume of complex experimental fMRI data, we hope that researchers in the field will find our empirical helpful in understanding and assessing the similarities and differences among data-driven analysis methods applied to fMRI data. We also hope this research will lead to building more sophisticated data-driven analysis methods for identification of functional connectivity in fMRI.

Appendix A

Tables

This section presents the tables that contain the results of the component-wise comparison of resulting ICA and GMM components of the real fMRI study for K = 15, referred in Section 5.3.3.

		GMM														
	#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ICA	1	0.386	0.236	0.094	0.065	0.029	0.138	0.118	0.082	0.129	0.120	0.104	0.132	0.108	0.042	0.020
	2	0.186	0.181	0.133	0.108	0.136	0.093	0.108	0.129	0.122	0.082	0.080	0.112	0.070	0.073	0.068
	3	0.163	0.142	0.152	0.171	0.050	0.119	0.148	0.101	0.115	0.057	0.045	0.087	0.063	0.079	0.061
	4	0.106	0.074	0.207	0.090	0.268	0.089	0.105	0.124	0.098	0.074	0.054	0.078	0.065	0.112	0.072
	5	0.119	0.118	0.120	0.050	0.091	0.091	0.136	0.141	0.095	0.126	0.050	0.073	0.136	0.051	0.086
	6	0.106	0.078	0.057	0.131	0.037	0.119	0.063	0.057	0.123	0.175	0.135	0.063	0.072	0.118	0.058
	7	0.103	0.114	0.047	0.076	0.092	0.090	0.029	0.125	0.136	0.063	0.097	0.110	0.084	0.126	0.079
	8	0.064	0.033	0.067	0.161	0.056	0.064	0.060	0.042	0.058	0.081	0.182	0.057	0.082	0.033	0.027
	9	0.054	0.082	0.102	0.102	0.110	0.070	0.084	0.088	0.081	0.058	0.088	0.049	0.077	0.110	0.071
	10	0.089	0.078	0.066	0.065	0.075	0.086	0.057	0.060	0.043	0.093	0.093	0.079	0.093	0.052	0.065
	11	0.068	0.084	0.082	0.056	0.078	0.061	0.079	0.100	0.069	0.073	0.065	0.059	0.042	0.070	0.041
	12	0.071	0.047	0.055	0.036	0.128	0.093	0.025	0.052	0.107	0.048	0.046	0.065	0.083	0.048	0.057
	13	0.064	0.052	0.055	0.032	0.081	0.058	0.043	0.034	0.085	0.083	0.046	0.038	0.076	0.045	0.045
	14	0.026	0.045	0.031	0.045	0.050	0.034	0.106	0.042	0.055	0.057	0.044	0.052	0.038	0.035	0.041
	15	0.035	0.037	0.074	0.049	0.050	0.057	0.045	0.042	0.022	0.082	0.053	0.032	0.064	0.055	0.030

Table A.1: Average of the absolute values of correlations coefficient between resulting ICA and GMM spatial maps for real fMRI study. K = 15.

		GMM														
	#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ICA	1	0.918	0.714	0.453	0.442	0.418	0.509	0.638	0.499	0.768	0.550	0.499	0.709	0.488	0.175	0.178
	2	0.552	0.652	0.375	0.393	0.491	0.335	0.426	0.505	0.541	0.394	0.358	0.514	0.407	0.370	0.343
	3	0.597	0.482	0.430	0.543	0.366	0.539	0.645	0.391	0.537	0.256	0.347	0.460	0.248	0.348	0.310
	4	0.519	0.409	0.441	0.474	0.769	0.395	0.420	0.536	0.498	0.279	0.286	0.383	0.368	0.390	0.222
	5	0.422	0.354	0.430	0.271	0.286	0.285	0.480	0.401	0.296	0.374	0.345	0.253	0.473	0.206	0.412
	6	0.372	0.237	0.316	0.326	0.233	0.324	0.216	0.270	0.322	0.474	0.345	0.228	0.250	0.417	0.306
	7	0.437	0.417	0.230	0.243	0.283	0.411	0.252	0.394	0.503	0.312	0.276	0.510	0.202	0.465	0.339
	8	0.321	0.228	0.211	0.393	0.231	0.231	0.278	0.211	0.322	0.324	0.610	0.185	0.262	0.178	0.157
	9	0.286	0.282	0.307	0.234	0.390	0.238	0.236	0.304	0.329	0.174	0.354	0.291	0.327	0.361	0.359
	10	0.404	0.309	0.293	0.273	0.286	0.222	0.312	0.297	0.301	0.321	0.333	0.424	0.391	0.182	0.267
	11	0.391	0.338	0.276	0.235	0.352	0.264	0.354	0.391	0.366	0.349	0.325	0.406	0.161	0.264	0.173
	12	0.316	0.185	0.282	0.290	0.416	0.332	0.192	0.345	0.335	0.262	0.207	0.286	0.336	0.325	0.308
	13	0.357	0.273	0.174	0.242	0.285	0.312	0.275	0.209	0.409	0.275	0.211	0.277	0.260	0.157	0.215
	14	0.183	0.233	0.162	0.151	0.223	0.158	0.231	0.241	0.255	0.181	0.225	0.243	0.195	0.148	0.197
	15	0.215	0.167	0.252	0.098	0.146	0.153	0.200	0.206	0.211	0.227	0.201	0.187	0.210	0.149	0.158

Table A.2: Average of the absolute values of correlations coefficient between resulting ICA and GMM time courses for real fMRI study. K = 15.

