Comparing Statistically Pooled Brain Maps in FMRI studies using Parametric and Non-parametric Methods

by

Taniya Mandal

(Under the direction of Nicole Lazar)

Abstract

In a typical functional brain imaging experiment, scientists aim to map the specific areas of the brain that are activated while subjects perform a designated task (which may be cognitive, motor, or other). For comparison purposes (comparing patients vs. controls, females vs. males, as some examples), combining the brain maps from the subjects in an efficient way becomes imperative so that we can get an overall picture of activity for each group. We use statistical tests that have been developed historically for combining independent sources of information to create maps for each group of subjects in a neuroimaging study. These statistical tests follow two basic approaches - combining p-values and meta-analysis. Through these methods we aim to draw conclusions about the behavioral pattern of two or more groups with respect to each other. We also want to compare the performance of the different methods. Group comparisons have been done in the past using "group maps" for each population through a fixed effects model or a random effects model. This dissertation explores some pre-existing statistical combination techniques used in combining and interpreting "group maps". We will use parametric and non-parametric approaches to compare between two or more populations. While combining and comparing brains, there are two aspects that arise - spatial and statistical. We will only focus on the latter aspect. We will assume that the voxels of the brain are independent of each other. However, as we conduct

various tests to compare group maps at each voxel, to minimize false positives i.e. voxels declared active when they are not, we will threshold at each voxel. In this dissertation we will explore thresholding through false discovery rate, permutation tests and bootstrapping and we compare these methods to draw a conclusion about which one would be apt to use.

INDEX WORDS: group comparison, combination tests, thresholding, permutation, bootstrapping

Comparing Statistically Pooled Brain Maps in FMRI studies using Parametric and Non-parametric Methods

by

Taniya Mandal

B.S., University of Calcutta, India, 2001

M.S., University of Calcutta, India, 2003

M.S., The University of Georgia, USA, 2005

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2010

Comparing Statistically Pooled Brain Maps in FMRI studies using Parametric and Non-parametric Methods

by

Taniya Mandal

Approved:

Major Professor: Nicole Lazar

Committee: Jaxk Reeves
Jeongyoun Ahn
Lynne Seymour
Paul Schliekelman

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2010

Acknowledgments

I would like to dedicate this thesis to my beloved parents and my wonderful husband, Sudhin. This thesis would not be complete without recognizing my parents' perseverance and guidance through out my formative years and my husband's abound support, encouragement and help to keep me focused during my graduate years.

I would also like to thank Dr. Jennifer McDowell for providing me with the data and Dr. Nicole Lazar for her immense support and understanding as an advisor. I would also take this opportunity to thank my advisory committee for their help and encouragement through out all the years I have been in the Department of Statistics at University of Georgia, Athens.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to Medical Imaging

The recent discovery that magnetic resonance imaging can be used to map changes in flow of blood in the brain that correspond to mental operations extends traditional anatomical imaging to include maps of human brain function. The ability to observe both the structures and also which structures participate in specific functions is due to a new technique called functional magnetic resonance imaging, fMRI, that provides high resolution, noninvasive reports of neural activity detected by a blood oxygen level dependent signal (Ogawa, et al, 1990 a and b). This new ability to observe brain function indirectly opens an array of new opportunities to advance our understanding of brain organization, as well as a potential new standard for assessing neurological status and neurosurgical risk thereby forming an important component of much of the current research in cognitive, clinical and social psychology.

Statisticians play a key role in this research, since the data that are obtained from these studies are remarkably complex (correlated in time and in space in ways that are still not fully understood) and massive (a typical number may be hundreds of thousands of time series which again is composed of possibly hundreds of time points for a single subject, one for each "voxel" or volume element of the brain). The number of subjects on the other hand is comparatively small, a situation that creates challenges for statistical inference. A brief introduction is necessary to understand the use of statistics in this dissertation.

1.1 Science of Medical Imaging

Imaging is the representation or reproduction of an object's outward form; especially a visual representation (i.e., the formation of an image). Imaging methodologies and technologies have been used in various fields - chemical imaging to take simultaneous measurements of spectra, medical imaging to create images of the human body or parts of it to diagnose and examine a disease and so on.

Medical imaging is part of biological imaging and incorporates radiology, nuclear medicine, investigative radiological sciences, endoscopy, (medical) thermography, medical photography and microscopy (e.g. for human pathological investigations). Measurement and recording techniques which are not primarily designed to produce images, such as electroencephalography (EEG), magnetoencephalography (MEG), electrocardiography (EKG) and others, but which produce data amenable to be represented as maps (i.e. containing positional information), can also be seen as forms of medical imaging (Robb, 1999).

In the clinical context, medical imaging is generally equated to radiology or "clinical imaging" and the medical practitioner responsible for interpreting (and sometimes acquiring) the images is a radiologist. Diagnostic radiography designates the technical aspects of medical imaging and in particular the acquisition of medical images. The radiographer or radiologic technologist is usually responsible for acquiring medical images of diagnostic quality, although some radiological interventions are performed by radiologists. While radiology is an evaluation of anatomy, nuclear medicine provides functional assessment.

As a field of scientific investigation, medical imaging constitutes a sub-discipline of biomedical engineering, medical physics or medicine depending on the context: research and development in the area of instrumentation, image acquisition (e.g. radiography), modeling and quantification are usually the preserve of biomedical engineering, medical physics and computer science; research into the application and interpretation of medical images is usually the preserve of radiology and the medical sub-discipline relevant to medical condition or area of

medical science (neuroscience, cardiology, psychiatry, psychology, etc) under investigation. Many of the techniques developed for medical imaging also have scientific and industrial applications.

Medical imaging encompasses a set of techniques that produce images of the internal aspect of the body. Medical imaging can be seen as the solution of mathematical inverse problems - in order to determine the location of the activity within a specific organ, advanced signal processing techniques are used to estimate the location of that activity's source, which is referred to as the inverse problem. (The forward problem is the situation where we know where the sources are and we are estimating the field at a given distance from them). This means that cause (the properties of living tissue) is inferred from effect (the observed signal). In the case of ultrasonography, for instance, the probe consists of ultrasonic pressure waves and echoes inside the tissue show the internal structure. In the case of projection radiography, the probe is X-ray radiation which is absorbed at different rates in different tissue types such as bone, muscle and fat.

Some of the prominent imaging technologies that are available are:

1) Electron microscopy: The electron microscope is a microscope that can magnify very small details with high resolving power due to the use of electrons as the source of illumination, magnifying at levels up to 2,000,000 times. Electron microscopy is employed in anatomic pathology to identify organelles within the cells. Its usefulness has been greatly reduced by immunohistochemistry but it is still irreplaceable for the diagnosis of kidney disease, identification of immotile cilia syndrome and many other tasks.

2) Radiographic: Two forms of radiographic images are in use in medical imaging; projection radiography and fluoroscopy, with the latter being useful for intra operative and catheter guidance. These 2D techniques are still in wide use despite the advance of 3D tomography due to the low cost, high resolution, and, depending on application, lower radiation dosages.

This imaging modality utilizes a wide beam of X-rays for image acquisition and was the first imaging technique available in modern medicine.

3) Ultrasound: Medical ultrasonography (Udupa and Herman, 2000) uses high frequency broadband sound waves in the megahertz range that are reflected by tissue to varying degrees to produce (up to 3D) images. This is commonly associated with imaging the fetus in pregnant women. Uses of ultrasound are much broader, however. Other important uses include imaging the abdominal organs, heart, breast, muscles, tendons, arteries and veins. While it may provide less anatomical detail than techniques such as CT or MRI, it has several advantages which make it ideal in numerous situations, in particular that it studies the function of moving structures in real-time, emits no ionizing radiation, and contains speckle (a random, deterministic, interference pattern in an image formed with coherent radiation of a medium). It is very safe to use and does not appear to cause any adverse effects, although information on this is not well documented. It is also relatively inexpensive and quick to perform. Ultrasound scanners can be taken to critically ill patients in intensive care units, avoiding the danger caused while moving the patient to the radiology department. The real time moving image obtained can be used to guide drainage and biopsy procedures. Doppler capabilities on modern scanners allow the blood flow in arteries and veins to be assessed.

4) Electroencephalography: Electroencephalography (EEG) (Swartz and Goldenson, 1998) is the recording of electrical activity along the scalp produced by the firing of neurons within the brain. In clinical contexts, EEG refers to the recording of the brain's spontaneous electrical activity over a short period of time, usually 20 to 40 minutes, as recorded from multiple electrodes placed on the scalp. In neurology, the main diagnostic application of EEG is in the case of epilepsy, as epileptic activity can create clear abnormalities on a standard EEG recording. A secondary clinical use of EEG is in the diagnosis of coma, encephalopathies, and brain death. EEG used to be a first-line method for the diagnosis of tumors, stroke and other focal brain disorders, but this use has decreased with the advent of anatomical imaging techniques such as MRI and CT. Derivatives of the EEG technique include evoked

potentials (EP), which involves averaging the EEG activity time-locked to the presentation of a stimulus of some sort (visual, somatosensory, or auditory). Event-related potentials refer to averaged EEG responses that are time-locked to more complex processing of stimuli; this technique is used in cognitive science, cognitive psychology, and psycho-physiological research.

5) Magnetoencephalography: Magnetoencephalography (MEG) (Cohen, 1972) is an imaging technique used to measure the magnetic fields produced by electrical activity in the brain via extremely sensitive devices such as superconducting quantum interference devices (SQUIDs). These measurements are commonly used in both research and clinical settings. There are many uses for MEG, including assisting surgeons in localizing a pathology, assisting researchers in determining the function of various parts of the brain, neurofeedback, and others. In research, MEG's primary use is the measurement of time courses of activity. MEG can resolve events with a precision of 10 milliseconds or less. MEG also accurately pinpoints sources in primary auditory, somatosensory and motor areas, whereas its use in creating functional maps of human cortex during more complex cognitive tasks is more limited; in those cases MEG should preferably be used in combination with fMRI. It should be noted, however, that neuronal (MEG) and hemodynamic (fMRI) data do not necessarily agree and the methods complement each other. However, the two signals may have a common source: it is known that there is a tight relationship between LFP (local field potentials) and BOLD (blood oxygenation level dependent) signals. Since the LFP is the source signal of MEG/EEG, MEG and BOLD signals may derive from the same source (though the BOLD signals are filtered through the hemodynamic response).

6) Tomography: Tomography is the method of imaging a single plane, or slice, of an object resulting in a tomogram. There are several forms of tomography: linear tomography, zonography, orthopantomography (OPT or OPG) and computed tomography (CT), or computed axial tomography (CAT).

7) Computed tomography: Computed tomography (CT) is a medical imaging method employing tomography created by computer processing. Digital geometry processing is used to generate a three-dimensional image of the inside of an object from a large series of two-dimensional X-ray images taken around a single axis of rotation. MRI scanners on the other hand produce about the same quality of images without using X-rays, thereby removing the increased risk of cancer.

8) Positron emission tomography: Positron emission tomography (PET) (Herman, 2009) is a nuclear medicine imaging technique which produces a three-dimensional image or picture of functional processes in the body. The system detects pairs of gamma rays emitted indirectly by a positron-emitting radionuclide (tracer), which is introduced into the body on a biologically active molecule. Images of tracer concentration in 3-dimensional or 4-dimensional space (the 4th dimension being time) within the body are then reconstructed by computer analysis. In modern scanners, this reconstruction is often accomplished with the aid of a CT X-ray scan performed on the patient during the same session, in the same machine. PET scans are increasingly read alongside CT or magnetic resonance imaging (MRI) scans, the combination ("co-registration") giving both anatomic and metabolic information (i.e., what the structure is, and what it is doing biochemically). Because PET imaging is most useful in combination with anatomical imaging, such as CT, modern PET scanners are now available with integrated high-end multi-detector-row CT scanners. Because the two scans can be performed in immediate sequence during the same session, with the patient not changing position in between, the two sets of images are more-precisely registered, so that areas of abnormality on the PET image can be more perfectly correlated with anatomy on the CT images. This is very useful in showing detailed views of moving organs or structures with higher anatomical variation, which is more common outside the brain. At the Julich Institute of Neurosciences and Biophysics, the world's largest PET/MRI device began operation in April 2009: a 9.4-Tesla magnetic resonance tomograph (MRT) combined with a positron

emission tomograph (PET). Presently, only the head and brain can be imaged at these high magnetic field strengths.

9) Magnetic resonance imaging (MRI): Magnetic Resonance Imaging (MRI)(Udupa and Herman, 2000), or nuclear magnetic resonance imaging (NMRI), is a medical imaging technique most commonly used in radiology to visualize the internal structure and function of the body. MRI provides much greater contrast between the different soft tissues of the body than computed tomography (CT) does, making it especially useful in neurological (brain), musculoskeletal, cardiovascular, and oncological (cancer) imaging. Unlike CT, it uses no ionizing radiation, but uses a powerful magnetic field to align the nuclear magnetization of (usually) hydrogen atoms in water in the body. Radiofrequency (RF) fields are used to systematically alter the alignment of this magnetization, causing the hydrogen nuclei to produce a rotating magnetic field detectable by the scanner. This signal can be manipulated by additional magnetic fields to build up enough information to construct an image of the body. Since MR forms the base of the dissertation, I will discuss it more in details in the next two sections.

10) Nuclear medicine: Nuclear medicine encompasses both diagnostic imaging and treatment of disease, and may also be referred to as molecular medicine or molecular imaging and therapeutics. Nuclear medicine uses certain properties of isotopes and the energetic particles emitted from radioactive material to diagnose or treat various pathologies. Different from the typical concept of anatomic radiology, nuclear medicine enables assessment of physiology. This function-based approach to medical evaluation has useful applications in most subspecialties, notably oncology, neurology, and cardiology.

1.2 History of Magnetic Resonance Imaging

Critical to making Magnetic Resonance Imaging (MRI) a reality was the advent of the high speed computers needed to handle the enormous quantity and complexity of the computations involved in imaging. In addition to the necessary computing power, three other

developments contributed to the birth of MRI. One was the work of British electronics engineer Godfrey Hounsfield, who in 1971 built an instrument that combined an X-ray machine and a computer and used certain principles of algebraic reconstruction to scan the body from many directions–manipulating the images to produce a kind of cutaway view of the interior. Unknown to Hounsfield, South African nuclear physicist Allan Cormack had published essentially the same idea in 1963, using a reconstruction technique called the Radon transform. Although Cormack's work was not widely circulated, in 1979 he and Hounsfield shared the Nobel Prize in physiology or medicine for the development of computerized tomography, or CT. The principles underlying CT are the foundation of many sophisticated imaging methods in use today.

The other two developments essential to MRI were related to nuclear magnetic resonance (NMR). One was the conceptualization of NMR as a medical diagnostic tool; the other was the invention of a practical method for producing useful images from NMR data.

As early as 1959, J. R. Singer at the University of California, Berkeley, proposed that NMR could be used as a non-invasive tool to measure in vivo blood flow (Singer, 1959). Then in 1969, Raymond Damadian, a physician at Downstate Medical Center in Brooklyn, New York, began to think of a way to use the technique to probe the body for early signs of cancer. In a 1970 experiment he surgically removed fast-growing tumors that had been implanted in lab rats and showed that the tumors' NMR signals differed from those of normal tissue. Damadian published the results of his experiments in 1971 in the journal *Science* (Damadian, 1971).

An essential technical advance that opened up the ensuing widespread application of NMR to produce useful images was due to chemist Paul Lauterbur, who was then at the State University of New York at Stony Brook. In 1971, he watched a chemist named Leon Saryan repeat Damadian's experiments with tumors and healthy tissues from rats. Lauterbur con-

cluded that the technique was insufficiently informative for locating and diagnosing tumors and went on to devise a practical way to use NMR to make images (Oransky, 2007).

Lauterbur's groundbreaking idea was to superimpose on the spatially uniform static magnetic field a second, weaker, magnetic field that varied with position in a controlled fashion, creating what is known as a magnetic field gradient. At one end of a sample the graduated magnetic field would be strong, becoming weaker in a precisely calibrated way down to the other end. Because the resonance frequency of nuclei in an external magnetic field is proportional to the strength of the field, different parts of the sample would have different resonance frequencies. Thus, a given resonance frequency could be associated with a given position. Moreover, the strength of the resonance signal at each frequency would indicate the relative size of volumes containing nuclei at different frequencies and thus at the corresponding position. Subtle variations in the signals could then be used to map the positions of the molecules and construct an image. Today's magnetic resonance imaging devices impose three sets of electromagnetic gradient coils on the subject to encode the three spatial coordinates of the signals.

Across the Atlantic in Britain, Peter Mansfield at the University of Nottingham, England, had a similar idea. He was looking into using NMR to obtain structural details of crystalline materials. In work published in 1973, Mansfield and his colleagues also used a field gradient scheme to develop a MRI technique known as echo-planar imaging, which can rapidly scan a whole brain (Mansfield et al., 1973).

Meanwhile, Lauterbur's results, published in 1973, included an image of his test sample: a pair of small glass tubes immersed in a vial of water. Working with the small NMR scanner he had created (and using a technique called back projection borrowed from CT scanning), he continued to image small objects. By 1974, using a larger NMR device, he produced an image of the thoracic cavity of a living mouse. Mansfield, for his part, had imaged a number of plant stems and a dead turkey leg by 1975, and by the next year he had captured the

first human NMR image - a finger. Damadian also was at work producing images. In 1977, he produced an image of the chest cavity of a live man (Oransky, 2007).

By the early 1980s the flurry of activity around MRI had given rise to a burgeoning commercial enterprise. ("Nuclear" had been quietly dropped from the name in the meantime because of its unfavorable connotations.) Advances in high-speed computing and superconductive magnets allowed researchers to build larger MRI machines with enormously improved sensitivity and resolution and made possible many new applications.

1.3 Science of Magnetic Resonance Imaging

A magnetic resonance imaging instrument (MRI scanner) uses powerful magnets to polarize and excite hydrogen nuclei (single proton) in water molecules in human tissue. This produces a detectable signal which is spatially encoded, resulting in images of the body. MRI uses three electromagnetic fields: a very strong (on the order of units of Teslas, which is the derived unit of magnetic inductivity) static magnetic field to polarize the hydrogen nuclei, called the static field; a weaker time-varying (on the order of 1 kHz) field(s) for spatial encoding, called the gradient field(s); and a weak radiofrequency (RF) field for manipulation of the hydrogen nuclei to produce measurable signals, collected through an RF antenna (Hacke et al., 1999).

Like CT, MRI traditionally creates a two dimensional image of a thin "slice" of the body and is therefore considered a tomographic imaging technique. Modern MRI instruments are capable of producing images in the form of 3D blocks, which may be considered a generalization of the single-slice, tomographic, concept. Unlike CT, MRI does not involve the use of ionizing radiation and is therefore not associated with the same health hazards. However, MRI has only been in use since the early 1980s; although there are no currently known long-term effects of exposure to strong static magnetic fields, that has been a subject of some debate. At present there is no limit to the number of scans to which an individual can be subjected, in contrast with X-ray and CT. However, there are well-identified health

risks associated with tissue heating from exposure to the RF field and the presence of any metal in the body, such as pace makers. These risks are strictly controlled as part of the design of the instrument and the scanning protocols used.

Because CT and MRI are sensitive to different tissue properties, the appearance of the images obtained with the two techniques differ markedly. In CT, X-rays must be blocked by some form of dense tissue to create an image, so the image quality when looking at soft tissues will be poor. In MRI, while any nucleus with a net nuclear spin can be used, the proton of the hydrogen atom remains the most widely used, especially in the clinical setting, because it is so ubiquitous and returns a large signal. This nucleus, present in water molecules, allows the excellent soft-tissue contrast achievable with MRI.

Chapter 2

Functional Brain Imaging

It has long been of interest to researchers to learn about the human brain and its functioning from the perspectives of development, cognition, social behavior and others. Only in recent years have technologies been developed that enable observing a working brain, collecting the data on it and performing analysis to comprehend and address various issues regarding the location or intensity or extent of the areas that are stimulated in response to a particular task. Functional Brain Imaging, more specifically Functional Magnetic Resonance Imaging (fMRI) is one such technique and is the focus of this dissertation.

Functional Magnetic Resonance Imaging (fMRI) is a type of specialized MRI scan for imaging brain activity and studying the processes underlying the changes in the hemodynamic response (changes in blood oxygenation and flow) related to response in the neural activity in the human brain.

2.1 History of Functional Brain Imaging

Nearly 30 years ago the introduction of X-ray computed tomography (CT) set in motion a revolution in medical imaging that changed forever the practice of medicine. The introduction of CT also proved to be an immediate and powerful catalyst for the development of other imaging techniques, particularly positron emission tomography (PET) and magnetic resonance imaging (MRI) [for historical perspectives see Webb (1990) and Kevles (1997)]. With the development of PET and MRI came the opportunity to not only look at the anatomy of organs within the living human but also to evaluate their function (Raichle, 2000).

With these new imaging techniques, researchers interested in the function of the human brain were presented with an unprecedented opportunity to examine the neurobiological correlates of human behaviors. This opportunity along with prescient early support from the combined resources of some government funding agencies contributed significantly to the development of the field of cognitive neuroscience, a field of research that combines the experimental strategies of psychology with various techniques to actually examine how brain function supports mental activities.

The field of cognitive neuroscience, particularly related to studies involving functional imaging techniques, has experienced explosive growth over the past 15 years. This is exemplified not only by a plethora of published papers in established as well as new journals, some devoted exclusively to imaging (e.g., *NeuroImage*, *Human Brain Mapping*), but also by the formation of new societies (e.g., the Organization for Human Brain Mapping and the Cognitive Neuroscience Society). Equally remarkable has been a worldwide movement to establish research-imaging centers in which expensive imaging equipment (primarily MRI), along with teams of investigators, is devoted exclusively to research.

The research on brain images relates not only to the scientific importance of the work itself but also to the fact that the subject matter of cognitive neuroscience touches on subjects of importance to everyone (e.g., normal as well as disordered memory, attention, language, motivation, emotion, decision making, and even consciousness). In addition, the imaging data produced by cognitive neuroscientists are often quite intriguing; observing the brain of another human at work seems to fascinate scientists and nonscientists alike.

Despite these advances and new findings, researchers and scientists have questioned the ability of this approach to provide analyses of brain function that are sufficiently refined to truly enlighten us about the relationship between human behavior and brain function (Nichols and Newsome, 1999). One of the keys to evaluating such concerns is the ability to

relate work in cognitive neuroscience and imaging to that which parallels it in other areas of neuroscience.

Among the most important questions are how to relate functional imaging to the cell biology and neurophysiology of brain cells and their microvasculature. Additionally, it seems like a good time to ask whether cognitive neuroscience with its imaging tools has provided us with new insights into brain function and organization or merely confirmed what we have known all along (Raichle, 2003). Such issues have been explored for a long period of time. It is useful to consider the intended goal of functional brain imaging. This may seem self-evident to most, yet interpretations frequently stated or implied about functional imaging data suggest that, if we are not careful, functional brain imaging could be viewed as no more than a modern and extraordinarily expensive version of 19th century phrenology (Nichols and Newsome, 1999).

It is Korbinian Brodmann, one of the pioneers of cytoarchitectonic parcellation of the cerebral cortex, whose perspective is appealing even though it was written well in advance of the discovery of modern imaging technology (Brodmann, 1909). He said "... Indeed, recently theories have abounded which, like phrenology, attempt to localize complex mental activity such as memory, will, fantasy, intelligence or spatial qualities such as appreciation of shape and position to circumscribed cortical zones." He went on to say "... these mental faculties are notions used to designate extraordinarily involved complexes of mental functions... one cannot think of their taking place in any other way than through an infinitely complex and involved interaction and cooperation of numerous elementary activities... in each particular case (these) supposed elementary functional loci are active in differing numbers, in differing degrees and in differing combinations... Such activities are... always the result... of the function of a large number of suborgans distributed more or less widely over the cortical surface.... " (for these English translations see Garey (1994), pages 254 to 255).

With this prescient admonition in mind, the assignment of functional brain imaging becomes clear: identify multiple regions and their temporal relationships associated with the performance of a well designed task. The brain instantiation of the job will emerge from an understanding of the elementary operations performed within such a network. The great strength of functional brain imaging is that it can contribute uniquely to such an assignment by providing a broad and detailed view of the processing architecture of cognitively engaged networks. Importantly, this can be accomplished in the brain of most interest to us, the human brain. It is fair to say that functional brain imaging, using increasingly sophisticated experimental and analytical strategies and ever more powerful imaging devices, will contribute significantly to this important enterprise in studies of humans as well as experimental animals.

A second general point relates to the nature of the functional imaging signal. Functional brain imaging with fMRI is based on a remarkably consistent relationship between regional changes in the cellular activity of the brain and changes in the circulation and metabolism of that region. Scientists have known since the late 1800s that local blood flow in the brain changes in parallel with changes in cellular activity (Raichle, 1998, 2000). More surprisingly, it was discovered in the 1980s with PET that these changes in blood flow are not accompanied by comparable changes in local oxygen consumption (Fox and Raichle, 1986; Fox et al., 1988). This discrepancy between changes in blood flow and changes in oxygen consumption results in changes in the local concentration of oxygen in the micro circulation of the brain. Because MRI signals are sensitive to the oxygenation of blood (Ogawa et al., 1990a,b), this discovery paved the way for the introduction of fMRI (Bandettini et al., 1992; Kwong et al., 1992; Ogawa et al., 1992).

Much discussion and debate have centered on the cellular events underlying this apparent change in brain metabolism (i.e., a relative increase in energy released when glucose is being converted in the body – this process is termed glycolysis) (Raichle, 1998; Buxton and Frank, 1997; Buxton et al.,1998). Presently, the most parsimonious explanation for this observation

is that the increase in glycolysis is related to metabolic changes in astrocytes (the star-shaped neuro-cells in the brain and spinal cord) associated with increased clearance of glutamate – which plays an important role in human metabolism as it is the most common amino acid and the main component in many proteins in the human tissues – from the synapse (the tiny gap between the ends of nerve fibers across which nerve impulses pass from one neuron, a nerve cell, to another.) (Magistretti et al., 1999; Mintun et al., 2001; Shulman et al., 2001). It has been suggested recently that the astrocyte is also a critical link between neurons and blood vessels in orchestrating the changes in blood flow associated with changes in neuronal activity (Zonta et al., 2003). Present information thus leaves little doubt about the central importance of the astrocyte in the cell biology of functional brain imaging signals.

Regardless of how the energy consumption of the brain is altered locally to meet its changing demands, it is critically important to know from a neurophysiological perspective just what cellular events are associated with the local changes in blood flow, metabolism, and tissue oxygenation. To many it may seem obvious that these changes must relate to the spiking activity of neurons because that is what neurophysiologists most commonly measure in relation to behavior (Kirsten et al., 2004). In fact, the spiking activity of neurons has been used as the gold standard in assessing the ability of functional imaging signals to track events of interest within the brain (Hyder et al., 2003; Smith et al., 2003). Surprisingly for some, it is not the spiking activity of neurons that is important. Rather, synaptic events, as reflected in local field potentials, are most influential in determining the signals obtained with functional imaging, a result anticipated by early tissue autoradiographic studies of metabolism (Sharp, 1976; Sharp et al., 1977; Schwartz et al., 1979). These new findings obviously pose an interesting challenge when interpreting correspondences within the brain, or their absence, in the results of functional imaging research.

In closing, it is important to maintain a sense of proportion when it comes to viewing functional imaging signals. In the average adult human, the brain represents 2 percent of the body weight. Remarkably, despite its relatively small size, the brain accounts for 20 percent

of the oxygen, and hence calories, consumed by the body (Clark and Sokoloff, 1999), which is 10 times that predicted by its weight alone. In relation to this very high rate of baseline metabolism, functional imaging signals are remarkably small, in metabolic terms usually less than 5 percent of the ongoing metabolism of the brain, truly modest modulations in ongoing or baseline activity. Evidence now suggests that this baseline activity may instantiate important components of brain function (for an introduction to these issues see Gusnard and Raichle (2001) and Raichle and Gusnard (2002)).

2.2 Science of fMRI

In order to understand the statistical issues inherent in fMRI, it is important to understand the mechanics, physics and biophysics underlying the data acquisition process.

To start, let's look at the parts of the MRI machine.

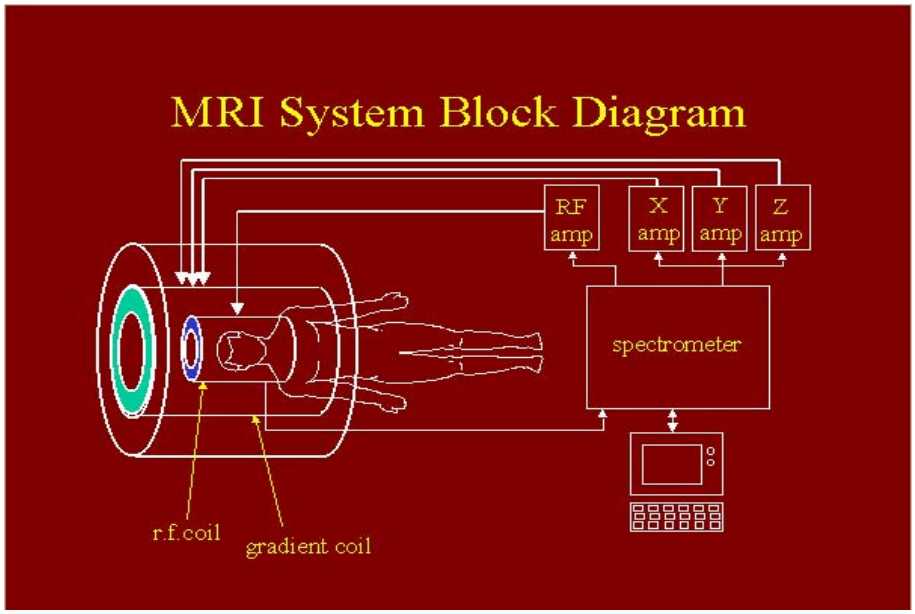

Figure 2.1: MRI system (Jezzard, 1999).

From Figure 2.1, we can see that the three basic components of the MRI machine are:

1) The primary magnet: This is the largest part of the scanner. A permanent magnet, powerful enough to use in an MRI scanner would be too costly to produce and too cumbersome to

store. The other way to make a magnet is to coil electrical wire and run a current through it. This creates a magnetic field within the center of the coil. In order to create a strong enough magnetic field to perform MRI, the coils of the wire must have no resistance; therefore they are bathed in liquid helium at a temperature of 450 degrees Fahrenheit below zero. This allows the coils to develop magnetic fields of 1.5 to 3 Tesla (the strength of most medical MRI machines).

2) The gradient magnets: There are three smaller magnets within an MRI machine called gradient magnets or coil (as referred in Fig 2.1). These magnets are much smaller that the primary magnet (about 1/1000 as strong), but they allow the magnetic field to be altered very precisely. They are the "fine-tuning" part of the MRI machine. The magnet produces the B_0 field (where B_0 is the constant, homogeneous magnetic field used to polarize spins, creating magnetization; the direction of this field defines the longitudinal axis) for the imaging procedure. Within the magnet are the gradient coils for producing a gradient in B_0 in the X, Y, and Z directions. It is these gradient magnets that allow image "slices" of the body to be created. By altering the gradient magnets, the magnetic field can be specifically focused on a selected part of the body. The gradient amplifiers (X, Y and Z amplifiers shown in Figure 2.1) increase the power of the gradient pulses to a level sufficient to drive the gradient coils. They are also responsible for the "clanging" noise heard during a scan.

3) The radiofrequency coil: Within the gradient coils is the RF coil and it is next to the part of the body being imaged. There are coils made for shoulders, knees, and other body parts. The human body is composed primarily of hydrogen atoms (63%); other common elements are oxygen (26%), carbon (9%), nitrogen (1%), and relatively small amounts of phosphorus, calcium, and sodium. MRI uses a property of hydrogen atoms called "spin" to distinguish differences between tissues such as muscle, fat, and tendon. The radiofrequency coil produces the B_1 magnetic field necessary to rotate the atoms (see Figure 2.3) by any degree selected by the radiofrequency pulse sequence which makes MRI possible. The radiofrequency coil also detects the signal from the spins within the body. The patient is positioned within the

magnet by a computer controlled patient table. The table has a positioning accuracy of 1 mm. The scan room is surrounded by radiofrequency shield. The shield prevents the high power radiofrequency pulses from radiating.

The heart of the imager is the computer. It controls all components on the imager. The computer interprets the data, and creates images that display the different resonance characteristics of different tissue types. The radiofrequency amplifier increases the pulse's power from milliwatts to kilowatts. The computer also controls the gradient pulse programmer which sets the shape and amplitude of each of the three gradient fields. The gradient amplifier increases the power of the gradient pulses to a level sufficient to drive the gradient coils. For the purpose of brain imaging, the machine has additional features like headphones to cut out the noise, video screen for display of the visual tasks and so on (Figure 2.2).

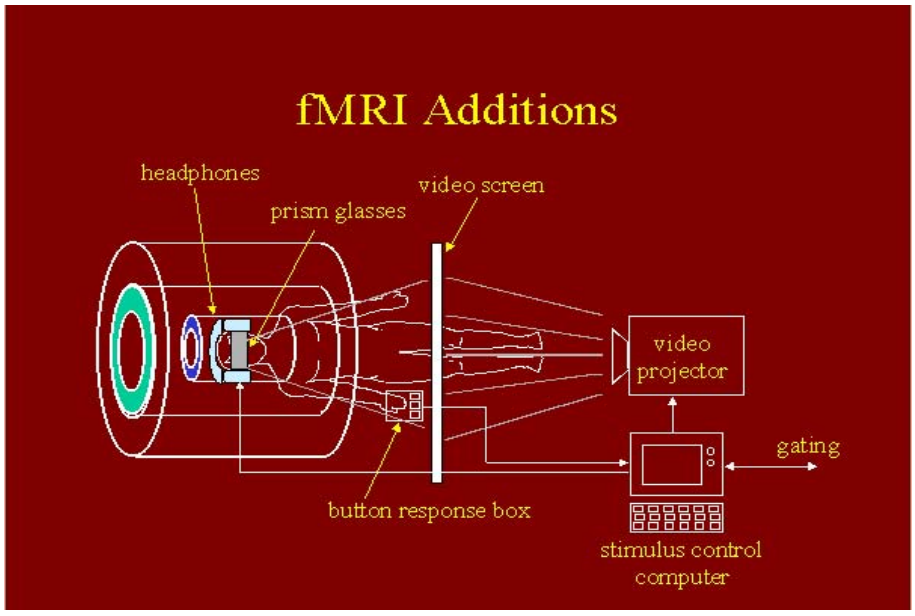

Figure 2.2: fMRI system (Jezzard, 1999).

