Spatio-Temporal *f*MRI Signal Analysis Using Information Theory

by

Junmo Kim

B.S., Seoul National University (1998)

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

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2000
[September 2000]

Author .....
Department of Electrical Engineering and Computer Science
August 4, 2000

Certified by .....
John W. Fisher
Research Scientist, Laboratory for Information and Decision Systems
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Spatio-Temporal *f*MRI Signal Analysis Using Information Theory

by

Junmo Kim

Submitted to the Department of Electrical Engineering and Computer Science
on August 4, 2000, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering

Abstract

Functional MRI is a fast brain imaging technique which measures the spatio-temporal neuronal activity. The development of automatic statistical analysis techniques which calculate brain activation maps from *f*MRI data has been a challenging problem due to the limitation of current understanding of human brain physiology. In previous work a novel information-theoretic approach was introduced for calculating the activation map for *f*MRI analysis [Tsai *et al* , 1999]. In that work the use of mutual information as a measure of activation resulted in a nonparametric calculation of the activation map. Nonparametric approaches are attractive as the implicit assumptions are milder than the strong assumptions of popular approaches based on the general linear model popularized by Friston *et al* [1994]. Here we show that, in addition to the intuitive information-theoretic appeal, such an application of mutual information is equivalent to a hypothesis test when the underlying densities are unknown. Furthermore we incorporate local spatial priors using the well-known Ising model thereby dropping the implicit assumption that neighboring voxel time-series are independent. As a consequence of the hypothesis testing equivalence, calculation of the activation map with local spatial priors can be formulated as mincut/maxflow graph-cutting problem. Such problems can be solved in polynomial time by the Ford and Fulkerson method. Empirical results are presented on three *f*MRI datasets measuring motor, auditory, and visual cortex activation. Comparisons are made illustrating the differences between the proposed technique and one based on the general linear model.

Thesis Supervisor: John W. Fisher
Title: Research Scientist, Laboratory for Information and Decision Systems

Acknowledgments

At this point of finishing my Master's thesis and making a new start as a more mature student and researcher, there are many people to whom I would like to express my sincere gratitude. First and foremost, I would like to thank my research supervisor, Dr. John Fisher, for his advice, encouragement and sense of humor. He introduced me to the field of information theoretic signal processing and has been an endless source of interesting idea and insight since I took my first step in research as a 6.961 student until I finished this thesis. He also gave me the opportunity to meet with Prof. Ramin Zabih, which motivated me to search after graph theoretic approach for the MAP estimation problem I formulated. I would also like to thank Prof. Alan Willsky for his enthusiastic teaching of 6.432 which was filled with both intuition and mathematical rigor and his thought-provoking advice in the grouplet which I have attended since the fall term, 1999. In addition, I am grateful to Andrew Kim for his willingness to review the draft of this thesis and to Sandy Wells for his valuable discussion. I would like to acknowledge Andy Tsai and Cindy Wible for providing me with the *f*MRI data.

My graduate life at MIT has been enjoyable and valuable thanks to the opportunity of various kinds of meetings and interaction I had with members of the SSG. I especially thank: John Richard for his encouragement and help from the time I was new at MIT; Andy Tsai for his introduction to the *f*MRI analysis and advice on the graduate life; Andrew Kim for his help on every questions on SSG network and sparing his time for discussion of my research; Ron Dror for his informing me an opportunity of work at MERL; Martin Wainwright who gave me a nick name "apprentice"; Erik Sudderth for his excellence as a new SSG member; Mike Schneider for his smile and humor; Dewy Tucker for his kindness in making me feel at home at

MIT; Alex Ihler for his cheerfulness and the time we studied and discussed together. Especially, the discussion with him was helpful to clarifying the idea of Chapter 3.

I acknowledge the Korea Foundation for Advanced Studies(KFAS) for providing me with the full financial support for the first year of my graduate study. I wish to thank those who brought me up in academia. Especially, my interest in the theory of probability is attributed to Prof. Taejeong Kim. I also would like to thank Prof. Amos Lapidoth, who showed me the beauty of information theory.

Finally, I would like to express my appreciation to those who have made me who I am, my father and mother. They have been my great teachers both intellectually and spiritually. I would like to give my thanks to them for their unconditional love in this small thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Functional magnetic resonance imaging (*f*MRI) is a recently developed fast brain imaging technique which takes a series of 3D MR images of the brain in real time. Since an *f*MRI signal represents the change of the blood oxygenation level induced by neuronal activity, it is used to detect the regions of the brain activated by a specific cognitive function. It is a very promising functional imaging technique. Not only can it measure dynamic neuronal activity with a spatial resolution equivalent to that of positron emission tomography (PET), the current standard for functional analysis, but it also has several advantages over PET such as a higher temporal resolution, an absence of radioactive compounds, and a relatively low cost. It is clinically used to detect the causes of behavioral malfunctions or the effects of tumors in certain locations of the brain. It is also of interest to cognitive scientists because of its possibility to reveal secrets of human brain.

Functional MRI measures neuronal activity indirectly through the blood oxygenation level dependent (BOLD) response. Since the underlying physiology is not

thoroughly understood, automated statistical analysis of the *f*MRI signal is a challenging problem. Conventional methods of detecting activated regions include direct subtraction, correlation coefficient, and the general linear model [1]. Direct subtraction and correlation coefficient assume a linear relationship between the protocol and *f*MRI temporal response, while the general linear model assumes that the *f*MRI temporal response is a linear combination of basis signals such as the protocol, cardiovascular response, and others.

In contrast to the above linear techniques, mutual information measures nonlinear relationships beyond second order statistics [2]. With this as a motivation, Tsai *et al* [2] propose a novel information theoretic approach for calculating *f*MRI activation maps, where activation is as quantified by the mutual information between the protocol signal and the *f*MRI time-series at a given voxel. In that work, it is empirically shown that the information-theoretic approach can be as effective as other conventional methods of calculating brain activation maps.

In this thesis, we extend the ideas first proposed in Tsai *et al* . Specifically, we incorporate a spatial prior using Markov Random Fields (MRF). Additionally, in support of the MRF approach, we reinterpret mutual information (MI) in the context of hypothesis testing. This allows for a fast Maximum a Posteriori (MAP) solution of the *f*MRI activation map.

Figure 1.1 gives an overview of this thesis where our contributions are illustrated by the branches in the diagram. Starting from our previous work [2] on the information theoretic approach, Arrow (1) indicates the mathematical interpretation of the previous work to enable the extension to the Bayesian framework. Motivated by the work of Descombes [3], we apply an MRF prior to the *f*MRI analysis in the Bayesian framework as illustrated by Arrow (2). The method developed by Greig

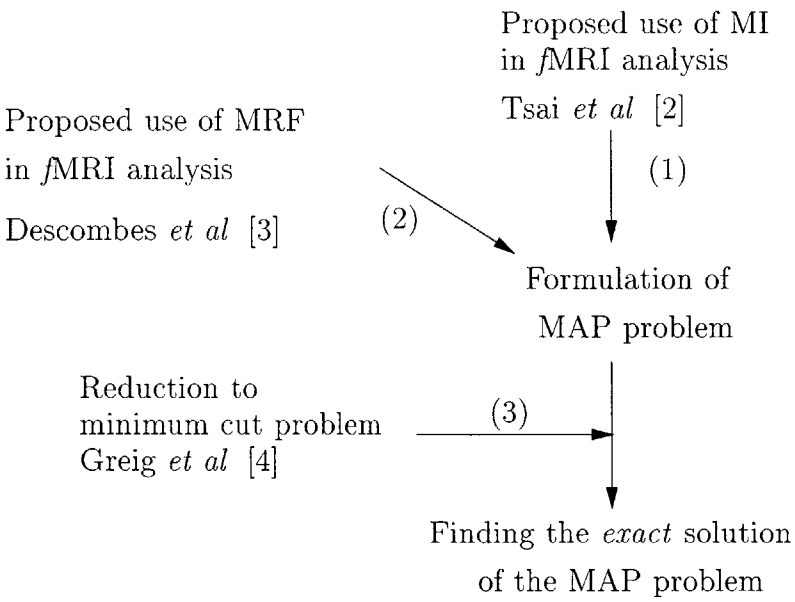


Figure 1.1: Overview of this thesis.

et al [4] is then applicable to the MAP problem we have formulated. This link is denoted by Arrow (3). This method solves the MAP problem *exactly* in polynomial time by reducing the MAP problem to minimum cut problem in flow network.

1.1 Contributions

As stated, there are two contributions of this thesis. The first is the interpretation of MI in the context of hypothesis testing. This enables the second and primary contribution: formulation of *f*MRI analysis as a MAP estimation problem which subsequently leads to an *exact* solution of the MAP problem in *f*MRI analysis by solving a minimum cut problem.

Information Theoretic Approach: Nonparametric Detection

The use of MI as a statistic can be naturally interpreted in the hypothesis testing context, which will be demonstrated in Chapter 3. The natural hypothesis testing problem is to test whether the *f*MRI signal is independent of the protocol signal. We will show that the likelihood ratio can be approximated as an exponential of the mutual information estimate. This reveals that our information theoretic approach is asymptotically equivalent to a likelihood ratio test for the nonparametric hypothesis testing problem. Therefore, it suggests that the information theoretic approach has high detection power considering Neyman-Pearson lemma.

Bayesian Framework with a Spatial Prior and Its Exact Solution

It is well accepted that there are spatial dependencies in *f*MRI data. This spatial information can be exploited by modeling voxel-dependency with an Ising prior which is a binary Markov Random Field with the nearest neighborhood system. Since the likelihood ratio can be approximated in terms of the estimated mutual information, the Ising prior is easily combined with the information theoretic approach. Interestingly, the resulting MAP problem derived from the Ising prior and approximated likelihood ratio function was found to be solvable by Greig's method [4] which reduces the MAP problem to minimum cut problem in network flow graph. The significance of this reduction is that it gives an *exact* solution to the MAP problem with an Ising prior in polynomial time.

Analysis of Kernel Size with Regard to *f*MRI Analysis

The estimation of MI involves the use of a Parzen density estimator. A fundamental issue in the use of the Parzen estimator is the choice of a regularizing kernel size parameter. We propose and evaluate an automated way of choosing this kernel size parameter for estimating MI in *f*MRI analysis.

1.2 Organization

The remainder of this thesis is organized as follows. Chapter 2 provides background on the physiology of *f*MRI, various *f*MRI signal analysis techniques including the general linear model (GLM), and the information theoretic approach proposed by Tsai *et al* [2]. In Chapter 3, we present a mathematical interpretation of the information theoretic approach in the hypothesis testing context. This will then be combined with Markov random field prior casting the problem in a Bayesian framework in Chapter 4. In Chapter 5, we present experimental results of the method developed in this thesis, discuss its significance, and compare it with conventional methods such as the general linear model. We conclude with brief summary of the work and directions for future research in Chapter 6.

Chapter 2

Background

This chapter introduces preliminary knowledge on *f*MRI, our previous information theoretic approach, and statistical concepts such as detection and Markov random fields (MRF) which are used in Chapter 3 and Chapter 4. In Section 2.1, we describe the current understanding of *f*MRI and the challenges in *f*MRI signal analysis. In Section 2.2, we discuss several conventional methods with an emphasis on the general linear model which is the current standard. The information theoretic approach of Tsai *et al* is presented in Section 2.3. We conclude with a brief introduction to binary hypothesis testing theory in Section 2.4 and Markov Random Field theory in Section 2.5.

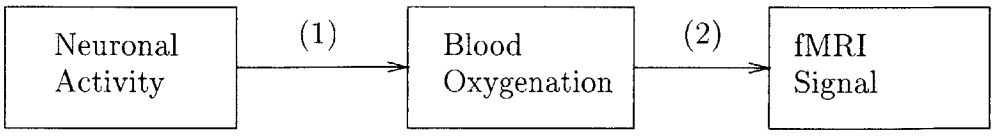


Figure 2.1: Route from neuronal activity to *f*MRI signal.

2.1 A Brief Discussion of *f*MRI

In this section, we provide background on *f*MRI such as the physiology of human brain relevant to *f*MRI, the physical meaning of *f*MRI signal, and common experiment design. In addition, the limitation and challenges of *f*MRI signal analysis are presented.

2.1.1 Characteristics of *f*MRI Signals

An *f*MRI signal is a collection of 3D brain images taken periodically (for example, every 3 seconds) while a subject is performing an experimental task in the imaging device. The obtained *f*MRI signal is thus a 4 dimensional discrete-space-time signal, which can be denoted by $X(i, j, k, t)$. The significance of the *f*MRI signal can be explained as follows and is depicted in Figure 2.1. The item of interest is the neuronal activity in the brain associated with a particular task. This activity is not directly measurable but is related to the blood oxygenation level which can be measured with *f*MRI.

Neuronal activity results in oxygen consumption to which the body reacts with a highly localized oxygen delivery and overcompensates for the oxygen consumption. As a result, a substantial rise in oxyhemoglobin is seen, where the rise in relative oxyhemoglobin is maximal after 4–10 seconds [5]. This phenomena is called hemo-

dynamic response. The hemodynamic response is a delayed and dispersed version of neuronal activity limiting both the spatial and temporal resolution of the imaging.

*f*MRI measures the change of blood oxygenation level. Specifically the difference in T_2 (transverse or spin-spin decay time constant) between oxygenated(red) blood and deoxygenated(blue) blood. The signal from a long T_2 substance (such as red blood) is stronger than that from a short T_2 substance (such as blue blood) and so a locality with red blood appears brighter than a locality with blue blood [6]. The observed *f*MRI signal is thus called a *blood oxygenation level dependent (BOLD) signal.*

There is uncertainty in both the hemodynamic response and the imaging process. Not only may different parts of brain exhibit different hemodynamic behavior, but noise is also introduced during imaging process. Therefore, the *f*MRI signal is reasonably modeled as a stochastic process.

A fundamental objective is to design an experiment that allows us to detect which regions of the brain are functionally related to a given stimulus. The typical approach is the so called block experimental design. In this method, a subject is asked to perform a task for a specific time (for example, 30 seconds) and then to rest for another period of time. This procedure is then repeated several times. As a result of the block experimental design, each voxel of the brain has its own *f*MRI signal called the *f*MRI temporal response. The idea is to compare the observed temporal response of that voxel during the task state and the temporal response during the rest state. If there is significant difference between those two, the voxel is considered to be activated during the experiment.

All data used in this thesis has following formats:

- The image is taken every 3 seconds for 180 seconds resulting in 60 images.
- Each image has 64 by 64 by 21 voxels.
- The lengths of the task and rest states are both 30 seconds.

2.2 Conventional Statistical Techniques of *f*MRI Signal Analysis

In this section, we describe several popular techniques for *f*MRI analysis. In all the methods, the decision of the activation state of voxel (i, j, k) is made based on $X(i, j, k, \cdot)$, the discrete-time temporal response of that voxel.

Direct Subtraction

The direct subtraction method tests whether the mean intensity of the temporal response during the task state is different from the mean intensity of the temporal response during the rest state. Student's t-test is typically used to test this with the following statistic:

$$t = \frac{\bar{x}_{on} - \bar{x}_{off}}{\sqrt{\frac{\sigma^2_{x_{on}}}{N_{on}-1} + \frac{\sigma^2_{x_{off}}}{N_{off}-1}}} \tag{2.1}$$

where x_{on} and x_{off} are the sets of *f*MRI temporal responses corresponding to the task state and the rest state respectively, $\bar{x}_{on}$ and $\bar{x}_{off}$ are the averages of the set x_{on} and x_{off}, N_{on} and N_{off} are the cardinalities of the sets, and $\sigma^2_{x_{on}}$ and $\sigma^2_{x_{off}}$ are the

variances of the sets respectively.

This test is widely used because it is simple to understand and implement. However, this only tests whether or not the 2 source distributions have the same mean and furthermore optimality of the test only holds for Gaussian distributions.

Cross correlation

This method calculates the cross correlation between the *f*MRI temporal response $X(i,j,k,\cdot)$ and a reference signal designed to reflect the change of task and rest states. The cross correlation between two signals (x_i) and (y_i) are given by

$$\rho_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{2.2}$$

By Schwartz inequality, $-1 \leq \rho_{xy} \leq 1$ and a high $|\rho_{xy}|$ means a tight linear relationship exists between the *f*MRI temporal response and the reference signal thereby suggesting activation of the voxel considered. The weak point of this method is that the performance depends on an accurate reference signal which is difficult to model due to the lack of knowledge on the characteristics of the *f*MRI temporal response.

Kolmogorov-Smirnov

The idea in the Kolmogorov-Smirnov test is similar to that of the direct subtraction method. The major distinction of this method is that it is nonparametric so it does not make an assumption that *f*MRI signal is Gaussian. Specifically, this test

decides whether the two sets of data corresponding to the task state and the rest state respectively, were drawn from same distribution or two different distributions. This test calculates a kind of distance between two empirical distributions[1] obtained from two data sets as follows:

$$D_{n,n} = \sup_x |F_{n,on}(x) - F_{n,off}(x)| \tag{2.3}$$

where $F_{n,on}(x)$ and $F_{n,off}(x)$ are the empirical distributions obtained from the two sets x_{on} and x_{off}.

Under the condition that $x_{on} \cup x_{off}$ are independent identically distributed (i.i.d.), this statistic has the property that "$Pr\{D_{n,n} < \frac{r}{n}\}$ equals the probability in a symmetric random walk that a path of length $2n$ starting and terminating at the origin does not reach the points $\pm r$. [7, pages 36–39]"

General Linear Model

In the general linear model, a subspace representing an active response and a subspace representing a nuisance signal are designed, then the voxel is declared to be active if significant power of the observed *f*MRI temporal response is in the subspace of the active response. In this section, this approach is presented mathematically.

Let $Y_1, \ldots, Y_n$ be observed *f*MRI temporal response of a certain voxel. This

[1]Empirical distribution is defined as follows:
$F_n(x) = \frac{1}{n}[$ number of observations $\leq x$ among $X_1, \ldots, X_n]$

method assumes the linear model

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_t \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1i} & \cdots & X_{1L} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{t1} & \cdots & X_{ti} & \cdots & X_{tL} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{ni} & \cdots & X_{nL} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_L \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_t \\ \vdots \\ e_n \end{pmatrix},$$

or more compactly,

$$\begin{aligned} Y &= X\beta + e && (2.4) \\ &= [X_1 \vdots X_2] \begin{bmatrix} \beta_1 \\ \cdots \\ \beta_2 \end{bmatrix} + e && (2.5) \\ &= X_1\beta_1 + X_2\beta_2 + e && (2.6) \end{aligned}$$

where $e_1, \ldots, e_n$ are i.i.d. according to $N(0, \sigma^2)$ and σ is unknown, X_1 is a matrix whose range is a subspace of interest, X_2 is a matrix whose range is a nuisance subspace, and $X = [X_1 \vdots X_2]$ is a design matrix whose columns are the explanatory time series.

Before proceeding, let us define some notation.

- $p = \text{rank}(X)$
- $p_2 = \text{rank}(X_2)$

- $P_X = X(X^TX)^{-1}X^T$: a projection matrix of X
- P_{X_2}: a projection matrix of X_2
- $S(\beta) = \|Y - P_X Y\|^2$
- $S(\beta_2) = \|Y - P_{X_2} Y\|^2$

Since $(I - P_X)Y = (I - P_X)e$ and $(I - P_{X_2})Y = (I - P_{X_2})(X_1\beta_1 + e)$,

$$\begin{aligned} S(\beta) &= \|(I - P_X)Y\|^2 \\ &= \|(I - P_X)e\| \\ &= e^T(I - P_X)^2 e \\ &= e^T(I - P_X)e \qquad (2.7) \\ S(\beta_2) - S(\beta) &= Y^T(I - P_{X_2})Y - Y^T(I - P_X)Y \\ &= Y^T(P_X - P_{X_2})Y \\ &= (X_1\beta_1 + e)^T(P_X - P_{X_2})(X_1\beta_1 + e) \qquad (2.8) \end{aligned}$$

where the properties of projection matrix, $P_X^2 = P_X$ and $P_X^T = P_X$, were used.

In the general linear model, the hypothesis testing problem is as follows:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0.$$

The Statistic is

$$F = \frac{\frac{S(\beta_2)-S(\beta)}{p-p_2}}{\frac{S(\beta)}{n-p}}. \tag{2.9}$$

Property 2.2.1. *F is distributed according to an* F*-distribution with degree of freedom $p - p_2$ and $n - p$ and noncentrality parameter $d^2 = \|X_1\beta_1/\sigma\|^2$, i.e. $F \sim \mathrm{F}_{p-p_2,n-p}(\cdot; d^2)$*

Proof. It is sufficient to show that $S(\beta)/\sigma^2$ is a central χ^2 random variable with degree of freedom $n - p$ and $(S(\beta_2) - S(\beta))/\sigma^2$ is a noncentral χ^2 random variable with degree of freedom $p - p_2$ and noncentrality parameter $d^2 = \|X_1\beta_1/\sigma\|^2$.[2] This can be seen as follows:

Let us make an orthogonal matrix Q such that the first p_2 columns are an orthonormal basis for Range(X_2) and the first p columns are an orthonormal basis for Range(X). Then the projection of Y on Range(X) is found by considering $Y = Qz = \sum_{i=1}^{n} q_i z_i$. Then

$$P_X Y = \sum_{i=1}^{p} q_i z_i = Q \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} z = Q \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} Q^T Y.$$

Thus

$$P_X = Q \begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix} Q^T.$$

Similarly,

$$P_{X_2} = Q \begin{bmatrix} I_{p_2} & 0 \\ 0 & 0 \end{bmatrix} Q^T.$$

[2]See Appendix A for definitions of the F and χ^2 distributions.

Thus

$$S(\beta)/\sigma^2 = e^T(I - P_X)e/\sigma^2 = (Q^T e/\sigma)^T \begin{bmatrix} 0 & 0 \\ 0 & I_{n-p} \end{bmatrix} (Q^T e/\sigma)$$

is the sum of $n - p$ independent random variables drawn from $N(0, 1)$. Therefore, $S(\beta)/\sigma^2$ is a central χ^2 random variable with degree of freedom $n - p$. Similarly

$$(S(\beta_2) - S(\beta))/\sigma^2 = (Q^T(X_1\beta_1 + e)/\sigma)^T \begin{bmatrix} 0 & 0 & 0 \\ 0 & I_{p-p_2} & 0 \\ 0 & 0 & 0 \end{bmatrix} (Q^T(X_1\beta_1 + e)/\sigma)$$

is a sum of $p - p_2$ independent random variables drawn from $N(\mu_i, I)$, where $\mu = [\mu_1 \ldots \mu_n]^T = Q^T X_1\beta_1/\sigma$. Therefore, $(S(\beta_2) - S(\beta))/\sigma^2$ is a noncentral χ^2 random variable with degree of freedom $p - p_2$ and noncentrality parameter $d^2 = \|X_1\beta_1/\sigma\|^2$.