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *Neural Information Processing Systems*, pages 757–763, 1996.

[3] K. Baek, B. Draper, J.R. Beveridge, and K. She. PCA vs. ICA: a comparison on the FERET data set. In *Proc. of the Fourth International Conference on Computer Vision, Pattern Recognition and Image Processing*, pages 824–827, 2002.

[4] P. Bandettini, A. Jesmanowicz, E. Wong, and J. Hyde. Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance Imaging*, 30(2):161–173, 1993.

[5] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski. Face recognition by independent component analysis. *IEEE Trans. on Neural Networks*, 13(6):1450–1464, 2002.

[6] R. Baumgartner, L. Ryner, W. Richter, R. Summers, M. Jarmasz, and R. Somorjai. Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs. principal component analysis. *Magnetic Resonance Imaging*, pages 89–94, 2000.

[7] C. Beckmann, M. DeLuca, J. Devlin, and S. Smith. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:1001–1013, 2005.

[8] A. Bell and T. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[9] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] B. Biswal, F. Z. Yetkin, V. M. Haughton, and J. S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34:537–541, 1995.

[11] V. D. Calhoun, J. Pekar, V. McGinty, T. Adali, T. Watson, and G. Pearlson. Different activation dynamics in multiple neural systems during simulated driving. *Human Brain Mapping*, 16:158–167, 2002.

[12] V.D. Calhoun and T. Adali. Unmixing fMRI with independent component analysis. *Engineering in Medicine and Biology Magazine, IEEE*, 25(2):79–90, 2006.

[13] V.D. Calhoun, T. Adali, G. Pearlson, and J. Pekar. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Human Brain Mapping*, 13:43–53, 2001.

[14] V.D. Calhoun, P. K. Maciejewski, G.D. Pearlson, and K. A. Kiehl. Temporal lobe and "default" hemodynamic brain modes discriminate between schizophrenia and bipolar disorder. *Human Brain Mapping*, Preprint, 2007.

[15] V.D. Calhoun, G. Pearlson, and T. Adali. Independent component analysis applied to fMRI data: a generative model for validating results. *The Journal of VLSI Signal Processing*, 37:281–291, 2004.

[16] J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114, 1997.

[17] D. Cordes, V. Haughton, J. Carew, K. Arfanakis, and K. Maravilla. Hierarchical clustering to measure connectivity in fMRI resting-state data. *Magnetic Resonance in Medicine*, 20(4):305–317, 2002.

[18] T. M. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2001.

[19] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[20] C. Davatzikos, K. Ruparel, Y. Fan, D.G. Shen, M. Acharyya, J.W. Loughead, R.C. Gur, and D.D. Langleben. Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *Neuroimage*, 28(3):663–668, 2005.

[21] K. Delac, M. Grgic, and S. Grgic. Independent comparative study of PCA, ICA, and LDA on the FERET data set. *International Journal of Imaging Systems and Technology*, 15(5):252–260, 2006.

[22] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[23] B. Draper, K. Baek, M.S. Bartlett, and J.R. Beveridge. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding*, 91:115–137, 2003.

[24] F. Esposito, E. Formisano, E. Seifritz, R. Goebel, R. Morrone, G. Tedeschi, and F. Di Salle. Spatial independent component analysis of functional MRI time-series: to what extent do results depend on the algorithm used? *Human Brain Mapping*, 16:146–157, 2002.

[25] O. Friman, J. Carlsson, P. Lundberg, M. Borga, and H. Knutsson. Detection of neural activity in functional MRI using canonical correlation analysis. *Magnetic Resonance Imaging*, 45(2):323–330, 2001.

[26] K. Friston. Functional and effective connectivity in neuroimaging: a systhesis. *Human Brain Mapping*, 2:56–78, 1994.

[27] K. Friston. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

[28] K. Friston, J. Ashburner, J. Poline, C. Frith, J. Heather, and R. Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 2:165–189, 1995.

[29] W. Ganong. *Review of Medical Physiology*. Prentice-Hall, 1995.

[30] P. Golland, Y. Golland, and R. Malach. Detection of spatial activation patterns as unsupervised segmentation of fMRI data. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2007.

[31] P. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. Springer, 2005.

[32] C. Goutte, P. Toft, E. Rostrup, F. Nielsen, and L. K. Hansen. On clustering fMRI time series. *NeuroImage*, 9:298–319, 1999.

[33] P. Hall. On representatives of subsets. *Journal of London Mathematical Society*, 10:26–30, 1935.