Magnetic fields affect or attract atomic nuclei with an odd number of protons or an odd number of neutrons. These nuclei will try to align themselves parallel or anti-parallel to the field when they are exposed to a strong magnetic field. The parallel orientation has slightly lower energy than the anti-parallel orientation. Hence more nuclei align themselves in the parallel orientation, thereby causing an overall magnetization of the object in the field. The

alignment of the nuclei is not perfect in either direction. If one places an atom within a magnetic field plane, i.e, subjects it to magnetic forces along two of the three dimensions, then the nucleus will orbit around the third (vertical) axis. These atoms *precess* about the field at a fixed frequency (precession refers to the revolution of the axis of rotation of the atoms). Each type of nucleus has its own precession frequency and it is linearly proportional to the strength of the magnetic field. In other words, when one causes nuclei to precess their spin (Figure 2.3) will cause them to align themselves with the magnetic field. The spin of a nucleus is just like the ends of a bar magnet in that it can have a positive or negative value. Two negative or two positive ends of a magnet repel one another, but negative and positive ends attract each other. Similarly, all the negative spin atoms align themselves downward towards the feet of the subject, and all the positive atoms align upward towards the subject's head. Each atom with a positive spin cancels out (render undetectable) an atom with a negative spin. There remain, however, a few atoms that do not cancel one another out. At room temperature, there are always more positive spin atoms than negative spin atoms. Recorded is also the amplitude of the signal at each voxel in each image.

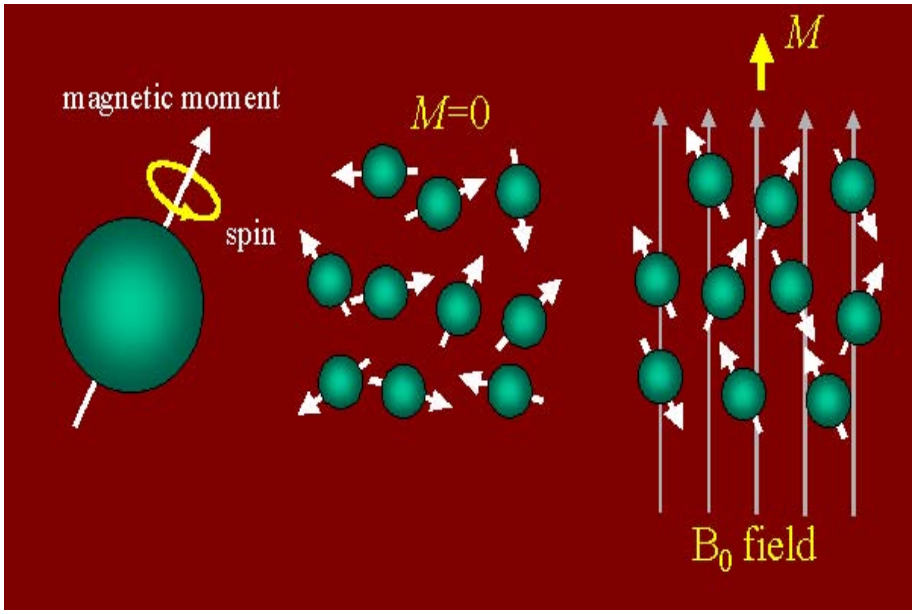

Figure 2.3: Alignment of spins in a magnetic field, M (Jezzard, 1999).

Positive spin atoms are in a low energy state. The atoms achieve an equilibrium magnetization value along the direction of the magnetic field, i.e., the Z axis. When radiofrequency energy (a pulse of magnetic energy) is injected into the system at the frequency of precession of the nuclei and perpendicular to the main magnetic field, it excites the nuclei, which are the unmatched ones, temporarily and then the nuclei return to equilibrium state. Associated with this is the energy which is emitted in order for the nuclei to return to the equilibrium state, which is at the frequency of precession. So with the radiofrequency signals, only the nuclei that are near the appropriate precession frequencies will get affected; hence absorption and emission of energy are termed as *selective*. It is this key concept of *selectivity* that provides magnetic resonance signals. The signal that the fMRI machine detects is the energy emitted by these unmatched atoms as they make a transition from the higher energy state to the lower energy state after the radiofrequency pulse. fMRI involves the use of the precession concept to collect high-resolution images. The strength of the signal is proportional to the number of nuclei of a specific type. Hence the method allows us to count the nuclei with particular properties (Lazar, 2008).

The signal intensity of the MR image is determined by four basic parameters: 1) proton density, 2) T1 relaxation time, 3) T2 relaxation time, and 4) flow. Proton density is the concentration of protons in the tissue in the form of water and macromolecules (proteins, fat, etc). The T1 and T2 relaxation times define the way that the atoms revert back to their resting states after the initial radiofrequency pulse. The most common effect of flow is loss of signal from rapidly flowing arterial blood.

The amount of time it takes for the atoms to return to their equilibrium value is called the "spin lattice relaxation time" or T1. T1 is, thus, a measure of the half-life of inverted spins. If one uses the gradient magnets inside the machine to alter the local net magnetization so that it is in the XY plane (cutting a very thin virtual slice across the patient), the local net magnetization rotates the Z axis (takes on positive and negative X and Y values) at a frequency called the Larmor frequency. The Larmor frequency equals the frequency of the

atom which would cause a transition between the two energy levels of the nucleic spin. Atoms that are placed in a magnetic field of given strength, usually denoted by B_0, absorb photons of frequency ω if the atoms have nonzero spin (Lazar, 2008). The frequency of absorption depends on the *gyromagnetic ratio*, γ, of the nucleus and is expressed by the Larmor equation:

$$\omega = \gamma B_0$$

By again introducing a pulse of magnetic energy in the form of a radiofrequency pulse that is specific to the type of atom, the fMRI machine causes the unmatched atoms to resonate. The resonating atoms absorb the radio energy and go to the higher energy state, i.e., they become negative spin atoms relative to the XY axis (the transverse axis). The amount of time it takes for the atoms to return to their equilibrium magnetization value along XY axis (transverse axis) is called the "spin-spin relaxation time" or T2. T2, as a result, measures the rate of change of spin phases. Whereas a typical T1 (spin lattice relaxation time) is approximately 1 second, the T2 (spin-spin relaxation time) is usually less than 100ms. This difference in the relative times is what makes T2 better suited than T1 for functional metabolic imaging. The T1 relaxation curve can be described by an exponential function, $1 - \exp^{-t/T1}$, where t is the elapsed time; if M_0 is the original magnetization, then M_z, the amount of longitudinal magnetization at time t following an excitation pulse, is given by

$$M_z = M_0 \left(1 - \exp^{-t/T1}\right)$$

T2 relaxation is the result of gradual loss of phase coherence, and hence it is a result of heterogeneities in the tissue which is described by an exponential function,$\exp^{-t/T2}$, or

$$M_{xy} = M_0 \exp^{-t/T2}$$

where M_0 is as before and M_{xy} is the signal loss at time t. Since loss of phase coherence must occur before equilibrium can be reached, T2 is usually shorter than T1 (Lazar, 2008).

Particularly important for fMRI is the measure of decay of transverse magnetization, T2* which takes into account two important factors: molecular interactions and inhomogeneities

in the magnetic field. fMRI creates the images or brain maps of brain function by setting up and utilizing an advanced MRI scanner in such a way that increased blood flow to the activated areas of the brain shows up on the MRI scans. The MRI scanners do not actually detect blood flow or other metabolic processes. Rather, blood flow alterations and/or associated metabolic processes in brain areas are indirectly inferred from the signal intensity contrast for a given brain region relative to both normal levels and levels immediately adjacent to the area in question. The intensity of an MRI signal is determined by the level of magnetic resonance, which is the BOLD (blood-oxygen level dependent) effect on T2* (Wallis, 2009).

2.3 Data Collection in fMRI

In fMRI, a series of magnetic resonance images is collected over time to gather information about the neuronal activation in the brain during the course of the scan. These images are typically three-dimensional as they are divided into volume elements or voxels. A scan session involves placing a subject inside the MR machine and asking him/her to perform some particular task of interest. While the subject performs the task, the scanner takes images of the working brain, and regions that are activated in response to the stimulus can be detected using statistical methods that will be discussed below. Magnetic fields are altered to some extent by the presence of any substance. Many materials exhibit pronounced polarization in a magnetic field. The degree of this effect is referred to as the "magnetic moment" or "magnetic susceptibility". The magnetic properties of oxygenated and deoxygenated hemoglobin differ (Pauling, 1935). Spatial and temporal variation in local concentrations of deoxygenated hemoglobin to oxygenated hemoglobin result in corresponding changes in magnetic susceptibility, which in turn cause the local T2* values to fluctuate. Oxygenated hemoglobin is diamagnetic (i.e., tends to take a position at right angles to the lines of magnetic force, and is repelled by either pole of the magnet), while deoxygenated hemoglobin is paramagnetic (i.e., takes a position parallel and proportional to the intensity of the magnetizing field). Thus, MRI is able to detect a small difference (a signal of the order of 3 percent) between

the two types of hemoglobin (Thulborn et al., 1982). This is called a blood-oxygen level dependent, or "BOLD" signal. Researchers are currently exploring the precise relationship between neural activity and the BOLD signal.

As the basic story goes: blood is delivered to the brain by arteries and transported from the brain by veins. Not only is the actual blood volume relatively low in the brain, but the majority of blood volume is in the capillary bed–the very small vessels that connect arteries and veins. Capillaries are often so small that hemoglobin travels in single file. Whereas arterial blood has a high concentration of oxygenated hemoglobin, as the blood cells pass through the capillary bed the concentration of deoxygenated hemoglobin increases relative to oxygenated hemoglobin. Thus, a gradient of highly oxygenated hemoglobin to highly deoxygenated hemoglobin runs across the capillary bed from arteriole to venulle. As a result, a corresponding gradient in T2* ranges from longer T2* (diamagnetic oxygenated hemoglobin-rich) to shorter T2* values (paramagnetic deoxygenated hemoglobin-rich). That is to say, when resting neurons become active, the rate of blood flow to the neighborhood of these neurons increases as glucose is being delivered to the regions of interest. This is known as the *hemodynamic response*. There is a rise in the metabolism of these neurons as the rate of firing increases which in result enhances the influx of oxygenated blood to the affected regions. Since active neurons do not consume much more oxygen than resting neurons, oxygen levels rise in the nearby blood vessels too. Due to the difference in the magnetic properties of the oxygenated and deoxygenated blood, the magnetic resonance signal from the neighborhood of a neuron should change as the concentration of oxygenated blood around the neuron changes. The idea that there exists a correlation between blood flow changes and changes in brain function had been toyed with for a considerable amount of time (Raichle, 1994). Magnetic resonance imaging is sensitive enough to detect these functionally induced changes in blood oxygenation in the human brain (Ogawa et al., 1990; Kwong et al., 1992) and functional magnetic resonance imaging is a step in further understanding of this process.

The relevant spatial unit for measuring local T2* for fMRI contrast is called a "voxel". A voxel is the smallest unit of MRI reconstruction, and corresponds to a single pixel in an MRI display image. The relative ratio of deoxygenated to oxygenated hemoglobin within a voxel determines the T2* value for that voxel. The increase in T2* resulting from increase in metabolic function causes a corresponding increase in image intensity.

The raw data from an MR scanner are spatial frequency data. Spatial information is determined from both the phase (longitudinal i.e. T1 and transverse i.e. T2) of the magnetization and the frequency of the MR signal. By using a gradient magnetic field, the phases and frequencies of protons in different locations can be localized. The two methods that can localize the signal being detected by the receiving coil are phase encoding and frequency encoding, which can be separated according to their timing during the image formation process. Information from three dimensions, provided by 1) slice excitation, 2) frequency encoding, and 3) phase encoding, is used to create an image of spatial location based on frequencies and phases. This is encoded, in that a Fourier transform is used to create the image itself from the raw data–a Fourier transform is the conversion of data in the time domain (which is how it is collected), into the frequency domain, by modeling it as a sine wave. The magnetic gradients and radiofrequency pulses focused on a slice allow differentiation of locations based on the phase and the frequency data collected from each coordinate. The Fourier transform is needed to convert the raw values into phase and frequency information that can be displayed as a wave. That is, the data are the coefficients of the Fourier representation of the object being imaged. Alternatively we can say that the data are the inverse Fourier transform of the object. The spatial frequency domain has been called "Fourier Space", "frequency space", or most popularly "k-space". The process of taking the inverse Fourier transform to obtain the image (image space as shown in Figure 2.4) has been termed "data reconstruction". Letting Ψ be the Fourier transform, we have for an $n \times m$ pixel image I,

$$\psi \quad = \quad \hat{I}(k_x, k_y)$$

$$= \quad n^{-1}m^{-1}\sum_{x}\sum_{y} I(x,y)\exp(-i2\pi(xk_x + yk_y))$$

So, k-space is a graph of spatial frequency. In k-space, the low frequencies (large values) are located in the center of the image with the frequencies increasing outward with distance from the origin i.e. the smaller the value the further from the center the coordinate will be located. The larger the dimensions to the k-space image (the larger the amount of information in the periphery), the more detailed the resulting image will appear. In short, the intensity of the signal determines the central part of k-space (as seen in Figure 2.4 with the darker pixels indicating larger values), and the level of detail determines the periphery. In fMRI both low and high frequency information is important.

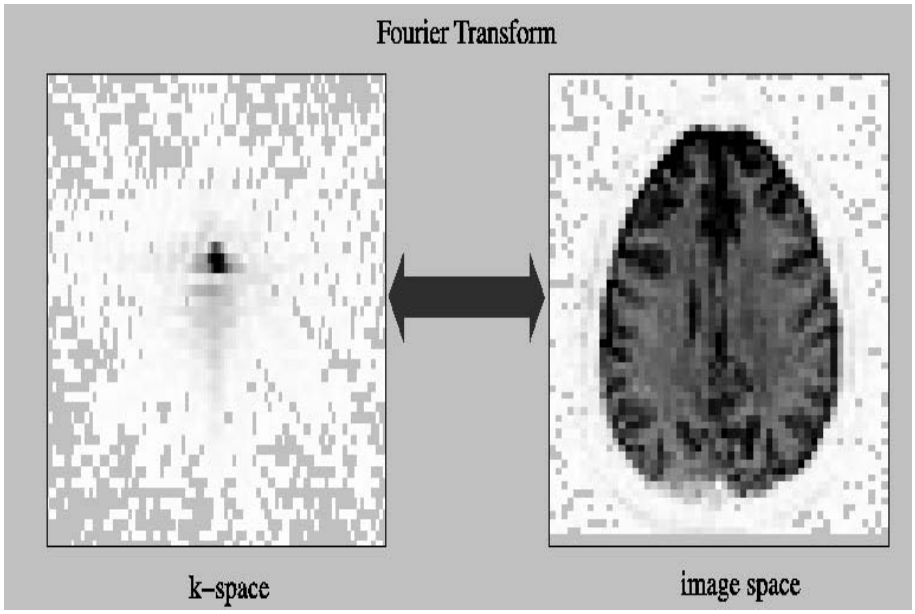

Figure 2.4: Collected fMRI data (Eddy and McNamee, 2004). The plot on the left shows the modulus of the k-space data, and the plot on the right shows the modulus of the image. Darker pixels indicate larger values (the opposite of the "radiological convention" derived from X-ray images on photographic film).

A typical fMRI data set might consist of a 64 by 64 array of 16 bit complex values recorded for each of 32 two-dimensional slices, each ranging between 80-200 time points spaced a few seconds apart. This yields a huge quantity of data collected in a small amount of time. If many

runs are performed on a single subject over the course of an hour or so, and several subjects are examined over time, the necessary storage requirements can become quite extensive.

2.3.1 An Example

For purposes of illustrating how the cognitive process works in relation to an activity and what can be statistically deduced from the images while the subjects perform that activity, we consider a simple example:

Table 2.1: Examples of a cognitive task - different sentence types for language study.

Sentence type	**Example**
common-simple	*the writer attacked the king and admitted the mistake*
common-hard	*the writer that the king attacked admitted the mistake*
rare-simple	*the pundit attacked the regent and admitted the gaffe*
rare-hard	*the pundit that the regent attacked admitted the gaffe*

An experiment was carried out to examine the role of syntactic and lexical difficulty in the comprehension of sentences (Keller et al., 1998). There were four experimental conditions (as shown in Table 2.1): common words (high lexical frequency) in sentences with simple syntax, common words in sentences with difficult syntax, rare words (low lexical frequency) in sentences with simple syntax and rare words in sentences with difficult syntax.

Subjects were all native English speakers. The subjects were visually presented with the sentences and had to answer comprehension (true/false) questions regarding who-did-what-to-whom. The main interest of the researchers was in the measured activation in the left temporal language area which would be a direct effect of difficulty in comprehension. This activity enabled researchers to understand how human beings process and then comprehend information and extract meaning out of it.

Analysis of the subjects revealed a main effect of lexical frequency, a main effect of syntactic complexity and an interaction between the two i.e. lexical and syntactic frequencies (Lazar et al., 1999). Sentences using rare words induce more stimulation in the language area, as

do sentences with greater syntactic difficulty (main effect). This type of finding is not only intriguing but is also indicative of the progress in sophistication regarding the information that can be attained using fMRI.

Chapter 3

Statistics in Functional Magnetic Resonance Imaging

The features that characterize the fMRI data acquired on single subjects are abundance, noise and high correlation both temporally and spatially. We can think of the data as a time series, or more generally a movie, of the human brain in action. The complexity as well as the volume of the data makes statistics an integral part in their analysis, comprehension and information extraction.

The field of statistics makes valuable contributions to functional imaging research by establishing procedures for the design of brain imaging experiments and providing tools for objectively quantifying and measuring the strength of scientific evidence provided by the data. Common research objectives include detecting brain regions that reveal task-related alterations in measured brain activity (detection) and identifying highly correlated brain regions that exhibit similar patterns of activity over time.

Statistical issues stem from the nature of the problem addressed, for example, what specific regions in the brain from different subjects in the same cohort are activated while performing a particular task? Or what is their intensity of activation? Data are collected as images of the human brain that are acquired over time as the experiment progresses. Since the same person is involved, that is, the brain of a single individual is doing all the work, all the voxels are correlated with each other (spatial correlation). And since the images are of the same person over a short period of time, there is also temporal correlation. Added to these complications, the data also include random noise (plausible causes would be equipment variability and so on). Hence the signal to noise ratio becomes very small, making it important to look at the processes closely. It then becomes imperative to use statistics to analyze data from the cohort

in totality and also to use risk and probability based statements to address the questions of interest.

This chapter deals with the way in which the raw data from an fMRI experiment are analyzed. The aim of such analysis is to determine those regions in the image in which the signal changes upon stimulus presentation. Although it is possible to devise many different techniques for detecting activation, if these techniques are to be used in practice it is necessary to know how much confidence can be placed in the results. That is to say, what is the probability that a purely random response could be falsely labeled as activation. This requires an understanding of the statistics behind the technique used.

Many of the statistically robust techniques used to analyze fMRI data have been developed from PET. These try to model the time course that is expected, and determine how well each voxel's temporal response fits this model. However, since fMRI experiments allow good time resolution, it is possible to carry out experiments which determine the order in which different cognitive events occur. Analysis of the data from such an experiment requires a "non-directed" technique which makes few assumptions about the timings of the activation responses expected.

There are three stages to the analysis of the data from any fMRI experiment (Figure 3.1). Firstly there are the preprocessing steps, which are applied to the data to improve the detection of activation events. These include registering the images, to correct for subject movement during the experiment, and so on. There are some well-known and understood sources of noise, for example scanner drift, head motion, differences in timing of slice acquisition which need to be corrected for. If we analyze the data without correcting for these, we will get false impressions of the activation patterns in the data. So we make corrections like smoothing in the preprocessing stage to help clean up the sources of noise that are understood and can be easily corrected for. This has the effect of improving our ability to detect relevant activation because it increases the signal to noise ratio. Next, the statistical

analysis, which detects the pixels in the image which show a response to the stimulus, is carried out. Finally the activation images must be displayed, and probability values, which give the statistical confidence that can be placed in the result, quoted.

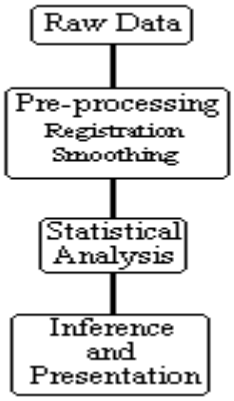

Figure 3.1: Steps involved in the processing of fMRI data.

3.1 Preparing MR Images for Statistical Analysis

There are a number of steps that can be carried out prior to the statistical analysis of the data. Each of these steps is independent and offers different benefits. The flowchart diagram in Figure 3.1 shows the sequence in which fMRI data are prepared for analysis.

3.1.1 Preprocessing

Preprocessing is an essential step before analyzing functional brain imaging data because the measured signal change is very small compared to the total intensity of the functional magnetic resonance signal and the task related signal change is very small compared to the total spatial and temporal variability across images - hence changes due to non-task related sources need to be removed as much as possible prior to statistical analysis.

Areas of brain activity that are found due to specific tasks are dependent on the image to image changes in the measurements within a voxel. Therefore, to produce valid results these

changes must be specifically attributable to functional differences in the brain elicited by the task. Unfortunately, fMRI data are beset with many sources of bias and variability, which can lead to erroneous detection of regions of brain activity and false conclusions about the study. Problems in the data can arise from many sources including the MR scanner itself, the experimental subject, and external interference. The sources of noise in fMRI data can be quite extensive (Eddy and McNamee, 2004). Although many are covered here, this summary is not exhaustive. Preprocessing comes into play in minimizing the variability and bias due to these sources of noise.

(i) Noise from the equipment: One main source of bias and systematic variation in fMRI data arises from the MR scanner. The performance of an MR scanner can vary, which can introduce fluctuations in the data, even when the stability measures are well within the instrumental norms (Weisskoff, 1996). Noise from the equipment can occur as systematic or random errors.

Sources of systematic error in the data from the equipment include DC shifts and Nyquist ghosts. DC shifts are also known as baseline errors. This source of data bias is caused by the miscalibration of the analog-to-digital (A/D) converter; the baseline value is not reported as zero. Nyquist ghosts, which are present only in echo-planar imaging, also produce systematic bias in the data. Echo-planar pulse sequences traverse k-space on a boustrophedonic path (back-and-forth as the ox plows the field). Nyquist ghosts are introduced through the mist-iming of the oscillating magnetic gradients. The exact time at which the gradient crosses zero is incorrect. This timing error causes an aliasing effect in the reconstructed image and is most prominent in the phase-encode or direction of the fMRI scan (leading to a ghost of the image repeated at the top and bottom of the true image). Both DC shift errors and Nyquist ghosts that are present in the fMRI data can be corrected to a reasonable extent.

Random errors from the equipment can also introduce problems in the fMRI data. One source of unpredictable instability results from inhomogeneities in the static magnetic field

of the equipment. Magnetic field inhomogeneities have been reported as one of the most prominent sources of distortion in fMRI studies (Jezzard, 1999)). Local variations in the static magnetic field during fMRI will lead to blurring and pixel shifts, which can introduce gross geometric distortions in the images. This problem is especially prominent at regional boundaries in the sample containing different magnetic susceptibility properties, for example, air-tissue interfaces around the frontal lobes and bone-tissue interfaces.

Additionally, random instability in the MR machine can result from imperfections in the B_1 field. The B_1 field is ideally a linear magnetic gradient that selects certain regions of tissue to be excited, thereby leading to the collection of single slices. Again, problems with this linear magnetic field can lead to blurring and geometric distortions in the data.

(ii) Noise from the experimental subject:

As with other types of human studies, the experimental subjects can lead to large amounts of bias and variability in the data. While the subjects themselves have a great deal of intrinsic variability due to differences in brain sizes, shapes, and functionality in general, the subjects can also introduce additional variability that will "drown out" the desired results from brain activity if the investigator is not careful.

One important source of noise from the experimental subject is due to head motion. As previously described, BOLD fMRI studies compare very small regions of brain tissue across a sequence of images that are taken over the course of several minutes. While BOLD has the advantage that it requires no exogenous contrast agents, its measurable effects are very small. Typical changes in the MR signal due to BOLD are on the order of 1-5%, making this technique highly susceptible to noise. If the subject makes a small movement during the scan, which can vary in signal value by more than 10%, adjacent voxels cause distortions in the recorded signal information and can lead to false negative and false positive regions of activation (Eddy et al., 1996). A typical voxel measures $3mm$ by $3mm$ by $3mm$, so even a small amount of motion will "shift" the measured signal into a different voxel.

Thus, to obtain valid fMRI data, the subject must remain motionless throughout the scanning period. Motion has been shown to be correlated with stimulus related events during visual stimulation, thereby contributing to the likelihood that the computed regions of activation are due to motion artifact rather than neural activity (Hajnal et al.,1994). The amount of subject motion has also been shown to increase over time during the course of a scanning session (Green et al., 1996). Additionally, children, elderly subjects, and subjects with mental disorders tend to move more than healthy young adults, thereby increasing the difficulty of studying these subjects using fMRI.

A second source of error from the experimental subject is due to "physiological noise", which is noise that results from the subject's heart beat and respiration. This type of complex noise is thought to interfere with the MR data through various mechanisms. For example, the pulsatile motions of the brain and cerebral spinal fluid (CSF) induced from pressure changes during both the cardiac and respiratory cycle lead to volume changes within the head which cause displacement of tissue (Dagli et al., 1999). Large organ movements due to respiration are also thought to cause fluctuations in the magnetic field, and effects of the oscillating cardiac cycle on the BOLD signal response are unknown (Dagli et al., 1999).

There are many other sources of noise associated with the experimental subject. Thermal noise is caused by atomic vibration that occurs at any temperature above absolute zero. Susceptibility artifacts arise from local sharp changes in magnetic susceptibility; these occur at the boundaries of tissue types and are typically greatest at air/tissue boundaries. Chemical shift artifacts arise from small changes in the Larmor frequency caused by the local chemical environment.

(iii) External noise:

Interference from outside sources can also lead to distortions and artifacts in the data. Examples of interference sources include mechanical vibrations from other equipment in the building or passing vehicles, and 60 (or 50) Hertz RF noise from other nearby electrical

equipment. These sources are usually considered before installing the MR machines, and precautions are normally taken. For example, an isolated foundation will reduce the effect of external sources of vibration; copper shielding will reduce the effect of nearby sources of microwave radiation, and iron shielding will reduce the effect of nearby electrical equipment (and help contain the magnetic field itself).

Typical preprocessing steps include (Lindquist, 2008):

(a) Slice timing correction: Corrects for differences in acquisition time within a TR (in conventional imaging, TR is the time between phases where the effective TR is the time between images; typical values range from 100 to 4000 ms). This is a temporal correction. To correct slice-timing errors, experimental analysis modifies the predicted hemodynamic response so that each slice is compared to a hemodynamic response function with slightly different timing. Temporal interpolation techniques are used to estimate the amplitude of the MR signal. The compromise we make here is that the efficiency of temporal interpolation reduces with increased TR as the sampling frequency is reduced.

(b) Motion correction: Motion correction is done first to minimize motion effects (eg. head motion) associated with interpolation across adjacent voxels at a cost of slight timing uncertainty. There are two types of artifacts due to motion viz. random movements which produce a blurry and noisy image and periodic motion which creates ghost images. This correction is achieved through transformations and/or minimization of squared differences (eg. sum of squared differences). However the drawbacks with motion correction are: loss of data at edges of imaging volume, ghosts in image do not change in the same manner as the real data, distortions may be due to position in field and not position in head, etc.

(c) Coregistration: Coregistration is the process of aligning all image volumes to a reference volume i.e. coregistration of functional T1-weighted fMRI images is a necessary step for combining functional information with anatomical information. The benefit of having coregistered images is that it enables the visualization of functional data by superimposing

it on a high-resolution anatomical MR image and therefore improves localization of neural activation loci. The advantages of coregistration are: aids in normalization, allows display of activation on anatomical images, allows comparison across modalities and necessary if there are no anatomical images. The disadvantages of coregistration are: may severely distort functional data and may reduce correspondence between functional and anatomical images.

(d) Normalization: Normalization is a form of coregistration, except that here the image volumes to be coregistered differ in shape and not as a result of distortion. Normalization enables us to compare fMRI results across subjects and across different studies. The most widely used stereotaxic (an external three dimensional frame of reference to locate voxels within the brain) space is Talairach space (Talairach and Tournoux, 1988). Normalization algorithms determine the overall size of the brain as well its gross anatomical features to warp it to the common template. The advantages of normalization are: allows generalization of results to larger population, improves comparison with other studies, provides coordinate space for reporting results and enables averaging across subjects. The disadvantages are: reduces spatial resolution, may reduce activation strength by subject averaging, time consuming and could be potentially problematic (doing bad normalization is much worse than not normalizing).

(e) Spatial smoothing: Smoothing is a process by which the data points are averaged with their neighbors in an image. Spatial smoothing reduces the high frequency spatial component in the images. The technique used is normally a Gaussian kernel i.e the normal distribution function is used to take the weighted average of neighboring points. This is because it has a very narrow pass band (frequency of interest) and it attenuates all frequency outside it. The advantages of spatial smoothing are: increase in signal to noise ratio (SNR) and may improve comparisons across subjects. However the disadvantages are: reduces spatial resolution and it is challenging to smooth accurately if size/shape of signal is not known.

(f) Segmentation: Classifies voxels within an image into different anatomical divisions for easy visual representation, for example, in Figure 3.2 we see gray matter (yellow), white matter (pink) and cerebro-spinal fluid(CSF)(green).

Figure 3.2: Segmenting into 3 anatomical divisions viz. gray matter, white matter and CSF (Bizzell and Belgar, 2002).

(g) Identification of regions of interest (ROI): The ROI-based approach is used to allow direct, unbiased measurement of activity in an anatomical region, to improve ability to identify topographic changes (eg., motor mapping, social perception mapping, etc.) and to complement voxel-based analyzes. This is not strictly a preprocessing step as others and is heavily dependent on the problem of interest.

(h) Bias field correction: Corrects for intensity nonuniformities (bias) in fMRI images. There are two models in the literature that correct for bias. The first model involves resetting the image to a hypothetical ideal. The second model uses a template image with known voxel intensities and compares the subject image to it. The idea here is that deviations in the ratios of corresponding voxel intensities identify nonuniformities which can be removed from the subject image.

3.1.2 Detection and Estimation

Detection, that is, locating which voxels are activated in response to a given task, and estimation of the hemodynamic response function are two different topics of interest. Most of the commonly employed statistical techniques have as their end product a map, usually referred to as a statistical parametric/non-parametric map of the brain. These maps are a graphical representation of the output of the statistical analysis at each voxel of the brain: a map of t-statistics, or F statistics, and so forth. These maps enable us to locate the voxels that become activated while performing a particular task. Efficient estimation of the hemodynamic response function not only can be used to arrive at conclusions as to which voxels should be regarded as *active* but it might also be of interest on its own. Event-related designs offer maximum estimation efficiency but poor detection power, while block designs offer good detection power at the cost of minimum estimation efficiency. However, both these designs can simultaneously achieve the estimation efficiency of randomized designs and the detection power of block designs at the cost of increasing the length of an experiment (Liu et al., 2001).

3.1.3 Thresholding

Acquisition of the statistical parametric/non-parametric maps involves a combination of statistical models, statistical tests, and corrections for multiple testing. When we aim at answering the question "Which voxels show significant levels of activation during task compared to the control condition?", the task of classifying each of the enormous number of voxels in a typical study as "significantly active" or "not significantly active" is a formidable problem of multiplicity. This issue of multiplicity is known as the problem of *thresholding* in the neuroimaging community.

In any statistical test, a binary decision (fail to reject null/reject null) is made. The true state of nature is also binary (null is true/null is false). Thus, underlying a series of statistical

tests and decisions, such as might be taken regarding the voxels in an fMRI dataset, is a simple two-by-two table (Table 3.1).

Table 3.1: The binary decision of a statistical test - reject the null hypothesis or fail to reject the null hypothesis - in conjunction with the true state of nature: null is true or null is false.

Choices	**Fail to reject null**	**Reject null**	**Total**
Null True	m_{00}	m_{01}	$m_{0.}$
Null False	m_{10}	m_{11}	$m_{1.}$
Total	$m_{.0}$	$m_{.1}$	m

In total there are m voxels, where m numbers in the hundreds of thousands for fMRI studies. Of these, $m_{.1}$ are declared active, that is, the null hypothesis is rejected. In reality, $m_{0.}$ voxels are inactive, that is, the null hypothesis is true, and $m_{1.}$ are active. We are truly interested in the m_{11} voxels for which the null is rejected, when in fact it is false. These are the *true activations*. Different approaches to correcting for multiple testing aim at different types of control of this unknown number.

A standard quantity to control is the *familywise error rate* (FWER), which is the probability of having even one false discovery across all the tests. More recently, methods that control the *false discovery rate* (FDR), or the expected proportion of incorrectly rejected null hypotheses, out of all the voxels that have been declared active, have become increasingly popular. FDR is less conservative, with greater power than FWER control, at the cost of increasing the likelihood of obtaining Type 1 errors (Lazar, 2008).

3.2 Statistical Analysis of fMRI Images

In this section, I will give a very brief overview of different approaches to obtaining activation maps, followed by a slightly more detailed introduction to analysis via the general linear model (GLM; currently the most popular statistical approach).

After the pre-processing steps, statistical analysis is carried out to determine which voxels are activated by the stimulus or task. Many techniques have been proposed for statistically

analyzing fMRI data, and a variety of these are in general use. These techniques can be simple correlation analysis or more advanced modeling of the expected hemodynamic response to the stimulus or task. Various possible statistical corrections can be included, such as correction for smoothness of the measured time series at each voxel. The aim of these various analyses is to produce an image that shows regions with significant signal change in response to the task. Each voxel is assigned a value, dependent on the likelihood of the null hypothesis, a relevant test statistic or equivalently a p-value. Such an image is called a statistical parametric map.

It is most common to analyze each voxel's time series independently ("univariate analysis"). For example, the standard GLM analysis is univariate (although cluster-based thresholding, commonly used at the final inference stage, does use spatial neighborhood information and is therefore not univariate). There are also "multivariate" methods which process all the data together; these methods make more use of spatial relationships within the data than does the univariate analysis (Jezzard et al., 2001).