□

2.3 Information Theoretic Approach

This section presents the information theoretic approach to *f*MRI analysis developed by Tsai *et al* [2]. In this method, mutual information is used to quantify the degree of dependency between an *f*MRI temporal response and a protocol signal. This is attractive in that MI can capture *nonlinear* dependencies which cannot be detected by the traditional methods that assume a linear Gaussian structure for the *f*MRI signal or measure linear dependency. Furthermore, this nonparametric approach makes few assumptions on the structure of the *f*MRI temporal response. Instead of making strong functional assumptions, it treats the signal as stochastic entity and learns its underlying distribution from the observed data.

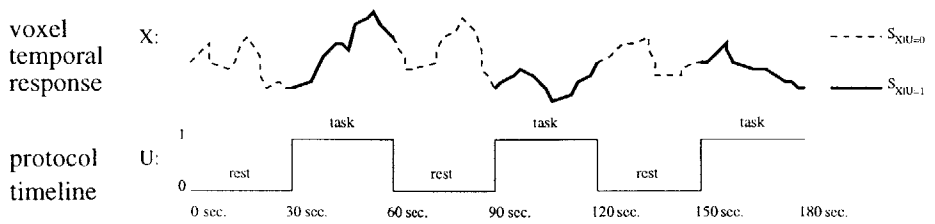


Figure 2.2: Illustration of the protocol time-line, $S_{X|U=0}$, and $S_{X|U=1}$.

This method is *voxelwise* in that it decides whether a voxel is activated based solely on the temporal response of that voxel without considering the temporal response of neighboring voxels. Specifically, the probability density function (pdf) of the *f*MRI signal for each voxel, and hence its entropy, are all estimated independently. The MI between each voxel and the protocol signal is then estimated and used as a statistic for the decision on activation. Estimating the pdf for each voxel is necessary considering that different parts of brain have different behaviors when they are activated. However, this does not take the spatial dependency into account.

2.3.1 Calculation of Brain Activation Map by MI

In order to calculate the MI between a protocol signal and voxel signal, we let $X(\cdot,\cdot,\cdot,\cdot) = \{X(i,j,k,t)|1 \leq t \leq n\}$ denote the observed *f*MRI signal, where i,j,k are spatial coordinates and t is a time coordinate. Each voxel (i,j,k) has an associated discrete-time temporal response, $X_1,\ldots,X_n$, where X_t is defined as $X(i,j,k,t)$ for notational convenience.

Figure 2.2 illustrates the protocol time-line and an associated temporal response. $S_{X|U=0}$ denotes the set of X_i's where the protocol is 0 while $S_{X|U=1}$ denotes the set of X_i's where the protocol is 1. *It is implicitly assumed in this approach that* $S_{X|U=[0,1]}$ *are i.i.d. according to* $p_{X|U=[0,1]}(x)$. We treat the protocol U as a dis-

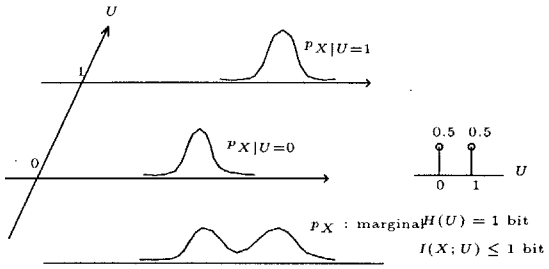


Figure 2.3: Illustration of $p_{X|U=0}, p_{X|U=1}, p_X$, and p_U

crete random variable taking 0 and 1 with equal probability. Figure 2.3 shows this situation. In this case, the MI between X and U is as follows:

$$\begin{aligned} I(X;U) &= H(U) - H(U|X) && (2.10)\\ &= h(X) - h(X|U) && (2.11)\\ &= h(X) - \frac{1}{2}h(X|U=0) - \frac{1}{2}h(X|U=1) && (2.12)\\ &= h(p_X(\cdot)) - \frac{1}{2}h(p_{X|U=0}(\cdot)) - \frac{1}{2}h(p_{X|U=1}(\cdot)) && (2.13) \end{aligned}$$

where $H(U)$ is the entropy of the discrete random variable U and $h(X)$ is the differential entropy of the continuous random variable X. It is an elementary information theory fact that $H(U) \leq 1$ bit and that $0 \leq H(U|X) \leq H(U)$ consequently $0 \leq I(X;U) \leq 1$. Thus, MI is a normalized measure of dependency between X and U with a high mutual information near 1 bit indicating that the voxel is activated.

2.3.2 Estimation of Differential Entropy

The differential entropy of random variable X is estimated as follows:

$$\begin{aligned} h(X) &= -\int_S p_X(x)\log p_X(x)dx && (2.14)\\ &= -E[\log p_X(X)] && (2.15)\\ &\approx -\frac{1}{n}\sum_{i=1}^{n}\log p_X(X_i) \text{ (Weak law of large numbers)} && (2.16)\\ &\approx -\frac{1}{n}\sum_{i=1}^{n}\log \hat{p}_X(X_i) \text{ (Parzen density estimator)} && (2.17)\\ &= -\frac{1}{n}\sum_{i=1}^{n}\log \frac{1}{n}\sum_{j=1}^{n} k(X_i - X_j, \sigma) && (2.18) \end{aligned}$$

where $\hat{p}_X(X_i)$ is the Parzen density estimator [8], defined as

$$\begin{aligned} \hat{p}_X(x) &= \frac{1}{n}\sum_{j} k(x - X_j, \sigma) && (2.19)\\ &= \frac{1}{n}\sum_{j}\frac{1}{\sigma}k\left(\frac{x - X_j}{\sigma}\right). && (2.20) \end{aligned}$$

The kernel $k(x)$ must be a valid pdf (a double exponential kernel[3] in our case). We will use $k(x,\sigma)$ and $\frac{1}{\sigma}k(\frac{x}{\sigma})$ interchangeably for notational convenience.

Eqs. (2.18) – (2.20) show that the estimate of entropy is affected by the choice of σ, which is commonly called the kernel size, bandwidth, or smoothing parameter. A larger kernel size necessarily produces a "flatter" more "spread out" pdf estimate $\hat{p}_X(x)$, as is directly evident from (2.20). This "flatter" pdf can be shown to directly lead to larger entropy estimates. Therefore, too large kernel size leads to overestima-

[3] $k(x,\sigma) = \frac{1}{2\sigma}e^{-\frac{|x|}{\sigma}}$

tion of entropy going to ∞ in the extreme case of $\sigma \rightarrow \infty$. On the other hand, the estimate of the pdf approaches a set of impulses located at each data point X_j, as the kernel size goes to zero. In this case, the estimate of entropy goes to $-\infty$. Therefore, the estimate of entropy can take value from $-\infty$ to ∞ as the kernel size changes.

Since we estimate entropy from at most 60 data points in our application, the choice of the kernel size becomes more important. Intuitively, the kernel size should be large if the underlying pdf p_X has high variance or the number of data points is small. This suggests that the kernel size should be calculated from the actual data since the data has information on how sparse p_X is.

A principled way to do this is to use kernel size that maximizes the likelihood of the data in terms of the estimate of the pdf as follows:

$$\hat{\sigma}_{ML} = \arg\max_{\sigma} \log \prod_i \hat{p}_X(X_i) \tag{2.21}$$

$$= \arg\max_{\sigma} \sum_i \log \frac{1}{n} \sum_j k(X_i - X_j, \sigma). \tag{2.22}$$

However, in this case, the maximum likelihood (ML) kernel size is trivially zero, which makes the log likelihood infinite. This motivates using the Parzen window density estimator with leave-one-out (2.23) as $\hat{p}_X(x)$. The ML kernel size in terms of the Parzen density estimate with leave-one-out is defined as follows:

$$\hat{p}_X(X_i) = \frac{1}{n-1}\sum_{j\neq i} k(X_i - X_j, \sigma) \tag{2.23}$$

$$\hat{\sigma}_{ML} = \arg\max_{\sigma} \log \prod_i \hat{p}_X(X_i) \tag{2.24}$$

$$= \arg\max_{\sigma} \sum_i \log \frac{1}{n-1}\sum_{j\neq i} k(X_i - X_j, \sigma). \tag{2.25}$$

Thus, our entropy estimator is

$$\hat{h}(X) = \min_{\sigma} -\frac{1}{n}\sum_i \log \frac{1}{n-1}\sum_{j\neq i} k(X_i - X_j, \sigma). \tag{2.26}$$

This entropy estimator with the leave-one-out method was discussed by Hall *et al* [9], where a motivation for using the kernel size that minimizes the entropy estimate is given. Further discussion on this entropy estimator is given in Section 3.3.

Figure 2.4 shows an example of the Parzen window density estimates of an activated voxel and a non-activated voxel. Figure 2.4(a) shows the Parzen estimate of two conditional pdf's $\hat{p}_{X|U=0}$ and $\hat{p}_{X|U=1}$ of a voxel declared to be activated. Visually it is clear that these two pdf's are very distinct and thus the *f*MRI signal is highly affected by the state of protocol signal. In contrast, the two conditional pdf's of a non-activated voxel shown in Figure 2.4(b) are very similar implying that the voxel response and stimulus are practically independent.

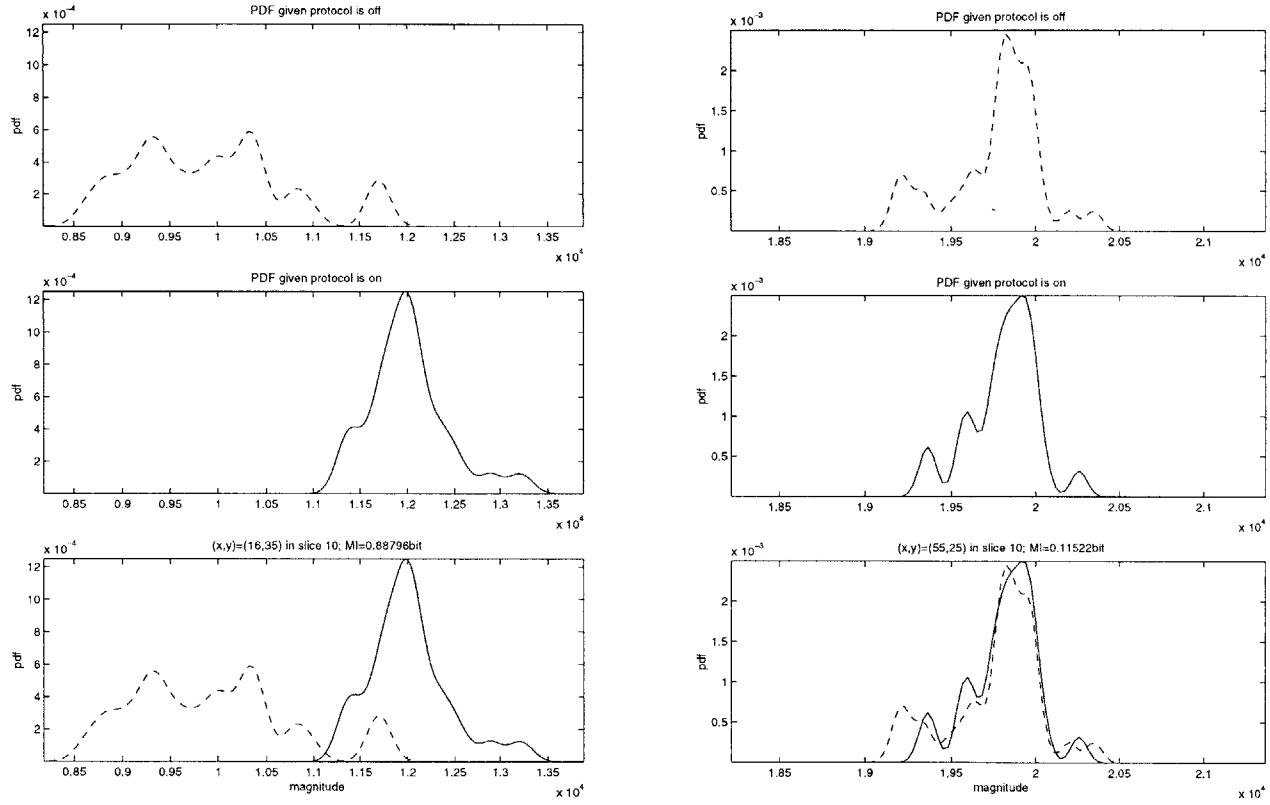
Figure 2.5 is generated from actual *f*MRI data to show how the choice of the kernel size affects the estimates of the entropy and the mutual information. Figure 2.5(a) and Figure 2.5(c) are zoomed-up versions of Figure 2.5(b) and Figure 2.5(d) respectively. The solid curves represent the entropy $h(X)$ and the dashed curves rep-

resent the conditional entropy $h(X|U)$ in Figure 2.5(a) and Figure 2.5(b). The dashed vertical line denotes the ML kernel size defined by (2.25). The solid vertical lines denote the domain inspected for the minimum of $\hat{h}(X)$.[4] Figure 2.5(c) and Figure 2.5(d) gives important information on the stability of the estimate of MI. We know that true MI, $I(X;U)$ is between 0 and 1. You can see that there is a region where the estimate of MI is far below 0 or far above 1. In this example, the ML kernel size is in the region where the estimate of the mutual information is fairly stable. In contrast, small kernel sizes, as evidenced in the figure, give a highly variable estimate of the mutual information. This example also illustrates how choosing a kernel size which maximizes estimated MI could lead to an excessively small kernel size and hence a high variance estimate.

2.3.3 Preliminary Results

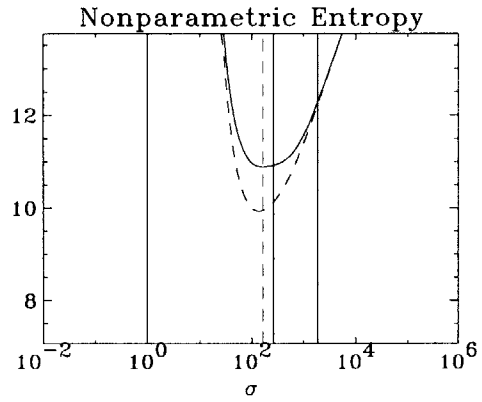
Figure 2.6 shows an *f*MRI image of the 10th coronal slice obtained from a motor cortex experiment. In this block protocol experiment, the subjects move their right hands during the task block (a period of 30 seconds) and rests during the rest block (also 30 seconds). White spots are voxels declared active by each analysis method. For the conventional methods (DS, CC, GLM), thresholds were chosen based on the prior expectation that activation will be primarily in the motor cortex. In the MI approach, 0.7 bit was used as a threshold of MI in Figure 2.6(d). At this point, it is hard to tell that which method is best since the ground truth is unknown. However, this can serve as preliminary evidence that MI is comparable to the other methods.

[4]One can use several techniques to search the minimum. Appendix B describes our method.

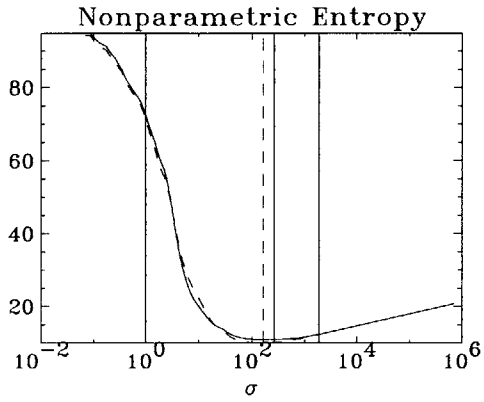


(a) Parzen pdf estimate of an activated voxel (b) Parzen pdf estimate of an non-activated voxel

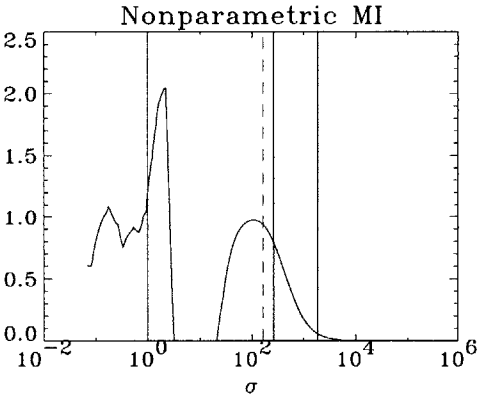
Figure 2.4: Estimates of pdf's



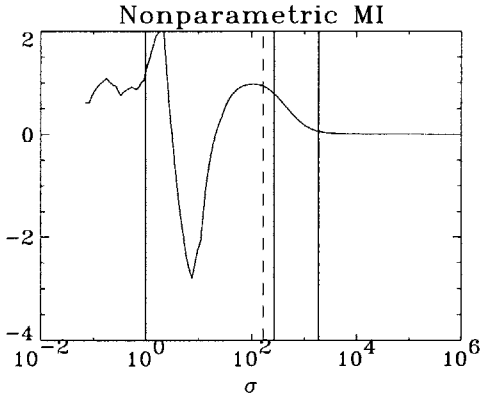
(a) Entropy vs kernel size (small scale)



(b) large scale

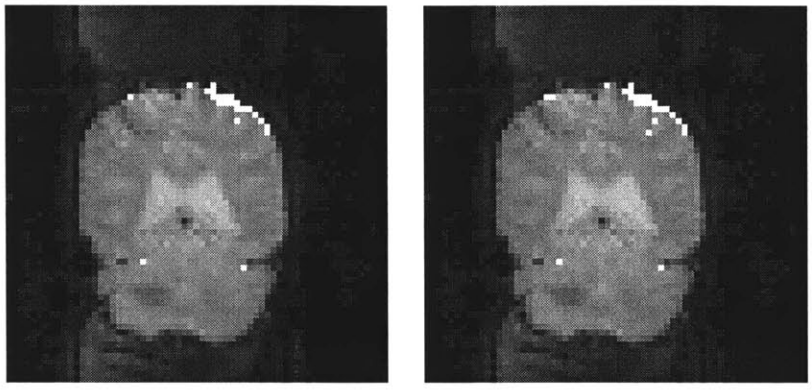


(c) Mutual information vs kernel size



(d) large scale

Figure 2.5: Illustration of the effect of kernel size



(a) DS (b) CC

(c) GLM (d) MI

Figure 2.6: Comparison of fMRI analysis techniques. Detections are denoted as white pixels.

2.4 Binary Hypothesis Testing

This section presents background on binary hypothesis testing, which will be extensively used in Chapter 3.

2.4.1 Bayesian Framework

Let the prior model be $P_0 = Pr(H = H_0)$ and $P_1 = Pr(H = H_1)$ and the measurement model be $p_{\underline{Y}|H}(\underline{y}|H_0)$ and $p_{\underline{Y}|H}(\underline{y}|H_1)$. Let us define the following performance criterion. C_{ij}: Cost of saying $H = H_i$ when in truth $H = H_j$. Then the problem is to determine the decision rule $y \rightarrow \hat{H}(y)$ that minimizes the expected cost. Using standard statistical decision theory and assuming a diagonal cost matrix, $C_{ij} = 1 - \delta_{ij}$, this results in

$$Pr(H_1|\underline{y}) \underset{H_0}{\overset{H_1}{\gtrless}} Pr(H_0|\underline{y}) \tag{2.27}$$

which is called the Maximum a Posteriori (MAP) rule.

2.4.2 Neyman-Pearson Lemma

Proposition 2.4.1 (Neyman-Pearson lemma). *[10] Let $X_1, \ldots, X_n$ be observations for a binary hypothesis testing. For $T \geq 0$, define a decision region for $\hat{H} = H_1$*

$$A_n(T) = \left\{ \frac{p(x_1, \ldots, x_n|H_1)}{p(x_1, \ldots, x_n|H_0)} > T \right\}$$

Let $P_F^* = Pr(A_n(T)|H_0)$, $P_D^* = Pr(A_n(T)|H_1)$, be the corresponding probability of false alarm and probability of detection corresponding to decision region A_n. Let B_n be any other decision region with the associated P_F and P_D. If $P_F \leq P_F^*$, then $P_D \leq p_D^*$.

Proof. See [10, page 305]. □

2.5 Markov Random Fields

Markov random field models have been widely used in various engineering problems, especially in image processing problems. In particular, MRFs can be used to smooth an image while preserving the distinctness of an edge [3]. As mentioned before, we will use an Ising model as a spatial prior to implement the idea that the brain activation map has a degree of regularity or smoothness.

Following the description of Geman and Geman [11], let us briefly introduce MRFs.

2.5.1 Graphs and Neighborhoods

Let $S = \{s_1, s_2, \ldots, s_N\}$ be a set of sites and $\mathcal{G} = \{\mathcal{G}_s, s \in S\}$ be a *neighborhood system* for S where $\mathcal{G}_s$ is the set of *neighbors* of site s. The neighborhood system should satisfy following:

- $s \notin \mathcal{G}_s$
- $s \in \mathcal{G}_r \Leftrightarrow r \in \mathcal{G}_s$

Then, $\{S, \mathcal{G}\}$ is an adjacency-list representation [12, pages 465–466] of an undirected graph (V, E). A subset $C \subseteq S$ is a *clique* if every pair of distinct sites in C are neighbors; $\mathcal{C}$ denotes the set of cliques.

2.5.2 Markov Random Fields and Gibbs Distributions

Let $X = \{X_s, s \in S\}$ denote a family of random variables indexed by S, where $X_s \in \Lambda, \Lambda = \{0, 1, 2, \ldots, L-1\}$. Let Ω be the set of all possible *configurations*: $\Omega = \{\omega : (x_{s_1}, \ldots, x_{s_N}) : x_{s_i} \in \Lambda, 1 \leq i \leq N\}$

Definition 2.5.1. *X is an MRF with respect to $\mathcal{G}$ if $P(X = \omega) > 0$ for all $\omega \in \Omega$;*

$$P(X_s = x_s | X_r = x_r, r \neq s) = P(X_s = x_s | X_r = x_r, r \in \mathcal{G}_s)$$

for every $s \in S$ and $(x_{s_1}, \ldots, x_{s_N}) \in \Omega$.