[34] J. Haxby, M. I. Gobbini, M. Furey, A. Ishai, J. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:242–2430, 2001.

[35] B. Horwitz. The elusive concept of brain connectivity. *Neuroimage*, 19:466–470, 2003.

[36] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer, 2004.

[37] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.

[38] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[39] F. Jezzard, P. Matthews, and S. Smith. *Functional MRI: An introduction to methods*. Oxford University Press, 2002.

[40] I. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[41] T. Jung, C. Humphries, M. Lee, V. Iragui, S. Makeig, and T. Sejnowski. Removing electroencephalographic artifacts: Comparison between ICA and PCA. In *IEEE International Workshop on Neural Networks for Signal Processing*, pages 63–72, 1998.

[42] V. Kiviniemi, J. Kantola, J. Jauhiainen, A. Hyvarinen, and C. Tervonena. Independent component analysis of nondeterministic fMRI signal sources. *Neuroimage*, 19:253–260, 2003.

[43] T. Kohonen. *Self-organizing maps*. Springer Verlag, 1995.

[44] L. Lee, L. Harrison, and A. Mechelli. The functional brain connectivity workshop: report and commentary. *Neuroimage*, 2003.

[45] C. Liu and H. Wechsler. Comparative assessment of independent component analysis for face recognition. In *Second International Conference on Audio- and Video- based Biometric Person Authentication*, 1999.

[46] M. J. Lowe, B. Mock, and J. A. Sorenson. Functional connectivity in single and multislice echoplanar imaging using resting state fluctuations. *Neuroimage*, 7:119–132, 1998.

[47] L. Ma, B. Wang, X. Chen, and J. Xiong. Detecting functional connectivity in the resting brain: a comparison between ICA and CCA. *Magnetic Resonance Imaging*, 25(1):47–56, 2007.

[48] M. J. McKeown, L. K. Hansen, and T. J. Sejnowski. Independent component analysis of functional MRI: what is signal and what is noise? *Current Opinion in Neurobiology*, 13(5):620–629, 2003.

[49] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, A. A. Kindermann, A. Bell, and T. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6:160–188, 1998.

[50] A. Meyer-Baese, A. Wismueller, and O. Lange. Comparison of two exploratory data analysis methods for fMRI: unsupervised clustering versus independent component analysis. *IEEE Transactions on Information Technology in Medicine*, 8(3):387–398, 2004.

[51] S. Ogawa, T. Lee, A. Kay, and D. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87:9868–9872, 1990.

[52] T. Parrish, D. Gitelman, K. LaBar, and M.-M. Mesulam. Impact of signal-to-noise on functional MRI. *Magnetic Resonance Imaging*, 44:925–932, 2000.

[53] S. J. Peltier, T. A. Polk, and D. C. Noll. Detecting low-frequency functional connectivity in fMRI using a self-organizing map (SOM) algorithm. *Human Brain Mapping*, 20:220–226, 2003.

[54] W. Penny, K. Stephan, A. Mechelli, and K. Friston. Modelling functional integration: a comparison of structural equation and dynamic models. *NeuroImage*, 23:264–274, 2004.

[55] S. Polyn, V. Natu, J. Cohen, and K. Norman. Category-specific cortical activity precedes retrieval during memory search. *Science*, 310:1963–1966, 2005.

[56] N. Ramnani, L. Lee, A. Mechelli, C. Phillips, A. Roebroeck, and E. Formisano. Exploring brain connectivity: a new frontier in systems neuroscience. *Trends in Neurosciences*, 25(10):496–497, 2002.

[57] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[58] A. Smolders, F. De Martino, N. Staeren, P. Scheunders, J. Sijbers, R. Goebel, and E. Formisano. Dissecting cognitive stages with time-resolved fMRI data: a comparison of fuzzy clustering and independent component analysis. *Magnetic Resonance Imaging*, 25(6):860–868, 2007.

[59] J. Stone. *Independent component analysis: a tutorial introduction*. MIT Press, 2004.

[60] J. Talairach and P. Tournoux. *Referentially Oriented Cerebral MRI Anatomy*. Thieme, 1993.

[61] B. Thirion, S. Dodel, and J. B. Poline. Detection of signal synchronizations in resting-state fMRI datasets. *Neuroimage*, 1:321–327, 2006.

[62] M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of Royal Statistical Society, Series B*, 61:611–622, 1999.

[63] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.

[64] R. Turner, D. L. Bihan, C. Moonen, D. Despres, and J. Frank. Echo-planar time course MRI of cat brain deoxygenation changes. *Magnetic Resonance Imaging*, 22:159–166, 1991.

[65] V. van de Ven, E. Formisano, D. Prvulovic, C. Roeder, and D. Linden. Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. *Human Brain Mapping*, 22:165–178, 2004.

[66] K. Worsley. An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, 5:254–258, 1997.

[67] K. Yang and J. C. Rajapakse. ICA gives higher-order functional connectivity of brain. *Neural Information Processing - Letters and Review*, 2:27–32, 2004.