There is also a distinction between model-based and model-free methods. In a model-based method, a model of the expected response is generated and compared with the data. In a model-free method, effects or components of interest in the data are found on the basis of some specific criteria (for example, the spatial or temporal components should be statistically independent of each other). This allows for "surprise" in the data, and also the analysis of data where it is difficult to generate a good model. There are also a few methods which lie between model-based and model-free, for example (Clare et al., 1999), where the only "model" information given is the time of the beginning of each stimulation period (the actual time-course within each period is not pre-specified). A statistical map is generated by comparing the variance within periods with the variance across periods.

In this section, I demonstrate analysis techniques on an example data set. The experiment performed was intended to detect activations resulting from a visually cued motor task. The whole brains of the subjects were imaged, in 16 coronal slices of resolution 3 x 3 x 10 mm^3,

every four seconds. As cued by an LED display, subjects were required to squeeze a ball at the rate of 2 Hz. The experiment involved 16 s of rest, followed by 16 s of task performance, repeated 32 times (Zar, 1996).

3.2.1 Subtraction Techniques

One of the simplest methods for obtaining results from a two state fMRI experiment is to perform a simple subtraction. This is carried out by averaging together all the images acquired during the "on" phase of the task, and subtracting the average of all the "off" images. The disadvantage of such a technique is that it is extremely sensitive to head motion, leading to large amounts of artifact in the image. Figure 3.3a shows a single slice through the motor cortex from the example data set, and Figure 3.3b shows the result of subtracting the "off" images from the "on" images. Although signal increase can be seen in the primary motor cortex, there is also a large amount of artifact, particularly at the boundaries in the image. This analysis ignores both spatial and temporal correlation.

Such a method does not yield a statistic that can be tested against the null hypothesis, so instead of straight subtraction it is more common to use a Student's t-test. This weights the difference in means, by the standard deviation in "off" or "on" values, giving high t-scores to large differences with small standard deviations, and low t-scores to small differences with large standard deviations. The t-score is calculated on a pixel by pixel basis, for a time series X, using the formula

$$t = \frac{\overline{X_1} - \overline{X_2}}{S}$$

where

$$S = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

and S_p^2 is the pooled variance

$$S_p^2 = \frac{\sum \left(X_1 - \overline{X_1}\right)^2 + \sum \left(X_2 - \overline{X_2}\right)^2}{n_1 + n_2 - 2}$$

The suffix "1" refers to the n_1 images acquired during the "on" period of the task, and "2" refers to the n_2 images acquired during the rest period. Figure 3.3c shows the statistical parametric map of t-scores for the sample data set. Again motor cortex activation is clearly seen, but the movement artifact in the figure 3.3b is slightly reduced compared to the subtraction technique.

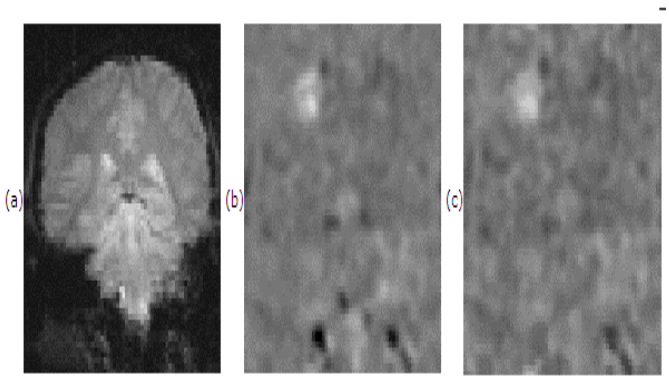

Figure 3.3: Use of subtraction techniques to analyze fMRI data. (a) A single slice coronal image through the primary motor cortex. (b) The mean of the images acquired during the "off" period of the fMRI experiment subtracted from the mean of the images acquired during the "on" period. (c) The t-statistical parametric map corresponding to image (b)(Stuart, 1997).

3.2.2 Correlation Techniques

Since we know that the BOLD response is mediated by blood flow, it is possible to improve the detection of activations by predicting the shape of the response to the stimulus, and calculating correlation coefficients between each pixel time course and the reference waveform. This is less sensitive to other physiological changes during the experiment, and to movement. For a time course X, and a reference waveform Y, we calculate the usual Pearson's correlation coefficient, denoted by r.

The choice of an appropriate reference waveform is vital for the success of this technique in finding activations. The first approximation might be a square wave, which is high for scans acquired during the task, and low for scans acquired during rest (Figure 3.4a). Such a waveform however takes no account of the delay and smoothness of the hemodynamic

response which regulates the BOLD contrast. An improvement to this would be to change the phase of the square wave (Figure 3.4b), with the delay being between 3 and 6 seconds. This corresponds with the delay in the onset of the hemodynamic response.

To improve the reference waveform further, it is necessary to look more closely at the actual hemodynamic response. In an experiment such as the one used for the example data set, where there is both visual and motor activation, it is possible to use the response to one type of stimulus to form the reference waveform for finding the other. In this case the time series for one or more pixels in, say the visual cortex is extracted (Figure 3.4c), and correlation coefficients are calculated between this waveform and that of every other pixel in the image. Such an analysis detects only those regions in the brain which respond to the stimulus in the same way as the visual cortex. The major disadvantage of this technique is that it is particularly sensitive to motion artifact, since if such artifact is present in the reference waveform then the movement of other regions will be highly correlated. In an attempt to reduce this, the response in the visual cortex to each stimulus can be averaged together, producing a mean response to the single cycle. The reference waveform is then made up of a repetition of these single cycle average responses (Figure 3.4d).

To be more general in predicting the hemodynamic response, so that a reference waveform can be constructed for any length of stimulus, it is necessary to know the response to a single stimulus. Friston et al. (1995) suggested that a hemodynamic response function could be considered as a point spread function (point spread function is the Fourier transform of any distribution kernel—we usually use Gaussian distribution kernel; point spread function answers the question, how much blurring would occur if you are trying to image a point), which smooths and shifts the input function. By deconvolving the response from a known area of activation with the stimulus function, the hemodynamic response function can be obtained. The hemodynamic response function is not completely uniform across the entire brain however, and the shape obtained from one region may not be optimal for another. As an alternative, the response can be modeled by a mathematical function, such as a Poisson

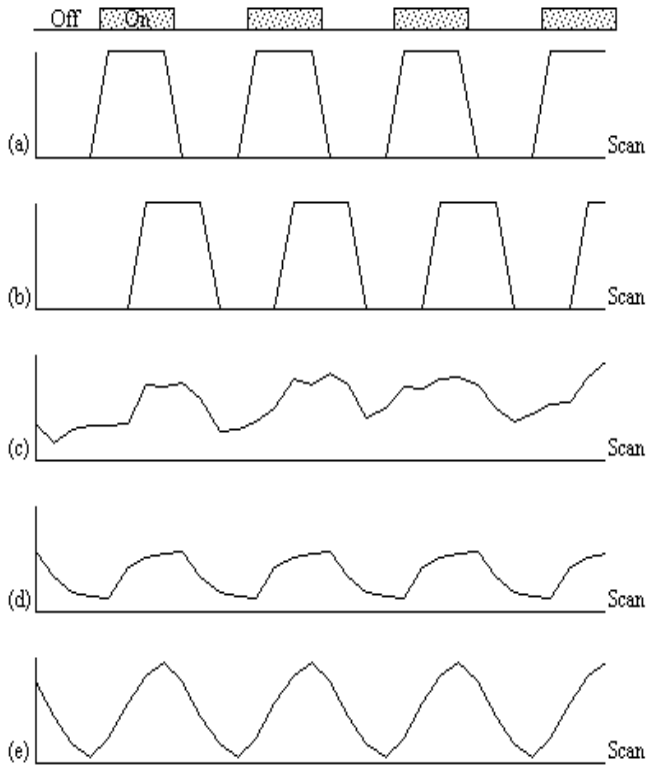

Figure 3.4: Various reference functions that can be used to correlate with a pixel time course to detect activations (Stuart, 1997).

function (Friston et al., 1994), Gamma-variate function (Lange and Zeger, 1996), Gaussian function (Rajapakse et al., 1998) or difference of two gamma variate functions (Glover, 1999) and so on. The Poisson function (see Figure 3.4e) gave a reasonable fit to this example dataset and hence was used to fit the observed hemodynamic response.

Since in general each slice of the volume imaged is not acquired at the same instant, it is necessary to accommodate timing differences in the correlation with the reference waveform (one of the steps in preprocessing involving slice time correction would have eradicated this issue but that was not done in this example). In order to do this, the relative magnitude of the activation at the time each slice was acquired is predicted, by convolving the input stimulus with a Poisson function. Then from this series, the correlation coefficients can be calculated on a slice by slice basis, constructing the reference waveform from the appropriate points in the predicted time series.

Examples of the effect of the reference waveform on the resultant analysis are shown in Figure 3.5. This example is used only to illustrate how the correlation technique works. Here, pixels in the head which correlate to the reference waveforms (shown in Figure 3.4), with $r > 0.40$ are shown in red, on top of the base image. The square wave correlation is the least effective in detecting activations (a), however a considerable improvement is obtained by delaying the waveform by 4 seconds (b). The correlation of the visual cortex with itself (c) is, not surprisingly, high, but using the average visual cortex response (d) improves the correlation in the motor cortex. The Poisson function model of the hemodynamic response (e) improves slightly on the delayed square wave, and is a good model.

3.2.3 General Linear Model

The statistical techniques described above are both parametric and specifically assume that the observations are taken from normal populations. Most parametric modeling techniques are special cases of the general linear model. This framework for analyzing functional imaging

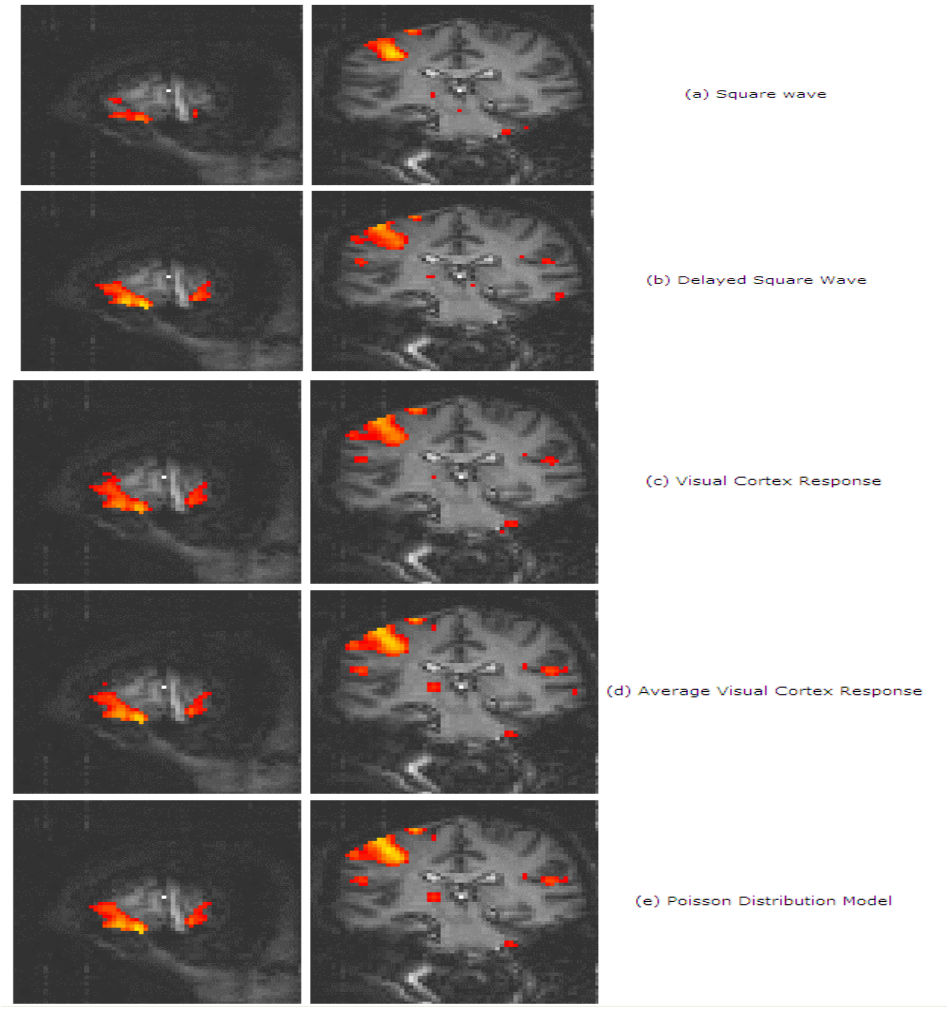

Figure 3.5: Activation images obtained by correlating the test data sets with the reference waveforms shown in Figure 3.4 (Stuart, 1997).

data was first developed for PET and then extended for fMRI. Here I will give a brief overview.

The aim of the general linear model is to explain the variation of the time course $y_1, ..., y_i, ..., y_n$, in terms of a linear combination of explanatory variables and an error term. We will first use GLM in a univariate way, where we consider one voxel only, and the fitting of models to a single voxel's time-course. Hence we will consider that the data of interest comprise a single 1D vector of intensity values. For a simple model with only one explanatory variable $x_1, ..., x_i, ..., x_n$, the general linear model can be written

$$y_i = x_i\beta + \epsilon_i$$

where β is the scaling, or slope parameter, and ϵ_i is the error term. If the model includes more variables it is convenient to write the general linear model in matrix form

$$Y = X\beta + E$$

where now Y is the vector of observed pixel values, β is the vector of parameters and E is the vector of error terms. The matrix X is known as the design matrix. It has one row for every time point in the original data, and one column for every explanatory variable in the model. In analyzing an fMRI experiment, the columns of X contain vectors corresponding to the "on" and "off" elements of the stimulus presented. By finding the magnitude of the parameter in β corresponding to these vectors, the presence or absence of activation can be detected.

β can be determined by solving the normal equations

$$X^T Y = \left(X^T X\right)\widehat{\beta}$$

where $\widehat{\beta}$ is the best linear estimate of β. Provided that $(X^T X)$ is invertible then $\widehat{\beta}$ is given by

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

Such parameter estimates are normally distributed (if error terms are normally distributed), and since the error term can be determined, statistical inference can be made as to whether the β parameter corresponding to the model of an activation response is significantly different from the null hypothesis that there is no contribution of a component in the design matrix to the value of response.

The general linear model provides a framework for most kinds of modeling of the data, and can eliminate effects that may confound the analysis, such as drift or respiration, provided that these can be modeled. The successive steps of every GLM-based method are: modeling the response at each voxel, then testing a hypothesis (about the parameters of the model) and representing the observed statistic map thresholded at a given level according to the point distribution (distribution at a given voxel i.e one random variable) of the statistic or according to the field distribution (continuous version of a multivariate distribution i.e a distribution of a vector of random variables) of the statistic (Leibovici and Smith, 2001). The GLM can refer to a single subject, to a single group or to multiple groups which can represent different subjects (e.g. male, female) or the same subjects (e.g. placebo, drug A, drug B in a cross-over design). As stated earlier, the modeling part is univariate, i.e. it is done separately for each voxel. The paradigm applied is usually well balanced (e.g. the same number of 'rest' and 'stimulation' scans in a block-design). The simplest approach uses only a t-test, and this analysis can be embedded in a general linear model to be able to take into account covariates in the analysis (Friston et al., 1995). Multi-subject fMRI experiments can also be expressed in a GLM framework with different forms according to the approach taken: fixed or random subject analysis which will be discussed in details in the next chapter. Estimation can be improved using spatial consideration (smoothing) in order to regain some quality of the estimation (spatially smoothed autocorrelation). In the context of multi-subject analysis a similar spatial consideration is investigated (Worsley et al., 2000) to "diminish between-subject variability".

Single-Subject Analysis

To illustrate the single-subject fMRI study, one collects, at each voxel, a time series of responses (intensities of the BOLD signal) to a stimulus, e.g. an ON (condition B) and OFF (condition A) experiment. This, however, is only one type of design, a block design with two sets of conditions. There are other types of designs, e.g., more than two sets of conditions in a block design, event-related, and so on. The same general principles apply in any case, with appropriate modifications to take account of the design matrix. Let y_t, be the time series observed from $t = 1,...,T$, at a given voxel. Among the T values observed at this voxel, T_A of them were recorded while under condition A (OFF or rest condition) and T_B of them were recorded while under condition B (ON or stimulation condition) according to the paradigm. The observations are assumed to be independent and identically distributed (i.i.d.) and to come from a Normal distribution with the same variance σ^2 and means μ_A and μ_B for the two conditions. To decide if there was an activation (at this voxel) during the experiment, one has to compare the means in the two conditions. If the difference of the means is big enough relative to its dispersion, one will assume activation. Under the null hypothesis (of no activation $\mu_A = \mu_B$):

$$t_0 \sim t_{dist}\,(T_A + T_B - 2)$$

In fMRI the data is often balanced i.e. $T_A = T_B = T/2$. The null is rejected at a chosen level of α to conclude that the voxel was activated if

$$p\,(t_{dist}\,(T_A + T_B - 2) \geq t_0) < \alpha$$

Multi-Subject Analysis

One easy implementation of this analysis is to perform a two-stage procedure. At the first stage the analysis explained in the previous section is performed for each of the n subjects. Hence for each subject, we create a statistical map of t-values. The second stage deals with

combining these maps to answer the question whether there is activation in the population from which these subjects have been drawn. There are two methods that can be chosen for this analysis and each gives a different conclusion for the population: fixed subject-effect analysis allows a conclusion limited to the sample studied and random subject-effect analysis allows a conclusion that can be extended to the population at large. If the sample is small then the latter will typically be conservative due to fairly large subject to subject variability; some alternative approaches have also been developed: "conjunction analysis" (Friston et al., 1995) and "variance ratio smoothing" (Worsley et al., 2000). Both fixed effect and random effect analysis are discussed in detail in the next chapter.

Alternatives to the Random or Fixed effect Analysis

Conjunction Analysis

Sometimes it gets difficult to choose either of the two approaches, fixed-effect model or random-effect model. Also, random-effect model is difficult to apply usually because of over-estimation of variances due to small samples. To overcome these drawbacks, some alternatives have been investigated. One is the "conjunction analysis" (Friston et al., 1995), which uses statistical maps from all the subjects to localize where all the subjects activated (at a chosen level i.e. it is a thresholded map of the *minimum* map over the subjects (Worsley and Friston, 2000)). This is to say that in order to show activation at a voxel at the group level, all subjects have to show activation at that particular voxel. This means that for p-value group maps, the "least significant" activation has to be below a threshold. Then using simple probability theory, we can find the proportion of the sample which shows activation at the previously defined level (in the single-subject analysis). Let (t) denote the tested status of activation with the experiment and (a) denote the true status of activation, while + indicates the status that it is activated and - indicates the status that it is not activated, then α_c= proportion of the population which shows activation at the single-subject level

$$\alpha_c \;=\; P\left(\text{all} t_+\right)$$

$$
\begin{aligned}
&= [P(t_+)]^n \\
&= [P(t_+/a_-)\,P(a_-) + P(t_+/a_+)\,P(a_+)]^n \\
&= [\alpha(1-\gamma) + \beta\gamma]^n
\end{aligned}
$$

where α is the chosen single-subject level of activation, β is the power or sensitivity of the experiment, which is not known but can be set at 1 to provide a lower bound of the proportion γ of the population showing the effect. Setting $\beta = 1$ gives

$$\gamma \geq \gamma_1 = \frac{\alpha_c^{1/n} - \alpha}{1 - \alpha}$$

Thus the conclusion about the population is, for example, with certainty of $0.95\,(1-\alpha_c)$, we can say that at least $80\%(\gamma_1)$ of the population would show activation at level of significance 0.001. However, this is not only stringent but the requirement to show consistent activation at the voxel level is unrealistic.

Variance Ratio Smoothing

Smoothing the observed variance over the whole brain would produce a better estimate of the random variance, but that assumes constant underlying variance over the brain. At times the homoscedasticity assumption will be violated, e.g. difference in white matter and gray matter (Woolrich et al., 2000). Worsley et al. (2000) suppose that the ratio of random-effects and fixed-effects variance is locally constant, so that smoothing the ratio would produce a pooled estimate. The method consists of first performing random and fixed effects analyses, then of spatially smoothing the ratio of variances obtained with the two methods (random/fixed), and finally returning to random-effects variance by multiplying the smoothed ratio with the fixed effects variance before performing the group t test. They estimate the degrees of freedom for the test as:

$$
\begin{aligned}
dfWratio &= 1/\left[1/df_{ratio} + 1/df_{fixed}\right] \\
&= df_{ratio}/\left(1 + df_{ratio}/df_{fixed}\right)
\end{aligned}
$$

where

$$df_{ratio} = df_{random}\left[2\left(\frac{FWHM_s}{FWHM_{data}}\right)^2 + 1\right]^{3/2}$$

and $FWHM_s$ is the Gaussian smoothing parameter which enables one to move between a random analysis if set to 0 (no smoothing) and a fixed analysis if set to ∞ (smoothing the variance ratio to one everywhere). Sensible choices for $FWHM_s$ would be not to increase too much the degrees of freedom comparatively to its no smoothing situation (df_{random}), an obvious limit being the degrees of freedom of a single subject experiment i.e. df_{fixed}/n. The value recommended (Worsley et al., 2000) is 15mm for an original smoothness of 6mm ($FWHM_{data}$).

One or two Groups of Subjects

A one-group analysis is a multi-subjects analysis with the obvious restriction that every subject of the random sample studied is a member of this group. Two-group analysis is carried out with a two-sample t-test. Under the fixed-effects approach, we have:

$$\widehat{\beta}_{1i} \sim N\left(m_{\beta_{1i}}, \widehat{\sigma}^2_{\beta\epsilon_{1i}}\right) i = 1, ..., n_1$$

$$\widehat{\beta}_{2i} \sim N\left(m_{\beta_{2i}}, \widehat{\sigma}^2_{\beta\epsilon_{2i}}\right) i = 1, ..., n_2$$

$m_{\beta_{1i}}$ and $m_{\beta_{2i}}$ represent the population mean activation in the two groups. Under the random-effects approach one will consider estimated mean activation, so we have:

$$\widehat{\beta}_{1i} \sim N\left(\beta_{1i}, \widehat{\sigma}^2_{\beta\epsilon_{1i}}\right) i = 1, ..., n_1$$

$$\widehat{\beta}_{2i} \sim N\left(\beta_{2i}, \widehat{\sigma}^2_{\beta\epsilon_{2i}}\right) i = 1, ..., n_2$$

We will consider the activation for a given subject as a random observation of the activation for the population. $\beta_{1i} = m_{\beta_1} + \eta_i, \beta_{1i}$ is random $\sim N\left(m_{\beta_1}, \sigma^2_{\beta_1\eta}\right)$ and $\beta_{2i} = m_{\beta_2} + \eta_i, \beta_{2i}$ is random $\sim N\left(m_{\beta_2}, \sigma^2_{\beta_2\eta}\right)$.

For each of the groups, it is a *two* levels variation (within subject and between subject). So now we have:
$\widehat{\beta_{1i}} = m_{\beta_1} + \eta_i$ and $\widehat{\beta_{2i}} = m_{\beta_2} + \eta_i$, i.e. two error terms in each group.

All the first level analysis has been done for every subject in each group. This looks very similar to the original simple single-subject analysis, but here the sample size of groups may be quite different. We will test $\theta = m_{\beta_1} - m_{\beta_2} = 0$, i.e. the population mean activation in both the groups are same, using a two sample t-test. So we have:

$$t_0\left(fixed\right) \approx t_{dist}\left(\left(n_1 + n_2\right)\left(T_A + T_B - 2\right)\right)$$

or

$$t_0\left(random\right) \approx t_{dist}\left(\left(n_1 + n_2\right) - 2\right)$$

Here no conjunction analysis approach can be made unless the same subjects are in both groups (for example, before and after medical treatment) as pairing of subjects across the groups would be required - in fact this then ends up reverting to a one-group analysis (with a paired -test). The "variance-ratio method" can, however, be performed.

So, $\widehat{m}_{\beta_1} = \widehat{\beta} \sim N\left(m_{\beta_1}, 1/n\sigma^2_{\beta_1\eta+\beta_1\epsilon}\right)$ and $\widehat{m}_{\beta_2} = \widehat{\beta} \sim N\left(m_{\beta_2}, 1/n\sigma^2_{\beta_2\eta+\beta_2\epsilon}\right)$,
$\widehat{\sigma}^2_{\beta_1\eta+\beta_1\epsilon} = \widehat{\sigma}^2_{\beta_1\eta} + \widehat{\sigma}^2_{\beta_1\epsilon}$ and $\widehat{\sigma}^2_{\beta_2\eta+\beta_2\epsilon} = \widehat{\sigma}^2_{\beta_2\eta} + \widehat{\sigma}^2_{\beta_2\epsilon}$.

On a side note, when g groups are studied one can also compare them two by two, then introducing a multiple comparison. If equal variances are assumed, a pooled variance of all g samples must be used for either method. To do more advanced comparisons, one has to return to the GLM or ANOVA to be able to test, for example, if all the groups have the same activation, or if there is a trend in the groups, as one would expect for groups defined by increasing doses of a treatment. These would involve either F statistics and/or using a linear function of the parameters estimated in the model, i.e. contrasts.

3.2.4 Serial t Test

For many experiments, the use of rapid imaging, and carefully designed paradigms, makes the separation of the order of cognitive events possible. One such example is a paradigm involving the initiation of movement. In this experiment, the subject is required to respond, by pressing a hand held button, to the visual presentation of the number '5', and to make no response to the presentation of a '2'. This paradigm presents two differences to conventional, epoch-based experiments. Firstly, the activations of interest, which are those responsible for the button pressing, occur at an irregular rate. Secondly, all the cognitive processes involved in the task, including both the planning and execution of the movement (event-related design), occur in a time period of a few hundred milliseconds, as opposed to the sustained activation used in epoch-based paradigms. Such an experiment requires a new form of analysis. Here two techniques are chosen, both of which make no assumptions about the time course of activations during the task: the serial t-test, described here, and an analysis of variance technique, explained in the next sub-section.

The basis of the serial t-test is to define a resting state baseline, and compare the images acquired at each time point before, during and after the task with this baseline. Figure 3.6 illustrates the technique. For each time point following the stimulus, a mean and standard deviation image is constructed, as is a baseline mean and standard deviation image. Then a set of t-statistical parametric maps is formed by calculating, on a pixel by pixel basis, the t-score (as described in the subsection on Subtraction Techniques) for the difference between mean image one and the mean baseline image, mean image two and baseline, and so on.
The technique has two major disadvantages. The first is that, in order to achieve a sufficient signal to noise ratio, it is necessary to have many more cycles than in an epoch-based paradigm (block design with alternating stimulation and rest), thus leading to longer experiments. This can be uncomfortable for the subject, and puts additional demands on the scanner hardware. There is some scope for bringing the single event tasks closer together, but there must be a sufficient interval to allow the BOLD signal to return to baseline. This

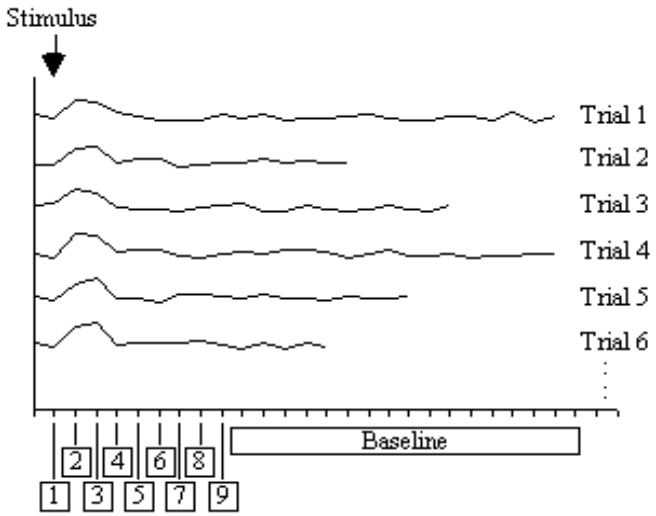

Figure 3.6: fMRI analysis using the serial t test (Stuart, 1997).

delay is at least ten seconds in length depending on the individual or the task given (Pollmann et al., 2000). The second disadvantage is that the analysis results in many statistical parametric maps, which have to be interpreted as a whole. However the fact that the technique makes few assumptions about the data time course makes it a strong technique, and opens up the possibility of more diverse experimental design, and a move away from the epoch-based paradigms.

3.2.5 Analysis of Variance

A second technique which does not require any assumptions about the shape of the activation time course, looks at the changes in variance upon averaging. The technique is based on simple signal averaging theory. Take, for example, the response measured to a repeated signal as shown in Figure 3.7. The time series contains two components, one is a genuine response to the signal, and the other is the random fluctuations due to uncorrelated physiological events and noise in the image. Upon averaging 32 cycles together, the magnitude of the

noisy component is reduced but that of the repeated signal is not. The reduction of the noisy component can be measured by calculating the variance of both the unaveraged and averaged data set.

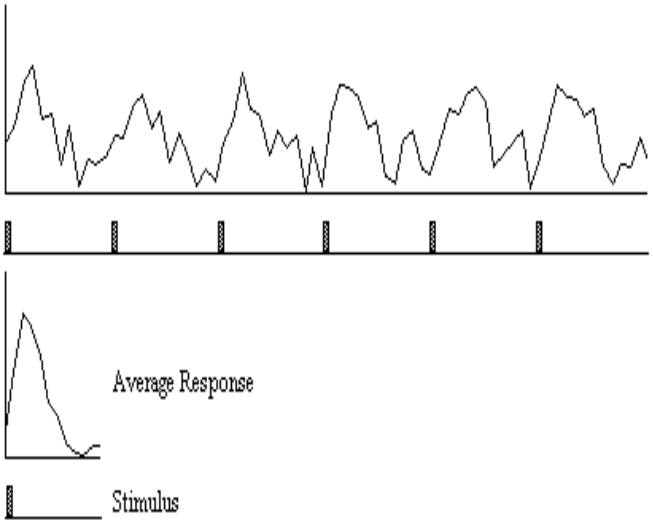

Figure 3.7: Signal averaging. The variance of the noise in the average signal is n times less than it is in the original signal, where n is the number of cycles (Stuart, 1997).

To detect regions of activation, the ratio of the variance of the averaged data set to the variance of the unaveraged data set is calculated for each pixel in the image. For pixels in regions of purely random intensity variations, this ratio will be around $\frac{1}{n}$, where n is the number of cycles averaged together. Pixels in regions of activation, however, will have a significantly higher ratio than this, since the variance of both unaveraged and averaged data sets is dominated by the stimulus locked intensity variations of the BOLD effect, which does not reduce upon averaging.

The technique is more formally explained as an analysis of variance (ANOVA). If X_{ij} refers to the i^{th} time point after the stimulus, of the j^{th} cycle of an experiment:

$$\begin{array}{cccccccc}
\text{time}t & X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1n} & \mathbf{X_1} \\
\text{time2}t & X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2n} & \mathbf{X_2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{time}it & X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{in} & \mathbf{X_i} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\text{time}kt & X_{k1} & X_{k2} & \dots & X_{kj} & \dots & X_{kn} & \mathbf{X_k} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \mathbf{X}
\end{array}$$

with n cycles and k points per cycle. The null hypothesis is that there is no significant difference in the means, $\bar{X}_i$. This can be tested by comparing two estimates of the population variance, σ^2, one based on variations in measurements of the same time point, and one based on the variance between time points.

The variance within measurement of any time point can be calculated by

$$s_i^2 = \sum_{j=1}^{n} \frac{\left(X_{ij} - \overline{X_i}\right)^2}{n-1}$$

and so the mean variance within time points is given by

$$\begin{aligned}
\widehat{\sigma}_W^2 &= \sum_{i=1}^{k} \frac{s_i^2}{k} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{\left(X_{ij} - \overline{X_i}\right)^2}{k\,(n-1)}
\end{aligned}$$

and is based on $k\,(n-1)$ degrees of freedom. The variance of the time point means is given by

$$s_{\overline{X}}^2 = \sum_{i=1}^{k} \frac{\left(\overline{X_i} - \overline{X}\right)^2}{k-1}$$

and since

$$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}$$

then σ^2 can be estimated by

$$\begin{aligned} \widehat{\sigma}_B^2 &= n \cdot s_{\overline{X}}^2 \\ &= n \cdot \sum_{i=1}^{k} \frac{\left(\overline{X_i} - \overline{X}\right)^2}{k-1} \end{aligned}$$

which is based on $k-1$ degrees of freedom. Under the null hypothesis, both $\widehat{\sigma}_W^2$ and $\widehat{\sigma}_B^2$ independently estimate the population variance σ^2. This means that the ratio

$$F = \frac{\widehat{\sigma}_B^2}{\widehat{\sigma}_W^2}$$

will have an F distribution with $k-1$ and $k(n-1)$ degrees of freedom. If there is any signal change that is time locked to the stimulus, the value of $\widehat{\sigma}_B^2$ will be larger than expected under the null hypothesis. In the analysis of fMRI data, all the above equations are used to form an F-statistical parametric map.

3.2.6 The statistical mapping approach

Statistical mapping is used for identifying regionally specific effects (e.g., brain activations) recorded during functional neuroimaging experiments using neuroimaging technologies such as fMRI to characterize functional anatomy and disease-related changes. It entails the characteristics below to perform the analysis:

Unit of measurement:

The fMRI scanner produces a "map" of the area being scanned that is represented as an array of voxels. Each voxel represents the activity of a particular coordinate in three dimensional space. The exact size of a voxel will vary depending on the technology used, although fMRI voxels typically represent a volume of 27 mm^3 (a cube with 3 mm length sides)

Experimental design:

Researchers are often interested in examining brain activity linked to a specific psychological

process or processes. An experimental approach to this problem might involve asking the question "which areas of the brain are significantly more active when a person is doing task A compared to task B?". The brain is likely to show changes in activity between tasks due to factors other than task differences (as the brain is involved with co-ordinating a whole range of parallel functions unrelated to the experimental task). Furthermore, the signal may contain noise from the imaging process itself as explained earlier in this Chapter.

To accommodate these random effects, and to highlight the areas of activity linked specifically to the process under investigation, statistics are used to look for the most significant difference above and beyond background brain activity. This involves a multi-stage process to prepare the data, and to subsequently analyze it using some statistical method.

Image pre-processing:

Images from the brain scanner may be pre-processed before any statistical comparison takes place to remove noise or correct for sampling errors. This has been explained in detail earlier in this Chapter.

Statistical comparison:

Parametric statistical models are assumed at each voxel, most commonly using the general linear model to describe the variability in the data in terms of experimental and confounding effects, and residual variability. Hypotheses expressed in terms of the model parameters are assessed at each voxel with univariate statistics.