Thus, in an MRF each point X_s, its neighborhood $\mathcal{G}_s$ conveys all the relevant

information from the other $r \in S - \{s\}$.

Definition 2.5.2. *Gibbs distribution relative to* $\{S, \mathcal{G}\}$ *is a probability measure* π *on* Ω *with*

$$\pi(\omega) = \frac{1}{Z} e^{-U(\omega)/T}$$

where U*, called the* energy function*, is of the form*

$$U(\omega) = \sum_{C \in \mathcal{C}} V_C(\omega)$$

and

$$Z = \sum_{\omega} e^{-U(\omega)/T}$$

is a normalizing constant. Here, V_C*, called the* potential*, is a function on* Ω *with the property that* $V_C(\omega)$ *depends only on those coordinates* x_s *of* ω *for which* $s \in C$*.*

Proposition 2.5.1 (Hammersley-Clifford). *Let* $\mathcal{G}$ *be a neighborhood system. Then* X *is an MRF with respect to* $\mathcal{G}$ *if and only if* $\pi(\omega) = P(X = \omega)$ *is a Gibbs distribution with respect to* $\mathcal{G}$*.*

The Ising prior used in Chapter 4 is represented as a Gibbs distribution, which is equivalent to an MRF by the Hammersley-Clifford theorem.

Chapter 3

Interpretation of the Mutual Information in *f*MRI Analysis

In the previous work [2], the mutual information between the protocol and *f*MRI signals is used to detect an activated voxel. This is motivated by the idea that the larger the dependency between an *f*MRI signal and protocol signal, the more likely the voxel is activated. However, the use of mutual information in this context is heuristic with little statistical interpretation.

In this chapter, we specify a binary hypothesis testing problem where the null hypothesis is that an *f*MRI signal is independent of the protocol signal indicating the voxel is not activated and vice versa. The hypothesis testing problem is motivated naturally from the property of mutual information, namely the mutual information quantitatively measures the amount of dependency between two random variables. Interestingly, it turns out that in the previous detection method, thresholding the estimate of the mutual information is an asymptotic likelihood ratio test for that

hypothesis testing model. Therefore, the information theoretic detection method approximates the optimal test in the sense of Neyman-Pearson lemma. Furthermore, this enables the extension of incorporating a spatial prior within a Bayesian framework, which is discussed in Chapter 4.

3.1 Nonparametric Hypothesis Testing Problem

In this section, a hypothesis testing problem for the *f*MRI analysis is specified and the likelihood ratio for the hypothesis testing is approximated using estimates of pdf's.

Remember that $S_{X|U=0}$ and $S_{X|U=1}$ are partition of the time series $\{X_1, \ldots, X_n\}$, such that $\{X_1, \ldots, X_n\} = S_{X|U=0} \cup S_{X|U=1}$, $S_{X|U=0} \cap S_{X|U=1} = \emptyset$, and $|S_{X|U=0}| = |S_{X|U=1}| = \frac{n}{2}$.

Assumption 3.1.1. *The following conditions are assumed on the X_i:*

- *(a) $X_i \in S_{X|U=0}$ are i.i.d. and $X_i \in S_{X|U=1}$ are i.i.d.*
- *(b) $X_i \in S_{X|U=0}$ and $X_i \in S_{X|U=1}$ are independent.*

We propose a mathematical model for two hypotheses, namely, H_0: the voxel is not activated and H_1: the voxel is activated. When the voxel is not activated, the *f*MRI signal X is not related to the state of the protocol U. The following hypothesis testing model implements this idea. In this hypothesis testing, the decision will be

made based on the observation of X_i's in $S_{X|U=0}$ and X_i's in $S_{X|U=1}$.

$$H_0 \ : \ p_{X,U} = p_X p_U, \text{ i.e. } p_{X|U=0} = p_{X|U=1} \text{ (Independent)} \tag{3.1}$$

$$H_1 \ : \ p_{X,U} \neq p_X p_U, \text{ i.e. } p_{X|U=0} \neq p_{X|U=1} \text{ (Dependent)} \tag{3.2}$$

Now, the above hypothesis testing model can be interpreted as follows considering Assumption 3.1.1:

- If H_0 is true, $X_i \in S_{X|U=0} \cup S_{X|U=1}$ are i.i.d. by Assumption 3.1.1. Thus, there exists p_X such that $X_1, \ldots, X_n$ are i.i.d. according to p_X.
- If H_1 is true, there exists $p_{X|U=0}$ and $p_{X|U=1}$ such that $X_i \in S_{X|U=0} \overset{i.i.d.}{\sim} p_{X|U=0}$ and $X_i \in S_{X|U=1} \overset{i.i.d.}{\sim} p_{X|U=1}$, where the marginal pdf p_X is $p_X = (p_{X|U=0} + p_{X|U=1})/2$.
- Note that we consider $p_X, p_{X|U=0}$, and $p_{X|U=1}$ to be *unknown* without making any parametric restrictions on the form of these density functions.

The likelihood of $(X_1, \ldots, X_n)$ given that H_0 is true is

$$p_{\underline{X}|H_0}(X_1, \ldots, X_n|H_0) = \prod_i p_X(X_i). \tag{3.3}$$

The likelihood of $(X_1, \ldots, X_n)$ given that H_1 is true is

$$\begin{aligned} p_{\underline{X}|H_1}(X_1, \ldots, X_n|H_1) &= p(S_{X|U=0}|H_1)p(S_{X|U=1}|H_1) \text{ by Assumption 3.1.1 (b)} \\ &= \prod_{X_t \in S_{X|U=0}} p_{X|U=0}(X_t) \prod_{X_t \in S_{X|U=1}} p_{X|U=1}(X_t) \end{aligned} \tag{3.4}$$

by Assumption 3.1.1 (a).

If the likelihood ratio

$$\frac{p_{\underline{X}|H_1}(X_1,\ldots,X_n|H_1)}{p_{\underline{X}|H_0}(X_1,\ldots,X_n|H_0)} = \frac{\prod_{X_t\in S_{X|U=0}} p_{X|U=0}(X_t)\prod_{X_t\in S_{X|U=1}} p_{X|U=1}(X_t)}{\prod_i p_X(X_i)} \tag{3.5}$$

can be calculated, we can use it as a statistic for an optimal test as a result of the Neyman-Pearson lemma. Suppose we know the densities $p_{X|U=0}$, $p_{X|U=1}$, and p_X. Then the log likelihood ratio is related to $I(X;U)$ as follows:

$$\begin{aligned}
& \log \frac{\prod_{X_t\in S_{X|U=0}} p_{X|U=0}(X_t)\prod_{X_t\in S_{X|U=1}} p_{X|U=1}(X_t)}{\prod_i p_X(X_i)} \\
= & \sum_{X_t\in S_{X|U=0}} \log p_{X|U=0}(X_t) + \sum_{X_t\in S_{X|U=1}} \log p_{X|U=1}(X_t) - \sum_i \log p_X(X_i) \\
= & \frac{n}{2}\frac{1}{|S_{X|U=0}|}\sum_{X_t\in S_{X|U=0}} \log p_{X|U=0}(X_t) + \frac{n}{2}\frac{1}{|S_{X|U=1}|}\sum_{X_t\in S_{X|U=1}} \log p_{X|U=1}(X_t) \\
& - n\frac{1}{|S_{X|U=0}\cup S_{X|U=1}|}\sum_i \log p_X(X_i) \\
\approx & -\frac{n}{2}h(X|U=0) - \frac{n}{2}h(X|U=1) + nh(X) \text{ (Weak law of large numbers)} \\
= & nI(X;U).
\end{aligned} \tag{3.6}$$

A natural question is if we can approximate the log likelihood ratio using estimates of pdf's. We address this question in Section 3.2.

We note that in the case when U is binary with equal probability, the mutual

information can be simplified as follows:

$$\begin{aligned} I(X;U) &= h(X) - \frac{1}{2}h(X|U=0) - \frac{1}{2}h(X|U=1) \\ &= -\int \frac{p_{X|U=0} + p_{X|U=1}}{2} \log p_X dx \\ &\quad + \frac{1}{2}\int p_{X|U=0} \log p_{X|U=0} dx + \frac{1}{2}\int p_{X|U=1} \log p_{X|U=1} dx \\ &= \frac{1}{2}\int p_{X|U=0} \log \frac{p_{X|U=0}}{p_X} dx + \frac{1}{2}\int p_{X|U=1} \log \frac{p_{X|U=1}}{p_X} dx \\ &= \frac{1}{2}D(p_{X|U=0}\|p_X) + \frac{1}{2}D(p_{X|U=1}\|p_X) \end{aligned} \tag{3.7}$$

where $D(\cdot\|\cdot)$ is the Kullback-Leibler divergence and can be loosely interpreted as a distance between 2 pdf's. This is another view of the mutual information in the context of hypothesis testing.

It is interesting to note that this hypothesis testing model is more general than that of the direct subtraction method, where the null hypothesis $E[X|U=0] = E[X|U=1]$ only considers means.

3.2 Nonparametric Likelihood Ratio

In the previous section, some arguments on the log likelihood ratio were made assuming the densities were known. In our case, we do not assume that all of the densities are known. Thus we replace the densities p_X and $p_{X|U}$ with estimates $\hat{p}_X$ and $\hat{p}_{X|U}$.

Let $X_1, \ldots, X_n$ be i.i.d. according to unknown $p_X(x)$. We are going to calculate an approximation of the likelihood of the sequence, $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ which will be used to calculate an approximation of the likelihood ratio mentioned in (3.5).

The likelihood of the sequence, $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$ is

$$\begin{aligned} p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= \prod_{t=1}^{n} p_X(x_t) \\ &= \exp(\sum_t \log p_X(x_t)) \\ &= \exp(\sum_t \log \hat{p}_X(x_t) + \sum_t \log \frac{p_X(x_t)}{\hat{p}_X(x_t)}) \\ &= \exp\left(-n[\hat{h}(X) + \frac{1}{n}\sum_t \log \frac{\hat{p}_X(x_t)}{p_X(x_t)}]\right) \end{aligned} \tag{3.8}$$

where $\hat{h}(X) \stackrel{\text{def}}{=} -\frac{1}{n}\sum_t \log \hat{p}_X(x_t)$.

Let $f(x) = \frac{1}{n}\sum_t \delta(x - x_t)$ denote a train of impulses at data points and $k(x)$ be the kernel used in $\hat{p}_X(x)$. Then,[1]

$$\hat{p}_X(x) = \frac{1}{n}\sum_t k(x - x_t) = \left(\frac{1}{n}\sum_t \delta(x - x_t)\right) * k(x) = f * k(x). \tag{3.9}$$

We make following approximation on the term in (3.8).

$$\frac{1}{n}\sum_t \log \frac{\hat{p}_X(x_t)}{p_X(x_t)} = \int f(x) \log \frac{f * k(x)}{p_X(x)} dx \tag{3.10}$$

$$\approx \int f * k(x) \log \frac{f * k(x)}{p_X(x)} dx \tag{3.11}$$

$$= D(\hat{p}_X \| p_X). \tag{3.12}$$

The difference between (3.10) and (3.11) is that the impulse train $f(x)$ in the integrand was replaced by a smoothed version, $f * k(x)$. Note that the area under $k(x)$ is exactly 1 and area of the impulse is also 1. Thus, if $\log \frac{f*k(x)}{p_X(x)}$ is smooth, i.e. slowly

[1] In this section, we use the standard form of Parzen window density estimator instead of the leave-one-out method for the purpose of analysis.

varying relative to the size of the kernel $k(x)$ then (3.11) is a good approximation of (3.10). Therefore, (3.8) can be approximated as follows:

$$p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{t=1}^{n} p_X(x_t) \tag{3.13}$$

$$\approx e^{-n[\hat{h}(X)+D(\hat{p}_X\|p_X)]}. \tag{3.14}$$

This is an extension of the notion of typicality in that for a "typical" sequence, its likelihood is approximately $e^{-n\hat{h}(X)}$. Furthermore, this reminds us the well-known theorem of the method of types [10, page 281] since $\hat{p}_X$ behaves like the type of the sequence. Intuitively, the term $e^{-nD(\hat{p}_X\|p_X)}$ indicates that the larger the distance between the empirical distribution and the true distribution, the lower the likelihood of the sequence.

Proposition 3.2.1. *Assume that* (3.14) *is true, then the likelihood ratio can be approximated as follows:*

$$\frac{p_{\underline{X}|H_1}(X_1,\ldots,X_n|H_1)}{p_{\underline{X}|H_0}(X_1,\ldots,X_n|H_0)} \approx e^{n(\hat{I}(X;U)-\gamma)}$$

where $\gamma = \frac{1}{2}[D(\hat{p}_{X|U=0}\|p_{X|U=0}) + D(\hat{p}_{X|U=1}\|p_{X|U=1})] - D(\hat{p}_X\|p_X)$.

Proof.

$$p_{\underline{X}|H_0}(X_1,\ldots,X_n|H_0) = \prod_i p_X(X_i) \text{ by (3.3)}$$

$$\approx e^{-n[\hat{h}(X)+D(\hat{p}_X\|p_X)]} \text{ by (3.14)}$$

$$p_{\underline{X}|H_1}(X_1,\ldots,X_n|H_1) = \prod_{X_t\in S_{X|U=0}} p_{X|U=0}(X_t) \prod_{X_t\in S_{X|U=1}} p_{X|U=1}(X_t) \text{ by (3.4)}$$

$$\approx e^{-\frac{n}{2}[\hat{h}(X|U=0)+D(\hat{p}_{X|U=0}\|p_{X|U=0})]} e^{-\frac{n}{2}[\hat{h}(X|U=1)+D(\hat{p}_{X|U=1}\|p_{X|U=1})]} \text{ by (3.14)}$$

Thus the likelihood ratio is

$$\begin{aligned} & \frac{p_{\underline{X}|H_1}(X_1, \ldots, X_n|H_1)}{p_{\underline{X}|H_0}(X_1, \ldots, X_n|H_0)} \\ \approx\ & e^{n[\hat{h}(X) - \frac{1}{2}\hat{h}(X|U=0) - \frac{1}{2}\hat{h}(X|U=1) + D(\hat{p}_X \| p_X) - \frac{1}{2}D(\hat{p}_{X|U=0} \| p_{X|U=0}) - \frac{1}{2}D(\hat{p}_{X|U=1} \| p_{X|U=1})]} \\ =\ & e^{n(\hat{I}(X;U) - \gamma)}. \end{aligned}$$

□

This proposition shows how the true likelihood described in (3.5) can be approximated using estimates of the pdf's when the underlying pdf's are unknown. $\gamma = \frac{1}{2}[D(\hat{p}_{X|U=0}\|p_{X|U=0}) + D(\hat{p}_{X|U=1}\|p_{X|U=1})] - D(\hat{p}_X\|p_X)$ can be considered as the unestimated residual term in this estimation. Let us discuss the properties of γ. One property of γ is that it is nonnegative because of the convexity of Kullback-Leibler divergence since $p_X = \frac{1}{2}p_{X|U=0} + \frac{1}{2}p_{X|U=1}$ and $\hat{p}_X = \frac{1}{2}\hat{p}_{X|U=0} + \frac{1}{2}\hat{p}_{X|U=1}$.[2] Another property of γ is that it is unknown since it is a function of both true pdf's and estimates of pdf's.

Since γ is unknown, we choose a nonnegative value for γ when we use this approximation of the likelihood ratio. This corresponds to choosing a suitable threshold of the estimate of the mutual information in our previous approach. Then our information theoretic approach, where a voxel is declared to be active if the estimated mutual information of the voxel is above a positive threshold, can be viewed as an asymptotic likelihood ratio test for the hypothesis testing problem. Considering that likelihood ratio test is optimal in the sense of the Neyman-Pearson lemma mentioned in Section 2.4, we can say that the test based on the estimate of the mutual information is close to an optimal test for this hypothesis testing problem.

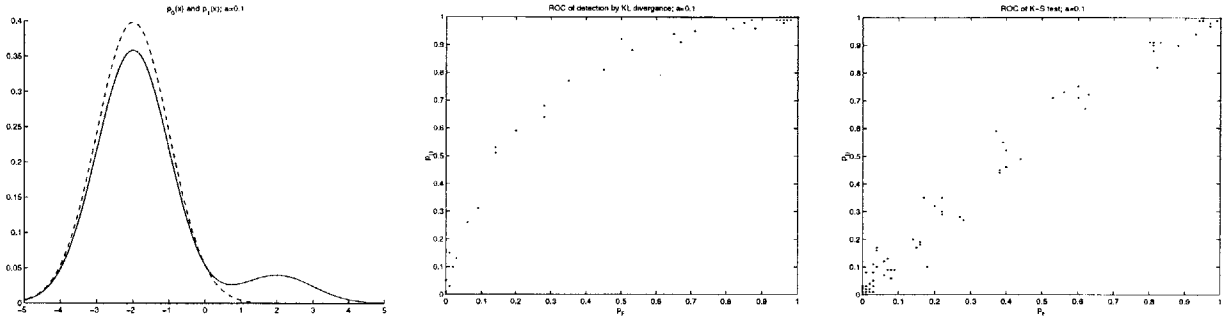
[2]For the case of leave-one-out method, $\hat{p}_X \approx \frac{1}{2}\hat{p}_{X|U=0} + \frac{1}{2}\hat{p}_{X|U=1}$.

A Simple Example Comparing Nonparametric MI with the Kolmogorov-Smirnov Test

It is interesting to note that the Kolmogorov-Smirnov test deals with the same hypothesis testing problem specified in (3.1) and (3.2). For this hypothesis testing, the Kolmogorov-Smirnov test is conventionally used. Proposition 3.2.1 combined with the Neyman-Pearson lemma suggests that our information theoretic test is likely to outperform the conventional Kolmogorov-Smirnov test. As an example, we made a problem of discriminating two close pdf's, $p_0 = N(x; -2, 1)$ and $p_1 = 0.9N(x; -2, 1) + 0.1N(x; 2, 1)$. Figure 3.1 shows the empirical ROC for the case when the tests are based on the observation of two sets of 30 i.i.d samples. For each choice of a threshold, rough estimates of probability of detection, P_D was obtained from 100 Monte Carlo trials with each one basing the estimates of p_0 and p_1 on 30 data points draw from the true respective distributions. For each Monte Carlo trial, if the calculated statistic was above the threshold, it added to the count of detection. P_F is estimated in a similar fashion but drawing the two sets of samples from the single pdf p_0. As expected, Figure 3.1(b) and (c) show that the test based on MI performs better than the KS test in this problem.

3.3 Nonparametric Estimation of Entropy

This section discusses the effect of using the ML kernel size formulated in (2.25). Since this section is somewhat of an aside, readers can continue to Chapter 4 without loss of understanding.



(a) two pdf's; p_0 and p_1 (b) ROC of MI test (c) ROC of K-S test

Figure 3.1: Empirical ROC curves for the test deciding whether two pdf's are same or not

Empirical results presented by Hall *et al* [9] suggest that entropy estimators in the form of $-\frac{1}{n}\sum_i \log \frac{1}{n-1}\sum_{j\neq i} k(X_i - X_j, \sigma)$ are likely to overestimate the unknown entropy. The result favors the use of the ML kernel size since the ML kernel size minimizes the entropy estimate.

3.3.1 Estimation of Entropy and Mutual Information

We first explain why our entropy estimator based on Parzen pdf estimator is likely to overestimate the unknown entropy. By the law of large numbers, our entropy estimate

can be approximated as

$$\begin{aligned}
\hat{h}(X) &= -\frac{1}{n}\sum_i \hat{p}_X(X_i) \\
&\approx -\int p_X(x)\log \hat{p}_X(x)dx \\
&= -\int p_X(x)\log p_X(x)dx + \int p_X(x)\log\frac{p_X(x)}{\hat{p}_X(x)}dx \\
&= h(X) + D(p_X\|\hat{p}_X) \\
&\geq h(X). \qquad (3.15)
\end{aligned}$$

Note that $D(p_X\|\hat{p}_X)$ appears because we use $\hat{p}_X$ instead of p_X while X_i is drawn from p_X.

On the other hand, our estimate of the mutual information $\hat{I}(X;U)$ is likely to underestimate $I(X;U)$. This can be understood as follows:

$$\begin{aligned}
\hat{I}(X;U) &= \hat{h}(X) - \frac{1}{2}\hat{h}(X|U=0) - \frac{1}{2}\hat{h}(x|U=1) \\
&\approx h(X) + D(p_X\|\hat{p}_X) - \frac{1}{2}(h(X|U=0) + D(p_{X|U=0}\|\hat{p}_{X|U=0})) \\
&\quad -\frac{1}{2}(h(X|U=1) + D(p_{X|U=1}\|\hat{p}_{X|U=1})) \\
&= I(X;U) + D(p_X\|\hat{p}_X) - \frac{1}{2}(D(p_{X|U=0}\|\hat{p}_{X|U=0}) + D(p_{X|U=1}\|\hat{p}_{X|U=1})) \\
&\leq I(X;U)
\end{aligned}$$

where the inequality comes from the convexity of Kullback-Leibler divergence.

3.3.2 Lower Bound on the Bias of the Entropy Estimator

Let us begin with providing background on the mean of the Parzen density estimator $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x - X_i)$.

$$\begin{aligned} E[\hat{p}(x)] &= E[\frac{1}{n}\sum_{i=1}^{n} k(x - X_i)] \\ &= E[k(x - X)] \\ &= \int k(x - y)p_X(y)dy \\ &= p_X(x) * k(x) \end{aligned} \tag{3.16}$$

and thus the pdf estimate is clearly biased for any kernel that is not the Dirac impulse.