Graphical representations:

Differences in measured brain activity can be represented in a number of ways. Most simply, they can be presented as a table, displaying coordinates that show the most significant differences in activity between tasks. However, differences in brain activity are more often shown as patches of color on an MRI brain "slice", with the colors representing the location of voxels that have shown statistically significant differences between conditions. The gradient of color is mapped to statistical values, such as t-values or z-scores. This creates an intuitive

and visually appealing means of delineating the relative statistical strength of a given area of activation. Recently, an alternative approach has been suggested, in which the statistical map is combined or overlaid with the map of the original difference in brain activity (or, more generally speaking, with the original contrast) and color codes are attributed to the latter (Reimold et al., 2006). Differences in activity may also be represented as a "glass brain", a representation of three outline views of the brain as if it were transparent. Only the patches of activation are visible as areas of shading. This is useful as a quick means of summarizing the total area of significant change in a given statistical comparison (Reimold et al., 2006).

3.3 Introduction to the Thesis

The human brain controls all activities ranging from heart rate, breathing, motor activities, senses to learning, language, emotion, mood and behavior. This powerful, highly sophisticated organization and system of communication and co-ordinated functionality is achieved through networks of interconnected neurons which aid in transmitting signals. The human brain comprises of four lobes: frontal, parietal, temporal and occipital (Figure 3.8). The surface of the brain is referred to as the cortex and there are further subdivisions within each lobe.

Functional brain imaging has proliferated cognitive research, hence enabling us to understand differences between two subject populations, between healthy and abnormal brains, between males and females and so on. Data collection from an fMRI experiment is driven by MR physics (Buxton, 2002). The MR physics component depends on the nature of the psychological stimuli, the particular MR scanner being used, and the location of the expected response. Scanning parameters, such as the number and orientation of slices to be collected, the time of repetition, and others, must be chosen before carrying out an experiment. The physics of the scan will depend on these chosen parameters. The physical method for collecting each slice of data, termed the pulse sequence, must also be selected. The pulse sequence is a small program run in the scanner to manipulate the magnetic and radiofrequency fields, and is

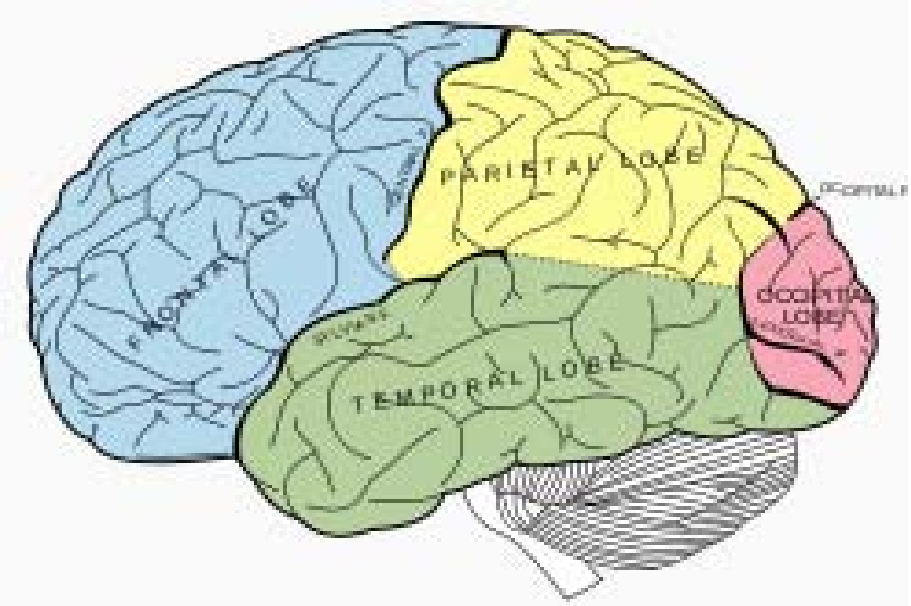

Figure 3.8: The human brain with four lobes (Mysid, 2006).

thus dependent on the type of MR scanner and the pulse sequences available for use. An fMRI data set from a single session can either be thought of as t volumes, one taken every few seconds, or as ν voxels, each with an associated time series of t time points (Jezzard et al., 2001).

For calculation purposes, a voxel is a small square that constitutes only a small fraction of a slice of the brain in the two-dimensional image produced by the scanner (we will consider the height of the slice negligible). Each slice is acquired in a grid form, for example each voxel being *4mm* by *4mm* of data; each voxel represents the strength or intensity of the magnetic resonance signal relating to the changes in brain activity as the task is being performed. The goal of fMRI analysis is to detect, in a robust, sensitive and valid way, those parts of the brain that show increased intensity at the points in time that stimulation was applied.

It is common to run an experiment multiple times, either on the same subject, or with multiple different subjects, or both. This can both increase the sensitivity of the overall

experiment (as more data can lead to increased sensitivity to an effect) or allow the generalization of any conclusions to the whole population. In order to combine statistics across different sessions or subjects, the first necessary step is to align the brain images from all sessions into some common space. This leads to a problem which has two aspects. One is the spatial aspect since different people have different sizes and shapes of the brain and hence the need to standardize them on a common spatial scale in such a way so as to extract maximum information. The standard practice to address this problem is to warp and smooth the brain images of the subjects on to a common atlas such as Talairach coordinates (Talairach and Tournoux, 1988) or The Montreal Neurological Institute (MNI) Broadmann atlas (Evans et al., 1992). However, any such warping leads to tremendous loss in real spatial resolution (Woods, 1996). The second aspect of the problem is statistical and involves methods of combining information from different subjects in an efficient fashion. In order to avoid this problem, historically researchers would take average of each slices over all the subjects. Averaging may not yield a great loss of power in some cases (Cochran, 1954) but some times it may not truly represent the population. This is because the subjects are picked *randomly* and a result from one subject which does not conform with the rest will *skew* the conclusion significantly. Averaging allows for accumulation of evidence arising from multiple subjects who display *similar* patterns of activation.

In many fMRI experiments, it is desirable to directly compare and contrast different conditions occurring in a voxel within or between subjects. Methods for performing such comparisons have been built on the GLM to assess activity in an fMRI study, resulting in a statistical parametric map. As we have already seen, the GLM can refer to a single subject, one group of subjects or multiple groups representing different subjects (for e.g. male, female) or same subjects (for e.g. placebo, drug A, drug B as in cross-over design). Usually multi-subjects is desirable for any statistical analysis. Hence the question arises as to how we can combine the data from the brain scans to conclude that there is activation in the population from which these subjects come from. There are two approaches based on GLM: fixed effects

approach and random effects approach (explained in details in the "Combined Estimation" section of Chapter 4). The shortcoming of the former approach is that it considers the errors of measurement estimated for the subjects as the only source of variation when estimating the population mean. That is to say that only *within-subject* variation is accounted for and there is no consideration of *between-subject*; therefore it is valid for only subjects chosen for a particular experiment (no sampling variation). The latter approach is a valid one, however it is difficult to be confident in as the sample size used is generally small compared to what is necessary in estimation (at least 30). This leads to over-estimation of variances. Hence, conjunction analysis (Friston et al., 1995) or combination methods have been developed for analyzing a single group of subjects.

This thesis is an extension of analysis on statistical parametric group maps using combination methods. It develops statistical techniques for comparison of brain function, as measured by common functional brain imaging procedures, across different groups of experimental subjects. A crucial question that evolves with multi-subject group comparison based on the GLM (which is performed at each voxel for each subject) is: "If we find a significant difference in the two groups then does a truly differentially active voxel come from the same predictor variables across the two groups or from the same location of the activated voxel in the group maps?" In order to make valid inter-subject or inter-group comparisons, we will also assume that the data have been mapped onto a common co-ordinate system. Hence the images from different subjects are comparable and thereby combinable. This thesis considers the spatial issue as having been corrected and hence will focus on addressing the statistical issue only.

Ad-hoc graphical procedures, averaging method and the overly-conservative random effects model are the only resorts for psychologists to understand the differences in brain function between two or more groups of subjects. Several methods have been developed to combine data from independent studies which in our case would be each subject. If there is a reliable effect that can be detected, combining the data from many subjects in a statistically reasonable way will yield a stronger signal. One of the more difficult and interesting problems in

this context is comparing the images obtained through functional brain imaging from two or more different subject populations and deriving conclusions regarding the *intensity, magnitude* and *location* of their regions of activation while performing the same cognitive task. This thesis aims to develop sound statistical methods for such comparisons and will provide an exploratory tool to contrast the "level of activation" and the "proportion of population which shows this activation at a specifically defined level" between two or more groups. That is to say that this thesis will aid researchers in the field of functional neuroimaging to better answer their real questions of interest regarding differences in brain function and activation for different groups of subjects. Although the techniques that will be proposed here have precedence in the neuroimaging literature, the development of efficient, small-sample methods for the comparison of groups of subjects will represent a significant conceptual and practical breakthrough in the analysis and effective utilization of fMRI data. The use of statistical methods that are familiar to researchers in the field, albeit in a new way, will enhance the attractiveness of these methodologies.

Chapter 4

Combination Methods used in Analyzing Single Group Maps

Group maps created from individual functional maps provide useful summaries of patterns of brain activation. In this chapter, we will review a number of ideas suggested in the statistics literature for combining information across studies, where we think of each individual subject as a study. The goal of all these methods is to combine the information acquired from different studies in a statistically relevant and meaningful way so that we can make statements about the population and draw power from accumulation of evidence. As pointed out in Hedges (1992), there are two main approaches to this problem - combining hypothesis tests and combining estimates of treatment effects.

4.1 Combining Hypothesis Tests

Combined maps are useful for examining activation regions that are "typical" of a certain subject population or for comparing average effects of one group of subjects to another (Friston et al., 1999). The group maps that are created from combining maps of individuals in that particular group are maps made up of statistics. These statistical parametric (can be a t-statistic as described in the section "Subtraction Techniques", an F-statistic where we take ratios instead of differences as in t-maps, etc.) or non-parametric (Wilcoxon rank sum, etc.) maps are used for examining differences in brain activity recorded using neuroimaging technologies such as fMRI.

4.1.1 Combination Tests for Single Groups

Throughout this thesis, I will demonstrate combining methods with t-statistic maps calculated for each individual subject as a first step from which we will derive the p-value maps. This is only for the purpose of demonstration since most methods in this section actually combine the p-values. Any statistic (for example F-statistic) used in creating individual maps which will yield p-values is a valid input in the combination techniques.

Suppose we have k independent tests of a null hypothesis of observing any kind of true activation in response to a particular task, with values of a t-test statistic T_1,T_2,....,T_k and corresponding p-values p_1,p_2,.....,p_k. A low p-value would indicate a stronger or a more intense signal. The tests here are one-sided since we are looking for areas of activation. In our context, k is the number of subjects in an fMRI study. Even though the voxels within a subject are dependent, the subjects themselves are still independent. Hence independent information is being combined to form the combination test statistic on each voxel in Talairach space. Since the brain maps of the individual subjects are corrected for spatial disparity, it is sensible to combine them and analyze the resultant group map.

The statistical literature has many reasonable methods for combining independent sources of information. We will describe a few of the simpler combination methods following the discussion in Lazar et al. (2002) relevant to the work addressed here in the thesis.

(i) The most popular is perhaps Fisher's combination method (1950) which uses the test statistic

$$T_F = -2\sum_{i=1}^{k} \log p_i$$

T_F is compared to a χ^2 distribution with *2k* degrees of freedom. The null hypothesis of no activation during the imaging study is rejected at a large value of T_F relative to the tabulated χ^2 distribution. The p-values are one-sided, the null being absence of any activation, and are calculated from the t-statistic at *each voxel* for each of the subject's t-maps. The combined

test statistic used to create the group map is a function of p-values of each subject at each voxel; hence changes in a p-value for any subject will have an impact on the combined test statistic. If an individual p-value is near 1, then the statistic remains unchanged but the degrees of freedom increase by 2. If an individual p-value is near 0, then a small change in the p-value changes the statistic by $\frac{-2}{p_i}$ (Lazar et al., 2002).

(ii) Another p-value combination method put forth by Stouffer et al. (1949) is defined as

$$T_S = \sum_{i=1}^{k} \frac{\Phi^{-1}(1-p_i)}{\sqrt{k}}$$

where Φ^{-1} is the inverse normal cumulative distribution function. The null hypothesis of no activation during the imaging study is rejected at a large value of T_S relative to the tabulated standard normal distribution. If an individual p-value is 1, then the statistic goes to infinity and hence it does not carry any information. If an individual p-value is near 0, then a small change in the p-value changes the statistic by $-[\sqrt{k}\varphi(\Phi^{-1}(1-p_i))]^{-1}]$ (Lazar et al., 2002).

(iii) Mudholkar and George (1979) proposed another combination method which is defined as

$$T_M = -c\sum_{i=1}^{k} \log\left(\frac{p_i}{1-p_i}\right)$$

where $c = \sqrt{3(3k+4)/k\pi^2(5k+2)}$. The null hypothesis of no activation during the imaging study is rejected at a large value of T_M relative to the tabulated t-distribution with 5k+4 degrees of freedom. If an individual p-value is 1, then the statistic goes to infinity and hence it does not carry any information. If an individual p-value is near 0, then a small change in the p-value changes the statistic by $\frac{-c}{p_i} - \frac{c}{(1-p_i)}$ (Lazar et al., 2002).

(iv) A commonly used but ad hoc method of combining the brain maps of all the subjects in a particular study is to average the t-statistics computed for each subject voxel-wise. The combined statistic is defined as

$$T_A = \sum_{i=1}^{k} T_i/\sqrt{k}$$

The null hypothesis of no activation during the imaging study is rejected at a large value of T_A relative to the tabulated standard normal distribution. Under the null hypothesis, T_A is approximately equal to T_S.

These combination tests, with the exception of T_A, are based on combining the p-values; hence the statistics used to derive the p-values need not be similar in any way and can be based on different kind of measurements. This makes these combination tests very appealing and quite general in their applicability. Furthermore, the combination tests suggested by Fisher (1950) and by Mudholkar and George (1979) satisfy an optimality criteria, Bahadur efficiency (Bahadur, 1967, 1971), related to effective use of data as the number of subjects increase. The pitfall of these combination tests is that we cannot obtain any information on the size, direction (from a two-sided test) or consistency of effects (here effect being activation with respect to a particular task) across the different studies. In our single group studies, we will concentrate on one-sided tests since we will be looking at areas of activation while we take less interest in the areas of deactivation. This helps us eliminate one of the drawbacks as we can draw conclusions based on direction (tests being one-sided).

It is to be noted that the hypotheses based on the combination tests are:
H_0: there is no activation in a particular voxel of the brain during an imaging study across all the subjects i.e effect$= 0$.
H_a: there exists activation in a particular voxel of the brain for every subject i.e effect> 0.
Hence the null hypothesis can be rejected on the basis of a non-zero effect in any one of the subjects, which can also be a false positive effect.

4.2 Combining Estimates of Treatment Effects

"Meta-analysis" refers to the statistical technique for amalgamating, summarizing, and reviewing previous quantitative studies that have similar designs and measure the same outcome of interest. By using meta-analysis, a wide variety of questions can be investigated,

as long as a reasonable body of primary research studies exists. Selected parts of the reported results of primary studies are entered into a database, and these "meta-data" are "meta-analyzed", in similar ways to working with other data — descriptively and then inferentially to test certain hypotheses. Meta-analysis allows for inference about size, direction and consistency of effects, unlike the combination methods described in the previous section (Cohen, 1988).

Meta-analysis can be used as a guide to answer the question 'does what we are doing make a difference to X?', even if 'X' has been measured using different instruments across a range of different people. Meta-analysis provides a systematic overview of quantitative research which has examined a particular question. The appeal of meta-analysis is that it in effect combines all of the published and relevant research on one topic into one large study with many participants. However publication bias is a thorny issue in meta-analysis (Cohen, 1988) which may seriously distort attempts to estimate the effect under investigation. Publication bias is the term for what occurs whenever the research that appears in the published literature is systematically unrepresentative of the population of completed studies. Simply put, when the research that is readily available differs in its results from the results of all the research that has been done in an area, readers and reviewers of that research are in danger of drawing the wrong conclusion about what that body of research shows (Rothstein et al., 2005). Also it is difficult to find all the relevant studies needed to conduct a particular meta-analysis. The danger is that in amalgamating a large set of different studies the construct (response variable) definitions can become imprecise and the results difficult to interpret meaningfully.

Not surprisingly, as with any research technique, meta-analysis has its advantages and disadvantages. Advantages lie in derivation and statistical testing of overall factors (overall effectiveness of interventions), generalization to the population of studies, the impact of independent variables and the strength of relationship between variables, ability to control for between-study variation and higher statistical power to detect an effect than in "$n = 1$ sized study sample", and yet like any research, ultimately its value depends on making some

qualitative-type contextualizations and understandings of the objective data where the conclusions are based on exploratory analysis and hence are not quantitative in nature. One erroneous or poorly conducted study can place the results of the entire meta-analysis at risk. On the other hand, setting almost unattainable criteria and criteria for inclusion can leave the meta-study with too small a sample size to be statistically relevant. Striking a balance can be a little tricky, but the whole field is in a state of constant development.

The steps in conducting meta-analysis are:

1. Developing a research question and identifying studies.

2. Selecting studies ("incorporation criteria"): based on quality criteria (e.g. the requirement of randomization and blinding in a clinical trial), selecting specific studies on a well-specified subject (e.g. the treatment of breast cancer) and/or deciding whether unpublished studies are included to avoid publication bias.

3. Deciding which dependent variables or summary measures are allowed: for instance, differences (discrete data), means (continuous data), etc.

4: Selecting a model: fixed effects model or random effects model.

5. Calculating a summary effect and interpreting the results in the light of findings.

The last two steps are relevant to this thesis. The most common meta-analysis models are the one-factor fixed effects and random effects models. Generally, a neuroimaging study with more than one or two subjects will have a place for both types of analysis. For our purpose, a "subject" in an fMRI experiment is a "study" in a meta-analysis. The fixed effect model answers the question whether the studies included in the meta-analysis show that the treatment or exposure produced the effect on average. Methods of fixed effect meta-analysis are based on the mathematical assumption that a single common (or "fixed") effect underlies every study in the meta-analysis. In other words, if we were doing a meta-analysis of odds ratios, we would assume that every study is estimating the same odds ratio. Under this assumption, if every study were infinitely large, every study would yield an identical result. This is the same as assuming there is no (statistical) heterogeneity among the studies

i.e. we are measuring the same effect. The random effects model answers the question, on the basis of the studies that are examined, is it possible to comment that the treatment or the exposure will produce a result? A random effects analysis makes the assumption that individual studies are estimating different treatment effects. In order to make sense of the different effects we assume they have a distribution with some central value and a degree of variability. The idea of a random effects meta-analysis is to learn about this distribution of effects across different studies. By convention most interest is focused on the central value, or mean, of the distribution of effects. It is also important to know the variability of effects. A random effects model is computationally more intense than a fixed effects model.

If important diversity or heterogeneity in the review is identified or suspected, there are several options. One option is that of not performing a meta-analysis. An unwise meta-analysis can lead to highly misleading conclusions. If we have clinical, methodological or statistical heterogeneity it may be better to present as a systematic review using a more qualitative approach to combining results, or to combine studies only for some comparisons or outcomes. The decision should, of course, be made at the question formulation stage (Cochrane Collaboration, 2002). A random effects model is a better approach in case of heterogeneity between the studies in the review. However there is a great deal of debate between statisticians about whether it is better to use a fixed or random effect meta-analysis. The debate is not about whether the underlying assumption of a fixed effect is likely but more about which is the better trade off, stable robust techniques with an unlikely underlying assumption (fixed effect) or less stable, sometimes unpredictable techniques based on a somewhat more likely assumption (random effects). Sometimes the point estimate of the treatment effect differs between fixed and random effects because of publication or quality related bias. This may indicate that careful investigations are required, perhaps with expert methodological input. Whatever statistical model we choose, we have to be confident that clinical and methodological diversity is not so great that we should not be combining studies at all. This is a judgment, based on evidence, about how we think the treatment effect might vary in

different circumstances. This judgment is a common source of disagreement about the results of meta-analyses.

The typical study proceeds with a type of model called the hierarchical model, in which both fixed and random effects are considered. Single-subject analyses are generally carried out with a fixed-effects model, where only the scan-to-scan variance is considered. Those analyses generally yield some type of summary measure of activation, for example a t statistic. Once those summary measures are collected for each subject, then, a random-effects analysis can be performed on the summaries, looking at the variance between effect sizes as a random effect. Again, only a single source of variance is considered at a single time. For the most part, the rule of thumb is: use fixed effects model for single-subject analyses (to leverage the greater power of this approach) and any analysis involving a group of subjects that we would like to express something about the population should be random-effects (Cochrane Collaboration, 2002).

If we can consider that the studies are measuring the same effect and the variances of the subjects are homogeneous, then a suitable statistical model would be

$$y_i = \theta + \epsilon_i$$

where y_i is the effect observed in study i or one subject in an fMRI study, θ is the common mean effect and ϵ_i is the error in the ith study. The errors are independent and normally distributed with mean 0 and variance σ_θ^2.

Under this *fixed effects model*, $\hat{\theta}$ provides an estimate for the common mean θ through a weighted average of the y_i.

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i},$$

Here the weights, w_i, are inversely proportional to the estimated variance of each study i.e. $w_i = 1/\widehat{\sigma}_\theta^2$. The estimate $\hat{\theta}$ is approximately Normally distributed with mean θ and

estimated variance $1/\sum w_i$. We can test whether the common mean, θ, is different from zero or not through a t-type statistic given by

$$T_X = \frac{\widehat{\theta}}{\sqrt{1/\sum w_i}}$$

and rejecting the null hypothesis for large values of T_X relative to the tabulated t-distribution with $k-1$ degrees of freedom. If the errors have a variance of 1, then the fixed effects model is equivalent to Stouffer's method, that is, T_S is the unweighted version of T_X. An extreme subject might be one that shows the same level of activity in both conditions, and almost no variation. For this model, such a subject will have a very large weight and an effect size of 0, so will contribute nothing to the numerator of $\hat{\theta}$, but the denominator will be very large, hence T_X will be close to 0 (Lazar et al., 2002).

The *random effects model* may be used for heterogeneous studies (for example each subject performed different cognitive tasks) or as a result of rejecting homogeneity of the effect θ for the fixed effects model. In any situation, most of the time between subject variability is much greater than within subject variability. Furthermore, as stated earlier in this section, the interest lies in making inference about the hypothetical population at large from which the subjects were drawn rather than restricting conclusions to the sample. Even though we have to compromise on power, the above reasons provide a good base for considering the random effects model over the fixed effects model. The differences between the observed y_i are now assumed to come from experimental error, as before, and from real differences among the studies, so that the true effect in study i is a sample from a "hyperpopulation" of treatment effects. The model has the form

$$\begin{aligned} y_i &= \theta_i + \epsilon_i \\ \theta_i &= \theta + e_i \end{aligned}$$

where the ϵ_i are usually taken to be normal with mean 0 and variance σ_i^2, the e_i are taken to be normal with mean 0 and variance σ_θ^2, and all e_i and ϵ_i are independent. We can use other

distributions of the error terms but the Normal distribribution is not only convenient but also the most widely used. When $\sigma_\theta^2 = 0$, this model reduces to the fixed effects model. Under this *random effects model*, $\hat{\theta}^*$ provides an estimate for θ_i through a weighted average of the y_i:

$$\hat{\theta}^* = \frac{\sum_{i=1}^{k} w_i^* y_i}{\sum_{i=1}^{k} w_i^*},$$

Here the weights, $w_i^* = 1/(s_i^2 + \hat{\sigma}_\theta^2)$ and s_i^2 is an estimate of $\sigma_i^2 = E[(y_i - \theta_i)^2]$. Now there are two sources of uncertainty, whose sum has to estimated; as a result, random effects models are more complicated than fixed effects models. There are different ways of estimating σ_θ^2, one of which was proposed by Hedges (1992)

$$\hat{\sigma}_\theta^2 = S^2 - \frac{\sum s_i^2}{k},$$

where S^2 is the sample variance of $y_1, y_2, ..., y_k$. The drawback of this estimator is that it can be negative, in which case the standard recommendation is to truncate to 0. There are other estimators that do not possess this undesirable property but they are computationally complicated requiring an iterative procedure (for example, Rao and Kleffe, 1988). Any method of estimating the variance components, together with $\hat{\theta}^*$, can be used to build a test for the hypothesis that $\theta = 0$. For the random effects model, an unusual subject such as the one described for the fixed effects will influence the estimate of σ_θ^2, but will otherwise not contribute to the outcome.

The standard errors for the fixed effects estimate, $\hat{\theta}$, tend to be smaller than those for the random effects estimate, $\hat{\theta}^*$ since the latter takes into account the variability across studies. The variance of $\hat{\theta}$ is

$$\frac{1}{\sum_i 1/\sigma_i^2}$$

whereas the variance of $\hat{\theta}^*$ is

$$\frac{1}{\sum_i 1/(\sigma_i^2 + \sigma_\theta^2)}$$

For $\sigma_\theta^2 = 0$, the variances of the above two models coincide, but otherwise, the random effects estimate has a larger variance.

4.3 Multiple Testing in fMRI

Regardless of which specific methods are used to combine functional images across subjects, it is important to adjust for multiplicity and significance level due to the involvement of a large number of voxels when creating a statistical parametric or non-parametric map. In fMRI studies, as stated earlier, data analysis is usually done voxel-wise with all statistical tests conducted separately and simultaneously. These voxel-by-voxel tests increase of the chance that at least one of the conclusions is wrong. Therefore a family of statistical tests suffers one serious problem: type I error rate is greater than that of the error on an individual test. To control this problem, a multiple testing correction (similar to multiple testing correction in the traditional ANOVA sense) is desirable during group analysis. Different approaches to correcting for multiple testing aim at different types of control of this (unknown) number.

A standard quantity to control is *familywise error rate* (FWER), which is the probability of having even one false discovery among all the tests. The familywise approach fixes α for the whole family (brain) of tests. For example in a brain with 10,000 voxels, a fixed type I error of 0.05 would lead to false detection of 500 active voxels on average simply by chance even if the null hypotheses were true everywhere. By significantly lowering the individual type I error, we can achieve control of the total type I error. A corrected type I error of p means that among 100 such voxels on average $100p\%$ of them would have a false detection. The cost of the approach is a loss of power on the individual tests, increasing type II error. The Bonferroni correction (Miller, 1981) is such a familywise multiple-comparison correction used when multiple dependent or independent statistical tests are being performed

simultaneously, but it is overly conservative in the case of fMRI analysis. For example, in fMRI analysis, tests are done on over 100000 voxels in the brain. For $\alpha = 0.05$, the Bonferroni method would require p-values to be smaller than .05/100000 to declare significance. A further complication in neuroimaging studies is that the data are available in two coordinates — the original acquisition space and Talairach space. It would be wrong to calculate the Bonferroni correction on the original data and then apply this criterion to the transformed voxels in Talairach coordinates, or vice versa, because the number of voxels changes when we go from one set of coordinates to the other and the correlation structure of the data is also affected (the switch to Talairach space induces dependence among voxels beyond that already present in the data). The extreme conservativeness of the Bonferonni method, coupled with its inability to take into consideration the particular features of fMRI data (such as switching between the original and smoothed co-ordinate systems), requires other techniques for error control.

There are five thresholding methods that are currently implemented in many fMRI studies: cluster thresholds, in which a contiguous collection of voxels all need to be declared significant at a prespecified level (also the number of voxels in a cluster has to be prespecified) in order for the cluster as a whole to be retained; random field methods, which use the theoretical behavior of random fields to determine deviations from null behavior; thresholds obtained by permutations, in which the theoretical results of the random field theory are replaced by empirical ones; procedures for controlling the false discovery rate (FDR) instead of FWER, which is advantageous in terms of power, ease of use, adaptability, and interpretability; and an ad hoc method, which involves setting the threshold by eye estimation, based on the practitioner's experience and knowledge. In this thesis, we will restrict ourselves to multiplicity adjustment through control of false discovery rate.

4.3.1 Multiple Testing of voxels through control of the False Discovery Rate

Whereas the Bonferroni correction controls for the familywise error rate, other types of error control are possible. In the context of fMRI specifically, it is not reasonable to control the familywise error rate, since scientists care about the overall picture of activation, and not any one particular voxel. The drawbacks of the Bonferroni correction addressed in the previous section contribute to its unsuitability as a multiplicity adjustment. Another correction, of increasing popularity in the recent statistical literature at large, is to control the False Discovery Rate (FDR). This is a rate for the proportion of tests falsely declared significant, out of all tests declared significant (Benjamini and Hochberg, 1995). It quickly became apparent that this powerful, intuitive, and easy to implement procedure would have widespread applicability for the analysis of large datasets. For testing k hypotheses $H_1, H_2, ..., H_k$, the first step is to order the p-values corresponding to the hypotheses from the smallest to largest. The ordered p-values are written as $p_{(1)}, p_{(2)}, ..., p_{(k)}$, with $H_{(i)}$ denoting the null hypothesis with p-value $p_{(i)}$. Let q be the desired false discovery rate, that is, the rate of false discovery that the researcher is willing to tolerate, and let r be the largest i for which

$$p_{(i)} \leq \frac{iq}{kc(k)},$$

where $c(k)$ differs according to the correlation structure of the tests. For independent tests, or when the tests follow a technical condition (positive dependence), $c(k) = 1$. The form of $c(k)$ used to accommodate an arbitrary joint distribution of p-values is

$$c(k) = \sum_{i=1}^{k} 1/i \approx \log k + \gamma,$$

where $\gamma \approx 0.577$, Euler's constant. Then reject $H_1, H_2, ..., H_r$. If no hypotheses are rejected, that is, the criterion defined by the inequality equation is not met for any i, then the false discovery rate is zero. While there is a tendency to set q values similar to standard p-values, such as 0.05, there is in fact no special reason for doing so. We can choose the value of

q depending on the problem. Naturally, researchers tend to want low proportions of false discoveries, preferring that all discoveries be scientifically meaningful and real, but it should be emphasized that this is not an achievable goal. The q value provides a straightforward way of exploring this point together with the scientist.

An interesting aspect of this latter procedure is that it does not require that the test statistics be independent or even of the same kind. Since the voxels in an individual subject's brain are almost certainly not independent, this is a desirable feature. When all null hypotheses are true, the false discovery rate is the same as the familywise error rate, hence the controlling parameter q may be chosen at conventional levels for significance testing. Furthermore, the method is adaptive, in the sense that the chosen thresholds change (become more or less conservative) with the strength of the signal. This would solve the difficult problem of finding thresholds that work for all subjects under all conditions - instead of trying to find such a threshold, which is likely to be arbitrary and ad hoc, the researcher can keep the tolerated level of false discoveries at a constant across subjects and experiments, and the appropriate thresholds will be determined by the data.

Chapter 5

Comparing Two Group Maps using Combination Methods

Extending the use of combination methods to compare two different sample groups and generalize to their respective populations is the main focus of this thesis. The previous chapter explains in detail various methods used for creating statistical maps for a single group of subjects. As discussed before, these group maps combine the information from each subject in a sample to form a summary of the overall pattern of activation in the brain, in response to a particular task. Areas of activation can be identified and we can also quantify the sensitivity of different methods to unusual subjects (those whose level or extent of activation is extremely high or extremely low compared to rest of the subjects in the group; those who exhibit activity in an aberrant location in the brain; and so forth) (McNamee and Lazar, 2004).

A common approach for data analysis in fMRI literature, which is used for everything from analyzing the data from a single subject to comparing the results across groups of subjects, is the *linear model*, and in particular the *random effect model*. However this method has a few drawbacks:

1. Since we are estimating two variances, if we have fewer than approximately 10-12 subjects per group, we will lack sufficient power for a realistic analysis. Since the random effects model uses up more degrees of freedom, it will have much less power for very small groups of subjects.

2. This model will answer the question if all subjects "activate" in the same location and with roughly the same magnitude. That is to say that the effect that the researcher desires to detect should be consistently exhibited across subjects at the same location in the brain

for the random effect model to be implemented.

3. It also assumes that the design is balanced, or the design matrices are identical for each subject. This implies that the length of time and design used should be the same for each subject, particularly the number of scans per subject. The reason for this is that the contribution to the first-level estimate of the variance should be comparable across subjects. However, the actual order of scans can be randomized or changed across subjects (unless we have reason to believe that a particular order invokes a reaction not otherwise seen) without invalidating the "sameness" assumption.

4. It is computationally intensive and estimating the variances is not easily done.

5. It is conservative.

6. It requires means and standard deviations of each condition from each subject; the individual t-maps, for instance, do not suffice for these operations (Beckmann et al., 2003). Hence, both in terms data storage and computation, this procedure is very demanding.

None of the conditions required for the random effects approach are likely to hold in *all* fMRI studies. First, it is often difficult to recruit subjects, particularly from clinical (patient) populations, and to retain them throughout the course of the study. Furthermore, scanning each individual subject is time-consuming and expensive. As a result, many fMRI studies include a relatively small number of subjects. Finding statistically powerful methods for small samples that will be applicable in this context is therefore critical. Due to individual differences, even after the brain images of all subjects have been transformed into a common coordinate system, it is also unrealistically optimistic to expect to find consistent activation at the voxel level. Hence, methods that are more "liberal" in their approach to combining information are required. We want to answer the slightly less stringent question "do the subjects all activate in the same general location?". Several methods at the single group level were introduced in the previous chapter. I will use the techniques already listed and extend them to compare the group maps derived from two different samples. I will explore these through distribution theory or parametric methods as well as through non-parametric

methods, namely, *permutation tests* and *bootstrapping*. In order to compare the group maps through distribution theory, I will either take the ratio or difference of the combined test statistics and look at their distributional form. After comparing the group maps I will use multiplicity corrections when testing at each voxel, so as to minimize false positives and conclude which voxels are active.

Here the inference will be based on the combined statistic of the group maps at each voxel. All we can test in a voxel by voxel analysis is whether the level of activation is the same in both the groups. Since we will not be testing at the level of the image as a whole, we will not be able to draw any conclusion regarding the overall activation "pattern". This analysis will not examine the effect of particular subjects on the group comparison. At each voxel we are testing:

H_0: there is no difference in the activation of the two groups.

H_1: there is a difference in the activation of the two groups.

5.1 Combination Methods using Distribution Theory

In comparing two classes of subjects, taking ratios or differences (perhaps with appropriate modification) of the two group maps using the combination methods introduced in the previous chapter often leads to a known distribution whose results can be easily interpreted. All the methods listed below are calculated at each voxel of the group maps. Here the tests are two sided as we are looking for the effect of activation in one group compared to another in terms of location, size and intensity of the signal.

(i) Using the test statistic from Fisher's combination method (1950), we get two group maps

$$\begin{aligned} T_{F1} &= -2\sum_{i=1}^{k_1} \log p_{1i} \\ T_{F2} &= -2\sum_{i=1}^{k_2} \log p_{2i} \end{aligned}$$

p_{1i} and p_{2i} are p-values derived from the t-test statistic group maps of two samples with k_1 and k_2 number of subjects respectively. T_{F1} and T_{F2} compare to χ^2 distribution with $2k_1$ and $2k_2$ degrees of freedom respectively. These two test statistics can be compared through a known distribution, namely the F-distribution, with degrees of freedom related to the sizes of the two group maps. The null hypothesis of no difference between the two groups is rejected at a large value of the above calculated F relative to the tabulated F distribution. More precisely, the F-test is usually performed to only look at the upper tail, but in our case a small p-value resulting from taking the ratio of two χ^2 could mean that the group in the denominator shows more activation than the group in the numerator at a particular voxel. But it could also be due to random noise which can persist even after noise correction. Hence we need to perform this test twice for each comparison - once with the "first" group in the numerator, and once with the "second" group in the numerator.

(ii) An addendum to Fisher's method can be used to analyze two group maps. The drawback of the F-test is that it will not be able to pick up the tail probabilities of the group in the denominator or to separate effect from random noise. So in order to analyze the effect of one group relative to the other, we need to calculate two F ratios thereby doubling the number of tests. The proposed modification addresses that drawback, however, it works *only* for two equal sample sizes. After we combine the data into two group maps from two samples of equal size k, we have

$$T_{F1} \sim \chi^2_{2k}$$

$$T_{F2} \sim \chi^2_{2k}$$

Define $R = \frac{T_{F1}}{T_{F2}}$, and

$V = \max\left(\log(1+R), \log(1+1/R)\right) - \log(2k)$; then

$V + \log(2k) \sim$ Beta prime distribution or Beta distribution of the second kind (Johnson, 1995) with location parameter k and shape parameter k.

Also, $V + \log(2k) \sim$ *log-F* distribution (Jones, 2006) where $R \sim F(2k, 2k)$.

(iii) Using the test statistic from the p-value combination method that was put forth by Stouffer et al. (1949), we get two group maps

$$\begin{aligned} T_{S1} &= \sum_{i=1}^{k_1} \frac{\Phi^{-1}(1-p_{1i})}{\sqrt{k_1}} \\ T_{S2} &= \sum_{i=1}^{k_2} \frac{\Phi^{-1}(1-p_{2i})}{\sqrt{k_2}} \end{aligned}$$

T_{S1} and T_{S2} both follow a standard Normal distribution. Therefore we can compare them through their difference, which will follow a Normal distribution with mean 0 and variance 2. The null hypothesis of no difference between the two groups is rejected if there is a big departure from the mean in either tail of the above calculated Normal relative to the tabulated *Normal(0,2)* distribution.

(iv) Using the test statistic of Mudholkar and George (1979) we get

$$\begin{aligned} T_{M1} &= -c_1 \sum_{i=1}^{k_1} \log\left(\frac{p_{1i}}{1-p_{1i}}\right) \\ T_{M2} &= -c_2 \sum_{i=1}^{k_2} \log\left(\frac{p_{2i}}{1-p_{2i}}\right) \end{aligned}$$

where $c_1 = \sqrt{3(3k_1+4)/k_1\pi^2(5k_1+2)}$ and $c_2 = \sqrt{3(3k_2+4)/k_2\pi^2(5k_2+2)}$. T_{M1} and T_{M2} follow t distributions with $5k_1+4$ and $5k_2+4$ degrees of freedom respectively. Press (1969) and Garthwaite and Crawford (2004) considered ratios and differences of t-distributions, respectively. However in neither case is there a closed form for the resultant distribution; they can be studied using non-parametric techniques. In this thesis, I have compared the results from the test statistics from distribution theory derived to compare the group maps with the non-parametric methods. I have not looked at any test statistic which does not have a closed form and I will consider it as part of my future work. In addition, there are a few well known transformations to the t distribution and specifically, Chu's transformation

(Bailey, 1980) leads to upper bounds and approximate values for tail probabilities that are reasonably accurate for degrees of freedom as low as two. These two test statistics, T_{M1} and T_{M2}, can be compared by applying Chu's transformation and then taking the difference of the transformed maps. This difference follows a *Normal* distribution with mean 0 and variance 2.

Chu's transformation is given as

$$\begin{aligned} Z_1 &= \pm(k_1 - 0.5)\log(1 + \frac{X_1^2}{k_1}) \\ Z_2 &= \pm(k_2 - 0.5)\log(1 + \frac{X_2^2}{k_2}) \end{aligned}$$

where $X_1 \sim t_{k_1}$ and $X_2 \sim t_{k_2}$.
Then $Z_1 \sim Normal(0,1)$ and $Z_2 \sim Normal(0,1)$.
The null hypothesis of no difference between the two groups is rejected if there is a big departure from the mean in either tail of the above calculated *Normal* relative to the tabulated *Normal(0,2)* distribution.

(v) A commonly used but ad hoc method of combining the brain maps of all the subjects in a particular study is to average the t-statistics from the t-maps computed for each subject voxel-wise. The combined statistic is defined as

$$\begin{aligned} T_{A1} &= \sum_{i=1}^{k_1} T_{1i}/\sqrt{k_1} \\ T_{A2} &= \sum_{i=1}^{k_2} T_{2i}/\sqrt{k_2} \end{aligned}$$

T_{1i} and T_{2i} are t-values from the t-test statistic group maps of two samples with k_1 and k_2 number of subjects respectively. T_{A1} and T_{A2} both follow a standard Normal distribution and hence they can also be compared through their difference which will follow a *Normal* distribution with mean 0 and variance 2. The null hypothesis of no difference between the

two groups is rejected if there is a big departure from the mean in either tail of the above calculated *Normal* relative to the tabulated *Normal(0,2)* distribution.

(vi) Another technique popularized by Haar-Fisz (Fryzlewicz et al., 2006) uses the basics of Fisher's combination method. After we combine the data into two group maps from two samples with sizes k_1 and k_2, we have $T_{F1} \sim \chi^2_{2k_1}$ and $T_{F2} \sim \chi^2_{2k_2}$
Define,

$$R = \frac{T_{F1} - T_{F2}}{T_{F1} + T_{F2}}$$

Then, $R \sim 2Y - 1$, where $Y \sim \text{Beta}(k_1, k_2)$. Here the null hypothesis of no difference of two groups as a proportion of the total effect from the two groups is rejected at a large value of calculated R relative to tabulated $2\text{Beta}(k_1, k_2) - 1$.

5.2 Combination Methods using Non-parametric Methods

Limiting ourselves to conclusions derived from tests related to distribution theory takes away the flexibility of interpreting results. Furthermore distribution theory restricts us from performing extensive exploratory analysis and looking at the data from various angles using any type of statistic. Employing a different mathematical operation, to compare the two groups, without the restrictions imposed by distribution theory, may give better results as well as an easy interpretation. The non-parametric methods implemented here are conceptually simple, rely only on minimal assumptions, deal with the multiple testing issue, and can be used when the assumptions of a parametric approach are untenable. These are some of the motivations for using non-parametric methods. Here we will explore the use of computationally intensive statistical methods, permutation tests and bootstrapping, to address the question of comparing fMRI group maps across groups of subjects without taking into consideration typical assumptions underlying a parametric statistical analysis.

The overly conservative nature of the Bonferroni correction was noted by Blair and Karniski (1994). As an alternative they proposed the permutation test. The permutation test is one

of a family of methods known collectively as resampling procedures (Efron, 1982). Taking advantage of the speed of modern computing systems, these methods construct an explicit, non-parametric model of the actual distribution from which a set of observations has been drawn. The reasoning behind the non-parametric methods implemented in this thesis can be developed from an examination of more traditional tests. For example, we consider a single subject fMRI experiment, where a single subject is scanned repeatedly under "rest" and "activation" conditions. Here a parametric method of fMRI data analysis with null hypothesis being no difference between the two conditions is measured through, for example, the t-statistic. If the null is true (and assuming that the effect of autocorrelation is negligible), the two conditions are interchangeable; any ordering of the conditions will produce similar results.

The permutation test uses such re-orderings (or "resamplings") of the conditions to construct an empirical estimate of the distribution from which the test statistic has been drawn. All the conditions (here the subjects) are pooled together and re-assigned labels as "control" and "patients" at random. On each of a large number of iterations, the sequence of the observations is randomized, and the test statistic is calculated with respect to the data in this randomized sequence. Each iteration produces one point in the empirical distribution. The probability that the test statistic will be less than or equal to a certain value k under the null hypothesis can then be computed as the rank of k within the empirical distribution, divided by the number of points in the distribution (Blair and Karniski, 1994).

Another resampling technique implemented here is bootstrapping where the re-assignment of labels is done with replacement. In bootstrapping, the random sampling is done *with replacement* from each of the two groups and we do not implement "prior mixing" of the samples as in permutation tests. However there are some differences between the two resampling techniques. A permutation test exploits special symmetry that exists under the null hypothesis to create a permutation distribution of the test statistic. For example, in comparing two groups, when testing whether they come from the same distribution, all permutations

of the order statistic of the combined sample are equally probable. As a result of this symmetry, the actual significance level ($Prob_{H_0}(t_{observed} >= t_{tabulated})$) from a permutation test is exact: in our case of the two sample problem, it is the exact probability of obtaining a test statistic as extreme as the one observed, having fixed the data values of the combined sample. In contrast, the bootstrap explicitly estimates the probability mechanism under the null hypothesis, and then samples from it to estimate the actual significance level. This estimate has no interpretation as an exact probability, but like all bootstrap estimates is only guaranteed to be accurate as the bootstrap resamples goes to infinity. On the other hand, the bootstrap hypothesis does not require special symmetry that is required for a permutation test, and so can be applied much more generally. For example, in our two sample case, the permutation test can only test for equality in two population distributions while the bootstrap can test equality of means and equality of variances, or equality of means with possibly unequal variances (Efron and Tibshirani, 1993). The bootstrap method shows superior performance over permutation method for fMRI data analysis when the statistic in question is a simple linear "smooth" statistic such as the mean (Moonen and Bandetti, 1999). However we will use bootstrapping for thresholding purpose too.

Given the assumptions used for the data permutation/bootstrapping approach, it can only be applied to some data sets. It is not always practical to randomly allocate experimental conditions and many times the data at hand are not randomized. For example, we can not randomly assign subjects to be patients or controls. Hence in such instances where there is no explicit randomization of conditions to scans, it is imperative to make weak distributional assumptions to justify bootstrapping or permuting the labels on the data. Usually, the assumption that is made is that the samples drawn for the two groups come from populations whose distributions have the same shape or are symmetric. In general practice, one of the two cases in which permutation/bootstrapping can be used is across-subject (independent data) which we have been using here to combine our datasets. Map-wise

threshold provided by the discussed two non-parametric methods will always be less than or equal to that provided by FDR theory or Bonferroni correction. The additional benefits are:

1. Non-parametric methods do not require assumptions regarding the smoothness of the data or their stationarity. These assumptions are required for FDR approaches, and are often invalid for data sets with limited degrees of freedom. Therefore, non-parametric approaches are "robust".

2. Methods to account for cluster-size in map-wise thresholds are not available for low-degree of freedom t or F maps. Non-parametric approaches can use cluster size as a criterion without resort to additional assumptions.

3. Analyses conducted with low degrees of freedom (such as across-subject random effect analyses) suffer from poor accuracy in specifying the error. Improved error estimation can be obtained by performing local spatial averaging of the error term across voxels, creating a pseudo t-map (instead of using the square root of the estimated variance of the change in BOLD contrast signals as the denominator while performing the t-statistics. This suggests pooling the variance estimate at a voxel with those of its neighbors to give a locally pooled variance estimate as a better estimate of the actual variance. The model has the same form at all voxels, hence the variance estimates have the same degrees of freedom at all voxels, and the locally pooled variance estimate is simply the average of the variance estimates in the neighborhood of the voxel in question.) Solutions do not exist within distribution theory to asses the map-wise significance of pseudo t-maps or any statistic that does not conform to any distribution theory. Non-parametric approaches can handle this variance smoothing approach without additional assumptions.

We will apply permutation tests and bootstrapping to all the combination methods listed in Section 6.2. In this thesis, we calculate a two sample t-statistic at each voxel; two samples being "rest" and "activation" conditions when a single subject performs a particular task. Hence we create t-maps for the individual subjects in both the groups. Using the combination

method, we get a combined test statistic and then using distributional theory, we get a test statistic to compare the groups at each voxel. Under the null hypothesis of no difference between the two groups, it should not matter if a subject belongs to one group or the other. Since we assume independence between the subjects, we can interchange subject labels. We re-assign the subjects randomly into two groups and calculate the test statistic used to compare the two groups. We repeat the resampling and re-allocation procedure 5000 times to get 5000 such test statistics that compare the two groups at each voxel. Hence at each voxel, we get an empirical distribution from those test statistics and can calculate the p-value for the original test statistic based on this permutation or bootstrapped based empirical distribution. If the omnibus null hypothesis is true (and assuming that the effect of autocorrelation is negligible), the labels on the subjects are interchangeable with respect to the voxel statistic under consideration, then the labels are exchangeable with respect to any statistic summarizing the voxel statistics or any statistic summarizing the subjects in each group at each voxel. Hence given the empirical distribution, voxels with statistics exceeding the threshold provided by the original statistic exhibit evidence against the voxel's null hypothesis. The irregularities of the observed data are maintained in the permuted data sets and are included in the estimation of permuted probability.

Chapter 6

A Data example

6.1 Description of an fMRI Experiment

The University of Georgia Psychology Department permitted us to statistically evaluate and analyze fMRI data collected from 15 healthy subjects (controls), 16 schizophrenia patients and 13 relatives of the patients. Data were collected in the scanner while the subjects performed antisaccade tasks. To illustrate an antisaccade task, we will consider a block design constituting alternate periods of fixation and periods of task. A study of antisaccade (a saccade is a rapid eye movement) performance requires the subject to fixate on a central target; a novel visual stimulus appears unpredictably in the left or right periphery, and the task is to inhibit a reflexive saccade, in response to the stimulus, toward the target and then to make a voluntary saccade to the opposite periphery. Failure to inhibit a reflexive saccade toward the peripheral target is considered an error. Analysis focused on p-value maps derived from t-statistic maps, comparing activation during time periods of performing the task versus time periods of rest (in this case, fixation).

Thirty-eight axial slices for each subject were collected from the bottom of the brain to the top, each composed of a field of view of 24cm by 24cm. Stimulus was presented and taken away at regular alternating intervals (22.4 seconds of fixation alternating with 22.4 seconds of antisaccades) in blocks of 9 time points. This paradigm was repeated 9 times. All of the data were placed on to the common Talairach atlas. Each slice is 40x48 voxels, and the time course for each voxel is 81 TRs long. Figures 6.1, 6.2, 6.3 and 6.4 give a general idea about the data matrix and the image pattern indicating fluctuations in brain activation due to the presence or absence of the stimulus.

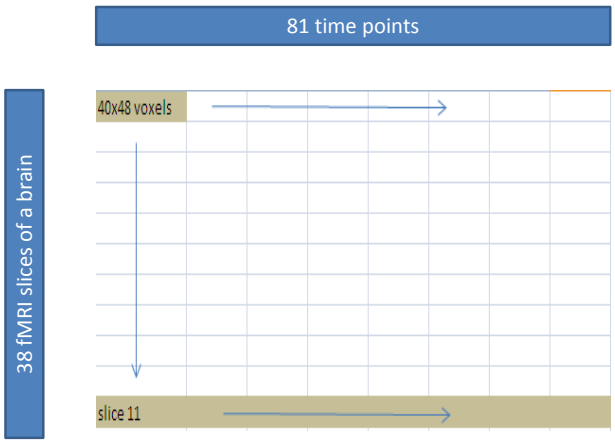

Figure 6.1: Layout of the acquisition matrix for each subject. Each square represents 40x48 voxels in the data matrix.

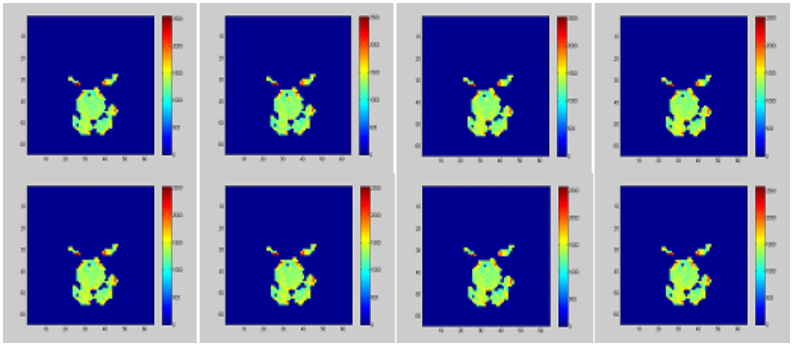

Figure 6.2: Image of activation of the voxels across 8 time points for the first slice for the first control subject at every tenth time point. Highest voxel values are exhibited through brightest red colors.

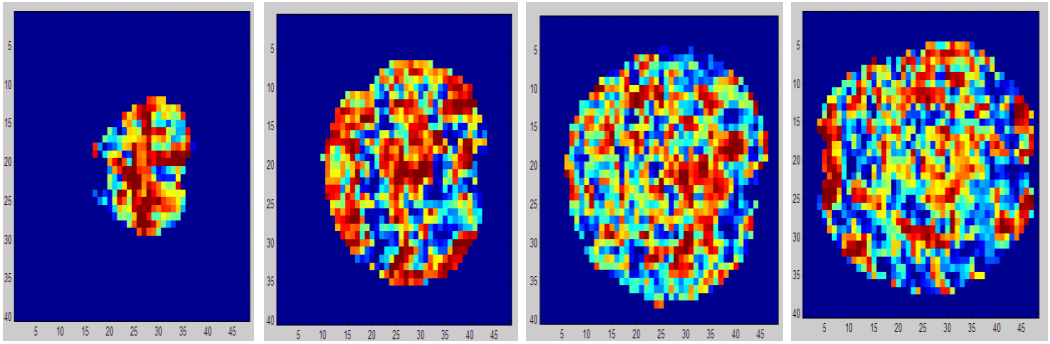

Figure 6.3: Image of activation of the voxels for the fifth, tenth, fifteenth and twentieth slices for the first control subject. The images shown spans nearly the whole brain: the bottom of the brain, near the brain stem, to the top of the brain.

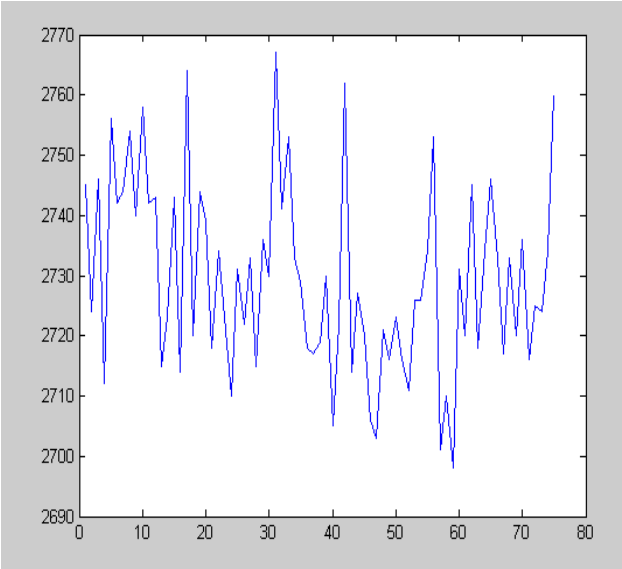

Figure 6.4: When graphed over time, a particular voxel increases or decreases its value based on the response triggered by the stimulus. This figure illustrates a typical voxel, 32^{nd} by 32^{nd} voxel of the tenth slice for the first control subject across all the time points is indicative of the wavelength function of activation with alternating presence of stimulus.

6.2 Comparison of Combining Methods through Distribution Theory

In this section we present the results of six methods for combining information on the data from 15 control subjects (Group A) and 16 schizophrenic patients (Group B) and analysing differences between these two groups. We examine the behavior of the following procedures:

- Fisher's method;
- Stouffer's method;
- Mudholkar and George's method;
- Average t method;
- Haar-Fisz method;

- random effects model;

In this comparison we ignore the time series nature of the data by collapsing the measurements at each voxel into a t-statistic, which averages over time assuming independence. A slice from the middle of the brain was chosen (Slice 23) to represent each technique, the same slice for all cases. Here we are calculating the t-statistic at each voxel for each subject where the numerator is the difference of the BOLD signal change. This difference is the product of the baseline MR signal intensity and the percentage BOLD signal change, hence it is imperative that we choose the slice showing most contrast. In other words, slices of interest are chosen according to where the task related activation should be the greatest. Not all the slices of the brain contain voxels that respond to the particular task; hence in order to perform any analysis we choose a slice where we can actually see activation. Slice 23 seems to be one such slice.

6.2.1 Results

First, we calculated t-maps (two sample t-test based on alternating periods of activation and periods of rest) and subsequently p-value maps (we assume that the tests are two-sided since we are looking for areas of activation), on the individual subjects in each of the groups for slice 23. Plotting the p-value maps for each individual showed that only about 990 out of the 1920 voxels in the slice represent brain; the rest are air, and these are often clipped out, however we do not do this here. It has been noted that the air voxels serve a very useful (statistical) purpose (Lazar et al., 2002). After thresholding, if there is much apparent activation detected outside the brain, in the air, then the threshold is set very high (since small p-values are significant) because it is permitting too many false discoveries (outside the brain should all be in the null hypothesis by definition). If there is little or no activation outside the brain after thresholding then greater confidence is placed in the analysis techniques.

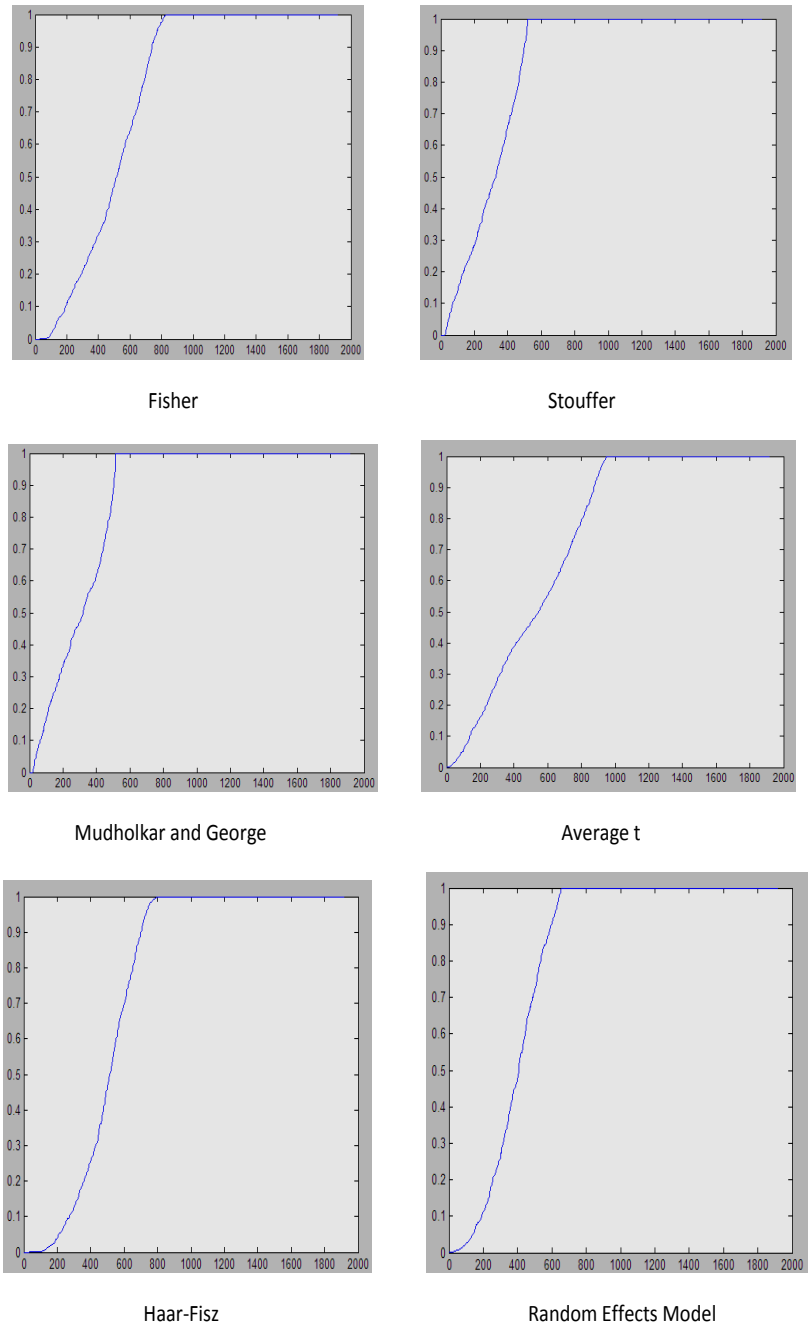

Figure 6.5: Ordered p-values from the distributions of the different comparison methods, derived from the combination techniques in Section 6.2, used to compare two groups.

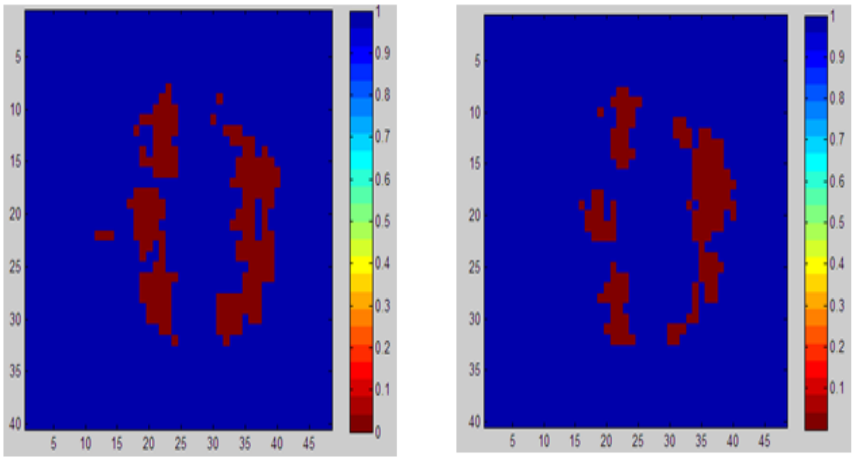

Figure 6.6: Chi-square maps of control and patient groups derived using Fisher's combination.

From the individual p-value maps, we created group maps for controls and patients (each group separately) using the combination methods described in Section 6.2. Figure 6.6 shows examples of group maps using Fisher's combination technique. Then using distribution theory as described earlier, we performed the tests to compare the two groups at each voxel. Here we performed two-sided tests. The largest p-values in the upper right corner of Figure 6.5 most likely come from those voxels outside the brain i.e. the air voxels. Small values, at the bottom left corner of each plot, correspond to the voxels for which the null hypothesis of no difference between the two groups should be rejected. Now arises the question of thresholding to decide that the given voxel is differentially active in the presence of multiplicity. The question becomes complicated if we are interested in comparing different methods because it is further evident from inspection of the graphs that this can be done in at least two ways in addition to the various more complex methods stated in Section 4.3. We can either pick a significance level and all the voxels below that line will be declared active (equivalent to doing Bonferroni multiplicity adjustment) or we can pick some number of voxels and all

voxels to the left of the line would be considered active (this is similar to using false discovery rate procedure).

There are a number of characteristics that can be observed from Figure 6.5. The curves representing comparative methods for average t and random effects model are similar to each other but different from the rest which are similar in appearance among themselves. The methods differ in how long the initial plateau near zero lasts and how quickly the value 1 is reached. They have similar qualitative properties.

Table 6.1: The table displays the p-value of the r_{th} ordered voxel for each of six methods and various values of r. For instance, using Fisher's method to combine and compare the two groups of subjects, the 48th most significant voxel has a p-value of approximately 0, while using Stouffer's method, the 48th most significant voxel has a p-value of 0.0691.

Method with rth value	**Fisher**	**Stouffer**	**Mudholkar-George**	**Average t**	**Haar-Fisz**	**Random Effect**
48	0	0.05	0.0691	0	0.0000087	0.02129
96	0.0000002	0.1333	0.1588	0.00001695	0.0145	0.04081
192	0.000526	0.2737	0.3221	0.0108	0.1006	0.0884
384	0.0877	0.6171	0.5927	0.4043	0.3004	0.18013
768	1.0	1.0	1.0	0.2043	0.5	0.40175

Table 6.1 shows the significance levels corresponding to designating a fixed number of voxels to be active, with that number increasing from 48 voxels, or 2.5 percent of all voxels, to 768, or 40 percent of all voxels. For all methods, the most significant voxels are very highly significant, with p-values near 0. The effect of averaging in the Average-t method is to smooth out any signal that is not too strong and consistent across subjects. All the methods can pick up patterns of activity that are manifested by some, but not all, subjects.

Figure 6.7 displays each of the comparison methods (the test statistics derived from comparing the two groups using the combination methods) and the areas of activation that are found in the brain thresholded at $\alpha = 0.01$, corrected for multiplicity using FDR. The images produced by Fisher, Stouffer and Mudholkar-George are quite close to each other but they are not very informative regarding the existence of any difference between the two groups.

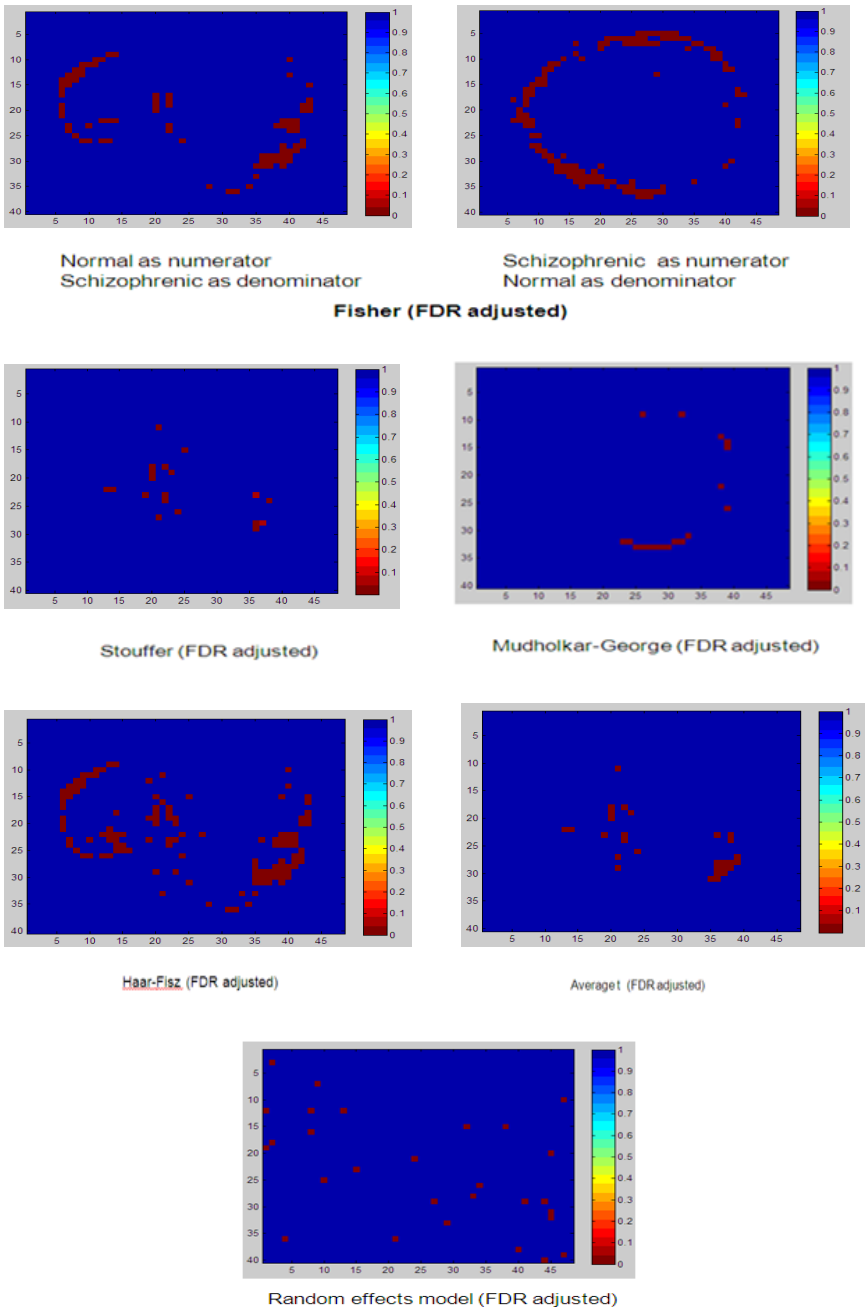

Figure 6.7: Thresholded two group comparative maps using all combining methods, for a threshold corresponding to the 1920 voxels. The top rows shows Fisher with controls as numerator and schizophrenic patients as denominator and vice-versa; the second row shows Stouffer and Mudholkar and George; the third row shows Haar-Fisz and Average t; the fourth row shows the random effects model.

Edge effect exists in all the images. The Haar-Fisz image seems to have more activation that the others.

Table 6.2 presents the second way of comparing the procedures used for comparing the two group maps, that is, the numbers of voxels that would be picked up by each of the methods, for a given significance level. This is equivalent to drawing a horizontal line across each of the plots in Figure 6.5 at the same height. The first column in the table gives the significance level that was applied to the individual voxel tests based on false discovery rate procedure over 1920 brain voxels in slice 23.

Table 6.2: The table displays the number of voxels declared differentially active for each of the six methods and various significance levels. For instance, using Fisher's method of combining data and then taking the ratio of the two group maps, and a significance level of 0.05, 197 voxels will be declared differentially active, whereas by using the Average-t method, 57 voxels will be picked out as differentially active. This has been corrected for multiple testing using FDR.

q-value FDR adjusted	**Fisher**	**Stouffer**	**Mudholkar-George**	**Average t**	**Haar-Fisz**	**Random Effect**
0.05	197	43	48	57	173	105
0.01	132	29	28	32	125	20

Based on Table 6.2, Fisher and Haar-Fisz methods can be grouped together and Stouffer and Mudholkar-George can be grouped together in terms of picking out nearly the same number of "differentially active" voxels. We can see the clear distinction between the ratio based comparison methods versus difference based procedures.