Lemma 3.3.1. For any fixed σ,

$$E[\hat{h}(X)] \geq h(X) + D(p\|E[\hat{p}]) \geq h(X).$$

Proof.

$$
\begin{aligned}
E[\hat{h}(X)] &= -E[\frac{1}{n}\sum_i \log \frac{1}{n-1}\sum_{j\neq i} k(X_i - X_j, \sigma)] && (3.17)\\
&= -\frac{1}{n}\sum_i E[\log \frac{1}{n-1}\sum_{j\neq i} k(X_i - X_j, \sigma)] && (3.18)\\
&= -\frac{1}{n}\sum_i E[E[\log \frac{1}{n-1}\sum_{j\neq i} k(X_i - X_j, \sigma)|X_i]] && (3.19)\\
&= -E[E[\log \overbrace{\frac{1}{n-1}\sum_{j\neq 1} k(X_1 - X_j, \sigma)}^{Y} |X_1]] && (3.20)\\
&\geq -E[\log E[\overbrace{\frac{1}{n-1}\sum_{j\neq 1} k(X_1 - X_j, \sigma)}^{Y} |X_1]] && (3.21)\\
&= -E[\log E[k(X_1 - X_2, \sigma)|X_1]] && (3.22)\\
&= -E[\log(p * k(X_1))] \text{ by (3.16)} && (3.23)\\
&= -\int p(x)\log(p * k(x))dx && (3.24)\\
&= \int p(x)\log\frac{1}{p(x)}dx + \int p(x)\log\frac{p(x)}{p * k(x)}dx && (3.25)\\
&= h(X) + D(p||p * k) && (3.26)\\
&= h(X) + D(p||E[\hat{p}]) \text{ by (3.16)} && (3.27)
\end{aligned}
$$

where the inequality comes from Jensen's inequality, $E[\log Y] \leq \log E[Y]$. □

Consequently,

$$\inf_\sigma E[\hat{h}(X)] \geq h(X) + \inf_\sigma D(p||E[\hat{p}]) \geq h(X).$$

Note that this does not show that the bias of the entropy estimator using the ML kernel size is positive, since σ was fixed, i.e. independent of the data. The following conjecture is stronger argument than the Lemma 3.3.1.

Conjecture 3.3.1.

$$E[\min_{\sigma>0} -\frac{1}{n}\sum_{i}\log\frac{1}{n-1}\sum_{j\neq i}k(X_i - X_j, \sigma)] \geq h(X)$$

If this conjecture is true, the ML kernel size is the kernel size which achieves the minimum bias. However, the verification of this conjecture is beyond the scope of this thesis.

3.3.3 Open Questions

How to choose the kernel shape and size which minimizes the mean square error,

$$E[(\hat{h}(X) - h(X))^2]$$

is still an open question. Another question is whether an unbiased estimator of $h(X)$ exists. Our conjecture is that there is no such estimator, considering that there is no unbiased estimator of pdf $p_X(x)$ [13].

Chapter 4

Bayesian Framework Using the Ising Model as a Spatial Prior

In the research in [2], activated voxels are detected without considering the spatial dependency among neighboring voxels. However, we know *a priori* that activated regions are localized and neighboring voxels are likely to have the same activation states. In this chapter, this knowledge is incorporated in our information theoretic approach by using an Ising model, a simple Markov random field (MRF), as a spatial prior of the binary activation map. This enables the removal of isolated spurious responses to be done automatically.

Descombes *et al* [3] also use such an MRF, specifically Potts model[1], as a spatial prior of a ternary activation map. However, the MRF prior model is combined with a heuristic data attachment term in place of the likelihood ratio. The main

[1]Potts model is an M-ary ($M \geq 3$) Gibbs field with the same lattice structure as Ising model.

component of the data attachment term is a potential for the Gibbs field designed to enforce the idea that a voxel is likely to be activated if the norm of the estimated hemodynamic function is above a threshold. In addition, they use simulated annealing to solve the estimation problem which is an energy minimization problem with no guarantee of an exact solution.

In contrast, our method uses the asymptotic likelihood ratio developed in Chapter 3 as a principled data attachment term. Furthermore, the MAP estimation problem in this method can be reduced to a minimum cut problem in a flow network, which can be solved in polynomial time by the well-known Ford-Fulkerson method. This reduction from MAP estimation of the binary image to a minimum cut problem was found by Greig [4]. Using Greig's result, the Ising model can be efficiently incorporated in the *f*MRI analysis.

4.1 Ising Model

The Ising model captures the notion that neighboring voxels of an activated voxel are likely to be activated and similarly for nonactivated voxels. Specifically, let $y(i,j,k)$ be a binary activation map such that $y(i,j,k) = 1$ if voxel (i,j,k) is activated and 0, otherwise. Then this idea can be formulated using an Ising model as the prior probability of the activation map $y(\cdot,\cdot,\cdot)$. Let $\Omega = \left\{\omega : \omega \in \{0,1\}^{N_1 \times N_2 \times N_3}\right\}$ be the set of all possible [0-1] configurations and let $\omega(i,j,k)$ be a component of any one sample configuration $\omega(\cdot,\cdot,\cdot)$.

The Ising prior on $y(\cdot,\cdot,\cdot)$ penalizes every occurrence of neighboring voxels

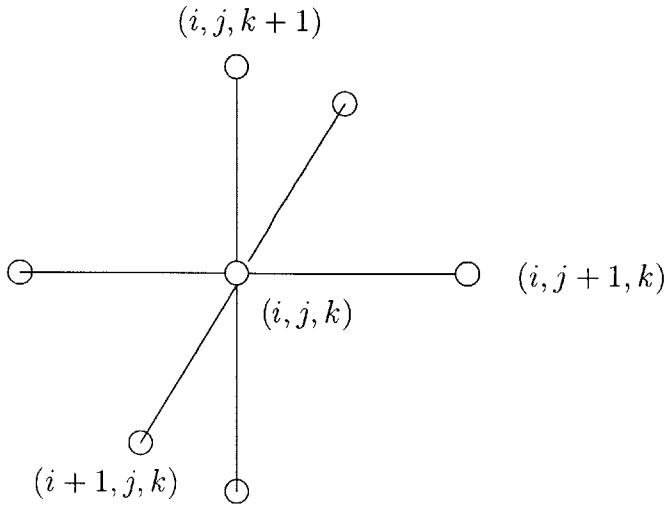


Figure 4.1: Lattice structure of the Ising model

with different activation states as follows:

$$P(y(\cdot,\cdot,\cdot) = \omega) = \frac{1}{Z} e^{-U(\omega)} \qquad Z = \sum_{\omega \in \Omega} e^{-U(\omega)}$$

$$\begin{aligned} U(\omega) &= \beta \sum_{i,j,k} (\omega(i,j,k) \oplus \omega(i+1,j,k) + \omega(i,j,k) \oplus \omega(i,j+1,k) \\ &\quad + \omega(i,j,k) \oplus \omega(i,j,k+1)), \end{aligned} \tag{4.1}$$

where $\beta > 0$. Note that in the summation in (4.1), each pair of adjacent voxels is counted exactly once.

4.2 Formulation of Maximum a Posteriori Detection Problem

As in [3], we assume that

$$p(X(\cdot,\cdot,\cdot,\cdot)|Y(\cdot,\cdot,\cdot)) = \prod_{i,j,k} p(X(i,j,k,\cdot)|Y(i,j,k)).$$

That is, conditioned on the activation map, voxel time-series are independent. The MAP estimate of the activation is then

$$\begin{aligned}
\hat{Y}(\cdot,\cdot,\cdot) &= \arg\max_{y(\cdot,\cdot,\cdot)} p(Y = y(\cdot,\cdot,\cdot)) p(X(\cdot,\cdot,\cdot,\cdot)|Y(\cdot,\cdot,\cdot) = y(\cdot,\cdot,\cdot)) \\
&= \arg\max_{y(\cdot,\cdot,\cdot)} p(Y = y(\cdot,\cdot,\cdot)) \prod_{i,j,k} [p(X(i,j,k,\cdot)|Y(i,j,k) = 1)^{y(i,j,k)} \\
&\quad p(X(i,j,k,\cdot)|Y(i,j,k) = 0)^{1-y(i,j,k)}] \\
&= \arg\max_{y(\cdot,\cdot,\cdot)} \log p(Y = y(\cdot,\cdot,\cdot)) + \sum_{i,j,k} y(i,j,k) \log \frac{p(X(i,j,k,\cdot)|Y(i,j,k) = 1)}{p(X(i,j,k,\cdot)|Y(i,j,k) = 0)} \\
&= \arg\max_{y(\cdot,\cdot,\cdot)} \sum_{i,j,k} \lambda_{i,j,k} y(i,j,k) - \beta \sum_{i,j,k} (Y(i,j,k) \oplus Y(i+1,j,k) \\
&\quad + Y(i,j,k) \oplus Y(i,j+1,k) + Y(i,j,k) \oplus Y(i,j,k+1)),
\end{aligned} \tag{4.2}$$

where $\lambda_{i,j,k} = \ln \frac{p(X(i,j,k,\cdot)|Y(i,j,k)=1)}{p(X(i,j,k,\cdot)|Y(i,j,k)=0)} = n(\hat{I}_{i,j,k}(X;U) - \gamma)$ is the log-likelihood ratio at voxel (i,j,k) and $\hat{I}_{i,j,k}(X;U)$ is the mutual information estimated from time-series $X(i,j,k,\cdot)$. The previous use of MI as the activation statistic fits readily into the MAP formulation. Note that H_0 and H_1 in Chapter 3 correspond to $Y(i,j,k) = 0$ and $Y(i,j,k) = 1$ respectively.

4.3 Exact Solution of the MAP Problem

There are 2^{N_v} possible configurations of $y(\cdot,\cdot,\cdot)$ (or equivalently elements of the set Ω) where $N_v = N_1 N_2 N_3$ is the number of voxels. It has been shown by Greig *et al* [4] that this seemingly NP-complete problem can be solved *exactly* in polynomial time (order N_v) by reducing the MAP estimation problem to the minimum cut problem in a flow network [4]. Greig *et al* accomplished this by demonstrating that under certain conditions, the binary image MAP estimation problem (using an MRF prior) can be reduced to the minimum cut problem of a network flow. Consequently, the methodology of Ford and Fulkerson for such problems can be applied directly. We are able to employ the same technique as a consequence of demonstrating the equivalence of MI to the log-likelihood ratio of a binary hypothesis testing problem.

This approach has several advantages over simulated annealing. In simulated annealing, it is difficult to make concrete statements about the final solution with regard to the optimization criterion, since it may be a local minimum. However, if we know the exact solution of the optimization criterion, we can ignore issues of local minima and solely consider the quality of optimization criterion [4]. Specifically, we can explore the effect of varying parameter β of the prior and obtain some intuition on it.

In this section, we describes how the reduction of the MAP problem to the mincut problem is possible based on the Greig's work. For notational convenience, let us use a 1 dimensional spatial coordinate i instead of 3 D spatial coordinate (i, j, k). Let $y = (y_1, \ldots, y_m)$ denote the brain activation map where y_i takes on 0 or 1 and m is the number of voxels.

- The prior on Y is

$$p(Y = y) = exp[-\sum_{i<j} \beta_{ij}(y_i - y_j)^2]$$

where β_{ij} is β if i and j are neighbors and 0, otherwise.

- The likelihood ratio is

$$\frac{p(X_i(\cdot)|Y_i = 1)}{p(X_i(\cdot)|Y_i = 0)} = e^{n(\hat{I}_i(X;U)-\gamma)}.$$

The MAP problem is

$$\begin{aligned}
\hat{Y} &= \arg\max_y p(y) \prod_{i=1}^{m} p(X_i(\cdot)|Y_i = 1)^{y_i} p(X_i(\cdot)|Y_i = 0)^{1-y_i} \\
&= \arg\max_y \log p(y) + \sum_{i}^{m} y_i \log \frac{p(X_i(\cdot)|Y_i = 1)}{p(X_i(\cdot)|Y_i = 0)} \\
&= \arg\max_y \sum_{i=1}^{m} \lambda_i y_i - \sum_{i<j} \beta_{ij}(y_i - y_j)^2 \qquad (4.3)
\end{aligned}$$

where $\lambda_i = \ln p(x_i|1)/p(x_i|0) = n(\hat{I}_i(X;U) - \gamma)$ is the log-likelihood ratio at voxel i.

4.3.1 Preliminaries of Flow Networks

Here we repeat some definitions from [12], which are essential to understand how the flow network can be designed such that minimum capacity cut is the solution of the MAP problem.

Definition 4.3.1 (flow network). *A* flow network $G = (V, E)$ *is a directed graph with two special vertices, a* source s *and a* sink t *in which each edge* $(u, v) \in E$ *has a*

nonnegative capacity $c(u,v) \geq 0$ *and* $c(u,v) = 0$ *if* $(u,v) \notin E$.

Definition 4.3.2 (flow). *A* flow *in* G *is a real-valued function* $f : V \times V \rightarrow R$ *that satisfies the following three properties:*

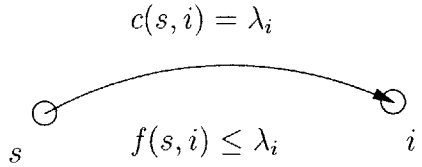
- *Capacity constraint: For all* $u, v \in V$, *we require* $f(u,v) \leq c(u,v)$.
- *Skew symmetry: For all* $u, v \in V$, $f(u,v) = -f(v,u)$.
- *Flow conservation: For all* $u \in V - \{s,t\}$, $\sum_{v \in V} f(u,v) = 0$.

Definition 4.3.3.

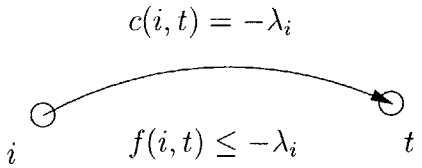
- *The* value *of a flow is defined as* $|f| = \sum_{v \in V} f(s,v)$, *that is, the total net flow out of the source.*
- $$f(X,Y) = \sum_{x \in X} \sum_{y \in Y} f(x,y)$$
- $$c(X,Y) = \sum_{x \in X} \sum_{y \in Y} c(x,y)$$
- *The residual capacity:* $c_f(u,v) = c(u,v) - f(u,v)$
- *The residual network of* G *induced by* f *is* $G_f = (V, E_f)$, *where* $E_f = \{(u,v) \in V \times V : c_f(u,v) > 0\}$.

Definition 4.3.4 (cut). *A* cut (S,T) *of flow network* $G = (V,E)$ *is a partition of* V *into* S *and* $T = V - S$ *such that* $s \in S$ *and* $t \in T$.

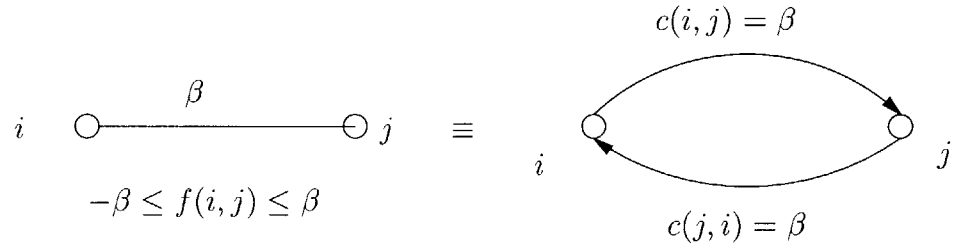
Definition 4.3.5 (cut capacity). *If* f *is a flow, then the net flow across the cut* (S,T) *is defined to be* $f(S,T)$. *The* capacity *of the cut* (S,T) *is* $c(S,T)$.



(a) when $\lambda_i > 0$



(b) when $\lambda_i \leq 0$



(c) when i and j are neighbors

Figure 4.2: Constructing a capacitated network

4.3.2 Reduction of the binary MAP problem to the Minimum Cut Problem in a Flow Network

In this section, we discuss how the minimum capacity cut corresponds to the solution of the MAP problem elaborating Greig's work [4].

We form a capacitated network composed of $m + 2$ vertices, which consist of a source s, a sink t, and the m voxels, as follows:

- As illustrated in Figure 4.2(a), if the log-likelihood ratio $\lambda_i = \hat{I}_i(X;U) - \gamma >$

0, i.e. the estimate of the mutual information of the ith voxel is above the threshold, then there is a directed edge (s, i) from s to voxel i with capacity $c(s, i) = \lambda_i = n(\hat{I}_i(X; U) - \gamma)$. A directed edge (s, i) implies that $c(i, s) = 0$ because if there is no edge in the flow network, the capacity is zero. Then $f(s, i) \leq \lambda_i$ by the capacity constraint and $-f(s, i) = f(i, s) \leq c(i, s) = 0$ by skew symmetry and the capacity constraint. Thus, $0 \leq f(s, i) \leq \lambda_i$.

- As illustrated in Figure 4.2(b), if $\lambda_i \leq 0$, i.e. the ith voxel has an estimate mutual information less than or equal to the threshold, then there is a directed edge (i, t) from i to t with capacity $c(i, t) = -\lambda_i = -n(\hat{I}_i(X; U) - \gamma)$. Similarly, $0 \leq f(i, t) \leq -\lambda_i$.

- As illustrated in Figure 4.2(c), if voxel i and voxel j are neighbors, then there is an undirected edge (i, j), i.e. two directed edges (i, j) and (j, i) between two internal vertices (voxels) i and j with capacity $c(i, j) = c(j, i) = \beta$.[2] And the flows must satisfy $f(i, j) \leq c(i, j) = \beta, -f(i, j) = f(j, i) \leq c(j, i) = \beta \rightarrow -\beta \leq f(i, j) \leq \beta$.

Note that the resulting flow network has information relevant to the MAP problem. Voxels with a positive log-likelihood ratio are connected to the source node. The higher the mutual information of such a voxel, the higher the capacity of the edge from the source to the voxel. Similarly, voxels with a negative log-likelihood ratio are connected to the sink node with a capacity such that the less likely the voxel is to be active, the higher the capacity from the voxel to the sink. The prior is implemented as undirected edges between neighboring voxles. Cutting such edges corresponds to the occurence of two neighboring voxels with different activation states. Suppose we have an activation map $y = (y_1, \ldots, y_m)$ and make a corresponding cut of the set of

[2] the parameter of the Ising prior

nodes, i.e. a partition of the nodes into two sets where the source node is in one set and the sink node is in the other such that active voxels are in the set with source node and nonactive voxels are in the other set. Then, for the cut, the cut capcity is sum of the capacities of the edges where cutting is applied. Intuitively, high capacity between two nodes means high bonding power between the two nodes and cutting an edge of high capacity results in high penalty thus making cutting such an edge less desirable. A more formal description showing the equivalence of the minimum cut to the MAP estimate is as follows.

For *any* binary image $y = (y_1, \ldots, y_m)$ let $S = \{s\} \cup \{i : y_i = 1\}$ and $T = \{i : y_i = 0\} \cup \{t\}$ be *the corresponding cut.* Then the corresponding cut capacity is

$$\begin{aligned} C(y) = C(S,T) &= \sum_{i \in S} \sum_{j \in T} c(i,j) && (4.4) \\ &= \sum_{i \in T} c(s,i) + \sum_{i \in S} c(i,t) + \sum_{i;y_i=1} \sum_{j;y_j=0} c(i,j) \\ &= \sum_{i} (1 - y_i) c(s,i) + \sum_{i} y_i c(i,t) + \sum_{i;y_i=1} \sum_{j;y_j=0} \beta_{ij} && (4.5) \end{aligned}$$

Note that $c(s,i) = \max(0, \lambda_i)$ since there is an edge (s,i) if and only if $\lambda_i > 0$ as illustrated in Figure 4.2(a). Thus, $\sum_i c(s,i)(1-y_i) = \sum_i (1-y_i) \max(0, \lambda_i)$. Similarly, $\sum_i c(i,t) y_i = \sum_i y_i \max(0, -\lambda_i)$. Because $\beta_{ij} = \beta_{ji}$, we have that

$$\begin{aligned} 2 \sum_{i;y_i=1} \sum_{j;y_j=0} \beta_{ij} &= \sum_{i;y_i=1} \sum_{j;y_j=0} \beta_{ij} + \sum_{i;y_i=0} \sum_{j;y_j=1} \beta_{ij} \\ &= \sum_{i} \sum_{j} \beta_{ij} (y_i - y_j)^2 \\ &= 2 \sum_{i<j} \beta_{ij} (y_i - y_j)^2. && (4.6) \end{aligned}$$

Therefore,

$$C(y) = \sum_i (1 - y_i) \max(0, \lambda_i) + \sum_i y_i \max(0, -\lambda_i) + \sum_{i<j} \beta_{ij}(y_i - y_j)^2 \quad (4.7)$$

Finally, $\arg\min_y C(y)$ is the MAP solution specified by (4.3), since

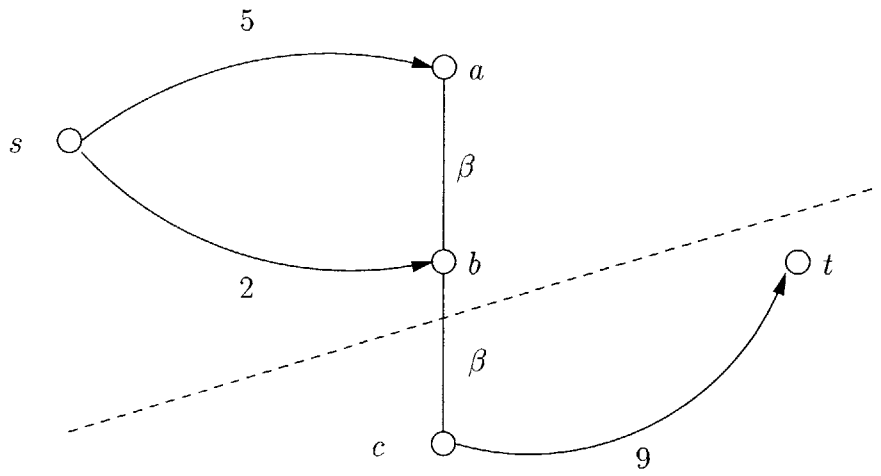
$$\begin{aligned} C(y) &= \sum y_i \frac{|\lambda_i| - \lambda_i}{2} + \sum (1 - y_i) \frac{|\lambda_i| + \lambda_i}{2} + \sum_{i<j} \beta_{ij}(y_i - y_j)^2 \\ &= -\sum \lambda_i y_i + \sum_{i<j} \beta_{ij}(y_i - y_j)^2 + \text{terms independent of } y. \end{aligned} \quad (4.8)$$

Note that Greig's method can be applied to an optimization problem if and only if it can be cast in the form of (4.3). As a consequence, this method can solve not only problems with Ising model but also problems with more complex lattice structures as long as the prior has only terms of *pairwise* interaction, i.e. doubleton potential, and Y_i is *binary*. Furthermore, it has been proven that the optimization problem is NP-hard if the Y_i takes on more than two values [14].