6.3 Comparison of Combining Methods through Non-parametric Methods

6.3.1 Using Permutation Tests

In contrast to the theoretical approach, which relies on various rather strong assumptions such as having images that are smooth, and uniformly so, as well as the need to make multiplicity adjustments, the permutation method relies solely on exchangeability under the

null hypothesis. The main motivation for using the permutation approach in this thesis is to study the behavior of the test statistics derived from the comparative methods, i.e. to understand the empirical distributions under the null hypothesis that there is no difference between the two groups. I will conduct a comparative study of the non-parametric methods with FDR.

The basic idea, even in the neurological context, is a familiar one. If there is no difference between experimental conditions (that is, under the null hypothesis), then the labels "controls" and "patients" can be thought of as arbitrary in the sense that any observation arising from the "patients" group could just as readily have been an observation from the control group, and vice versa. Thus, in order to assess the significance of the difference actually observed in the data at a particular voxel, one can create an empirical distribution by permuting the labels among the subjects in the two groups. For each such permutation the relevant test statistic is computed, and the observed value of the statistic is then compared to the permutation distribution.

Our main interest lies in finding out whether there is a functional anatomical difference between two groups of subjects. So we resort to consideration of the null hypothesis that there is no difference between groups. Two related questions then have to be addressed. What measure(s) of difference between two groups are likely to be most informative about departure from the null hypothesis? How can we ascertain the distribution of a potential test statistic under the null hypothesis?

These two questions leads us towards multiple testing. To obtain an adjusted threshold, all voxels need to be considered simultaneously, that is, the permutations need to be carried out at the level of images. Ideally, we would like to look at the level of entire images, or perhaps images dissected into regions. In relation to the first of these questions, Nichols and Holmes (2001) describe a way of doing this through the use of the *maximal statistic*. Essentially, the maximal statistic is as the name implies, the maximum voxel value in an image which acts as

the summary of the image as a whole. Nichols and Holmes considered two types of threshold: *a single threshold test* and *suprathreshold cluster test*. The former thresholds data at a global or overall level through computing the permutation distribution of the maximal voxel statistic over the volume of interest i.e. it is a measure of the difference between groups in volume of a given region of interest, while the latter thresholds at a cluster level through computing the permutation distribution of the sum of maximal voxel statistics in a cluster or a group of clusters. Though both tests consider the entire image volume, we will not apply either the global test or the cluster level test because in this thesis we do not have apriori knowledge of any areas of interest or contiguous voxels of activation (clusters). Also in the above two cases, functionally distinct regions may be merged and activations may be missed, the threshold must be chosen in advance, the choice of cluster-forming threshold is arbitrary and trying several thresholds increases the number of performed tests, cluster size tests favor big regions over small regions and activated voxels are not localized within detected clusters. We will test voxel-by-voxel to pick out differences between the two groups assuming independence among subjects. This voxel-by-voxel test has good localization power but relatively poor sensitivity to the cluster level test. However it serves our purpose of detecting the difference in two groups.

Multiple Testing using Permutation Tests

After combining the subjects from each group using any of the combination statistics, we have devised tests for comparing the groups. Hence at each voxel we have produced a p-value related to that particular statistic. Let that p-value be p^k, for the null hypothesis H_0^k, where the superscript k indexes the voxels. In these group comparison maps we do not have any a priori knowledge as to which voxels will exhibit differential activation to a particular task. If we would have known that, we could have simply tested at that voxel using an appropriate level α test. Hence arises the problem of multiplicity of testing. If we consider $\alpha = 0.05$, clearly, 5% of the voxels are expected to have p-values less than $\alpha = 0.05$. However these p-values are uncorrected p-values since we require a test procedure maintaining strong

control over map-wise Type I error. An intuititve and easily implemented solution is through non-parametric resampling (Westfall and Young, 1993).

Frequently, non-parametric approaches are less powerful than their parametric counterparts when the assumptions for the latter are true. However in the context of assessing statistical images from fMRI, permutation methods perform at least as well as parametric methods on real data (Arndt et al., 1996). For noisy statistic images, such as t-statistic images with low degrees of freedom, the ability to consider pseudo t-statistics constructed with locally pooled (smoothed) variance estimates affords the permutation approach with some additional power (Holmes, 1994).

There are, however, additional considerations when using the non-parametric approach with a maximal statistic to account for multiple testing. For the single threshold test to be equally sensitive at all voxels, the (null) sampling distribution of the chosen statistic should be similar across voxels. For instance, the simple mean difference statistic could be considered as a voxel statistic, but areas where the mean difference is highly variable will dominate the permutation distribution for the maximal statistic. The test will still be valid, but will be less sensitive at those voxels with lower variability. So, for an individual voxel a permutation test on group mean differences is equivalent to one using a two-sample t-statistic (Edgington, 1995).

Results

Assumptions:

For a valid permutation test the only assumptions required are those to justify permuting the labels. Clearly the experimental design, model, statistic and permutations must also be appropriate for the question of interest. For a randomization test the probabilistic justification follows directly from the initial randomization of condition labels to scans. In the

absence of an initial randomization, permutation of the labels can be justified via weak distributional assumptions. Thus, only minimal assumptions are required for a valid test. In contrast to parametric approaches where the statistic must have a known null distributional form, the permutation approach is free to consider any statistic summarizing evidence for the effect of interest at each voxel. The consideration of the maximal statistic over the volume of interest then deals with the multiple testing problem.

The key steps in performing a permutation analysis:

1. Null Hypothesis: Specify the null hypothesis.

2. Exchangeability: Specify exchangeability of observations under the null hypothesis.

3. Statistic: Specify the statistic of interest, usually broken down into specifying a voxel-level statistic deduced from comparison methods using the combination methods.

4. Relabeling: Determine all possible relabelings given the exchangeability scheme under the null hypothesis.

5. Permutation Distribution: Calculate the value of the statistic at each voxel for each relabeling, building the permutation distribution voxel-by-voxel.

6. Significance: Use the permutation distribution and the critical value derived from the comparative test statistic from the original data, thresholding for the statistical image is done at voxel-level.

Null hypothesis:

The permutation test considers the data to be a random realization from some distribution, which is the same approach used in a parametric test (except that a particular parametric distribution is specified). Here the data are fixed and we use the randomness of the experimental design to perform the test. Although the machinery of the permutation and randomization tests are the same, the assumptions and scope of inference differ.

Each subject has an image expressing activation to a particular task, the difference of the task and the rest condition estimates. We make the weak distributional assumption that the

values of the subject difference images at any given voxel (across subjects) are drawn from a symmetric distribution (the distribution may be different at different voxels, provided it is symmetric). The null hypothesis is that these distributions are centered on zero.

Exchangeability:

The conventional assumption of independent subjects implies exchangeability. We consider subject labels of +1 and -1, indicating unflipped or flipped labels of the data. Under the null hypothesis, we have data symmetric about zero, and hence for a particular subject the label of the observed data can be flipped without altering its distribution. With exchangeable subjects, we can flip the labels of any or all subjects' data and the joint distribution of all of the data will remain unchanged.

Statistic:

In this example we use a permutation thresholding test.

Voxel-level summary statistic:

We are interested in searching over the whole brain for significant changes. We will use the comparative methods described in Section 6.2. This will enable us to compare parametric and non-parametric methods. However we can use any statistic to look at the relationship between the two groups.

Relabeling enumeration:

Based on our exchangeability under the null hypothesis, we can flip the sign on all of our subjects' data. We chose 10,00 random permutations of a total of 31 subjects from two groups, randomly re-assigned their labels on 31 subjects and subdivided them into two groups of 15 controls and 16 patients.

Permutation distribution:

For each of the 10,000 relabelings, we computed the comparative test statistics at each voxel. So for each permutation or relabeling, we have a comparative test statistics map for two groups and draw our conclusion based on the value of the comparative statistics from the

original data as the critical value. The test is made at each voxel based on the permutation distribution. Figure 6.8 shows the permutation distributions for each comparison method, based on the 10000 relabelings.

Significance threshold:

P-values are calculated in order to declare a voxel as "active" based on the permutation distribution and the critical value from the original comparative maps. The empirical distribution and the critical values are calculated at each voxel and hence the thresholding is done at the voxel-level; hence a p-value is assigned at each voxel for the comparative methods.

The overall permutation distribution from all voxels of the various comparison methods for comparing two groups under H_0 is shown in Figure 6.8. It is overlaid with the "true" theoretical distribution for that particular comparison method.

Both theoretical and permutation tests were used to assess the probability of each observed voxel statistic under the null hypothesis, and the observed effects maps were thresholded at a size of a two-tailed tests with $\alpha = 0.01$. Since no true difference is expected to exist between the two groups under the null distribution, all significant voxels identified by the size of the two-tailed tests at 0.01 should be false positives. In other words, the observed number of positive tests (significant voxels) should equal the predicted number of false positive tests. Figure 6.8 shows that for all the comparison methods, the number of positive tests observed by the theoretical distribution is generally less than the expected number given by the permutation distribution, implying that permutation tests may be more efficient in finding false positives. We can also say that the theoretical distribution overestimates the probability of positives. Correspondence between observed and expected positive tests was closer for comparative methods using Fisher's and Haar-Fisz's combination techniques than for the other combination techniques using differences.

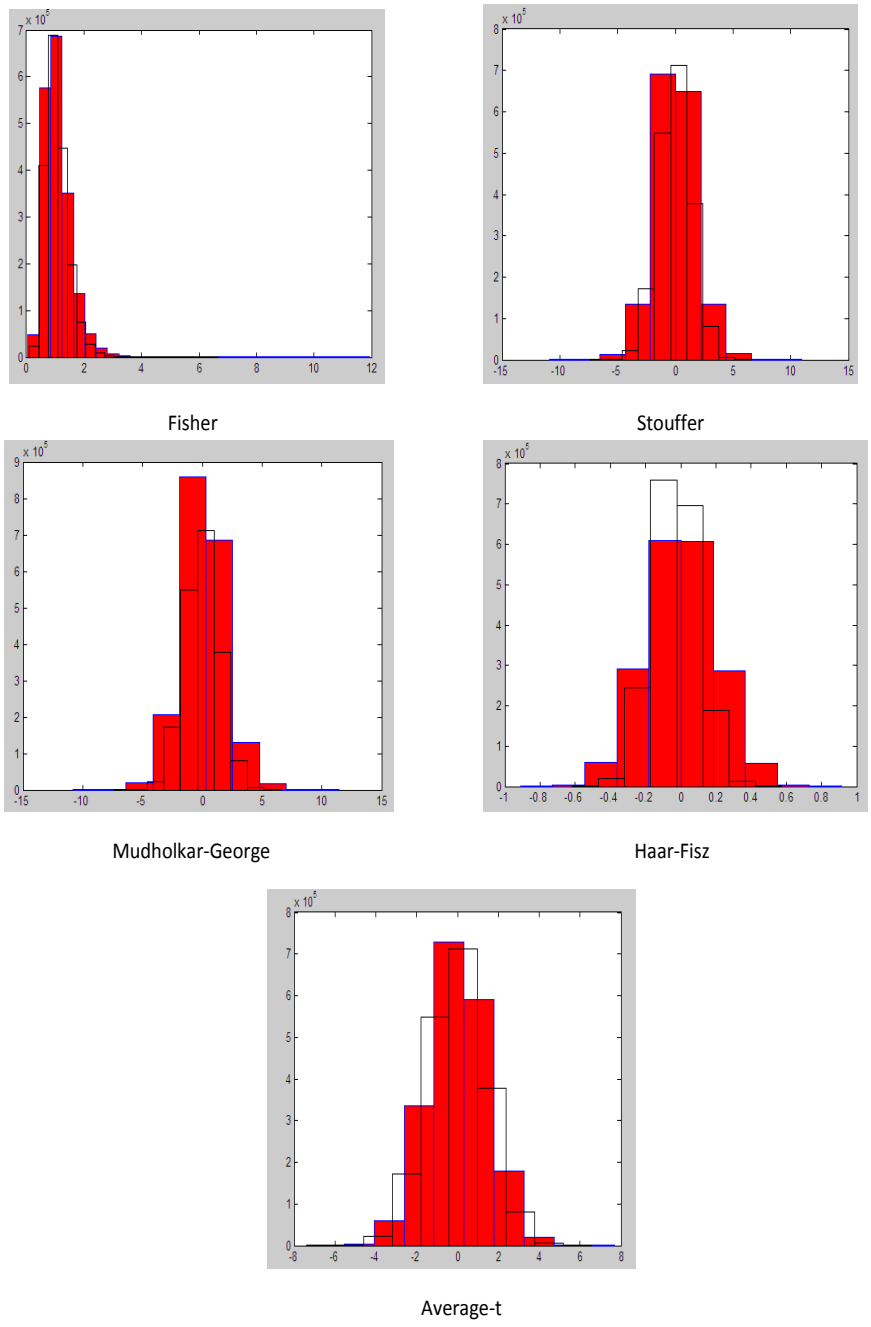

Figure 6.8: Histograms depicting the permutation distribution derived from the combined statistics used to compare two groups (colored in red). The clear histograms represent the true distributions of the combined statistic used to compare two groups. The top rows shows Fisher and Stouffer; the second row shows Mudholkar and George and Haar-Fisz; the third row shows Average t.

Figure 6.9 shows the quantile-quantile plot of true distribution with the permuted distribution. It shows that Haar-Fisz's methods conform more to the true distribution than rest of the methods. Stouffer and Mudholkar-George both look similar.

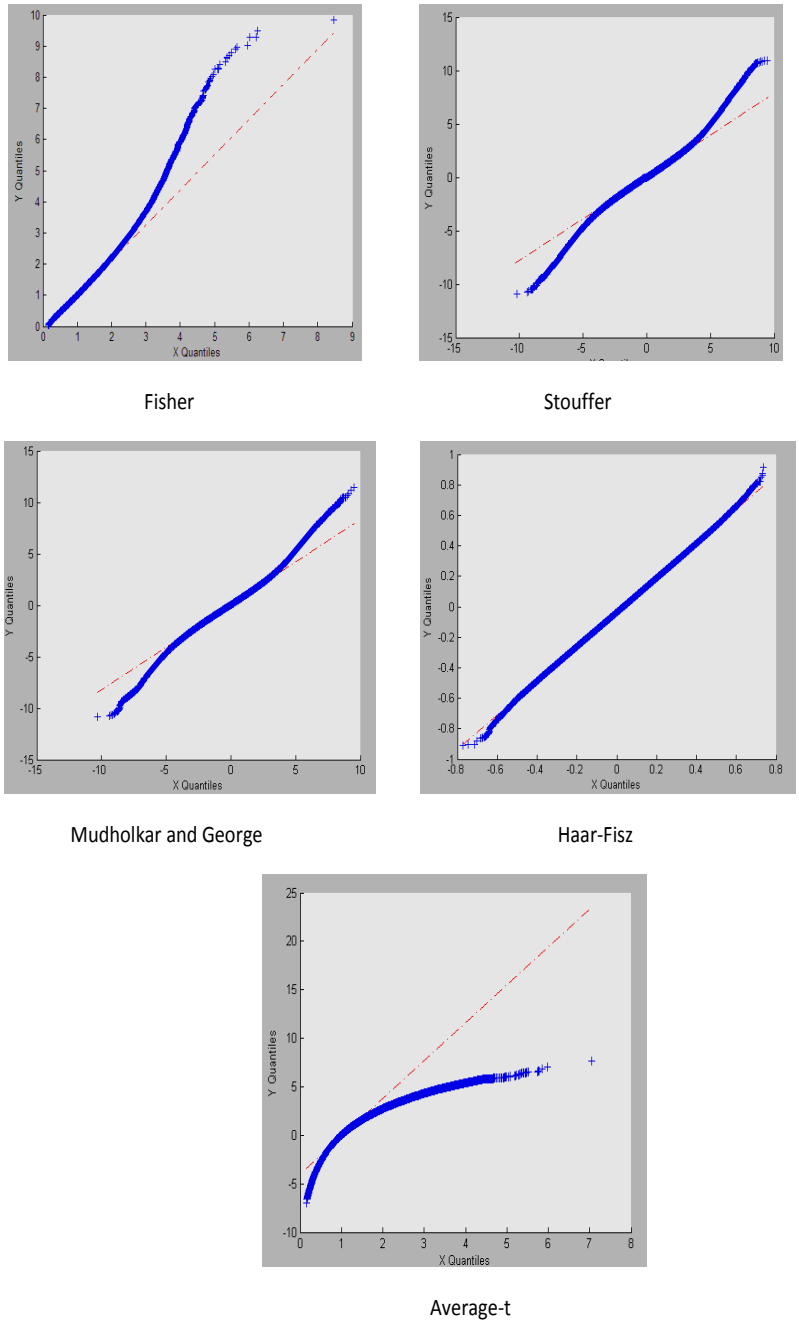

Figure 6.9: QQ plot of the true distribution versus the permuted distribution.

6.3.2 Using Bootstrapping

In bootstrapping there are two sources of error:

1. Error caused by resampling from an empirical cumulative distribution function formed from the initial data set.

2. Error caused by carrying out only a finite number of resamples. For messier problems when the test statistic has a complicated, analytically intractable, distribution the bootstrap would be a reasonable way to address that.

The key steps in performing a bootstrap analysis:

1. Null Hypothesis: Specify the null hypothesis.

2. Exchangeability: Specify exchangeability of observations under the null hypothesis.

3. Statistic: Specify the statistic of interest, usually broken down into specifying a voxel-level statistic deduced from comparison methods using the combination methods.

4. Relabeling: Determine all possible relabeling given the exchangeability scheme under the null hypothesis.

5. Bootstrap Distribution: Calculate the value of the statistic for each relabeling, building the bootstrap distribution.

6. Significance: Use the bootstrap distribution and the critical value derived from the comparative test statistic from the original data, thresholding for the statistical image is done.

Results

Null hypothesis:

H_0: The distributions of the (voxel values of the) groups' difference images have zero mean.

Exchangeability:

The conventional assumption of independent subjects implies exchangeability under the null hypothesis.

Statistic:

In this example we use a bootstrap thresholding test.

Voxel-level summary statistic:

We are interested in searching over the whole brain for significant changes. We will use the comparative methods described in Section 6.2. This will enable us to compare parametric and non-parametric methods. However we can use any statistic to look at the relationship between the two groups.

Relabeling enumeration:

Based on our exchangeability under the null hypothesis, we can randomly sample subjects with replacement from each of the two groups so that we have 15 controls and 16 patients; hence a subject may be repeated more than once in a group. We repeat this random allocation 10,00 times.

Bootstrap distribution:

For each of the 10,000 bootstrappings at each voxel, we compute the comparative test statistics. So for each bootstrap, we have a comparative test statistic image or map. We note the value of the comparative statistics for the original data at each voxel as the critical value. Figure 6.9 shows the bootstrap distributions for each comparison method, based on the 10000 relabelings.

Significance threshold:

P-values are calculated in order to declare a voxel as "active" based on our the bootstrap distribution and the critical value from the original comparative maps. The empirical distribution and the critical values are calculated at each voxel and hence the thresholding is done at the voxel-level; hence a p-value is assigned at each voxel for the comparative methods.

Figure 6.10 shows the bootstrap distribution against the "true" theoretical distribution of the statistic derived from comparing the two groups using the combination tests.

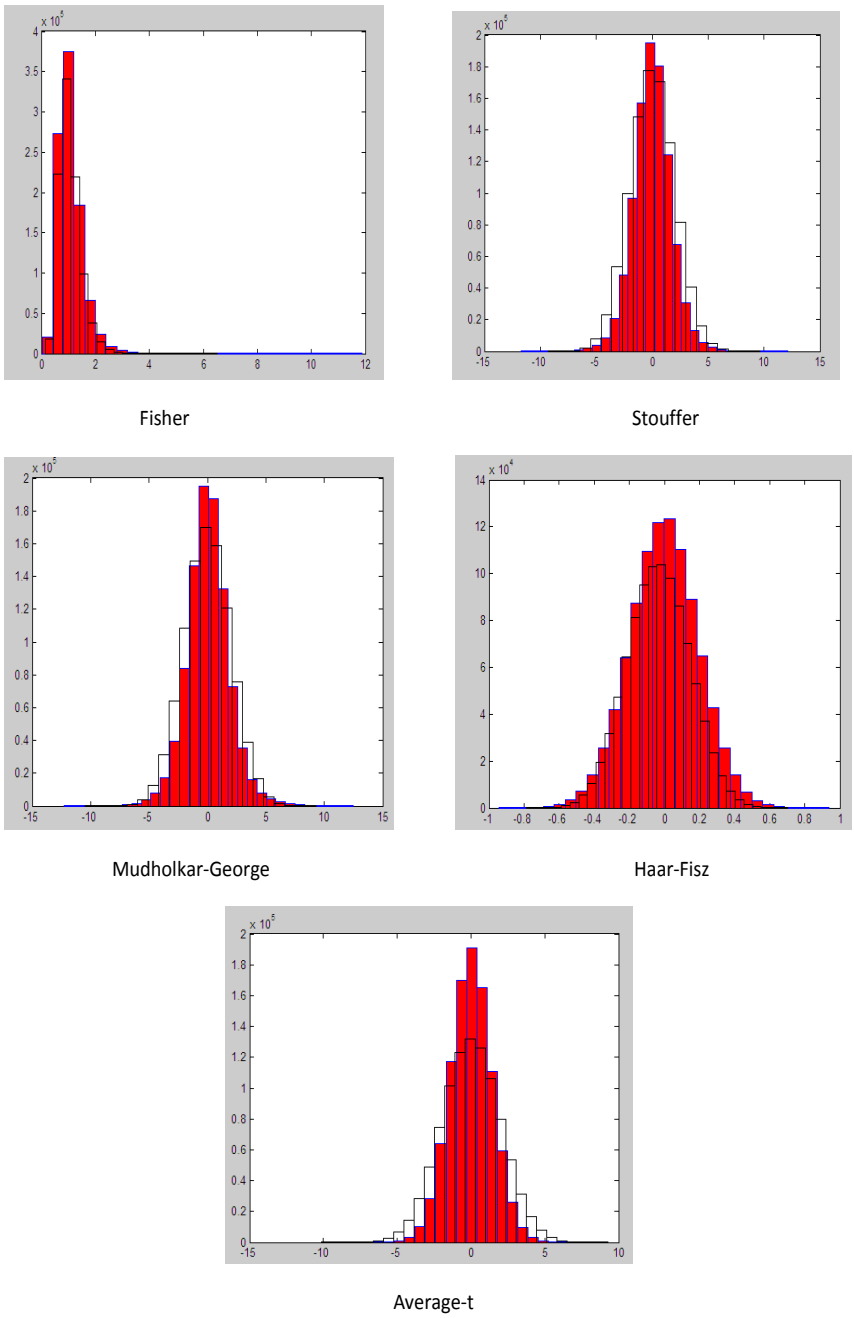

Figure 6.10: Histograms depicting the bootstrapped distribution derived from the combined statistics used to compare two groups (colored in red). The clear histograms represent the true distributions of the combined statistic used to compare two groups. The top rows shows Fisher and Stouffer; the second row shows Mudholkar and George and Haar-Fisz; the third row shows Average-t.

Both theoretical and bootstrap tests were used to assess the probability of each observed voxel statistic under the null hypothesis, and the observed effects maps were thresholded at a size of a twotailed tests with $\alpha = 0.01$. Since no true difference is expected to exist between the two groups, all significant voxels identified by the size of the two-tailed tests at 0.01 should be false positives. In other words, the observed number of positive tests (significant voxels) should equal the predicted number of false positive tests. Figure 6.10 shows that for all the comparison methods, the number of positive tests observed by the theoretical distribution is slightly more than the expected number given by the bootstrap distribution, implying that bootstrap tests are a little more conservative in finding false positives. We can also say that the theoretical distribution underestimates the probability of positives. Correspondence between observed and expected positive tests was nearly the same across all the comparative methods.

Figure 6.11 shows the quantile-quantile plot of the true distribution with the permuted distribution. It shows that Haar-Fisz's methods conform more to the true distribution than do the rest of the methods. Stouffer and Mudholkar-George again have similar performance.

6.4 Discussion

The aim of this thesis is to compare groups of subjects, each subject performing the same task, through distribution theory and non-parametric methods. Theoretically, parametric methods are not suitable since time points are known to be highly correlated and there is usually a small number of subjects (Lage-Castellanos et al., 2009). However the appropriateness of the use of univariate analysis and the treatment of the temporal dimension was discussed in Kiebel and Friston (2004), concluding this approach is a useful strategy; the univariate t-test does not allow identification of time points where relevant differences between the experimental conditions appear but it can identify locations where there is different response to the stimulus. These t-maps form the basis of this thesis to enable us to compare the groups. Specifically, we have estimated comparative test statistics at each voxel

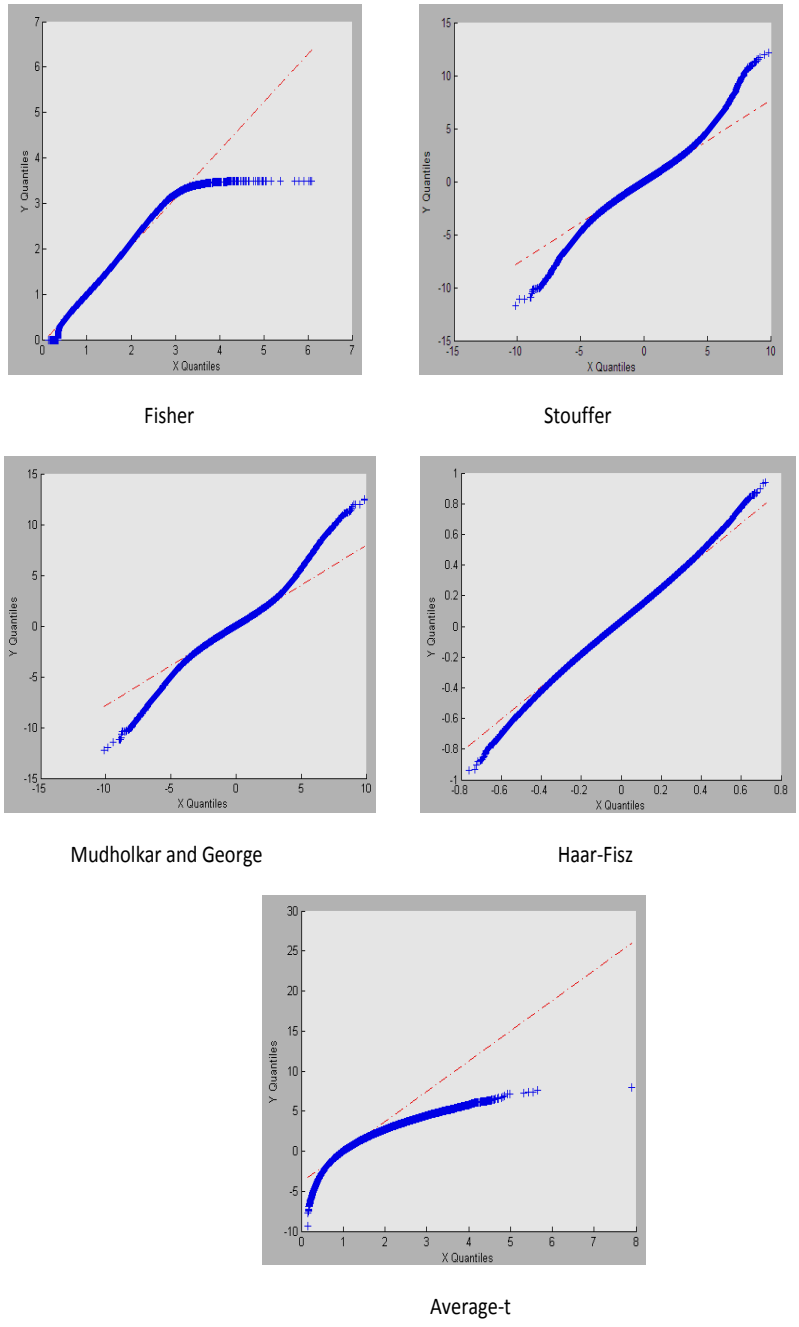

Figure 6.11: QQ plot of the true distribution versus the bootstrap distribution.

and tested each of these against null distributions derived from distribution theory, repeated permutation of the observed data (which in our case is the individual p-value map) and resampling using bootstrapping. We applied FDR correction to the original data set when comparing groups using distribution theory. GLM is the most prevalent method to do so but for reasons stated in Section 5.1, we have explored other methods. The combination techniques addressed here not only enables us to combine p-values from any statistic to create group maps but also they are easy to calculate (even for large number of subjects) and tend to be more liberal in admitting activation, which in our case is ideal since we are exploring the data rather than predicting it. However, we need to keep in mind that the combination methods using p-value maps are often influenced by individual subject behavior.

In order to compare the efficacy and sensitivity of various comparative tests, derived from different combination techniques, in detecting a difference between the two groups, we have created "overlap" maps. These maps show regions of activation during a particular task for patients and controls. If they do not follow the same activation pattern, it will enable us to visually detect the differences between the two groups and corroborate our results from the parametric and the non-parametric approaches with this visual tool for detecting differences between the two groups. Figure 6.12 is a color coded map that shows where the two groups are individually activated in response to the same task and where they overlap. For this figure, individual group maps were created using Fisher's method of combining data. Our aim in this study is to look for differences in the maps i.e. where they do *not* overlap. The figure only looks at the comparison methods using Fisher's and Haar-Fisz's methods of combining data. Similarly, Figure 6.13 is a color coded map where individual group maps were created using Stouffer's method of combining data. This figure only looks at the comparison methods using Stouffer's and Average-t methods of combining data. Figure 6.14 is a color coded map where individual group maps were created using Mudholkar-George's method of combining data. This figure only looks at the comparison methods using Mudholkar-George's method of combining data.

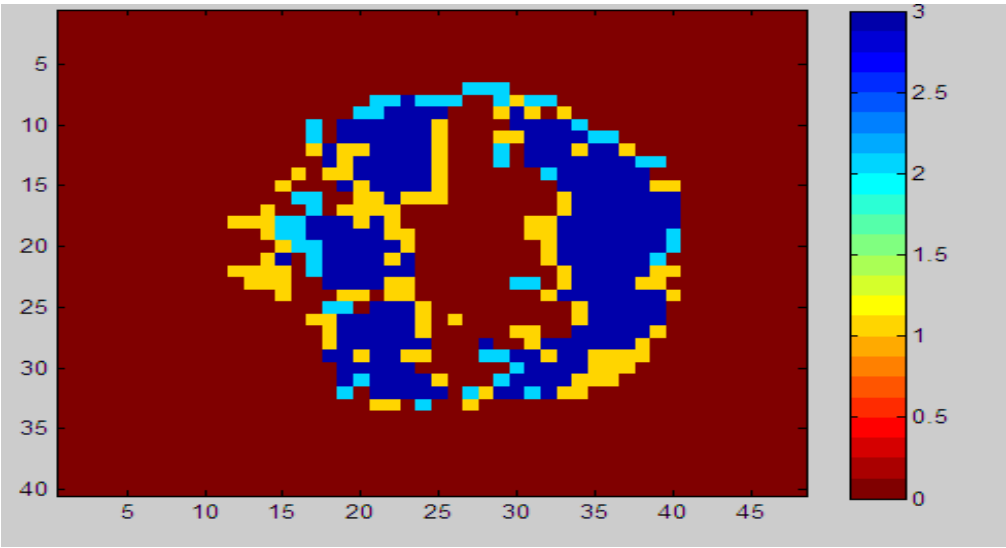

Figure 6.12: The number '1' or the color yellow shows the significantly active voxels for controls. The number '2' or the color sky blue shows the significantly active voxels for schizophrenic patients. The number '3' or the color dark blue shows the significantly active voxels where the two groups overlap. The map is created using Fisher's combination method.

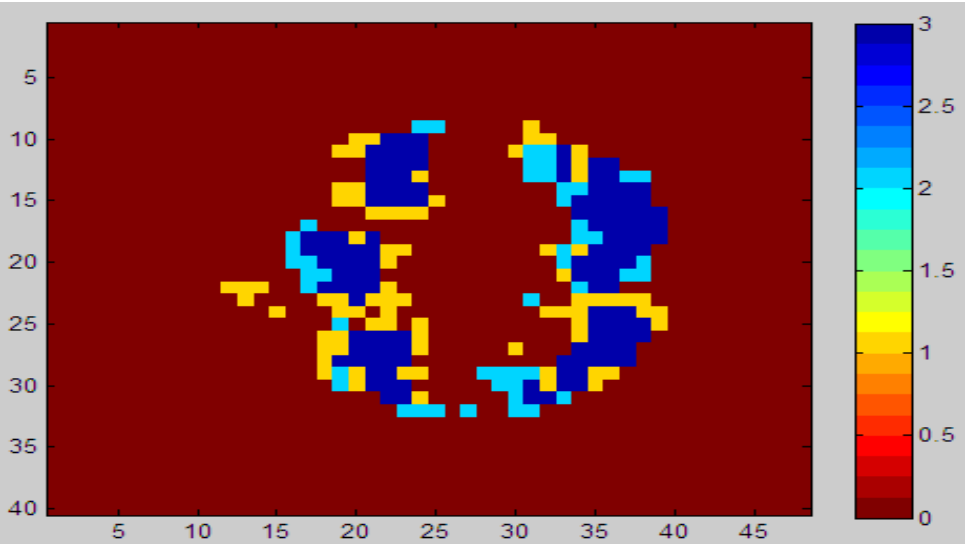

Figure 6.13: The number '1' or the color yellow shows the significantly active voxels for controls. The number '2' or the color sky blue shows the significantly active voxels for schizophrenic patients. The number '3' or the color dark blue shows the significantly active voxels where the two groups overlap. The map is created using Stouffer's combination method.

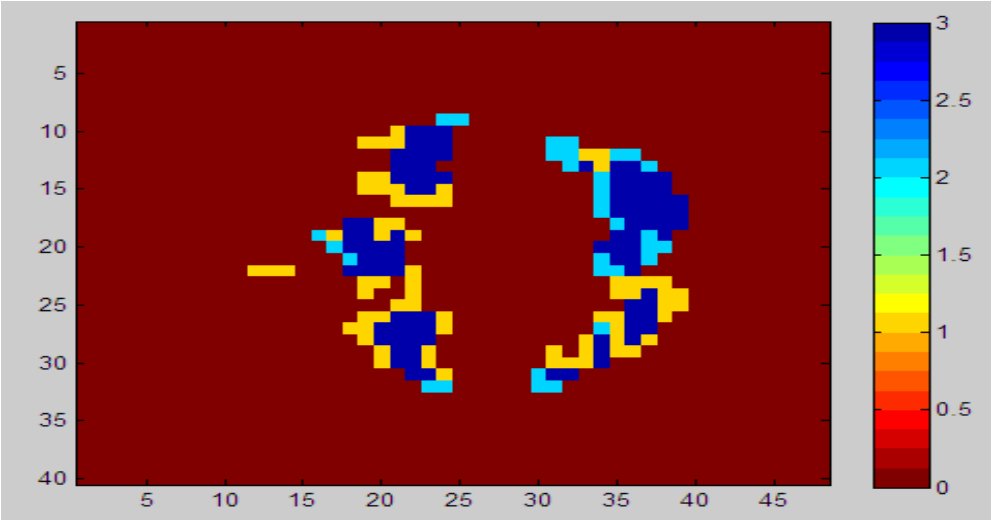

Figure 6.14: The number '1' or the color yellow shows the significantly active voxels for controls. The number '2' or the color sky blue shows the significantly active voxels for schizophrenic patients. The number '3' or the color dark blue shows the significantly active voxels where the two groups overlap. The map is created using Mudholkar-George's combination method.

Here we will compare the figures described in the previous paragraph with the FDR adjusted comparative maps using distribution theory. Comparing Figures 6.12 and 6.7, we can see that Fisher and Haar-Fisz have been able to detect the signals within the brain fairly well. Comparing Figures 6.13 and 6.7 and Figures 6.14 and 6.7 respectively, we can see that Stouffer, Average-t and Mudholkar-George detect the signals towards the periphery but can not detect much towards the inside of the brain. Figures 6.8 and 6.10 showing the distribution of real data after performing the comparison tests against the true distribution also indicate that the true distribution fits better to Fisher and Haar-Fisz than the others. From the real data comparative study, we can say that the ratio methods for comparing two groups give better evidence at looking at the difference in two groups than the difference methods. Mudholkar-George underwent two transformations (logit and Chu's normalizing)and hence a

lot of signal loss is expected in that regard. Table 6.6 also shows that Fisher's and Haar-Fisz's comparison methods are more liberal in detecting any difference between the two groups than Stouffer's, Mudholkar-George's and Average-t, both at FDR adjusted q-values of 0.05 and 0.01.

Looking at differences between the two groups at each voxel involves a high number of hypotheses being tested simultaneously which increases the risk of Type I error. Thresholding or multiple testing, referring to identifying the voxels which are "truly" activated in response to a task and not falsely triggered, is also considered in this thesis. Some works have proposed Bonferroni correction (Yandell, 1997), FDR (Genovese, Lazar and Nichols, 2002) and the use of computer-intensive methods based on permutation test (Westfall and Young, 1993). FDR correction is a common practice in this regard and it is addressed for the comparative test statistics using distribution theory. We also addressed non-parametric methods like permutation tests and bootstrapping to look at the problem of thresholding and compare these with FDR across all the comparative tests using the various combination techniques. The map-wise threshold provided by a permutation analysis is more lenient than that provided by Bonferroni correction or even FDR. In some cases, it can be substantially more lenient, while still providing appropriate tabular control of the false positive rate (Nichols and Holmes, 2001). The non-parametric methods actually control family-wise error rate in our case.

Under the null hypothesis of no difference between the two groups, the comparison among FDR and non-parametric methods controls the same test statistics. In this scenario the power is zero by definition since there cannot be any Type II error, i.e. none of the hypotheses can be mistakenly classified as null since all of them are truly null hypotheses. But from Figures 6.12, 6.13 and 6.14, we see that this is not the case. Table 6.3 and Figure 6.15 compares the differentially active voxels in the two groups using False Discovery Rate, permutation tests and bootstrapping for all the combination methods used to compare two groups of subjects. The α value selected is 0.01.

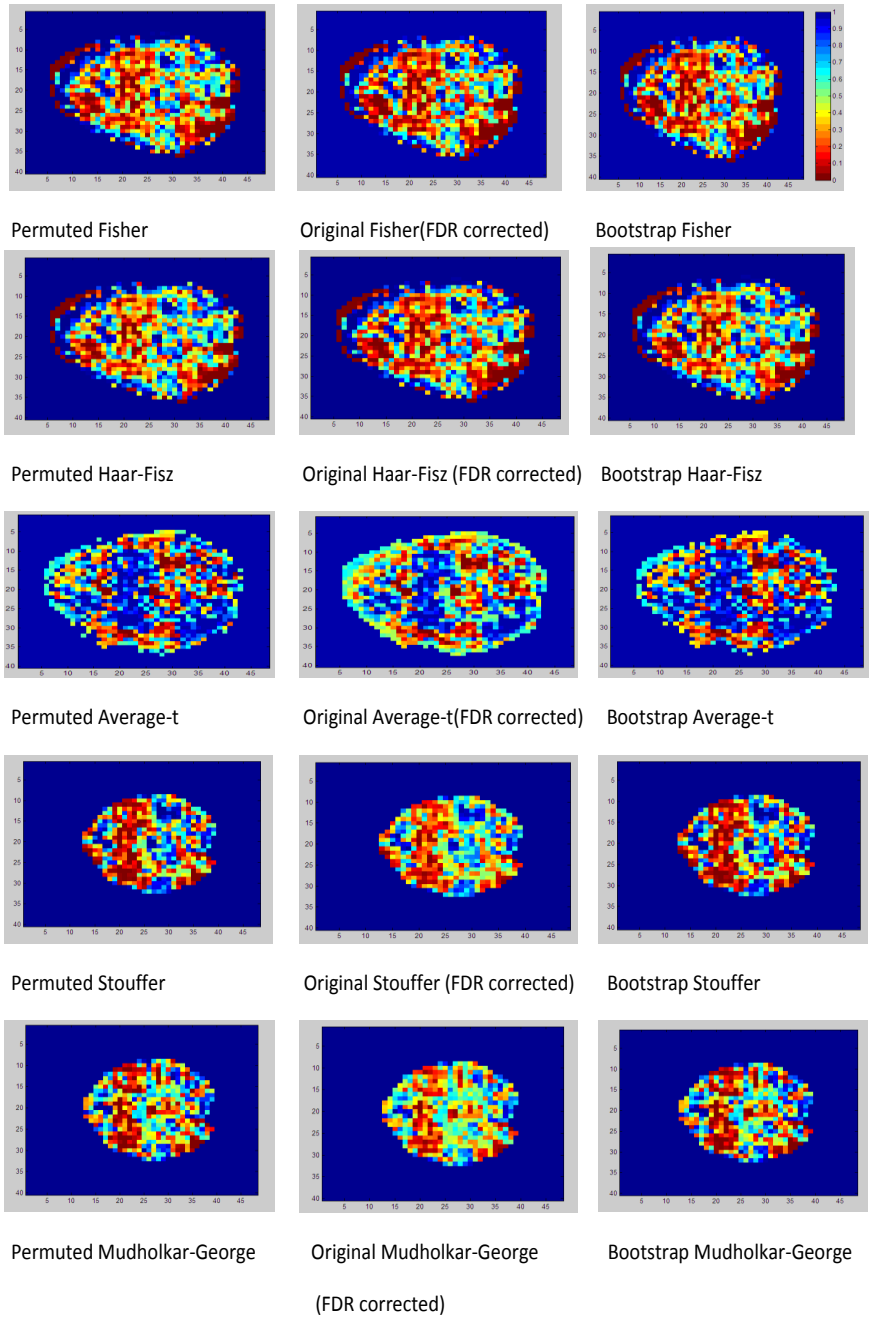

Figure 6.15: Thresholded maps comparing two groups for all the combining methods. The maps on the left are the empirical distributions from permutation tests thresholded with the observed ones. The maps in the middle are derived from comparing the two groups using the original data and FDR corrected. The maps on the right are the empirical distributions from bootstrapping thresholded with the observed ones.

Table 6.3: The table displays the number of voxels declared significant for each of the five methods using three methods for multiple testing at 0.01 level of significance. For instance, using Fisher's method of combining data and then taking the ratio of the two group maps, and using FDR approach for multiple testing, 132 voxels will be declared differentially active, whereas by using the Bootstrap for the same method, 34 voxels will be declared differentially active.

Combination Methods	**Number of significant voxels from Permutation tests**	**Number of significant voxels from Bootstrap**	**Number of significant voxels from FDR**
Fisher	56	34	132
Stouffer	26	17	29
Mudholkar-George	25	19	28
Haar-Fisz	60	43	125
Average-t	23	12	32

Table 6.3 shows the number of rejected null hypothesis (out of 1920 voxels) or significant voxels indicating existence of a difference between two groups for three choices of thresholds. There is a consistency of results among the three thresholding choices across all the comparison methods, FDR being the most liberal and bootstrapping being the least in declaring a voxel differentially active. This could be because the non-parametric methods control FWER which is known to be more conservative than controlling the FDR. All the voxels declared differentially active by bootstrapping subset are in both the permutation and the FDR subsets. However not all the voxels declared differentially active by permutation thresholding are in the FDR subset - most of them are but not all. The critical value for rejecting the null hypothesis being the same for the two non-parametric methods, we can make a statement about the variability of the comparative test statistics; we can say that the variability of the comparative test statistics in permutation distribution is more than that of bootstrapping and hence, we see more differentially active voxels for permutation than bootstrapping. This

is also vindicated in Figures 6.8 and 6.10 as we compare the permutation distributions from various comparison methods with their true null distribution respectively. We believe that the permutation subset is likely to be closer to the true subset, not only because it is an exact test while bootstrapping is an approximate test but also since it makes use of a pooled variance estimate of the comparative test statistics measured at each voxel and the two sample sizes are nearly the same. Figures 6.9 and 6.11 shows how close the permuted and bootstrap distribution are close to the true distribution. Fisher's and Haar-Fisz's comparison methods conform more to the true distribution than the rest of the methods. Also, we can see that the lower quantiles for the bootstrap distribution is much lower than the permuted distribution when both are compared to the true distribution suggesting that bootstrap is more conservative in finding true positives than permutation tests.

Chapter 7

Simulation Study

To compare patterns of activation in two groups, I have considered 10 subjects as controls (Group 1) and 7 as schizophrenic patients (Group 2). P-value maps are simulated from Uniform(0,1) distribution under the null hypothesis; these p-values are assumed to have been generated from the t-statistic maps comparing activation during time periods of stimulation versus time periods at rest. Each p-value map is composed of 64 by 64 voxels for each subject. Three types of simulations were performed for each comparison method based on their *size, intensity* and *location* of activation using both distributional theory and non-parametric methods. We looked at the performance of different comparison methods and the multiple testing methods. We will look at Simulation 1 more extensively than the other two since lessons learned are similar in all cases.

1. Simulation 1: The two groups vary in their intensity of activation when performing a particular task while their location and size of activation remains the same.
2. Simulation 2: The two groups vary in their size (the extent or the number of voxels activated in a particular region) of activated region when performing a particular task while location and intensity of activation remain the same.
3. Simulation 3: The two groups vary in their location of activation when performing a particular task while intensity and size of activation remain the same.

7.1 Results from Simulations

7.1.1 Simulation 1

Three regions of interest are arbitrarily chosen to represent activation in two groups which vary in their intensity only. The intensity, measured in p-values, in the planted patches of activation for the schizophrenic patients varied between (0,0.02) and the intensity of the controls varied between (0,0.04). We explored the simulated data of the two groups looking for differentially active voxels through distributional theory, permutation tests and bootstrapping using the comparison methods described in Section 6.2.

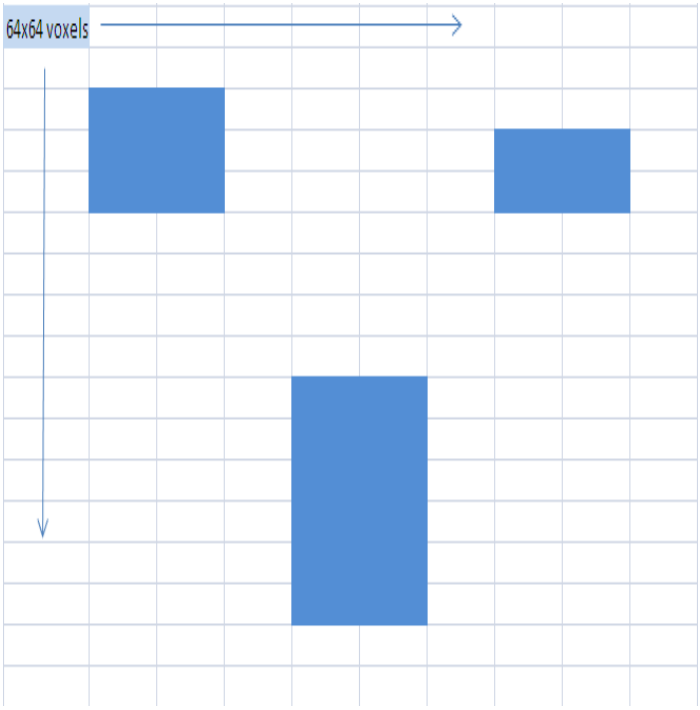

Figure 7.1: Template showing how the stimulation regions are planted in the first simulation study.

Figure 7.1 shows the planted stimulation regions for patients and controls in the first simulation study. The stimulated regions and their size remain the same for two groups while only the magnitude of stimulation varies. Figure 7.2 shows the activated voxels for the two groups; we can detect the locations where the voxels are activated for each group and hence

where the difference in activation lies when thresholded at 0.01. This figure does not reveal whether one group is performing better than the other in terms or intensity or magnitude of the signal strength.

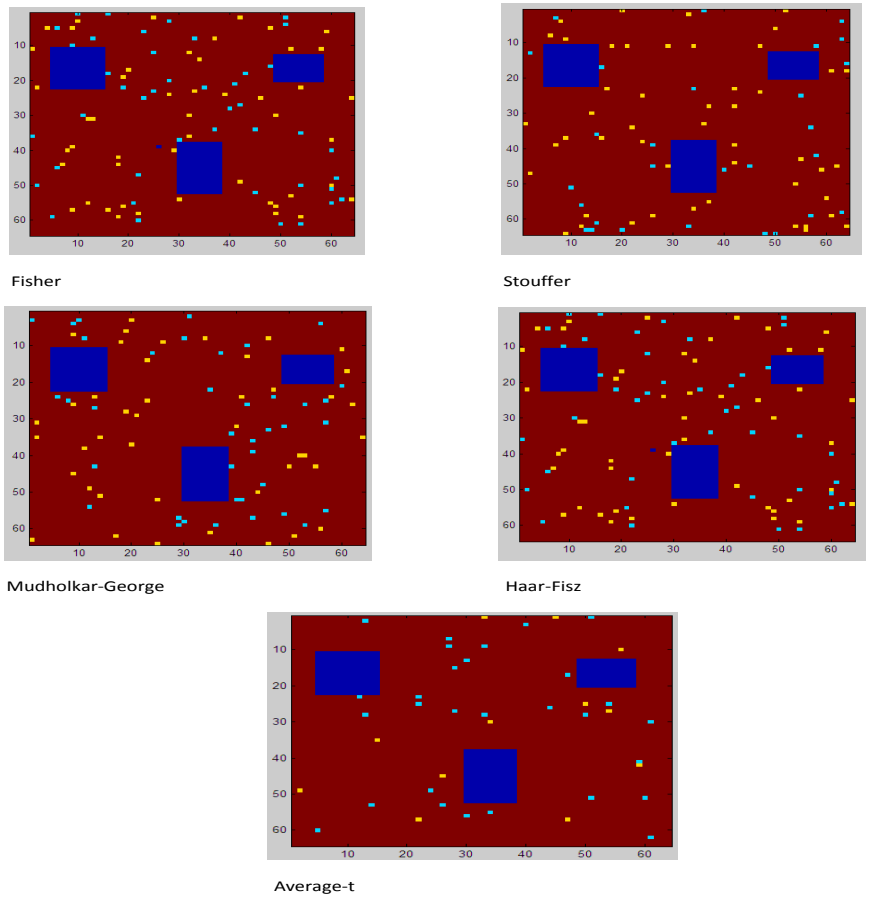

Figure 7.2: This is for Simulation 1. The color yellow shows the significantly active voxels for patients. The color sky blue shows the significantly active voxels for controls. The color dark blue shows the significantly active voxels where the two groups overlap.

Table 7.1 shows number of voxels detected to be differentially active using FDR to correct for multiplicity. The q-values chosen as thresholds to detect false positives are chosen to be 0.05 and 0.01. Table 7.2 illustrates the comparison of the three multiple testing methods through the five comparison methods used to compare two groups. Figures 7.3 and 7.4 demonstrate how each of the comparison methods behave for the three multiple testing methods.

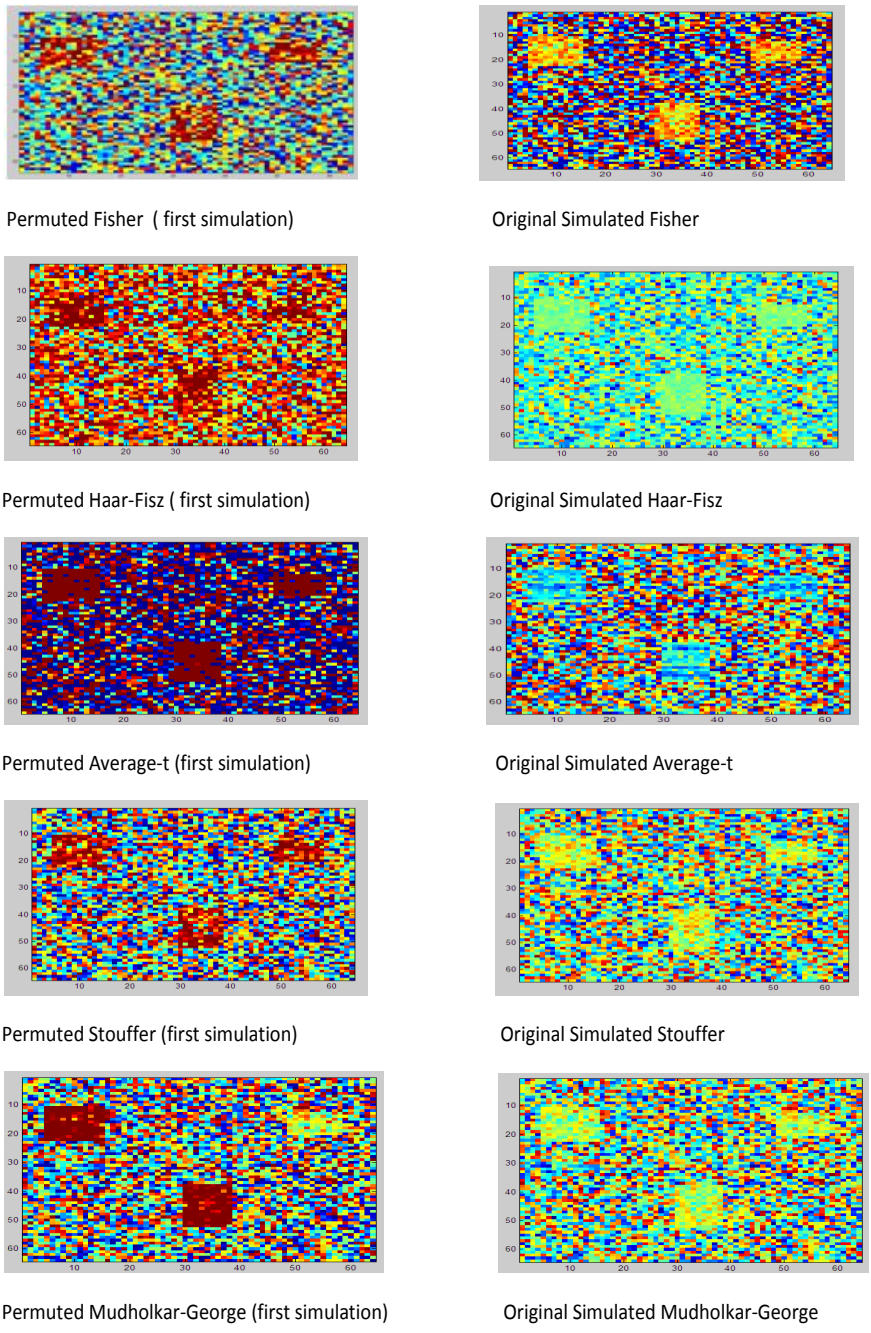

Figure 7.3: Thresholded maps comparing two groups for all the combining methods. The original maps are simulated for 10 controls in one group and 7 patients in the other. The maps on the left are the empirical distributions from permutation tests thresholded with the observed ones. The maps on the right are derived from comparing the two groups using the original data and FDR corrected.

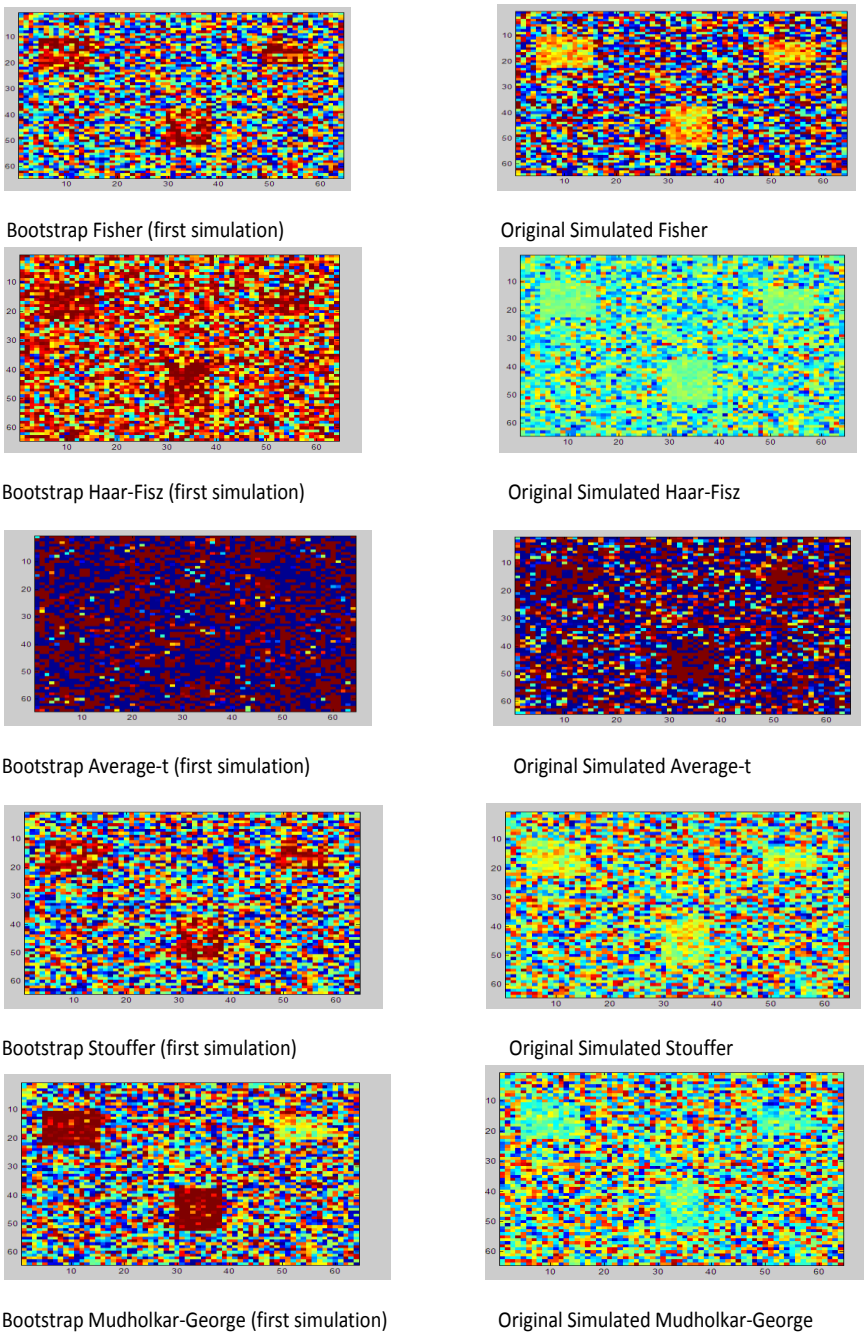

Figure 7.4: Thresholded maps comparing two groups for all the combining methods. The original maps are simulated for 10 controls in one group and 7 patients in the other. The maps on the left are the empirical distributions from bootstrapping thresholded with the observed ones. The maps on the right are derived from comparing the two groups using the original data and FDR corrected.

Table 7.1: The table displays the number of voxels declared differentially active for each of the five methods at various significance levels. This is for the first simulation. For instance, using Fisher's method of combining data and then using ratio of the two group maps, and a significance level of 0.05, 481 voxels will be declared differentially active, whereas by using the Stouffer method, 1108 voxels will be regarded as differentially active.

q-value FDR adjusted	**Fisher**	**Stouffer**	**Mudholkar-George**	**Average t**	**Haar-Fisz**
0.05	481	1108	1348	2911	457
0.01	396	993	1179	2935	385

7.1.2 Simulation 2

The planted stimulation regions for patients and controls in the second simulation study are similar to Figure 7.1, the differences being that the two groups varied in size of activation to a particular task, though hovering around the same region as shown in the template, and the magnitude of stimulation is the same in both the groups unlike the first simulation study. Figure 7.5 shows the activated voxels for the two groups; we can detect the locations where the voxels are activated for each group and hence where the difference in activation lies when thresholded at 0.01. This figure does not reveal whether one group is performing better than the other in terms or intensity or magnitude of the signal strength.

To show that the results are consistent over various simulations, I am furnishing the values and images (Figure 7.6) for the comparison done with Fisher's method combining data and then using the ratio of the two group maps to compare the two groups. For the Fisher's method, at q-value of 0.01, 150 voxels will be declared differentially active using FDR approach for multiple testing, 123 voxels will be declared differentially active using permutation tests for multiple testing and 107 voxels will be declared differentially active using bootstrapping for multiple testing. Figure 7.6 demonstrates how sensitive the methods used to correct for multiple testing are with respect to each of the comparison methods.

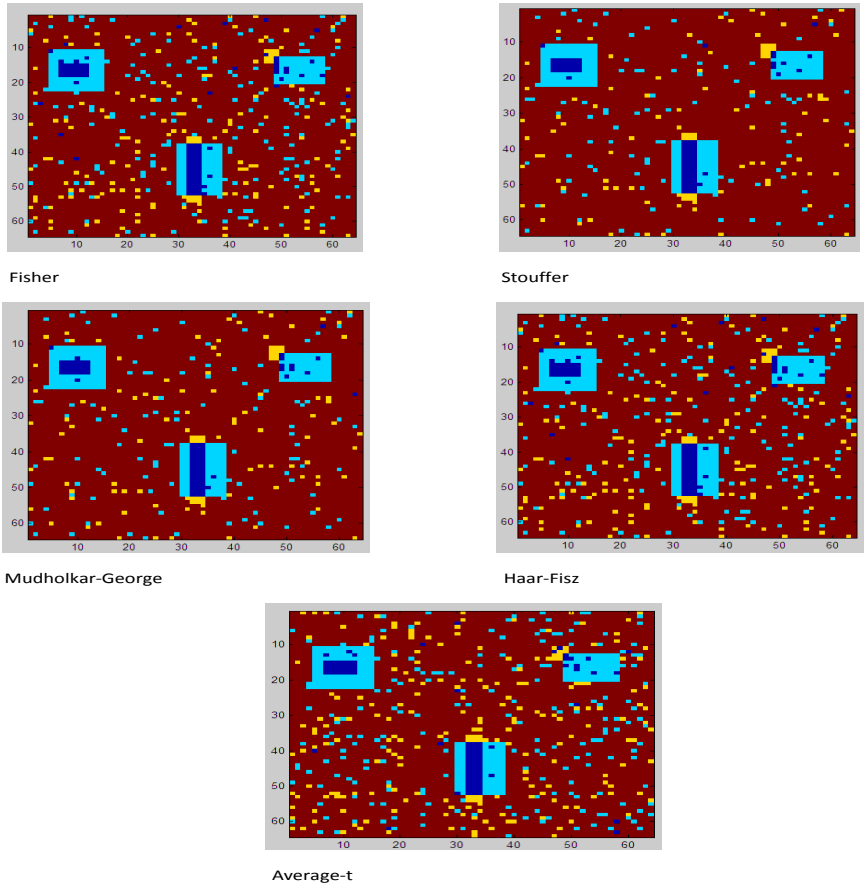

Figure 7.5: The color yellow shows the significantly active voxels for patients. The color sky blue shows the significantly active voxels for controls. The color dark blue shows the significantly active voxels where the two groups overlap.

Table 7.2: The table displays the number of voxels declared significant for each of the five methods using three methods for multiple testing at 0.01 level of significance. This is for the first simulation. For instance, using Fisher's method of combining data and then using the ratio of the two group maps, and using FDR approach for multiple testing, 396 voxels will be declared differentially active, whereas by using the Bootstrap for the same method, 384 voxels will be declared differentially active.

Combination Methods	**Number of significant voxels from Permutation tests**	**Number of significant voxels from Bootstrap**	**Number of significant voxels from FDR**
Fisher	387	384	396
Stouffer	859	776	993
Mudholkar-George	851	705	1179
Haar-Fisz	366	366	385
Average-t	2332	1879	2935

7.2 Discussion

The tables and images presented from the simulation studies demonstrates the sensitivity and ability of each of the procedures discussed in Section 6.2 in comparing the two groups. The first simulation studied here extensively looks at the sensitivity of each of the comparison methods in detecting a difference between the groups since all that was varied was the intensity while the location and size of the planted patches remained the same. Images of the original simulated data from Figures 7.3 and 7.4 (for first simulation study) showing the five comparison methods and Figure 7.6 (for second simulation study) showing the Fisher's comparison method illustrate that the comparison methods were able to detect the planted patches where the difference between the two groups actually lies even though the difference was small, on average. The various colors represent the difference in intensity or magnitude of the signal between the two groups. When we corroborate that information with Figure 7.2 for simulation study 1 and Figure 7.5 for simulation study 2, we can see that all the

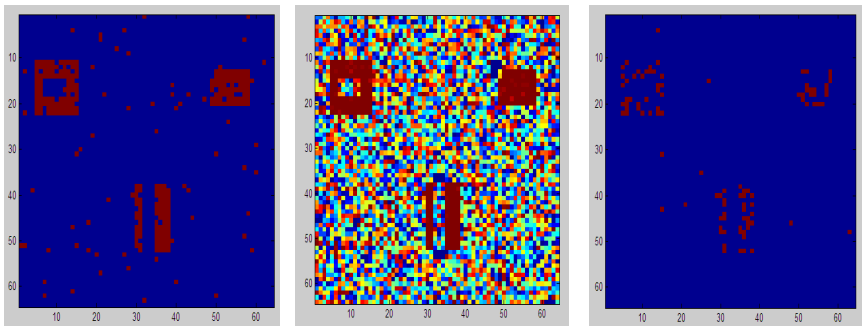

Figure 7.6: Thresholded maps comparing two groups for all the combining methods. This is for the second simulation data. The original maps are simulated for 10 controls in one group and 7 patients in the other. The maps on the left are the empirical distributions from permutation tests thresholded with the observed ones. The maps in the middle are derived from comparing the two groups using the original data and FDR corrected. The maps on the right are the empirical distributions from bootstrapping thresholded with the observed ones.

comparison methods were able to detect the planted patches but the level of sensitivity in detecting the differences between the two groups varied. This figure was used to only get an overview whether there was a difference in activation between the two groups or not. It does not reveal whether one group was better than another in terms of intensity or magnitude of signal strength. Table 7.3 and Table 7.4 puts a perspective to that level of sensitivity with regard to how many true differential activations each comparison method has been able to detect since we know the truth from our simulated data.

In Table 7.3 and 7.4, the first and last rows depict the accuracy of detecting a true differential activation for each of the comparison methods while the second and the third rows depict the mistakes made with each of those methods. The tables suggest that comparison using Fisher's and Haar-Fisz's techniques made the least mistake in declaring a voxel differentially active while the numbers of voxels that were not detected to be differentially active when they are truly so was worst for average-t. The comparison method using Haar-Fisz's and

Table 7.3: The "truth" table for simulation study 1 displays true difference and false difference between the two groups at 0.01 level of significance and FDR adjusted for multiple testing.

True discovery vs. False discovery	**Fisher**	**Stouffer**	**Mudholkar-George**	**Average-t**	**Haar-Fisz**
Number of voxels that were truly different were discovered	347	234	216	110	347
Number of voxels that were truly different were not discovered	35	148	166	272	35
Number of voxels outside the planted patches were discovered as different	49	759	963	2825	38
Number of voxels outside the planted patches were not discovered as different	3665	2955	2751	889	3676

Fisher's method of combining data discovered most truly differentially active voxels and less number of false alarms. So overall comparison method using ratio techniques have been able to balance the Type I and the Type II error and hence been able to detect the maximum number of true differentially active voxels.

It is a very powerful fact that whenever the two sample sizes are approximately equal, the permutation method provides an estimated test statistic null distribution which is asymptotically correct and may in fact be more efficient for small sample sizes (by using pooled estimates of the covariance matrix). However, the permutation method suffers from a bias where the estimated null distributions of the mean of the test-statistics used to compare the two groups do not have their means as zero if the two sample sizes are unequal, unless the

Table 7.4: The "truth" table for simulation study 2 displays true difference and false difference between the two groups at 0.01 level of significance and FDR adjusted for multiple testing.

True discovery vs. False discovery	**Fisher**	**Stouffer**	**Mudholkar-George**	**Average-t**	**Haar-Fisz**
Number of voxels that were truly different were discovered	58	36	27	10	59
Number of voxels that were truly different were not discovered	3	25	34	51	2
Number of voxels outside the planted patches were discovered as different	92	201	257	1283	79
Number of voxels outside the planted patches were not discovered as different	3943	3834	3778	2752	3956

observed difference in means is zero (Pollard and van der Laan, 2003). However for bootstrap method the bias is independent of the observed difference but can be anti-conservative if the errors are not homogeneous. From my limited simulation, we can also see from Table 7.2 that none of the multiple testing methods were able to detect "all" the true differentially active voxels. However all the voxels that were detected by each of the multiple testing methods represent true differential activation. Similar to real data, FDR was least conservative and bootstrap the most.

Chapter 8

Conclusion and Recommendations

In this thesis I have discussed five methods (namely, Fisher, Stouffer, Mudholkar-George, Average-t and Haar-Fisz) used to combine data from multiple subjects and then compare two group maps. I have made a comparative study of these with the popular random effects model using real data only for distribution theory. All the five combination methods are based on combining p-values obtained from a combined statistical hypothesis test. The random effects model is based on the data itself in the form of t-tests.