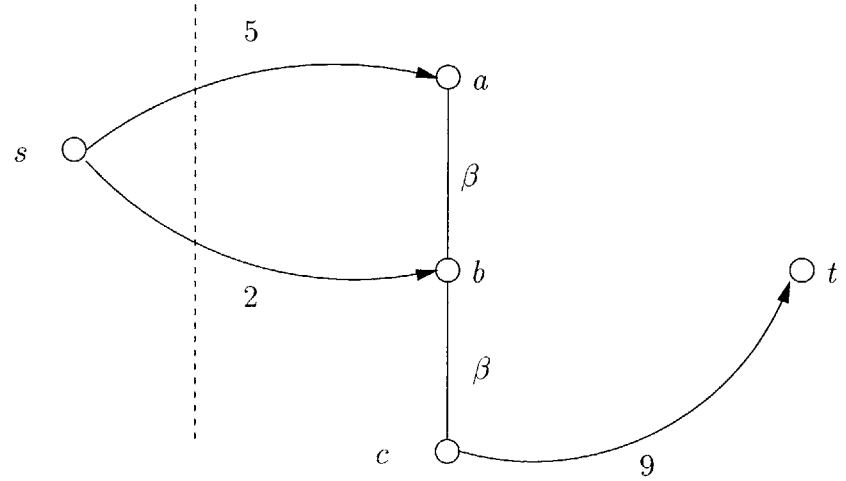
Illustrating Example

Figure 4.3 illustrates the cut that achieves the minimum cut capacity in a simple flow network. You can consider nodes a and b as voxels with mutual information estimates above the threshold and c as a voxel with mutual information estimate below the threshold.

As a special case, when $\beta = 0$, the minimum cut capacity is trivially 0 and the cut is shown in Figure 4.3(a). This situation corresponds to the case when there is no prior and consequently voxel a and b are declared to be active and c is declared



(a) $0 \leq \beta < 7$



(b) $7 < \beta$

Figure 4.3: Examples of the cuts that minimize cut capacities

to be nonactive. Increasing beta up to 7 does not change the cut.

If $\beta = 8$, for instance, the minimum cut capacity is 7 as shown in Figure 4.3(b), and all the voxels are declared to be nonactive. This is because node c is connected to sink node with high capacity and nodes a and b are tied to node c with the high capacity β.

4.3.3 Solving Minimum Cut Problem in Flow Network: Ford and Fulkerson Method

The binary activation map which minimizes the cut capacity can be calculated by the Ford and Fulkerson method which finds the flow that maximizes the net flow from source to sink. The following lemmas and proposition relate the cut to the flow.

Lemma 4.3.1. *[12] Let f be a flow in a flow network G with source s and sink t, and let (S,T) be a cut of G. Then, the net flow across (S,T) is $f(S,T) = |f|$.*

Proof. See [12, page 592] for the proof. □

Lemma 4.3.2. *[12] The value of any flow f in a flow network G is bounded from above by the capacity of any cut of G.*

Proof. See [12, page 592] for the proof. □

Proposition 4.3.1 (Max-flow min-cut theorem). *[12]*

If f is a flow in a flow network $G = (V,E)$ with source s and sink t, then the following conditions are equivalent:

1. *f is a maximum flow in G*
2. *the residual network G_f contains no path from s to t.*
3. *$|f| = c(S,T)$ for some cut (S,T) of G.*

Proof. We repeat the proof of [12, page 593].

(1) $\Rightarrow$ (2): Suppose that f is a maximum flow in G but that G_f has an augmenting path p. Let's define a function $f_p : V \times V \rightarrow R$ by

$$f_p(u,v) = \begin{cases} c_f(p) & \text{if } (u,v) \text{ is on } p, \\ -c_f(p) & \text{if } (v,u) \text{ is on } p, \\ 0 & \text{otherwise.} \end{cases}$$

where $c_f(p) = \min\{c_f(u,v) : (u,v) \text{ is on } p\}$. Then, the flow sum $f + f_p$ is a flow in G with value strictly greater than $|f|$, contradicting the assumption that f is a maximum flow.

(2) $\Rightarrow$ (3): Suppose G_f has no path from s to t. Let $S = \{v \in V :$ there exists a path from s to v in $G_f\}$ and $T = V - S$. Then the partition (S,T) is a cut, since $s \in S$ and $t \notin S$. For $\forall u \in S$ and $\forall v \in T$, $f(u,v) = c(u,v)$, since otherwise $(u,v) \in E_f$ and $v \in S$. Thus $|f| = f(S,T) = c(S,T)$ by Lemma 4.3.1.

(3) $\Rightarrow$ (1): By Lemma 4.3.2, $|f| \leq c(S,T)$ for all cuts (S,T). The condition $|f| = c(S,T)$ thus implies both that f is a maximum flow in G and that $c(S,T)$ is a minimum cut capacity. □

Description of the Ford-Fulkerson Method

We use the Ford-Fulkerson method to solve the max-flow problem. Figure 4.4 describes the Ford-Fulkerson method. As a result of the max-flow min-cut theorem, this method also finds the min-cut of a given flow network.

When the method finishes calculating the maximum flow, the residual network

G_f gives the minimum cut as follows. We make two sets $S = \{v \in V$: there exists a path from s to v in $G_f\}$ and $T = V - S$. Then the partition (S, T) is the minimum capacity cut considering the proof of (2) $\Rightarrow$ (3) and (3) $\Rightarrow$ (1) of the max-flow min-cut theorem.

In implementing the Ford-Fulkerson method, we use the breadth-first search (BFS) to find a path p from s to t in the residual network G_f. The computation time of the Ford-Fulkerson method in our *f*MRI application depends mainly on the number of voxels whose MI estimates are above γ, i.e. the voxels connected to the source because the BFS starts from the source and searches over all the nodes connected to the source. Note that most of the voxels have low MI estimates and are connected to sink node t since the protocol is designed to activate a localized region of brain such as the motor cortex, auditory cortex, or visual cortex. This makes the BFS (the bottleneck of our algorithm) fast. However, if γ is low, for example 0.4 bit, then BFS part of the algorithm can become extensive.

It is interesting to note that if $\hat{I}_i(X;U) > \gamma + \frac{6\beta}{n}$, i.e. $\lambda_i > 6\beta$, then voxel i must be active by inspection of (4.3). Suppose, k voxels among 6 neighboring voxels of voxel i are nonactive. Then when $y_i = 1$, (4.3) has $\lambda_i - k\beta$ and other terms independent of y_i. When $y_i = 0$, (4.3) has $-(6-k)\beta$ and terms independent of y_i. The condition $\lambda_i > 6\beta$ thus implies that $\lambda_i - k\beta - \{-(6-k)\beta\} = \lambda_i + (6-2k)\beta > \lambda_i - 6\beta > 0$ and consequently $y_i = 1$. Similarly, if $\hat{I}_i(X;U) < \gamma - \frac{6\beta}{n}$, i.e. $\lambda_i < -6\beta$, the voxel i must be nonactive. We have not used this information in the Ford-Fulkerson method but it may be exploited as the initialization of the method.

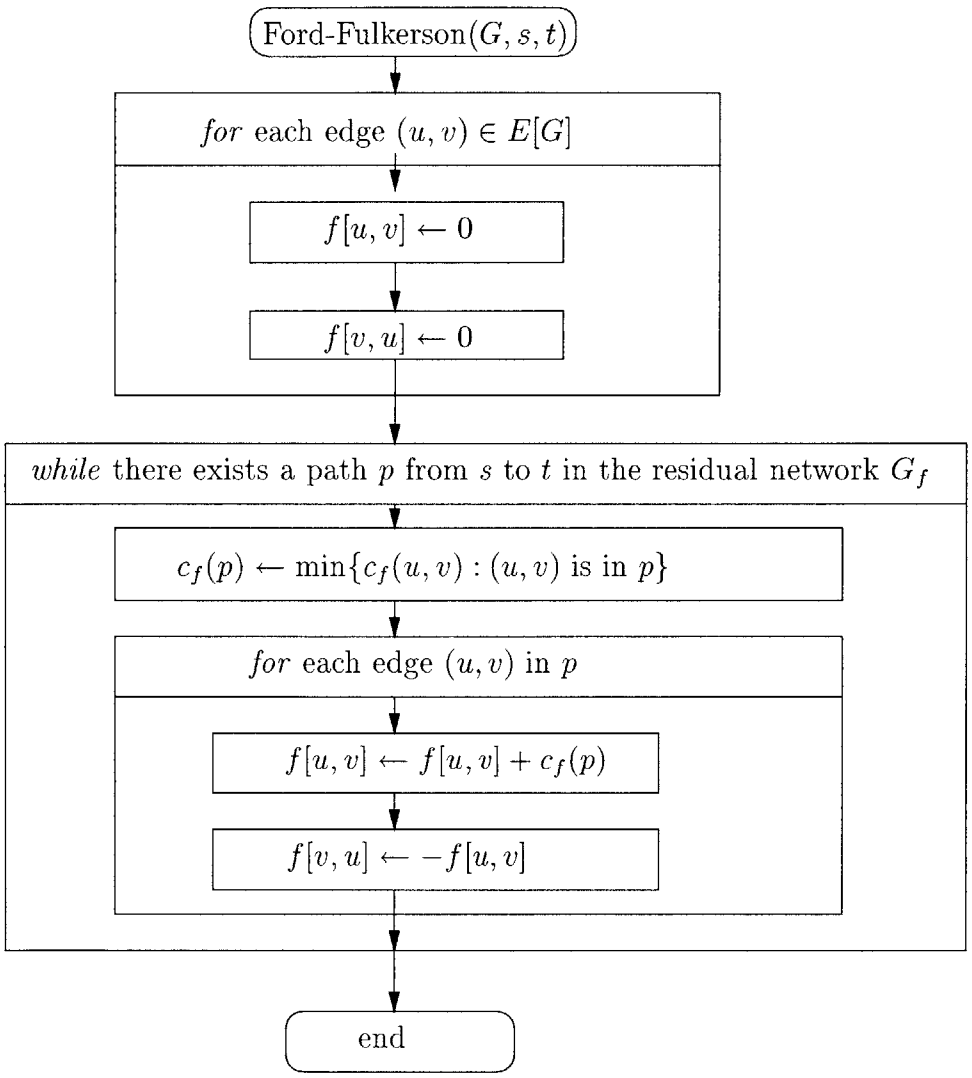


Figure 4.4: Flow chart of the Ford-Fulkerson method

Chapter 5

Experimental Results

This chapter presents the results of the MAP detection method developed in Chapter 3 and Chapter 4 which simultaneously detects active voxels and removes isolated spurious responses. With traditional techniques, activation decisions at each voxel are independent of neighboring voxels. Spurious responses are then removed by *ad hoc* techniques (e.g. morphological operators). The weakness of this technique is that a morphological operator does not consider the evidence in the data. In contrast, via the Ising prior, our method does consider the evidence in the data when removing spurious responses. Section 5.1 demonstrates this idea. The performance of our novel method is then compared with conventional methods such as the GLM and Kolmogorov-Smirnov test. The direct comparison of the performance of these methods in *f*MRI analysis is difficult without absolute ground truth. To deal with this difficulty, the experiments are designed to demonstrate the *relative modeling capacity* of each standard approach to our new approach. In Section 5.2 and Section 5.3, an artificial ground truth is constructed from which ROC curves showing the relative modeling capacity are generated. More discussion on the comparison of the GLM

and our method is made in Section 5.4

5.1 Effect of Ising Prior

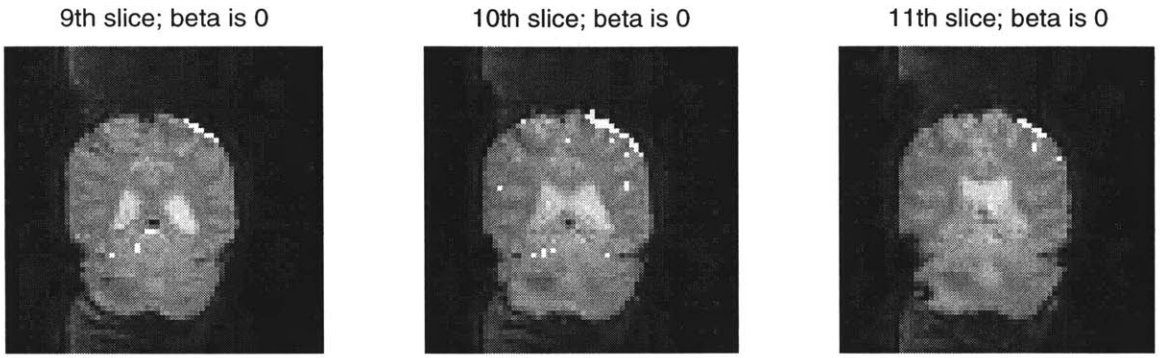
In this section, we present experimental results on three *f*MRI data sets. The protocols are designed to activate the motor cortex (via ball-squeezing protocol), auditory cortex (via word association protocol), and visual cortex (via visual stimulation with alternating checkerboard pattern), respectively. Each data set contains 60 whole brain acquisitions taken three seconds apart. Figure 5.1, Figure 5.2 and Figure 5.3 are the result of MAP segmentation for different values of β for the motor cortex, auditory cortex, and visual cortex experiments respectively.

In the 10th slice of Figure 5.1(a), there are many isolated voxels declared to be active when β is set to 0. Most of these isolated voxels were removed in Figure 5.1(b) while retaining the main part of motor cortex by increasing β to 1. Further increasing β to 2 in Figure 5.1(c), removes more of the remaining isolated spurious responses.

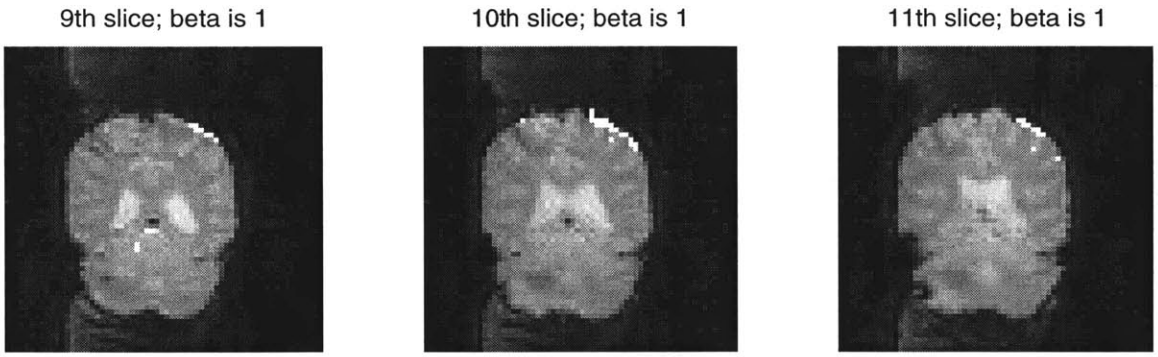
To discuss the effect of the Ising prior quantitatively, let us start by repeating (4.2).

$$\begin{aligned}\hat{Y}(\cdot,\cdot,\cdot) &= \arg\max_{y(\cdot,\cdot,\cdot)} \sum_{i,j,k} n(\hat{I}_{i,j,k}(X;U)-\gamma)y(i,j,k) - \beta \sum_{i,j,k}(y(i,j,k)\oplus y(i+1,j,k) \\ &\quad + y(i,j,k)\oplus y(i,j+1,k) + y(i,j,k)\oplus y(i,j,k+1)) \end{aligned} \tag{5.1}$$

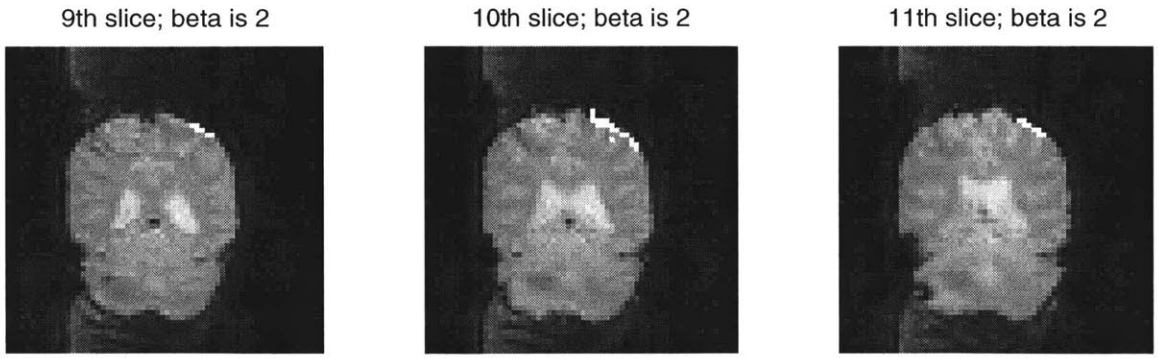
An intuitive understanding of the relationship of γ and β is obtained from (5.1) as follows. If $\beta = 0$, then there is no prior and the method reduces to MI with independent voxels. For $\beta \neq 0$ the interpretation is not as simple, but we can consider



(a) $\beta = 0$

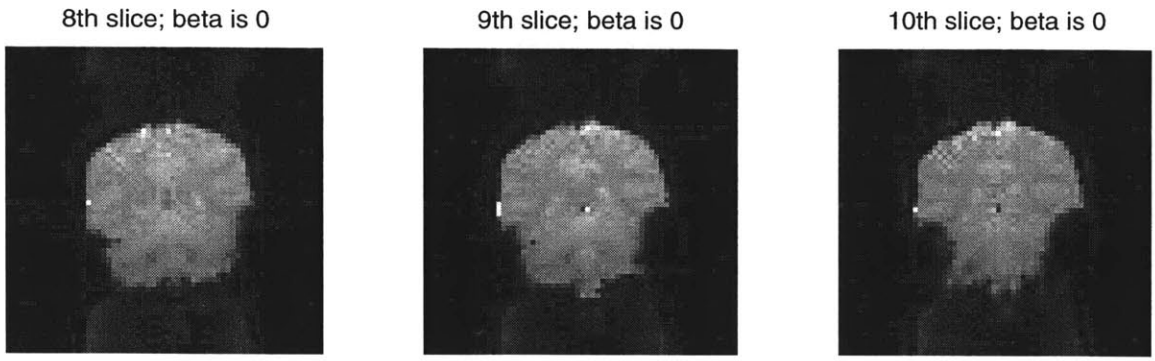


(b) $\beta = 1$

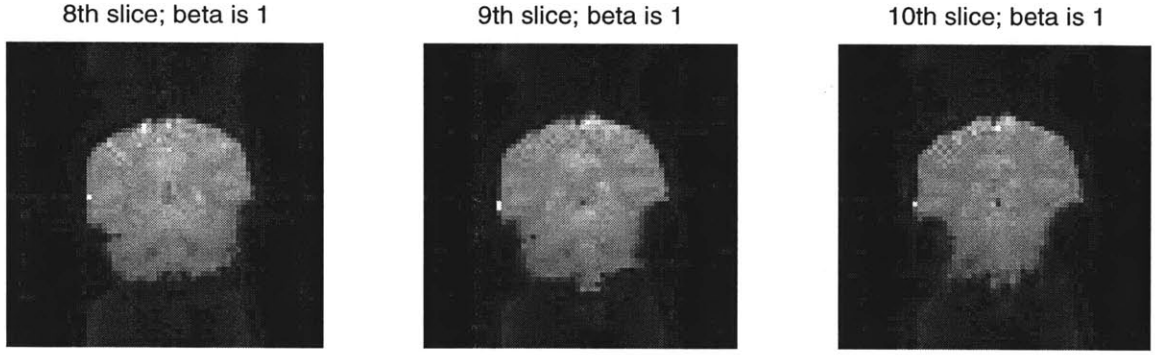


(c) $\beta = 2$

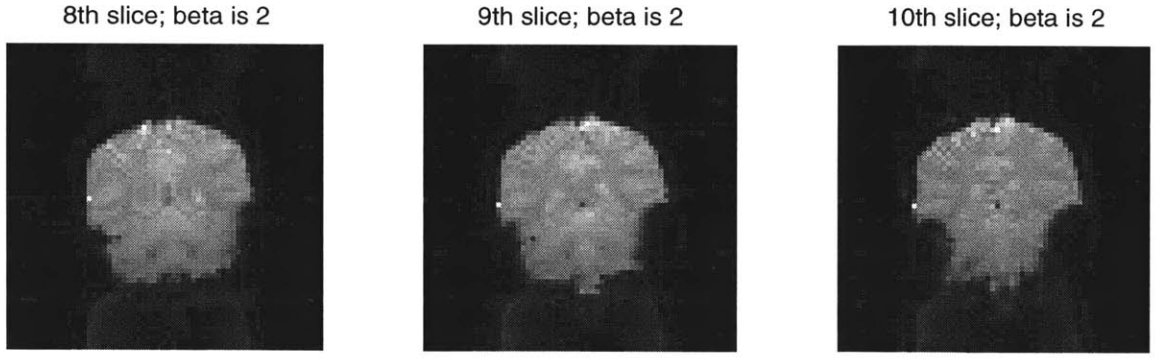
Figure 5.1: 9th, 10th, and 11th slices of the motor cortex experiments for different values of β and $\gamma = 0.6$ bit. Detections are denoted as white pixels.



(a) $\beta = 0$



(b) $\beta = 1$



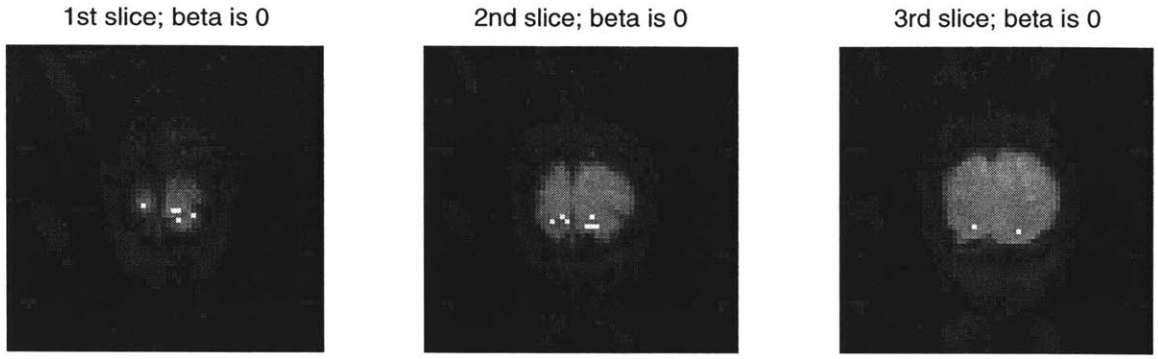
(c) $\beta = 2$

Figure 5.2: 8th, 9th, and 10th slices of the auditory cortex experiments for different values of β and $\gamma = 0.6$ bit. Detections are denoted as white pixels.