The comparative maps and tables showing the performance of each of these methods in comparing two groups leads to the issue of which method should be used. Averaging of t-statistics (or other statistics) at each voxel over multiple subjects uses the data efficiently but it smooths away much of the signal and hence a lot of information across the subjects is lost. However usage of pseudo-t will give a better variance estimation but we will not be able to use distribution theory in this regard because there does not exist any known reference distribution. Among the other p-value based methods, Fisher and Haar-Fisz takes a logarithm, Stouffer takes an inverse cumulative distribution and Mudholkar-George takes a logit transform and then normal approximation to t-distribution to combine and compare two group maps. The random effect model is based on a weighted average which is done to smooth over the realigned voxels (Lazar et al., 2002). Working with transformations on real data always leads to some loss of information as in the case of the combination techniques explored here. But their advantage over the random effects model is that the computation and mathematical manipulations are not complicated.

Fisher and Haar-Fisz use a ratio of statistics to compare the two groups; hence these tests do not give us the magnitude of the difference in the combined signal strength for the two groups as do Stouffer, Mudholkar-George and Average-t but rather a proportion of the numerator group combined signal relative to the denominator. In other words, under alternative hypothesis, the ratio tests look at the relative difference in the strength of the signal i.e. difference in intensity of one group with respect to another while the difference tests look at the absolute difference in magnitude of the signal strength of one group compared to another. In some cases not related to fMRI, it has been shown that a closer fit to reality is sometimes achieved if comparisons are a function of differences than ratios (Hirshleifer, 1989). On the other hand for another unrelated example, it has been shown that if the data are linear with slope which changes under motivational bias strongly argues against difference comparator mechanisms and in favor of ratio comparators (Gibbon and Fairhurst, 1994).

Fisher's method is easy to implement and has a convenient distributional form but it has some disadvantages. It can yield inconsistent results with even simple overall tests like the sign test of the null hypothesis of a 50:50 split (Siegel, 1956). Thus for a large number of studies, if the vast majority showed results in one direction then by the sign test, we would reject the null hypothesis even if the consistent p-values were not very much below 0.5. However in these situations, the Fisher method would not yield an overall significant p-value (Mosteller and Bush, 1954). Another problem with Fisher's method is that if two studies with equally and strongly significant results in opposite directions are obtained, this method supports the significance of either outcome. Thus p-values of 0.001 for $A > B$ and 0.001 for $B > A$ combine to a $p < .01$ for $A > B$ or $B > A$ (Adcock, 1960). Despite these limitations Fisher's method still remains the best known and most discussed among all the methods for combining probabilities (Rosenthal, 1978). Many authors and most notably Rosenthal concluded that there does not exist any best method under all conditions (Birnbaum, 1954), but the one that could be used under the largest range of conditions is the method of adding

Z scores. Hence the statement points towards Stouffer's method and somewhat towards Average-t.

Fisher and Haar-Fisz were able to detect the signals within the brain fairly well. Comparing Figures 6.11 and 6.7 and Figures 6.12 and 6.7 respectively, we can see that Stouffer, Average-t and Mudholkar-George detected the signals towards the periphery which were likely artifacts but did not detect much inside the brain. The true distribution fitted better to Fisher and Haar-Fisz comparison methods than the others. From the real data comparative study, we can say that the ratio methods for comparing two groups were better at detecting differences than the difference methods. Mudholkar-George underwent two transformations, logit and Chu's normalizing, and hence a lot of signal loss is expected in that regard. Fisher's and Haar-Fisz's comparison methods were liberal in detecting any difference between the two groups more than Stouffer's, Mudholkar-George's and Average-t's, both at FDR adjusted q-values of 0.05 and 0.01. We derive similar conclusions from looking at the simulation studies. Since we had planted areas of activation which would be our regions of interest, we did not expect all the voxels to conform to the null distribution. The tables and images from the simulation studies showed that Fisher and Haar-Fisz were able to detect the difference in two groups better than others. However it can be noted that Stouffer, Mudholkar-George and Average-t furnish near-similar results and seemed to be quite close in their approach. With the evidence from real data and simulated study, it would be more appropriate to compare two groups using Fisher's combination technique to get group maps and then taking the ratio of the two Fisher maps.

From the real data study, there is a consistency of results among the three thresholding choices across all the comparison methods: FDR being the most liberal and bootstrapping being the least in declaring a voxel differentially active i.e. bootstrapping exerts a stronger control of the type I error than the FDR method. Using real data, all the voxels declared differentially active by bootstrapping subset were in both the permutation and the FDR subset. However all the voxels declared differentially active by permutation subset were not

in the FDR subset - most of them were but not all. The variability of the comparative test statistics in permutation distribution was more than that of bootstrapping and hence, we saw more differentially active voxels for permutation than bootstrapping. Using simulated data, we saw the similar trend in detecting differentially active voxels among the three multiple testing methods and all the voxels declared differentially active by the non-parametric methods were in the FDR subset. We believe that the permutation subset is likely to be closer to the true subset, not only because it is an exact test while bootstrapping is an approximate test but also since it makes use of a pooled variance estimate of the comparative test statistics measured at each voxel and the two sample sizes are nearly the same. The computation complexity of the non-parametric methods were also higher than that of FDR and the estimation usually took 20 times as long to get thresholds from a permutation method than to do the FDR thresholding. The FDR method may be a good alternative but with the advent of fast computing, permutation test gives a better result. Permutations tests are very flexible and intuitive with requirement of minimum assumptions for valid inference and will work under heterogeneous error. Also permutation tests provide a viable alternative analysis method to parametric approaches when the assumptions of the latter are not met.

Chapter 9

Future work

9.1 Extension to three or more groups

We have explored the possibility of comparing three groups without having to do pairwise comparisons. We have two methods to do so:

1. The group maps overlaid so that we have a qualitative but not quantitative way of understanding where the difference in activation lies with respect to each group.

2. Use of non-parametric one-way ANOVA: Kruskal Wallis method.

The source populations for the three or more group maps may not be normal in our case of fMRI study. An appropriate non-parametric alternative to the one-way independent-samples ANOVA can be found in the Kruskal Wallis test. It does, however, assume that the observations in each group come from populations with the same shape of distribution, so if different groups have different shapes (one is skewed to the right and another is skewed to the left, for example, or they have different variances), the Kruskal Wallis test may give inaccurate results (Fagerland and Sandvik 2009). But that should not be our case since we will use the same combination method to combine the data in each group with the belief that individual maps in a group are comparable and hence combinable. Like many non-parametric tests, the Kruskal-Wallis test is performed on ranked data, so the measurement observations are converted to their ranks in the overall data set. The loss of information involved in substituting ranks for the original values can make this a less powerful test than its parametric counterpart. The null hypothesis is that the samples come from populations such that the probability that a random observation from one group is greater than a random observation

from another group is 0.5 i.e. the group maps comes from the same population and hence there is no significant difference among them. The Kruskal Wallis test does not test the null hypothesis that the populations have identical means or have equal medians. The Kruskal Wallis test statistic, H is given by:

$$H = \frac{12}{n\,(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\,(n+1)$$

where R_i = sum of the ranks for group i for $i = 1, 2, ..., k$. H represents the variance of the ranks among groups, with an adjustment for the number of ties. H is approximately chi-square distributed. If the sample sizes are too small, H does not follow a chi-squared distribution very well, and the results of the test should be used with caution. N less than 5 in each group seems to be the accepted definition of 'too small'.

The real data set also contains a third group: relatives of schizophrenic patients comprising 13 subjects who performed the same anti-saccade task. Figure 9.1 is the color coded map showing the difference in activation patterns in the three groups. The combining method used to derive each group map is Fisher's.

Figure 9.2 demonstrates the capability of the Kruskal Wallis test to detect the difference in activation patterns in the above mentioned three groups at 0.01 level of significance and FDR corrected.

Comparing figures 9.1 and 9.2, we want to see Kruskal Wallis being able to detect the signals represented by '2' (orange),'3' (yellow) or '4' (light green). We can see that Kruskal Wallis has been able to detect signals more towards inside of the brain than towards the periphery. However, it does not provide a conclusive evidence that it has been able to fully detect the activation patterns in the three groups that differs one from the other. This area can be explored more and will be the focus of more of my future research.

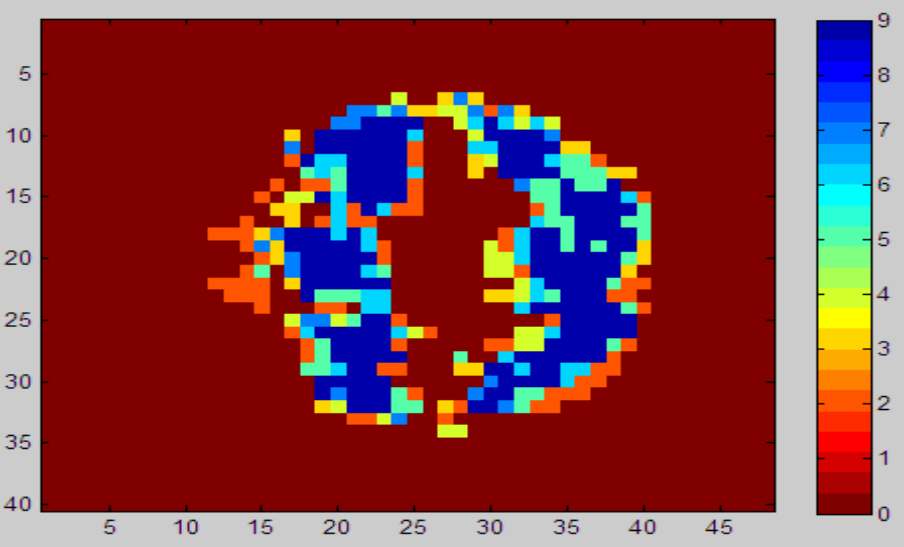

Figure 9.1: The number '2' or the color orange shows the significantly active voxels for controls. The number '3' or the color yellow shows the significantly active voxels for schizophrenic patients. The number '4' or the color light green shows the significantly active voxels for relatives of schizophrenic patients. The number '5' or the color turquoise blue shows the significantly active voxels where the controls and schizophrenic patients overlap. The number '6' or the color sky blue shows the significantly active voxels where the controls and their relatives of schizophrenic patients overlap. The number '7' or the color light blue shows the significantly active voxels where the schizophrenic patients and their relatives overlap. The number '9' or the color dark blue shows the significantly active voxels where the three groups overlap.

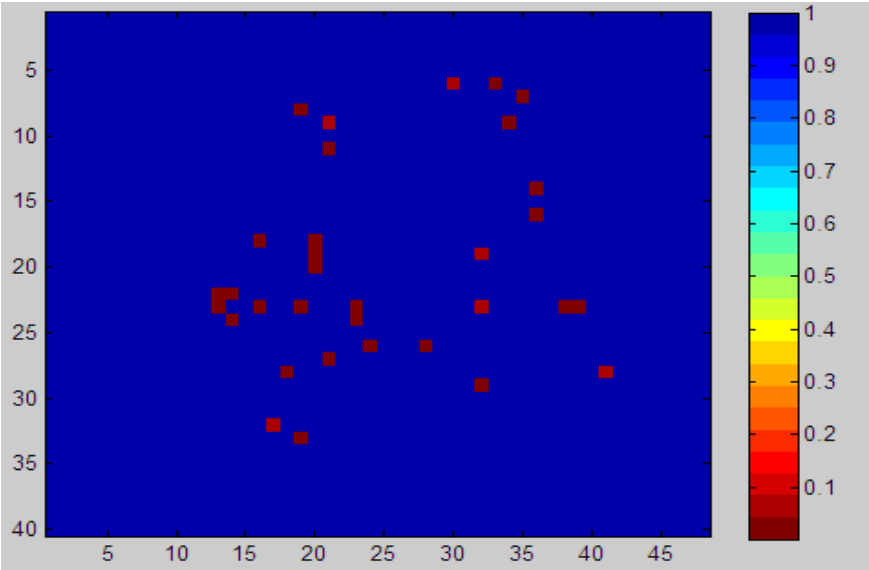

Figure 9.2: The image depicting the p-values derived from Kruskal-Wallis test which is FDR corrected at 0.01 level of significance.

9.2 Addendum to Fisher's

An addendum to Fisher's method (H. Ombao; communicated personally to N.Lazar) can be used to analyze two group maps. The drawbacks of Fisher's method detailed in the previous chapter can be corrected using this method. After we combine the data into two group maps from two samples of equal size k, we have

$$T_{F1} \sim \chi^2_{2k}$$

$$T_{F2} \sim \chi^2_{2k}$$

We define, $R = \frac{T_{F1}}{T_{F2}}$

We define, $V = \max\left(\log(1+R), \log(1+1/R)\right) - \log(2k)$

Then, $V + \log(2k) \sim$ Beta prime distribution or Beta distribution of second kind (Johnson,

1995) with location parameter k and shape parameter k.

Also, $V + \log(2k) \sim$ log-F distribution (Jones, 2006) where $R \sim F(2k, 2k)$.

The disadvantage of this method is that it works *only* for two equal sample sizes. I have tried derivations to get a closed theoretical distribution for this so that we can apply distribution theory to compare two group maps but that was not feasible. However, with the usage of permutation tests and bootstrapping for comparison purposes, once again we need not be bound by a statistic that has a closed or known distributional form and hence it will be feasible for us to perform comparative study using non-parametric methods.

9.3 Edge Effect correction

The "edge effect" refers to the phenomenon that the comparison methods seem to mostly pick out voxels along the edge of the brain as being different between the two groups, but this is an artifact of the Talairach transformation not warping the individual maps to be exactly the same size and shape. So all that is being picked up are the edges, where one group map has voxels and the other doesn't. Edge effects complicate the analysis of fMRI data where we need to consider spatial processes. Many formulas relating to spatial processes are affected by the finiteness of the region of interest.

We tried a few methods to correct it:

1. Truncating the area outside the brain or the air voxels and re-calculating our combined statistic and subsequently the statistics to compare the group maps. We assumed that all the activity is concentrated within the brain and hence extraneous information from the air voxels will only inflate our results. However this method did not yield any satisfactory result because the tests were done at each voxel and not on clusters or regions of interest.

2. Following the footsteps of spatial statistics where edge effect can be minimized by statistical tests based on the distance to the apparent nearest neighbor, we weighted the individual p-value maps so that less weight was placed on the periphery of the brain and more towards

the center. The weighting was adhoc. However in the process of doing so, we have warped the brain by a large margin and hence much of the real information was lost and hence after comparing the group maps, we have not been able to see any satisfactory result. This is a serious issue which can be addressed through spatio-temporal modeling of the fMRI data.

9.4 Assessing effect of individual subjects on group comparison

It is informative to look at fMRI group maps which provide a collective summary of a specific subject population and to use them to compare activation regions between two or more groups. However, variability exists among subjects belonging to one particular group — some individuals may exhibit higher levels and more pervasive regions of activity than others. These individuals may significantly influence the group maps and we would want to know whether and by what degree the high-activators affect the group maps than low-activators. If the population is not homogeneous as in this case, the scientific conclusion may be biased by one or a few subjects in a group. McNamee and Lazar's (2004) paper states that " A 1992 National Research Council report on combining information compared, from a statistical perspective, various popular methods based on p-values (such as Fisher's method) and on statistical models (such as fixed and random effects). The report indicated that some methods may indeeed be sensitive to the results of a single study, although the issue was not addressed in much detail. Researchers in statistics and in psychology who have compared combining methods have tended to do so from perspective of power and basic attributes such as ease of implementation. Thus from statistical point of view, not much is known that will provide a practical guidance to experimenters who wish to understand the robustness of their results. With respect to fMRI data, research in this area is also rather sparse (Page 920)."

In order to explore the sensitivity of the comparative statistics, as addressed in Section 6.2, to individual data values from two groups and the effect of an individual on the statistic, we will approach it using "jackknife". The jackknife method or "delete one diagnostic" was

first introduced by Quenouille (1949) as a method to estimate the bias of a one-sample estimator and thus obtain a bias-reduced jackknife estimator. From this method, we not only gain a better understanding of how the influence of one or more subjects manifests itself in comparison maps but we also learn which methods are particularly susceptible to such influence. This procedure is implemented by first calculating the complete comparative statistical maps (CCSM) based on all subjects in two groups and then calculating "leave-one-out" comparative statistics maps (LOOCSM) leaving out each of the individual subjects from any one of the two groups in turn. We can then quantify and define the impact through four disparity measures derived from taking the difference between the former and the latter maps and define it — the value of the difference is binary because after FDR correction, the comparative maps show voxels of activation, denoted by 1, and voxels of inactivation, denoted by 0. Thus the possible values in our case will be 1, 0 and -1.

The four disparity measures are defined as follows (McNamee and Lazar, 2004):

1. Enhancing voxels: counts the number of voxels that were added (or enhanced) in the CCSM when the subject was included in the analysis.

2. Diminishing voxels: counts the number of voxels that were taken away (or diminished) from the CCSM when the subject was included in the analysis.

3. Relative effect: computes the ratio of the sum of the enhancing and diminishing voxels to the total number of active voxels in the CCSM.

4. Percent overlap: measures the proportion of voxels in the individual subject's FDR thresholded t-map that are present in CCSM. This measure is not based on LOOCSM.

9.5 Modeling variance in combination tests

Random effect model models variance. In a multi-subject group setting, the subject to subject variability is accounted for in the random effect model but in order to make more statistically sound comparison between it and the combination tests, we can modify the p-value combining methods such that we are looking at the absolute difference or proportion of the signals

between two groups in units of standard deviation. Hence each combining techniques will need to be uniquely standardized so that we can incorporate the notion of between-subject variability.

Bibliography

Adcock CH (1960) A note on combining probabilities. *Psychometrika*, **25**, 303–305.

Arndt S, Cizadlo T, Andreasen NC, Heckel D, Gold S, O'Leary DS (1996) Tests for comparing images based on randomization and permutation methods. *Journal of Cerebral Blood Flow Metabolism*, **16**, 1271–1279.

Bahadur RR (1967) Rates of convergence of estimates and test statistics. *Annals Of Mathematical Statistics*, **28**, 303–324.

Bahadur RR (1971) Some Limit Theorems in Statistics *Regional Conference Series in Applied Mathematics* Philadelphia: SIAM.

Bailey BJR (1980) Accurate normalizing transformations of a t-variate. *Applied Statistics*, **29**, 304–306.

Bandettini PA, Wong EC, Hinks RS, Tikofsky RS, Hyde JS (1992) Time course EPI of human brain function during task activation. *Magnetic Resonance in Medicine*, **25**, 390–397.

Beckmann CF, Jenkinson M, Smith SM (2003) Generalized multilevel linear modeling for group analysis in fMRI. *NeuroImage*, **20**, 1052–1063.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, **57**, 289–300.

Birnbaum A (1954) Combining independent tests of significance. *Journal of the American Statistical Association*, **49**, 559-574.

Blair RC and Karniski W (1994) Distribution-Free Statistical Analyses of Surface and Volumetric Maps. *Functional Neuroimaging: Technical Foundations*, San Diego: Academic Press.

Efron B and Tibshirani RJ (1993) An Introduction to the Bootstrap. *Chapman and Hall* CRC.

Brodmann K (1909) Vergleichende Lokalisationlehre der Grosshirnrinde. *Leipzig: J. A. Barth.*

Buckner RL (2003) Functional-anatomic correlates of control processes in memory. *Journal of Neuroscience*, **23**, 3999–4004.

Buxton RB (2002) Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques. *Cambridge University Press*, Cambridge.

Buxton RB, Frank LR (1997) A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *Journal of Cerebral Blood Flow and Metabolism*, **17**, 64–72.

Buxton RB, Wong EC, Frank LR (1998) Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic Resonance in Medicine*, **39**, 855–864.

Clare S, Humberstone M, Hykin J, Blumhardt L, Bowtell R, Morris P (1999) Detecting activations in event-related fMRI using analysis of variance. *Magnetic Resonance in Medicine*, **42**, 1117–1122.

Clark DD, Sokoloff L (1999) Circulation and energy metabolism of the brain. *Basic Neurochemistry: Molecular, cellular and medical aspects* Ed 6 (Siegel GJ, Agranoff BW, Albers RW, Fisher SK, Uhler MD, eds), Philadelphia: Lippincott-Raven.

Cochrane Collaboration (2002) Diversity and Heterogeneity. Web: http://www.cochrane-net.org/openlearning/HTML/mod13-4.htm.

Cohen D (1972) Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer. *Science*, **175**, 664–666.

Cohen J (1988) Statistical power analysis for the behavioral sciences. New York: Academic Press (2nd ed.).

Dagli MS, Ingeholm JE, Haxby JV (1999) Localization of cardiac-induced signal change in fMRI. *NeuroImage*, **9**, 407–415.

Damadian RV (1971) Tumor detection by nuclear magnetic resonance. *Science*, **171**, 1151–1153.

Eddy WF, Fitzgerald M, Noll DC (1996) Improved image registration by using Fourier interpolation. *Magnetic Resononance in Medicine*, **36**, 923–931.

Eddy WF, McNamee RL (2004) Functional magnetic resonance imaging. *Handbook of Computational Statistics* (J.E. Gentle, W. Hardle, Y. Mori, eds.) Springer Verlag, 1002–1027.

Edgington ES (1995) *Randomization Tests.* New York:Marcel Dekker(3rd edition).

Efron B (1982) *The Jackknife, the Bootstrap, and other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Evans AC, Biel C, Marrett S, Thompson CJ, Hakim A (1992) Anatomical-functional correlation using an adjustable MRI-based region of interest atlas with positron emission tomography. *Journal of Cerebral Blood Flow and Metabolism*, **8**, 513–530.

Fagerland MW, Sandvik L (2009) The Wilcoxon-Mann-Whitney test under scrutiny. *Statistics in Medicine*, **28**, 1487–1497.

Fisher RA (1950) *Statistical Methods for Research Workers.* London:Oliver and Boyd (11th edition).

Fox PT, Raichle ME (1986) Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Sciences*, USA, **83**, 1140–1144.

Fox PT, Raichle ME, Mintun MA, Dence C (1988) Nonoxidative glucose consumption during focal physiologic neural activity. *Science*, **241**, 462-464.

Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, Frackowiak RSJ (1995) Statistical parametric maps in functional imaging - A general linear approach. *Human Brain Mapping*, **2**, 189–210.

Friston KJ, Holmes AP, Worsley KJ (1999) How many subjects constitute a study?. *NeuroImage*, **10**, 1–5.

Friston KJ, Jezzard P, Turner R (1994) Analysis of function MRI time series. *Human Brain Mapping*, **1**, 153-171.

Garey LJ (1994) *Brodmanns Localization in the Cerebral Cortex.* London:Smith-Gordon.

Garthwaite PH, Crawford JR (2004) The distribution of the differences of two t-variates. *Biometrika*, **91**, 987–994.

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, **15**, 870–878.

Gibbon J, Fairhurst S (1994) Ratio versus difference comparators in choice. *Journal of Experimental Analysis of Behavior*, **62**, 409–434.

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

Green MV, Seidel J, Stein SD, Tedder TE, Kempner KM, Kertzman C, Zeffiro TA (1994) Head movement in normal subjects during simulated PET brain imaging with and without head restraint. *Journal of Nuclear Medicine*, **35**, 1538–1546.

Gusnard DA, Raichle ME (2001) Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience*, **2**, 685-694.

Haacke EM, Brown RF, Thompson M, Venkatesan R (1999) *Magnetic Resonance Imaging: Physical Principles and Sequence Design.* New York: J. Wiley and Sons.

Hajnal JV, Myers R, Oatridge A, Schwieso JE, Young IR, Bydder GM (1994) Artifacts due to stimulus correlated motion in functional imaging of the brain. *Magnetic Resonance in Medicine*, **31,**, 283–291.

Holmes AP (1994) Statistical Issues in Functional Brain Mapping. Web: http://fil.ion.ucl.ac.uk/spm/papers/APH_thesis.

Holmes AP, Blair RC,watson JDG, Ford I (1996) Nonparametric analysis of statistical images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism*, **16**, 7–22.

Hedges LV (1992) Meta-analysis. *Journal Of Educational Statistics*, **17**, 279-296.

Herman GT (2009) *Fundamentals of Computerized Tomography: Image Reconstruction from Projections.* (2nd ed.), Springer.

Hirschleifer J (1989) Conflict and rent-seeking success functions: Ratio vs. difference models of relative success. *Public Choice*, **63**, 101–112.

Hyder F, Rothman DL, Shulman RG (2002) Total neuroenergetics support localized brain activity: implications for the interpretation of fMRI. *Proceedings of the National Academy of Sciences*, USA, **99**, 10771-10776.

Jezzard P (1999) *Basic Physics of fMRI.*
Web: http://users.fmrib.ox.ac.uk/∼peterj/lectures/hbm_1/index.htm.

Jezzard P (1999) Sources of distortion in functional MRI data. *Human Brain Mapping*, **8**, 80–85.

Jezzard P, Matthews P, Smith S (2001) *Functional MRI: An Introduction to Methods.* Oxford: OUP.

Jonhnson NL, Kotz S, Balakrishnan N (1995) *Continuous Univariate Distributions.* Wiley, Volume 2 (2nd Edition):248.

Jones MC (2006) The logistic and log-F distribution. *Technical Reports*, UK Open University.

Kevles BH (1997) *Naked to the Bone: Medical Imaging in the Twentieth Century.* New Brunswick, NJ: Rutgers UP.

Kiebel SJ, Friston K (2004) Statistical parametric mapping for event-related potentials: One generic consideration. *NeuroImage*, **22**, 492–502.

Kirsten T, Nikolas O, Martin L (2004) Principal neuron spiking: neither necessary nor sufficient for cerebral blood flow in rat cerebellum. *The Journal of Physiology*, **560**, 181–189.

Kwong KK, Belliveau JW, Chesler DA, Goldberg IE, Weiskoff RM, Poncelet BP, Kennedy DN, Hoppel BE, Cohen MS, Turner R, Cheng HM, Brady TJ, Rosen BR (1992) Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, USA, **89**, 5675-5679.

Lage–Castellanos A, Martinez–Montes E, Hernandez–Cabrera JA, Galan Lidice (2010) False discovery rate and permutation test: An evaluation in ERP data analysis. *Statistics in Medicine*, **29**, 63-74.

Lange N, Zeger SL (1996) Non-linear Fourier analysis of magnetic resonance functional neuroimage time series. *Applied Statistics*, **46**, 1-29.

Lauritzen M, Gold L (2003) Brain function and neurophysiological correlates of signals used in functional imaging. *Journal of Neurosciences*, **23**, 3972–3980.

Lazar NA (2008) *The Statistical Analysis of Functional MRI Data.* Springer.

Lazar NA, Eddy WF, Genovese CR, Welling J (1999) Statistical issues in fMRI for brain imaging. *International Statistical Review*, **69**, 105–127.

Lazar NA, Luna B, Sweeney CR, Eddy W (2002) Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, **16**, 538–550.

Lindquist MA (2008) The statistical analysis of fMRI data. *Statistical Science*, **3**, 439–454.

Liu TT, Frank LR, Wong EC, Buxton RB (2001) Detection power, estimation efficiency, and predictability in event-related fMRI. *Neuroimage*, **13**, 759–773.

Magistretti PJ, Pellerin L, Rothman DL, Shulman RG (1999) Energy on demand. *Science*, **283**, 496-497.

Mansfield P, Coxon R, Glover P (1973) Echo-planar imaging of the brain at 3.0T: First normal volunteer results. *Journal of Computer Assisted Tomography*, **18**, 339-507.

McNamee RL, Lazar NA (2004) Assessing the sensitivity of fMRI group maps. *NeuroImage*, **22**, 920-931.

Miller RG (1981) Simultaneous Statistical Inference, 2nd edition New York: Springer-Verlag.

Mintun MA, Lundstrom BN, Snyder AZ, Vlassenko AG, Shulman GL,Raichle ME (2001) Blood flow and oxygen delivery to human brain during functional activity: theoretical modeling and experimental data. *Proceedings of the National Academy of Sciences*, USA, **98**, 6859-6864.

Moonen CTW, Bandetti PA (1999) *Functional MRI.* Springer:Germany.

Mosteller F, Bush R (1954) *Selected Quantitative Techniques.* Lindzey G, editor, Handbook of Social Psychology, volume I: 289-334. Addison-Wesley, Cambridge, Mass.

Mudholkar GS, George EO (1979) The logit method for combining probabilities. *Symposium On Optimizing Methods In Statistics* (J. Rustagi, ed.), 345–366. New York: Academic Press.

Nichols TE, Holmes AP (2001) Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, **15**, 1–25.

Nichols MJ, Newsome WT (1999) The neurobiology of cognition. *Nature*, **402**, C35-C38.

Leibovici DG, Smith S, Radcliffe J (2001) Comparing groups of subjects in fMRI studies: A review of GLM approach. *FMRIB Technical Report*, TR00DL1.

Ogawa S, Lee TM, Kay AR, Tank DW (1990a) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, USA, **87**, 9868–9872.

Ogawa S, Lee TM, Naycik AS, Glynn P (1990b) Oxygenation-sensitive contrast in magnetic resonance imaging of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, **16**, 9-18.

Ogawa S, Tank DW, Menon R, Ellermann JM, Kim S-G, Merkle H, Ugurbil K (1992) Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, USA, **89**, 5951-5955.

Oransky I (2007) Paul C Lauterbur. *The Lancet*, **369**, 1686.

Pollard KS, van der Laan MJ (2003) Resampling-based multiple testing: Asymptotic control of Type I error and applications to gene expression data. *U.C. Berkeley Division of Biostatistics working paper series*, **Paper 121**.

Pollmann S, Dove A, Yves von Cramon D, Wiggins CJ (2000) Event-related fMRI: Comparison of conditions with varying BOLD overlap. *Human Brain Mapping*, **9**, 26–37.

Press SJ (1969) The t-ratio distribution. *Journal of American Statistical Association*, **64**, 242-252.

Quenouille MH (1949) Approximate tests of correlation in time series. *Journal of Royal Statistical Society*, **11**, 68-84.

Raichle ME (1998) Behind the scenes of functional brain imaging: A historical and physiological perspective. *Proceedings of the National Academy of Sciences*, USA, **95**, 765-772.

Raichle ME (2000) A brief history of human functional brain mapping. *Brain Mapping: The Systems* (Toga AW, Mazziotta JC, eds),San Diego: Academic Press.

Raichle ME, Gusnard DA (2002) Appraising the brains energy budget. *Proceedings of the National Academy of Sciences*, USA, **99**, 10237-10239.

Raichle ME (2003) Functional brain imaging and human brain function. *The Journal of Neuroscience*, **23**, 3959–3962.

Rajapakse JC, Kruggel F, Cramon DYV (1998) Modeling hemodynamic response for analysis of functional MRI time-series. *Human Brain Mapping*, **6**, 283-300.

Rao CR, Kleffe J (1988) *Estimation of Variance Components and Applications.* (2nd edition) New York: John Wiley and Sons.

Reimold M, Slifstein M, Heinz A, Mller-Schauenburg W, Bares R (2006) Effect of spatial smoothing on t-maps: Arguments for going back from t-maps to masked contrast images. *Journal of Cerebral Blood Flow and Metabolism*, **26**, 751-759.

Robb RA (1999) *Biomedical Imaging, Visualization, and Analysis.* John Wiley and Sons, Inc.

Rosenthal R (1978) Combining results of independent studies. *Psychological Bulletin*, **85**, 185–193.

Rothstein HR, Sutton AJ, Borenstein M (2005) *Publication Bias in Meta-Analysis Prevention, Assessment and Adjustments.* John Wiley and Sons, Ltd.

Schwartz WJ, Smith CB, Davidsen L, Savaki H, Sokoloff L, Mata M, Fink DJ, Gainer H (1979) Metabolic mapping of functional activity in the hypothalamo-neurohypophysial system of the rat. *Science*, **205**, 723-725.

Sharp FR (1976) Relative cerebral glucose uptake of neuronal perikarya and neuropil determined with 2-deoxyglucose in resting and swimming rat. *Brain Research*, **110**, 127-139.

Sharp FR, Kauer JS, Shepherd GM (1977) Laminar analysis of 2-deoxyglucose uptake in olfactory bulb and olfactory cortex of rabbit and rat. *Journal of Neurophysiology*, **40**, 800-813.

Shulman RG, Hyder F, Rothman DL (2001) Cerebral energetics and the glycogen shunt: Neurochemical basis of functional imaging. *Proceedings of the National Academy of Sciences*, USA, **98**, 6417-6422.

Siegel S(1956) *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, Inc., Tokyo, Japan.

Singer JR (1959) Blood flow rates by nuclear magnetic resonance measurements. *Science*, **130**, 1652–1653.

Smith AJ, Blumenfeld H, Behar KL, Rothman DL, Shulman RG, Hyder F (2002) Cerebral energetics and spiking rate: the neurophysiological basis of fMRI. *Proceedings of the National Academy of Sciences*, USA, **99**, 10765-10770.

Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RM (1949) *Adjustment During Army Life* The American Soldier: Vol.I., Princeton: Princeton University Press.

Stuart Clare (1997) The analysis of fMRI data.
Web: http://users.fmrib.ox.ac.uk/∼stuart/thesis/chapter_6/contents.html.

Swartz BE, Goldensohn ES (1998) Timeline of the history of EEG and associated fields. *Electroencephalography and Clinical Neurophysiology*, **106**, 1173-176.

Udupa JK, Herman GT (2000) *3D Imaging in Medicine.* 2nd Edition, CRC Press.

Wallis C (2009) *fMRI.* Web: http://www.csulb.edu/∼cwallis/482/fmri/fmri.html.

Webb S (1990) *From the Watching Of Shadows.* New York: Adam Hilger.

Westfall PH, Young SS (1993) *Resampling based Multiple Testing: Examples and Methods for p-value Adjustments.* New York: Wiley.

Weisskoff RM (1996) Simple measurement of scanner stability for functional NMR imaging of activation in the brain. *Magnetic Resonance in Medicine*, **36**, 643–645.

Woolrich M, Ripley BD, Brady JM, and Smith S (2000) Temporal autocorrelation in univariate linear modelling of fMRI Data. *Human Brain Mapping*, Abstract:S610

Worsley KJ, Friston KJ (2000) A test for a conjunction. *Statistics and Probability Letters*, **47**, 135–140.

Worsley KJ, Liao C, Grabove M, Petre V, Ha B, and Evans AC (2000) General statistical analysis for fMRI data. *Human Brain Mapping*, Abstract:S648.

Yandell BS (1997) *Practical Data Analysis for Designed Experiments.* Chapman and Hall: London (first edition).

Zar JH (1996) *Biostatistical Analysis.* Prentice-Hall.

Zonta M, Angulo MC, Gobbo S, Rosengarten B, Hossmann K-A, Pozzan T, Carmignoto G (2003) Neurontoastrocyte signaling is central to the dynamic control of brain micro circulation. *Nature Neuroscience*, **6**, 43-50.