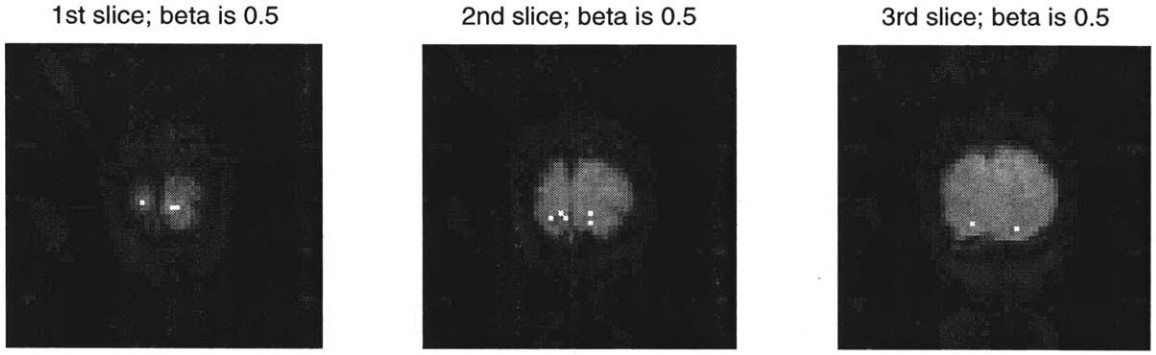
a special case. Suppose all the neighbors of a voxel are declared to be active (in our case there are six neighbors for every voxel), then the effective MI activation threshold γ for that voxel has been reduced by $6\beta/n$. By this, the Ising prior makes it more likely to fill the hole in activation map. Conversely, if all of the neighbors are inactive then the effective threshold is increased by the same amount. For these experiments, $n = 60$ and a change of β by 1 equates to a 0.1 nat (0.14 bit) change in the MI activation threshold for the special cases described. The value of β that affects the effective threshold of MI by 1 bit can be considered too large for a voxel surrounded by all active voxels or all nonactive voxels. That case corresponds to $\beta = 10 \ln 2 \approx 7$.

A good example of this idea is seen in Figure 5.2. One can see an isolated detection (white spot) at the center of the 9th slice in Figure 5.2(a) with its time series and pdf's displayed in Figure 5.4(a). The estimate of mutual information of that voxel is 0.656 bit, which is higher than the threshold $\gamma = 0.6$ bit. Since the voxel is surrounded by voxels declared non-active, increasing β by 1 acts like increasing γ to 0.74 bits. Thus the voxel is removed in Figure 5.2(b). On the other hand, the only detection of 10th slice in Figure 5.2(a) whose time series and pdf's are displayed in Figure 5.4(b) has mutual information estimate 0.845 bit. In this case, even increasing β by 2 does not remove the voxel. If all the neighbors of the voxel were declared non-activated, that change of β would correspond to increasing γ to 0.88 bit and removal of the voxel. However, the voxel is connected to the active voxel in slice 9, so in this case, the effective change of γ is $5\beta/n = 0.167$ nat $= 0.240$ bit and the voxel survives the increase of β. Therefore, we can say intuitively that changing β behaves like changing the threshold γ in a way which is adaptive to the state of its neighborhood.

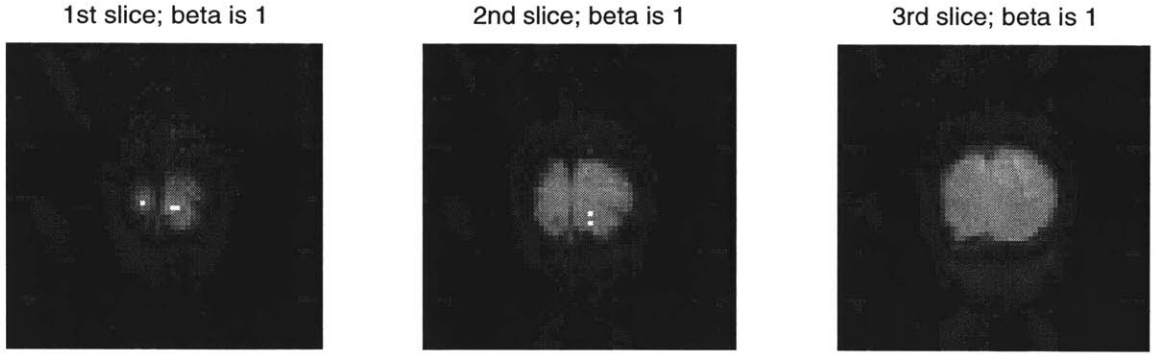
Figure 5.3 shows the effect of the Ising prior for the case of the visual cortex experiments. The two detections in the 3rd slice of Figure 5.3(b) disappear when β increases to 1 in Figure 5.3(c). Their time series and pdf's are displayed in Figure 5.5(a)



(a) $\beta = 0$



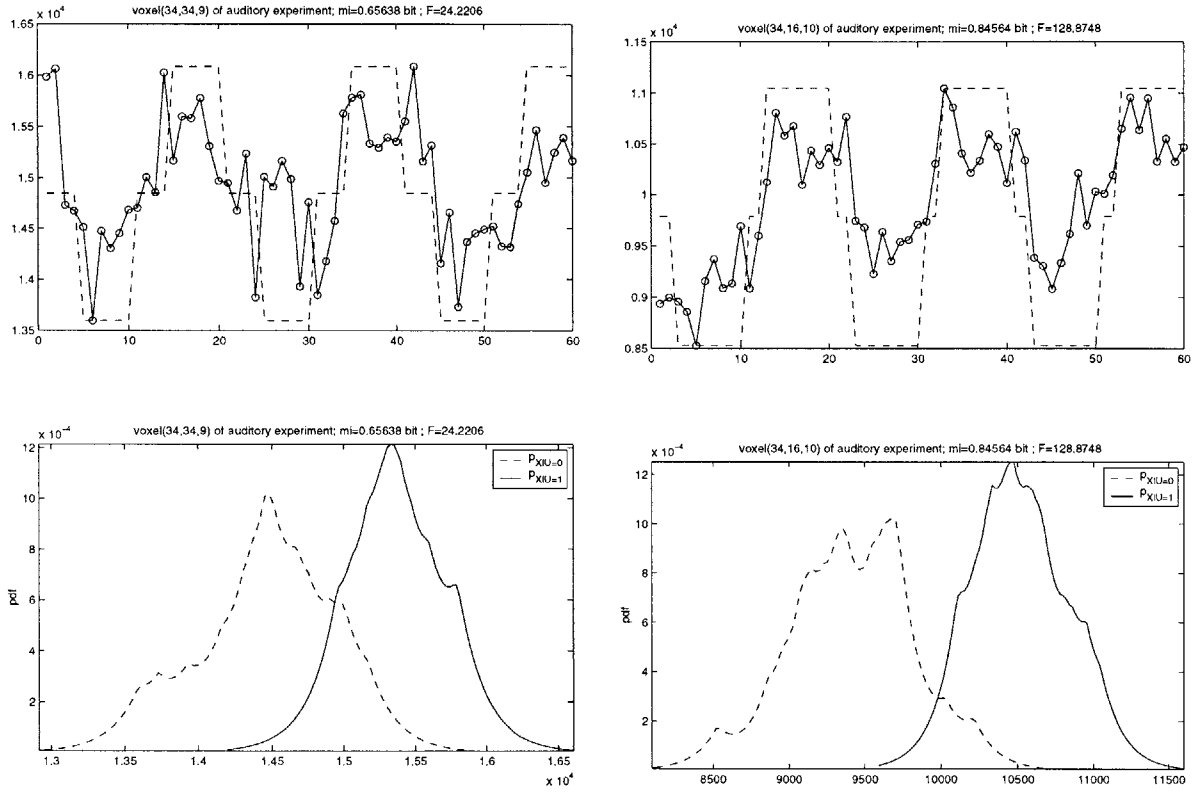
(b) $\beta = 0.5$



(c) $\beta = 1$

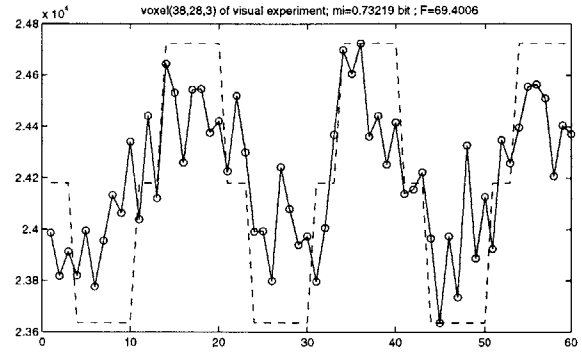
Figure 5.3: 1st, 2nd, and 3rd slices of the visual cortex experiments for different values of β and $\gamma = 0.6$ bit. Detections are denoted as white pixels.

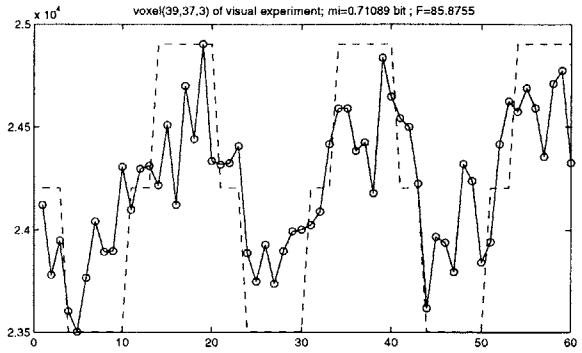
and (b). On the other hand, the isolated voxel in the 1st slice of Figure 5.3 survives the case $\beta = 1$ since its estimate of mutual information is 0.987 bit. The associated time series and pdf's are displayed in Figure 5.5(c).

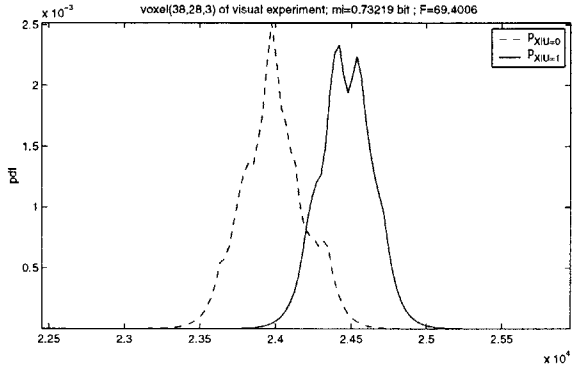


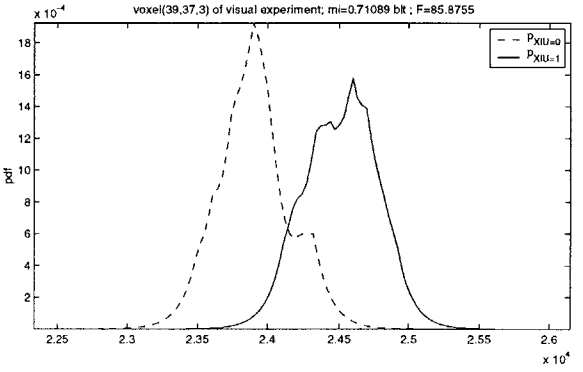
(a) the isolated voxel in the 9th slice (b) the active voxel in the 10th slice

Figure 5.4: Time series and pdf's of the voxels of interest; The auditory cortex experiments

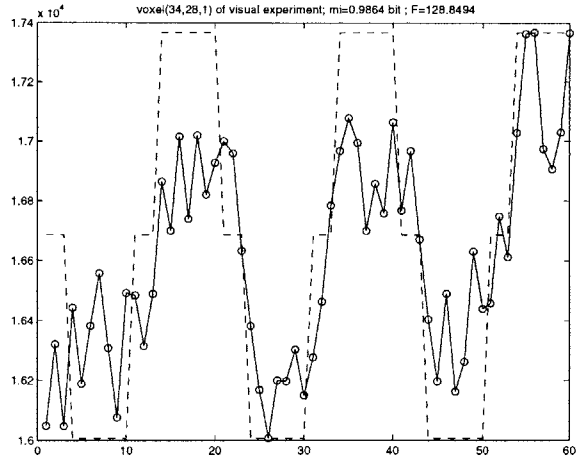


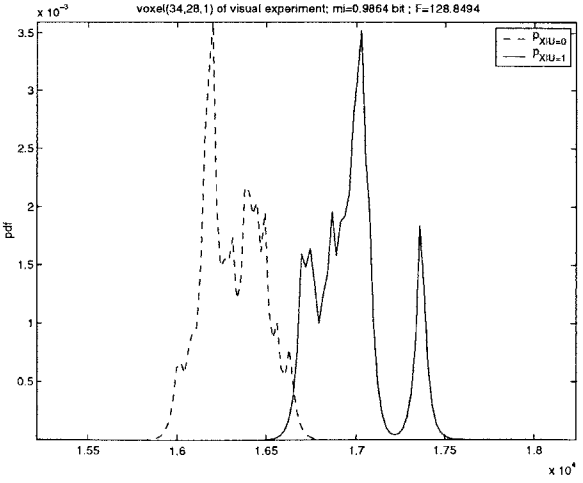






(a) the voxel in the left half of the 3rd slice (b) the voxel in the right half of the 3rd slice





(c) the isolated voxel in the 1st slice

Figure 5.5: Time series of the voxels of interest; The visual cortex experiments

5.2 ROC Curve Assuming Ground Truth

In this section, we compare the MAP detection method with other conventional methods such as the GLM and Kolmogorov-Smirnov test. The problem here is that the ground truth is unknown. One possible comparison is to observe ROC curves assuming that one method gives the truth. In the absence of known ground truth, this is an indirect way of gauging the relative modeling capacities of these methods. Note that these ROC curves do not necessarily indicate absolute performance.

Figure 5.6 shows the activation maps assumed as truth. Note that the isolated voxel in the top left of the 10th slice, the so called sympathetic response, is expected to be active in motor cortex experiments and is included in the assumed truth. Figure 5.7 shows the ROC curves that compare the GLM, MI, KS, and MI &MRF methods for the case of motor cortex experiments. The assumed truth used to generate Figure 5.7 is determined as follows. We start with the activation map obtained by the MAP method (MI & MRF) with $\beta = 1$ and $\gamma = 0.6$ bit. Then we chose the threshold of the GLM such that the number of activated voxels is same as that of the MAP result. The same thing is done for the MI case. These three activation maps, obtained from GLM, MI and MI & MRF are used as the assumed truth in Figure 5.7(a), Figure 5.7(b) and Figure 5.7(c) respectively.

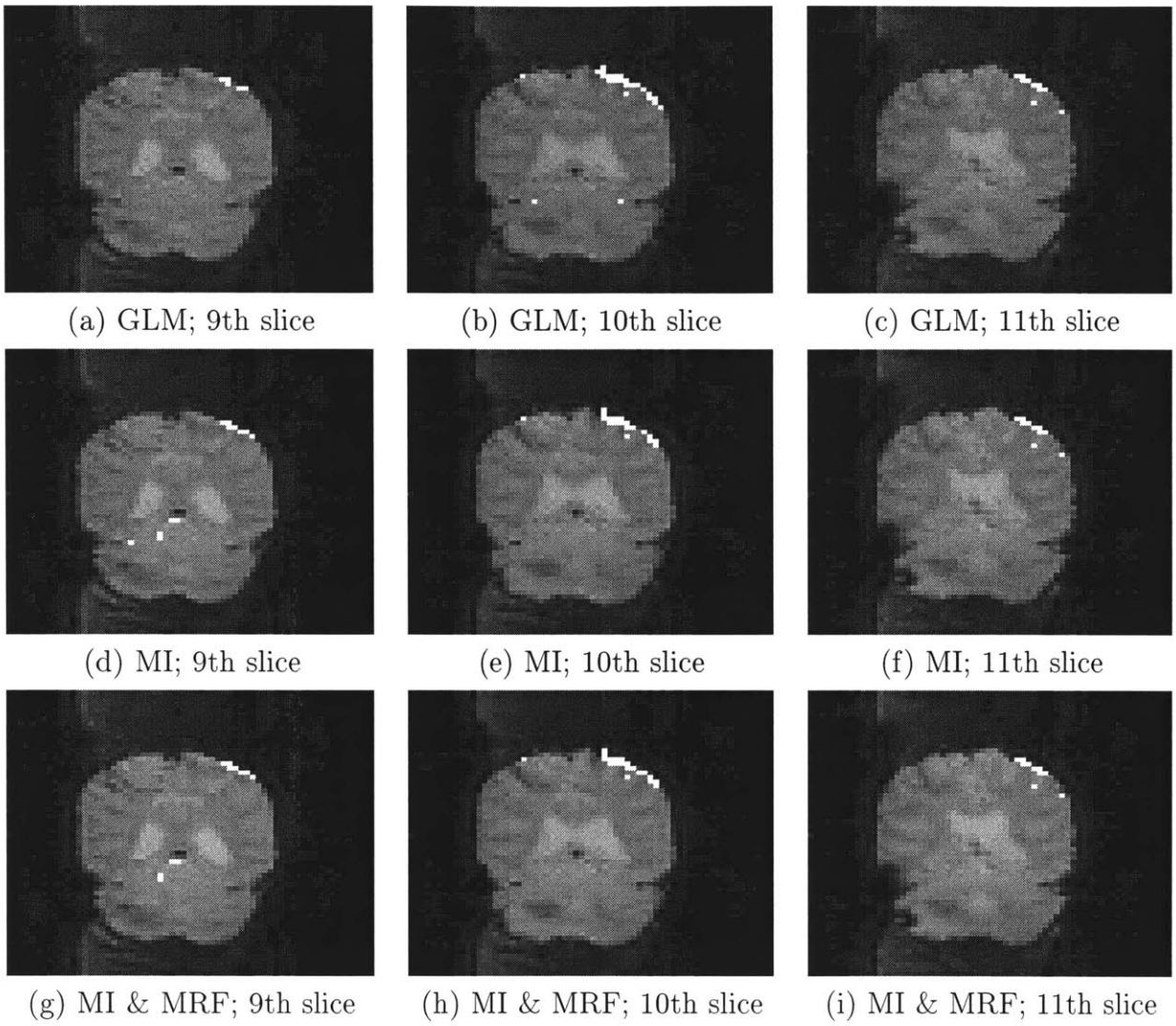
Let us describe how the ROC curve is generated. As an example, see Figure 5.7(a), which is the ROC curve for MI when GLM is assumed as the truth. Each choice of the threshold of MI gives a corresponding activation map, from which P_F and P_D are calculated relative to the assumed truth from the the GLM. If MI detects a voxel that is not detected by GLM, it is regarded as a false alarm. Similarly, if MI detects a voxel that is detected by GLM, it is regarded as a detection. In this way, the

probabilities of false alarm and detection are calculated for each choice of threshold of MI. Figure 5.7(b) and Figure 5.7(c) are generated in the same way. Note that the ROC curve for MI & MRF contains a small number of (P_D, P_F) pairs, which are generated from each pair of (β, γ) from the set $\{(\beta, \gamma) | \beta = 0, 0.5, 1, 1.5, 2$ and $\gamma = 0.5, 0.6, 0.7$ bit$\}$.

Now let us discuss the results. Figure 5.7(a) suggests that if the GLM gives the true activation map, MI performs better than KS and MI & MRF works as well as MI. This is empirical evidence that MI is more robust than KS, which is also supported by the result of Chapter 3. Figure 5.7(b) shows that the KS test is better than the GLM if MI gives the true activation map. This is the consequence of the similarity between MI and KS in that KS test solves the same hypothesis testing problem as MI. As expected, MI & MRF is close to the ideal ROC curve. Figure 5.7(c), where MI & MRF is taken as truth, can be understood in the same way as Figure 5.7(b). Finally, Figure 5.7(d) overlays the ROC for MI in Figure 5.7(a) and the ROC for GLM in Figure 5.7(b). This shows that the MI method reliably captures what the GLM considers activated, but that the GLM does not reliably capture what the MI considers activated. This also suggests the existence of phenomena that are not modeled simply by the GLM basis functions.[1] Perhaps additional bases can be chosen to correct this; however, this presumes that they are known in advance. Despite the simple basis, this demonstrates to some degree the broader modeling capacity of the nonparametric approach. Furthermore, this also demonstrates that the MI approach can uncover "new" phenomenology, which might later inspire other bases.

It is difficult to repeat the same analysis for the visual and auditory cortex experiments due to the nature of the protocols. In particular, we expect very few

[1] In the case of the GLM, we used a simple design matrix where the basis for the subspace of interest is a square wave like the protocol signal and the basis for a nuisance subspace is a DC wave.



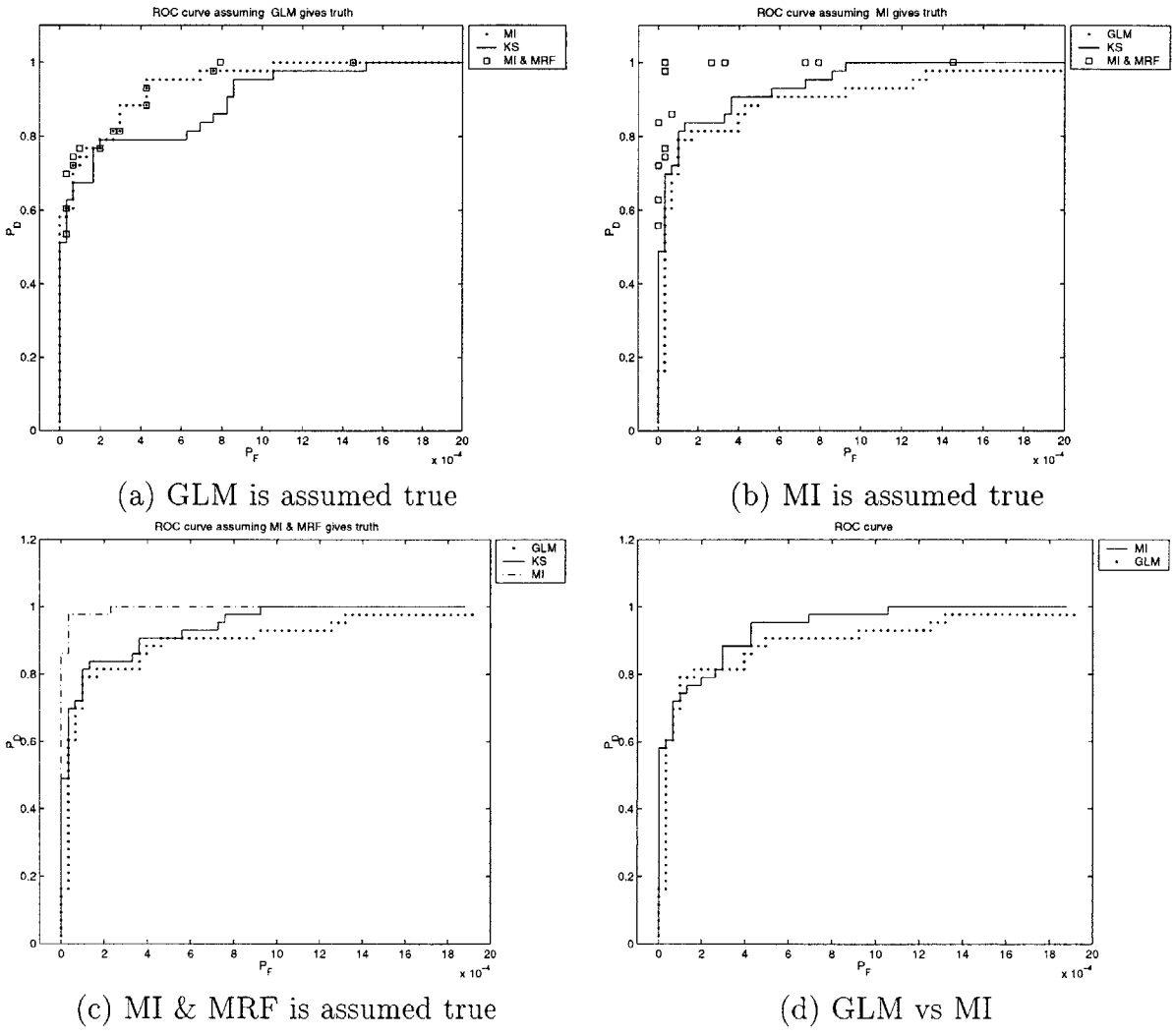
(a) GLM; 9th slice (b) GLM; 10th slice (c) GLM; 11th slice

(d) MI; 9th slice (e) MI; 10th slice (f) MI; 11th slice

(g) MI & MRF; 9th slice (h) MI & MRF; 10th slice (i) MI & MRF; 11th slice

Figure 5.6: The assumed truth obtained from GLM, MI, and MI & MRF

activations, so P_D is hard to gauge.



(a) GLM is assumed true

(b) MI is assumed true

(c) MI & MRF is assumed true

(d) GLM vs MI

Figure 5.7: Comparison of GLM, MI, KS, and MI & MRF via ROC curves with an assumed truth in motor cortex experiments

P value	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
F statistic	2.79	7.09	12.02	17.46	23.42	29.92	36.98	44.65	52.97	61.98

Table 5.1: P value and corresponding F statistic with degree of freedom (1,58)

5.3 Using Dilated Activation Maps as Assumed Truth

Another way of constructing an artificial truth is to construct an activation map by setting a very high threshold for GLM or MI and to perform a dilation operation for the activation map. This has both a biological and statistical motivation. First, dilation is motivated by prior knowledge from brain studies of the local nature of some cognitive functions in the human brain [15], particularly cortical activity. Second, we use imperfect statistical tests but high thresholds have high confidence. Thus we can get approximate ground truth by combining these two ideas. We do this because in some sense this approximates model mismatch. However, this is only an approximation and does not suffice for actual ground truth, which is in general too difficult to obtain.

Two kinds of dilation are used in this experiment. The first one dilates the activation map by including 6 nearest neighbors of all active voxels[2]. The second one dilates the activation map by including 27 neighbors within a cube whose center is an active voxel. Figure 5.8 shows ROC curves when the assumed truth is dilated with 6 neighbors and Figure 5.12 shows the case of dilation with 27 neighbors.

The threshold in Figure 5.8 was chosen such that the number of activated voxels is one third of that in the case of Figure 5.7 (1/3 of 43 voxels $\approx$ 14 voxels). This

[2]This of course closely matches the Ising model, so it will not be a surprise if MI & MRF performs well.

corresponds to an MI threshold of 0.96 bit and an F threshold of 141.3 for GLM[3]. Thus the dilation operation constructs a main localized region of the activation. Figure 5.9 shows the activation map from GLM before the dilation with 6 neighbors and the resulting "ground truth" after the dilation. Due to using a very high threshold before dilation, the sympathetic area is not in the assumed ground truth in this case.

Figure 5.8(a) compares the ROC curves of MI, KS, GLM and MI & MRF assuming the dilated activation map obtained from GLM is truth. In the low probability of false alarm regime, GLM is better than MI while MI is better than GLM when $P_F > 2 \cdot 10^{-3}$. MI looks better than KS in general. The most interesting point is that MI & MRF outperforms the GLM even though the truth was constructed from GLM. Though dilation has a role in making MI & MRF look better, this result is evidence of the viability of MI & MRF considering that GLM is assumed true. In addition, as will be shown, this improvement carries over to the 27-neighbor dilation as well.

Let us further discuss Figure 5.8(a) comparing the point $(P_F, P_D) = (0.0011, 0.5147)$ from MI & MRF and the point $(P_F, P_D) = (0.0011, 0.4853)$ from GLM. These two points demonstrate that MI & MRF performs better than GLM at $P_F = 0.0011$. The point $(P_F, P_D) = (0.0011, 0.5147)$ of MI & MRF corresponds to the pair $(\beta, \gamma) = (1, 0.5 \text{ bit})$ and the point of GLM corresponds to an F-threshold= 62.37. There are 5 voxels which are detected correctly with respect to the assumed truth by MI & MRF with the (β, γ) pair but not detected by GLM with the F-threshold. Figure 5.11 shows the temporal response of those voxels with protocol signal overlaid and there is clear structure related to the protocol signal suggesting that those voxels are active.

In Figure 5.8(b) where a dilated MI map is assumed as truth, GLM and MI

[3]This is very high considering that its P value is below 10^{-10}. See Table 5.1.

are equally good and better than KS. Not surprisingly, MI & MRF works best in this setting. Figure 5.8(c) is an overlay of the ROC curve of MI in Figure 5.8(a) and that of GLM in Figure 5.8(b). The overall impression is that MI is better than GLM. This may be more of an indication that for strongly active voxels GLM is a good model.

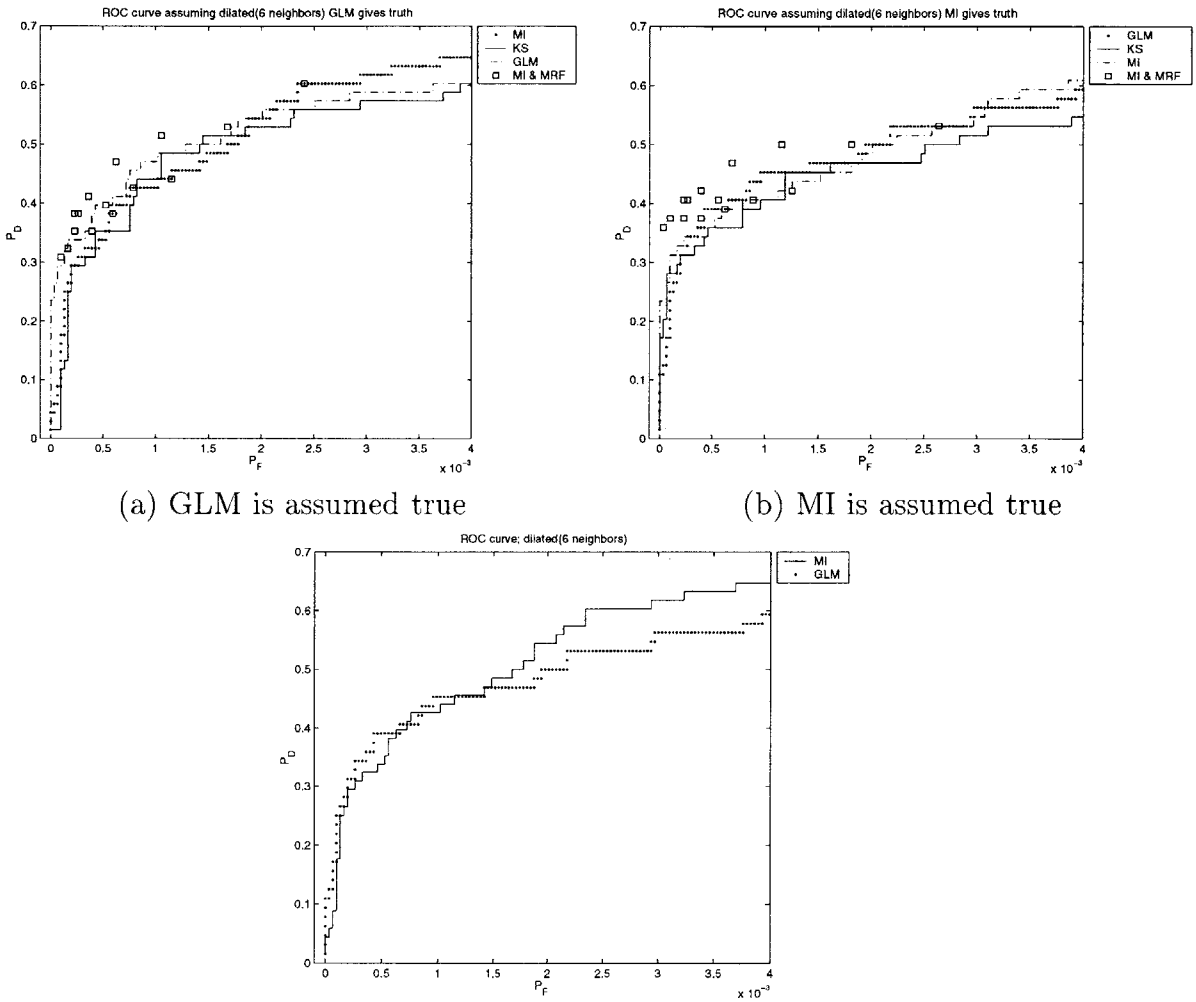
Figure 5.10 shows the method used to construct assumed truth via dilation with 27 neighbors, where the threshold is chosen such that the number of activated voxels are one tenth of that in the case of Figure 5.7 (1/10 of 43 voxels $\approx$ 4 voxels). This corresponds to an MI threshold of 1 bit and an F threshold of 217.8 for GLM. Again, the sympathetic area is not included in the assumed truth.

The resulting ROC curves are given in Figure 5.12. In the case of Figure 5.12(a), performance is in the order of MI & MRF, GLM, MI and KS. Again, MI & MRF still looks the best even though the truth was constructed from GLM. In Figure 5.12(b), performance is in the order of MI & MRF, MI, KS and GLM. The overlaid version Figure 5.12(c) also shows that MI is more robust than GLM. In other words, MI has a higher modeling capacity than GLM.

5.4 Comparison with GLM

This section presents another way to compare the activation map computed by three methods: GLM, nonparametric MI, nonparametric MI with an Ising prior.

We first apply the GLM method to each data set. The coronal slice exhibiting the highest activation for each data set is shown in the first column of Figure 5.13 with the GLM activation map overlaid in white for each data set. The F-statistic threshold

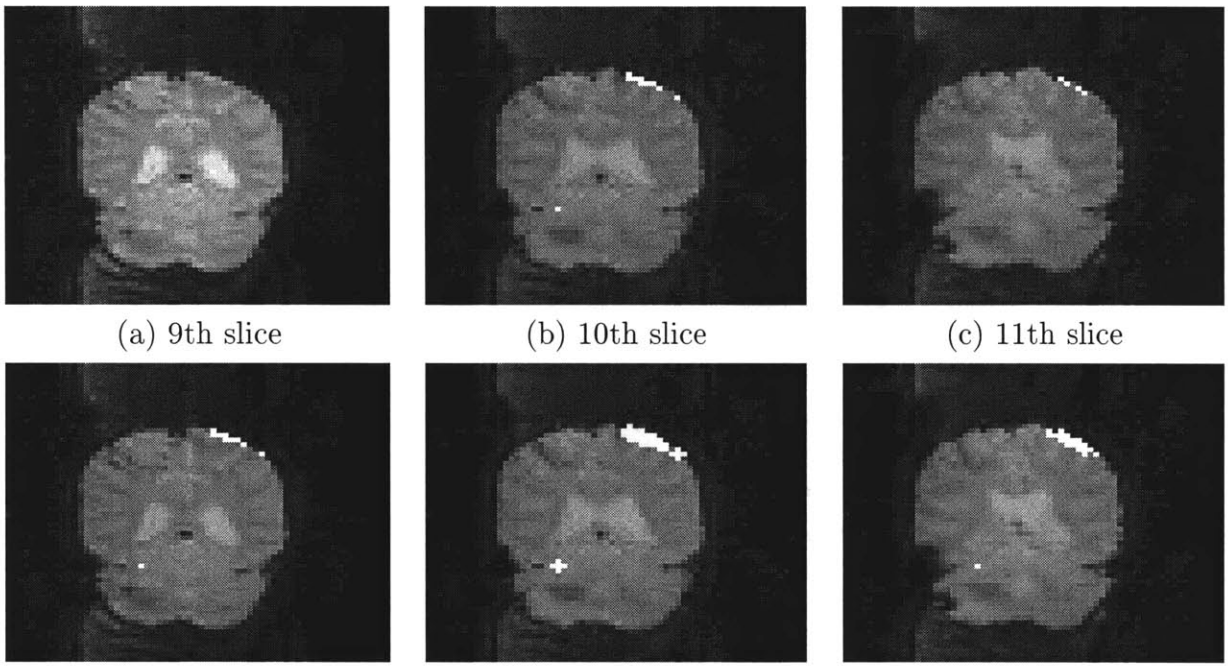


(a) GLM is assumed true

(b) MI is assumed true

(c) GLM vs MI

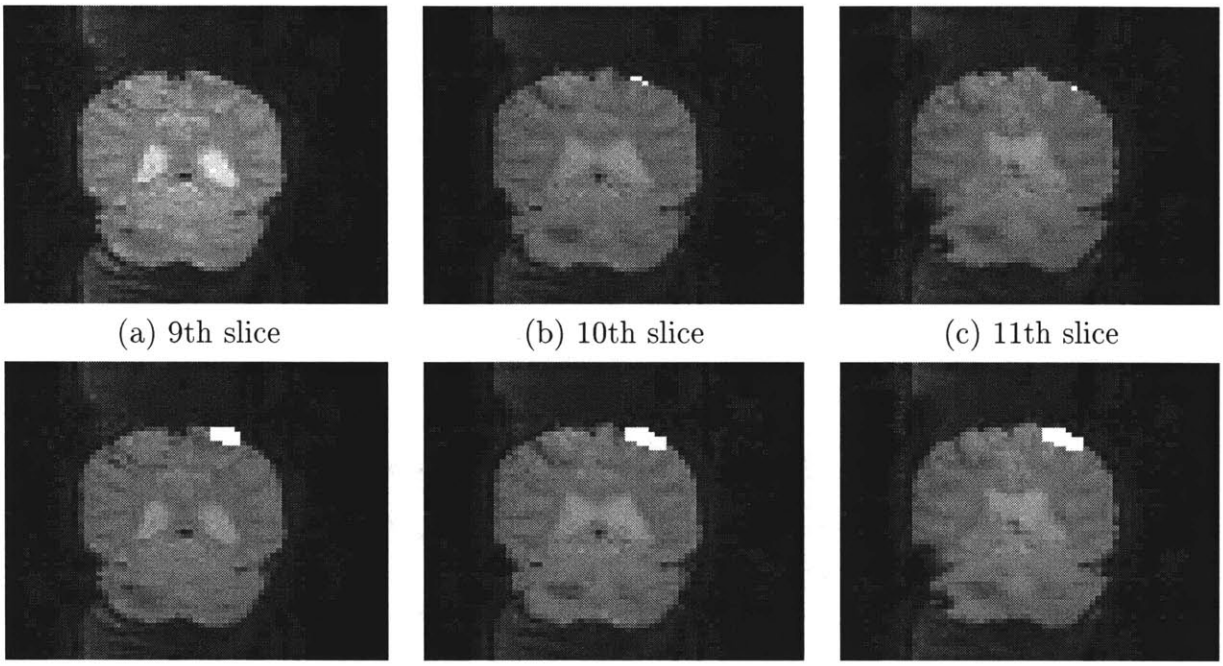
Figure 5.8: Comparison of GLM, MI, KS, and MI & MRF via ROC curves with an assumed truth in motor cortex experiments; dilation with 6 neighbors



(a) 9th slice (b) 10th slice (c) 11th slice

(d) 9th slice (e) 10th slice (f) 11th slice

Figure 5.9: The activation map from GLM; (a), (b), and (c) are before dilation; (d), (e), and (f) are after dilation with 6 neighbors



(a) 9th slice (b) 10th slice (c) 11th slice

(d) 9th slice (e) 10th slice (f) 11th slice

Figure 5.10: The activation map from GLM; (a), (b), and (c) are before dilation; (d), (e), and (f) are after dilation with 27 neighbors

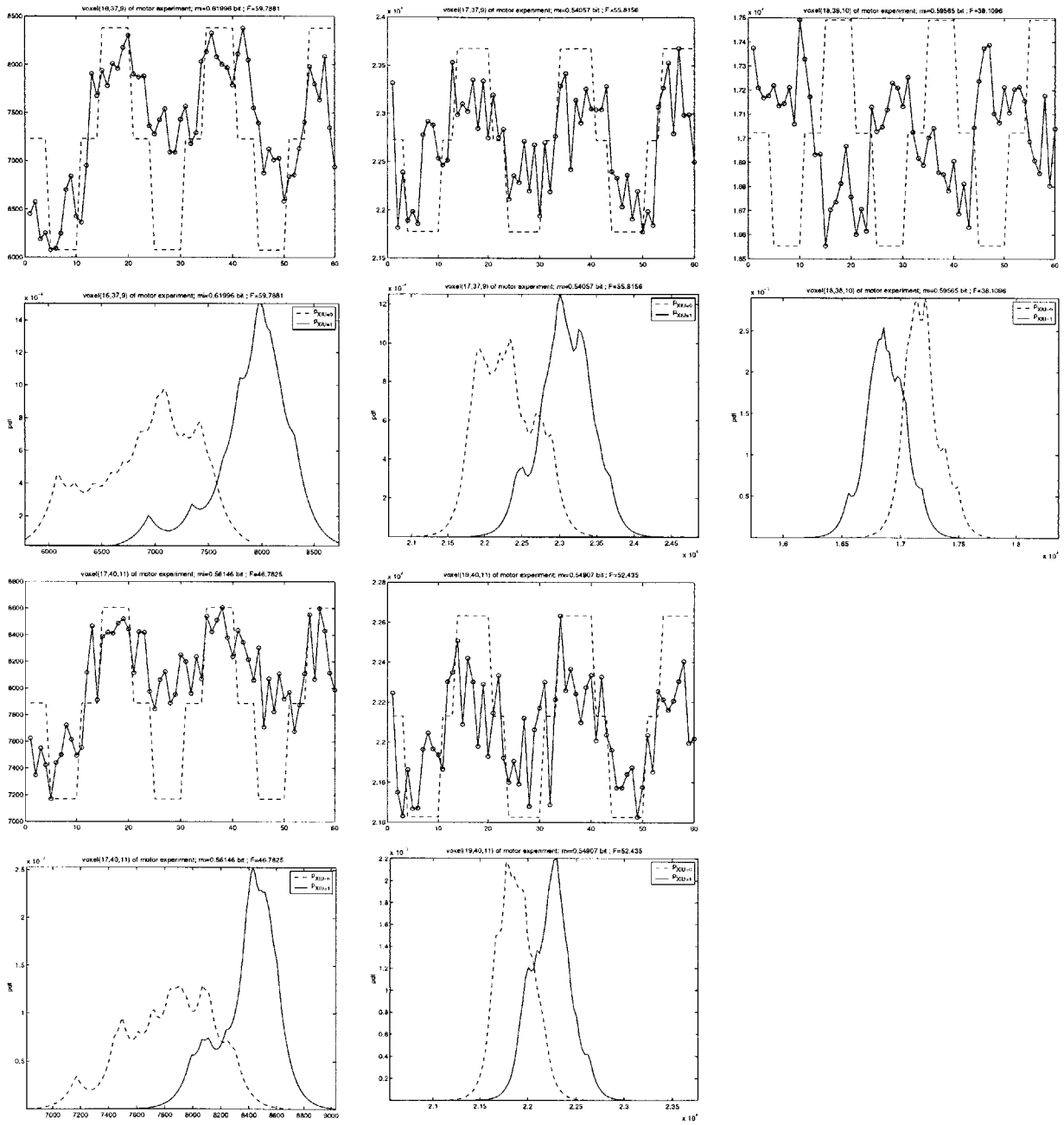
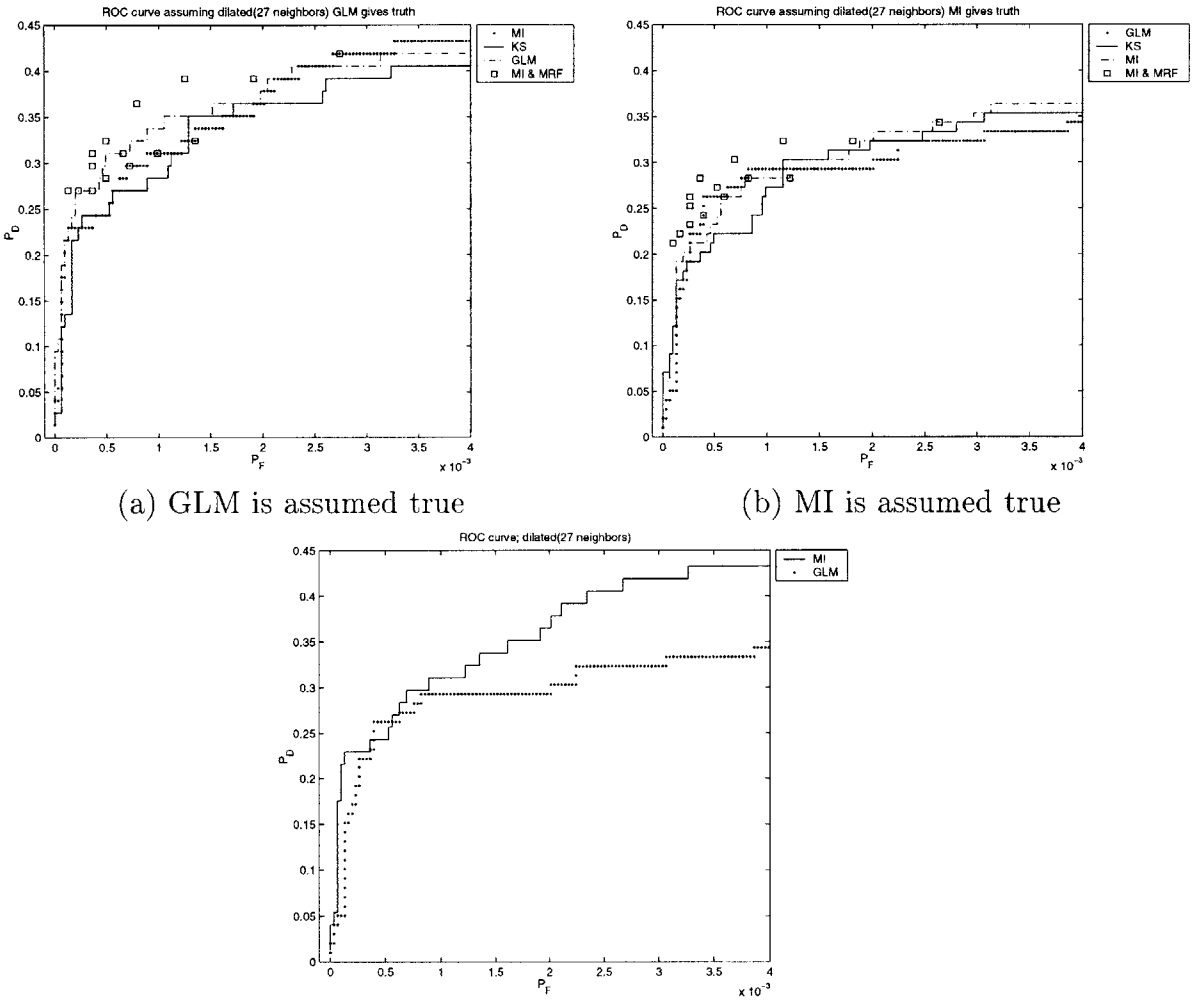


Figure 5.11: Temporal responses of voxels detected by MI & MRF

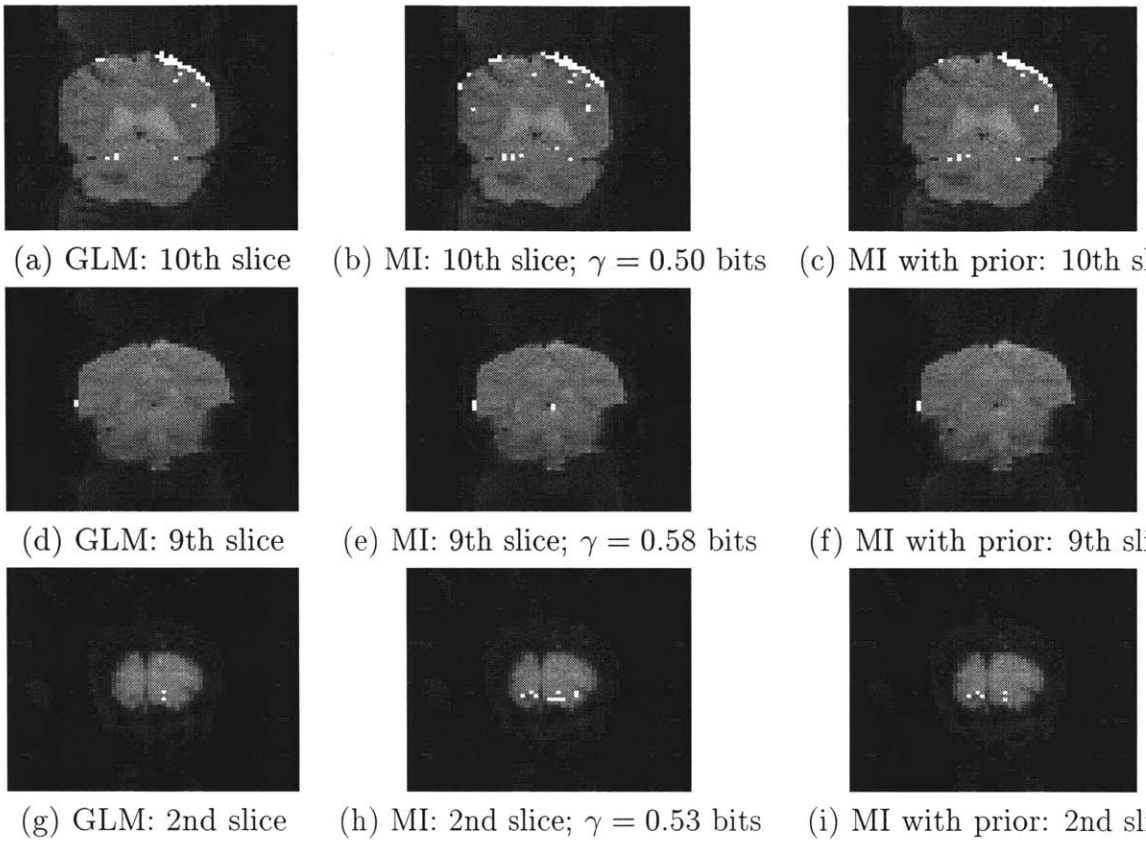


(a) GLM is assumed true

(b) MI is assumed true

(c) GLM vs MI

Figure 5.12: Comparison of GLM, MI, KS, and MI & MRF via ROC curves with an assumed truth in the motor cortex experiments; dilation with 27 neighbors



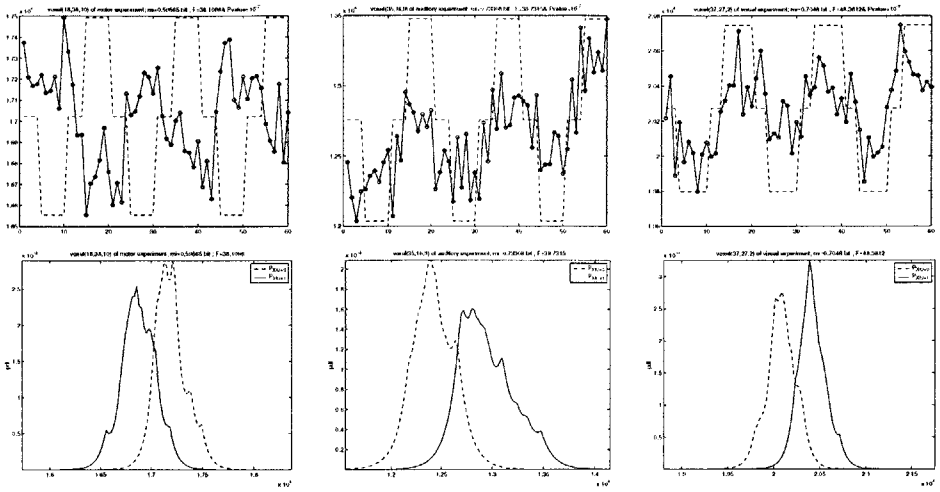
(a) GLM: 10th slice (b) MI: 10th slice; $\gamma = 0.50$ bits (c) MI with prior: 10th slice

(d) GLM: 9th slice (e) MI: 9th slice; $\gamma = 0.58$ bits (f) MI with prior: 9th slice

(g) GLM: 2nd slice (h) MI: 2nd slice; $\gamma = 0.53$ bits (i) MI with prior: 2nd slice

Figure 5.13: Comparison of *f*MRI analysis results from motor, auditory and visual experiments

for GLM is set such that the visual inspection of the activation map is consistent with our prior expectation for the number of activated voxels which corresponds to a p-value of 10^{-10}. In the next column of the figure, the same slices are shown using MI to compute the activation map. In this case, the MI threshold γ was set such that all of the voxels detected by the GLM were detected by MI. Consequently, Figures 5.13 (b), (e) and (h) contain additional activations when compared to GLM. Some of these additional activations are spurious and some are not. Finally, the Ising prior is applied to the MI activation map with $\beta = 1$. An intuitive argument on the relationship of γ and β was given previously.



(a) motor (b) auditory (c) visual

Figure 5.14: Temporal responses of voxels newly detected by the MI with the Ising prior method

Comparison of Figures 5.13 (b), (e) and (h) to Figures 5.13 (c), (f), and (i) shows that many of the isolated activations are removed by the Ising prior, but some of the new MI activations remain. Figure 5.14 shows the temporal responses of the voxels with the lowest GLM score which are detected by MI with an MRF prior but not by GLM. Examination of these temporal responses (with protocol signal overlaid) reveals obvious structure related to the protocol signal.

A reasonable question is whether this result is due to an unusually high threshold set for GLM. In order to address this, we next lower the GLM threshold such that the voxels of Figure 5.14 are detected by GLM. We then consider regions of the resulting activation map where new activations have appeared in Figure 5.15. The activations of Figure 5.15(a) and Figure 5.15(b) (motor cortex experiments and auditory cortex experiments), would be considered spurious in light of the region in which they occur. The result for Figure 5.15(c) is not so clear as these activations are most likely spurious, but might possibly be related to higher-ordered visual processing.

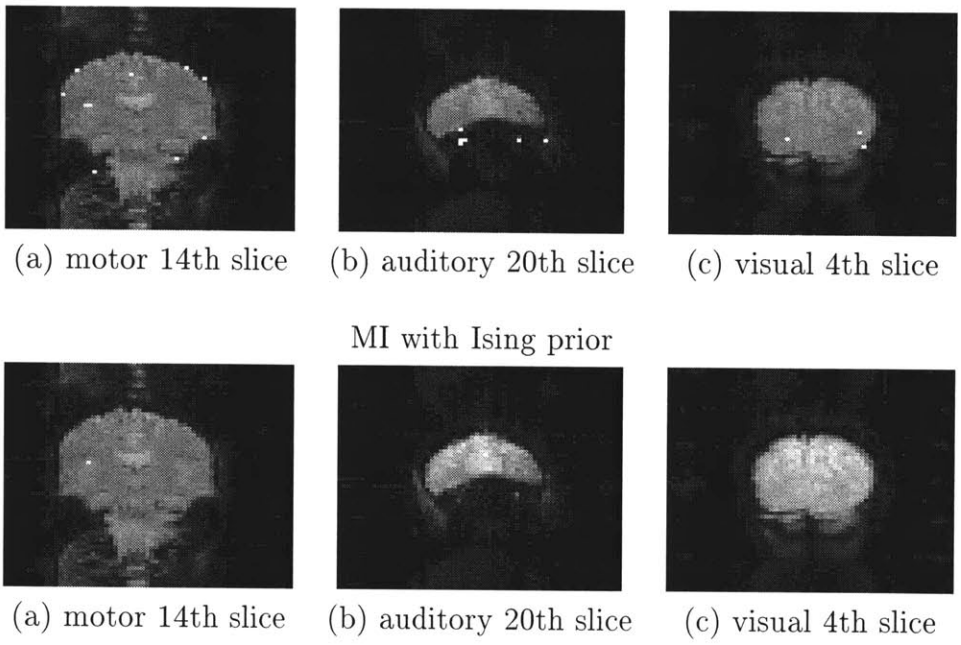


Figure 5.15: Comparison of *f*MRI Analysis results from motor, auditory and visual experiments with lowered GLM threshold

Chapter 6

Conclusions

In Section 6.1 of this chapter, we briefly summarize the major part of the work demonstrated in Chapter 3, Chapter 4, and Chapter 5 mentioning the contributions of this thesis. We then summarize the remaining issues that the thesis does not address and suggest possible extensions of this work.

6.1 Brief Summary

In this thesis, we develop an *f*MRI signal analysis algorithm that detects activated voxels due to specific experimental stimuli and simultaneously removes spurious isolated responses. Specifically, we combine a previously developed information theoretic approach [2] with an MRF prior within a Bayesian framework and used Greig's [4] efficient algorithm to solve the MAP problem *exactly.*

6.1.1 Nonparametric Hypothesis Testing

The mutual information between the *f*MRI signal and the protocol signal suggests a hypothesis testing problem which tests if two random variables are independent. Under this hypothesis space, we derive a relationship between the likelihood ratio and the mutual information, and show that the test using an estimate of mutual information is close to likelihood ratio test, which is optimal by Neyman-Pearson lemma. This theoretical interpretation of the use of MI by Tsai *et al* [2] allows the extension to the Bayesian framework.

Using the nonparametric estimate of mutual information, or equivalently Kullback-Leibler divergence, is an alternative solution for the hypothesis testing problem which tests whether two sets of data are drawn from same distribution, for which Kolmogorov-Smirnov test is conventionally used. Information theoretic quantities like entropy and mutual information are used in this way for various hypothesis testing problems.

6.1.2 Applying MRF to *f*MRI Analysis

Using the MRF prior in *f*MRI analysis is not new as Descombes *et al* [3] propose using it for *f*MRI signal restoration and detection of active voxels. However, in that approach, the data attachment term, which is also modeled as an MRF, is heuristic in that it is not a rigorous likelihood ratio. Consequently, the formulation of the hypothesis testing problem is not rigorous within the Bayesian framework.

In contrast, applying an MRF prior to the information theoretic method in this thesis is rather straightforward due to the result of Chapter 3. Furthermore, the

resulting MAP problem meets the conditions for using Greig's method [4] making the *exact* solution of the MAP problem possible in polynomial time.

6.1.3 Experimental results

We present the activation map, which is the exact solution of the MAP estimation problem with the Ising prior. The experimental results show that this method can remove isolated responses effectively within a rigorous framework. This has an advantage over using morphological operator in that it considers the evidence in the data as well as the spatial dependency.

We also compare our approach with the GLM method. While *f*MRI analysis of patient data is always faced with the difficulty that exact truth is unknown, our results indicate that the MI approach with spatial priors is able to detect "true" activations with a significantly smaller number of spurious responses. However, more validation is necessary.

6.2 Extensions

This thesis proposes and reports on the use of MI and a MRF prior for analyzing *f*MRI data. The work here focuses on the development of the theoretical foundation of this analysis, and we have made several simplifying assumptions in the process. In the following subsections, we briefly discuss how some of these restrictions can be relaxed as avenues of further research.

More Complex Graph Structure in Prior model

As stated in Chapter 4, the method of applying the MRF to *f*MRI analysis can be extended to any prior model allowing graph structures more complex than the Ising model as long as the prior is a binary Gibbs field whose energy is composed of only doubleton potentials. Taking advantage of this, one may extend this model to take advantage of other priors that incorporate existing knowledge of the anatomical structure of the brain. For example, it is generally known that the auditory cortex is related to the temporal lobe though these two lie in disjoint regions. We expect that an appropriate graphical model can take such structure into account.

Issues on Nonparametric Estimation of Entropy

For sample sets of limited size, such as in the case for *f*MRI data, the nonparametric estimate of entropy is significantly affected by the choice of kernel parameters. The area of kernel width selection is a broad and active branch of research in nonparametric statistics whose results can be applied here. While the kernel shape also, impacts estimator performance, its impact is significantly less than kernel width.

On the Structure of *f*MRI Temporal Response

We assume that $S_{X|U=0}$ and $S_{X|U=1}$ are i.i.d. in Assumption 3.1.1. This assumption is useful in estimating entropy and mutual information. It is natural to consider the impact of relaxing this assumption. The *f*MRI temporal response then is modeled as a stochastic process and as a consequence, the entropy rate of a stochastic process

arises instead of the entropy of a random variable. An example of nonparametric treatment of stochastic processes can be found in [16].

Appendix A

χ^2 and F Distribution

The definitions of χ^2 and F distribution are quoted from [17].

Definition A.0.1 (Central χ^2). *The distribution of the sum $Y = \sum_{n=1}^{N} X_n^2$ is central χ_N^2 when the X_n are i.i.d. N(0,1) random variables. The density for y is*

$$p_Y(y) = \frac{1}{\Gamma(N/2)2^{N/2}} y^{(N/2)-1} e^{-y/2}; y \geq 0$$

We say that Y is a central χ^2 random variable with N degrees of freedom.

Definition A.0.2 (Noncentral χ^2). *When the independent random variables X_n in the sum $Y = \sum_{n=1}^{N} X_n^2$ are distributed according to $N(\mu_n, 1)$, then the distribution of Y is noncentral χ^2 with noncentrality parameter $d^2 = \sum_{n=1}^{N} \mu_n^2$.*

Definition A.0.3 (Central F). *Let $Y : \chi_p^2$ and $Z : \chi_{N-p}$ denote independent central χ^2 random variables with respective degrees of freedom p and $N - p$. The ratio*

$$F = \frac{\frac{Y}{p}}{\frac{Z}{N-p}}$$

is called an F-statistic, and the distribution of F is called an F-distribution. The density function for F is

$$p_F(f) = \frac{\Gamma(N/2)[p/(N-p)]^{p/2}}{\Gamma(p/2)\Gamma[(N-p)/2]} \frac{f^{p/2-1}}{[1+(p/(N-p))f]^{N/2}}; f \geq 0.$$

Definition A.0.4 (Noncentral F). *When $Y : \chi^2_p$ is replaced by noncentral χ^2 random variable $Y : \chi^2_p(d^2)$ in the definition of central F distribution, the distribution of the ratio $F = \frac{\frac{Y}{p}}{\frac{Z}{N-p}}$ is called noncentral F-distribution.*

$$p_F(f) = \sum_{n=0}^{\infty} e^{-d^2/2} \frac{(d^2/2)^n}{n!} \frac{\Gamma(N/2+n)}{\Gamma(p/2+n)\Gamma[(N-p)/2]} (\frac{p}{N-p})^{p/2+n} \frac{f^{p/2-1+n}}{[1+(p/(N-p))f]^{N/2+n}}; f \geq 0.$$

Appendix B

Finding Maximum Likelihood Kernel Size

In this appendix, we present the mathematical specifics necessary in finding the ML kernel size for the cases of a Gaussian and double-exponential kernel.

B.1 Gaussian Kernel

$$k(x,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{B.1}$$

$$\frac{\partial}{\partial\sigma} k(x,\sigma) = \frac{1}{\sigma^3} k(x,\sigma)(x^2-\sigma^2) = \frac{1}{\sigma} k(x,\sigma)(\frac{x^2}{\sigma^2}-1) \tag{B.2}$$

The likelihood of the data in terms of density estimate is

$$L = \prod_i \hat{p}(X_i) = \prod_i \left(\frac{1}{n-1}\sum_{j\neq i} k(x_i - x_j, \sigma)\right) \tag{B.3}$$

$$\log L = \sum_i \log\left(\frac{1}{n-1}\sum_{j\neq i} k(x_i - x_j, \sigma)\right). \tag{B.4}$$

Then the score function with respect to the kernel size is

$$S = \frac{\partial \log L}{\partial \sigma} = \sum_i \left[\frac{1}{\hat{p}(x_i)}\frac{1}{n-1}\sum_{j\neq i}\frac{\partial}{\partial\sigma}k(x_i - x_j, \sigma)\right] \tag{B.5}$$

$$= \sum_i \left[\frac{1}{\hat{p}(x_i)}\frac{1}{n-1}\sum_{j\neq i}\frac{1}{\sigma}k(x_i - x_j, \sigma)\left(\frac{(x_i - x_j)^2}{\sigma^2} - 1\right)\right]. \tag{B.6}$$

Calculation of the score function requires $O(n^2)$ time. Using this score function and the root search method, the ML kernel size can be found as a root of $S = 0$ if the score function is concave[1].

B.2 Double Exponential Kernel

$$k(x, \sigma) = \frac{1}{2\sigma}e^{-\frac{|x|}{\sigma}} \tag{B.7}$$

$$\frac{\partial}{\partial\sigma}k(x, \sigma) = \frac{1}{\sigma^2}k(x, \sigma)(|x| - \sigma) = \frac{1}{\sigma}k(x, \sigma)\left(\frac{|x|}{\sigma} - 1\right) \tag{B.8}$$

[1]The concavity of the score function is not proved, but it is supported by empirical results.

Then the score function with respect to the kernel size is

$$S = \frac{\partial \log L}{\partial \sigma} = \sum_i \left[\frac{1}{\hat{p}(x_i)} \frac{1}{n-1} \sum_{j \neq i} \frac{\partial}{\partial \sigma} k(x_i - x_j, \sigma) \right] \tag{B.9}$$

$$= \sum_i \left[\frac{1}{\hat{p}(x_i)} \frac{1}{n-1} \sum_{j \neq i} \frac{1}{\sigma} k(x_i - x_j, \sigma) \left(\frac{|x_i - x_j|}{\sigma} - 1 \right) \right] \tag{B.10}$$

Again, calculation of the score function requires $O(n^2)$ time.

Bibliography

[1] K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapping*, vol. 2, pp. 189–210, 1995.

[2] A. Tsai, J. W. Fisher, C. Wible, W. M. Wells, J. Kim, and A. S. Willsky, "Analysis of fmri data using mutual information," in *Second International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 473–480, Sep 1999.

[3] X. Descombes, F. Kruggel, and D. Y. von Cramon, "Spatio-temporal fmri analysis using markov random fields," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 1028–1029, Dec 1998.

[4] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society. Series B(Methodological)*, vol. 51, no. 2, pp. 271–279, 1989.

[5] K. J. Friston, P. Jezzard, and R. Turner, "The analysis of functional mri time-series," *Human Brain Mapping*, vol. 1, pp. 153–171, 1994.

[6] V. Solo, E. Brown, and R. Weisskoff, "A signal processing approach to functional mri for brain mapping," in *Image Processing, 1997. Proceedings., International Conference on*, vol. 2, pp. 121–123, 1997.

[7] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2. John Wiley & Sons, second ed., 1970.

[8] E. Parzen, "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[9] P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, no. 1, pp. 69–88, 1993.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley-Interscience, 1991.

[11] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, pp. 721–741, Nov 1984.

[12] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to algorithms.* The MIT Press, 1990.

[13] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, vol. 27, pp. 832–837, Sep 1956.

[14] Y. Boykov, O. Veksler, and R. Zabih, "Energy minimization with discontinuities."

[15] D. J. Amit, *Modeling Brain Function.* Cambridge University Press, 1989.

[16] J. W. Fisher, A. T. Ihler, and P. Viola, "Learning informative statistics: A nonparametric approach," in *Advances in Neural Information Processing Systems*, (Denver, Colorado), November 1999.

[17] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis.* Addison-Wesley, 1991.